Learning and Control of Dynamical Systems

Thesis by

Ali Sahin Lale

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2023
Defended May 3rd, 2023

Ali Sahin Lale
ORCID: 0000-0002-7191-346X

Anneme ve Babama.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my heartfelt gratitude to my amazing advisors Prof. Anima Anandkumar and Prof. Babak Hassibi. I am forever indebted to both of you for giving me the opportunity to enjoy doing research at Caltech.

I cannot thank Anima enough for the unwavering support she has provided me since day one in the lab. Her guidance and belief in my abilities have given me the confidence to pursue research directions that truly ignite my passion. Throughout every stage of my PhD journey, her constant support and emphasis on seeing the bigger picture in research problems have been invaluable. Her guidance not only shaped this thesis but also transformed me into a better researcher. Thanks to her persistent encouragement, I was able to translate my research from theoretical concepts on pen and paper to impactful real-world applications, forever altering my perspective on research problems. I am also deeply grateful to Babak, whose exceptional ability to tackle every technical problem still amazes me to this day. His uncompromising rigor and meticulous attention to detail have profoundly influenced the way I approach research problems.

I extend my sincere gratitude to the members of my defense and candidacy committees: Prof. P.P. Vaidyanathan, Prof. Yaser Abu-Mostafa, and Prof. Adam Wierman. I truly appreciate your support, valuable feedback, and guidance on my research and future plans. I would like to extend special thanks to Adam for his accessibility and support throughout every stage of my PhD journey.

One of the greatest aspects of being part of two research groups is the privilege to have twice as many group members to learn from and be inspired by. From Anima's group, I extend my heartfelt gratitude to Kamyar Azizzadenesheli, Angie Liu, Yuanyuan Shi, Jiawei Zhao, Hongkai Zheng, Zongyi Li, Guanzhi Wang, Rafal Kocielnik, Or Sharir, and Kaiyu Yang for their encouragement and invaluable feedback. The paper reading sessions we shared before deadlines will always hold a special place in my memory. From Babak's group, I would like to thank Hikmet Yildiz, Oron Sabag, Ehsan Abbasi, Fariborz Salehi, Navid Azizan, Ahmed Douik, and Taylan Kargin for the countless thought-provoking discussions and the extended Friday group meetings. The friendly and intellectually stimulating atmosphere of our lab has truly been one of the most fulfilling experiences during my time at Caltech.

I am deeply indebted to my collaborators, Kamyar Azizzadenesheli, Adam Wierman,

Mory Gharib, Yuanyuan Shi, Guannan Qu, Navid Azizan, Oron Sabag, Oguzhan Teke, Taylan Kargin, Peter Renn, and Nithin Varma, for their patience and for all the things I learned from them. I will forever cherish the fun and long hours that we spent thinking and learning together. Some of the contributions and ideas of this thesis developed during long whiteboard discussions with Kamyar, and I am grateful to him for introducing me to the world of reinforcement learning. Although our collaborative work with Oron is not directly incorporated into this thesis, I wish to extend my thanks to him for his inspiring productivity and for being a supportive friend.

During my time at Caltech, I have had the privilege of forming incredible bonds with special friends: Hikmet Yildiz, Oguzhan Teke, Halime Teke, Utkan Candogan, Corina Panda, Sinan Kefeli, Zeynep Turan, Hoang Le, Kordag Kilic, Recep Can Yavas, Taylan Kargin, and Helen Ha. I cannot fathom how this thesis would have been accomplished without their unconditional support. They have made me feel at home and loved, standing by me through both the best and the most challenging times. I am immensely grateful for the adventures, trips, and cherished memories we have shared, and for becoming my chosen family away from home. I want to extend a special thank you to my roommate of seven years, Hikmet, for his patience and brotherhood. I also want to express my gratitude to Oguzhan and Utkan, who have embraced my quirks from day one, and to Helen, who has made the past three years truly wonderful. Thank you all for making my PhD journey an unforgettable one.

I would like to extend my heartfelt gratitude to my dear friends back home: Hakan Akar, Cihan Ceyhan, Can Gursoy, Mert Degirmenci, Alper Tastemur, Feyzullah Alim Kalyoncu. Throughout these years, they have consistently provided emotional support, from the games we played together to the profound conversations we shared, despite the challenges of different time zones. Their presence has kept me motivated and encouraged me to remain true to myself. I genuinely appreciate the strong bond of brotherhood we share. Although we are now scattered across the world, I find solace in knowing that we are friends for life, always ready to stand by each other's side.

At the core of every achievement I have made, I owe an immeasurable debt of gratitude to my family. They say it takes a village to raise a child, and in my case, this sentiment rings true. I consider myself incredibly fortunate to be surrounded by a large and unwaveringly supportive family. To my beloved uncles and aunts, I express my heartfelt appreciation for treating me as your own child. To my cousins, you have been like siblings to me, providing companionship and support that I will forever treasure. To my beloved grandparents, I am eternally grateful for your

unconditional love, support, and the countless ways you have cared for me since the day I was born. You have been a constant source of strength and inspiration.

Lastly, but most importantly, I extend my deepest gratitude to my mother Semra, and my father Niyazi. Thank you for placing my personal and academic growth above all else, and for sacrificing so much for my success. Your unwavering support has been the driving force behind my ambition and motivation since my early years in primary school. Words cannot adequately convey the depth of my gratitude for your unconditional love, the countless precious memories we have shared, and the values you have instilled in me. It is with immense gratitude and love that I dedicate this thesis to both of you.

ABSTRACT

Despite the remarkable success of machine learning in various domains in recent years, our understanding of its fundamental limitations remains incomplete. This knowledge gap poses a grand challenge when deploying machine learning methods in critical decision-making tasks, where incorrect decisions can have catastrophic consequences. To effectively utilize these learning-based methods in such contexts, it is crucial to explicitly characterize their performance. Over the years, significant research efforts have been dedicated to learning and control of dynamical systems where the underlying dynamics are unknown or only partially known a priori, and must be inferred from collected data. However, much of these classical results have focused on asymptotic guarantees, providing limited insights into the amount of data required to achieve desired control performance while satisfying operational constraints such as safety and stability, especially in the presence of statistical noise.

In this thesis, we study the statistical complexity of learning and control of unknown dynamical systems. By utilizing recent advances in statistical learning theory, high-dimensional statistics, and control theoretic tools, we aim to establish a fundamental understanding of the number of samples required to achieve desired (i) accuracy in learning the unknown dynamics, (ii) performance in the control of the underlying system, and (iii) satisfaction of the operational constraints such as safety and stability. We provide finite-sample guarantees for these objectives and propose efficient learning and control algorithms that achieve the desired performance at these statistical limits in various dynamical systems. Our investigation covers a broad range of dynamical systems, starting from fully observable linear dynamical systems to partially observable linear dynamical systems, and ultimately, nonlinear systems.

We deploy our learning and control algorithms in various adaptive control tasks in real-world control systems and demonstrate their strong empirical performance along with their learning, robustness, and stability guarantees. In particular, we implement one of our proposed methods, Fourier Adaptive Learning and Control (FALCON), on an experimental aerodynamic testbed under extreme turbulent flow dynamics in a wind tunnel. The results show that FALCON achieves state-of-the-art stabilization performance and consistently outperforms conventional and other learning-based methods by at least 37%, despite using 8 times less data. The superior performance of FALCON arises from its physically and theoretically accurate modeling of the underlying nonlinear turbulent dynamics, which yields rigorous

finite-sample learning and performance guarantees. These findings underscore the importance of characterizing the statistical complexity of learning and control of unknown dynamical systems.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] Taylan Kargin*, Sahin Lale*, Kamyar Azizzadenesheli, Animashree Anandkumar, and Babak Hassibi. Thompson sampling achieves $\tilde{O}(\sqrt{T})$ regret in linear quadratic control. In *Conference on Learning Theory*, pages 3235–3284. PMLR, 2022.
S.L. is the co-first author of this paper. He participated in the conception of the project, the analysis of the problem, and the writing of the manuscript.

[2] Taylan Kargin, Sahin Lale, Kamyar Azizzadenesheli, Animashree Anandkumar, and Babak Hassibi. Thompson sampling for partially observable linear-quadratic control. *2023 American Control Conference (ACC)*, 2023.
S.L. participated in the conception of the project, the analysis of the problem, and the writing of the manuscript.

[3] Sahin Lale, Kamyar Azizzadenesheli, Anima Anandkumar, and Babak Hassibi. Stochastic linear bandits with hidden low rank structure. *arXiv preprint arXiv:1901.09490*, 2019.
S.L. participated in the conception of the project, performed the analysis, and participated in the writing of the manuscript.

[4] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. *Advances in Neural Information Processing Systems*, 33:20876–20888, 2020.
S.L. participated in the conception of the project, performed the analysis, and participated in the writing of the manuscript.

[5] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Regret minimization in partially observable linear quadratic control. *arXiv preprint arXiv:2002.00082*, 2020.
S.L. participated in the conception of the project, performed the analysis, and participated in the writing of the manuscript.

[6] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Adaptive control and regret minimization in linear quadratic gaussian (lqg) setting. In *2021 American Control Conference (ACC)*, pages 2517–2522. IEEE, 2021.
S.L. participated in the conception of the project, performed the analysis, and participated in the writing of the manuscript.

[7] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Finite-time system identification and adaptive control in autoregressive exogenous systems. In *Learning for Dynamics and Control*, pages 967–979. PMLR, 2021.
S.L. participated in the conception of the project, performed the analysis, and participated in the writing of the manuscript.

[8] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Model learning predictive control in nonlinear dynamical systems. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 757–762. IEEE, 2021.
S.L. participated in the conception of the project, performed the analysis, and participated in the writing of the manuscript.

[9] Sahin Lale*, Oguzhan Teke*, Babak Hassibi, and Anima Anandkumar. Stability and identification of random asynchronous linear time-invariant systems. In *Learning for Dynamics and Control*, pages 651–663. PMLR, 2021.
S.L. is the co-first author of this paper. He participated in the conception of the project, the analysis of the problem, and the writing of the manuscript.

[10] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Animashree Anandkumar. Reinforcement learning with fast stabilization in linear dynamical systems. In *International Conference on Artificial Intelligence and Statistics*, pages 5354–5390. PMLR, 2022.
S.L. participated in the conception of the project, performed the analysis, and participated in the writing of the manuscript.

[11] Sahin Lale, Yuanyuan Shi, Guannan Qu, Kamyar Azizzadenesheli, Adam Wierman, and Anima Anandkumar. Kcrl: Krasovskii-constrained reinforcement learning with guaranteed stability in nonlinear dynamical systems. *arXiv preprint arXiv:2206.01704*, 2022.
S.L. participated in the conception of the project, performed the analysis, and participated in the writing of the manuscript.

[12] Sahin Lale, Peter I. Renn, Kamyar Azizzadenesheli, Babak Hassibi, Morteza Gharib, and Anima Anandkumar. Falcon: Fourier adaptive learning and control for disturbance rejection under extreme turbulence. *arXiv preprint*, 2023. *Spotlight Presentation at Neurips 2022 AI for Science Workshop*
S.L. participated in the conception of the project, performed the analysis, and participated in the writing of the manuscript.

[13] Guannan Qu*, Yuanyuan Shi*, Sahin Lale*, Anima Anandkumar, and Adam Wierman. Stable online control of linear time-varying systems. In *Learning for Dynamics and Control*, pages 742–753. PMLR, 2021.
S.L. is the co-first author of this paper. He participated in the conception of the project, the analysis of the problem, and the writing of the manuscript.

[14] K Nithin Varma, Sahin Lale, and Anima Anandkumar. Stochastic linear bandits with unknown safety constraints and local feedback. *arXiv preprint*, 2023.
S.L. participated in the conception of the project, the analysis of the problem, and the writing of the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

Controlling dynamical systems involves designing a control policy or set of actions to govern the behavior of the system in a desired manner. Typically, this problem has three components: a dynamical system/environment, a controlling agent, and a control objective. When the dynamics of the system are well-understood and can be modeled mathematically, the problem of controlling the system can be formulated as an optimization problem. Optimal control theory has a long history of applications and success in solving these optimization problems [19, 28, 105].

However, in most real-world applications, the underlying dynamics are either unknown or only partially known a priori, making the problem of controlling the system significantly more challenging. In this setting, learning and control are intertwined processes, where the agent uses feedback from the environment to update its knowledge of the dynamics and adjust its actions accordingly. To optimize the overall control objective, the controlling agent must "explore" the environment to gain a better understanding of the system dynamics, which is often called system identification in control theory. The agent then uses this understanding to design a set of controllers that simultaneously reduce the possible future costs, i.e., "exploit", and also enable the agent to explore the important and unknown aspects of the system. This process is often referred to as adaptive control design.

This procedure captures the fundamental trade-off in decision-making under uncertainty tasks: exploration vs. exploitation. In recent decades, this challenging problem has been extensively studied and resulted in a set of foundational steps to study the consistency of the model estimates and asymptotic convergence to optimal controllers. In particular, in system identification of dynamical systems, the primary focus is on the asymptotic recovery guarantees of the underlying system [89, 126, 130] or the practical aspects of the proposed methods [53, 282, 296]. Similarly, classical works in adaptive control provide asymptotic performance guarantees of the designed controllers, assuming that the system is perfectly recovered asymptotically [88, 152, 157, 158], or they design new practical methods [145, 154, 221].

While asymptotic analyses set the ground for the design of optimal control, understanding the finite-sample behavior of adaptive control algorithms is critical for real-

world applications. This is particularly important in safety-critical decision-making tasks where incorrect actions can have catastrophic consequences. Moreover, the emergence of autonomous and data-driven agents in challenging interactive tasks, such as self-driving vehicles and agile robotic systems, requires the joint study of system identification and adaptive control termed as learning and control of dynamical systems [227]. Recent developments in the fields of statistics and machine learning along with control theory [158, 218, 280] empowers us to not only advance the study of the asymptotic efficiency of learning and control algorithms but also to analyze their finite-sample behavior [2, 88].

In this thesis, we study the statistical complexity of learning and control of dynamical systems. Combining the modern advancements in statistics, machine learning, and reinforcement learning, with well-established control theoretic ideas, we take steps toward developing theoretical foundations on finite-sample behavior of learning and control algorithms. Throughout the thesis, we aim to shed light on several fundamental research questions on learning and control of dynamical systems.

1.1 Research Questions Studied in This Thesis

"What is the required amount of data for learning and control algorithms to achieve the desired accuracy in learning the unknown dynamics?"

This question focuses on the system identification aspect of learning and control and seeks sample complexity guarantees for learning the unknown dynamics. In this thesis, we study the problem of learning the underlying system dynamics from a single trajectory in various dynamical systems. In linear dynamical systems, we investigate the learning of model parameters that govern the system dynamics.

Addressing this question involves careful considerations due to the inherent dynamics of the systems. The data collected from dynamical systems are not identically and independently distributed (i.i.d.) but instead highly correlated, as observations at any given time may carry the effects of prior observations and inputs. These correlations pose a challenge for system identification, even in open-loop scenarios where control inputs are chosen i.i.d. These difficulties are further amplified in closed-loop scenarios where a feedback controller designs inputs based on previous observations. The conventional statistical learning theory tools developed for i.i.d. data are not suitable for handling these correlations, requiring novel approaches to address these issues.

Learning the dynamics in nonlinear dynamical systems is even more challenging since local behavior in some parts of the state space does not determine global behavior, unlike in linear systems. To address this, one can select suitable nonlinear basis functions for learning the system dynamics. While this approach is feasible, it requires new tools for its approximation theoretic analysis to provide finite sample learning guarantees.

Despite these statistical and structural challenges, on a positive note, the dynamical system construction can be leveraged to learn underlying dynamics. System identification methods could exploit control-theoretic concepts such as the controllability or stability of the system to simplify the learning task.

"What is the required amount of data for learning and control algorithms to achieve the desired performance in the control of an unknown dynamical system?"

This research question focuses on the adaptive control aspect of learning and control and seeks finite-time performance guarantees in the control of unknown dynamical systems. In this thesis, we study the problem of regret minimization as the performance metric for learning and control algorithms. Regret measures the performance of the learning agent as the difference between the cumulative cost encountered by the learning agent and that of a baseline policy, such as an optimal controller that knows the system dynamics. Thus, it measures the sub-optimality gap in the control performance due to the lack of knowledge of system dynamics.

In the regret minimization framework, the desirable behavior of the learning agent is to achieve sublinear regret, i.e., a regret of $o(T)$ after T time steps. This scaling shows that the performance of the learning agent approaches that of the optimal controller as more data is collected, indicating learning to control behavior. However, achieving this requires a careful balance of the exploration vs. exploitation trade-off. Too much exploration can result in linear regret due to not being able to exploit the gathered information toward optimizing the control objective. In contrast, too little exploration can lead to linear regret due to being stuck at sub-optimal policies.

To address this fundamental problem, various methodologies to balance exploration and exploitation have been proposed and studied in reinforcement learning literature, as shown in Figure 1.1. One promising methodology is to use the *optimism in the face of uncertainty* (OFU) principle [156]. OFU-based methods estimate the environment

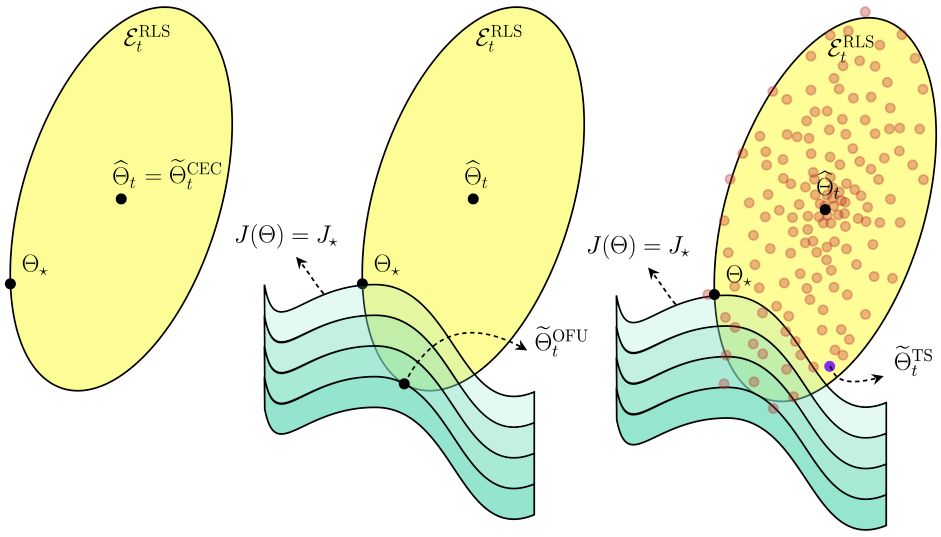

Figure 1.1: Methodologies to balance the exploration vs. exploitation trade-off: Certainty Equivalence (CE), Optimism in the Face of Uncertainty (OFU), Thompson Sampling (TS).

model up to a confidence interval and construct a set of plausible models within this interval. Among those models, they choose the one with the lowest expected cost, i.e., the optimistic one, and follow the optimal behavior suggested by the selected model. Therefore, these methods are also referred to as the upper-confidence bound algorithms. The intuition behind these methods is that if the chosen model is in fact a low-cost achieving model, then we minimize the regret and achieve desirable exploitation. On the other hand, if it is not, then we reduce the uncertainty around these models, providing desirable exploration.

Another prominent methodology is Thompson Sampling (TS) [263]. In TS, the agent samples a model from a distribution computed based on prior actions and observations, and then takes the optimal action for this sampled model and updates the distribution based on its new observation. Similar to OFU, it balances the exploration and exploitation by refining the uncertainties in the estimates, thus skewing the sampling distribution toward low-cost achieving models.

Moreover, several control theoretic concepts unique to dynamical systems, such as stability and stabilizability, also impact the regret of learning and control in dynamical systems. Understanding the roles of these concepts and problem-dependent constants, such as dimension dependencies, is crucial for deriving finite-time performance guarantees of learning and control in dynamical systems.

"What is the required amount of data for learning and control algorithms to satisfy operational constraints such as stabilization of the underlying dynamics?"

This research question focuses on the finite-time stabilization and safety aspects in the learning and control of dynamical systems. While learning-based control methods show promise for improved control performance, there are significant challenges that must be addressed before they can be deployed in real-world systems. Safety-critical systems in particular have high standards for stability, and learning and control algorithms must meet operational constraints while also achieving good performance guarantees.

Achieving stability in an unknown dynamical system is one of the fundamental challenges in control systems engineering. This involves finding a stabilizing controller with minimal interactions with the system to perform tasks like maintaining a dynamical system around a desired equilibrium point or tracking a reference signal. Stability is crucial not only for minimizing regret, as discussed earlier, but also for avoiding saturation and maintaining the validity of linearization around a certain point for nonlinear dynamical systems [258]. Therefore, stabilizing the system dynamics is often the first step toward solving more complex control tasks, and learning and control algorithms need to quickly stabilize the underlying system dynamics. In this thesis, we investigate the sample complexity of designing stabilizing controllers for the underlying dynamical systems, either as the primary goal or as an intermediate step toward regret minimization.

"Can we design computationally efficient learning and control algorithms that achieve these desired finite-time performances in unknown dynamical systems?"

In this thesis, besides obtaining statistically efficient learning and control in dynamical systems, i.e., state-of-the-art finite-sample guarantees for the performance metrics above, we aim to achieve these results via computationally efficient methods. To address this research question, we either provide computationally efficient learning and control algorithms for various dynamical systems or propose and discuss ways to improve possible computational inefficiencies to obtain effective implementations.

"Can we design learning and control algorithms that not only have strong theoretical guarantees but also perform well empirically in real-world systems?"

The ultimate goal of studying the finite-sample guarantees in the learning and control of dynamical systems is to design and deploy robust and high-performing control systems in practice. While statistical learning guarantees are essential for robustness purposes, they may not always guarantee practical performance. Complex algorithms that achieve strong statistical guarantees may not perform well in practice due to their added complexity to handle all possible situations. Therefore, it is important to design simpler algorithms with strong empirical performance. In real-world applications, besides statistical efficiency and robustness, other practical aspects such as latency, scalability, interpretability, and implementability are also crucial. These aspects may be hard to quantify statistically, so statistically efficient learning and control algorithms should aim to achieve efficiency in the simplest way possible to allow for real-world implementation and strong empirical performance.

1.2 Outline and Scope of the Thesis

This thesis studies the learning and control of various dynamical systems ranging from the most basic stateless dynamical systems of stochastic linear bandits to the most complicated partially observable nonlinear dynamical systems. In our investigation, we follow the guideline of the fundamental research questions posed in the previous section and try to answer and shed light on them in each dynamical system setting. Our central goal is to provide finite-time performance characterization of learning and control of dynamical systems and propose efficient algorithms that achieve these fundamental limits. The chapters of this thesis are organized in order of increasing complexity of learning and control tasks. In particular, as the initial setting, we study stochastic linear bandits in Chapter 2. In Chapter 3, we consider learning and control of canonical linear time-invariant dynamical systems, namely, linear quadratic regulators. Then, we focus on linear time-varying systems in Chapter 4. Chapter 5 studies the measurement feedback setting of linear dynamical systems, where the dynamical system evolves with respect to a latent state and the decision-making agent only observes noisy linear measurements of this state. Finally, Chapter 6 considers the most general setting of learning and control in partially observable nonlinear dynamical systems. We end the thesis with some interesting future research directions in Chapter 7.The following briefly outlines the scope of each chapter.

Stochastic Linear Bandits with Practical Concerns (Chapter 2)

We start our study with stochastic linear bandits (SLB). The standard formulation of SLB is the simplest sequential decision-making under uncertainty problem and is often regarded as a "stateless" dynamical system since the feedback/reward of each action is independent and does not impact the overall condition of the environment. In this setting, at each round of SLB, an agent chooses an action x_t from the given set of actions D_t and receives a stochastic reward from the environment,

$$r_t = x_t^\top \theta_* + \eta_t,$$

whose expected value is an *unknown* linear function θ_* of the d-dimensional action representation vector $x_t \in D_t$, i.e., η_t is a random variable with zero mean. The agent's goal is to maximize its cumulative reward. Clearly, with the knowledge of θ_*, the optimal strategy is to choose $\text{argmax}_{x \in D_t} x^\top \theta_*$. However, due to the lack of knowledge of the true environment model, the agent makes mistakes by picking sub-optimal actions. Thus, the goal in SLB is to design a strategy for the agent to minimize the cumulative cost of these mistakes, known as *regret*. For this task, the agent needs to dedicate its actions to not only maximize the immediate reward but also to explore other actions to build a better estimate of the unknown linear function and guarantee higher future rewards. Therefore, despite their simplicity, stochastic linear bandits fundamentally capture the trade-off between *exploration* and *exploitation*, which is the crux of sequential decision-making under uncertainty problems. For this very reason, they have been widely adopted as modeling tools to improve decision making in clinical trials, ad display optimization, energy management, and recommendation systems.

The fundamental understanding and finite-time performance (regret) guarantees have been developed in the canonical setting described above using different policy design strategies such as the OFU principle or Thompson Sampling. However, real-world decision-making tasks often involve high-dimensional action representations and unknown safety constraints, which pose additional challenges that are not addressed by the standard algorithms. This chapter addresses these practical aspects of modern-day decision-making under uncertainty tasks in the context of stochastic linear bandits: (i) high-dimensional feature representations for the action vectors and (ii) unknown (nonlinear) safety constraints.

For the first aspect, we propose an algorithm called **P**rojected **S**tochastic **L**inear **B**andits (PSLB) that leverages principal component analysis-based projection to ef-

ficiently recover the low-dimensional subspace structure of high-dimensional action representations. PSLB exploits the underlying hidden structure to guide exploration and exploitation, which yields improved performance compared to existing algorithms. We prove that PSLB obtains a regret upper bound that scales with the intrinsic dimension of the subspace, rather than the large ambient dimension of the action space. Empirical studies on image classification tasks show that PSLB significantly reduces regret and converges faster to an accurate model compared to state-of-the-art bandit algorithms.

For the second aspect, we study stochastic linear bandits with unknown safety constraints and local feedback. We propose optimism (upper confidence bound) and Thompson Sampling-based algorithms that carefully incorporate an additional exploration incentive to ensure the selection of high-reward actions that are also safe and encourage exploration in the relevant constraint sets to recover the optimal safe action. We provide tight regret bounds for these algorithms, showing that they achieve optimal sublinear regret without any safety violations. Empirical studies on various safety constraints and a real-world credit dataset demonstrate the effectiveness of the proposed algorithms in safely exploring and recovering optimal safe actions quickly.

Learning and Control in Linear Quadratic Regulator (Chapter 3)

With this chapter, we begin our study of learning and control of dynamical systems with an underlying state that evolves over time based on the executed actions. We consider the canonical setting of learning and control in dynamical systems: fully observable (state-feedback) linear time-invariant systems

$$x_{t+1} = A_* x_t + B_* u_t + w_t, \tag{1.1}$$

where x_t is the state of the system, u_t is the control input, w_t is the stochastic process noise at time t, and A_*, B_* are the model parameters. In particular, we study the problem of adaptive control of the linear time-invariant systems given in (1.1) with quadratic regulatory cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$, commonly referred to as linear-quadratic regulators (LQRs) [134], with *unknown parameters A_* and B_*, and without a priori known stabilizing controller*. This framework has been the main focus of providing finite-time guarantees in learning and control of dynamical systems due to its simplicity and ability to capture the crux of the problem in terms of system identification and adaptive control synthesis [2, 7, 71, 85, 137, 166, 191, 242]. When

the model dynamics are known, the LQR control problem becomes an optimization problem with a scalar cost objective, whose optimal solution is given as a linear state-feedback controller [134]. However, when the model is *unknown*, the learning agent needs to learn the dynamics and find the optimal control policy online, i.e., online LQR, which is a core challenge in reinforcement learning and control theory.

The ultimate goal of the online LQR problem is to design RL agents that can adapt autonomously to unknown environments with minimal information while ensuring finite-time stability and performance guarantees. Despite significant research interest, there are few approaches that provide a complete treatment of the problem without initial model estimates or stabilizing controllers, and the existing methods that learn from scratch often suffer from exponential regret and unstable dynamics, limiting their practical deployment.

To address these challenges, we propose two learning and control algorithms for online LQRs. The first algorithm, **Stab**ilizing **L**earning (StabL), incorporates the optimism principle into the online LQ control problem. StabL achieves fast stabilization of the system by effectively exploring the environment with an improved exploration strategy, resulting in $\mathcal{O}(\sqrt{T})$ regret after T time steps, which is optimal for online LQR. We also show that the regret of StabL has only a polynomial dependence on the problem dimensions, which is an exponential improvement over prior methods. The key ingredient that allows these results is our exploration strategy that combines the sophisticated exploration approach of optimism with isotropic exploration to achieve fast system identification and stabilization, hence, optimal regret. We also demonstrate that StabL outperforms prior algorithms empirically in various adaptive control tasks.

The second algorithm, **T**hompson **S**ampling-based **A**daptive **C**ontrol (TSAC), addresses possible computational inefficiencies of StabL due to optimism by using Thompson Sampling (TS) to balance exploration and exploitation trade-off in controller design. Despite the computational efficiency of TS, prior works in online LQR were able to attain optimal regret only for scalar systems, whose extension to multi-dimensional LQR systems has been proposed as an open problem in [6]. We design TSAC according to the algorithmic insights obtained from StabL, and show that it achieves the optimal $\mathcal{O}(\sqrt{T})$ regret even for multidimensional systems, thereby solving the open problem posed in prior work. Similar to StabL, TSAC does not require a known stabilizing controller and achieves fast stabilization through effective exploration in the early stages. Our breakthrough in TSAC lies in developing

a novel lower bound on the probability of obtaining an optimistic sample with TS. By carefully prescribing an early exploration strategy and policy update rule, we show that TSAC achieves optimal regret in adaptive control of multidimensional stabilizable LQRs. Our empirical results demonstrate the effective performance and efficiency of TSAC in adaptive control of Boeing 747 with linearized dynamics.

These results make StabL and TSAC the first algorithms that achieve optimal regret in stabilizable LQRs without an initial stabilizing policy. This feat highlights the benefit of early improved exploration to achieve fast stabilization and reduce the cumulative regret at the expense of a slight increase in regret in the early stages. Moreover, our optimal regret guarantee on TSAC also shows that a simple sampling strategy based on confidence sets provides effective exploration to recover low-cost and eventually optimal controllers in adaptive control of LQRs.

Learning and Control in Linear Time-Varying Systems (Chapter 4)

Time-invariant systems such as LQRs considered in Chapter 3 have traditionally been the main focus in the learning and control perspective by the reinforcement learning and control communities. However, real-world dynamical systems are often time-varying, e.g., power systems, autonomous vehicles, and financial markets. While not all time-varying systems have linear dynamics, many applications with nonlinear dynamics can be approximated by linear time-varying (LTV) systems via a local linear approximation at each time step. In this chapter, we study the learning and control of LTV systems through the lenses of stability and online stabilization.

There are several notions of stability considered in the study of stability in LTV systems. Of all these notions, we focus on the input-to-state stability (ISS) and mean-square stability of LTV systems in this chapter. ISS aims to guarantee the boundedness of the state given bounded initial conditions, which is crucial for many applications of LTV systems to avoid saturation, and maintain the robustness, and validity of linearization around a certain point. Mean-square stability, on the other hand, is crucial in stochastic systems and implies that the system converges to its fixed point asymptotically in a mean-square sense.

While there is considerable prior work focused on stabilizing the LTV systems, most of these works study stability in the offline setting, where the sequence of system parameters is known or has a particular variation pattern. Maintaining stability guarantees becomes significantly harder in the online setting, where the system parameters are observed in real-time and may have arbitrary variations,

which is most relevant to many real-world applications. Moreover, when designing a controller, the designer must not only stabilize the dynamics but also aim to have a low cost, as controllers that only focus on stability may result in sub-optimal cost. Yet the converse may also be true, as controllers that only focus on minimizing the regulatory cost myopically may result in unstable dynamics.

In this chapter, we aim to answer the question of whether it is possible for an online controller to guarantee stability and maintain low costs in LTV systems. We propose an efficient online control algorithm, **Co**variance **C**onstrained **O**nline **L**inear **Q**uadratic (COCO-LQ) control, that guarantees ISS for a large class of LTV systems while also minimizing the control cost. COCO-LQ incorporates a novel state covariance constraint into the semidefinite programming (SDP) formulation of the optimal LQ control problem. We show that this constraint promotes a joint stabilization property for the sequence of controllers designed by COCO-LQ, which gives the desired ISS property even under modeling errors in the online setting which we quantify precisely. We empirically demonstrate the performance of COCO-LQ in both synthetic experiments and a real-world power system frequency control setting.

Next, we investigate the effect of asynchrony and randomization on the stability of linear dynamical systems. Asynchrony and randomization are inherent in many computational tasks and dynamical systems and have been considered as means to increase computation speed and reduce cost, albeit at the expense of accuracy and convergence rate. Motivated by this, we propose random asynchronous LTI systems, a novel LTV system model, that generalizes the standard "synchronous" LTI systems, i.e., (1.1). In this model, each state variable is updated randomly and asynchronously with some probability, following the underlying LTI system structure.

We first explore the mean-square stability properties of these systems and analyze how stability varies with respect to randomization and asynchrony. Surprisingly, we show that the stability of random asynchronous LTI systems does not necessarily imply or is not necessarily implied by the stability of the synchronous variant of the system. We also demonstrate that an unstable synchronous system can be stabilized via randomization and/or asynchrony. This result highlights the novel challenges and opportunities brought about by the new dynamical system evolution formulation based on randomness and asynchrony.

We further investigate a special case of the introduced model, namely randomized LTI systems, where each state element is updated randomly with some fixed but unknown probability. We consider the problem of system identification for unknown

randomized LTI systems, utilizing the precise characterization of mean-square stability via the extended Lyapunov equation. We propose a system identification method to recover the underlying dynamics of unknown randomized LTI systems, including the model parameters, update probabilities of state variables, and noise covariance. Finally, we present empirical results that demonstrate the effectiveness of our proposed method in consistently recovering the underlying dynamics at the optimal rate. Our findings highlight the potential of using randomness and asynchrony in dynamical systems to achieve improved performance and shed light on the challenges and opportunities associated with analyzing input/output data from linear dynamical systems with a fixed network structure and random asynchronous updates.

Learning and Control in Partially Observable Linear Dynamical Systems (Chapter 5)

The focus of this chapter is on understanding the challenges posed by partial observability in learning and control of linear dynamical systems. Unlike the dynamical systems studied in Chapters 3 and 4, in partially observable linear dynamical systems, the decision-making agent does not have direct access to the state of the system. Instead, the learning agent is only able to observe a noisy linear measurement of the underlying latent state:

$$\begin{aligned} x_{t+1} &= A_* x_t + B_* u_t + w_t \\ y_t &= C_* x_t + z_t, \end{aligned} \tag{1.2}$$

where y_t is the measurement from the latent state x_t, and z_t is the measurement noise. Thus, these systems are commonly referred to as measurement-feedback systems. In this chapter, we study finite-time system identification, stabilization, and control performance of learning and control in partially observable linear dynamical systems given in (1.2) with quadratic regulatory costs $c_t = y_t^\top Q y_t + u_t^\top R u_t$ for $Q \succeq 0$ and $R \succ 0$, commonly referred to as Linear Quadratic Gaussian (LQG) control systems with Gaussian w_t and z_t, for *unknown parameters* $A_*, B_*,$ *and* C_*.

To begin, we explore the problem of learning unknown system dynamics and highlight the limitations of existing finite-time estimation techniques in the literature. We then present the first system identification method that enables model parameter estimation with finite-time guarantees in both open and closed-loop control settings. The key idea behind this learning method is to represent the dynamics in (1.2) in its predictor form introduced by Kalman in his seminal paper [134], which

is an input-output parametrization of the dynamics. We show that this learning method successfully overcomes the dependencies of covariates and noise terms, recovers a balanced realization of the model parameters with confidence intervals, and achieves optimal estimation error rate for both i.i.d. Gaussian control inputs and measurement-feedback controllers as long as the inputs are persistently exciting. The persistence of excitation (PE) condition refers to linear scaling of the smallest singular value of the Gram matrix, i.e., the sample covariance matrix of the covariates. At a high level, this condition refers to a full-row rank mapping of a short history of process and measurement noises to the covariates. Using this definition, we characterize the PE condition for the underlying system when it is controlled by its optimal controller.

Building on this novel closed-loop system identification method, we investigate the adaptive control problem in unknown partially observable linear dynamical systems. Our study mostly focuses on the canonical setting of LQG control systems, yet, we extend our results in various directions along the way. Overall, we propose three new algorithmic frameworks for learning and control of unknown partially observable linear dynamical systems. The first framework, **LQG** control via **Opt**imism (LqgOpt), employs the OFU principle to strike a balance between exploration and exploitation when designing controllers. Our analysis reveals that LqgOpt retains the PE condition, despite model estimation errors, if the optimal controller of the underlying system meets the aforementioned condition. After establishing the continuous optimal rate of improvement of model parameter estimates in LqgOpt, we determine the number of samples required to ensure closed-loop stability of its optimistic policies. Finally, we analyze the regret of LqgOpt against the optimal average expected cost and show that it achieves a state-of-the-art regret of $\tilde{\mathcal{O}}(\sqrt{T})$ after T interactions with the system. Our study introduces new learning and control theoretic techniques for analyzing the finite-time performance of adaptive control algorithms, which can be used to derive new regret guarantees in partially observable linear systems. These tools are of independent interest for future work in this area. Finally, we extend the results of LqgOpt to ARX systems.

Next, we delve into our second algorithmic framework, namely Thompson Sampling under Partial Observability (TSPO). TSPO uses our novel closed-loop system identification method to continuously improve the model parameter estimates and their confidence intervals and employs Thompson Sampling with these confidence intervals for control design in adaptive control of unknown LQG control systems.

Although TSPO has a larger problem-dependent constant in regret compared to LqgOpt due to replacing the optimization procedure for finding optimistic models with an efficient sampling method, we demonstrate that it still attains the state-of-the-art regret upper bound of $\tilde{O}(\sqrt{T})$ after T time steps.

Lastly, we study a more general setting of partially observable linear systems with strongly convex cost functions, which can be possibly time-varying. For this challenging learning and control problem, we study the finite-time regret performance of learning agents against the best controller, in hindsight, from a given set of controllers. For this learning and control problem, we propose an efficient adaptive control algorithm, Adaptive Control Online Learning (AdaptOn). AdaptOn turns the learning and control problem in partially observable linear dynamical systems into an online convex optimization problem. It adaptively learns the model dynamics via our novel model learning method, which overcomes the dependencies in data due to closed-loop control, and continuously optimizes the controller using a convex policy parameterization of measurement feedback controllers, which alleviates the highly nonlinear dependencies due to the feedback loop and gives a linear map that is computationally and statistically efficient to optimize. We show that this unique combination of tools from control theory and online learning allows AdaptOn to achieve *optimal logarithmic regret* in learning and control of partially observable linear dynamical systems with strongly convex cost. This is the first logarithmic regret bound for partially observable linear dynamical systems with unknown dynamics, which include the canonical setting of LQG control systems, and it improves the prior $\tilde{O}(\sqrt{T})$ regret bounds of LqgOpt and TSPO. We show that the strong convexity of the cost functions plays a major role in this improved regret performance, so much so that under (weakly) convex functions AdaptOn attains $\tilde{O}(\sqrt{T})$ regret, matching LqgOpt and TSPO. Finally, we extend our study of AdaptOn to the adaptive control of ARX systems and demonstrate that the results hold even in these more general systems with relaxed assumptions on system dynamics.

Learning and Control in Nonlinear Dynamical Systems (Chapter 6)

In the penultimate chapter of this thesis, we examine learning and control in dynamical systems in their most general setting: partially observable nonlinear dynamical systems. These nonlinear dynamical systems are commonly encountered in real-world applications as they capture complex systems with nonlinear hidden dynamics and observations. Drawing inspiration from the results of partially observable linear

dynamical systems, we propose novel system identification methods with finite-time learning guarantees for partially observable nonlinear systems. We consider two function classes for the system dynamics: systems that live in Reproducing Kernel Hilbert Spaces (RKHS) or Sobolev space of periodic functions. For nonlinear systems in RKHS, we utilize Random Fourier Features (RFF) [226] to represent and learn the system dynamics. Specifically, by generating a D-dimensional RFF basis, we learn the nonlinear system dynamics as a linear system over this basis. We provide a precise characterization of the required number of samples to optimally learn the nonlinear dynamics up to the desired error. To this end, we derive a novel function approximation theoretic guarantee for RFF learning, which can be of independent interest. In particular, we show that the best RFF approximation of a nonlinear system has an approximation error of $\tilde{O}(1/\sqrt{D})$, where D is the dimension of RFF representation. By combining this result with the linear system learning guarantees from previous chapters, we offer rigorous system identification guarantees for the underlying nonlinear system that lives in an RKHS, using RFF-based model learning.

Building upon this method, we present an efficient online control framework, **M**odel **L**earning **P**redictive **C**ontrol (MLPC), that learns to control unknown partially observable nonlinear systems using the estimated system dynamics via RFF-based system identification. MLPC deploys a model predictive control (MPC) method with the estimated dynamics for planning, and occasionally updates the model estimates to enhance the accuracy and effectiveness of the control policies. We provide stability guarantees for single trajectory online control via MLPC, presuming that the given MPC method stabilizes the underlying system for sufficiently small estimation errors. Finally, we prove that MLPC attains $\tilde{O}(T^{2/3})$ regret with respect to the agent that uses the same MPC policy with the true system dynamics. To demonstrate the efficacy of MLPC empirically, we showcase its performance on the classical inverted pendulum task using two different MPC methods.

We then shift our attention to the nonlinear systems that live in the Sobolev space of periodic functions. For such systems, we introduce a model learning method that employs a finite order Fourier series basis. Similar to the RFF setting, we propose to learn the nonlinear system dynamics as a linear system over the selected Fourier basis. We establish that this approach estimates the underlying nonlinear system with a near-optimal estimation error of $\tilde{O}(T^{\varepsilon-0.5})$, after T samples, where ε depends on the smoothness of the Sobolev space and the order of the Fourier basis.

Inspired by this effective modeling strategy, we tackle the challenging real-world

aerodynamic control problem of disturbance rejection in extreme turbulence. Controlling aerodynamic forces in gusty and turbulent conditions is essential for the safe and effective operation of unmanned aerial vehicles (UAVs). However, since these extreme flow conditions are difficult to predict and model explicitly, it is challenging to design effective flow-informed controllers beyond traditional reactive control methods. Moreover, the noise and error in onboard sensor measurements bring further uncertainties in designing control policies. To address these challenges, we introduce **F**ourier **A**daptive **L**earning and **Con**trol (FALCON), which is the first model-based reinforcement learning method capable of efficiently learning to control aerodynamic forces acting on an airfoil under extreme turbulence. Using the Fourier basis-based model learning strategy, FALCON achieves effective modeling and control of the aerodynamic forces due to turbulent flow dynamics and achieves state-of-the-art disturbance rejection performance.

FALCON builds on two key observations: that the chaotic dynamics involved in turbulent flows are well-modeled in the frequency domain and that most of the energy in turbulent flows is stored in low-frequency components. Leveraging these observations, FALCON selects a concise Fourier basis to learn the underlying system dynamics using only 35 seconds of flow data. To address the issue of partial observability due to sensor measurements, FALCON uses a short history of actions and measurements to model the system dynamics. With this physically sound and accurate model-learning approach, FALCON employs an MPC method similar to MLPC for safe and efficient control design. When evaluated under highly turbulent wind conditions generated in Caltech's closed-loop wind tunnel, FALCON learns the underlying nonlinear dynamics and adapts to the changing flow conditions with less than 9 minutes of data and consistently outperforms the state-of-the-art methods by at least 37%.

In addition to strong empirical performance, FALCON comes with performance guarantees which certify the stability and robustness of the proposed framework. In particular, building upon the near-optimal estimation error rate of finite-order Fourier-basis learning, we show that FALCON can achieve desired estimation error for stability in finite time and provide closed-loop stability with a given MPC policy. We then show that FALCON follows a trajectory close to the agent that uses the same MPC method with the true system dynamics, and attains $\tilde{O}(\sqrt{T})$ regret against this agent. To the best of our knowledge, FALCON is the *first* efficient RL algorithm that achieves $O(\sqrt{T})$ regret in online control of nonlinear dynamical systems, and its

guarantees extend to a wide range of partially observable nonlinear dynamical systems such as real-world dynamical systems governed by partial differential equations.

Lastly, we focus on addressing the limitations of our previous results in this chapter, which assumed that given MPC methods stabilize the underlying system for small enough estimation errors. To overcome this limitation, we incorporate a control theoretic stability verification method in the policy design for adaptive control of nonlinear dynamical systems. In particular, we design a new policy optimization problem that adopts Krasovskii's construction of quadratic Lyapunov functions as a stability constraint in the control design. We prove that if the underlying system satisfies Krasovskii's Lyapunov function construction for a class of controllers $g_\theta(\cdot)$ parameterized by θ, then the new constrained policy optimization problem is guaranteed to stabilize the underlying system even under modeling errors, e.g., estimation errors. Furthermore, we provide a characterization of the required estimation errors to achieve stabilization. This leads to a policy design method with a stability margin, which can be used for a wide range of nonlinear dynamical systems.

To deploy this policy optimization problem in an RL pipeline, we propose a primal-dual method to solve this problem and show that at convergence this method guarantees the recovery of a stabilizing policy even under modeling error. Combining this result with the RFF model learning method, we propose a novel model-based RL framework called Krasovskii-Constrained RL (KCRL). We formally guarantee that KCRL learns a stabilizing policy in finite time/samples with an explicit characterization of the required number of samples. Finally, we evaluate the performance of the KCRL framework in a real-world application of voltage control in a distributed power system under different operating conditions. We show that KCRL guarantees stability under all operating conditions, whereas standard RL methods fail to stabilize.

Chapter 2

STOCHASTIC LINEAR BANDITS WITH PRACTICAL CONCERNS

In this chapter, we study stochastic linear bandits (SLB) which can be regarded as the most basic dynamical system without an evolving state over time, yet with the control goal of maximizing the reward[1]. In this classical setting of decision-making, we consider 2 practical aspects which are present in modern-day decision-making under uncertainty tasks: (i) high-dimensional feature representations for the action vectors and (ii) unknown (nonlinear) safety constraints.

For setting (i), we propose Projected Stochastic Linear Bandits (PSLB), a sequential decision-making algorithm for high-dimensional stochastic linear bandits. We show that when the representation of actions has an underlying unknown low-dimensional subspace structure, PSLB deploys principal component analysis-based projection to efficiently recover this structure. PSLB exploits this hidden structure to better guide the exploration and exploitation, resulting in a significant improvement in performance. We prove that PSLB notably advances the previously known regret upper bounds and obtains a regret upper bound, which scales with the intrinsic dimension of the subspace, rather than the large ambient dimension of the action space. We empirically study PSLB on a range of image classification tasks formulated as bandit problems. We show that when a pre-trained deep neural network provides the high-dimensional action (label) representations, deploying PSLB results in a significant reduction of regret and faster convergence to an accurate model compared to the state-of-art algorithm.

In many real-world decision-making tasks, e.g., clinical trials, the agents must satisfy a diverse set of unknown safety constraints at all times while getting feedback only on the safety constraints relevant to the chosen action, e.g., the ones close to the violation. For setting (ii), we study stochastic linear bandits with such unknown safety constraints and local safety feedback. The agent's goal is to maximize the cumulative reward while satisfying *multiple unknown affine or nonlinear* safety constraints. At each time step, the agent receives noisy feedback on a particular safety constraint *only if* the chosen action belongs to the associated constraint set,

[1]This chapter is based on [14, 159].

i.e., local safety feedback. For this setting, we design upper confidence bound and Thompson Sampling-based algorithms. In the design of these algorithms, we carefully prescribe an additional exploration incentive that guarantees the selection of high-reward actions that are also safe and ensures sufficient exploration in the relevant constraint sets to recover the optimal safe action. We show that for M distinct constraints, both of these algorithms attain $\tilde{O}(\sqrt{MT})$ regret after T time steps without any safety violations. We empirically study the performance of the proposed algorithms under various safety constraints and with a real-world credit dataset. We show that both algorithms safely explore and quickly recover the optimal safe actions.

2.1 Motivation and Background

Fundamentals of SLB problem: In the SLB problems, the goal is to strike a balance between exploration and exploitation such that the decision-making agent minimizes the regret with respect to the optimal policy. One promising approach is to utilize the *optimism in the face of uncertainty* (OFU) principle [156]. OFU-based methods estimate the environment model up to a confidence interval and construct a set of plausible models within this interval. Among those models, they choose the most optimistic one and follow the optimal behavior suggested by the selected model. Therefore, these methods are also referred to as the upper-confidence bound algorithms.

Another prominent strategy to balance the exploration vs. exploitation trade-off is Thompson Sampling (TS). In TS, the agent samples a model from a distribution computed based on prior action and reward pairs, and then selects the optimal action for this sampled model and updates the distribution based on its novel reward. Since it relies solely on sampling, this approach provides polynomial-time algorithms and can be more efficient than OFU-based methods.

For general stochastic linear bandit problems, Abbasi-Yadkori et al. [3] deploys the OFU principle, proposes OFUL algorithm, and for a d-dimensional stochastic linear bandit and T time steps of agent-environment interaction, derives a regret upper bound of $\widetilde{O}(d\sqrt{T})$. These types of regret bounds in high-dimensional problems, especially when d and T are about the same order are not practically tolerable. Fortunately, real-world problems may contain hidden low intrinsic dimensional structures. For example, in classical recommendation systems, each item is represented by a large and highly detailed hand-engineered feature vector; however, not all the components of the features are helpful for the recommendation task. Therefore, the true underlying linear function in stochastic linear bandits can be considered to be

highly sparse. Abbasi-Yadkori et al. [4] and Carpentier and Munos [47] show how to exploit this additional structure and, under slightly different assumptions, derive regret upper bounds of $\widetilde{O}\left(\sqrt{sdT}\right)$ and $\widetilde{O}\left(s\sqrt{T}\right)$ respectively where s is the sparsity level of the true underlying linear function.

High-dimensional SLB framework: The recent successes of deep neural networks (DNN) in representation learning provide significant promises in advancing machine learning to high-dimensional real-world tasks [26, 172, 173], where SLB is one of them. DNNs receive the raw features of the input and pass them through a variety of potentially nonlinear layers and construct new feature representations which can arguably replace the hand-engineered feature vectors. When a DNN provides the feature representations for actions, the sparse structure is not relevant anymore; instead, the low-rank sub-space structure might be considered as suitable.

At each round of SLB, the agent chooses an action from a given decision set, and the environment reveals the reward associated with that action. Therefore, the chosen action is assigned a *supervised* reward signal, while other actions in the decision sets remain *unsupervised*. One of the primary motivations of the SLB framework is the study of decision-making under uncertainty in large decision sets with potentially possible intrinsic hidden structures. For the SLB framework, the majority of the prior works are mainly devoted to the general cases where no possible hidden or low intrinsic dimensional structures in the decision sets are considered [67, 169, 180, 232]. For example, the groundbreaking work by [3] utilizes only the supervised actions, i.e., the actions selected by the algorithm, to estimate the environment model. It ignores all other unsupervised actions in the decision set. On the contrary, in the presence of such latent structures, the large number of actions in the decision sets can be utilized to understand the latent structure, reducing the dimension of the problem, and improving the learning and sample complexity.

Safety in SLB setting: In many real-world decision-making problems, the agents require satisfaction of some safety/operational constraints while aiming to maximize the cumulative reward. Thus, the tools developed for unconstrained stochastic linear bandit framework do not directly apply to real-world safety-critical decision-making tasks such as clinical trials [286]. There have been several new frameworks with different forms of safety constraints proposed to model these tasks. Some of these frameworks include safety constraints through stage-wise reward [143, 202], while some of them focus on cumulative or policy-based constraints [140, 187, 216].

Another line of work considers a more challenging setting of hard constraints on the actions, where the safety constraint needs to be met at every time step (stage-wise) [12, 203]. This setting is more suitable for safety-critical tasks where executing even one unsafe action may lead to catastrophic results. However, prior works in this setting only consider very simple models where there is a single unknown linear constraint depending on the reward function that the agent observes feedback from at every time step. Despite giving an initial understanding of safety in the stochastic linear bandit, these works do not capture the complex constraint and feedback structure of real-world decision-making tasks. The following considers a safety-critical decision-making scenario in which the prior works fail to model.

2.2 Stochastic Linear Bandits with Hidden Low-Rank Structure

In this section, we demonstrate a method that utilizes unsupervised actions in the decision sets to improve the performance in stochastic linear bandits. We deploy the principal component analysis (PCA), one of the highly celebrated subspace recovery methods, to exploit the subspace structure of the action set using the massive number of unsupervised actions observed in the decision sets and finally reduce the dimensionality and the complexity of stochastic linear bandits. We propose Projected Stochastic Linear Bandits (PSLB), an algorithm for high-dimensional stochastic linear bandits, and show that if the actions come from a perturbed m-dimensional subspace, deploying PSLB improves the regret upper bound to $min\left\{\widetilde{O}\left(\Upsilon\sqrt{T}\right), \widetilde{O}\left(d\sqrt{T}\right)\right\}$. Here Υ captures the effect of the difficulty of the subspace recovery in stochastic linear bandit as a function of the problem structure. If the underlying subspace is easily identifiable, *e.g.*, large decision sets per round, recovering the subspace provides faster learning of the underlying linear function; therefore, resulting in a smaller regret. In contrast, if learning the subspace is hard, *e.g.*, the number of actions (unsupervised signals) in each round is small, then subspace recovery-based approaches might not provide many benefits in learning the underlying system; therefore, resulting in a similar performance of the plain OFUL.

We test the performance of PSLB both on artificial data and real-world data and compare it with OFUL. We first generate a stochastic linear bandit problem with a hidden low-rank structure with different levels of perturbation. We empirically demonstrate that as the subspace recovery gets easier due to a decreased level of perturbations, PSLB explores more efficiently and learns the underlying model faster, resulting in smaller regret compared to OFUL. We then adapt the image classification tasks on

MNIST [172], CIFAR-10 [150], and ImageNet [151] datasets to the stochastic linear bandit framework and apply both PSLB and OFUL on these datasets. We first verify that the feature representation of the DNNs exhibits a low-dimensional subspace structure when a pre-trained DNN produces the d-dimensional feature vectors. We empirically show that PSLB learns the underlying model significantly faster than OFUL and provides orders of magnitude smaller regret in stochastic linear bandits obtained from MNIST, CIFAR-10, and ImageNet datasets.

2.2.1 Problem Formulation

Model: At each round t, the agent is given a decision set D_t with K actions, $\hat{x}_{t,1}, \ldots, \hat{x}_{t,K} \in \mathbb{R}^d$. Let V be an $d \times m$ orthonormal matrix with $m \leq d$, where $\text{span}(V)$ defines a m-dimensional subspace in $\mathbb{R}^d$. The followings are the translation of the standard considerations in the PCA-based subspace recovery methods [204, 284] to stochastic linear bandits. Consider a zero mean true action vector, $x_{t,i} \in \mathbb{R}^d$, such that $x_{t,i} \in \text{span}(V)$ for all $i \in [K]$. Let $\psi_{t,i} \in \mathbb{R}^d$ be zero mean random vectors which are uncorrelated with true action vectors, i.e., $\mathbb{E}[x_{t,i}\psi_{t,i}^T] = 0$ for all $i \in [K]$. Each action vector $\hat{x}_{t,i}$ is generated as follows,

$$\hat{x}_{t,i} = x_{t,i} + \psi_{t,i}. \tag{2.1}$$

This model states that each $\hat{x}_{t,i}$ in D_t is a perturbed version of the true underlying $x_{t,i}$. Denote the covariance matrix of $x_{t,i}$ by Σ_x. Notice that Σ_x is rank-m. Perturbation vectors, $\psi_{t,i}$, are isotropic, thus covariance matrix $\Sigma_\psi = \sigma^2 I_d$. Let $\lambda_+ \coloneqq \lambda_1(\Sigma_x)$ and $\lambda_- \coloneqq \lambda_m(\Sigma_x)$.

Assumption 2.2.1 (Bounded Action and Perturbation Vectors)**.** *There exists finite constants, d_x and d_ψ, such that for all $i \in [K]$,*

$$\|x_{t,i}\|_2^2 \leq d_x \lambda_+, \quad \|\psi_{t,i}\|_2^2 \leq d_\psi \sigma^2.$$

Both d_x and d_ψ can be dependent on m or d and they can be interpreted as the effective dimensions of the corresponding vectors. At each round t, the agent chooses an action, $\hat{X}_t \in D_t$ and observes a reward r_t such that

$$r_t = \hat{X}_t^T \theta_* + \eta_t \qquad \forall t \in [T], \tag{2.2}$$

where $\theta_* \in \text{span}(V)$ is the unknown parameter vector and η_t is the random noise at round t. Notice that since $\theta_* \in \text{span}(V)$, $r_t = \hat{X}_t^T \theta_* + \eta_t = (P\hat{X}_t)^T \theta_* + \eta_t$, where $P = VV^T$ is the projection matrix to the m-dimensional subspace $\text{span}(V)$. We later

exploit this equivalence in finding the optimistic action. Consider $\{F_t\}_{t=0}^{\infty}$ as any filtration of σ-algebras such that for any $t \geq 1$, $\hat{X}_t$ is F_{t-1} measurable and η_t is F_t measurable.

Assumption 2.2.2 (Sub-Gaussian Noise)**.** *For all t, η_t is conditionally R-sub-Gaussian where $R \geq 0$ is a fixed constant, i.e., $\forall \lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda \eta_t} | F_{t-1}] \leq e^{\frac{\lambda^2 R^2}{2}}$.*

The goal of the agent is to maximize the total expected reward accumulated in any T rounds, $\sum_{t=1}^{T} \hat{X}_t^T \theta_*$. With the knowledge of θ_*, the oracle chooses the action $\hat{X}_t^* = \arg\max_{x \in D_t} x^T \theta_*$ at each round t. We evaluate the agent's performance against this oracle's performance. Define the *regret* at round T as the difference between the expected reward of the oracle and the agent,

$$R_T \coloneqq \sum_{t=1}^{T} \hat{X}_t^{*T} \theta_* - \sum_{t=1}^{T} \hat{X}_t^T \theta_* = \sum_{t=1}^{T} (X_t^* - \hat{X}_t)^T \theta_*. \tag{2.3}$$

The agent aims to minimize this quantity over time. Similar to the problem of sparse linear bandit where the sparsity level is known to the agent [4, 169], in the setting described above, the dimensionality of subspace m is known to the agent. This assumption is also needed and standard to PCA analyses [204, 284]. In practice, m, along with other problem-dependent quantities, can be estimated and updated at each round. Finally, we define the following quantities about the structure of the problem. For all $\delta \in (0, 1)$:

$$g_x = \frac{\lambda_+}{\lambda_-}, \quad g_\psi = \frac{\sigma^2}{\lambda_-}, \quad \Gamma = 2g_\psi + 4\sqrt{g_x g_\psi}, \quad \alpha = \max(d_x, d_\psi),$$
$$n_\delta = 4\alpha \left(\Gamma \sqrt{\log \frac{2d}{\delta}} + \sqrt{2 g_x \log \frac{m}{\delta}} \right)^2. \tag{2.4}$$

2.2.2 Projected Stochastic Linear Bandits (PSLB)

We propose Projected Stochastic Linear Bandits (PSLB), a stochastic linear bandit algorithm that employs subspace recovery to extract information from the unsupervised data accumulated in the stochastic linear bandit. The PSLB is illustrated in Algorithm 1. PSLB consists of three key elements: subspace estimation, creating confidence sets, and acting optimistically. In the following, we discuss each of these elements.

Subspace estimation: At each round t, the agent exploits the action vectors observed up to round t, $\uplus_{i=1}^{t} D_i$, to estimate the underlying m-dimensional subspace. In particular, the agent performs PCA on action vectors and computes $\hat{V}_t$, the matrix of

Algorithm 1 PSLB

Input: m, λ_+, λ_-, σ^2, α, δ
for t = 1 to T **do**
  Compute PCA over $\uplus_{i=1}^{t} D_i$
  Create $\hat{P}_t$ with first m eigenvectors
  Construct $C_{p,t}$, high probability confidence set on $\hat{P}_t$
  Construct $C_{m,t}$, high probability confidence set for θ_* using subspace recovery
  Construct $C_{d,t}$, high probability confidence set for θ_* without using subspace recovery
  Construct $C_t = C_{m,t} \cap C_{d,t}$
  $(\tilde{\theta}_t, \hat{X}_t) = \text{argmax}_{(\theta,x)\in C_t \times D_t} x^T\theta$
  Play $\hat{X}_t$ and observe r_t

top m eigenvectors of $\frac{1}{tK}\sum_{\hat{x}\in\uplus_{i=1}^{t} D_i} \hat{x}\hat{x}^T$. span($\hat{V}_t$) is the estimate of the underlying m-dimensional subspace. The agent uses $\hat{V}_t$ to compute $\hat{P}_t \coloneqq \hat{V}_t\hat{V}_t^T$, the projection matrix onto span($\hat{V}_t$), and constructs a confidence set $C_{p,t}$ around $\hat{P}_t$ which contains both $\hat{P}_t$ and P with high probability.

Next, we demonstrate the construction of $C_{p,t}$, and show that as the agent observes more action vectors, $C_{p,t}$ shrinks and the estimation error on $\hat{P}_t$ vanishes.

Confidence set construction: At the beginning of each round t, the agent uses $\hat{P}_t$ and projects the supervised actions, the actions that have been chosen, and assigned rewards in the previous rounds, onto the estimated m-dimensional subspace. The d-dimensional linear bandit problem reduces to a m-dimensional problem. The agent then estimates the model parameter θ_*, as θ_t, up to a high probability confidence set $C_{m,t}$. The tightness of this confidence interval, besides the action-reward pairs, depends on the confidence in the estimation of the subspace and its confidence interval $C_{p,t}$.

Simultaneously, relying *only* on the history of action-reward pairs, the supervised actions, the agent estimates the model parameter θ_*, as $\hat{\theta}_t$, up to another high-probability confidence set $C_{d,t}$. This is the same confidence set generation subroutine as OFUL [3]. Since θ_* lives in both of these sets with high probability, it lies in the intersection of them with high probability. Finally, the agent takes the intersection of the constructed confidence sets to create the main confidence set, $C_t = C_{m,t} \cap C_{d,t}$. If an efficient recovery of the subspace is possible, then the plausible parameter set of $C_{m,t}$ is significantly smaller than the set of $C_{d,t}$, resulting in smaller C_t as well as more confident parameter estimation. If the subspace recovery is hard, then $C_{m,t}$ might not provide much information, and the intersection would mainly result in $C_{d,t}$.

Acting optimistically: The agent finally searches for the highest reward-bringing model and action pair from $C_t \times D_t$ and executes the reward maximizing action, line 9 in Algorithm 1.

2.2.3 Theoretical Analysis of PSLB

In this section, we state the regret upper bound of PSLB and provide the theoretical components that build up to this result. Recalling the quantities defined in (2.4), define Υ such that

$$\Upsilon = O\left(\left(1 + \Gamma\sqrt{\frac{\alpha}{K}}\right)\left(\frac{\Gamma\sqrt{m\alpha}}{\sqrt{K}\sqrt{\lambda_- + \sigma^2}} + m\right)\right). \tag{2.5}$$

It represents the overall effect of the deploying subspace recovery on the regret in terms of structural properties of the stochastic linear bandit setting.

Theorem 2.2.3 (Regret Upper Bound of PSLB)**.** *Fix any $\delta \in (0, 1)$. Assume that for all $\hat{x}_{t,i} \in D_t$, $\hat{x}_{t,i}^T \theta_* \in [-1, 1]$. Under Assumptions 2.2.1 & 2.2.2, $\forall t \geq 1$, with probability at least $1 - 6\delta$, the regret of PSLB satisfies*

$$R_t = \min\left\{\widetilde{O}\left(\Upsilon\sqrt{t}\right), \widetilde{O}\left(d\sqrt{t}\right)\right\}. \tag{2.6}$$

The proof of the theorem involves three main pieces: the projection error analysis, the construction of projected confidence sets, and the regret analysis.

Projection Error Analysis

Consider the matrix $\hat{V}_t^T V$ and its ith singular value denoted as $\sigma_i(\hat{V}_t^T V)$, such that $\sigma_1(\hat{V}_t^T V) \geq \ldots \geq \sigma_m(\hat{V}_t^T V)$. Using the definition of the aperture of two linear manifolds [11], we write the following equivalence:

$$\begin{aligned}
\|\hat{P}_t - P\|_2 &= \max\left\{\max_{x \in \text{span}(V), \|x\|_2 = 1} \|(I_d - \hat{P}_t)x\|_2, \max_{y \in \text{span}(\hat{V}_t), \|y\|_2 = 1} \|(I_d - P)y\|_2\right\} \\
&= \max\left\{\|(I - \hat{V}_t\hat{V}_t^T)V\|_2, \ \|(I - VV^T)\hat{V}_t\|_2\right\} \\
&= \sqrt{\lambda_{max}\left(V^T(I_d - \hat{V}_t\hat{V}_t^T)(I_d - \hat{V}_t\hat{V}_t^T)V\right)} \qquad (2.7) \\
&= \sqrt{\lambda_{max}\left(I_m - (\hat{V}_t^T V)^T(\hat{V}_t^T V)\right)} \\
&= \sqrt{1 - \sigma_m^2} \qquad \text{where } \sigma_m \text{ is the smallest singular value of } \hat{V}_t^T V, \\
&= \sqrt{1 - \cos^2\Theta_m(\text{span}(V), \text{span}(\hat{V}))} = \sin\Theta_m, \qquad (2.8)
\end{aligned}$$

where (2.7) follows since V, and $\hat{V}_t$ have same dimensions, and (2.8) follows from the fact that $\cos\Theta_i(\text{span}(V), \text{span}(\hat{V})) = \sigma_i(\hat{V}_t^T V)$ where Θ_m is the largest principal angle between the column spans of V and $\hat{V}_t$. Thus, bounding the projection error between two projection matrices is equivalent to bounding the sine of the largest principal angle between the subspaces that they project. In light of this relation, and the prior analysis of Davis-Kahan $\sin\Theta$ Theorem [68], we provide the following Lemma on the concentration of the sine and the finite sample projection error.

Lemma 2.2.4 (Finite Sample Projection Error)**.** *Fix any $\delta \in (0, 1)$. Let $t_{w,\delta} = \frac{n_\delta}{K}$. Suppose Assumption 2.2.1 holds. Then with probability at least $1 - 3\delta$, $\forall t \geq t_{w,\delta}$,*

$$\|\hat{P}_t - P\|_2 \leq \frac{\phi_\delta}{\sqrt{t}} \quad , \text{ where } \phi_\delta = 2\Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}}. \tag{2.9}$$

Lemma 2.2.4 improves the existing bounds on the projection error (Corollary 2.9 in Vaswani and Narayanamurthy [284]) by using the matrix Chernoff inequality [268]. It also provides the precise problem-dependent quantities in the bound which are required for defining the minimum number of samples required to construct tight confidence sets by using subspace estimation. The formal and detailed version of the Lemma 2.2.4 and the details of the proof are provided in Appendix A.1.1.

Note that, as discussed in Section 2.2.2, we define the confidence set $C_{p,t}$ in (2.9) for all $t \geq t_{w,\delta}$. Due to the equivalence $\|\hat{P}_t - P\|_2 = \sin\Theta_m$, $\forall t \geq 1$ we have that $\|\hat{P}_t - P\|_2$ is always less than or equal to 1, i.e., $\|\hat{P}_t - P\|_2 \leq 1$. Therefore, any projection error bound greater than 1 is vacuous. Consequently, we state that, with high probability, the bound on the projection error in (2.9) becomes less than 1 when $t \geq t_{w,\delta}$. After round $t_{w,\delta}$, PSLB starts to produce non-trivial confidence sets $C_{p,t}$ around $\hat{P}_t$. However, note that $t_{w,\delta}$ can be significantly large for problems that have latent structures that are hard to recover, e.g., having α linear in d.

The term ϕ_δ in Lemma 2.2.4 also provides several important intuitions about the subspace estimation problem in terms of the problem structure. Recalling the definition of Γ in (2.4), as g_ψ decreases, the projection error shrinks since the underlying subspace becomes more distinguishable. Conversely, as g_x diverges from 1, it becomes harder to recover the underlying m-dimensional subspace. Additionally, since α is the maximum of the effective dimensions of the true action vector and the perturbation vector, having large α makes the subspace recovery harder and the projection error bound looser, whereas observing more action vectors, K, in each round produces tighter bound on $\|\hat{P}_t - P\|_2$. The effects of these structural properties

on the subspace estimation translate to confidence set construction and ultimately to the regret upper bound.

Projected Confidence Sets

In this section, we analyze the construction of $C_{m,t}$ and $C_{d,t}$. For any round $t \geq 1$, define $\hat{\Sigma}_t \coloneqq \sum_{i=1}^{t} \hat{X}_i \hat{X}_i^T = \hat{\mathbf{X}}_t \hat{\mathbf{X}}_t^T$. At round t, let $A_t \coloneqq \hat{P}_t(\hat{\Sigma}_{t-1}+\lambda I_d)\hat{P}_t$ for $\lambda > 0$. Let B_t be a symmetric matrix such that $A_t = \hat{V}_t B_t \hat{V}_t^T$. Notice that B_t is a full rank $m \times m$ matrix. The rewards obtained up to round t are denoted as $\mathbf{r}_{t-1}$. At round t, after estimating the projection matrix $\hat{P}_t$ associated with the underlying subspace, PSLB finds θ_t, an estimate of θ_*, while having θ_* living within the estimated subspace with high probability. Therefore, θ_t is the solution to the following Tikhonov-regularized least squares problem with regularization parameters $\lambda > 0$ and $\hat{P}_t$,

$$\theta_t = \underset{\theta}{\operatorname{argmin}} \|(\hat{P}_t \hat{\mathbf{X}}_{t-1})^T \theta - \mathbf{r}_{t-1}\|_2^2 + \lambda \|\hat{P}_t \theta\|_2^2.$$

Notice that regularization is applied along the estimated subspace. Solving for θ gives $\theta_t = A_t^{\dagger}\left(\hat{P}_t \hat{\mathbf{X}}_{t-1} \mathbf{r}_{t-1}\right)$. Let $S_t \coloneqq \sum_{i=1}^{t} \hat{P}_t \hat{X}_{i-1} \eta_{i-1} = \hat{P}_t \mathbf{X}_{t-1} \boldsymbol{\eta}_{t-1}$. Before presenting the confidence set construction, we provide a self-normalized bound on S_t.

Theorem 2.2.5 (Self-Normalized Bound for Vector-Valued Martingales)**.** *For any $\delta \in (0,1)$, with probability at least $1-\delta$, for all $t \geq 1$,*

$$\|S_t\|_{A_t^{\dagger}}^2 \leq 2R^2 \log\left(\frac{\det(B_t)^{1/2} \det(\lambda I_m)^{-1/2}}{\delta}\right).$$

This result is a similar self-normalized bound for vector-valued martingales in Abbasi-Yadkori et al. [3], and it can be considered as the projected version of their Theorem 1. The proof of Theorem 2.2.5 is given in Appendix A.1.2. Define L such that for all $t \geq 1$ and $i \in [K]$, $\|\hat{x}_{t,i}\|_2 \leq L$ and let $\gamma = \frac{L^2}{\lambda \log\left(1+\frac{L^2}{\lambda}\right)}$. Consider the following lemmas that will be useful in proving confidence set construction, and their proofs are given in Appendix A.1.3.

Lemma 2.2.6. *Suppose Assumptions 2.2.1 & 2.2.2 hold. Then,* $\det(B_t) \leq \left(\lambda + \frac{tL^2}{m}\right)^m$.

Lemma 2.2.7. *Suppose Assumptions2.2.1 & 2.2.2 hold. Then,*

$$\|(A_t^{\dagger})^{1/2} \hat{P}_t \hat{\Sigma}_{t-1}\|_2 \leq L\sqrt{t}\sqrt{\gamma m}\sqrt{\log\left(1 + \frac{tL^2}{m\lambda}\right)}.$$

The following theorem gives the construction of the projected confidence set, $C_{m,t}$.

Theorem 2.2.8 (Projected Confidence Set Construction, $C_{m,t}$)**.** *Fix any $\delta \in (0,1)$. Let Assumptions 2.2.1 & 2.2.2 hold, and $\forall t \geq 1$ and $i \in [K]$, $\|\hat{x}_{t,i}\|_2 \leq L$. If $\|\theta_*\|_2 \leq S$ then, with probability at least $1-4\delta$, $\forall t \geq t_{w,\delta}$, θ_* lies in the set $C_{m,t} = \left\{\theta \in \mathbb{R}^d : \|\theta_t - \theta\|_{A_t} \leq \beta_{t,\delta}\right\}$, where*

$$\beta_{t,\delta} = R\sqrt{2\log\left(\frac{1}{\delta}\right) + m\log\left(1+\frac{tL^2}{m\lambda}\right)} + LS\phi_\delta\sqrt{\gamma m\log\left(1+\frac{tL^2}{m\lambda}\right)} + S\sqrt{\lambda}. \quad (2.10)$$

Proof. From the definition of θ_t and r_t, we get the following:

$$\begin{aligned}
\theta_t &= A_t^\dagger S_t + A_t^\dagger \hat{P}_t\hat{\Sigma}_{t-1}P\theta_* \quad \text{since } \theta_* \in \text{span}(V)\\
&= A_t^\dagger S_t + A_t^\dagger\left(\hat{P}_t\hat{\Sigma}_{t-1}(\hat{P}_t + P - \hat{P}_t) + \lambda\hat{P}_t - \lambda\hat{P}_t\right)\theta_*\\
&= A_t^\dagger S_t + \hat{P}_t\theta_* + A_t^\dagger(\hat{P}_t\hat{\Sigma}_{t-1}(P-\hat{P}_t))\theta_* - \lambda A_t^\dagger\theta_*.
\end{aligned}$$

Using this, we derive the following for $x = A_t(\theta_t - \theta_*)$:

$$\begin{aligned}
x^T\theta_t - x^T\theta_* &= x^TA_t^\dagger S_t + x^TA_t^\dagger(\hat{P}_t\hat{\Sigma}_{t-1}(P-\hat{P}_t))\theta_* - \lambda x^TA_t^\dagger\theta_*\\
&= \langle x, S_t\rangle_{A_t^\dagger} + \langle x, \hat{P}_t\hat{\Sigma}_{t-1}(P-\hat{P}_t)\theta_*\rangle_{A_t^\dagger} - \lambda\langle x,\theta_*\rangle_{A_t^\dagger}.
\end{aligned}$$

Using Cauchy-Schwarz inequality, we can upper bound the magnitude of the difference as follows:

$$\begin{aligned}
|x^T\theta_t - x^T\theta_*| &\leq \|x\|_{A_t^\dagger}\left(\|S_t\|_{A_t^\dagger} + \|\hat{P}_t\hat{\Sigma}_{t-1}(P-\hat{P}_t)\theta_*\|_{A_t^\dagger} + \lambda\|\theta_*\|_{A_t^\dagger}\right)\\
&\leq \|x\|_{A_t^\dagger}\left(\|S_t\|_{A_t^\dagger} + \|(A_t^\dagger)^{1/2}\hat{P}_t\hat{\Sigma}_{t-1}(P-\hat{P}_t)\theta_*\|_2 + \sqrt{\lambda}\|\theta_*\|_2\right) \quad (2.11)\\
&\leq \|x\|_{A_t^\dagger}\left(\|S_t\|_{A_t^\dagger} + \|(A_t^\dagger)^{1/2}\hat{P}_t\hat{\Sigma}_{t-1}\|_2\|P-\hat{P}_t\|_2\|\theta_*\|_2 + \sqrt{\lambda}\|\theta_*\|_2\right).
\end{aligned}$$

Plugging in $x = A_t(\theta_t - \theta_*)$, we get

$$\|\theta_t-\theta_*\|_{A_t}^2 \leq \|A_t(\theta_t-\theta_*)\|_{A_t^\dagger}\left(\|S_t\|_{A_t^\dagger}+\|(A_t^\dagger)^{1/2}\hat{P}_t\hat{\Sigma}_{t-1}\|_2\|P-\hat{P}_t\|_2\|\theta_*\|_2+\sqrt{\lambda}\|\theta_*\|_2\right).$$

Since $\|A_t(\theta_t-\theta_*)\|_{A_t^\dagger} = \|\theta_t-\theta_*\|_{A_t}$, dividing both sides with $\|\theta_t-\theta_*\|_{A_t}$ gives and using the fact that $\|\theta_*\| \leq S$,

$$\|\theta_t-\theta_*\|_{A_t} \leq \|S_t\|_{A_t^\dagger} + S\|(A_t^\dagger)^{1/2}\hat{P}_t\hat{\Sigma}_{t-1}\|_2\|P-\hat{P}_t\|_2 + S\sqrt{\lambda}. \quad (2.12)$$

Notice that the first term is the projected version of Theorem 1 in [3] and the second term is the additional term appearing in the confidence interval construction due to non-zero projection error. As it can be seen with the knowledge of true projection

matrix, the confidence interval reduces to the one in [3] with replacement of d with m.

Using Theorem 2.2.5 and Lemma 2.2.4, we get:

$$\|\theta_t - \theta_*\|_{A_t} \leq R\sqrt{2\log\frac{\det(B_t)^{1/2}\det(\lambda I_m)^{-1/2}}{\delta}} + \frac{S\phi_\delta}{\sqrt{t}}\|(A_t^\dagger)^{1/2}\hat{P}_t\hat{\Sigma}_{t-1}\|_2 + S\sqrt{\lambda}.$$

Finally, combining this with Lemma 2.2.6 and Lemma 2.2.7 gives the statement of the Theorem 2.2.8. □

Notice that the overall proof follows a similar machinery used by [3]. Specifically, the first term of $\beta_{t,\delta}$ in (2.10) is derived similarly by a self-normalized tail inequality, Theorem 2.2.5. However, since at each round PSLB projects the supervised actions to an estimated m-dimensional subspace to estimate θ_*, d is replaced by m in the bound using Lemma 2.2.6. While enjoying the benefit of projection, this construction of the confidence set suffers from the finite sample projection error, i.e., uncertainty in the subspace estimation. This effect is observed via the second term in (2.10). The second term involves the confidence bound for the estimated projection matrix, ϕ_δ. This is critical in determining the tightness of the confidence set on θ_*. As discussed before, ϕ_δ reflects the difficulty of subspace recovery of the given problem and it depends on the underlying structure of the problem and SLB. This shows that as estimating the underlying subspace gets easier, having a projection-based approach in the construction of the confidence sets on θ_* provides tighter bounds.

In order to tolerate the possible difficulty of subspace recovery, PSLB also constructs $C_{d,t}$, which is the confidence set for θ_* without having subspace recovery. The construction of $C_{d,t}$ follows OFUL [3]. Let $Z_t = \hat{\Sigma}_{t-1} + \lambda I_d$. The algorithm tries to find $\hat{\theta}_t$ which is the ℓ^2-regularized least squares estimate of θ_* in the ambient space. Construction of $C_{d,t}$ is done under the same assumptions of Theorem 2.2.8, such that with probability at least $1-\delta$, θ_* lies in the set $C_{d,t} = \left\{\theta \in \mathbb{R}^d : \|\hat{\theta}_t - \theta\|_{Z_t} \leq \Omega_{t,\delta}\right\}$, where $\Omega_{t,\delta} = R\sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{tL^2}{m\lambda}\right)} + S\sqrt{\lambda}$. The search for an optimistic parameter vector happens in $C_{m,t} \cap C_{d,t}$. Notice that $\theta_* \in C_{m,t} \cap C_{d,t}$ with probability at least $1 - 5\delta$. Optimistically choosing the pair, $(\hat{X}_t, \tilde{\theta}_t)$, within the described confidence sets gives PSLB a way to tolerate the possibility of failure in recovering an underlying structure. If confidence set $C_{m,t}$ is loose or PSLB is not able to recover an underlying structure, then $C_{d,t}$ provides the useful confidence set to obtain desirable learning behavior.

Regret Analysis

PSLB uses the intersection of $C_{m,t}$ and $C_{d,t}$ as the confidence set at round t. Using only $C_{d,t}$ is equivalent to following OFUL and the regret analysis can be found in [3]. The regret analysis of using only the projected confidence set $C_{m,t}$ is the main contribution of this work. The following lemmas will be key in obtaining the regret analysis.

Lemma 2.2.9. *At round k, for any $\hat{x} \in D_k$, if $\nu \in C_k$, then $|(\hat{P}_k\hat{x})^T(\nu - \theta_k)| \leq \beta_{k,\delta}\|\hat{x}\|_{A_k^\dagger}$.*

Define $t_{r,\delta}$ such that $t_{r,\delta} = 1 + \left(\left(\frac{2m-1}{2m}\right)\frac{4L^2\Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}}+\sqrt{2L(\lambda_-+\sigma^2)\log\frac{m}{\delta}}}{\lambda_-+\sigma^2}\right)^2$.

Lemma 2.2.10. *For all $t \geq t_{w,\delta}$, with probability at least $1-\delta$,*

$$\lambda_m(\hat{P}_t\hat{\Sigma}_{t-1}\hat{P}_t) \geq (t-1)(\lambda_-+\sigma^2) - \sqrt{t-1}\left(4L^2\Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}} + \sqrt{2L(\lambda_-+\sigma^2)\log\frac{m}{\delta}}\right). \tag{2.13}$$

Also, for all $t \geq t_{r,\delta}$, with probability at least $1-\delta$,

$$\lambda_m(\hat{P}_t\hat{\Sigma}_{t-1}\hat{P}_t) \geq \frac{(\lambda_- + \sigma^2)}{2m}(t-1). \tag{2.14}$$

The proofs of Lemma 2.2.9 and 2.2.10 are in Supplementary Material A.1.4. The following theorem gives the regret upper bound for using only the projected confidence set $C_{m,t}$.

Theorem 2.2.11 (Regret Upper Bound of using only $C_{m,t}$)**.** *Fix any $\delta \in (0,1)$. Assume that for all $\hat{x}_{t,i} \in D_t$, $\hat{x}_{t,i}^T\theta_* \in [-1,1]$. Under Assumptions 1 and 2, $\forall t \geq 1$, with probability at least $1-6\delta$, the regret of using only $C_{m,t}$ satisfies*

$$R_{t,C_{m,t}} \leq \widetilde{O}\left(\left(1+\Gamma\sqrt{\frac{\alpha}{K}}\right)\left(\frac{\Gamma\sqrt{m\alpha}}{\sqrt{K}\sqrt{\lambda_-+\sigma^2}} + m\right)\sqrt{t}\right). \tag{2.15}$$

Proof. The instantaneous regret, $l_i = \hat{X}_i^{*T}\theta_* - \hat{X}_i^T\theta_*$, of the algorithm at ith round can be decomposed as follows:

$$\begin{aligned}
&\hat{X}_i^{*T}\theta_* - \hat{X}_i^T\theta_* \\
&\leq (\tilde{P}_i\hat{X}_i)^T\tilde{\theta}_i - (P\hat{X}_i)^T\theta_* \qquad (2.16)\\
&= \hat{X}_i^T(\tilde{P}_i - \hat{P}_i + \hat{P}_i)\tilde{\theta}_i - \hat{X}_i^T(\hat{P}_i + P - \hat{P}_i)\theta_* \\
&= (\hat{P}_i\hat{X}_i)^T(\tilde{\theta}_i - \theta_i) + (\hat{P}_i\hat{X}_i)^T(\theta_i - \theta_*) + ((\hat{P}_i - P)\hat{X}_i)^T\theta_* + ((\tilde{P}_i - \hat{P}_i)\hat{X}_i)^T\tilde{\theta}_i \\
&\leq 2\beta_{i,\delta}\|\hat{X}_i\|_{A_i^\dagger} + 2LS\|\hat{P}_i - P\|_2, \qquad (2.17)
\end{aligned}$$

where (2.16) follows since $(\tilde{P}_i, \hat{X}_i, \tilde{\theta}_i)$ is optimistic and (2.17) holds for all i with probability at least $1 - 4\delta$ due to Lemma 2.2.9 and Theorem 2.2.8. Combining this decomposition with the fact that $l_i \leq 2$, we get

$$\begin{aligned} l_i &\leq 2 \min\left(\beta_{i,\delta}\|\hat{X}_i\|_{A_i^\dagger} + LS\|\hat{P}_i - P\|_2, 1\right) \\ &\leq 2\beta_{i,\delta} \min(\|\hat{X}_i\|_{A_i^\dagger}, 1) + 2LS \min(\|\hat{P}_i - P\|_2, 1). \end{aligned} \tag{2.18}$$

Now we can provide an upper bound on the regret. For all $t \geq 1$, with probability at least $1 - 5\delta$,

$$\begin{aligned} R_t &\leq \sum_{i=1}^{t} 2\beta_{i,\delta} \min(\|\hat{X}_i\|_{A_i^\dagger}, 1) + 2LS \min(\|\hat{P}_i - P\|_2, 1) \\ &= 2LS \sum_{i=1}^{t} \min(\|\hat{P}_i - P\|_2, 1) + \sum_{i=1}^{t} 2\beta_{i,\delta} \min(\|\hat{X}_i\|_{A_i^\dagger}, 1) \\ &\leq 2LS \sum_{i=1}^{t} \min(\|\hat{P}_i - P\|_2, 1) + 2\beta_{t,\delta} \sum_{i=1}^{t} \min(\|\hat{X}_i\|_{A_i^\dagger}, 1) \qquad (2.19) \\ &\leq 2LS \sum_{i=1}^{t} \min(\|\hat{P}_i - P\|_2, 1) + 2\beta_{t,\delta} \sqrt{t \sum_{i=1}^{t} \min(\|\hat{X}_i\|^2_{A_i^\dagger}, 1)} \\ &\leq 2LS \sum_{i=1}^{t} \min(\|\hat{P}_i - P\|_2, 1) + 2\sqrt{t}\beta_{t,\delta} \sqrt{\sum_{i=1}^{t} \min\left(\lambda_{\max}(A_i^\dagger) L^2, 1\right)} \qquad (2.20) \\ &\leq 2LS \sum_{i=1}^{t} \min(\|\hat{P}_i - P\|_2, 1) + 2\sqrt{t}\beta_{t,\delta} \sqrt{\sum_{i=1}^{t} \min\left(\frac{L^2}{\lambda + \lambda_m(\hat{P}_i \hat{\Sigma}_{i-1} \hat{P}_i)}, 1\right)} \qquad (2.21) \\ &\leq 2LS\left(t_{w,\delta} + 2\Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}} \sum_{i=t_{w,\delta}}^{t} \frac{1}{\sqrt{i}}\right) + 2L\sqrt{t}\beta_{t,\delta} \sqrt{\frac{t_{r,\delta}}{\lambda} + \frac{2m}{\lambda_- + \sigma^2} \sum_{i=t_{r,\delta}}^{t} \frac{1}{i}}, \qquad (2.22) \end{aligned}$$

where (2.19) follows from the fact that $\beta_{1,\delta} \leq \cdots \leq \beta_{t,\delta}$. Since $\|x\|_M \leq \lambda_{\max}(M)\|x\|_2$, we get (2.20). The maximum eigenvalue of $A_t^\dagger$ is equivalent to mth eigenvalue of A_t, thus (2.21) is obtained. Recall that $\|\hat{P}_i - P\|_2 < 1$ for $t \geq t_{w,\delta}$. Using Lemma 2.2.4 and the second statement of Lemma 2.2.10 we get (2.22). Finally, Lemma

A.1.5 provides the following regret upper bound

$$R_t \leq 2LSt_{w,\delta} + 4LS\Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}}(2\sqrt{t} - 2\sqrt{t_{w,\delta}+1} + 1) \\ + 2L\sqrt{t}\beta_{t,\delta}\sqrt{\frac{t_{r,\delta}}{\lambda} + \frac{2m + 2m\log t - 2m\log(t_{r,\delta}+1)}{\lambda_- + \sigma^2}}. \quad (2.23)$$

Recall that $\beta_{t,\delta} = O\left(\Gamma\sqrt{\frac{\alpha m}{K}\log t} + \sqrt{m\log t}\right)$. Therefore, last term dominates the asymptotic upper bound on regret. Using the definition of $t_{r,\delta}$ and the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b > 0$, we get that the regret of the algorithm is

$$R_t = O\left(\frac{\sqrt{m}}{\lambda_- + \sigma^2}\left(\Gamma\sqrt{\frac{\alpha}{K}} + \frac{\alpha\Gamma^2}{K}\right)\sqrt{t\log t} + \frac{m}{\sqrt{\lambda_- + \sigma^2}}\left(1 + \Gamma\sqrt{\frac{\alpha}{K}}\right)\sqrt{t}\log t\right) \\ = \tilde{O}\left(\left(1 + \Gamma\sqrt{\frac{\alpha}{K}}\right)\left(\frac{\Gamma\sqrt{m\alpha}}{\sqrt{K}\sqrt{\lambda_- + \sigma^2}} + m\right)\sqrt{t}\right) = \widetilde{O}\left(\Upsilon\sqrt{t}\right).$$

□

Proof of Theorem 2.2.3: Using the intersection of $C_{m,t}$ and $C_{d,t}$ as the confidence set at round t, gives PSLB the ability to obtain the lowest possible instantaneous regret among both confidence sets. Therefore, the regret of PSLB is upper bounded by the minimum of the regret upper bounds on the individual strategies. Thus, Theorem 2.2.11 and Theorem 3 of Abbasi-Yadkori et al. [3] give the statement of Theorem 2.2.3.

Interpreting the Regret Bound

Υ is the reflection of the finite sample projection error at the beginning of the algorithm. It captures the difficulty of subspace recovery based on the structural properties of the problem and determines the regret of deploying projection-based methods in SLBs. Recall that α is the maximum of the effective dimensions of the true action vectors and the perturbation vectors. Depending on the structure of the problem, α can be $O(d)$, e.g., the perturbation can be uniform in all dimensions, which prevents the projection error from shrinking; thus, causes $\Upsilon = O(d\sqrt{m})$ resulting in $\widetilde{O}(d\sqrt{mt})$ regret. The eigengap within the true action vectors g_x and the eigengap between the true action vectors and the perturbation vectors g_ψ are critical factors that determine the identifiability of the hidden subspace. As σ^2 increases, the subspace recovery becomes harder since the effect of perturbation increases.

Conversely, as λ_- increases, the underlying subspace becomes easier to identify. These effects are significant and translate to the regret of PSLB via Γ in Υ.

Moreover, having finite samples to estimate the subspace affects the regret bound through Υ. Due to the nature of SLB, i.e., finite action vectors in decision sets, this is unavoidable. Note that if the decision set contained infinitely many actions, the subspace recovery would be accomplished perfectly. Thus, the problem would reduce to a m-dimensional SLB which has a regret upper bound of $\widetilde{O}(m\sqrt{t})$. This behavior can be seen in Υ. As $K \to \infty$, $\Upsilon = O(m)$ which gives the regret upper bound of $\widetilde{O}(m\sqrt{t})$ as expected.

Theorem 2.2.3 states that if the underlying structure is easily recoverable, e.g., $\Upsilon = O(m)$, then using PCA-based dimension reduction and construction of confidence sets provides substantially better regret upper bound for large d. If that is not the case, then due to the best-of-the-both-worlds approach provided by PSLB, the agent obtains the best possible regret upper bound. Note that the bound for using only $C_{m,t}$ is a worst-case bound, and as we present in Section 2.2.4, in practice PSLB can give significantly better results.

2.2.4 Experiments

Synthetic example: We study PSLB on 50 dimensional SLBs with 4-dimensional hidden subspace structure. At each round t, there are $K = 200$ actions in D_t. Each action is generated as $\hat{x}_{t,i} = x_{t,i} + \psi_{t,i}$. $\psi_{t,i} \in \mathbb{R}^d$ is drawn from Normal distribution but rejected if $\|\psi_{t,i}\|_2^2 > d_\psi$ for some d_ψ. We picked an orthonormal matrix $V \in \mathbb{R}^{50\times 4}$ and generate $x_{t,i}$ such that $x_{t,i} = V\epsilon$ where $\epsilon \sim uniform([-1, 1])^4$. For $T = 10,000$ rounds, we generate 3 different decision sets using $d_\psi = 1, 10$ and 20.

Figure 2.1(A) is a 2-D representation of the effect of increasing perturbation level, d_ψ. Assume that the underlying subspace, span(V), is the horizontal line segment. The blue, orange, and green data points represent $\hat{x}_{t,i}$ obtained by $d_\psi = 1, 10,$ and 20 respectively. As we increase the perturbation level, the hidden structure is concealed. Using these SLB settings, we studied the performance of PSLB and OFUL. Figure 2.1(B) provides the change in regrets as we increase the perturbation level from $d_\psi = 1$ to $d_\psi = 20$. Note that d_ψ can be interpreted as the effective dimension of the perturbation vectors. As d_ψ increases, the perturbations become more dominant, the subspace recovery becomes harder and PSLB loses its advantage of recovering the underlying subspace. The size of the $C_{m,t}$ increases and PSLB starts performing similar to OFUL. As suggested in the analysis through α, having d_ψ close to the

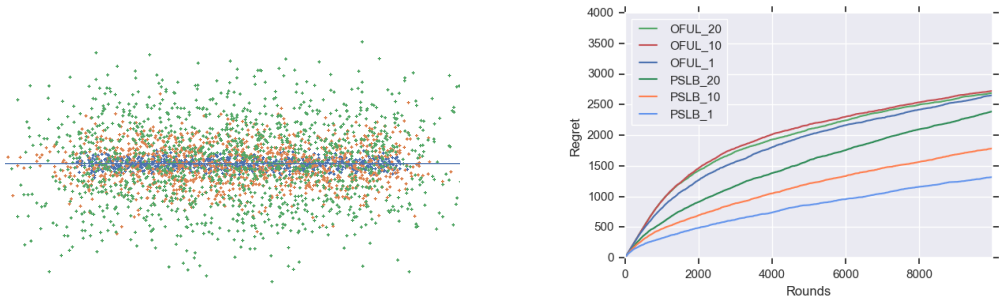

Figure 2.1: **(A)** 2-D representation of the effect of increasing perturbation level in concealing the underlying subspace. **(B)** Regrets of PSLB vs. OFUL under $d_\psi = 1, 10$ and 20. As the effect of perturbation increases, PSLB's performance approaches the performance of OFUL.

dimension of the ambient space leads to poor subspace recovery performance, thus resulting in higher regret in SLBs. This example demonstrates the overall effect of perturbation level on the subspace estimation, confidence set construction, and ultimately regret.

Image Classification in SLB Setting: In the image classification experiments, we study MNIST, CIFAR-10, and ImageNet datasets and use them to create the decision sets for the SLB setting. A simple 5-layer CNN, a pre-trained ResNet-18, and a pre-trained ResNet-50 are deployed respectively for MNIST, CIFAR-10, and ImageNet. Before training, we modify the architecture of the representation layer (the layer before the final layer) to make it suitable for the SLB study and obtain decision sets for each image.

Consider a standard network whose dimension of the representation layer is d. Therefore, the final layer for K class classification is fully connected and it is a $d \times K$ matrix that outputs K values to be used for classification. In this study, instead of having a final layer of $d \times K$ matrix, we construct the final layer as a d-dimensional vector and make the feature representation layer a Kd dimensional vector. We treat this vector as the concatenation of K contexts of d-dimensions i.e., $[\hat{x}_1, \ldots, \hat{x}_K]$. The final d-dimensional layer is θ_* of the SLB, where the logit for each class is computed as an inner product of the class context $\hat{x}_i$ and θ_*. We train these architectures for different d values using cross entropy loss.

Removing the final layer, the resulting trained networks are used to generate the feature representations of each image for each class which produces the decision sets at each time step of SLB. Since MNIST and CIFAR-10 have 10 classes, in each

decision set we obtain 10 action vectors where each of them are segments in the representation layer. On the other hand, from the ImageNet dataset we get 1000 actions per decision set due to 1000 classes in the datasets. In the SLB setting, the agent receives a reward of 1 if it chooses the right action, which is the segment in the representation layer corresponding to correct label according to trained network, and 0 otherwise. We apply both PSLB and OFUL on these SLBs. We measure the regret by counting the number of mistakes each algorithm makes.

Through computing PCA of the empirical covariance matrix of the action vectors, surprisingly we found that projecting action vectors onto the 1-dimensional subspace defined by the dominant eigenvector is sufficient for these datasets in the SLB setting. While surprising, a similar observation is obtained in [50] that the diffusion matrix which depends on the architecture, weights, and the dataset has a significantly low-rank structure for MNIST and CIFAR-10 datasets. Yet, in order to display the learning behavior, we present the regret obtained by PSLB and OFUL for the MNIST dataset with $d = 1000$ and $m = 8$, CIFAR-10 dataset with $d = 1000$ and $m = 2$ and ImageNet with $d = 100$ and $m = 8$ in Figure 2.2(a), (c), and (e) respectively.

With the help of subspace recovery and projection, PSLB provides a massive reduction in the dimensionality of the SLB problem and immediately estimates a fairly accurate model for θ_*. On the other hand, OFUL naively tries to sample from all dimensions in order to learn θ_*. This difference yields orders of magnitude improvement in regret in high-dimensional SLB problems. During the SLB experiment, we also sample the optimistic models that are chosen by PSLB and OFUL. We use these models to test the model accuracy of the algorithms, i.e., perform classification over the entire dataset. The model accuracy comparisons for the aforementioned experiment settings are depicted in Figure 2.2 (b), (d), (f). This portrays the learning behavior of PSLB and OFUL. Using projection, PSLB learns the underlying linear model in the first few rounds, whereas OFUL suffers from the high dimension of the SLB framework and lack of knowledge besides chosen action-reward pairs. An extensive study of PSLB and OFUL in these datasets with different subspace constructions is given in Lale et al. [159].

We would like to particularly highlight Figure 2.2 (e), (f). Since there are 1000 different classes in the dataset, the SLB framework synthesized from the ImageNet dataset has 1000 actions in each decision set. Therefore, even if $d = 100$ is not a fairly high-dimensional feature space, having 1000 actions makes the learning task harder. Thus, SLB algorithms are expected to have higher regrets and slower

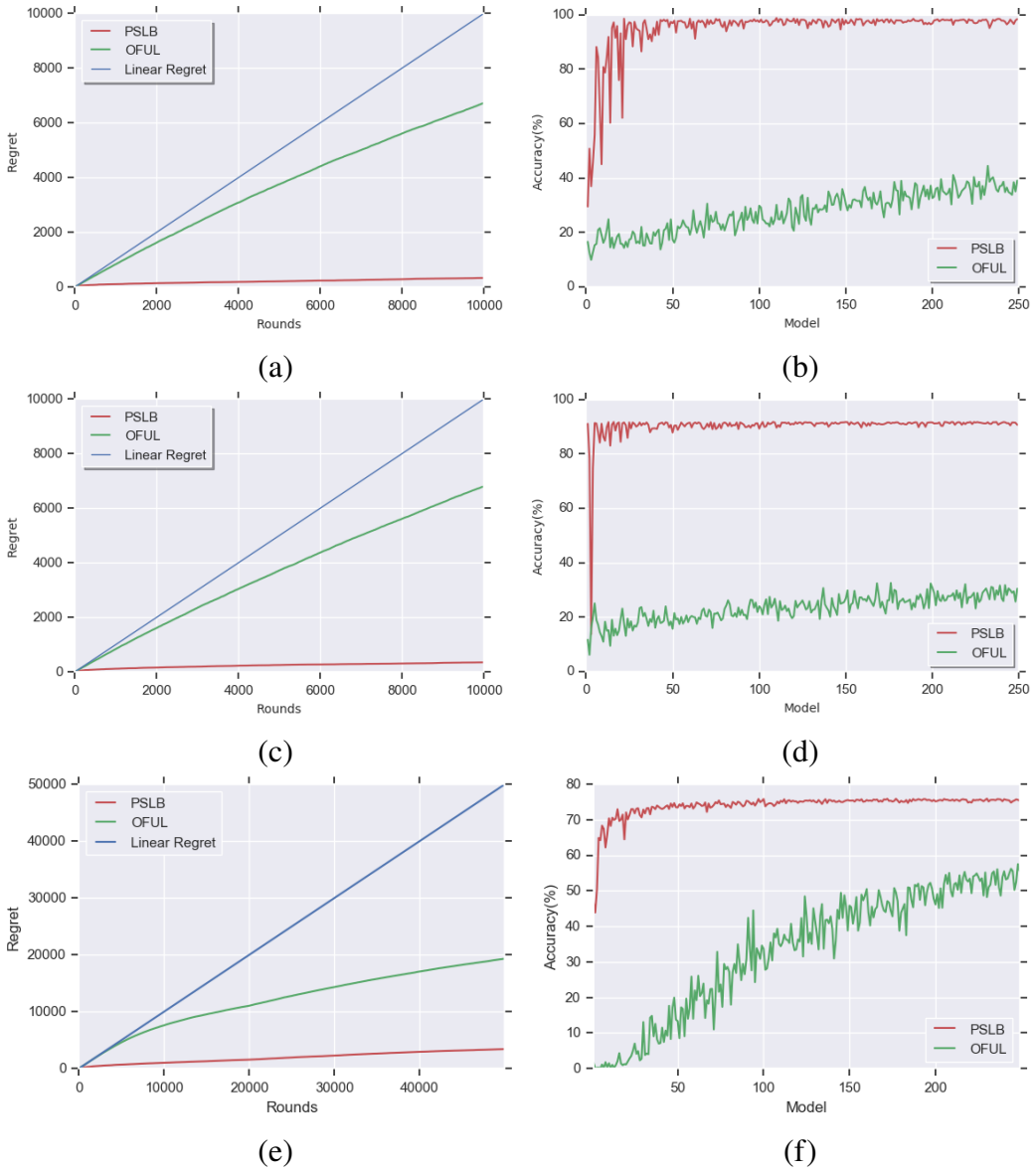

Figure 2.2: Regret of PSLB vs. OFUL in a stochastic linear bandit setting created using MNIST, CIFAR-10, and ImageNet datasets in (a), (c), and (e) respectively. Image classification accuracy of periodically sampled optimistic models of PSLB and OFUL in MNIST, CIFAR-10, and ImageNet datasets in (b), (d), and (f) respectively.

convergence to the underlying model. However, a large number of actions is key to having lower regret for PSLB. Instead of ignoring actions that are not chosen at the current round, PSLB uses them to get an idea about the structure of the action vectors. This setting clearly points out the advantage of PSLB over OFUL. While OFUL obtains linear regret in the beginning and struggles to construct a meaningful confidence set, PSLB uses hidden information in the massive number

of action vectors and reduces the dimensionality of the SLB framework. Then it exploits this information and converges to the accurate model without committing many mistakes.

2.3 Stochastic Linear Bandits with Unknown Safety Constraints

In this section, we study a novel stochastic linear bandit framework with unknown safety constraints where we model safety constraints as either multiple unknown affine or nonlinear functions, which generalizes the constraints considered in prior works. Given a set of actions, the goal of the decision-making agent is to maximize its reward by selecting safe actions defined by these constraints at every time step. Since the safety functions are unknown, the agent needs to learn them through feedback. To design a realistic feedback mechanism, we model these safety constraints locally in the action space and associate feedback sets for each constraint. The agent receives a noisy observation of the constraint function only when it picks actions from the corresponding constraint feedback set. This local feedback mechanism captures the feedback structure in many applications where choosing actions outside of a known safe set is subject to safety constraints.

We first consider the framework with multiple unknown affine safety constraints. For this setting, we propose two novel safe stochastic linear bandit algorithms. The first algorithm is the safe version of the linear upper confidence bound (UCB) algorithm [3] (OFUL): Safe-OFUL. In the design of Safe-OFUL, we decouple the exploration for learning the reward parameter and the safety constraints. This is in contrast to prior works which rely on the same exploration strategy for both reward and safety which fails in the affine safety function setting. The main technical challenge in the design of Safe-OFUL is to carefully prescribe an additional exploration within the UCB framework which guarantees the selection of optimistic safe actions and ensures sufficient exploration of the relevant constraint sets. For M distinct unknown affine constraints, we prove that Safe-OFUL attains $\tilde{O}(\sqrt{MT})$ regret after T time steps without violating any safety constraints.

We propose the safe version of the well-known Thompson Sampling algorithm [5] (LinTS): Safe-LinTS. Safe-LinTS is a computationally efficient alternative to Safe-OFUL which can possibly have computational challenges in finding optimistic actions similar to the other UCB algorithms. In the design of Safe-LinTS, unlike prior works, we also decouple the exploration for the reward parameter and safety functions such that the agent chooses the optimal action with respect to the sampled

Table 2.1: Comparison with prior works on safe stochastic linear bandit with $\tilde{O}(\sqrt{T})$ Regret. These works achieve this result using different methods for different safety aspects with different constraint types and for different numbers of constraints.

Work	Safety Aspect	Constraint Type	# of Constraints	Method
[140]	Reward	Cumulative	Single	UCB
[143]	Reward	Stage-wise - Linear	Single	UCB
[202]	Reward	Stage-wise - Linear	Single	UCB + TS
[216]	Policy	Stage-wise - Linear	Multiple	UCB
[12]	Action	Stage-wise - Linear	Single	UCB
[203]	Action	Stage-wise - Linear	1	TS
Our Work	**Action**	**Stage-wise - Affine/Nonlinear**	**Multiple**	**UCB + TS**

reward parameter from the estimated safe action sets at every time step.

The main technical challenge in the design of Safe-LinTS is to lower bound the probability of being optimistic for the sampled reward parameter while satisfying the safety constraints. To this end, we carefully design the sampling distributions for the reward parameter and safety functions such that the sampled parameters satisfy certain concentration and anti-concentration properties, and give a novel lower bound for this probability tailored for our stochastic linear bandit framework. For M distinct unknown affine constraints, we also show that Safe-LinTS attains $\tilde{O}(\sqrt{MT})$ regret.

Finally, we study the setting of multiple unknown nonlinear constraints. We extend Safe-OFUL and Safe-LinTS for this setting via a novel initial exploration strategy. We propose learning Taylor approximations of the underlying safety constraints and designing a new initial exploration phase that uses a priori known safe action per constraint to achieve uniform exploration, i.e., the persistence of excitation. We show that this exploration strategy allows the error in the estimates of the safety functions to be well-controlled and guarantees the identification of a safety set that contains the optimal safe action with high probability. We eventually show that the proposed method also attains $\tilde{O}(\sqrt{T})$ regret.

We empirically study all of these algorithms on both synthetic and real-world datasets. On the synthetic dataset with various safety constraints, we observe that both algorithms achieve sublinear regret without any safety violations, concurring with our theoretical results. We then modify a credit classification task on the German Credit dataset [141] into the loan approval stochastic linear bandit setting with two safety constraints by featurizing each individual as discussed in the case study above. We demonstrate the benefit of additional exploration in achieving improved regret while maintaining zero safety violations.

Our results subsume and generalize the state-of-the-art algorithms for stochastic linear bandit with stage-wise safe action constraints, see Table 2.1 for comparison.

2.3.1 Problem Formulation

Reward Model: At each time step t, the agent plays an action $x_t \in D_0$, where D_0 denotes the fixed decision set. Subsequently, the agent observes the reward $r_t = \mu^\top x_t + \eta_t^r$, where $\mu \in \mathbb{R}^d$ is unknown and η_t^r is random noise.

Safety Constraints: The environment is subjected to M distinct safety constraints, where $\mathbf{M} \coloneqq [M]$ is the index set of the constraints. We model these constraints as affine functions unknown to the agent (they will be modeled as nonlinear functions in Section 2.3.5). We consider localized safety constraints, where we define associated constraint feedback sets $\Gamma_i \subseteq D_0, \forall i \in \mathbf{M}$. At each time step, the agent needs to satisfy all the constraints corresponding to the feedback sets that the chosen action x_t belongs to. More precisely, if $x_t \in \Gamma_i$, the agent needs to have

$$\gamma_i^\top x_t + c_i \leq \tau, \quad \forall t, \tag{2.24}$$

for some γ_i and c_i are *unknown* and τ known to the agent $\forall i \in \mathbf{M}$. These constraints, therefore, form a region of safe actions $D_0^{\text{safe}} \subseteq D_0$, where

$$D_0^{\text{safe}} \coloneqq \bigcup\nolimits_{i \in M} \left\{ x \in \Gamma_i : \gamma_i^\top x_t + c_i \leq \tau \right\}. \tag{2.25}$$

We consider the setting where the agent is subject to hard constraints, i.e., the agent needs to play actions that belong to D_0^{safe} with high probability at all time steps. This safety constraint formulation captures many safety-critical real-world decision-making applications. Since the safety constraints are unknown to the agent, the agent needs to learn them via feedback and conservatively pick actions to ensure that the constraints are satisfied. In particular, we consider localized feedback such that the agent gets noisy observations of the constraint functions only when it picks actions from their corresponding constraint feedback set, i.e.,

$$\tilde{y}_t^i = \gamma_i^\top x_t + c_i + \eta_t^i \quad \text{if} \quad x_t \in \Gamma_i. \tag{2.26}$$

Figure 2.3 illustrates an example safety constraint structure.

Regret: We study the (pseudo) regret of the agent

$$R_T = \sum\nolimits_{t=1}^{t=T} \mu^\top x^* - \mu^\top x_t,$$

for T time steps, where $x^* = \arg\max_{x \in D_0^{\text{safe}}} \mu^\top x$, i.e., the optimal safe action. The goal of the agent is to minimize the regret over time and achieve sublinear regret

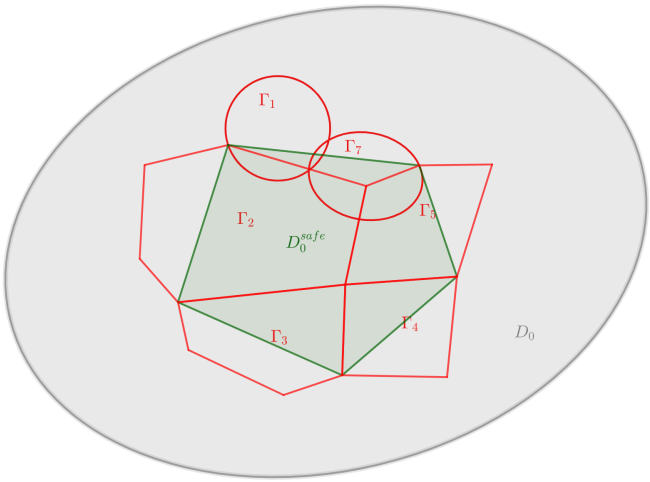

Figure 2.3: Illustration of the safety constraints: D_0 represents all actions. Γ_i represents the constraint feedback regions where the affine constraints (2.24) need to be satisfied. D_0^{safe} is the safe set of actions formed by the union of the safe regions from each Γ_i.

while satisfying the safety constraints at all time steps. Let F_t denote the σ-algebra (history) up to time t, such that x_t is F_{t-1} measurable and the noise terms, i.e., η_t^r and $\eta_t^{i_t}$, are F_t measurable. Before describing our first algorithm for this setting, we adopt some technical assumptions, which are standard in the literature, [3, 5, 12, 143].

Assumption 2.3.1 (Subgaussian Noise)**.** *For all $t \in [T]$ and $i \in \mathbf{M}$, η_t^r,η_t^i are conditionally R-sub-Gaussian where $R \geq 0$ is a fixed constant, i.e., $\forall \lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda \eta_t^r} | F_{t-1}] \leq \exp(\lambda^2 R^2/2), \mathbb{E}[e^{\lambda \eta_t^i} | F_{t-1}] \leq \exp(\lambda^2 R^2/2)$.*

Assumption 2.3.2 (Boundedness)**.** *$s < \|\mu\|_2, \|\gamma_i\|_2 < S, \|x\|_2 < L, \|x - x_i^s\|_2 < L^c \leq L$, if $x \in \Gamma_i$ and $\mu^\top x \in [-1, 1], \forall x \in D_0$ for some $s, S, L, L^c > 0$.*

Assumption 2.3.3 (Known safe actions)**.** *For every constraint $i \in \mathbf{M}$, the agent knows a safe action $x_i^s \in \Gamma_i$ such that $x_i^s \in D_0^{safe}$ and $\gamma_i^\top x_i^s + c_i = \tau_i^s < \tau$ where τ, τ_i^s are known.*

Note that Assumption 2.3.3 holds in many real-world decision-making tasks such as robotics and clinical trials where there are known safe actions. Note that the known safe actions do not need to be unique. If τ_i^ss are unknown, the agent can sample the known safe actions to estimate the values of τ_i^ss.

2.3.2 Safe-OFUL

In this section, we propose Safe-OFUL, the safe version of the well-known linear upper confidence bound algorithm studied in the literature [3], (also named OFUL as discussed in Section 2.2). Similar to OFUL, Safe-OFUL deploys the OFU principle to balance the exploration vs. exploitation trade-off. This algorithmic approach proposes constructing confidence sets for the underlying parameter μ using the history of actions and rewards and playing the optimal action for the most optimistic model within these sets. However, unlike the unconstrained setting of OFUL, the agent in our stochastic linear bandit framework needs to satisfy the unknown safety constraints at every time step.

To address this, Safe-OFUL conservatively explores starting around the known safe actions to learn the safety constraints as well as the underlying reward parameter while avoiding safety violations. During the course of interaction, besides the confidence set for the underlying reward parameter μ, it forms confidence sets for the unknown safety functions, i.e., affine parameters γ_i, and includes this information to safely expand its estimate of the safety set D_0^{safe}. In deploying the OFU principle, it includes an additional exploration to tolerate the uncertainty in the safety set estimate which enforces the algorithm to pick conservatively to avoid safety violations. Safe-OFUL is given in Algorithm 2. Safe-OFUL consists of 3 key elements: parameter estimation, safety construction, and acting optimistically.

Parameter Estimation: At each time step t, Safe-OFUL uses the history of action-reward pairs to obtain a ℓ_2-regularized (for some $\lambda > 0$) least squares (RLS) estimate of the underlying reward parameter μ via

$$\hat{\mu}_t = V_t^{-1} \sum_{k=1}^{t-1} r_k x_k, \tag{2.27}$$

where $V_t = \lambda I + \sum_{k=1}^{t-1} x_k x_k^\top$. Safe-OFUL then builds a confidence set around this RLS estimate

$$C_t = \left\{ v \in \mathbb{R}^d : \|v - \hat{\mu}_t\|_{V_t} \leq \beta_t \right\}, \tag{2.28}$$

where $\beta_t = R\sqrt{d \log\left((1+(t-1)L^2/\lambda)/\delta\right)} + \sqrt{\lambda} S$, for $\delta \in (0,1)$. The choice of β_t follows from Theorem 2 of Abbasi-Yadkori et al. [3], such that $\mu \in C_t$ with probability at least $1-\delta$, for all $t > 0$. Thus, Safe-OFUL guarantees that the event $\mathcal{E}_\mu = \{\mu \in C_t\}$ holds with high probability.

Similarly, Safe-OFUL estimates the unknown safety functions, i.e., parameters γ_i for all $i \in \mathbf{M}$, via RLS as

$$\hat{\gamma}_{i,t} = A_{i,t}^{-1} \sum_{k=1}^{N_i(t)} y_k^i (x_k - x_i^s), \tag{2.29}$$

Algorithm 2 Safe-OFUL

Input: $\tau_i^s, x_i^s, \tau, L, S, R$
for t=1,. . . ,T **do**
  Compute $\hat{\mu}_t$ via (2.27) & $\hat{\gamma}_{i,t}$ via (2.29)
  Construct β_t in (2.28) & β_t^i $\forall i \in \mathbf{M}$ in (2.30)
  Construct D_t^{safe} according to (2.31)
  Find $x_t = \text{argmax}_{i \in \mathbf{M}, x \in \hat{\Gamma}_{i,t}} \mathbf{ucb}(x, i, t-1)$ via (2.32)
  Play x_t and observe reward r_t

for $y_t^i = \tilde{y}_t^i - \tau_i^s, \forall t$, where $A_{i,t} = \lambda I + \sum_{k=1}^{N_i(t)} (x_k - x_i^s)(x_k - x_i^s)^\top$ and $N_i(t)$ denotes the number of times the agent has gotten feedback from the constraint set Γ_i until time t. It also builds confidence sets around these estimates

$$C_t^i = \left\{ v \in \mathbb{R}^d : \left\| v - \hat{\gamma}_{i,t} \right\|_{A_t^i} \leq \beta_t^i \right\}, \tag{2.30}$$

with $\beta_t^i = R\sqrt{d \log\left(|M|(1 + N_i(t)L^2/\lambda)/\delta\right)} + \lambda^{1/2} S_\gamma$, such that the event $\mathcal{E}_{\gamma_i} = \{\gamma_i \in C_t^i\}$ holds with probability at least $1 - \delta$, for all $t > 0$ and $i \in \mathbf{M}$.

Safety Construction: Next, conditioned in the joint event $\mathcal{E} \coloneqq \mathcal{E}_\mu \cup \bigcup_{i \in M} \mathcal{E}_{\gamma_i}$, Safe-OFUL aims to satisfy the unknown safety constraints when picking actions. To achieve this, it conservatively constructs a safe set of actions $\hat{\Gamma}_{i,t} \coloneqq \{x \in \Gamma_i : \hat{\gamma}_{i,t}^\top (x - x_i^s) + \beta_t^i ||x - x_i^s||_{A_{i,t}^{-1}} \leq \tau - \tau_i^s\}$, where

$$D_t^{\text{safe}} = \bigcup\nolimits_{i \in M} \hat{\Gamma}_{i,t}. \tag{2.31}$$

For this constructed safety set, we have the following result.

Lemma 2.3.4. *Conditioned on $\mathcal{E}$, $D_t^{safe} \subseteq D_0^{safe}$, for all $t > 0$.*

The proof is given in Appendix A.2.1, where we show that conditioned on $\mathcal{E}$, $\hat{\gamma}_{i,t}^\top (x - x_i^s) + \beta_t^i ||x - x_i^s||_{A_{i,t}^{-1}}$ is an upper bound on $\gamma_i^\top (x - x_i^s)$, $\forall i \in M$. This ensures that D_t^{safe} is a conservative estimate of D_0^{safe}, such that Safe-OFUL satisfies the safety constraints with high probability.

Acting Optimistically: At the final step, Safe-OFUL picks an action x_t from the constructed safe set D_t^{safe} which maximizes the Upper Confidence Bound (**ucb**) defined as

$$\mathbf{ucb}(x, i, t) = \hat{\mu}_t^\top x + \beta_t ||x||_{V_t^{-1}} + k_i \beta_t^i ||x - x_i^s||_{A_{i,t}^{-1}}, \tag{2.32}$$

$\forall i \in \mathbf{M}$, where $k_i \geq 2LS/(\tau - \tau_i^s)$. In the following, we show that this construction of **ucb** ensures sufficient exploration of the safety constraint set in order to balance exploration vs. exploitation via optimistic action selection.

2.3.3 Theoretical guarantees for Safe-OFUL

Before presenting the theoretical guarantees, we place the following technical assumption on the safety feedback sets that the optimal safe action belongs to, denoted as Γ_{i^*}.

Assumption 2.3.5 (Star convex optimal constraint sets)**.** *Γ_{i^*} is star convex around the safe known action $x^s_{i^*}$ such that the convex combination $\alpha x^* + (1-\alpha)x^s_i \in \Gamma_{i^*}, \forall \alpha \in [0,1]$.*

Note that since the constraint sets are localized around a particular safe action x^s_i, this assumption is reasonable in the safe stochastic linear bandit framework, and weaker than the prior work, e.g., [13], where the entire space of actions is considered to be star convex. In the regret analysis of Safe-OFUL, we follow the standard analysis of UCB and decompose the regret R_T into two terms: (i) $\sum_{t=1}^{t=T}(\mu^\top x^* - \mathbf{ucb}(x_t, i_t, t))$ and (ii) $\sum_{t=1}^{t=T}(\mathbf{ucb}(x_t, i_t, t) - \mu^\top x_t)$.

In the unconstrained setting, the optimism principle is satisfied by construction, since the optimal action belongs to the decision set D_0, yielding **(i)** to be non-positive. However, in the safe stochastic linear bandit framework, the optimal safe action x^* may not belong to the constructed safe set D^{safe}_t where optimistic action selection happens. Thus, we first show that the new construction of **ucb** in (2.32) still provides optimistic actions.

Theorem 2.3.6 (Optimism)**.** *For all $i \in \mathbf{M}$, setting $k_i \geq 2LS/(\tau - \tau^s_i)$ guarantees optimism with high probability:*

$$\max_{i \in M, x \in \hat{\Gamma}_{i,t}} \boldsymbol{ucb}(x, i, t) \geq \mu^\top x^* \quad \forall t.$$

The proof is given in Appendix A.2.1. To sketch the proof idea, we consider two cases of whether $x^* \in D^{\text{safe}}_t$ or not. If yes, via standard UCB arguments, we guarantee that Safe-OFUL selects optimistic actions. If not, we show that the additional exploration bonus $k_i\beta^i_t||x - x^s_i||_{A^{i-1}_t}$ ensures optimistic action selection for the given choice of k_i. This shows that adjusting the additional exploration bonus around the known safe actions ensures that the relevant constraint regions are well-explored, i.e., x^* eventually belongs to D^{safe}_t.

The choice of k_i highlights the key challenge in our proposed stochastic linear bandit framework. In contrast to prior works, the agent gets feedback from a constraint only if it plays an action within the associated feedback set. Therefore, while aiming to

learn the underlying reward function, Safe-OFUL needs to cautiously choose actions from the constraint sets where it wants to learn the constraints at the cost of not receiving any feedback from the non-active constraints. The new **ucb** term in (2.32) captures this trade-off and selecting k_i as in Theorem 2.3.6 balances it effectively. In particular, we see that this exploration bonus is inversely proportional to the gap between the safety threshold and the value of the known safe action. Intuitively, this means that if the known safe action is close to violation, Safe-OFUL needs to explore more/act more optimistically to learn the optimal safe action. We pay an extra price in regret due to this additional effort.

Theorem 2.3.7 (Regret Bound)**.** *Suppose Assumptions 2.3.1–2.3.3 and 2.3.5 hold. Then for any* $\delta \in (0,1)$ *and* $k_i = 2LS/(\tau - \tau_i^s)$*, with probability at least* $1-2\delta$*, the regret of Safe-OFUL is* $R_T \leq R_\mu + R_\gamma$*, where* $R_\mu = 2\beta_T\sqrt{2Td\log((1+TL^2/(d\lambda))/\delta)}$ *and* $R_\gamma = (k_{i_{max}}\beta_T^{i_{max}} + 2)\sqrt{2|M|Td\log((1+TL^2/(d\lambda))/\delta)}$*, for* $\beta_T^{i_{max}} = \max_{j\in\mathbf{M}}\beta_T^j$ *and* $k_{i_{max}} = \max_{j\in\mathbf{M}} k_j$.

The proof is given in Appendix A.2.1. In the proof, since **(i)** is non-positive via Theorem 2.3.6, we study **(ii)** and decompose it into 2 terms. R_μ results from learning the unknown reward parameter and R_γ is due to learning M different constraints. Notice that R_γ scales with the hardest, i.e., the most exploration needed, constraint through $\beta_T^{i_{max}}$ and $k_{i_{max}}$. Moreover, the regret rate of Safe-OFUL matches the prior unconstrained UCB results [3] and single linear constrained UCB results [12, 216], where the additional price of learning under M distinct constraints with local safety feedback, which generalizes the prior work, appears as $\sqrt{M}$.

2.3.4 Safe-LinTS

In many scenarios, solving the bilinear optimization problem of UCB-type algorithms, i.e., Line 6 of Algorithm 2, can be computationally challenging. To this end, Thompson Sampling (TS)-based methods are proposed, e.g., LinTS [5, 10]. These approaches sample a model within the constructed confidence set of plausible models and find the optimal action with respect to this sampled model. Therefore, they consider a linear optimization problem for decision-making, which can be solved efficiently. Because of this computational efficiency, simplicity, and possibly better empirical performance, they are adopted in many decision-making scenarios [7, 137]. In this section, we propose Safe-LinTS, the safe version of LinTS. The pseudocode of Safe-LinTS is given in Algorithm 3.

Algorithm 3 Safe-LinTS

1: **Input:** $\tau_i^s, x_i^s, \tau, L, S, R, \mathcal{P}^{TS}, \mathcal{P}_c^{TS}$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Compute $\hat{\mu}_t$ via (2.27) & $\hat{\gamma}_{i,t}$ via (2.29)
4: Construct β_t in (2.28) & β_t^i $\forall i \in M$ in (2.30)
5: Construct D_t^{safe} according to (2.31)
6: Sample $\eta_t \sim \mathcal{P}^{TS}$ and $\eta_t^c \sim \mathcal{P}_c^{TS}$
7: Compute $\tilde{\mu}_t = \hat{\mu}_t + \beta_t V_t^{-1/2} \eta_t$
8: Compute $\tilde{\omega}_{i,t} = \beta_t^i A_{i,t}^{-1/2} \eta_t^c, \forall i \in \mathbf{M}$
9: Find $x_t = \text{argmax}_{i \in \mathbf{M}, x \in \hat{\Gamma}_{i,t}} \tilde{\mu}_t^\top x + \tilde{\omega}_{i,t} \top (x - x_i^s)$
10: Play x_t and observe reward r_t

The construction of Safe-LinTS follows similarly to Safe-OFUL regarding the estimation of the reward parameter and safety parameters and safety construction (Lines 3–5). After this, it draws two random perturbations $\eta_t \in \mathbb{R}^d$ and $\eta_t^c \in \mathbb{R}^d$ from i.i.d. distributions $\mathcal{P}^{TS}$ and $\mathcal{P}_c^{TS}$ respectively (will be characterized shortly). Among these perturbations, while Safe-LinTS uses η_t in a standard way to sample a reward parameter, it uses η_t^c in a novel way to expand the estimated safe set to satisfy optimistic action selection.

The main novelty in the design of Safe-LinTS lies in this decoupling of the exploration for the reward parameter and the safety functions. In particular, the prior work in safe linear bandits [203] relies on using the same Gram matrix to learn both the safety and reward parameters simultaneously. However, learning in the affine setting involves separate Gram matrices, thus, Safe-LinTS explicitly balances the exploration trade-off between learning the unknown reward parameter and the safety parameters, ensuring safety and optimism for the entire horizon.

To this end, $\mathcal{P}^{TS}$ and $\mathcal{P}_c^{TS}$ are chosen to satisfy certain concentration and anti-concentration properties. In particular, for some $\delta \in (0, 1)$ and constants c, c', Safe-LinTS selects $\mathcal{P}^{TS}$ such that $\mathbb{P}(\|\eta_t\|_2 \leq \sqrt{cd \log(c'd/\delta)}) \geq 1 - \frac{\delta}{2}$, and $\mathbb{P}(u^\top \eta_t \geq 1) = p_1 > 0$, for any $u \in \mathbb{R}^d$ with $\|u\| = 1$. Similarly, $\mathcal{P}_c^{TS}$ is chosen such that $\mathbb{P}(\|\eta_t^c\|_2 \leq \frac{2LS^\gamma}{\tau - \tau_*^s} \sqrt{cd \log(c'd/\delta)})) \geq 1 - \frac{\delta}{2}$ and $\mathbb{P}(u^\top \eta_t^c \geq \frac{2LS^\gamma}{\tau - \tau_*^s}) = p_2 > 0$, where $S^\gamma > \max_{i \in \mathbf{M}} \|\gamma_i\|$ and $\tau_*^s = \max_{i \in \mathbf{M}} \tau_i^s$. These requirements imply that these distributions with high probability should concentrate, yet, still provide a certain amount of exploration (anti-concentration), which is crucial in achieving low regret. Natural candidates for $\mathcal{P}^{TS}$ and $\mathcal{P}_c^{TS}$ are $\mathcal{N}(0, I)$ and $\mathcal{N}(0, \frac{2LS^\gamma}{\tau - \tau_*^s} I)$, respectively.

Safe-LinTS uses $\eta_t \sim \mathcal{P}^{TS}$ to sample $\tilde{\mu}_t$ around $\hat{\mu}_t$ which provides the balance

between exploration and exploitation while learning the unknown reward parameter, i.e., $\tilde{\mu}_t = \hat{\mu}_t + \beta_t V_t^{-1/2}\eta_t$. It then uses $\eta_t^c \sim \mathcal{P}_c^{TS}$ to sample $\tilde{\omega}_{i,t} = \beta_t^i A_{i,t}^{-1/2}\eta_t^c, \forall i \in \mathbf{M}$, which will be used to provide the exploration needed to expand the estimated safe set to include higher rewarding actions, i.e., optimistic actions.

At the final step, Safe-LinTS picks an action x_t from D_t^{safe} by maximizing $\tilde{\mu}_t^\top x + \tilde{\omega}_{i,t}^\top (x - x_i^s)$. Note that this is a linear objective with transparent exploration goals. In particular, the reward exploration is similar to LinTS in Abeille and Lazaric [5], whereas the second term adds exploration along the safety constraints using the known safe actions. Notice that this approach generalizes the algorithm proposed in [203] whose setting is a special case of the SLB framework considered in this study.

Theorem 2.3.8. *Suppose Assumptions 2.3.1–2.3.3 and 2.3.5 hold. Then for any* $\delta \in (0, 1)$*, with probability at least* $1–\delta$*, the regret of Safe-LinTS is* $R_T = \tilde{O}(d^{3/2}\sqrt{|M|T})$.

The proof and the exact expressions are given in Appendix A.2.2. In the proof, we first show that Safe-LinTS selects safe (via Lemma 2.3.4) optimistic actions with at least $p_1 p_2$ probability by showing that η_t and η_t^c provide sufficient exploration. Finally, we use the regret decomposition in [5] to give the regret upper bound. Notably, this result matches the regret upper bound in [203] for their setting. In the exact regret expression, the leading term has $1/(p_1 p_2)$ i.e., the inverse of optimistic action probability. This relation is similar to that of $k_{i_{max}}$ in Safe-OFUL. In particular, similar to Safe-OFUL, for a smaller worst-case safety gap of known safe actions, Safe-LinTS needs to explore more to learn the optimal safe action which results in increased regret through p_2.

2.3.5 Linear Bandits with Nonlinear Constraints

In this section, we consider the most general setting of multiple nonlinear safety constraints, which captures the most diverse class of decision-making scenarios.

Safety Constraints: The environment is subject to M distinct nonlinear safety constraints, such that if $x_t \in \Gamma_i$, the agent needs to have $f_i(x_t) \leq \tau, \forall t$ for some unknown f_i and known $\tau, \forall i \in \mathbf{M}$. The region of safe actions corresponds to $D_0^{\text{safe}} := \bigcup_{i \in \mathbf{M}} \{x \in \Gamma_i : f_i(x) \leq \tau\}$. Similar to the affine case, we consider localized feedback for the agent: $\tilde{y}_t^i = f_i(x_t) + \eta_t^i$ if $x_t \in \Gamma_i$. Moreover, without loss of generality, we consider $\Gamma_i = \{x \in \mathbb{R}^d : \|x - x_i^s\|_2 \leq \delta_f\}$ for some $\delta_f > 0$ for all $i \in \mathbf{M}$. In parallel with Assumption 2.3.3, we assume that the agent knows a safe action for each

constraint, $x_i^s \in D_0^{\text{safe}}, \forall i \in \mathbf{M}$, as well as their safety values $f_i(x_i^s) = \tau_i^s < \tau$. Finally, we adopt the following simple regularity assumption on the nonlinear constraints.

Assumption 2.3.9 (Smooth & Lipschitz Safety Constraints)**.** *$f_i(x)$ is ζ-smooth and S-Lipschitz, $\forall x \in \Gamma_i$, $\forall i \in \mathbf{M}$.*

The local smoothness assumption is a significantly weak assumption [24], while the local Lipschitzness is the nonlinear analog to Assumption 2.3.2 with affine constraints. The setting characterized above subsumes and generalizes the affine case in Section 2.3.1. Using the first-order Taylor expansion about the known safe actions, we obtain $f_i(x) = f_i(x_i^s) + \nabla f_i(x_i^s)^\top (x - x_i^s) + \epsilon_i(x)$, where $\epsilon_i(x)$ represents the remainder terms. Notice that for small enough δ_f, this expansion behaves very similarly to affine functions studied in previous sections, which motivates the following our algorithm design. To avoid any further structural assumptions and keep the setting as general as possible, while keeping the problem tractable, we assume some safety gap for the optimal safe action to account for the function approximation errors.

Assumption 2.3.10 (Safety gap for optimal action)**.** *The optimal safe action x^* has at least Δ safety gap from constraint boundary, i.e., $f_{i^*}(x^*) \leq \tau - \Delta$, such that $\Delta > \zeta \delta_f^2$.*

This is a mild assumption since for a nonlinear function the optimal action need not be at the boundary, unlike linear constraints. Moreover, this assumption holds in many safe decision-making tasks, where the optimal safe action might be a significantly safe one, yet, to learn this action one might need to consider a higher threshold in the learning process.

Safe-OFUL/LinTS with Pure Exploration

We propose an extension of our prior algorithms to achieve safe and effective decision-making for the SLB with multiple nonlinear safety constraints. Due to Assumption 2.3.9, we know that there exists a safe ball of actions around each $x_i^s, \forall i \in \mathbf{M}$, i.e., $f_i(x_t) \leq \tau$ if $x_t \in \{x \in \Gamma_i : ||x - x_i^s||_2 < \delta_r\}$ for $\delta_r \leq (\tau - \tau_i^s)/(S + \zeta \delta_f)$. The existence of this ball helps the agent to estimate the gradient of the nonlinear function around the known safe actions x_i^s. The main idea in our algorithm design is to learn the first-order function approximation in each Γ_i while taking into account the estimation error so that the agent can eventually get to the optimal action x^*

Algorithm 4 Safe-LinUCB/LinTS with Pure Exploration

Input: $\tau_i^s, x_i^s, \tau, \zeta, S, \Delta, \delta_f$
for $i \in \mathbf{M}$ **do**
  for $t = 1, 2, \ldots, T'$ **do**
    Play $x_t = \arg\max_{x \in D_i^w} \|x - x_i^s\|_{A_{i,t}^{-1}}$
Construct $D_{MT'}^{\text{safe}}$
Run Safe-LinUCB/LinTS for the remainder with $D_{MT'}^{\text{safe}}$

without violating safety. The algorithm consists of two phases: (i) Pure Exploration and (ii) Safe-LinUCB/LinTS. The pseudocode is given in Algorithm 4.

Pure Exploration: In this phase, the agent samples T' actions from each constraint set Γ_i. It uniformly excites all the directions by playing $x_t = \arg\max_{x \in D_i^w} \|x - x_i^s\|_{A_{i,t}^{-1}}$ for T' steps, where D_i^w is the $d-1$ dimensional boundary surface of the δ_r-ball around the known safe actions x_i^s defined as $D_i^w = \{x \in \Gamma_i : ||x - x_i^s||_2 = \delta_r\}$, and $A_{i,t} = \lambda I + \sum_{k=1}^{N_i(t)} (x_k - x_i^s)(x_k - x_i^s)^\top$. By construction of D_i^w, the agent achieves safe exploration. Moreover, this exploration strategy ensures that the agent always picks the direction of the smallest eigenvalue, resulting in persistent excitation in all directions since actions in D_i^w have the same norm.

At the end of this phase, the algorithm estimates the gradient of the constraint functions using RLS such that $\nabla \hat{f}_{it} = A_{i,t}^{-1} \sum_{k=1}^{N_i(t)} y_k^i (x_k - x_i^s)$ for $y_t^i = \tilde{y}_t^i - \tau_i^s, \forall t$. Note that $N_i(t)$ is equal to T' for all i at the end of this phase. Next, the algorithm further decomposes

$$\nabla \hat{f}_{it} = \nabla \hat{f}_{it}^{LS} + \hat{\epsilon}_{it},$$

where $\nabla \hat{f}_{it}^{LS} = A_{i,t}^{-1} \sum_{\tau=1}^{N_i(t)} (x_\tau - x_i^s)(\nabla f(x_i^s)^\top (x_\tau - x_i^s) + \eta_t^i)$ and $\hat{\epsilon}_{it} = A_{i,t}^{-1} \sum_{\tau=1}^{N_i(t)} (x_\tau - x_i^s)\epsilon_i(x_\tau)$. Notice that the expression for $\nabla \hat{f}_{it}^{LS}$ is the nonlinear analog of (2.29). Thus, the algorithm builds confidence sets around the estimates $\nabla \hat{f}_{it}^{LS}, \forall i \in \mathbf{M}$:

$$C_t^i = \{v \in \mathbb{R}^d : \|v - \nabla \hat{f}_{it}^{LS}\|_{A_t^i} \leq \beta_t^i\},$$

with $\beta_t^i = R\sqrt{d \log\left(|M|(1 + T'L^2/\lambda)/\delta\right)} + \lambda^{1/2} S$. It also defines the event $\mathcal{E}_{\nabla f_i} = \{\nabla f_i(x_i^s) \in C_t^i\}$ which holds with probability at least $1 - \delta$, for all $t > 0$ and $i \in \mathbf{M}$.

Safety Construction: Next, conditioned in the joint event $\mathcal{E}_{\nabla f_t} := \bigcup_{i \in M} \mathcal{E}_{\nabla f_{it}}$, the algorithm aims to satisfy safety constraints when picking actions. To achieve this, it conservatively constructs a safe set of actions $\hat{\Gamma}_t^i = \{x \in \Gamma_i : \nabla \hat{f}_{it}^\top (x - x_i^s) + \frac{\Delta}{2} \leq \tau - \tau_i^s\}$, where $D_{MT}^{\text{safe}} = \bigcup_{i \in M} \hat{\Gamma}_{i,t}$.

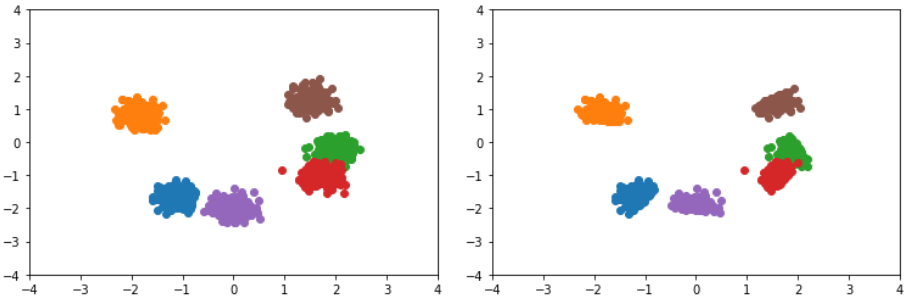

Figure 2.4: D_0 and D_0^{safe} respectively for affine constraints. Different colors represent different feedback sets Γ_i.

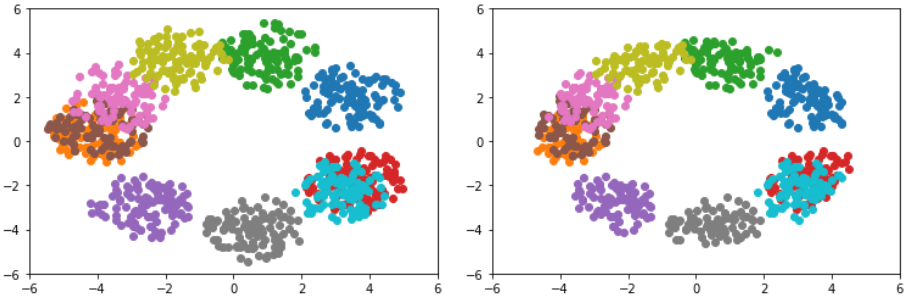

Figure 2.5: D_0 and D_0^{safe} respectively for nonlinear (ℓ_2 norm bound) constraints. Different colors represent different feedback sets Γ_i.

Theorem 2.3.11. *Suppose Assumptions 2.3.9 & 2.3.10 hold. For any $\delta \in (0,1)$, after T' time steps of pure exploration per constraint set, we have i) $x^* \in D_{MT'}^{safe}$ and ii) $D_{MT'}^{safe} \subseteq D_0^{safe}$ with probability at least $1-\delta$, if $\frac{T'}{\log^2 T'} \geq \left(2d\frac{4\delta_f^2}{(\Delta-\zeta\delta_f^2)^2}\right)^2$.*

The proof is in Appendix A.2.3. The main idea of the proof is to show that we can control the error from non-linearity using smoothness and simultaneously learn the gradient at that point by uniformly playing actions around and close to the known safe actions. We then build $D_{|M|T'}^{\text{safe}}$ using $\hat{\nabla} f_i(x_i^s)$ and add error margin to compensate for smoothness approximation error, away from x_i^s. After this phase, the agent executes the previously proposed algorithms using $D_{|M|T'}^{\text{safe}}$.

Corollary 2.3.12 (Regret Bound)**.** *Suppose 2.3.9 & 2.3.10 hold. Then for the given duration of T' in Theorem 2.3.11, for any $\delta \in (0,1)$, with probability at least $1-2\delta$, the regret of Algorithm 4 the above algorithm is $\tilde{O}(|M|T' + \sqrt{T})$.*

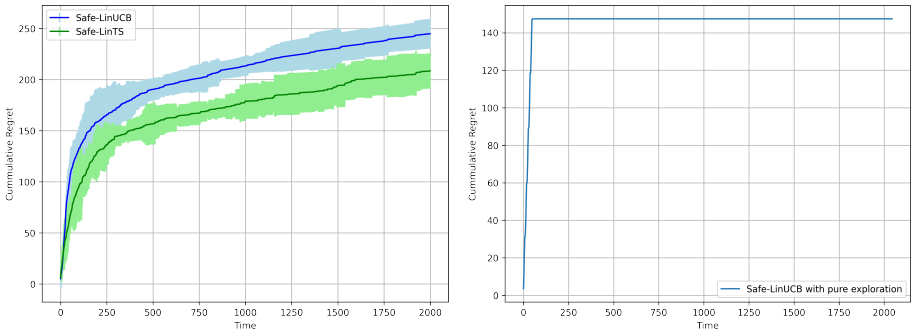

Figure 2.6: **Left:** Cumulative regret of Safe-OFUL (Safe-LinUCB) and Safe-LinTS for the setting in Figure 2.4 (Solid line is the average, shaded region is one std), **Right:** Cumulative regret of Algorithm 4 (Safe-OFUL with initial pure exploration) for the setting in Figure 2.5.

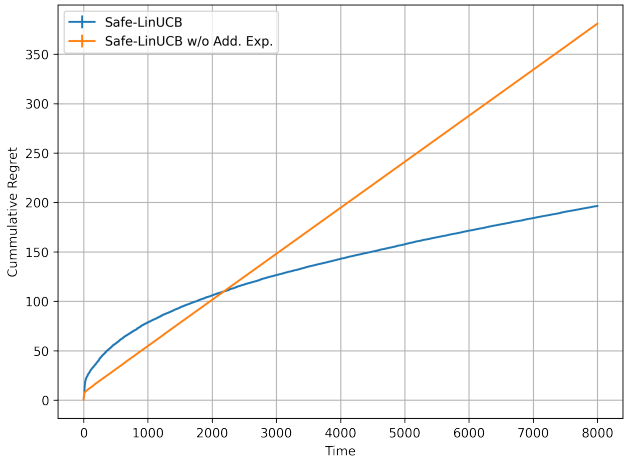

Figure 2.7: Cumulative regret in loan approval problem. Comparison of Safe-OFUL(Safe-LinUCB) with and without additional exploration.

2.3.6 Experiments

Illustrative 2D Simulations: We first empirically study the proposed algorithms in 2D action space. In the setting with 6 unknown affine constraints and feedback regions, we perform 5 independent runs of Safe-LinUCB and Safe-LinTS for 2000 time steps and report their performance. An example of the decision set D_0 with different (color) feedback regions and the region of safe actions determined by the affine constraints are shown in Figure 2.4. The cumulative regret of the algorithms in this setting is given in the first plot of Figure 2.6. We observe that both of the algorithms achieve competitive, i.e., sublinear, regret without any safety violations. We show that Safe-LinTS achieves improved practical performance in this setting with optimized exploration parameters η_t and η_t^c, which further motivates the use

of sampling-based methods in practice.

Next, we study the setting with 10 unknown nonlinear constraints and feedback regions. We model the constraints as ℓ_2-norm bound constraints. An example of D_0 and D_0^{safe} are given in Figure 2.5. We consider an optimal action with a safety gap in parallel with Assumption 2.3.10. We implement Algorithm 4 using Safe-LinUCB and provide the cumulative regret in Figure 2.6. As predicted by the theory, algorithm attains linear regret during its orthogonal pure exploration phase. However, this phase allows sufficient exploration of the safety sets and unknown reward function such that Algorithm 4 discovers a safe action that achieves at least as high reward as the optimal action, yielding constant regret after pure exploration. This shows that the novel initial exploration strategy in Section 2.3.5 is effective in uniformly exploring the decision set without any safety violations.

Loan Approval as a Safe SLB Problem: We consider the German Credit Dataset from Keogh et al. [141]. The data classify customers as good or bad for credit for loan approval and provide 24 attributes per user. To turn this into a safe SLB problem, we featurize the user attributes using a neural network and pose the bandit problem as a regression problem with affine safety constraints in the feature space. We impose two safety violations as picking bad customers with 1) high credit and 2) with high age, where the last one is a surrogate to retirement discussed in the case study at the beginning. We compare Safe-LinUCB with a naive version which does not include the additional exploration bonus needed to ensure optimism under safety.

Figure 2.7 gives the cumulative regret comparison. Initially, Safe-LinUCB attains higher regret than the naive version due to additional exploration incentives as expected. However, this additional exploration provides the sufficient exploration needed in the relevant constraint regions and allows Safe-LinUCB to achieve lower cumulative regret in the long run with no safety violations, concurring with the theory. The naive method, on the other hand, does not select optimistic actions and fails to explore efficiently, resulting in sub-optimal actions. This result highlights the importance of the carefully tuned exploration bonus under safety constraints to recover the underlying reward parameter.

2.4 Conclusion and Future Directions

In this chapter, we studied decision-making under uncertainty in the context of "stateless" dynamical systems, i.e., stochastic linear bandits, by addressing the challenges of high-dimensional feature representations and unknown safety constraints.

The proposed algorithms, PSLB, and safety-constrained stochastic linear bandits algorithms, Safe-OFUL and Safe-LinTS provide novel solutions to these challenges and have the potential to improve the performance and safety of decision-making tasks in various real-world applications. The empirical studies validate the effectiveness of the proposed algorithms and highlight their applicability to practical decision-making scenarios.

The PSLB framework can be generalized in various ways. Even though we chose PCA for the subspace recovery method for this study, it would be an interesting future directions to study the PSLB framework under robust principal component analysis or dictionary learning settings. Extending this line of study to the general class of low-dimensional manifold structured problems is also another interesting future direction. Moreover, deploying Thompson Sampling instead of optimism would mitigate the computational complexity of PSLB as shown in the safe bandit study.

In our study of stochastic linear bandits with unknown safety constraints and local safety feedback, our main contribution is to decouple the safety exploration from the reward exploration and find the right balance to ensure effective recovery of the safety functions while not overly exploring the system. This dichotomized approach captures the essence of safe decision-making under uncertainty problems. Extending this approach to other RL and safety frameworks is an important future direction. Another interesting future direction is to establish lower bounds for the multiple constraint setting. Even though our regret results are tight in terms of the horizon T, the tightness of $\sqrt{M}$ scaling of the regret upper bound requires further investigation.

Chapter 3

LEARNING AND CONTROL IN LINEAR QUADRATIC REGULATOR (LQR)

In this chapter, we study the problem of learning and control of unknown state-feedback linear time-invariant systems with quadratic cost, i.e., adaptive control of stabilizable linear-quadratic regulators (LQRs) without an a priori known stabilizing controller[1]. LQR is the canonical setting for linear dynamical systems with quadratic regulatory costs and observable state evolution. For a known LQR model, the optimal control policy is given by a stabilizing linear state feedback controller [28]. When the underlying model is unknown, the learning agent needs to learn the dynamics in order to (1) stabilize the system and (2) find the optimal control policy. This online control task is one of the core challenges in RL and control theory.

The ultimate goal in online control is to design learning agents that can autonomously adapt to the unknown environment with minimal information and also enjoy finite-time stability and performance guarantees. This problem has sparked a flurry of research interest in the control and RL communities. However, there are only a few approaches that provide a complete treatment of the problem and strive for learning from scratch with no initial model estimates [2, 7, 56]. Other than these, the prior works focus either on the problem of finding a stabilizing policy while ignoring the control costs [82], or on achieving low control costs while assuming access to an initial stabilizing controller [8, 242].

The existing works [2, 6, 7] that learn from scratch in LQRs aim to minimize the regret, which is the additional cumulative control cost of an agent compared to the expected cumulative cost of the optimal policy. These algorithms suffer from regret that has an exponential dependence in the LQR dimensions since they do not assume access to an initial stabilizing policy. They also face system blow-ups due to unstable system dynamics. Besides poor regret performance, the uncontrolled dynamics prevent the deployment of these learning algorithms in practice.

In this chapter, we design model-based RL agents for online LQRs that achieve low regret and fast stabilization. To design stabilizing policies without prior knowledge, the agent needs to effectively explore the environment and estimate the system

[1] This chapter is based on [137, 166].

dynamics. However, in order to achieve low regret, the agent should also strategically exploit the gathered knowledge. Thus, the agent requires balancing exploration and exploitation such that it designs stabilizing policies to avoid dire consequences of unstable dynamics and minimize regret.

We consider the stabilizable multi-dimensional LQR setting. Stabilizability is the necessary and sufficient condition to have a well-posed LQR control problem [132]. On the contrary, prior works usually consider the controllable LQR setting, which is a subclass of stabilizable LQRs [56, 62] or their algorithms have guarantees only for scalar systems [6, 7]. While the controllability condition simplifies the learning and control problem, it is also often violated in many real-world control systems [91].

For this setting, we propose 2 model-based reinforcement learning algorithms. The first algorithm, **Stab**ilizing **L**earning, StabL, deploys the optimism principle, discussed in Chapter 2 for Safe-OFUL, into online LQ control problem. StabL certifies fast stabilization of the underlying system by effectively exploring the environment with an improved exploration strategy. We show that StabL attains $\tilde{O}(\sqrt{T})$ regret after T time steps of agent-environment interaction. Here $\tilde{O}(\cdot)$ presents the order up to logarithmic terms. We also show that the regret of the proposed algorithm has only a polynomial dependence in the problem dimensions, which gives an exponential improvement over the prior methods. Our improved exploration method is simple, yet efficient, and it combines a sophisticated exploration policy of optimism with an isotropic exploration strategy to achieve fast stabilization and improved regret. We empirically demonstrate that the proposed algorithm outperforms other popular methods in several adaptive control tasks.

The second algorithm, **T**hompson **S**ampling-based **A**daptive **C**ontrol, TSAC, overcomes possible computational inefficiencies of StabL by using TS to balance exploration vs. exploitation trade-off and design the controllers. Despite the computational efficiency of TS, prior work [7] was able to achieve the optimal $\tilde{O}(\sqrt{T})$ regret only for scalar systems. TSAC builds on the algorithmic intuitions gathered in StabL and achieves $\tilde{O}(\sqrt{T})$ regret even for multidimensional systems, thereby solving the open problem posed in [7]. Similar to StabL, TSAC does not require an a priori known stabilizing controller and achieves fast stabilization of the underlying system by effectively exploring the environment in the early stages. Our breakthrough hinges on developing a novel lower bound on the probability that the TS provides an optimistic sample. By carefully prescribing an early exploration strategy and a policy update rule, we show that TS achieves order-optimal regret in adaptive

Table 3.1: Works that attain $\tilde{O}(\sqrt{T})$ regret on LQR, † = 1-dim LQRs.

Work	Setting	Stabilizing Controller	Computation
Mania et al. [191]	Controllable	Required	P
Cohen et al. [62]	Controllable	Required	P
Abbasi-Yadkori and Szepesvári [2]	Controllable	Not required	NP
Chen and Hazan [56]	Controllable	Not required	P
Faradonbeh et al. [83]	Stabilizable	Required	P
Faradonbeh et al. [81]	Stabilizable	Required	NP
Simchowitz and Foster [242]	Stabilizable	Required	P
Abeille and Lazaric [7]	Stabilizable†	Not Required	P
StabL (Algorithm 5)	Stabilizable	Not required	NP
TSAC (Algorithm 6)	Stabilizable	Not required	P

control of multidimensional stabilizable LQRs. We empirically demonstrate the performance and the efficiency of TSAC in the adaptive control task of Boeing 747.

3.1 Motivation and Background

The learning and control problem in LQRs have been studied in an array of prior works [1, 7, 8, 56, 81, 85, 191, 242]. These works provide finite-time performance guarantees of adaptive control algorithms in terms of *regret*. In particular, they show that $\tilde{O}(\sqrt{T})$ regret after T time steps is optimal in adaptive control of LQRs. They utilize several different paradigms for algorithm design such as Certainty Equivalence, Optimism or Thompson Sampling, yet, they suffer either from the inherent algorithmic drawbacks or limited applicability in practice. Table 3.1 gives an overall comparison of these works in terms of their setting, the requirement of stabilizing controllers, and computational complexities. In the following, we compare these works in more detail.

Certainty equivalent control: Certainty equivalent control (CEC) is one of the most straightforward paradigms for control design in adaptive control of dynamical systems. In CEC, an agent obtains a nominal estimate of the system, and executes the optimal control law for this estimated system (Figure 1.1). Even though Mania et al. [191] and Simchowitz and Foster [242] show that this simple approach attains order-optimal regret in LQRs, the proposed algorithms have several drawbacks. First and foremost, CEC is sensitive to model mismatch and requires significantly small model estimation errors to the point that exploration of the system dynamics is not required. Since this level of refinement is challenging to obtain for an unknown system, these methods rely on access to an initial stabilizing controller to enable a long exploration. In practice, such a priori known controllers may not be available,

which hinders the deployment of these algorithms.

Moreover, CEC-based approaches follow the given non-adaptive initial stabilizing policy for a *long* period of time with isotropic perturbations. Thus, they provide an order-optimal theoretical regret upper bound with an additional large constant regret. However, in many applications such as medical, such constant regret and non-adaptive controllers are not tolerable. Our methods StabL and TSAC aim to address these challenges and provide learning and control algorithms that can be deployed in practice. As we will show in Sections 3.2.3 and 3.3.4, they achieve significantly improved performance over the prior baseline RL algorithms in various adaptive control tasks.

Optimism in the face of uncertainty (OFU) principle: As we have seen in Chapter 2, one of the most prominent methods to effectively balance exploration and exploitation is optimism in the face of uncertainty (OFU) principle [156]. An agent that follows the OFU principle deploys the optimal policy of the model with the lowest optimal cost within the set of plausible models (Figure 1.1). This guarantees the asymptotic convergence to the optimal policy for the LQR [30]. Using the OFU principle, the learning algorithms of [2, 86] attain order-optimal $O(\sqrt{T})$ regret after T time steps, but their regret upper bounds suffer from an *exponential* dependence in the LQR model dimensions.

This is due to the fact that the OFU principle relies heavily on the confidence-set constructions. An agent following the OFU principle mostly explores parts of state space with the lowest expected cost and with higher uncertainty. When the agent does not have reliable model estimates, this may cause a lack of exploration in certain parts of the state space that are important in designing stabilizing policies. This problem becomes more evident in the early stages of agent-environment interactions due to the lack of reliable knowledge about the system. Note that this issue is unique to control problems and not as common in other RL settings, e.g., bandits and gameplay. With the early improved exploration strategy of StabL, we alleviate this problem, achieve fast stabilization, and thus attain $\tilde{O}(\sqrt{T})$ regret upper bound with polynomial dimension dependency.

Thompson Sampling: Thompson Sampling (TS) is one of the oldest strategies to balance the exploration vs. exploitation trade-off [263]. In TS, the agent samples a model from a distribution computed based on prior control input and observation pairs, and then takes the optimal action for this sampled model and updates the

distribution based on its novel observation (Figure 1.1). Since it relies solely on sampling, this approach provides polynomial-time algorithms for adaptive control. Therefore, it is a promising alternative to overcome the possible computational burden of optimism-based methods, since they solve a non-convex optimization problem to find the optimistic controllers, which is an NP-hard problem in general [9].

For this reason, [6, 7] propose adaptive control algorithms using TS. In particular, Abeille and Lazaric [7] provide the first TS-based adaptive control algorithm for LQRs that attains optimal regret of $\tilde{O}(\sqrt{T})$. However, their result *only holds for scalar* stabilizable systems, since they were able to show that TS samples optimistic parameters with constant probability in only scalar systems. Further, they conjecture that this is true in multidimensional systems as well and TS-based adaptive control can provide optimal regret in multidimensional LQRs, and provide a simple numerical example to support their claims. In the analysis of TSAC, we derive a new lower bound which in fact proves that TS samples optimistic parameters with a constant probability for all stabilizable LQRs. Further, we design TSAC with the required algorithmic improvements to attain $\tilde{O}(\sqrt{T})$ in multidimensional stabilizable LQRs.

Finding a stabilizing controller: Similar to the regret minimization, there has been a growing interest in finite-time stabilization of linear dynamical systems [72, 82, 84]. Among these works, Faradonbeh et al. [82] is the closest to our study. However, there are significant differences in the methods and the span of the results. In Faradonbeh et al. [82], random linear controllers are used solely for finding a stabilizing set without a control goal. This results in the explosion of state, presumably exponentially in time, leading to a regret that scales exponentially in time. The proposed method provides many insightful aspects for finding a stabilizing set in finite time, yet a cost analysis of this process or an adaptive control policy have not been provided. Moreover, the stabilizing set in [82] relates to the minimum value that satisfies a specific condition for the roots of a polynomial. This results in a somewhat implicit sample complexity for constructing such a set. On the other hand, in this chapter, we provide a complete study of an autonomous learning and control algorithm for the online LQR problem. Among our results, we give an explicit formulation of the stabilizing set and a sample complexity that only relates to the minimal stabilizability information of the system.

Generalized LQR setting: Another line of research considers the generalizations of the online LQR problem under partial observability [160–162, 191, 245] or ad-

versarial disturbances [56, 108]. These works either assume a given stabilizing controller or open-loop stable system dynamics, except Chen and Hazan [56]. Independently and concurrently, the recent work by Chen and Hazan [56] designs an autonomous learning algorithm and regret guarantees that are similar to StabL and TSAC. However, the approaches and the settings have major differences. Chen and Hazan [56] consider the restrictive setting of *controllable* systems, yet with adversarial disturbances and general cost functions. They inject *significantly* big inputs, *exponential in system parameters*, with a pure exploration intent to guarantee the recovery of system parameters and stabilization. This negatively affects the practicality of the algorithm. On the other hand, in StabL and TSAC, we inject isotropic Gaussian perturbations to improve the exploration in the stochastic (sub-Gaussian noise) *stabilizable* LQR while still aiming to control, i.e., no pure exploration phase. This yields practical RL algorithms that attain state-of-the-art performance.

Notation

We denote the Euclidean norm of a vector x as $\|x\|_2$. For a matrix $A \in \mathbb{R}^{n\times d}$, we denote $\rho(A)$ as the spectral radius of A, $\|A\|_F$ as its Frobenius norm and $\|A\|$ as its spectral norm. $\mathrm{tr}(A)$ denotes its trace, $A^\top$ is the transpose. For any positive definite matrix V, $\|A\|_V = \|V^{1/2}A\|_F$. For matrices $A, B \in \mathbb{R}^{n\times d}$, $A \bullet B = \mathrm{tr}(AB^\top)$ denotes their Frobenius inner product. The j-th singular value of a rank-n matrix A is $\sigma_j(A)$, where $\sigma_{\max}(A) := \sigma_1(A) \geq \ldots \geq \sigma_{\min}(A) := \sigma_n(A)$. I represents the identity matrix with the appropriate dimensions. $\mathcal{M}_n = \mathbb{R}^{n\times n}$ denotes the set of n-dimensional square matrices. $\mathcal{N}(\mu, \Sigma)$ denotes normal distribution with mean μ and covariance Σ. $Q(\cdot)$ denotes the Gaussian Q-function. $O(\cdot)$ and $o(\cdot)$ denote the standard asymptotic notation and $f(T) = \omega(g(T))$ is equivalent to $g(T) = o(f(T))$. $\tilde{O}(\cdot)$ presents the order up to logarithmic terms.

Problem Setting

Consider a discrete-time linear time-invariant system,

$$x_{t+1} = A_* x_t + B_* u_t + w_t, \tag{3.1}$$

where $x_t \in \mathbb{R}^n$ is the state of the system, $u_t \in \mathbb{R}^d$ is the control input, $w_t \in \mathbb{R}^n$ is the process noise at time t. We consider the systems with sub-Gaussian noise.

Assumption 3.1 (Sub-Gaussian Noise)**.** *The process noise w_t is a martingale difference sequence with respect to the filtration $(\mathcal{F}_{t-1})$. Moreover, it is component-wise*

conditionally σ_w^2-sub-Gaussian and isotropic such that for any $s \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(sw_{t,j}\right)|\mathcal{F}_{t-1}\right] \leq \exp\left(s^2\sigma_w^2/2\right)$$

and $\mathbb{E}\left[w_t w_t^\top|\mathcal{F}_{t-1}\right] = \bar{\sigma}_w^2 I$ for some $\bar{\sigma}_w^2 > 0$.

Note that the results of this chapter only require the conditional covariance matrix $W = \mathbb{E}[w_t w_t^\top|\mathcal{F}_{t-1}]$ to be full rank. The isotropic noise assumption is chosen to ease the presentation, and similar results can be obtained with upper and lower bounds on W, i.e., $W_{up} > \sigma_{\max}(W) \geq \sigma_{\min}(W) > W_{low} > 0$.

At each time step t, the system is at state x_t. After observing x_t, the agent applies a control input u_t and the system evolves to x_{t+1} at time $t+1$. At each time step t, the agent pays a cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$, where $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{d\times d}$ are positive definite matrices such that $\|Q\|, \|R\| < \overline{\alpha}$ and $\sigma_{\min}(Q), \sigma_{\min}(R) > \underline{\alpha}$. The problem is to design control inputs based on past observations in order to minimize the average expected cost

$$J_* = \lim_{T\to\infty} \min_{u=[u_1,\ldots,u_T]} \frac{1}{T}\mathbb{E}\Big[\sum\nolimits_{t=1}^{T} x_t^\top Q x_t + u_t^\top R u_t\Big], \tag{3.2}$$

by designing control inputs based on past observations. This problem is the canonical example for the control of linear dynamical systems and is termed as linear quadratic regulator (LQR). The system (3.1) can be represented as $x_{t+1} = \Theta_*^\top z_t + w_t$, where $\Theta_*^\top = [A_* \;\; B_*]$ and $z_t = [x_t^\top \;\; u_t^\top]^\top$. Knowing Θ_*, the optimal control policy, is a linear state feedback control $u_t = K(\Theta_*)x_t$ with $K(\Theta_*) = -(R+B_*^\top P_* B_*)^{-1}B_*^\top P_* A_*$, where P_* is the unique solution to the discrete-time algebraic Riccati equation (DARE) [28]:

$$P_* = A_*^\top P_* A_* + Q - A_*^\top P_* B_*(R + B_*^\top P_* B_*)^{-1}B_*^\top P_* A_*. \tag{3.3}$$

The optimal cost for Θ_* is denoted as $J(\Theta_*) = J_* = \mathrm{Tr}(\bar{\sigma}_w^2 P_*)$. In this work, unlike the controllable LQR setting of the prior adaptive control algorithms without a stabilizing controller [2, 56], we study the online LQR problem in the general setting of *stabilizable* LQR.

Definition 3.1 (Stabilizability vs. Controllability)**.** *The linear dynamical system Θ_* is stabilizable if there exists K such that $\rho(A_* + B_* K) < 1$. On the other hand, the linear dynamical system Θ_* is controllable if the controllability matrix $[B_* \;\; A_*B_* \;\; A_*^2B_* \;\; \ldots \;\; A_*^{n-1}B_*]$ has full row rank.*

Note that the stabilizability condition is the minimum requirement to define the optimal control problem. It is *strictly weaker than controllability*, i.e., all controllable systems are stabilizable but the converse is not true [28]. Similar to Cohen et al. [62], we quantify the stabilizability of Θ_* for the finite-time analysis.

Definition 3.2 ((κ, γ)-Stabilizability)**.** *The linear dynamical system Θ_* is (κ, γ)-stabilizable for ($\kappa \geq 1$ and $0 < \gamma \leq 1$) if $\|K(\Theta_*)\| \leq \kappa$ and there exists L and $H \succ 0$ such that $A_* + B_* K(\Theta_*) = HLH^{-1}$, with $\|L\| \leq 1 - \gamma$ and $\|H\|\|H^{-1}\| \leq \kappa$.*

Note that this is merely a quantification of stabilizability. In other words, any stabilizable system is also (κ, γ)-stabilizable for some κ and γ and conversely (κ, γ)-stabilizability implies stabilizability. In particular, for all stabilizable systems, by setting $1 - \gamma = \rho(A_* + B_* K(\Theta_*))$ and κ to be the condition number of $P(\Theta_*)^{1/2}$ where $P(\Theta_*)$ is the positive definite matrix that satisfies the following Lyapunov equation:

$$(A_* + B_* K(\Theta_*))^\top P(\Theta_*)(A_* + B_* K(\Theta_*)) \leq P(\Theta_*), \tag{3.4}$$

one can show that $A_* + B_* K(\Theta_*) = HLH^{-1}$, where $H = P(\Theta_*)^{-1/2}$ and $L = P(\Theta_*)^{1/2}(A_* + B_* K(\Theta_*))P(\Theta_*)^{-1/2}$ with $\|H\|\|H^{-1}\| \leq \kappa$, and $\|L\| \leq 1 - \gamma$, (see Lemma B.1 of Cohen et al. [61]).

Assumption 3.2 (Stabilizable Linear Dynamical System)**.** *The unknown parameter $\Theta_* \in \mathcal{S}$ such that $\mathcal{S} = \left\{\Theta' = [A', B'] \;\middle|\; \Theta' \text{ is } (\kappa, \gamma)\text{-stabilizable, } \|\Theta'\|_F \leq S\right\}$.*

Notice that $\mathcal{S}$ denotes the set of all bounded systems that are (κ, γ)-stabilizable, where Θ_* is an element of, and the membership to $\mathcal{S}$ can be easily verified. Moreover, the proposed algorithm in this work only requires the upper bounds on these relevant control-theoretic quantities κ, γ, and S, which are also standard in prior works, e.g., [2, 62]. In practice, when there is a total lack of knowledge about the system, one can start with conservative upper bounds and adjust these based on the behavior of the system, e.g., the growth of the state. From (κ, γ)-stabilizability, we have that $\rho(A' + B' K(\Theta')) \leq 1 - \gamma$, and $\sup\{\|K(\Theta')\| \mid \Theta' \in \mathcal{S}\} \leq \kappa$. The following lemma shows that for any (κ, γ)-stabilizable system the solution of (3.3) is bounded.

Lemma 3.1 (Bounded DARE Solution)**.** *For any Θ that is (κ, γ)-stabilizable and has bounded regulatory cost matrices, i.e., $\|Q\|, \|R\| < \overline{\alpha}$, the solution of* (3.3)*, P, is bounded as $\|P\| \leq D := \overline{\alpha}\gamma^{-1}\kappa^2(1 + \kappa^2)$.*

Proof. The solution of DARE in (3.3) corresponds to recursively applying the following

$$\begin{aligned}
\|P_*\| &= \left\| \sum_{t=0}^{\infty} \left((A_* + B_* K(\Theta_*))^t \right)^\top \left(Q + K(\Theta_*)^\top R K(\Theta_*) \right) (A_* + B_* K(\Theta_*))^t \right\| \\
&= \left\| \sum_{t=0}^{\infty} \left(H L^t H^{-1} \right)^\top \left(Q + K(\Theta_*)^\top R K(\Theta_*) \right) \left(H L^t H^{-1} \right) \right\| \\
&\leq \overline{\alpha}(1 + \|K(\Theta_*)\|^2) \|H\|^2 \|H^{-1}\|^2 \sum_{t=0}^{\infty} \|L\|^{2t} \qquad (3.5) \\
&\leq \overline{\alpha} \gamma^{-1} \kappa^2 (1 + \kappa^2), \qquad (3.6)
\end{aligned}$$

where (3.5) follows from the upper bound on $\|Q\|, \|R\| \leq \overline{\alpha}$ and (3.6) follows from the definition of (κ, γ)-stabilizability. □

This lemma also shows that for the underlying stabilizable system, $J(\Theta_*) < \infty$.

Finite-Time Adaptive Control/Model-based RL Problem in LQRs

In our study, we consider the adaptive control setting where the model parameters A_* and B_*, i.e., Θ_*, are unknown. In this scenario, the learning agent needs to interact with the environment to learn these parameters and aims to minimize the cumulative cost $\sum_{t=1}^{T} c_t$. Note that the cost matrices Q and R are the designer's choice and given. After T time steps, we evaluate the regret of the learning agent as

$$R_T = \sum_{t=0}^{T} (c_t - J(\Theta_*)),$$

which is the difference between the performance of the agent and the expected performance of the optimal controller.

3.2 Optimism-Based Adaptive Control

In this section, we describe Stabilizing Learning algorithm, StabL, for the online LQR problem and study its performance both theoretically and empirically. We carefully prescribe an early exploration strategy and a policy update rule in the design of StabL. We show that StabL quickly stabilizes the underlying system, and henceforth certifies the stability of the dynamics with high probability in the stabilizable LQRs. We show that StabL attains $\tilde{O}(\text{poly}(n, d)\sqrt{T})$ regret in the online control of unknown stabilizable LQRs. This makes StabL the first RL algorithm to achieve order-optimal regret in all stabilizable LQRs without a given initial stabilizing policy.

The design of StabL is motivated by the importance of stabilizing the unknown dynamics and the need for exploration in the early stages of agent-environment

interactions. StabL deploys the OFU principle to balance the exploration vs. exploitation trade-off. Due to the lack of reliable estimates in the early stages of learning, an optimistic controller, guided by OFU, neither provides sufficient exploration required to achieve stabilizing controllers nor achieves sub-linear regret. Therefore, StabL uses isotropic exploration along with the optimistic controller in the early stages to achieve an improved exploration strategy. This allows StabL to excite all dimensions of the system uniformly as well as the dimensions that have a more promising impact on the control performance. By carefully adjusting the early improved exploration, we guarantee that the inputs of StabL are persistently exciting the system under the sub-Gaussian process noise. We show that using this improved exploration quickly results in stabilizing policies with high probability, therefore a much smaller regret in the long term.

We conduct extensive experiments to verify the theoretical claims about StabL and study the performance of StabL in various adaptive control tasks. We empirically show that the improved exploration strategy of StabL persistently excites the system in the early stages and achieves effective system identification required for stabilization. We observe that the optimism-based learning algorithm of Abbasi-Yadkori and Szepesvári [2] fails to achieve effective exploration in the early stages and suffers from unstable dynamics and high regret. In contrast, StabL obtains reliable model estimates for stabilization, and the balanced strategy prescribed by the OFU principle effectively guides StabL to regret-minimizing policies, resulting in-orders-of magnitude improvement in regret compared to the existing certainty equivalent and optimism-based methods in all settings.

3.2.1 StabL

We present StabL, a sample efficient stabilizing RL algorithm for the online stabilizable LQR problem. The algorithmic outline is provided in Algorithm 5. StabL only requires minimal information about the stabilizability of the underlying system and *does not* need a stabilizing controller. Therefore, along with the ultimate goal of minimizing regret, StabL puts its primary focus on achieving stabilizing controllers for unknown system dynamics.

Algorithm 5 StabL

Input: κ, γ, Q, R, σ_w^2 $\bar{\sigma}_w^2$, $V_0 = \mu I$, $\hat{\Theta}_0 = 0$, $\tau = 0$
for $t = 0, \ldots, T$ **do**
  if $(\det(V_t) > 2\det(V_0))$ **and** $(t - \tau > H_0)$ **then**
    Estimate $\hat{\Theta}_t$ & find optimistic $\tilde{\Theta}_t \in C_t(\delta) \cap \mathcal{S}$
    Set $V_0 = V_t$ and $\tau = t$.
  else
    $\tilde{\Theta}_t = \tilde{\Theta}_{t-1}$
  if $t \leq T_w$ **then**
    $u_t = K(\tilde{\Theta}_{t-1})x_t + \nu_t$ ▷ Improved Exploration
  else
    $u_t = K(\tilde{\Theta}_{t-1})x_t$ ▷ Stabilizing Control
  Pay cost c_t & Observe x_{t+1}
  Update $V_{t+1} = V_t + z_t z_t^\top$ for $z_t = [x_t^\top \;\; u_t^\top]^\top$

Adaptive Control with Improved Exploration

In order to quickly design stabilizing controllers, StabL needs to explore the system dynamics effectively. To this end, StabL solves

$$\min_{\Theta} \sum_{s=0}^{t-1} \|x_{s+1} - \Theta^\top z_s\|^2 + \mu\|\Theta\|_F^2, \tag{3.7}$$

using the past state-input pairs to estimate the system dynamics as $\hat{\Theta}_t$. Using this estimate, StabL constructs a high probability confidence set $C_t(\delta)$ that contains the underlying parameter Θ_* with high probability. In particular, for $\delta \in (0, 1)$, at time step t, it forms

$$C_t(\delta) = \{\Theta : \|\Theta - \hat{\Theta}_t\|_{V_t} \leq \beta_t(\delta)\}, \tag{3.8}$$

for $\beta_t(\delta) = \sigma_w\sqrt{2n\log(\delta^{-1}\sqrt{\det(V_t)/\det(\mu I)})} + \sqrt{\mu}S$ and $V_t = \mu I + \sum_{i=0}^{t-1} z_i z_i^\top$ such that $\Theta_* \in C_t(\delta)$ with probability at least $1 - \delta$ for all time steps t. Note that this estimation method and the learning guarantee is standard in learning linear dynamical systems since Abbasi-Yadkori and Szepesvári [2]. Instead of solving (3.7) from scratch, the model estimate updates can be done via batch or online updates using the standard linear regression techniques.

The confidence set given in (3.8) provides a self-normalized bound on the model parameter estimates via design matrix V_t. StabL uses the OFU principle in this confidence set to design a policy. In particular, it chooses an optimistic parameter $\tilde{\Theta}_t$ from $C_t \cap \mathcal{S}$, which has the lowest expected optimal cost, and constructs the optimal linear controller $K(\tilde{\Theta}_t)$ for $\tilde{\Theta}_t$, i.e., the optimistic controller. At time t,

StabL uses the optimistic controller $K(\tilde{\Theta}_{t-1})$. This choice is for technical reasons to guarantee the persistence of excitation (Appendix B.1.1).

The optimistic controllers allow StabL to adaptively balance exploration and exploitation. They guide the exploration toward the region of state space with the lowest expected cost. The key idea in this design is that as the confidence set shrinks, the performance of StabL improves over time [30].

Due to lack of an initial stabilizing policy, StabL aims to rapidly stabilize the system to avoid the consequences of unstable dynamics. To stabilize an unknown LQR, one requires sufficient exploration in all directions of the state-space (Lemma 3.3). Unfortunately, due to lack of reliable estimates in the early stages, the optimistic policies come short to guarantee such an effective exploration.

Therefore, StabL deploys an adaptive control policy with an improved exploration in the early stages of interactions with the system. In particular, for the first T_w time-steps, StabL uses isotropic perturbations along with the optimistic controller. For $t \leq T_w$, it injects an i.i.d. Gaussian vector $v_t \sim \mathcal{N}(0, \sigma_v^2 I)$ to the system besides the optimistic policy $K(\tilde{\Theta}_{t-1})x_t$, where $\sigma_v^2 = 2\kappa^2\bar{\sigma}_w^2$.

StabL effectively excites and explores all dimensions of the system via this improved exploration strategy (Theorem 3.1). The duration of the adaptive control with improved exploration phase is chosen such that StabL quickly finds a stabilizing controller. In particular, after $T_w := poly(\sigma_w, \sigma_v, n, d, \gamma^{-1}, \kappa, \overline{\alpha}, \log(1/\delta))$ time steps, StabL has the guarantee that the linear controllers $K(\tilde{\Theta}_{t-1})$ stabilize Θ_* for all $t \geq T_w$ with high probability (Lemma 3.2 & 3.3).

Moreover, StabL avoids frequent updates in the system estimates and the controller. It uses the same controller at least for a fixed time period of $H_0 = O(\gamma^{-1}\log(\kappa))$ and also waits for a significant improvement in the estimates. The latter is achieved by updating the controller if the determinant of the design matrix V_t is doubled since the last update. This update rule is chosen such that policy changes do not cause unstable dynamics for the stabilizable LQR. The effects of this update rule on maintaining a bounded state for StabL are studied in detail in Section 3.2.2.

Stabilizing Adaptive Control

After guaranteeing the stabilizing policy design, StabL starts the adaptive control that stabilizes the underlying system. In this phase, StabL stops injecting isotropic perturbations and relies on the balanced exploration and exploitation via the op-

timistic controller design. The stabilizing optimistic controllers further guide the exploration to adapt the structure of the problem and fine-tune the learning process to achieve optimal performance. However, note that frequent policy changes can still cause unbounded growth of the state even though the policies are stabilizing. Therefore, StabL continues the same policy update rule in this phase to maintain a bounded state.

Unlike the prior works that constitute two distinct phases, StabL has a very subtle two-phase structure. In particular, the same subroutine (optimism) is applied continuously with the aim of balancing exploration and exploitation. An additional isotropic perturbation is only deployed for an improved exploration in the early stages to achieve stable learning for the autonomous agent.

3.2.2 Theoretical Analysis of StabL

In this section, we study the main theoretical guarantees of StabL. We first discuss the challenges that the stabilizability setting brings compared to the setting of the prior learning algorithms for the online LQR. We then introduce our approaches to overcome these challenges in the design of StabL. Later in the section, we provide the formal statements for the theoretical guarantees of StabL and, finally, we give the regret upper bound of StabL.

Challenges in the Online Stabilizable LQR Problem

The main challenge for learning algorithms in control problems is to achieve input-to-state stability (ISS), which requires having a well-bounded state in future time steps via using bounded inputs. Achieving this becomes significantly more challenging in the setting of stabilizable LQR compared to their controllable counterpart considered in many recent works [2, 56, 191]. A controllable system can be brought to $x_t = 0$ in finite time steps. Furthermore, some of these works assume that the underlying system is closed-loop contractible, i.e., $\|A_* - B_* K(\Theta_*)\| < 1$. These facts significantly simplify the overall stabilization problem. Moreover, recalling Definition 3.1, for controllable systems, the controllability matrix is full row rank. In prior works, this has been a prominent factor in guaranteeing the persistence of excitation (PE) of the inputs, identifying the system, and deriving regret bounds, e.g., [56, 108].

Unfortunately, we do not have these properties in the general stabilizable LQR setting. Recall Assumption 3.2 that states the system is (κ, γ)-stabilizable, which yields $\rho(A_* + B_* K(\Theta_*)) \leq 1 - \gamma$ for the optimal policy $K(\Theta_*) \leq \kappa$. Therefore, even

if the optimal policy of the underlying system is chosen by the learning algorithm, it may not produce a contractive closed-loop system, i.e., we can have $\rho(A_* + B_*K(\Theta_*)) < 1 < \|A_* + B_*K(\Theta_*)\|$ since for any matrix M, $\rho(M) \leq \|M\|$.

Moreover, from the definition of stabilizability in Definitions 3.1 and 3.2, we know that for any stabilizing controller K', there exists a similarity transformation $H' \succ 0$ such that it makes the closed-loop system contractive, i.e., $A_* + B_*K' = H'LH'^{-1}$, with $\|L\| < 1$. However, even if all the policies that StabL executes stabilize the underlying system, these different similarity transformations of different policies can further cause an explosion of state during the policy changes. If policy changes happen frequently, this may even lead to linear growth of the state over time.

In order to resolve these problems, StabL carefully designs the timing of the policy updates and applies all the policies long enough, so that the state stays well controlled, i.e., ISS is achieved. To this end, StabL applies the same policy at least for $H_0 = 2\gamma^{-1}\log(2\kappa\sqrt{2})$ time steps. This particular choice prevents state blow-ups due to policy changes in the optimistic controllers in the stabilizable LQR setting (see Appendix B.1.3).

To achieve persistence of excitation and consistent model estimates under the stabilizability condition, we leverage the early improved exploration strategy which does not require controllability. Using the isotropic exploration in the early stages, we derive a novel lower bound for the smallest eigenvalue of the design matrix V_t in the stabilizable LQR with sub-Gaussian noise setting. Moreover, we derive our regret results using the fast stabilization and the optimistic policy design of StabL. The results only depend on the stabilizability and other trivial model properties such as the LQR dimensions.

Benefits of Early Improved Exploration

To achieve effective exploration in the early stages, StabL deploys isotropic perturbations along with the optimistic policy for $t \leq T_w$. Define $\sigma_\star > 0$ where $\sigma_\star$ is a problem and in particular $\bar{\sigma}_w, \sigma_w, \sigma_v$-dependent constant (See Appendix B.1.1 for exact definition). The following shows that for a long enough improved exploration, the inputs are persistently exciting the system.

Theorem 3.1 (Persistence of Excitation During the Improved Exploration)**.** *If StabL follows the early improved exploration strategy for $T \geq poly(\sigma_w^2, \sigma_v^2, n, \log(\frac{1}{\delta}))$ time steps, then with probability at least $1-\delta$, StabL has $\sigma_{\min}(V_T) \geq \sigma_\star^2 T$.*

This theorem shows that having isotropic perturbations along with the optimistic controllers provides persistence excitation of the inputs, i.e., linear scaling of the smallest eigenvalue of the design matrix V_t. This result is quite technical and its proof is given in Appendix B.1.1. At a high level, we show that isotropic perturbations allow the covariates to have a Gaussian-like tail lower bound even in the stabilizable LQR with sub-Gaussian process noise setting. Using the standard covering arguments, we prove the statement of the theorem. This result guarantees that the inputs excite all dimensions of the state space and allows StabL to obtain uniformly improving estimates at a faster rate.

Lemma 3.2 (Parameter estimation error)**.** *Suppose Assumptions 3.1 and 3.2 hold. For $T \geq poly(\sigma_w^2, \sigma_\nu^2, n, \log(1/\delta))$ time steps of adaptive control with improved exploration, with probability at least 1–2δ, StabL achieves $\|\hat{\Theta}_T - \Theta_*\|_2 \leq \beta_t(\delta)/(\sigma_\star\sqrt{T})$.*

This lemma shows that early improved exploration strategy using $\nu_t \sim \mathcal{N}(0, \sigma_\nu^2)$ for $\sigma_\nu^2 = 2\kappa^2\bar{\sigma}_w^2$ enables guarantee of the consistency of the parameter estimation. The proof is in Appendix B.1.2, where we combine the confidence set construction in (3.8) with Theorem 3.1. This bound is utilized to guarantee stabilizing controllers after early improved exploration. However, first we have the following lemma, which shows that there is a stabilizing neighborhood around Θ_*, such that $K(\Theta')$ stabilizes Θ_* for any Θ' in this region.

Lemma 3.3 (Strongly Stabilizable Neighborhood)**.** *For $D = \overline{\alpha}\gamma^{-1}\kappa^2(1+\kappa^2)$, let $C_0 = 142D^8$ and $\epsilon = 1/(54D^5)$. For any (κ, γ)-stabilizable system Θ_* and for any $\varepsilon \leq \min\{\sqrt{\bar{\sigma}_w^2 nD/C_0}, \epsilon\}$, such that $\|\Theta' - \Theta_*\| \leq \varepsilon$, $K(\Theta')$ produces (κ', γ')-stable closed-loop dynamics on Θ_* where $\kappa' = \kappa\sqrt{2}$ and $\gamma' = \gamma/2$.*

Proof. For stabilizable systems, we know that the solution of (3.3) is unique and positive definite. Let $J_* \leq \mathcal{J}$. The following lemma is adapted from Simchowitz and Foster [242] and shows that if the estimation error on the system parameters is small enough, then the performance of the optimal controller synthesized by these model parameter estimates scales quadratically with the estimation error.

Lemma 3.4 (Adapted from [242])**.** *For constants $C_0 = 142\|P_*\|^8$ and $\epsilon = \frac{54}{\|P_*\|^5}$, such that, for any $0 \leq \varepsilon \leq \epsilon$ and for $\|\Theta' - \Theta_*\| \leq \varepsilon$, the infinite horizon performance of the policy $K(\Theta')$ on Θ_* obeys the following $J(K(\Theta'), A_*, B_*, Q, R) - J_* \leq C_0\varepsilon^2$.*

This result shows that there exists a ϵ-neighborhood around the system parameters that stabilizes the system. One can extend this result further to quantify

the stability as in Cassel et al. [48]. In particular, for bounded infinite horizon cost obtained by a policy $K(\Theta')$ on Θ_*, Lemma 41 of Cassel et al. [48] shows that $K(\Theta')$ produces (κ', γ')-stable closed-loop dynamics where $\kappa' = \sqrt{\frac{\mathcal{J}'}{\underline{\alpha}\tilde{\sigma}_w^2}}$ and $\gamma' = 1/2\kappa'^2$. Under Assumptions 3.1 & 3.2, for $\varepsilon \leq \min\{\sqrt{\mathcal{J}/C_0}, \epsilon\}$, we obtain $J(K(\Theta'), A_*, B_*, Q, R) \leq 2\mathcal{J}$. Plugging this in $\mathcal{J}'$ gives the advertised result. □

This lemma shows that to guarantee the stabilization of the unknown dynamics, a learning agent should have uniformly sufficient exploration in all directions of the state-space. By the choice of T_w (precise expression given in Appendix B.1.3) and using Lemma 3.2, StabL guarantees quickly finding this stabilizing neighborhood with high probability due to the adaptive control with improved exploration phase of T_w time steps. For the remaining time steps, $t \geq T_w$, StabL starts redressing the possible state explosion due to unstable controllers and the perturbations in the early stages. Define T_{base} and T_r such that $T_{base} = (n+d)\log(n+d)H_0$ and $T_r = T_w + T_{base}$. Recall that H_0 is the minimum duration for a controller such that the state is well-controlled despite the policy changes. The following shows that the stabilizing controllers are applied long enough that the state stays bounded for $T > T_r$.

Lemma 3.5 (Bounded states)**.** *Suppose Assumption 3.1 & 3.2 hold. For given T_w and T_{base}, StabL controls the state such that $\|x_t\| = O((n+d)^{n+d})$ for $t \leq T_r$, with probability at least $1 - 2\delta$ and $\|x_t\| \leq (12\kappa^2 + 2\kappa\sqrt{2})\gamma^{-1}\sigma_w\sqrt{2n\log(n(t-T_w)/\delta)}$ for $T \geq t > T_r$, with probability at least $1 - 4\delta$.*

In the proof (Appendix B.1.3), we show the policies seldom change via the determinant doubling condition or the lower bound of H_0 for the adaptive control with improved exploration phase to keep the state bounded. For the stabilizing adaptive control, we show that deploying stabilizing policies for at least H_0 time-steps provides an exponential decay on the state and after T_{base} time-steps brings the state to an equilibrium.

Regret Upper Bound of StabL

After showing the effect of fast stabilization, we can finally present the regret upper bound of StabL.

Theorem 3.2 (Regret of StabL)**.** *Suppose Assumptions 3.1 and 3.2 hold. For the given choices of T_w and T_{base}, with probability at least $1 - 4\delta$, StabL achieves regret of $O\left(poly(n,d)\sqrt{T\log(1/\delta)}\right)$, for long enough T.*

Table 3.2: Regret Performance after 200 Time Steps in Marginally Unstable Laplacian System. StabL outperforms other algorithms by a significant margin.

Algorithm	**Average Regret**	**Top 90%**	**Top 75%**	**Top 50%**
StabL	$\mathbf{1.5 \times 10^4}$	$\mathbf{1.3 \times 10^4}$	$\mathbf{1.1 \times 10^4}$	$\mathbf{8.9 \times 10^3}$
OFULQ	6.2×10^{10}	4.0×10^6	3.5×10^5	4.7×10^4
CEC-Fix	3.7×10^{10}	2.1×10^4	1.9×10^4	1.7×10^4
CEC-Dec	4.6×10^4	4.0×10^4	3.5×10^4	2.8×10^4

The proofs and the exact expressions are presented in Appendix B.1.5. Here, we provide a proof sketch. The regret decomposition leverages the optimistic controller design. Recall that for the early improved exploration, StabL applies independent perturbations through the controller yet still deploys the optimistic policy. Thus, we consider this external perturbation as a part of the underlying system and study the regret obtained by the improved exploration strategy separately.

In particular, denote the system evolution noise at time t as ζ_t. For $t \leq T_w$, system evolution noise can be considered as $\zeta_t = B_* \nu_t + w_t$ and for $t > T_w$, $\zeta_t = w_t$. We denote the optimal average cost of system $\tilde{\Theta}$ under ζ_t as $J_*(\tilde{\Theta}, \zeta_t)$. Using the Bellman optimality equation for LQR [28], we consider the system evolution of the optimistic system $\tilde{\Theta}_t$ using the optimistic controller $K(\tilde{\Theta}_t)$ in parallel with the true system evolution of Θ_* under $K(\tilde{\Theta}_t)$ such that they share the same process noise (see details in Appendix B.1.5). Using the confidence set construction, optimistic policy, Lemma 3.5, Assumption 3.2, and Lemma 3.1, we get a regret decomposition and bound each term separately.

At a high level, the exact regret expression has a constant regret term due to early additional exploration for T_w time-steps with exponential dimension dependency and a term that scales with the square root of the duration of stabilizing adaptive control with polynomial dimension dependency, i.e., $(n + d)^{n+d} T_w + \text{poly}(n, d)\sqrt{T - T_w}$. Note that T_w is a problem-dependent expression. Thus, for large enough T, the polynomial dependence dominates, giving Theorem 3.2.

3.2.3 Experiments

In this section, we evaluate the performance of StabL in three adaptive control tasks: **(1)** a marginally unstable Laplacian system [71], **(2)** the longitudinal flight control of Boeing 747 with linearized dynamics [123], and **(3)** a stabilizable but not controllable linear dynamical system. For each task, we compare StabL with three RL algorithms: (i) OFULQ of Abbasi-Yadkori and Szepesvári [2]; (ii) certainty

Table 3.3: Maximum State Norm in the Laplacian System.

Algorithm	**Average max$\|\mathbf{x}\|_2$**	**Worst 5%**	**Worst 10%**	**Worst 25%**
StabL	$\mathbf{1.3 \times 10^1}$	$\mathbf{2.2 \times 10^1}$	$\mathbf{2.1 \times 10^1}$	$\mathbf{1.9 \times 10^1}$
OFULQ	9.6×10^3	1.8×10^5	9.0×10^4	3.8×10^4
CEC-Fix	3.3×10^3	6.6×10^4	3.3×10^4	1.3×10^4
CEC-Dec	2.0×10^1	3.5×10^1	3.3×10^1	2.9×10^1

equivalent controller with fixed isotropic perturbations (CEC-Fix), which is the standard baseline in control theory; and (iii) certainty equivalent controller with decaying isotropic perturbations (CEC-Dec), which is shown to *achieve optimal regret with a given initial stabilizing policy* [71, 191, 242]. In the implementation of CEC-Fix and CEC-Dec, the optimal control policies of the estimated model are deployed. Furthermore, in finding the optimistic parameters for StabL and OFULQ, we use projected gradient descent within the confidence sets. We perform 200 independent runs for each algorithm for 200 time steps starting from $x_0 = 0$. We present the performance of the best parameter choices for each algorithm. For further details and the experimental results please refer to [166].

Before discussing the experimental results, we would like to highlight the baseline choices. Unfortunately, there are only a few works in literature that consider RL in LQRs without a stabilizing controller. These works are OFULQ of [2], [7], and [56]. Among these, [56] considers LQRs with adversarial noise setting and deploys *impractically large inputs*, e.g., 10^{28} for task **(1)**, whereas the algorithm of [7] only works in the scalar setting. These prohibit meaningful regret and stability comparisons, thus, we compare StabL against the only relevant comparison of OFULQ among these. Moreover, there are only a few limited experimental studies in the literature of RL in LQRs. Among these, [71, 83, 85] highlight the superior performance of CEC-Dec. Therefore, we compare StabL against CEC-Dec with the best-performing parameter choice, as well as the standard baseline of CEC-Fix.

(1) Laplacian system (Appendix I.1 of [166]). Table 3.2 provides the regret performance for the average, top 90%, top 75%, and top 50% of the runs of the algorithms. We observe that StabL attains at least an order of magnitude improvement in regret over OFULQ and CECs. This setting combined with the unstable dynamics is challenging for the *solely* optimism-based learning algorithms. Our empirical study indicates that, at the early stages of learning, the smallest eigenvalue of the design matrix V_t for OFULQ is much smaller than that of StabL as shown in Figure 3.1. The early improved exploration strategy helps StabL achieve linear

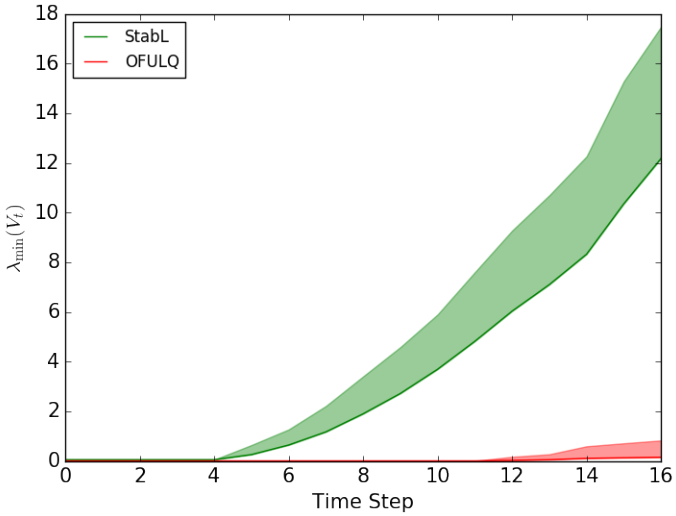

Figure 3.1: Evolution of the smallest eigenvalue of the design matrix for StabL and OFULQ in the Laplacian system. The solid line is the mean and the shaded region is one standard deviation. StabL attains linear scaling whereas OFULQ suffers from the lack of early exploration.

scaling in $\lambda_{\min}(V_t)$, thus persistence of excitation and identification of stabilizing controllers. In contrast, the only OFU-based controllers of OFULQ fail to achieve persistence of excitation and accurate estimate of the model parameters. Therefore, due to lack of reliable estimates and the skewed cost, OFULQ cannot design effective strategies to learn model dynamics and results in unstable dynamics (see Table 3.3). Table 3.3 displays the stabilization capabilities of the deployed RL algorithms. In particular, it provides the averages of the maximum norms of the states for all runs, the worst 5%, 10% and 25% runs. Of all algorithms, StabL keeps the state smallest.

(2) Boeing 747 (Appendix I2 of [166]). In practice, nonlinear systems, like Boeing 747, are modeled via local linearizations which hold as long as the states are within a certain region. Thus, to maintain the validity of such linearizations, the state of the underlying system must be well-controlled, i.e., stabilized. Table 3.4 provides the regret performances and Table 3.5 displays the stabilization capabilities of the deployed RL algorithms similar to **(1)**. Once more, among all algorithms, StabL maintains the maximum norm of the state smallest and operates within the smallest radius around the linearization point of origin.

(3) Stabilizable but not controllable system (Appendix I4 of [166]). We consider

Table 3.4: Regret Performance after 200 Time Steps in Boeing 747 Flight Control.

Algorithm	**Average Regret**	**Top 90%**	**Top 75%**	**Top 50%**
StabL	$\mathbf{1.3 \times 10^4}$	$\mathbf{9.6 \times 10^3}$	$\mathbf{7.6 \times 10^3}$	$\mathbf{5.3 \times 10^3}$
OFULQ	1.5×10^8	9.9×10^5	5.6×10^4	8.9×10^3
CEC-Fix	4.8×10^4	4.5×10^4	4.3×10^4	3.9×10^4
CEC-Dec	2.9×10^4	2.5×10^4	2.2×10^4	1.9×10^4

Table 3.5: Maximum State Norm in Boeing 747 Control.

Algorithm	**Average max$\|\mathbf{x}\|_2$**	**Worst 5%**	**Worst 10%**	**Worst 25%**
StabL	$\mathbf{3.4 \times 10^1}$	$\mathbf{7.5 \times 10^1}$	$\mathbf{7.0 \times 10^1}$	$\mathbf{5.2 \times 10^1}$
OFULQ	1.6×10^3	2.2×10^4	1.4×10^4	6.3×10^3
CEC-Fix	5.0×10^1	7.8×10^1	7.3×10^1	6.5×10^1
CEC-Dec	4.6×10^1	8.0×10^1	7.3×10^1	6.3×10^1

Table 3.6: Regret after 200 Time Steps in Stabilizable but Not Controllable System.

Algorithm	**Average Regret**	**Top 90%**	**Top 75%**	**Top 50%**
StabL	$\mathbf{1.68 \times 10^6}$	$\mathbf{7.21 \times 10^5}$	$\mathbf{3.72 \times 10^5}$	$\mathbf{1.29 \times 10^5}$
OFULQ	5.20×10^{12}	8.27×10^{11}	2.13×10^{11}	4.51×10^{10}
CEC w/t Decay	1.56×10^7	9.75×10^6	5.96×10^6	2.33×10^6

the online LQ control problem with the following parameters:

$$A_* = \begin{bmatrix} -2 & 0 & 1.1 \\ 1.5 & 0.9 & 1.3 \\ 0 & 0 & 0.5 \end{bmatrix}, \quad B_* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, Q = I, \quad R = I, \quad w \sim \mathcal{N}(0, I), \tag{3.9}$$

This problem is particularly challenging in terms of system identification and controller design since the system is not controllable but stabilizable. As expected, besides StabL, which is tailored for the general stabilizable setting, other algorithms perform poorly in this challenging setting. In particular, CEC-Fix drastically blows the state up due to significantly unstable dynamics for the uncontrollable part of the system. Therefore, the performances of only StabL, OFULQ, and CEC-Dec are presented. Table 3.6 provides the regret of the algorithms after 200 time steps, while Table 3.7 displays the maximum norm of the state during the execution of the algorithms. This setting is where OFULQ fails dramatically due to not being tailored for the stabilizable systems. The evolution of the regret performance of the algorithms is provided in Figure 3.2. Note that Figure 3.2 is in a semi-log scale. StabL provides an order-of-magnitude improved regret compared to the best performing state-of-the-art baseline CEC-Dec.

Table 3.7: Maximum State Norm in Stabilizable but Not Controllable System.

Algorithm	**Average** $\max \|x\|_2$	**Worst 5%**	**Worst 10%**	**Worst 25%**
StabL	$\mathbf{3.02 \times 10^2}$	$\mathbf{1.04 \times 10^3}$	$\mathbf{8.88 \times 10^2}$	$\mathbf{6.68 \times 10^2}$
OFULQ	4.39×10^5	3.10×10^6	2.40×10^6	1.39×10^6
CEC w/t Decay	1.37×10^3	4.07×10^3	3.54×10^3	2.78×10^3

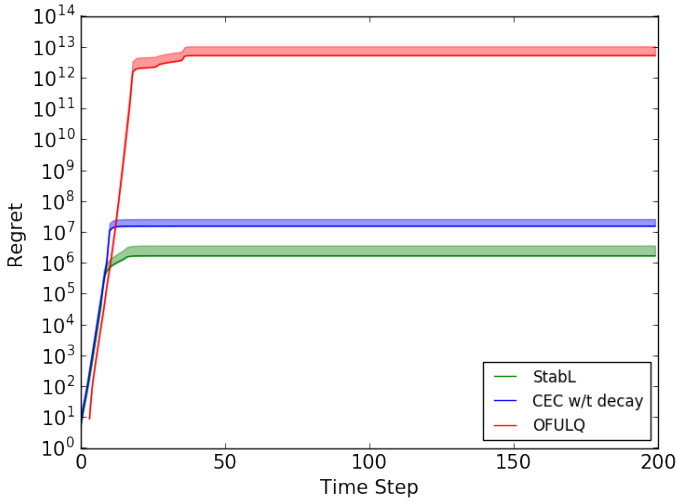

Figure 3.2: Regret Comparison of three algorithms in controlling a stabilizable but not controllable system (3.9). The solid lines are the average regrets and the shaded regions are the quarter standard deviations.

3.3 Thompson Sampling-Based Adaptive Control

Even though in the prior section we showed that StabL achieves optimal regret without a stabilizing controller, the adaptive control procedure of StabL requires solving a non-convex optimization problem to find the optimistic controller. Unfortunately, finding the optimistic parameters among the plausible models is an NP-hard problem in general, and requires computational heuristics for large-scale dynamical systems [9]. This computational inefficiency severely limits the practicality of the optimistic controller design approach. Even though [8] recently proposed a relaxation to the optimistic controller computation, which makes the optimism-based controllers efficient, their approach requires significantly well-refined model estimates and a given initial stabilizing policy similar to certainty equivalence-based controllers discussed before.

This section provides a computationally efficient adaptive control algorithm alternative to StabL with similar regret upper bound guarantees. In particular, we study Thompson Sampling-based Adaptive Control (TSAC), which attains $\tilde{O}(\sqrt{T})$ regret

in multidimensional stabilizable LQRs. This makes TSAC the first efficient adaptive control algorithm to achieve order-optimal regret in all stabilizable LQRs without the prior knowledge of a stabilizing policy, thereby solving the open problem posed in [7]. We empirically demonstrate the performance of TSAC and compare it to StabL and TS-based methods that do not require an initial stabilizing policy in flight control of Boeing 747 with linearized dynamics considered in the previous section. We show that TSAC effectively explores the system to find a stabilizing policy and achieves competitive regret performance while being computationally feasible.

The design of TSAC and our regret guarantee hinge on three important pieces missing in prior works: fixed policy update rule, improved exploration in early stages of adaptive control, and a novel lower bound that shows that TS samples optimistic parameters with non-zero probability in multidimensional LQRs. Unlike the frequent policy update rule of Abeille and Lazaric [7] in scalar LQRs, TSAC updates its policy with fixed time periods. This policy update rule prevents fast policy changes that would cause state blow-ups in stabilizable LQRs.

Due to the results of StabL, at the beginning of agent-environment interaction, TSAC focuses on quickly finding a stabilizing controller to avoid state blow-ups due to the lack of a known initial stabilizing policy. By using isotropic exploration in the early stages along with the exploration of TS policy, we show that TSAC achieves fast stabilization. After stabilizing the unknown system dynamics, TSAC relies on the effective exploration of the TS to find desirable controllers. In particular, we show that the TS samples optimistic parameters with a constant probability in any LQR setting. This novel lower bound shows that the TS is an efficient alternative to optimism in all adaptive control problems in LQRs. Combining this lower bound with the fixed policy update rule, we derive the optimal regret guarantee for TSAC.

3.3.1 TSAC

We present TSAC, a sample efficient TS-based adaptive control algorithm for the unknown stabilizable LQRs. The algorithm is summarized in Algorithm 6. It has two phases: 1) TS with improved exploration and 2) stabilizing TS.

TS with Improved Exploration

Due to the lack of an a priori known stabilizing controller, TSAC focuses on rapidly learning stabilizing controllers in the early stages of the algorithm. To achieve this, TSAC explores the system dynamics effectively in this phase. At any time-step

Algorithm 6 TSAC

Input: κ, γ, Q, R, σ_w^2 , $V_0 = \mu I$, $\hat{\Theta}_0 = 0$
for $i = 0, 1, \ldots$ **do**
  Estimate $\hat{\Theta}_i$ using (3.7) & Sample $\tilde{\Theta}_i = \mathcal{R}_{\mathcal{S}}(\hat{\Theta}_i + \beta_t V_t^{-1/2}\eta_t)$
  for $t = i\tau_0, \ldots, (i+1)\tau_0 - 1$ **do**
    if $t \leq T_w$ **then**
      Deploy $u_t = K(\tilde{\Theta}_i)x_t + \nu_t$ ▷ TS with Improved Explorationn
    else
      Deploy $u_t = K(\tilde{\Theta}_i)x_t$ ▷ Stabilizing TS

t, given the RLS estimate $\hat{\Theta}_t$ and the design matrix V_t as described in (3.8) for StabL, TSAC samples a perturbed model parameter $\tilde{\Theta}_t = \mathcal{R}_{\mathcal{S}}(\hat{\Theta}_t + \beta_t(\delta)V_t^{-1/2}\eta_t)$, where $\mathcal{R}_{\mathcal{S}}$ denotes the rejection sampling operator associated with the set $\mathcal{S}$ given in Assumption 3.2 and $\eta_t \in \mathbb{R}^{(n+d)\times n}$ is a matrix with independent standard normal entries. Here $\mathcal{R}_{\mathcal{S}}$ guarantees that $\tilde{\Theta}_t \in \mathcal{S}$ and $\beta_t(\delta)V_t^{-1/2}\eta_t$ randomizes the sampled parameter coherently with the RLS estimate and the uncertainty associated with it. Using this sampled model parameter, TSAC constructs the optimal linear controller $\bar{u}_t = K(\tilde{\Theta}_t)x_t$ for $\tilde{\Theta}_t$.

However, to obtain stabilizing controllers for an unknown linear dynamical system, one needs to explore the state-space in all directions, Lemma 3.3. Unfortunately, due to the lack of reliable estimates in the early stages, deploying the policy achieved via TS, $\bar{u}_t$, may not achieve such effective exploration. Therefore, in the early stages of interactions with the underlying system, TSAC deploys isotropic perturbations along with the sampled policy. In particular, for the first T_w time-steps, TSAC uses $u_t = \bar{u}_t + \nu_t$ as the control input where $\nu_t \sim \mathcal{N}(0, 2\kappa^2\sigma_w^2 I)$. This improved exploration policy effectively excites and explores all dimensions of the system to certify the design of stabilizing controllers. TSAC sets T_w such that all the sampled controllers $K(\tilde{\Theta}_t)$ are guaranteed to stabilize the underlying system Θ_* for all $t > T_w$ (Appendix B.2.1).

Unlike most of the popular RL strategies that follow lazy updates, TSAC updates its sampled policy in every fixed τ_0 steps, i.e., the same sampled policy $K(\tilde{\Theta}_t)$ is deployed for τ_0 time-steps. This update rule is carefully chosen such that TSAC samples enough optimistic policies to reduce the cumulative regret and avoids too frequent policy changes which would cause state blow-ups.

Stabilizing TS

After guaranteeing the design of stabilizing policies with improved exploration in the first phase, TSAC starts the adaptive control with only TS. In particular, for the remaining time-steps, TSAC deploys $u_t = K(\tilde{\Theta}_t)x_t$ for $\tilde{\Theta}_t = \mathcal{R}_{\mathcal{S}}(\hat{\Theta}_t + \beta_t(\delta)V_t^{-1/2}\eta_t)$ and updates the sampled model parameter in every τ_0 time-steps. Note that, even though all the policies during this phase are stabilizing, frequent policy changes can still cause undesirable state growth. TSAC prevents this possibility by applying the same control policy for τ_0 time-steps in this phase as well. During this phase, TSAC decays the possible state blow-ups in the first phase and maintains stable dynamics.

3.3.2 Theoretical Analysis of TSAC

In this section, we study the theoretical guarantees of TSAC. For simplicity of presentation, we consider the Gaussian process noise for the system dynamics. In particular, we assume that there exists a filtration $\mathcal{F}_t$ such that for all $t \geq 0$, x_t, z_t are $\mathcal{F}_t$-measurable and $w_t|\mathcal{F}_t = \mathcal{N}(0, \sigma_w^2 I)$ for some known $\sigma_w > 0$. The following results can be extended to sub-Gaussian process noise setting, i.e., Assumption 3.1, using the techniques developed in the previous section (see Lemma 3.1 and its proof in Appendix B.1.1). The following states the first order-optimal frequentist regret bound for TS in multidimensional stabilizable LQRs, our main result.

Theorem 3.3 (Regret of TSAC)**.** *Suppose Assumption 3.2 holds and set* $\tau_0 = 2\gamma^{-1}\log(2\kappa\sqrt{2})$ *and* $T_0 = \mathrm{poly}(\log(1/\delta), \sigma_w^{-1}, n, d, \bar{\alpha}, \gamma^{-1}, \kappa)$*. Then, for long enough* T*, TSAC achieves the regret* $R_T = \widetilde{O}\left((n+d)^{(n+d)}\sqrt{T\log(1/\delta)}\right)$ *w.p. at least* $1 - 10\delta$*, if* $T_w = \max\left(T_0,\, c_1(\sqrt{T}\log T)^{1+o(1)}\right)$ *for a constant* $c_1 > 0$*. Furthermore, if the closed loop matrix of the optimally controlled underlying system,* $A_{c,*} := A_* + B_*K_*$*, is non-singular, i.e.,* A_* *is non-singular, w.p. at least* $1 - 10\delta$*, TSAC achieves the regret* $R_T = \widetilde{O}\left(\mathrm{poly}(n,d)\sqrt{T\log(1/\delta)}\right)$ *if* $T_w = \max\left(T_0,\, c_2(\log T)^{1+o(1)}\right)$ *for a constant* $c_2 > 0$*.*

This makes TSAC the *first efficient* adaptive control algorithm that achieves optimal regret in adaptive control of all LQRs *without an initial stabilizing policy*. To prove this result, we follow a similar approach as StabL in the previous section and [7], and define the high probability joint event $E_t = \hat{E}_t \cap \tilde{E}_t \cap \bar{E}_t$, where $\hat{E}_t$ states that the RLS estimate $\hat{\Theta}$ concentrates around Θ_*, $\tilde{E}_t$ states that the sampled parameter $\tilde{\Theta}$ concentrates around $\hat{\Theta}$, and $\bar{E}_t$ states that the state remains bounded respectively. Conditioned on this event, we decompose the frequentist regret as

$R_T \mathbb{1}_{E_T} \leq R_{T_w}^{\text{exp}} + R_T^{\text{RLS}} + R_T^{\text{mart}} + R_T^{\text{TS}} + R_T^{\text{gap}}$, where $R_{T_w}^{\text{exp}}$ accounts for the regret attained due to improved exploration, R_T^{RLS} represents the difference between the value function of the true next state and the predicted next state, R_T^{mart} is a martingale with bounded difference, R_T^{TS} measures the difference in optimal average expected cost between the true model Θ_* and the sampled model $\tilde{\Theta}$, and R_T^{gap} measures the regret due to policy changes. The decomposition and expressions are given in Appendix B.2.3. In the analysis, we bound each term separately (Appendix B.2.4). Note that R_T^{RLS} and R_T^{mart} appear in the regret analysis of StabL due to algorithmic and problem setting construction, thus, follow directly from the prior analysis. Before discussing the further details of the analysis, we first consider the prior works that use TS for adaptive control of LQRs and discuss their shortcomings. Further, we highlight the challenges in adaptive control of multidimensional stabilizable LQRs using TS and present our approaches to overcome these.

Prior Work on TS-based Adaptive Control and Challenges

For the frequentist regret minimization problem, the state-of-the-art adaptive control algorithm that uses TS is Abeille and Lazaric [7]. They consider the "contractible" LQR systems, i.e. $|A_* + B_* K(\Theta_*)| < 1$, and provide $\tilde{O}(\sqrt{T})$ regret upper bound for scalar LQRs, i.e. $n = d = 1$. Notice that the set of contractible systems is a small subset of the set $\mathcal{S}$ defined in Assumption 3.2 and they are only equivalent for scalar systems since $\rho(A_* - B_* K(\Theta_*)) = |A_* - B_* K(\Theta_*)|$. This simplified setting allow them to reduce the regret analysis into the trade-off between $R_T^{\text{TS}} = \sum_{t=0}^{T} \{J(\tilde{\Theta}_t) - J(\Theta_*)\}$ and $R_T^{\text{gap}} = \sum_{t=0}^{T} \mathbb{E}[x_{t_1}^\top (P(\tilde{\Theta}_{t+1}) - P(\tilde{\Theta}_t) x_{t+1} \,|\, \mathcal{F}_t]$.

These regret terms are central in the analysis of several adaptive control algorithms. In the certainty equivalent control approaches, R_T^{TS} is bounded by the quadratic scaling of model estimation error after a significantly long exploration with a known stabilizing controller [191, 242]. In the optimism-based algorithms such as StabL, R_T^{TS} is bounded by 0 by design [2, 81]. Similarly, in the Bayesian regret setting, [212] assume that the underlying parameter Θ_* comes from a known prior that the expected regret is computed with respect to. This true prior yields $\mathbb{E}[R_T^{\text{TS}}] = 0$ in certain restrictive LQRs. Whereas the conventional approach in the analysis of R_T^{gap} is to have lazy policy updates, i.e., $O(\log T)$ policy changes such as StabL, via doubling the determinant of V_t or exponentially increasing epoch durations [48, 85].

On the other hand, Abeille and Lazaric [7] bound R_T^{TS} by showing that TS samples the optimistic parameters, $\tilde{\Theta}_t$ such that $J(\tilde{\Theta}_t) \leq J(\Theta_*)$, with a constant probability,

which reduces the regret of non-optimistic steps. Unlike the conventional policy update approaches, the key idea in Abeille and Lazaric [7] is to update the control policy every time-steps via TS, which increases the number of optimistic policies during the execution. They show that while this frequent update rule reduces R_T^{TS}, it only results with $R_T^{\mathrm{gap}} = \widetilde{O}(\sqrt{T})$. However, they were only able to show that this constant probability of optimistic sampling holds for scalar LQRs.

The difficulty of the analysis for the probability of optimistic parameter sampling lies in the challenging characterization of the optimistic set. Since $J(\tilde{\Theta}) = \sigma_w^2 \operatorname{tr}(P(\tilde{\Theta}))$, one needs to consider the spectrum of $P(\tilde{\Theta})$ to define optimistic models, which makes the analysis difficult. In particular, decreasing the cost along one direction may result in an increase in other directions. However, for the scalar LQR setting considered in Abeille and Lazaric [7], $J(\tilde{\Theta}) = P(\tilde{\Theta})$ and using standard perturbation results on DARE suffices. As mentioned in Abeille and Lazaric [7], one can naively consider the surrogate set of being optimistic in all directions, i.e., $P(\tilde{\Theta}) \preccurlyeq P(\Theta_*)$. Nevertheless, this would result in a probability that decays linearly in time and does not yield sub-linear regret. In this study, we propose new surrogate sets to derive a lower bound on the probability of having optimistic samples and show that TS in fact samples optimistic model parameters with constant probability.

In designing TS-based adaptive control algorithms for multidimensional stabilizable LQRs, one needs to maintain a bounded state. In bounding the state, Abeille and Lazaric [7] rely on the fact that the underlying system is contractive, $\|\tilde{A} + \tilde{B}K(\tilde{\Theta})\| < 1$. However, under Assumption 3.2, even if the optimal policy of the underlying system is chosen by the learning agent, the closed-loop system may not be contractive since for any symmetric matrix M, $\rho(M) \leq \|M\|$. Thus, to avoid dire consequences of unstable dynamics, TS-based adaptive control algorithms should focus on finite-time stabilization of the system dynamics in the early stages.

Moreover, the lack of contractive closed-loop mappings in stabilizable LQRs prevents frequent policy changes used in Abeille and Lazaric [7]. From the definition of (κ, γ)-stabilizability, for any stabilizing controller K', we have that $A_* + B_* K' = H' L H'^{-1}$, with $\|L\| < 1$ for some similarity transformation H'. Thus, as noted in the analysis of StabL, even if all the policies are stabilizing, changing the policies at every time step could cause couplings of these similarity transformations and result in linear growth of the state over time. Thus, TS-based adaptive control algorithms need to find the balance in the rate of policy updates, so that frequent policy switches are avoided, yet, enough optimistic policies are sampled. In light of

these observations, our results hinge on the following:

1) Improved exploration that allows fast stabilization of the dynamics;

2) Fixed policy update rule that prevents state blow-up and reduces R_T^{gap} and R_T^{TS};

3) A novel result that shows TS samples optimistic model parameters with a constant probability for multidimensional LQRs and gives a novel bound on R_T^{TS}.

Details of the analysis

The improved exploration along with TS in the early stages allows TSAC to effectively explore the state space in all directions. The following shows that for a long enough improved exploration phase, TSAC achieves consistent model estimates and guarantees the design of stabilizing policies.

Lemma 3.6 (Model Estimation Error and Stabilizing Policy Design)**.** *Suppose Assumption 3.2 holds. For* $t \geq 200(n+d)\log\frac{12}{\delta}$ *time-steps of TS with improved exploration, with probability at least* $1-2\delta$*, TSAC obtains model estimates such that* $\|\hat{\Theta}_t - \Theta_*\|_2 \leq 7\beta_t(\delta)/(\sigma_w\sqrt{t})$*. Moreover, after* $T_w \geq T_0 := \text{poly}(\log(1/\delta), \sigma_w^{-1}, n, d, \bar{\alpha}, \gamma^{-1}, \kappa)$ *length TS with improved exploration phase, with probability at least* $1-3\delta$*, TSAC samples controllers* $K(\tilde{\Theta}_t)$ *such that the closed-loop dynamics on* Θ_* *is* $(\kappa\sqrt{2}, \gamma/2)$ *strongly stable for all* $t > T_w$*, i.e., there exists* L *and* $H \succ 0$ *such that* $A_* + B_*K(\tilde{\Theta}_t) = HLH^{-1}$*, with* $\|L\| \leq 1-\gamma/2$ *and* $\|H\|\|H^{-1}\| \leq \kappa\sqrt{2}$*.*

The proof and the precise expression of T_w can be collected in Appendix B.2.1. In the proof, we show that the inputs $u_t = K(\tilde{\Theta}_i)x_t + \nu_t$ for $\nu_t \sim \mathcal{N}(0, 2\kappa^2\sigma_w^2 I)$ guarantee the persistence of excitation with high probability, i.e., the smallest eigenvalue of the design matrix V_t scales linearly over time. Combining this result, with the confidence set construction in (3.8), we derive the first result. Using the first result and the fact that there exists a stabilizing neighborhood around the model parameter Θ_*, such that all the optimal linear controllers of the models within this region stabilize Θ_*, we derive the final result. Due to early improved exploration, TSAC stabilizes the system dynamics after T_w samples and starts stabilizing adaptive control with only TS. Using the stabilizing controllers for fixed $\tau_0 = 2\gamma^{-1}\log(2\kappa\sqrt{2})$ time-steps, TSAC decays the state magnitude and remedy possible state blow-ups in the first phase. To study the boundedness of state, define $T_r = T_w + (n+d)\tau_0\log(n+d)$. The following shows that the state is bounded and well-controlled.

Lemma 3.7 (Bounded states)**.** *Suppose Assumption 3.2 holds. For given T_w and T_r, TSAC controls the state such that $\|x_t\| = O((n+d)^{n+d})$ for $t \le T_r$, with probability at least $1-3\delta$ and $\|x_t\| \le (12\kappa^2+2\kappa\sqrt{2})\gamma^{-1}\sigma_w\sqrt{2n\log(n(t-T_w)/\delta)}$ for $T \ge t > T_r$, with probability at least $1-4\delta$.*

This result is a trivial extension of Lemma 3.5 for StabL since rejection sampling guarantees that the sampled model is an element of $\mathcal{S}$, thus, it is $(\kappa,\gamma)-$stabilizable by its corresponding optimal controller, $1-\gamma \ge \max_{t\le T}\rho(\tilde{A}_t + \tilde{B}_t K(\tilde{\Theta}_t))$. Using this fact, following the proof Lemma 3.5 in Appendix B.1.3 one can show that for $t \le T_r$, deploying the same policy for τ_0 time-steps in the first phase maintains a well-controlled state except for $n+d$ time-steps, under the high probability event of $\hat{E}_t \cap \tilde{E}_t$. For bounding the state after $t > T_r$, the proof of Lemma 3.5 follows directly such that after $(n+d)\log(n+d)$ policy updates, the state is well-controlled and brought to equilibrium. This result shows that the joint event $E_t = \hat{E}_t \cap \tilde{E}_t \cap \bar{E}_t$ holds with probability at least $1-4\delta$ for all $t \le T$.

Conditioned on this event, we analyze the regret terms individually (Appendix B.2.4). We show that with probability at least $1-\delta$, $R^{\text{exp}}_{T_w}$ yields $\widetilde{O}((n+d)^{n+d}T_w)$ regret due to isotropic perturbations. R^{RLS}_T and R^{mart}_T are $\widetilde{O}((n+d)^{n+d}\sqrt{T_r} + \text{poly}(n,d)\sqrt{T-T_r})$ with probability at least $1-\delta$ due to standard arguments based on the event E_T. More importantly, conditioned on the event E_T, we prove that $R^{\text{gap}}_T = \widetilde{O}((n+d)^{n+d}\sqrt{T_r} + \text{poly}(n,d)\sqrt{T-T_r})$ with probability at least $1-2\delta$, and $R^{\text{TS}}_T = \widetilde{O}(nT_w + \text{poly}(n,d)\sqrt{T-T_w})$ with probability at least $1-2\delta$, whose analyses require several novel fundamental results.

To bound on R^{gap}_T, we extend the results in Abeille and Lazaric [7] to multidimensional stabilizable LQRs and incorporate the slow update rule and the early improved exploration. We show that while TSAC enjoys well-controlled state with polynomial dimension dependency on regret due to slow policy updates, it also maintains the desirable $\widetilde{O}(\sqrt{T})$ regret of frequent updates with only a constant τ_0 scaling. As discussed before, bounding R^{TS}_T requires selecting optimistic models with constant probability, which has been an open problem in the literature for multidimensional systems. In this study, we provide a solution to this problem and show that TS indeed selects optimistic model parameters with a constant probability for multidimensional LQRs. The precise statement of this result and its proof outline are given in Section 3.3.3. Leveraging this result, we derive the upper bound on R^{TS}_T. Combining all these terms yields the regret upper bound of TSAC given in Theorem 3.3.

3.3.3 Proof Outline of Sampling Optimistic Models with Constant Probability

In this section, we provide the precise statement that the probability of sampling an optimistic parameter is lower bounded by a fixed constant with high probability. Then we give the proof outline with the main steps. The complete proof with the intermediate results is given in Appendix B.2.2.

Theorem 3.4 (Optimistic probability)**.** *Let $\mathcal{F}_t^{cnt} := \sigma(F_{t-1}, x_t)$ be the information available to the controller up to time t. Denote the optimistic set by $\mathcal{S}^{\text{opt}} := \left\{\Theta \in \mathbb{R}^{(n+d)\times n} \;\middle|\; J(\Theta) \leq J(\Theta_*)\right\}$. If $T_w = cn^2(\sqrt{T}\log T)^{1+o(1)}$ for a constant $c > 0$, then under the event E_T for large enough T, we have that*

$$p_t^{\text{opt}} := \mathbb{P}\left\{\tilde{\Theta}_t \in \mathcal{S}^{opt} \,\middle|\, \mathcal{F}_t^{cnt}, \hat{E}_t\right\} \geq \frac{Q(1)}{1+o(1)},$$

for any $T_r < t \leq T$. Furthermore, if the closed-loop matrix, $A_{c,*} = A_* + B_* K_*$, is non-singular, then the bound above still holds when $T_w = c(\log T)^{1+o(1)}$ for a constant $c > 0$.

Surrogate Set Definition

First, we define a surrogate subset $\mathcal{S}^{\text{surr}}$ to the optimistic set $\mathcal{S}^{\text{opt}}$. The construction of $\mathcal{S}^{\text{surr}}$ is important as the geometry of $\mathcal{S}^{\text{opt}}$ is complicated to study due to (3.3) that controls the spectrum of $P(\Theta)$.

Lemma 3.8 (Surrogate set)**.** *Let $J(\Theta, K) := \text{tr}\left((Q + K^\top R K)\Sigma(\Theta, K)\right)$ be the expected average cost of controlling a system $\Theta \in \mathcal{S}$ by a fixed stabilizing control policy $K \in \mathbb{R}^{d\times n}$ where $\Sigma(\Theta, K) := \lim_{t\to\infty} \mathbb{E}\left[x_t x_t^\top\right]$ is the covariance of the state. The following surrogate set is a subset of $\mathcal{S}^{\text{opt}}$:*

$$\mathcal{S}^{surr} := \left\{\Theta = (A,\ B)^\top \in \mathbb{R}^{(n+d)\times n} \,\middle|\, J(\Theta, K(\Theta_*)) \leq J(\Theta_*, K(\Theta_*)) = J(\Theta_*)\right\} \subset \mathcal{S}^{\text{opt}}.$$

Note that $\Sigma(\Theta, K)$ satisfies the Lyapunov equation $\Sigma(\Theta, K) - \Theta^\top H_K \Sigma(\Theta, K) H_K^\top \Theta = \sigma_w^2 I$, where $H_K^\top := [I,\ K^\top]$, and $\Theta^\top H_K = A + BK$, given that K stabilizes the system Θ. We can analytically express $\Sigma(\Theta, K)$ as a converging infinite sum $\Sigma(\Theta, K) = \sigma_w^2 \sum_{t=0}^{\infty} (A+BK)^t (A^\top + K^\top B^\top)^t$ [132]. Using the properties of the trace operator, one can write $J(\Theta, K(\Theta_*)) = L(\Theta^\top H_*)$, where $L(A_c) := \sigma_w^2 \sum_{t=0}^{\infty} \left\|A_c^t\right\|_{Q_*}^2$ for any stable matrix A_c, $Q_* := Q + K(\Theta_*)^\top R K(\Theta_*)$, and $H_*^\top := [I, K(\Theta_*)^\top]$.

Therefore, we can lower bound the probability of being optimistic as

$$p_t^{\text{opt}} \geq \mathbb{P}\left\{\tilde{\Theta}_t \in \mathcal{S}^{\text{surr}} \,\middle|\, \mathcal{F}_t^{\text{cnt}}, \hat{E}_t\right\} = \mathbb{P}\left\{L(\tilde{\Theta}_t^\top H_*) \leq L(\Theta_*^\top H_*) \,\middle|\, \mathcal{F}_t^{\text{cnt}}, \hat{E}_t\right\}$$

$$\geq \min_{\hat{\Theta} \in \mathcal{E}_t^{\text{RLS}}} \mathbb{P}_t\{L(\hat{\Theta}^\top H_* + \eta^\top \beta_t V_t^{-\frac{1}{2}} H_*) \leq L(\Theta_*^\top H_*)\} \tag{3.10}$$

$$= \min_{\hat{\Theta} \in \mathcal{E}_t^{\text{RLS}}} \mathbb{P}_t\{L(\hat{\Theta}^\top H_* + \Xi\sqrt{F_t}) \leq L(\Theta_*^\top H_*)\}, \tag{3.11}$$

where $\mathbb{P}_t\{\cdot\} \coloneqq \mathbb{P}\{\cdot \mid \mathcal{F}_t^{\text{cnt}}\}$, $F_t \coloneqq \beta_t^2 H_*^\top V_t^{-1} H_*$ and Ξ is a matrix of size $n \times n$ with iid $\mathcal{N}(0,1)$ entries. Here (3.10) considers the worst possible estimate within $\mathcal{E}_t^{\text{RLS}}$ and (3.11) is the whitening transformation.

Reformulation in Terms of Closed-Loop Matrix

In the second step, we reformulate the probability of sampling optimistic parameters in terms of closed-loop system matrix $\tilde{A}_c \coloneqq \tilde{\Theta}^\top H_* = \tilde{A} + \tilde{B}K(\Theta_*)$ of the sampled system $\tilde{\Theta} = (\tilde{A}, \tilde{B})^\top$ driven by the policy $K(\Theta_*)$. Transitioning to the closed-loop formulation allows tighter bounds on the optimistic probability. To complete this reformulation, we need to construct an estimation confidence set for the closed-loop system matrix $\hat{A}_c \coloneqq \hat{\Theta}^\top H_* = \hat{A} + \hat{B}K(\Theta_*)$ of the RLS-estimated system $\hat{\Theta} = (\hat{A}, \hat{B})^\top$ and show that the constructed confidence set is a superset to $\mathcal{E}_t^{\text{RLS}}$.

Lemma 3.9 (Closed-loop confidence)**.** *Let $F_t(\delta) \coloneqq \beta_t^2(\delta) H_*^\top V_t^{-1} H_*$. For any $t \geq 0$, define by*

$$\mathcal{E}_t^{cl}(\delta) \coloneqq \left\{\hat{\Theta} \in \mathbb{R}^{(n+d)\times n} \;\middle|\; \operatorname{tr}\left[(\hat{\Theta}^\top H_* - \Theta_*^\top H_*) F_t^{-1}(\delta)(\hat{\Theta}^\top H_* - \Theta_*^\top H_*)^\top\right] \leq 1\right\},$$

the closed-loop confidence set. Then, for all times $t \geq 0$ and $\delta \in (0,1)$, we have that $\mathcal{E}_t^{RLS}(\delta) \subseteq \mathcal{E}_t^{cl}(\delta)$.

Note that the definition of $\mathcal{E}_t^{\text{cl}}(\delta)$ *only* involves closed-loop matrices $\hat{A}_c \coloneqq \hat{\Theta}^\top H_*$ and $A_{c,*} \coloneqq \Theta_*^\top H_*$. We can use the result of Lemma 3.9 to reformulate the probability of sampling optimistic parameters, $\tilde{\Theta} = (\tilde{A}, \tilde{B})$, as sampling optimistic closed-loop system matrices, $\tilde{A}_c$. We bound p_t^{opt} from below as

$$p_t^{\text{opt}} \geq \min_{\hat{\Theta} \in \mathcal{E}_t^{\text{cl}}} \mathbb{P}_t\{L(\hat{\Theta}^\top H_* + \Xi\sqrt{F_t}) \leq L(A_{c,*})\} \tag{3.12}$$

$$= \min_{\hat{A}_c \,:\, \|\hat{A}_c^\top - A_{c,*}^\top\|_{F_t^{-1}} \leq 1} \mathbb{P}_t\{L(\hat{A}_c + \Xi\sqrt{F_t}) \leq L(A_{c,*})\} \tag{3.13}$$

$$= \min_{\hat{\Upsilon} \,:\, \|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t\{L(A_{c,*} + \hat{\Upsilon}\sqrt{F_t} + \Xi\sqrt{F_t}) \leq L(A_{c,*})\}, \tag{3.14}$$

where (3.12) is due to Lemma 3.9 and (3.13) follows from the fact that H_* has full column rank. Observe that, in Equation (3.14), $\hat{\Upsilon}$ is a unit Frobenius norm matrix of size $n \times n$ and the term $A_{c,*} + \hat{\Upsilon}\sqrt{F_t}$ accounts for the confidence ellipsoid for the estimated closed-loop matrix, $\hat{A}_c$. The event in (3.14) corresponds to finding the closed-loop matrix, $A_{c,*} + (\Xi + \hat{\Upsilon})\sqrt{F_t}$ of the TS sampled system in the sublevel manifold $\mathcal{M}_* := \left\{A_c \in \mathcal{M}_n \mid L(A_c) \leq L(A_{c,*})\right\}$ as illustrated in Figure 3.3.

Local Geometry of Optimistic Set under Perturbations

Next, we further simplify the form of the probability in (3.14) by exploiting the local geometric structure of the function $L : A_c \mapsto \sigma_w^2 \sum_{t=0}^{\infty} \left\|A_c^t\right\|_{Q_*}^2$ defined over the set of (Schur-)stable matrices, $\mathcal{M}_{\text{Schur}} := \{A_c \in \mathcal{M}_n \mid \rho(A_c) < 1\}$. The following lemma characterizes perturbative properties of L.

Lemma 3.10 (Perturbations)**.** *The function $L : \mathcal{M}_{Schur} \to \mathbb{R}_+$ defined as $L(A_c) = \sigma_w^2 \sum_{t=0}^{\infty} \left\|A_c^t\right\|_{Q_*}^2$ is smooth in its domain. For any $A_c \in \mathcal{M}_{Schur}$, there exists $\epsilon > 0$ such that for any perturbation $\|G\|_F \leq \epsilon$, the function L admits a quadratic Taylor expansion as*

$$L(A_c + G) = L(A_c) + \nabla L(A_c) \bullet G + \frac{1}{2} G \bullet \mathcal{H}_{A_c + sG}(G) \tag{3.15}$$

for an $s \in [0, 1]$ where $\mathcal{H}_{A_c} : \mathcal{M}_n \to \mathcal{M}_n$ is the Hessian operator evaluated at a point $A_c \in \mathcal{M}_{Schur}$. In particular, we have that $\nabla L(A_{c_*}) = 2P(\Theta_*)A_{c,*}\Sigma_*$. Furthermore, there exists a constant $r > 0$ such that $\left|G \bullet \mathcal{H}_{A_c+sG}(G)\right| \leq r\|G\|_F^2$ for any $s \in [0, 1]$ and $\|G\|_F \leq \epsilon$.

Lemma 3.10 guarantees that if a perturbation is sufficiently small, the perturbed function can be locally expressed as a quadratic function of the perturbation. Since the set of stable matrices, $\mathcal{M}_{\text{Schur}}$, is globally non-convex and Taylor's theorem only holds in convex domains, we restrict the perturbations in a ball of radius $\epsilon > 0$. The fact that there is a neighborhood of stable matrices around a matrix A_c enables us to apply Taylor's theorem in this neighborhood.

Given the optimal closed-loop system matrix $A_{c,*}$, let $\epsilon_* > 0$ be chosen such that the expansion in (3.15) holds for perturbations $\|G\|_F \leq \epsilon_*$ around $A_{c,*}$. Denote the perturbation due to Thompson sampling and estimation error as $G_t = (\Xi + \hat{\Upsilon})\sqrt{F_t}$

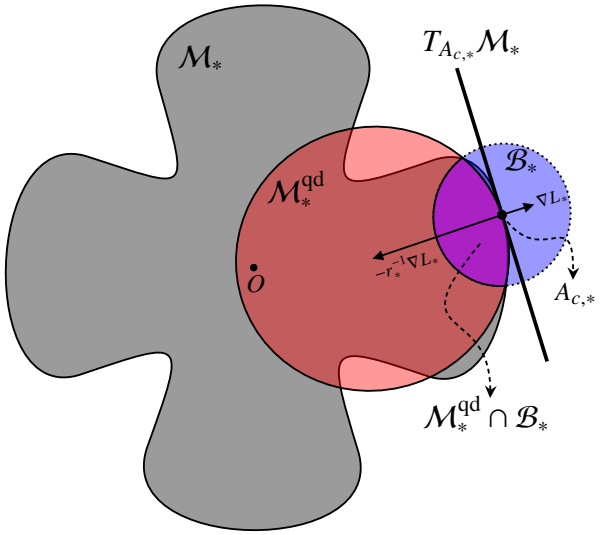

Figure 3.3: A visual representation of sublevel manifold $\mathcal{M}_*$. O is the origin and $A_{c,*}$ is the optimal closed-loop system matrix. $T_{A_{c,*}}\mathcal{M}_*$ is the tangent space to the manifold $\mathcal{M}_*$ at the point $A_{c,*}$ and ∇L_* is the Jacobian of the function L at $A_{c,*}$. $\mathcal{M}_*^{\text{qd}}$ is the sublevel manifold of the quadratic approximation to L and $\mathcal{B}_*$ is a small ball of stable matrices around $A_{c,*}$. The intersection $\mathcal{M}_*^{\text{qd}} \cap \mathcal{B}_*$ is a subset of $\mathcal{M}_*$.

and let $\|G_t\|_F \leq \epsilon_*$. Then, we can write

$$\begin{aligned} L(A_{c,*} + G_t) &= L(A_{c,*}) + \nabla L(A_{c,*}) \bullet G_t + \frac{1}{2} G_t \bullet \mathcal{H}_{A_{c,*}+sG_t}(G_t) \\ &\leq L(A_{c,*}) + \nabla L(A_{c,*}) \bullet G_t + \frac{r_*}{2}\|G_t\|_F^2, \end{aligned} \tag{3.16}$$

where $r_* > 0$ is a constant due to Lemma 3.10. Using (3.16), we have the following lower bound on (3.14),

$$p_t^{\text{opt}} \geq \min_{\hat{\Upsilon} \,:\, \|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t \Big\{ \frac{r_*}{2} \|(\Xi + \hat{\Upsilon}) F_t^{\frac{1}{2}}\|_F^2 + \nabla L_* \bullet (\Xi + \hat{\Upsilon}) F_t^{\frac{1}{2}} \leq 0, \text{ and } \|(\Xi + \hat{\Upsilon}) F_t^{\frac{1}{2}}\|_F \leq \epsilon_* \Big\}, \tag{3.17}$$

where $\nabla L_* \coloneqq \nabla L(A_{c,*})$. The event in (3.17) corresponds to finding $A_{c,*} + (\Xi + \hat{\Upsilon})\sqrt{F_t}$ at the intersection of the stable ball $\mathcal{B}_* \coloneqq \{A_c \in \mathcal{M}_n \mid \|A_c - A_{c,*}\|_F \leq \epsilon_*\}$ and the sublevel manifold $\mathcal{M}_*^{\text{qd}} \coloneqq \{A_c \in \mathcal{M}_n \mid \|A_c - A_{c,*} + r_*^{-1}\nabla L_*\|_F \leq \|r_*^{-1}\nabla L_*\|_F\}$ as illustrated in Figure 3.3.

The intersection $\mathcal{M}_*^{\text{qd}} \cap \mathcal{B}_* \subset \mathcal{M}_*$ serves as another surrogate to sublevel manifold $\mathcal{M}_*$. Switching to the new surrogate $\mathcal{M}_*^{\text{qd}}$ helps us overcome the issue of working with intractable and complicated geometry of $\mathcal{M}_*$ due to infinite sum in $L(A_c)$. We

can utilize techniques relating to Gaussian probabilities as the geometry of $\mathcal{M}_*^{\mathrm{qd}}$ is described by a quadratic form.

Final Bound

Equipped with the preceding results, we can bound the optimism probability tractably from below by the probability of a TS sampled closed-loop system matrix lying inside the intersection of two balls $\mathcal{M}_*^{\mathrm{qd}} \cap \mathcal{B}_*$ as given in (3.17). By bounding the weighted Frobenius norms in (3.17) from above by $\lambda_{\max,t}$, the maximum eigenvalue of F_t, and normalizing the matrix $\nabla L_* \sqrt{F_t}$, we can write

$$\begin{aligned} p_t^{\mathrm{opt}} &\geq \min_{\|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t \left\{ \frac{r_*}{2} \lambda_{\max,t} \|\Xi + \hat{\Upsilon}\|_F^2 + (\nabla L_* \sqrt{F_t}) \bullet (\Xi + \hat{\Upsilon}) \leq 0, \text{ and } \lambda_{\max,t} \|\Xi + \hat{\Upsilon}\|_F^2 \leq \epsilon_*^2 \right\} \\ &= \min_{\|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t \left\{ \frac{(\nabla L_* F_t^{1/2}) \bullet (\Xi + \hat{\Upsilon})}{\|\nabla L_* F_t^{1/2}\|_F} \leq \frac{-\lambda_{\max,t} r_* \|\Xi + \hat{\Upsilon}\|_F^2}{2\|\nabla L_* F_t^{1/2}\|_F}, \text{ and } \|\Xi + \hat{\Upsilon}\|_F^2 \leq \frac{\epsilon_*^2}{\lambda_{\max,t}} \right\}. \end{aligned} \tag{3.18}$$

Observe that the inner product $(\nabla L_* F_t^{1/2}) \bullet \hat{\Upsilon}$ is maximized by $\Upsilon_\# := \frac{(\nabla L_* F_t^{1/2})}{\|\nabla L_* F_t^{1/2}\|_F}$ subject to $\|\hat{\Upsilon}\|_F \leq 1$. Since the probability distribution of $\|\Xi + \hat{\Upsilon}\|_F^2$ is invariant under orthogonal transformation of Ξ and $\hat{\Upsilon}$, (3.18) also attains its minimum at $\Upsilon_\#$. Thus, we can rewrite (3.18) as

$$\begin{aligned} p_t^{\mathrm{opt}} &\geq \mathbb{P}_t \left\{ \frac{(\nabla L_* F_t^{1/2}) \bullet \Xi}{\|\nabla L_* F_t^{1/2}\|_F} + 1 \leq \frac{-\lambda_{\max,t} r_*}{2\|\nabla L_* F_t^{1/2}\|_F} \|\Xi + \Upsilon_\#\|_F^2, \text{ and } \|\Xi + \Upsilon_\#\|_F^2 \leq \frac{\epsilon_*^2}{\lambda_{\max,t}} \right\} \\ &= \mathbb{P}_t \left\{ \xi + 1 \leq -\frac{\lambda_{\max,t} r_*}{2\|\nabla L_* F_t^{1/2}\|_F} \left((\xi+1)^2 + X \right), \text{ and } (\xi+1)^2 + X \leq \frac{\epsilon_*^2}{\lambda_{\max,t}} \right\}, \end{aligned} \tag{3.19}$$

where $\xi \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_{n^2-1}$ are independent standard normal and chi-squared distributions, and (3.19) is derived by rotating Ξ so that its first element is along the direction of $\nabla L_* F_t^{1/2}$. We use the following lemma to characterize the eigenvalues of F_t and control the lower bound (3.19) on p_t^{opt}.

Lemma 3.11 (Bounded eigenvalues)**.** *Suppose $T_w = O((\sqrt{T})^{1+o(1)})$. Denote the minimum and maximum eigenvalues of F_t by $\lambda_{\min,t}$ and $\lambda_{\max,t}$, respectively. Under the event E_T, for large enough T, we have that $\lambda_{\max,t} \leq C \frac{\log T}{T_w}$ and $\frac{\lambda_{\max,t}}{\lambda_{\min,t}} \leq C \frac{T \log T}{T_w}$ for any $T_r < t \leq T$ for a constant $C = \mathrm{poly}(n, d, \log(1/\delta))$.*

Lemma 3.11 states that maximum eigenvalue and the condition number of F_t are controlled inversely by the length of initial exploration phase T_w and proportionally

by $\log T$ and $T \log T$ given that exploration time is bounded by a certain amount. The length of initial exploration T_w relative to the horizon T is critical in guaranteeing asymptotically constant optimistic probability p_t^{opt}. Although more lengthy initial exploration will lead to better convergence to constant optimistic probability, it also incurs higher asymptotic regret due to linear scaling of exploration regret with T_w.

Using the relation $\|\nabla L_* F_t^{\frac{1}{2}}\|_F \geq \max(\sigma_{\min,*}\|F_t^{\frac{1}{2}}\|_F, \lambda_{\min,t}^{\frac{1}{2}}\|\nabla L_*\|_F)$ where $\sigma_{\min,*}$ is the minimum singular value of ∇L_*, we can further bound (3.19) from below. From Lemma 3.10, we can write $\nabla L_* = 2P(\Theta_*)A_{c,*}\Sigma_*$ where $P(\Theta_*) > 0$ is the solution to the DARE in (3.3) and $\Sigma_* = \Sigma(\Theta_*, K_*) \succ 0$ is the stationary state covariance matrix. Notice that the minimum singular value of ∇L_* is positive (i.e., ∇L_* is full-rank) if and only if the closed-loop system matrix, $A_{c,*}$, is non-singular.

In general, $A_{c,*}$ can be singular. Assuming that $T_w = O((\sqrt{T})^{1+o(1)})$, under the event E_T, we can use $\|\nabla L_* F_t^{\frac{1}{2}}\|_F \geq \sqrt{\lambda_{\min,t}}\|\nabla L_*\|_F$ to obtain the following lower bound on p_t^{opt} for $T_r < t \leq T$:

$$\begin{aligned} p_t^{\text{opt}} &\geq \mathbb{P}_t\left\{\xi + 1 \leq -\frac{\sqrt{\lambda_{\max,t}}}{2\rho_*}\sqrt{\frac{\lambda_{\max,t}}{\lambda_{\min,t}}}\left((\xi+1)^2 + X\right), \text{ and } (\xi+1)^2 + X \leq \frac{\epsilon_*^2}{\lambda_{\max,t}}\right\}, \\ &\geq \mathbb{P}\left\{\xi + 1 \leq -\frac{C}{2\rho_*}\frac{\sqrt{T}\log T}{T_w}\left((\xi+1)^2 + X\right), \text{ and } (\xi+1)^2 + X \leq \frac{\epsilon_*^2 T_w}{C\log T}\right\}, \end{aligned}$$

where $\rho_* := \|r_*^{-1}\nabla L_*\|_F$. Choosing the exploration time as $T_w = \omega(\sqrt{T}\log T)$ makes the coefficients $\frac{\sqrt{T}\log T}{T_w} = o(1)$ to be very small and $\frac{T_w}{\log T}$ to be very large, leading to constant lower bound on limiting optimistic probability $\liminf_{T\to\infty} p_T^{\text{opt}} \geq \mathbb{P}\{\xi + 1 \leq 0\} =: Q(1)$.

On the other hand, if $A_{c,*}$ is non-singular, then we can use the alternative bound $\|\nabla L_* \sqrt{F_t}\|_F \geq \sigma_{\min,*}\|\sqrt{F_t}\|_F \geq \sigma_{\min,*}\sqrt{\lambda_{\max,t}}$ to obtain the following lower bound for $T_r < t \leq T$:

$$\begin{aligned} p_t^{\text{opt}} &\geq \mathbb{P}_t\left\{\xi + 1 \leq -\frac{\sqrt{\lambda_{\max,t}}}{2\sigma_{\min,*}}\left((\xi+1)^2 + X\right), \text{ and } (\xi+1)^2 + X \leq \frac{\epsilon_*^2}{\lambda_{\max,t}}\right\}, \\ &\geq \mathbb{P}\left\{\xi + 1 \leq -\frac{\sqrt{C}}{2\sigma_{\min,*}}\sqrt{\frac{\log T}{T_w}}\left((\xi+1)^2 + X\right), \text{ and } (\xi+1)^2 + X \leq \frac{\epsilon_*^2 T_w}{C\log T}\right\}. \end{aligned}$$

Similarly, choosing the exploration time as $T_w = \omega(\log T)$ makes the coefficients $\sqrt{\frac{\log T}{T_w}} = o(1)$ to be very small and $\frac{T_w}{\log T} = \omega(1)$ to be very large, leading to constant lower bound on limiting optimistic probability $\liminf_{T\to\infty} p_T^{\text{opt}} \geq Q(1)$.

Table 3.8: Regret and Maximum State Norm in Boeing 747 Flight Control.

Algorithm	Average Regret	Top 95%	Top 90%	Average $\max \|x\|_2$	Top 95%	Top 90%
TSAC	$\mathbf{4.58 \times 10^7}$	$\mathbf{1.43 \times 10^5}$	$\mathbf{9.49 \times 10^4}$	$\mathbf{1.23 \times 10^3}$	$\mathbf{1.07 \times 10^2}$	$\mathbf{9.77 \times 10^1}$
StabL	1.34×10^4	1.05×10^3	9.60×10^3	3.38×10^1	3.14×10^1	2.98×10^1
OFULQ	1.47×10^8	4.19×10^6	9.89×10^5	1.62×10^3	5.21×10^2	2.78×10^2
TS-LQR	5.63×10^{11}	3.07×10^7	5.33×10^6	6.26×10^4	1.08×10^3	6.39×10^2

In both cases, the optimistic probability achieves a constant lower bound for large enough T as $p_T^{\text{opt}} \geq Q(1)(1 + o(1))^{-1}$. This result can be interpreted in a geometric way as follows. As the time passes, the estimates of the system become more accurate in the sense that the confidence region of the estimate shrinks very quickly as controlled by the eigenvalues of F_t. Similarly, the high-probability region of TS samples also shrink very fast controlled by the covariance matrix F_t. Therefore, for large enough T, the confidence region of the model estimate and the high-probability region of TS samples get significantly smaller compared to the surrogate optimistic set $\mathcal{M}_*^{\text{qd}} \cap \mathcal{B}_*$. This size difference effectively reduces the probability of finding a sampled system in $\mathcal{M}_*^{\text{qd}} \cap \mathcal{B}_*$ to the probability of finding a sampled system in the half-space separated by the tangent space $T_{A_{c,*}} \mathcal{M}_*$.

3.3.4 Numerical Experiments

Finally, we evaluate the performance of TSAC in longitudinal flight control of Boeing 747 with linearized dynamics [123]. We compare TSAC with three adaptive control algorithms in the literature that do not require an initial stabilizing policy: (i) OFULQ of Abbasi-Yadkori and Szepesvári [2], (ii) TS-LQR of Abeille and Lazaric [7], and (iii) StabL. We perform 200 independent runs for 200 time steps for each algorithm and report their average, top 95% and top 90% regret, and maximum state norm performances. We present the performance of the best parameter choices for each algorithm. For a fair comparison, we also adopt slow policy updates in OFULQ and TS-LQR. For further details and the experimental results please refer to [166]. The results are presented in Table 3.8. Notice that TSAC achieves the second-best performance after StabL. As expected, StabL outperforms TSAC since it performs much heavier computations to find the optimistic controller in the confidence set, whereas TSAC samples optimistic parameters only with some fixed probability. However, TSAC compares favorably against both OFULQ and TS-LQR, making it the best-performing computationally efficient algorithm.

3.4 Conclusion and Future Directions

In this chapter, we proposed two learning and control algorithms StabL and TSAC that both attain optimal regret of $\tilde{O}(\sqrt{T})$ in stabilizable LQRs without an initial stabilizing policy. StabL follows the OFU principle to balance exploration and exploitation in interaction with LQRs. We showed that if an additional random exploration is enforced in the early stages of the agent's interaction with the environment, StabL has the guarantee to design a stabilizing controller sooner. We then show that while the agent enjoys the benefit of stable dynamics in further stages, the additional exploration does not alter the early performance of the agent considerably. Finally, we prove that the regret upper bound of StabL is $O(\sqrt{T})$ with polynomial dependence in the problem dimensions of the LQRs in stabilizable systems.

Using the idea of early improved exploration to reduce regret, we designed TSAC to alleviate the possible computational difficulties of StabL. TSAC follows Thompson Sampling to balance exploration and exploitation in interaction within LQRs. Quickly stabilizing the system dynamics and relying only on sampling from the confidence sets make TSAC the first efficient adaptive control algorithm that attains optimal regret of $\tilde{O}(\sqrt{T})$ in stabilizable LQRs without an initial stabilizing policy. The main technical contribution in the development of TSAC is to show that TS samples optimistic parameters with constant probability in all LQRs, thereby resolving the conjecture in Abeille and Lazaric [7] and achieving optimal regret performance, similar to StabL. Note that our numerical experiments show that TSAC performs slightly worse on regret and stabilization than StabL, corroborating our theoretical results (StabL uses an optimistic parameter whereas TSAC samples one with at least a non-zero probability). However, the computational efficiency of TSAC makes it a viable option in practice where complicated projected gradient descent surrogate of optimistic model selection of StabL is not feasible all the time.

Our results highlight the benefit of early improved exploration to achieve improved regret at the expense of a slight increase in regret in the early stages. An important future direction is to study this phenomenon in more challenging online control problems in linear systems, e.g., under partial observability. Another interesting direction is to combine this mindset with the existing state-of-the-art model-based RL approaches for the general systems and study their performance.

Moreover, our results with TS show that a simple sampling strategy provides effective exploration to recover low-cost achieving controllers in adaptive control of LQRs which yields order optimal regret. An important future direction is to in-

vestigate whether TS achieves optimal regret in partially observable LTI systems, e.g., [160, 163]. Even though we will provide some preliminary results in Chapter 5 relying on the persistence of excitation, the general question of applying TS in the measurement-feedback setting is still an open question. Moreover, to obtain a constant probability of sampling optimistic parameters for general LQRs, TSAC requires $T_w = \omega(\sqrt{T}\log T)$ time-steps of improved exploration (Theorem 3.4), which causes the regret to be dominated by this phase. This long exploration is avoided in LQRs with non-singular optimal closed-loop matrix, which results in regret that scales polynomially in system dimensions (Theorem 3.3). It remains an open problem whether this polynomial dimension dependency in regret can be achieved via TS in general LQRs.

Chapter 4

LEARNING AND CONTROL IN LINEAR TIME-VARYING SYSTEMS

Time-invariant systems such as the ones considered in Chapter 3 have traditionally been the main focus of the study for the linear dynamical systems community [305]. However, real-world systems are often *time-varying*. For example, consider a power system that includes renewable generation (e.g., solar/wind). Due to the intermittency of renewable energy, the system dynamics for frequency regulation in the power system are time-varying. Applying a time-invariant controller in this setting may lead to frequency instability and line failures [276]. Time-varying systems are also crucial for many other applications, such as autonomous vehicles and aircraft control [79]. While not all time-varying systems have linear dynamics, many applications can be approximated by linear time-varying (LTV) systems via a local linear approximation at each time step [265], e.g., the frequency control example described above. As a result, LTV systems are widely-used and there is a large literature focused on designing controllers for LTV systems [14, 212].

Perhaps the most fundamental challenge in dynamical systems is stability. As discussed in Chapter 3, the design of stable linear time-invariant (LTI) systems is well understood, on the other hand, the same cannot be said for LTV systems. To this point, several notions of stability have received attention, e.g., input-to-state stability (ISS), mean-square stability, and Lyapunov stability. In this chapter, we will study ISS and mean-square stability in two different LTV system examples[1].

ISS is one of the most widely adopted notions in the stability of LTV systems. It aims to guarantee the boundedness of the state given bounded initial conditions [114]. In most applications of LTV systems, it is crucial to guarantee ISS both in order to avoid saturation and maintain the robustness and validity of linearization [142, 258].

In stochastic systems, mean-square stability is the crucial notion of stability. For a system with a fixed-point, mean-square stability means that the system converges to its fixed point asymptotically in the mean-square sense. For a noisy system, it implies that the covariance matrix of the state vector stays finite and converges to the solution of the Lyapunov equation of the system, i.e., the steady-state covariance matrix.

[1] This chapter is based on [165, 225].

While there is considerable prior work focused on stability in LTV systems, most prior work studies stability in the offline setting where either the sequence of system parameters are known, e.g., [14, 182], or the system parameters have a particular variation pattern, e.g., [94]. Maintaining stability guarantees becomes significantly harder in the online setting where the system parameters are observed in real-time and may have arbitrary variations. This online setting is the most relevant to many real-world applications, e.g., frequency regulation.

Though stability is crucial, it is not enough for a controller to be stable. A controller must also have low-cost. For instance, in order to stabilize the dynamics, a controller may use arbitrarily big control inputs, which may result in sub-optimal costs. In classical optimal control problems, e.g., the LQR in the previous chapter, the goal is to design a stabilizing controller that minimizes the cost for a particular finite horizon while assuming access to the whole trajectory for that duration. It is possible to characterize the optimal policy in such settings [28]; however, in the online setting when only current or short-termed system information is available, these methods may not guarantee stability, e.g., see Section 4.2.2. There have been recent efforts to provide sub-optimality guarantees on the acquired cost in the online LTV setting, e.g., [98], but it is unclear if the proposed controllers maintain stability for all time-steps since the main focus is on minimizing the cumulative cost.

Thus, in this chapter, we first consider the classical LTV system formulation and aim to answer the following question:

Is it possible for an online controller to guarantee stability and maintain low cost in LTV systems?

We propose an efficient online control algorithm, **CO**variance **C**onstrained **O**nline **L**inear **Q**uadratic (**COCO-LQ**) control, that guarantees input-to-state stability for a large class of LTV systems while also minimizing the control cost. The proposed method incorporates a state covariance constraint into the semi-definite programming (SDP) formulation of the LQ optimal controller. We empirically demonstrate the performance of COCO-LQ in both synthetic experiments and a power system frequency control example.

After studying the classical LTV system setting, we study the effect of asynchrony and randomization in the dynamical systems. In many computational tasks and dynamical systems, asynchrony and randomization are naturally present and have

been considered as ways to increase the speed and reduce the cost of computation while compromising the accuracy and convergence rate. With this motivation, we introduce a natural model for random asynchronous linear time-invariant (LTI) systems which generalizes the standard (synchronous) LTI systems considered in Chapter 3, and gives a new LTV system construction.

In this model, each state variable is updated randomly and asynchronously with some probability according to the underlying system dynamics. We examine how the mean-square stability of random asynchronous LTI systems vary with respect to randomization and asynchrony. Surprisingly, we show that the stability of random asynchronous LTI systems does not imply or is not implied by the stability of the synchronous variant of the system and an unstable synchronous system can be stabilized via randomization and/or asynchrony. We further study a special case of the introduced model, namely randomized LTI systems, where each state element is updated randomly with some fixed but unknown probability.

We consider the problem of system identification of unknown randomized LTI systems using the precise characterization of mean-square stability via the extended Lyapunov equation. For unknown randomized LTI systems, we propose a system identification method to recover the underlying dynamics. Given a single input/output trajectory, our method estimates the model parameters that govern the system dynamics, the update probability of state variables, and the noise covariance using the correlation matrices of collected data and the extended Lyapunov equation. Finally, we empirically demonstrate that the proposed method consistently recovers the underlying system dynamics with the optimal rate.

4.1 Related Work and Background

This chapter builds on the design of linear time-invariant controllers to provide a new approach for the design of stable controllers for linear-time-varying (LTV) systems. As such, we describe related work on both LTI and LTV systems below.

In the study of the control of LTI systems, linear quadratic regulator (LQR) has been considered in detail. In the classical setting where the underlying system is known, the optimal control law is given by a linear feedback controller obtained by solving Riccati equations [28]. Alternatively, the optimal control problem can also be posed via semi-definite programming (SDP) [283], which is the approach we build on in the current study.

Recently, there has been growing interest in online control of these linear systems

when the underlying dynamics are unknown. Most of these works study the problem with a regret minimization perspective, e.g., [2, 71, 162, 166]. However, these methods have so far only been applied in LTI systems with time-varying costs and disturbances. Extensions to LTV dynamics, which are the focus of this chapter, are not known.

As in the case of LTI systems, optimal control of LTV systems where the sequence of system parameters can be obtained by solving backwards Riccati equations [28]. However, in the online case when the sequence of systems is unknown, the design of controllers is challenging. There are several lines of work in adaptive control and model-predictive control (MPC) that have been studied to this point. In adaptive control of LTV systems, the underlying systems are unknown and the results generally assume slow and bounded or fixed systematic variation of dynamics with bounded disturbances [193, 199, 212]. In MPC of LTV systems, a finite horizon of sequence of systems (predictions) is known and the system is again assumed to be slowly varying or open-loop stable, e.g., [78, 304]. Different from prior works, in this chapter we consider the online problem and make no assumptions about how the system varies over time. As in the LTI setting, the study of regret minimization in LTV systems has recently received attention. Gradu et al. [98] are most related to the current study. Gradu et al. [98] studies the adaptive regret of online control in LTV systems with bounded cost. Note that when the cost is bounded, a finite regret need not guarantee stability. In contrast, we use a quadratic (unbounded) cost and we can guarantee stability.

Randomization and asynchrony are crucial to many computational tasks that involve a large number of agents working cooperatively with each other [77, 135]. They allow speed-ups and cost reductions in many artificial and biological processes by removing the synchronization time, relaxing the communication bottlenecks, minimizing the cost of cooperation, and increasing efficiency. For example, large-scale control systems with multiple sensors adopt random asynchronous updates from their sensors due to power saving and difficulty of synchronization [110]. Similarly, asynchrony and randomization are central elements in the dynamical systems of biological neural networks [246, 273]. In various studies, e.g., [40, 75], researchers have found that the synchrony/asynchrony balance phenomenon is ubiquitous in cortical networks. They show the existence of a delicate equilibrium between synchrony and asynchrony of neural firings in many cognitive tasks and regions of the brain such as visual, auditory, and memory maintenance to obtain

stable dynamics during computations. Any disturbance to this natural equilibrium may result in neurological disorders [274].

In modeling stochastic or varying dynamics like random asynchronous LTI systems, there has been a strong interest in switching linear systems/Markov jump systems [63, 210, 254, 257, 300, 301], in which state variables evolve according to a randomly selected model among all possible models. Although randomized linear models can be studied under this framework, the number of possible models becomes exponential in the number of state variables, making this approach prohibitive for large-scale systems. Moreover, the connections between the nodes in randomized systems are mostly fixed without any switching between different systems. Having these connections on or off is the main cause of randomization within the system. Thus, in these dynamical systems, the underlying system is time-invariant while the active interaction within the system is time-varying. Prior works that adopt switching linear systems fail to capture the nature of random asynchronous LTI systems.

In addition to the switching systems viewpoint, the statistical behavior of LTV systems can be also studied from the product of random matrices perspective [21, 74, 100, 104, 131, 219]. However, these frameworks usually come with additional constraints on the state transition matrix, e.g., Hartfiel [104] requires it to be element-wise nonnegative and Avron et al. [21] requires the state transition matric to be positive definite. Similarly, approaches based on *joint spectral radius* are too restrictive to reveal the effect of randomization [131].

In modeling the dynamics of a system, the underlying system is usually unknown and only a sequence of inputs and outputs is available. This raises the system identification problem which aims to recover the parameters that govern the dynamics from the data collected. The classical and recent system identification methods mainly focus on linear dynamical systems (LDS) and consider stable synchronous LTI systems or switching linear systems [160, 189, 214, 233]. For switching linear systems, the system identification methods require the knowledge of the order of switched systems, otherwise, they become computationally intractable and sample inefficient due to exponential dimension dependency [170]. Thus, they have limited applicability to large-scale practical random asynchronous LTI systems. This highlights the necessity of a careful and systematic approach in deriving stability conditions and system identification framework of random asynchronous LTI systems.

Notation. We denote the Euclidean norm of a vector x as $\|x\|$. For a matrix A, $\|A\|$ is its spectral norm, $A^\top$ is its transpose, and $\mathrm{Tr}(A)$ is its trace, $\rho(A)$ denotes the spectral radius of A, i.e., the largest absolute value of its eigenvalues. $\delta(t)$ denotes the unit impulse function. The Kronecker product is denoted as $\otimes$ and $\odot$ denotes the Hadamard product. $\mathcal{N}(\mu, \Sigma)$ denotes normal distribution with mean μ and covariance Σ. $A \succ B$ and $A \succeq B$ denote that $A - B$ is positive definite and positive semi-definite respectively. $A \bullet B$ denotes the element-wise inner product of A and B, i.e., $\mathrm{Tr}(A^\top B)$. $\mathbf{I}_d$ denotes $d \times d$ identity matrix.

4.2 Stable Online Control of Linear Time-Varying Systems

In this section, we design an online controller which guarantees ISS and maintains low regulating cost in LTV systems. Specifically, we propose **Co**variance **C**onstrained **O**nline **L**inear **Q**uadratic (COCO-LQ) control, a novel online control algorithm that aims to minimize the control cost while ensuring provable stability guarantees in LTV systems without restricting how slow or fast the underlying system changes. Further, we demonstrate the performance of the proposed method in various synthetic LTV systems and in the power system frequency control example that motivated our study.

The main technical contribution of our study is a stability guarantee for COCO-LQ in LTV systems. Specifically, we show that COCO-LQ guarantees ISS in online time-varying systems. The key technique that underpins the proposed algorithm is the addition of a novel semi-definiteness constraint on the state covariance matrix into the standard online semi-definite programming (SDP) formulation of linear quadratic optimal control. We show that this constraint promotes the sequential strong stability of the controllers [61], which in turn guarantees ISS with a proper choice of an algorithm hyperparameter. Adding this additional constraint is simple and does not result in a significant increase of computational complexity compared to the standard LQ formulation. Moreover, we prove that if the proposed SDP is not directly feasible, short-term predictions on the future system parameters are necessary and can be used in COCO-LQ in order to ensure ISS.

4.2.1 Problem Setting

We consider the following linear time-varying (LTV) system,

$$x_{t+1} = A_t x_t + B_t u_t + w_t, \tag{4.1}$$

where $x_t \in \mathbb{R}^d$ is the system state, $u_t \in \mathbb{R}^p$ is the control input and $w_t \in \mathbb{R}^d$ is the disturbance at time t. The system is stochastic, i.e., $w_t \sim \mathcal{N}(0, W)$ for $W \succ 0$. The cost at each time-step is a quadratic function of the state and control, $x_t^\top Q x_t + u_t^\top R u_t$, where $Q, R \succ 0$. The decision maker operates in an online setting. That is, at each time-step t, the learner observes the state x_t and system matrix (A_t, B_t) before choosing action u_t and suffering cost $x_t^\top Q x_t + u_t^\top R u_t$. We assume that the cost matrices (Q, R) are time-invariant and known to the learner. However, future system matrices $(A_{t+1}, ..., A_T)$ and $(B_{t+1}, ..., B_T)$ are unknown to the learner and are chosen by the environment, potentially stochastically or adversarially.

Stability. One of the most central goals for controller design is to ensure stability. In this section, we focus on the notion of input-to-state stability (ISS) and strive to design controllers that provide ISS. ISS has been the main notion of stability considered in designing stabilizing controllers both in linear and nonlinear systems [114, 127, 249]. To formally define ISS, let $\mathcal{K}_\infty$ be the set of functions from nonnegative reals to nonnegative reals that are continuous, strictly increasing, and bijective. Then, ISS is defined as follows.

Definition 4.1 (ISS)**.** *A LTV system with deterministic policy $\mathcal{A}$ is said to be input to state stable if there exists functions $\beta_1 : [0, \infty) \times \mathbb{N} \to [0, \infty)$ and $\beta_2 \in \mathcal{K}_\infty$ that satisfy $\beta_1(\cdot, t) \in \mathcal{K}_\infty$ for any $t \in \mathbb{N}$, $\lim_{t\to\infty} \beta_1(a, t) = 0$ for any $a \geq 0$ such that, for any disturbance sequence $\{w_t\}_{t=0}^\infty$, any initial time t_0, any initial state x_{t_0}, and any $t \geq t_0$, we have $\|x_t\| \leq \beta_1(\|x_{t_0}\|, t - t_0) + \beta_2(\sup_{t' \in \mathbb{N}} \|w_{t'}\|)$.*

Cost. In addition to stability, another important objective for controller design is maintaining a small, near-optimal control cost. Here we adopt the standard linear quadratic (LQ) cost model, i.e.,

$$J_T(\mathcal{A}) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^T x_t^\top Q x_t + u_t^\top R u_t\right], \tag{4.2}$$

where $u_1, \ldots, u_t$ are chosen according to policy $\mathcal{A}$, and the expectation is taken with respect to the randomness of noise sequence w_t.

In this section, our goal is to ensure both stability and near-optimal cost. It should be noted that there is a trade-off between these two goals. On the one hand, a stabilizing controller without cost-awareness may produce arbitrarily large control inputs and induce high cost, which is impractical to implement. On the other hand, a greedy approach that merely focuses on cost minimization may lead to instability, as we highlight in the Section 4.2.2 below.

Though our focus is on LTV systems, our approach builds on the SDP formulation of the optimal controller for LTI systems in [283].

Proposition 4.1. *[283] When $A_t = A, B_t = B$ and (A, B) is controllable, the optimal $K^* = LQR(A, B, Q, R)$ where $u_t = K^* x_t$, can be obtained by the following SDP*

$$\min_{\Sigma \succeq 0} \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \bullet \Sigma \qquad s.t. \qquad \Sigma_{xx} = \begin{bmatrix} A_t & B_t \end{bmatrix} \Sigma \begin{bmatrix} A_t & B_t \end{bmatrix}^\top + W,$$

which has a unique symmetric solution Σ^* that decomposes to the following blocks $\Sigma^* = \begin{bmatrix} \Sigma^*_{xx} & \Sigma^*_{xu} \\ {\Sigma^*_{xu}}^\top & \Sigma^*_{uu} \end{bmatrix}$, where $\Sigma^*_{xx} \in \mathbb{R}^{d \times d}$, $\Sigma^*_{xu} \in \mathbb{R}^{d \times p}$ and $\Sigma^*_{uu} \in \mathbb{R}^{p \times p}$. Then, the optimal controller is $K^* = {\Sigma^*_{xu}}^\top (\Sigma^*_{xx})^{-1}$.

The optimal LQR controller described above both stabilizes the system and achieves the minimum cost. The current study makes a step toward understanding if it is possible to extend this formulation to the case of LTV systems.

4.2.2 Naive Approach

How to achieve stable, cost-optimal control of LTI systems is well-known; however this is not the case in LTV systems. To illustrate the challenge of online control of LTV systems, we start by studying the performance of a naive "plug-in" approach where upon receiving (A_t, B_t), an optimal controller for A_t, B_t is computed under the assumption that the system is time-invariant. Due to its simplicity, this approach has been employed in many contexts, e.g., Li et al. [183] for a Markov decision process setting. In this section we provide an example that shows that such a myopic approach based on optimal LTI control described above fails to stabilize the system even in simple settings where A_t can only switch between two possible choices and B_t is fixed. This highlights that one cannot naively apply LTI design approaches in LTV systems and expect to maintain stability.

Example 4.1. *Consider a system with $Q = \epsilon I$, $R = I$, $w_t = 0$, and*

$$A = \begin{bmatrix} \rho & 0 \\ a & \rho \end{bmatrix}, \quad A' = \begin{bmatrix} \rho & a \\ 0 & \rho \end{bmatrix},$$

where $0 < \rho < 1$, and $a > \sqrt{2}$. Suppose A_t alternates between A and A' and $B_t = B = I$. Define the optimal LTI controllers for A and A' as $K := LQR(A, B, Q, R)$ and $K' := LQR(A', B, Q, R)$. To show that the optimal LTI controllers will not stabilize the system, we consider a case where $\epsilon \to 0$. In this case, one can check

that $K, K' \to 0$. Since A_t alternates between A, A', K_t also alternates between K and K' under the myopic design we are considering. Thus, the system state follows $x_{t+2} = (A+K)(A'+K')x_t$. Notice that as $\epsilon \to 0$,

$$(A+K)(A'+K') \to AA' = \begin{bmatrix} \rho^2 & a\rho \\ a\rho & a^2+\rho^2 \end{bmatrix}.$$

Here, AA' is unstable since its largest eigenvalue is greater than $\frac{1}{2}\mathrm{Tr}(AA') = \rho^2 + \frac{a^2}{2} > 1$. Thus, for small enough ϵ, the naive strategy that uses the LTI controller at each time-step leads to instability.

4.2.3 Main Result

The previous section highlights that a naive application of LTI control cannot guarantee stability for LTV systems. We now propose a new approach, COvariance Constrained Online LQ (COCO-LQ) control. Our main technical result shows that COCO-LQ provably guarantees stability in LTV systems when the SDP is feasible. Then, we discuss how to handle the situation when the SDP is infeasible. Finally, we discuss the effect of model estimation error.

COvariance Constrained Online LQ (COCO-LQ)

The naive approach discussed in Section 4.2.2 seeks to solve the LTI problem at every time step, which is equivalent to solving the SDP in Proposition 4.1 for every (A_t, B_t). The reason this method fails is that it only considers cost minimization without explicitly considering stability. The main idea of COCO-LQ is to enforce stability via a state covariance constraint embedded into the SDP framework. The proposed algorithm is stated formally in Algorithm 7.

COCO-LQ solves an SDP (4.3) at each time step that is similar to that in Proposition 4.1. The crucial difference is the new constraint (4.3c), which involves parameter α. Plugging (4.3a) into constraint (4.3c) yields the following:

$$\Sigma_{xx} \preceq \frac{1}{1-\alpha} W.$$

This highlights that constraint (4.3c) can be interpreted as an upper bound on the state covariance matrix Σ_{xx}. When $\alpha = 0$, the controller essentially cancels out the dynamics, without taking into account the cost of doing so. This ensures stability but can lead to large cost. At another extreme, when $\alpha \to 1$, the SDP solved at each time step is the same as for the LTI setting, and so COCO-LQ matches the naive

Algorithm 7 COCO-LQ: COvariance Constrained Online LQ

1: **Input:** $\alpha \in [0, 1)$, $Q, R, W \succ 0$
2: **for** $t = 1, 2, \ldots$ **do**
3: Receive state x_t, and system parameter A_t, B_t
4: Solve the following SDP for $\Sigma_t \in \mathbb{R}^{(d+p)\times(d+p)}$:

$$\begin{aligned} \text{minimize} \quad & \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \bullet \Sigma \\ \text{subject to} \quad & \Sigma_{xx} = \begin{bmatrix} A_t & B_t \end{bmatrix} \Sigma \begin{bmatrix} A_t & B_t \end{bmatrix}^\top + W && \text{(4.3a)} \\ & \Sigma \succeq 0 && \text{(4.3b)} \\ & \begin{bmatrix} A_t & B_t \end{bmatrix} \Sigma \begin{bmatrix} A_t & B_t \end{bmatrix}^\top \preceq \alpha \Sigma_{xx} && \text{(4.3c)} \end{aligned}$$

5: Compute the control gain $K_t = \Sigma_{xu}^\top \Sigma_{xx}^{-1}$, where $\Sigma_t = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xu} \\ \Sigma_{xu}^\top & \Sigma_{uu} \end{bmatrix}$
6: Execute control action $u_t = K_t x_t$

approach. Thus, α trades off between stability and cost. In the following section, we show that this novel state covariance constraint promotes sequential strong stability [61], which in turn guarantees ISS with a proper choice of α.

Stability

We now state our main technical result, which provides a formal stability guarantee for COCO-LQ.

Theorem 4.1. *Let $0 \leq \alpha < 1/2$, and suppose* (4.3) *is feasible for all t, then the resulting dynamical system satisfies ISS in the sense that for any disturbance sequence $\{w_t\}_{t=0}^{\infty}$ and for any $t \geq t_0$,*

$$\|x_t\| \leq \rho^{t-t_0} \|x_{t_0}\| + \frac{\kappa\rho}{1-\rho} \sup_{t_0 \leq k < t} \|w_k\|$$

for $\rho = \sqrt{\frac{\alpha}{1-\alpha}} \in [0, 1)$ and $\kappa = \frac{\kappa_W}{\sqrt{1-\alpha}}$, where $\kappa_W = \|W\|\|W^{-1}\|$ is the condition number of W.

The key intuition underpinning this result is that the additional state covariance constraint (4.3c) implicitly enforces sequential strong stability [61], which in turn ensures ISS. These results are formally stated in Lemma 4.1 and Lemma 4.2 respectively. First, we formally define sequential strong stability.

Definition 4.2 (Sequential Strong Stability)**.** *A sequence of policies $K_1, K_2, \ldots$, such that $u_t = K_t x_t$ is (κ, γ, ρ)-sequential strongly stable (for $\kappa > 0$, $0 < \gamma \leq 1$ and*

$0 \leq \rho < 1$*) if there exist matrices* $H_1, H_2, ...,$ *and* $L_1, L_2 ...,$ *such that* $A_t + B_t K_t = H_t L_t H_t^{-1}$ *for all t, with the following properties: (a)* $\|L_t\| \leq 1 - \gamma$*; (b)* $\|H_t\| \leq \beta_1$ *and* $\|H_t^{-1}\| \leq 1/\beta_2$ *with* $\kappa = \beta_1/\beta_2$*; (c)* $\|H_{t+1}^{-1} H_t\| \leq \frac{\rho}{1-\gamma}$.

With this definition in place, we present the connection between the condition in (4.3c) and sequential strong stability.

Lemma 4.1. *Under the conditions in Theorem 4.1, the policies designed by COCO-LQ are* (κ, γ, ρ)*-sequential strongly stability for* $\kappa = \frac{\kappa_W}{\sqrt{1-\alpha}}, \gamma = 1 - \sqrt{\alpha}, \rho = \sqrt{\frac{\alpha}{1-\alpha}}$ *where* $\kappa_W = \|W\|\|W^{-1}\|$.

Proof. To prove Lemma 4.1, we first show that the optimal solution to the SDP in (4.3) has a specific low rank structure. In particular, we claim that if (4.3) has a minimizer Σ^*, then there exists a $K \in \mathbb{R}^{p \times d}$ s.t. Σ^* can be written as

$$\Sigma^* = \begin{bmatrix} \Sigma_{xx}^* & \Sigma_{xu}^* \\ (\Sigma_{xu}^*)^\top & \Sigma_{uu}^* \end{bmatrix} = \begin{bmatrix} \Sigma_{xx}^* & \Sigma_{xx}^* K^\top \\ K\Sigma_{xx}^* & K\Sigma_{xx}^* K^\top \end{bmatrix}. \tag{4.4}$$

To see this, first note that $\Sigma_{xx}^* \succeq W \succ 0$, and therefore we can simply define $K = (\Sigma_{xu}^*)^\top (\Sigma_{xx}^*)^{-1}$, and the only thing we need to show is $\Sigma_{uu}^* = K\Sigma_{xx}^* K^\top$. Suppose this is not true, then since $\Sigma^* \succeq 0$, there must exist $D \neq 0, D \geq 0$ s.t.

$$\Sigma_{uu}^* = K\Sigma_{xx}^* K^\top + D. \tag{4.5}$$

Then, by (4.3a), we also have,

$$\Sigma_{xx}^* = (A_t + B_t K)\Sigma_{xx}^* (A_t + B_t K)^\top + B_t D B_t^\top + W. \tag{4.6}$$

Viewing the above as a Lyapunov equation in terms of Σ_{xx}^*, and since $B_t D B_t^\top + W \succ 0$ and $\Sigma_{xx}^* \succ 0$, we get $A_t + B_t K$ is a stable matrix. Next, since $A_t + B_t K$ is stable, we can construct $\tilde{\Sigma}_{xx}$ to be the unique solution to the following Lyapunov equation,

$$\tilde{\Sigma}_{xx} = (A_t + B_t K)\tilde{\Sigma}_{xx} (A_t + B_t K)^\top + W. \tag{4.7}$$

We further define,

$$\tilde{\Sigma} = \begin{bmatrix} \tilde{\Sigma}_{xx} & \tilde{\Sigma}_{xx} K^\top \\ K\tilde{\Sigma}_{xx} & K\tilde{\Sigma}_{xx} K^\top \end{bmatrix},$$

and we claim that $\tilde{\Sigma}$ is a feasible solution to (4.3). Clearly $\tilde{\Sigma}$ is positive semi-definite, and further, it satisfies (4.3a). We now check that (4.3c) is met below.

$$
\begin{aligned}
&\begin{bmatrix} A_t & B_t \end{bmatrix} \tilde{\Sigma} \begin{bmatrix} A_t & B_t \end{bmatrix}^\top - \alpha \tilde{\Sigma}_{xx} \\
&= (A_t + B_t K) \tilde{\Sigma}_{xx} (A_t + B_t K)^\top - \alpha \tilde{\Sigma}_{xx} \\
&= (A_t + B_t K) \Sigma^*_{xx} (A_t + B_t K)^\top - \alpha \Sigma^*_{xx} + (A_t + B_t K)(\tilde{\Sigma}_{xx} - \Sigma^*_{xx})(A_t + B_t K)^\top - \alpha(\tilde{\Sigma}_{xx} - \Sigma^*_{xx}) \\
&= [A_t, B_t] \Sigma^* [A_t, B_t]^\top - B_t D B_t^\top - \alpha \Sigma^*_{xx} + (A_t + B_t K)(\tilde{\Sigma}_{xx} - \Sigma^*_{xx})(A_t + B_t K)^\top - \alpha(\tilde{\Sigma}_{xx} - \Sigma^*_{xx}) \\
&\preceq -B_t D B_t^\top + (A_t + B_t K)(\tilde{\Sigma}_{xx} - \Sigma^*_{xx})(A_t + B_t K)^\top - \alpha(\tilde{\Sigma}_{xx} - \Sigma^*_{xx}), \qquad (4.8)
\end{aligned}
$$

where (4.8) is due to Σ^* must satisfy (4.3c). Subtracting (4.6) from (4.7), we have,

$$
\tilde{\Sigma}_{xx} - \Sigma^*_{xx} = (A_t + B_t K)(\tilde{\Sigma}_{xx} - \Sigma^*_{xx})(A_t + B_t K)^\top - B_t D B_t^\top. \qquad (4.9)
$$

Plugging (4.9) into (4.8), we have,

$$
\begin{bmatrix} A_t & B_t \end{bmatrix} \tilde{\Sigma} \begin{bmatrix} A_t & B_t \end{bmatrix}^\top - \alpha \tilde{\Sigma}_{xx} \preceq (1 - \alpha)(\tilde{\Sigma}_{xx} - \Sigma^*_{xx}).
$$

Therefore, to check (4.3c) it remains to show $\tilde{\Sigma}_{xx} - \Sigma^*_{xx} \preceq 0$. To see this, we view (4.9) as a Lyapunov equation in terms of $\tilde{\Sigma}_{xx} - \Sigma^*_{xx}$, and as $A_t + B_t K$ is stable, and $B_t D B_t^\top \succeq 0$, we have $\tilde{\Sigma}_{xx} - \Sigma^*_{xx} \preceq 0$. As a result, (4.3c) holds and $\tilde{\Sigma}_{xx}$ is indeed a feasible solution to (4.3).

Further, we can show $\tilde{\Sigma}$ achieves a strictly lower cost than Σ^*. To see this, note we have already shown $\tilde{\Sigma}_{xx} \preceq \Sigma^*_{xx}$, which also implies $\tilde{\Sigma}_{uu} = K \tilde{\Sigma}_{xx} K^\top \preceq K \Sigma^*_{xx} K^\top = \Sigma^*_{uu} - D$ with $D \succeq 0$, $D \neq 0$. Coupled this with the fact that Q, R are strictly positive definite, we deduce that $\tilde{\Sigma}$ must achieve a lower cost. So we get a contradiction, and verified that (4.4) holds.

Now that we established the decomposition in (4.4) for the solution of (4.3), we proceed with the proof of Lemma 4.1. Let Σ_t be the solution to the SDP (4.3) at step t. Then, Σ_t can be rewritten as,

$$
\Sigma_t = \begin{bmatrix} \Sigma_{t,xx} & \Sigma_{t,xx} K_t^\top \\ K_t \Sigma_{t,xx} & K_t \Sigma_{t,xx} K_t^\top \end{bmatrix},
$$

for some $\Sigma_{t,xx} \succ 0$, and K_t is the linear controller at step t. With this, we can re-write the left side of constraint (4.3c) as follows,

$$
\begin{aligned}
\begin{bmatrix} A_t & B_t \end{bmatrix} \Sigma_t \begin{bmatrix} A_t & B_t \end{bmatrix}^\top &= \begin{bmatrix} A_t & B_t \end{bmatrix} \begin{bmatrix} \Sigma_{t,xx} & \Sigma_{t,xx} K_t^\top \\ K_t \Sigma_{t,xx} & K_t \Sigma_{t,xx} K_t^\top \end{bmatrix} \begin{bmatrix} A_t^\top \\ B_t^\top \end{bmatrix} \\
&= A_t \Sigma_{t,xx} A_t^\top + A_t K_t \Sigma_{t,xx} B_t^\top + B_t \Sigma_{t,xx} K_t^\top A_t^\top + B_t K_t \Sigma_{t,xx} K_t^\top B_t^\top \\
&= (A_t + B_t K_t) \Sigma_{t,xx} (A_t + B_t K_t)^\top.
\end{aligned}
$$

As a result, (4.3c) can be equivalently expressed as,

$$(A_t + B_t K_t)\Sigma_{t,xx}(A_t + B_t K_t)^\top \preceq \alpha \Sigma_{t,xx}. \tag{4.10}$$

Left and right multiplying the above by $\Sigma_{t,xx}^{-1/2}$, we get

$$\Sigma_{t,xx}^{-1/2}(A_t + B_t K_t)\Sigma_{t,xx}^{1/2}\Sigma_{t,xx}^{1/2}(A_t + B_t K_t)^\top \Sigma_{t,xx}^{-1/2} \preceq \alpha I\,. \tag{4.11}$$

Let $H_t = \Sigma_{t,xx}^{1/2}$, $L_t = \Sigma_{t,xx}^{-1/2}(A_t + B_t K_t)\Sigma_{t,xx}^{1/2}$. We get the following decomposition,

$$(A_t + B_t K_t) = H_t L_t H_t^{-1}\,. \tag{4.12}$$

We now show that this decomposition yields the desired sequential stability property. To do so, we need to check the three conditions in Definition 4.2.

To check condition (a) in Definition 4.2, note that (4.11) provides an upper bound on $\|L_t\|$, that is $L_t L_t^\top \preceq \alpha I$. Therefore, $\|L_t\| \leq \sqrt{\alpha} = 1 - \gamma$ with $\gamma = 1 - \sqrt{\alpha}$.

To check condition (b) which is an upper bound on $\|H_t\|$ and $\|H_t^{-1}\|$, we recall constraint (4.3a),

$$\begin{bmatrix} A_t & B_t \end{bmatrix} \Sigma_t \begin{bmatrix} A_t & B_t \end{bmatrix}^\top = \Sigma_{t,xx} - W.$$

As the left-hand side of the above is positive semi-definite, we must have $\Sigma_{t,xx} \succeq W$. Further, plugging (4.3a) into (4.3c), we can get, $\Sigma_{t,xx} - W \preceq \alpha \Sigma_{t,xx}$. These can be equivalently written as

$$W \preceq \Sigma_{t,xx} \preceq \frac{W}{1-\alpha}. \tag{4.13}$$

Therefore, $\|H_t\| = \|\Sigma_{t,xx}^{1/2}\| \leq \frac{\|W^{1/2}\|}{\sqrt{1-\alpha}}$, $\|H_t^{-1}\| = \|\Sigma_{t,xx}^{-1/2}\| \leq \|W^{-1/2}\|$, $\|H_t\|\|H_t^{-1}\| \leq \frac{\|W^{1/2}\|\|W^{-1/2}\|}{\sqrt{1-\alpha}} = \frac{\kappa_W}{\sqrt{1-\alpha}}$. Therefore, condition (b) holds with $\kappa = \frac{\kappa_W}{\sqrt{1-\alpha}}$.

Lastly, we check condition (c), which is an upper bound on $\|H_{t+1}^{-1} H_t\|$. Note that,

$$H_{t+1}^{-1} H_t H_t^\top (H_{t+1}^{-1})^\top = H_{t+1}^{-1}\Sigma_{t,xx}(H_{t+1}^{-1})^\top \preceq \frac{1}{1-\alpha}(\Sigma_{t+1,xx})^{-1/2} W (\Sigma_{t+1,xx})^{-1/2} \preceq \frac{1}{1-\alpha} I,$$

where the two inequalities in the above derivations are due to (4.13). As a result, we have $\|H_{t+1}^{-1} H_t\| \leq \frac{1}{\sqrt{1-\alpha}} = \frac{\rho}{1-\gamma}$ with $\rho = \frac{\sqrt{\alpha}}{\sqrt{1-\alpha}}$. As $\alpha < 1/2$, we have $\rho < 1$. As such, condition (c) holds. In summary, the (κ, γ, ρ) strong sequential stability holds, with $\kappa = \frac{\kappa_W}{\sqrt{1-\alpha}}$, $\gamma = 1 - \sqrt{\alpha}$, and $\rho = \sqrt{\frac{\alpha}{1-\alpha}}$, which concludes the proof. □

Now that we established the sequential strong stability of the policies obtained via COCO-LQ, we require relating this fact to ISS. The following provides the remaining piece in the proof of Theorem 4.1.

Lemma 4.2. *Suppose a sequence of policies $K_1, \ldots, K_t$ is (κ, γ, ρ)-sequential strongly stable. Then, the closed-loop system obtained via the sequence of these policies is input-to-state stable in the sense that for any $t \geq t_0 \geq 1$,*

$$\|x_t\| \leq \kappa \rho^{t-t_0} \|x_{t_0}\| + \frac{\kappa \rho}{1-\rho} \max_{t_0 \leq s < t} \|w_s\|.$$

Proof. For notational simplicity, we only prove the case with $t_0 = 1$; the general case follows similarly, at the cost of heavier notations. Let $x_1, x_2, \ldots$ be a sequence of states starting from initial state x_1, and generated by the dynamics $x_{t+1} = A_t x_t + B_t u_t + w_t = (A_t + B_t K_t) x_t + w_t$. Hence, $x_t = M_1 x_1 + \sum_{s=1}^{t-1} M_{s+1} w_s$, where

$$M_t = I; M_s = M_{s+1}(A_s + B_s K_s) = (A_{t-1} + B_{t-1} K_{t-1}) \cdots (A_s + B_s K_s).$$

By the sequential strong stability, there exist matrices $H_1, H_2, \ldots$ and $L_1, L_2, \ldots$ such that $A_j + B_j K_j = H_j L_j H_j^{-1}$, and H_j, L_j satisfy the properties specified in Definition 4.2. Thus we have for all $1 \leq s < t$,

$$\begin{aligned}
\|M_s\| &= \|H_{t-1} L_{t-1} H_{t-1}^{-1} H_{t-2} L_{t-1} H_{t-2}^{-1} \cdots H_s L_s H_s^{-1}\| \\
&\leq \|H_{t-1}\| \left(\prod_{j=s}^{t-1} \|L_j\| \right) \left(\prod_{j=s}^{t-2} \|H_{j+1}^{-1} H_j\| \right) \|H_s^{-1}\|, \\
&\leq \beta_1 (1-\gamma)^{t-s} (\frac{\rho}{1-\gamma})^{t-s-1} (1/\beta_2) \leq \frac{\kappa(1-\gamma)}{\rho} \rho^{t-s}.
\end{aligned}$$

As $\kappa \geq 1$, the same holds for M_t. Thus, we have,

$$\begin{aligned}
\|x_t\| &\leq \|M_1\| \|x_1\| + \sum_{s=1}^{t-1} \|M_{s+1}\| \|w_s\| \\
&\leq \kappa \rho^{t-1} \|x_1\| + \kappa \sum_{s=1}^{t-1} \rho^{t-s-1} \|w_s\| \leq \kappa \rho^{t-1} \|x_1\| + \frac{\kappa \rho}{1-\rho} \max_{1 \leq s < t} \|w_s\|.
\end{aligned}$$

□

Proof of Theorem 4.1. Combining Lemma 4.1 with Lemma 4.2 gives the advertised result of stability for COCO-LQ. □

A critical assumption in Theorem 4.1 is that (4.3) is feasible for $0 \leq \alpha < 1/2$. The following result provides a condition when the problem is always feasible.

Lemma 4.3. *When B_t is full row rank, then the SDP* (4.3) *is always feasible.*

Proof. We prove this result by explicitly constructing a feasible solution for (4.3). As B_t has full row rank, $B_t B_t^\top$ is invertible. Let

$$\begin{aligned}\Sigma_0 &= \begin{bmatrix} W & -WA_t^\top (B_t B_t^\top)^{-1} B_t \\ -B_t^\top (B_t B_t^\top)^{-1} A_t W & B_t^\top (B_t B_t^\top)^{-1} A_t W A_t^\top (B_t B_t^\top)^{-1} B_t \end{bmatrix} \\ &= \begin{bmatrix} I \\ -B_t^\top (B_t B_t^\top)^{-1} A_t \end{bmatrix} W \begin{bmatrix} I \\ -B_t^\top (B_t B_t^\top)^{-1} A_t \end{bmatrix}^\top.\end{aligned}$$

It suffices to show that Σ_0 satisfies (4.3a), (4.3b), and (4.3c). Notice that

$$[A_t, B_t]\Sigma_0[A_t, B_t]^\top = (A_t - B_t B_t^\top (B_t B_t^\top)^{-1} A_t) W (A_t - B_t B_t^\top (B_t B_t^\top)^{-1} A_t)^\top = 0.$$

Therefore, constraint (4.3a) is equivalent to $\Sigma_{xx} = W$, which holds by the construction of Σ_0. Next, as $W \succ 0$, Σ_0 is clearly positive semi-definitive and (4.3b) holds. Finally, note the left-hand side of (4.3c) is 0, and the right-hand side of (4.3c) is positive semi-definite. As a result, (4.3c) holds. □

Note that having B_t full row rank is a sufficient but not necessary condition for the feasibility of (4.3) of COCO-LQ. When B_t is not full row rank, the feasibility assumption may still hold, and therefore our assumption is weaker than the invertibility assumption used in the literature, e.g., Lai [155]. More broadly, in Theorem 4.1, $\alpha < 0.5$ is a sufficient condition for stability. For $\alpha \geq 0.5$, stability may still hold for some problem instances (A_t, B_t) as will be shown in the simulations in Section 4.2.4. How to provide a more refined instance-dependent threshold on α is an interesting future direction.

Infeasibility and the Role of Predictions

We now turn our attention to the case when the SDP given in (4.3) is infeasible. In this case it is necessary for the controller to use additional information in order to stabilize the system. In particular, the following example shows that when B_t is not full row rank, for any (deterministic) online control algorithm that has causal access to system matrices, there exists a future sequence of (A_t, B_t) such that the system state will blow up from a given initial state.

Example 4.2. *Set $d = 2, p = 1$, i.e., A_t is 2-by-2 and B_t is 2-by-1. For all t, we set $B_t = [1, 0]^\top$, and for even time indices $t = 2k$, we set $A_t = I$. Further, assume that $x_0 = [1, 1]^\top$ and $w_t = 0$. Our construction is based on induction. Suppose*

Algorithm 8 COCO-LQ with Predictions

Input: $\alpha \in [0, 1)$, $Q, R, W \succ 0$
for $t = 1, 2, \ldots$ **do**
  Receive state x_t
  if $t \equiv 1 \pmod{H}$ **then**
    Receive system parameters $(A_t, B_t), \ldots, (A_{t+H-1}, B_{t+H-1})$
    Solve (4.3) by replacing R with $\tilde{R} = \mathbf{I}_H \otimes R$, A_t with $\tilde{A}_t := [A_{t+H-1} \cdots A_t]$, and B_t with $\tilde{B}_t := [B_{t+H-1}, A_{t+H-1}B_{t+H-2}, \ldots, A_{t+H-1} \cdots A_{t+1}B_t]$ to recover Σ_t
    Compute $\mathbf{K_t} = [K_t^\top, \ldots, K_{t+H-1}^\top]^\top = \Sigma_{xu}^\top \Sigma_{xx}^{-1}$, where $\Sigma_t = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xu} \\ \Sigma_{xu}{}^\top & \Sigma_{uu} \end{bmatrix}$
  Execute control action $u_t = K_t x_t$

at an even index $t = 2k$, $x_{2k} = [x_{2k,1}, x_{2k,2}]^\top$ with $x_{2k,2} \geq 0$. Then, after u_{2k} is determined, we set

$$A_{2k+1} = \begin{bmatrix} 1 & 0 \\ \epsilon & 2 \end{bmatrix}$$

with $\epsilon = 0.5 \frac{|x_{2k,2}|}{\max(|x_{2k,1}|, 1)} \operatorname{sign}(u_{2k})$ Then, after (A_{2k+1}, B_{2k+1}) is revealed to the agent, it takes an action u_{2k+1}, resulting in $x_{2k+2} = A_{2k+1}A_{2k}x_{2k} + A_{2k+1}B_{2k}u_{2k} + B_{2k+1}u_{2k+1}$. Since $A_{2k} = I$ and $B_{2k} = B_{2k+1} = [1, 0]^\top$, the second coordinate of x_{2k+2} can then be written as $x_{2k+2,2} = \epsilon x_{2k,1} + 2x_{2k,2} + \epsilon u_{2k}$. By the way ϵ is chosen, we have $\epsilon u_{2k} \geq 0$, and $|\epsilon x_{2k,1}| \leq 0.5|x_{2k,2}|$, and therefore, $x_{2k+2,2} \geq 1.5x_{2k,2}$. Since the system starts with $x_0 = [1, 1]^\top$, applying the argument above recursively, we have for any k, $x_{2k,2} \geq 1.5^k$, i.e., the state blows up.

In this section, we show that using (A_t, B_t) together with short-term predictions of future system matrices is enough to stabilize the system under standard controllability assumptions. Specifically, we extend COCO-LQ to include future H steps of predictions in Algorithm 8. The key idea is to rewrite the dynamics as

$$x_{t+H} = \tilde{A}_t x_t + \tilde{B}_t \bar{u}_t + [I, A_{t+H-1}, \cdots, A_{t+H-1} \ldots A_{t+1}] \bar{w}_t, \tag{4.14}$$

where $\tilde{A}_t := A_{t+H-1} \cdots A_t$, $\tilde{B}_t := [B_{t+H-1}, A_{t+H-1}B_{t+H-2}, \ldots, A_{t+H-1} \cdots A_{t+1}B_t]$, $\bar{u}_t := [u_{t+H-1}^\top, u_{t+H-2}^\top, \ldots, u_t^\top]^\top$ and $\bar{w}_t := [w_{t+H-1}^\top, w_{t+H-2}^\top, \ldots, w_t^\top]^\top$. Notice that $\tilde{B}_t$ is in the form of the controllability matrix for this LTV system. When H is long enough such that $\tilde{B}_t$ is full row rank, i.e., LTV system is H-controllable, we can use Algorithm 8 on $\tilde{A}_t$ and $\tilde{B}_t$ and avoid the infeasibility issue, and we have the stability guarantee.

Theorem 4.2. *Suppose for each t, matrix $\tilde{B}_t$ defined above satisfies $\tilde{B}_t\tilde{B}_t^\top \geq \sigma I$ for some $\sigma > 0$, and $\|A_t\| \leq a, \|B_t\| \leq b$ for some $a, b > 0$. Then, the SDP in Algorithm 8 is always feasible. Further, when $\alpha < 1/2$, the closed-loop system is ISS for any t,*

$$\|x_t\| \leq \kappa_A' \rho^{\frac{t}{H}-1}\|x_1\| + \kappa_A'\kappa_A\kappa \max(1, \frac{\rho}{1-\rho}) \sup_{1\leq s<t} \|w_s\|,$$

where the same as Theorem 4.1, $\rho = \sqrt{\frac{\alpha}{1-\alpha}} \in [0,1)$ and $\kappa = \frac{\kappa_W}{\sqrt{1-\alpha}}$ with $\kappa_W = \|W\|\|W^{-1}\|$ being the condition number of W; further, $\kappa_A = 1 + a + \ldots + a^{H-1}$, and $\kappa_A' = a^{H-1} + b^2\kappa_A^2\kappa_R\frac{\kappa+a^H}{\sigma}$ with κ_R being the condition number of R.

Proof. The feasibility directly follows Lemma 4.3. For stability, using the dynamics given in (4.14), from Theorem 4.1, we have, $\forall k \geq 1$

$$\|x_{kH+1}\| \leq \rho^k\|x_1\| + \frac{\kappa\rho}{1-\rho} \sup_{0\leq\tau<k} \|\tilde{w}_{\tau H+1}\|.$$

Note that, for any τ,

$$\begin{aligned}\|\tilde{w}_{\tau H+1}\| &\leq \sum_{\ell=1}^{H} \|A_{\tau H+H}\| \cdots \|A_{\tau H+\ell+1}\|\|w_{\tau H+\ell}\| \\ &\leq (1 + a + \ldots + a^{H-1}) \sup_{\tau H+1\leq s\leq(\tau+1)H} \|w_s\| := \kappa_A \sup_{\tau H+1\leq s\leq(\tau+1)H} \|w_s\|.\end{aligned}$$

These show that,

$$\|x_{kH+1}\| \leq \rho^k\|x_1\| + \frac{\kappa_A\kappa\rho}{1-\rho} \sup_{1\leq s\leq kH} \|w_s\|.$$

The above already shows the boundedness of states for time indexes $t = 1 \mod (H)$.

To show the boundedness of the states for all t, we need to consider the effect of control inputs in all H sequences. Note that Theorem 4.1 shows that the policies K_{kH+1} is (κ, γ, ρ)-sequential strongly stable, with $\gamma = 1 - \sqrt{\alpha}$ and $\kappa = \frac{\kappa_W}{\sqrt{1-\alpha}}$. This implies $\|\tilde{A}_{kH+1} + \tilde{B}_{kH+1}K_{kH+1}\| \leq (1-\gamma)\kappa$, showing

$$\|\tilde{B}_{kH+1}K_{kH+1}\| \leq (1-\gamma)\kappa + a^H,$$

which further leads to

$$\|K_{kH+1}\| \leq K_{\max} = \kappa_R\frac{b(1 + a + \ldots + a^{H-1})}{\sigma}((1-\gamma)\kappa + a^H).$$

Therefore, we have

$$\begin{aligned}
\|x_{kH+\ell}\| &\leq \|A_{kH+\ell-1}\cdots A_{kH+1}\|\|x_{kH+1}\| \\
&\quad + \| \begin{bmatrix} B_{kH+\ell-1}, A_{kH+\ell-1}B_{kH+\ell-2}, \ldots, A_{kH+\ell-1}\cdots A_{kH+2}B_{kH+1} \end{bmatrix} \begin{bmatrix} u_{kH+\ell-1} \\ \vdots \\ u_{kH+1} \end{bmatrix} \| \\
&\quad + \| [I, A_{t+H-1}, \cdots, A_{kH+\ell-1}\ldots A_{kH+1}] \begin{bmatrix} w_{kH+\ell-1} \\ \vdots \\ w_{kH+1} \end{bmatrix} \| \\
&< \kappa'_A \|x_{kH+1}\| + \kappa_A \sup_{kH+1\leq s<kH+\ell} \|w_s\|.
\end{aligned}$$

For any t, let k be such that $t = kH + \ell$ with $1 \leq \ell \leq H$. Then, we have,

$$\begin{aligned}
\|x_t\| &\leq \kappa'_A \|x_{kH+1}\| + \kappa_A \sup_{kH+1\leq s<t} \|w_s\| \\
&\leq \kappa'_A \rho^k \|x_1\| + \kappa'_A \frac{\kappa_A \kappa \rho}{1-\rho} \sup_{1\leq s\leq kH} \|w_s\| + \kappa_A \sup_{kH+1\leq s<t} \|w_s\| \\
&\leq \kappa'_A \rho^{\frac{t}{H}-1} \|x_1\| + \kappa'_A \kappa_A \kappa \max(1, \frac{\rho}{1-\rho}) \sup_{1\leq s\leq t} \|w_s\|.
\end{aligned}$$

□

Estimation Error

In both Algorithm 7 and Algorithm 8, the exact knowledge of state-transition matrices (A_t, B_t) or the extended state-transition matrices $(\tilde{A}_t, \tilde{B}_t)$ are needed when deriving the control actions. In this section, we show that COCO-LQ can still obtain a stabilizing controller in the case where only approximations are known, if the estimation error is controlled. Our main result is the following.

Theorem 4.3. *Let $(\hat{A}_t,\ \hat{B}_t)$ be an estimate of $(A_t,\ B_t)$. Given $\alpha \in [0, \frac{1}{2})$, let $\rho = \sqrt{\frac{\alpha}{1-\alpha}}$, $\kappa = \frac{\|W\|\|W^{-1}\|}{\sqrt{1-\alpha}}$ and $\gamma = 1 - \sqrt{\alpha}$. Let $K_1, K_2, \ldots$ be the policies designed by COCO-LQ for $(\hat{A}_t,\ \hat{B}_t)$ with parameter α. When the estimation error satisfies,*

$$\max\{||\hat{A}_t - A_t||_2, ||\hat{B}_t - B_t||_2\} \leq \delta \frac{\gamma}{\kappa(1 + K_{max})}, \tag{4.15}$$

where δ can be any number in $(0, \frac{\sqrt{1-\alpha}-\sqrt{\alpha}}{1-\sqrt{\alpha}})$, and K_{max} is any uniform upper bound on $\|K_t\|$. Then, the policies K_t are ISS when applied to the system (A_t, B_t),

$$\|x_t\| \leq (\rho')^{t-t_0} \|x_{t_0}\| + \frac{\kappa\rho'}{1-\rho'} \sup_{t_0\leq k<t} \|w_k\|,$$

where $\rho' = \frac{1-(1-\delta)\gamma}{1-\gamma}\rho \in (0,1)$. *Finally, when* $\|\hat{A}_t\| \le \bar{\sigma}_A$, $\|\hat{B}_t\| \le \bar{\sigma}_B$ *and* $\hat{B}_t\hat{B}_t^\top \succeq \underline{\sigma}_B^2 I$, *one uniform upper bound for* $\|K_t\|$ *is* $K_{max} = \kappa_R \frac{\bar{\sigma}_B}{\underline{\sigma}_B^2}(\kappa(1-\gamma)+\bar{\sigma}_A)$ *with* $\kappa_R = \|R\|\|R^{-1}\|$.

Proof. The proof is divided by two parts. For the first part, we prove the ISS property given the upper bound K_{max} on the controllers $\|K_t\|$. In the second part, we provide such an upper bound K_{max}.

Proof of ISS. By Lemma 4.2, to show ISS we only need to show that $(K_t)_{t=0}^\infty$ is sequential strongly stable for system $(A_t, B_t)_{t=0}^\infty$. $\{K_t\}_{t=0}^\infty$ is (κ,γ,ρ)-sequential strongly stable for the system $(\hat{A}_t, \hat{B}_t)_{t=0}^\infty$ with (κ,γ,ρ) defined by Lemma 4.1 as $\kappa = \frac{\kappa_W}{\sqrt{1-\alpha}}, \gamma = 1-\sqrt{\alpha}, \rho = \sqrt{\frac{\alpha}{1-\alpha}}$ where $\kappa_W = \|W\|\|W^{-1}\|$. Thus, there exist matrices $H_1, H_2, ...,$ and $L_1, L_2 ...,$ such that $\hat{M}_t := \hat{A}_t + \hat{B}_t K_t = H_t L_t H_t^{-1}$ with the following properties: (a) $\|L_t\| \le 1-\gamma$; (b) $\|H_t\| \le \beta_1$ and $\|H_t^{-1}\| \le 1/\beta_2$ with $\kappa = \beta_1/\beta_2$; (c) $\|H_{t+1}^{-1}H_t\| \le \frac{\rho}{1-\gamma}$. With this decomposition for $\hat{M}_t$, we show that $M_t := A_t + B_t K_t$ can be decomposed similarly. Let $\Delta_t = M_t - \hat{M}_t$. Then,

$$M_t = \hat{M}_t + \Delta_t = H_t L_t H_t^{-1} + H_t H_t^{-1}\Delta_t H_t H_t^{-1} = H_t(L_t + H_t^{-1}\Delta_t H_t)H_t^{-1} = H_t L_t' H_t^{-1}. \tag{4.16}$$

We next upper bound $\|L_t'\|$. Notice that

$$\begin{aligned}
\|\Delta_t\| = \|M_t - \hat{M}_t\| &= \|A_t + B_t K_t - \hat{A}_t - \hat{B}_t K_t\| \\
&\le \|A_t - \hat{A}_t\| + \|B_t - \hat{B}_t\|\|K_t\| \\
&\le \max\{\|A_t - \hat{A}_t\|, \|B_t - \hat{B}_t\|\}(1 + \|K_t\|) \\
&\le \delta \frac{\gamma}{\kappa(1+K_{max})}(1+\|K_t\|) \le \delta\frac{\gamma}{\kappa}.
\end{aligned}$$

Then, we have the following bound on $\|L_t'\|$,

$$\|L_t'\| = \|L_t + H_t^{-1}\Delta_t H_t\| \le \|L_t\| + \|H_t^{-1}\|\|\Delta_t\|\|H_t\| \le 1-\gamma+\delta\gamma = 1-(1-\delta)\gamma. \tag{4.17}$$

Define $\gamma' = (1-\delta)\gamma$ and $\rho' = \frac{1-(1-\delta)\gamma}{1-\gamma}\rho$. We claim that $(K_t)_{t=0}^\infty$ is (κ,γ',ρ') sequential strongly stable for system $(A_t, B_t)_{t=0}^\infty$. In the decomposition (4.16), L_t' satisfies $\|L_t'\| \le 1-\gamma'$ (which we have proved in (4.17)), and by definition H_t satisfies $\|H_t\| \le \beta_1$ and $\|H_t^{-1}\| \le 1/\beta_2$ with $\kappa = \beta_1/\beta_2$; $\|H_{t+1}^{-1}H_t\| \le \frac{\rho}{1-\gamma} = \frac{\rho'}{1-\gamma'}$. Therefore, the only conditions we need to verify are (a) $0 < \gamma' \le 1$ and (b) $0 \le \rho' < 1$. Condition (a) is trivial since for any $\alpha \in [0, \frac{1}{2})$, we have $\delta \in (0, \frac{\sqrt{1-\alpha}-\sqrt{\alpha}}{1-\sqrt{\alpha}}) \subset (0,1)$ and hence $\gamma' = (1-\delta)\gamma$ lies in $(0,1]$. For condition (b), as $\rho' = \frac{1-(1-\delta)\gamma}{1-\gamma}\rho$, where $\rho =$

$\sqrt{\frac{\alpha}{1-\alpha}}$ and $\gamma = 1 - \sqrt{\alpha}$, it follows that $\rho' < 1$ is equivalent to the following condition:

$$\rho' = \frac{1-(1-\delta)\gamma}{1-\gamma}\rho = \frac{\delta + (1-\delta)\sqrt{\alpha}}{\sqrt{1-\alpha}} < 1. \tag{4.18}$$

Reorganizing the terms we have $\delta < \frac{\sqrt{1-\alpha}-\sqrt{\alpha}}{1-\sqrt{\alpha}}$, which is satisfied by our condition on δ.

Upper bounding $\|K_t\|$. The only thing that remains to be shown is that there exists an upper bound on the controller gain matrix under the conditions stated in the theorem.

Note that K_t must satisfy $\hat{M}_t = \hat{A}_t + \hat{B}_t K_t$. On the other hand, by the COCO-LQ formulation, K_t solves the following problem

$$\begin{aligned} \min_{K_t} \quad & \mathrm{Tr}(RK_t\Sigma_{xx}K_t^\top) \\ \text{s.t. } & \hat{M}_t = \hat{A}_t + \hat{B}_t K_t, \end{aligned}$$

where Σ_{xx} is the solution to the COCO-LQ formulation at time t. The Lagrangian of the above optimization is

$$L(K_t, \Gamma) = \mathrm{Tr}(RK_t\Sigma_{xx}K_t^\top) + \mathrm{Tr}(\Gamma^\top(\hat{A}_t + \hat{B}_t K_t - \hat{M}_t)).$$

The optimizer K_t must satisfy the following condition,

$$\nabla_{K_t} L(K_t, \Gamma) = 2RK_t\Sigma_{xx} + \hat{B}_t^\top \Gamma = 0.$$

$$\hat{M}_t = \hat{A}_t + \hat{B}_t K_t.$$

Together, we obtain $K_t = R^{-1}\hat{B}_t^\top(\hat{B}_t R^{-1}\hat{B}_t^\top)^{-1}(\hat{M}_t - \hat{A}_t)$ and

$$\|K_t\| \le \|R^{-1}\|\|\hat{B}_t^\top\| \frac{1}{\sigma_{\min}(R^{-1})\sigma_{\min}(\hat{B}_t\hat{B}_t^\top)}(\kappa(1-\gamma) + \|\hat{A}_t\|), \tag{4.20}$$

where we have used by the (κ, γ, ρ) sequential strong stability of $(K_t)_{t=0}^\infty$ for system $(\hat{A}_t, \hat{B}_t)_{t=0}^\infty$, we have $\|\hat{M}_t\| = \|\hat{A}_t + \hat{B}_t K_t\| \le \kappa(1-\gamma)$. By our condition, we have $\|\hat{A}_t\| \le \bar{\sigma}_A$, $\|\hat{B}_t\| \le \bar{\sigma}_B$, $\hat{B}_t\hat{B}_t^\top \succeq \underline{\sigma}_B^2$, and $\kappa_R = \|R\|\|R^{-1}\|$. Then, we have $\|K_t\|$ upper bounded as follows,

$$\|K_t\| \le \kappa_R \frac{\bar{\sigma}_B}{\underline{\sigma}_B^2}(\kappa(1-\gamma) + \bar{\sigma}_A).$$

□

This result highlights the tradeoff between the estimation error and the algorithm performance. If we choose a small α, the algorithm can tolerate a larger estimation error (i.e., larger right-hand side of (4.15) can be obtained) but may lead to high control cost due to the tight state covariance constraint. If we choose a larger α, the algorithm tolerates smaller estimation errors while its performance improves due to the less strict state covariance constraint.

4.2.4 Experiments

The results in the previous section focus on stability of the COCO-LQ approach. Here, we use experimental results to highlight that COCO-LQ also performs near-optimally in terms of cost while also stabilizing systems that the naive approach based on LTI control cannot. We first test our method on random, synthetic linear time-varying systems, and then demonstrate the performance of COCO-LQ in real-world power system frequency control settings. Finally, we test the performance of COCO-LQ for online control of linear time varying systems that are derived from local linearization of nonlinear systems.

Synthetic Time-Varying Systems

We first consider the control of switching and time-varying systems. The cost function is set as $Q = 0.2I, R = I$, and system is subject to Gaussian disturbance $w_t \sim N(0, 0.1^2)$. We average the simulation results over 5 runs and visualize the mean performance and standard deviation.

a) We consider a switching system similar to Example 4.1 in Section 4.2.2, where A_t alternates between

$$A = \begin{bmatrix} 0.99 & 0 \\ a & 1.5 \end{bmatrix}, \quad A' = \begin{bmatrix} 0.99 & 1.5 \\ 0 & 0.99 \end{bmatrix},$$

and $B_t = I$.

b) We consider a system with

$$A_t = \begin{bmatrix} 0.99 & \sin(\frac{\pi t}{2})|e^{t/60} \\ |\cos(\frac{\pi t}{2})|e^{t/60} & 0.99 \end{bmatrix}$$

that is continually changing over time, and $B_t = I$.

As we can see in Figure 4.1, COCO-LQ is able to quickly and effectively stabilize the system under various time-varying scenarios, which corraborates our theoretical findings. As α increases, the acquired cost of COCO-LQ first decreases and then increases (explosion of state), highlighting that α can explicitly control the tradeoff between cost and stability. With proper selection of α, COCO-LQ achieves near-optimal cost (within 30% of the offline optimal for both system a) and b).

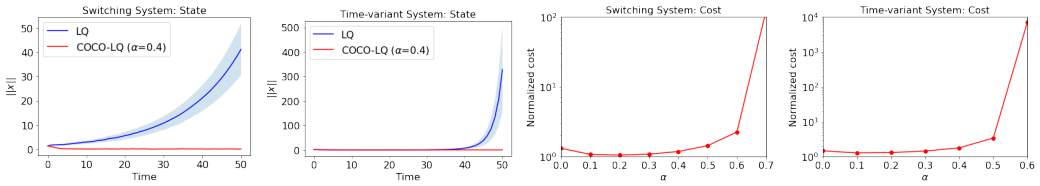

Figure 4.1: Performance comparison of COCO-LQ and naive LQ control on synthetic systems given in a) and b). The left two figures show the state evolution, and right two figures show the normalized cost (cost of COCO-LQ divided by cost of the offline optima) under different α.

Frequency Control with Renewable Generation

We now consider a power system frequency control problem on standard IEEE WECC 3-machine 9-bus system (Figure 4.2**(Left)**), which is a widely adopted system used in frequency stability studies. The state space model of power system frequency dynamics follows [111],

$$\underbrace{\begin{bmatrix}\dot{\theta}\\ \dot{\omega}\end{bmatrix}}_{\dot{\mathbf{x}}} = \underbrace{\begin{bmatrix}0 & I\\ -M_t^{-1}L & -M_t^{-1}D\end{bmatrix}}_{A_t}\begin{bmatrix}\theta\\ \omega\end{bmatrix} + \underbrace{\begin{bmatrix}0\\ M_t^{-1}\end{bmatrix}}_{B_t}\underbrace{p_{in}}_{\mathbf{u_t}}, \tag{4.21}$$

where the state variable is defined as the stacked vector of the voltage angle θ and frequency ω. $M_t = \text{diag}(m_{t,i})$ is the inertia matrix, where $m_{t,i}$ represents the equivalent rotational inertia at bus i and time t. M_t is time-varying and depends on the mix of online generators, since only thermal generators provide rotational inertia and renewable generation does not [276]. $D = \text{diag}(d_i)$ is the damping matrix, where d_i is the generator damping coefficient. L is the network susceptance matrix. The control variable p_{in} corresponds to the electric power generation.

We assume the system is changing between two states: a high renewable generation scenario where $m_{t,i} = 2$ (i.e., 80 percent renewable with zero inertia and 20 percent of thermal generation with 10s inertia), and a low renewable generation scenario where $m_{t,i} = 8$ (i.e., 20 percent renewable and 80 percent thermal generation), with additional random fluctuations between $[0, 0.2]$. This setup represents the real-world situation where we have high solar output during the daytime, and low output in the morning/evening, with intra-day variations due to clouds and weather changes. Notice that B_t is not full rank, thus we need to leverage predictions, i.e., A_{t+1} and B_{t+1}. For fair comparison, we compete against the H-horizon optimal control in [28], which is the extension of naive LTI controller to use H-step predictions. In both

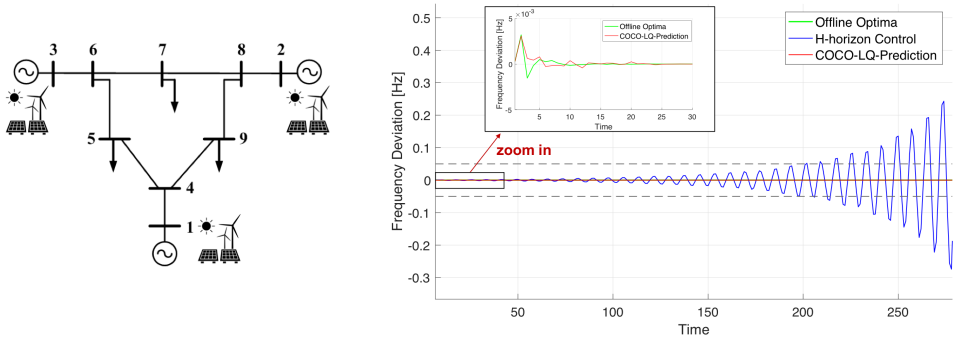

Figure 4.2: **(Left)** IEEE WECC 3-machine 9-bus system schematic with generators at bus 1, 5, 9 are mixture of thermal generation and renewable. **(Right)** Frequency dynamics under offline optima, baseline H-horizon control, and COCO-LQ. The dotted grey lines (±0.05Hz) are the safety margin of power system frequency variation.

cases, we assume the prediction is accurate and use the exact value of A_{t+1} and B_{t+1} for computing control actions.

Figure 4.2**(Right)** visualizes the power system frequency dynamics under three controllers: the offline optimal control, the baseline H-horizon optimal controller, and the proposed COCO-LQ-Prediction method. We ideally desire a controller that is able to maintain the frequency variation within ±0.05Hz and eventually stabilize the system. It can be observed that our algorithm succeeds at maintaining the frequency stability under random, time-varying renewable generations. Furthermore, the performance of COCO-LQ is very close to the offline optimal, while the system frequency diverges under the baseline H-horizon optimal control.

Inverted Pendulum Swingup Task

We further test COCO-LQ for online control of linear time-varying systems that are derived from the local linearization of nonlinear systems. In particular, we consider the pendulum swing-up task, where the system dynamics are described by

$$\begin{bmatrix} \dot{\theta} \\ \ddot{\theta} \end{bmatrix} = \begin{bmatrix} \dot{\theta} \\ \frac{g}{l} sin\theta \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{ml^2} \end{bmatrix} u = f(x, u), \tag{4.22}$$

where $x = [\theta, \dot{\theta}]$ represents the system state, in which θ, $\dot{\theta}$ and $\ddot{\theta}$ are the angle, angular velocity and angular acceleration. g is the gravitational acceleration, l and m are the length and mass of the pendulum. The control goal is to stabilize the pendulum at the straight up position with $\theta = 0$ and $\dot{\theta} = 0$, starting from any initial

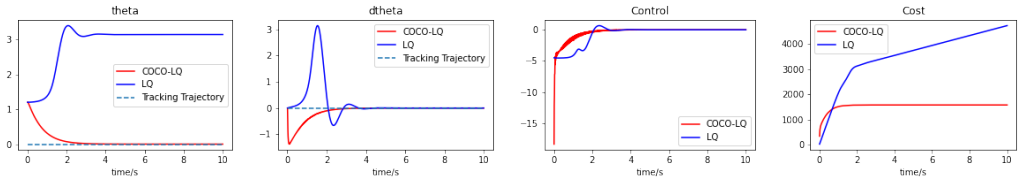

Figure 4.3: Performance comparison of COCO-LQ and LQ on inverted pendulum via locally linearization. The left two figures show the state evolution of angle θ and angular velocity $\dot{\theta}$. Initial angle is set as $\theta = 1.2$rad, and the desired state is $\theta = \dot{\theta} = 0$. The right two figures show the control action and cost comparison, with $Q = R = I$.

angle and velocity. The pendulum dynamics described by (4.22) is a nonlinear system, by taking local linear approximation at each x_t, we can approximate the nonlinear system via a linear time varying system,

$$\begin{bmatrix} \dot{\theta} \\ \ddot{\theta} \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 \\ \frac{g}{l}\cos\theta_t & 0 \end{bmatrix}}_{A_t} \begin{bmatrix} \theta \\ \dot{\theta} \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ \frac{1}{ml^2} \end{bmatrix}}_{B_t} u, \tag{4.23}$$

Fig 4.3 compares the performance of the proposed COCO-LQ approach and baseline LQ approach. As we can observe from the left two plots, COCO-LQ is able to stabilize the pendulum at the desired position due to its robustness under model estimation error (caused by local linear approximation) and time-varying dynamics, while the naive LQR approach fails to achieve the swing-up task and stabilize the system. The right two plots show the evolution of the control efforts and cost. COCO-LQ initially outputs significantly larger control actions compared to the baseline LQ controller, to stop the pendulum from falling over. The overall control cost of COCO-LQ quickly converges once the swing-up task is finished, while the cost of LQ control keeps growing.

4.2.5 Conclusion

In this section, we studied the stability of LTV systems. Our results demonstrate the challenge of ensuring stability for LTV systems compared to LTI systems. Motivated by this challenge, we propose a COCO-LQ/COCO-LQ-prediction policy that can guarantee stability for LTV systems under certain assumptions. There are many interesting open questions that remain. For example, the bound $\alpha < 1/2$ in Theorem 4.1 is a sufficient condition, and studying how to relax the bound and how to derive instance-dependent bounds is an interesting future question. Another

important direction is to analyze the performance (e.g., the regret) of the proposed approach in order to quantify the trade-off between stability and performance.

4.3 Stability and Identification of Random Asynchronous LTI Systems

In this section, we introduce the random asynchronous LTI systems, a new form of linear time-varying systems, and study the stability characterization and the system identification problem in that setting. We empirically show that the new dynamical system evolution formulation based on randomness and asynchrony brings novel challenges and opportunities in terms of stabilization of linear dynamical systems (LDS). We empirically demonstrate that the mean-square stability of random asynchronous LTI systems has a delicate dependency on the update and asynchrony probabilities. Furthermore, the stability of synchronous LDS does not imply stability of asynchronous LDS, and vice versa. For a randomly generated system, by changing the asynchrony or the update probability in the system, we show that unstable synchronous LTI systems may be stabilized or stable synchronous LTI systems may have unstable dynamics.

We consider the mean-square stability characterization of randomized LTI systems, which are special cases of random asynchronous LTI systems with no delay. We show that this setting corresponds to the model introduced in Teke and Vaidyanathan [261], and the stability is governed by an extended Lyapunov equation. Relating the extended Lyapunov equation to standard Lyapunov equation of synchronous LTI system, we discuss the precise characterization of the mean-square stability of randomized LTI systems.

We propose a novel method to recover the impulse response of the “average system", as well as the true underlying system parameters, the update probability, noise, and input covariances for unknown stable randomized LTI systems. In order to achieve this, we first visit the well-known central limit theorem for Markov chains and solve a least-squares problem to obtain an estimate of the average system parameters of the underlying system. Then, we propose an optimization problem to estimate the update probability and noise covariances that optimally satisfy the extended Lyapunov equation for the estimated average system parameters. By solving the optimization problem analytically, we present a closed-form expression for an estimate of the update probability and noise covariances for the system. The underlying true system parameters that govern the dynamics are ultimately recovered via the estimates of the update probability and the average system parameters.

Finally, we empirically demonstrate the performance of the novel method on a simulated randomized LTI system and show that our proposed method reliably and efficiently recovers the underlying dynamics with the optimal rate. This shows that the presented model and the system identification framework provide a more realistic and computationally efficient alternative to the general switching linear systems in analyzing input/output data collected from an LDS that has a fixed network structure with random asynchronous updates.

4.3.1 Problem Setting

Recall the state-space model of LTI systems introduced in Chapter 3:

$$x_{t+1} = \mathbf{A}x_t + \mathbf{B}u_t + w_t, \tag{4.24}$$

where x_t, u_t, and w_t are the state, input, and noise vectors respectively. In this section, we refer to this model as "synchronous LTI system" since all state elements are updated in every time step using the most recent information on all state elements. We study a random asynchronous variant of this LTI system where the states evolve randomly and asynchronously. The model studied here generalizes (4.24) in two different directions:

1) The state elements get updated randomly in every iteration. If a state element gets updated, it follows the state dynamics. Otherwise, its value remains the same;

2) When a state element is being updated, it may use out-dated information regarding the other state elements. We assume that the delay in information flow is also probabilistic.

More precisely, given a node update probability $0 < p \leq 1$, we consider the following *random and asynchronous* state-space model for all state variables i at each time step t:

$$(x_{t+1})_i = \begin{cases} \sum_{j=1}^{n} A_{i,j}\,(x_{t-k_{i,j}})_j + (\mathbf{B}\,u_t + w_t)_i, & \text{w.p.} \quad p, \\ (x_t + w_t)_i, & \text{w.p.} \quad 1-p, \end{cases} \tag{4.25}$$

where w_t denotes the noise component for the state elements. More importantly, $k_{i,j} \geq 0$ denotes the delay in information observed by the i^{th} element regarding the j^{th} element. So, state variables are allowed to observe different amount of delay regarding other state variables. In our model, we will consider *random and*

independent delays with the following distribution:

$$\mathbb{P}[k_{i,j} = \tau] = \begin{cases} q\,(1-q)^{\tau}, & \tau = \{0, \ldots, h-1\}, \\ (1-q)^{h}, & \tau = h, \end{cases} \tag{4.26}$$

for some fixed delay probability $0 < q \leq 1$ and finite h. So, higher values of q implies lower amount of delay in information flow. In summary, in every time-step, each state element is updated with probability p using linear dynamics based on the most recent data available from other state variables, or its value remains the same (up to an input noise) with probability $1 - p$.

The model (4.25) captures the random asynchrony of large-scale LDS (e.g., networked control systems, social networks, and biological networks), where each state variable could be considered as a sensor/node. At any time step, the update on the node happens randomly. The node may not have the most recent data from the other nodes, and it updates its state based on the available (possibly outdated) information. It is possible to extend this model in such a way that each state variable has a different update probability, a different delay probability or updated and non-updated state elements have different noise components. However, for the sake of simplicity, we assume that all the state variables are updated with the same probability, the same delay scheme, and the same noise characteristics. Notice that the model (4.25) reduces to the standard *synchronous* state-space model (4.24) when the probabilities are selected as $p = 1$ and $q = 1$. As a result, one can consider (4.25) as a random asynchronous extension of synchronous LTI systems that are studied in the last century.

For the given system in (4.25), $x_t, w_t \in \mathbb{R}^{d_x}$, $u_t \in \mathbb{R}^{d_u}$, and the system matrices $\mathbf{A}$ and $\mathbf{B}$ have appropriate dimensions accordingly. Note that this work considers the problem in real domain for a simpler presentation. Nevertheless, the results can be easily extended to complex domain with proper conjugation operations. Without loss of generality, we assume that $x_0 = \mathbf{0}$. Furthermore, we assume that the input and noise vectors are zero mean and have unknown variance, i.e.,

$$\mathbb{E}[u_t] = \mathbb{E}[w_t] = \mathbf{0}, \;\; \mathbb{E}[u_t\, u_k^{\top}] = \mathbf{U}, \;\; \mathbb{E}[w_t\, w_k^{\top}] = \delta(t-k)\;\sigma_w^2\, \mathbf{I}_{d_x} \tag{4.27}$$

for some *unknown* $\mathbf{U} \succeq 0$ and $\sigma_w^2 > 0$.

4.3.2 Stability of Random Asynchronous LTI Systems

Since the random asynchronous LTI systems have stochastic behavior, the stability of the state vector x_k should be considered statistically. Therefore, we numerically

study the mean-square stability of (4.25). Note that for this setting, the mean-square stability implies almost-sure stability [147].

Consider the random asynchronous systems given in (4.25) with the following transition matrices,

$$\mathbf{A}_1 = \begin{bmatrix} 0.05 & 0.36 & 0.39 \\ 0.01 & -0.37 & 0.23 \\ 0.23 & 0.23 & -0.98 \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} 0.78 & -1.45 & -1.12 \\ -0.20 & -0.25 & 0.36 \\ 0.49 & -0.58 & 0.20 \end{bmatrix}, \tag{4.28}$$

with $h = 2$. Note that the spectral radius of $\rho(\mathbf{A}_1) \approx 1.1065$, i.e., the synchronous LTI system with $\mathbf{A}_1$ is unstable. In order to numerically examine the mean-square stability of the random asynchronous system, we study the system from Markov-jump linear system perspective. In particular, the system in (4.25) can be represented as a Markov-jump linear system of $d_x(h+1)$ dimensions that switches between $(1 + (h+1)^{d_x})^{d_x}$ possible systems. Let $\mathbf{S}_h \in \mathbb{R}^{(d_x(h+1))^2 \times (d_x(h+1))^2}$ be the matrix that governs the evolution of the correlation matrix of $d_x(h+1)$ variables for the Markov-jump system. The stability of $\mathbf{S}_h$ is equivalent to the mean-square stability of the Markov-jump linear system [63] and thus the mean-square stability of (4.25). Therefore, for all p and q values, we construct the corresponding $\mathbf{S}_h$ and consider the spectral radius of $\mathbf{S}_h$. The first figure in Figure 4.4 depicts the mean-square stable and unstable regions for all p and q values for the random asynchronous system with $\mathbf{A}_1$.

In this figure, we observe that the synchronous variant ($p = q = 1$) of this system is unstable as expected. However, it also shows that by randomizing the updates, i.e., decreasing p, or increasing the asynchrony in the updates, i.e., decreasing q, one can achieve stable system dynamics from this unstable system. This observation provides a novel perspective to the common perception of randomization and asynchrony. Even though they are usually considered as the ways to decrease the cost or delays in the expense of accuracy and convergence, this result shows that they can be also utilized as the mechanisms to stabilize the dynamical systems. On the other hand, we should note that for this particular system, a drastic increase in asynchrony still results in unstable system dynamics (top left region on the figure). Moreover, for totally non-random system ($p = 1$), moderate levels of asynchrony ($q \in [0.25, 0.65]$) stabilizes the system but we obtain unstable dynamics toward both extremes of asynchrony. This highlights the fact that a careful study is required for understanding the precise effects of randomization and asynchrony on the stability of LTI systems. Next, we consider a stable synchronous LTI system with the state transition matrix of $\mathbf{A}_2$, i.e., $\rho(\mathbf{A}_2) \approx 0.9778$. Similar to the unstable case, we compute $\mathbf{S}_h$ for this random

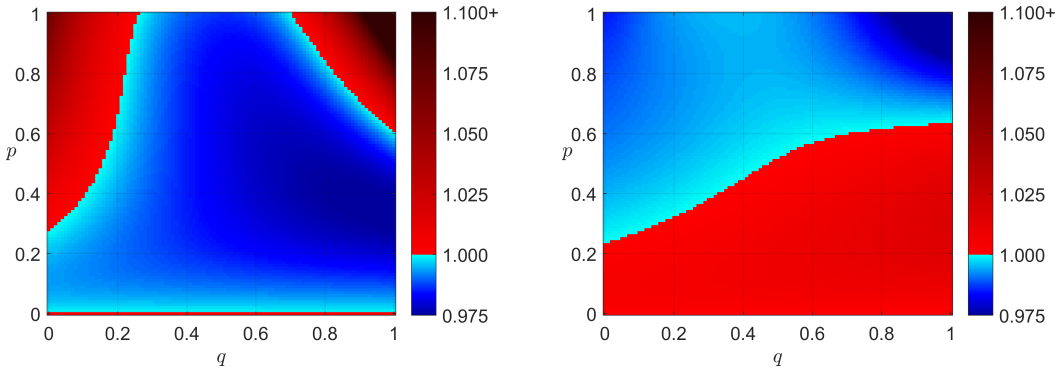

Figure 4.4: The spectral radius of $\mathbf{S}_h$ that represents mean-square stability of the random asynchronous systems with $h = 2$ and (a) unstable $\mathbf{A}_1$ and (b) stable $\mathbf{A}_2$ state transition matrices

asynchronous system and the second plot in Figure 4.4 shows the spectral radius of it for all p and q values. The stability behavior of this random asynchronous system is significantly different. Surprisingly, increasing randomization for this system provides mean-square unstable dynamics in all asynchrony conditions. In addition, for moderately non-random updates ($p > 0.6$), all levels of asynchrony provides mean-square stability.

In these settings, the numerical computation of $\mathbf{S}_h$ was feasible since the systems in (4.28) have $d_x = 3$ and $h = 2$, yielding 21952 possible systems to switch between. For large-scale LDS, one needs to compute the closed-form expression of $\mathbf{S}_h$ in order to characterize the mean-square stability of random asynchronous LTI systems. The theoretical analysis of the mean-squared stability of the general system given in (4.25) is left for future work. However, in the following section, we provide the precise closed-form characterization of $\mathbf{S}_h$ for $h = 0$ which is a special case of random asynchronous LTI systems where $q = 1$ and $0 < p \leq 1$, named as randomized LTI systems.

4.3.3 Randomized LTI Systems

Since $q = 1$, randomized LTI systems do not have any delays or asynchrony in the system, i.e., if the state element getting updated it has access to the most recent information on all states. The random delay probabilities reduces to $\mathbb{P}[k_j = 0] = 1$ for all j. Thus, we get the following model:

$$(x_{t+1})_i = \begin{cases} (\mathbf{A}\, x_t + \mathbf{B}\, u_t + w_t)_i, & \text{w.p.} \quad p, \\ (x_t + w_t)_i, & \text{w.p.} \quad 1 - p. \end{cases} \tag{4.29}$$

This model corresponds to the setting first introduced in Teke and Vaidyanathan [261]. In the following, we consider the properties of randomized LTI systems.

Markov Parameters

Markov parameters of an LDS is the unique matrix impulse response of the system. For a synchronous LTI system, the Markov parameters of the system $(\mathbf{A}, \mathbf{B})$ are given as $\mathbf{H}_k = \mathbf{A}^{k-1}\mathbf{B}$ for $k \geq 1$. From the input-output viewpoint, the randomized updates on the system with $(\mathbf{A}, \mathbf{B})$ can be represented in an average sense as a synchronous LTI system with parameters $(\overline{\mathbf{A}}, \overline{\mathbf{B}})$ where the average state-transition matrix $\overline{\mathbf{A}}$ and the average input matrix $\overline{\mathbf{B}}$ are given as follows:

$$\overline{\mathbf{A}} = p\,\mathbf{A} + (1-p)\,\mathbf{I}_{d_x}, \qquad \overline{\mathbf{B}} = p\,\mathbf{B}. \tag{4.30}$$

As a result, Markov parameters of the average system can be obtained as $\overline{\mathbf{H}}_k = \overline{\mathbf{A}}^{k-1}\overline{\mathbf{B}}$, for $k \geq 1$. Notice that Markov parameters of the underlying system and the randomized system can be directly obtained from each other. The k^{th} Markov parameter of the randomized system, $\overline{\mathbf{H}}_k$, can be written as a linear combination of the first k Markov parameters of the synchronous system:

$$\overline{\mathbf{H}}_k = \left(p\mathbf{A} + (1-p)\mathbf{I}_{d_x}\right)^{k-1} p\mathbf{B} = \sum_{i=1}^{k} \binom{k-1}{i-1} p^i (1-p)^{k-i} \mathbf{H}_i.$$

More generally, define the first K Markov parameters matrices $\mathbf{G} = [\mathbf{H}_1\ \mathbf{H}_2\ \cdots\ \mathbf{H}_K]$ and $\overline{\mathbf{G}} = [\overline{\mathbf{H}}_1\ \ \overline{\mathbf{H}}_2\ \ \cdots\ \ \overline{\mathbf{H}}_K]$. We have $\overline{\mathbf{G}} = \mathbf{G}(\mathbf{T} \otimes \mathbf{I}_{d_u})$, where $\mathbf{T} \in \mathbb{R}^{K \times K}$ is an upper triangular matrix with $\mathbf{T}_{i,j} = \binom{j-1}{i-1} p^i (1-p)^{j-i}$ for $j \geq i \geq 1$. Notice that $\mathbf{T}$ does not depend on the system parameters $(\mathbf{A}, \mathbf{B})$, and it is determined solely by the probability p. Moreover, $\mathbf{T}$ has diagonal entries (thus eigenvalues) of $\mathbf{T}_{i,i} = p^{i-1}$. Thus, $\mathbf{T}$ is always invertible since we trivially assumed that $p > 0$. This shows that once the average system behavior and the rate of updates are known, one can identify the underlying system parameters exactly. When the update probability is $p = 1$ (synchronous), we get $\mathbf{T} = \mathbf{I}_K$ so that $\overline{\mathbf{G}} = \mathbf{G}$. Note that the properties above could be trivially extended to measurement feedback systems, i.e., randomized partially observed LTI systems.

Mean-Squared Stability

In this section, we precisely characterize the mean-square stability of randomized LTI systems which corresponds to the rightmost vertical axes of plots in Figure 4.4. As discussed above, the dynamics of the randomized LTI system is determined by

the matrix $\overline{\mathbf{A}}$ in an average sense. The stability of the matrix $\overline{\mathbf{A}}$ is necessary, but not *sufficient* for the stability of the system. In order to analyze the mean-square stability, we look for the condition that ensures $\mathbb{E}[x_t\, x_t^\top]$ stays finite as $t \to \infty$. In fact, the steady-state covariance matrix, i.e., $\lim_{t\to\infty} \mathbb{E}[x_t\, x_t^\top] = \mathbf{\Gamma}$, can be found as the solution of the following *extended* Lyapunov equation introduced in Teke and Vaidyanathan [262]:

$$\mathbf{\Gamma} = \phi(\mathbf{\Gamma}) + \overline{\mathbf{B}}\,\mathbf{U}\,\overline{\mathbf{B}}^\top + \left(\frac{1}{p} - 1\right)\left(\overline{\mathbf{B}}\,\mathbf{U}\,\overline{\mathbf{B}}^\top\right) \odot \mathbf{I} + \sigma_w^2 \mathbf{I}, \tag{4.31}$$

where the function $\phi(\cdot)$ is defined as follows:

$$\phi(x) = \overline{\mathbf{A}} x \overline{\mathbf{A}}^\top + (p - p^2)\big((\mathbf{A} - \mathbf{I}) x (\mathbf{A}^\top - \mathbf{I})\big) \odot \mathbf{I} = \overline{\mathbf{A}} x \overline{\mathbf{A}}^\top + \left(\frac{1}{p} - 1\right)\big((\overline{\mathbf{A}} - \mathbf{I}) x (\overline{\mathbf{A}}^\top - \mathbf{I})\big) \odot \mathbf{I}.$$

The function $\phi(\cdot)$ is a *positive linear map* that controls the evolution of the state covariance matrix in the extended Lyapunov equation. It can be vectorized as $\mathrm{vec}(\phi(x)) = \mathbf{S}\ \mathrm{vec}(x)$ where

$$\mathbf{S} = \overline{\mathbf{A}} \otimes \overline{\mathbf{A}} + (p - p^2)\,\mathbf{J}\,(\mathbf{A} - \mathbf{I}) \otimes (\mathbf{A} - \mathbf{I}), \tag{4.32}$$

for $\mathbf{J} = \sum_{i=1}^{d_x} \left(e_i e_i^\top\right) \otimes \left(e_i e_i^\top\right)$. $\mathbf{S} \in \mathbb{R}^{d_x^2 \times d_x^2}$ is the matrix representation of the linear map $\phi(\cdot)$ and corresponds to $\mathbf{S}_h$ introduced in Section 4.3.2 at $h = 0$. Note that to extend this to complex valued systems, (4.32) needs element-wise conjugate operations on the left-side of Kronecker products.[2] Recall that in Section 4.3.2, the mean-square stability of the random asynchronous LTI systems is demonstrated numerically. However, with the closed-form expression of $\mathbf{S}$ in (4.32), we can analytically characterize the stability of the randomized LTI system.

Lemma 4.4. *[260, 262] The randomized LTI systems given in (4.29) are mean-square stable if and only if* $\rho(\mathbf{S}) < 1$.

This result is due to the fact that one can recursively represent the covariance matrix of the state variables at time $t + 1$ as a function of the covariance matrix at time t. Since $\mathbf{S}$ represents this mapping, stability of $\mathbf{S}$ is a necessary and sufficient condition for the convergence of the covariance matrix, which implies mean-square stability for the state variables. The key observation in Lemma 4.4 is that the

[2] Element-wise conjugation ensures that $\mathbf{S}$ always has a real nonnegative eigenvalue that is equal to its spectral radius, and the corresponding eigenvector is the vectorized version of a positive semidefine matrix. This follows from the extensions of the Perron-Frobenius theorem to positive maps in more general settings, Theorem 5 of Karlin [139].

mean-square stability of the randomized LTI system and the stability of $\mathbf{A}$ do not imply each other, i.e., $\rho(\mathbf{S}) < 1$ and $\rho(\mathbf{A}) < 1$ are not equivalent in general. Note that Lemma 4.4 provides the precise characterization of the rightmost axes of the plots in Figure 4.4.

In order to visualize the convergence behavior of the randomized updates, we consider a numerical test example of size $d_x = 2$ with a constant input (i.e., fixed-point iteration), and initialize x_0 with independent Gaussian random variables (left-most block in Figure 4.5). Then, the distribution of the state vector x_t (at time t) follows a Gaussian mixture model (GMM) due to the randomized nature of the updates (See Figure 4.5). Furthermore, the stability of the matrix $\mathbf{S}$ ensures that the mean of x_t converges to the fixed-point of the system while the variance of x_t converges to zero.

The key insight to the convergence behavior in Figure 4.5 is as follows: When represented as a switching system, the randomized LTI model (4.29) switches between 2^{d_x} systems randomly, and it can be shown that all these 2^{d_x} systems (including the original system) have the same fixed-point. It should also be noted that not all 2^{d_x} systems are stable by themselves, and an arbitrary switching does not necessarily ensure the convergence. Nevertheless, with a careful selection of the probability, the randomized model can obtain convergence even when the synchronous system is unstable.

When the system is mean-square stable, the steady-state covariance matrix, $\mathbf{\Gamma}$, is given as

$$\operatorname{vec}(\mathbf{\Gamma}) = (\mathbf{I} - \mathbf{S})^{-1} \left(\left(p^2 \, \mathbf{I} + (p - p^2) \, \mathbf{J} \right) \operatorname{vec}(\mathbf{B} \, \mathbf{U} \, \mathbf{B}^\top) + \sigma_w^2 \operatorname{vec}(\mathbf{I}) \right). \tag{4.33}$$

When $p = 1$, we have $\phi(x) = \mathbf{A} \, x \, \mathbf{A}^\top$, which implies that $\mathbf{\Gamma} = \mathbf{A} \, \mathbf{\Gamma} \, \mathbf{A}^\top + \mathbf{B} \, \mathbf{U} \, \mathbf{B}^\top + \sigma_w^2 \, \mathbf{I}$ and $\rho(\mathbf{S}) = \rho^2(\mathbf{A})$. So, we have $\rho(\mathbf{S}) < 1$ if and only if $\rho(\mathbf{A}) < 1$ for synchronous LTI systems, which recovers the well-known stability result in the classical systems theory.

4.3.4 System Identification for Randomized LTI Systems

In this section, we propose a system identification method for learning unknown mean-square stable randomized LTI systems from a single input-output trajectory. Regarding the underlying system, we do not have any other assumptions besides stability, *i.e.* $\rho(\mathbf{S}) < 1$, and the assumptions in (4.27).

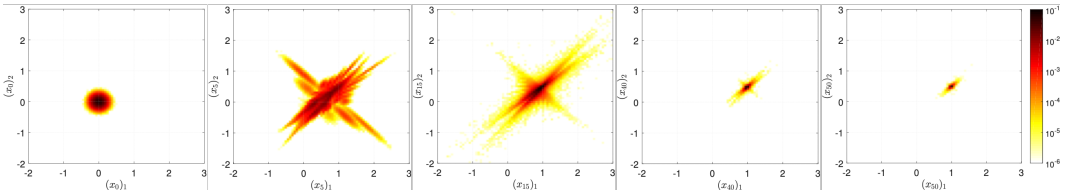

Figure 4.5: Evolution of the state vector for a mean-square stable (but synchronously unstable) 2-dimensional randomized LTI system with a fixed input and Gaussian initialization

First, recall the Markov chain central limit theorem (MC-CLT). Assume that we have a Markov chain at its stationary distribution. MC-CLT states that, the sample average of any measurable, finite-variance and real-valued function of a sequence of n variables from this Markov chain converge to a Gaussian distribution as $n \to \infty$, where mean is the expected value of this function at the stationary distribution and the variance linearly decays in n [129].

Notice that the randomized updates of (4.29) form an ergodic Markov chain (due to independent selection in every iteration) and the stability of the system guarantees the stationary distribution. We also know that the stable systems converge exponentially fast to their steady state, i.e., Markov chain formed by (4.29) quickly approaches to its stationary distribution. In light of these observations, we can deduce that, as the number of collected input-output samples T increases, the sample state correlation and input-output cross correlation matrices converge to their expected values with the rate of $1/\sqrt{T}$. In particular, given a sequence of inputs and outputs $\{x_0, u_0, x_1, \ldots, u_{T-1}, x_T\}$, let

$$\mathbf{C}_0 = \frac{1}{T}\sum_{t=0}^{T-1}\begin{bmatrix} x_t \\ u_t \end{bmatrix}\begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top, \qquad \mathbf{C}_1 = \frac{1}{T}\sum_{t=1}^{T} x_t \begin{bmatrix} x_{t-1} \\ u_{t-1} \end{bmatrix}^\top. \tag{4.34}$$

According to MC-CLT, as $T \to \infty$, $\mathbf{C}_0$ and $\mathbf{C}_1$ converge to $\mathbb{E}[\mathbf{C}_0]$ and $\mathbb{E}[\mathbf{C}_1]$ respectively, where

$$\mathbb{E}\left[\mathbf{C}_0\right] = \begin{bmatrix} \mathbf{\Gamma} & \mathbf{0} \\ \mathbf{0} & \mathbf{U} \end{bmatrix}, \qquad \mathbb{E}\left[\mathbf{C}_1\right] = \begin{bmatrix} \overline{\mathbf{A}}\mathbf{\Gamma} & \overline{\mathbf{B}}\mathbf{U} \end{bmatrix}. \tag{4.35}$$

Therefore, using $\mathbf{C}_1\mathbf{C}_0^{-1}$ converges to the average state transition and input matrices $\begin{bmatrix} \overline{\mathbf{A}} & \overline{\mathbf{B}} \end{bmatrix}$. Notice that $\mathbf{C}_1\mathbf{C}_0^{-1}$ is in fact the solution of the following least squares problem:

$$\arg\min_{\Theta} \sum_{t=1}^{T} \operatorname{tr}\left(\left(x_t - \Theta\begin{bmatrix} x_{t-1} \\ u_{t-1} \end{bmatrix}\right)\left(x_t - \Theta\begin{bmatrix} x_{t-1} \\ u_{t-1} \end{bmatrix}\right)^\top\right). \tag{4.36}$$

Thus, we are guaranteed to recover the average system consistently via (4.36). This result could be extended to recover first K Markov parameters of the randomized partially observable LTI systems. Define $\mathbf{E} = [\mathbf{I}_{d_x} \quad \mathbf{0}] \in \mathbb{R}^{d_x \times (d_x + d_u)}$. Then, the extended Lyapunov equation can be written as

$$\mathbf{E}\begin{bmatrix}\mathbf{\Gamma} & \mathbf{0}\\ \mathbf{0} & \mathbf{U}\end{bmatrix}\mathbf{E}^\top = \begin{bmatrix}\overline{\mathbf{A}} & \overline{\mathbf{B}}\end{bmatrix}\begin{bmatrix}\mathbf{\Gamma} & \mathbf{0}\\ \mathbf{0} & \mathbf{U}\end{bmatrix}\begin{bmatrix}\overline{\mathbf{A}} & \overline{\mathbf{B}}\end{bmatrix}^\top + \Big(1/p - 1\Big)\Bigg(\begin{bmatrix}\overline{\mathbf{A}} - \mathbf{I} & \overline{\mathbf{B}}\end{bmatrix}\begin{bmatrix}\mathbf{\Gamma} & \mathbf{0}\\ \mathbf{0} & \mathbf{U}\end{bmatrix}\begin{bmatrix}\overline{\mathbf{A}} - \mathbf{I} & \overline{\mathbf{B}}\end{bmatrix}^\top\Bigg) \odot \mathbf{I} + \sigma_w^2 \mathbf{I}, \tag{4.37}$$

We know that covariance matrices of the state variables $\mathbf{\Gamma}$ and inputs $\mathbf{U}$ must satisfy (4.37) for a stable randomized LTI system. The central idea for our system identification method is to exploit this fact and recover the randomization probability p, the noise covariance σ_w^2 and the system parameters $\mathbf{A}, \mathbf{B}$ of a stable randomized LTI system. Therefore, we can write extended Lyapunov equation (4.37) in terms of $\mathbf{C}_0$ and $\mathbf{C}_1$ and due to (4.35) expect to have $\mathtt{ly}(\mathbf{C}_0, \mathbf{C}_1) = 0$, where

$$\mathtt{ly}(\mathbf{C}_0, \mathbf{C}_1) \coloneqq \mathbf{E}\mathbf{C}_0\mathbf{E}^\top - \mathbf{C}_1\mathbf{C}_0^{-1}\mathbf{C}_1^\top - \left(\frac{1}{p} - 1\right)\Big((\mathbf{C}_1\mathbf{C}_0^{-1} - \mathbf{E})\mathbf{C}_0(\mathbf{C}_1\mathbf{C}_0^{-1} - \mathbf{E})^\top\Big) \odot \mathbf{I} - \sigma_w^2\mathbf{I}.$$

Thus, to identify the underlying system dynamics, we propose to solve the following:

$$\widehat{p}, \widehat{\sigma}_w^2 = \arg\min_{p, \sigma_w^2} \|\mathtt{ly}(\mathbf{C}_0, \mathbf{C}_1)\|_{\mathrm{F}}^2 . \tag{4.38}$$

This problem can be further simplified to

$$\widehat{p}, \widehat{\sigma}_w^2 = \arg\min_{p, \sigma_w^2} \left\| \boldsymbol{M}_1 - (1/p)\, \boldsymbol{M}_2 - \sigma_w^2\, \mathbf{I} \right\|_{\mathrm{F}}^2 , \tag{4.39}$$

where $\boldsymbol{M}_2 = \Big((\mathbf{C}_1\mathbf{C}_0^{-1} - \mathbf{E})\mathbf{C}_0(\mathbf{C}_1\mathbf{C}_0^{-1} - \mathbf{E})^\top\Big) \odot \mathbf{I}$ and $\boldsymbol{M}_1 = \mathbf{E}\,\mathbf{C}_0\,\mathbf{E}^\top - \mathbf{C}_1\mathbf{C}_0^{-1}\,\mathbf{C}_1^\top + \boldsymbol{M}_2$. Notice that p and σ_w^2 appear decoupled in (4.39). Therefore, we can first solve (4.39) for $\widehat{\sigma}_w^2$ for a fixed value of p to get an optimal solution. Then, substituting $\widehat{\sigma}_w^2$ into the problem and solving for $\widehat{p}$ we obtain the optimal estimate for p. The described procedure yields the following optimal estimates:

$$\widehat{p} = \frac{d_x\ \mathrm{tr}(\boldsymbol{M}_2^\top\, \boldsymbol{M}_2) - \mathrm{tr}^2(\boldsymbol{M}_2)}{d_x\, \mathrm{tr}(\boldsymbol{M}_1^\top\, \boldsymbol{M}_2) - \mathrm{tr}(\boldsymbol{M}_1)\, \mathrm{tr}(\boldsymbol{M}_2)}, \qquad \widehat{\sigma}_w^2 = \frac{\mathrm{tr}(\boldsymbol{M}_1) - (1/\widehat{p})\ \mathrm{tr}(\boldsymbol{M}_2)}{d_x}. \tag{4.40}$$

Using the estimate of randomization probability $\widehat{p}$ and $\mathbf{C}_1\mathbf{C}_0^{-1} = [\widehat{\overline{\mathbf{A}}}\ \widehat{\overline{\mathbf{B}}}]$, i.e., the estimate of average system transition parameters, the underlying system parameters could be recovered as $\widehat{\mathbf{A}} = (1/\widehat{p})\widehat{\overline{\mathbf{A}}} + (1 - 1/\widehat{p})\mathbf{I}_{d_x}$ and $\widehat{\mathbf{B}} = (1/\widehat{p})\widehat{\overline{\mathbf{B}}}$. To study the performance of the proposed system identification method, we consider a randomized LTI system with state transition matrix of $\mathbf{A}_1$ and a random $\mathbf{B}$ with $p = 0.5$

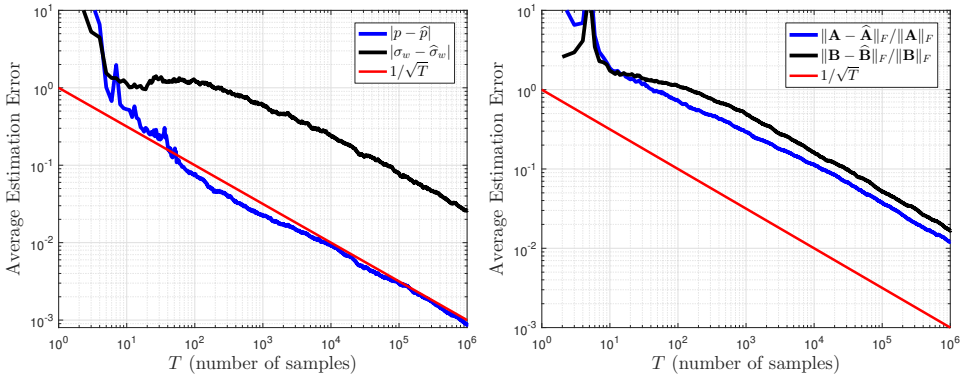

Figure 4.6: Average estimation error for the unknown system parameters of the stable randomized LTI system with state transition matrix of $\mathbf{A}_1$ and random $\mathbf{B}$ for 100 independent single trajectories

which guarantees the stability (verified by Lemma 4.4) and set $\sigma_w = 1$. We run 100 independent single trajectories and present the average rate of decay for the estimation errors of p, σ_w^2 in the first plot and $\mathbf{A}$ and $\mathbf{B}$ in the second plot of Figure 4.6. Notice that the estimation errors behave irregularly at the beginning where there are few samples, corresponding to burn-in period to converge to steady-state. On the other hand, Figure 4.6 show that, as predicted by MC-CLT, the estimation errors decay with $1/\sqrt{T}$ rate as we get more samples. This estimation error rate is the optimal behavior in linear regression problems with independent noise and covariates [106]. This depicts the consistency and efficiency of the proposed system identification method for randomized LTI systems.

4.3.5 Conclusion

In this study, we introduced a natural model of random asynchronous LTI systems that can be used to model various LDS with randomized and asynchronous updates. We numerically and analytically studied the mean-square stability of these systems and showed that the stability of random asynchronous systems is governed by the matrix representation of a positive linear map that controls the evolution of the state covariance matrix, rather than the state transition matrix. We proposed a system identification method for stable randomized LTI systems and observed that the method is consistent and efficient based on the empirical study.

In future work, we aim to derive the precise characterization of mean-square stability and extend the proposed system identification method in the general random asynchronous systems. We also plan to study the finite-time adaptive control and stabilization of random asynchronous LTI systems.

Chapter 5

LEARNING AND CONTROL IN PARTIALLY OBSERVABLE LINEAR DYNAMICAL SYSTEMS

In this chapter, we study the problem of learning and control of unknown partially observable linear dynamical systems, also known as measurement-feedback systems[1]. Unlike the dynamical systems considered in Chapters 3 and 4, the state of the dynamical system in this setting is not directly available to the decision-making agent. Instead, the learning agent observes a noisy (Gaussian or sub-Gaussian) linear measurement of the underlying hidden state. Due to this partial observability, learning the system dynamics with finite time guarantees brings substantial challenges, making it a long-lasting problem in adaptive control. In particular, when the latent states of a system are not fully observable, future observations are correlated with the past inputs and observations through the latent states. These correlations are even magnified when closed-loop controllers, i.e., those that naturally use past experiences to come up with control inputs, are deployed. Therefore, more sophisticated estimation methods that consider these complicated and unknown correlations are required for learning the dynamics.

In recent years, a series of works have studied this learning problem and presented a range of novel methods with finite-sample learning guarantees. These studies propose to employ i.i.d. Gaussian excitation as the control input, collect system outputs, and estimate the model parameters using the data collected through the Markov parameters of the system. The use of i.i.d. Gaussian noise as the open-loop control input is effective in mitigating the correlation between the inputs and output observations. For stable systems, these methods provide efficient ways to learn the model dynamics with confidence bounds of $\tilde{O}(1/\sqrt{T})$, after T times step of agent-environment interaction [213, 234, 245, 269]. Since then, this approach has become the standard practice in adaptive control, as it is challenging to learn the model dynamics when closed-loop controllers are used [223]. Closed-loop controllers that design inputs based on the history of inputs and observations result in highly correlated inputs with past process noise sequences, preventing reliable finite-time estimation of Markov parameters using the available system identification methods in the literature.

[1] This chapter is based on [138, 160–163].

These open-loop identification methods later have been deployed to propose explore-then-commit-based RL algorithms to minimize regret. In particular, these algorithms deploy i.i.d. Gaussian noise as the control input to learn the model parameters in the *explore* phase and then exploit these estimates during the *commit* phase to minimize regret. Among these works, Simchowitz et al. [245] propose using online convex optimization [17] during the commit phase. They show that their approach attains regret of $\tilde{O}(T^{2/3})$ when in the case of convex cost functions. Moreover, in the case of strongly convex cost functions, Mania et al. [191], Simchowitz et al. [245] show that exploiting the strong convexity allows guaranteeing regret of $\tilde{O}(\sqrt{T})$. These methods heavily rely on the lack of correlation achieved by using i.i.d. Gaussian noise as the open-loop control input to estimate the model. Therefore, they do not generalize to the adaptive settings, where the past observations are used to continuously improve both model estimates and the controllers. These challenges pose the following two open problems:

"*Can we learn the underlying model parameters in closed-loop setting with finite-sample guarantees?*"

"*Can we leverage such learning method to design adaptive control algorithms with improved regret guarantees in partially observable linear dynamical systems?*"

In the following sections, we give **affirmative** answers to both of these questions.

We introduce the first system identification method that allows estimating the model parameters with finite-time guarantees in both open and closed-loop settings. We exploit the classical predictive form representation of the system that goes back to Kalman [134] and reformulate each output as a linear function of previous control inputs and outputs with an additive i.i.d. Gaussian noise, named as the innovation process. This reformulation allows for addressing the limitations of the prior estimation methods in handling the correlations in inputs and outputs. We state a novel least squares problem to recover the Markov parameters of the system and propose a subspace identification method SysId, to obtain a balanced realization of the model parameters. We show that when the controllers persistently excite the system, i.e., the smallest singular value of the Gram matrix of the covariates scales linearly, the parameter estimation error of this novel closed-loop system identification method is $\tilde{O}(1/\sqrt{T})$ after T samples. This approach allows updating the model estimates while controlling the system with an adaptive controller.

Table 5.1: Comparison with prior works in learning and control of partially observable linear dynamical systems.

Work	Regret	Cost	Identification
Mania et al. [191]	$\sqrt{T}$	Strongly Convex	Open-Loop
Simchowitz et al. [245]	$\sqrt{T}$	Strongly Convex	Open-Loop
AdaptOn	$\sqrt{T}$	Strongly Convex	Closed-Loop (No PE)
AdaptOn	polylog(T)	Strongly Convex	Closed-Loop
Simchowitz et al. [245]	$T^{2/3}$	Convex	Open-Loop
LqgOpt & TSPO & AdaptOn	$T^{2/3}$	Convex	Closed-Loop (No PE)
LqgOpt & TSPO & AdaptOn	$\sqrt{T}$	Convex	Closed-Loop

Using this effective closed-loop system identification method, we focus on answering the second question and design adaptive control algorithms with improved regret guarantees. To this end, we mostly investigate the adaptive control of Linear Quadratic Gaussian (LQG) control systems and design three novel algorithmic frameworks: **LQG** control via **Opt**imism (LqgOpt), **T**hompson **S**ampling under **P**artial **O**bservability (TSPO), and **Adap**tive Control **On**line Learning (AdaptOn). These algorithms use three different methodologies to balance the exploration vs. exploitation trade-off and design adaptive controllers: optimism, Thompson Sampling, and online learning, respectively. We show the benefit of our closed-loop system identification method in achieving continuous model updates throughout the entire timeline of the adaptive control process, which yields significantly improved regret guarantees compared to prior work, see Table 5.1. In particular, due to the unique combination of the efficient closed-loop system identification method, strong convexity of the cost functions, convex policy parametrization, and the statistical efficiency of online gradient descent, we show that AdaptOn achieves the first logarithmic regret in learning and control of unknown partially observable linear dynamical systems. This surprising result sheds light on an interesting phenomenon that the learning and control in partially observable linear systems can be statistically easier than in the fully observable setting discussed in Section 3, see Remark 5.2. Furthermore, we also extend the guarantees of these algorithms to the more general setting of ARX systems with sub-Gaussian noise which are discussed in the subsequent sections.

The rest of the chapter is organized as follows: in the following, we review the prior work on finite-time guarantees of learning and control in partially observable linear dynamical systems. In Section 5.1, we introduce the relevant concepts and the prob-

lem setting. In Section 5.2, we review an open-loop system identification method adopted widely in the literature and explain its shortcomings. Section 5.3 introduces our novel closed-loop system identification method with finite-time guarantees and highlights how it overcomes the dependencies of covariates and noise terms by adopting a simple reparametrization of the system dynamics. In Section 5.4 and Section 5.5, we introduce LqgOpt and TSPO with their corresponding learning and control guarantees, respectively. Finally, in Section 5.6, we present AdaptOn and highlight the contributing pieces to its logarithmic regret. We end our discussion with Section 5.7 where we present several interesting future directions building on the results presented in this chapter.

Background and Motivation

Our study lies at the intersection of statistical learning, control, and reinforcement learning. Recently, there have been considerable efforts to give finite-time regret guarantees for adaptive control algorithms in linear dynamical systems (see Section 3.1 for the rich finite-time learning and control literature in fully observable linear dynamical systems). On the other hand, in partially observable linear dynamical systems the finite-time learning and control literature is more sparse due to the challenges of partial observability.

Learning with Finite-time Guarantees: The classical open or closed-loop system identification methods either consider the asymptotic behavior of the estimators or demonstrate the positive and negative empirical performances without theoretical guarantees [89, 188, 189, 224, 281, 282, 285]. Most of the prior work exploits the state-space form or the innovations form representation of the system, while only a handful consider the predictor form representation for system identification [58, 126], which will be the system dynamics representation used in our novel learning method. Interested readers can refer to [223] for an extensive overview.

In contrast to classical results in both of these problems that analyze the asymptotic performances, recently, there has been a flurry of studies that consider the finite-time learning guarantees. In finite-time system identification setting pioneered by Campi and Weyer [43, 44], currently, the main focus has been on obtaining the optimal learning rate of $1/\sqrt{T}$ after T samples. Using open-loop data collection to avoid correlations in the inputs and outputs, Oymak and Ozay [213], Sarkar et al. [234], Simchowitz et al. [244], Tsiamis and Pappas [269] suggest methods

that achieve this rate for stable dynamical systems. However, due to the difficulty in handling the correlations caused by the feedback controller, finite-time closed-loop system identification guarantees are scarce in the literature. To the best of our knowledge, only [174] considers finite-time closed-loop system identification. However, they analyze the output estimation error rather than explicitly recovering the model parameters as presented in our study.

Adaptive Control with Finite-time Guarantees: Besides the results presented in this chapter, there are only a couple of works that study the challenging topic of adaptive control with finite-time guarantees in partially observable linear dynamical systems [191, 245]. Among these, the CE-based method in [191] attains $\tilde{O}(\sqrt{T})$ regret if the quadratic cost in the LQG control systems is *strongly convex* ($Q, R \succ 0$). Similarly, under strongly convex cost condition, [245] show that $O(\sqrt{T})$ regret is attainable using online learning, while in the setting of convex cost they show sub-optimal regret of $O(T^{2/3})$. In our results for LqgOpt, TSPO, and AdaptOn we match these results in the most general setting, see Corollaries 5.5.1, 5.8.1, and 5.10.3, respectively. However, when a PE condition is satisfied for the underlying system, then we significantly advance the known regret results in the literature due to our novel closed-loop system identification method. In particular, we show that for partially observable systems with convex cost, our algorithms can attain $O(\sqrt{T})$ regret rate, see Theorem 5.5, 5.8, and Corollary 5.10.3, respectively. Perhaps surprisingly, for the setting of strongly convex cost functions, we show that AdaptOn achieves logarithmic regret in Theorem 5.10. Furthermore, we extend these results to the more general setting of ARX systems with sub-Gaussian noise and relax certain restrictive assumptions such as controllability and observability of the system dynamics to the stabilizability and detectability, see Sections 5.4.6 and 5.6.5.

5.1 Preliminaries

5.1.1 Notation

We denote the Euclidean norm of a vector x as $\|x\|_2$. For a given matrix A, $\|A\|_2$ denotes the spectral norm, $\|A\|_F$ denotes the Frobenius norm while $A^\top$ is its transpose, $A^\dagger$ is the Moore-Penrose inverse, and $\mathrm{Tr}(A)$ gives the trace of matrix A. The j-th singular value of a rank-n matrix A is denoted by $\sigma_j(A)$, where $\sigma_{\max}(A) := \sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_{\min}(A) := \sigma_n(A) > 0$. I is the identity matrix with the appropriate dimension. In the following, $\mathcal{N}(\mu, \Sigma)$ denotes a multivariate normal distribution with mean vector μ and covariance matrix Σ.

5.1.2 Partially Observable Linear Dynamical Systems

In this chapter, we will first study the canonical measurement-feedback linear control systems known as Linear Quadratic Gaussian (LQG) control systems. As their name suggests, these systems have linear dynamics, quadratic control costs, and Gaussian noise disturbances. These systems will be our starting point, and we will consider various generalizations such as without the exact knowledge of the noise covariance, sub-Gaussian noise, and general (strongly) convex cost functions in our results.

In the LQG control systems, we have $\Theta = (A, B, C)$ with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times p}$, $C \in \mathbb{R}^{m\times n}$ as the model parameters of a partially observable linear time-invariant dynamical system in the state-space form:

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + w_t \\ y_t &= Cx_t + z_t, \end{aligned} \tag{5.1}$$

where $x_t \in \mathbb{R}^n$ is the (latent) state of the system, $u_t \in \mathbb{R}^p$ is the control input, the observation $y_t \in \mathbb{R}^m$ is the output of the system, $w_t \sim \mathcal{N}(0, W)$, and $z_t \sim \mathcal{N}(0, Z)$ are i.i.d. process noise and measurement noise, respectively. Note that for simplicity of presentation, in LQG control systems, we will consider isotropic Gaussian process and measurement noise, i.e., $W = \sigma_w^2 I$ and $Z = \sigma_z^2 I$. At each time step t, the system is at state x_t and the agent observes y_t, i.e., imperfect state information. Then, the agent applies a control input u_t and the system evolves to x_{t+1} at time step $t + 1$. We will assume that the underlying system is controllable and observable.

Definition 5.1. *A linear system $\Theta = (A, B, C)$ is (A, B) controllable if the controllability matrix,*

$$\mathbf{C}(A, B, n) = [B \;\; AB \;\; A^2B \ldots A^{n-1}B]$$

has full row rank. For all $H \geq n$, $\mathbf{C}(A, B, H)$ defines the extended (A, B) controllability matrix. Similarly, a linear system $\Theta = (A, B, C)$ is A, C observable if the observability matrix,

$$\mathbf{O}(A, C, n) = [C^\top \;\; (CA)^\top \;\; (CA^2)^\top \ldots (CA^{n-1})^\top]^\top$$

has full column rank. For all $H \geq n$, $\mathbf{O}(A, C, H)$ defines the extended (A, C) observability matrix.

By assuming controllability and observability of the underlying system with state dimension n, we implicitly assume the order of the underlying system is also n, i.e.,

the system is in its minimal representation. We adopt this assumption for ease of presentation. There are several efficient algorithms that find the order of an unknown linear dynamical system [234]. Using these techniques, we can lift the assumption on the order of the system without jeopardizing any performance guarantees.

Notice that unlike the dynamical systems studied in Chapters 3 and 4, in this system the agent does not observe the state, thus it is needed to be estimated. For this setting, in his seminal work, Kalman derived a closed-form expression for $\hat{x}_{t|t,\Theta}$, the minimum mean squared error (MMSE) estimate of the underlying state x_t using the past information of control inputs and observations, and the model parameters Θ, where $\hat{x}_{0|-1,\Theta} = 0$. This formulation is denoted as the Kalman filter and is efficiently obtained via

$$\hat{x}_{t|t,\Theta} = (I - LC)\,\hat{x}_{t|t-1,\Theta} + Ly_t, \tag{5.2}$$

$$\hat{x}_{t|t-1,\Theta} = (A\hat{x}_{t-1|t-1,\Theta} + Bu_{t-1}), \tag{5.3}$$

$$L = \Sigma C^\top \left(C\Sigma C^\top + \sigma_z^2 I\right)^{-1}, \tag{5.4}$$

where Σ is the unique positive semidefinite solution to the following Discrete Algebraic Riccati Equation (DARE):

$$\Sigma = A\Sigma A^\top - A\Sigma C^\top \left(C\Sigma C^\top + \sigma_z^2 I\right)^{-1} C\Sigma A^\top + \sigma_w^2 I. \tag{5.5}$$

Σ can be interpreted as the steady state error covariance matrix of state estimation under Θ. There are various equivalent characterizations of the dynamics of the discrete-time linear time-invariant system Θ besides the state-space form given in (5.1) [132, 160, 270]. Note that these representations all have the same second-order statistics. One of the most common forms is the innovations form[2] of the system characterized as

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + Fe_t \\ y_t &= Cx_t + e_t, \end{aligned} \tag{5.6}$$

where $F = AL$ is the Kalman gain in the observer form and e_t is the zero mean white innovation process. In this equivalent representation of the system, the state x_t can be seen as the estimate of the state in the state space representation, which is the expression stated in (5.3), i.e., the MMSE estimate of state x_t given

[2]For simplicity, all of the system representations are presented for the steady-state of the system. Note that the system converges to the steady state exponentially fast [211].

$(y_{t-1}, \ldots, y_0, u_{t-1}, \ldots, u_0)$. In the steady state, $e_t \sim \mathcal{N}\left(0, C\Sigma C^\top + \sigma_z^2 I\right)$. Using the relationship between e_t and y_t, we obtain the following characterization of the system Θ, known as the predictor form of the system,

$$\begin{aligned} x_{t+1} &= \bar{A}x_t + Bu_t + Fy_t \\ y_t &= Cx_t + e_t, \end{aligned} \tag{5.7}$$

where $\bar{A} = A - FC$ and $F = AL$. Notice that at steady state, the predictor form allows the current output y_t to be described by the history of inputs and outputs with an i.i.d. Gaussian disturbance $e_t \sim \mathcal{N}\left(0, C\Sigma C^\top + \sigma_z^2 I\right)$. In our results, we exploit these fundamental properties to estimate the underlying system, even with feedback control.

The predictor form dynamics given in (5.7) belong to a larger class of dynamical systems named Autoregressive Exogenous (ARX) systems. ARX systems are central dynamical systems in time-series modeling due to input-output form representation as given in (5.7). Due to their ability to approximate linear systems in a parametric model structure, they have been crucial in many areas including chemical engineering, power engineering, medicine, economics, and neuroscience [23, 42, 87, 118, 209]. These models provide a *general* representation of LDS with *arbitrary* stochastic disturbances. In our study, besides LQG control systems in predictor form (5.7) with e_t being the innovation process, we will consider the general setting of dynamical systems of the form (5.7) with sub-Gaussian e_t and arbitrary $\bar{A}$ and F.

Definition 5.2. *A matrix $M \in \mathbb{R}^{n\times n}$ (κ, γ)-stable for $\kappa \geq 0$ and $0 < \gamma \leq 1$ if there exists a similarity transformation $M = S\Lambda S^{-1}$ where $\|S\|\|S^{-1}\| \leq \kappa$ and $\|\Lambda\| \leq 1 - \gamma$.*

We will consider (κ_1, γ_1) open-loop stable LQG control systems. From the definition above, this means that for all k, $\|A^k\| \leq \kappa_1(1 - \gamma_1)^k$. Notice that Definition 5.2 is the stability corresponding to the stabilizability definition given in Definition 3.2 in Chapter 3. The stability of A is required to have a simply bounded state in the analysis and is not a fundamental requirement in the predictor form nor for the ARX systems. In particular, in the predictor form of LQG, $\bar{A}$ is stable due to observability assumption and in the ARX systems, we will explicitly assume the stability of $\bar{A}$ which captures an extensive number of systems including all detectable partially observable linear dynamical systems [132]. Thus, for the LQG control systems, one can show exponential in dimension bound on state x_t similar to the analysis provided in Chapter 3. We leave this analysis for future work.

To summarize, we assume that the underlying system lives in the following set.

Assumption 5.1. *The unknown system $\Theta = (A, B, C)$ is a member of a set $\mathcal{S}$, such that,*

$$\mathcal{S} \subseteq \left\{ \Theta' = (A', B', C', F') \;\middle|\; \begin{array}{l} A' \text{ is } (\kappa_1, \gamma_1)\text{-stable,} \\ (A', B') \text{ is controllable,} \\ (A', C') \text{ is observable,} \\ (A', F') \text{ is controllable,} \\ \max(\|A\|, \|B\|, \|C\|, \|F\|) \leq \psi. \end{array} \right\},$$

where $\psi > 0$, $\kappa_1 > 0$, and $\gamma_1 \in (0, 1]$. In particular, we assume that there exist constants $\kappa_2, \kappa_3 > 0$ and $\gamma_2, \gamma_3 \in (0, 1]$ such that the systems are (κ_2, γ_2)-stabilizable as defined in Definition 3.2 and $\bar{A}' := A' - F'C'$ is (κ_3, γ_3)-stable for all $\Theta' \in \mathcal{S}$.

Note that (κ_2, γ_2)-stabilizability follows directly from the controllability of the system and the closed-loop stability of $A' - F'C'$ also follows from the observability of the system, in other words, it can be considered as the stabilizability property with respect to the filtering problem.

The behavior of an LQG control system or an ARX system is uniquely governed by its Markov parameters, i.e., impulse response.

Definition 5.3 (Markov Parameters)**.** *The set of matrices that maps the previous inputs to the output is called input-to-output Markov parameters and the ones that map the previous outputs to the output are denoted as output-to-output Markov parameters of the system Θ. In particular, for the dynamics in* (5.1)*, the set of Markov parameters is defined as $G^i_{u\to y} = CA^{i-1}B, \forall i \geq 1$. For the predictor form or ARX systems given in* (5.7)*, the matrices that map inputs and outputs to the output are the elements of the Markov operator, $\mathbf{G} = \{G^i_{u\to y}, G^i_{y\to y}\}_{i\geq 1}$ where $\forall i \geq 1$, $G^i_{u\to y} = C\bar{A}^{i-1}B$ and $G^i_{y\to y} = C\bar{A}^{i-1}F$ which are unique.*

In learning the system dynamics, we will aim to learn the Markov parameters of the system since they are uniquely identifiable.

5.1.3 Control Problem Under Quadratic Costs

To define the control problem in the partially observable linear dynamics introduced in the previous section, we will consider two different cost functions. The first one is the canonical quadratic cost on the inputs and outputs of the LQG control system

setting. In this setting, at time t, when the agent applies a control input it receives a cost of

$$c_t = y_t^\top Q y_t + u_t^\top R u_t,$$

where Q and R are positive semidefinite (psd) and positive definite (pd) matrices, respectively. We denote this cost as the convex quadratic cost since Q is psd. The goal of the controlling agent is to reduce the cumulative cost $\sum_{t=0}^{T} c_t$ by deploying control actions after $T \geq 0$ number of interactions with the environment. This can be achieved by finding the best *control policy* that minimizes the average expected cost subject to the dynamical constraints in (5.1) as

$$J_*(\Theta) = \lim_{T\to\infty} \min_{u=[u_1,\ldots,u_T]} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} y_t^\top Q y_t + u_t^\top R u_t\right]. \tag{5.8}$$

This problem is known as the LQG control. Here $J_*(\Theta)$ is the optimal average expected cost of the system Θ. The optimal solution to the LQG control problem is obtained by estimating the latent state of the system and deploying the optimal controller synthesis described in Chapter 3. This is known as the separation principle in control theory [19]. In particular, the optimal controller is given as the linear feedback policy

$$u_t = -K\hat{x}_{t|t,\Theta}, \tag{5.9}$$

where K is the optimal feedback gain matrix,

$$K = \left(R + B^\top P B\right)^{-1} B^\top P A,$$

as defined in Chapter 3 and P is the unique positive semidefinite solution to the following discrete-time algebraic Riccati equation (DARE):

$$P = A^\top P A + C^\top Q C - A^\top P B \left(R + B^\top P B\right)^{-1} B^\top P A, \tag{5.10}$$

and $\hat{x}_{t|t,\Theta}$ is the MMSE estimate given in (5.2). The optimal average expected cost of LQG control system that satisfies Assumption 5.1 takes a finite value and can be computed as

$$J_*(\Theta) = \mathrm{Tr}\left(\left(Q + L^\top P L - L^\top C^\top Q C L\right)\left(C\Sigma C^\top + \sigma_z^2 I\right)\right). \tag{5.11}$$

The majority of this chapter studies adaptive control under quadratic cost for LQG control systems and ARX systems. However, in Section 5.6, we consider general strongly convex cost functions which can be adversarially chosen and provide new sets of results using online gradient descent for the aforementioned settings.

5.1.4 Regret

Due to a lack of knowledge of model parameters, i.e., system dynamics, the learning and control agent cannot design the discussed optimal control law. Therefore, the agent needs to learn them through interaction with the system with the aim of minimizing the cumulative costs $\sum_{t=1}^{T} c_t$ after T time steps. We measure the performance of the agent using regret, *i.e.,* the difference between the agent's cost and the optimal expected cost:

$$\textsc{Regret}(T) = \sum_{t=0}^{T} (c_t - J_*(\Theta)). \tag{5.12}$$

In the majority of this chapter, we consider this regret metric for adaptive control performance in LQG control and ARX systems. However, in Section 5.6 we consider the regret with respect to the best controller in the hindsight from a given class of controllers and show improved regret guarantees.

5.2 Open-Loop System Identification

In this section, we study the open-loop system identification methods that are adopted in the literature. In order to minimize the regret given in (5.12), the learning agent needs to efficiently *explore* the environment to learn the system dynamics, and *exploit* the gathered experiences to minimize overall cost [171]. However, since the underlying states of the systems are not fully observable, learning the system dynamics with finite time guarantees brings substantial challenges, making it a long-lasting problem in adaptive control. In particular, when the latent states of a system are not fully observable, future observations are correlated with the past inputs and observations through the latent states. These correlations are even magnified when closed-loop controllers, those that naturally use past experiences to come up with control inputs, are deployed. Therefore, more sophisticated estimation methods that consider these complicated and unknown correlations are required for learning the dynamics.

An Open-loop System Identification Method

In recent years, a series of works have studied this learning problem and presented a range of novel methods with finite-sample learning guarantees. These studies propose to employ i.i.d. Gaussian excitation as the control input, i.e., open-loop control, collect system outputs, and estimate the model parameters using the data collected. These methods study the system identification problem using the state-space representation (5.1) and aim to recover the input-to-output Markov parameters

$G^i_{u\to y} = CA^{i-1}B$ introduced in Definition 5.3. The use of i.i.d. Gaussian noise as the open-loop control input (not using past experiences) mitigates the correlation between the inputs and the output observations. For stable systems, these methods provide efficient ways to learn the model dynamics with confidence bounds of $\tilde{O}(1/\sqrt{T})$, after T times step of agent-environment interaction [166, 213, 234, 245, 269]. Here $\tilde{O}(\cdot)$ denotes the order up to logarithmic factors. Deploying i.i.d. Gaussian noise for a long period of time to estimate the model parameters has been the common practice in adaptive control since incorporating a closed-loop controller introduces significant challenges to learning the model dynamics [223]. In this section, we review one of such open-loop system identification methods and discuss the reason why methods that use the state-space representation of the system (5.1) cannot provide reliable estimates in closed-loop estimation problems.

Using the state-space representation in (5.1), for any positive integer H, one can rewrite the output at time t as follows,

$$y_t = \sum\nolimits_{i=1}^{H} CA^{i-1}Bu_{t-i} + CA^H x_{t-H} + z_t + \sum\nolimits_{i=0}^{H-1} CA^i w_{t-i-1}. \tag{5.13}$$

Recalling Definition 5.3, for $\kappa_{\mathbf{G}} \geq 1$, let the Markov operator of Θ be bounded, i.e., $\sum_{i\geq 0} \|G^i_{u\to y}\| \leq \kappa_{\mathbf{G}}$. Due to Assumption 5.1, i.e., the stability of A, the second term in (5.13) decays exponentially, and for large enough H it becomes negligible. Therefore, we obtain the following for the output at time t,

$$y_t \approx \sum\nolimits_{i=1}^{H} G^i_{u\to y}u_{t-i} + z_t + \sum\nolimits_{i=0}^{H-1} CA^i w_{t-i-1}. \tag{5.14}$$

From this formulation, a least squares estimation problem can be formulated using outputs as the dependent variable and the concatenation of H input sequences $\bar{u}_t = [u_{t-1}, \ldots, u_{t-H}]$ as the regressor to recover the Markov parameters of the system:

$$\widehat{\mathbf{G}}_{u\to y} = [\widehat{G}^1_{u\to y}, \ldots, \widehat{G}^H_{u\to y}] = \operatorname*{argmin}_X \sum\nolimits_{t=H}^{T} \|y_t - X\bar{u}_t\|_2^2. \tag{5.15}$$

Prior finite-time system identification algorithms propose using i.i.d. zero-mean Gaussian noise for the input, to make sure that the two noise terms in (5.14) are not correlated with the inputs. In particular, exciting the system with i.i.d. $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ for $1 \leq t \leq T_{exp}$ provides a lack of correlation between the regressor and the noise components in (5.14) and allows solving (5.15) in closed-form with finite-time estimation error guarantees for the unknown input-to-output Markov parameters [161, 166, 213, 234, 244]. Note that besides lack of correlation, the i.i.d Gaussian control inputs persistently excite the system allows consistent estimation

of the Markov parameters. Interested readers can find the general analysis in [213] where Oymak and Ozay, show that using i.i.d. Gaussian control inputs allows estimating the Markov parameters with the optimal rate of $\tilde{O}(1/\sqrt{T_{exp}})$, i.e.,

$$\|\widehat{\mathbf{G}}_{u\to y} - \mathbf{G}_{u\to y}\| \leq \frac{c}{\sigma_u \sqrt{T_{exp}}} \tag{5.16}$$

for some problem-dependent constant c after large enough T_{exp} time steps. This rate is the same error rate one would get from solving a linear regression problem with independent noise and independent covariates [106].

Even though Markov parameters uniquely determine the underlying system, to design the controller for the underlying system as described in Section 5.1.3, one needs to find a balanced realization of Θ from $\widehat{\mathbf{G}}_{u\to y}$. To achieve this, the well-known subspace method Ho-Kalman algorithm is the primary choice [112]. The Ho-Kalman algorithm is given in Algorithm 9. It takes the Markov parameter matrix estimate $\widehat{\mathbf{G}}_{u\to y}$, H, the systems order n, and dimensions d_1, d_2, as the input and computes an order n system $\hat{\Theta} = (\hat{A}, \hat{B}, \hat{C})$. It is worth restating that the dimension of the latent state, n, is the order of the system for observable and controllable dynamics. With the assumption that $H \geq 2n + 1$, we pick $d_1 \geq n$ and $d_2 \geq n$ such $d_1 + d_2 + 1 = H$. This guarantees that the system identification problem is well-conditioned.

Algorithm 9 Ho-Kalman Algorithm

1: **Input:** $\widehat{\mathbf{G}}_{u\to y}$, H, system order n, d_1, d_2 such that $d_1 + d_2 + 1 = H$
2: Form the Hankel Matrix $\hat{\mathcal{H}} \in \mathbb{R}^{md_1 \times p(d_2+1)}$ from $\widehat{\mathbf{G}}_{u\to y}$
3: Set $\hat{\mathcal{H}}^- \in \mathbb{R}^{md_1 \times pd_2}$ as the first pd_2 columns of $\hat{\mathcal{H}}$
4: Using SVD obtain $\hat{\mathcal{N}} \in \mathbb{R}^{md_1 \times pd_2}$, the rank-$n$ approximation of $\hat{\mathcal{H}}^-$
5: Obtain $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V} = \text{SVD}(\hat{\mathcal{N}})$
6: Construct $\hat{\mathbf{O}} = \mathbf{U}\mathbf{\Sigma}^{1/2} \in \mathbb{R}^{md_1 \times n}$
7: Construct $\hat{\mathbf{C}} = \mathbf{\Sigma}^{1/2}\mathbf{V} \in \mathbb{R}^{n \times pd_2}$
8: Obtain $\hat{C} \in \mathbb{R}^{m \times n}$, the first m rows of $\hat{\mathbf{O}}$
9: Obtain $\hat{B} \in \mathbb{R}^{n \times p}$, the first p columns of $\hat{\mathbf{C}}$
10: Obtain $\hat{\mathcal{H}}^+ \in \mathbb{R}^{md_1 \times pd_2}$, the last pd_2 columns of $\hat{\mathcal{H}}$
11: Obtain $\hat{A} = \hat{\mathbf{O}}^\dagger \hat{\mathcal{H}}^+ \hat{\mathbf{C}}^\dagger \in \mathbb{R}^{n \times n}$

Since only the order n input-output response of the system is uniquely identifiable [188], the system parameters Θ (even with the correct Markov parameters matrix $\mathbf{G}_{u\to y}$) are recovered up to similarity transformation. More generally, for any invertible $\mathbf{T} \in \mathbb{R}^{n\times n}$, the system $A' = \mathbf{T}^{-1}A\mathbf{T}, B' = \mathbf{T}^{-1}B, C' = C\mathbf{T}$ gives the

same Markov parameters matrix $\mathbf{G}_{u\to y}$, equivalently, the same input-output impulse response.

For $H \geq 2n+1$, using $[\widehat{G}^1_{u\to y}, \ldots, \widehat{G}^H_{u\to y}] \in \mathbb{R}^{m\times Hp}$, the Ho-Kalman algorithm constructs a $(n \times n+1)$ block Hankel matrix $\hat{\mathcal{H}} \in \mathbb{R}^{nm\times(n+1)p}$ such that (i,j)th block of Hankel matrix is $\hat{G}^{i+j-1}_{u\to y}$. It is worth noting that if the input to the algorithm was $\mathbf{G}_{u\to y}$ then the corresponding Hankel matrix, $\mathcal{H}$ is rank n, more importantly,

$$\mathcal{H} = [C^\top \ (CA)^\top \ldots (CA^{n-1})^\top]^\top [B \ AB \ldots A^nB] = \mathbf{O}[\mathbf{C} \ \ A^nB] = \mathbf{O}[B \ \ A\mathbf{C}],$$

where $\mathbf{O}$ and $\mathbf{C}$ are observability and controllability matrices respectively. Essentially, the Ho-Kalman algorithm estimates these matrices using $\widehat{\mathbf{G}}_{u\to y}$. In order to estimate $\mathbf{O}$ and $\mathbf{C}$, the algorithm constructs $\hat{\mathcal{H}}^-$, the first np columns of $\hat{\mathcal{H}}$ and calculates $\hat{\mathcal{N}}$, the best rank-n approximation of $\hat{\mathcal{H}}^-$. Therefore, the singular value decomposition of $\hat{\mathcal{N}}$ provides us with the estimates of $\mathbf{O}, \mathbf{C}$, i.e., $\hat{\mathcal{N}} = \mathbf{U}\mathbf{\Sigma}^{1/2} \ \mathbf{\Sigma}^{1/2}\mathbf{V} = \hat{\mathbf{O}}\hat{\mathbf{C}}$. From these estimates, the algorithm recovers $\hat{B}$ as the first $n \times p$ block of $\hat{\mathbf{C}}$, $\hat{C}$ as the first $m \times n$ block of $\hat{\mathbf{O}}$, and $\hat{A}$ as $\hat{\mathbf{O}}^\dagger \hat{\mathcal{H}}^+ \hat{\mathbf{C}}^\dagger$ where $\hat{\mathcal{H}}^+$ is the submatrix of $\hat{\mathcal{H}}$, obtained by discarding the left-most $nm \times p$ block.

Note that if we feed $\mathbf{G}_{u\to y}$ to the Ho-Kalman algorithm, the $\mathcal{H}^-$ is the first np columns of $\mathcal{H}$, it is rank-n, and $\mathcal{N}=\mathcal{H}^-$. Using the outputs of the Ho-Kalman algorithm, *i.e.,* $(\hat{A}, \hat{B}, \hat{C})$, we can construct confidence sets centered around these outputs that contain a similarity transformation of the system parameters $\Theta = (A, B, C)$ with high probability. Theorem 5.1 states the construction of confidence sets and it is a slight modification of Corollary 5.4 of Oymak and Ozay [213].

Theorem 5.1 (Confidence Set Construction)**.** *Suppose $\mathcal{H}$ is the rank-n Hankel matrix obtained from $\mathbf{G}_{u\to y}$. Let $\bar{A}, \bar{B}, \bar{C}$ be the system parameters that Ho-Kalman algorithm provides for $\mathbf{G}_{u\to y}$. Define the rank-n matrix $\mathcal{N}$ such that it is the submatrix of $\mathcal{H}$ obtained by discarding the last block column of $\mathcal{H}$. Suppose $\sigma_n(\mathcal{N}) > 0$ and $\|\hat{\mathcal{N}} - \mathcal{N}\| \leq \frac{\sigma_n(\mathcal{N})}{2}$. Then, there exists a unitary matrix $\mathbf{T} \in \mathbb{R}^{n\times n}$ such that, $\bar{\Theta} = (\bar{A}, \bar{B}, \bar{C}) \in (C_A \times C_B \times C_C)$ for*

$$C_A = \left\{ A' \in \mathbb{R}^{n\times n} : \|\hat{A} - \mathbf{T}^\top A' \mathbf{T}\| \leq \left(\frac{31n\|\mathcal{H}\|}{\sigma_n^2(\mathcal{H})} + \frac{13n}{2\sigma_n(\mathcal{H})} \right) \|\widehat{\mathbf{G}}_{u\to y} - \mathbf{G}_{u\to y}\| \right\}$$

$$C_B = \left\{ B' \in \mathbb{R}^{n\times p} : \|\hat{B} - \mathbf{T}^\top B'\| \leq \frac{7n}{\sqrt{\sigma_n(\mathcal{H})}} \|\widehat{\mathbf{G}}_{u\to y} - \mathbf{G}_{u\to y}\| \right\}$$

$$C_C = \left\{ C' \in \mathbb{R}^{m\times n} : \|\hat{C} - C'\mathbf{T}\| \leq \frac{7n}{\sqrt{\sigma_n(\mathcal{H})}} \|\widehat{\mathbf{G}}_{u\to y} - \mathbf{G}_{u\to y}\| \right\},$$

where $\hat{A}, \hat{B}, \hat{C}$ *obtained from the Ho-Kalman algorithm using the least squares estimate of the Markov parameter matrix* $\widehat{\mathbf{G}}_{u\to y}$.

Proof. The proof is similar to the proof of Theorem 4.3 in [213]. The difference in the presentation arises due to providing a different characterization of the dependence on $\|\mathcal{N} - \hat{\mathcal{N}}\|$ and centering the confidence ball over the estimations rather than the output of Ho-Kalman algorithm with the input of G. In Oymak and Ozay [213], from the inequality

$$\|\bar{B} - \mathbf{T}^\top \hat{B}\|_F^2 \leq \frac{2n\|\mathcal{N} - \hat{\mathcal{N}}\|^2}{(\sqrt{2}-1)\left(\sigma_n(\mathcal{N}) - \|\mathcal{N} - \hat{\mathcal{N}}\|\right)},$$

the authors use the assumption $\|\mathcal{N} - \hat{\mathcal{N}}\| \leq \frac{\sigma_n(\mathcal{N})}{2}$ to cancel out numerator and denominator. In this presentation, we define T_N such that for large enough exploration time T_{exp} such that $T_{exp} \geq T_N$, we have $\|\mathcal{N} - \hat{\mathcal{N}}\| \leq \frac{\sigma_n(\mathcal{N})}{2}$ with high probability. See [166] for the precise expression of T_N. Note that T_N depends on $\sigma_n(H)$, due to the fact that singular values of submatrices by column partitioning are interlaced, i.e., $\sigma_n(\mathcal{N}) = \sigma_n(\mathcal{H}^-) \geq \sigma_n(\mathcal{H})$. Then, we redefine the denominator based on $\sigma_n(\mathcal{N})$ and again use the fact $\sigma_n(\mathcal{N}) = \sigma_n(\mathcal{H}^-) \geq \sigma_n(\mathcal{H})$. Following the proof steps provided in Oymak and Ozay [213] and combining with the fact that $\|\mathcal{N} - \hat{\mathcal{N}}\| \leq 2\left\|\mathcal{H}^- - \hat{\mathcal{H}}^-\right\| \leq 2\sqrt{\min\{d_1, d_2\}}\|\widehat{\mathbf{G}}_{u\to y} - \mathbf{G}_{u\to y}\|$ (see Lemma B.1 of [213]), we obtain the presented theorem. □

Combining Theorem 5.1 with (5.16) shows that using the open-loop system identification method, a balanced realization of Θ could be recovered with the optimal estimation rate with high probability. However, when a controller designs the inputs based on the history of inputs and observations, the inputs become highly correlated with the past process noise sequences, $\{w_i\}_{i=0}^{t-1}$. This correlation prevents the consistent and reliable estimation of Markov parameters using (5.15). Therefore, these prior open-loop estimation methods do not generalize to the systems that adaptive controllers generate the inputs for estimation, i.e., closed-loop estimation. For this very reason, the open-loop system identification techniques have been only deployed to propose explore-then-commit-based adaptive control algorithms to minimize regret as discussed at the beginning of this chapter. In the following section, we provide a closed-loop system identification algorithm that alleviates the correlations in the covariates and the noise sequences by considering the predictor form of the system dynamics (5.7) rather than the state space form (5.1).

5.3 A Novel Closed-Loop System Identification Method

In this section, we introduce the first system identification method that allows estimating the model parameters with finite-time guarantees in both open and closed-loop settings. Without loss of generality, since the Kalman filter converges exponentially fast to the steady-state gain in observer form, we assume that $x_0 \sim \mathcal{N}(0, \Sigma)$, i.e., the system starts at the steady state. This consideration eases the presentation of our method. To analyze any arbitrary and almost surely finite initialization we refer the interested reader to Appendix G of [162].

For the LQG control systems or ARX systems, using the predictor form representation (5.7), for a positive integer H, the output at time t can be rewritten as follows,

$$y_t = \sum_{k=0}^{H-1} C\bar{A}^k \left(Fy_{t-k-1} + Bu_{t-k-1}\right) + e_t + C\bar{A}^H x_{t-H}. \tag{5.17}$$

Using the open- or closed-loop generated input-output sequences up to time τ, $\{y_t, u_t\}_{t=1}^{\tau}$, we construct subsequences of H input-output pairs for $H \le t \le \tau$,

$$\phi_t = \left[y_{t-1}^\top, \ldots, y_{t-H}^\top, u_{t-1}^\top, \ldots, u_{t-H}^\top\right]^\top \in \mathbb{R}^{(m+p)H}.$$

Recall the predictor form Markov parameters defined in Definition 5.3, i.e., input-to-output Markov parameters $G^i_{u\to y} = C\bar{A}^{i-1}B$ and output-to-output Markov parameters $G^i_{y\to y} = C\bar{A}^{i-1}F$. Then, the output of the system, y_t can be represented using ϕ_t as:

$$y_t = \mathcal{G}_{\mathbf{yu}}\phi_t + e_t + C\bar{A}^H x_{t-H} \tag{5.18}$$

for

$$\mathcal{G}_{\mathbf{yu}} = [G^1_{y\to y}, \ldots, G^H_{y\to y}, G^1_{u\to y}, \ldots, G^H_{u\to y}]. \tag{5.19}$$

Notice that $\bar{A}$ is stable due to (A, F)-controllability and observability of Θ for LQG systems (in fact it only requires the weaker conditions of stabilizability and detectability), and in the ARX systems, we explicitly assume the stability of $\bar{A}$. Therefore, with a similar argument used in (5.13), for $H = O(\log(T))$, the last term in (5.17) is negligible. This yields into a linear model of the dependent variable y_t and the regressor ϕ_t with additive i.i.d. Gaussian noise e_t:

$$y_t \approx \mathcal{G}_{\mathbf{yu}}\phi_t + e_t. \tag{5.20}$$

For this model, we achieve consistent and reliable estimates by solving the following regularized least squares problem,

$$\widehat{\mathcal{G}}_{\mathbf{yu}} = \operatorname*{argmin}_X \lambda\|X\|_F^2 + \sum\nolimits_{t=H}^{\tau} \|y_t - X\phi_t\|_2^2. \tag{5.21}$$

Algorithm 10 SysId

Input: $\widehat{\mathcal{G}}_{\mathbf{yu}}$, H, system order n, d_1, d_2 such that $d_1 + d_2 + 1 = H$
Form two $d_1 \times (d_2+1)$ Hankel matrices $\hat{\mathcal{H}}_{\mathbf{y}\to\mathbf{y}}$ and $\hat{\mathcal{H}}_{\mathbf{u}\to\mathbf{y}}$ from $\widehat{\mathcal{G}}_{\mathbf{yu}}$ and construct $\hat{\mathcal{H}} = \begin{bmatrix}\hat{\mathcal{H}}_{\mathbf{y}\to\mathbf{y}}, & \hat{\mathcal{H}}_{\mathbf{u}\to\mathbf{y}}\end{bmatrix} \in \mathbb{R}^{md_1\times(m+p)(d_2+1)}$
Obtain $\hat{\mathcal{H}}^-$ by discarding (d_2+1)th and $(2d_2+2)$th block columns of $\hat{\mathcal{H}}$
Using SVD obtain $\hat{\mathcal{N}} \in \mathbb{R}^{md_1\times(m+p)d_2}$, the best rank-$n$ approximation of $\hat{\mathcal{H}}^-$
Obtain $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V} = \text{SVD}(\hat{\mathcal{N}})$
Construct $\hat{\mathbf{O}}(\bar{A}, C, d_1) = \mathbf{U}\mathbf{\Sigma}^{1/2} \in \mathbb{R}^{md_1\times n}$
Construct $[\hat{\mathbf{C}}(\bar{A}, F, d_2+1), \quad \hat{\mathbf{C}}(\bar{A}, B, d_2+1)] = \mathbf{\Sigma}^{1/2}\mathbf{V} \in \mathbb{R}^{n\times(m+p)d_2}$
Obtain $\hat{C} \in \mathbb{R}^{m\times n}$, the first m rows of $\hat{\mathbf{O}}(\bar{A}, C, d_1)$
Obtain $\hat{B} \in \mathbb{R}^{n\times p}$, the first p columns of $\hat{\mathbf{C}}(\bar{A}, B, d_2+1)$
Obtain $\hat{F} \in \mathbb{R}^{n\times m}$, the first m columns of $\hat{\mathbf{C}}(\bar{A}, F, d_2+1)$
Obtain $\hat{\mathcal{H}}^+$ by discarding 1st and (d_2+2)th block columns of $\hat{\mathcal{H}}$
Obtain $\hat{\bar{A}} = \hat{\mathbf{O}}^\dagger(\bar{A}, C, d_1) \;\; \hat{\mathcal{H}}^+ \;\; [\hat{\mathbf{C}}(\bar{A}, F, d_2+1), \quad \hat{\mathbf{C}}(\bar{A}, B, d_2+1)]^\dagger$
Obtain $\hat{A} = \hat{\bar{A}} + \hat{F}\hat{C}$
Obtain $\hat{L} \in \mathbb{R}^{n\times m}$, as the first $n \times m$ block of $\hat{A}^\dagger\hat{\mathbf{O}}^\dagger(\bar{A}, C, d_1)\hat{\mathcal{H}}^-$

Notice that the noise e_t and the covariates of the estimation problem in (5.21), ϕ, are independent, and the dependencies in the prior least squares methods (5.15) are alleviated. In particular, this problem does not require any specification on how the inputs are generated and therefore can be deployed in both open- and closed-loop estimation problems.

Exploiting the specific structure of $\mathcal{G}_{\mathbf{yu}}$ in (5.19), we design a procedure named SysId, given in Algorithm 10, which recovers model parameters from $\widehat{\mathcal{G}}_{\mathbf{yu}}$. SysId is a variant of the Ho-Kalman procedure. Similar to the standard Ho-Kalman algorithm, SysId takes $\widehat{\mathcal{G}}_{\mathbf{yu}}$, the two sets of Markov parameters estimates of the predictor form, H, the systems order n, and dimensions d_1, d_2, as the inputs and computes an order n system $\hat{\Theta} = (\hat{A}, \hat{B}, \hat{C}, \hat{L})$. Note that $\hat{L}$ is an estimate of the optimal Kalman gain L given in (5.4) for the LQG control systems, and our novel estimation method allows us to estimate it, which will be also useful in control design in the upcoming sections. SysId constructs two separate $d_1 \times (d_2+1)$ Hankel matrices from the Markov parameter estimates and similar to the Ho-Kalman algorithm, since the order of the system is n, by choosing $H \geq 2n+1$ and picking $d_1 \geq n$ and $d_2 \geq n$ such $d_1+d_2+1 = H$, we guarantee a well-conditioned system identification problem.

From the blocks of $\widehat{\mathcal{G}}_{\mathbf{yu}} \in \mathbb{R}^{m\times H(m+p)}$, SysId constructs two $(n\times n+1)$ block Hankel matrices $\hat{\mathcal{H}}_{\mathbf{y}\to\mathbf{y}} \in \mathbb{R}^{nm\times(n+1)m}$ and $\hat{\mathcal{H}}_{\mathbf{u}\to\mathbf{y}} \in \mathbb{R}^{nm\times(n+1)p}$, such that (i,j)th block of $\hat{\mathcal{H}}_{\mathbf{y}\to\mathbf{y}}$ is $\hat{G}_{y\to y}^{i+j-1}$ and (i,j)th block of $\hat{\mathcal{H}}_{\mathbf{u}\to\mathbf{y}}$ is $\hat{G}_{u\to y}^{i+j-1}$ from $\widehat{\mathcal{G}}_{\mathbf{yu}}$ due to the structure

in (5.19). Then, it forms the following matrix $\hat{\mathcal{H}}$:

$$\hat{\mathcal{H}} = \left[\hat{\mathcal{H}}_{\mathbf{y}\to\mathbf{y}},\ \hat{\mathcal{H}}_{\mathbf{u}\to\mathbf{y}}\right].$$

Notice that from the Definition 5.1, if the input to the SysId was $\mathcal{G}_{\mathbf{yu}}$ then constructed Hankel matrix, $\mathcal{H}$ would be rank n,

$$\begin{aligned}\mathcal{H} &= [C^\top,\ \ldots,\ (C\bar{A}^{n-1})^\top]^\top[F,\ \ldots,\ \bar{A}^nF,\ B,\ \ldots,\ \bar{A}^nB]\\ &= \mathbf{O}(\bar{A}, C, n)\ \ [\mathbf{C}(\bar{A}, F, n+1),\quad \bar{A}^nF,\quad \mathbf{C}(\bar{A}, B, n+1),\quad \bar{A}^nB]\\ &= \mathbf{O}(\bar{A}, C, n)\ \ [F,\quad \bar{A}\mathbf{C}(\bar{A}, F, n+1),\quad B,\quad \bar{A}\mathbf{C}(\bar{A}, B, n+1)].\end{aligned}$$

Notice that $\mathcal{G}_{\mathbf{yu}}$ and $\mathcal{H}$ are uniquely identifiable for a given system Θ, whereas for any invertible $\mathbf{T} \in \mathbb{R}^{n\times n}$, the system resulting from

$$A' = \mathbf{T}^{-1}A\mathbf{T},\ B' = \mathbf{T}^{-1}B,\ C' = C\mathbf{T},\ F' = \mathbf{T}^{-1}F$$

gives the same $\mathcal{G}_{\mathbf{yu}}$ and $\mathcal{H}$. Similar to the Ho-Kalman algorithm, SysId computes the SVD of $\hat{\mathcal{H}}$ and estimates the extended observability and controllability matrices and eventually system parameters up to similarity transformation. To this end, SysId constructs $\hat{\mathcal{H}}^-$ by discarding $(n+1)$th and $(2n+2)$th block columns of $\hat{\mathcal{H}}$, i.e., if it was $\mathcal{H}$ then we have,

$$\mathcal{H}^- = \mathbf{O}(\bar{A}, C, n)\ \ [\mathbf{C}(\bar{A}, F, n+1),\quad \mathbf{C}(\bar{A}, B, n+1)].$$

The SysId procedure then calculates $\hat{\mathcal{N}}$, the best rank-n approximation of $\hat{\mathcal{H}}^-$, obtained by setting its all but top n singular values to zero. The estimates of $\mathbf{O}(\bar{A}, C, n)$, $\mathbf{C}(\bar{A}, F, n+1)$ and $\mathbf{C}(\bar{A}, B, n+1)$ are given as

$$\hat{\mathcal{N}} = \mathbf{U}\mathbf{\Sigma}^{1/2}\ \mathbf{\Sigma}^{1/2}\mathbf{V}^\top = \hat{\mathbf{O}}(\bar{A}, C, n)\ \ [\hat{\mathbf{C}}(\bar{A}, F, n+1),\quad \hat{\mathbf{C}}(\bar{A}, B, n+1)].$$

From these estimates SysId recovers $\hat{C}$ as the first $m \times n$ block of $\hat{\mathbf{O}}(\bar{A}, C, n)$, $\hat{B}$ as the first $n \times p$ block of $\hat{\mathbf{C}}(\bar{A}, B, n+1)$ and $\hat{F}$ as the first $n \times m$ block of $\hat{\mathbf{C}}(\bar{A}, F, n+1)$. Let $\hat{\mathcal{H}}^+$ be the matrix obtained by discarding 1st and $(n+2)$th block columns of $\hat{\mathcal{H}}$, i.e., if it was $\mathcal{H}$ then

$$\mathcal{H}^+ = \mathbf{O}(\bar{A}, C, n)\ \ \bar{A}\ \ [\mathbf{C}(\bar{A}, F, n+1),\quad \mathbf{C}(\bar{A}, B, n+1)].$$

Therefore, SysId recovers $\hat{\bar{A}}$ as,

$$\hat{\bar{A}} = \hat{\mathbf{O}}^\dagger(\bar{A}, C, n)\ \ \hat{\mathcal{H}}^+\ \ [\hat{\mathbf{C}}(\bar{A}, F, n+1),\quad \hat{\mathbf{C}}(\bar{A}, B, n+1)]^\dagger.$$

For identifying the model parameters of an ARX model, these are all needed by the learning agent. For the LQG control systems, the learning agent requires recovering A and L. Using the definition of $\bar{A} = A - FC$, the algorithm obtains $\hat{A} = \hat{\bar{A}} + \hat{F}\hat{C}$. For an estimate of L, SysId selects the first $n \times m$ block of $\hat{A}^{\dagger}\hat{\mathbf{O}}^{\dagger}(\bar{A}, C, n)\hat{\mathcal{H}}^{-}$. For the persistently exciting inputs, the following gives the first finite-time system identification guarantee in both open and closed-loop estimation problems.

Theorem 5.2 (Estimation Error Guarantees of the Novel System Identification Method)**.** *If the inputs are persistently exciting, then for T input-output pairs, as long as T is large enough, solving the least squares problem in (5.21) provides Markov parameter estimates $\widehat{\mathbf{G}}_{y\to y}, \widehat{\mathbf{G}}_{u\to y}$ and deploying SysId procedure gives model parameter estimates $(\hat{A}, \hat{B}, \hat{C}, \hat{F}, \hat{L})$ in which there exists a similarity transformation $\mathbf{T} \in \mathbb{R}^{n\times n}$ such that, with high probability, $\|\widehat{\mathbf{G}}_{y\to y} - \mathbf{G}_{y\to y}\|, \|\widehat{\mathbf{G}}_{u\to y} - \mathbf{G}_{u\to y}\|$, as well as $\|\hat{A} - \mathbf{T}^{-1}A\mathbf{T}\|, \|\hat{B} - \mathbf{T}^{-1}B\|, \|\hat{C} - C\mathbf{T}\|, \|\hat{F} - \mathbf{T}^{-1}F\|$, and $\|\hat{L} - \mathbf{T}^{-1}L\|$ are all $\tilde{\mathcal{O}}(1/\sqrt{T})$.*

The proof is given in Section 5.3.1. The above theorem shows that our proposed estimation method overcomes the correlations in the dynamics estimation problem, does not care about the way control inputs are generated, and provides the optimal estimation error rate of linear regression with independent noise and independent covariates. One key aspect to highlight in the above result is the persistence of excitation (PE) condition. This condition refers to the characterizations of the covariates in the least squares problem and has been the key element of the system identification algorithms to allow reliable and consistent estimation [18, 37, 38, 206]. Precisely it means that

$$\sigma_{\min}\left(\sum_{i=1}^{\tau} \phi_i \phi_i^{\top}\right) \geq \sigma_{\star}^2 \tau, \tag{5.22}$$

for some constant $\sigma_{\star}^2$, i.e., for some time τ, the minimum singular value of the design matrix scales linearly. Even though one can recover the model parameters in a self-normalized way, i.e., via quantifying the uncertainties in the estimates based on inputs, to obtain uniformly improving estimates, the persistence of excitation is required. The i.i.d. Gaussian inputs utilized in the previous section for open-loop system identification and in Chapter 3 during the improved exploration phases for StabL and TSAC allow such results. In the closed-loop control setting, one approach is to explicitly inject isotropic perturbations along with the control actions. However, as one can expect, this approach will give sub-optimal control performance. Luckily, in partially observable dynamical systems like LQG control systems or ARX systems, the closed-loop systems can reflect the effect of measurement noise

onto the regressors by construction which gives the PE condition for free. For instance, the PE condition holds in many well-known controllers such as optimal $\mathcal{H}_2$ and $\mathcal{H}_\infty$ controllers. Moreover, if a controller is PE, then there exists a neighborhood around it consisting of persistently exciting controllers, which is important to achieve consistent estimation under modeling error. Therefore, it is a mild condition to hold and it only requires a significantly wide matrix that maps a short history of noise sequences to the inputs, to be full row rank. The precise characterization of the PE condition in the open-loop setting is given in Section 5.3.2, while the precise characterization and analysis of how the PE condition is met for closed-loop controllers in the adaptive control setting are given in Sections 5.4-5.6.

5.3.1 Proof of Theorem 5.2

In this section, we first present the proof of Theorem 5.2 under the PE assumption with precise expressions. In particular, we show the self-normalized error bound on the (5.21), Theorem 5.3. Then, assuming the PE condition, we convert the self-normalized bound into a Frobenius norm bound to be used for parameter estimation error bounds in Theorem 5.4, which concludes the proof of Theorem 5.2.

First, consider the effect of the truncation bias term, $C\bar{A}^H x_{t-H}$ in (5.18). From Assumption 5.1, we have that $\bar{A}$ is (κ_3, γ_3) stable. Thus, $C\bar{A}^H x_{t-H}$ scales with the order of $(1-\gamma_3)^H$ for bounded x. In order to get consistent estimation, for some problem-dependent constant c_H, we set $H \geq \frac{\log(c_H T\sqrt{m}/\sqrt{\lambda})}{\log(1/(1-\gamma_3))}$, resulting in a negligible bias term of order $1/T$. Note that c_H is determined by the underlying system and the control policy since it is related to the scaling of the latent state. Using this we first obtain a self-normalized finite sample estimation error of (5.21):

Theorem 5.3 (Self-normalized Estimation Error)**.** *Let $\widehat{\mathcal{G}}_{\mathbf{yu}}$ be the solution to (5.21) at time τ. For $H \geq \frac{\log(c_H T\sqrt{m}/\sqrt{\lambda})}{\log(1/(1-\gamma_3))}$, define $V_\tau = \lambda I + \sum_{i=H}^{\tau} \phi_i \phi_i^\top$. Let $\|\mathcal{G}_{\mathbf{yu}}\|_F \leq S$. For $\delta \in (0,1)$, with probability at least $1-\delta$, for all $t \leq \tau$, $\mathcal{G}_{\mathbf{yu}}$ lies in the set $\mathcal{C}_{\mathcal{G}_{\mathbf{yu}},t}$, where*

$$\mathcal{C}_{\mathcal{G}_{\mathbf{yu}},t} = \{\mathcal{G}_{\mathbf{yu}}' : \mathrm{Tr}((\widehat{\mathcal{G}}_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}')V_t(\widehat{\mathcal{G}}_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}')^\top) \leq \beta_\tau^2\},$$

for $\beta_\tau = \sqrt{m\Sigma_e \log\left(\frac{\det(V_\tau)^{1/2}}{\delta \det(\lambda I)^{1/2}}\right)} + S\sqrt{\lambda} + \frac{\tau\sqrt{H}}{T}$, *where* $\Sigma_e := \|C\Sigma C^\top + \sigma_z^2 I\|_F$.

The proof is given in Appendix C.1. Note that the above result holds under sub-Gaussian e_t and is satisfied in both LQG control systems and ARX systems. Using

this result, we have

$$\sigma_{\min}(V_\tau)\|\widehat{\mathcal{G}}_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\|_F^2 \leq \mathrm{Tr}((\widehat{\mathcal{G}}_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}})V_t(\widehat{\mathcal{G}}_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}})^\top) \leq \beta_\tau^2,$$

Assume that ϕ_i is bounded (which will be rigorously shown for different adaptive control algorithms, i.e., Sections 5.4–5.6) such that $\max_{i\leq\tau}\|\phi_i\| \leq \Upsilon\sqrt{H}$. For persistently exciting inputs, *i.e.*, $\sigma_{\min}(V_\tau) \geq \sigma_\star^2\tau$ for $\sigma_\star > 0$, we get, with probability at least $1-\delta$,

$$\|\widehat{\mathcal{G}}_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\|_F \leq \frac{\sqrt{m\Sigma_e\left(\log(\frac{1}{\delta}) + \frac{H(m+p)}{2}\log\left(\frac{\lambda(m+p)+\tau\Upsilon^2}{\lambda(m+p)}\right)\right)} + S\sqrt{\lambda} + \sqrt{H}}{\sigma_\star\sqrt{\tau}} \tag{5.23}$$

after τ time steps. Note that $\widehat{\mathcal{G}}_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}} = [\widehat{\mathbf{G}}_{y\to y}, \widehat{\mathbf{G}}_{u\to y}] - [\mathbf{G}_{y\to y}, \mathbf{G}_{u\to y}]$, thus (5.23) translates to the same error bounds for $\|\widehat{\mathbf{G}}_{y\to y} - \mathbf{G}_{y\to y}\|$ and $\|\widehat{\mathbf{G}}_{u\to y} - \mathbf{G}_{u\to y}\|$, proving the first part of Theorem 5.2. This result shows that the novel least squares problem provides consistent estimates and the estimation error is $\tilde{O}(1/\sqrt{T})$ after T samples.

For the second part of Theorem 5.2, we show that SysId provides a balanced realization of Θ such that we have confidence sets around the estimated model parameters in which a similarity transformation of Θ lives in with high probability similar to Theorem 5.1. For this, define $T_{\mathcal{G}_{\mathbf{yu}}}$ as the number of samples required such that $\|\widehat{\mathcal{G}}_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\| \leq 1$ in (5.23). Let

$$T_N = T_{\mathcal{G}_{\mathbf{yu}}}\frac{8H}{\sigma_n^2(\mathcal{H})}, \quad T_B = T_{\mathcal{G}_{\mathbf{yu}}}\frac{20nH}{\sigma_n(\mathcal{H})}. \tag{5.24}$$

We have the following result on the model parameter estimates.

Theorem 5.4 (Model Parameters Estimation Error)**.** *Let $\mathcal{H}$ be the concatenation of two Hankel matrices obtained from $\mathcal{G}_{\mathbf{yu}}$. Let $\bar{A}, \bar{B}, \bar{C}, \bar{F}, \bar{L}$ be the system parameters that SysId provides for $\mathcal{G}_{\mathbf{yu}}$. At time step t, let $\hat{A}_t, \hat{B}_t, \hat{C}_t, \hat{F}_t, \hat{L}_t$ denote the system parameters obtained by SysId using $\widehat{\mathcal{G}}_{\mathbf{yu}}$. For all $t \geq \max\{T_{\mathcal{G}_{\mathbf{yu}}}, T_N, T_B\}$, for $H \geq \max\left\{2n+1, \frac{\log(c_H T\sqrt{m}/\sqrt{\lambda})}{\log(1/(1-\gamma_3))}\right\}$, there exists a unitary matrix $\mathbf{T} \in \mathbb{R}^{n\times n}$ such that, $\bar{\Theta} = (\bar{A}, \bar{B}, \bar{C}, \bar{F}, \bar{L}) \in (C_A \times C_B \times C_C \times C_F \times C_L)$ where*

$$\begin{aligned}
C_A(t) &= \left\{A' \in \mathbb{R}^{n\times n} : \|\hat{A}_t - \mathbf{T}^\top A'\mathbf{T}\| \leq \beta_t^A\right\}, \quad C_B(t) = \left\{B' \in \mathbb{R}^{n\times p} : \|\hat{B}_t - \mathbf{T}^\top B'\| \leq \beta_t^B\right\}, \\
C_C(t) &= \left\{C' \in \mathbb{R}^{m\times n} : \|\hat{C}_t - C'\mathbf{T}\| \leq \beta_t^C\right\}, \quad C_F(t) = \left\{F' \in \mathbb{R}^{n\times m} : \|\hat{F}_t - \mathbf{T}^\top F'\| \leq \beta_t^F\right\}, \\
C_L(t) &= \left\{L' \in \mathbb{R}^{n\times m} : \|\hat{L}_t - \mathbf{T}^\top L'\| \leq \beta_L(t)\right\},
\end{aligned} \tag{5.25}$$

for

$$\beta_t^A = c_1\left(\frac{\sqrt{nH}(\|\mathcal{H}\|+\sigma_n(\mathcal{H}))}{\sigma_n^2(\mathcal{H})}\right)\|\widehat{\mathcal{G}}_{\mathbf{yu}}-\mathcal{G}_{\mathbf{yu}}\|,\ \ \beta_t^B=\beta_t^C=\beta_t^F=\sqrt{\frac{20nH}{\sigma_n(\mathcal{H})}}\|\widehat{\mathcal{G}}_{\mathbf{yu}}-\mathcal{G}_{\mathbf{yu}}\|,$$
$$\beta_t^L = \frac{c_2\|\mathcal{H}\|}{\sqrt{\sigma_n(\mathcal{H})}}\beta_A + c_3\frac{\sqrt{nH}(\|\mathcal{H}\|+\sigma_n(\mathcal{H}))}{\sigma_n^{3/2}(\mathcal{H})}\|\widehat{\mathcal{G}}_{\mathbf{yu}}-\mathcal{G}_{\mathbf{yu}}\|,$$

for some problem-dependent constants c_1, c_2 and c_3.

Before presenting the proof, we state the following lemmas which are adapted from Oymak and Ozay [213] with slight modifications to fit our setting. In particular, they are originally used for the Ho-Kalman algorithm and SysId is a variant of this algorithm. These results will be useful in proving error bounds on system parameters.

Lemma 5.1. *$\mathcal{H}$, $\hat{\mathcal{H}}_t$ and $\mathcal{N}, \hat{\mathcal{N}}_t$ satisfies the following perturbation bounds,*

$$\max\left\{\left\|\mathcal{H}^+-\hat{\mathcal{H}}_t^+\right\|,\left\|\mathcal{H}^--\hat{\mathcal{H}}_t^-\right\|\right\}\le\|\mathcal{H}-\hat{\mathcal{H}}_t\|\le\sqrt{\min\{d_1,d_2+1\}}\|\widehat{\mathcal{G}}_{\mathbf{yu}}-\mathcal{G}_{\mathbf{yu}}\|$$
$$\|\mathcal{N}-\hat{\mathcal{N}}_t\|\le 2\left\|\mathcal{H}^--\hat{\mathcal{H}}_t^-\right\|\le 2\sqrt{\min\{d_1,d_2\}}\|\widehat{\mathcal{G}}_{\mathbf{yu}}-\mathcal{G}_{\mathbf{yu}}\|.$$

Lemma 5.2. *Suppose $\sigma_{\min}(\mathcal{N})\ge 2\|\mathcal{N}-\hat{\mathcal{N}}\|$ where $\sigma_{\min}(\mathcal{N})$ is the smallest nonzero singular value (i.e., nth largest singular value) of $\mathcal{N}$. Let rank-n matrices $\mathcal{N}, \hat{\mathcal{N}}$ have singular value decompositions $\mathbf{U\Sigma V}^\top$ and $\hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^\top$. There exists an $n\times n$ unitary matrix $\mathbf{T}$ so that*

$$\left\|\mathbf{U\Sigma}^{1/2}-\hat{\mathbf{U}}\hat{\mathbf{\Sigma}}^{1/2}\mathbf{T}\right\|_F^2+\left\|\mathbf{V\Sigma}^{1/2}-\hat{\mathbf{V}}\hat{\mathbf{\Sigma}}^{1/2}\mathbf{T}\right\|_F^2\le\frac{5n\|\mathcal{N}-\hat{\mathcal{N}}\|^2}{\sigma_n(\mathcal{N})-\|\mathcal{N}-\hat{\mathcal{N}}\|}.$$

Proof. For brevity, we note $\mathbf{O}=\mathbf{O}(\bar{A},C,d_1)$, $\mathbf{C_F}=\mathbf{C}(\bar{A},F,d_2+1)$, $\mathbf{C_B}=\mathbf{C}(\bar{A},B,d_2+1)$, $\hat{\mathbf{O}}_\mathbf{t}=\hat{\mathbf{O}}_\mathbf{t}(\bar{A},C,d_1)$, $\hat{\mathbf{C}}_{\mathbf{F_t}}=\hat{\mathbf{C}}_\mathbf{t}(\bar{A},F,d_2+1)$, $\hat{\mathbf{C}}_{\mathbf{B_t}}=\hat{\mathbf{C}}_\mathbf{t}(\bar{A},B,d_2+1)$. In the definition of T_N, we use $\sigma_n(H)$, due to the fact that singular values of submatrices by column partitioning are interlaced, i.e., $\sigma_n(\mathcal{N})=\sigma_n(\mathcal{H}^-)\ge\sigma_n(\mathcal{H})$. Directly applying Lemma 5.2 with the condition that for given $t\ge T_N$, we have $\sigma_{\min}(\mathcal{N})\ge 2\|\mathcal{N}-\hat{\mathcal{N}}\|$, we can guarantee that there exists a unitary transform $\mathbf{T}$ such that

$$\left\|\hat{\mathbf{O}}_\mathbf{t}-\mathbf{OT}\right\|_F^2+\left\|[\hat{\mathbf{C}}_{\mathbf{F_t}}\ \ \hat{\mathbf{C}}_{\mathbf{B_t}}]-\mathbf{T}^\top[\mathbf{C_F}\ \ \mathbf{C_B}]\right\|_F^2\le\frac{10n\|\mathcal{N}-\hat{\mathcal{N}}_t\|^2}{\sigma_n(\mathcal{N})}. \tag{5.26}$$

Since $\hat{C}_t-\bar{C}\mathbf{T}$ is a submatrix of $\hat{\mathbf{O}}_\mathbf{t}-\mathbf{OT}$, $\hat{B}_t-\mathbf{T}^\top\bar{B}$ is a submatrix of $\hat{\mathbf{C}}_{\mathbf{B_t}}-\mathbf{T}^\top\mathbf{C_B}$ and $\hat{F}_t-\mathbf{T}^\top\bar{F}$ is a submatrix of $\hat{\mathbf{C}}_{\mathbf{F_t}}-\mathbf{T}^\top\mathbf{C_F}$, we get the same bounds for them stated in (5.26). Using Lemma 5.1, with the choice of $d_1, d_2\ge\frac{H}{2}$, we have

$$\|\mathcal{N}-\hat{\mathcal{N}}_t\|\le\sqrt{2H}\|\widehat{\mathcal{G}}_{\mathbf{yu}}-\mathcal{G}_{\mathbf{yu}}\|.$$

This provides the advertised bounds in the theorem:

$$\|\hat{B}_t - \mathbf{T}^\top \bar{B}\|, \|\hat{C}_t - \bar{C}\mathbf{T}\|, \|\hat{F}_t - \mathbf{T}^\top \bar{F}\| \leq \frac{\sqrt{20nH}\|\widehat{\mathcal{G}}_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\|}{\sqrt{\sigma_n(\mathcal{N})}}.$$

Notice that for $t \geq T_B$, we have all the terms above to be bounded by 1. In order to determine the closeness of $\hat{A}_t$ and $\bar{A}$ we first consider the closeness of $\hat{\bar{A}}_t - \mathbf{T}^\top \bar{\bar{A}}\mathbf{T}$, where $\bar{\bar{A}}$ is the output obtained by SysId for $\bar{A}$ when the input is $\mathcal{G}_{\mathbf{yu}}$. Let $X = \mathbf{OT}$ and $Y = \mathbf{T}^\top [\mathbf{C_F} \;\; \mathbf{C_B}]$. Thus, we have

$$\begin{aligned}
\|\hat{\bar{A}}_t - \mathbf{T}^\top \bar{\bar{A}}\mathbf{T}\|_F &= \|\hat{\mathbf{O}}_{\mathbf{t}}^\dagger \hat{\mathcal{H}}_t^+ [\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger - X^\dagger \mathcal{H}^+ Y^\dagger\|_F \\
&\leq \left\| \left(\hat{\mathbf{O}}_{\mathbf{t}}^\dagger - X^\dagger \right) \hat{\mathcal{H}}_t^+ [\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger \right\|_F + \left\| X^\dagger \left(\hat{\mathcal{H}}_t^+ - \mathcal{H}^+ \right) [\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger \right\|_F \\
&\quad + \left\| X^\dagger \mathcal{H}^+ \left([\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger - Y^\dagger \right) \right\|_F.
\end{aligned}$$

For the first term, we have the following perturbation bound [197, 291],

$$\begin{aligned}
\|\hat{\mathbf{O}}_{\mathbf{t}}^\dagger - X^\dagger\|_F &\leq \|\hat{\mathbf{O}}_{\mathbf{t}} - X\|_F \max\{\|X^\dagger\|^2, \|\hat{\mathbf{O}}_{\mathbf{t}}^\dagger\|^2\} \\
&\leq \|\mathcal{N} - \hat{\mathcal{N}}_t\| \sqrt{\frac{10n}{\sigma_n(\mathcal{N})}} \max\{\|X^\dagger\|^2, \|\hat{\mathbf{O}}_{\mathbf{t}}^\dagger\|^2\}.
\end{aligned}$$

Since we already had $\sigma_n(\mathcal{N}) \geq 2\|\mathcal{N} - \hat{\mathcal{N}}\|$, we have $\|\hat{\mathcal{N}}\| \leq 2\|\mathcal{N}\|$ and $2\sigma_n(\hat{\mathcal{N}}) \geq \sigma_n(\mathcal{N})$. Thus,

$$\max\{\|X^\dagger\|^2, \|\hat{\mathbf{O}}_{\mathbf{t}}^\dagger\|^2\} = \max\left\{ \frac{1}{\sigma_n(\mathcal{N})}, \;\; \frac{1}{\sigma_n(\hat{\mathcal{N}})} \right\} \leq \frac{2}{\sigma_n(\mathcal{N})}. \tag{5.27}$$

Combining these and following the same steps for $\|[\hat{\mathbf{C}}_{\mathbf{F_t}} \; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger - Y^\dagger\|_F$, we get

$$\left\| \hat{\mathbf{O}}_{\mathbf{t}}^\dagger - X^\dagger \right\|_F, \;\; \left\| [\hat{\mathbf{C}}_{\mathbf{F_t}} \; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger - Y^\dagger \right\|_F \leq \left\| \mathcal{N} - \hat{\mathcal{N}}_t \right\| \sqrt{\frac{40n}{\sigma_n^3(\mathcal{N})}}. \tag{5.28}$$

The following individual bounds obtained by using (5.27), (5.28) and triangle inequality:

$$\begin{aligned}
\left\| \left(\hat{\mathbf{O}}_{\mathbf{t}}^\dagger - X^\dagger \right) \hat{\mathcal{H}}_t^+ [\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger \right\|_F &\leq \|\hat{\mathbf{O}}_{\mathbf{t}}^\dagger - X^\dagger\|_F \|\hat{\mathcal{H}}_t^+\| \|[\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger\| \\
&\leq \frac{4\sqrt{5n} \left\| \mathcal{N} - \hat{\mathcal{N}}_t \right\|}{\sigma_n^2(\mathcal{N})} \left(\|\mathcal{H}^+\| + \|\hat{\mathcal{H}}_t^+ - \mathcal{H}^+\| \right) \\
\left\| X^\dagger \left(\hat{\mathcal{H}}_t^+ - \mathcal{H}^+ \right) [\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger \right\|_F &\leq \frac{2\sqrt{n}\|\hat{\mathcal{H}}_t^+ - \mathcal{H}^+\|}{\sigma_n(\mathcal{N})} \\
\left\| X^\dagger \mathcal{H}^+ \left([\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger - Y^\dagger \right) \right\|_F &\leq \|X^\dagger\| \|\mathcal{H}^+\| \|[\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger - Y^\dagger\| \\
&\leq \frac{2\sqrt{10n} \left\| \mathcal{N} - \hat{\mathcal{N}}_t \right\|}{\sigma_n^2(\mathcal{N})} \|\mathcal{H}^+\|.
\end{aligned}$$

Combining these we get

$$\|\hat{\bar{A}}_t - \mathbf{T}^\top \bar{\bar{A}} \mathbf{T}\|_F \le \frac{31\sqrt{n}\|\mathcal{H}^+\| \left\|\mathcal{N} - \hat{\mathcal{N}}_t\right\|}{2\sigma_n^2(\mathcal{N})} + \|\hat{\mathcal{H}}_t^+ - \mathcal{H}^+\| \left(\frac{4\sqrt{5n}\left\|\mathcal{N} - \hat{\mathcal{N}}_t\right\|}{\sigma_n^2(\mathcal{N})} + \frac{2\sqrt{n}}{\sigma_n(\mathcal{N})} \right)$$
$$\le \frac{31\sqrt{n}\|\mathcal{H}^+\|}{2\sigma_n^2(\mathcal{N})} \left\|\mathcal{N} - \hat{\mathcal{N}}_t\right\| + \frac{13\sqrt{n}}{2\sigma_n(\mathcal{N})} \|\hat{\mathcal{H}}_t^+ - \mathcal{H}^+\|.$$

These results give the estimation error guarantees for the ARX systems. For LQG control systems we additionally need to recover A and L. Now consider $\hat{A}_t = \hat{\bar{A}}_t + \hat{F}_t \hat{C}_t$. Using Lemma 5.1,

$$\begin{aligned}
&\|\hat{A}_t - \mathbf{T}^\top \bar{A} \mathbf{T}\|_F \\
&= \|\hat{\bar{A}}_t + \hat{F}_t \hat{C}_t - \mathbf{T}^\top \bar{\bar{A}} \mathbf{T} - \mathbf{T}^\top \bar{F} \bar{C} \mathbf{T}\|_F \\
&\le \|\hat{\bar{A}}_t - \mathbf{T}^\top \bar{\bar{A}} \mathbf{T}\|_F + \|(\hat{F}_t - \mathbf{T}^\top \bar{F}) \hat{C}_t\|_F + \|\mathbf{T}^\top \bar{F} (\hat{C}_t - \bar{C}\mathbf{T})\|_F \\
&\le \|\hat{\bar{A}}_t - \mathbf{T}^\top \bar{\bar{A}} \mathbf{T}\|_F + \|(\hat{F}_t - \mathbf{T}^\top \bar{F})\|_F \|\hat{C}_t - \bar{C}\mathbf{T}\|_F \\
&\qquad + \|(\hat{F}_t - \mathbf{T}^\top \bar{F})\|_F \|\bar{C}\| + \|\bar{F}\| \|(\hat{C}_t - \bar{C}\mathbf{T})\|_F \\
&\le \frac{31\sqrt{2nH}\|\mathcal{H}\|}{2\sigma_n^2(\mathcal{N})} \|\widehat{\mathcal{G}}_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\| + \frac{13\sqrt{nH}}{2\sqrt{2}\sigma_n(\mathcal{N})} \|\widehat{\mathcal{G}}_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\| + \frac{20nH \|\widehat{\mathcal{G}}_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\|^2}{\sigma_n(\mathcal{N})} \\
&\qquad + (\|\bar{F}\| + \|\bar{C}\|) \|\widehat{\mathcal{G}}_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\| \sqrt{\frac{20nH}{\sigma_n(\mathcal{N})}}.
\end{aligned}$$

Using the result above, to obtain an estimation error bound for $\hat{L}_t$, we define T_A as the samples required to have $\|\hat{A}_t - \mathbf{T}^\top \bar{A} \mathbf{T}\| \le \sigma_n(\bar{A})/2$ for all $t \ge T_A$, i.e., $T_A = T_{\mathcal{G}_{\mathbf{yu}}} \left(\frac{\frac{62\sqrt{2nH}\|\mathcal{H}\|}{2\sigma_n^2(\mathcal{N})} + \frac{26\sqrt{nH}}{2\sqrt{2}\sigma_n(\mathcal{N})} + (\|\bar{F}\|\|\bar{H}\|\|\bar{C}\|) \sqrt{\frac{80nH}{\sigma_n(\mathcal{N})}} + \sqrt{\frac{40nH\sigma_n(\bar{A})}{\sigma_n(\mathcal{N})}}}{\sigma_n(\bar{A})} \right)^2$. From Weyl's inequality, we have $\sigma_n(\hat{A}_t) \ge \sigma_n(\bar{A})/2$. Recalling that $X = \mathbf{O}(\bar{A}, C, d_1)\mathbf{T}$, under Assumption 5.1, we consider $\hat{L}_t$:

$$\begin{aligned}
&\|\hat{L}_t - \mathbf{T}^\top \bar{L}\|_F \\
&= \|\hat{A}_t^\dagger \hat{\mathbf{O}}_{\mathbf{t}}^\dagger \hat{\mathcal{H}}_t^- - \mathbf{T}^\top \bar{A}^\dagger \mathbf{O}^\dagger \mathcal{H}^-\|_F \\
&\le \|(\hat{A}_t^\dagger - \mathbf{T}^\top \bar{A}^\dagger \mathbf{T}) \hat{\mathbf{O}}_{\mathbf{t}}^\dagger \hat{\mathcal{H}}_t^-\|_F + \|\mathbf{T}^\top \bar{A}^\dagger \mathbf{T} (\hat{\mathbf{O}}_{\mathbf{t}}^\dagger - X^\dagger) \hat{\mathcal{H}}_t^-\|_F + \|\mathbf{T}^\top \bar{A}^\dagger \mathbf{T} X^\dagger (\hat{\mathcal{H}}_t^- - \mathcal{H}^-)\|_F \\
&\le \|\hat{A}_t^\dagger - \mathbf{T}^\top \bar{A}^\dagger \mathbf{T}\|_F \|\hat{\mathbf{O}}_{\mathbf{t}}^\dagger\| \|\hat{\mathcal{H}}_t^-\| + \|\hat{\mathbf{O}}_{\mathbf{t}}^\dagger - X^\dagger\|_F \|\bar{A}^\dagger\| \|\hat{\mathcal{H}}_t^-\| + \sqrt{n}\|\hat{\mathcal{H}}_t^- - \mathcal{H}^-\| \|\bar{A}^\dagger\| \|X^\dagger\| \\
&\le \left(\|\hat{A}_t^\dagger - \mathbf{T}^\top \bar{A}^\dagger \mathbf{T}\|_F \sqrt{\frac{2}{\sigma_n(\mathcal{N})}} + \left\|\mathcal{N} - \hat{\mathcal{N}}_t\right\| \sqrt{\frac{40n}{\sigma_n^3(\mathcal{N})}} \|\bar{A}^\dagger\| \right) \left(\|\mathcal{H}^-\| + \|\hat{\mathcal{H}}_t^- - \mathcal{H}^-\| \right) \\
&\qquad + \sqrt{n}\|\bar{A}^\dagger\| \frac{1}{\sqrt{\sigma_n(\mathcal{N})}} \|\hat{\mathcal{H}}_t^- - \mathcal{H}^-\|.
\end{aligned}$$

Again using the perturbation bounds of the Moore–Penrose inverse under the Frobenius norm [197], we have $\|\hat{A}_t^\dagger - \mathbf{T}^\top \bar{A}^\dagger \mathbf{T}\|_F \leq \frac{2}{\sigma_n^2(\bar{A})} \|\hat{A}_t - \mathbf{T}^\top \bar{A} \mathbf{T}\|$. Notice that the similarity transformation that transfers A to $\bar{A}$ is bounded since $S = \left([C^\top \ (C\bar{A})^\top \dots (C\bar{A}^{d_1-1})^\top]^\top\right)^\dagger \mathbf{O}(\bar{A}, C, d_1)$. Combining all and using Lemma 5.1, we obtain the confidence set for $\hat{L}_t$ given in Theorem 5.4. □

Combining Theorem 5.4 with the guarantee that $\|\widehat{\mathcal{G}}_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\| = \tilde{O}(1/\sqrt{T})$ given in (5.23), finishes the proof of the second part of Theorem 5.2. Overall, we showed that our novel system identification method allows closed-loop and open-loop estimation in both LQG and ARX systems. This method will be the key piece in our adaptive control design.

Remark 5.1. *Note that to recover $\mathcal{G}_{\mathbf{yu}}$ using the closed-loop system identification method presented in this section, we only require stabilizability condition on (A, B) and detectability conditions on (A, C), i.e., there exists a matrix K and F such that $A - BK$ and $A - FC$ are stable, rather than controllability and observability conditions provided in Assumption 5.1. Stabilizability and detectability are necessary and sufficient conditions to have a well-defined learning and control problem in partially observable linear dynamical systems, and they provide the conditions required for our novel closed-loop system identification method to work, i.e., stable $\bar{A}$. However, controllability and observability assumptions are required for the subspace identification method SysId, since it requires rank-n observability and controllability matrices to achieve a balanced realization. If the goal is to recover the Markov parameters of the system or if one can design adaptive control methods only using Markov parameter estimates, e.g., Section 5.6.5, stabilizability and detectability of the underlying system are sufficient to have reliable estimates as in Theorem 5.3 and* (5.23).

5.3.2 PE Condition in the Open-Loop Setting

Before studying the adaptive control problem in partially observable linear dynamical systems, at the end of this section, we show that the PE condition required for consistent estimation is satisfied for the open-loop control, i.e., i.i.d. Gaussian control inputs. To this end, we introduce the truncated open-loop noise evolution parameter $\mathcal{G}^{ol}$. $\mathcal{G}^{ol}$ represents the effect of noises in the system on the outputs. We define $\mathcal{G}^{ol}$ for $2H$ time steps back in time and show that the last $2H$ process and measurement noises provide sufficient persistent excitation for the covariates in the estimation problem. In the following, we show that there exists a positive σ_o such that $\sigma_o < \sigma_{\min}(\mathcal{G}^{ol})$, *i.e.,* $\mathcal{G}^{ol}$ is full row rank. Let $\bar{\phi}_t = P\phi_t$ for a permutation

matrix P that gives

$$\bar{\phi}_t = \begin{bmatrix} y_{t-1}^\top & u_{t-1}^\top \dots y_{t-H}^\top & u_{t-H}^\top \end{bmatrix}^\top \in \mathbb{R}^{(m+p)H}.$$

We will consider the state space representation for the analysis for LQG control systems given in (5.1), but one can apply the same analysis for predictor form/ARX systems (see [163] for the details). For the control input of $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$, let $f_t = [y_t^\top u_t^\top]^\top$. From the evolution of the system with given input we have the following:

$$f_t = \mathbf{G^o} \begin{bmatrix} w_{t-1}^\top & z_t^\top & u_t^\top & \dots & w_{t-H}^\top & z_{t-H+1}^\top & u_{t-H+1}^\top \end{bmatrix}^\top + \mathbf{r_t^o}$$

where

$$\mathbf{G^o} := \begin{bmatrix} 0_{m\times n}\, I_{m\times m}\, 0_{m\times p} & C\, 0_{m\times m}\, CB & \dots & CA^{H-2}\, 0_{m\times m}\, CA^{H-2}B \\ 0_{p\times n}\, 0_{p\times m}\, I_{p\times p} & 0_{p\times n}\, 0_{p\times m}\, 0_{p\times p} & \dots & 0_{p\times n}\, 0_{p\times m}\, 0_{p\times p} \end{bmatrix},$$

and $\mathbf{r_t^o}$ is the residual vector that represents the effect of $[w_{i-1}\ z_i\ u_i]$ for $0 \le i < t - H$, which are independent. Notice that $\mathbf{G^o}$ is full row rank even for $H = 1$, due to first $(m + p) \times (m + n + p)$ block. Using this, we can represent $\bar{\phi}_t$ as follows

$$\bar{\phi}_t = \underbrace{\begin{bmatrix} f_{t-1} \\ \vdots \\ f_{t-H} \end{bmatrix}}_{\mathbb{R}^{(m+p)H}} + \begin{bmatrix} \mathbf{r_{t-1}^o} \\ \vdots \\ \mathbf{r_{t-H}^o} \end{bmatrix} = \mathcal{G}^{ol} \underbrace{\begin{bmatrix} w_{t-2} \\ z_{t-1} \\ u_{t-1} \\ \vdots \\ w_{t-2H-1} \\ z_{t-2H} \\ u_{t-2H} \end{bmatrix}}_{\mathbb{R}^{2(n+m+p)H}} + \begin{bmatrix} \mathbf{r_{t-1}^o} \\ \vdots \\ \mathbf{r_{t-H}^o} \end{bmatrix} \quad \text{where}$$

$$\mathcal{G}^{ol} := \begin{bmatrix} [\ \mathbf{G^o}\ ] & 0_{(m+p)\times(m+n+p)} & 0_{(m+p)\times(m+n+p)} & 0_{(m+p)\times(m+n+p)} & \dots \\ 0_{(m+p)\times(m+n+p)} & [\ \mathbf{G^o}\ ] & 0_{(m+p)\times(m+n+p)} & 0_{(m+p)\times(m+n+p)} & \dots \\ & & \ddots & & \\ 0_{(m+p)\times(m+n+p)} & 0_{(m+p)\times(m+n+p)} & \dots & [\ \mathbf{G^o}\ ] & 0_{(m+p)\times(m+n+p)} \\ 0_{(m+p)\times(m+n+p)} & 0_{(m+p)\times(m+n+p)} & 0_{(m+p)\times(m+n+p)} & \dots & [\ \mathbf{G^o}\ ] \end{bmatrix}. \tag{5.29}$$

Recall Assumption 5.1. The following lemma shows that covariates ϕs are bounded for the given system under open-loop control.

Lemma 5.3. *After applying the control inputs of $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ for T_w time steps for all $1 \le t \le T_w$, with probability $1 - \delta/2$,*

$$\|x_t\| \le X_w := \frac{(\sigma_w + \sigma_u\|B\|)\kappa_1(1-\gamma_1)}{\sqrt{1-(1-\gamma_1)^2}}\sqrt{2n\log(12nT_w/\delta)}, \tag{5.30}$$

$$\|z_t\| \le Z := \sigma_z\sqrt{2m\log(12mT_w/\delta)}, \tag{5.31}$$

$$\|u_t\| \le U_w := \sigma_u\sqrt{2p\log(12pT_w/\delta)}, \tag{5.32}$$

$$\|y_t\| \le \|C\|X_w + Z. \tag{5.33}$$

Thus, we have $\max_{i\le t\le T_w}\|\phi_i\| \le \Upsilon_w\sqrt{H}$, *where* $\Upsilon_w = \|C\|X_w + Z + U_w$.

Proof. For all $1 \le t \le T_w$, $\Sigma(x_t) \preccurlyeq \mathbf{\Gamma}_\infty$, where $\mathbf{\Gamma}_\infty$ is the steady state covariance matrix of x_t such that,

$$\mathbf{\Gamma}_\infty = \sum\nolimits_{i=0}^{\infty} \sigma_w^2 A^i (A^\top)^i + \sigma_u^2 A^i B B^\top (A^\top)^i.$$

From the Assumption 5.1, we have $\|A^\tau\| \le \kappa_1(1-\gamma_1)^\tau$ for all $\tau \ge 0$. Thus, $\|\mathbf{\Gamma}_\infty\| \le (\sigma_w^2 + \sigma_u^2\|B\|^2)\frac{\kappa_1^2(1-\gamma_1)^2}{1-(1-\gamma_1)^2}$. Notice that each x_t is component-wise $\sqrt{\|\mathbf{\Gamma}_\infty\|}$-sub-Gaussian random variable. Using standard sub-Gaussian vector norm upper bound with a union bound argument, we get the advertised result. □

The following lemma shows that the i.i.d. Gaussian inputs uniformly excite the system and satisfy the PE condition after enough interactions.

Lemma 5.4 (Persistence of Excitation in Open-Loop Control Setting)**.** *$\mathcal{G}^{ol}$ is full row-rank such that $\sigma_{\min}(\mathcal{G}^{ol}) > \sigma_o > 0$. For some $\delta \in (0,1)$, and Υ_w defined in Lemma 5.3, let $T_o = 32\Upsilon_w^4\sigma_o^{-4}\log^2\left(\frac{2H(m+p)}{\delta}\right)\max\{\sigma_w^{-4}, \sigma_z^{-4}, \sigma_u^{-4}\}$. After applying the control inputs of $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ for $T_w \ge T_o$ time steps, with probability at least $1-\delta$ we have $\sigma_{\min}\left(\sum_{i=1}^t \phi_i\phi_i^\top\right) \ge t\frac{\sigma_o^2}{2}\min\{\sigma_w^2, \sigma_z^2, \sigma_u^2\}$.*

Proof. Let $\bar{\mathbf{0}} = 0_{(m+p)\times(m+n+p)}$. Since each block row is full row-rank, we get the following decomposition using QR decomposition for each block row:

$$\mathcal{G}^{ol} = \underbrace{\begin{bmatrix} Q^o & 0_{m+p} & 0_{m+p} & 0_{m+p} & \dots \\ 0_{m+p} & Q^o & 0_{m+p} & 0_{m+p} & \dots \\ & & \ddots & & \\ 0_{m+p} & 0_{m+p} & \dots & Q^o & 0_{m+p} \\ 0_{m+p} & 0_{m+p} & 0_{m+p} & \dots & Q^o \end{bmatrix}}_{\mathbb{R}^{(m+p)H\times(m+p)H}} \underbrace{\begin{bmatrix} R^o & \bar{\mathbf{0}} & \bar{\mathbf{0}} & \bar{\mathbf{0}} & \dots \\ \bar{\mathbf{0}} & R^o & \bar{\mathbf{0}} & \bar{\mathbf{0}} & \dots \\ & & \ddots & & \\ \bar{\mathbf{0}} & \bar{\mathbf{0}} & \dots & R^o & \bar{\mathbf{0}} \\ \bar{\mathbf{0}} & \bar{\mathbf{0}} & \bar{\mathbf{0}} & \dots & R^o \end{bmatrix}}_{\mathbb{R}^{(m+p)H\times 2(m+n+p)H}},$$

where $R^o = \begin{bmatrix} \times & \times & \times & \times & \times & \times & \dots \\ 0 & \times & \times & \times & \times & \times & \dots \\ & \ddots & & & & & \\ 0 & 0 & 0 & \times & \times & \times & \dots \end{bmatrix} \in \mathbb{R}^{(m+p)\times H(m+n+p)}$ where the elements in the diagonal are positive numbers. Notice that the first matrix with Q^0 is full rank. Also, all the rows of the second matrix are in row echelon form and the second matrix is full row-rank. Thus, we can deduce that $\mathcal{G}^{ol}$ is full row-rank, i.e., $\sigma_{\min}(\mathcal{G}^{ol}) > \sigma_o > 0$. Since $\mathcal{G}^{ol}$ is full row rank, we have that

$$\mathbb{E}[\bar{\phi}_t \bar{\phi}_t^\top] \geq \mathcal{G}^{ol} \Sigma_{w,z,u} \mathcal{G}^{ol\top},$$

where $\Sigma_{w,z,u} \in \mathbb{R}^{2(n+m+p)H \times 2(n+m+p)H} = \text{diag}(\sigma_w^2, \sigma_z^2, \sigma_u^2, \dots, \sigma_w^2, \sigma_z^2, \sigma_u^2)$. This gives us

$$\sigma_{\min}(\mathbb{E}[\bar{\phi}_t \bar{\phi}_t^\top]) \geq \sigma_o^2 \min\{\sigma_w^2, \sigma_z^2, \sigma_u^2\}$$

for all t. From Lemma 5.3, we have $\max_{i \leq \tau} \|\phi_i\| \leq \Upsilon_w \sqrt{H}$ with probability at least $1 - \delta/2$. Given this holds, one can use Matrix Azuma inequality in [267], to obtain the following which holds with probability $1 - \delta/2$:

$$\lambda_{\max}\left(\sum_{i=1}^{t} \phi_i \phi_i^\top - \mathbb{E}[\phi_i \phi_i^\top]\right) \leq 2\sqrt{2t}\Upsilon_w^2 H \sqrt{\log\left(\frac{2H(m+p)}{\delta}\right)}.$$

Using Weyl's inequality, during the warm-up period with probability $1 - \delta$, we have

$$\sigma_{\min}\left(\sum_{i=1}^{t} \phi_i \phi_i^\top\right) \geq t\sigma_o^2 \min\{\sigma_w^2, \sigma_z^2, \sigma_u^2\} - 2\sqrt{2t}\Upsilon_w^2 H \sqrt{\log\left(\frac{2H(m+p)}{\delta}\right)}.$$

For all $t \geq T_o$, we have the stated lower bound. □

This result verifies that the PE condition holds in the open-loop control setting, which shows that the estimation error guarantees given in Theorem 5.2 hold for open-loop data collection. Therefore, even if the closed-loop PE condition is not satisfied such that we cannot guarantee the estimation error guarantees given in Theorem 5.2 for the closed-loop control, one can use the novel system identification method with the i.i.d. control inputs to obtain state-of-the-art guarantees. However, if one has PE in the closed-loop setting we can further guarantee consistent improvement of estimates which would not be possible with prior methods. This novelty will be crucial in the adaptive control tasks discussed next.

5.4 Optimism-Based Adaptive Control

After studying the novel system identification method in Section 5.3, we study adaptive control of partially observable linear dynamical systems in this section. In particular, we will use the principle of optimism in the face of uncertainty (OFU) to design the controllers using the confidence sets given by the system identification method and balance exploration and exploitation trade-off. Recall that the optimism principle has been used in Chapter 2 and Chapter 3 in the design of StabL to achieve state-of-the-art regret guarantees in the learning and control of various systems. In this section, we will first consider this control design method in obtaining regret guarantees for learning and control of LQG control systems and then we will extend these results to the ARX systems.

We propose **LQG** control via **Opt**imism (LqgOpt), an adaptive control algorithm for learning and controlling unknown LQG control systems. LqgOpt interacts with the system, collects samples, estimates the model parameters, and adapts accordingly. LqgOpt deploys OFU principle to balance the *exploration* vs. *exploitation* trade-off. Using the predictor form of the state-space equations of the partially observable linear systems, we deploy the least-squares estimation problem introduced in Section 5.3 and obtain confidence sets on the system parameters.

LqgOpt then uses these confidence sets to find the optimistic model and use the optimal controller for the chosen model for further exploration-exploitation. To analyze the finite-time regret of LqgOpt, we first provide a stability analysis for the sequence of optimistic controllers. We then present a novel way of regret decomposition by deriving the Bellman optimality equation for average cost per stage LQG control. Utilizing the OFU principle, we prove that LqgOpt achieves a regret upper bound of $\tilde{O}(\sqrt{T})$ for adaptive control of partially observable linear dynamical systems with *convex* quadratic cost if the underlying optimal controller (5.9) allows PE condition, where T is the number of total interactions. If the PE condition does not hold for the underlying optimal LQG controller, which results in inconsistent system identification under closed-loop control, then we show that LqgOpt achieves a regret upper bound of $\tilde{O}(T^{2/3})$. These results make LqgOpt the state-of-the-art learning and control algorithm for partially observable linear dynamical systems with *convex* quadratic cost.

Finally, we consider the ARX models with a convex quadratic cost function. We show that with appropriate changes in the control design of LqgOpt, the same regret guarantees hold for ARX systems with sub-Gaussian noise. In particular, we show

Algorithm 11 LqgOpt

Input: T_w, H, $\delta > 0$, Q, R
—— Warm-Up ——
for $t = 1, \ldots, T_w$ **do**
  Deploy $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ and store $\mathcal{D}_1 = \{y_t, u_t\}_{t=1}^{T_w}$
—— Adaptive Control in Epochs ——
for $i = 1, \ldots$ **do**
  Calculate $\widehat{\mathcal{G}}_{\mathbf{yu}}^i$ using $\mathcal{D}_i = \{y_t, u_t\}_{t=1}^{2^{i-1}T_w}$ via (5.21)
  Deploy SysId $(H, \widehat{\mathcal{G}}_{\mathbf{yu}}^i, n)$ for $\hat{A}_i, \hat{B}_i, \hat{C}_i, \hat{L}_i$
  Construct the confidence sets $C_A(i), C_B(i), C_C(i), C_L(i)$ given in (5.25)
  Find $\tilde{\Theta}_i = (\tilde{A}_i, \tilde{B}_i, \tilde{C}_i, \tilde{L}_i) \in C_i \cap \mathcal{S}$ for $C_i := (C_A(i) \times C_B(i) \times C_C(i) \times C_L(i))$ s.t.
    $J(\tilde{\Theta}_i) \leq \inf_{\Theta' \in C_i \cap \mathcal{S}} J(\Theta') + T^{-1}$
  for $t = 2^{i-1}T_w, \ldots 2^i T_w - 1$ **do**
    Execute the optimal controller for $\tilde{\Theta}_i$

that without the closed-loop PE condition for the optimal ARX controller, LqgOpt yields regret of $\tilde{O}(T^{2/3})$ and with the PE condition which allows continuously updated model estimates via closed-loop data, we attain regret of $\tilde{O}(\sqrt{T})$ in adaptive control of ARX systems.

The rest of this section is organized as follows: in Section 5.4.1, we describe the algorithm of LqgOpt and provide the main regret guarantees for the adaptive control of LQG control systems. In Section 5.4.2 we provide the PE condition in the closed-loop control using the underlying optimal controller and show that this condition can be satisfied by LqgOpt with small enough estimation error. Then, we show that LqgOpt keeps the measurements and the state estimates bounded in Section 5.4.3, and in Sections 5.4.4 and 5.4.5 we give the regret decomposition and the proofs of the main results respectively. Finally, in Section 5.4.6 we extend the prior results to the ARX systems.

5.4.1 Adaptive Control via LqgOpt

In this section, we present LqgOpt, and describe its compounding components. The outline of LqgOpt is given in Algorithm 11. The early stage of deploying LqgOpt involves a fixed warm-up period dedicated to pure exploration using Gaussian excitation. In particular it excites the system with $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ for $1 \leq t \leq T_w$. LqgOpt requires this exploration period to estimate the model parameters reliably enough that the controller designed based on the parameter estimation and their confidence set results in a stabilizing controller on the real system. The duration of this period T_w depends on how stabilizable the true parameters are and how accurate the model

estimations should be, i.e., characterizations provided in Assumption 5.1. We will formally quantify these statements and the length of the warm-up period shortly.

After the warm-up period, LqgOpt utilizes the model parameter estimations and their confidence sets to design a controller corresponding to an optimistic model in the confidence sets, obtained by following the OFU principle. Due to the reliable estimation from the warm-up period, this controller and all the future designed controllers stabilize the underlying true unknown model. The agent deploys the prescribed controller on the real system for exploration and exploitation. The agent collects samples throughout its interaction with the environment and uses these samples for further improvement in model estimation, confidence interval construction, and design of the controller regarding an optimistic model. This process functions in epochs of doubling length until the end of execution. In particular, in an epoch, the agent uses the most recent optimistic controller to control the underlying system Θ for twice as long as the duration of the previous control policy, i.e., each epoch i for $i = \{1, 2, \dots, \}$ is of length $2^{i-1}T_w$ time steps. This technique is known as "the doubling trick" in reinforcement learning and online learning which prevents frequent policy updates and balances the policy changes so that the overall regret of the algorithm is affected by a constant factor only.

System Identification

LqgOpt uses the novel system identification procedure described in Section 5.3, which allows both open-loop and closed-loop data collection to obtain consistent estimates of the dynamics. In particular, at the beginning of each epoch i, it solves the regularized least squares problem given in (5.21) to recover input-to-output and output-to-output Markov parameters of Θ in predictor form using the entire history of data up to the current time-step, $\mathcal{D}_i = \{y_t, u_t\}_{t=1}^{2^{i-1}T_w}$. The estimated Markov parameters $\widehat{\mathcal{G}}^i_{\mathbf{yu}}$ are then used with SysId to obtain a balanced realization of the underlying model parameters, $\hat{A}_i, \hat{B}_i, \hat{C}_i, \hat{L}_i$ with corresponding confidence sets $C_A(i), C_B(i), C_C(i), C_L(i)$ as presented in (5.25). This confidence set $C_i := (C_A(i) \times C_B(i) \times C_C(i) \times C_L(i))$ contains the underlying parameters $\Theta = (A, B, C, L)$ up to a similarity transformation with high probability. LqgOpt uses this confidence set with the set $\mathcal{S}$ defined in Assumption 5.1 to select the optimistic model among the plausible models. As LqgOpt collects more data, the confidence sets shrink with the rate given in Theorem 5.2, providing significantly refined estimates of the model parameters. LqgOpt adapts and updates its policy by

deploying the OFU principle on the new confidence sets.

Recall that in Section 5.3.2, we showed that using control inputs of $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ allows PE condition and consistent estimation. In particular, we defined $\mathcal{G}^{ol}$ which encodes the open-loop evolution of the disturbances in the system and represents the responses to these disturbances on the *batch* of observations and actions *history* and showed that $\mathcal{G}^{ol}$ is full row-rank, i.e., $\sigma_{\min}(\mathcal{G}^{ol}) > \sigma_o > 0$ for some known σ_o, allowing PE condition. Thus, we have the guarantee that after the warm-up period of LqgOpt, the estimation error of model parameters is $\tilde{O}(1/\sqrt{T_w})$, due Theorem 5.2.

Similarly, in order to obtain the PE condition and the consistent system identification during the adaptive control which happens with a closed-loop controller, we define the truncated closed-loop noise evolution parameter $\mathcal{G}^{cl}$. When the controller is set to be the optimal policy for the underlying system in (5.9), i.e., closed-loop system, $\mathcal{G}^{cl} \in \mathbb{R}^{H(m+p)\times 2H(n+m)}$ represents the translation of the truncated history of process and measurement noises on the inputs, ϕ's. The exact construction of $\mathcal{G}^{cl}$ is provided in detail in Equation (5.38) of the next section. Briefly, it is formed by shifting a block matrix $\bar{\mathbf{G}} \in \mathbb{R}^{(m+p)\times 2H(n+m)}$ by $m+n$ in each block row where $\bar{\mathbf{G}}$ is constructed by H $(m+p)\times(n+m)$ matrices. Assuming that H used in LqgOpt is large enough such that $\bar{\mathbf{G}}$ is full row rank for the given system, we will show that $\mathcal{G}^{cl}$ is also full row rank. Thus, we have that for the choice of H in LqgOpt, $\sigma_{\min}(\mathcal{G}^{cl})$ is lower bounded by some positive value, i.e., $\sigma_{\min}(\mathcal{G}^{cl}) > \sigma_c > 0$, where LqgOpt only knows σ_c and searches for an optimistic system whose closed-loop noise evolution parameter satisfies this lower bound. Note that we define $\mathcal{G}^{cl}$ based on the optimal closed-loop system and we need to make sure that our model parameter estimates are close enough to the true ones such that the PE condition is also satisfied for the constructed controller. This analysis is provided in the next section, Lemma 5.5. With the guarantee of the PE condition in the closed-loop setting, LqgOpt is guaranteed to continuously refine the model parameter estimates, thus it improves the controllers and effectively balances the exploration-exploitation trade-off.

Adaptive Control

After estimating the model parameters effectively at the beginning of epoch i, LqgOpt uses these confidence sets along with the set $\mathcal{S}$ to implement the OFU principle. In particular, at time $t = 2^{i-1}T_w$, the algorithm chooses a system $\tilde{\Theta}_i = (\tilde{A}_i, \tilde{B}_i, \tilde{C}_i, \tilde{L}_i)$

from $C_i \cap \mathcal{S}$ such that

$$J(\tilde{\Theta}_i) \leq \inf_{\Theta' \in C_i \cap \mathcal{S}} J(\Theta') + 1/T. \tag{5.34}$$

LqgOpt then designs the optimal feedback policy $(\tilde{K}_t, \tilde{L}_t)$ for the chosen system $\tilde{\Theta}_t$ as shown in (5.9), i.e., it uses $\tilde{A}_t, \tilde{B}_t, \tilde{C}_t$, and $\tilde{L}_t$ for estimating the underlying state and deploys the feedback gain matrix of $\tilde{K}_t$ to design the control inputs. This measurement feedback policy is executed until the end of the epoch, whose duration is twice the previous epoch. The following gives the regret guarantee for LqgOpt.

Theorem 5.5 (Regret of LqgOpt with closed-loop PE condition)**.** *Given an LQG control system $\Theta = (A, B, C)$, and regulating parameters $Q \succeq 0$ and $R \succ 0$, suppose Assumptions 5.1 and 5.2 hold such that the underlying system satisfies the PE condition with its optimal policy, i.e., $\sigma_{\min}(\mathcal{G}^{cl}) > \sigma_c > 0$. Fixing a horizon T, let $H \geq \max\left\{2n+1, \frac{\log(c_H T\sqrt{m}/\sqrt{\lambda})}{\log(1/(1-\gamma_3))}\right\}$ and*

$$T_w = poly\left(H, \sigma_o, \sigma_c, \kappa_1, \kappa_2, \kappa_3, \frac{1}{1-\gamma_1}, \frac{1}{1-\gamma_2}, \frac{1}{1-\gamma_3}, \psi, m, n, p\right).$$

Then, with high probability, the regret of LqgOpt with a warm-up duration of T_w is $\textsc{Regret}(T) = \tilde{O}\left(\sqrt{T}\right)$.

The proof of this result will be presented in Section 5.4.5 with intermediate results given in Sections 5.4.2–5.4.4. Here T_w is chosen to guarantee well-refined model estimates, the PE condition during the warm-up and adaptive control periods, the stability of the optimistic controllers, and the boundedness of the measurements and state estimations. The exact requirements on T_w are given in the following sections with detailed expressions. Nevertheless, the warm-up duration is a fixed problem-dependent constant. This result shows that LqgOpt achieves the same regret rate of LQR systems shown in Chapter 3 in the challenging partially observable LQG control system setting. Moreover, this makes LqgOpt the first adaptive control algorithm to attain $\tilde{O}\left(\sqrt{T}\right)$ regret for partially observable linear dynamical systems with convex cost. The following corollary is the direct extension of the result above and considers the case when the underlying optimal controller does not satisfy the PE condition. In this case, closed-loop system identification cannot provide reliable and consistent estimates and LqgOpt relies solely on the warm-up duration with i.i.d. Gaussian inputs, i.e., the open-loop control. Therefore, throughout the adaptive control process all the model parameter estimation errors scale with $\tilde{O}\left(1/\sqrt{T_w}\right)$.

Corollary 5.5.1 (Regret of LqgOpt without the closed-loop PE condition)**.** *For the system given in Theorem 5.5 with the choices of H and T_w, if the underlying system is not persistently excited with its optimal policy, LqgOpt incurs the following regret with high probability, $\textsc{Regret}(T) = \tilde{O}\left(T_w + \frac{T-T_w}{\sqrt{T_w}}\right)$. Therefore, the optimal regret upper bound of this setting is obtained with a warm-up duration of $T_w = O(T^{2/3})$, which gives the regret of $\textsc{Regret}(T) = \tilde{O}\left(T^{2/3}\right)$ for LqgOpt.*

This result shows that if the PE condition does not hold for the underlying system with the optimal controller, then LqgOpt requires a longer open-loop exploration, i.e., warm-up, to compensate for this lack of improvement during the adaptive control.

5.4.2 PE Condition in the Closed-Loop Setting

As the previous result shows, having the PE condition satisfied during the adaptive control provides a significant improvement in regret due to the novel closed-loop system identification method proposed in Section 5.3. In this section, we explore this condition and give its precise characterization in the closed-loop control setting when the system is controlled by its optimal controller. Then, we show that this condition holds under small enough estimation errors which can be guaranteed with the warm-up period of LqgOpt.

After the warm-up period, for $t \geq T_w$, at epoch i, LqgOpt executes the control input of $u_t = -\tilde{K}_t \hat{x}_{t|t,\tilde{\Theta}}$ using the optimistic model $\tilde{\Theta}_i$. Using the state estimation update equations given in (5.2)-(5.4) with the optimistic parameters, we have:

$$\begin{aligned}
\hat{x}_{t|t-1,\tilde{\Theta}} &= \tilde{A}_{t-1}\hat{x}_{t-1|t-1,\tilde{\Theta}} - \tilde{B}_{t-1}\tilde{K}_{t-1}\hat{x}_{t-1|t-1,\tilde{\Theta}} \\
\hat{x}_{t|t,\tilde{\Theta}} &= \hat{x}_{t|t-1,\tilde{\Theta}} + \tilde{L}_t(y_t - \tilde{C}_t\hat{x}_{t|t-1,\tilde{\Theta}}) \\
&= (\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1})\hat{x}_{t-1|t-1,\tilde{\Theta}} + \tilde{L}_t(Cx_t + z_t - \tilde{C}_t(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1})\hat{x}_{t-1|t-1,\tilde{\Theta}}) \\
&= (I - \tilde{L}_t\tilde{C}_t)(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1})\hat{x}_{t-1|t-1,\tilde{\Theta}} + \tilde{L}_t(C(Ax_{t-1} - B\tilde{K}_{t-1}\hat{x}_{t-1|t-1,\tilde{\Theta}} + w_{t-1}) + z_t).
\end{aligned} \tag{5.35}$$

Similar to Section 5.3.2, let $f_t = [y_t^\top, u_t^\top]^\top$ and $\mathbf{x}_t = [x_t^\top, \hat{x}_{t|t,\tilde{\Theta}}^\top]^\top$. Using (5.1) and (5.35), we have the following equations,

$$\mathbf{x}_t = \underbrace{\begin{bmatrix} A & -B\tilde{K}_{t-1} \\ \tilde{L}_tCA & (I - \tilde{L}_t\tilde{C}_t)(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1}) - \tilde{L}_tCB\tilde{K}_{t-1} \end{bmatrix}}_{\tilde{\mathbf{G}}_2^{(t)}} \mathbf{x}_{t-1} + \underbrace{\begin{bmatrix} I & 0 \\ \tilde{L}_tC & \tilde{L}_t \end{bmatrix}}_{\tilde{\mathbf{G}}_3^{(t)}} \begin{bmatrix} w_{t-1} \\ z_t \end{bmatrix}$$

$$f_t = \underbrace{\begin{bmatrix} C & 0 \\ 0 & -\tilde{K}_t \end{bmatrix}}_{\tilde{\mathbf{\Gamma}}_\mathbf{t}} \tilde{\mathbf{G}}_\mathbf{2}^{(\mathbf{t})} \begin{bmatrix} x_{t-1} \\ \hat{x}_{t-1|t-1,\tilde{\Theta}} \end{bmatrix} + \underbrace{\begin{bmatrix} C & 0 \\ 0 & -\tilde{K}_t \end{bmatrix}}_{\tilde{\mathbf{\Gamma}}_\mathbf{t}} \underbrace{\begin{bmatrix} I & 0 \\ \tilde{L}_t C & \tilde{L}_t \end{bmatrix}}_{\mathbf{G}_\mathbf{3}^{(\mathbf{t})}} \begin{bmatrix} w_{t-1} \\ z_t \end{bmatrix} + \begin{bmatrix} z_t \\ 0 \end{bmatrix}.$$

Rolling back in time for H time steps we get the following,

$$f_t = \tilde{\mathbf{\Gamma}}_\mathbf{t} \left(\sum_{i=t-H+1}^{t} \left(\prod_{j=i}^{t} \tilde{\mathbf{G}}_\mathbf{2}^{(\mathbf{j})} \right) \tilde{\mathbf{G}}_\mathbf{3}^{(\mathbf{i-1})} \begin{bmatrix} w_{i-2} \\ z_{i-1} \end{bmatrix} \right) + \underbrace{\begin{bmatrix} C & I \\ -\tilde{K}_t \tilde{L}_t C & -\tilde{K}_t \tilde{L}_t \end{bmatrix}}_{\tilde{\mathbf{G}}_\mathbf{1}^{(\mathbf{t})}} \begin{bmatrix} w_{t-1} \\ z_t \end{bmatrix} + \mathbf{r}_\mathbf{t}^\mathbf{c}, \quad (5.36)$$

where $\mathbf{r}_\mathbf{t}^\mathbf{c}$ is the residual vector that represents the effect of $[w_{i-1} \;\; z_i]$ for $0 \leq i < t-H$, which are independent. Let $\bar{\phi}_t = P\phi_t$ for a permutation matrix P that gives

$$\bar{\phi}_t = \begin{bmatrix} y_{t-1}^\top & u_{t-1}^\top \dots y_{t-H}^\top & u_{t-H}^\top \end{bmatrix}^\top \in \mathbb{R}^{(m+p)H}.$$

Using (5.36), we can represent $\bar{\phi}_t$ as follows

$$\bar{\phi}_t = \underbrace{\begin{bmatrix} f_{t-1} \\ \vdots \\ f_{t-H} \end{bmatrix}}_{\mathbb{R}^{(m+p)H}} + \begin{bmatrix} \mathbf{r}_\mathbf{t-1}^\mathbf{c} \\ \vdots \\ \mathbf{r}_\mathbf{t-H}^\mathbf{c} \end{bmatrix} = \mathcal{G}_t^{cl} \underbrace{\begin{bmatrix} w_{t-2} \\ z_{t-1} \\ \vdots \\ w_{t-2H-1} \\ z_{t-2H} \end{bmatrix}}_{\mathbb{R}^{2(n+m)H}} + \begin{bmatrix} \mathbf{r}_\mathbf{t-1}^\mathbf{c} \\ \vdots \\ \mathbf{r}_\mathbf{t-H}^\mathbf{c} \end{bmatrix} \text{ where}$$

$$\mathcal{G}_t^{cl} = \begin{bmatrix} [\quad \bar{\mathbf{G}}_\mathbf{t-1} \quad] & 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & \dots \\ 0_{(m+p)\times(m+n)} & [\quad \bar{\mathbf{G}}_\mathbf{t-2} \quad] & 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & \dots \\ & & \ddots & & \\ 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & \dots & [\quad \bar{\mathbf{G}}_\mathbf{t-H+1} \quad] & 0_{(m+p)\times(m+n)} \\ 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & \dots & [\quad \bar{\mathbf{G}}_\mathbf{t-H} \quad] \end{bmatrix} \text{ for} \quad (5.37)$$

$$\bar{\mathbf{G}}_\mathbf{t} = \left[\tilde{\mathbf{G}}_\mathbf{1}^{(\mathbf{t})}, \tilde{\mathbf{\Gamma}}_\mathbf{t} \tilde{\mathbf{G}}_\mathbf{2}^{(\mathbf{t})} \tilde{\mathbf{G}}_\mathbf{3}^{(\mathbf{t-1})}, \tilde{\mathbf{\Gamma}}_\mathbf{t} \tilde{\mathbf{G}}_\mathbf{2}^{(\mathbf{t})} \tilde{\mathbf{G}}_\mathbf{2}^{(\mathbf{t-1})} \tilde{\mathbf{G}}_\mathbf{3}^{(\mathbf{t-2})}, \dots, \tilde{\mathbf{\Gamma}}_\mathbf{t} \tilde{\mathbf{G}}_\mathbf{2}^{(\mathbf{t})} \tilde{\mathbf{G}}_\mathbf{2}^{(\mathbf{t-1})} \cdots \tilde{\mathbf{G}}_\mathbf{3}^{(\mathbf{t-H})} \right] \in \mathbb{R}^{(m+p)\times H(n+m)}$$

Notice that if the agent knows the underlying model parameters, it can deploy the optimal control policy. Therefore, we denote $\mathcal{G}^{cl}$ as the translation of the process

and measurement noises into $\bar{\phi}_t$ while using the optimal policy in (5.9):

$$\mathcal{G}^{cl} = \begin{bmatrix} [\quad \bar{\mathbf{G}} \quad] & 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & \cdots \\ 0_{(m+p)\times(m+n)} & [\quad \bar{\mathbf{G}} \quad] & 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & \cdots \\ & & \ddots & & \\ 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & \cdots & [\quad \bar{\mathbf{G}} \quad] & 0_{(m+p)\times(m+n)} \\ 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & \cdots & [\quad \bar{\mathbf{G}} \quad] \end{bmatrix} \tag{5.38}$$

where

$$\bar{\mathbf{G}} = \begin{bmatrix} \mathbf{G_1}, & \mathbf{\Gamma G_2 G_3}, & \mathbf{\Gamma G_2}^2\mathbf{G_3}, & \ldots, & \mathbf{\Gamma G_2}^{H-1}\mathbf{G_3} \end{bmatrix} \in \mathbb{R}^{(m+p)\times H(n+m)} \tag{5.39}$$

for $\mathbf{G_1} = \begin{bmatrix} C & I \\ -KLC & -KL \end{bmatrix}$, $\mathbf{G_2} = \begin{bmatrix} A & -BK \\ LCA & (I-LC)(A-BK)-LCBK \end{bmatrix}$, $\mathbf{G_3} = \begin{bmatrix} I & 0 \\ LC & L \end{bmatrix}$, and $\mathbf{\Gamma} = \begin{bmatrix} C & 0 \\ 0 & -K \end{bmatrix}$.

Assumption 5.2 (PE structure of the underlying system with its optimal control)**.** *H is large enough such that $\bar{\mathbf{G}}$ given in* (5.39) *is full row rank.*

Note that this assumption is solely dependent on the underlying system and for long enough H one can show that it holds. Similar to the case with truncated open-loop noise evolution parameter in Section 5.3.2, having full row rank block rows provides a full row rank $\mathcal{G}^{cl}$ via the same QR decomposition argument. Therefore, under Assumption 5.2, we have a lower bound of the smallest singular value of the $H-$length truncated closed-loop noise evolution parameter, $\sigma_{\min}(\mathcal{G}^{cl}) > \sigma_c > 0$. However, we need to show that LqgOpt can satisfy this condition even under modeling errors, which is shown in the following.

Due to Assumption 5.1, i.e., the boundedness of the set $\mathcal{S}$ that LqgOpt is searching on, let $\|\tilde{\mathcal{G}^{cl}}\|_F \leq G$ for all model in $\mathcal{S}$. For bounded covariates ϕ, $\max_{T_w \leq t \leq T} \|\phi_i\| \leq \Upsilon_c\sqrt{H}$, which will be rigorously shown shortly in Section 5.4.3, define $G_r = G + \frac{\sigma_c\sqrt{H(m+p)}}{2}$, $\eta_T = \sigma_w\sqrt{2n\log\left(\frac{2nT}{\delta}\right)} + \sigma_z\sqrt{2m\log\left(\frac{2mT}{\delta}\right)}$, and

$$T_c = \frac{2048\Upsilon_c^4 H^2\left(\log(\frac{H(m+p)}{\delta}) + H^2(m+p)(m+n)\log(G_r + \frac{32H\Upsilon_c\sqrt{2}\eta_T + 32H\eta_T^2 + 16\max\{\sigma_w^2,\sigma_z^2\}}{\sigma_c^2\min\{\sigma_w^2,\sigma_z^2\}})\right)}{\sigma_c^4\min\{\sigma_w^4,\sigma_z^4\}}.$$

With these definitions, the following lemma shows that LqgOpt persistently excites the system during the adaptive control period after a long enough warm-up period.

Lemma 5.5. *Let* $T_w \geq T_{\mathcal{G}} \coloneqq T_B \left(\frac{2H+2H\kappa_2\psi+2H(H-1)\kappa_2\psi}{\sigma_c}\right)^2$, *where* T_B *is defined in* (5.24). *Suppose Assumption 5.2 holds. After* T_c *time steps in adaptive control period, with probability* $1-3\delta$, *we have the following for all* $t \geq T_c$,

$$\sigma_{\min}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top\right) \geq t\frac{\sigma_c^2\min\{\sigma_w^2,\sigma_z^2\}}{16}. \tag{5.40}$$

Proof. Define $\tilde{\mathcal{G}^{cl}}$, which is the translation parameter for the process and measurement noises into $\bar{\phi}_t$ for the system that is governed by the *optimistically chosen parameter by* LqgOpt *while using the optimal optimistic controller*. Recall that we are searching for the optimistic system model which attains the optimal LQG cost over the set of $\mathcal{C}_t \cap \mathcal{S}$ and whose closed-loop noise evolution parameter satisfies the lower bound on the smallest singular value of the $H-$length truncated closed-loop noise evolution parameter, σ_c. Therefore, LqgOpt has the guarantee that $\sigma_{\min}(\tilde{\mathcal{G}^{cl}}) \geq \sigma_c$. Picking $T_w \geq T_{\mathcal{G}}$, guarantees that in adaptive control period for all $t \geq T_w$, $\|\mathcal{G}_t^{cl} - \tilde{\mathcal{G}^{cl}}\| \leq \frac{\sigma_c}{2}$. Using Weyl's inequality on singular values, we have that $\sigma_{\min}(\mathcal{G}_t^{cl}) \geq \frac{\sigma_c}{2}$. Hence, for all $t \geq T_w$, we have that

$$\mathbb{E}[\bar{\phi}_t\bar{\phi}_t^\top] \succeq \mathcal{G}_t^{cl}\Sigma_{w,z}\mathcal{G}_t^{cl\top},$$

where $\Sigma_{w,z} \in \mathbb{R}^{2(n+m)H\times 2(n+m)H} = \mathrm{diag}(\sigma_w^2,\sigma_z^2,\ldots,\sigma_w^2,\sigma_z^2)$. This gives us

$$\sigma_{\min}(\mathbb{E}[\bar{\phi}_t\bar{\phi}_t^\top]) \geq \frac{\sigma_c^2}{4}\min\{\sigma_w^2,\sigma_z^2\}$$

for $t \geq T_w$. Let ϕ be bounded during the adaptive control as $\|\phi_t\| \leq \Upsilon_c\sqrt{H}$ with probability at least $1-2\delta$. Given this holds, *for a given optimistic model*, one can use Matrix Azuma inequality [267] similar to the proof of $\mathcal{G}^{ol}$, to obtain the following which holds with probability $1-\delta$:

$$\mathbf{\Lambda}_t = \lambda_{\max}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top - \mathbb{E}[\phi_i\phi_i^\top]\right) \leq 2\sqrt{2t}\Upsilon_c^2 H\sqrt{\log\left(\frac{H(m+p)}{\delta}\right)}. \tag{5.41}$$

Notice that this upper bound holds only for a single model. However, we need to show that for any random model within the confidence set, it holds. Thus, we use a standard covering argument. Using the perturbation result that holds for all $t \geq T_w$, we have $\|\mathcal{G}_t^{cl}\|_F \leq G_r$. We have the following upper bound on the covering number:

$$\mathcal{N}(B(G_r),\|\cdot\|_F,\epsilon) \leq \left(G_r+\frac{2}{\epsilon}\right)^{(m+p)(n+m)H^2}.$$

Thus, the following holds for all the centers of ϵ-balls in $\|\mathcal{G}_t^{cl}\|_F$, for all $t \geq T_w$, with probability $1-\delta$:

$$\mathbf{\Lambda}_t \leq 2\sqrt{2t}\Upsilon_c^2 H \sqrt{\log\left(\frac{H(m+p)}{\delta}\right) + H^2(m+p)(m+n)\log\left(G_r + \frac{2}{\epsilon}\right)}. \quad (5.42)$$

Considering all the systems in the ϵ-balls, during the adaptive control period with probability $1-3\delta$, we have

$$\sigma_{\min}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top\right) \geq t\left(\frac{\sigma_c^2}{4}\min\{\sigma_w^2,\sigma_z^2\} - 2\epsilon\left(H\Upsilon_c\sqrt{2}\eta_T + H\eta_T^2 + \max\{\sigma_w^2/2, \sigma_z^2/2\}\right)\right)$$
$$- 2\sqrt{2t}\Upsilon_c^2 H\sqrt{\log\left(\frac{H(m+p)}{\delta}\right) + H^2(m+p)(m+n)\log\left(G_r + \frac{2}{\epsilon}\right)}.$$

Let $\epsilon = \frac{\sigma_c^2\min\{\sigma_w^2,\sigma_z^2\}}{16\left(H\Upsilon_c\sqrt{2}\eta_T + H\eta_T^2 + \max\{\sigma_w^2/2,\sigma_z^2/2\}\right)}$. This gives the following bound

$$\sigma_{\min}\left(\textstyle\sum_{i=1}^{t}\phi_i\phi_i^\top\right) \geq t\left(\frac{\sigma_c^2}{8}\min\{\sigma_w^2,\sigma_z^2\}\right)$$
$$-2\sqrt{2t}\Upsilon_c^2 H\sqrt{\log\left(\tfrac{H(m+p)}{\delta}\right) + H^2(m+p)(m+n)\log\left(G_r + \tfrac{32H\Upsilon_c\sqrt{2}\eta_T + 32H\eta_T^2 + 16\max\{\sigma_w^2,\sigma_z^2\}}{\sigma_c^2\min\{\sigma_w^2,\sigma_z^2\}}\right)}.$$

For all $t \geq T_c$, we have the stated lower bound. □

Since the first adaptive control epoch is of length T_w, due to doubling epoch durations, by setting $T_w \geq T_c$, we guarantee that the estimates obtained at the end of the first epoch are consistent and have the estimation error rate of $\tilde{O}(1/\sqrt{2T_w})$, including the data of the warm-up phase while solving (5.21) for system identification. This result verifies that the PE condition holds in the closed-loop control setting for LqgOpt as long as Assumption 5.2 holds. This shows that the estimation error guarantees given in Theorem 5.2 hold for closed-loop data collection as well, which allows LqgOpt to update its model parameter estimates and reduce the regret consistently. Note that this consistent improvement is made possible by the novel system identification method discussed in Section 5.3, since even if the PE condition holds, the prior works which only use input-to-output Markov parameters for system identification would not be able to achieve this improvement due to the correlations in the covariates mentioned in Section 5.2.

5.4.3 Boundedness of the Output and State Estimation

In the previous sections, we established the warm-up duration to achieve persistence of excitation throughout the entire execution of LqgOpt, in both open and closed-loop settings, which verified the consistent estimation property in Theorem 5.2

for LqgOpt. The only thing that remains is to analyze the regret of LqgOpt. In Section 5.4.5, we will show that the regret of the warm-up phase scales linearly as expected due to i.i.d. Gaussian inputs. In order to analyze the regret obtained during the adaptive control period, we first need to show that system will be well-controlled during the adaptive control period.

This result is critical for regret analysis due to the nature of the adaptive control problem in partially observable environments. The inaccuracies in the system parameter estimates affect both the optimal feedback gain synthesis and the estimation of the underlying state. If these inaccuracies are not tolerable in the adaptive control of the system, they will accumulate fast and cause explosion and unboundedness in the input and output of the system. This would result in linear, and potentially super-linear regret. Therefore, we need to show that the policies designed by LqgOpt stabilize the system dynamics right after the warm-up period. To this end, we need to set the warm-up duration such that the model estimation error after the warm-up phase is small enough to yield stable closed-loop dynamics and yield bounded inputs and outputs. To this end, define

$$T_{\max} = \text{poly}\left(H, \sigma_o, \sigma_c, \kappa_1, \kappa_2, \kappa_3, \frac{1}{1-\gamma_1}, \frac{1}{1-\gamma_2}, \frac{1}{1-\gamma_3}, \psi, m, n, p\right). \tag{5.43}$$

The following lemma shows that for a long enough warm-up, we have this desired closed-loop stability.

Lemma 5.6. *Suppose Assumptions 5.1 holds. After the warm-up period of $T_w \geq T_{\max}$, LqgOpt has the $(\kappa', \gamma')-$stable closed-loop dynamics when applied to the underlying system with probability $1-\delta$ where*

$$\kappa' = poly(\kappa_1, \kappa_2, \kappa_3, \psi, (1-\gamma_2)^{-1}, (1-\gamma_3)^{-1}), \quad \gamma' = poly(\gamma_2, \gamma_3). \tag{5.44}$$

Moreover, for all $T_w \leq t \leq \tau$ and $\delta \in (0,1)$, with probability $1-\delta$, we have that

$$\begin{aligned} \|x_t\| \leq \bar{X}_\tau, \quad \|y_t\| \leq \bar{Y}_\tau, \\ \|\hat{x}_{t|t,\tilde{\Theta}}\| \leq \bar{\mathcal{X}}_\tau, \quad \|u_t\| \leq \bar{U}_\tau, \end{aligned} \tag{5.45}$$

for $\bar{X}_\tau, \bar{Y}_\tau, \bar{\mathcal{X}}_\tau, \bar{U}_\tau = O(\sqrt{\log(\tau/\delta)})$. Here, O hides the problem-dependent constants.

The proof of the lemma with the precise expressions is given in Appendix C.2.1. Here, we provide a proof sketch. Given the optimal control policy of $\tilde{\Theta}_i$ at epoch i, we construct a $2n$-dimensional autonomous linear dynamical system of the joint evolution of the state x_t and the MMSE estimate using $\tilde{\Theta}_i$, $\hat{x}_{t|t,\tilde{\Theta}}$. By showing that

the joint evolution is stable when the system Θ is controlled by its own optimal controller, we can create a neighborhood (ρ-ball) around Θ such that any model in the proximity yields a (κ', γ')-stable joint evolution. We use the following result in our proof, whose proof is also provided in the Appendix.

Lemma 5.7 (Strong stability of perturbation)**.** *Suppose the matrix $A \in \mathbb{R}^{n\times n}$ is (κ, γ)-stable for $\kappa \geq 1$ and $\gamma \in (0, 1]$. For $\gamma' \in (0, \gamma]$ and perturbation $\Delta \in \mathbb{R}^{n\times n}$, the perturbed matrix $A + \Delta$ is (κ, γ')-stable whenever $\|\Delta\| \leq \kappa^{-1}(\gamma - \gamma')$.*

Similar to finding a stabilizing neighborhood in Chapter 3, i.e., Lemma 3.3, we use Lemma 5.7 to deduce the estimation error that we can tolerate such that the optimistic controllers for the systems selected within the confidence sets stabilize the underlying system. By setting T_w long enough such that the first adaptive control epoch is long enough to provide bounded inputs and outputs we conclude the proof.

Lemma 5.5 and Lemma 5.6 together show that by setting $T_w \geq H + T_{\max}$, LqgOpt guarantees that while the system parameter estimates are continuously refining, the input to the system and the system's output stay bounded during the adaptive control period, i.e., LqgOpt stabilizes and persistently excites the underlying system right after the warm-up during the first epoch in the adaptive control period. Note that collecting more data in the subsequent epochs does not decrease the design matrix, i.e., making it less positive definite, after the warm-up period even if the PE condition does not hold. Therefore, we can argue that the model estimation error in the adaptive control epochs is worst-case $\tilde{O}(1/\sqrt{T_w})$ even if Assumption 5.2 does not hold. Therefore, all controllers designed by LqgOpt in the adaptive control epochs after a warm-up period of T_w stabilize the underlying system due Lemma 5.6.

5.4.4 Regret Decomposition

Given the verification of stability in the adaptive control period in Lemma 5.6, we now focus on the regret of adaptive control. To analyze the regret expression given in (5.12), we leverage the optimistic controller design. Recall that in Chapter 3 for the analysis of StabL and TSAC, we use the Bellman optimality equation for LQR [28], and consider the system evolution of the optimistic system using the optimistic controller in parallel with the true system evolution under the optimistic controller such that they share the same process noise. With this approach, for StabL and TSAC, we divide the regret into several pieces which account for the model estimation error, the martingale property of the value function with a bounded difference, the policy updates, and the difference in optimal average expected cost

between the underlying model $J_*(\Theta)$ and the model parameter used in the controller design $J_*(\tilde{\Theta})$. Since StabL uses optimistic controllers similar to LqgOpt, we get $J_*(\tilde{\Theta}) - J_*(\Theta) \leq 0$.

In order to provide a similar analysis to the LQR case, we first derive the Bellman optimality equation for the average cost-per-step LQG control problem. Surprisingly, this result was not stated in the literature and can be of independent interest. For infinite state and control space average cost per step problems, e.g., the LQG control system Θ with regulating parameters Q and R, the optimal average cost per stage $J_*(\Theta)$ as given in (5.8) and the differential(relative) cost satisfy Bellman optimality equation [28]. In the following lemma, we identify the correct differential cost for LQG systems and derive the Bellman optimality equation.

Lemma 5.8 (Bellman Optimality Equation for LQG)**.** *Given state estimation $\hat{x}_{t|t-1} \in \mathbb{R}^n$ and an observation $y_t \in \mathbb{R}^m$ pair at time t, the Bellman optimality equation of average cost per stage control of LQG system $\Theta = (A, B, C)$ with regulating parameters Q and R is*

$$\begin{aligned} J_*(\Theta) + \hat{x}_{t|t}^\top (P - C^\top Q C)\hat{x}_{t|t} + y_t^\top Q y_t \qquad (5.46)\\ = \min_u \left\{ y_t^\top Q y_t + u^\top R u + \mathbb{E}\left[\hat{x}_{t+1|t+1}^{u\top} \left(P - C^\top Q C\right) \hat{x}_{t+1|t+1}^u + y_{t+1}^{u\top} Q y_{t+1}^u \right] \right\}, \end{aligned}$$

where P is the unique solution to DARE of Θ (5.10), $\hat{x}_{t|t} = (I - LC)\hat{x}_{t|t-1} + Ly_t$, $y_{t+1}^u = C(Ax_t + Bu + w_t) + z_{t+1}$, and $\hat{x}_{t+1|t+1}^u = (I - LC)\left(A\hat{x}_{t|t} + Bu\right) + Ly_{t+1}^u$. The equality is achieved by the optimal controller of Θ given in (5.9)*.*

Proof. Define $\hat{\omega}_t = Ax_t - A\hat{x}_{t|t} + w_t$. Notice that $\hat{\omega}_t$ is independent of the policy used and depends only on the estimation error and noise in the steady state. Also notice that $\mathbb{E}\left[\hat{\omega}_t \hat{\omega}_t^\top\right] = \Sigma$ where Σ is the positive semidefinite solution to the algebraic Riccati equation given in (5.5):

$$\Sigma = A\bar{\Sigma}A^\top + \sigma_w^2 I, \qquad \bar{\Sigma} = \Sigma - \Sigma C^\top \left(C\Sigma C^\top + \sigma_z^2 I\right)^{-1} C\Sigma. \qquad (5.47)$$

Recall the state estimation updates given in (5.2)-(5.4) such that, for any given y_t and $\hat{x}_{t|t-1}$ at time t, the optimal state estimation and the output at $t+1$ can be written as

$$\hat{x}_{t|t} = (I - LC)\,\hat{x}_{t|t-1} + Ly_t, \quad \hat{x}_{t+1|t,u} = A\hat{x}_{t|t} + Bu, \quad y_{t+1,u} = CA\hat{x}_{t|t} + CBu + C\hat{\omega}_t + z_{t+1}. \qquad (5.48)$$

Since the aim is to minimize the average cost per stage of controlling Θ, the optimal control input is given as $u = -K\hat{x}_{t|t}$. Recall that optimal average stage cost of LQG

is $J_*(\Theta) = \text{Tr}(C^\top Q C \bar{\Sigma}) + \text{Tr}(P(\Sigma - \bar{\Sigma})) + \text{Tr}(\sigma_z^2 Q)$. Suppose the differential cost h is a quadratic function of s_t where $s_t = [\hat{x}_{t|t-1}^\top \; y_t^\top]^\top \in \mathbb{R}^{n+m}$, i.e.,

$$h(s_t) = s_t^\top \begin{bmatrix} G_1 & G_2 \\ G_2^\top & G_3 \end{bmatrix} s_t = \hat{x}_{t|t-1}^\top G_1 \hat{x}_{t|t-1} + 2\hat{x}_{t|t-1}^\top G_2 y_t + y_t^\top G_3 y_t.$$

One needs to verify that there exists G_1, G_2, G_3 such that they satisfy Bellman optimality equation for the chosen differential cost:

$$\begin{aligned} &J_*(\Theta) + \hat{x}_{t|t-1}^\top G_1 \hat{x}_{t|t-1} + 2\hat{x}_{t|t-1}^\top G_2 y_t + y_t^\top G_3 y_t = \\ &y_t^\top Q y_t + \hat{x}_{t|t}^\top K^\top R K \hat{x}_{t|t} + \mathbb{E}\left[\hat{x}_{t+1|t}^\top G_1 \hat{x}_{t+1|t} + 2\hat{x}_{t+1|t}^\top G_2 y_{t+1} + y_{t+1}^\top G_3 y_{t+1}\right] \end{aligned}$$

Using the fact that $\bar{\Sigma} = \Sigma - L\left(C\Sigma C^\top + \sigma_z^2 I\right) L^\top$, we can write the optimal average cost as $J_*(\Theta) = \text{Tr}\left((Q + L^\top P L - L^\top C^\top Q C L)\left(C\Sigma C^\top + \sigma_z^2 I\right)\right)$. Expanding the expectation given $\hat{x}_{t|t-1}$, y_t and using (5.48) , we get

$$\begin{aligned} &\hat{x}_{t|t-1}^\top G_1 \hat{x}_{t|t-1} + 2\hat{x}_{t|t-1}^\top G_2 y_t + y_t^\top G_3 y_t \qquad (5.49) \\ &\quad = y_t^\top Q y_t + \hat{x}_{t|t}^\top K^\top R K \hat{x}_{t|t} + \hat{x}_{t|t}^\top (A - BK)^\top G_1 (A - BK)\hat{x}_{t|t} \\ &\quad + 2\hat{x}_{t|t}^\top (A - BK)^\top G_2 C (A - BK)\hat{x}_{t|t} + \hat{x}_{t|t}^\top (A - BK)^\top C^\top G_3 C (A - BK)\hat{x}_{t|t} \\ &\quad + \mathbb{E}\left[\hat{\omega}_t^\top C^\top G_3 C \hat{\omega}_t + z_{t+1}^\top G_3 z_{t+1}\right] - \text{Tr}\left(\left(Q + L^\top P L - L^\top C^\top Q C L\right)\left(C\Sigma C^\top + \sigma_z^2 I\right)\right). \end{aligned}$$

Notice that $\mathbb{E}\left[\hat{\omega}_t^\top C^\top G_3 C \hat{\omega}_t + z_{t+1}^\top G_3 z_{t+1}\right] = \text{Tr}\left(G_3\left(C\Sigma C^\top + \sigma_z^2 I\right)\right)$. In order to match with the last term of (5.49), set $G_3 = Q + L^\top (P - C^\top Q C) L$. Inserting G_3 to (5.49), we get following 3 equations to solve for G_1 and G_2:

1) From quadratic terms of y_t:

$$\begin{aligned} &L^\top P L - L^\top C^\top Q C L \\ &\quad = L^\top K^\top R K L + L^\top (A - BK)^\top G_1 (A - BK) L + 2L^\top (A - BK)^\top G_2 C (A - BK) L \\ &\quad + L^\top (A - BK)^\top C^\top \left(Q + L^\top P L - L^\top C^\top Q C L\right) C (A - BK) L; \end{aligned}$$

2) From quadratic terms of $x_{t|t-1}$:

$$\begin{aligned} G_1 &= (I - LC)^\top K^\top R K (I - LC) + (I - LC)^\top (A - BK)^\top G_1 (A - BK)(I - LC) \\ &\quad + 2(I - LC)^\top (A - BK)^\top G_2 C (A - BK)(I - LC) \\ &\quad + (I - LC)^\top (A - BK)^\top C^\top \left(Q + L^\top P L - L^\top C^\top Q C L\right) C (A - BK)(I - LC); \end{aligned}$$

3) From bilinear terms of $x_{t|t-1}$ and y_t:

$$\begin{aligned} G_2 &= (I - LC)^\top K^\top R K L + (I - LC)^\top (A - BK)^\top G_1 (A - BK) L \\ &\quad + 2(I - LC)^\top (A - BK)^\top G_2 C (A - BK) L \\ &\quad + (I - LC)^\top (A - BK)^\top C^\top \left(Q + L^\top P L - L^\top C^\top Q C L\right) C (A - BK) L. \end{aligned}$$

$G_1 = (I-LC)^\top (P - C^\top QC)(I-LC)$ and $G_2 = (I-LC)^\top (P - C^\top QC) L$ satisfies all 3 equations. Thus one can write the Bellman optimality equation as

$$
\begin{aligned}
&J_*(\Theta)+\hat{x}_{t|t-1}^\top (I-LC)^\top \left(P-C^\top QC\right)(I-LC)\hat{x}_{t|t-1} \\
&+2\hat{x}_{t|t-1}^\top (I-LC)^\top \left(P-C^\top QC\right) L y_t + y_t^\top \left(Q+L^\top \left(P-C^\top QC\right)L\right) y_t = \\
&y_t^\top Q y_t + \hat{x}_{t|t}^\top K^\top R K \hat{x}_{t|t} + \mathbb{E}\left[\hat{x}_{t+1|t}^\top (I-LC)^\top \left(P-C^\top QC\right)(I-LC)\hat{x}_{t+1|t}\right] \\
&+2\mathbb{E}\left[\hat{x}_{t+1|t}^\top (I-LC)^\top \left(P-C^\top QC\right) L y_{t+1} + y_{t+1}^\top \left(Q+L^\top \left(P-C^\top QC\right)L\right) y_{t+1}\right].
\end{aligned}
$$

Combining terms using (5.48) gives the advertised result. □

Using this result, we decompose the regret similar to the optimistic controllers in the LQR setting. In particular, we use (5.8) to study the one-step (instantaneous) system evolution of the optimistic system $\tilde{\Theta}_t$ using the optimistic controller $u_t = -\tilde{K}_t \hat{x}_{t|t,\tilde{\Theta}_t}$ in parallel with the true system Θ evolution under the same optimistic controller such that they share the exact process and *measurement* noises, which was not present in the LQR setting. Notice that, this is equivalent to providing the regret decomposition for a system that is obtained via similarity transformation $\mathbf{S}$, i.e., $A' = \mathbf{S}^{-1}A\mathbf{S}$, $B' = \mathbf{S}^{-1}B$, $C' = C\mathbf{S}$. Therefore, without loss of generality, we will assume that $\mathbf{S} = I$ in the regret decomposition and the concentration bounds used in the regret analysis. To formally define two different instantaneous system revolutions described above, for given $\hat{x}_{t|t-1}$ and y_t at time t, we define the following expressions for time step $t+1$ using the model specified as a subscript:

$$\hat{x}_{t|t,\tilde{\Theta}_t} = \left(I - \tilde{L}_t \tilde{C}_t\right)\hat{x}_{t|t-1} + \tilde{L}_t y_t \tag{5.50}$$

$$y_{t+1,\tilde{\Theta}_t} = \tilde{C}_t\left(\tilde{A}_t - \tilde{B}_t \tilde{K}_t\right)\hat{x}_{t|t,\tilde{\Theta}_t} + \tilde{C}_t \tilde{A}_t\left(x_t - \hat{x}_{t|t,\tilde{\Theta}_t}\right) + \tilde{C}_t w_t + z_{t+1} \tag{5.51}$$

$$\hat{x}_{t+1|t+1,\tilde{\Theta}_t} = \left(\tilde{A}_t - \tilde{B}_t \tilde{K}_t\right)\hat{x}_{t|t,\tilde{\Theta}} + \tilde{L}_t \tilde{C}_t \tilde{A}_t\left(x_t - \hat{x}_{t|t,\tilde{\Theta}}\right) + \tilde{L}_t \tilde{C}_t w_t + \tilde{L}_t z_{t+1} \tag{5.52}$$

$$y_{t+1,\Theta} = CA\hat{x}_{t|t,\tilde{\Theta}} - CB\tilde{K}_t \hat{x}_{t|t,\tilde{\Theta}} + Cw_t + CA(x_t - \hat{x}_{t|t,\tilde{\Theta}}) + z_{t+1} \tag{5.53}$$

$$\hat{x}_{t+1|t+1,\Theta} = (A-B\tilde{K}_t)\hat{x}_{t|t,\tilde{\Theta}} + LCw_t + LCA(x_t - \hat{x}_{t|t,\tilde{\Theta}_t}) + (I-LC)A(\hat{x}_{t|t,\Theta} - \hat{x}_{t|t,\tilde{\Theta}_t}) + Lz_{t+1}. \tag{5.54}$$

Notice that given $x_{t|t-1}$ and y_t first and fourth terms in (5.54) are deterministic. Using this fact and the definitions given in (5.50)–(5.54), we obtain the following decomposition starting from the Bellman optimality equation for the optimistic system:

$$J_*(\tilde{\Theta}) + R_1 + R_2 - R_{\mathrm{u}} = \left(y_t^\top Q y_t + u_t^\top R u_t\right) + R_3 + R_4 + R_5 + R_6 + R_7 + R_8 + R_9 + R_{10} + R_{11}, \tag{5.55}$$

where

$$
\begin{aligned}
R_1 &= \hat{x}_{t|t,\tilde{\Theta}}^\top \left(\tilde{P}_t - \tilde{C}_t^\top Q \tilde{C}_t\right) \hat{x}_{t|t,\tilde{\Theta}} - \mathbb{E}\left[\hat{x}_{t+1|t+1,\Theta}^\top \left(\tilde{P}_{t+1} - \tilde{C}_{t+1}^\top Q \tilde{C}_{t+1}\right) \hat{x}_{t+1|t+1,\Theta} \middle| \hat{x}_{t|t-1}, y_t, u_t\right], \\
R_u &= \mathbb{E}\left[\hat{x}_{t+1|t+1,\Theta}^\top \left((\tilde{P}_t - \tilde{C}_t^\top Q \tilde{C}_t) - (\tilde{P}_{t+1} - \tilde{C}_{t+1}^\top Q \tilde{C}_{t+1})\right) \hat{x}_{t+1|t+1,\Theta} \middle| \hat{x}_{t|t-1}, y_t, u_t\right], \\
R_2 &= y_t^\top Q y_t - \mathbb{E}\left[y_{t+1,\Theta}^\top Q y_{t+1,\Theta} \middle| \hat{x}_{t|t-1}, y_t, u_t\right], \\
R_3 &= \hat{x}_{t|t,\tilde{\Theta}}^\top (\tilde{A}_t - \tilde{B}_t \tilde{K}_t)^\top \tilde{C}_t^\top Q \tilde{C}_t (\tilde{A}_t - \tilde{B}_t \tilde{K}_t) \hat{x}_{t|t,\tilde{\Theta}} - \hat{x}_{t|t,\tilde{\Theta}}^\top (A - B\tilde{K}_t)^\top C^\top Q C (A - B\tilde{K}_t) \hat{x}_{t|t,\tilde{\Theta}}, \\
R_4 &= \hat{x}_{t|t,\tilde{\Theta}}^\top (\tilde{A}_t - \tilde{B}_t \tilde{K}_t)^\top (\tilde{P}_t - \tilde{C}_t^\top Q \tilde{C}_t)(\tilde{A}_t - \tilde{B}_t \tilde{K}_t) \hat{x}_{t|t,\tilde{\Theta}} \\
&\qquad - \hat{x}_{t|t,\tilde{\Theta}}^\top (A - B\tilde{K}_t)^\top (\tilde{P}_t - \tilde{C}_t^\top Q \tilde{C}_t)(A - B\tilde{K}_t) \hat{x}_{t|t,\tilde{\Theta}}, \\
R_5 &= -2\hat{x}_{t|t,\tilde{\Theta}}^\top (A - B\tilde{K}_t)^\top (\tilde{P}_t - \tilde{C}_t^\top Q \tilde{C}_t)(I - LC) A (\hat{x}_{t|t,\Theta} - \hat{x}_{t|t,\tilde{\Theta}}), \\
R_6 &= -(\hat{x}_{t|t,\Theta} - \hat{x}_{t|t,\tilde{\Theta}})^\top A^\top (I - LC)^\top (\tilde{P}_t - \tilde{C}_t^\top Q \tilde{C}_t)(I - LC) A (\hat{x}_{t|t,\Theta} - \hat{x}_{t|t,\tilde{\Theta}}), \\
R_7 &= \mathbb{E}\left[w_t^\top \tilde{C}_t^\top Q \tilde{C}_t w_t\right] - \mathbb{E}\left[w_t^\top C^\top Q C w_t\right], \\
R_8 &= \mathbb{E}\left[w_t^\top \tilde{C}_t^\top \tilde{L}_t^\top \left(\tilde{P}_t - \tilde{C}_t^\top Q \tilde{C}_t\right) \tilde{L}_t \tilde{C}_t w_t\right] - \mathbb{E}\left[w_t^\top C^\top L^\top \left(\tilde{P}_t - \tilde{C}_t^\top Q \tilde{C}_t\right) L C w_t\right], \\
R_9 &= \mathbb{E}\left[\left(x_t - \hat{x}_{t|t,\tilde{\Theta}}\right)^\top \tilde{A}_t^\top \tilde{C}_t^\top Q \tilde{C}_t \tilde{A}_t \left(x_t - \hat{x}_{t|t,\tilde{\Theta}}\right) \middle| \hat{x}_{t|t-1}, y_t\right] \\
&\qquad - \mathbb{E}\left[\left(x_t - \hat{x}_{t|t,\tilde{\Theta}}\right)^\top A^\top C^\top Q C A \left(x_t - \hat{x}_{t|t,\tilde{\Theta}}\right) \middle| \hat{x}_{t|t-1}, y_t\right], \\
R_{10} &= \mathbb{E}\left[\left(x_t - \hat{x}_{t|t,\tilde{\Theta}}\right)^\top \tilde{A}_t^\top \tilde{C}_t^\top \tilde{L}_t^\top \left(\tilde{P}_t - \tilde{C}_t^\top Q \tilde{C}_t\right) \tilde{L}_t \tilde{C}_t \tilde{A}_t \left(x_t - \hat{x}_{t|t,\tilde{\Theta}}\right) \middle| \hat{x}_{t|t-1}, y_t\right] \\
&\qquad - \mathbb{E}\left[\left(x_t - \hat{x}_{t|t,\tilde{\Theta}}\right)^\top A^\top C^\top L^\top \left(\tilde{P}_t - \tilde{C}_t^\top Q \tilde{C}_t\right) L C A \left(x_t - \hat{x}_{t|t,\tilde{\Theta}}\right) \middle| \hat{x}_{t|t-1}, y_t\right], \\
R_{11} &= 2\mathbb{E}\left[z_{t+1}^\top L^\top (\tilde{P}_t - \tilde{C}_t^\top Q \tilde{C}_t)\left(\tilde{L}_t - L\right) z_{t+1}\right] + \mathbb{E}\left[z_{t+1}^\top (\tilde{L}_t - L)^\top (\tilde{P}_t - \tilde{C}_t^\top Q \tilde{C}_t)\left(\tilde{L}_t - L\right) z_{t+1}\right].
\end{aligned}
$$

Recall the regret definition in (5.12):

$$\textsc{Regret}(T) = \sum\nolimits_{t=0}^{T} \left(y_t^\top Q y_t + u_t^\top R u_t - J_*(\Theta)\right).$$

Due to optimistic parameter selection such that $J_*(\tilde{\Theta}) \leq J_*(\Theta) + T^{-1}$, from (5.55), we get

$$y_t^\top Q y_t + u_t^\top R u_t - J_*(\Theta) \leq R_1 + R_2 - R_{\mathrm{u}} - R_3 - R_4 - R_5 - R_6 - R_7 - R_8 - R_9 - R_{10} - R_{11} + T^{-1}.$$

Therefore, we can bound the regret of LqgOpt as

$$\textsc{Regret}(T) < R_{\text{warm-up}} + \sum_{t=T_w}^{T} R_1 + R_2 - R_{\mathrm{u}} - R_3 - R_4 - R_5 - R_6 - R_7 - R_8 - R_9 - R_{10} - R_{11}, \quad (5.56)$$

where $R_{\text{warm-up}}$ is the regret of the warm-up phase. In the following, we bound each term separately.

5.4.5 Regret Analysis and Proofs of Theorem 5.5 and Corollary 5.5.1

In this section, we provide the regret analysis of LqgOpt that leads to the guarantees in Theorem 5.5 and Corollary 5.5.1. At first, we discuss the regret due to exploring the system with $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ in *warm-up* phase for $1 \leq t \leq T_w$, i.e., $R_{\text{warm-up}}$. Then, we provide the regret of deploying the optimistic controllers in *adaptive control in epochs* phase for $T_w + 1 \leq t \leq T$, i.e., R_u, and R_{1-11}.

Lemma 5.9. *Suppose Assumption 5.1 holds. Given an LQG control system $\Theta = (A, B, C)$, the regret of deploying $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ for $1 \leq t \leq T_w$ is upper bounded as follows with high probability*

$$R_{warm\text{-}up} = c_* T_w + \tilde{O}\left(\sqrt{T_w}\right), \tag{5.57}$$

where c_* is a problem-dependent constant.

This lemma might feel intuitive to many readers. One of the main reasons we provide Lemma 5.9 is the importance and contribution of $\tilde{O}\left(\sqrt{T_w}\right)$ terms in (5.57) to the final regret bound in particular to Corollary 5.5.1. The proof and the precise expressions are in Appendix C.2.3. Before presenting the upper bounds on the regret components in the adaptive control phase, we provide the following lemma on $\|\tilde{\Sigma} - \Sigma\|$ where $\tilde{\Sigma}_i$ is the solution to DARE given in (5.5) for the system $\tilde{\Theta}_i$ in epoch i and Σ is the solution to (5.5) for the underlying system Θ.

Lemma 5.10. *Suppose Assumption 5.1 holds. For $\delta \in (0, 1)$ and long enough warm-up duration T_w, there exists a similarity transformation $\mathbf{S} \in \mathbb{R}^{n \times n}$ such that, with probability at least $1 - 7\delta$,*

$$\|\tilde{\Sigma}_i - \mathbf{S}^{-1}\Sigma\mathbf{S}\| \leq \Delta\Sigma := \frac{\kappa_1^2(8\psi + 4)D^2 + \sigma_z^2(8\kappa_1 + 4)D}{\sigma_z^2(1 - (1 - \gamma_3)^2)} \max\left\{\beta_i^A, \beta_i^C\right\}$$

for all adaptive control epochs, where D is an upper bound on the operator norm of the solution to the DARE (5.5) for Θ, i.e., $\|\Sigma\| \leq D = \sigma_z^2 \gamma_3^{-1} \kappa_3^2 (1 + \kappa_3^2)$, and β_i^A, β_i^C are defined in Theorem 5.4.

The proof is given in Appendix C.2.2, where we deploy a fixed point argument for the solution of (5.5). With this result at the hand, we are ready to bound individual terms in the regret decomposition.

Lemma 5.11. *Under the setting of Theorem 5.5, for $\delta \in (0, 1)$, with probability at least $1 - 5\delta$,*

$$|R_i| = \tilde{O}\left(\frac{T_w \log(1/\delta)}{\sqrt{T_w}} + \frac{2T_w \log(1/\delta)}{\sqrt{2T_w}} + \frac{4T_w \log(1/\delta)}{\sqrt{4T_w}} + \dots\right)$$

for $i = 1, 3 \ldots, 11$*, and* $R_2 = \tilde{O}(\sqrt{T - T_w}\log(1/\delta))$*. Moreover,* LqgOpt *makes at most* $O(\log(T/\delta))$ *policy changes which yields* $|R_u| = O(\log(T))$ *with the same probability.*

The proof of this lemma studies each component individually and can be collected from Appendix H of [161] by combining the doubling update rule of LqgOpt. Combining all the results above we give the proof of Theorem 5.5.

Proof of Theorem 5.5. Using the doubling trick, i.e., Lemma C.3.11, on the results of Lemma 5.11, we have $|R_i| = \tilde{O}(\sqrt{T})$ for all $i = 1, \ldots, 11$, with probability at least $1 - 5\delta$. Therefore, for the regret of adaptive control in epochs phase we have $\textsc{Regret}(T - T_w) = \tilde{O}\left(\sqrt{T}\right)$. Combining this with Lemma 5.9, which states that warm-up period has regret that scales linearly with T_w which depends on the horizon as $O(\log(T))$, we conclude with the advertised result of Theorem 5.5. □

Next, we bound the regret of the adaptive control phase when the underlying system and its optimal controller do not satisfy the PE condition, i.e., Assumption 5.2, and LqgOpt solely relies on the warm-up period.

Lemma 5.12. *Under the setting of Corollary 5.5.1, for* $\delta \in (0, 1)$*, with probability at least* $1 - 5\delta$*,*

$$|R_i| = \tilde{O}\left(\frac{(T - T_w)\log(1/\delta)}{\sqrt{T_w}}\right)$$

for $i = 1, 3 \ldots, 11$*, and* $R_2 = \tilde{O}(\sqrt{T - T_w}\log(1/\delta))$*.*

The proof of this simply uses the same regret decomposition as Lemma 5.11 with the observation that the agent's model parameter estimates have the error of $\tilde{O}(1/\sqrt{T_w})$ during the entire adaptive control phase due to the lack of persistence of excitation.

Proof of Corollary 5.5.1. Combining Lemma 5.12 and 5.9, the first statement follows trivially, since $\frac{T - T_w}{\sqrt{T_w}}$ is the dominating term in the regret of the adaptive control phase. Taking the derivative of the total regret expression with respect to T_w and finding its roots gives the minimizing solution of $T_w = T^{2/3}$. Inserting it to the overall regret expression proves the final statement. □

5.4.6 Extensions to the ARX systems

In this section, we generalize the results for LqgOpt to the ARX systems, i.e., the systems with the dynamics of the form (5.7) with sub-Gaussian e_t with covariance matrix of Σ_E and arbitrary $\bar{A}$ and F. First, for the given quadratic control problem in (5.8) in an ARX system that satisfies Assumption 5.1, we derive the optimal control law, i.e., the analog of (5.9), using the average cost optimality equation. From the first principles [28], the value function of the given ARX system is quadratic, and due to stochasticity we have the following format:

$$V(x, y) = \begin{bmatrix} x \\ y \end{bmatrix}^\top \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \lambda.$$

Using the average cost optimality equation, we can determine the value function for the given system Θ as follows:

$$\begin{bmatrix} x \\ y \end{bmatrix}^\top \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \lambda = \min_u \left\{ y^\top Q y + u^\top R u + \mathbb{E}\left[\begin{bmatrix} Ax + Bu + Fy \\ CAx + CBu + CFy + e \end{bmatrix}^\top \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} Ax + Bu + Fy \\ CAx + CBu + CFy + e \end{bmatrix} \right] \right\}.$$

Expanding all and minimizing for u gives the optimal control of

$$u = -(R + B^\top \mathbf{P} B)^{-1} \left[B^\top \mathbf{P} A x + B^\top \mathbf{P} F y \right],$$

where $\mathbf{P} = P_{11} + P_{12}C + C^\top P_{21} + C^\top P_{22} C$. Inserting the expression for u, we get $\lambda = \mathrm{Tr}(P_{22}E)$ where

$$\begin{aligned} &\begin{bmatrix} x \\ y \end{bmatrix}^\top \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\ &= \begin{bmatrix} x \\ y \end{bmatrix}^\top \begin{bmatrix} A^\top \big(\mathbf{P} - \mathbf{P}B(R + B^\top \mathbf{P} B)^{-1} B^\top \mathbf{P}\big) A & A^\top \big(\mathbf{P} - \mathbf{P}B(R + B^\top \mathbf{P} B)^{-1} B^\top \mathbf{P}\big) F \\ F^\top \big(\mathbf{P} - \mathbf{P}B(R + B^\top \mathbf{P} B)^{-1} B^\top \mathbf{P}\big) A & Q + F^\top \big(\mathbf{P} - \mathbf{P}B(R + B^\top \mathbf{P} B)^{-1} B^\top \mathbf{P}\big) F \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \end{aligned}$$

This must hold for all x and y. Therefore, using the definition of $\mathbf{P}$, we conclude that $\mathbf{P}$ satisfies the DARE

$$\mathbf{P} = C^\top Q C + (A + FC)^\top \mathbf{P} (A + FC) - (A + FC)^\top \mathbf{P} B (R + B^\top \mathbf{P} B)^{-1} B^\top \mathbf{P} (A + FC), \quad (5.58)$$

and the infinite horizon optimal cost of this system is

$$J_*(\Theta) = \mathrm{Tr}\left(\Sigma_E \left(Q + F^\top \left(\mathbf{P} - \mathbf{P} B (R + B^\top \mathbf{P} B)^{-1} B^\top \mathbf{P} \right) F \right) \right). \quad (5.59)$$

Therefore, we can write the optimal control law of ARX systems, π^*, as a linear feedback policy,

$$u_t^* = K_x^* x_t + K_y^* y_t = -(R + B^\top \mathbf{P} B)^{-1} B^\top \mathbf{P} \left(A x_t + F y_t \right), \quad (5.60)$$

where $\mathbf{P}$ is the unique positive semidefinite solution to the discrete-time algebraic Riccati equation given in (5.58). We assume that the systems in the set $\mathcal{S}$ are *stabilizable* such that the optimal controller produces stable closed-loop system dynamics for the state and the output, i.e., $\rho(A + BK_x^*) < 1$ and $\rho(F + BK_y^*) < 1$. In the regret metric for this problem we consider the optimal cost given in (5.59) as the baseline.

LqgOpt for ARX systems: The only algorithmic change required to LqgOpt is the confidence set construction since for an ARX system, the learning agent needs to construct the confidence sets of $C_A(i), C_B(i), C_C(i)$, and $C_F(i)$. Moreover, in the optimistic model selection $\tilde{\Theta}_i$, we consider the sublevel sets of optimal cost with the structure of (5.59). Note that the open-loop PE condition of Section 5.3.2 still holds in ARX systems. In the following, we characterize the closed-loop PE condition for the underlying ARX system with its optimal controller. This will allow us to have PE in the adaptive control phase of LqgOpt, under small enough estimation errors which can be guaranteed with the long enough warm-up duration in ARX systems, as in Lemma 5.5.

PE Condition in the Closed-Loop Setting: After the warm-up phase, for $t \geq T_w$, Algorithm LqgOpt executes the input of $u_t = \tilde{K}_t^x x_t + \tilde{K}_t^y y_t$. Let $f_t = [y_t^\top u_t^\top]^\top$. Using the state-space representation of ARX model, we get

$$f_t = \underbrace{\begin{bmatrix} C \\ \tilde{K}_t^x + \tilde{K}_t^y C \end{bmatrix}}_{\tilde{\mathbf{\Gamma}}_\mathbf{t}} x_t + \underbrace{\begin{bmatrix} I \\ \tilde{K}_t^y \end{bmatrix}}_{\tilde{\mathbf{\Omega}}_\mathbf{t}} e_t.$$

Moreover, $x_t = \underbrace{[A + B\tilde{K}_{t-1}^x + FC + B\tilde{K}_{t-1}^y C]}_{\tilde{\mathbf{\Lambda}}_\mathbf{t-1}} x_{t-1} + \underbrace{[F + B\tilde{K}_{t-1}^y]}_{\tilde{\mathbf{\Xi}}_\mathbf{t-1}} e_{t-1}$. Thus for f_t, we get:

$$f_t = \tilde{\mathbf{\Gamma}}_\mathbf{t}\tilde{\mathbf{\Lambda}}_\mathbf{t-1} x_{t-1} + \tilde{\mathbf{\Gamma}}_\mathbf{t}\tilde{\mathbf{\Xi}}_\mathbf{t-1} e_{t-1} + \tilde{\mathbf{\Omega}}_\mathbf{t} e_t.$$

Rolling back for H time steps, we get the following,

$$f_t = \tilde{\mathbf{\Gamma}}_\mathbf{t}\left(\sum_{i=t-H+1}^{t}\left(\prod_{j=i}^{t-1}\tilde{\mathbf{\Lambda}}_\mathbf{j}\right)\tilde{\mathbf{\Xi}}_\mathbf{i-1} e_{i-1}\right) + \tilde{\mathbf{\Omega}}_\mathbf{t} e_t + \mathbf{r}_\mathbf{t}^\mathbf{c},$$

where $\mathbf{r}_\mathbf{t}^\mathbf{c}$ is the residual vector that represents the effect of e_i for $0 \leq i < t - H$, which are independent. Using this, one can write the full characterization of $\bar{\phi}_t$ as

follows

$$\bar{\phi}_t = \begin{bmatrix} f_{t-1} \\ \vdots \\ f_{t-H} \end{bmatrix} + \begin{bmatrix} \mathbf{r^c_{t-H}} \\ \vdots \\ \mathbf{r^c_{t-H}} \end{bmatrix} = \mathcal{G}^{cl}_t \underbrace{\begin{bmatrix} e_{t-1} \\ e_{t-2} \\ \vdots \\ e_{t-2H-1} \\ e_{t-2H} \end{bmatrix}}_{\mathbb{R}^{2Hm}} + \begin{bmatrix} \mathbf{r^c_{t-1}} \\ \vdots \\ \mathbf{r^c_{t-H}} \end{bmatrix},$$

where

$$\mathcal{G}^{cl}_t = \begin{bmatrix} [\quad \bar{\mathbf{G}}_{\mathbf{t-1}} \quad] & 0_{(m+p)\times m} & 0_{(m+p)\times m} & 0_{(m+p)\times m} & \cdots \\ 0_{(m+p)\times m} & [\quad \bar{\mathbf{G}}_{\mathbf{t-2}} \quad] & 0_{(m+p)\times m} & 0_{(m+p)\times m} & \cdots \\ & & \ddots & & \\ 0_{(m+p)\times m} & 0_{(m+p)\times m} & \cdots & [\quad \bar{\mathbf{G}}_{\mathbf{t-H+1}} \quad] & 0_{(m+p)\times m} \\ 0_{(m+p)\times m} & 0_{(m+p)\times m} & 0_{(m+p)\times m} & \cdots & [\quad \bar{\mathbf{G}}_{\mathbf{t-H}} \quad] \end{bmatrix} \tag{5.61}$$

for

$$\bar{\mathbf{G}}_{\mathbf{t}} = \left[\tilde{\mathbf{\Omega}}_{\mathbf{t}},\ \tilde{\mathbf{\Gamma}}_{\mathbf{t}}\tilde{\mathbf{\Xi}}_{\mathbf{t-1}},\ \tilde{\mathbf{\Gamma}}_{\mathbf{t}}\tilde{\mathbf{\Lambda}}_{\mathbf{t-1}}\tilde{\mathbf{\Xi}}_{\mathbf{t-2}}, \ldots,\ \tilde{\mathbf{\Gamma}}_{\mathbf{t}}\tilde{\mathbf{\Lambda}}_{\mathbf{t-1}}\tilde{\mathbf{\Lambda}}_{\mathbf{t-2}} \ldots \tilde{\mathbf{\Lambda}}_{\mathbf{t+H-1}}\tilde{\mathbf{\Xi}}_{\mathbf{t-H}} \right] \in \mathbb{R}^{(m+p)\times hm}.$$

If the underlying system is known, then the optimal control law for the ARX system could be applied to control the system. In the following, $\mathcal{G}^{cl}$ is the closed-loop mapping of noise process to the covariates $\bar{\phi}$ via optimal policy

$$\mathcal{G}^{cl} = \begin{bmatrix} [\quad \bar{\mathbf{G}} \quad] & 0_{(m+p)\times m} & 0_{(m+p)\times m} & 0_{(m+p)\times m} & \cdots \\ 0_{(m+p)\times m} & [\quad \bar{\mathbf{G}} \quad] & 0_{(m+p)\times m} & 0_{(m+p)\times m} & \cdots \\ & & \ddots & & \\ 0_{(m+p)\times m} & 0_{(m+p)\times m} & \cdots & [\quad \bar{\mathbf{G}} \quad] & 0_{(m+p)\times m} \\ 0_{(m+p)\times m} & 0_{(m+p)\times m} & 0_{(m+p)\times m} & \cdots & [\quad \bar{\mathbf{G}} \quad] \end{bmatrix} \tag{5.62}$$

for

$$\bar{\mathbf{G}} = \left[\mathbf{\Omega},\ \mathbf{\Gamma\Xi},\ \mathbf{\Gamma\Lambda\Xi}, \ldots,\ \mathbf{\Gamma\Lambda}^{H-1}\mathbf{\Xi} \right]$$

where

$$\mathbf{\Omega} = \begin{bmatrix} I \\ K^*_y \end{bmatrix},\ \mathbf{\Gamma} = \begin{bmatrix} C \\ K^*_x + K^*_y C \end{bmatrix},\ \mathbf{\Lambda} = [A + BK^*_x + FC + BK^*_y C],\ \mathbf{\Xi} = [F + BK^*_y].$$

Note that $\bar{\mathbf{G}}$ corresponds to truncated closed-loop noise to covariate Markov operator. Notice that if $\bar{\mathbf{G}}$ is full row rank, following a similar approach with Section 5.3.2, $\mathcal{G}^{cl}$ is also full row rank. Thus, we have the following persistence of excitation condition on the optimal control law for the ARX system:

Assumption 5.3 (PE structure of the underlying ARX system with its optimal control)**.** *H is large enough such that $\bar{\mathbf{G}}$ formed via optimal control policy of the given ARX system is full row rank.*

Therefore, under Assumption 5.3, we have a lower bound of the smallest singular value of the $H-$length truncated closed-loop noise evolution parameter, $\sigma_{\min}(\mathcal{G}^{cl}) > \sigma_c > 0$. Using the same definitions of T_c and G_r given in Section 5.4.2, Lemma 5.5 guarantees that under Assumption 5.3, after T_c time steps of adaptive control period of LqgOpt, with probability $1-3\delta$, the following holds for the remainder of adaptive control period,

$$\sigma_{\min}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top\right) \geq t\frac{\sigma_c^2\sigma_e^2}{16}. \tag{5.63}$$

After establishing, the closed-loop PE condition in ARX systems, we are ready to state the regret guarantee of LqgOpt in ARX systems. Note that for a long enough warm-up period, we have the stability of the closed-loop system dynamics via LqgOpt, hence bounded x_t and y_t throughout the entire horizon, i.e., Lemma 5.6.

Theorem 5.6 (Regret of LqgOpt in ARX systems with closed-loop PE condition)**.** *Let $\delta \in (0,1)$. For the unknown ARX system Θ and long enough warm-up phase, if Assumption 5.3 holds such that the optimal controller for Θ persistently excites the underlying system, then LqgOpt attains regret of*

$$\textsc{Regret}(T) = \tilde{O}\left(\sqrt{T}\right), \tag{5.64}$$

with probability at least $1-5\delta$ after T time steps.

Corollary 5.6.1 (Regret of LqgOpt in ARX Sytems without closed-loop PE condition)**.** *For the system given in Theorem 5.6 with the choices of H and T_w, if the underlying system is not persistently excited with its optimal policy, LqgOpt incurs the following regret with high probability, $\textsc{Regret}(T) = \tilde{O}\left(T_w + \frac{T-T_w}{\sqrt{T_w}}\right)$. Therefore, the optimal regret upper bound of this setting is obtained with a warm-up duration of $T_w = O(T^{2/3})$, which gives the regret of $\textsc{Regret}(T) = \tilde{O}\left(T^{2/3}\right)$ for LqgOpt.*

The proof of these results follows similarly to Theorem 5.5 and Corollary 5.5.1. In particular, we use the Bellman optimality equation for the average cost-per-step ARX LQ control problem and consider the system evolution of the optimistic system using the optimistic controller in parallel with the true system evolution under the optimistic controller such that they share the same process noise. The following gives the Bellman optimality equation for the ARX linear quadratic control problem.

Lemma 5.13 (Bellman Optimality Equation for ARX System)**.** *Given state $x_t \in \mathbb{R}^n$ and an observation $y_t \in \mathbb{R}^m$ pair at time t, Bellman optimality equation of average cost per stage control of the system $\Theta = (A, B, C, F)$ with regulating parameters Q and R is*

$$J_*(\Theta) + \bar{x}_t^\top \left(\mathbf{P} - \mathbf{P}B(R + B^\top \mathbf{P}B)^{-1}B^\top \mathbf{P}\right)\bar{x}_t + y_t^\top Q y_t$$
$$= y_t^\top Q y_t + u_t^\top R u_t + \mathbb{E}\left[\bar{x}_{t+1}^\top (\mathbf{P} - \mathbf{P}B(R + B^\top \mathbf{P}B)^{-1}B^\top \mathbf{P})\bar{x}_{t+1} + y_{t+1}^\top Q y_{t+1}\right] \quad (5.65)$$

for $\bar{x}_t = Ax_t + Fy_t$ and $\bar{x}_{t+1} = Ax_{t+1} + Fy_{t+1}$.

The proof follows similarly to the proof of Lemma 5.8 and can be found in [163]. Following the decomposition steps given in Section 5.4.4, we get

$$J_*(\tilde{\Theta}_t) + R_1 + R_2 - R_u = \left(y_t^\top Q y_t + u_t^\top R u_t\right) + R_3 + R_4 + R_5,$$

where

$$R_1 = (\tilde{A}_t x_t + \tilde{F}_t y_t)^\top \tilde{\tilde{\mathbf{P}}}_t (\tilde{A}_t x_t + \tilde{F}_t y_t) - \mathbb{E}\left[(Ax_{t+1,\Theta} + Fy_{t+1,\Theta})^\top \tilde{\tilde{\mathbf{P}}}_{t+1}(Ax_{t+1,\Theta} + Fy_{t+1,\Theta}) | x_t, y_t, u_t\right]$$
$$R_u = \mathbb{E}\left[(Ax_{t+1,\Theta} + Fy_{t+1,\Theta})^\top \left(\tilde{\tilde{\mathbf{P}}}_t - \tilde{\tilde{\mathbf{P}}}_{t+1}\right)(Ax_{t+1,\Theta} + Fy_{t+1,\Theta}) | x_t, y_t, u_t\right],$$
$$R_2 = y_t^\top Q y_t - \mathbb{E}\left[y_{t+1,\Theta}^\top Q y_{t+1,\Theta} \middle| x_t, y_t, u_t\right]$$
$$R_3 = \bar{x}_{t,\tilde{\Theta}_t}^\top (I - \tilde{B}_t \tilde{M}_t)^\top \tilde{C}_t^\top Q \tilde{C}_t (I - \tilde{B}_t \tilde{M}_t)\bar{x}_{t,\tilde{\Theta}} - x_{t+1,\Theta}^\top C^\top Q C x_{t+1,\Theta}$$
$$R_4 = \bar{x}_{t,\tilde{\Theta}_t}^\top (I - \tilde{B}_t \tilde{M}_t)^\top (\tilde{A}_t + \tilde{F}_t \tilde{C}_t)^\top \tilde{\tilde{\mathbf{P}}}_t (\tilde{A}_t + \tilde{F}_t \tilde{C}_t)(I - \tilde{B}_t \tilde{M}_t)\bar{x}_{t,\tilde{\Theta}_t} - x_{t+1,\Theta}^\top (A + FC)^\top \tilde{\tilde{\mathbf{P}}}_t (A + FC) x_{t+1,\Theta}$$
$$R_5 = 2\mathbb{E}\left[e_{t+1}^\top F^\top \tilde{\tilde{\mathbf{P}}}_t (\tilde{F}_t - F) e_{t+1}^\top\right] + \mathbb{E}\left[e_{t+1}^\top (\tilde{F}_t - F)^\top \tilde{\tilde{\mathbf{P}}}_t (\tilde{F}_t - F) e_{t+1}^\top\right],$$

where $M = (R + B^\top \mathbf{P}B)^{-1}B^\top \mathbf{P}$, $\tilde{M}_t = (R + \tilde{B}_t^\top \tilde{\mathbf{P}}_t \tilde{B}_t)^{-1}\tilde{B}_t^\top \tilde{\mathbf{P}}_t$, and $\tilde{\tilde{\mathbf{P}}}_t = \tilde{P}_t - \tilde{P}_t \tilde{B}_t (R + \tilde{B}_t^\top \tilde{P}_t \tilde{B}_t)^{-1}\tilde{B}_t^\top \tilde{P}_t$. Using the fact that $\tilde{\Theta}_t$ is optimistically chosen, we have

$$\textsc{Regret}(T) < R_{\text{warm-up}} + \sum_{t=T_w}^{T} R_1 + R_2 - R_{\text{u}} - R_3 - R_4 - R_5,$$

for the regret of LqgOpt in the given ARX system. Following the proof steps of Lemma 5.11, we have

$$R_i = \tilde{O}\left(\frac{T_w}{\sqrt{T_w}} + \frac{2T_w}{\sqrt{2T_w}} + \frac{4T_w}{\sqrt{4T_w}} + \dots\right) = \tilde{O}\left(\sqrt{T}\right),$$

for $i = 1, 3, 4, 5$, $R_2 = \tilde{O}\left(\sqrt{T - T_w}\right)$, and $|R_u| = O(\log(T))$, which gives the desired result of Theorem 5.6. Corollary 5.6.1 is proved similarly, i.e., through the steps of Lemma 5.12.

5.5 Thompson Sampling-Based Adaptive Control

Even though in the prior section we showed that LqgOpt achieves $\tilde{O}(\sqrt{T})$ regret in learning and control of LQG control systems, the adaptive control procedure of LqgOpt requires solving a non-convex optimization problem to find the optimistic controller. Unfortunately, finding the optimistic parameters among the plausible models is an NP-hard problem in general, and requires computational heuristics for large-scale dynamical systems [9]. This computational inefficiency severely limits the practicality of the optimistic controller design approach.

As discussed in Chapter 2 and 3, Thompson Sampling (TS) is a promising alternative to overcome the computational burden of finding the optimistic policy [252]. In TS, the agent samples at random from the posterior distribution of models computed from a given prior distribution and the observed data and executes the corresponding LQG-optimal policy for this model [263]. This approach replaces the cumbersome optimization in optimism with straightforward sampling and results in a polynomial-time method. Motivated by the Thompson Sampling-based Adaptive Control (TSAC) algorithm in Chapter 3 for LQRs, we theoretically and empirically study TS in adaptive control of unknown partially observable LQG control systems.

We propose an efficient TS-based adaptive control algorithm, **T**hompson **S**ampling under **P**artial **O**bservability, TSPO, for learning and controlling unknown LQG control systems. We show that TSPO attains $O(\sqrt{T})$ regret after T time steps, which makes TSPO the first efficient adaptive control algorithm to achieve this regret rate for adaptive control of partially observable LQ control systems with convex cost (Table 5.1). Furthermore, we empirically study the performance of TSPO in the measurement-feedback control of a $2^{\text{nd}}-$order SISO system. We show that TSPO effectively explores the model dynamics and achieves competitive regret performance in a computationally efficient way.

TSPO starts with a short warm-up period to gather data to generate an initial model estimate. It then interacts with the system in epochs where it uses a fixed controller throughout each epoch. At the beginning of each epoch, TSPO uses the closed-loop system identification method introduced in Section 5.3 and estimates the underlying model parameters along with confidence intervals. Using these estimates and associated uncertainties, TSPO constructs a posterior distribution on the model parameters and randomly samples a model from it. Throughout the epoch, it uses the optimal LQG control policy for this sampled model. Similar to LqgOpt, TSPO uses epochs with doubling length, and adaptively improves the model estimates and the

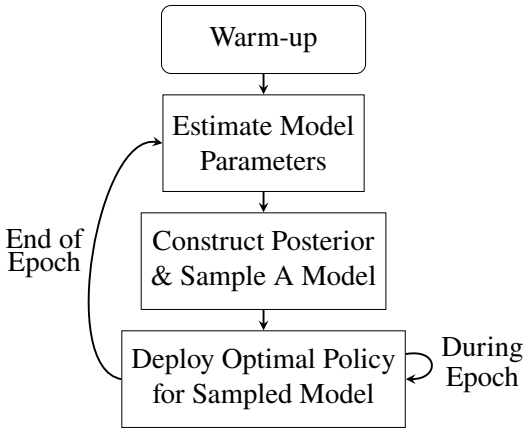

Figure 5.1: TSPO Framework.

controllers. The outline of TSPO is given in Figure 5.1.

Conceptually, TSPO is the Thompson Sampling extension of LqgOpt and builds upon its analysis. Therefore, the algorithmic methodology may not seem very surprising. What is surprising is that the simple TSPO also achieves $O(\sqrt{T})$ regret. The main technical challenge of this work is to establish this fact. To do so, we first show that the regret of a fixed TS policy scales linearly over time with respect to the estimation error in the model parameters. Further, we prove that TS policies maintain stable system dynamics and bounded inputs/outputs provided a long enough warm-up duration. Finally, we show that model the estimates and the TS samples jointly concentrate around the true model parameters over time. Combining these results with the logarithmic policy updates of TSPO, we prove that the regret of TSPO is $O(\sqrt{T})$.

5.5.1 TSPO

In this section, we present our proposed algorithm TSPO, the first computationally efficient and regret optimal RL algorithm for partially observable linear-quadratic control systems with convex instantaneous cost. TSPO is provided in Algorithm 12. It consists of two phases: (i) warm-up period for pure exploration, (ii) adaptive control using TS.

Warm-up: In the early stages, TSPO excites the system by injecting i.i.d. isotropic Gaussian noise $u_t \sim \mathcal{N}(0, \sigma_u^2)$ for a duration of $T_w \geq 0$ and collects samples of observed output and control input, $\mathcal{D}_0 = \{(y_t, u_t)\}_{t=0}^{T_w}$. By exciting the system with i.i.d. inputs, TSPO explores the system effectively and generates a reliable initial estimate of the underlying model using the data collected. The warm-up duration, T_w, is set to meet a desired estimation accuracy so that any policy designed from

Algorithm 12 TSPO

Input: (n, m, d), (Q, R), T_w, H, $\delta > 0$, $\lambda > 0$, $\psi > 0$
—— Warm-Up ——
for $t = 0, 1, \ldots, T_w$ **do**
  Deploy $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ and store $\mathcal{D}_0 = \{y_t, u_t\}_{t=1}^{T_w}$
—— Adaptive Control ——
for $i = 1, \ldots$ **do**
  Calculate $\widehat{\mathcal{G}}_{\mathbf{yu}}^i$ using $\mathcal{D}_i = \{y_t, u_t\}_{t=1}^{2^{i-1}T_w}$ via (5.21)
  Sample $\widetilde{\mathcal{G}}_{\mathbf{yu}}^i \leftarrow \mathcal{R}_{\mathcal{S}}(\widehat{\mathcal{G}}_{\mathbf{yu}}^i + \beta_i V_i^{-\frac{1}{2}} \Xi)$, $[\Xi]_{ij} \sim \mathcal{N}(0, 1)$
  $\tilde{\Theta}_i \leftarrow$ SysId $(\widetilde{\mathcal{G}}_{\mathbf{yu}}^i, H, n)$
  for $t = 2^{i-1}T_w, \ldots, 2^i T_w - 1$ **do**
    Execute the optimal controller for $\tilde{\Theta}_i$

the confidence set is guaranteed to stabilize and persistently excite the underlying system, following the guarantees derived in Lemma 5.6 and Lemma 5.5, respectively.

Adaptive Control: After guaranteeing the design of stabilizing and persistently exciting policies during the warm-up phase, TSPO proceeds to the adaptive control phase. In this phase, TSPO cycles through epochs of fixed-policy control with doubling duration. At the beginning of each epoch, TSPO updates its policy based on input-output data gathered up to that time. The policy design involves three steps.

In the first step, TSPO deploys regularized least squares (RLS) subroutine of (5.21) to perform a closed-loop model estimation from the collected input-output data. In the second step, TSPO calls subroutine of Thompson Sampling, similar to TSAC introduced in Section 3.3, to further explore the unknown system by sampling a random model from a distribution incorporating the estimated model and the associated uncertainty in the estimation. Given the estimate $\widehat{\mathcal{G}}_{\mathbf{yu}}^i$ and the design matrix V_t at time $t \geq T_w$, the TS samples a perturbed truncated ARX model $\widetilde{\mathcal{G}}_{\mathbf{yu}}^i$ as follows

$$\widetilde{\mathcal{G}}_{\mathbf{yu}}^i = \mathcal{R}_{\mathcal{S}}(\widehat{\mathcal{G}}_{\mathbf{yu}}^i + \beta_t V_t^{-\frac{1}{2}} \Xi), \tag{5.66}$$

where $\mathcal{R}_{\mathcal{S}}$ denotes the rejection sampling operator associated with the set $\mathcal{S}$ given in Assumption 5.1, β_t is the confidence ellipsoid bound in Theorem 5.3, and $\Xi \in \mathbb{R}^{m \times (m+d)H}$ is the random perturbation matrix with iid standard normal entries, $[\Xi]_{ij} \sim \mathcal{N}(0, 1)$. The perturbation $\beta_t V_t^{-\frac{1}{2}} \Xi$ randomizes the RLS estimate coherently with the uncertainty conveyed by the design matrix. The rejection sampling operator $\mathcal{R}_{\mathcal{S}}$ keeps sampling independent random perturbations Ξ until the perturbed model $\widehat{\mathcal{G}}_{\mathbf{yu}}^i + \beta_t V_t^{-\frac{1}{2}} \Xi$ lies in set $\mathcal{S}$. The following lemma gives the confidence set for the

sampled $\widetilde{\mathcal{G}}^i_{\mathbf{yu}}$.

Lemma 5.14 (TS confidence set)**.** *For all $t \geq H$, the sampled ARX model $\widetilde{\mathcal{G}}^i_{\mathbf{yu}}$ lies in the set C_t defined as*

$$C_t := \left\{ \mathcal{G}_{\mathbf{yu}} \,\middle|\, \operatorname{tr}((\widehat{\mathcal{G}}^i_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}) V_t (\widehat{\mathcal{G}}^i_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}})^\top) \leq \nu_t^2 \right\}, \tag{5.67}$$

with probability at least $1 - \delta$ where

$$\nu_t := \beta_t m \sqrt{2(m+d)H \log\left(2m(m+d)HT\delta^{-1}\right)}. \tag{5.68}$$

Proof. We bound the probability of belonging to C_t as

$$\begin{aligned}
\mathbb{P}(\forall t \leq T, \widetilde{\mathcal{G}}^i_{\mathbf{yu}} \in C_t) &= 1 - \mathbb{P}(\exists t \leq T, \widetilde{\mathcal{G}}^i_{\mathbf{yu}} \notin C_t) && (5.69)\\
&\geq 1 - \sum\nolimits_{i=0}^{T} \mathbb{P}(\widetilde{\mathcal{G}}^i_{\mathbf{yu}} \notin C_t) && (5.70)\\
&\geq 1 - \sum\nolimits_{i=0}^{T} \mathbb{P}(\|\Xi\|_F \geq \nu_t/\beta_t) && (5.71)\\
&\geq 1 - \delta, && (5.72)
\end{aligned}$$

where (5.70) is due to union bound, (5.71) is due to rejection sampling, and (5.72) is due to Gaussian norm bound. □

In the last step of policy design, TSPO deploys subroutine SysId introduced in Section 5.3 to obtain a balanced state-space realization from the sampled truncated ARX matrix. By taking in $\widetilde{\mathcal{G}}^i_{\mathbf{yu}}$, SysId recovers model parameters $\tilde{\Theta}_t := (\tilde{A}_t, \tilde{B}_t, \tilde{C}_t)$. The propagation of error from the truncated ARX model to the state-space realization designed by SysId is linear as shown in Theorem 5.4.

In the rest of the epoch, TSPO deploys the optimal control policy of the sampled model $\tilde{\Theta}_t$ given as $u_t = -\tilde{K}_t \hat{x}_{t|t,\tilde{\Theta}}$ where $\tilde{K}_t$ is the optimal feedback matrix of $\tilde{\Theta}_t$ and $\hat{x}_{t|t,\tilde{\Theta}}$ is the MMSE estimate of the state assuming system $\tilde{\Theta}_t$. Repeating this, TSPO keeps collecting samples during each epoch and uses the gathered data for refined model estimation, uncertainty quantification, and uncertainty-informed model sampling to further improve controller design in the next epoch. Due to reliable model estimation from the warm-up period, the controller designed right after the warm-up and all subsequently designed controllers stabilize and persistently excite the underlying model (Lemma 5.6 and Lemma 5.5)

5.5.2 Algorithmic Guarantees

In this section, we derive the regret guarantees of TSPO. We use asymptotic notation and hide problem-dependent constants to streamline the exposition as we are mainly interested in the regret rate with respect to the horizon, T. We also note that all the constants in this manuscript, where some are omitted to ease the presentation, have polynomial dependence in the problem-dependent constants. As shown in Lemma 5.4, the underlying LQG control system is persistently excited, i.e., $\sigma_{\min}(V_{T_w}) = \Omega(T_w)$ during the warm-up period by injection of Gaussian input, which yields the estimation error given in (5.23), i.e., $\|\widehat{\mathcal{G}}^1_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\|_F = \tilde{O}(1/\sqrt{T_w})$. Similarly, for the perturbation error in TS, we have $\|\widetilde{\mathcal{G}}^1_{\mathbf{yu}} - \widehat{\mathcal{G}}^1_{\mathbf{yu}}\|_F = \tilde{O}(1/\sqrt{T_w})$. Combining these gives $\|\widetilde{\mathcal{G}}^1_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\|_F = \tilde{O}(1/\sqrt{T_w})$, which also translates to the same behaving bounds on $\|\tilde{A}_t - \mathbf{T}^\top \bar{A}\mathbf{T}\|$, $\|\tilde{B}_t - \mathbf{T}^\top \bar{B}\|$, and $\|\tilde{C}_t - \bar{C}\mathbf{T}\|$ from Theorem 5.4, where $\mathbf{T}$ is a unitary matrix and $(\bar{A}, \bar{B}, \bar{C})$ are the model parameters obtained by SysId using true Markov parameters, $\mathcal{G}_{\mathbf{yu}}$. Therefore, from the appropriate selection of T_w, i.e., long enough warm-up period, we can use Lemma 5.6 and Lemma 5.5 for the sampled model parameters $(\bar{A}, \bar{B}, \bar{C})$ to get closed-loop stability and if Assumption 5.2 holds closed-loop persistence of excitation for TSPO. In particular, with closed-loop PE condition and the doubling epoch duration, for epoch i, we have $\|\widetilde{\mathcal{G}}^i_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\|_F = \tilde{O}(1/\sqrt{2^{i-1}T_w})$.

The following meta-theorem provides an upper bound on the regret of controlling a system Θ by deploying the optimal policy of another system $\tilde{\Theta}_t$ for a fixed period of time. This result shows that inaccuracies due to model mismatch are propagated linearly in regret with linear-time growth. By controlling the model mismatch error in each fixed-policy epoch, we can reduce the regret to a desired level.

Theorem 5.7 (Regret of Model Mismatch)**.** *Suppose that the LQG control system $\Theta \in \mathcal{S}$, whose Markov parameters are $\mathcal{G}_{\mathbf{yu}}$, is controlled by the optimal policy of a model $\tilde{\Theta} \in \mathcal{S}$, which is obtained as an output of SysId for $\widetilde{\mathcal{G}}_{\mathbf{yu}}$. For $\delta \in (0,1)$, the regret incurred due to model mismatch after $\tau \geq 0$ steps is bounded as*

$$R_{\tilde{\Theta}_t}(\tau) \leq \tilde{O}\left(\tau \|\widetilde{\mathcal{G}}^i_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\|_F\right), \tag{5.73}$$

with probability at least $1-\delta$, whenever the model mismatch error satisfies the conditions of Lemma 5.6 and Lemma 5.5, i.e., after a long enough warm-up period.

Proof. We split the regret as follows:

$$R_{\tilde{\Theta}_t}(\tau) = \sum_{t=0}^{\tau} (c_t - J_*(\tilde{\Theta})) + (J_*(\tilde{\Theta}) - J_*(\Theta)), \tag{5.74}$$

where $J_*(\tilde{\Theta})$ is the optimal average expected cost of LQG control system $\tilde{\Theta}$. Note that the dynamical variables are all bounded by Lemma 5.6 due to closed-loop stability achieved after the warm-up period. Recall the Bellman optimality equation-based decomposition given in (5.55), and consider it for the sampled $\tilde{\Theta}$. Via Lemma 5.11, we can bound the first term as $\tilde{O}\left(\tau\|\widetilde{\mathcal{G}}^i_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\|_F\right)$. The second regret term in the analysis of LqgOpt is trivially zero due to optimistic model selection. However, in TSPO, we sample a model parameter and the sampled parameter is not necessarily optimistic. In fact in TSAC for LQRs, we show that Thompson Sampling selects optimistic parameters with only a non-zero probability. Instead, for the second term, we consider $\delta\Theta := \tilde{\Theta} - \Theta$, i.e., the difference between the models. Without loss of generality, due to similarity transformations, we can argue

$$\epsilon := \max(\|\delta A\|_F, \|\delta C\|_F, \|\delta C\|_F) = \max\left\{\|\tilde{A}_t - \mathbf{T}^\top \bar{A}\mathbf{T}\|, \|\tilde{B}_t - \mathbf{T}^\top \bar{B}\|, \|\tilde{C}_t - \bar{C}\mathbf{T}\|\right\},$$

where $\mathbf{T}$ is a unitary matrix and $(\bar{A}, \bar{B}, \bar{C})$ are the model parameters obtained by SysId using true Markov parameters, $\mathcal{G}_{\mathbf{yu}}$. Recall that the optimal average expected cost function, $J_*(\Theta') = \mathrm{Tr}\left((Q + L'^\top P' L' - L'^\top C'^\top Q C' L')\left(C'\Sigma' C'^\top + \sigma_z^2 I\right)\right)$, for $\Theta' = (A', B', C')$. This function is a smooth function of its parameters, Θ' within the highly non-convex domain $\mathcal{S}$. In order to obtain an error bound on the difference $J_*(\tilde{\Theta}) - J_*(\Theta)$, we can use linearized Taylor expansion in the close vicinity of Θ. In other words, there exists a problem dependent constant $\epsilon_r > 0$ such that for $\epsilon \le \epsilon_r$, we have that

$$\begin{aligned} J_*(\tilde{\Theta}) - J_*(\Theta) &= \nabla_{\Theta'} J(\Theta') \bullet \delta\Theta \\ &\le \max(\|\nabla_A J(\Theta')\|, \|\nabla_B J(\Theta')\|, \|\nabla_C J(\Theta)\|)\epsilon := \Gamma_{\mathcal{S}}\epsilon, \end{aligned} \tag{5.75}$$

where $\Theta_1 \bullet \Theta_2 := \mathrm{Tr}(A_1 A_2^\top) + \mathrm{Tr}(B_1 B_2^\top) + \mathrm{Tr}(C_1 C_2^\top)$ is the Euclidean inner product and $\Theta' = \Theta + t\delta\Theta$ for $t \in [0, 1]$. Taking the supremum of the last inequality over all $\Theta \in \mathcal{S}$ and noting that $\nabla_{\Theta'} J(\Theta')$ is a continuous function over the compact set $\mathcal{S}$, we obtain the error bound $J_*(\tilde{\Theta}) - J_*(\Theta) \le \Gamma_{\mathcal{S}}\epsilon$ where $\Gamma_{\mathcal{S}}$ is the maximum norm of $\nabla_{\Theta'} J(\Theta')$ attained in $\mathcal{S}$. Substituting this result into the regret decomposition yields the desired regret bound. □

With these results in hand, we give the upper bound on the overall regret of TSPO.

Theorem 5.8 (Regret of TSPO)**.** *Suppose Assumptions 5.1 and 5.2 hold such that the underlying system satisfies the PE condition with its optimal policy, i.e.,* $\sigma_{\min}(\mathcal{G}^{cl}) > \sigma_c > 0$*. Fixing a horizon* $T > 0$*, let* $H = \max(2n + 1, \Omega(\log T))$ *and*

$$T_w = poly\left(H, \sigma_o, \sigma_c, \kappa_1, \kappa_2, \kappa_3, \frac{1}{1-\gamma_1}, \frac{1}{1-\gamma_2}, \frac{1}{1-\gamma_3}, \psi, m, n, p\right).$$

The regret incurred by TSPO up to horizon T is bounded with high probability as

$$\textsc{Regret}(T) = \tilde{O}(\sqrt{T}). \tag{5.76}$$

Proof. We split the overall regret into individual regrets incurred during the warm-up period and each of the epochs in the adaptive control period as

$$\textsc{Regret}(T) = R_{\text{warm-up}} + \sum_{i=0}^{i_T} R_{\tilde{\Theta}_i}(\tau_{i+1} - \tau_i), \tag{5.77}$$

where $R_{\tilde{\Theta}_i}$ is the regret incurred during epoch i and τ_i denotes the start time of epoch i. From Lemma 5.9, we have that $R_{\text{warm-up}} = \tilde{O}(T_w)$. From Theorem 5.7, we can bound each regret term as

$$R_{\tilde{\Theta}_i}(\tau_{i+1} - \tau_i) \leq O((\tau_{i+1} - \tau_i)\|\widetilde{\mathcal{G}}^i_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\|_F). \tag{5.78}$$

Noting that due to the closed-loop PE condition and the doubling epoch duration, for epoch i, we have $\|\widetilde{\mathcal{G}}^i_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\|_F = \tilde{O}(1/\sqrt{2^{i-1}T_w})$ and $\tau_{i+1} - \tau_i = 2^{i-1}T_w$, we get

$$R_{\tilde{\Theta}_i}(\tau_{i+1} - \tau_i) = O\left(\sqrt{2^{i-1}T_w}\right). \tag{5.79}$$

Summing all these terms for $i_T = \lfloor \log(T/T_w) \rfloor$, we obtain $\textsc{Regret}(T) = O(\sqrt{T})$. □

For the systems which do not satisfy Assumption 5.2, i.e., whose optimal policy does not satisfy the PE condition, we have the following regret bound.

Corollary 5.8.1 (Regret of TSPO without the closed-loop PE condition)**.** *For the system given in Theorem 5.8 with the choices of H and T_w, if the underlying system is not persistently excited with its optimal policy, TSPO incurs the following regret with high probability, $\textsc{Regret}(T) = \tilde{O}\left(T_w + \frac{T - T_w}{\sqrt{T_w}}\right)$. Therefore, the optimal regret upper bound of this setting is obtained with a warm-up duration of $T_w = O(T^{2/3})$, which gives the regret of $\textsc{Regret}(T) = \tilde{O}\left(T^{2/3}\right)$ for TSPO.*

Proof. Similar to the proof of Theorem 5.8, TSPO incurs $O(T_w)$ regret during warm-up. Since the system is not guaranteed to be persistently excited, the best error bound for model mismatch error is attained right after warm-up. In other words, $\|\widetilde{\mathcal{G}}^i_{\mathbf{yu}} - \mathcal{G}_{\mathbf{yu}}\|_F = \tilde{O}(1/\sqrt{T_w})$ for all epochs. By substituting this error result in the regret decomposition by invoking Theorem 5.7, the desired bound is obtained. Substituting $T_w = O(T^{2/3})$ yields the specified bound. □

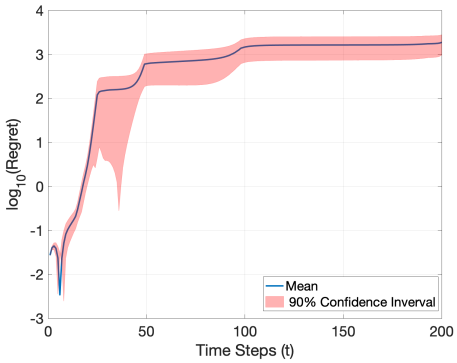

Figure 5.2: Regret Performance of TSPO.

These results show that the regret guarantees obtained via optimism in Section 5.4 for LqgOpt, also hold adaptive control via Thompson Sampling for TSPO. Note that in our presentation we hide the constants. Intuitively, due to the sampling nature, the constants of TSPO are larger than that of LqgOpt, e.g., $\Gamma_{\mathcal{S}}$ for all $\Theta' \in \mathcal{S}$ in (5.75) can be arbitrarily big, whereas in LqgOpt $J_*(\tilde{\Theta}) - J_*(\Theta) \leq 0$ by design. Thus, if the optimistic model selection can be achieved efficiently, LqgOpt comes with a tighter regret upper bound. However, in certain cases, the nonconvex optimization problem to find the optimistic parameters in the LQG control setting yields an NP-hard problem. In such situations, TSPO provides a more effective and efficient solution to adaptive control.

5.5.3 Numerical Simulations

In this section, we evaluate the performance of TSPO in a simulated adaptive measurement-feedback control task. In the simulations, we used state-space parameters given as

$$A = \begin{bmatrix} 0.9 & 0 \\ 0 & 0.7 \end{bmatrix}, \; B = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, C = \begin{bmatrix} 2 & 1 \end{bmatrix}, \tag{5.80}$$

with $Q = R = I$ and isotropic Gaussian process and measurement noise with standard deviations as $\sigma_w = \sigma_v = 0.05$. We set the hyperparameters of TSPO as follows: ARX model truncation length $H = 10$, warm-up period $T_w = 12$, Gaussian excitation covariance $\sigma_u = 0.01$, RLS regularization parameter $\lambda = 0.01$, and $\delta = 0.05$.

We perform 100 independent runs for 200 time-steps for TSPO and report their average and 90% confidence interval. The results are presented in Figure 5.2. The simulation results demonstrate that the regret over time almost stabilizes for the given system and the growth is sub-quadratic, matching the theoretical findings.

5.6 Online Gradient Descent-Based Adaptive Control

In this section, we consider another generalization of the adaptive control problem in LQG control systems which has been the main focus in Sections 5.4 and 5.5. In particular, we consider the system given in (5.1) with general strongly convex cost functions, which can possibly be chosen adversarially at each time step. For this general model, we consider the regret minimization problem against the best controller in hindsight from a given set of controllers, known as policy regret [17].

Leveraging the closed-loop model estimation method introduced in Section 5.3, we propose **Adapt**ive Control **On**line Learning algorithm (AdaptOn) that *adaptively* learns the model dynamics and efficiently uses the model estimates to continuously optimize the controller and reduce the cumulative cost. Similar to prior algorithms in this setting, AdaptOn operates in growing size epochs and, at the beginning of each epoch, estimates the model parameters using our novel model estimation method. However, different from the prior algorithms such as LqgOpt and TSPO, during each epoch, AdaptOn follows an online learning procedure. It utilizes a convex policy reparameterization of linear dynamic (feedback) controllers and the estimated model dynamics to construct counterfactual loss functions. AdaptOn then deploys online gradient descent on these counterfactual loss functions to gradually optimize the controller. This additional optimization within the given set of convex parameterized controllers provides continuous improvement during the adaptive control epochs. We show that as the model estimates improve, the gradient updates become more accurate, resulting in improved controllers.

We show that AdaptOn attains a regret upper bound of polylog(T) after T time steps of agent-environment interaction when the cost functions are strongly convex and the given set of controllers satisfies the PE condition. To the best of our knowledge, this is the first logarithmic regret bound for partially observable linear dynamical systems with unknown dynamics which include the canonical LQG setting studied in this chapter. The presented regret bound improves $\tilde{O}(\sqrt{T})$ regret of LqgOpt and TSPO, as well as prior works in stochastic partially observable linear dynamical systems [191, 245] with the help of novel estimation method which allows updating model estimates during control.

We also show that if the cost function is convex, e.g., the convex quadratic function studied in prior sections, then AdaptOn attains $\tilde{O}(\sqrt{T})$, which matches the prior results in this chapter, i.e., Theorems 5.5 and 5.8. Furthermore, we show that if the PE condition does not hold under the closed-loop setting then AdaptOn

attains $\tilde{O}(\sqrt{T})$ regret for strongly convex cost function and $\tilde{O}(T^{2/3})$ for convex cost function, which also matches the regret upper bound presented in Corollaries 5.5.1 and 5.8.1. Finally, we prove that the results above also extend to the more general setting of ARX systems with sub-Gaussian perturbations and general (strongly) convex cost functions. Along the way, we relax several restrictive assumptions on the system dynamics such as controllability and observability to stabilizability and detectability, which are the necessary conditions to have meaningful learning and control problems.

5.6.1 Preliminaries

Partially Observable LTI System: We consider the unknown discrete-time linear time-invariant system Θ introduced in 5.1. At each time step t, the system is at state x_t and the agent observes y_t, i.e., an imperfect state information. Then, the agent applies a control input u_t, observes the loss function ℓ_t, pays the cost of $c_t = \ell_t(y_t, u_t)$, and the system evolves to a new x_{t+1} at time step $t + 1$. Let $(\mathcal{F}_t; t \geq 0)$ be the corresponding filtration. For any t, conditioned on $\mathcal{F}_{t-1}$, w_t and z_t are $\mathcal{N}(0, \sigma_w^2 I)$ and $\mathcal{N}(0, \sigma_z^2 I)$ respectively. In this section, in contrast to the standard assumptions on the process and measurement noises in Sections 5.4 and 5.5 such that both σ_w^2 and σ_z^2 are known, we only assume the knowledge of their upper and lower bounds, i.e., $\overline{\sigma}_w^2, \underline{\sigma}_w^2, \overline{\sigma}_z^2$, and $\underline{\sigma}_z^2$, such that, $0 < \underline{\sigma}_w^2 \leq \sigma_w^2 \leq \overline{\sigma}_w^2$ and $0 < \underline{\sigma}_z^2 \leq \sigma_z^2 \leq \overline{\sigma}_z^2$, for some finite $\overline{\sigma}_w^2, \overline{\sigma}_z^2$. Note that this is a significant generalization compared to the previous sections. We would like to highlight that this relaxation can be achieved since we do not explicitly construct a Kalman filter for state estimation in this section. Instead, we will consider the online learning setting where a set of possible controller parameters are given for the decision-making algorithm to select. Finally, we assume that the underlying system Θ satisfies Assumption 5.1, which guarantees that the adaptive control problem is well-defined.

Strongly Convex Cost Function: For the system above, we study the case when the cost function at time t, ℓ_t, is smooth and strongly convex for all t, i.e., $0 \prec \underline{\alpha}_{loss} I \preceq \nabla^2 \ell_t(\cdot, \cdot) \preceq \overline{\alpha}_{loss} I$ for a finite constant $\overline{\alpha}_{loss}$. Note that this is also a generalization of the standard quadratic regulatory costs studied in previous sections where $\ell_t(y_t, u_t) = y_t^\top Q_t y_t + u_t^\top R_t u_t^\top$ with bounded positive definite matrices Q_t and R_t are special cases of the current setting. For all t, the unknown lost function $c_t = \ell_t(\cdot, \cdot)$ is non-negative strongly convex and associated with a parameter L, such

that for any R with $\|u\|, \|u'\|, \|y\|, \|y'\| \leq R$, we have,

$$|\ell_t(y,u) - \ell_t(y',u')| \leq LR(\|y - y'\| + \|u - u'\|) \quad \text{and} \quad |\ell_t(y,u)| \leq LR^2. \tag{5.81}$$

Linear dynamic controller (LDC): An LDC policy π is a linear controller with internal state dynamics governed by $(A_\pi, B_\pi, C_\pi, D_\pi)$ such that

$$s^\pi_{t+1} = A_\pi s^\pi_t + B_\pi y_t, \qquad u^\pi_t = C_\pi s^\pi_t + D_\pi y_t, \tag{5.82}$$

where $s^\pi_t \in \mathbb{R}^s$ is the state of the controller, y_t is the input to the controller, *i.e.*, observation from the system that controller is designing a policy for, and u^π_t is the output of the controller. LDC controllers provide a large class of controllers. For instance, the optimal controller for the LQG control systems given in (5.9) is an LDC policy. Deploying a LDC policy π on the system $\Theta = (A, B, C)$ with dynamics 5.1 induces the following joint dynamics of the x^π_t, s^π_t and the observation-action process:

$$\begin{bmatrix} x^\pi_{t+1} \\ s^\pi_{t+1} \end{bmatrix} = \underbrace{\begin{bmatrix} A + BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix}}_{A'_\pi} \begin{bmatrix} x^\pi_t \\ s^\pi_t \end{bmatrix} + \underbrace{\begin{bmatrix} I_n & BD_\pi \\ 0_{s\times n} & B_\pi \end{bmatrix}}_{B'_\pi} \begin{bmatrix} w_t \\ z_t \end{bmatrix},$$

$$\begin{bmatrix} y^\pi_t \\ u^\pi_t \end{bmatrix} = \underbrace{\begin{bmatrix} C & 0_{m\times s} \\ D_\pi C & C_\pi \end{bmatrix}}_{C'_\pi} \begin{bmatrix} x^\pi_t \\ s^\pi_t \end{bmatrix} + \underbrace{\begin{bmatrix} 0_{m\times n} & I_n \\ 0_{p\times n} & D_\pi \end{bmatrix}}_{D'_\pi} \begin{bmatrix} w_t \\ z_t \end{bmatrix},$$

where $(A'_\pi, B'_\pi, C'_\pi, D'_\pi)$ are the associated parameters of induced closed-loop system. We define the Markov operator for the system $(A'_\pi, B'_\pi, C'_\pi, D'_\pi)$, as $\mathbf{G}'_\pi = \{{G'_\pi}^{[i]}\}_{i=0}$, where ${G'_\pi}^{[0]} = D'_\pi$, and $\forall i > 0$, ${G'_\pi}^{[i]} = C'_\pi {A'_\pi}^{i-1} B'_\pi$. Note that the Markov operator $\mathbf{G}'_\pi$ has the structure of the input-to-output Markov parameters $\{G^i_{u\to y}\}_{i\geq 1}$ introduced in Definition 5.3, yet for the closed-loop system created by the LDC policy.

Proper Decay Function: Let $\psi : \mathbb{N} \to \mathbb{R}_{\geq 0}$ be a proper decay function, such that ψ is non-increasing and $\lim_{h'\to\infty} \psi(h') = 0$. For a input-to-output Markov operator $\mathbf{G}$, $\psi_{\mathbf{G}}(h)$ defines the induced decay function on $\mathbf{G} = \{G^i_{u\to y}\}_{i\geq 1}$, i.e., $\psi_{\mathbf{G}}(h) \coloneqq \sum_{i\geq h} \|G^{[i]}\|$. $\Pi(\psi)$ denotes the class of LDC policies associated with a proper decay function ψ, such that for all $\pi \in \Pi(\psi)$, and all $h \geq 0$, $\sum_{i\geq h} \|{G'_\pi}^{[i]}\| \leq \psi(h)$. Let $\kappa_\psi \coloneqq \psi(0)$.

Nature's output: Using (5.13), we can further decompose the generative components of y_t to obtain,

$$y_t = z_t + CA^t x_0 + \sum_{i=0}^{t-1} CA^{t-i-1} w_i + \sum_{i=0}^{t} G^i_{u\to y} u_{t-i}$$

Notice that first three components generating y_t are derived from the uncontrollable noise processes in the system, while the last one is a linear combination of control inputs. The first three components are known as Nature's y, i.e., Nature's output [245, 299], of the system,

$$b_t(\mathbf{G}) \coloneqq y_t - \sum_{i=0}^{t-1} G^i_{u\to y} u_{t-i} = z_t + CA^t x_0 + \sum_{i=0}^{t-1} CA^{t-i-1} w_i. \tag{5.83}$$

The ability to define Nature's y is a unique characteristic of linear dynamical systems. At any time step t, after following a sequence of control inputs $\{u_i\}_{i=0}^{t}$, and observing y_t, we can compute $b_t(\mathbf{G})$ using (5.83). This quantity allows for counterfactual reasoning about the outcome of the system. Particularly, having access to $\{b_{\tau-t}(\mathbf{G})\}_{t\geq 0}$, we can reason what the outputs $y'_{\tau-t}$ of the system would have been, if the agent, instead, had taken other sequence of control inputs $\{u'_i\}_{i=0}^{\tau-t}$, i.e.,

$$y'_{\tau-t} = b_{\tau-t}(\mathbf{G}) + \sum_{i=0}^{\tau-t-1} G^i_{u\to y} u'_{\tau-t-i}.$$

This property indicates that we can use $\{b_{\tau-t}(\mathbf{G})\}_{t\geq 0}$ to evaluate the quality of any other potential input sequences and build a desirable controller, as elaborated in the following.

Disturbance feedback control (DFC): In this section, we adopt a convex policy reparametrization called DFC introduced by Simchowitz et al. [245]. A DFC policy of length H' is defined as a set of parameters, $\mathbf{M}(H') := \{M^{[i]}\}_{i=0}^{H'-1}$, prescribing the control input of

$$u_t^{\mathbf{M}} = \sum_{i=0}^{H'-1} M^{[i]} b_{t-i}(\mathbf{G}) \tag{5.84}$$

for Nature's y, $\{b_{t-i}(\mathbf{G})\}_{i=0}^{H'-1}$, and resulting in state $x_{t+1}^{\mathbf{M}}$ and observation $y_{t+1}^{\mathbf{M}}$. The DFC policy construction is in parallel with the classical Youla parametrization [299] which states that any linear controller can be prescribed as acting on past noise sequences. Thus, DFC policies can be regarded as truncated approximations to LDCs, or any stabilizing LDC policy can be well-approximated as a DFC policy. More formally, in the following, we show that for any $\pi \in \Pi(\psi)$ and any input u_t^{π} at time step t, there is a set of parameters $\mathbf{M}(H')$, such that $u_t^{\mathbf{M}}$ is sufficiently close to u_t^{π}, and the resulting y_t^{π} is sufficiently close to $y_t^{\mathbf{M}}$.

Lemma 5.15. *Suppose $\|b_t(\mathbf{G})\| \le \kappa_b$ for all $t \le T$ for some κ_b. For any LDC policy $\pi \in \Pi(\psi)$, there exists a H' length DFC policy $\mathbf{M}(H')$ such that $\|u_t^\pi - u_t^{\mathbf{M}}\| \le \psi(H')\kappa_b$, and $\|y_t^\pi - y_t^{\mathbf{M}}\| \le \psi(H')\kappa_{\mathbf{G}}\kappa_b$. One of the DFC policies that satisfies this is $M^{[0]} = D_\pi$, and $M^{[i]} = C'_{\pi,u}A'^{i-1}_{\pi}B'_{\pi,z}$ for $0 < i < H'$.*

The proof is provided in Appendix C.3. This lemma further entails that any stabilizing LDC can be well approximated by a DFC that belongs to a bounded set of DFCs such as

$$\mathcal{M} = \left\{\mathbf{M}(H') := \{M^{[i]}\}_{i=0}^{H'-1} : \sum\nolimits_{i\ge 0}^{H'-1} \|M^{[i]}\| \le \kappa_{\mathcal{M}}\right\},$$

indicating that using the class of DFC policies as an approximation to LDC policies is justified. Using this fact, define the convex compact sets of DFCs, $\mathcal{M}_\psi$ and $\mathcal{M}$ such that the DFC controllers $\mathbf{M}(H'_0) \in \mathcal{M}_\psi$ are bounded i.e., $\sum_{i\ge 0}^{H'_0-1} \|M^{[i]}\| \le \kappa_\psi$ and $\mathcal{M}$ is an r-expansion of $\mathcal{M}_\psi$, i.e., $\mathcal{M} = \{\mathbf{M}(H') = \mathbf{M}(H'_0) + \Delta : \mathbf{M}(H'_0) \in \mathcal{M}_\psi, \sum_{i\ge 0}^{H'-1} \|\Delta^{[i]}\| \le r\kappa_\psi\}$ where $H'_0 = \lfloor \frac{H'}{2} \rfloor - H$. Thus, all controllers $\mathbf{M}(H') \in \mathcal{M}$ are also bounded $\sum_{i\ge 0}^{H'-1} \|M^{[i]}\| \le \kappa_{\mathcal{M}}$ where $\kappa_{\mathcal{M}} = \kappa_\psi(1 + r)$. Throughout the interaction with the system, we assume that the agent has access to $\mathcal{M}$.

Policy Regret: We evaluate the agent's performance by its regret with respect to $\mathbf{M}_\star$, the optimal, in hindsight, DFC policy in the given set $\mathcal{M}_\psi$, i.e., $\mathbf{M}_\star = \operatorname{argmin}_{\mathbf{M}\in\mathcal{M}_\psi} \sum_{t=1}^T \ell_t(y_t^{\mathbf{M}}, u_t^{\mathbf{M}})$. After T step of interaction, the agent's regret is denoted as

$$\textsc{Regret}(T) = \sum\nolimits_{t=1}^T c_t - \ell_t(y^{\mathbf{M}_\star}, u^{\mathbf{M}_\star}). \tag{5.85}$$

Note that we assume the agent has access to an overparameterized set of controllers $\mathcal{M}$ while competing against the best controller in hindsight in $\mathcal{M}_\psi$. This overparameterization is required to obtain the persistence of excitation and desirable approximation of the underlying optimal controller even under model learning error. The rest of this section is organized as follows: in Section 5.6.2, we describe the algorithm of AdaptOn and provide the main regret guarantees, in particular the first logarithmic regret bound for the adaptive control of LQG control systems. In Section 5.6.3, we provide the PE condition in the closed-loop control and show that this condition can be satisfied by AdaptOn with a small enough estimation error throughout the entire adaptive control. Then, we present the key pieces that allow the superior regret of AdaptOn in Section 5.6.4. Finally, in Section 5.6.5 we extend the guarantees of AdaptOn to the ARX systems with sub-Gaussian noise.

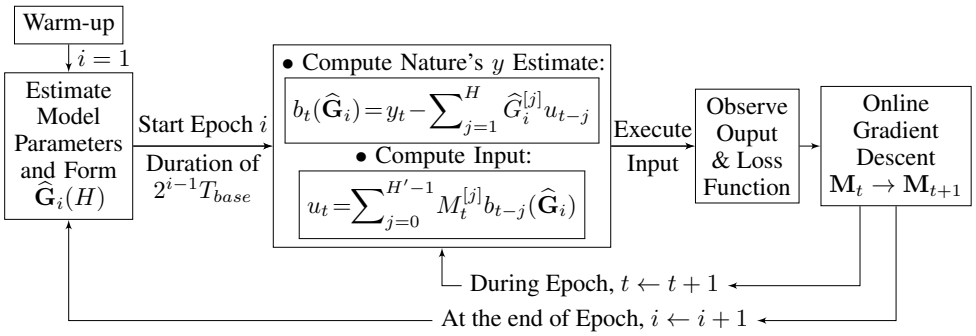

Figure 5.3: AdaptOn.

5.6.2 Adaptive Control via AdaptOn

In this section, we present AdaptOn, a sample efficient **adapt**ive control **on**line learning algorithm which learns the model dynamics through interaction with the environment and continuously deploys online convex optimization to improve the control policy. AdaptOn is illustrated in Figure 5.3 and the detailed pseudo-code is provided in Algorithm 13.

Warm-up: AdaptOn starts with a fixed warm-up period and applies $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ for the first T_w time steps. The length of the warm-up period is chosen to guarantee an accountable first estimate of the system, the persistence of excitation during the adaptive control period, and the stability of the online learning algorithm on the underlying system. The details are provided in the following section.

Adaptive control in epochs: After the warm-up, AdaptOn starts controlling the system and operates in epochs with doubling length. AdaptOn sets the base period to the initial value of T_w and for each epoch i, it runs for $2^{i-1}T_w$ time steps. Note that this doubling update rule is the same as LqgOpt.

Model estimation in the beginning of epochs: At the beginning of each epoch i, AdaptOn exploits the past experiences up to epoch i. It deploys the proposed closed-loop estimation method and solves (5.21) to estimate $\mathcal{G}_{\mathbf{yu}}$. AdaptOn then exploits the construction of true $\mathcal{G}_{\mathbf{yu}}$ to estimate model parameter estimates $\hat{A}_i, \hat{B}_i, \hat{C}_i$ via SysId as described in Section 5.3 and constructs an estimate of H-length input-to-output Markov parameters matrix, $\widehat{\mathbf{G}}_i(h) = [\widehat{G}^1_{u\to y}, \ldots, \widehat{G}^h_{u\to y}]$.

Control input, output and loss during the epochs: AdaptOn utilizes $\widehat{\mathbf{G}}_i(H)$ and the past inputs to estimate the Nature's outputs, $b_t(\widehat{\mathbf{G}}_i) = y_t - \sum_{j=1}^{h} \widehat{G}^j_{u\to y} u_{t-j}$. Using these estimates, AdaptOn executes a DFC policy $\mathbf{M}_t \in \mathcal{M}$ such that $u_t^{\mathbf{M}_t} =$

$\sum_{j=0}^{H'-1} M_t^{[j]} b_{t-j}(\widehat{\mathbf{G}}_i)$ and observes the output of $y_t^{\mathbf{M}_t}$. Finally, AdaptOn receives the loss function ℓ_t, pays a cost of $\ell(y_t, u_t^{\mathbf{M}_t})$.

Counterfactual input, output, loss: AdaptOn uses counterfactual reasoning introduced in Simchowitz et al. [245] to update its controller. After observing the loss function ℓ_t, it constructs,

$$\tilde{u}_{t-j}(\mathbf{M}_t, \widehat{\mathbf{G}}_i) = \sum\nolimits_{l=0}^{H'-1} M_t^{[l]} b_{t-j-l}(\widehat{\mathbf{G}}_i), \tag{5.86}$$

the counterfactual inputs, which are the recomputations of past inputs as if the current DFC policy is applied using Nature's y estimates. Then, AdaptOn reasons about the counterfactual output of the system. Using the current Nature's y estimate and the counterfactual inputs, the agent approximates what the output of the system could be, if counterfactual inputs had been applied,

$$\tilde{y}_t(\mathbf{M}_t, \widehat{\mathbf{G}}_i) = b_t(\widehat{\mathbf{G}}_i) + \sum\nolimits_{j=1}^{h} \widehat{G}^j_{u\to y} \tilde{u}_{t-j}(\mathbf{M}_t, \widehat{\mathbf{G}}_i). \tag{5.87}$$

Using the counterfactual inputs, output, and the revealed loss function ℓ_t, AdaptOn finally constructs,

$$f_t(\mathbf{M}_t, \widehat{\mathbf{G}}_i) = \ell_t(\tilde{y}_t(\mathbf{M}_t, \widehat{\mathbf{G}}_i), \tilde{u}_t(\mathbf{M}_t, \widehat{\mathbf{G}}_i)), \tag{5.88}$$

which is termed counterfactual loss. It is AdaptOn's approximation of what the cost would have been at time t, if the current DFC policy was applied until time t. It gives a performance evaluation of the current DFC policy to AdaptOn for updating the policy. Note that the Markov parameter estimates are crucial in the accuracy of this performance evaluation.

Online convex optimization: In order to optimize the controller during the epoch, at each time step, AdaptOn runs online gradient descent on the counterfactual loss function $f_t(\mathbf{M}_t, \widehat{\mathbf{G}}_i)$ while keeping the updates in the set $\mathcal{M}$ via projection [245],

$$\mathbf{M}_{t+1} = proj_{\mathcal{M}}\left(\mathbf{M}_t - \eta_t \nabla_{\mathbf{M}} f_t\left(\mathbf{M}, \widehat{\mathbf{G}}_i\right)\Big|_{\mathbf{M}_t}\right).$$

Notice that if AdaptOn had access to the underlying input-to-output Markov operator $\mathbf{G}$, the counterfactual loss would have been the true loss of applying the current DFC policy until time t, up to truncation. By knowing the exact performance of the DFC policy, online gradient descent would have obtained accurate updates. Using the counterfactual loss for optimizing the controller causes an error in the gradient updates which is a function of the estimation error of $\widehat{\mathbf{G}}_i$. Therefore, as the Markov

Algorithm 13 AdaptOn

Input: T, h, H, H', T_w, $\mathcal{M}$

—— Warm-Up ——

for $t = 1, \ldots, T_w$ **do**

  Deploy $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$

Store $\mathcal{D}_{T_w} = \{y_t, u_t\}_{t=1}^{T_w}$, set $t_1 = T_w$, $t = T_w + 1$, and $\mathbf{M}_t$ as any member of $\mathcal{M}$

—— Adaptive Control ——

for $i = 1, 2, \ldots$ **do**

  Solve (5.21) using $\mathcal{D}_t$, estimate $\hat{A}_i, \hat{B}_i, \hat{C}_i$ using SysId and construct $\widehat{\mathbf{G}}_i(h)$

  Compute $b_\tau(\widehat{\mathbf{G}}_i) := y_\tau - \sum_{j=1}^{h} \widehat{G}^j_{u\to y} u_{\tau-j}, \forall \tau \le t$

  while $t \le t_i + 2^{i-1}T_w$ *and* $t \le T$ **do**

    Observe y_t, and compute $b_t(\widehat{\mathbf{G}}_i) := y_t - \sum_{j=1}^{h} \widehat{G}^j_{u\to y} u_{t-j}$

    Commit to $u_t = \sum_{j=0}^{H'-1} M_t^{[j]} b_{t-j}(\widehat{\mathbf{G}}_i)$, observe ℓ_t, and pay a cost of $\ell_t(y_t, u_t)$

    Update $\mathbf{M}_{t+1} = proj_{\mathcal{M}}(\mathbf{M}_t - \eta_t \nabla f_t(\mathbf{M}_t, \widehat{\mathbf{G}}_i))$, $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{y_t, u_t\}$, set $t = t+1$

  $t_{i+1} = t_i + 2^{i-1}T_w$

estimates improve via our closed-loop estimation method, the gradient updates get more and more accurate.

For the convex compact DFC policy sets, $\mathcal{M}_\psi$ and $\mathcal{M}$, one can have all the controllers in these sets persistently excite the system Θ. The precise definition of the persistence of excitation condition is given in Section 5.6.3. As discussed in previous sections, the PE condition is a mild condition and in this case, it only requires a significantly wide matrix that maps past H' noise sequences to input to be full row rank. This condition holds in many well-known controllers such as $\mathcal{H}_2, \mathcal{H}_\infty$, as well as their DFC approximations. Moreover, for a DFC policy that satisfies the PE condition, there exists a neighborhood around it consisting of persistently exciting controllers such that the convex compact sets of $\mathcal{M}_\psi$ and $\mathcal{M}$ could be characterized as. Given the construction of AdaptOn and $\mathcal{M}$, we have the following results.

Theorem 5.9. *Suppose for an unknown partially observable linear dynamical system $\Theta = (A, B, C)$ with strongly convex cost function, Assumption 5.1 holds. Given $\mathcal{M}$, a closed, compact, and convex set of DFC policies with the persistence of excitation, i.e., policies that satisfy Assumption 5.4, for long enough warm-up period of T_w, with high probability, AdaptOn achieves the optimal logarithmic regret, i.e., Regret$(T) = polylog(T)$. For the given setting if the policies in $\mathcal{M}$ do not satisfy the PE condition, then for a warm-up duration of $T_w = \tilde{\mathcal{O}}(\sqrt{T})$, AdaptOn attains the regret of Regret$(T) = \tilde{\mathcal{O}}(\sqrt{T})$, with high probability.*

The proof of this result is given in Appendix C.3.4, while the pieces leading up to

this result and a precise statement of Theorem 5.9 will be presented in Section 5.6.4. This result highlights that the learning and control procedure of AdaptOn turns the adaptive control problem into an online convex optimization problem, which yields the optimal regret result in learning and control of partially observable linear dynamical systems. Here the warm-up duration T_w is chosen as $T_w \geq T_{\max}$ where

$$T_{\max} := \max\{h, H, H', T_o, T_A, T_B, T_c, T_{\epsilon_G}, T_{cl}, T_{cx}, T_r\}. \tag{5.89}$$

The detailed expressions for these values are presented throughout this section. In particular, this choice of warm-up guarantees an accountable first estimate of the system (T_o, see Lemma 5.4), the stability of the online learning algorithm on the underlying system (T_A, T_B, see Appendix C.3.2), the stability of the inputs and outputs ($T_{\epsilon_{\mathbf{G}}}$, see Section 5.6.3), the persistence of excitation during the adaptive control period (T_{cl}, see Section 5.6.3), an accountable estimate at the first epoch of adaptive control (T_c, see Section 5.6.3), the conditional strong convexity of expected counterfactual losses (T_{cx}, see Appendix C.3.3), and the existence of a good comparator DFC policy in $\mathcal{M}$ (T_r, see Appendix C.3.3). Note that h is the length of the estimated input-to-output Markov operator for Nature's output computation, H is the length of history used for closed-loop system identification, and H' is the length of the DFC policy. Note that the warm-up duration is a fixed problem-dependent constant. In the following, we first characterize the PE condition needed for the given policy set and then show that if it holds, then the DFC policies constructed by AdaptOn still provide persistence of excitation even under small enough model estimation errors. This result will help AdaptOn refine its model estimates, in particular, input-to-output Markov parameter estimate, during the adaptive control period due to Theorem 5.2 and allows logarithmic regret.

5.6.3 PE Condition in the Closed-Loop Setting

In this section, we provide the persistence of excitation of AdaptOn policies that is required for consistent estimation of system parameters as pointed out in Theorem 5.2. Note that during the warm-up phase, AdaptOn uses i.i.d. Gaussian excitations as inputs which are shown to achieve the persistence of excitation in the covariates for the closed-loop system identification as shown in Section 5.3.2. Therefore, we focus on the persistence of excitation in the adaptive control phase. Throughout the adaptive control phase, we assume that the agent has access to a convex compact set of DFCs, $\mathcal{M}$ which is an r-expansion of $\mathcal{M}_\psi$, such that $\kappa_{\mathcal{M}} = \kappa_\psi(1+r)$ and all controllers $\mathbf{M} \in \mathcal{M}$ are persistently exciting the system Θ. In the following, we formally

define the persistence of excitation condition for the given set $\mathcal{M}$. Then, we show that the persistence of excitation is achieved by the policies that AdaptOn deploys.

Persistence of Excitation Condition of $\mathbf{M} \in \mathcal{M}$

Assume that the underlying system Θ is known by the agent. Then, during the adaptive control, the following expressions give the inputs of AdaptOn and the outputs of the system:

$$u_t = \sum_{j=0}^{H'-1} M_t^{[j]} b_{t-j}(\mathbf{G})$$
$$y_t = [G_{u\to y}^0 \;\; G_{u\to y}^1 \ldots \;\; G_{u\to y}^h] \left[u_t^\top \;\; u_{t-1}^\top \ldots u_{t-h}^\top\right]^\top + b_t(\mathbf{G}) + \mathbf{r_t},$$

where $\mathbf{r_t} = \sum_{k=h+1}^{t-1} G_{u\to y}^k u_{t-k}$. For H defined in Section 5.3, $H \geq \max\{2n+1, \frac{\log(c_H T^2\sqrt{m}/\sqrt{\lambda})}{\log(1/\upsilon)}\}$, we have the following decompositions for ϕ_t:

$$\phi_t = \underbrace{\begin{bmatrix} G_{u\to y}^0 & G_{u\to y}^1 & \ldots & \ldots & \ldots & G_{u\to y}^h & 0_{m\times p} & 0_{m\times p} & \ldots & 0_{m\times p} \\ 0_{m\times p} & G_{u\to y}^0 & \ldots & \ldots & \ldots & G_{u\to y}^{h-1} & G_{u\to y}^h & 0_{m\times p} & \ldots & 0_{m\times p} \\ & & \ddots & & & & & \ddots & & \\ 0_{m\times p} & \ldots & 0_{m\times p} & G_{u\to y}^0 & G_{u\to y}^1 & \ldots & \ldots & \ldots & G_{u\to y}^{h-1} & G_{u\to y}^h \\ I_{p\times p} & 0_{p\times p} & 0_{p\times p} & 0_{p\times p} & 0_{p\times p} & 0_{p\times p} & \ldots & \ldots & \ldots & 0_{p\times p} \\ 0_{p\times p} & I_{p\times p} & 0_{p\times p} & 0_{p\times p} & 0_{p\times p} & 0_{p\times p} & \ldots & \ldots & \ldots & 0_{p\times p} \\ & & \ddots & & & & & & & \\ 0_{p\times p} & 0_{p\times p} & \ldots & I_{p\times p} & 0_{p\times p} & \ldots & \ldots & \ldots & \ldots & 0_{p\times p} \end{bmatrix}}_{\mathcal{T}_{\mathbf{G}} \in \mathbb{R}^{H(m+p)\times(H+H)p}} \underbrace{\begin{bmatrix} u_{t-1} \\ \vdots \\ u_{t-h} \\ \vdots \\ u_{t-h-H} \end{bmatrix}}_{\mathcal{U}_t} + \underbrace{\begin{bmatrix} b_{t-1} \\ \vdots \\ b_{t-H} \\ 0_p \\ \vdots \\ 0_p \end{bmatrix}}_{B_y(\mathbf{G})(t)} + \underbrace{\begin{bmatrix} \mathbf{r_{t-1}} \\ \vdots \\ \mathbf{r_{t-H}} \\ 0_p \\ \vdots \\ 0_p \end{bmatrix}}_{\mathbf{R}_t}$$

$$\mathcal{U}_t = \underbrace{\begin{bmatrix} M_{t-1}^{[0]} & M_{t-1}^{[1]} & \ldots & \ldots & M_{t-1}^{[H'-1]} & 0_{p\times m} & 0_{p\times m} & \ldots & 0_{p\times m} \\ 0_{p\times m} & M_{t-2}^{[0]} & \ldots & \ldots & M_{t-2}^{[H'-2]} & M_{t-2}^{[H'-1]} & 0_{p\times m} & \ldots & 0_{p\times m} \\ & & \ddots & & & & & \ddots & \\ 0_{p\times m} & \ldots & 0_{p\times m} & M_{t-h-H}^{[0]} & \ldots & \ldots & \ldots & \ldots & M_{t-h-H}^{[H'-1]} \end{bmatrix}}_{\mathcal{T}_{\mathbf{M}_t} \in \mathbb{R}^{(h+H)p\times m(H+H'+h-1)}} \underbrace{\begin{bmatrix} b_{t-1}(\mathbf{G}) \\ b_{t-2}(\mathbf{G}) \\ \vdots \\ b_{t-H'+1}(\mathbf{G}) \\ \vdots \\ b_{t-h-H-H'+1}(\mathbf{G}) \end{bmatrix}}_{B(\mathbf{G})(t)}$$

$$B(\mathbf{G})(t) = \underbrace{\begin{bmatrix} I_m & 0_m & \dots & 0_m & C & CA & \dots & \dots & \dots & CA^{t-3} \\ 0_m & I_m & & 0_m & 0_{m\times n} & C & \dots & \dots & \dots & CA^{t-4} \\ & & \ddots & & & & \ddots & & \ddots & \\ 0_m & 0_m & \dots & I_m & 0_{m\times n} & \dots & \dots & C & \dots & CA^{t-h-H-H'-1} \end{bmatrix}}_{O_t} \underbrace{\begin{bmatrix} z_{t-1} \\ z_{t-2} \\ \vdots \\ z_{t-h-H-H'+1} \\ w_{t-2} \\ w_{t-3} \\ \vdots \\ w_1 \end{bmatrix}}_{\boldsymbol{\eta}_t}$$

$$\text{and } B_y(\mathbf{G})(t) = \underbrace{\begin{bmatrix} I_m & 0_m & \dots & \dots & 0_m & C & \dots & \dots & \dots & CA^{t-3} \\ & \ddots & & & \vdots & & \ddots & & \ddots & \\ 0_m & \dots & I_m & \dots & 0_m & 0_{m\times n} & \dots & C & \dots & CA^{t-H-2} \\ & & & & \mathbf{0}_{(pH)\times((h+H+H'-1)m+(t-2)n)} & & & & & \end{bmatrix}}_{\bar{O}_t} \boldsymbol{\eta}_t.$$

Combining all gives

$$\phi_t = \left(\mathcal{T}_{\mathbf{G}}\mathcal{T}_{\mathbf{M}_t}O_t + \bar{O}_t\right)\boldsymbol{\eta}_t + \mathbf{R}_t.$$

Therefore, similar to Assumption 5.2, we require the truncated closed-loop noise evolution parameter to be full-row rank, which is stated in the following assumption.

Assumption 5.4. *For the given system Θ, for $t \geq H + H' + H$, $\mathcal{T}_{\mathbf{G}}\mathcal{T}_{\mathbf{M}_t}O_t + \bar{O}_t$ is full row rank for all $\mathbf{M} \in \mathcal{M}$, i.e.,*

$$\sigma_{\min}(\mathcal{T}_{\mathbf{G}}\mathcal{T}_{\mathbf{M}_t}O_t + \bar{O}_t) > \sigma_c > 0. \tag{5.90}$$

Satisfying the PE Condition in the Adaptive Control Period

In this section, we show that the input-to-output Markov parameter estimates of AdaptOn are well-refined and that, the controller of AdaptOn constructed by using a DFC policy in $\mathcal{M}$ still provides persistence of excitation. In other words, we will show that the inaccuracies in the model parameter estimates do not cause a lack of persistence of excitation in the adaptive control period.

First, we have the following lemma which shows that inputs have persistence of excitation during the adaptive control period. Let $d = \min\{m, p\}$. Using (C.30) and

(C.32), define

$$T_{\epsilon_{\mathbf{G}}} = 4c_1^2\kappa_{\mathcal{M}}^2\kappa_{\mathbf{G}}^2\gamma_{\mathbf{G}}^2\gamma_{\mathcal{H}}^2 T_{\mathcal{G}_{\mathbf{yu}}} \quad T_{cl} = \frac{T_{\epsilon_{\mathbf{G}}}}{\left(\frac{3\sigma_c^2\min\{\underline{\sigma}_w^2,\underline{\sigma}_z^2\}}{8\kappa_u^2\kappa_y H} - \frac{1}{10T}\right)^2}, \tag{5.91}$$

$$T_c = \frac{2048\Upsilon_c^4 H^2\log\left(\frac{H(m+p)}{\delta}\right) + H'mp\log\left(\kappa_{\mathcal{M}}\sqrt{d}+\frac{2}{\epsilon}\right)}{\sigma_c^4\min\{\sigma_w^4,\sigma_z^4\}}, \tag{5.92}$$

for

$$\epsilon = \min\left\{1, \frac{\sigma_c^2\min\{\underline{\sigma}_w^2,\underline{\sigma}_z^2\}\sqrt{\min\{m,p\}}}{68\kappa_b^3\kappa_{\mathbf{G}}H\left(2\kappa_{\mathcal{M}}^2+3\kappa_{\mathcal{M}}+3\right)}\right\}$$

Lemma 5.16. *Let h chosen such that $\psi_{\mathbf{G}}(h+1) \le 1/10T$. Suppose Assumptions 5.1 and 5.4 hold. For a warm-up period of $T_w \ge T_{\max}$, for $T_{\max}$ in* (5.89)*, after T_c time steps in the adaptive control period, with probability $1-3\delta$, we have persistence of excitation for the remainder of the adaptive control epochs of* AdaptOn*, i.e.,*

$$\sigma_{\min}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top\right) \ge t\frac{\sigma_c^2\min\{\sigma_w^2,\sigma_z^2\}}{16}. \tag{5.93}$$

Proof. During the adaptive control, at time t, the input of AdaptOn is given by

$$u_t = \sum\nolimits_{j=0}^{H'-1} M_t^{[j]} b_{t-j}(\mathbf{G}) + M_t^{[j]}\left(b_{t-j}(\widehat{\mathbf{G}}_i) - b_{t-j}(\mathbf{G})\right),$$

where

$$b_{t-j}(\mathbf{G}) = y_{t-j} - \sum_{k=1}^{t-j-1} G_{u\to y}^k u_{t-j-k} = z_{t-j} + \sum_{k=1}^{t-j-1} CA^{t-j-k-1}w_k \tag{5.94}$$

$$b_{t-j}(\widehat{\mathbf{G}}_i) = y_{t-j} - \sum_{k=1}^{h}\widehat{G}_{u\to y}^{[k]} u_{t-j-k}. \tag{5.95}$$

Thus, we obtain the following for u_t and y_t,

$$u_t = \sum_{j=0}^{H'-1} M_t^{[j]} b_{t-j}(\mathbf{G}) + \underbrace{\sum_{j=0}^{H'-1} M_t^{[j]}\left(\sum_{k=1}^{t-j-1}[G_{u\to y}^k - \widehat{G}_{u\to y}^k]u_{t-j-k}\right)}_{u_{\Delta b}(t)}$$

$$y_t = [G_{u\to y}^0 \;\; G_{u\to y}^1 \ldots \;\; G_{u\to y}^h]\left[u_t^\top \;\; u_{t-1}^\top \ldots u_{t-h}^\top\right]^\top + b_t(\mathbf{G}) + \mathbf{r_t},$$

where $\mathbf{r_t} = \sum_{k=h+1}^{t-1} G_{u\to y}^k u_{t-k}$ and $\sum_{k=h}^{t-1}\|G_{u\to y}^k\| \le \psi_{\mathbf{G}}(H+1) \le 1/10T$ which is bounded by assumption. Notice that $\|u_{\Delta b}(t)\| \le \kappa_{\mathcal{M}}\kappa_u\epsilon_{\mathbf{G}}(1,\delta)$ after warm-up

period, where κ_u is a bound on $\|u\|$, $\sum_{i\geq 0}^{H'-1} \|M^{[i]}\| \leq \kappa_{\mathcal{M}}$, and $\epsilon_{\mathbf{G}}(1,\delta)$ is the bound on the estimation error of input-to-output Markov operator after the warm-up period, which holds for the entire adaptive control period Using the definitions from previous section, ϕ_t can be written as,

$$\phi_t = \left(\mathcal{T}_{\mathbf{G}}\mathcal{T}_{\mathbf{M}_t}O_t + \bar{O}_t\right)\boldsymbol{\eta}_t + \mathbf{R}_t + \mathcal{T}_{\mathbf{G}}\mathcal{U}_{\Delta b}(t) \tag{5.96}$$

where

$$\mathcal{U}_{\Delta b}(t) = \begin{bmatrix} u_{\Delta b}(t-1) \\ u_{\Delta b}(t-2) \\ \vdots \\ u_{\Delta b}(t-H) \\ \vdots \\ u_{\Delta b}(t-H-H) \end{bmatrix}.$$

Consider the following,

$$\begin{aligned}\mathbb{E}\left[\phi_t\phi_t^\top\right] = \mathbb{E}\Big[&\left(\mathcal{T}_{\mathbf{G}}\mathcal{T}_{\mathbf{M}_t}O_t + \bar{O}_t\right)\boldsymbol{\eta}_t\boldsymbol{\eta}_t^\top\left(\mathcal{T}_{\mathbf{G}}\mathcal{T}_{\mathbf{M}_t}O_t + \bar{O}_t\right)^\top + \boldsymbol{\eta}_t^\top\left(\mathcal{T}_{\mathbf{G}}\mathcal{T}_{\mathbf{M}_t}O_t + \bar{O}_t\right)^\top\left(\mathcal{T}_{\mathbf{G}}\mathcal{U}_{\Delta b}(t) + \mathbf{R}_t\right) \\ &+ \left(\mathcal{T}_{\mathbf{G}}\mathcal{U}_{\Delta b}(t) + \mathbf{R}_t\right)^\top\left(\mathcal{T}_{\mathbf{G}}\mathcal{T}_{\mathbf{M}_t}O_t + \bar{O}_t\right)\boldsymbol{\eta}_t + \left(\mathcal{T}_{\mathbf{G}}\mathcal{U}_{\Delta b}(t) + \mathbf{R}_t\right)^\top\left(\mathcal{T}_{\mathbf{G}}\mathcal{U}_{\Delta b}(t) + \mathbf{R}_t\right)\Big]\end{aligned}$$

$$\begin{aligned}\sigma_{\min}\left(\mathbb{E}\left[\phi_t\phi_t^\top\right]\right) &\geq \sigma_c^2\min\{\underline{\sigma}_w^2, \underline{\sigma}_z^2\} \\ &\quad -2\kappa_b(\kappa_{\mathcal{M}}+\kappa_{\mathcal{M}}\kappa_{\mathbf{G}}+1)\sqrt{H}((1+\kappa_{\mathbf{G}})\kappa_{\mathcal{M}}\kappa_u\epsilon_{\mathbf{G}}(1,\delta)\sqrt{H}+\sqrt{H}\kappa_u/10T) \\ &\geq \sigma_c^2\min\{\underline{\sigma}_w^2, \underline{\sigma}_z^2\} - 2\kappa_u^2\kappa_y H(2\kappa_{\mathbf{G}}\kappa_{\mathcal{M}}\epsilon_{\mathbf{G}}(1,\delta) + 1/10T).\end{aligned}$$

Note that for $T_w \geq T_{cl}$, $\epsilon_{\mathbf{G}}(1,\delta) \leq \frac{1}{2\kappa_{\mathcal{M}}\kappa_{\mathbf{G}}}\left(\frac{3\sigma_c^2\min\{\underline{\sigma}_w^2,\underline{\sigma}_z^2\}}{8\kappa_u^2\kappa_y H} - \frac{1}{10T}\right)$ with probability at least $1-2\delta$. Thus, we get

$$\sigma_{\min}\left(\mathbb{E}\left[\phi_t\phi_t^\top\right]\right) \geq \frac{\sigma_c^2}{4}\min\{\underline{\sigma}_w^2, \underline{\sigma}_z^2\}, \tag{5.97}$$

for all $t \geq T_w$. Using Lemma C.3.2, we have that for $\Upsilon_c := (\kappa_y + \kappa_u)$, $\|\phi_t\| \leq \Upsilon_c\sqrt{H}$ with probability at least $1-2\delta$. Therefore, for a chosen $\mathbf{M} \in \mathcal{M}$, using Matrix Azuma inequality [267], we have the following with probability $1-3\delta$:

$$\lambda_{\max}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top - \mathbb{E}[\phi_i\phi_i^\top]\right) \leq 2\sqrt{2t}\Upsilon_c^2 H\sqrt{\log\left(\frac{H(m+p)}{\delta}\right)}. \tag{5.98}$$

In order to show that this holds for any chosen $\mathbf{M} \in \mathcal{M}$, we adopt a standard covering argument. We know that from Lemma 5.4 of Simchowitz et al. [245], the Euclidean

diameter of $\mathcal{M}$ is at most $2\kappa_{\mathcal{M}}\sqrt{\min\{m,p\}}$, *i.e.* $\|\mathbf{M}_t\|_F \leq \kappa_{\mathcal{M}}\sqrt{\min\{m,p\}}$ for all $\mathbf{M}_t \in \mathcal{M}$. Thus, we can upper bound the covering number as follows,

$$\mathcal{N}(B(\kappa_{\mathcal{M}}\sqrt{\min\{m,p\}}), \|\cdot\|_F, \epsilon) \leq \left(\kappa_{\mathcal{M}}\sqrt{\min\{m,p\}} + \frac{2}{\epsilon}\right)^{H'mp}.$$

The following holds for all the centers of ϵ-balls in $\|\mathbf{M}_t\|_F$, for all $t \geq T_w$, with probability $1-3\delta$:

$$\lambda_{\max}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top - \mathbb{E}[\phi_i\phi_i^\top]\right) \leq 2\sqrt{2t}\Upsilon_c^2 H\sqrt{\log\left(\tfrac{H(m+p)}{\delta}\right) + H'mp\log\left(\kappa_{\mathcal{M}}\sqrt{\min\{m,p\}}+\tfrac{2}{\epsilon}\right)}. \tag{5.99}$$

Consider all $\mathbf{M}$ in the ϵ-balls, *i.e.* effect of epsilon perturbation in $\|\mathbf{M}\|_F$ sets, using Weyl's inequality we have with probability at least $1-3\delta$,

$$\sigma_{\min}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top\right) \geq t\left(\frac{\sigma_c^2}{4}\min\{\underline{\sigma}_w^2,\underline{\sigma}_z^2\} - \frac{8\kappa_b^3\kappa_{\mathbf{G}}H\epsilon\left(2\kappa_{\mathcal{M}}^2+3\kappa_{\mathcal{M}}+3\right)}{\sqrt{\min\{m,p\}}}\left(1+\frac{1}{10T}\right)\right)$$
$$- 2\sqrt{2t}\Upsilon_c^2 H\sqrt{\log\left(\frac{H(m+p)}{\delta}\right) + H'mp\log\left(\kappa_{\mathcal{M}}\sqrt{\min\{m,p\}}+\frac{2}{\epsilon}\right)}.$$

for $\epsilon \leq 1$. Let $\epsilon = \min\left\{1, \frac{\sigma_c^2\min\{\underline{\sigma}_w^2,\underline{\sigma}_z^2\}\sqrt{\min\{m,p\}}}{68\kappa_b^3\kappa_{\mathbf{G}}H\left(2\kappa_{\mathcal{M}}^2+3\kappa_{\mathcal{M}}+3\right)}\right\}$. For this choice of ϵ, we get

$$\sigma_{\min}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top\right) \geq t\left(\tfrac{\sigma_c^2}{8}\min\{\underline{\sigma}_w^2,\underline{\sigma}_z^2\}\right) - 2\sqrt{2t}\Upsilon_c^2 H\sqrt{\log\left(\tfrac{H(m+p)}{\delta}\right) + H'mp\log\left(\kappa_{\mathcal{M}}\sqrt{\min\{m,p\}}+\tfrac{2}{\epsilon}\right)}.$$

By picking $T_w \geq T_c$, we can guarantee that after T_c time steps in the first epoch, we have the advertised lower bound. □

Note that combining Lemma 5.16 with Theorem 5.3 gives

$$\|\widehat{\mathcal{G}}_{\mathbf{yu},\mathbf{i}} - \mathcal{G}_{\mathbf{yu}}\| \leq \frac{\kappa_e}{\sigma_c\sqrt{2^{i-1}T_w}\sqrt{\frac{\min\{\underline{\sigma}_w^2,\underline{\sigma}_z^2\}}{16}}}, \tag{5.100}$$

for each epoch i of AdaptOn, with probability at least $1-4\delta$, for

$$\kappa_e = \sqrt{m\Sigma_e\left(\log(\frac{1}{\delta}) + \frac{H(m+p)}{2}\log\left(\frac{\lambda(m+p)+\tau\Upsilon^2}{\lambda(m+p)}\right)\right)} + S\sqrt{\lambda} + \sqrt{H}.$$

Setting $\sigma_\star^2 := \min\left\{\frac{\sigma_o^2\underline{\sigma}_w^2}{2}, \frac{\sigma_o^2\underline{\sigma}_z^2}{2}, \frac{\sigma_o^2\sigma_y^2}{2}, \frac{\sigma_c^2\underline{\sigma}_w^2}{16}, \frac{\sigma_c^2\underline{\sigma}_z^2}{16}\right\}$, provides the guarantee in Theorem 5.2 as in (5.23) for both warm-up and adaptive control periods.

5.6.4 Regret Analysis

From the persistence of excitation guarantee given in Lemma 5.16, we have the conditions for Theorem 5.2 to hold for the entire duration of AdaptOn, i.e., our proposed adaptive control algorithm estimates the system dynamics consistently. Thus, using the analysis provided in Appendix C.3.2, we have the same consistent estimation guarantee for the constructed input-to-output Markov operator $\widehat{\mathbf{G}}_i(h)$ using the model parameter estimates obtained from SysId. In particular, using our novel closed-loop system identification method at the beginning of any epoch i ensures that during the epoch, $\|\widehat{\mathbf{G}}_i(h) - \mathbf{G}(h)\| \leq \epsilon_{\mathbf{G}}(i, \delta) = \tilde{O}(1/\sqrt{2^{i-1}T_w})$.

Stable system dynamics with AdaptOn: Since w_t and z_t are Gaussian disturbances, from standard concentration results we have that Nature's y is bounded with high probability for all t (see Appendix C.3.3). Thus, let $\|b_t(\mathbf{G})\| \leq \kappa_b$ for some κ_b. The following lemma shows that during AdaptOn, Markov parameter estimates are well-refined such that the inputs, outputs, and the Nature's y estimates of AdaptOn are uniformly bounded with high probability. The proof is in Appendix C.3.3.

Lemma 5.17. *For all t during the adaptive control epochs, $\|u_t\| \leq \kappa_{\mathcal{M}}\kappa_b$, $\|y_t\| \leq \kappa_b(1 + \kappa_{\mathbf{G}}\kappa_{\mathcal{M}})$ and $\|b_t(\widehat{\mathbf{G}})\| \leq 2\kappa_b$ with high probability.*

Regret upper bound of AdaptOn: Before presenting a precise version of Theorem 5.9, we provide our intuition in bounding the regret of AdaptOn. The regret decomposition of AdaptOn includes 3 main pieces: (R_1) : Regret due to warm-up; (R_2) : Regret due to online learning controller; (R_3) : Regret due to lack of system dynamics knowledge. R_1 gives constant regret for a fixed short warm-up period. R_2 results in $O(\log(T))$ regret. Note that this regret decomposition and these results follow and adapt Theorem 5 of Simchowitz et al. [245]. The key difference is in R_3, which scales quadratically with the Markov parameter estimation error $\|\widehat{\mathbf{G}}_i(h) - \mathbf{G}(h)\|$. Simchowitz et al. [245] deploys an open-loop estimation such as the one presented in Section 5.2 and does not update the model parameter estimates during adaptive control and attains $R_3 = \tilde{O}(\sqrt{T})$ which dominates the regret upper bound. However, using our novel system identification method with the closed-loop learning guarantees of Markov parameters and the doubling epoch lengths AdaptOn gets $R_3 = O(\text{polylog}(T))$.

Theorem 5.10 (Precise Version of Theorem 5.9)**.** *Suppose Assumption 5.1 holds and H' and h are chosen to satisfy $H' \geq 3h \geq 1$, $\psi(\lfloor H'/2 \rfloor - h) \leq \kappa_{\mathcal{M}}/T$ and*

$\psi_{\mathbf{G}}(h+1) \leq 1/10T$. *For the described strongly convex cost function in Section 5.6.1, after a warm-up period time* $T_w \geq T_{\max}$, *if* AdaptOn *runs with step size* $\eta_t = \frac{12}{\alpha t}$, *then with probability at least* $1 - 5\delta$, *the regret of* AdaptOn *is bounded as*

$$\textsc{Regret}(T) \lesssim T_w L \kappa_y^2 + \frac{L^2 H'^3 \min\{m,p\} \kappa_b^4 \kappa_{\mathbf{G}}^4 \kappa_{\mathcal{M}}^2}{\min\{\alpha, L\kappa_b^2 \kappa_{\mathbf{G}}^2\}} \left(1 + \frac{\overline{\alpha}_{loss}}{\min\{m,p\} L \kappa_{\mathcal{M}}}\right) \log\left(\frac{T}{\delta}\right)$$
$$+ \sum_{t=T_w+1}^{T} \epsilon_{\mathbf{G}}^2 \left(\left\lceil \log_2\left(\frac{t}{T_w}\right)\right\rceil, \delta\right) H' \kappa_b^2 \kappa_{\mathcal{M}}^2 \left(\frac{\kappa_{\mathbf{G}}^2 \kappa_b^2 \left(\overline{\alpha}_{loss} + L\right)^2}{\alpha} + \kappa_y^2 \max\left\{L, \frac{L^2}{\alpha}\right\}\right).$$

Here $\lesssim$ *denotes that the inequality holds up to polynomial functions of problem quantities and logarithmic factors. If Assumption 5.4 holds for the given* $\mathcal{M}$, *then with the stated choice of* T_w, *then with the same probability, we have* $\epsilon_{\mathbf{G}}(i,\delta) = \tilde{O}(1/\sqrt{2^{i-1}T_w})$, *which would yield* $\textsc{Regret}(T) = polylog(T)$. *If the given set of controllers does not satisfy Assumption 5.4, then* $\epsilon_{\mathbf{G}}(i,\delta) = \tilde{O}(1/\sqrt{T_w})$, *for all i, and with the choice of* $T_w = O(\sqrt{T})$, *we get* $\textsc{Regret}(T) = \tilde{O}(\sqrt{T})$.

The proof is given in Appendix C.3.4. The regret decomposition and individual upper bounds build on the reduction of the adaptive control problem, more specifically the regret minimization problem, to the online convex optimization with memory setting introduced in [17]. Notice that due to the open-loop stability of the system dynamics, one can easily find h and H' that satisfy the conditions stated at the beginning of the theorem. The main ingredient that allows logarithmic regret is the combining strong convexity of the cost function, which allows the policy regret to scale quadratically with the estimation error, and our novel closed-loop system identification method, which allows continuously improving model estimates.

Remark 5.2. *Initially, the logarithmic regret upper bound of* AdaptOn *may seem surprising since LQG control systems can be seen as strict partial observability generalizations of LQR considered in Section 3, whose regret is proven to be lower bounded by* $\sqrt{T}$ *[242]. However, this is not entirely the case. The logarithmic regret of* AdaptOn *is possible since the LQG control system studied here has isotropic, or in general non-degenerate, measurement noise, which allows the PE condition for a closed-loop controller (Lemma 5.16) without additional perturbations to the system, such as i.i.d. Gaussian inputs. Unfortunately, this is not the case for LQRs, since the state is fully observable. Therefore, in order to achieve PE condition, algorithms like StabL and TSAC in Section 3, have to inject i.i.d. Gaussian perturbations to the system. As shown in* (5.100), *this persistence of excitation allows continuous exploration, thus continuous improvement of the model estimates while simultaneously exploiting near-optimal policies. In other words, the measurement noise,*

which is present for both the learning agent and the best controller in hindsight, makes the control problem harder for the optimal controller, thus, making the regret minimization problem easier.

In minimizing the regret, AdaptOn competes against the best DFC policy in the set $\mathcal{M}_\psi$. Recall that any stabilizing LDC policy can be well-approximated as a DFC policy, Lemma 5.15. Therefore, for any LDC policy π whose DFC approximation lives in the given $\mathcal{M}_\psi$, Theorem 5.9 can be extended to achieve the first logarithmic regret in the LQG control setting which has been studied in Sections 5.4 and 5.5.

Corollary 5.10.1. *Let $\pi_\star$ be the optimal linear controller for the underlying system Θ in the LQG control setting. If the DFC approximation of $\pi_\star$ is in $\mathcal{M}_\psi$, such that it satisfies Assumption 5.2, then the regret of AdaptOn with respect to $\pi_\star$ is $\sum_{t=1}^{T} c_t - \ell_t(y_t^{\pi_\star}, u_t^{\pi_\star}) = polylog(T)$.*

As stated in the second half of Theorem 5.10, without any consistent closed-loop model estimate updates during the adaptive control, AdaptOn reduces to a variant of the algorithm given in Simchowitz et al. [245], yielding $\tilde{O}(\sqrt{T})$ regret. This is equivalent to having no model updates in the adaptive control phase even if the control set $\mathcal{M}$ satisfies Assumption 5.2, i.e., explore-and-commit approach. While the doubling epoch length update rule of AdaptOn results in $\left\lceil \log\left(\frac{T}{T_w}\right) \right\rceil$ updates in the adaptive control period, one can follow different update schemes as long as AdaptOn obtains enough samples at the beginning of the adaptive control period to obtain persistence of excitation. The following is an immediate corollary of Theorem 5.9 which considers the case when the number of epochs or estimations are limited during the adaptive control period.

Corollary 5.10.2. *If enough samples are gathered in the adaptive control period such that the closed-loop persistence excitation condition is satisfied for AdaptOn, i.e., after T_c time steps, AdaptOn with any update scheme less than $\left\lceil \log\left(\frac{T}{T_w}\right) \right\rceil$ updates has Regret$(T) \in [polylog(T), \tilde{O}(\sqrt{T})]$.*

If the control cost functions c_t for the given system are (weakly) convex $0 \preceq \nabla^2 \ell_t(\cdot, \cdot)$, the online gradient descent procedure of AdaptOn does not enjoy the quadratic scaling of regret with respect to the estimation error $\epsilon_{\mathbf{G}}$ in input-to-output Markov parameters, and settles down with linear scaling, i.e., the last term in the regret expression of Theorem 5.10 scales with $\epsilon_{\mathbf{G}} \left(\left\lceil \log_2\left(\frac{t}{T_w}\right) \right\rceil, \delta \right)$. Notice that this was

the case for LQG control and ARX systems studied in Sections 5.4 and 5.5, due to positive semidefinite Q matrix in the state cost. For this setting AdaptOn attains the same regret guarantees of LqgOpt and TSPO.

Corollary 5.10.3. *For the system given in Theorem 5.9 with convex cost functions, $0 \preceq \nabla^2 \ell_t(\cdot,\cdot)$, if $\mathcal{M}$ satisfies Assumption 5.4, then the regret of AdaptOn is $\textsc{Regret}(T) = \tilde{O}(\sqrt{T})$, with high probability. Furthermore, if the policies in $\mathcal{M}$ do not satisfy the PE condition, then for a warm-up duration of $T_w = O(T^{2/3})$, AdaptOn attains the regret of $\textsc{Regret}(T) = \tilde{O}\left(T^{2/3}\right)$, with high probability.*

5.6.5 Extensions to the ARX systems

In this section, similar to Section 5.4, we extend AdaptOn for LQG control to the ARX systems, i.e., the systems with the dynamics of the form (5.7) with sub-Gaussian e_t with the covariance matrix of $\Sigma_E \succeq \sigma_e^2$ and arbitrary F and a stable $\bar{A}$. Moreover, we will relax the conditions on the system dynamics given in Assumption 5.1. In particular, we will consider the stabilizable and detectable ARX systems, since the control design procedure of AdaptOn will not require identification of model parameters and utilize only the input-to-output and output-to-output Markov parameter estimates $\widehat{\mathbf{G}}_{\mathbf{u}\to\mathbf{y}}(h)$ and $\widehat{\mathbf{G}}_{\mathbf{y}\to\mathbf{y}}(h)$, respectively. Additionally, we will consider the strongly convex cost functions introduced in Section 5.6.1 in the adaptive control problem. For the ARX systems, we define a new Nature's output.

Output uncertainties $\bar{b}_t(\mathcal{G}_{\mathbf{yu}})$: Recall the output rollout given in (5.17) using $\mathcal{G}_{\mathbf{yu}}$. The output uncertainties of the ARX system at time t is denoted as follows:

$$\bar{b}_t(\mathcal{G}_{\mathbf{yu}}) = y_t - \left(\sum_{k=0}^{t-1} G_{u\to y}^{k+1} u_{t-k-1} + G_{y\to y}^{k+1} y_{t-k-1}\right) = CA^t x_0 + e_t. \tag{5.101}$$

This definition is similar to Nature's output $b_t(\mathbf{G})$ adopted for LQG control systems in previous sections. It represents the only unknown components of the output. Notice that, one can identify the uncertainty in the output at any time step uniquely using the history of inputs, outputs, and the Markov parameters. This gives the ability of counterfactual reasoning, i.e., consider what the output would have been if the agent had taken a different sequence of inputs and observed different outputs.

Disturbance feedback control (DFC): As in the previous sections, we consider the DFC controllers but this time using the output uncertainties $\bar{b}_t()$, i.e., for the set of parameters $\mathbf{M}(H') := \{M^{[i]}\}_{i=0}^{H'-1}$, the control input is

$$u_t^{\mathbf{M}} = \sum_{i=0}^{H'-1} M^{[i]} \bar{b}_{t-i}(\mathcal{G}_{\mathbf{yu}}). \tag{5.102}$$

We will adopt the same construction of convex compact sets $\mathcal{M}_\psi$ and $\mathcal{M}$ introduced in Section 5.6.1, such that $\mathcal{M}$ is an r-expansion of $\mathcal{M}_\psi$ and the regret of AdaptOn is measured against the optimal, in hindsight, DFC policy in $\mathcal{M}_\psi$, (5.85).

Adaptive Control of ARX systems via AdaptOn

In the ARX systems, we mainly follow AdaptOn given in Algorithm 13. However, we have several changes. Note that to compute the output uncertainties, we use $\widehat{\mathcal{G}}_{\mathbf{yu}}$, which is obtained via solving (5.21). To utilize this fact, we set the regularized least squares roll-out horizon for the closed-loop system identification H equal to the number of Markov parameters used to compute the output uncertainties h. Therefore, in the ARX systems, AdaptOn does not need to construct a balanced realization of the model parameters via SysId. This allows AdaptOn to learn and control a general class of partially observable linear dynamical systems, i.e., stabilizable and detectable system dynamics with sub-Gaussian disturbances. For the rest of the steps of AdaptOn, we follow the counterfactual input, output, and loss construction presented in Section 5.6.2 but this time using both input-to-output and output-to-output Markov parameter estimates $\widehat{\mathbf{G}}_{\mathbf{u}\to\mathbf{y}}(h)$ and $\widehat{\mathbf{G}}_{\mathbf{y}\to\mathbf{y}}(h)$. Finally, we update the DFC controller using the online projected gradient descent at each time step. Similar to the derivation given in Section 5.6.3, in particular Lemma 5.16, one can show that after T_c time steps, where we replace $\min\{\sigma_w^4, \sigma_z^4\}$ with σ_e^4 in (5.92), we have closed-loop PE condition for AdaptOn in ARX systems such that $\sigma_{\min}\left(\sum_{i=1}^{t} \phi_i \phi_i^\top\right) \geq t\frac{\sigma_c^2\sigma_e^2}{16}$. Moreover, the boundedness guarantees of AdaptOn translate to the ARX systems due to the concentration properties of sub-Gaussian noise. Thus, for large enough h and a proper decay function ψ, such that $\psi_{\mathcal{G}_{\mathbf{yu}}}(h) \coloneqq \max\{\psi_{\mathbf{G}_{u\to y}}(h), \psi_{\mathbf{G}_{y\to y}}(h)\} \leq 1/10T$, we have the same results provided for LQG control systems in Theorem 5.10 and Corollaries 5.10.1, 5.10.2, and 5.10.3 for the stabilizable and detectable open-loop stable ARX systems.

5.7 Conclusion and Future Directions

In this chapter, we covered several aspects of learning and control in partially observable linear dynamical systems. We started our study with the problem of learning the underlying unknown system dynamics. After explaining the drawbacks of open-loop estimation methods, we provided the first system identification method that allows estimating the model parameters with finite-time guarantees in both open and closed-loop settings. We showed that this system identification method achieves optimal learning rates for both i.i.d. Gaussian control inputs and measurement-

feedback controllers and recovers a balanced realization of the model parameters via a new subspace identification approach SysId. Building upon this method, we then proposed three novel algorithmic frameworks: LqgOpt, TSPO, and AdaptOn, for learning and control of unknown partially observable linear dynamical systems, namely LQG control systems and ARX systems.

We developed LqgOpt using the OFU principle to design the controllers and balance the exploration vs. exploitation trade-off. We derived the precise characterization of the persistence of excitation condition for the underlying system with its optimal controller and showed that LqgOpt achieves persistence of excitation even under model estimation errors. We proved that LqgOpt achieves stable closed-loop dynamics using the optimistic controllers to refine the model parameter estimates and studied the regret of LqgOpt. In our regret analysis, we derived the Bellman optimality equation for LQG control systems with convex quadratic cost function and a novel regret decomposition based on this which can be of independent interest in deriving diverse regret guarantees in partially observable linear systems in future studies. In fact, we used this regret decomposition in providing regret guarantees for Thompson Sampling-based adaptive control in LQG control systems via TSPO. Our analysis showed that LqgOpt attains regret of $\tilde{O}(\sqrt{T})$ under if PE condition holds for the underlying system with its optimal controller and $\tilde{O}(T^{2/3})$ without PE. Finally, we extended the framework of LqgOpt to the ARX systems.

Secondly, we studied adaptive control of unknown partially observable linear dynamical systems using Thompson Sampling and proposed TSPO. By replacing the cumbersome optimization procedure to find optimistic models in LqgOpt with an efficient Markov parameter sampling method, we showed that TSPO achieves the same regret rates as LqgOpt, using Thompson Sampling. We remarked that this computational efficiency does not come freely. We showed that even though the order-wise regret of TSPO matches LqgOpt, it suffers from larger problem-dependent constants which can be arbitrarily big due to the sampling nature of the algorithm.

Finally, we studied the learning and control of partially observable linear dynamical systems with general strongly convex cost functions and proposed an efficient learning and control algorithm AdaptOn. In this challenging setting, we considered the policy regret as the performance metric and used online convex optimization for the adaptive controller design. Leveraging our novel closed-loop system identification method, we deployed a convex policy reparametrization that approximates the linear measurement-feedback controllers well and aims to reduce the cost caused by

uncontrollable nature disturbances. We showed that by continuously updating the model estimates in adaptive control epochs and running projected online gradient descent to further improve the control policies during the epochs, AdaptOn achieves optimal logarithmic regret. This surprising result makes AdaptOn the first algorithm to achieve optimal regret in the challenging setting of learning and control in partially observable linear dynamical systems, and in fact, shows that the finite-time adaptive control problem or specifically regret minimization problem is simpler in partially observable systems compared to the fully observable setting. We showed that for (weakly) convex cost functions AdaptOn recovers the results of LqgOpt and TSPO, hinting that logarithmic regret is due to strong convexity. Finally, we extended the guarantees of AdaptOn to ARX systems with minimal assumptions which is due to the convex policy parameterization adopted in the control design, alleviating the need for SysId.

One of the most important future directions is to further investigate the role of persistence of excitation (PE). In the LQR setting studied in Section 3, we showed that StabL and TSAC attain $O(\sqrt{T})$ regret using a self-normalized construction in the analysis, without PE condition. It remains an open problem if these results could be extended to the LQG control problems. We conjecture that this is the case. Informally, using the predictor form representation of the system dynamics, one can study the truncated online LQG control problem as an LQR problem. This would allow transferring the self-normalized results in the LQR setting to partially observable linear dynamical systems and by showing that this truncation does not result in a linear scaling of regret, one can show that the requirement of PE condition to achieve $O(\sqrt{T})$ would be eliminated.

Another important direction is to study whether it is possible to design controllers directly from the Markov parameter estimates without deploying subspace identification algorithms like SysId. These methods require controllability and observability of the underlying system, which are usually restrictive for most of the partially observable linear dynamical systems and are quite sensitive to the minimum singular values of the constructed Hankel matrices as discussed in Theorem 5.4. Designing controllers directly from data, i.e., using the solution of (5.21), would alleviate such needs and provide more robust learning and control algorithms. An example of such construction is given in ARX systems via AdaptOn in Section 5.6.5.

Another interesting research question is to investigate the role of open-loop stability in learning and control of partially observable linear dynamical systems. In our

study, we assumed that the state evolution matrix A is stable. This allowed bounded input and outputs and simplified the analysis. One can replace this assumption with a known stabilizing controller assumption, e.g., [245]. However, this does not fully solve the problem. For example, in the LQR setting of Section 3, we showed that StabL and TSAC achieve optimal regret guarantee with polynomial dimension dependency without any assumption on A, even though the state might have an exponential dependency in dimension until finding a stabilizing neighborhood. It is an open problem, due to its partially observable nature, if we can have the same analysis in the LQG control setting and remove the stability assumption overall while suffering from an exponential in-dimension state until the recovery of stabilizing measurement-feedback controllers.

Another important direction is to see whether LqgOpt and TSPO can achieve polylogarithmic regret under strongly convex cost conditions, i.e., $Q, R > 0$. Moreover, even though it is a computationally efficient alternative to LqgOpt, the larger constants in the regret upper bound of TSPO, make it undesirable in complicated scenarios. Recently, [8] showed a computationally efficient relaxation on the optimistic model selection problem in the LQR setting with some performance degradation, yet order-wise the same regret as StabL. Studying such a relaxation to LqgOpt, thus making the optimistic model selection computationally efficient, is a direction that would yield tighter regret guarantees than TSPO. The final major future direction is to consider the constrained learning and control setting in partially observable linear dynamical systems, e.g., safety, which would yield adaptive algorithms that are suitable in more challenging control scenarios.

Chapter 6

LEARNING AND CONTROL IN NONLINEAR DYNAMICAL SYSTEMS

In this chapter, we study the learning and control in dynamical systems problem in its most general setting: (partially observable) nonlinear systems[1]. As highlighted in the previous sections, there has been a flurry of studies that consider obtaining finite-time performance guarantees for learning and control of linear dynamical systems. Even though these remarkable efforts provide a foundational understanding and aim to shed light on the learning and control of more complex systems, they generally do not extend to nonlinear systems since they heavily rely on the simplicity of linear system modeling. This is disappointing since most of the systems encountered in practice are nonlinear dynamical systems. To make the learning and control design problem tractable, the recent studies in finite-time learning to control in nonlinear dynamical systems consider identifying the system with known nonlinearities [133, 192, 235]. However, the question of providing learning and control guarantees for unknown nonlinear systems is currently less explored due to its difficulty in modeling the system dynamics and control design.

In this chapter, we take the first steps toward providing finite-time learning and control guarantees in the adaptive control of unknown nonlinear dynamical systems. We study partially observable nonlinear dynamical systems, where the learning agent has only access to the system outputs. In particular, we consider two function classes for the system dynamics: systems that live in Reproducing Kernel Hilbert Spaces (RKHS) or Sobolev space of periodic functions. For these settings, we provide novel system identification methods with finite-time learning guarantees and adaptive control methods with state-of-the-art real-world performance, as well as regret and stability guarantees.

For nonlinear systems in RKHS, we use Random Fourier Features (RFF) [226] to represent and learn the underlying system up to a confidence interval with optimal estimation error rate. In the analysis of this system identification method, we derive a novel function approximation theoretic guarantee for RFF which proves that the best RFF approximation of a nonlinear system has an approximation error of $\tilde{O}(1/\sqrt{D})$,

[1] This chapter is based on [164, 167, 168].

where D is the dimension of RFF representation. Using this method, we propose an efficient online control framework, Model Learning Predictive Control (MLPC), that learns to control the unknown system and minimizes the overall control cost. Once a reliable estimate of the dynamics is obtained through RFF-based system identification, MLPC deploys a model predictive control (MPC) method with the estimated system dynamics for planning. MLPC occasionally updates the underlying model estimates and improves the accuracy and effectiveness of the MPC policies. We provide stability guarantees for single trajectory online control and show that MLPC attains $\tilde{O}(T^{2/3})$ regret after T time steps in online control of stable partially observable nonlinear systems against the controller that uses the same MPC oracle with the true system dynamics. We empirically demonstrate the performance of MLPC on the inverted pendulum task and show the flexibility of the proposed general framework via deploying different planning strategies for the controller design to achieve low-cost control policies.

For nonlinear systems in the Sobolev space of periodic functions, we use the Fourier basis to represent and learn the underlying system. We show that using an nth order Fourier basis, our model learning approach estimates the underlying system with the near-optimal estimation error rate of $\tilde{O}(T^{\varepsilon-0.5})$, after T samples where ε depends on the smoothness of the Sobolev space and the order n, such that $0 \leq \varepsilon < 0.5$. Using this model learning approach, we propose an efficient model-based RL algorithm, **F**ourier **A**daptive **L**earning and **Con**trol (FALCON), for online control of unknown partially observable nonlinear dynamical systems.

We study FALCON for disturbance rejection under unknown extreme turbulence. We show that FALCON allows effective modeling and control of the aerodynamic forces due to turbulent flow dynamics and achieves state-of-the-art disturbance rejection performance. FALCON builds on two key observations that the chaotic dynamics involved in turbulent flows are well-modeled in the frequency domain and that most of the energy in turbulent flows is stored in low-frequency components. To this end, FALCON cleverly chooses a concise Fourier basis for learning the underlying system dynamics only using 35 seconds of flow data. To overcome the problem of partial observability due to sensor measurements, FALCON uses a short history of actions and measurements to model the system dynamics. With this physically sound and accurate model learning approach, FALCON deploys a model predictive control (MPC) method for safe and efficient control design. When evaluated under highly turbulent wind conditions generated in Caltech closed-loop wind tunnel,

Table 6.1: Comparison of Works with Regret Guarantees in Nonlinear Systems

Work	Regret Result	Learning Basis	Computational Efficiency	Memory Efficiency
Kakade et al. [133]	$\sqrt{T}$	Known	No	No
Boffi et al. [34]	$\sqrt{T}$	Known	Yes	Yes
MLPC	$T^{2/3}$	Unknown	Yes	Yes
FALCON	$\sqrt{T}$	Unknown	Yes	Yes

FALCON learns the underlying nonlinear dynamics and adapts to the changing flow conditions with less than 9 minutes of data and consistently outperforms the state-of-the-art methods. In addition to strong empirical performance, FALCON comes with learning and performance guarantees which certify the stability and robustness of the proposed framework. In particular, we show that FALCON attains $\mathcal{O}(\sqrt{T})$ regret against the agent who has access to the underlying dynamics and uses the same control design. To the best of our knowledge, FALCON is the *first* efficient RL algorithm that achieves $\mathcal{O}(\sqrt{T})$ regret in online control of nonlinear dynamical systems, Table 6.1.

Finally, to end this chapter, we consider the online stabilization of unknown nonlinear dynamical systems. Note that in both MLPC and FALCON, while deriving the performance guarantees, we assume that the deployed MPC method achieves stabilization under small enough modeling error. Even though this is verified in practice for the challenging settings we considered, it may not hold in general. To this end, we propose a novel policy optimization method that adopts Krasovskii's family of Lyapunov functions as a stability constraint. We show that solving this stability-constrained optimization problem using a primal-dual approach recovers a stabilizing policy for the underlying system even under modeling error. Combining this method with model learning, e.g., RFF-based system identification, we propose a model-based RL framework with formal stability guarantees, Krasovskii-Constrained Reinforcement Learning (KCRL). We theoretically study KCRL with RFF representation in model learning and provide a sample complexity guarantee to learn a stabilizing controller for the underlying system. Further, we empirically demonstrate the effectiveness of KCRL in learning stabilizing policies in online voltage control of a distributed power system. We show that KCRL stabilizes the system under various real-world solar and electricity demand profiles, whereas standard RL methods often fail to stabilize.

Motivation and Background

Reinforcement Learning (RL) has been recognized as a promising alternative for traditional decision-making and control tasks in engineering systems, e.g., robotics [178], energy systems [55], and transportation [297]. However, despite the promise, major hurdles remain before deployment in such systems is feasible. One of the key challenges is that many real-world systems are safety-critical and have high standards for stability, thus requiring finite-time guarantees. Even though RL algorithms outperform classical control methods in complex and uncertain dynamical environments, they often do not provide formal stability guarantees outside of simple systems, e.g., COCO-LQ in Chapter 4 for linear time-varying systems which are usual models for nonlinear dynamics with disturbances [225]. In particular, the most popular RL algorithms for the control of nonlinear systems follow model-free gradient-based policies that focus on minimizing the control cost and do not explicitly consider stability or regret performance [185].

As displayed throughout this thesis, the model-based methods hold promise to achieve significant empirical success while providing sample complexity guarantees for learning and control. Moreover, most real-world systems are governed by physics, which can be incorporated into model learning. This also enables them to generalize better to out-of-distribution samples, which is crucial in safety-critical tasks. However, despite these promises, most of the current model-based RL methods are rarely implemented in real-world systems due to their need for highly accurate models and the challenges that partial observability brings.

In most real-world dynamical systems, the system state is hidden, and instead, the controlling agent observes a nonlinear and noisy measurement of the state, e.g., through sensors. This partial observability brings uncertainties in modeling the system dynamics and designing the policies [60]. It also violates the common design assumption of the Markov property in the collected samples, which significantly complicates the modeling task [22]. These challenges make model-based RL remarkably difficult in real-world systems. To remedy these challenges, the majority of the model-based RL methods rely on the expressive power of deep neural networks (DNN) in modeling the dynamics. However, these approaches require a vast number of samples and usually yield black-box models, which are notoriously challenging to dissect and provide theoretical guarantees. Thus, these methods are most relevant in stationary and safe environments such as robotic manipulations [205]. However, under unsteady conditions such as complex turbulent flow fields, we require efficient

and adaptive modeling so that we can provide performance guarantees for generalizable learning. This lack of guarantees currently prevents the deployment of RL algorithms in real-world problems, where the dynamics are usually nonlinear and instabilities are costly, e.g., voltage instability in power systems [240].

Most of the model-based RL methods with guarantees are developed for linear systems due to their simplicity [56, 62, 85, 160, 166, 213, 242, 243, 270]. The central goal of these works is to derive finite-time learning and regret guarantees. Recently, there has been a growing interest to extend these results to nonlinear systems. [192, 235] consider the model learning problem by modeling the underlying system as a linear function of a *known* nonlinear basis. [133] study the regret minimization in this setting and propose an approach which attains $O(\sqrt{T})$ regret, but is *not* computationally or memory efficient. [34] studies a slightly different setting where the model dynamics are known but the system is subject to unmodeled disturbances and shows that their certainty equivalent controller attain $O(\sqrt{T})$ regret. Note that, these works' study of empirical performances is limited to simulations and they do not consider the challenges of real-world applications. Our goal is to improve upon these prior works in terms of regret guarantee and efficiency, as well as practical implementation. To the best of our knowledge, FALCON is the first efficient RL algorithm to attain $O(\sqrt{T})$ regret in partially observable nonlinear systems and achieve effective performance in a challenging real-world task.

6.1 Preliminaries

Notations: The Euclidean norm of a vector x is denoted as $\|x\|_2$. For a given matrix A, $\|A\|_2$ denotes its spectral norm, $\|A\|_F$ is its Frobenius norm, $A^\top$ is its transpose, $\mathrm{Tr}(A)$ is the trace, and $\sigma_{\min}(A)$ is the smallest singular value. I is the identity matrix with appropriate dimensions. $\mathcal{N}(\mu, \Sigma)$ denotes a multivariate normal distribution with mean vector μ and covariance matrix Σ. $x_{a:b}$ denotes the sequence of vectors between indices a and b, $[x_a, \ldots, x_b]$, where the index order can be increasing or decreasing depending on the choices of a and b.

6.1.1 Setting

Consider an unknown discrete-time partially observable nonlinear dynamical system

$$x_{t+1} = f(x_t, u_t) + w_t, \quad y_t = g(x_t) + z_t, \tag{6.1}$$

where $x_t \in \mathbb{R}^n$ is the state of the system, $u_t \in \mathbb{R}^p$ is the control input, $y_t \in \mathbb{R}^m$ is the output of the system, $w_t \in \mathbb{R}^n$ is the process noise, and $z_t \in \mathbb{R}^m$ is the

measurement noise. The system dynamics are governed by *unknown* nonlinear functions $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ which respectively live within some known Reproducing Kernel Hilbert Spaces (RKHS). The additive noise processes are independent at each time step. At the time t, the system is at state x_t and the agent observes y_t. Then, the agent applies a control input u_t, pays a cost $C_t(y_t, u_t)$ and the system evolves to x_{t+1} at time $t+1$. The sequence of cost functions is known a priori and we have the following assumption.

Assumption 6.1 (Cost Functions)**.** *For any y, y' and u, u' such that* $\max\{\|y - y'\|, \|u - u'\|\} \leq \Gamma$*, for all t,*

$$|C_t(y, u) - C_t(y', u')| \leq R(\|y - y'\|^2 + \|u - u'\|^2).$$

The following proposition states that the system in (6.1) can be equivalently represented in infinite-order nonlinear autoregressive (NARX) form.

Proposition 6.1.1 (System Equivalence)**.** *There exists an infinite-order NARX system with transition function $\bar{f}$ and measurement function $\bar{g}$:*

$$\bar{x}_{t+1} = \bar{f}(\bar{x}_t, u_t, y_t), \quad y_t = \bar{g}(\bar{x}_t) + e_t, \tag{6.2}$$

for some process e_t that depends on $f, g, \bar{f}, \bar{g}, w_t, z_t$, such that the input-to-output impulse response of the system in (6.1) *is equivalent to input-to-output impulse response of* (6.2)*.*

The proposition follows trivially using the "kernel trick" (which is further discussed in Section 6.1.4) to write nonlinear functions f and g in terms of linear mapping of some basis functions of feature maps and then shaping them up into new nonlinear functions $\bar{f}$ and $\bar{g}$ to satisfy the system dynamics. This proposition can be seen as the nonlinear counterpart of standard reparametrization from state-space representation to predictor form representation in LQG control systems given in Section 5.1. With this line of reasoning $\bar{x}_t$ can be thought of as the observer form state estimate of x_t whereas e_t is the innovation process, yet, clearly not i.i.d. due to nonlinearities.

Fading memory systems are the systems where the effects of past inputs on the output decay asymptotically [36]. In the identification of unknown nonlinear dynamical systems, they are usually considered due to the ability to approximate them arbitrarily close [64]. Note that finite-order nonlinear autoregressive (NARX) systems, i.e., nonlinear version of ARX systems studied in Sections 5.4.6 and 5.6.5, are a subset

of fading memory systems. Therefore, for simplicity of exposition and due to Proposition 6.1.1, we study finite order NARX systems:

$$y_t = F(u_{t-1}, \dots, u_{t-h}, y_{t-1}, \dots, y_{t-h}) + e_t, \tag{6.3}$$

where $F : \mathbb{R}^{hp+hm} \to \mathbb{R}^m$ is an unknown nonlinear function of past h inputs and outputs, i.e., model order is h, and $e_t \in \mathbb{R}^m$ are i.i.d. Gaussian, i.e., $e_t \sim \mathcal{N}(0, \sigma_e^2 I)$ for all t. Note that the systems with the model given in (6.3) are fundamental in many industrial applications and the isotropic assumption on e_t is for simplicity. In general for a stable infinite-order NARX system in (6.2), the dynamics given in (6.3) include an additive exponentially decaying term, which should be considered in a more general study. Therefore, our results can be extended to exponentially fading NARX systems without the finite order assumption.

Let $F_i(\cdot) : \mathbb{R}^{h(p+m)} \to \mathbb{R}$ denote the ith mapping of F from input to output, i.e., $y_{t,i} = F_i(\phi_t) + e_{t,i}$, for all $i \in 1, \dots, m$, where $\phi_i = [y_{i-1}^\top, \dots, y_{i-h}^\top, u_{i-1}^\top, \dots, u_{i-h}^\top]^\top \in \mathbb{R}^{h(p+m)}$. We have the following assumption on the function class of $F(\cdot)$.

Assumption 6.2 (Stable & Lipschitz System)**.** *The system $F(\cdot)$ is exponentially input-to-output stable (e-IOS), i.e., for $t > t_0$*

$$\| |\mathbb{E}[y_t | y_{t_0}, u_t, \dots u_{t_0}]\| \leq \lambda \alpha^{t-t_0} \|y_{t_0}\| + K \sup_{i \in [t_0:t]} \|u_i\|,$$

for $\lambda, K > 0$ and $0 < \alpha < 1$. Moreover, $F(\cdot)$ is L-Lipschitz.

The stability assumption is required to avoid output blow-up due to unmodeled system dynamics. Moreover, without the Lipschitz assumption, the noise term might affect the system in arbitrary ways, regardless of the input.

6.1.2 Control Problem

We will study the problem of online control of the unknown system given in (6.3). In the stochastic optimal control setting, the goal is to minimize the control cost starting from y_0, i.e.,

$$\min_{u_0, u_1, \dots, u_T} \mathbb{E}\left[\sum_{t=0}^{T} C_t(y_t, u_t) \,\middle|\, y_0 \right],$$

subject to dynamics given in (6.3) with initial condition of y_0 and where u_t is chosen causally. For nonlinear dynamical systems such as (6.3), finding the optimal solution to this problem is usually challenging [142]. As a practical and efficient

alternative, model predictive control (MPC) has been adopted for designing controllers in nonlinear dynamical systems [54]. In MPC, at any time step t, given the initial conditions, the transition dynamics (can be an estimated model $\hat{f}$), running and terminal cost functions $C_{t:t+\tau}(\cdot,\cdot)$, the planner solves:

$$\begin{aligned} &\min_{u_t,\ldots,u_{t+\tau}} \sum_{s=t}^{t+\tau} C_s(y_s, u_s) \\ &\text{s.t.} \qquad y_{t+1} = \hat{f}(u_t, \ldots, u_{t-h+1}, y_t, \ldots, y_{t-h+1}), \end{aligned} \tag{6.4}$$

a short τ-step optimal control problem, and executes the first action u_t and continues this process as it gathers new observations. Intuitively, instead of trying to solve the challenging global optimal control problem, MPC myopically solves a locally optimal control problem (6.4). Note that (6.4) presents an unconstrained MPC problem, and usually physical or safety constraints are added to the formulation. This makes MPC a viable approach for control design in model-based RL, thus, we will adopt it in our control design. In particular, we will utilize the sampling-based MPC called Cross-Entropy Method.

Cross-Entropy Method (CEM):

CEM is a sampling-based (zeroth-order) MPC policy to solve the problem given in (6.4) [35]. CEM maintains a distribution, predominantly Gaussian, to sample action roll-outs for the planning horizon and iteratively updates this distribution to assign a higher probability near lower-cost action sequences based on the estimated dynamics. After a certain number of updates (once it converges), it executes the first action on the lowest cost-achieving action sequence in the sampled roll-outs. The CEM algorithm is given in full detail in Algorithm 14. Despite the simple structure, one can show that CEM converges to a local optima [116]. However, similar to other sampling-based MPC methods, CEM can be computationally inefficient in high-dimensional control problems. To this end, one can use more efficient frameworks such as CEM-GD [117] which combines zeroth- and first-order optimization methods. Our studies in [117] show that incorporating the gradient information within the CEM framework improves the performance with 100× fewer samples per time step, resulting in around 25% less computation time and 10% less memory usage. Moreover, the local convergence guarantees of CEM can also translate to CEM-GD.

Algorithm 14 Cross Entropy Method (CEM)

Input: τ, K, M, $0 < \gamma < 1$, N, σ_{init}, $\hat{F}(\cdot)$, $y_{t:t-h+1}$, $u_{t-1:t-h+1}$, $C_{t:(t+\tau-1)}$, $\mathcal{N}(\mu, \sigma^2 I)$
for $i = 1, 2, \ldots, M$ **do**
  if $i = 1$ **then**
    Set the mean μ to the best action sequence from the previous time-step by shifting (Warm-Start)
    Set the variance $\sigma = \sigma_{init}$
  Sample $K\ \gamma^{i-1}$ action sequences $u^j_{t:t+\tau-1}$ of τ length using $\mathcal{N}(\mu, \sigma^2 I)$, $j \in \{1, \ldots, K\ \gamma^{i-1}\}$
  Compute the trajectory roll-outs $\forall u^j_{t:t+\tau-1}$ using $\hat{F}(\cdot)$ with initial $y_{t:t-h+1}$, $u_{t-1:t-h+1}$
  Compute the cost of each trajectory roll-out using $C_{t:(t+\tau-1)}$
  Sample best N action sequences according to their acquired costs
  Update μ and σ to fit the Gaussian distribution to the best N action sequences
Execute the first action of (i) the best action sequence of the Mth iteration or (ii) a newly sampled action sequence using $\mathcal{N}(\mu, \sigma^2 I)$

6.1.3 Regret

Similar to previous chapters, we evaluate the performance of our adaptive control algorithms by their regret. However, in this setting, we consider a slightly different regret definition and compute the regret with respect to the policy $\pi_\star$ which uses the MPC oracle at each time step with the true transition dynamics F to design its control inputs. Thus, the goal of the online control algorithm is to minimize the following:

$$\textsc{Regret}(T) = \sum\nolimits_{t=1}^{T} \left(C_t(y_t, u_t) - C_t(y_t^{\pi_\star}, u_t^{\pi_\star}) \right), \tag{6.5}$$

after T time steps of interaction with (6.3). Since the learning agent does not know the underlying dynamics f, it requires learning it from the data collected in the system. Notice that in the prior chapters of this thesis, the standard basis has been used in learning the underlying linear dynamical systems. In this chapter, learning the dynamics of the underlying system is done on two different nonlinear bases, which allows characterizing the prior knowledge of the system dynamics and deriving finite-time learning guarantees as discussed shortly.

6.1.4 Random Fourier Features (RFF)

Kernel methods are powerful tools used in modeling complicated functional relationships in many problems in machine learning. They are mainly built upon the kernel trick, *i.e.*, for some positive definite kernel $\kappa(\cdot, \cdot)$, the kernel evaluation of data points x_1 and x_2 are equivalent to the inner product between possibly infinite-

dimensional feature representations $\psi(\cdot)$ of the data points in a Hilbert space $\mathcal{H}$: $\kappa(x_1, x_2) = \langle \psi(x_1), \psi(x_2) \rangle_{\mathcal{H}}$, and the representer theorem in Kimeldorf and Wahba [144]. Given collected data pairs $\mathcal{D} = (x_i, y_i)_{i=1}^n$ for $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}$, kernel methods allow construction of nonlinear models as $\hat{f}(\cdot) = \sum_{i=1}^n \alpha_i \kappa(x_i, \cdot)$, such that α_i are chosen to explain the relationship in $\mathcal{D}$ for some positive definite kernel κ. However, for a large amount of data, solving for α is computationally expensive. For this reason, Rahimi and Recht [226] proposed to approximate the kernel, in particular, $\psi(\cdot)$, with finite-dimensional features $z(\cdot)$, named as Random Fourier Features, which provides an unbiased estimate of kernel κ. More formally, they defined RFF as a D-dimensional feature representation $z(x)$ of x such that

$$z(x) \coloneqq \sqrt{\frac{2}{D}} \begin{bmatrix} \cos\left(\omega_1^\top x + b_1\right) \\ \vdots \\ \cos\left(\omega_D^\top x + b_D\right) \end{bmatrix}, \tag{6.6}$$

where ω_i are drawn i.i.d. from distribution $p(\omega)$ which is the normalized Fourier transform of the kernel κ, and b_i are drawn i.i.d. from uniform distribution on $[0, 2\pi]$. This finite-dimensional construction motivated the use of RFF for function approximation in practice [133]:

$$\hat{f}(\cdot) = \sum_{i=1}^n \alpha_i \langle \psi(x_i), \psi(\cdot) \rangle_{\mathcal{H}} \approx \sum_{i=1}^n \alpha_i z(x_i)^\top z(\cdot), \tag{6.7}$$

and achieved significant empirical success in supervised learning setting [65, 119] (see [186] for a review). [226] showed that the approximated kernel converges to the true kernel exponentially fast in the number of features D. In this work, we provide a theoretical guarantee to approximate functions that live in an RKHS of a known kernel using RFF. In particular, we have the following assumption and theorem.

Assumption 6.3 (Nonlinear system in an RKHS)**.** *$F_i(\cdot)$ lives within the RKHS of infinitely smooth functions defined by a known positive definite continuous kernel $\kappa(\cdot, \cdot)$, e.g. Gaussian kernel, for all $i \in 1, \ldots, m$.*

Theorem 6.1 (Function Approximation Theory of RFF)**.** *Suppose $F : \Omega \to \mathbb{R}^m$ for $\Omega \subset \mathbb{R}^{h(p+m)}$ and Assumption 6.3 holds. For a given choice of D, let $\bar{F}(\cdot) = \Theta_* z(\cdot)$ be the best D-dimensional RFF approximation of F for $\Theta_* \in \mathbb{R}^{m \times D}$. Then, for some bounded region of state $\|\phi\| \leq \Gamma$, that depends on the function properties, with high probability we have $\sup_{\|\phi\| \leq \Gamma} \|\bar{F}(\phi) - F(\phi)\| \leq \tilde{O}(1/\sqrt{D})$. Here $\tilde{O}(\cdot)$ denotes the order up to logarithmic factors of D and hides the dependencies on n and Γ.*

The proof can be collected from the extended version of [164] online. At a high level, the proof combines the spectral convergence guarantee for kernel approximation in Theorem 6.1 of Rieger and Zwicknagl [229] which shows that the approximation error of a function that satisfies Assumption 6.3 decays exponentially with the number of kernel evaluations and a union bound argument over Claim 1 of [226]. This result will be the key to the theoretical guarantees of our adaptive control algorithm MLPC, in particular, to derive the finite-time learning and stabilization guarantees in Section 6.2.

Remark 6.1. *Note that Theorem 6.1 derives an approximation guarantee within a bounded region for a vector-valued nonlinear function that lives in a known RKHS. This setting is significantly more general than the kernelized nonlinear systems considered in [133, 192], which restricts the systems to be characterized as a finite sum of kernel evaluations. In our result, we do not make this restrictive assumption, instead, we consider the function class of infinite sum and study the explicit approximation error due to finite kernel evaluations.*

6.1.5 Fourier Series

A Fourier series is an expansion of a periodic function in terms of an infinite sum of complex exponentials, or sines and cosines. They are one of the most popular choices of the set of basis in representing periodic functions or periodic extensions of functions in a bounded domain due to their ability to approximate functions arbitrarily [45]. Consider the domain $\Omega = (0, 2\pi)^d$ in $\mathbb{R}^d$. Let $W_p^{m,2}(\Omega)$ denote the Sobolev space of order m for periodic functions. For a nonlinear function (or *its periodic extension*), $\bar{F}(\cdot) : \mathbb{R}^d \to \mathbb{R}$, that lives in $W_p^{m,2}(\Omega)$, one can write its Fourier series as

$$\bar{F}(x) = a_0 + \sum\nolimits_{\omega} \left[a_\omega \cos\left(\omega^\mathsf{T} x\right) + b_\omega \sin\left(\omega^\mathsf{T} x\right) \right], \tag{6.8}$$

where $\omega = [\omega_1, \ldots, \omega_d]$, $\omega_j \in \{1, 2, \ldots\}$, $1 \leq j \leq d$ and a_ω, b_ω are Fourier series coefficients. Note that this representation can be on infinitely many bases. However, in approximating $\bar{F}(\cdot)$, one can choose only a finite number of basis among ω and find the best approximation on this basis. To this end, the popular choice is to consider the nth order Fourier expansion and approximate $\bar{F}(x)$ in ω where $\omega_j \in \{1, \ldots, n\}$. This corresponds to $D = 1 + 2n^d$ basis functions and results in a D-dimensional Fourier series feature representation:

$$\phi(x) = [1, \cos(\omega_1^\mathsf{T} x), \sin(\omega_1^\mathsf{T} x), \ldots, \cos(\omega_{(D-1)/2}^\mathsf{T} x), \sin(\omega_{(D-1)/2}^\mathsf{T} x)]^\top. \tag{6.9}$$

One can choose the truncated Fourier series representation to approximate $\bar{F}(x)$ such that for

$$\mathbf{w} = [a_0, a_{\omega_1}, b_{\omega_1}, ..., a_{\omega_{(D-1)/2}}, b_{\omega_{(D-1)/2}}]^\top,$$

the approximation is $\mathbf{w}^\top \phi(x)$. However, this does *not* correspond to the best approximation in this basis in L^p-norms for $1 \leq p \leq \infty$ [45]. For the best L^p-norm approximation using (6.9), one needs to solve for the optimal coefficients $a_0, a^*_{\omega_i}$ and $b^*_{\omega_i}$ for $i \in \{1, \ldots, (D-1)/2\}$. Under the following assumption of the system dynamics, we have the approximation theoretic guarantee on using order-n Fourier basis in representing the underlying system.

Assumption 6.4 (Nonlinear system in Sobolev space of periodic functions)**.** *$F_i(\cdot)$ (or its periodic extension) lives in $W_p^{k,2}([0, 2\pi]^{h(m+p)})$, i.e., Sobolev space of order k for periodic functions, for all $1 \leq i \leq m$.*

Theorem 6.2 (Theorem 4.3 of [237]))**.** *Suppose $F : \Omega \to \mathbb{R}^m$ for $\Omega \subset \mathbb{R}^{h(p+m)}$ and Assumption 6.4 holds. Let $\theta_*^{[i]\top}\phi(\cdot)$ denote the best Fourier series approximation of F_i, using the nth order Fourier expansion for large enough n, i.e., D-dimensional Fourier basis $\phi(\cdot)$. Then, we have*

$$\sup_{\|s\| \leq S} \left| F_i - \theta_*^{[i]\top}\phi(s) \right| \leq Cn^{-k} \|\partial^k F_i(\cdot)\|_{L^\infty}.$$

for problem-dependent constant $C = poly(S, e^m, n)$, where $\|\partial^k F_i(\cdot)\|_{L^\infty}$ determines the smoothness of the system.

This result is the multivariate analog of Jackson's Theorem [124]. The exact values of required n and C can be collected from Sections 2 and 3 of [237]. This result will be the key to the theoretical guarantees of our adaptive control algorithm FALCON, in particular, to derive the finite-time learning and stabilization guarantees in Section 6.3.

6.2 Model Learning and Control with Random Fourier Features

In this section, we study online learning and control of an unknown partially observable nonlinear dynamical system given in (6.3) and design a sample efficient and practical RL framework with finite-time learning to control guarantees. Using Proposition 6.1.1, we propose our framework, Model Learning Predictive Control (MLPC), for practical, finite order, stable, and smooth NARX systems.

MLPC uses kernel-based feature representation methods, i.e., Random Fourier Features (RFF) [226], to represent the unknown nonlinear dynamics. It deploys uniform

exploratory inputs to excite the underlying system and gather information about the system dynamics. Using the data collected, MLPC tackles the system identification problem via RFF representation and solves a regularized least-squares problem to estimate the underlying system with finite-time guarantees. The proposed estimation method works under both open-loop and closed-loop controllers as long as the inputs persistently excite the system. This allows for model estimate updates while controlling the system. In the analysis, we use the novel function approximation theoretic guarantee for RFF given in Theorem 6.1 which proves that the best RFF approximation of a nonlinear system has an approximation error of $\tilde{O}(1/\sqrt{D})$, where D is the dimension of RFF representation. Then, we show that under the persistence of excitation, the estimation error of the underlying nonlinear system attains the rate of $\tilde{O}(1/\sqrt{D} + \sqrt{D/T})$ after T samples.

In order to obtain a practical online control algorithm, the MLPC framework uses an MPC oracle to design the adaptive controller. In particular, MLPC provides the estimated system dynamics to the MPC oracle, which uses it to design the control inputs. In our analysis, we compare the performance of MLPC to the controller, π_*, that uses the same MPC oracle with the true system dynamics. We show that if the true model is well-approximated by the estimated model, then the controllers designed by the MPC oracle via estimated dynamics achieve comparable trajectory and cost with respect to π_*. Thus, we show that using all data gathered and occasionally updating the model estimates, MLPC attains regret upper bound of $\tilde{O}(T^{2/3})$.

Finally, we deploy the MLPC framework in online control of the inverted pendulum task. We use two different sampling-based MPC oracles in planning: MPPI [294] and CEM [35]. We show that in both variants MLPC quickly learns the underlying unknown nonlinear system which yields nearly optimal MPC policies and low regret. This demonstrates the flexibility and the modularity of the proposed framework which allows tuning the algorithm for task-specific constraints.

6.2.1 Model Learning Predictive Control Framework

In this section, we present Model Learning Predictive Control (MLPC), an efficient online learning and control algorithm that learns the model dynamics through interaction with the system and deploys an MPC oracle-based controller using the learned model. The pseudo-code of MLPC is provided in Algorithm 15. It has two phases: Exploration and Adaptive Control.

Exploration: MLPC starts with an exploration period of length T_w time-steps to

Algorithm 15 MLPC

Input: $T, T_w, h, t_{up}, t_p, D, n, m, p, C_{0:T}$
—— Exploration ——
for $t = 1, 2, \ldots, T_w$ **do**
  Deploy P.E. inputs u_t and store $\mathcal{D}_0 = \{y_t, u_t\}_{t=0}^{T_w}$
—— Adaptive Control ——
Form $\phi_t = [y_{t-1:t-h}^\top, u_{t-1:t-h}^\top]^\top$ using $\mathcal{D}_0$ for $h \le t \le T_w$
Compute $z(\phi_t)$, D-dim RFF vector for all ϕ_t
for $i = 0, \ldots$ **do**
  Solve (6.12) for $\hat{\Theta}_i$ & Form $\hat{F}_i(\cdot) = \hat{\Theta}_t^\top z(\cdot)$
  for $t = T_w + it_{up} + 1, \ldots, T_w + (i+1)t_{up}$ **do**
    $u_t = \text{MPC-Oracle}\left(\hat{F}_i, y_{t:t-h+1}, u_{t-1:t-h+1}, C_{t:(t+t_p)}\right)$
    Observe y_{t+1} & Form ϕ_{t+1} and $z(\phi_{t+1})$

collect some data about the unknown system. Due to Assumption 6.2, MLPC uses bounded, persistently exciting inputs, without aiming to control the system. The goal is to guarantee an accountable first estimate of the model for reliable controller design. In Section 6.2.2, we discuss the choice of T_w in order to provide finite-time estimation and controller guarantees. Note that in practice if Assumption 6.2 does not hold, MLPC can use a known stabilizing controller of the system to safely collect data.

Adaptive Control: After the exploration period, MLPC starts controlling the underlying system. It operates in epochs with user-defined parameter t_{up}, *i.e.*, each epoch lasts for t_{up} time steps. Note that, unlike the standard RL setting, MLPC is a *single* trajectory algorithm, i.e., there is no reset at the end of epochs. At the beginning of each epoch, MLPC uses the history of interactions with the system to identify the system dynamics. In this regard, in epoch k, it constructs subsequences of length h input-output pairs using *all* collected data $\mathcal{D}_k = \{y_t, u_{t-1}, y_{t-1}, \ldots, u_0, y_0\}$:

$$\phi_i = [y_{i-1}^\top, \ldots, y_{i-h}^\top, u_{i-1}^\top, \ldots, u_{i-h}^\top]^\top \in \mathbb{R}^{h(p+m)}, \tag{6.10}$$

for all $h \le i \le t$ where $t = T_w + (k-1)t_{up}$. From the system model given in (6.3), we know that $y_i = F(\phi_i) + e_i$. Using the known kernel κ and the RFF generation procedure described in Section 6.1.4, MLPC computes D-dimensional RFF representation of ϕ_i: $z(\phi_i)$ for all i. The number of features D is a user-specified parameter, which can be adjusted depending on the difficulty of the learning and control task, as well as the computational budget. As Section 6.2.2 demonstrates, the choice of D also brings a theoretical trade-off between approximating the nonlinear system dynamics, system identification, and the regret guarantees.

Equation (6.7) shows that $F(\phi_i)$ is linear in $\psi(\phi_i)$, the possibly infinite-dimensional feature representation of ϕ_i, which can be approximated via $z(\phi_i)$ to obtain

$$y_i \approx \Theta_*^\top z(\phi_i) + e_t, \tag{6.11}$$

for some unknown $\Theta_* \in \mathbb{R}^{D\times m}$. Thus, MLPC considers the model in (6.11) for system identification. At the beginning of epoch k, MLPC obtains an estimate of Θ_* by solving the following regularized least squares problem:

$$\min_{\Theta} \lambda\|\Theta\|_F^2 + \sum_{i=h}^{t} \|y_i - \Theta^\top z(\phi_i)\|_2^2, \tag{6.12}$$

for some $\lambda > 0$. The closed-form solution of (6.12) is given as

$$\hat{\Theta}_k = (Z_t Z_t^\top + \lambda I)^{-1} Z_t Y_t^\top, \tag{6.13}$$

where $Y_t = [y_t, \ldots, y_h] \in \mathbb{R}^{m\times N}$, $Z_t = [z(\phi_t), \ldots, z(\phi_h)] \in \mathbb{R}^{D\times N}$ for $N = t - h + 1$. Using $\hat{\Theta}_k$, MLPC forms an estimate of the system dynamics as $\hat{F}_k(\cdot) = \hat{\Theta}_k^\top z(\cdot)$. This system identification process is repeated to obtain improved estimates of (6.3) at the beginning of each epoch, i.e., in every t_{up} time steps. As mentioned, the update frequency of t_{up} is a user-specified parameter and it can vary according to computational complexity and task. Moreover, instead of using the closed-form solution given (6.13), the model estimate updates can be done via batch or online updates using the standard linear regression techniques.

Once MLPC has an estimated system dynamics at the beginning of the epoch, it uses an MPC oracle to design the control inputs during the epoch. Let t_p denote the planning horizon of the MPC oracle. At any time step t, MLPC provides the recent estimated system dynamics, $\hat{F}_k(\cdot)$, last h input-output pairs as the initial state, $(u_{t-1:t-h+1}, y_{t:t-h+1})$, and the next t_p cost functions, $C_{(t+t_p):t}$, to the MPC oracle. Upon receiving these, the MPC oracle solves the t_p time step optimal control problem for the given system with initial conditions and it returns the control action u_t to be taken by MLPC on the underlying system $F(\cdot)$. In the analysis and practical implementations, efficient oracles are considered, e.g., constrained optimization based [76] or sampling-based [35, 294] methods. Note that if there are any control constraints, MLPC can include them in its query to the MPC oracle. Upon deploying the received control input u_t, the system gives output y_{t+1}. Using u_t and the history of input-output pairs, MLPC constructs ϕ_{t+1} and $z(\phi_{t+1})$. This control input generation process repeats until the end of the epoch.

6.2.2 Regret Analysis

In this section, we will provide the learning and regret guarantees of MLPC. We will first discuss the finite-time system identification guarantee of Θ_* via the exploration phase. This result will be used to show that after sufficient exploration, the NARX system is well-approximated. We have the following assumption in the exploration phase.

Assumption 6.5 (Exploratory Inputs)**.** *We have access to a set of bounded persistently exciting (PE) inputs that can be used for exploration and excite the system uniformly. In other words, the smallest eigenvalue of the design (sample covariance) matrix $Z_t Z_t^\top$ scales linearly over time.*

This assumption is fairly standard and it guarantees the consistent and reliable estimation of the underlying system.

6.2.3 Learning System Dynamics

Since RFF representation turns the NARX system into a linear system form in (6.11), we can use the analysis of online linear least squares studied in Section 5.3.

Lemma 6.1. *Let $\hat{\Theta}_1$ be the solution to* (6.12) *at the end of exploration phase, i.e., at $t = T_w$. Let $V_t = \lambda I + Z_t Z_t^\top$ and $\|\Theta_*\|_F \leq S$. For $\delta \in (0, 1)$, with probability $1 - \delta$, we have*

$$\mathrm{Tr}((\hat{\Theta}_1 - \Theta_*) V_t (\hat{\Theta}_1 - \Theta_*)^\top) \leq \beta_t^2, \tag{6.14}$$

where $\beta_t = \sigma_e \sqrt{mD \log\left(\frac{1+2t/\lambda D}{\delta}\right)} + \sqrt{\lambda} S$ since $\|z(\phi)\|_2^2 \leq 2$ due to RFF construction. If Assumption 6.5 holds, then

$$\|\hat{\Theta}_1 - \Theta_*\| = \tilde{O}(\sqrt{D/T_w}).$$

Lemma 6.1 shows that the estimation error on Θ_* decays with the optimal $1/\sqrt{t}$ rate and with $\sqrt{D}$ scaling from the number of RFF. This result shows that as the number of features increases, the estimation of the best linear system, Θ_* becomes harder. However, the number of features shouldn't be too small, since it directly affects how well the underlying nonlinear function is approximated. This brings a trade-off in the number of features, which is soon analyzed in the approximation of the underlying NARX system and obtaining a reliable system model over time. The following result trivially combines Lemma 6.1 with the approximation theoretical guarantee on RFF approximation, i.e., Theorem 6.1.

Corollary 6.2.1. *Suppose Assumptions 6.2 and 6.3 hold. For a given choice of D, for some bounded region of state $\|\phi\| \leq \Gamma$, that depends on the function properties λ, K, α, using the estimate of $\hat{\Theta}_1$ to construct an estimate of the underlying system $\hat{F}_1(\cdot) = \hat{\Theta}_1^\top z(\cdot)$ at the end of the exploration phase, with high probability, we have*

$$\sup_{\|\phi\| \leq \Gamma} \|F(\phi) - \hat{F}_1(\phi)\| = O(1/\sqrt{D} + \sqrt{D/T_w}), \tag{6.15}$$

6.2.4 Boundedness of state

Next, we show that the MPC oracle keeps the state bounded during the adaptive control period using the refined estimate of the system dynamics according to Corrollary 6.2.1. To this end, we provide the following condition on the MPC oracle which allows us to quantify stabilization behavior and the regret guarantee of MLPC.

Assumption 6.6 (MPC Oracle)**.** *The MPC oracle that uses $F(\cdot)$ to design inputs achieves e-IOS closed-loop dynamics on the underlying system such that*

$$\|y_t\| \leq (1-\rho)^{t-t_0}\|y_{t_0}\| + L \sup_{i \in [t_0:t]} \|e_i\|,$$

for $t > t_0$, $L > 0$ and $0 < \rho < 1$. Moreover, for any NARX $\hat{F}(\cdot)$ such that $\sup_{\|\phi\| \leq B} \|F(\phi) - \hat{F}(\phi)\| \leq \epsilon$, the MPC oracle that uses $\hat{F}$ to design inputs also achieves e-IOS on the underlying system for $\rho' = \rho/2$ and $L' = 2L$ such that

$$\|y_t\| \leq (1-\rho')^{t-t_0}\|y_{t_0}\| + L' \sup_{i \in [t_0:t]} \|e_i\|,$$

and the inputs designed by the MPC oracle are L_o-Lipschitz in the planning model within this neighborhood.

The first statement says the MPC oracle stabilizes the underlying system. The second statement states that if the estimated model dynamics is within a neighborhood of the underlying system in a bounded region of inputs ϕ, then the MPC oracle stabilizes that system as well but at a slower rate. Note that this assumption is mild. For instance, as discussed in Chapters 3 and 5, one can show that this assumption holds for linear dynamical systems. Thus, intuitively, it holds for nonlinear systems where the validity of the linearization is defined via the bounded region of inputs. Based on Assumption 6.6, if the MPC oracle stabilizes the system then we obtain bounded outputs y_t. Thus, we further consider that the inputs generated by the MPC oracle are bounded in this stabilizing regime due to the choice of control costs $C(\cdot, \cdot)$ and the MPC oracle construction. With this, we are ready to state the required amount of

exploration and the number of features MLPC requires to maintain stable closed-loop dynamics in adaptive control.

Lemma 6.2. *For the ϵ defined in Assumption 6.6, if the number of RFF $D = O(1/\epsilon^2)$ and the warm-up duration $T_w = O(1/\epsilon^4)$, then MLPC using the transition dynamics $\hat{F}_1 = \hat{\Theta}_1^\top z(\cdot)$ in the MPC oracle achieves e-IOS with ρ' and L'.*

The result follows directly from optimizing D and T_w for ϵ error given in Assumption 6.6. Note that with this new e-IOS, we can bound the outputs of the system, thus the covariates ϕ, during the adaptive control phase similar to the exploration phase, i.e., $\|\phi_t\| \leq \Gamma'$ for Γ' determined by σ_e, L', ρ' for $t > T_w$.

6.2.5 Regret Guarantees

Once the stabilization of MPC oracle with the estimated system dynamics is satisfied, MLPC can safely deploy the control inputs which are bounded as described and attain bounded outputs. Finally, we have the following assumption on the inputs designed by the MPC oracle.

Assumption 6.7. *For any nonlinear system $\hat{F}(\cdot)$ such that $\sup_{\|\phi\| \leq B} \|F(\phi) - \hat{F}(\phi)\| \leq \epsilon$, the MPC oracle that uses $\hat{F}$ to design inputs persistently excites the underlying system F.*

In practice, the condition above can be satisfied by the combination of unmodelled system dynamics and system noise. Moreover, using sampling-based MPC oracles to design inputs would also improve the randomness in the dynamics which may contribute toward Assumption 6.7. In particular, in our experiments in Section 6.2.6, we use sampling-based MPC oracles and we observe that the system identification errors decrease consistently indicating the validity of Assumption 6.7. The following states the regret upper bound of MLPC.

Theorem 6.3. *Let T_w be chosen such that it satisfies the condition in Theorem 6.2. If Assumption 6.7 holds, then for $D = \max\{O(1/\epsilon^2), O(T^{1/3})\}$ MLPC attains $\textsc{Regret}(T) = \tilde{O}(T^{2/3})$.*

The proof can be collected from the extended version of [164] online. At a high level, it combines all the results derived in Section 6.1 and the current section. We first show the closeness of the trajectory of inputs and outputs generated by MLPC that uses the estimated system dynamics to the trajectory of MLPC with the true

underlying system dynamics using the stability of the designed controllers. We then invoke Assumption 6.1 and show that regret of MLPC scales quadratically with $\sup_{\|\phi\|\leq B} \|F(\phi) - \hat{F}_t(\phi)\| = O(1/\sqrt{D} + \sqrt{D/t})$, since the adaptive control inputs satisfy the PE condition through Assumption 6.7, we get

$$\textsc{Regret}(T) = T_w + \frac{T}{D} + D\log(T) + \sqrt{T}.$$

From Lemma 6.2 with $D = \max\{O(\frac{1}{\epsilon^2}), O(T^{\frac{1}{3}})\}$ gives the advertised result.

6.2.6 Experiments

In this section, we present our empirical study of MLPC. We evaluate the MLPC framework on the inverted pendulum task from OpenAI Gym [39] using two different sampling based MPC oracles: MPPI [294] and CEM [35]. Note that these MPC oracles are computationally efficient and can handle non-convexities in the system dynamics and the cost functions. Moreover, they can be parallelized to sample large amount of trajectories to achieve improved performance. For each variant, Algorithm 15 is implemented and the MPC oracle is called for planning at each time-step, i.e., generate the control input. MLPC is run for 1000 time steps and with three randomly chosen seeds. The figures provide the cumulative cost and they are obtained by averaging over three runs.

The inverted pendulum task is an NARX system of order $h = 1$, where the observations are $m = 3$-dimensional, cosine and sine of the angle and the angular velocity, and the input is the torque applied at the contact point. We use RFF to represent the mapping from $\phi \in \mathbb{R}^4$ to the next observation. Note that with known dynamics, the optimal controller rapidly achieves zero instantaneous cost, thus the cumulative cost plots given below are equivalent to the regret behavior.

MLPC with MPPI Oracle

In the experiments, we tested the effect of number of RFF D on the regret and learning the underlying system. To this end, we run MLPC with $D = 10, 30, 100, 1000$. Figure 6.1 shows the results. It can be seen that 10 Random Fourier features are not enough to represent the underlying NARX. On the other hand, once enough RFF is provided MLPC learns to control the underlying system dynamics very quickly. Moreover, in parallel to the theoretical trade-off discussed in Section 6.2.2, there is a trade-off in D experimentally. Larger D values result in harder learning tasks for simple models and can obtain inferior performance. In Figure 6.1, it can be seen that the best performance is obtained with $D = 50$.

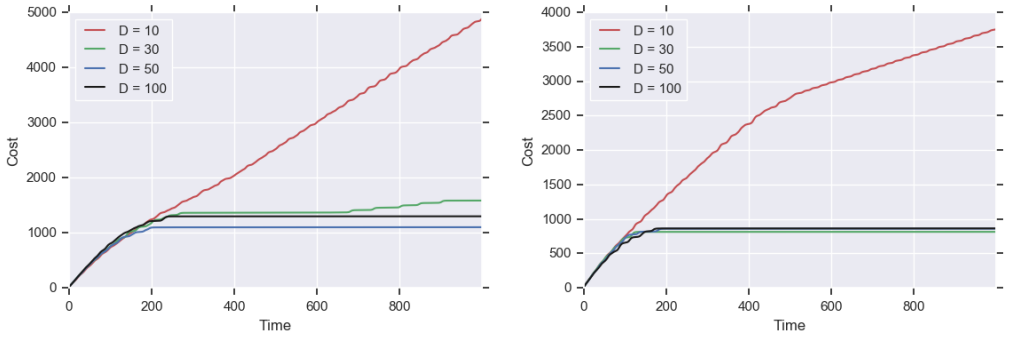

Figure 6.1: Cumulative Cost of MLPC with MPPI and CEM respectively for Different Number of RFF.

MLPC with CEM Oracle

Similar to the MPPI variant, $D = 10$ is not enough to represent the nonlinear dynamics and for other D values the NARX is learned rapidly. Note that the optimal performance is achieved with $D = 30$. This shows that the number of features can depend on the choice of MPC oracle in practice.

6.3 Fourier Adaptive Learning and Control for Disturbance Rejection Under Extreme Turbulence

Turbulent atmospheric winds often contain transient flow disturbances and aerodynamic forces that can affect a variety of systems and structures [32]. These forces are particularly significant for aerodynamic technologies like unmanned aerial vehicles (UAVs) and wind turbines, which rely on fluid interaction for regular operation and can be damaged when operating in turbulent conditions [136, 241, 290]. Developing active control strategies to mitigate the effects of these turbulent forces is one of the most important challenges in the safe deployment of UAV technologies or extending the lifetime and reliability of wind turbines [46, 109, 128]. Developing and utilizing complex flow models for control is challenging in real-time due to sensor noise, low-latency, and high-frequency control requirements [41, 115, 220].

Conventional control strategies for UAVs, such as proportional-integral-derivative (PID) controllers, are designed to reactively correct inertial deviations from the desired trajectory without taking into consideration the underlying flow dynamics or the source of the disturbance [103, 107, 250]. These approaches are often insufficient for maintaining stability in extreme atmospheric turbulence, which prevents

deployment of these technologies in safety-critical scenarios such as deploying unmanned aerial vehicles (UAV) in densely populated urban areas [113, 195, 200], such as Figure 6.2A.

In contrast, biological swimmers and flyers have the ability to directly observe and respond to the physics responsible for changes in motion [31, 73, 251, 266]. By drawing inspiration from these biological systems, there have been considerable efforts to improve control strategies for UAVs by using easily measurable flow quantities, such as pressure, to anticipate and mitigate the effects of turbulent disturbances [96, 149, 194, 201, 215, 239]. The majority of these works again utilize the flow-sensing information within PID control frameworks which limits their desirable performance to low velocities [96, 149, 194, 201] or they consider uniform wind/flow scenarios where the eddies and gusts have smaller scale than the UAVs which result in small aerodynamic disturbances [215, 239].

To tap into the potential of flow-sensing in designing disturbance rejection policies, reinforcement learning (RL), a machine learning area, has been recognized as a promising framework, due to its ability to learn and adapt to the unmodeled dynamics and design nonlinear policies with various objectives. Most of the prior works on RL for flow control have focused on model-free RL techniques and developed in computational fluid dynamic (CFD) simulations. Model-free RL methods do not construct an explicit model of the system dynamics and aim to learn the control policies directly through interactions with the system [255]. Therefore, they are the most intuitive choices for policy design in environments difficult to model such as turbulent flow dynamics. Among these model-free RL works, Bieker et al. [29] introduced a novel framework with online learning to predict and control flow in a 2D CFD simulation. Gunnarson et al. [99] introduced an algorithm to navigate a simulated "swimmer" across an unsteady flow in a 2D simulation. In the experimental studies, Fan et al. [80] demonstrated the first experimental applications of model-free RL in fluid mechanics. Recently, Renn and Gharib [228] used a model-free RL method for controlling the aerodynamic forces on an airfoil under turbulent flow in an experimental setting (similar to the one considered in this work) and achieved state-of-the-art disturbance rejection performance, outperforming PID control. They also documented that the power spectrum of the turbulent flow at high Reynolds numbers is dominated by the low-frequency components which inspired the development of the algorithm in this work. However, despite this strong empirical performance, their method suffers from well-known limitations of the model-free

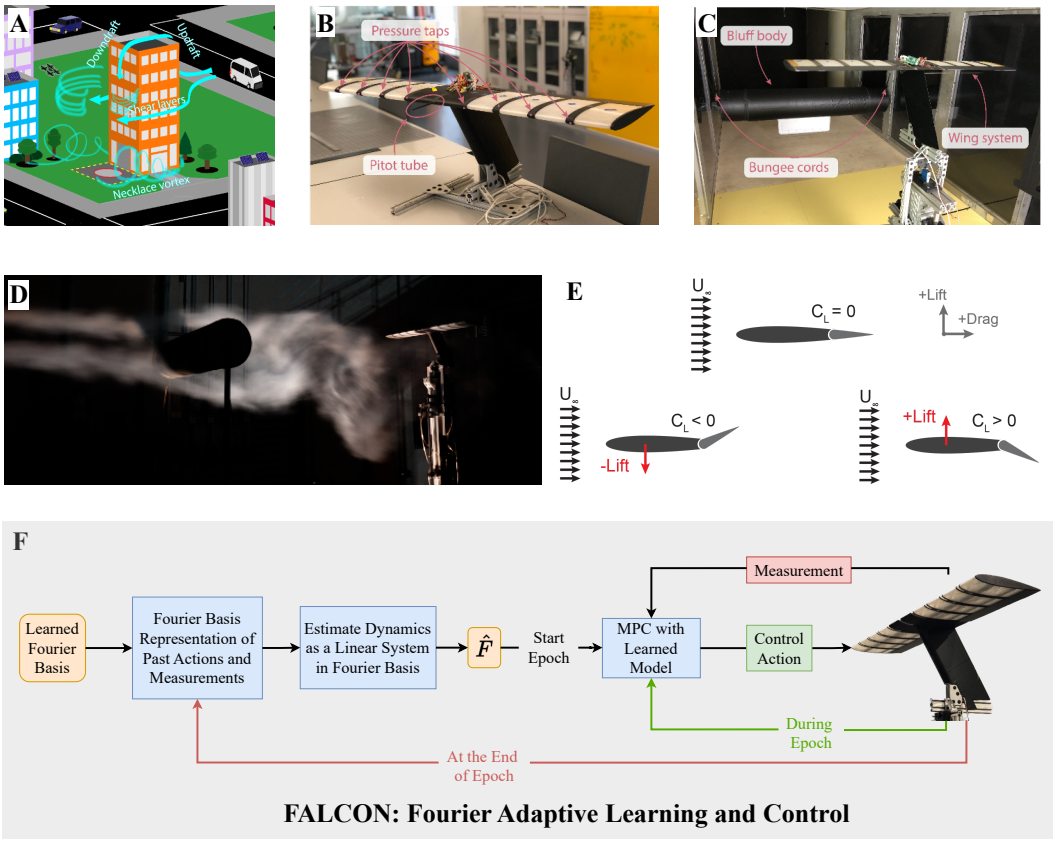

Figure 6.2: **(A)** Complex airflow structures in urban environments. **(B)** The wing has 9 sensors to measure the airflow (8 equally-spaced pressure taps and 1 pitot tube) and is mounted on a one-dimensional load cell to measure the lift. Trailing-edge flaps change orientation to manipulate the aerodynamic forces. **(C)** Experiment setup to create irregular turbulent wake of a bluff body under high wind speeds. **(D)** Smoke visualization of the turbulent wake of a cylinder at a smaller Reynolds number. This image is obtained at the Caltech Real Weather Wind Tunnel system at a significantly lower flow speed than the experiments conducted in this work for visualization purposes. The actual flow conditions used in our studies were too turbulent to have clear smoke visualization. **(E)** Under a uniform flow U_∞, symmetric airfoils do not have any vertical aerodynamic forces on them when they are aligned with the airflow. However, altering the position of a trailing edge flap on the airfoil can modify the lift coefficient C_L, yielding an upward or downward aerodynamic lift force. **(F)** Outline of FALCON, a model-based reinforcement learning framework that allows effective modeling and control of the aerodynamic forces due to turbulent flow dynamics and achieves state-of-the-art disturbance rejection performance.

RL methods, namely, extensive and laborious data collection, and brittle policies.

Overview of Contributions

In this section, we take on the challenge of designing a model-based RL framework for flow-informed aerodynamic control in a highly turbulent and vortical environment to overcome these limitations. We propose an efficient model-based RL algorithm, **F**ourier **A**daptive **L**earning and **Con**trol (FALCON), for online control of unknown partially observable nonlinear dynamical systems, in particular for disturbance rejection under unknown extreme turbulence (Figure6.2F). FALCON leverages the domain knowledge that the underlying turbulent flow dynamics are well-modeled in the frequency domain and that most of the energy in the turbulent flows is present in low-frequency components [222, 228]. Therefore, it learns the underlying partially observable system in a succinct Fourier Series basis. FALCON consists of two main parts: a warm-up phase and adaptive control in epochs phase. In the warm-up phase, using only a small amount of flow data (35 seconds-equivalent to approximately 85 vortex-shedding interactions), FALCON recovers a succinct Fourier basis that explains the collected data and enforces that this learned basis is mostly composed of low-frequency components following the prior observations on turbulent flow dynamics. It then uses this basis to learn the unknown linear coefficients that best fit the acquired data on the learned Fourier basis during the adaptive control phase.

In the control design, FALCON uses model predictive control (MPC) and efficiently solves a short-horizon planning problem at every time step with the learned system dynamics. This recurrent short-horizon planning approach allows FALCON to adapt to the sudden changes in the flow while designing more sophisticated policies that consider future flow effects in contrast to the conventional purely reactive controllers. Moreover, the simple yet physically accurate dynamics learning approach of FALCON further facilitates the effective control design, which results in a sample-efficient and high-frequency control policy. During the adaptive control phase, FALCON refines its model estimate, i.e., the linear coefficients, in epochs in order to improve the learned model, which in turn improves the performance of the MPC policy. Overall, FALCON provides a simple, efficient, and interpretable dynamics modeling and an adaptive policy design method for the flow-predictive aerodynamic control problem and significantly outperforms the state-of-the-art model-free RL methods and conventional control strategies, i.e., PID, using only a total of 9 min-

utes of training data representative of approximately 1300 vortex-shedding cycles. FALCON easily incorporates the physical and safety constraints in the policy design and builds on a fundamental understanding of how well nonlinear systems can be approximated and how these approximation errors affect the control performance, which we support with rigorous theory.

We implement FALCON on an experimental aerodynamic testbed that abstracts the fundamental physics involved in flight through a turbulent atmospheric flow and is specifically relevant for fixed-wing UAV applications. This testbed consists of a 3D-printed airfoil with actuated trailing edge flaps and an array of pressure sensors to measure the surrounding flow (Figure 6.2B). The system is mounted in a closed-loop wind tunnel on a load cell measuring the aerodynamic lifting force acting on the airfoil. The testbed is placed in the wake of a bluff body at a Reynolds number of $230,000$, which generates a highly turbulent and vortical environment (Figure 6.2C). The aerodynamic control goal is set to minimize the standard deviation of the lift forces by adjusting the position of the trailing-edge flaps in response to incoming disturbances with the help of flow sensors (Figure 6.2E). In free flight, this would be equivalent to minimizing the inertial deviations along the lifting axis.

Through these wind tunnel experiments, we report that FALCON achieves 37% better disturbance rejection performance than the state-of-the-art model-free RL method [228], using only a single trajectory and 8 times less data. Moreover, we document a performance improvement of 45% over the conventional reactive PID controller. Overall, we find that the superior performance of FALCON is consistent over independent runs in the highly irregular unsteady turbulent flow dynamics, demonstrating the adaptation and generalization capability of FALCON to the unseen conditions.

6.3.1 Results

In this section, we first explain the key insights behind the design of our model-based reinforcement learning algorithm, FALCON. Second, we discuss the experimental platform for data collection and highlight key aspects of the experiments. In particular, we present the flow-informed aerodynamic testbed and the characterization of the turbulent environment used in our aerodynamic control experiments. Finally, we discuss the training and test performance of FALCON and compare it with several state-of-the-art baseline methods. We observe that FALCON consistently outperforms the prior learning-based and conventional controllers by a significant margin

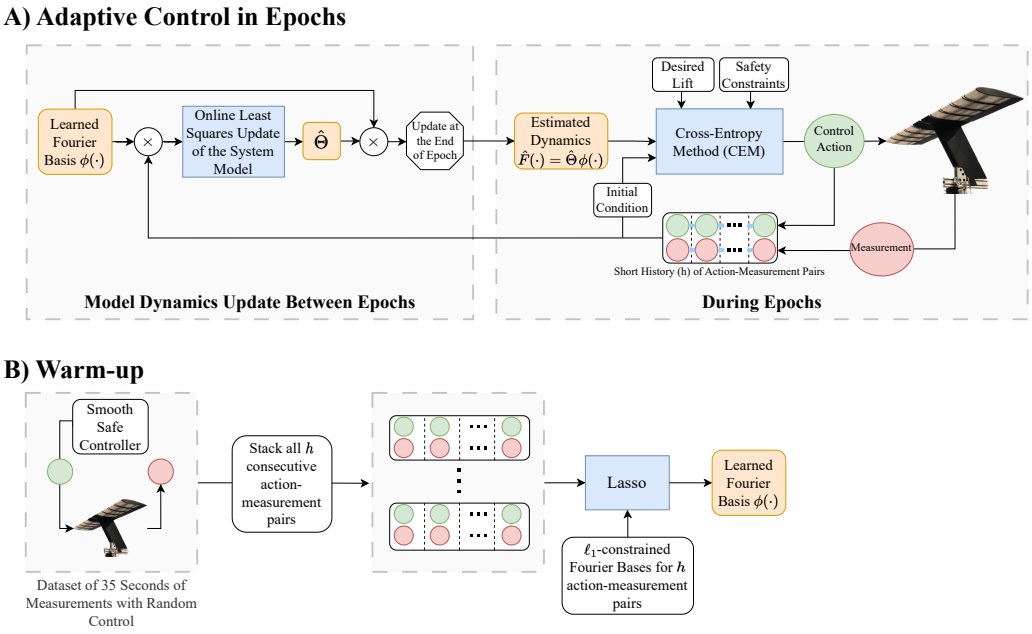

Figure 6.3: **FALCON Framework.** It consists of two phases: Warm-Up and Adaptive Control in Epochs. **(A) Adaptive Control in Epochs:** FALCON models the system dynamics as a linear map of the representation of a short history (h-length) of action-measurement pairs in the succinct Fourier basis learned in the warm-up phase. FALCON learns the unknown linear coefficients that best model the dynamics via online least squares. It updates the estimated system dynamics, i.e., the linear coefficients, at the end of each epoch, and during the epochs, it uses Cross-Entropy Method (CEM), a sampling-based MPC method, to control the airfoil under extreme turbulence using the estimated system dynamics while satisfying desired lift and safety requirements. **(B) Warm-up:** It is a one-time 35-second process before starting the adaptive control phase for safely collecting some exploratory data about the unknown system to recover a relevant Fourier basis to be used in learning and adaptive control. To achieve this, FALCON forms h-length subsequences of action-measurement pairs (a short history) from the safely collected dataset and solves the Lasso problem on the ℓ_1-constrained Fourier basis representation of these subsequences. FALCON selects the Fourier basis vectors that correspond to non-zero coefficients in the solution of the Lasso problem as the succinct Fourier basis $\phi(\cdot)$ for the entire adaptive control in epochs phase for learning and control of the system.

while requiring order-of-magnitude fewer samples.

Novel Model-based RL Framework: FALCON

FALCON has two main phases: a warm-up and an adaptive control in epochs phase (Figure 6.3). The warm-up phase is a short initial period, where FALCON collects initial data about the fully unknown system. The goal is to purely explore the system and recover a coarse model of the dynamics. To that end, FALCON executes smooth and safe actions, i.e., time-correlated Gaussian inputs, that safely excite the system (see Methods for other variants). Since FALCON relies on pressure sensors on the airfoil to measure the system dynamics, it operates under partial observability. To overcome the uncertainties that partial observability brings, FALCON uses a short history of actions and measurements to model the system dynamics. At the end of the warm-up phase, FALCON uses the data collected to carefully learn the most relevant Fourier basis that explains the observed turbulent dynamics. FALCON is data efficient and requires only 35 seconds of flow data during the warm-up phase to recover an informative Fourier basis. This 35 seconds of flow data include fewer than 1500 samples taken over a period spanning approximately 85 vortex-shedding events.

FALCON incorporates several key features that achieve data and computational efficiency in the basis learning process. In particular, FALCON uses ℓ_1-constrained (sparse) Fourier basis, as well as Least Absolute Shrinkage and Selection Operator (Lasso) [264] to recover a succinct basis representation (see details in Methods). This improved basis selection yields a significantly compact model representation while allowing physically accurate modeling of the underlying system dynamics due to the low-frequency dominant choice of Fourier basis, i.e., sparse Fourier basis vectors [25]. Indeed, spectral methods and modal analyses for modeling turbulent fluid dynamics are well-established concepts [121, 256], and it is known that large eddies with low frequencies contain the most energy in turbulent flows [222]. This inductive bias in modeling via sparse Fourier basis reduces the number of samples required to learn the turbulent dynamics with small modeling errors and alleviates the computational burden in the predictive control design, facilitating high-frequency control actions. FALCON allows flexibility in the basis learning procedure such that the number of Fourier basis used in model learning could be easily adjusted based on the prior knowledge of the system dynamics, the difficulty of the learning task, and the computational budget.

After recovering a succinct Fourier basis for model learning, FALCON starts the adaptive control in epochs phase, Figure 6.3A. It estimates the model dynamics as a linear model in the learned Fourier basis and aims to learn the unknown linear coefficients that best fit the acquired data onto this basis. In particular, FALCON solves an online least-squares problem that has a closed-form solution to learn these linear coefficients. This interpretable and lightweight model learning allows online and/or batch updates for computational efficiency and comes with strong learning theoretical guarantees for the robustness of modeling (see Materials).

During this phase, FALCON designs an online control policy based on this learned model while improving the system dynamics model in an online fashion over time. This process goes in epochs with doubling duration, i.e., each epoch is double the length in seconds of the previous epoch, where at the end of each epoch FALCON updates its linear coefficient estimates on the model for better-refined dynamics modeling and control. This epoch schedule reduces the number of model updates toward the later stages of adaptive control where the dynamics are already well-modeled and only small tuning is required to further improve. We would like to highlight that FALCON is a single trajectory algorithm in the sense that it does not require a reset between epochs, which makes it efficient in the data collection process.

As the online control policy, FALCON uses model predictive control (MPC) with the estimated model dynamics to design the control inputs during the adaptive control phase. For controlling nonlinear dynamical systems such as aerodynamic control in turbulent flow considered in this work, finding the optimal solution to the control problem is usually challenging [142]. As a practical and efficient alternative, MPC policies have been the dominant choice for designing controllers in nonlinear dynamical systems [54]. Given the initial conditions, the transition dynamics (can be an estimated model or a nominal model), the running costs, and the terminal costs at any given time step, the objective in MPC is to solve a short horizon optimal control problem and execute the first action of the solution sequence. This process is then continued as we gather new observations. Intuitively, instead of trying to solve the challenging global optimal control problem, MPC myopically solves a locally optimal control problem. Usually physical or safety constraints on actions, observations, and dynamics are added in the MPC formulation due to its simplicity of implementation. The choice of the MPC policy depends on the control task. In general, the MPC policies are either optimization-based [76] or

sampling-based [35]. However, sampling-based methods are usually preferred in model-based RL due to challenging nonlinear system dynamics and complicated cost and constraint functions [294].

Therefore, at every time step of the adaptive control phase, FALCON deploys a Cross-Entropy Method (CEM) policy, a sampling-based MPC policy [35], to design control actions using the most recent system dynamics estimate as the transition dynamics. CEM maintains a distribution, predominantly Gaussian, to sample action roll-outs for the short planning horizon and iteratively updates this distribution to assign a higher probability near lower cost action sequences based on the estimated system dynamics. After a certain number of updates, it executes the first action on the lowest cost-achieving action sequence in the sampled roll-outs (see further details of MPC design and CEM, Section 6.1, and in Methods 6.3.3). FALCON takes in the most recent short history of actions and measurements as the initial condition for short-horizon MPC objective. For the running and terminal control costs FALCON can utilize any kind of cost functions as long as they can be evaluated efficiently depending on the control task. In our experiments, we design the cost function of FALCON based on our aerodynamic control objective such that FALCON avoids large lift forces and prevents rapid changes in lift forces and fast/jittery action changes. The first two design choices are clear from the control goal, i.e., minimizing the mean and the standard deviation of the overall lift forces, whereas, the last one is more subtle. In our experiments, we observed that non-smooth changes in actions cause additional lift forces on the airfoil (see further discussion on cost design choice in Materials). Furthermore, in the policy design FALCON includes action constraints due to mechanical restraints of the aerodynamics testbed as we shortly discuss in the next section.

FALCON can easily include further safety or physical constraints within its MPC framework. This makes FALCON a reliable algorithm for safety-critical tasks such as free flight through turbulence. Moreover, the recurrent short-horizon planning approach through CEM allows FALCON to design sophisticated policies that consider future flow effects through the use of estimated model dynamics. Thus, rather than designing purely reactive policies that cancel out the instantaneous aerodynamic forces, FALCON designs flow-predictive disturbance rejection policies which aim to minimize the lift forces while accounting for unsteady flow dynamics. In this way, FALCON adapts to sudden changes in the flow while avoiding overcompensation of the aerodynamic disturbances by maintaining an overall understanding

of the flow field. The simple yet physically sound and accurate dynamics learning approach of FALCON facilitates this effective control design, which results in a sample-efficient and high-frequency (42 Hz) state-of-the-art control policy with generalizable performance.

The construction of FALCON is modular such that different basis functions, e.g., wavelets [236], could be utilized in learning the underlying system dynamics depending on the domain knowledge about the system, while the MPC framework could be selected based on the specific needs, e.g., optimization-based MPC for simpler model dynamics. This interchangeable design of FALCON makes it a viable model-based RL method for designing diverse online/adaptive control strategies for various tasks (see Discussion). Moreover, it also allows for the derivation of strong theoretical guarantees for the robustness of model learning and the control performance under modeling error (see the Methods section). In particular, we prove that a wide range of partially observable nonlinear dynamical systems such as dynamical systems governed by partial differential equations could be learned with arbitrary modeling error using the Fourier basis for FALCON. We also show that this effective model learning allows stable control design for robust MPC frameworks and the systems controlled by FALCON follow a trajectory close to the systems regulated with the same MPC policy that has access to *perfect* system dynamics information. Finally, we formalize the performance guarantee of FALCON such that the control performance of FALCON converges to the idealized MPC controller that knows the perfect system dynamics. These rigorous theoretical results display the reliability of FALCON while attaining state-of-the-art performance in predictive flow control.

Experimental aerodynamic testbed

In this work, we abstract the problem of stabilizing an aerodynamic system under turbulence to basic components while maintaining the core complexity of the physics involved. We utilize an experimental aerodynamics testbed that captures and generalizes the fundamental physics involved in flight through a turbulent environment [228]. The aerodynamic testbed consists of a symmetric generic airfoil with motorized trailing-edge flaps and integrated flow sensors (Figure 6.2B). The trailing-edge flaps have an actuation range of [-40°, 40°], and are mapped linearly from the action space of $[-1, 1]$, yielding 1-dimensional control action per time step. Similar to flap systems on conventional airplanes, actuating the trailing-edge flaps generates a lifting force that can offset the aerodynamic forces associated with

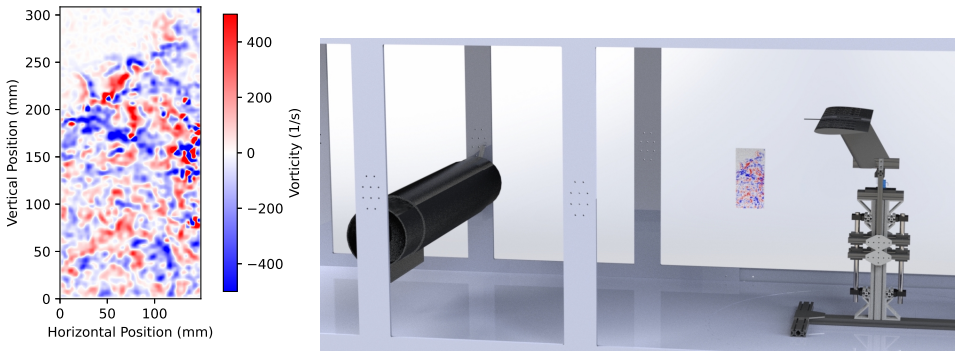

Figure 6.4: Particle Image Velocimetry (PIV) Visualization of the Turbulent Flow Field

flow disturbances as shown in Figure 6.2E. The testbed is equipped with 9 sensors which are placed 10cm apart along the spanwise axis. Observations of the surrounding flow are measured through a series of eight pressure tap flow sensors built into the body of the airfoil, with a single pitot-static tube located at the center of the airfoil. The pressure taps, placed near the leading edge of the airfoil, provide valuable information on incoming pressure differentials between the upper and lower surface of the airfoil. The pitot-static tube measures the total pressure of the incoming flow which is approximately proportional to the mean velocity of the flow. The aerodynamic testbed is mounted on a one-dimensional load cell which is used to observe the lift forces acting on the airfoil, which serves the same effective role as an inertial measurement unit on conventional UAVs. Combined with 9 flow sensors, we obtain 10-dimensional measurements per time step. Further details on aerodynamics testbed design are provided in Methods.

Turbulent environment

We study control in the context of a canonical problem in fluid dynamics: the turbulent wake of a bluff body. When placed incident to winds, bluff bodies produce an oscillating vortical wake commonly known as a Kármán vortex street [230]. At sufficient wind speeds, this wake becomes highly turbulent and can result in significant forces [25], as visualized with smoke in Figure 6.2D. This photo is captured at Caltech Center for Autonomous Systems and Technologies (CAST) fan-array wind tunnel with a standard cylinder at a lower wind speed than the experiments presented in this work. Figure 6.2D depicts the turbulent wake of the cylinder where the vortex shedding is irregular. This phenomenon is famously responsible for the 1940 collapse of the Tacoma Narrows Bridge [288].

All of the quantitative results in this work were obtained from the experiments conducted in Caltech's Lucas Adaptive Wall Wind Tunnel, a closed-loop wind tunnel. As discussed previously, our experiments were performed in the wake of a bluff body at a mean flow speed of 6.81 m/s, which corresponds to a Reynolds number of $Re_D = 230,000$ over the bluff body. The bluff body consisted of a cylinder with a diameter of 30 cm with a normal flat plate fixed asymmetrically that increased the effective diameter to 53 cm, as shown in Figure 6.2C. This construction is used to encourage vortex dislocation which results in less regular vortex shedding events [295]. Further, the bluff body was mounted to the walls of the tunnel with elastic cords to allow for dynamic oscillations which may also encourage less regularity in shedding events (see Methods for details).

We used particle image velocimetry (PIV) to visualize a portion of the turbulent flow field in our experimental environment (see Methods for details). Figure 6.4 presents the PIV measurements of the vorticity field contextualized in the wind tunnel. The complex vorticity patterns clearly demonstrate the chaotic and unsteady turbulent flow dynamics, with strong three-dimensional effects likely present in our experimental setting. Moreover, through hot-wire anemometer measurements in the wind tunnel, we record a turbulence intensity of 10.8%.

Baseline Control Methods

To test the performance of FALCON, we deploy several RL baselines and the industry-standard responsive control strategy of PID (Proportional - Integral - Derivative) control in our aerodynamics testbed. In particular, we compare FALCON with the twin delayed deep deterministic policy gradient algorithm TD3 [92], its variant known as LSTM-TD3 [196], and soft actor-critic algorithm SAC [102] (see Methods for a detailed overview). These methods are the state-of-the-art off-the-shelf model-free RL methods deployed in many real-world control tasks [80, 95, 101, 228]. They are off-policy actor-critic algorithms that utilize neural networks for control policies. Off-policy methods are usually preferred over on-policy methods in real-world dynamical systems with unsteady dynamics since they can learn from a wide range of experiences, including observations from previous policies, which makes them more robust to changes in the environment. Another advantage of off-policy methods is that they can learn an optimal policy even when the current policy is significantly sub-optimal, which is usually the case for challenging real-world control tasks due to the lack of clearly superior expert policy. These all combined

allow off-policy methods to be more stable during the learning process, which leads to better convergence and generalization performance [255].

The TD3 algorithm has previously demonstrated success in experimental flow control in different settings [80]. To improve performance in partially observable systems, such as the turbulent flow dynamics measured by sensors, LSTM-TD3 utilizes recurrent long-short-term memory (LSTM) cells in the neural network structure of TD3. The addition of LSTM cells has been previously shown to improve the performance in prediction and control of highly unsteady stochastic environments like turbulent flow fields [287]. In particular, recently, LSTM-TD3 has been demonstrated to achieve state-of-the-art performance in disturbance rejection in a similar experimental setting studied in this work [228]. Therefore, LSTM-TD3 provides the ultimate baseline for FALCON. In their implementations, both TD3 and LSTM-TD3 have nearly identical parameters besides the additional LSTM structure of LSTM-TD3 for an additional memory element in the policy. While TD3 and LSTM-TD3 provide deterministic policies, SAC designs stochastic policies, which are shown to achieve significant success in various real-world tasks such as quadrupedal robots and voltage control [102, 289]. It provides a sample efficient alternative policy design method compared to TD3 and LSTM-TD3.

Due to the stochasticity in the process of training RL algorithms, we trained each of the agents presented here with three independent random seeds and present the average training results to display their performances. Unlike FALCON, the model-free methods work in episodes with reset for retraining. We train the model-free methods for 200 episodes of 800 samples per episode and use the best-performing agent for each algorithm in presenting their final performance. The feedback gains of the PID controller were tuned manually to achieve constant zero lift using the readings of load cell measurements. In our experiments, we run an exhaustive grid search over the feedback coefficients and report the best-performing controller. All methods, including FALCON, are implemented with 42-Hz sensing and control frequency.

Superior Performance and Sample Efficiency of FALCON

First, we provide the results on the training of the methods. In presenting the training behavior, we exclude the warm-up of each method. Note that the warm-up phase of FALCON requires only 1500 samples, which corresponds to approximately 85 vortex-shedding cycles from the upstream bluff body. Figure 6.5 shows the moving

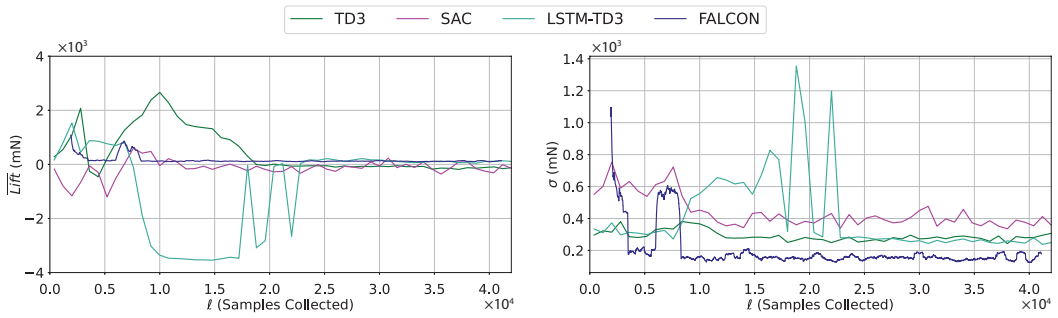

Figure 6.5: Evolution of the mean ($\overline{\text{Lift}}$) and standard deviation (σ) of the lift forces for the best-performing agents of each algorithm shown over the first 40,000 samples. The full training performance for the model-free algorithms can be found in Figure 6.6.

average of the mean and standard deviation of the lift forces on the airfoil for the best-performing policy of each method over the first 42,000 samples collected. In this plot, while the data collection procedure of FALCON does not pause in between model updates (epochs), the model-free algorithms pause (end of their episode) and train for some time to update their policy. From these plots, we observe that FALCON quickly finds the unknown linear coefficients to represent the system dynamics in the learned Fourier basis and achieves significantly better performance than model-free methods with fewer samples. We also note that during the 40,000 sample period shown, FALCON has only 25 learning updates while the other algorithms shown have 50. As the model-free algorithms used for comparison typically require more data, we trained these algorithms for a total of 200 episodes (equivalent to 160,000 samples), the remainder of which can be seen in Figure 6.6. The training behavior of FALCON indicates that FALCON agents consistently improve throughout training and require significantly less training time to outperform state-of-the-art model-free methods due to its physically accurate model learning procedure and efficient control design.

Even though FALCON can suffer from model uncertainty and execute sub-optimal actions at the beginning of training due to unsteady flow dynamics (see the outlier bump in the standard deviation of lift in Figure 6.5 around $t = 6,000$), FALCON effectively explores the state-space to improve the accuracy of the model and hence the performance of the controller to bring the standard deviation in the lift forces to desirable values. After 10,000 samples, i.e., 4 minutes of training of FALCON, the average standard deviation of lift forces on the aerodynamics testbed remains stable at a level significantly better than other tested methods. Similarly, the mean lift

forces achieved via FALCON consistently outperform that of model-free methods.

Among the model-free methods, Figure 6.5 shows that LSTM-TD3 is the second best-performing algorithm by outperforming TD3 slightly, while SAC fails to achieve acceptable performance. Note that LSTM-TD3 and TD3 share the same policy constructions except the LSTM part that adopts latent states in the policy. Combined with the superior performance of FALCON, this highlights the importance of a latent state representation in achieving desirable learning and control performance in partially observable real-world settings. Similar to FALCON, LSTM-TD3 achieves consistent performance after sufficient samples, yet it requires an order of magnitude more than FALCON. Despite our significant efforts in hyperparameter tuning, SAC agents failed to learn desirable policies that minimize the aerodynamic forces. Even the best-performing agent significantly underperformed compared to other model-free policies, which indicates that stochastic policies such as SAC might not be suitable for controlling unsteady dynamical systems.

From our experiments, we observe that FALCON is more robust to hyperparameter tuning and has a notably more stable training process (Figure 6.5) compared to model-free methods. In particular, in our training process, tuning FALCON requires only a few trials on the history of modeling (number of past observation-action pairs used), the sparsity weight in Lasso for recovering a succinct basis, and the planning horizon for CEM. On the other hand, model-free methods require extensive hyperparameter search to achieve some learning behavior. This extensive search is significantly time-consuming and laborious in real-world problems, e.g., the training process of model-free methods corresponds to an hour of training for each hyperparameter configuration in our setting. This process becomes unfeasible in online resource-constrained settings, which are typically the scenarios for adaptive control systems.

Our experiments overall showed that FALCON consistently achieves better and more stable training performance than model-free methods while converging to its optimal policy with orders of magnitude fewer samples. This superior performance and sample efficiency show that the simple yet efficient partially observable dynamics modeling approach of FALCON reduces the complexity of the aerodynamic control under turbulence problem significantly. Combined with the efficient predictive control design, FALCON agents learn to reject the flow disturbance effectively. The model learned by FALCON also aligns well with established knowledge regarding spectral energy content; in particular, the improved basis selection with ℓ_1-constraint

and Lasso enforces the model to have relatively low frequencies corresponding to the dominant energy-containing eddies. In our experiments, we observe that FALCON recovers model estimates which put significant weight on the low-frequency basis, e.g., DC components, and some high-frequency components (see Methods). This shows that via using the relevant basis for learning, FALCON learns a physically meaningful model, which contributes to training stability and disturbance rejection performance of FALCON.

To further investigate the effect of the concise Fourier basis-based model learning approach of FALCON, we implemented a reasonably-sized deep neural network for model learning and combined it with CEM-based policy design as FALCON. However, despite our significant efforts in tuning, this approach failed to learn a reliable model for control design due to unsteady and chaotic flow conditions, which resulted in performance significantly worse than the reported algorithms in this work. This outcome highlights that black-box dynamics modeling methods such as deep neural networks can fail in unsteady systems such as turbulent flow dynamics.

Consistent and Generalizable Performance of FALCON

Next, we study the generalization performance of FALCON and other methods including the PID controller. Table 6.2 presents the average performance of the best-performing policies by each method in 10 independent 90-second length runs, i.e., 4000 samples, as well as the number of samples required to train the respective "best" policies. We report the mean and standard deviation of the lift forces on the airfoil averaged over these runs. As discussed before, the standard deviation of the lift forces is the key metric for disturbance rejection and aerodynamic flow control. Table 6.2 shows that FALCON improves upon the prior state-of-the-art performance in flow disturbance rejection under extreme turbulence by 37%. Further, FALCON achieves this performance using only 8.7 minutes of data, whereas the model-free algorithms can take hours to train in the same setting [228]. This significant improvement with 8 times fewer samples shows that FALCON adapts and generalizes across independent runs despite remarkably different training conditions due to unsteady and chaotic turbulent flow dynamics. Moreover, FALCON outperforms the industry standard PID controller as well as TD3 policy by more than 45%. Similar to the training performance, we have observed that the SAC policy failed to achieve acceptable performance in disturbance rejection while requiring less training data to converge

Control Method	Training Samples	Average Over 10 Tests		Std Over 10 Tests	
		Absolute Mean Lift (mN)	Std in Lift (mN)	Absolute Mean Lift (mN)	Std in Lift (mN)
PID	N/A	6	271	1	8
TD3 [92]	1.71×10^5	183	267	4	9
SAC [102]	1.51×10^5	268	395	213	64
LSTM-TD3 [228]	1.76×10^5	139	236	16	5
FALCON	$\mathbf{2.20 \times 10^4}$	**2**	**148**	**10**	**13**

Table 6.2: Disturbance rejection performance of the methods over 10 independent 90-second test runs.

compared to other model-free methods. This result also suggests that stochastic policies might not be effective in controlling unknown unsteady or chaotic systems.

We also document that FALCON achieves the best average absolute mean performance. In particular, it outperforms PID control which is designed to keep the absolute mean close to zero. Even though it is a secondary and easy-to-offset metric in aerodynamic control, we observe that model-free methods attain significantly higher absolute mean lift compared to FALCON and PID controllers. Among these methods, LSTM-TD3 also achieve the lowest absolute mean.

Finally, we present the standard deviation in these performance metrics over 10 independent runs to test the consistency of the control methods' performance in Table 6.2. We see that the performance of FALCON is consistent over the unsteady dynamics with minimal change over the runs and it almost matches the consistency of the non-learning-based PID controller. Notably, besides SAC, other RL methods also perform consistently over these independent runs, where LSTM-TD3 which uses memory units in their policy construction, akin to FALCON, outperforms TD3. These results overall show that FALCON is able to generalize its performance to unseen disturbances and consistently provides state-of-the-art predictive-flow disturbance rejection in extreme turbulent flow dynamics.

6.3.2 Discussion

We have designed and demonstrated FALCON, the first model-based RL method that can effectively learn to control aerodynamic forces acting on an airfoil under extreme turbulence with which conventional methods struggle. Our results indicate that combining flow sensing with physically sound model learning and efficient control design allows state-of-the-art disturbance rejection despite the chaotic nonlinear turbulent dynamics. Besides the superior performance, the physics-informed

lightweight design for learning and control of FALCON allows an order-of-magnitude improvement over the number of samples required to achieve desirable control performance compared to prior RL methods. Further, we document that our method has a stable training procedure and a consistent performance even under highly irregular unsteady dynamics of turbulent flow. These results indicate the potential to use this method to stabilize systems such as UAVs under extreme turbulence in free-flight scenarios.

We observed that FALCON improves the aerodynamic control performance of prior state-of-the-art LSTM-TD3 [228] by 37%. Even though LSTM-TD3 is a model-free RL algorithm, it shares a similarity with FALCON that it uses recurrent LSTM cells to utilize a history of observations in the policy design. With this construction, LSTM-TD3 is able to capture the latent state dynamics in designing policies. This allows LSTM-TD3 to handle partial observability in system dynamics due to sensor measurements, and design policies based on modeled latent states. Due to the structural similarities of LSTM-TD3 and TD3 algorithms, the superior performance of LSTM-TD3 over TD3 should be attributed to the latent state modeling via LSTM cells.

In contrast to FALCON, the modeling of latent states in LSTM-based policy design is mostly black box and without any physical interpretation. On the other hand, in FALCON, the flow dynamics are captured by significant low-frequency and some high-frequency model components with learned linear mixing coefficients using a history of observations and actions. Our proposed modeling approach in FALCON is motivated by the prior studies on turbulent flow dynamics that observe a well-defined frequency spectrum with significant low-frequency energy content for highly turbulent flows [222, 228]. The interpretable dynamics modeling of flow disturbances of FALCON simplifies the disturbance rejection task to a low-dimensional learning-to-control problem, where learning and control design are executed efficiently. Moreover, this principled approach allows theoretical guarantees on sample-efficient model learning, robustness against imperfect learning, and control performance under modeling error which are derived for FALCON in the Methods section. All these results highlight the importance of deploying domain knowledge in model learning for unsteady and chaotic systems such as turbulent flow fields.

Modeling

FALCON exhibits a modular structure, where the model learning and control components could be replaced depending on the task. The Fourier basis is deployed in

FALCON and our experiments due to prior studies which showed that turbulent flow dynamics have a well-defined power spectrum dominated by the low-frequency components. Due to such underlying physics, the choice of Fourier basis allows theoretically guaranteed learning of the underlying system (see Methods). In particular, we rigorously show that the modeling error of such underlying systems could be made arbitrarily small with sufficient basis and data points from the system. This approach is in contrast to black-box modeling of the system dynamics, e.g., via deep neural networks, which naively uses purely data-driven basis functions, which may cause instability and fragility in model learning and control. In fact, our experiments with deep neural network modeling showed that in such temporally unsteady systems, it is hard to fully characterize the system dynamics without incorporating domain knowledge. This insufficient learning caused significantly inferior control policies.

The modeling capabilities of FALCON could be improved by adding nonlinearities to the modeling via Fourier basis. The composition of Fourier basis learning with nonlinear functions has shown success in learning the solution operators of partial differential equations [184]. Adopting such a modeling approach could further extend the model learning capabilities of FALCON and improve its aerodynamic control performance. Different basis vectors such as Random Fourier Features (RFF) have been deployed in prior model-based RL works[164]. Incorporating them along the Fourier basis can also extend the class of systems that could be learned via FALCON. Finally, different modeling approaches, such as modeling the pressure/lift differences on the sensors via the history of observations and actions, could be deployed to improve the sample efficiency and performance further. In our experiments, we tried this approach but did not observe a change. Yet, this approach could be helpful in deploying FALCON in more challenging turbulent environments.

Control design

In the control design, FALCON adopts CEM, a sampling-based MPC method, to exploit the learned accurate model. By design, CEM provides a transparent control design method in terms of what control cost is to be minimized and for how long of a trajectory should be considered in planning. This transparency is significant when incorporating domain knowledge and experimental observations directly in the control design. In particular, during our experiments, we observed that having rapid changes in the flap angles, i.e., too much variation in consecutive actions, results in a slight increase in lift forces on the airfoil. With this observation, we

added a term in the cost design of FALCON which prevents these changes to a certain extent and improves aerodynamic control performance. This cost design is also intuitive for general flight control since it also reduces the wear and tear on the actuators. Moreover, due to this transparency, FALCON includes safety and physical constraints easily in the control design problem by simply eliminating trajectories or action sequences that violate these constraints.

This is in stark contrast with the black-box controllers provided by the model-free RL algorithms. These methods are very sensitive to many hyperparameters which control the neural network architecture and training procedure, yet, the effect of each hyperparameter on the performance is unclear. This lack of transparency leads to a reliance on intuition, experience, and trial and error when tuning these hyperparameters, making the process time-consuming and frustrating. Even though the data presented for each model-free method took around 6 hours in the wind tunnel through training and testing, the actual process of hyperparameter tuning required dozens of additional hours for each algorithm. This presents a challenge in dynamic experimental environments, such as aerodynamic control under turbulence. On the other hand, the whole process of tuning, training, and testing of FALCON took around 9 wind tunnel hours in total.

In the aerodynamic control problem studied in this work, we considered one-dimensional control actions with a 5 time-step planning horizon. This results in a relatively small search space to find optimal actions for CEM. This was particularly important in the control design of FALCON since the sampling-based MPC methods as CEM can be inefficient in longer planning horizons or larger action spaces. One can increase the number of samples per iteration in higher dimensional control problems, yet this might cause delays in control and poor performance. In order to deploy FALCON in higher dimensional control settings, utilizing a more efficient model predictive control method based on first-order optimization might be useful. This can be easily achieved with the same model learning module of FALCON and replacing CEM due to the modular structure of FALCON. Interior-Point Methods (IP) and Sequential Quadratic Programming (SQP) are two well-known algorithms for numerically solving these nonlinear optimization problems [208]. For high-dimensional problems, they can be further improved to exploit the sparsity in the control design and achieve desirable performance without sacrificing efficiency [181]. SQP methods are particularly good candidates in the control design since they use the result of the previous iteration to warm-start the next iteration of

the control design similar to CEM.

Computational Efficiency

In the warm-up phase, the proposed approach of solving the Lasso problem over ℓ_1-constrained Fourier basis is able to learn a concise and effective basis for representing the system dynamics in a data-efficient way. This requires only 35 seconds of flow data at 42 Hz which is collected with time-correlated Gaussian inputs. This is equivalent to approximately 85 vortex-shedding cycles, however, due to the irregular shape and dynamic motion of the bluff body, it is likely that the wing-vortex interactions varied significantly during this period. Finding the solution of Lasso takes about 7 minutes on a standard desktop computer, and this problem is solved only once at the end of the warm-up phase. This effective succinct basis representation for the underlying dynamics allows FALCON to design control actions in less than 10 ms within the CEM model predictive control framework which yields 42 Hz sensing and control frequency. The fast adaptive control approach allows FALCON react to the changes in the flow field rapidly while still reasoning about how to mitigate upcoming turbulent disturbances on the system via the learned and updated model dynamics. To achieve this fast control design, FALCON leverages the parallel computing on GPU and samples a significant amount of initial action roll-out in CEM to overcome possible local minima in designing control actions. In this end-to-end control loop, serial communications between the controller and sensors, and actuators are the main bottleneck.

To further improve the disturbance rejection performance of FALCON, increasing the control frequency is one of the future developments to focus on. This could be achieved by reducing the code execution time and communication delays. Deploying a faster implementation of FALCON using C++ or utilizing a more computationally efficient MPC framework such as CEM-GD [117] which combines zeroth and first-order optimization methods could allow us to achieve sub-5ms control design duration. Moreover, having a streamlined communication layer could also reduce the latency between the controller and sensors or actuators significantly.

Generalization to New Tasks and Free-Flight

In this work, we have developed a model-based reinforcement learning method, FALCON, on a generic aerodynamic testbed for flow-informed aerodynamic control under extreme turbulence. FALCON was tested on a single-dimensional aerodynamic

control system, yet it can be extended and adapted to systems with higher degrees of freedom. In particular, we can consider other forces and moments in three dimensions, besides the vertical lift forces acting on the testbed. Our experiments are conducted under the Reynolds number of $Re_D = 230,000$. In order to ensure the generalizable performance of FALCON across a range of Reynolds numbers, i.e., different turbulence characteristics, and different geometries of airfoils, further investigation is required.

The findings of this work hold potential in the deployment of next-generation technologies, including but not limited to flow-sensing UAVs capable of stable flight in windy urban areas and flow-informed wind turbines with gust protection. In the fixed-wing UAVs, FALCON is a promising algorithm for the inner-loop attitude control for fixed-wing vehicles. This will allow drones to maintain stable flight in extreme conditions by reducing the impact of turbulent disturbances. We believe that model-based RL methods, and FALCON in particular, could be used for full-stack control and navigation using flow information and simulated environments. The testbed in this work emulates stabilizing a UAV at a constant altitude. Future work will consider using FALCON with a trajectory planner such that FALCON aims to maintain the desired location coming from the planner and interacts with the planner to achieve energy efficient and safe navigation, similar to the prior work in computational fluid dynamics [99]. To accomplish this will require overcoming challenges such as sim-to-real transfer and distribution shift in data, where the data efficiency and fast adaptation capabilities of FALCON would be critical. We suggest using indicators of changes in turbulent conditions in hierarchical planning to control the frequency of model updates within FALCON. Another strategy would be using meta-learning to make the model learning process adaptive in basis selection and model updates for different turbulent conditions with different basis representations.

6.3.3 Methods

FALCON Algorithm

In this section, we present the methodology and the algorithmic details of our proposed model-based RL method Fourier Adaptive Learning and Control (FALCON). FALCON learns the model dynamics in Fourier basis through interaction with the system and deploys MPC using the learned model for control design. The outline of FALCON is given in Alg. 16. FALCON has two phases: Warm-up and Adaptive Control in Epochs.

Algorithm 16 FALCON

Input: T_w, h, t_{ep}, τ, D, $C_{0:T}$

—— Warm-up ——

for $t = 1, 2, \ldots, T_w$ **do**

Deploy exploratory u_t and store $\mathcal{D}_0 = \{y_t, u_t\}_{t=0}^{T_w}$

Form $s_t = [y_{t-1:t-h}^\top, u_{t-1:t-h}^\top]^\top$ for all t using $\mathcal{D}_0$

Compute $\phi'(s_t)$ via (6.9) $\forall s_t$, D'-dimensional Fourier Series representation

Solve Lasso (6.20) to learn succinct Fourier basis $\phi(\cdot)$ representation

—— Adaptive Control in Epochs ——

for $i = 1, \ldots$ **do**

Solve for $\hat{\Theta}_i$ via (6.17) & Form $\hat{F}_i(\cdot) = \hat{\Theta}_i^\top \phi(\cdot) \rightarrow$ Model Dynamics Updates

for $t = T_w + (i-1)t_{ep} + 1, \ldots, T_w + it_{ep}$ **do**

$u_t = \text{MPC}(\hat{F}_i, y_{t:t-h+1}, u_{t-1:t-h+1}, C_{t:(t+\tau)})$

Observe y_{t+1} & Form s_{t+1} and $\phi(s_{t+1})$ using the Learned Fourier Basis

Warm-Up: FALCON starts with a short warm-up period to collect some data about the unknown system. In this phase, the goal is to purely explore the system and recover a coarse model of the dynamics. Therefore, FALCON focuses on safely exciting the system for T_w time steps. The predominant choice for such a task is to use isotropic Gaussian inputs, $u_t \sim \mathcal{N}(0, \sigma_u I)$. However, for certain tasks, one may require smoother or safer exploration. This is usually the case in safety-critical control tasks like flight control under turbulence [228] or bipedal/quadrupedal walking [298]. In these situations, FALCON can use time-correlated inputs for smooth actions such that it avoids jerky and sudden changes in the actions. To this end, for some $\gamma \in [0, 1]$, FALCON can use the following control inputs

$$
\begin{aligned}
u_1 &= \boldsymbol{\eta}_1, \\
u_t &= \gamma u_{t-1} + \sqrt{1-\gamma^2}\boldsymbol{\eta}_t,
\end{aligned}
$$

where $\boldsymbol{\eta}_t \sim \mathcal{N}(0, \sigma_\eta I)$. We deploy this controller with $\gamma = 0.8$ during the warm-up phase in our experiments. Moreover, FALCON can deploy known safe nominal controllers, such as trajectory generators [122] or PID controller, accompanied with isotropic excitements, i.e., $u_t = K(y_t) + \boldsymbol{\eta}_t$ where $K(\cdot)$ is the nominal controller and $\boldsymbol{\eta}_t \sim \mathcal{N}(0, \sigma_\eta I)$.

Adaptive Control: After the warm-up, FALCON starts adaptive control of the underlying system. It uses epochs of doubling length starting with an initial epoch of t_{ep} time steps, i.e., each ith epoch is $2^{i-1}t_{ep}$ time steps for $i = 1, 2, \ldots$. FALCON is a single trajectory algorithm and does not require a reset between epochs. This makes FALCON efficient in data collection in the experiments.

Learning the Dynamics — At the end of the warm-up, FALCON estimates the model dynamics as a linear model in Fourier basis. To this end, it generates $T_w - h + 1$ subsequences of h input-output pairs,

$$s_i = [y_{i-1}^\top, \ldots, y_{i-h}^\top, u_{i-1}^\top, \ldots, u_{i-h}^\top]^\top \in \mathbb{R}^{h(d_y+d_u)}$$

for $h \leq i \leq T_w$. Using (6.3), one can write the system dynamics as $y_t = F(s_t) + e_t$. To estimate the unknown nonlinear function F, FALCON considers the nth order Fourier expansion of F and generates D-dimensional Fourier series representations of all s_t as given in (6.16), $\phi(s_t)$. The order of the Fourier expansion, thus the dimension D, is an important hyperparameter of FALCON. This choice depends on many factors including prior knowledge of the system dynamics, the difficulty of the learning task, and the computational budget. As explained in the Overview section of Methods, a wide range of nonlinear systems can be represented as linear models in the Fourier basis. Therefore, FALCON considers the following model for estimating the system dynamics F,

$$y_t \approx \Theta_*^\top \phi(s_t) + e_t, \tag{6.16}$$

for an unknown $\Theta_* \in \mathbb{R}^{D \times d_y}$. To recover an estimate of Θ_* solves a least-squares problem,

$$\min_\Theta \lambda \|\Theta\|_F^2 + \|Y_{T_w} - \Theta^\top \Phi_{T_w}\|_F^2, \tag{6.17}$$

for some $\lambda > 0$, where $Y_t = [y_t, \ldots, y_h] \in \mathbb{R}^{d_y \times t-h+1}$, $\Phi_t = [\phi(s_t), \ldots, \phi(s_h)] \in \mathbb{R}^{D \times t-h+1}$. The solution to this problem is given as $\hat{\Theta}_1 = (\Phi_{T_w} \Phi_{T_w}^\top + \lambda I)^{-1} \Phi_{T_w} Y_{T_w}^\top$. Using $\hat{\Theta}_1$, FALCON estimates the system dynamics as $\hat{F}_1(s) = \hat{\Theta}_1^\top \phi(s)$. FALCON repeats this dynamics estimation process at the beginning of each epoch using all the data gathered so far. Note that for large D, computing the closed-form solution could be computationally demanding or cause numerical errors. Instead, the model estimates can be updated recursively throughout the epochs using online updates, which we utilize in our implementation for the experiments. In particular, FALCON stores only the current model estimate, i.e., the model estimate at time step t in epoch i: $\hat{\Theta}_{i,t}$, and the inverse design matrix (sample covariance matrix), i.e., $V_{t-1}^{-1} = (\Phi_t \Phi_t^\top + \lambda I)^{-1}$. Using these the model estimates can be updated recursively throughout the epochs using online or batch updates via

$$\hat{\Theta}_{i,t} = \hat{\Theta}_{i,t-1} + \frac{V_{t-1}^{-1}\, \phi(s_t)\, (y_t - \hat{\Theta}_{i,t-1}^\top\, \phi(s_t))^\top}{1 + \phi(s_t)^\top V_{t-1}^{-1} \phi(s_t)}, \tag{6.18}$$

where V_{t-1}^{-1} is also updated recursively [238],

$$V_t^{-1} = V_{t-1}^{-1} - \frac{V_{t-1}^{-1}\,\phi(s_t)\phi(s_t)^\top V_{t-1}^{-1}}{1+\phi(s_t)^\top V_{t-1}^{-1}\phi(s_t)}.$$

Note that FALCON uses the initial most recent model estimate at the beginning of the epoch for the control design during the entire epoch. These online update rules are used to efficiently update the model estimates in the background at each time step with the new data such that at the beginning of the next epoch, there is no delay in updating the model estimate. This feature is important in real-time control systems where any delay in the system can cause further problems and compromise safety.

Improved Basis Selection — Note that as the order of the Fourier basis increases, D increases exponentially in the system's dimension. For large h, *i.e.*, higher order NARX models, this may cause an additional computational burden. To remedy this, we propose to use ℓ_1-constrained Fourier basis and Least Absolute Shrinkage and Selection Operator, *i.e.*, Lasso [264], for an improved basis selection in FALCON after the warm-up period. In particular, instead of generating the bases $\boldsymbol{\omega_i}$s for all n, we only consider the ℓ_1-constrained bases, i.e., $\|\boldsymbol{\omega_i}\|_1 \leq n$. The ℓ_1 constraint reduces the number of basis vectors from $1+2n^{h(d_y+d_u)}$ to $2D'$ basis vectors where

$$D' = \binom{h(d_y+d_u)+n}{n}. \tag{6.19}$$

We then solve the Lasso problem for the warm-up samples that are represented in these ℓ_1-constrained Fourier basis vectors. Lasso is the ℓ_1-regularized least squares method to recover sparse models, with few non-zero coefficients. Given the data points gathered during the warm-up period $\mathcal{D}_0$, using D' number basis vectors with $\|\boldsymbol{\omega_i}\|_1 \leq n$, FALCON forms the following feature representations for $s_h, \ldots, s_{T_w}$ generated via $\mathcal{D}_0$:

$$\phi(s_i) = [\cos(\boldsymbol{\omega_1}^\intercal s_i), \sin(\boldsymbol{\omega_1}^\intercal s_i), \ldots, \cos(\boldsymbol{\omega_{D'}}^\intercal s_i), \sin(\boldsymbol{\omega_{D'}}^\intercal s_i)]^\top.$$

FALCON then solves the Lasso problem:

$$\min_W \frac{1}{2T_w}\|Y_{T_w} - W^\top \Phi_{T_w}\|_F^2 + \alpha\|W\|_1, \tag{6.20}$$

for some $\alpha > 0$. FALCON then uses the basis vectors that have nonzero feature coefficients (entries) in the solution of (6.20), $W_\star$, for learning the model dynamics in the adaptive control phase. The choice of α determines the sparsity of the model

learned $W_\star$ which in turn determines the number of basis vectors, D, used in model learning, i.e., bigger α results fewer non-zero entries in $W_\star$ and fewer basis vectors for estimating the dynamics in the adaptive control period. This improved basis selection significantly decreases D and reduces the computational burden and the samples required to learn the dynamics.

Control Design — Once FALCON has an estimated model, it uses an MPC policy to design the control inputs during the epoch. The choice of the MPC policy depends on the control task. In general, the MPC policies are either optimization-based [76] or sampling-based [35]. However, sampling-based methods are usually preferred in model-based RL due to challenging nonlinear system dynamics and complicated cost functions [294]. Thus, FALCON uses Cross-Entropy Method (CEM) as the MPC policy. As described in Section 6.1, CEM is a sampling-based (zeroth order) MPC policy to solve the problem given in (6.4). CEM maintains a distribution, predominantly Gaussian, to sample action roll-outs for the planning horizon and iteratively updates this distribution to assign a higher probability near lower-cost action sequences based on the estimated dynamics. After a certain number of updates (once it converges), it executes the first action on the lowest cost-achieving action sequence in the sampled roll-outs.

At any time step t, FALCON uses the most recent dynamics estimate $\hat{F}_k(\cdot)$, the last h input-output pairs as the initial condition, and the next τ cost functions $C_{t:(t+\tau)}$ in solving the problem in (6.4) for the planning horizon τ. FALCON executes the first action u_t in the solution of (6.4), receives the output y_{t+1}, and constructs s_{t+1} and $\phi(s_{t+1})$. FALCON repeats this adaptive control process throughout the epoch. Note that any safety or physical constraint can be easily included in the MPC policy design problem (6.4), which makes FALCON a reliable algorithm for safety-critical environments.

Implementation Details of FALCON

We provide the implementation details of FALCON for the experiments.

NARX Modeling: We use an order-4 NARX model for learning the underlying system dynamics, $h = 4$. In our experiments, we deduce that this is optimal to overcome the uncertainties of partial observability and reasonable computational complexity. With this choice, s_t in the system modeling becomes 44-dimensional vector. To estimate the unknown nonlinear system F, we consider the 3rd-order Fourier expansion. However, to reduce the computational complexity for such a

high-dimensional learning problem, we only use $\|\omega_i\|_1 \leq 3$ constrained basis vectors and use Lasso to identify the most relevant basis vectors using the warm-up data as described in Appendix. At the end of this procedure, we obtain $D = 319$ dimensional Fourier series representation for learning the model dynamics.

Design Choices: The control goal in disturbance rejection is to minimize the mean and the standard deviation of the lift forces acting on the airfoil. Thus, we design our cost function to penalize large lift forces, rapid changes in lift forces, and fast/jittery action changes. FALCON has a warm-up duration of around 35 seconds, *i.e.*, $T_w = 1500$ samples, using the time-correlated sum of Gaussian inputs for smooth exploration to collect some data about the unknown system dynamics and recover the most relevant Fourier basis. The epochs of the adaptive control period are approximately 38 seconds, $t_{ep} = 1600$ samples per epoch. FALCON uses Cross-Entropy Method (CEM) as the MPC policy. CEM is a sampling-based MPC policy to solve the problem given in (6.4) [35]. CEM maintains a distribution, predominantly Gaussian, to sample action roll-outs for the planning horizon and iteratively updates this distribution to assign a higher probability near lower-cost action sequences based on the estimated dynamics. After a certain number of updates, it executes the first action on the lowest cost-achieving action sequence in the sampled roll-outs. The CEM algorithm is given in full detail in Algorithm 14.

Compared methods We compare FALCON with several model-free RL methods, including TD3 [92], LSTM-TD3 [196], SAC [102], and the industry-standard responsive control strategy of PID (Proportional–Integral–Derivative) controller. Of all these algorithms, LSTM-TD3 has been demonstrated to achieve state-of-the-art performance in disturbance rejection [228]. Unlike FALCON, the model-free methods work in episodes with reset for retraining. We train the model-free methods for 200 episodes of 800 samples per episode and test the best-performing policy in presenting the results. All methods, including FALCON, are implemented with 42 Hz sensing and control frequency.

Adaptive control design

For the planning horizon, FALCON uses $\tau = 5$ in CEM. Furthermore, FALCON samples $K = 1000$ trajectories in the first action roll-out of CEM and decays the number of samples in each update. In prior works, this sampling strategy has been observed as an efficient way to avoid possible local minima in finding the optimal action actions [117]. Table 6.3 summarizes the hyperparameters for FALCON in our exper-

iments. In order to achieve desired control and sensing frequency number of CEM samples (K) and iterations (M) create a trade-off in the implementation. Maintaining this control and sensing frequency is crucial in order to avoid *undersampling* the turbulent dynamics.

Table 6.3: Hyperparameters of FALCON in our experiments.

Hyperparameter	**Range**	**Best**
NARX-order (h)	3 – 5	4
Fourier Series Coef. (D)	100 – 700	319
Planning Horizon (τ)	3 – 8	5
CEM Samples (K)	150 – 1500	1000
CEM Iteration (M)	4 – 7	6
CEM Number of Elites (N)	10 – 30	30

Wing system design and manufacturing

The wing system was designed with a standard NACA0012 airfoil shape, which was previously studied for its dynamics in a bluff-body wake at a similar Reynolds number [176, 228]. The modular wing body was 3D printed using a combination of materials, allowing for various sensor configurations. The central module, made of micro carbon fiber-filled nylon (Markforged Onyx) reinforced with carbon fibers for added strength and rigidity, housed the primary electronics and secured the system to its mounting via a sweptback fairing. Spanwise sections designed to house individual pressure sensors were also printed using carbon fiber-filled nylon. Clear PLA was used to print the sections between the sensors, which were aligned and connected using carbon fiber spars to add rigidity. Trailing edge flaps were cut from insulation foam and covered with an adhesive-backed coating for protection and improved surface finish.

The wing had a spanwise length of 1 m and a total chord length of 25 cm with 5 cm trailing edge flaps. Using the mean flow velocity near the leading edge, the system had a Reynolds number of approximately $Re \approx 110,000$. There were 9 sensor locations distributed symmetrically about the wing, with exactly 10cm between each location. The central sensor location featured a pitot-static tube, with the rest of the sensor locations featuring surface pressure taps. The pressure taps and the pitot-static tube were printed using an SLA printer (Formlabs Form3) for improved surface feature accuracy. Pressure taps were located at 0.4%, 0.7%, 1.5%, and 6% of the chord length from the leading edge on both the pressure and suction sides of

the wing. The fairing on which the wing was mounted was reinforced with carbon fiber and aluminum and was set back with an angle of 60° to reduce aerodynamic interactions between the fairing and the wing. The fairing was connected to a set of vertically-aligned air bearings (New Way), which allowed for nearly frictionless motion along a single axis while constraining all other directions. The constrained fairing was mounted directly to a single-axis load cell (Interface SM-50) that passed the signal through an amplifier (Interface Model SGA) with a 50Hz low-pass filter, and the signal was read by a DAQ (NI USB-6008).

The wing had a total of 9 ultra-low range digital pressure sensors (Honeywell RSC-DRRM2.5MDSE3) to measure pressure values, which were communicated with a microcontroller (Teensy 4.0). The microcontroller also controlled the high-speed brushless servo motors (MKSHBL6625) that drove the trailing edge flaps. Due to mechanical restraints, the actuation for the servo motors had a maximum/minimum position of $+40°/-40°$. Both the microcontroller and the DAQ communicated with a desktop computer, which received measurements and sent actions. The full control loop operated at approximately 42 Hz.

Generation and characterizations of turbulence

The John W. Lucas Wind Tunnel (LWT) at Caltech was used to conduct all quantitative experiments discussed in this study. The wind tunnel is a closed-loop design with a test section size of 130 cm $\times$180 cm. To generate turbulence, an asymmetric bluff body was mounted to the wind tunnel using bungee cords, creating a wake of irregular and turbulent flow. The bluff body consisted of a large diameter cylinder (30 cm) with an asymmetrically mounted flat plate at the front, giving the entire body an effective diameter of 53 cm (Figure 6.2C). To encourage vortex dislocation and add irregularity to vortex shedding, the cylinder spanned the full width of the tunnel, while the flat plate only had a width of 60 cm. The bluff body was positioned 170 cm upstream of the wing system, with a vertical offset of 48 cm, and sparse elastic bands were placed horizontally across the test section immediately upstream of the bluff body to further increase turbulence intensity.

We used particle image velocimetry (PIV) (Fig 6.4) to visualize a limited portion of the bluff body wake. PIV is a quantitative flow visualization technique capable of measuring the velocity fields of fluid flows [293]. Here we performed two-dimensional, two-component (2D2C) PIV. This involves using a laser sheet to illuminate small, dense, neutrally buoyant seed particles. Recording the illuminated

particles with a high-speed camera, we can estimate velocity fields by calculating the effective inter-frame displacement of groups of particles via cross-correlation of subsequent images. Performing these experiments in air, we used a 200 mJ/pulse dual pulsed laser (Lumibird Evergreen) to illuminate soap bubbles (15-micron mean diameter) generated with a custom-built bubbler. The flow was recorded with a 4-MP CCD camera (IMPERX Bobcat B2401).

Characterization of the flow near the wing system was performed using a hot-wire anemometer (TSI IFA-300). The anemometer was mounted approximately 2 cm upstream of the leading edge of the airfoil, and measurements were taken at 1000 Hz for 120 seconds. The turbulence intensity was determined to be 10.8% using the hot-wire anemometer. We found the dominant shedding frequency to be 2.44 Hz, although as mentioned above our oscillating bluff body encouraged irregularities in the shedding process.

Baseline algorithms

In our experiments, besides FALCON, we test several model-free RL methods, including Twin Delayed DDPG (TD3) [92], LSTM-TD3 [196], Soft Actor-Critic (SAC) [102], and the industry-standard responsive control strategy of PID controller. Of all these algorithms LSTM-TD3 has recently demonstrated to achieve state-of-the-art performance in predictive flow disturbance rejection [228]. Note that both TD3 and LSTM-TD3 provide deterministic policies, whereas SAC designs stochastic policies. Unlike FALCON, these model-free methods work in episodes where the algorithms stop retraining the policy parameters (reset). We train the model-free methods for 200 episodes of 800 samples per episode and test the best-performing policy in presenting the results. The full 200 episodes of training for the best-performing policy in each method are shown in Figure 6.6. All methods were implemented using an NVIDIA GeForce RTX 3070 which enabled a 42 Hz frequency for sensing and control. The brief descriptions of the algorithms are given below with the relevant hyperparameters in Tables 6.4-6.7.

TD3: TD3 is a deterministic actor-critic type RL framework that builds on previous value-based methods. TD3 injects noise into actions to enable policy exploration (i.e., exploration noise), and injects noise into critic updates to regularize and smooth the resulting policy (i.e., policy smoothing noise). TD3 also uses delayed policy updates which decreases variance in value estimates. For this work, we

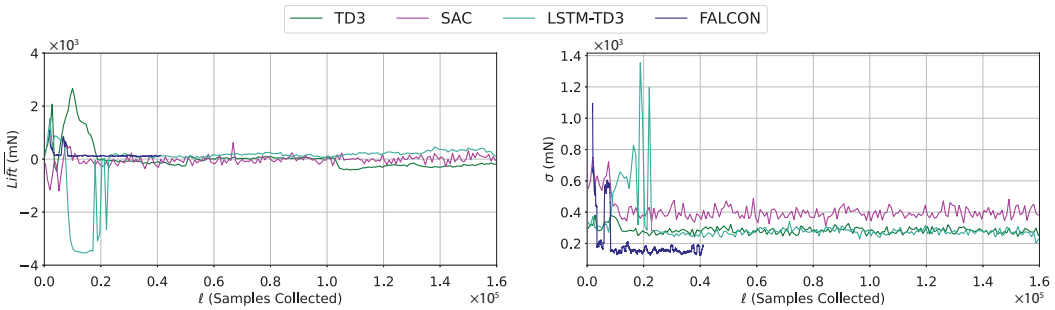

Figure 6.6: Evolution of the mean ($\overline{\text{Lift}}$) and standard deviation (σ) of the lift forces for the best-performing agents of each algorithm shown over the full 200 episodes for which the model-free algorithms were trained (160, 000 samples).

Table 6.4: Hyperparameters of TD3 in our experiments.

Hyperparameter	**Range**	**Best**
Discount factor	0.95 – 0.99	0.99
Batch size	16 – 128	50
Replay buffer size	$2 - 6 \times 10^4$	4×10^4
Target update rate	0.005	0.005
Actor learning rate	$10^{-5} - 10^{-2}$	10^{-4}
Critic learning rate	$10^{-5} - 10^{-2}$	10^{-4}
Exploration noise	0.025 – 0.2	0.05
Policy smoothing noise	0.025 – 0.2	0.05
Policy update delay	2 – 3	3
Target noise clip boundary	0.5	0.5
Actor gradient clip boundary	0.1 – 1	0.5
Critic gradient clip boundary	0.1 – 1	0.5

added gradient clipping to both actor and critic networks to encourage more stable learning in a real-world setting. TD3 has been proven effective in several simulated [92] environments and has previously been used for experimental flow control in a different setting [80]. For further implementation details of TD3 please refer to [92] and the code provided in the submission.

LSTM-TD3: LSTM-TD3 uses the same fundamental algorithm as TD3 but includes LSTM cells for a recurrent actor-critic framework. It was modified from TD3 to better address problems suffering from partial observability [196]. For further implementation details of LSTM-TD3 please refer to [196] and the code provided in the submission.

Table 6.5: Hyperparameters of LSTM-TD3 in our experiments.

Hyperparameter	**Range**	**Best**
Discount factor	0.95 – 0.99	0.99
Batch size	16 – 128	50
Replay buffer size	$2 - 6 \times 10^4$	4×10^4
Target update rate	0.005	0.005
Actor learning rate	$10^{-5} - 10^{-2}$	10^{-4}
Critic learning rate	$10^{-5} - 10^{-2}$	10^{-4}
Exploration noise	0.025 – 0.2	0.05
Policy smoothing noise	0.025 – 0.2	0.05
Policy update delay	2 – 3	3
Target noise clip boundary	0.5	0.5
Actor gradient clip boundary	0.1 – 1	0.5
Critic gradient clip boundary	0.1 – 1	0.5
LSTM Depth	3 – 15	10

SAC: This method is based on maximum entropy RL framework that wants to maximize the performance/reward concurrently maximizing the entropy of the policy, i.e., increase the randomness in the policy. Intuitively, this method results in a stochastic policy that achieves good performance and provides the most randomness in achieving this result. SAC uses a temperature parameter to weigh the entropy term relative to the reward function. The ideal temperature parameter can be automatically learned during training. This method is initially proposed for real-world robotic learning to facilitate smooth exploration, tolerate unexpected perturbations/changes during execution, and improve robustness to hyperparameters and sample efficiency. To this end, it requires relatively less hyperparameters compared to other model-free RL methods. For further implementation details of SAC please refer to [102] and the code provided in the submission.

PID: PID control is the most prevalent method found in industrial and commercial applications. PID controllers use a basic feedback control loop that attempts to minimize the error between an observed value and a desired setpoint. PID controllers weigh a proportional term, an integral term, and a derivative term, all of which are tuned for each specific application. As the name suggests, the proportional term contributes a control signal that is directly proportional to the magnitude of the error. The integral term provides a signal that corresponds to the running accumulated error but is slow to react. The derivative term sends a control signal that is proportional to the error rate of change, which effectively smooths behavior. All three of these

Table 6.6: Hyperparameters of SAC in our experiments.

Hyperparameter	Range	Best
Discount factor	0.95 – 0.99	0.99
Batch size	32 – 256	128
Replay buffer size	$2 \times 10^4 - 6 \times 10^4$	4×10^4
Actor learning rate	$10^{-5} - 10^{-2}$	5×10^{-4}
Critic learning rate	$10^{-5} - 10^{-2}$	10^{-3}
Target smoothing coefficient	$10^{-3} - 10^{-2}$	10^{-2}
Target entropy	$-2 - -0.5$	-0.5
Temperature learning rate	$10^{-5} - 10^{-2}$	5×10^{-4}

terms are weighed through corresponding constant values (i.e., K_P, K_I, K_D) that can be found through various tuning methods. For further implementation details of PID please refer to the code provided in the submission.

Table 6.7: Hyperparameters of PID in our experiments.

Hyperparameter	Range	Best
K_P	0 – 15	10
K_I	0 – 2	0.5
K_D	0 – 5	2

6.3.4 Stability and Performance Guarantees for FALCON

In this section, we provide the learning and regret guarantees of FALCON. The technical details and the proofs are given in Appendix D.1. First, let $F_i(\cdot) : \mathbb{R}^{h(d_y+d_u)} \to \mathbb{R}$ denote the ith mapping of F from input to output, i.e., $y_{t,i} = F_i(s_t) + e_{t,i}$. We assume that Assumptions 6.1, 6.2, and 6.4 hold. Assumption 6.2 is required to avoid blow-ups in the output due to noise and unmodeled dynamics and Assumption 6.4guarantees that the underlying system can be represented on a Fourier basis. For simplicity, we assume that $e_t \sim \mathcal{N}(0, \sigma_e^2 I)$, yet our technical results hold for sub-Gaussian noise. The regret of FALCON is computed as discussed in Section 6.1.3. For consistent and reliable initial estimation of the underlying system, we assume that FALCON uses bounded persistently exciting inputs during the warm-up period. Given these inputs, we have the following learning guarantee.

Proposition 6.1. *Let* $\alpha_m = \sup_{i,\|s\|\leq S} \|\partial^m F_i(s)\|_{L^\infty}$ *and* $d = h(d_y + d_u)$. *Using nth order Fourier basis for learning the model for sufficiently large n, after the warm-up of* T_w, *with high probability* $\sup_{\|s\|\leq S} \|F(s) - \hat{F}(s)\|_\infty = O(n^d T_w^{-0.5} + \alpha_m n^{-m})$.

Here $\tilde{O}(\cdot)$ presents the order up to logarithmic terms. The proof is given in Appendix D.1, where we use standard least-squares estimation error analysis and Theorem 6.2. This result shows that the underlying system can be identified with the optimal rate of $1/\sqrt{T}$, yet, due to the properties of the underlying system, there exists a constant term in the estimation error. Note that this constant term depends on the smoothness of the system. For nonlinear systems that live in high-order Sobolev spaces m, this constant term can be small. In the extreme case of infinitely differentiable systems, this constant term approaches to 0. Thus, we have $\sup_{\|s\|\leq S}\|F(s)-\hat{F}(s)\|_\infty = \tilde{O}(T_w^{\varepsilon-0.5})$ after warm-up, where ε depends on the smoothness of the system and the order of Fourier basis. Next, we focus on the adaptive control task. We have the following assumption on the MPC policy that FALCON deploys.

Assumption 6.8. *The MPC policy with $F(\cdot)$ achieves e-IOS, i.e., $\forall t > t_0$, $\|y_t\| \leq (1-\rho)^{t-t_0}\|y_{t_0}\| + M \sup_{i\in[t_0:t]} \|e_i\|$, for $M > 0$ and $0 < \rho < 1$. The MPC policy with planning model $\hat{F}(\cdot)$, such that $\sup_{\|s\|\leq S} \|F(s)-\hat{F}(s)\|_\infty \leq \epsilon$, achieves e-IOS with $\rho/2$ and $2M$ and synthesizes persistently exciting inputs which are locally L_o-Lipschitz in planning model.*

This assumption states that MPC stabilizes the underlying system by using any model within a neighborhood around the underlying system for planning. This assumption is mild and one can show that it holds for linear systems. Intuitively, this assumption holds for nonlinear systems with valid linearization for bounded inputs. Finally, the last statement allows consistent estimation during the adaptive control with reasonable variations in the input due to model dynamics used in planning. In practice, this condition is usually satisfied by the combination of unmodelled system dynamics, system noise, and sampling-based MPC policies [164]. Note that for a long enough warm-up T_w, the model estimation error can be made small enough to achieve stabilization via the MPC policy. Once FALCON is guaranteed to stabilize the dynamics, it safely regulates the system.

Theorem 6.4. *Suppose Assumptions 6.1, 6.2, 6.4, and 6.8 hold. Let T_w be chosen long enough such that the MPC policy of FALCON stabilizes the underlying system dynamics. Then, with high probability, for large enough T, FALCON attains regret of $\textsc{Regret}(T) = \tilde{O}(\sqrt{T} + \epsilon' T)$, where ϵ' depends on the smoothness of the underlying system. For sufficiently smooth systems, it achieves $\textsc{Regret}(T) = \tilde{O}(\sqrt{T})$.*

The proof is given in Appendix D.1. This shows that FALCON is the first efficient RL algorithm that attains $O(\sqrt{T})$ regret in adaptive control of partially observable nonlinear systems. Note that this result applies to various systems that are governed by partial differential equations since FALCON only requires the periodic extension of the model dynamics to live in the Sobolev space of periodic functions. Moreover, for infinitely smooth systems, e.g., sinusoidal systems, one can improve Proposition 6.1, and this in turn would give significantly improved regret upper bound for FALCON.

Corollary 6.4.1. *Under the setting of Theorem 6.4, for an infinitely smooth system, i.e.,* $F_i \in W_p^{\infty,2}([0, 2\pi]^{h(d_y+d_u)})$ *for* $1 \leq i \leq d_y$*, with high probability, FALCON attains* $\textsc{Regret}(T) = \mathit{polylog}(T)$.

This shows that for a certain class of dynamical systems, using the domain knowledge on the system dynamics FALCON can achieve almost logarithmic regret even if the underlying system is unknown.

6.4 Stability Constrained Model-Based RL

Finally, in this section, we study the problem of stabilizing policy design for unknown nonlinear dynamical systems. Note that this problem is avoided in Sections 6.2 and 6.3 by assuming that the underlying MPC policy stabilizes the system dynamics even under modeling error, namely Assumptions 6.6 and 6.8, respectively. In this section, we formally propose a policy optimization problem for a class of nonlinear dynamical systems, whose solution is guaranteed to stabilize the underlying system even under modeling error.

To achieve this, we integrate control theoretic tools into policy optimization in RL. In particular, we propose a policy optimization problem that adapts Krasovskii's construction of quadratic Lyapunov functions [148] as a stability constraint, which guarantees that the Lyapunov stability conditions are met by design for the solution of the policy optimization problem (Theorem 6.5). Further, we show that this stabilization guarantee holds for the controllers obtained using a learned model of system dynamics in the policy optimization for small enough modeling errors (Theorem 6.6).

To adapt this stability-constrained policy optimization problem into a reinforcement learning pipeline, we propose a primal-dual method. We show that the primal-dual method guarantees the satisfaction of the stability constraint and the design of a stabilizing policy for the underlying system after convergence (Theorem 6.7). This

allows us to design a novel model-based RL framework, Krasovskii-Constrained RL (KCRL), via combining model learning and the proposed policy optimization method. KCRL learns the unknown model dynamics in epochs and solves the proposed stability-constrained policy optimization problem via the primal-dual method using the learned model.

We study the KCRL framework both theoretically and empirically. On the theory side, we consider KCRL with kernel-based feature representations for model learning, i.e. Random Fourier Features (RFF) introduced in Section 6.1 and used in MLPC. We show that KCRL with RFF-learning formally guarantees the design of stabilizing control policies in finite time/samples (Theorem 6.8). On the empirical side, we study the performance of the KCRL framework in learning a stable policy for voltage control in a distributed power system with different operating conditions obtained via real-world operation data. We show that KCRL guarantees stability under all operating conditions, whereas the standard RL methods fail in stabilizing.

Related Work

Our work connects to a broad set of control and RL literature.

Lyapunov theory is a systematic framework to analyze the stability of a control system. To prove stability, Lyapunov's direct method aims to define a positive definite function, that decreases along the system trajectory, *i.e.*, a Lyapunov function. There is a large body of tools in control such as Krasovskii's method [148], contraction theory [190], feedback linearization [153], and passivity theory [247], which provide ways to construct Lyapunov function candidates and analyze the stability of the systems. In our work, we consider Krasovskii's method in designing stabilizing policies for systems with modeling errors. In the context of these control theoretic tools, our contributions bridge one of the classical tools in control with policy optimization in RL.

Control Lyapunov functions (CLFs) are popular tools in designing stabilizing controllers and they are also closely related to our framework [248]. In the construction of CLFs, it is often assumed that the system is control-affine, or more generally input-output linearizable [16]. For such systems, the Lyapunov function design problem simplifies due to linearized system dynamics [15]. However, to achieve such input-output linerization, existing works either assume the knowledge of the model dynamics or assume that the CLF constructed for the learned model is *also* a CLF for the underlying system [59, 259, 277, 292]. In this work, we do not have these

assumptions on the system dynamics or the constructed Lyapunov function, which are violated in many practical systems. Instead, we consider nonlinear systems that admit Krasovskii's family of Lyapunov functions and provide an end-to-end RL method, KCRL, which designs stabilizing controllers for the underlying system using model estimates. In particular, we quantify the amount of modeling error that KCRL can tolerate for stabilization.

Model-based RL in dynamical systems has been studied in many recent works due to its superior sample efficiency and interpretable guarantees. The main focus has been on learning the system dynamics and providing performance guarantees in finite-time for both linear [163] (and references within), and nonlinear systems [164]. While deriving these guarantees, the formal finite-time stability guarantees are also derived for linear systems [82]. However, these guarantees have only been *assumed to hold* with a stabilizing oracle for nonlinear systems [164]. Our work provides formal finite-time (sample) stabilization guarantees for nonlinear systems without these assumptions.

Stability Guarantees in Learning-based Control. What we present here is one among many directions on incorporating stability guarantees in learning-based control, with a focus on incorporating stability guarantees for policy optimization (PO) based RL algorithms. For the benefit of readers from both the learning and control communities, we highlight a few results from this vast and growing literature. The stability of learning-based MPC was established in [20, 231] and followed, for nonlinear systems, by efforts on joint learning of the controller and(or) Lyapunov functions [49, 51, 52, 66, 69]. [33, 70] studied learning of stability certificates and stable controllers from data, and [27] developed a provably stable data-driven algorithm based on system measurements and prior system knowledge. Another line of work considers incremental stability for nonlinear systems using contraction theory and convex optimization with modeling errors [271, 272]. Different from existing works, we construct the Lyapunov function based on Krasovskii's method (rather than learning the Lyapunov function from scratch or data) and train the policy network to satisfy the stability conditions derived from Krasovskii's method. In addition, to incorporate stability guarantees to policy optimization methods in RL, there have been works that proved stability and convergence for actor-critic based RL methods [179, 279] and Q-learning [278].

6.4.1 Preliminaries

Consider a discrete-time nonlinear system given as

$$x_{t+1} = f(x_t, u_t), \tag{6.21}$$

where $x_t \in \mathbb{R}^n$ is the state of the system, $u_t \in \mathbb{R}^p$ is the control input at time-step t. We study the discrete optimal control setting for the system given in (6.21). Suppose there is a class of controllers $g_\theta(\cdot)$, parameterized by $\theta \in \Theta$. The goal is to design a controller $g_\theta(\cdot)$ that minimizes a control cost,

$$\min_\theta J(\theta) = \sum\nolimits_{t=0}^{\infty} \gamma^t c(x_t, u_t), \tag{6.22a}$$

$$\text{s.t. } x_{t+1} = f(x_t, u_t)\,, u_t = g_\theta(x_t), \tag{6.22b}$$

where $c(x, u)$ is the cost and γ is the discounting factor. Note that there are many ways to solve or approximate the policy optimization problem (6.22). Generally speaking, the procedure is to run gradient methods on the policy parameter θ with step size η, $\theta \leftarrow \theta – \eta \nabla J(\theta)$. To approximate the gradient $\nabla J(\theta)$, one can use sampled trajectories such as REINFORCE or value function approximation such as actor-critic methods. As we are dealing with deterministic policies, one of the most popular choices is the Deep Deterministic Policy Gradient (DDPG) [185], where the policy gradient is approximated by $\nabla J(\theta) \approx \frac{1}{N} \sum_{i \in B} \nabla_u \hat{Q}(x, u)|_{x=x_i, u=g_\theta(x_i)} \nabla_\theta g_\theta(x)|_{x_i}$. Here $\hat{Q}(x, u)$ is the value (critic) network that can be learned via temporal difference learning, $g_\theta(x)$ is the actor network, and $\{x_i\}_{i \in B_J}$ are a batch of samples with batch size $|B_J| = N$ sampled from the replay buffer which stores historical state-action pairs. For further details on DDPG, please refer to [185].

Stability

In control systems, stability studies whether the state trajectory of the closed-loop system $x_{t+1} = f(x_t, g_\theta(x_t))$ asymptotically converges to the desired stationary point or a set of stationary points. The following formally defines stability in our context, using the notation $\text{dist}(x, S) := \inf_{y \in S} \|y - x\|$ to denote the distance between point x and set S.

Definition 6.1 (Asymptotically stable equilibrium)**.** *A dynamical system $x_{t+1} = f(x_t, g_\theta(x_t))$ is asymptotically stable around $x^{(e)}$ if $f(x^{(e)}, g_\theta(x^{(e)})) = x^{(e)}$, and further, there exists a region around $x^{(e)}$, $B_\delta(x^{(e)}) = \{x : \|x – x^{(e)}\| \leq \delta\}$ such that $\forall x_0 \in B_\delta(x^{(e)})$, we have $\lim_{t \to \infty} \|x_t - x^{(e)}\| = 0$. A dynamical system is globally asymptotically stable around $x^{(e)}$ if the same holds for all possible initial states $x_0 \in \mathbb{R}^n$.*

More generally, the following definition considers a set of equilibrium points, where we use the notation $\text{dist}(x, S) := \inf_{y \in S} \|y - x\|$ to denote the distance between point x and set S.

Definition 6.2. *(Asymptotically stable set) A dynamical system $x_{t+1} = f(x_t, g_\theta(x_t))$ is asymptotically stable around set S_e if $f(x^{(e)}, g_\theta(x^{(e)})) = x^{(e)}, \forall x^{(e)} \in S_e$, and further, there exists $B_\delta(S_e) := \{x : \text{dist}(x, S_e) \leq \delta\}$ such that $\forall x_0 \in B_\delta(S_e)$, we have $\lim_{t\to\infty} \text{dist}(x_t, S_e) = 0$. The system is globally asymptotically stable around S_e if the same holds for any initial state $x_0 \in \mathbb{R}^n$.*

A common approach to prove the stability of a dynamical system with respect to an equilibrium is via Lyapunov's direct method and a generalization of Lyapunov's method, known as LaSalle's Invariance Principle for proving stability to a set. Both involve defining a positive definite function that decreases along the system trajectory, *i.e.*, a Lyapunov function V. Please refer to [142] for a more complete overview.

In this section, we study problem (6.22) under unknown system dynamics. Note that for $\phi_t = [x_t^\top, u_t^\top]^\top$, one can write the system dynamics given in (6.21) as

$$x_{t+1} = F(\phi_t), \tag{6.23}$$

for some nonlinear function F. Further, we denote the closed-loop system dynamics obtained via the policy $u_t = g_\theta(x_t)$ as

$$x_{t+1} = F_\theta(x_t), \tag{6.24}$$

where $F_\theta(x_t) = F(\phi_t)$ for $\phi_t = [x_t^\top, g_\theta(x_t)^\top]^\top$. To ease the presentation, we use both notations interchangeably throughout this work. Suppose that F and g_θ are both continuously differentiable. Let $G(x, \theta)$ denote the true Jacobian of the closed-loop system with respect to state x, i.e., $G(x, \theta) = \frac{\partial F(\phi)}{\partial x} + \frac{\partial F(\phi)}{\partial u}\frac{\partial u}{\partial x}$. For discrete-time dynamical systems as in (6.24), Krasovskii's Lyapunov function candidate follows,

$$V(x) = (x - F_\theta(x))^\top M (x - F_\theta(x)), \tag{6.25}$$

such that there exists a pair (M, θ), where $M \succ 0$ and $G(x, \theta)^\top M G(x, \theta) - M \preceq 0$. In this work, we assume that the underlying system in (6.23) satisfies the Krasovskii's Lyapunov function construction for an (M, θ) pair with a stability margin, i.e., for some $\bar{\epsilon} > 0$,

$$G(x, \theta)^\top M G(x, \theta) - M \preceq -\bar{\epsilon} I. \tag{6.26}$$

Remark 6.2. *The stability margin is required to accommodate modeling errors in the dynamics. If one has access to the true model, $F(\cdot)$, $\bar{\epsilon} = 0$ would suffice, i.e., asymptotic stability.*

6.4.2 Krasovskii-Constrained Policy Optimization

In this section, we introduce our novel stability-constrained policy optimization problem and prove that its solution is a stabilizing policy under perfect model dynamics and also under modeling errors. We then provide a primal-dual policy gradient approach to solve this problem using a learned model and show that it finds a stabilizing policy.

Stabilizing Policy Design

Using Krasovskii's method of constructing Lyapunov functions for the underlying system described in Section 6.4.1, we add a stability constraint into the standard policy optimization problem in (6.22). In particular, for a given (estimated) model $\hat{F}(\cdot)$ on the true system dynamics $F(\cdot)$, we propose to solve the following constrained optimization problem

$$\min_{\theta} \sum_{t=0}^{T} \gamma^t c(x_t, u_t), \tag{6.27a}$$

$$\text{s.t. } x_{t+1} = F(\phi_t)\,, u_t = g_\theta(x_t), \tag{6.27b}$$

$$\hat{G}(x,\theta)^\top M \hat{G}(x,\theta) - M \prec -\epsilon_i I, \;\; \forall x \in \mathcal{X}, \tag{6.27c}$$

where $M \succ 0$, $\hat{G}(x,\theta) = \widehat{\frac{\partial F(\phi)}{\partial x}} + \widehat{\frac{\partial F(\phi)}{\partial u}} \frac{\partial u}{\partial x}$ for the Jacobian estimates $\widehat{\frac{\partial F(\phi)}{\partial x}}$ and $\widehat{\frac{\partial F(\phi)}{\partial u}}$ which can be computed via finite difference method using $\hat{F}(\cdot)$, and $\bar{\epsilon} \geq \epsilon_i > 0$, which is chosen based on the modeling error in $\hat{F}(\cdot)$ as discussed shortly.

Compared to (6.22), the formulation (6.27) incorporates an additional constraint (6.27c). This constraint adapts Krasovskii's method for Lyapunov function construction and enforces the stability of the learned policy. In what follows, we show that the solution of the novel stability-constrained policy optimization problem (6.27) using the true system $F(\cdot)$, particularly the true Jacobians in (6.27c), is a stabilizing policy by design.

Theorem 6.5 (Stability of the True Discrete-time System)**.** *Consider solving* (6.27) *with the knowledge of true model $F(\cdot)$, i.e.,* (6.27c) *is evaluated using the true Jacobian. Let $\theta_\star$ denote the solution of* (6.27)*, such that* (6.27c) *holds for some ϵ_i, where $\bar{\epsilon} \geq \epsilon_i \geq 0$. Then, we have the trajectory of $x_{t+1} = F_{\theta_\star}(x_t)$ is asymptotically stable around the origin, $F_{\theta_\star}(0) = 0$.*

Proof. Following Krasovskii's construction method gives the candidate Lyapunov function of

$$V_f(x) = (x - f(x))^\top M (x - f(x)), \tag{6.28}$$

for the underlying system $f(x)$. Note that we are considering the non-autonomous systems $F(\phi)$, where $\phi = [x^\top, u^\top]^\top$ given the controller $u = g_\theta(x)$. Let $F_\theta(x)$ denote the closed-loop system dynamics obtained via controller $g_\theta(x)$, i.e. $F_\theta(x) = F(\phi)$ with $\phi = [x^\top, g_\theta(x)^\top]^\top$. Therefore, we consider the following Lyapunov function:

$$V(x) = (x - F_\theta(x))^\top M(x - F_\theta(x)). \tag{6.29}$$

Firstly, note that $V(x) \geq 0$, and $V(x) = 0$ if and only if $x \in S_e$ as M is positive definite. Then, we consider

$$V(x_{t+1}) - V(x_t) = (x_{t+1} - F_\theta(x_{t+1}))^\top M(x_{t+1} - F_\theta(x_{t+1})) - (x_t - F_\theta(x_t))^\top M(x_t - F_\theta(x_t)). \tag{6.30}$$

First, consider the following for $h(x) = x - F_\theta(x)$. One can write $h(x_{t+1})$ in terms of $h(x_t)$ as follows,

$$h(x_{t+1}) = h(x_t) + \int_0^1 \frac{\partial h}{\partial x}(x_t + t(x_{t+1} - x_t))(x_{t+1} - x_t)dt.$$

Recall Kowalewski's Mean Value Theorem [146].

Proposition 6.2 (Kowalewski's Mean Value Theorem [146])**.** *Let $x_1, \ldots, x_n$ be continuous functions in a variable $t \in [a, b]$. There exists real numbers $t_1, \ldots, t_n$ in $[a, b]$ and non-negative $\lambda_1, \ldots, \lambda_n$, with $\sum_i^n \lambda_i = b - a$, such that*

$$\int_a^b x_k(t)dt = \sum_{i=1}^n \lambda_i x_k(t_i),$$

for each $k = 1, \ldots, n$.

From Proposition 6.2, we have

$$h(x_{t+1}) = h(x_t) + J_h(x_{t+1} - x_t),$$

where $J_h = \sum_{i=1}^n \lambda_i \frac{\partial h}{\partial x}(x_t + k_i(x_{t+1} - x_t))$ for $k_i \in [0, 1]$, $\lambda_i \geq 0$ for all i and $\sum_i^n \lambda_i = 1$. Plugging this in $V(x_{t+1})$, we get

$$V(x_{t+1}) = h(x_t)^\top M h(x_t) + 2(x_{t+1} - x_t)^\top J_h^\top M h(x_t) + (x_{t+1} - x_t)^\top J_h^\top M J_h(x_{t+1} - x_t)$$

Note that $x_{t+1} - x_t = F_\theta(x_t) - x_t = -h(x_t)$. Therefore, we get

$$\begin{aligned} V(x_{t+1}) &= h(x_t)^\top M h(x_t) - 2h(x_t)^\top J_h^\top M h(x_t) + h(x_t)^\top J_h^\top M J_h h(x_t) \\ &= h(x_t)^\top (I - J_h)^\top M(I - J_h)h(x_t). \end{aligned}$$

From the definition of J_h and $h(x) = x - F_\theta(x)$, we have $J_h = I - J_{F_\theta}$, where $J_{F_\theta} = \sum_{i=1}^{n} \lambda_i \frac{\partial F_\theta}{\partial x}(x_t + k_i(x_{t+1} - x_t))$, where k_i and λ_i follow from the definition of J_h. Thus we get

$$V(x_{t+1}) = (x_t - F_\theta(x_t))^\top J_{F_\theta}^\top M J_{F_\theta}(x_t - F_\theta(x_t)). \tag{6.31}$$

Plugging this in (6.30) gives

$$V(x_{t+1}) - V(x_t) = (x_t - F_\theta(x_t))^\top (J_{F_\theta}^\top M J_{F_\theta} - M)(x_t - F_\theta(x_t)). \tag{6.32}$$

For any $x \in \mathbb{R}^n$, we have

$$\begin{aligned}
x^\top J_{F_\theta}^\top M J_{F_\theta} x &= \left\| M^{1/2} J_{F_\theta} x \right\|^2 \\
&= \left\| \sum_{i=1}^{n} \lambda_i M^{1/2} \frac{\partial F_\theta}{\partial x}\left(x_t + k_i(x_{t+1} - x_t)\right) x \right\|^2 \\
&\leq \sum_{i=1}^{n} \lambda_i \left\| M^{1/2} \frac{\partial F_\theta}{\partial x}\left(x_t + k_i(x_{t+1} - x_t)\right) x \right\|^2 \\
&= \sum_{i=1}^{n} \lambda_i x^\top \frac{\partial F_\theta}{\partial x}\left(x_t + k_i(x_{t+1} - x_t)\right)^\top M \frac{\partial F_\theta}{\partial x}\left(x_t + k_i(x_{t+1} - x_t)\right) x,
\end{aligned}$$

where the inequality is due to Jensen's inequality. Due to the constraint (6.27c), we have that $\frac{\partial F_\theta}{\partial x}(x)^\top M \frac{\partial F_\theta}{\partial x}(x) \prec M - \epsilon I$ for all $x \in \mathbb{R}^n$. Thus, we have that

$$J_{F_\theta}^\top M J_{F_\theta} - M \preceq \tag{6.33}$$

$$\sum_i \lambda_i \left[\frac{\partial F_\theta}{\partial x}\left(x_t + k_i(x_{t+1} - x_t)\right)^\top M \frac{\partial F_\theta}{\partial x}\left(x_t + k_i(x_{t+1} - x_t)\right) - M \right] \preceq -\epsilon I. \tag{6.34}$$

This shows that the Lyapunov function is decreasing along the system trajectory, i.e. $V(x_{t+1}) - V(x_t) = (x_t - F_\theta(x_t))^\top (J_{F_\theta}^\top M J_{F_\theta} - M)(x_t - F_\theta(x_t)) < 0$.

Lastly, if a trajectory x_t is such that $V(x_{t+1}) - V(x_t) = 0, \forall t \geq 0$, then we must have $F_\theta(x_t) = x_t$ for all t, i.e. $x_t \in S_e$ for all t. Therefore, by LaSalle's Invariance Principle, we must have S_e asymptotically stable. □

Note that Theorem 6.5 uses the exact Jacobians rather than estimates obtained via the finite difference method.

Effect of Modeling Errors

We extend this result to tolerate modeling errors, i.e., errors in the Jacobian estimates. First, we quantify some regularity conditions of the system and the policy class.

Assumption 6.9 (Regularity Conditions)**.** *(i) F is L_F-Lipschitz, i.e., we have Jacobian of F, $\|J_F\| \le L_F$. (ii) $\|\nabla^2 F_i\| \le \mathbf{F_H}$, $\forall i$, where F_i denotes the mapping from ϕ_t to ith element of state vector x_{t+1}, i.e., $(x_{t+1})_i = F_i(\phi_t)$ for $i = 1, \ldots, n$. (iii) Policies in the policy class are L_u-Lipschitz, that is, $\|\frac{\partial g_\theta(x)}{\partial x}\| \le L_u$, $\forall \theta$.*

Note that in practice, one can use loose upper bounds for these system-related quantities and update them over time. The following shows that solving (6.27) using a well-refined model estimate $\hat{F}$ and an appropriate choice of ϵ_i guarantees the recovery of stabilizing policy for the underlying system.

Theorem 6.6 (Stability under Modeling Error)**.** *Suppose Assumption 6.9 holds and the Jacobian estimates obtained using a model estimate $\hat{F}(\cdot)$ satisfy*

$$\sup_x \max\left(\left\|\widehat{\frac{\partial F_i(\phi)}{\partial x}} - \frac{\partial F_i(\phi)}{\partial x}\right\|, \left\|\widehat{\frac{\partial F_i(\phi)}{\partial u}} - \frac{\partial F_i(\phi)}{\partial u}\right\|\right) \le \varepsilon_J < 1,$$

for all $i = 1, \ldots, n$. Let $\theta_\star$ be the solution of (6.27) *using these Jacobian estimates in* (6.27c) *for ϵ_i, such that $\bar{\epsilon} \ge \epsilon_i \ge 2\bar{G}\|M\|(1+L_u)\varepsilon_J + \|M\|(1+L_u)^2\varepsilon_J^2$, where $\bar{G} = (1+L_u)(L_F + \varepsilon_J)$. Then, we have the trajectory of $x_{t+1} = F_{\theta_\star}(x_t)$ is asymptotically stable around the origin.*

Proof. By Theorem 6.5, we only need to show the following: (here we drop θ dependence as it is fixed in the proof)

$$G(x)^\top M G(x) - M \prec 0, \forall x \in \mathcal{X}. \tag{6.35}$$

Let $\Delta G_i \coloneqq \hat{G}_i(x) - G_i(x)$. Using Assumption 6.9 and the construction of $G_i(x)$, we first bound $\|\Delta G_i\|$:

$$\|\Delta G_i\| \le (1+L_u)\varepsilon_J. \tag{6.36}$$

Next, note that the stability constraint (6.27c) indicates $\hat{G}_i(x)^\top M \hat{G}_i(x) - M \prec -\epsilon_i I$. Using this, we get

$$\begin{aligned} G_i(x)^\top M G_i(x) &= \hat{G}_i(x)^\top M \hat{G}_i(x) + \Delta G_i^\top M \hat{G}(x) \\ &\quad + \hat{G}_i(x)^\top M \Delta G_i + \Delta G_i^\top M \Delta G_i \\ &\le M - \epsilon_i I + 2\bar{G}\|M\|\|\Delta G_i\| I + \|M\|\|\Delta G_i\|^2 I \prec M, \end{aligned}$$

where in the final step, we use (6.36) and the choice of ϵ_i. This verifies (6.35) and gives the advertised result. □

Theorems 6.5 and 6.6 show that the solution of (6.27) stabilizes the underlying system even under modeling errors. To use this framework in online policy optimization, one requires to solve this constrained optimization effectively. To that end, we propose a primal-dual policy gradient technique.

Primal-Dual Approach

In the following, we describe the primal-dual technique to solve (6.27) and show that the convergence of this method guarantees the satisfaction of stability condition in (6.27c) with appropriate algorithmic choices. We use the following short-hand notation, $K(x,\theta) = \hat{G}(x,\theta)^\top M \hat{G}(x,\theta) - M + \epsilon_i I$. With this, (6.27) can be reformulated as

$$\min_{\theta} J(\theta) \text{ s.t. } \sup_x \lambda_{\max}(K(x,\theta)) < 0,$$

where $\lambda_{\max}(\cdot)$ is the largest eigenvalue. The Lagrangian for the problem is given as $L(\theta,\mu) = J(\theta) + \mu \sup_x \lambda_{\max}(K(x,\theta))$. The primal-dual algorithm then proceeds as follows [207],

$$\begin{aligned}
\theta &\leftarrow \theta - \eta_1 \Big[\nabla J(\theta) + \mu \nabla_\theta \sup_x \lambda_{\max}(K(x,\theta)) \Big], \\
\mu &\leftarrow \max(0, \mu + \eta_2 \sup_x \lambda_{\max}(K(x,\theta))).
\end{aligned}$$

Since it is not possible to evaluate $\sup_x$, we replace it with a supremum over a batch of representative points in the state space $\{x_i\}_{i\in\mathcal{B}}$. For the term $\nabla J(\theta)$, we use standard policy gradient estimators, e.g. DDPG [185], to evaluate the policy gradient and denote the estimated gradient as $\widehat{\nabla J(\theta)}$. Thus, the primal-dual algorithm is given as,

$$\begin{aligned}
\theta &\leftarrow \theta - \eta_1 \Big[\widehat{\nabla J(\theta)} + \mu \nabla_\theta \sup_{i\in\mathcal{B}} \lambda_{\max}(K(x_i,\theta)) \Big], \\
\mu &\leftarrow \max(0, \mu + \eta_2 \sup_{i\in\mathcal{B}} \lambda_{\max}(K(x_i,\theta) + \epsilon_{\text{pd}} I)),
\end{aligned} \tag{6.37}$$

where $\eta_1, \eta_2 > 0$ are the step sizes and $\mathcal{B}$ is a batch of representative points in the state space, and $\epsilon_{\text{pd}} > 0$ is a constant that is chosen to tolerate the possible representation incapability of $\mathcal{B}$. The following gives the characterization of ϵ_{pd} to verify that the solution obtained via the primal-dual method stabilizes the underlying system.

Theorem 6.7 (Convergence of Primal-Dual Algorithm Guarantees Stability)**.** *Suppose the primal-dual procedure converges, then the stability condition will be met for all samples in the batch of representative points $\mathcal{B}$ given in* (6.37)*. Suppose the batch $\mathcal{B} = \{x_i\}_{i=1}^N$ contains a finite set of points in $\mathcal{X}$ such that, $\forall x \in \mathcal{X}, \exists x_i \in \mathcal{B}, ||x - x_i|| < h$, for some $h > 0$. Under the conditions of Theorem 6.6, for $||\frac{\partial \hat{G}(x,\theta)}{\partial x}|| \leq M_G$, if ϵ_{pd} in* (6.37) *is set to $\epsilon_{\text{pd}} = 2\bar{G}||M||M_G h$, then the stability condition is met on the entire state space $\mathcal{X}$ for the choice of $\bar{\epsilon} - \epsilon_{\text{pd}} \geq \epsilon_i \geq 2\bar{G}\|M\|(1 + L_u)\varepsilon_J + \|M\|(1 + L_u)^2\varepsilon_J^2$ in* (6.27c)*.*

Remark 6.3. *The batch $\mathcal{B}$ constitutes arbitrary points in $\mathcal{X}$ to estimate the supremum of the stability constraint and does not correspond to data collected from the system. The primal-dual algorithm only requires the evaluation of $\lambda_{\max}(K(x_i, \theta))$ at these particularly chosen representative points in $\mathcal{X}$ using the estimated dynamics $\hat{F}(\cdot)$. Here h is the fill distance for the batch $\mathcal{B}$. This condition can be met by using $N = (\Gamma/h+1)^d$ samples in the batch $\mathcal{B}$. Note that this dependency is unavoidable to formally verify stability for the entire $\mathcal{X}$ using samples [97]. In practice, one can use falsifiers [93] to find states which violate the stability constraint and add them in $\mathcal{B}$, similar to [49]. Furthermore, in the expense of computational burden, N can be also picked larger which would reduce h and shrink ϵ_{pd} arbitrarily.*

Proof. Due to the dual variable update in (6.37), the convergence of (μ, θ) to (μ^*, θ^*) implies $\forall i \in \mathcal{B}, \lambda_{\max}(K(x_i, \theta^*))) \leq 0$, that is the Krasovskii's stability condition holds for all the samples in the batch. Since each batch $\mathcal{B}$ is drawn from the state space $\mathcal{X}$, as training time goes to infinity, the stability condition $\lambda_{\max}(K(x_i, \theta^*))) \leq 0$ also holds for all $x_i \in \mathcal{B}$. By the fill distance condition, i.e., $\forall x \in \mathcal{X}, \exists x_i \in \mathcal{B}, ||x - x_i|| < h$, (here we drop the dependence on θ^* as it is fixed in the proof)

$$\begin{aligned}
&\min_{x_i \in \mathcal{B}} ||K(x) - K(x_i)|| = \min_{x_i \in \mathcal{B}} ||\hat{G}(x)^\top M \hat{G}(x) - \hat{G}(x_i)^\top M \hat{G}(x_i)|| \\
&\leq \min_{x_i \in \mathcal{B}} ||(\hat{G}(x) - \hat{G}(x_i))^\top M \hat{G}(x)|| + ||\hat{G}(x_i)^\top M (\hat{G}(x) - \hat{G}(x_i))|| \\
&\leq 2\bar{G}||M||M_G h. \qquad (6.38)
\end{aligned}$$

Let $\epsilon_{pd} = 2\bar{G}||M||M_G h$, if $\forall x_i \in \mathcal{B}, K(x_i, \theta^*) + \epsilon_{pd} I \prec -\epsilon_i I$, then $K(x, \theta^*) \prec -\epsilon_i I$ holds for all x in the entire state space $\mathcal{X}$. By Theorem 6.6, stability holds for the true system, i.e., $G(x)^\top M G(x) - M \prec 0$ for all $x \in \mathcal{X}$. □

Remark 6.4. *The convergence of the primal-dual algorithm has been shown for linear systems in risk-constrained control [302, 303] and Q-learning [175]. Similar*

Algorithm 17 KCRL

Input: τ, g_{θ_0}, D, λ, ϵ_i, M, μ, η_1, η_2, ϵ_{pd}
for $i = 0, \ldots$ **do**
  for $t = i\tau, \ldots, (i+1)\tau$ **do**
    Execute $u_t = g_{\theta_i}(x_t)$
    Store $\phi_t = [x_t^\top, u_t^\top]^\top$ and x_{t+1}
  Estimate the model dynamics $\hat{F}_i(\cdot)$ ▷ **Learning**
  Solve (6.27) for θ_{i+1} using $\hat{F}_i(\cdot)$ via (6.37)
  Construct $g_{\theta_{i+1}}$ ▷ **Stable Policy Design**

convergence proofs translate to our setting for linear systems. Showing convergence for general nonlinear dynamics requires a piece of new machinery and is beyond the scope of this work.

Theorem 6.7 shows that solving (6.27) via primal-dual approach will recover a stabilizing solution within the given parameter space Θ, provided that the estimation error is sufficiently small. The parameter space Θ and its coupling with the underlying system in closed-loop form determine the level of estimation error required in the system dynamics through L_u, G, and M_G. For instance, in the LQR problem restricted to the linear state-feedback policy class Θ, the estimation error of the model parameters necessary for stable policy design depends on the maximum operator norms of the feedback controllers and their corresponding closed-loop matrices in the given parameter set Θ, while $M_G = 0$. In other words, the feasible set of $\theta \in \Theta$ for a fixed estimation error of the linear model parameters is determined by L_u and G, which can be upper-bounded using the continuous differentiability of F and g_θ. In the following, we design a model-based RL framework using the discussed primal-dual approach and show that for smooth dynamical systems it can be used for learning stabilizing controllers from scratch.

6.4.3 Krasovskii-Constrained RL Framework

In this section, we present the novel model-based RL framework: KCRL. The algorithm is outlined in Algorithm 17. KCRL works in epochs of length τ, where the controller is during the epoch is fixed. Each epoch consists of two parts: (i) *Model Learning*, where KCRL deploys the current controller in the underlying system to generate trajectories and update the model estimates, (ii) *Stable Policy Design*, where KCRL uses Krasovskii-constrained policy optimization approach to design the new controller for the next epoch.

Each epoch i of KCRL starts with a data collection from the underlying system for τ time-steps with the current controller, $g_{\theta_i}(\cdot)$. In each time step, KCRL takes the action $u_t = g_{\theta_i}(x_t)$, and stores the current state-action pair $\phi_t = [x_t^\top, u_t^\top]^\top$ and the observed next state x_{t+1}. Note that τ is a user-defined parameter and $g_{\theta_0}(\cdot)$ is the initial policy.

At the end of each epoch, KCRL uses all the data gathered to estimate a model of underlying system dynamics $\hat{F}_i(\cdot)$. This estimate can be obtained in various ways within a general supervised learning framework, e.g., through neural networks or system-dependent feature representations. Using neural networks, one can run a variant of gradient descent to update the model estimates. On the other hand, for system-dependent feature representations, one can consider the best linear approximation of the system dynamics on a nonlinear basis such as Random Fourier Features [226], wavelets, or more generally using an atomic norm minimization framework [57]. Once KCRL has a model estimate after the data collection, it aims to recover a stabilizing policy via solving (6.27) using (6.37) to obtain the controller for the next epoch.

6.4.4 KCRL with Random Fourier Features (RFF)

In this section, we theoretically analyze a variant of KCRL that uses RFF to learn the system dynamics. In particular, we give a sample complexity result to learn a stabilizing controller for the underlying system. To obtain such a result, we assume that the unknown nonlinear system F satisfies Assumption 6.3.

Remark 6.5. *Note that Gaussian kernels, which satisfy Assumption 6.3, are universal kernels such that they can approximate an arbitrary continuous target function uniformly on any compact subset of the input space using possibly infinite kernel evaluations [198]. Therefore, the class of nonlinear dynamics considered in this work constitutes a vast variety of nonlinear systems.*

For the details of RFF learning please refer to Section 6.1.4. From Theorem 6.1, we have

$$\sup_{\|\phi\| \leq \Gamma_\phi} \|\bar{F}(\phi) - F(\phi)\| \leq \tilde{O}(1/\sqrt{D}), \tag{6.39}$$

for the best D-dimensional RFF approximation of F, $\bar{F}(\cdot) = W_*^\top z(\cdot)$, where Γ_ϕ describes the bounded region. Here $\tilde{O}(\cdot)$ denotes the order up to logarithmic factors and hides the dependencies on n, Γ_ϕ, and the fill distance. W_* is the unique min-max optimal model and unique for the particular selection of RFF basis, i.e., a realization

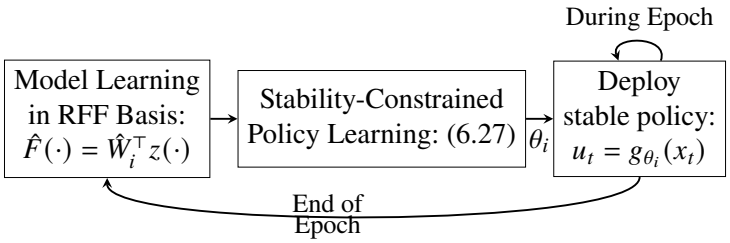

Figure 6.7: KCRL Framework with RFF Learning.

obtained via (6.6). This result is key to our analysis as we use it to derive the finite-time learning and stabilization guarantees of KCRL.

Using the best D-dimensional RFF approximation of F defined in (6.39), we can approximate (6.23) as $x_{t+1} \approx W_*^\top z(\phi_t)$, for some unknown $W_* \in \mathbb{R}^{D\times n}$. For model learning, KCRL considers this approximate model and tries to recover the best estimate for W_* using all the data gathered. In particular, after the data collection of epoch i, KCRL solves the following:

$$\min_W \lambda\|W\|_F^2 + \sum_{s=0}^{t=(i+1)\tau} \|x_{s+1} - W^\top z(\phi_s)\|_2^2, \tag{6.40}$$

for some $\lambda > 0$ to obtain an estimate of W_*. Note that $\hat{W}_i = (Z_t Z_t^\top + \lambda I)^{-1} Z_t X_t^\top$ gives the closed-form solution of (6.40) for $X_t = [x_{t+1}, \dots, x_1] \in \mathbb{R}^{n\times(t+1)}$, $Z_t = [z(\phi_t), \dots, z(\phi_0)] \in \mathbb{R}^{D\times(t+1)}$. Thus, at epoch i, the learned model by KCRL is given by $\hat{F}_i(\cdot) = \hat{W}_i^\top z(\cdot)$. Before we proceed, we have the following assumption on the initial policy of KCRL.

Remark 6.6. *We consider fully observable (state-feedback) nonlinear systems in this section. However, the results can be extended to partially observable nonlinear dynamical systems as in Section 6.2. More specifically, the results can be rewritten for an order-h nonlinear autoregressive exogenous system as depicted in previous sections, i.e., considering the last h input-output pairs as the state x_t.*

Assumption 6.10 (Exploratory and Bounded Initial Policy)**.** *The initial controller g_{θ_0} provides persistently exciting (PE) and bounded inputs that can be used for exploration and excite the system uniformly. In other words, the smallest eigenvalue of the design matrix $Z_t Z_t^\top$ scales linearly over time, and for $x_{t+1} = F_{\theta_0}(x_t)$, we have $\|\phi_t\| \le \Gamma_\phi$, for some finite Γ_ϕ.*

These assumptions are standard for consistent estimation of the model dynamics in statistical learning, Assumption 6.5. To achieve such initial controllers, recent tools

in control could be deployed [177, 217]. In practice, Dataset Aggregation methods could be used with policy $g_{\theta_0}(\cdot)$ for safe excitement of the systems coupled with randomized feedback policies.

Next, we focus on the learning guarantees of KCRL. We need to guarantee that the model estimation errors are small enough at the end of first epoch such that the controller obtained via solving (6.27) would stabilize the system. Using the result in (6.15), for large enough D, we get

$$\sup_{\|\phi\| \leq \Gamma_\phi} \|F(\phi) - \hat{F}_1(\phi)\| = \tilde{O}(1/\sqrt{D} + \sqrt{D/\tau}),$$

after τ time-steps, i.e., at the end of first epoch of KCRL. From (6.15), we derive the following novel finite sample approximation error guarantee on the Jacobian of the underlying function $F(\cdot)$ via the finite difference method. This result could be of independent interest in RFF learning and linearization of RFF-learned model dynamics for the study of different stability notions such as contraction theory or CLFs.

Proposition 6.3 (Approximation Error of Jacobian using RFF)**.** *Let J_F denote the Jacobian of the underlying system F given in* (6.23)*. Consider the finite difference approximation of J_F using $\hat{F}_1(\cdot) = \hat{W}_1^\top z(\cdot)$, such that*

$$\hat{J}_F^{(i,j)}(\phi) = \frac{\hat{F}_{1,i}(\phi + \varepsilon\, \mathbf{e_j}) - \hat{F}_{1,i}(\phi - \varepsilon\, \mathbf{e_j})}{2\varepsilon}, \tag{6.41}$$

where $\varepsilon > 0$, $\hat{F}_{1,i}(\cdot)$ is the mapping from input to the ith index of the output of $\hat{F}_1$ and $\mathbf{e_j}$ is the jth standard basis. Under Assumptions 6.9 & 6.10, for the choice of $\varepsilon = \tilde{O}\big((D^{-1/2} + \sqrt{D/\tau})^{1/3}\big)$, we have that $\sup_{\|\phi\| \leq B} \|\hat{J}_F(\phi) - J_F(\phi)\|_F = \tilde{O}(\varepsilon^2)$.

Proof. Consider the Taylor expansions of $F_i(\phi + \varepsilon\, \mathbf{e_j})$ and $F_i(\phi - \varepsilon\, \mathbf{e_j})$ at ϕ. From Assumption 6.9, we have

$$\frac{F_i(\phi + \varepsilon\, \mathbf{e_j}) - F_i(\phi - \varepsilon\, \mathbf{e_j})}{2\varepsilon} = \frac{\partial F_i}{\partial \phi_j} + \mathbf{F_H} O(\varepsilon^2). \tag{6.42}$$

Now consider (6.41). Let $\delta_\tau = \tilde{O}(1/\sqrt{D} + \sqrt{D/\tau})$. From (6.15), we have $\hat{F}_{1,i}(\phi + \varepsilon\, \mathbf{e_j}) = F_i(\phi + \varepsilon\, \mathbf{e_j}) + \epsilon_1$ and $\hat{F}_{1,i}(\phi - \varepsilon\, \mathbf{e_j}) = F_i(\phi - \varepsilon\, \mathbf{e_j}) + \epsilon_2$ for $0 \leq \epsilon_1, \epsilon_2 \leq \delta_\tau$. Combining this with (6.42), we obtain

$$\hat{J}_F^{(i,j)}(\phi) - \frac{\partial F_i}{\partial \phi_j} = \mathbf{F_H} O(\varepsilon^2) + \frac{\epsilon_1 - \epsilon_2}{2\varepsilon}.$$

This gives us that $|\hat{J}_F^{(i,j)}(\phi) - \frac{\partial F_i}{\partial \phi_j}| \leq \mathbf{F_H} O(\varepsilon^2) + \delta_\tau/\varepsilon$ for all i, j. Combining these yields $\|\hat{J}_F(\phi) - J_F(\phi)\|_F \leq n \left(c \mathbf{F_H} \varepsilon^2 + \frac{\delta_\tau}{\varepsilon}\right)$, for some constant c. Note that the optimal choice of ε is $\varepsilon = \tilde{O}\left((1/\sqrt{D} + \sqrt{D/\tau})^{1/3}\right)$, which finishes the proof. □

This result shows that the Jacobian of a vector-valued function in a known RKHS is well-approximated using the RFF representation of the function with finite samples. We finally provide the finite-sample stabilization guarantee of KCRL.

Theorem 6.8 (Finite Sample Stabilization via KCRL)**.** *Suppose Assumptions 6.9-6.10 hold and the batch $\mathcal{B}$ is informative enough that its fill distance h satisfies $\bar{\epsilon} - \epsilon_{pd} > 0$, for $\epsilon_{pd} = 2\bar{G}||M||M_G h$. Set $\epsilon_i = \bar{\epsilon} - \epsilon_{pd}$ in the constraint* (6.27c)*. If KCRL uses $D = \tilde{O}\Big(\Big(\frac{2\bar{G}\|M\|(1+L_u)+\|M\|(1+L_u)^2}{\bar{\epsilon}-\epsilon_{pd}}\Big)^3\Big)$ number of RFF in learning the system, after D^2 samples (time-steps), we have the trajectory of $x_{t+1} = f(x_t, g_\theta(x_t))$ is asymptotically stable around the origin, i.e., the solution of* (6.27) *after $\tau = D^2$ samples from the system gives a stabilizing controller g_{θ_1} for the unknown nonlinear dynamical system.*

Proof. Recall that the stability condition holds for the underlying system with $\bar{\epsilon}$ margin. Thus, combining Theorem 6.6 and Theorem 6.7, to guarantee the stabilization of the underlying system for the entire state-space, we require $\epsilon_i \leq \bar{\epsilon} - \epsilon_{pd}$, i.e.,

$$\frac{\bar{\epsilon} - \epsilon_{pd}}{2\bar{G}\|M\|(1 + L_u) + \|M\|(1 + L_u)^2} \geq \varepsilon_J, \tag{6.43}$$

since $\varepsilon_J < 1$. This gives an upper bound on the error of Jacobian estimates to guarantee stabilization. From Proposition 6.3, we also have that $\varepsilon_J = \tilde{O}\big((1/\sqrt{D} + \sqrt{D/\tau})^{2/3}\big)$, since $\widehat{\frac{\partial F(\phi)}{\partial x}} - \frac{\partial F(\phi)}{\partial x}$ and $\widehat{\frac{\partial F(\phi)}{\partial u}} - \frac{\partial F(\phi)}{\partial u}$ are submatrices of $\hat{J}_F(\phi) - J_F(\phi)$. The optimal choice of τ and D that minimizes this upper bound is $\tau = D^2$, which results that $\varepsilon_J = \tilde{O}\left(D^{-1/3}\right)$ after τ samples. Thus, for the stated choice of D, after $\tau = D^2$ time-steps, KCRL is guaranteed to stabilize the underlying system. □

This result shows that by setting the epoch length $\tau = D^2$, KCRL guarantees the recovery of a stabilizing controller at the end of first epoch, i.e. g_{θ_1}. The choice of ϵ_i also guarantees the recovery of stabilizing controllers for the subsequent epochs with the non-increasing estimation errors.

6.4.5 Case Study

We numerically study KCRL in learning stable policies for voltage control in a power distribution system [240]. Our case study focuses on the South California Edison 56-bus test feeder with high penetration of photovoltaic (PV) generations. The detailed system parameters follow the configuration in [240]. The system model

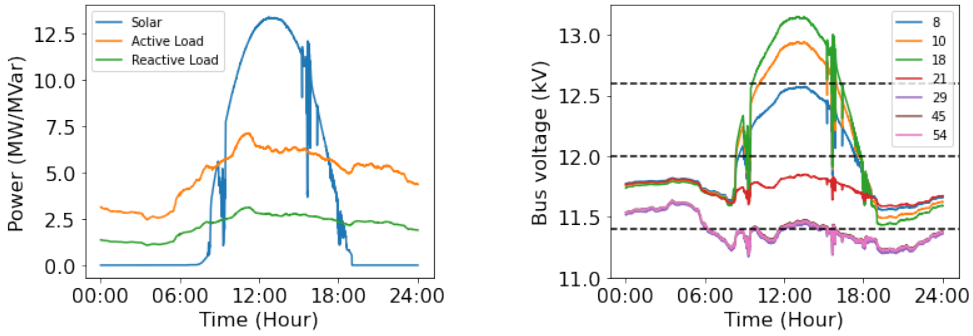

Figure 6.8: **(Left)** Real-world solar and load data across 24 hours with 6 seconds resolution; **(Right)** Serious voltage violations in the system without control.

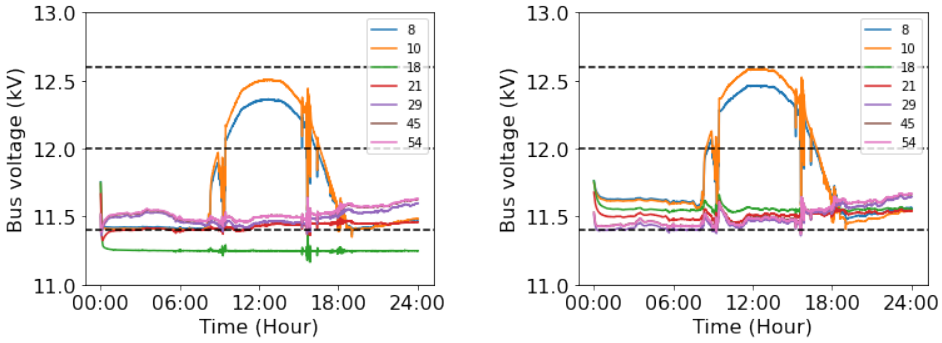

Figure 6.9: **(Left)** Standard DDPG [185] causes voltage violations in some nodes (e.g., node 18); **(Right)** KCRL can stabilize the system voltage within the nominal operation region (between the two dashed lines) under all conditions.

for the voltage control of this system is given by,

$$-p_j(t) = P_{ij}(t) - r_{ij}l_{ij}(t) - \sum_{k:(j,k)\in E} P_{jk}(t), \forall j, \tag{6.44a}$$

$$-q_j(t) = Q_{ij}(t) - x_{ij}l_{ij}(t) - \sum_{k:(j,k)\in E} Q_{jk}(t), \forall j, \tag{6.44b}$$

$$v_j(t) = v_i(t) - 2(r_{ij}P_{ij}(t) + x_{ij}Q_{ij}(t)) + (r_{ij}^2 + x_{ij}^2)l_{ij}(t), \forall (i,j) \in E, \tag{6.44c}$$

Here (6.44a) and (6.44b) represent the power conservation at node j, p_j denotes the real power injection at node j and q_j denotes the reactive power injection. (6.44c) represents the voltage drop from node i to node j. $l_{ij}(t) := |I_{ij}|^2 = (P_{ij}^2 + Q_{ij}^2)/v_i$ is the squared current, $v_i := |V_i|^2$ is the squared voltage, $P_{ij}(t)$ and $Q_{ij}(t)$ represent active and reactive power flow on line (i,j), respectively.

Consider the controller form $q(t+1) = q(t) + g_\theta(v(t))$, where $g_\theta(v(t))$ is represented as a neural network, that can be trained either by the proposed KCRL framework, or standard RL framework. We adopt DDPG [185], a commonly used RL algorithm for

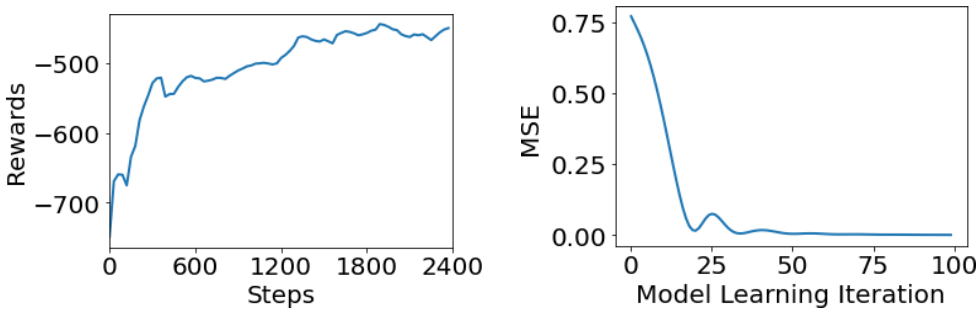

Figure 6.10: Model Performance vs. Iterations.

continuous control as a baseline. [240] shows that a Krasovskii's Lyapunov function exists for the voltage control system (6.44), where $M = X^{-1}$ with X representing the network reactance matrix. The desired stable set for the system is denoted as $\mathcal{S}_v = \{\mathbf{v} \in \mathbb{R}^n : \underline{v}_i \leq v_i \leq \overline{v}_i\}$, where $\underline{v}_i, \overline{v}_i$ are the lower and upper bound for the nominal voltage range. For the considered system, $\overline{v}_i = 12.6\text{kV}$ $\underline{v}_i = 11.4\text{kV}$ $\forall i$, that are plotted as the two dashed lines in Figure 6.8.

We simulate the performance of KCRL and DDPG using a real-world voltage control dataset and the results are presented in Figure 6.9. We also plot the training curve of the KCRL algorithm and the model learning error for the first 100 iterations in Figure 6.10. We observe that the model error keeps reducing and the policy performance keeps improving (measured by a lower cost $\sum_{t=0}^{T} \gamma^t c(x_t, u_t)$) per training iteration.

6.4.6 Conclusion and Future Work

In this chapter, we adapt the classical Krasovskii's Lyapunov function into policy optimization in RL. Using this method, we design stabilizing policies under modeling error with precise robustness guarantees. Furthermore, we propose a model-based RL framework, KCRL, that is guaranteed to design stabilizing controllers in online control of unknown nonlinear dynamical systems using finite samples. In future work, we aim to broaden the KCRL framework to incorporate other Lyapunov function constructions or learn and update the Lyapunov functions on the fly. For example, in closely related contraction analysis, the matrix M can be state-dependent. Extending our current results to such construction would allow KCRL to be deployed for different tasks.

Chapter 7

CONCLUDING REMARKS

In this thesis, we studied the learning and control of unknown dynamical systems ranging from linear time-invariant systems to time-varying or partially observable linear systems, as well as the most general setting of partially observable nonlinear dynamical systems. Blending recent advances in statistical learning theory and control theoretic tools, we provided an array of finite-sample guarantees for the accuracy in learning the unknown dynamics, regret performance in adaptive learning and control, and stabilization of the underlying dynamics. We proposed statistically and computationally efficient learning and control algorithms and demonstrated their strong empirical performance, robustness, and stability guarantees in various adaptive control tasks in real-world control systems.

In Chapter 2, we studied sample complexity and regret of learning and control in stochastic linear bandits with underlying low-rank structure and unknown safety constraints. In Chapter 3, we studied learning and control in linear time-invariant systems and proposed two efficient algorithms that attain optimal regret guarantees while achieving fast stabilization of the underlying dynamics along the way. Then, in Chapter 4, we studied the stability concepts of input-to-state stability and mean-square stability in classes of linear time-varying systems and showed finite-time learning and stabilization in these settings. In Chapter 5.1.2, we proposed the first closed-loop system identification method with optimal finite-time learning guarantees in partially observable linear dynamical systems. Combining this method with various adaptive control design strategies, we presented an array of efficient learning and control algorithms in this setting with optimal regret guarantees. Finally, in Chapter 6, we studied the learning and control of nonlinear dynamical systems. We proposed efficient learning methods for two classes of nonlinear systems and provided finite-sample guarantees. We designed effective learning and control algorithms by combining these methods with efficient model-predictive control algorithms or designed new finite-time stabilization methods using Krasovskii's Lyapunov function construction. We demonstrated the strong empirical performance of these methods in various real-world adaptive control and stabilization tasks. In particular, we deployed FALCON, on a wing setup under extreme turbulent flow dynamics in the Caltech wind tunnel and showed consistent and data-efficient

state-of-the-art disturbance rejection performance.

In addition to the results presented in this thesis, we believe that there are many interesting research problems for future studies on these dynamical systems.

7.1 Future Directions in Stochastic Linear Bandits

In our results, we primarily focused on worst-case guarantees in learning the underlying structure or unknown safety constraints. There are interesting future directions that could potentially improve the applicability of our work. One such direction is to consider data-dependent bounds, which may be more suitable for specific applications despite potentially resulting in worse statistical complexities. Another area for future research is multiplayer stochastic linear bandits with unknown safety constraints. Incorporating multiple players in our proposed unknown safety constraints with local feedback framework could further enhance real-world modeling capabilities, especially in scenarios such as economics and autonomous driving where a player's decision impacts the reward and safety of others. Analyzing regret and safety guarantees in this challenging setting would provide new insights into decision-making under uncertainty.

7.2 Future Directions in Linear Time-Invariant Systems

In Chapter 3, our focus was mainly on attaining $\tilde{O}(\sqrt{T})$ regret that scales with polynomial dimension dependencies. However, the exact polynomial dependencies and the effect of other control theoretic quantities are not further investigated. Therefore, one interesting future direction is to examine these dependencies more closely. Additionally, data-dependent bounds could be also desirable in learning and control in LQRs, similar to the bandit setting. Furthermore, our algorithms in Chapter 3 only consider unconstrained control, whereas many real-world systems have safety or control constraints. A possible future direction is to extend our algorithms to the control under safety or control constraints setting and study the effect of these constraints on regret.

Another important algorithmic future direction is to investigate the possibility of combining optimism and Thompson Sampling to achieve improved exploration. The current implementation of Thompson Sampling only incorporates uncertainties in the underlying model parameters in the sampling procedure to achieve effective exploration. However, in learning and control of dynamical systems, the ultimate goal is to achieve desirable control performance. Ideally, the exploration should aim to achieve uncertainty reduction in the low-cost achieving models to quickly

discover the optimal policy. This idea is the cornerstone of optimistic control design. Incorporating cost-based exploration to the sampling procedure of Thompson Sampling would intuitively lead to improved performance. Exploring this direction would be valuable in developing more effective learning and control algorithms for dynamical systems.

7.3 Future Directions in Linear Time-Varying Systems

In Section 4.2, we proposed COCO-LQ, a novel stabilizing policy design approach for linear time-varying systems with modeling mismatches. While our proposed approach minimizes the cost and stabilizes the system, an important research direction is to analyze its policy regret. By investigating the trade-off between the stabilization parameter of COCO-LQ and the cost it achieves compared to the best policy in hindsight, we can better understand the performance of the algorithm.

Another important research direction is to extend COCO-LQ to the learning and control of nonlinear dynamical systems. One of the conventional approaches in nonlinear control is to linearize the dynamics at the current state. By efficiently learning the underlying nonlinear dynamics via the methods proposed in Chapter 6, we can provide error bounds on the linearization error of the learned dynamics. Given the stability under the model mismatch guarantee of COCO-LQ, we can design stabilizing policies for nonlinear dynamical systems with finite-sample guarantees. While analyzing the evolution of the dynamics based on this control design requires careful analysis and developing new tools, conceptually this approach provides an efficient approach to designing policies that potentially provide stability in nonlinear systems.

In Section 4.3, we empirically studied the effect of randomization and asynchrony on the stability of random asynchronous LTI systems. Our investigation showed that by using randomization, we can stabilize otherwise unstable LTI systems. One exciting future direction is to study the problem of finite-time stabilization of LTI systems using randomization. This would introduce a new dimension of "control input" to learning and control of dynamical systems, which would unlock a plethora of new research problems. One motivating example of using randomization and asynchrony in stabilization is in biological neural networks in the brain. Recent studies show the existence of a delicate equilibrium between synchrony and asynchrony of neural firings in many cognitive tasks, where any disturbance to this natural equilibrium may result in neurological disorders [275]. Further understanding the effect of randomization and asynchrony in the stability of random asynchronous LTI systems may provide

new stabilization tools to deploy in complex dynamical systems such as the brain.

7.4 Future Directions in Partially Observable Linear Dynamical Systems

In Section 5.7, we provide several important future directions in detail. To summarize, investigating the role of PE conditions in partially observable linear dynamical systems is one of the research directions to consider. As shown in the LQR setting, the PE condition is not required for optimal regret. It remains an open problem whether this is possible in learning and control of LQG control systems. Another interesting problem to study in these dynamical systems is to design controllers directly from the Markov parameter estimates without relying on subspace identification algorithms such as SysId. These methods would be more robust and would not require controllability and observability of the underlying system, which is often restrictive in partially observable linear dynamical systems. Moreover, investigating the role of open-loop stability in the learning and control of partially observable systems is also an important future research topic. In addition to these, investigating the exact dimension dependencies and the effect of other control theoretic quantities in the regret are fundamental research topics that are needed to be considered in the future.

7.5 Future Directions in Nonlinear Dynamical Systems

Due to the notorious challenges that nonlinear dynamical systems bring, our understanding of learning and control of nonlinear dynamical systems is still limited. There are countless future directions worth exploring in this realm. In the following, we name a few of these directions. One fundamental problem is studying the persistence of excitation condition in nonlinear systems. In practice, deploying sufficiently random control inputs are the conventional ways to achieve this. However, the learning algorithms typically adopt various nonlinear basis functions in learning the underlying dynamics, and understanding how to achieve persistence of excitation, which is central in consistent learning, remains an important open problem.

Moreover, novel machine learning tools such as meta-learning and hierarchical learning are adopted in practice for learning and control of complex nonlinear systems [215]. These methods would allow learning different closed-loop systems obtained via different control policies on the same nonlinear dynamical system. Understanding the statistical complexity of these methods and improving their efficiency in learning remains an important challenge. In a similar vein, analyzing the closure models, where the agent learns the residual between the underlying dynamics and a low-fidelity physical model, is crucial for providing a statistical

understanding of practical modeling approaches.

Due to the difficulty of control design in nonlinear dynamical systems, the main focus of our study was on learning the dynamics and providing subsequent stability guarantees or assuming a stability margin to provide regret guarantees. To the best of our knowledge, there has been no end-to-end study of learning and control in nonlinear systems, as in linear systems. Studying the suboptimality guarantees in model predictive control or other explicit control strategies such as feedback linearization or backstepping control are important immediate research directions. These methods, as showcased within FALCON, achieve strong empirical success, yet, their statistical analysis is limited.

In the realm of finite-time stabilization, one important direction is generalizing Krasovskii's Lyapunov function construction to more diverse systems. Adopting a nonlinearity around this construction and providing learning guarantees with stability verification tools is an immediate research problem to be considered. The goal of this study would be to consider a more general class of Lyapunov functions, as the dynamical systems that satisfy Krasovskii's Lyapunov function are limited.

Bibliography

[1] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.

[2] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.

[3] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[4] Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pages 1–9, 2012.

[5] Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR, 2017.

[6] Marc Abeille and Alessandro Lazaric. Thompson sampling for linear-quadratic control problems. *arXiv preprint arXiv:1703.08972*, 2017.

[7] Marc Abeille and Alessandro Lazaric. Improved regret bounds for thompson sampling in linear quadratic control problems. In *International Conference on Machine Learning*, pages 1–9, 2018.

[8] Marc Abeille and Alessandro Lazaric. Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation. *arXiv preprint arXiv:2007.06482*, 2020.

[9] Shipra Agrawal. Recent advances in multiarmed bandits for sequential decision making. *Operations Research & Management Science in the Age of Analytics*, pages 167–188, 2019.

[10] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

[11] Naum Il'ich Akhiezer and Izrail' Markovich Glazman. *Theory of linear operators in Hilbert space*. Courier Corporation, 2013.

[12] Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] Sanae Amani, Christos Thrampoulidis, and Lin Yang. Safe reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 243–253. PMLR, 2021.

[14] Francesco Amato, Marco Ariola, and Carlo Cosentino. Finite-time control of discrete-time linear systems: analysis and design conditions. *Automatica*, 46(5):919–924, 2010.

[15] Aaron D Ames, Kevin Galloway, and Jessy W Grizzle. Control lyapunov functions and hybrid zero dynamics. In *IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 6837–6842, 2012.

[16] Aaron D Ames, Kevin Galloway, Koushil Sreenath, and Jessy W Grizzle. Rapidly exponentially stabilizing control lyapunov functions and hybrid zero dynamics. *IEEE Transactions on Automatic Control*, 59(4):876–891, 2014.

[17] Oren Anava, Elad Hazan, and Shie Mannor. Online learning for adversaries with memory: price of past mistakes. In *Advances in Neural Information Processing Systems*, pages 784–792, 2015.

[18] Brian DO Anderson and C Richard Johnson Jr. Exponential convergence of adaptive identification and control algorithms. *Automatica*, 18(1):1–13, 1982.

[19] Karl J Åström. *Introduction to stochastic control theory*. Courier Corporation, 2012.

[20] Anil Aswani, Humberto Gonzalez, S. Shankar Sastry, and Claire Tomlin. Provably safe and robust learning-based model predictive control. *Automatica*, 49(5):1216–1226, 2013. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2013.02.003. URL https://www.sciencedirect.com/science/article/pii/S0005109813000678.

[21] Haim Avron, Alex Druinsky, and Anshul Gupta. Revisiting asynchronous linear solvers: Provable convergence rate through randomization. *J. ACM*, 62(6):51:1–51:27, Dec. 2015. ISSN 0004-5411.

[22] Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement learning of pomdps using spectral methods. *arXiv preprint arXiv:1602.07764*, 2016.

[23] Peder Bacher, Henrik Madsen, and Henrik Aalborg Nielsen. Online short-term solar power forecasting. *Solar energy*, 83(10):1772–1783, 2009.

[24] Peter L Bartlett, Victor Gabillon, and Michal Valko. A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption. In *Algorithmic Learning Theory*, pages 184–206. PMLR, 2019.

[25] Peter W Bearman. On vortex shedding from a circular cylinder in the critical reynolds number regime. *Journal of Fluid Mechanics*, 37(3):577–585, 1969.

[26] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[27] Julian Berberich, Carsten W. Scherer, and Frank Allgöwer. Combining prior knowledge and data for robust controller design. *IEEE Transactions on Automatic Control*, pages 1–16, 2022. doi: 10.1109/TAC.2022.3209342.

[28] Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 2. Athena scientific Belmont, MA, 1995.

[29] Katharina Bieker, Sebastian Peitz, Steve L. Brunton, J. Nathan Kutz, and Michael Dellnitz. Deep model predictive flow control with limited sensor data and online learning. *Theoretical Computational Fluid Dynamics*, S.I. (34):577–591, 2020.

[30] Sergio Bittanti and Marco C Campi. Adaptive control of linear time invariant systems: the "bet on the best" principle. *Communications in Information & Systems*, 6(4):299–320, 2006.

[31] Horst Bleckmann, Joachim Mogdans, and Sheryl L Coombs. Flow sensing in air and water. *Berlin, Germany*, 976, 2014.

[32] F Boettcher, CH Renner, H-P Waldl, and J Peinke. On the statistics of wind gusts. *Boundary-Layer Meteorology*, 108(1):163–173, 2003.

[33] Nicholas Boffi, Stephen Tu, Nikolai Matni, Jean-Jacques Slotine, and Vikas Sindhwani. Learning stability certificates from data. In *Conference on Robot Learning*, pages 1341–1350. PMLR, 2021.

[34] Nicholas M Boffi, Stephen Tu, and Jean-Jacques E Slotine. Regret bounds for adaptive nonlinear control. In *Learning for Dynamics and Control*, pages 471–483. PMLR, 2021.

[35] Zdravko I Botev, Dirk P Kroese, Reuven Y Rubinstein, and Pierre L'Ecuyer. The cross-entropy method for optimization. In *Handbook of statistics*, volume 31, pages 35–59. Elsevier, 2013.

[36] Stephen Boyd and Leon Chua. Fading memory and the problem of approximating nonlinear operators with volterra series. *IEEE Transactions on circuits and systems*, 32(11):1150–1161, 1985.

[37] Stephen Boyd and Shankar Sastry. On parameter convergence in adaptive control. *Systems & control letters*, 3(6):311–319, 1983.

[38] Stephen Boyd and Sosale Shankara Sastry. Necessary and sufficient conditions for parameter convergence in adaptive control. *Automatica*, 22(6): 629–639, 1986.

[39] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[40] Michael Brosch, Eike Budinger, and Henning Scheich. Stimulus-related gamma oscillations in primate auditory cortex. *Journal of neurophysiology*, 87(6):2715–2725, 2002.

[41] Steven L Brunton and Bernd R Noack. Closed-loop turbulence control: Progress and challenges. *Applied Mechanics Reviews*, 67(5), 2015.

[42] Dave P Burke, Simon P Kelly, Philip De Chazal, Richard B Reilly, and Ciarán Finucane. A parametric feature extraction and classification strategy for brain-computer interfacing. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(1):12–17, 2005.

[43] Marco C Campi and Erik Weyer. Finite sample properties of system identification methods. *IEEE Transactions on Automatic Control*, 47(8):1329–1334, 2002.

[44] Marco C Campi and Erik Weyer. Guaranteed non-asymptotic confidence regions in system identification. *Automatica*, 41(10):1751–1764, 2005.

[45] Claudio Canuto, M Yousuff Hussaini, Alfio Quarteroni, and Thomas A Zang. *Spectral methods: fundamentals in single domains*. Springer Science & Business Media, 2007.

[46] CE Carcangiu, A Pujana-Arrese, A Mendizabal, I Pineda, and J Landaluze. Wind gust detection and load mitigation using artificial neural networks assisted control. *Wind Energy*, 17(7):957–970, 2014.

[47] Alexandra Carpentier and Rémi Munos. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *Artificial Intelligence and Statistics*, pages 190–198, 2012.

[48] Asaf Cassel, Alon Cohen, and Tomer Koren. Logarithmic regret for learning linear quadratic regulators efficiently. *arXiv preprint arXiv:2002.08095*, 2020.

[49] Ya-Chien Chang, Nima Roohi, and Sicun Gao. Neural lyapunov control. *Advances in Neural Information Processing Systems*, 2019.

[50] Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018.

[51] Shaoru Chen, Mahyar Fazlyab, Manfred Morari, George J Pappas, and Victor M Preciado. Learning lyapunov functions for piecewise affine systems with neural network controllers. *arXiv preprint arXiv:2008.06546*, 2020.

[52] Shaoru Chen, Mahyar Fazlyab, Manfred Morari, George J Pappas, and Victor M Preciado. Learning lyapunov functions for hybrid systems. In *Proceedings of the 24th International Conference on Hybrid Systems: Computation and Control*, pages 1–11, 2021.

[53] Sheng Chen, Stephen A Billings, and Wan Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of control*, 50(5):1873–1896, 1989.

[54] Wen-Hua Chen, Donald J Ballance, and Peter J Gawthrop. Optimal control of nonlinear systems: a predictive control approach. *Automatica*, 39(4): 633–641, 2003.

[55] Xin Chen, Guannan Qu, Yujie Tang, Steven Low, and Na Li. Reinforcement learning for decision-making and control in power systems: Tutorial, review, and vision. *arXiv preprint arXiv:2102.01168*, 2021.

[56] Xinyi Chen and Elad Hazan. Black-box control for linear dynamical systems. 2020.

[57] Yuejie Chi and Maxime Ferreira Da Costa. Harnessing sparsity over the continuum: Atomic norm minimization for superresolution. *IEEE Signal Processing Magazine*, 37(2):39–57, 2020.

[58] Alessandro Chiuso and Giorgio Picci. Consistency analysis of some closed-loop subspace identification methods. *Automatica*, 41(3):377–391, 2005.

[59] Jason Choi, Fernando Castaneda, Claire J Tomlin, and Koushil Sreenath. Reinforcement learning for safety-critical control under model uncertainty, using control lyapunov functions and control barrier functions. 2020.

[60] William R Clements, Bastien Van Delft, Benoît-Marie Robaglia, Reda Bahi Slaoui, and Sébastien Toth. Estimating risk and uncertainty in deep reinforcement learning. *arXiv preprint arXiv:1905.09638*, 2019.

[61] Alon Cohen, Avinatan Hassidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. *arXiv preprint arXiv:1806.07104*, 2018.

[62] Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. *arXiv preprint arXiv:1902.06223*, 2019.

[63] Oswaldo Luiz Valle Costa, Ricardo Paulino Marques, and Marcelo Dutra Fragoso. *Discrete-Time Markov Jump Linear Systems*. Springer, 2005.

[64] Munther A Dahleh, Eduardo D Sontag, NC David, and John N Tsitsiklis. Worst-case identification of nonlinear fading memory systems. *Automatica*, 31(3):503–508, 1995.

[65] Bo Dai, Bo Xie, Niao He, Yingyu Liang, Anant Raj, Maria-Florina F Balcan, and Le Song. Scalable kernel methods via doubly stochastic gradients. *Advances in neural information processing systems*, 27, 2014.

[66] Hongkai Dai, Benoit Landry, Lujie Yang, Marco Pavone, and Russ Tedrake. Lyapunov-stable neural-network control, 2021. URL `https://arxiv.org/abs/2109.14152`.

[67] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

[68] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[69] Charles Dawson, Sicun Gao, and Chuchu Fan. Safe control with learned certificates: A survey of neural lyapunov, barrier, and contraction methods. *arXiv preprint arXiv:2202.11762*, 2022.

[70] C. De Persis, M. Rotulo, and P. Tesi. Learning controllers from data via approximate nonlinearity cancellation. *IEEE Transactions on Automatic Control*, pages 1–16, 2023. doi: 10.1109/TAC.2023.3234889.

[71] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197, 2018.

[72] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 1–47, 2019.

[73] John Elder and Sheryl Coombs. The influence of turbulence on the sensory basis of rheotaxis. *Journal of comparative physiology A*, 201(7):667–680, 2015.

[74] Ludwig Elsner, Rafael Bru, and Michael M Neumann. Convergence of infinite products of matrices and inner-outer iteration schemes. *Electronic Transactions on Numerical Analysis*, 2, 1994.

[75] Andreas K Engel, Andreas K Kreiter, Peter König, and Wolf Singer. Synchronization of oscillatory neuronal responses between striate and extrastriate visual cortical areas of the cat. *Proceedings of the National Academy of Sciences*, 88(14):6048–6052, 1991.

[76] Tom Erez, Kendall Lowrey, Yuval Tassa, Vikash Kumar, Svetoslav Kolev, and Emanuel Todorov. An integrated system for real-time model predictive control of humanoid robots. In *2013 13th IEEE-RAS International conference on humanoid robots (Humanoids)*, pages 292–299. IEEE, 2013.

[77] Fabio Fagnani and Sandro Zampieri. Randomized consensus algorithms over large scale networks. *IEEE Journal on Selected Areas in Communications*, 26(4):634–649, 2008.

[78] Paolo Falcone, Manuela Tufo, Francesco Borrelli, Jahan Asgari, and H Eric Tseng. A linear time varying model predictive control approach to the integrated vehicle dynamics control problem in autonomous systems. In *2007 46th IEEE Conference on Decision and Control*, pages 2980–2985. IEEE, 2007.

[79] Paolo Falcone, Francesco Borrelli, H Eric Tseng, Jahan Asgari, and Davor Hrovat. Linear time-varying model predictive control and its application to active steering systems: Stability analysis and experimental validation. *International Journal of Robust and Nonlinear Control: IFAC-Affiliated Journal*, 18(8):862–875, 2008.

[80] Dixia Fan, Liu Yang, Zhicheng Wang, Michael S. Triatafyllou, and George Em Karniadakis. Reinforcement learning for bluff body active flow control in experiments and simulations. *Proceedings of the National Academy of Sciences*, 117(42):26091–26098, 2020.

[81] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Optimism-based adaptive regulation of linear-quadratic systems. *arXiv preprint arXiv:1711.07230*, 2017.

[82] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time adaptive stabilization of lq systems. *arXiv preprint arXiv:1807.09120*, 2018.

[83] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Input perturbations for adaptive regulation and learning. *arXiv preprint arXiv:1811.04258*, 2018.

[84] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Randomized algorithms for data-driven stabilization of stochastic linear systems. In *2019 IEEE Data Science Workshop (DSW)*, pages 170–174. IEEE, 2019.

[85] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. On adaptive linear–quadratic regulators. *Automatica*, 117:108982, 2020.

[86] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Optimism-based adaptive regulation of linear-quadratic systems. *IEEE Transactions on Automatic Control*, 2020.

[87] Barry Fetics, Erez Nevo, Chen-Huan Chen, and David A Kass. Parametric model derivation of transfer function for noninvasive estimation of aortic pressure by radial tonometry. *IEEE Transactions on Biomedical Engineering*, 46(6):698–706, 1999.

[88] Claude-Nicolas Fiechter. Pac adaptive control of linear systems. In *Annual Workshop on Computational Learning Theory: Proceedings of the tenth annual conference on Computational learning theory*, volume 6, pages 72–80. Citeseer, 1997.

[89] Urban Forssell and Lennart Ljung. Closed-loop identification revisited. *Automatica*, 35(7):1215–1241, 1999.

[90] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[91] Bernard Friedland. *Control system design: an introduction to state-space methods*. Courier Corporation, 2012.

[92] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, pages 1587–1596. PMLR, 2018.

[93] Sicun Gao, Jeremy Avigad, and Edmund M Clarke. δ-complete decision procedures for satisfiability over the reals. In *International Joint Conference on Automated Reasoning*, pages 286–300. Springer, 2012.

[94] Germain Garcia, Sophie Tarbouriech, and Jacques Bernussou. Finite-time stabilization of linear time-varying continuous systems. *IEEE Transactions on Automatic Control*, 54(2):364–369, 2009.

[95] Paul Garnier, Jonathan Viquerat, Jean Rabault, Aurélien Larcher, Alexander Kuhnle, and Elie Hachem. A review on deep reinforcement learning for fluid mechanics. *Computers & Fluids*, 225:104973, 2021.

[96] Nikola Gavrilovic, Murat Bronz, Jean-Marc Moschetta, and Emmanuel Bénard. Bioinspired wind field estimation—part 1: Angle of attack measurements through surface pressure distribution. *International Journal of Micro Air Vehicles*, 10(3):273–284, 2018.

[97] Peter Giesl. Construction of a local and global lyapunov function for discrete dynamical systems using radial basis functions. *Journal of Approximation Theory*, 153(2):184–211, 2008.

[98] Paula Gradu, Elad Hazan, and Edgar Minasyan. Adaptive regret for control of time-varying dynamics. *arXiv preprint arXiv:2007.04393*, 2020.

[99] Peter Gunnarson, Ioannis Mandralis, Guido Novati, Petros Koumoutsakos, and John O. Dabiri. Learning efficient navigation in vortical flow fields. *arXiv prepint arXive: 2102.10536*, 2021.

[100] Sy-Ming Guu and Chin-Tzong Pang. On the convergence to zero of infinite products of interval matrices. *SIAM journal on matrix analysis and applications*, 25(3):739–751, 2003.

[101] Tuomas Haarnoja, Sehoon Ha, Aurick Zhou, Jie Tan, George Tucker, and Sergey Levine. Learning to walk via deep reinforcement learning. *arXiv preprint arXiv:1812.11103*, 2018.

[102] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abishek Gupta, Pieter Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications. *arXiiv preprint arXiv:1812.05905*, 2019.

[103] Jingqing Han. From pid to active disturbance rejection control. *IEEE transactions on Industrial Electronics*, 56(3):900–906, 2009.

[104] DJ Hartfiel. On infinite products of nonnegative matrices. *SIAM Journal on Applied Mathematics*, 26(2):297–301, 1974.

[105] Babak Hassibi, Ali H Sayed, and Thomas Kailath. *Indefinite-Quadratic Estimation and Control: A Unified Approach to H2 and H-infinity Theories*, volume 16. SIAM, 1999.

[106] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[107] Erich Hau. *Wind turbines: fundamentals, technologies, application, economics*. Springer Science & Business Media, 2013.

[108] Elad Hazan, Sham M Kakade, and Karan Singh. The nonstochastic control problem. *arXiv preprint arXiv:1911.12178*, 2019.

[109] Benjamin Herrmann, Steven L Brunton, Johannes E Pohl, and Richard Semaan. Gust mitigation through closed-loop control. ii. feedforward and feedback control. *Physical Review Fluids*, 7(2):024706, 2022.

[110] Joo P Hespanha, Payam Naghshtabrizi, and Yonggang Xu. A survey of recent results in networked control systems. *Proceedings of the IEEE*, 95(1): 138–162, 2007.

[111] Patricia Hidalgo-Gonzalez, Rodrigo Henriquez-Auba, Duncan S Callaway, and Claire J Tomlin. Frequency regulation using data-driven controllers in power grids with variable inertia due to renewable energy. In *2019 IEEE Power & Energy Society General Meeting (PESGM)*, pages 1–5. IEEE, 2019.

[112] BL Ho and Rudolf E Kálmán. Effective construction of linear state-variable models from input/output functions. *at-Automatisierungstechnik*, 14(1-12): 545–548, 1966.

[113] Gabriel Hoffmann, Haomiao Huang, Steven Waslander, and Claire Tomlin. Quadrotor helicopter flight dynamics and control: Theory and experiment. In *AIAA guidance, navigation and control conference and exhibit*, page 6461, 2007.

[114] Yiguang Hong, Zhong-Ping Jiang, and Gang Feng. Finite-time input-to-state stability and applications to finite-time control design. *SIAM Journal on Control and Optimization*, 48(7):4395–4418, 2010.

[115] Wei Hou, Darwin Darakananda, and Jeff D Eldredge. Machine-learning-based detection of aerodynamic disturbances using surface pressure measurements. *AIAA Journal*, 57(12):5079–5093, 2019.

[116] Jiaqiao Hu, Ping Hu, and Hyeong Soo Chang. A stochastic approximation framework for a class of randomized optimization algorithms. *IEEE Transactions on Automatic Control*, 57(1):165–178, 2011.

[117] Kevin Huang, Sahin Lale, Ugo Rosolia, Yuanyuan Shi, and Anima Anandkumar. Cem-gd: Cross-entropy method with gradient descent planner for model-based reinforcement learning. *arXiv preprint arXiv:2112.07746*, 2021.

[118] Kuang Yu Huang and Chuen-Jiuan Jane. A hybrid model for stock market forecasting and portfolio selection based on arx, grey system and rs theories. *Expert systems with applications*, 36(3):5387–5392, 2009.

[119] Po-Sen Huang, Haim Avron, Tara N Sainath, Vikas Sindhwani, and Bhuvana Ramabhadran. Kernel methods match deep neural networks on timit. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 205–209. IEEE, 2014.

[120] AR Humphries and AM Stuart. Runge–kutta methods for dissipative and gradient dynamical systems. *SIAM journal on numerical analysis*, 31(5): 1452–1485, 1994.

[121] M Yousuff Hussaini and Thomas A Zang. Spectral methods in fluid dynamics. *Annual review of fluid mechanics*, 19(1):339–367, 1987.

[122] Atil Iscen, Ken Caluwaerts, Jie Tan, Tingnan Zhang, Erwin Coumans, Vikas Sindhwani, and Vincent Vanhoucke. Policies modulating trajectory generators. In *Conference on Robot Learning*, pages 916–926. PMLR, 2018.

[123] Tadashi Ishihara, Hai-Jiao Guo, and Hiroshi Takeda. A design of discrete-time integral controllers with computation delays via loop transfer recovery. *Automatica*, 28(3):599–603, 1992.

[124] Dunham Jackson. *The theory of approximation*, volume 11. American Mathematical Soc., 1930.

[125] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr): 1563–1600, 2010.

[126] Magnus Jansson. Subspace identification and arx modeling. *IFAC Proceedings Volumes*, 36(16):1585–1590, 2003.

[127] Zhong-Ping Jiang and Yuan Wang. Input-to-state stability for discrete-time nonlinear systems. *Automatica*, 37(6):857–869, 2001.

[128] Anya R Jones. Gust encounters of rigid wings: Taming the parameter space. *Physical Review Fluids*, 5(11):110513, 2020.

[129] Galin L Jones et al. On the markov chain central limit theorem. *Probability surveys*, 1:299–320, 2004.

[130] Anatoli Juditsky, Håkan Hjalmarsson, Albert Benveniste, Bernard Delyon, Lennart Ljung, Jonas Sjöberg, and Qinghua Zhang. Nonlinear black-box models in system identification: Mathematical foundations. *Automatica*, 31 (12):1725–1750, 1995.

[131] Raphaël Jungers. *The Joint Spectral Radius: Theory and Applications*. Springer, 2009.

[132] Thomas Kailath, Ali H Sayed, and Babak Hassibi. Linear estimation, 2000.

[133] Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *arXiv preprint arXiv:2006.12466*, 2020.

[134] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

[135] K. Kambatla, N. Rapolu, S. Jagannathan, and A. Grama. Asynchronous algorithms in mapreduce. In *2010 IEEE International Conference on Cluster Computing*, pages 245–254, 2010.

[136] Stoyan Kanev and Tim van Engelen. Wind turbine extreme gust control. *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, 13(1):18–35, 2010.

[137] Taylan Kargin*, Sahin Lale*, Kamyar Azizzadenesheli, Animashree Anandkumar, and Babak Hassibi. Thompson sampling achieves $\tilde{O}\sqrt{T}$ regret in linear quadratic control. In *Conference on Learning Theory*, pages 3235–3284. PMLR, 2022.

[138] Taylan Kargin, Sahin Lale, Kamyar Azizzadenesheli, Animashree Anandkumar, and Babak Hassibi. Thompson sampling for partially observable linear-quadratic control. *2023 American Control Conference (ACC)*, 2023.

[139] Samuel Karlin. Positive operators. *Journal of Mathematics and Mechanics*, 8(6):907–937, 1959.

[140] Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. *Advances in Neural Information Processing Systems*, 30, 2017.

[141] Eammon Keogh, C Blake, and Chris J Merz. Uci repository of machine learning databases. *Irvine, CA: Uni of California, Department of Information and Computer Science*, 1998.

[142] H.K. Khalil. *Nonlinear Systems*, volume 3. Prentice Hall, 2002.

[143] Kia Khezeli and Eilyan Bitar. Safe linear stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10202–10209, 2020.

[144] George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

[145] Petar V Kokotovic. The joy of feedback: nonlinear and adaptive. *IEEE Control Systems Magazine*, 12(3):7–17, 1992.

[146] G Kowalewski. Ein mittelwertsatz für ein system von n integralen. *Z. Math. Phys.(Schlömilch Z.)*, 42:153–157, 1895.

[147] Frank Kozin. On relations between moment properties and almost sure lyapunov stability for linear stochastic systems. *Journal of Mathematical Analysis and Applications*, 10(2):342–353, 1965.

[148] NN Krasovskii. Problems of the theory of stability of motion, 1963.

[149] Michael Krieg, Kevin Nelson, and Kamran Mohseni. Distributed sensing for fluid disturbance compensation and motion control of intelligent robots. *Nature machine intelligence*, 1(5):216–224, 2019.

[150] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[151] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[152] M Krstić, Ioannis Kanellakopoulos, and PV Kokotović. Adaptive nonlinear control without overparametrization. *Systems & Control Letters*, 19(3):177–185, 1992.

[153] Miroslav Krstic. Feedback linearizability and explicit integrator forwarding controllers for classes of feedforward systems. *IEEE Transactions on Automatic Control*, 49(10):1668–1682, 2004.

[154] Miroslav Krstic, Petar V Kokotovic, and Ioannis Kanellakopoulos. *Nonlinear and adaptive control design*. John Wiley & Sons, Inc., 1995.

[155] T.L Lai. Asymptotically efficient adaptive control in stochastic regression models. *Advances in Applied Mathematics*, 7(1):23 – 45, 1986.

[156] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[157] Tze Leung Lai and Ching-Zong Wei. Asymptotically efficient self-tuning regulators. *SIAM Journal on Control and Optimization*, 25(2):466–481, 1987.

[158] Tze Leung Lai, Ching Zong Wei, et al. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

[159] Sahin Lale, Kamyar Azizzadenesheli, Anima Anandkumar, and Babak Hassibi. Stochastic linear bandits with hidden low rank structure. *arXiv preprint arXiv:1901.09490*, 2019.

[160] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. *Advances in Neural Information Processing Systems*, 33:20876–20888, 2020.

[161] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Regret minimization in partially observable linear quadratic control. *arXiv preprint arXiv:2002.00082*, 2020.

[162] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Adaptive control and regret minimization in linear quadratic gaussian (lqg) setting. In *2021 American Control Conference (ACC)*, pages 2517–2522. IEEE, 2021.

[163] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Finite-time system identification and adaptive control in autoregressive exogenous systems. In *Learning for Dynamics and Control*, pages 967–979. PMLR, 2021.

[164] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Model learning predictive control in nonlinear dynamical systems. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 757–762. IEEE, 2021.

[165] Sahin Lale, Oguzhan Teke, Babak Hassibi, and Anima Anandkumar. Stability and identification of random asynchronous linear time-invariant systems. In *Learning for Dynamics and Control*, pages 651–663. PMLR, 2021.

[166] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Animashree Anandkumar. Reinforcement learning with fast stabilization in linear dynamical systems. In *International Conference on Artificial Intelligence and Statistics*, pages 5354–5390. PMLR, 2022.

[167] Sahin Lale, Yuanyuan Shi, Guannan Qu, Kamyar Azizzadenesheli, Adam Wierman, and Anima Anandkumar. Kcrl: Krasovskii-constrained reinforcement learning with guaranteed stability in nonlinear dynamical systems. *arXiv preprint arXiv:2206.01704*, 2022.

[168] Sahin Lale, Peter I. Renn, Kamyar Azizzadenesheli, Babak Hassibi, Morteza Gharib, and Anima Anandkumar. Falcon: Fourier adaptive learning and control for disturbance rejection under extreme turbulence. *arXiv preprint*, 2023.

[169] Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, 2018.

[170] Fabien Lauer and Gérard Bloch. *Hybrid system identification: Theory and algorithms for learning switching models*, volume 478. Springer, 2018.

[171] Steven M LaValle. *Planning algorithms*. Cambridge university press, 2006.

[172] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[173] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[174] Bruce Lee and Andrew Lamperski. Non-asymptotic closed-loop system identification using autoregressive processes and hankel model reduction. *arXiv preprint arXiv:1909.02192*, 2019.

[175] Donghwan Lee and Jianghai Hu. Primal-dual q-learning framework for lqr design. *IEEE Transactions on Automatic Control*, 64(9):3756–3763, 2018.

[176] Jonathan N Lefebvre and Anya R Jones. Experimental investigation of airfoil performance in the wake of a circular cylinder. *AIAA Journal*, 57(7):2808–2818, 2019.

[177] Tyler Lekang and Andrew Lamperski. Sufficient conditions for persistency of excitation with step and relu activation functions. *arXiv preprint arXiv:2209.06286*, 2022.

[178] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436, 2018.

[179] Frank L Lewis, Draguna Vrabie, and Kyriakos G Vamvoudakis. Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers. *IEEE Control Systems Mag.*, 32(6):76–105, 2012.

[180] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[181] Wei C Li and Lorenz T Biegler. A multistep, newton-type control strategy for constrained, nonlinear processes. In *1989 American Control Conference*, pages 1526–1527. IEEE, 1989.

[182] Xiaodi Li, Xueyan Yang, and Shiji Song. Lyapunov conditions for finite-time stability of time-varying time-delay systems. *Automatica*, 103:135–140, 2019.

[183] Yingying Li, Aoxiao Zhong, Guannan Qu, and Na Li. Online markov decision processes with time-varying transition probabilities and rewards. In *ICML Real-world Sequential Decision Making workshop*, 2019.

[184] Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, Anima Anandkumar, et al. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2020.

[185] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint*, 2015.

[186] Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan AK Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7128–7148, 2021.

[187] Xin Liu, Bin Li, Pengyi Shi, and Lei Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. *Advances in Neural Information Processing Systems*, 34:24075–24086, 2021.

[188] Lennart Ljung. System identification. *Wiley Encyclopedia of Electrical and Electronics Engineering*, pages 1–19, 1999.

[189] Lennart Ljung and Tomas McKelvey. Subspace identification from closed loop data. *Signal processing*, 52(2):209–215, 1996.

[190] Winfried Lohmiller and Jean-Jacques E Slotine. On contraction analysis for non-linear systems. *Automatica*, 34(6):683–696, 1998.

[191] Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalent control of lqr is efficient. *arXiv preprint arXiv:1902.07826*, 2019.

[192] Horia Mania, Michael I Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *arXiv preprint arXiv:2006.10277*, 2020.

[193] Riccardo Marino and Patrizio Tomei. Robust adaptive regulation of linear time-varying systems. *IEEE Transactions on Automatic Control*, 45(7):1301–1311, 2000.

[194] Lorenz Meier, Petri Tanskanen, Lionel Heng, Gim Hee Lee, Friedrich Fraundorfer, and Marc Pollefeys. Pixhawk: A micro aerial vehicle design for autonomous flight using onboard computer vision. *Autonomous Robots*, 33 (1):21–39, 2012.

[195] Daniel Mellinger, Nathan Michael, and Vijay Kumar. Trajectory generation and control for precise aggressive maneuvers with quadrotors. *The International Journal of Robotics Research*, 31(5):664–674, 2012.

[196] Lingheng Meng, Rob Gorbet, and Dana Kulić. Memory-based deep reinforcement learning for pomdps. *IROS 2021*, 2021.

[197] Lingsheng Meng and Bing Zheng. The optimal perturbation bounds of the moore–penrose inverse under the frobenius norm. *Linear algebra and its applications*, 432(4):956–963, 2010.

[198] Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.

[199] Richard H Middleton and Graham C Goodwin. Adaptive control of time-varying linear systems. *IEEE Transactions on Automatic Control*, 33(2): 150–155, 1988.

[200] Abdulghani Mohamed, Reece Clothier, Simon Watkins, Roberto Sabatini, and Mujahid Abdulrahim. Fixed-wing mav attitude stability in atmospheric turbulence, part 1: Suitability of conventional sensors. *Progress in Aerospace Sciences*, 70:69–82, 2014.

[201] Abdulghani Mohamed, Mujahid Abdulrahim, Simon Watkins, and Reece Clothier. Development and flight testing of a turbulence mitigation system for micro air vehicles. *Journal of Field Robotics*, 33(5):639–660, 2016.

[202] Ahmadreza Moradipari, Christos Thrampoulidis, and Mahnoosh Alizadeh. Stage-wise conservative linear bandits. *Advances in neural information processing systems*, 33:11191–11201, 2020.

[203] Ahmadreza Moradipari, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Safe linear thompson sampling with side information. *IEEE Transactions on Signal Processing*, 69:3755–3767, 2021.

[204] Boaz Nadler. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6): 2791–2817, 2008.

[205] Anusha Nagabandi, Kurt Konolige, Sergey Levine, and Vikash Kumar. Deep dynamics models for learning dexterous manipulation. In *Conference on Robot Learning*, pages 1101–1112. PMLR, 2020.

[206] Kumpati S Narendra and Anuradha M Annaswamy. Persistent excitation in adaptive systems. *International Journal of Control*, 45(1):127–160, 1987.

[207] Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, 2009.

[208] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[209] Sandra J Norquay, Ahmet Palazoglu, and JoséA Romagnoli. Model predictive control based on wiener models. *Chemical Engineering Science*, 53(1):75–84, 1998.

[210] Reza Olfati-Saber and Richard M Murray. Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on automatic control*, 49(9):1520–1533, 2004.

[211] S. Orfanidis. An exact solution of the time-invariant discrete kalman filter. *IEEE Transactions on Automatic Control*, 27(1):240–242, 1982.

[212] Yi Ouyang, Mukul Gagrani, and Rahul Jain. Learning-based control of unknown linear systems with thompson sampling. *arXiv preprint arXiv:1709.04047*, 2017.

[213] Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. *arXiv preprint arXiv:1806.05722*, 2018.

[214] Necmiye Ozay, Constantino Lagoa, and Mario Sznaier. Set membership identification of switched linear systems with known number of subsystems. *Automatica*, 51:180–191, 2015.

[215] Michael O'Connell, Guanya Shi, Xichen Shi, Kamyar Azizzadenesheli, Anima Anandkumar, Yisong Yue, and Soon-Jo Chung. Neural-fly enables rapid learning for agile flight in strong winds. *Science Robotics*, 7(66):eabm6597, 2022.

[216] Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 2827–2835. PMLR, 2021.

[217] Alberto Padoan, Giordano Scarciotti, and Alessandro Astolfi. A geometric characterisation of persistently exciting signals generated by continuous-time autonomous systems. *IFAC-PapersOnLine*, pages 826–831, 2016.

[218] Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2009.

[219] Z. Peng, Y. Xu, M. Yan, and W. Yin. Arock: An algorithmic framework for asynchronous parallel coordinate updates. *SIAM Journal on Scientific Computing*, 38(5):A2851–A2879, 2016.

[220] Johannes E Pohl, Rolf Radespiel, Benjamin Herrmann, Steven L Brunton, and Richard Semaan. Gust mitigation through closed-loop control. i. trailing-edge flap response. *Physical Review Fluids*, 7(2):024705, 2022.

[221] Marios M Polycarpou and Petros A Ioannou. A robust adaptive nonlinear control design. In *1993 American control conference*, pages 1365–1369. IEEE, 1993.

[222] Steven B. Pope. *Turbulent Flows*. Cambridge university press, 2000.

[223] S Joe Qin. An overview of subspace identification. *Computers & chemical engineering*, 30(10-12):1502–1513, 2006.

[224] S Joe Qin and Lennart Ljung. *Closed-loop subspace identification with innovation estimation*. Linköping University Electronic Press, 2003.

[225] Guannan Qu*, Yuanyuan Shi*, Sahin Lale*, Anima Anandkumar, and Adam Wierman. Stable online control of linear time-varying systems. *arXiv preprint arXiv:2104.14134*, 2021.

[226] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

[227] Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2: 253–279, 2019.

[228] Peter I Renn and Morteza Gharib. Machine learning for flow-informed aerodynamic control in turbulent wind conditions. *Communications Engineering*, 1(1):1–9, 2022.

[229] Christian Rieger and Barbara Zwicknagl. Sampling inequalities for infinitely smooth functions, with applications to interpolation and machine learning. *Advances in Computational Mathematics*, 32(1):103–129, 2010.

[230] Anatol Roshko. On the wake and drag of bluff bodies. *Journal of the aeronautical sciences*, 22(2):124–132, 1955.

[231] Ugo Rosolia and Francesco Borrelli. Learning model predictive control for iterative tasks. a data-driven control framework. *IEEE Transactions on Automatic Control*, 63(7):1883–1896, 2017.

[232] Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

[233] Tuhin Sarkar, Alexander Rakhlin, and Munther Dahleh. Nonparametric system identification of stochastic switched linear systems. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3623–3628. IEEE, 2019.

[234] Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Finite-time system identification for partially observed lti systems of unknown order. *arXiv preprint arXiv:1902.01848*, 2019.

[235] Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *arXiv preprint arXiv:2002.08538*, 2020.

[236] David A Schoenwald. System identification using a wavelet-based approach. In *Proceedings of 32nd IEEE Conference on Decision and Control*, pages 3064–3065. IEEE, 1993.

[237] Martin H Schultz. L^{∞}-multivariate approximation theory. *SIAM Journal on Numerical Analysis*, 6(2):161–183, 1969.

[238] Jack Sherman and Winifred J Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950.

[239] Xichen Shi, Patrick Spieler, Ellande Tang, Elena-Sorina Lupu, Phillip Tokumaru, and Soon-Jo Chung. Adaptive nonlinear control of fixed-wing vtol with airflow vector sensing. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5321–5327. IEEE, 2020.

[240] Yuanyuan Shi, Guannan Qu, Steven Low, Anima Anandkumar, and Adam Wierman. Stability constrained reinforcement learning for real-time voltage control. *American Control Conference*, 2022.

[241] Md Abu S Shohag, Emily C Hammel, David O Olawale, and Okenwa I Okoli. Damage mitigation techniques in wind turbine blades: A review. *Wind Engineering*, 41(3):185–210, 2017.

[242] Max Simchowitz and Dylan J Foster. Naive exploration is optimal for online lqr. *arXiv preprint arXiv:2001.09576*, 2020.

[243] Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. *arXiv preprint arXiv:1802.08334*, 2018.

[244] Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. *arXiv preprint arXiv:1902.00768*, 2019.

[245] Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. *arXiv preprint arXiv:2001.09254*, 2020.

[246] Wolf Singer. Neuronal synchrony: a versatile code for the definition of relations? *Neuron*, 24(1):49–65, 1999.

[247] Jean-Jacques E Slotine, Weiping Li, et al. *Applied nonlinear control*, volume 199. Prentice hall Englewood Cliffs, NJ, 1991.

[248] Eduardo D Sontag. A 'universal' construction of artstein's theorem on nonlinear stabilization. *Systems & control letters*, 1989.

[249] Eduardo D Sontag. Input to state stability: Basic concepts and results. In *Nonlinear and optimal control theory*, pages 163–220. Springer, 2008.

[250] Eduardo D Sontag. *Mathematical control theory: deterministic finite dimensional systems*, volume 6. Springer Science & Business Media, 2013.

[251] Susanne Sterbing-D'Angelo, Mohit Chadha, Chen Chiu, Ben Falk, Wei Xian, Janna Barcelo, John M Zook, and Cynthia F Moss. Bat wing sensors support flight control. *Proceedings of the National Academy of Sciences*, 108(27): 11291–11296, 2011.

[252] Malcolm Strens. A bayesian framework for reinforcement learning. In *ICML*, volume 2000, pages 943–950, 2000.

[253] Andrew Stuart and Anthony R Humphries. *Dynamical systems and numerical analysis*, volume 2. Cambridge University Press, 1998.

[254] Zhendong Sun. *Switched linear systems: control and design*. Springer Science & Business Media, 2006.

[255] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[256] Kunihiko Taira, Steven L. Brunton, Scott T.M. Dawson, Clarence W. Rowley, Tim Colonius, Beverley J. McKeon, Oliver T. Schmidt, Stanislav Gordeyev, Vassilios Theofilis, and Lawrence S. Ukeiley. Modal analysis of fluid flows: An overview. *AIAA Journal*, 55(12), 2017.

[257] Herbert G Tanner, Ali Jadbabaie, and George J Pappas. Flocking in fixed and switching networks. *IEEE Transactions on Automatic control*, 52(5): 863–868, 2007.

[258] Sophie Tarbouriech, Germain Garcia, and Adolf H Glattfelder. *Advanced Strategies in Control Systems with Input and Output Constraints*, volume 346. Springer Science & Business Media, 2006.

[259] Andrew J Taylor, Victor D Dorobantu, Hoang M Le, Yisong Yue, and Aaron D Ames. Episodic learning with control lyapunov functions for uncertain robotic systems. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6878–6884. IEEE, 2019.

[260] Oguzhan Teke. *Signals on Networks: Random Asynchronous and Multirate Processing, and Uncertainty Principles*. PhD thesis, California Institute of Technology, July 2020.

[261] Oguzhan Teke and Palghat P. Vaidyanathan. Random node-asynchronous updates on graphs. *IEEE Transactions on Signal Processing*, 67(11):2794–2809, June 2019. doi: 10.1109/TSP.2019.2910485.

[262] Oguzhan Teke and Palghat P. Vaidyanathan. Random node-asynchronous graph computations. *IEEE Signal Processing Magazine*, 37(6):64–73, 2020. doi: 10.1109/MSP.2020.3014442.

[263] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4): 285–294, 1933.

[264] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[265] Emanuel Todorov and Weiwei Li. A generalized iterative lqg method for locally-optimal feedback control of constrained nonlinear stochastic systems. In *Proceedings of the 2005, American Control Conference, 2005.*, pages 300–306. IEEE, 2005.

[266] Michael S Triantafyllou, Gabriel D Weymouth, and Jianmin Miao. Biomimetic survival hydrodynamics and flow sensing. *Annual Review of Fluid Mechanics*, 48:1–24, 2016.

[267] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[268] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[269] Anastasios Tsiamis and George J Pappas. Finite sample analysis of stochastic system identification. *arXiv preprint arXiv:1903.09122*, 2019.

[270] Anastasios Tsiamis, Nikolai Matni, and George J Pappas. Sample complexity of kalman filtering for unknown systems. *arXiv preprint arXiv:1912.12309*, 2019.

[271] Hiroyasu Tsukamoto and Soon-Jo Chung. Neural contraction metrics for robust estimation and control: A convex optimization approach. *IEEE Control Systems Letters*, 5(1):211–216, 2020.

[272] Hiroyasu Tsukamoto, Soon-Jo Chung, and Jean-Jacques Slotine. Learning-based adaptive control using contraction theory. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 2533–2538. IEEE, 2021.

[273] Peter Uhlhaas, Gordon Pipa, Bruss Lima, Lucia Melloni, Sergio Neuenschwander, Danko Nikolić, and Wolf Singer. Neural synchrony in cortical networks: history, concept and current status. *Frontiers in integrative neuroscience*, 3:17, 2009.

[274] Peter J. Uhlhaas and Wolf Singer. Neural synchrony in brain disorders: Relevance for cognitive dysfunctions and pathophysiology. *Neuron*, 52(1): 155 – 168, 2006. doi: https://doi.org/10.1016/j.neuron.2006.09.020.

[275] Peter J Uhlhaas and Wolf Singer. Neural synchrony in brain disorders: relevance for cognitive dysfunctions and pathophysiology. *neuron*, 52(1):155–168, 2006.

[276] Andreas Ulbig, Theodor S Borsche, and Göran Andersson. Impact of low rotational inertia on power system stability and operation. *IFAC Proceedings Volumes*, 47(3):7290–7297, 2014.

[277] Jonas Umlauft, Lukas Pöhler, and Sandra Hirche. An uncertainty-based control lyapunov approach for control-affine systems modeled by gaussian process. *IEEE Control Systems Letters*, 2:483–488, 2018.

[278] Kyriakos G Vamvoudakis. Q-learning for continuous-time linear systems: A model-free infinite horizon optimal control approach. *Systems & Control Letters*, 100:14–20, 2017.

[279] Kyriakos G Vamvoudakis and Frank L Lewis. Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica*, 46(5):878–888, 2010.

[280] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

[281] Peter Van Overschee and Bart De Moor. N4sid: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, 30 (1):75–93, 1994.

[282] Peter Van Overschee and Bart De Moor. Closed loop subspace system identification. In *Proceedings of the 36th IEEE Conference on Decision and Control*, volume 2, pages 1848–1853. IEEE, 1997.

[283] Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM review*, 38(1):49–95, 1996.

[284] Namrata Vaswani and Praneeth Narayanamurthy. Finite sample guarantees for pca in non-isotropic and data-dependent noise. In *Communication, Control, and Computing (Allerton), 2017 55th Annual Allerton Conference on*, pages 783–789. IEEE, 2017.

[285] Michel Verhaegen. Identification of the deterministic part of mimo state space models given in innovations form from input-output data. *Automatica*, 30(1):61–74, 1994.

[286] Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2): 199, 2015.

[287] Pantelis R Vlachas, Jaideep Pathak, Brian R Hunt, Themistoklis P Sapsis, Michelle Girvan, Edward Ott, and Petros Koumoutsakos. Backpropagation algorithms and reservoir computing in recurrent neural networks for the forecasting of complex spatiotemporal dynamics. *Neural Networks*, 126: 191–217, 2020.

[288] Theodore von Kármán. Collapse of the tacoma narrows bridge. *Resonance*, 10, 2005.

[289] Wei Wang, Nanpeng Yu, Yuanqi Gao, and Jie Shi. Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems. *IEEE Transactions on Smart Grid*, 11(4):3008–3018, 2019.

[290] Simon Watkins, Jane Burry, Abdulghani Mohamed, Matthew Marino, Samuel Prudden, Alex Fisher, Nicola Kloet, Timothy Jakobi, and Reece Clothier. Ten questions concerning the use of drones in urban environments. *Building and Environment*, 167:106458, 2020.

[291] Per-Åke Wedin. Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13(2):217–232, 1973.

[292] Tyler Westenbroek, Fernando Castañeda, Ayush Agrawal, S Shankar Sastry, and Koushil Sreenath. Learning min-norm stabilizing control laws for systems with unknown dynamics. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 737–744. IEEE, 2020.

[293] Christian E Willert and Morteza Gharib. Digital particle image velocimetry. *Experiments in fluids*, 10(4):181–193, 1991.

[294] Grady Williams, Nolan Wagener, Brian Goldfain, Paul Drews, James M Rehg, Byron Boots, and Evangelos A Theodorou. Information theoretic mpc for model-based reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1714–1721. IEEE, 2017.

[295] CH Williamson. K. vortex dynamics in the cylinder wake. *Annual Review of Fluid Mechanics*, 28(1):477–539, 1996.

[296] Keith Worden. *Nonlinearity in structural dynamics: detection, identification and modelling*. CRC Press, 2019.

[297] Cathy Wu, Aboudy Kreidieh, Kanaad Parvate, Eugene Vinitsky, and Alexandre M Bayen. Flow: Architecture and benchmarking for reinforcement learning in traffic control. *arXiv preprint arXiv:1710.05465*, 10, 2017.

[298] Yuxiang Yang, Ken Caluwaerts, Atil Iscen, Tingnan Zhang, Jie Tan, and Vikas Sindhwani. Data efficient reinforcement learning for legged robots. In *Conference on Robot Learning*, pages 1–10. PMLR, 2020.

[299] Dante Youla, Hamid Jabr, and Jr Bongiorno. Modern wiener-hopf design of optimal controllers–part ii: The multivariable case. *IEEE Transactions on Automatic Control*, 21(3):319–338, 1976.

[300] Liqian Zhang, Yang Shi, Tongwen Chen, and Biao Huang. A new method for stabilization of networked control systems with random delays. *IEEE Transactions on automatic control*, 50(8):1177–1181, 2005.

[301] Lixian Zhang and El-Kébir Boukas. Stability and stabilization of markovian jump linear systems with partly unknown transition probabilities. *Automatica*, 45(2):463–468, 2009.

[302] Feiran Zhao and Keyou You. Primal-dual learning for the model-free risk-constrained linear quadratic regulator. In *Learning for Dynamics and Control*, pages 702–714. PMLR, 2021.

[303] Feiran Zhao, Keyou You, and Tamer Başar. Global convergence of policy gradient primal-dual methods for risk-constrained lqrs. *IEEE Transactions on Automatic Control*, 2023.

[304] Alex Zheng and Manfred Morari. Robust control of linear time-varying systems with constraints. In *Proceedings of 1994 American Control Conference-ACC'94*, volume 3, pages 2416–2420. IEEE, 1994.

[305] Kemin Zhou, John Comstock Doyle, Keith Glover, et al. *Robust and optimal control*, volume 40. Prentice hall New Jersey, 1996.

Appendix A

FURTHER PROOFS FOR CHAPTER 2

A.1 Proofs of Section 2.2

A.1.1 Projection Error Analysis, Proof of Lemma 2.2.4

In this section, we provide the general version of Lemma 2.2.4 with the proof details. As stated in the main text, in order to bound the projection error, we will use Davis-Kahan $\sin\Theta$ theorem which states the following:

Theorem A.1.1 ([68])**.** *Let $S, H \in \mathbb{R}^{d\times d}$ be symmetric matrices, such that $\hat{S} = S+H$. The eigenvalues of S and $\hat{S}$ are $\lambda_1 \geq \ldots \geq \lambda_m \geq \ldots \geq \lambda_d$ and $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_m \geq \ldots \geq \hat{\lambda}_d$ respectively. Define the eigendecompositions of S and $\hat{S}$:*

$$S = [U \quad U_o] \begin{bmatrix} \Lambda & 0 \\ 0 & \Lambda_o \end{bmatrix} [U \quad U_o]^T, \quad \hat{S} = [\hat{U} \quad \hat{U}_o] \begin{bmatrix} \hat{\Lambda} & 0 \\ 0 & \hat{\Lambda}_o \end{bmatrix} [\hat{U} \quad \hat{U}_o]^T,$$

where Λ and $\hat{\Lambda}$ are diagonal matrices with first m eigenvalues of S and $\hat{S}$ respectively. $U = (u_1, \ldots, u_m) \in \mathbb{R}^{d\times m}$ and $\hat{U} = (\hat{u}_1, \ldots, \hat{u}_m) \in \mathbb{R}^{d\times m}$ denote the corresponding eigenvectors. Define $\delta := \inf\{|\hat{\lambda} - \lambda| : \lambda \in [\lambda_m, \lambda_1], \hat{\lambda} \in (-\infty, \hat{\lambda}_{m+1}]\}$. If $\delta > 0$, then $\sin\Theta_m$, sine of the largest principal angle between the column spans of U and $\hat{U}$, can be upper bounded as

$$\sin\Theta_m \leq \frac{\|\hat{S}U - U\Lambda\|_2}{\delta} = \frac{\|\hat{S}U - U\Lambda\|_2}{|\lambda_m - \hat{\lambda}_{m+1}|}. \tag{A.1}$$

Notice that in order to use Davis-Kahan $\sin\Theta$ theorem in our setting, we need to pick 2 symmetric matrices S and $\hat{S}$ such that their first m eigenvectors have the same span with the subspaces that P and $\hat{P}$ project to. Followed by these choices, in order to get a non-trivial bound we require a significant eigengap between λ_m and $\hat{\lambda}_{m+1}$, due to denominator in (A.1). Define $t_{\min,\delta} = \left(\sqrt{\frac{2d_x g_x}{K}\log\frac{m}{\delta}} + \Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}}\right)^2$. The following presents a more general version of Lemma 2.2.4.

Lemma A.1.2 (General version of Lemma 2.2.4)**.** *Fix any $\delta \in (0, 1/3)$. Suppose that Assumption 1 holds. Then with probability at least $1 - 3\delta$,*

$$\|\hat{P}_t - P\|_2 \leq \frac{\Gamma\sqrt{\frac{\alpha}{tK}\log\frac{2d}{\delta}}}{1 - \sqrt{\frac{2d_x g_x}{tK}\log\frac{m}{\delta}} - \Gamma\sqrt{\frac{\alpha}{tK}\log\frac{2d}{\delta}}}, \qquad \forall t \geq t_{w,\delta}. \tag{A.2}$$

Proof. We set $\hat{S} = \frac{1}{n}\sum_{i=1}^{n} \hat{x}_i \hat{x}_i^T$ and $S = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T + VV^T\Sigma_\psi VV^T$ where $n = tK$. Let U be the top m eigenvectors of S. Notice that $\text{span}(U) = \text{span}(V)$ and $\hat{V}_t$ is the matrix of top m eigenvectors of $\hat{S}$. Therefore, one can apply Theorem A.1.1 with given choices of S and $\hat{S}$, to bound $\|\hat{P}_t - P\|_2$. Since $\|\hat{S}U - U\Lambda\|_2 = \|(\hat{S} - S)V\|_2$,

$$\begin{aligned}\|\hat{P}_t - P\|_2 \leq \frac{\|(\hat{S} - S)V\|_2}{\lambda_m(S) - \lambda_{m+1}(\hat{S})} &\overset{(1)}{\leq} \frac{\|(\hat{S} - S)V\|_2}{\lambda_m(S) - \|\hat{S} - S\|_2} \\ &\overset{(2)}{\leq} \frac{\|\mathbb{E}[\hat{S} - S]V\|_2 + \|\hat{S} - S - \mathbb{E}[\hat{S} - S]\|_2}{\lambda_m(S) - \|\mathbb{E}[\hat{S} - S]\|_2 - \|\hat{S} - S - \mathbb{E}[\hat{S} - S]\|_2},\end{aligned}$$

where (1) follows from Weyl's inequality and the fact that S is rank m, $\lambda_{m+1}(S) = \ldots = \lambda_d = 0$, and (2) is due to triangle inequality. With the given choices of S and $\hat{S}$ and Assumption 2.2.1, we have the following:

- $\lambda_m(S) \geq \lambda_m(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T) + \lambda_{\min}(V^T\Sigma_\psi V) = \lambda_m(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T) + \sigma^2$
- $\hat{S} - S = \frac{1}{n}\sum_{i=1}^{n} \psi_i\psi_i^T + \frac{1}{n}\sum_{i=1}^{n} x_i\psi_i^T + \frac{1}{n}\sum_{i=1}^{n} \psi_i x_i^T - VV^T\Sigma_\psi VV^T$
- $\|\mathbb{E}[\hat{S} - S]\|_2 = \|\sigma^2 I_d - \sigma^2 P\|_2 = \sigma^2$
- $\mathbb{E}[\hat{S} - S]V = \Sigma_\psi V - VV^T\Sigma_\psi V = V_\perp V_\perp^T \Sigma_\psi V = 0$
- $\hat{S} - S - \mathbb{E}[\hat{S} - S] = \frac{1}{n}\sum_{i=1}^{n} \psi_i\psi_i^T - \Sigma_\psi + \frac{1}{n}\sum_{i=1}^{n} x_i\psi_i^T + \frac{1}{n}\sum_{i=1}^{n} \psi_i x_i^T$.

Inserting these expressions we get,

$$\|\hat{P}_t - P\|_2 \leq \frac{\|\frac{1}{n}\sum_{i=1}^{n} \psi_i\psi_i^T - \Sigma_\psi + \frac{1}{n}\sum_{i=1}^{n} x_i\psi_i^T + \frac{1}{n}\sum_{i=1}^{n} \psi_i x_i^T\|_2}{\lambda_m(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T) - \|\frac{1}{n}\sum_{i=1}^{n} \psi_i\psi_i^T - \Sigma_\psi + \frac{1}{n}\sum_{i=1}^{n} x_i\psi_i^T + \frac{1}{n}\sum_{i=1}^{n} \psi_i x_i^T\|_2}. \tag{A.3}$$

We first bound $\lambda_m(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T)$. From Assumption 1, $\lambda_{max}\left(x_i x_i^T\right) \leq d_x\lambda_+$ for all $i \in [n]$ and from the model properties, $\lambda_m\left(\sum_{i=1}^{n} \mathbb{E}[x_i x_i^T]\right) = n\lambda_-$. Using Matrix Chernoff Inequality [268], one can get that

$$\mathbb{P}\left[\lambda_m\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right) \leq \lambda_-\left(1 - \sqrt{\frac{2d_x g_x}{n}\log\frac{m}{\delta}}\right)\right] \leq \delta. \tag{A.4}$$

Now we consider $\|\frac{1}{n}\sum_{i=1}^{n} \psi_i\psi_i^T - \Sigma_\psi + \frac{1}{n}\sum_{i=1}^{n} x_i\psi_i^T + \frac{1}{n}\sum_{i=1}^{n} \psi_i x_i^T\|_2$. From triangle inequality, we have,

$$\left\|\frac{1}{n}\sum_{i=1}^{n} \psi_i\psi_i^T - \Sigma_\psi + \frac{1}{n}\sum_{i=1}^{n} x_i\psi_i^T + \frac{1}{n}\sum_{i=1}^{n} \psi_i x_i^T\right\|_2 \leq \left\|\frac{1}{n}\sum_{i=1}^{n} \psi_i\psi_i^T - \Sigma_\psi\right\|_2 + 2\left\|\frac{1}{n}\sum_{i=1}^{n} x_i\psi_i^T\right\|_2.$$

We will consider each term on the right-hand side separately. If Assumption 1 holds, then we have:

$$\mathbb{E}[\psi_i\psi_i^T] = \Sigma_\psi, \;\; \|\psi_i\psi_i^T\|_2 \leq d_\psi\sigma^2, \;\; \|\mathbb{E}[\psi_i\psi_i^T\psi_i\psi_i^T]\|_2 \leq d_\psi\sigma^4.$$

Applying Matrix Bernstein Inequality [268], we get

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\psi_i\psi_i^T - \Sigma_\psi\right\|_2 \geq 2\sigma^2\sqrt{\frac{d_\psi}{n}\log\frac{2d}{\delta}}\right] \leq \delta \quad \text{for } 2\sqrt{\frac{d_\psi}{n}\log\frac{2d}{\delta}} \leq 1.5. \quad \text{(A.5)}$$

Under the same assumption for the second term we have:

$$\begin{aligned}
\mathbb{E}[x_i\psi_i^T] &= 0, \\
\|x_i\psi_i^T\|_2 &= \sqrt{\lambda_{max}(\psi_i x_i^T x_i\psi_i^T)} \leq \sqrt{d_x\lambda_+ d_\psi\sigma^2}, \\
\|\mathbb{E}[x_i\psi_i^T\psi_i x_i^T]\|_2 &\leq d_\psi\sigma^2\|\mathbb{E}[x_i x_i^T]\|_2 = d_\psi\lambda_+\sigma^2, \\
\|\mathbb{E}[\psi_i x_i^T x_i\psi_i^T]\|_2 &\leq d_x\lambda_+\|\mathbb{E}[\psi_i\psi_i^T]\|_2 \leq d_x\lambda_+\sigma^2.
\end{aligned}$$

Once again applying Matrix Bernstein Inequality [268],

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}x_i\psi_i^T\right\|_2 \geq 2\sqrt{\lambda_+\sigma^2}\sqrt{\frac{\alpha}{n}\log\frac{2d}{\delta}}\right] \leq \delta \quad \text{for } 2\sqrt{\frac{\alpha}{n}\log\frac{2d}{\delta}} \leq 1.5. \quad \text{(A.6)}$$

Finally, combining (A.4), (A.5), (A.6) and using union bound, for any round $t \geq t_{\min,\delta}$, we get:

$$\|\hat{P}_t - P\|_2 \leq \min\left(\frac{\Gamma\sqrt{\frac{\alpha}{tK}\log\frac{2d}{\delta}}}{1 - \sqrt{\frac{2d_x g_x}{tK}\log\frac{m}{\delta}} - \Gamma\sqrt{\frac{\alpha}{tK}\log\frac{2d}{\delta}}}, 1\right) \quad \text{w.p. } 1 - 3\delta.$$

As explained in the main text, due to the equivalence between the projection error and the sine of the largest angle between the subspaces, the projection error is always bounded by 1. Thus, in our bound we impose that constraint. Notice that the lower bound on t is to satisfy that concentration inequalities provide meaningful results. In other words, $Kt_{\min,\delta}$ is the number of samples required to have a non-negative denominator to use the Davis-Kahan $\sin\Theta$ theorem. However, observe that we need $Kt_{w,\delta}$ samples to obtain a high probability error bound which is non-trivial, *i.e.* less than 1 and $t_{w,\delta} = 4t_{\min,\delta}$. Therefore, for any $t \geq t_{w,\delta}$ the stated bound (A.2) in the lemma holds with high probability, and for any $1 \leq t \leq t_{w,\delta}$ we bound the projection error by 1.

The only step remaining to show that the lemma holds $\forall t \geq t_{w,\delta}$. This requires an argument that shows that this bound is valid uniformly over all rounds. To this end,

we use stopping time construction, which goes back at least to [90]. Define the bad event,

$$E_\tau(\delta) = \left\{ \|\hat{P}_\tau - P\|_2 > \frac{\Gamma\sqrt{\frac{\alpha}{\tau K}\log\frac{2d}{\delta}}}{1 - \sqrt{\frac{2d_x g_x}{\tau K}\log\frac{m}{\delta}} - \Gamma\sqrt{\frac{\alpha}{\tau K}\log\frac{2d}{\delta}}} \right\}.$$

We are interested in the probability of $\bigcup_{t \geq t_{w,\delta}} E_t(\delta)$. Define $\tau(\omega) = \min\{t \geq t_{w,\delta} : \omega \in E_t(\delta)\}$, with the convention that $\min \emptyset = \infty$. Then, τ is a stopping time. Thus, $\bigcup_{t \geq t_{w,\delta}} E_t(\delta) = \{\omega : \tau(\omega) < \infty\}$. The Lemma A.1.2 can be obtained as follows:

$$\begin{aligned}\mathbb{P}\left[\bigcup_{t \geq t_{w,\delta}} E_t(\delta)\right] = \mathbb{P}[\tau < \infty] &= \mathbb{P}\left[\|\hat{P}_\tau - P\|_2 > \frac{\Gamma\sqrt{\frac{\alpha}{\tau K}\log\frac{2d}{\delta}}}{1 - \sqrt{\frac{2d_x g_x}{\tau K}\log\frac{m}{\delta}} - \Gamma\sqrt{\frac{\alpha}{\tau K}\log\frac{2d}{\delta}}}, \tau < \infty\right] \\ &= \mathbb{P}\left[\|\hat{P}_\tau - P\|_2 > \frac{\Gamma\sqrt{\frac{\alpha}{\tau K}\log\frac{2d}{\delta}}}{1 - \sqrt{\frac{2d_x g_x}{\tau K}\log\frac{m}{\delta}} - \Gamma\sqrt{\frac{\alpha}{\tau K}\log\frac{2d}{\delta}}}\right] \leq 3\delta.\end{aligned}$$

Finally, notice that Lemma 2.2.4 presented in the main text is direct consequence of having denominator at (A.2) greater than $\frac{1}{2}$ for all $t \geq t_{w,\delta}$. □

A.1.2 Proof of Theorem 2.2.5

Without loss of generality, assume that $R = 1$ since by appropriately scaling S_t, this can be achieved. Let $\lambda \in \mathbb{R}^d$ be a Gaussian random vector that is independent of all the other random variables and has covariance matrix $C^{-1} = \frac{1}{\lambda} I_d$. Consider for any $t \geq 0$,

$$M_t^\lambda = \exp\left(\lambda^T S_t - \frac{1}{2}\big(\lambda^T \sum_{i=1}^{t} \hat{P}_t \hat{X}_{i-1}\big)^2\right).$$

Define

$$M_t = \mathbb{E}_\lambda[M_t^\lambda | F_\infty],$$

where F_∞ is the tail σ-algebra of the filtration, *i.e.* the σ-algebra generated by the union of all the events in the filtration. Thus,

$$M_t = \int_{\mathbb{R}^d} \exp\left(\lambda^T S_t - \frac{1}{2}\lambda^T \hat{P}_t \hat{\Sigma}_{t-1} \hat{P}_t \lambda\right) f(\lambda) d\lambda$$

where $f(\lambda)$ is the pdf of λ. The following lemma will be crucial in proving the theorem.

Lemma A.1.3. $\mathbb{E}[M_t] \leq 1$ *for all* $t \geq 1$.

Proof.

$$\mathbb{E}[M_t] = \mathbb{E}\left[\int_{\mathbb{R}^d} \exp\left(\lambda^T S_t - \frac{1}{2}\lambda^T \hat{P}_t \hat{\Sigma}_{t-1} \hat{P}_t \lambda\right) f(\lambda) d\lambda \right]$$
$$\mathbb{E}[M_t] = \int_{\mathbb{R}^d} \mathbb{E}\left[\exp\left(\lambda^T S_t - \frac{1}{2}\lambda^T \hat{P}_t \hat{\Sigma}_{t-1} \hat{P}_t \lambda\right)\right] f(\lambda) d\lambda.$$

If one can show that $\mathbb{E}\left[\exp\left(\lambda^T S_t - \frac{1}{2}\lambda^T \hat{P}_t \hat{\Sigma}_{t-1} \hat{P}_t \lambda\right)\right] \leq 1$, then the claim follows. In the following, we use the law of total expectation.

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda^T S_t - \frac{1}{2}\lambda^T \hat{P}_t \hat{\Sigma}_{t-1} \hat{P}_t \lambda\right)\right] && \text{(A.7)}\\
&\quad = \mathbb{E}\left[\mathbb{E}_{\eta_{t-1}}\left[\exp\left(\lambda^T \sum_{i=1}^{t} \hat{P}_t \hat{X}_{i-1} \eta_{i-1} - \frac{1}{2}\lambda^T \hat{P}_t \Big(\sum_{i=1}^{t} \hat{X}_{i-1} \hat{X}_{i-1}^T \Big) \hat{P}_t \lambda\right) \Big| F_{t-1}\right]\right] \\
&\quad \leq \mathbb{E}\left[\exp\left(\lambda^T \sum_{i=1}^{t-1} \hat{P}_t \hat{X}_{i-1} \eta_{i-1} - \frac{1}{2}\lambda^T \hat{P}_t \Big(\sum_{i=1}^{t-1} \hat{X}_{i-1} \hat{X}_{i-1}^T \Big) \hat{P}_t \lambda\right)\right] && \text{(A.8)}\\
&\quad = \mathbb{E}\left[\mathbb{E}_{\eta_{t-2}}\left[\exp\left(\lambda^T \sum_{i=1}^{t-1} \hat{P}_t \hat{X}_{i-1} \eta_{i-1} - \frac{1}{2}\lambda^T \hat{P}_t \Big(\sum_{i=1}^{t-1} \hat{X}_{i-1} \hat{X}_{i-1}^T \Big) \hat{P}_t \lambda\right) \Big| F_{t-2}\right]\right] \\
&\quad \vdots \\
&\quad \leq 1,
\end{aligned}$$

where (A.8) follows from the assumption that η_t is conditionally R-sub-gaussian. □

We will use Lemma A.1.3 shortly but we first calculate M_t. For a positive definite matrix K, define $g(K) \coloneqq \sqrt{(2\pi)^m / \det(K)} = \int_{\mathbb{R}^m} \exp(-\frac{1}{2} x^T K x) dx$. One can

calculate M_t as follows,

$$
\begin{aligned}
M_t &= \int_{\mathbb{R}^d} \exp\left(\lambda^T S_t - \frac{1}{2}\lambda^T \bar{A}_t \lambda\right) f(\lambda) d\lambda \\
&= \int_{\mathbb{R}^m} \exp\left(\bar{\lambda}^T \hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t - \frac{1}{2}\bar{\lambda}^T \hat{V}_t^T \mathbf{X}_t \mathbf{X}_t^T \hat{V}_t \bar{\lambda}\right) f(\bar{\lambda}) d\bar{\lambda} \qquad \text{(A.9)} \\
&= \int_{\mathbb{R}^m} \exp\left(-\frac{1}{2}\|\bar{\lambda} - \bar{B}_t^{-1}\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t} + \frac{1}{2}\|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t^{-1}}\right) f(\bar{\lambda}) d\bar{\lambda} \\
&= \frac{\exp\left(\frac{1}{2}\|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t^{-1}}\right)}{g(\bar{C})} \int_{\mathbb{R}^m} \exp\left(-\frac{1}{2}\left(\|\bar{\lambda} - \bar{B}_t^{-1}\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t} + \|\bar{\lambda}\|^2_{\bar{C}}\right)\right) d\bar{\lambda} \qquad \text{(A.10)} \\
&= \frac{\exp\left(\frac{1}{2}\|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t^{-1}}\right)}{g(\bar{C})} \int_{\mathbb{R}^m} \exp\left(-\frac{1}{2}\left(\|\bar{\lambda} - (\bar{C} + \bar{B}_t)^{-1}\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{C}+\bar{B}_t} + \|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t^{-1}}\right.\right. \\
&\qquad \left.\left. - \|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{(\bar{C}+\bar{B}_t)^{-1}}\right)\right) d\bar{\lambda} \qquad \text{(A.11)} \\
&= \frac{\exp\left(\frac{1}{2}\|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{(\bar{C}+\bar{B}_t)^{-1}}\right)}{g(\bar{C})} \int_{\mathbb{R}^m} \exp\left(-\frac{1}{2}\left(\|\bar{\lambda} - (\bar{C} + \bar{B}_t)^{-1}\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{C}+\bar{B}_t}\right)\right) d\bar{\lambda} \\
&= \frac{\exp\left(\frac{1}{2}\|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{(\bar{C}+\bar{B}_t)^{-1}}\right)}{g(\bar{C})} g(\bar{C} + \bar{B}_t) = \left(\frac{\det(\bar{C})}{\det(\bar{C} + \bar{B}_t)}\right)^{1/2} \exp\left(\frac{1}{2}\|S_t\|^2_{(C+\bar{A}_t)^\dagger}\right),
\end{aligned}
$$

where in (A.9) there is a change of integration with $\bar{\lambda} = \hat{V}_t^T \lambda$, (A.10) follows from the fact that $f(\bar{\lambda}) = \frac{\exp(-\frac{1}{2}\bar{\lambda}^T \bar{C} \bar{\lambda})}{\sqrt{(2\pi)^m \det(\bar{C}^{-1})}}$ and defining $\bar{C} = \hat{V}_t^T C \hat{V}_t$ and finally (A.11) follows since

$$
\begin{aligned}
&\|\bar{\lambda} - \bar{B}_t^{-1}\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t} + \|\bar{\lambda}\|^2_{\bar{C}} \\
&\qquad = \|\bar{\lambda} - (\bar{C} + \bar{B}_t)^{-1}\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{C}+\bar{B}_t} + \|\bar{B}_t^{-1}\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t} - \|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{(\bar{C}+\bar{B}_t)^{-1}} \\
&\qquad = \|\bar{\lambda} - (\bar{C} + \bar{B}_t)^{-1}\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{C}+\bar{B}_t} + \|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{\bar{B}_t^{-1}} - \|\hat{V}_t^T \mathbf{X}_t \boldsymbol{\eta}_t\|^2_{(\bar{C}+\bar{B}_t)^{-1}}.
\end{aligned}
$$

Consider the following equivalence:

$$
\begin{aligned}
\mathbb{P}\left[\|S_t\|^2_{(C+\bar{A}_t)^\dagger} > 2\log\left(\frac{\det(\bar{C} + \bar{B}_t)^{1/2}}{\delta \det(\bar{C})^{1/2}}\right)\right] &= \mathbb{P}\left[\frac{\exp\left(\frac{1}{2}\|S_t\|^2_{(C+\bar{A}_t)^\dagger}\right)\delta}{\left(\frac{\det(\bar{C}+\bar{B}_t)}{\det(\bar{C})}\right)^{1/2}} > 1\right] \\
&\leq \mathbb{E}\left[\frac{\exp\left(\frac{1}{2}\|S_t\|^2_{(C+\bar{A}_t)^\dagger}\right)\delta}{\left(\frac{\det(\bar{C}+\bar{B}_t)}{\det(\bar{C})}\right)^{1/2}}\right] \qquad \text{(A.12)} \\
&= \mathbb{E}_{F_t}[M_t]\delta \leq \delta, \qquad \text{(A.13)}
\end{aligned}
$$

where (A.12) follows from Markov's inequality and (A.13) is due to Lemma A.1.3. Notice that, $A_t = \bar{A}_t + C$ and $B_t = \bar{B}_t + \bar{C}$. We will once again use a stopping-time construction. Define the bad event,

$$E_t(\delta) = \left\{ \|S_t\|^2_{A_t^\dagger} > 2R^2 \log\left(\frac{\det(B_t)^{1/2}}{\delta \det(\bar{C})^{1/2}} \right) \right\}.$$

We are interested in the probability of $\bigcup_{t\geq 0} E_t(\delta)$. Define $\tau(\omega) = \min\{t \geq 0 : \omega \in E_t(\delta)\}$, with the convention that $\min \emptyset = \infty$. Then, τ is a stopping time. Thus, $\bigcup_{t\geq 0} E_t(\delta) = \{\omega : \tau(\omega) < \infty\}$. The Theorem 2.2.5 can be obtained as follows:

$$\begin{aligned}\mathbb{P}\left[\bigcup_{t\geq 0} E_t(\delta) \right] = \mathbb{P}[\tau < \infty] &= \mathbb{P}\left[\|S_\tau\|^2_{A_\tau^\dagger} > 2R^2 \log\left(\frac{\det(B_\tau)^{1/2} \det(\bar{C})^{-1/2}}{\delta} \right), \tau < \infty \right] \\ &\leq \mathbb{P}\left[\|S_\tau\|^2_{A_\tau^\dagger} > 2R^2 \log\left(\frac{\det(B_\tau)^{1/2} \det(\bar{C})^{-1/2}}{\delta} \right), \right] \leq \delta.\end{aligned}$$

Since $C = \lambda I_d$, inserting $\bar{C} = \lambda I_m$ proves the theorem.

A.1.3 Proofs of Lemma 2.2.6 and Lemma 2.2.7

In this section, we provide the proofs or Lemma 2.2.6 and Lemma 2.2.7. First, recall that $A_t = \hat{P}_t(\hat{\Sigma}_{t-1} + \lambda I_d)\hat{P}_t$. Let B_t be a symmetric matrix such that $A_t = \hat{V}_t B_t \hat{V}_t^T$. Notice that B_t is a full rank $m \times m$ matrix. Also define $\bar{A}_t = A_t - \lambda \hat{P}_t = \hat{P}_t \hat{\Sigma}_{t-1} \hat{P}_t = \hat{V}_t \hat{V}_t^T \hat{\Sigma}_{t-1} \hat{V}_t \hat{V}_t^T = \hat{V}_t \bar{B}_t \hat{V}_t^T$ where $\bar{B}_t = \hat{V}_t^T \hat{\Sigma}_{t-1} \hat{V}_t = B_t - \lambda I_m$.

Proof of Lemma 2.2.6: $\det(B_t) = \det(\hat{V}_t^T \hat{\Sigma}_{t-1} \hat{V}_t + \lambda I_m) = \alpha_1 \alpha_2 \cdots \alpha_m$ where α_is are the eigenvalues of B_t. Notice that

$$\begin{aligned}\sum_{i=1}^m \alpha_i = m\lambda + \operatorname{tr}\left(\hat{V}_t^T \Big(\sum_{i=1}^t \hat{X}_{i-1} \hat{X}_{i-1}^T \Big) \hat{V}_t \right) &= m\lambda + \sum_{i=1}^t \operatorname{tr}\left(\hat{V}_t^T \hat{X}_{i-1} \hat{X}_{i-1}^T \hat{V}_t \right) \\ &\leq m\lambda + \sum_{i=1}^t \|\hat{X}_{i-1}\|_2^2 \leq m\lambda + tL^2\end{aligned}$$

from Assumptions 1 and 2. Using AM-GM inequality, *i.e*, $\sqrt[m]{\alpha_1 \alpha_2 \cdots \alpha_m} \leq \frac{1}{m} \sum_{i=1}^m \alpha_i$, we get $\alpha_1 \alpha_2 \cdots \alpha_m \leq \left(\lambda + \frac{tL^2}{m} \right)^m$. □

Lemma A.1.4. *Suppose Assumptions 2.2.1 and 2.2.2 hold. Then*

$$\sum_{i=1}^t \left\| \hat{V}_t^T \hat{X}_{i-1} \right\|^2_{B^{-1}_{t,i-1}} \leq \gamma m \log\left(1 + \frac{tL^2}{m\lambda} \right).$$

Proof. Analyzing $\det(B_t)$ at round t, we get the following:

$$\begin{aligned}\det\left(B_{t,t}\right) &= \det\left(B_{t,t-1} + \hat{V}_t^T \hat{X}_{t-1} \hat{X}_{t-1}^T \hat{V}_t\right) \\ &= \det\left(B_{t,t-1}^{1/2}\left(I_m + B_{t,t-1}^{-1/2} \hat{V}_t^T \hat{X}_{t-1} \hat{X}_{t-1}^T \hat{V}_t B_{t,t-1}^{-1/2}\right) B_{t,t-1}^{1/2}\right) \\ &= \det\left(B_{t,t-1}\right)\left(1 + \|\hat{V}_t^T \hat{X}_{t-1}\|^2_{B_{t,t-1}^{-1}}\right) = \lambda^m \prod_{i=1}^{t}\left(1 + \|\hat{V}_t^T \hat{X}_{i-1}\|^2_{B_{t,i-1}^{-1}}\right).\end{aligned}$$

Thus, $\sum_{i=1}^{t} \log(1 + \|\hat{V}_t^T \hat{X}_{i-1}\|^2_{B_{t,i-1}^{-1}}) = \log \frac{\det(B_t)}{\lambda^m} \leq m \log\left(1 + \frac{tL^2}{m\lambda}\right)$ where inequality follows from Lemma 2.2.6. Recall the definition of $\gamma = \frac{L^2}{\lambda \log\left(1+\frac{L^2}{\lambda}\right)}$. Since Assumption 1 and 2 hold, $\|\hat{V}_t^T \hat{X}_{i-1}\|^2_{B_{t,i-1}^{-1}} \leq \frac{L^2}{\lambda}$. Using $\|\hat{V}_t^T \hat{X}_{i-1}\|^2_{B_{t,i-1}^{-1}} \leq \gamma \log(1 + \|\hat{V}_t^T \hat{X}_{i-1}\|^2_{B_{t,i-1}^{-1}})$, which is true for $\|\hat{V}_t^T \hat{X}_{i-1}\|^2_{B_{t,i-1}^{-1}} \leq \frac{L^2}{\lambda}$, we get

$$\sum_{i=1}^{t} \left\|\hat{V}_t^T \hat{X}_{i-1}\right\|^2_{B_{t,i-1}^{-1}} \leq \gamma \sum_{i=1}^{t} \log(1 + \|\hat{V}_t^T \hat{X}_{i-1}\|^2_{B_{t,i-1}^{-1}})$$

The lemma follows immediately. □

Finally, we provide the bound on $\|(A_t^\dagger)^{1/2} \hat{P}_t \hat{\Sigma}_{t-1}\|_2$, *i.e.* Lemma 2.2.7 as follows.

Proof of Lemma 2.2.7: Recall that $\hat{\Sigma}_{t-1} = \sum_{i=1}^{t-1} \hat{X}_i \hat{X}_i^T$. With this, we get:

$$\begin{aligned}&\|(A_t^\dagger)^{1/2} \hat{P}_t \hat{\Sigma}_{t-1}\|_2 \\ &\leq \sum_{i=1}^{t} \left\|(A_t^\dagger)^{1/2} \hat{P}_t \hat{X}_{i-1} \hat{X}_{i-1}^T\right\|_2 \quad \text{Using Weyl's inequality for singular values} \\ &\leq \sum_{i=1}^{t} \left\|(A_t^\dagger)^{1/2} \hat{P}_t \hat{X}_{i-1}\right\|_2 \|\hat{X}_{i-1}\|_2 \quad \text{From Cauchy Schwarz} \\ &\leq L \sum_{i=1}^{t} \left\|\hat{P}_t \hat{X}_{i-1}\right\|_{A_t^\dagger} \quad \text{From Assumption 1} \\ &= L \sum_{i=1}^{t} \left\|\hat{V}_t^T \hat{X}_{i-1}\right\|_{B_t^{-1}} \quad \text{From the equality that } \hat{X}_{i-1}^T \hat{V}_t B_t^{-1} \hat{V}_t^T \hat{X}_{i-1} = \hat{X}_{i-1}^T \hat{P}_t A_t^\dagger \hat{P}_t \hat{X}_{i-1} \\ &\leq L \sum_{i=1}^{t} \left\|\hat{V}_t^T \hat{X}_{i-1}\right\|_{B_{t,i-1}^{-1}} \quad \text{Since at round t, } B_{t,i} = B_{t,i-1} + \hat{V}_t^T \hat{X}_i \hat{X}_i^T \hat{V}_t \\ &\leq L\sqrt{t}\sqrt{\sum_{i=1}^{t} \left\|\hat{V}_t^T \hat{X}_{i-1}\right\|^2_{B_{t,i-1}^{-1}}} \leq L\sqrt{\gamma m t}\sqrt{\log\left(1 + \frac{tL^2}{m\lambda}\right)} \quad \text{From Lemma A.1.4.}\end{aligned}$$

□

A.1.4 Proofs of Lemma 2.2.9 and Lemma 2.2.10

Proof of Lemma 2.2.9:

$$\begin{aligned}
|(\hat{P}_k\hat{x})^T(\nu-\theta_k)| &= |(\hat{P}_k\hat{x})^T(A_k^{\dagger})^{1/2}A_k^{1/2}(\nu-\theta_k)| \qquad \text{since } (A_k^{\dagger})^{1/2}A_k^{1/2}=\hat{P}_k\\
&= |(A_k^{\dagger})^{1/2}\hat{P}_k\hat{x})^T A_k^{1/2}(\nu-\theta_k)|\\
&\le \|(A_k^{\dagger})^{1/2}\hat{P}_k\hat{x}\|_2\|A_k^{1/2}(\nu-\theta_k)\|_2 \quad \text{by C.S.}\\
&\le \beta_{k,\delta}\|\hat{P}_k\hat{x}\|_{A_k^{\dagger}} \qquad \text{since } \nu\in C_k.\\
&= \beta_{k,\delta}\|\hat{V}_k^T\hat{x}\|_{B_k^{-1}} = \beta_{k,\delta}\|\hat{x}\|_{A_k^{\dagger}}
\end{aligned}$$

□

Proof of Lemma 2.2.10:

$$\begin{aligned}
\lambda_m(\hat{P}_t\hat{\Sigma}_{t-1}\hat{P}_t) &= \lambda_m\big((\hat{P}_t-P)\hat{\Sigma}_{t-1}\hat{P}_t + P\hat{\Sigma}_{t-1}(\hat{P}_t-P) + P\hat{\Sigma}_{t-1}P\big)\\
&\ge \lambda_m(P\hat{\Sigma}_{t-1}P) - 2(t-1)L^2\|\hat{P}_t-P\|_2\\
&\ge \lambda_{\min}(V^T\hat{\Sigma}_{t-1}V) - 4L^2\Gamma\sqrt{\frac{\alpha(t-1)}{K}\log\frac{2d}{\delta}} \quad \text{from Lemma 2.2.4.}
\end{aligned}$$

We also have that $\lambda_{\max}(V^T\hat{X}_j\hat{X}_j^TV) \le L, \forall j, \lambda_{\min}\left(\mathbb{E}\big[\sum_{j=1}^{t-1}V^T\hat{X}_j\hat{X}_j^TV\big]\right) = (t-1)(\lambda_- + \sigma^2)$. Applying Matrix Chernoff Inequality,

$$\mathbb{P}\left[\lambda_{\min}(V^T\hat{\Sigma}_tV) \le (t-1)(\lambda_-+\sigma^2) - \sqrt{2L(t-1)(\lambda_-+\sigma^2)\log\frac{m}{\delta}}\right] \le \delta.$$

Combining these with similar stopping time construction as described in previous sections we derive the first statement of lemma. Now for second statement with a constant C, observe that, $(t-1)(\lambda_- + \sigma^2) - \sqrt{t-1}\left(4L^2\Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}} + \sqrt{2L(\lambda_-+\sigma^2)\log\frac{m}{\delta}}\right) \ge C(t-1)$ holds if and only if

$$t \ge 1 + \left(\frac{4L^2\Gamma\sqrt{\frac{\alpha}{K}\log\frac{2d}{\delta}} + \sqrt{2L(\lambda_-+\sigma^2)\log\frac{m}{\delta}}}{\lambda_-+\sigma^2-C}\right)^2.$$

Choosing $C = \frac{\lambda_-+\sigma^2}{2m}$ proves the bound. □

Finally, we state the lemma which is used in (2.23) of Theorem 2.2.11.

Lemma A.1.5.

$$2\sqrt{t+1}-2 \le \sum_{i=1}^{t}\frac{1}{\sqrt{i}} \le 2\sqrt{t}-1 \qquad\qquad \log(t+1) \le \sum_{i=1}^{t}\frac{1}{i} \le 1+\log(t)$$

Proof. First, one can be obtained using integral estimates and the second one is due to harmonic sums. □

A.2 Proofs of Section 2.3

A.2.1 Proofs for Safe-OFUL - Section 2.3.2-2.3.3

Proof of Lemma 2.3.4

If $x \in D_t^{\text{safe}}$, we have following two cases:

Case 1 : $x \in D^w \rightarrow$ Trivially $x \in D_0^{\text{safe}}$

Case 2 : $x \in \Gamma_i$, then by definition

$$\begin{aligned}\tau_i - \tau_i^s &\geq \hat{\gamma}_{i,t}^\top (x - x_i^s) + \beta_t^i ||x - x_i^s||_{A_{i,t}^{-1}} \\ &= \gamma_i^\top (x - x_i^s) + (\hat{\gamma}_{i,t} - \gamma_i)^\top (x - x_i^s) + \beta_t^i ||x - x_i^s||_{A_{i,t}^{-1}} \\ &\geq \gamma_i^\top (x - x_i^s) \quad \text{(Conditioned on } \mathcal{E}_\mu \text{ and Cauchy Schwarz Inequality(CSI))}.\end{aligned}$$

Therefore $x \in D_0^{\text{safe}}$. □

Proof of Theorem 2.3.6

Recall $\mathbf{ucb}(x, i, t) = \hat{\mu}_t^\top x + \beta_t ||x||_{V_t^{-1}} + k_i \beta_t^i ||x - x_i^s||_{A_{i,t}^{-1}}$. We consider following cases:

Case 1 : $x^* \in D_t^{\text{safe}}$

$$\begin{aligned}\max_{i \in M, x \in \hat{\Gamma}_{i,t}} \mathbf{ucb}(x, i, t) &\geq \mathbf{ucb}(x^*, i^*, t) \\ &\geq \hat{\mu}^\top x^* + \beta_t ||x^*||_{V_t^{-1}} \quad (\text{Since } k_i \beta_t^i ||x - x_i^s||_{A_{i,t}^{-1}} \geq 0) \\ &= < \mu, x^* > + < \hat{\mu}_t - \mu, x^* > + \beta_t ||x^*||_{V_t^{-1}} \\ &\geq < \mu, x^* > + (1 - 1)\beta_t ||x^*||_{V_t^{-1}} \quad \text{(Conditioned on } \mathcal{E}_\mu \text{ \& CSI)} \\ &\geq \mu^\top x^*.\end{aligned}$$

Case 2 : $x^* \notin D_t^{\text{safe}}$

We consider the constraint set Γ_{i^*} in which x^* belongs to and define

$$\alpha_t = \max\{\alpha \in [0, 1] : \alpha \hat{\gamma}_{i^* t}^\top (x^* - x_{i^*}^s) + \alpha \beta_t^{i^*} ||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}} = \tau - \tau_{i^*}^s\}$$

This definition ensures $z_t = \alpha_t x^* + (1 - \alpha_t) x_{i^*}^s \in D_t^{\text{safe}}$. Now we have

$$\begin{aligned}\max_{i \in M, x \in \hat{\Gamma}_{i,t}} \mathbf{ucb}(x, i, t) &\geq \mu^\top z_t + (\hat{\mu}_t - \mu)^\top z_t + \beta_t ||z_t||_{V_t^{-1}} + k_{i^*} \beta_t^{i^*} ||z_t - x_{i^*}^s||_{A_{i^*,t}^{-1}} \\ &\geq \mu^\top z_t + k_{i^*} \beta_t^{i^*} ||z_t - x_{i^*}^s||_{A_{i^*,t}^{-1}} \quad \text{(Conditioned on } \mathcal{E}_\mu \text{ \& CSI)} \\ &\geq \alpha_t \mu^\top (x^* - x_{i^*}^s) + \mu^\top x_{i^*}^s + \alpha_t k_{i^*} \beta_t^{i^*} ||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}} \\ &\geq \alpha_t [\mu^\top (x^* - x_{i^*}^s) + k_{i^*} \beta_t^{i^*} ||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}}] + \mu^\top x_{i^*}^s.\end{aligned}$$

Define $B = \hat{\gamma}_{i^*t}^\top(x^* - x_{i^*}^s) + \beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}}$, by assumption $x^* \notin D_t^{\text{safe}}$ we have $B \geq \tau - \tau_{i^*}^s$

By definition, we have $\alpha_t B = \tau - \tau_{i^*}^s$. To lower bound α_t we first upper bound B

$$\begin{aligned}
B &= \hat{\gamma}_{i^*t}^\top(x^* - x_{i^*}^s) + \beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}} \\
&= \gamma_{i^*}^\top(x^* - x_{i^*}^s) + (\hat{\gamma}_{i^*t} - \gamma_{i^*})^\top(x^* - x_{i^*}^s) + \beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}} \\
&\leq \gamma_{i^*}^\top(x^* - x_{i^*}^s) + 2\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}} \quad \text{(Conditioned on } \mathcal{E}_\mu \text{ and CSI)} \\
&\leq \tau - \tau_{i^*}^s + 2\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}}.
\end{aligned}$$

Therefore, we have

$$\alpha_t \geq \frac{\tau - \tau_{i^*}^s}{\tau - \tau_{i^*}^s + 2\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}}}$$

If we choose k_i such that optimism is ensured for this lower bound, overall optimism is guaranteed. In particular, we need to satisfy the following

$$\begin{aligned}
&\max_{i \in M, x \in \hat{\Gamma}_{i,t}} \mathbf{ucb}(x, i, t) \geq \alpha_t[\mu^\top(x^* - x_{i^*}^s) + k_{i^*}\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}}] + \mu^\top x_{i^*}^s \\
&\geq \frac{\tau - \tau_{i^*}^s}{\tau - \tau_{i^*}^s + 2\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}}}[\mu^\top(x^* - x_{i^*}^s) + k_{i^*}\beta_t^{i^*}||x^* - x_{i^*}^s||_{A_{i^*,t}^{-1}}] + \mu^\top x_{i^*}^s \geq \mu^\top x^*.
\end{aligned}$$

Solving this gives the condition of $k_{i^*} \geq \frac{2LS}{\tau - \tau_{i^*}^s}$, which proves the theorem. □

Proof of Theorem 2.3.7

$$\begin{aligned}
R_T &= \sum_{t=1}^{T} \delta_t = \sum_{t=1}^{\top}(\mu^\top x^* - \mu^\top x_t) \\
&\leq \sum_{t=1}^{T}(\mathbf{ucb}(x_t, i_t, t) - \mu^\top x_t) \wedge 2 \qquad \text{(A.14)} \\
&= \sum_{t=1}^{T}(< \hat{\mu}_t - \mu, x_t > + \beta_t||x_t||_{V_t^{-1}} + k_{i_t}\beta_t^{i_t}||x - x_{i_t}^s||_{A_{i_t,t}^{-1}}) \wedge 2 \\
&\leq \sum_{t=1}^{T}(2\beta_t||x_t||_{V_t^{-1}} + k_{i_{max}}\beta_t^{i_t}||x - x_{i_t}^s||_{A_{i_t,t}^{-1}}) \wedge 2 \qquad \text{(A.15)} \\
&\leq \sum_{t=1}^{T}(2\beta_t||x_t||_{V_t^{-1}}) \wedge 2 + \sum_{t=1}^{T}(k_{i_{max}}\beta_t^{i_t}||x - x_{i_t}^s||_{A_{i_t,t}^{-1}}) \wedge 2.
\end{aligned}$$

Here (A.14) follows from optimism and (A.15) follows from Cauchy Schwarz Inequality conditioned on $\mathcal{E}_\mu$. Next, we analyze these self-normalized summations using the standard technique in [3]:

$$\sum_{t=1}^{\top} ||x_t||_{V_t^{-1}} \leq \sqrt{T \sum_{t=1}^{\top} ||x_t||^2_{V_t^{-1}}} \quad \text{(CSI)}$$

$$\leq \sqrt{2T \log(\frac{\det(A_T)}{\det(A_1)})} \tag{A.16}$$

$$\leq \sqrt{2Td \log(\frac{d\lambda + TL^2}{d\lambda})}. \tag{A.17}$$

In Inequality (A.16), we used the standard argument in regret analysis of linear bandits [3] (Lemma 11) as follows:

$$\sum_{t=1}^{n} \min\left(\|\mathbf{y}_t\|^2_{\mathbf{V}_t^{-1}}, 1\right) \leq 2 \log \frac{\det \mathbf{V}_{n+1}}{\det \mathbf{V}_1} \quad \text{where} \quad \mathbf{V}_n = \mathbf{V}_1 + \sum_{t=1}^{n-1} \mathbf{y}_t \mathbf{y}_t^\top.$$

In inequality (A.17), we used Assumption 2.3.2 and the fact that $\det(\mathbf{A}) = \prod_{i=1}^{d} \lambda_i(\mathbf{A}) \leq (\text{trace}(\mathbf{A})/d)^d$. Combining all these, we have with probability at least $1 - 2\delta$

$$R_T \leq 2\beta_T \sqrt{2Td \log(\frac{d\lambda + TL^2}{d\lambda})} + (k_{i_{max}} \beta_T^{i_{max}} + 2) \Sigma_{i \in M} \sqrt{2N_i(T) d \log(\frac{d\lambda + N_i(T)L^2}{d\lambda})}$$

$$\leq 2\beta_T \sqrt{2Td \log(\frac{d\lambda + TL^2}{d\lambda})} + (k_{i_{max}} \beta_T^{i_{max}} + 2) \sqrt{2|M|Td \log(\frac{d\lambda + TL^2}{d\lambda})}.$$

The last step follows from AM-QM inequality. □

A.2.2 Proofs of Section 2.3.4: Regret guarantee Safe-LinTS (Theorem 2.3.8)

The proof consists of two pieces. We will first show that Safe-LinTS selects optimistic actions with non-zero probability and then we use the regret decomposition in [5] to give the regret upper bound.

Optimism

Recall the following expressions:

$$\tilde{\mu}_t = \hat{\mu}_t + \beta_t (V_t)^{-\frac{1}{2}} \eta_t, \qquad \tilde{\omega}_t^i = \beta_t^i (A_{i,t})^{-\frac{1}{2}} \eta_t^c.$$

Moreover, we have the following concentration and anti-concentration properties:

$$\mathbb{P}(u^\top \eta_t \geq 1) = p_1, \qquad \mathbb{P}(u^\top \eta_t^c \geq \frac{2}{\tau - \tau_*^s} L S^\gamma) = p_2,$$

$$\mathbb{P}(\|\eta_t\|_2 \le \sqrt{cd\log(\frac{c'd}{\delta})}) \ge 1-\frac{\delta}{2}, \qquad \mathbb{P}(\|\eta_t^c\|_2 \le \frac{2LS^\gamma}{\tau-\tau_*^s}\sqrt{cd\log(\frac{c'd}{\delta})}) \ge 1-\frac{\delta}{2}.$$

For $\delta \in (0,1)$, $\delta' = \frac{\delta}{6T}$, we define the following high-probability events:

- $\mathcal{E}_{\mu,t}$ is the event that $\hat{\mu}_t$ concentrates around μ for all $s \le t$: $\mathcal{E}_{\mu,t} = \{\forall s \le t : \|\hat{\mu}_s - \mu\|_{V_s} \le \beta_s(\delta')\}$, then $\mathbb{P}(\mathcal{E}_{\mu,T}) \ge 1 - \frac{\delta}{6}$.
- $\mathcal{E}_{\gamma,t}$ is the event that $\hat{\gamma}_{i,t}$ concentrates around γ_i for all $s \le t$ and for all $i \in \mathbf{M}$: $\mathcal{E}_{\gamma,t} = \{\forall s \le t, \forall i \in M : \|\hat{\gamma}_{i,s} - \gamma_i\|_{A_{i,s}} \le \beta_s^i(\frac{\delta'}{|M|})\}$, then $\mathbb{P}(\mathcal{E}_{\gamma,T}) \ge 1 - \frac{\delta}{6}$.
- $\tilde{\mathcal{E}}$ be the event that such the sampled η_t and η_t^c are bounded for all $t \le T$: $\tilde{\mathcal{E}} = \{\forall t \le T, \|\eta_t\|_2 \le \sqrt{cd\log(\frac{12Tc'd}{\delta})}\} \cap \{\forall t \le T, \|\eta_t^c\|_2 \le \sqrt{cd\log(\frac{12Tc'd}{\delta})}\}$, then $\mathbb{P}(\tilde{\mathcal{E}}) \ge 1 - \frac{\delta}{6}$.
- Let $Z_t = \mathcal{E}_{\gamma,t} \cap \mathcal{E}_{\mu,t}$, then $\mathbb{P}(Z_T) \ge 1 - \frac{\delta}{3}$.
- Let $E_t = \tilde{\mathcal{E}} \cap \mathcal{E}_{\gamma,t} \cap \mathcal{E}_{\mu,t}$, then $\mathbb{P}(E_T) \ge 1 - \frac{\delta}{2}$.

Recall that $\hat{\Gamma}_{i,t} = \{x \in \Gamma_i : \hat{\gamma}_{i,t}^\top(x - x_i^s) + \beta_t^i ||x - x_i^s||_{A_{i,t}^{-1}} \le \tau - \tau_i^s\}$. Let

$$\alpha_t^* := \max\{\alpha \in [0,1] : z_t = \alpha x^* + (1-\alpha)x_{i^*}^s \in \hat{\Gamma}_{i^*,t}\},$$

then we can show that there exists $\alpha_t \le \alpha_t^*$ such that

$$\alpha_t \gamma_{i^*}^\top(x^* - x_i^s) + 2\alpha_t\beta_t^i ||x - x_i^s||_{A_t^{i-1}} = \tau - \tau_i^s.$$

Rearranging, we get $\frac{1}{\alpha_t} = 1 + \frac{2}{\tau - \tau_{i^*}^s}\beta_t^{i^*}\|x^* - x_{i^*}^s\|_{A_t^{i^*-1}}$. The goal is to show that playing the safe action $z_t = \alpha_t x^* + (1-\alpha_t)x_{i^*}^s$ is optimistic with constant probability. Define

$$J_t(\eta, \eta^c, i, x) = \tilde{\mu}_t^\top x + \tilde{\omega}_{i,t}^\top(x - x_i^s),\ J_t(\eta, \eta^c, i) = \max_{x \in \hat{\Gamma}_{i,t}} J_t(\eta, \eta^c, i, x),\ J_t(\eta, \eta^c) = \max_{i \in M} J_t(\eta, \eta^c, i),$$

and we analyze the probability with which the sampled parameters are optimistic, i.e., $J_t(\eta_t, \eta_t^c) \ge \mu^\top x^*$. Let $p_t = \mathbb{P}(J_t(\eta_t, \eta_t^c) \ge \mu^\top x^* | \mathcal{F}_t, Z_t)$. Then,

$$\begin{aligned}
p_t &= \mathbb{P}(J_t(\eta_t, \eta_t^c) \ge \mu^\top x^* | \mathcal{F}_t, Z_t) \\
&\ge \mathbb{P}(J_t(\eta_t, \eta_t^c, i^*, \alpha_t x^* + (1-\alpha_t)x_{i^*}^s) \ge \mu^\top x^* | \mathcal{F}_t, Z_t) \\
&= \mathbb{P}(\tilde{\mu}_t^\top z_t + \alpha_t \tilde{\omega}_{i^*,t}^\top(x^* - x_{i^*}^s) \ge \mu^\top x_{i^*}^s + \mu^\top(x^* - x_{i^*}^s) | \mathcal{F}_t, Z_t),
\end{aligned}$$

where $z_t = \alpha x^* + (1-\alpha)x_{i^*}^s$. Consider:

$$\tilde{\mu}_t^\top z_t = \hat{\mu}_t^\top z_t + z_t^\top \beta_t(V_t)^{-\frac{1}{2}}\eta_t \ge \mu^\top z_t - \beta_t\|z_t\|_{V_t^{-1}} + z_t^\top \beta_t(V_t)^{-\frac{1}{2}}\eta_t.$$

By construction of η_t we have:

$$\mathbb{P}(z_t^\top \beta_t(V_t)^{-\frac{1}{2}}\eta_t \geq \beta_t \|z_t\|_{V_t^{-1}} | \mathcal{F}_t, Z_t) = \mathbb{P}(u^\top \eta_t \geq 1) = p_1.$$

Using the fact that η_t, η_t^c are independent and then substituting in α_t we get

$$\begin{aligned}
p_t &\geq p_1 \mathbb{P}(\mu^\top z_t + \alpha_t \tilde{\omega}_{i^*,t}^\top (x^* - x_{i^*}^s) \geq \mu^\top x_{i^*}^s + \mu^\top (x^* - x_{i^*}^s) | \mathcal{F}_t, Z_t) \\
&\geq p_1 \mathbb{P}(\mu^\top x_{i^*}^s + \alpha_t \mu^\top (x^* - x_{i^*}^s) + \alpha_t \tilde{\omega}_{i^*,t}^\top (x^* - x_{i^*}^s) \geq \mu^\top x_{i^*}^s + \mu^\top (x^* - x_{i^*}^s) | \mathcal{F}_t, Z_t) \\
&= p_1 \mathbb{P}(\tilde{\omega}_{i^*,t}^\top (x^* - x_{i^*}^s) \geq \frac{1-\alpha_t}{\alpha_t} \mu^\top (x^* - x_{i^*}^s) | \mathcal{F}_t, Z_t) \\
&\geq p_1 \mathbb{P}(\tilde{\omega}_{i^*,t}^\top (x^* - x_{i^*}^s) \geq \frac{2LS}{\tau - \tau_{i^*}} \beta_t^{i^*} \|x^* - x_{i^*}^s\|_{A_{i^*,t}^{-1}} | \mathcal{F}_t, Z_t) \\
&= p_1 \mathbb{P}(\beta_t^{i^*} (A_t^{i^*})^{-\frac{1}{2}} \eta_t^{c\top} (x^* - x_{i^*}^s) \geq \frac{2LS}{\tau - \tau_{i^*}} \beta_t^{i^*} \|x^* - x_{i^*}^s\|_{A_{i^*,t}^{-1}} | \mathcal{F}_t, Z_t) \\
&= p_1 \mathbb{P}(u^\top \eta_t^c \geq \frac{2}{\tau - \tau_i} LS) \geq p_1 p_2.
\end{aligned}$$

Next, we need to show that conditioned on E_T, the algorithm is still optimistic. This is because the chosen confidence bound $\delta' = \frac{\delta}{6T}$ is small enough compared to the anti-concentration property. Moreover, we assume that $T \geq \frac{1}{3p_1p_2}$ which implies that $\delta' \leq \frac{p_1p_2}{2}$. We know that for any events A and B, we have $\mathbb{P}(A \cap B) = 1 - \mathbb{P}(A^c \cup B^c) \geq \mathbb{P}(A) - \mathbb{P}(B^c)$. Choosing $A = J_t(\eta_t, \eta_t^c) \geq \mu^\top x^*$ and $B = E_T$ we get

$$\mathbb{P}(J_t(\eta_t, \eta_t^c) \geq \mu^\top x^* | \mathcal{F}_t, Z_t) \geq p_1p_2 - \delta' \geq \frac{p_1p_2}{2}.$$

Regret

We can decompose the cumulative regret as follows:

$$R(T) = \sum_{t=1}^T \big(\underbrace{x^{*\top}\mu - J_t(\eta_t, \eta_t^c)}_{R_t^{TS}} \big) + \sum_{t=1}^T \big(\underbrace{J_t(\eta_t, \eta_t^c) - x_t^\top \mu}_{R_t^{RLS}} \big).$$

First, we consider

$$\begin{aligned}
R_t^{TS} &= x^{*T}\mu - J_t(\eta_t, \eta_t^c) \\
&\leq \mathbb{E}[J_t(\eta, \eta^c) - J_t(\eta_t, \eta_t^c) | (\eta, \eta^c) \in \Theta] \\
&\leq \mathbb{E}[J_t(\eta, \eta^c, i, x) - J_t(\eta_t, \eta_t^c, i, x) | (\eta, \eta^c, i, x) \in \Theta] \\
&\leq \mathbb{E}[(\tilde{\mu} - \tilde{\mu}_t)^\top x + (\tilde{\omega}^i - \tilde{\omega}_t^i)^\top (x - x_i^s) | (\eta, \eta^c, i, x) \in \Theta] \\
&\leq \mathbb{E}[\|\tilde{\mu} - \tilde{\mu}_t\|_{A_t} \|x\|_{V_t^{-1}} + \|\tilde{\omega}^i - \tilde{\omega}_t^i\|_{A_t^i} \|x - x_i^s\|_{A_{i,t}^{-1}} | (\eta, \eta^c, i, x) \in \Theta] \\
&\leq 2\sigma_t(\delta) \mathbb{E}[\|x\|_{V_t^{-1}} | (\eta, \eta^c, i, x) \in \Theta] + 2\sigma_t^i(\delta) \mathbb{E}[\|x - x_i^s\|_{A_{i,t}^{-1}} | (\eta, \eta^c, i, x) \in \Theta] \\
&\leq \frac{4}{p_1p_2} \{\sigma_t(\delta) \mathbb{E}[\|x\|_{V_t^{-1}}] + \sigma_t^i(\delta) \mathbb{E}[\|x - x_i^s\|_{A_{i,t}^{-1}}]\},
\end{aligned}$$

where $\sigma_t(\delta) = \beta_t(\delta)\sqrt{cd\log(\frac{c'd}{\delta})}$ and $\sigma_t^i(\delta) = \beta_t^i(\delta)\frac{2L^cS^c}{\tau-\tau_i}\sqrt{cd\log(\frac{c'd}{\delta})}$. Here $(\eta, \eta^c, i, x) \in \Theta$ denotes optimistic parameters. Next, consider the sum

$$\Sigma_{t=1}^{T}\mathbb{E}[\|x\|_{V_t^{-1}}] = \sum_{t=1}^{T}\|x\|_{V_t^{-1}} + \sum_{t=1}^{T}(\mathbb{E}[\|x\|_{V_t^{-1}}] - \|x\|_{V_t^{-1}}).$$

The second summation is a martingale sum, so we use Azuma's Inequality to get

$$\sum_{t=1}^{T}(\mathbb{E}[\|x\|_{V_t^{-1}}] - \|x\|_{V_t^{-1}}) \leq \sqrt{\frac{8TL^2}{\lambda}\log\frac{4}{\delta}},$$

with probability $1 - \frac{\delta}{2}$, since we have $\|x_t\|_2 \leq L$ and $V_t^{-1} \leq \frac{1}{\lambda}I$, which gives $\mathbb{E}[\|x\|_{A_s^{-1}}] - \|x\|_{A_s^{-1}} \leq \frac{2L}{\sqrt{\lambda}}$.

Now using standard analysis from previous sections and previous inequality we get

$$\begin{aligned} R^{TS}(T) \leq \frac{4\sigma_t(\delta)}{p_1p_2}\left(\sqrt{2dT\log(1+\frac{TL^2}{\lambda})} + \sqrt{\frac{8TL^2}{\lambda}\log\frac{8}{\delta}}\right) \\ + \frac{4\sigma_t^{i_{max}}(\delta)+2}{p_1p_2}\left(\sqrt{2d|M|T\log(1+\frac{TL^2}{\lambda})} + \sqrt{\frac{8TL^2}{\lambda}\log\frac{8}{\delta}}\right). \end{aligned}$$

Next, we consider

$$\begin{aligned} R_t^{RLS} &= J_t(\eta_t, \eta_t^c, i_t, x_t) - \mu^\top x_t \\ &= \tilde{\mu}_t^\top x + \tilde{\omega}_t^{iT}(x - x_i^s) - \mu^\top x_t \\ &\leq (\hat{\mu}_t - \mu)^\top x_t + \beta_t A_t^{-\frac{1}{2}}\eta_t^\top x_t + \beta_t^{i_t}A_t^{i_t-\frac{1}{2}}\eta_t^{cT}(x_t - x_{i_t}^s) \\ &\leq \beta_t\|x_t\|_{V_t^{-1}} + \sigma_t\|x_t\|_{V_t^{-1}} + \sigma_t^{i_t}\|x_t - x_{i_t}^s\|_{A_t^{i_t-1}}. \end{aligned}$$

So $R^{RLS}(T) \leq (\beta_T + \sigma_T)\sqrt{2dT\log(1+\frac{TL^2}{\lambda})} + \sigma_T^{i_{max}}\sqrt{2d|M|T\log(1+\frac{TL^2}{\lambda})}$. Combining these gives the advertised result in the theorem

$$R(T) \leq \tilde{O}(d^{3/2}\sqrt{|M|T}). \tag{A.18}$$

□

A.2.3 Proofs of Section 2.3.5: Regret Guarantee of Safe-OFUL/LinTS with Pure Exploration (Theorem 2.3.11)

First, we prove the following helper lemma.

Lemma A.2.1. *If we consider a δ_f-ball around a point, the approximation error for the first order Taylor expansion for a ζ-smooth function is bounded as*

$$|f(x) - f(a) - \nabla f(a)^\top (x - a)| \leq \frac{\zeta \delta^2}{2}.$$

Then, the least squares parameter of this approximation error is bounded as

$$\|\hat{\epsilon}_{i,T}\|_2 \leq 2\zeta\delta_r\sqrt{2d\log T} = O(\zeta\delta_r\sqrt{d\log T}).$$

Proof. Recall $\hat{\epsilon}_{it} = A_{i,t}^{-1} \sum_{\tau=1}^{N_i(t)} (x_\tau - x_i^s)\epsilon_i(x_\tau)$, where where $\epsilon_i(x_t) = f(x_t) - (f(x_i^s) + \nabla f(x_i^s)^\top (x_t - x_i^s)$. Define $Y_{\epsilon,t}$ as the column vector enumerating the approximation errors $y_{\epsilon,\tau} = \epsilon_i(x_\tau)$ for $0 < \tau \leq t$, and X_t corresponds to the matrix enumerating the shifted actions $x_\tau - x_i^s$ for $0 < \tau \leq t$. By definition we have

$$\hat{\epsilon}_{i,t} = \operatorname*{argmin}_\theta \|Y_{\epsilon,t} - X_t^\top \theta\|_2^2.$$

Next, define

$$T_1(\theta) := \|Y_{\epsilon,t} - X_t^\top \theta\|_2^2 = (Y_{\epsilon,t}^\top Y_{\epsilon,t} - 2Y_{\epsilon,t}^\top X_t^\top \theta + \theta^\top X_t X_t^\top \theta),$$

and

$$T_2(\theta) := \|y_{\epsilon,t+1} - x_{t+1}^\top \theta\|_2^2 = (y_{\epsilon,t+1}^2 - 2y_{\epsilon,t+1}x_{t+1}^\top \theta + \theta^\top x_{t+1}x_{t+1}^\top \theta).$$

Now, consider $\hat{\epsilon}_{i,t+1}$

$$\begin{aligned}\hat{\epsilon}_{i,t+1} &= arg\min_\theta T_1(\theta) + T_2(\theta) \\ &= arg\min_\theta (Y_{\epsilon,t}^\top Y_{\epsilon,t} - 2Y_{\epsilon,t}^\top X_t^\top \theta + \theta^\top X_t X_t^\top \theta) + (y_{\epsilon,t+1}^2 - 2y_{\epsilon,t+1}x_{t+1}^\top \theta + \theta^\top x_{t+1}x_{t+1}^\top \theta).\end{aligned}$$

For the minimizer, we have $\nabla T_1 + \nabla T_2 = 0$, $\nabla T_1 = 2X_t X_t^\top \theta - 2X_t Y_{\epsilon,t}$ and $\nabla T_2 = 2x_{t+1}x_{t+1}^\top \theta - 2x_{t+1}y_{\epsilon,t+1}$. If we re-parameterise $\theta = \hat{\epsilon}_{i,t} + w$, then at minima we get

$$w = (X_t X_t^\top + x_{t+1}x_{t+1}^\top)^{-1} x_{t+1}(y_{\epsilon,t+1} - x_{t+1}^\top \hat{\epsilon}_{i,t}).$$

Recall that in pure exploration we pick action such that $x_t = \max_{x \in \tilde{D}_i^w} \|x - x_i^s\|_{A_{i,t}^{-1}}$. As consequence, we pick orthogonal vectors in subsequent turns, which ensures that x_t is always an eigenvector, which implies

$$w = \frac{x_{t+1}}{\lambda_{t+1}}(y_{\epsilon,t+1} - x_{t+1}^\top \hat{\epsilon}_{i,t}).$$

Note that there is no rotation as the exploration strategy ensures that x_{t+1} are eigenvectors of $X_t X_t^\top + x_{t+1}x_{t+1}^\top$ with the eigenvalues of $\lambda_{t+1} \geq \delta_r^2$. Next, we upper bound

the magnitude difference at each step. By definition, we have $\|\hat{\epsilon}_{i,t+1}\|^2 = \|\hat{\epsilon}_{i,t} + w\|^2$. Substituting w and rearranging we get

$$\|\hat{\epsilon}_{i,t+1}\|_2^2 - \|\hat{\epsilon}_{i,t}\|_2^2 = \frac{\|x_{t+1}\|^2}{\lambda_{t+1}^2}(y_{\epsilon,t+1} - x_{t+1}^\top \hat{\epsilon}_{i,t})^2 + \frac{2x_{t+1}^\top \hat{\epsilon}_{i,t}}{\lambda_{t+1}}(y_{\epsilon,t+1} - x_{t+1}^\top \theta_{t+1}). \quad \text{(A.19)}$$

Now re-parameterizing: $\|\hat{\epsilon}_{i,t}\| = B$, $\|x_{t+1}\| = \delta_r$, $\cos\alpha = \frac{<x_{t+1}, \hat{\epsilon}_{i,t}>}{B\delta}$ and $L = \|\hat{\epsilon}_{i,t+1}\|_2^2 - \|\hat{\epsilon}_{i,t}\|_2^2$, we obtain

$$L = (\frac{B^2\delta^4}{\lambda_{t+1}^2} - \frac{2B^2\delta_r^2}{\lambda_{t+1}})\cos^2(\alpha) + (\frac{2y_{\epsilon,t+1}B\delta_r}{\lambda_{t+1}} - \frac{2y_{\epsilon,t+1}B\delta_r^3}{\lambda_{t+1}^2})\cos(\alpha) + \frac{\delta_r^2 y_{\epsilon,t+1}^2}{\lambda_{t+1}^2}.$$

To upper-bound L, we maximize L over α, which gives us the following condition

$$\frac{B\delta_r}{\lambda_{t+1}}\sin(\alpha)[\frac{2\lambda_{t+1} - \delta_r^2}{\lambda_{t+1}}B\delta_r\cos(\alpha) - \frac{\lambda_{t+1} - \delta_r^2}{\lambda_{t+1}}y_{\epsilon,t+1}] = 0.$$

Case 1 : $(\sin\alpha = 0)$

So $\cos(\alpha) = \pm 1$, which implies the increment w is along the direction of $\hat{\epsilon}_{i,t}$. Recall that $w = \frac{x_{t+1}}{\lambda_{t+1}}(y_{\epsilon,t+1} - B\delta_r\cos(\alpha))$. Since $\cos(\alpha) = \pm 1$, we get the following equality:

$$|\hat{\epsilon}_{i,t+1}| = |\hat{\epsilon}_{i,t}| + \frac{\delta_r\cos(\alpha)}{\lambda_{t+1}}(y_{\epsilon,t+1} - B\delta_r\cos(\alpha)).$$

If $B \geq \frac{\zeta\delta_r}{2}$, then from the smoothness assumption, we have $y_{\epsilon,t+1} \leq \frac{\zeta\delta_r^2}{2}$, which gives $|\hat{\epsilon}_{i,t+1}| - |\hat{\epsilon}_{i,t}| \leq 0$. If $B < \frac{\zeta\delta_r}{2}$, we obtain the following bound $|\hat{\epsilon}_{i,t+1}| \leq \frac{\zeta\delta_r}{2} + \frac{\zeta\delta_r^3}{\lambda_{t+1}}$. Recall that $\lambda_{t+1} > \delta_r^2$ by construction. So

$$|\hat{\epsilon}_{i,t+1}| \leq \frac{3\zeta\delta_r}{2} \leq 2\zeta\delta_r.$$

Case 2 : $(B\delta_r\cos(\alpha) = y_{\epsilon,t+1}\frac{\lambda_{t+1} - \delta_r^2}{2\lambda_{t+1} - \delta_r^2})$

Substituting, we get

$$L^* = \frac{\delta_r^2 y_{\epsilon,t+1}^2}{\lambda_t^2}(1 - \frac{\lambda_{t+1} - \delta_r^2}{2\lambda_{t+1} - \delta_r^2})^2 + \frac{2y_{\epsilon,t+1}^2}{\lambda_{t+1}}(1 - \frac{\lambda_{t+1} - \delta_r^2}{2\lambda_{t+1} - \delta_r^2})(\frac{\lambda_{t+1} - \delta_r^2}{2\lambda_{t+1} - \delta_r^2})$$

which simplifies to

$$L^* = \frac{y_{\epsilon,t+1}^2}{2\lambda_{t+1} - \delta_r^2} \leq \frac{y_{\epsilon,t+1}^2}{\lambda_{t+1}}$$

since $\lambda_{t+1} \geq \delta_r^2$.

Taking both cases into consideration and adding the telescopic series (A.19), we get

$$\begin{aligned}\|\hat{\epsilon}_{i,T}\|_2^2 \leq 4\zeta^2\delta_r^2 + \Sigma_{s=1}^{T}(\|\theta_s\|_2^2 - \|\theta_{s-1}\|_2^2) &\leq 4\zeta^2\delta_r^2 + \Sigma_{s=1}^{T}\frac{y_{\epsilon,s}^2}{\lambda_s} \\ &\leq 4\zeta^2\delta_r^2 + \Sigma_{s=1}^{\frac{T}{d}}\frac{y_{\epsilon,s}^2}{\delta_r^2}\frac{d}{s}, \qquad \text{(A.20)} \\ &\leq 4\zeta^2\delta_r^2 + \frac{d}{4}\zeta^2\delta_r^2\log T, \qquad \text{(A.21)}\end{aligned}$$

where (A.20), comes because we need to pick d orthogonal vectors before all the eigenvalues become equal again, and (A.21) is a standard bound on harmonic sum and because $y_{i,t}$ is bounded by $\frac{\zeta\delta_r^2}{2}$ by smoothness assumption. Therefore, we obtain the advertised bound: $\|\hat{\epsilon}_{i,T}\|_2 \leq 2\zeta\delta_r\sqrt{2d\log T} = O(\zeta\delta_r\sqrt{d\log T})$ □

Proof of Theorem 2.3.11

Recall $\Gamma_t^i = \{x \in \Gamma_i : \nabla\hat{f}_{it}^{\top}(x - x_i^s) + \frac{\Delta}{2} \leq \tau - \tau_i^s\}$. There exists $T'(\Delta)$ such that

$$\Delta \geq 2\zeta\delta_f^2\sqrt{2d\log T'(\Delta)} + S\beta_T\sqrt{\frac{d}{T'(\Delta)}},$$

where $T'(\Delta)$ is defined as $T'(x) = \min\{t > 0 : \Delta \geq 2\zeta\delta_f^2\sqrt{2d\log t} + S\beta_T\sqrt{\frac{d}{t}}\}$. Using Lemma A.2.1, we can upper bound $\nabla\hat{f}_{i^*t}^{\top}(x^* - x_{i^*}^s)$ as follows

$$\begin{aligned}\nabla\hat{f}_{i^*t}^{\top}(x^* - x_{i^*}^s) &= \nabla\hat{f}_{i^*}^{LS\top}(x^* - x_{i^*}^s) + \hat{\epsilon}_{it}^{\top}(x^* - x_{i^*}^s) \\ &\leq \nabla f_{i^*}^{\top}(x^* - x_{i^*}^s) + \beta_t^{i^*}||x - x_{i^*}^s||_{A_{i^*,t}^{-1}} + \hat{\epsilon}_{it}^{\top}(x^* - x_{i^*}^s) \\ &\leq \nabla f_{i^*}^{\top}(x^* - x_{i^*}^s) + \beta_t^{i^*}||x - x_{i^*}^s||_{A_{i^*,t}^{-1}} + 2\zeta\delta_r\delta_f\sqrt{2d\log T} \\ &\leq f_{i^*}(x^*) - f_{i^*}(x_{i^*}^s) + \frac{\zeta\delta_f^2}{2} + \beta_t^{i^*}||x - x_{i^*}^s||_{A_{i^*,t}^{-1}} + 2\zeta\delta_r\delta_f\sqrt{2d\log T},\end{aligned}$$

Since we pick actions as $x_t = \max_{x\in\tilde{D}_i^w}\|x - x_i^s\|_{A_{i,t}^{-1}}$, any d consecutive actions are orthogonal and uniformly expand the eigenspectrum, i.e., $\sum_{t=s}^{t=s+d} x_t x_t^{\top} = \delta_r^2\mathbf{I}$. Thus,

$$\beta_{T'}^{i^*}||x - x_{i^*}^s||_{A_{T'}^{i^*-1}} \leq \beta_{T'}^{i^*}\frac{\sqrt{d}}{\delta_r\sqrt{T'}}\|x - x_{i^*}^s\| \leq \beta_{T'}^{i^*}\frac{\sqrt{d}}{\delta_r\sqrt{T'}}\delta_f,$$

which gives us the following upper bound:

$$\begin{aligned}\nabla\hat{f}_{i^*T'}^{\top}(x^* - x_{i^*}^s) &\leq f_{i^*}(x^*) - f_{i^*}(x_{i^*}^s) + \frac{\zeta\delta_f^2}{2} + \beta_{T'}^{i^*}\frac{\sqrt{d}}{\delta_r\sqrt{T'}}\delta_f + 2\zeta\delta_r\delta_f\sqrt{2d\log T} \\ &= f_{i^*}(x^*) - f_{i^*}(x_{i^*}^s) + \frac{\Delta}{2}.\end{aligned}$$

The last inequality follows when T' is large enough such that

$$\frac{\Delta}{2} \geq \frac{\zeta\delta_f^2}{2} + 2\zeta\delta_r\delta_f\sqrt{2d\log T'} + \beta_T\sqrt{\frac{d}{T'}}\frac{\delta_f}{\delta_r}.$$

Moreover, we can scale $\delta_r \sim (\frac{1}{T'})^{0.25}$ to get

$$\frac{\Delta}{2} \geq \frac{\zeta\delta_f^2}{2} + 2\zeta\delta_f\sqrt{\frac{2d\log T'}{\sqrt{T'}}} + \beta_T\delta_f\sqrt{\frac{d}{\sqrt{T'}}} \sim O(1).$$

To find T', we show that the upper bound of RHS is less than LHS. In particular, we show the following

$$\frac{\zeta\delta_f^2}{2} + 2\zeta\delta_f\sqrt{\frac{2d\log T'}{\sqrt{T'}}} + \beta_T\delta_f\sqrt{\frac{d}{\sqrt{T'}}} \leq \frac{\zeta\delta_f^2}{2} + \delta_f\sqrt{\frac{2d\log T'}{\sqrt{T'}}}(2\zeta+\beta_T) \leq \frac{\Delta}{2}.$$

Rearranging the above, we get

$$\frac{T'}{\log^2 T'} \geq \left(2d\frac{4\delta_f^2}{(\Delta-\zeta\delta_f^2)^2}\right)^2.$$

Therefore for any $\Delta > \zeta\delta_f^2$, we can arbitrarily find large enough exploration time T' to satisfy the safety gap. Now plugging into the definition of $\hat{\Gamma}_{i,t}$, we get

$$f_{i^*}(x^*) - f_{i^*}(x_{i^*}^s) + \frac{\Delta}{2} + \frac{\Delta}{2} \leq \tau - \tau_i^s$$

which gives $f_{i^*}(x^*) \leq \tau - \Delta$ as desired.

Next, we use a similar argument to show that if $x \in \hat{\Gamma}_{i,t}$, then $f_i(x) \leq \tau$. To this we consider the following lower bound:

$$\begin{aligned}\nabla\hat{f}_{i,T'}^\top(x - x_{i^*}^s) &\geq f_i(x) - f_i(x_i^s) - \frac{\zeta\delta_f^2}{2} - \beta_{T'}^i\frac{\sqrt{d}}{\delta_r\sqrt{T'}}\delta_f + 2\zeta\delta_r\delta_f\sqrt{2d\log T} \\ &= f_i(x) - f_i(x_i^s) - \frac{\Delta}{2}.\end{aligned}$$

Plugging into definition of $\hat{\Gamma}_{i,t}$, the $\frac{\Delta}{2}$ terms cancel out to give $f_i(x) \leq \tau$ as desired.

□

Appendix B

FURTHER PROOFS FOR CHAPTER 3

B.1 Proofs of Section 3.2

In Appendix B.1.1, we show that due to improved exploration strategy, the regularized design matrix V_t has its minimum eigenvalue scaling linearly over time, which guarantees the persistently exciting inputs for finding the stabilizing neighborhood and stabilizing controllers after the adaptive control with improved exploration phase. The exact definition of $\sigma_\star$ from Lemma 3.1 is also given in Lemma B.1 in Appendix B.1.1. We provide the system identification and confidence set constructions with their guarantees (both in terms of the self-normalized and the spectral norm) in Appendix B.1.2. In Appendix B.1.3, we provide the boundedness guarantees for the system's state throughout the execution of StabL and provide the proof of Lemma 3.5. The precise definition of T_w, which was omitted in the main text, is also given in (B.24) in Appendix B.1.3. We provide the regret decomposition in Appendix B.1.4 and we analyze each term in this decomposition and give the proof of the main result of the paper in Appendix B.1.5.

B.1.1 Smallest Singular Value of V_t, Proof of Theorem 3.1

In this section, we show that improved exploration of StabL provides persistently exciting inputs, which will be used to enable reaching a stabilizing neighborhood around the system parameters. In other words, we will lower bound the smallest eigenvalue of the regularized design matrix, V_t. The analysis generalizes the lower bound on the smallest eigenvalue of the sample covariance matrix in Theorem 20 of [62] for the general case of subgaussian noise. For the state x_t, and input u_t, we have:

$$x_t = A_* x_{t-1} + B_* u_{t-1} + w_{t-1}, \quad \textit{and} \quad u_t = K(\tilde{\Theta}_{t-1}) x_t + \nu_t. \tag{B.1}$$

Let $\xi_t = z_t - \mathbb{E}\left[z_t | \mathcal{F}_{t-1}\right]$. Using the equalities in (B.1), and the fact that w_t and ν_t are $\mathcal{F}_t$ measurable, we write $\mathbb{E}\left[\xi_t \xi_t^\top | \mathcal{F}_{t-1}\right]$ as follows.

$$\mathbb{E}\left[\xi_t\xi_t^\top|\mathcal{F}_{t-1}\right] = \begin{pmatrix} I \\ K(\tilde{\Theta}_{t-1}) \end{pmatrix} \mathbb{E}\left[w_t w_t^\top|\mathcal{F}_{t-1}\right] \begin{pmatrix} I \\ K(\tilde{\Theta}_{t-1}) \end{pmatrix}^\top + \begin{pmatrix} 0 & 0 \\ 0 & \mathbb{E}\left[\nu_t\nu_t^\top|\mathcal{F}_{t-1}\right] \end{pmatrix}$$

$$= \begin{pmatrix} I \\ K(\tilde{\Theta}_{t-1}) \end{pmatrix} (\bar{\sigma}_w^2 I) \begin{pmatrix} I \\ K(\tilde{\Theta}_{t-1}) \end{pmatrix}^\top + \begin{pmatrix} 0 & 0 \\ 0 & \sigma_\nu^2 I \end{pmatrix} \tag{B.2}$$

$$= \begin{pmatrix} \bar{\sigma}_w^2 I & \bar{\sigma}_w^2 K(\tilde{\Theta}_{t-1})^\top \\ \bar{\sigma}_w^2 K(\tilde{\Theta}_{t-1}) & \bar{\sigma}_w^2 K(\tilde{\Theta}_{t-1})K(\tilde{\Theta}_{t-1})^\top + 2\kappa^2\bar{\sigma}_w^2 I \end{pmatrix} \tag{B.3}$$

$$\geq \bar{\sigma}_w^2 \begin{pmatrix} I & K(\tilde{\Theta}_{t-1})^\top \\ K(\tilde{\Theta}_{t-1}) & 2K(\tilde{\Theta}_{t-1})K(\tilde{\Theta}_{t-1})^\top + I/2 \end{pmatrix} \tag{B.4}$$

$$= \frac{\bar{\sigma}_w^2}{2} I + \bar{\sigma}_w^2 \begin{pmatrix} \frac{1}{\sqrt{2}} I \\ \sqrt{2}K(\tilde{\Theta}_{t-1}) \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} I \\ \sqrt{2}K(\tilde{\Theta}_{t-1}) \end{pmatrix}^\top \tag{B.5}$$

$$\geq \frac{\bar{\sigma}_w^2}{2} I, \tag{B.6}$$

where (B.3) follows from $\sigma_\nu^2 = 2\kappa^2\bar{\sigma}_w^2$ and (B.4) follows from the fact that $\kappa \geq 1$ and $\|K(\tilde{\Theta}_{t-1})\| \leq \kappa$ for all t. Let $s_t = v^\top\xi_t$ for any unit vector $v \in \mathbb{R}^{n+d}$. (B.6) shows that that Var $[s_t|\mathcal{F}_{t-1}] \geq \frac{\bar{\sigma}_w^2}{2}$.

Lemma B.1. *Suppose the system is stabilizable and we are in adaptive control with improved exploration phase of StabL. Denote $s_t = v^\top\xi_t$ where $v \in \mathbb{R}^{n+d}$ is any unit vector. Let $\bar{\sigma}_\nu := ((1+\kappa)^2 + 2\kappa^2)\sigma_w^2$. For a given positive σ_1^2, let E_t be an indicator random variable that equals 1 if $s_t^2 > \sigma_1^2$ and 0 otherwise. Then for any positive σ_1^2, and σ_2^2, such that $\sigma_1^2 \leq \sigma_2^2$, we have*

$$\mathbb{E}\left[E_t|\mathcal{F}_{t-1}\right] \geq \frac{\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 - 4\bar{\sigma}_\nu^2(1 + \frac{\sigma_2^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_2^2}{2\bar{\sigma}_\nu^2})}{\sigma_2^2}. \tag{B.7}$$

Note that, for any $\bar{\sigma}_\nu \geq \bar{\sigma}_w$, there is a pair (σ_1^2, σ_2^2) such that the right-hand side of (B.7) is positive.

Proof. Using the lower bound on the variance of s_t, we have,

$$\begin{aligned} \frac{\bar{\sigma}_w^2}{2} &\leq \mathbb{E}\left[s_t^2\mathbb{1}(s_t^2 < \sigma_1^2)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t^2\mathbb{1}(s_t^2 \geq \sigma_1^2)|\mathcal{F}_{t-1}\right] \\ &\leq \sigma_1^2 + \mathbb{E}\left[s_t^2\mathbb{1}(s_t^2 \geq \sigma_1^2)|\mathcal{F}_{t-1}\right]. \end{aligned}$$

Now, deploying the fact that both ν_t and w_t, for any t, are sub-Gaussian given $\mathcal{F}_{t-1}$, have that ξ_t is also sub-Gaussian vector. Therefore, s_t is a sub-Gaussian random

variable with parameter $\bar{\sigma}_\nu$, where $\bar{\sigma}_\nu := ((1+\kappa)^2 + 2\kappa^2)\sigma_w^2$:

$$\begin{aligned}\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 &\leq \mathbb{E}\left[s_t^2 \mathbb{1}(s_t^2 \geq \sigma_1^2)|\mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[s_t^2 \mathbb{1}(\sigma_2^2 \geq s_t^2 \geq \sigma_1^2)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t^2 \mathbb{1}(s_t^2 \geq \sigma_2^2)|\mathcal{F}_{t-1}\right]. \end{aligned} \tag{B.8}$$

For the second term in the right-hand side of the (B.8), under the considerations of Fubini's and Radon–Nikodym theorems, we derive the following equality,

$$\begin{aligned}\int_{s^2 \geq \sigma_2^2} \mathbb{P}(s_t^2 \geq s^2|\mathcal{F}_{t-1}) ds^2 &= \int_{s^2 \geq \sigma_2^2} \int_{s'^2 \geq s^2} -\frac{d\mathbb{P}(s_t^2 \geq s'^2|\mathcal{F}_{t-1})}{ds'^2} ds'^2 ds^2 \\ &= \int_{s'^2 \geq \sigma_2^2} \int_{s'^2 \geq s^2 \geq \sigma_2^2} -\frac{d\mathbb{P}(s_t^2 \geq s'^2|\mathcal{F}_{t-1})}{ds'^2} ds'^2 ds^2 \\ &= \int_{s'^2 \geq \sigma_2^2} \int_{s'^2 \geq s^2 \geq \sigma_2^2} -\frac{d\mathbb{P}(s_t^2 \geq s'^2|\mathcal{F}_{t-1})}{ds'^2} ds^2 ds'^2 \\ &= \int_{s'^2 \geq \sigma_2^2} -\frac{d\mathbb{P}(s_t^2 \geq s'^2|\mathcal{F}_{t-1})}{ds'^2} (s'^2 - \sigma_2^2) ds'^2 \\ &= \mathbb{E}\left[s_t^2 \mathbb{1}(s_t^2 \geq \sigma_2^2)|\mathcal{F}_{t-1}\right] - \sigma_2^2 \int_{s'^2 \geq \sigma_2^2} -\frac{d\mathbb{P}(s_t^2 \geq s'^2|\mathcal{F}_{t-1})}{ds'^2} ds'^2 \\ &= \mathbb{E}\left[s_t^2 \mathbb{1}(s_t^2 \geq \sigma_2^2)|\mathcal{F}_{t-1}\right] - \sigma_2^2 \, \mathbb{P}(s_t^2 \geq \sigma_2^2|\mathcal{F}_{t-1}),\end{aligned}$$

resulting in the following equality,

$$\mathbb{E}\left[s_t^2 \mathbb{1}(s_t^2 \geq \sigma_2^2)|\mathcal{F}_{t-1}\right] = \int_{s^2 \geq \sigma_2^2} \mathbb{P}(s_t^2 \geq s^2|\mathcal{F}_{t-1}) ds^2 + \sigma_2^2 \, \mathbb{P}(s_t^2 \geq \sigma_2^2|\mathcal{F}_{t-1}). \tag{B.9}$$

Using this equality, we extend the (B.8) as follows,

$$\begin{aligned}\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 &\leq \mathbb{E}\left[s_t^2 \mathbb{1}(\sigma_2^2 \geq s_t^2 \geq \sigma_1^2)|\mathcal{F}_{t-1}\right] + \int_{s^2 \geq \sigma_2^2} \mathbb{P}(s_t^2 \geq s^2|\mathcal{F}_{t-1}) ds^2 + \sigma_2^2 \, \mathbb{P}(s_t^2 \geq \sigma_2^2|\mathcal{F}_{t-1}) \\ &\leq \sigma_2^2 \, \mathbb{E}\left[\mathbb{1}(\sigma_2^2 \geq s_t^2 \geq \sigma_1^2)|\mathcal{F}_{t-1}\right] + \int_{s^2 \geq \sigma_2^2} \mathbb{P}(s_t^2 \geq s^2|\mathcal{F}_{t-1}) ds^2 + \sigma_2^2 \, \mathbb{P}(s_t^2 \geq \sigma_2^2|\mathcal{F}_{t-1}) \\ &\leq \sigma_2^2 \, \mathbb{E}\left[E_t|\mathcal{F}_{t-1}\right] + \int_{s^2 \geq \sigma_2^2} \mathbb{P}(s_t^2 \geq s^2|\mathcal{F}_{t-1}) ds^2 + \sigma_2^2 \, \mathbb{P}(s_t^2 \geq \sigma_2^2|\mathcal{F}_{t-1}).\end{aligned} \tag{B.10}$$

Rearranging this inequality, we have,

$$\begin{aligned}
\mathbb{E}\left[E_t|\mathcal{F}_{t-1}\right] &\geq \frac{\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 - \int_{s^2\geq\sigma_2^2} \mathbb{P}(s_t^2 \geq s^2|\mathcal{F}_{t-1})ds^2 - \sigma_2^2\,\mathbb{P}(s_t^2 \geq \sigma_2^2|\mathcal{F}_{t-1})}{\sigma_2^2} \\
&\geq \frac{\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 - 2\int_{s^2\geq\sigma_2^2} \exp(\frac{-s^2}{2\bar{\sigma}_v^2})ds^2 - 2\sigma_2^2\exp(\frac{-\sigma_2^2}{2\bar{\sigma}_v^2})}{\sigma_2^2} \\
&\geq \frac{\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 - 4\bar{\sigma}_v^2\exp(\frac{-\sigma_2^2}{2\bar{\sigma}_v^2}) - 2\sigma_2^2\exp(\frac{-\sigma_2^2}{2\bar{\sigma}_v^2})}{\sigma_2^2} \\
&= \frac{\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 - 4\bar{\sigma}_v^2(1+\frac{\sigma_2^2}{2\bar{\sigma}_v^2})\exp(\frac{-\sigma_2^2}{2\bar{\sigma}_v^2})}{\sigma_2^2}. \qquad \text{(B.11)}
\end{aligned}$$

The inequality in (B.11) holds for any $\sigma_1^2 \leq \sigma_2^2$, therefore, the stated lower-bound on $\mathbb{E}\left[E_t|\mathcal{F}_{t-1}\right]$ in the main statement holds. □

For the choices of σ_1^2 and σ_2^2 that makes right hand side of (B.7) positive, let c_p denote the right hand side of (B.7), $c_p = \frac{\frac{\bar{\sigma}_w^2}{2} - \sigma_1^2 - 4\bar{\sigma}_v^2(1+\frac{\sigma_2^2}{2\bar{\sigma}_v^2})\exp(\frac{-\sigma_2^2}{2\bar{\sigma}_v^2})}{\sigma_2^2}$.

Lemma B.2. *Consider $\bar{s}_t = v^\top z_t$ where $v \in \mathbb{R}^{n+d}$ is any unit vector. Let $\bar{E}_t$ be an indicator random variable that equals 1 if $\bar{s}_t^2 > \sigma_1^2/4$ and 0 otherwise. Then, there exist a positive pair σ_1^2, and σ_2^2, and a constant $c_p > 0$, such that $\mathbb{E}\left[\bar{E}_t|\mathcal{F}_{t-1}\right] \geq c_p' > 0$.*

Proof. Using the Lemma B.1, we know that for $s_t = v^\top\xi_t$, we have $|s_t| \geq \sigma_1$ with a non-zero probability c_p. On the other hand, we have that,

$$\bar{s}_t = v^\top z_t = v^\top\xi_t + v^\top\mathbb{E}\left[z_t|\mathcal{F}_{t-1}\right] = s_t + v^\top\mathbb{E}\left[z_t|\mathcal{F}_{t-1}\right].$$

Therefore, we have, $|\bar{s}_t| = |s_t + v^\top\mathbb{E}\left[z_t|\mathcal{F}_{t-1}\right]|$. Using this equality, if $|v^\top\mathbb{E}\left[z_t|\mathcal{F}_{t-1}\right]| \leq \sigma_1/2$, since $|s_t| \geq \sigma_1$ with probability c_p, we have $|\bar{s}_t| \geq \sigma_1/2$ with probability c_p.

In the following, we consider the case where $|v^\top\mathbb{E}\left[z_t|\mathcal{F}_{t-1}\right]| \geq \sigma_1/2$. For a constant σ_3, using a similar derivation as in (B.9) and (B.10), we have

$$\begin{aligned}
\mathbb{E}\left[s_t^2|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[s_t^2\mathbb{1}(\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t^2\mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t^2\mathbb{1}(s_t^2 \geq \sigma_3^2)|\mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[s_t^2\mathbb{1}(\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t^2\mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right] + 4\bar{\sigma}_v^2(1+\frac{\sigma_2^2}{2\bar{\sigma}_v^2})\exp(\frac{-\sigma_2^2}{2\bar{\sigma}_v^2}).
\end{aligned}$$

Using the lower bound in the variance results in,

$$\frac{\bar{\sigma}_w^2}{2} \leq \mathbb{E}\left[s_t^2 \mathbb{1}(\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t^2 \mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right] + 4\bar{\sigma}_\nu^2 (1 + \frac{\sigma_3^2}{2\bar{\sigma}_\nu^2}) \exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2}).$$

Therefore,

$$\begin{aligned}
\frac{\bar{\sigma}_w^2}{2} - 4\bar{\sigma}_\nu^2 (1 + \frac{\sigma_3^2}{2\bar{\sigma}_\nu^2}) \exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2}) &\leq \mathbb{E}\left[s_t^2 \mathbb{1}(\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[s_t^2 \mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right] \\
&= \sigma_3^2 \left(\mathbb{E}\left[\frac{s_t^2}{\sigma_3^2} \mathbb{1}(-\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[\frac{s_t^2}{\sigma_3^2} \mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right] \right) \\
&\leq \sigma_3^2 \left(\mathbb{E}\left[\frac{|s_t|}{\sigma_3} \mathbb{1}(-\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[\frac{s_t}{\sigma_3} \mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}\right] \right).
\end{aligned} \tag{B.12}$$

Note that for a large enough σ_3, the second term on the left-hand side vanishes. Since we have $\mathbb{E}[s_t|\mathcal{F}_{t-1}] = 0$, we write the following, to further analyze the right-hand side of (B.12),

$$\begin{aligned}
&\mathbb{E}[s_t|\mathcal{F}_{t-1}] = \mathbb{E}[s_t \mathbb{1}(s_t < 0)|\mathcal{F}_{t-1}] + \mathbb{E}[s_t \mathbb{1}(s_t > 0)|\mathcal{F}_{t-1}] = 0 \\
&\rightarrow \mathbb{E}[|s_t| \mathbb{1}(s_t < 0)|\mathcal{F}_{t-1}] = \mathbb{E}[s_t \mathbb{1}(s_t > 0)|\mathcal{F}_{t-1}].
\end{aligned}$$

Note that, since s_t is sub-Gaussian variable, and has bounded away from zero variance, we have $\mathbb{E}[\mathbb{1}(s_t < 0)|\mathcal{F}_{t-1}] + \mathbb{E}[\mathbb{1}(s_t > 0)|\mathcal{F}_{t-1}]$ is bounded away from zero. We write this equality as follows:

$$\begin{aligned}
&\mathbb{E}[|s_t| \mathbb{1}(-\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}] + \mathbb{E}[|s_t| \mathbb{1}(s_t \leq -\sigma_3)|\mathcal{F}_{t-1}] \\
&\qquad = \mathbb{E}[s_t \mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}] + \mathbb{E}[s_t \mathbb{1}(s_t \geq \sigma_3)|\mathcal{F}_{t-1}].
\end{aligned}$$

By rearranging this equality, and upper bounding the first term on the left-hand side, we have

$$\begin{aligned}
\mathbb{E}[|s_t| \mathbb{1}(-\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}] &\leq \mathbb{E}[s_t \mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}] + \mathbb{E}[s_t \mathbb{1}(s_t \geq \sigma_3)|\mathcal{F}_{t-1}] \\
&\leq \mathbb{E}[s_t \mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}] + \bar{\sigma}_\nu^2 \exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2}).
\end{aligned} \tag{B.13}$$

Similarly, we have

$$\mathbb{E}[s_t \mathbb{1}(\sigma_3 > s_t > 0)|\mathcal{F}_{t-1}] \leq \mathbb{E}[|s_t| \mathbb{1}(-\sigma_3 < s_t < 0)|\mathcal{F}_{t-1}] + \bar{\sigma}_\nu^2 \exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2}) \tag{B.14}$$

Using the inequality (B.13) on the right-hand side of (B.12), we have

$$\begin{aligned}\frac{\frac{\bar{\sigma}_w^2}{2}-4\bar{\sigma}_\nu^2(1+\frac{\sigma_3^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})}{\sigma_3^2} &\leq \mathbb{E}\left[\frac{|s_t|}{\sigma_3}\mathbb{1}(-\sigma_3<s_t<0)|\mathcal{F}_{t-1}\right]+\mathbb{E}\left[\frac{s_t}{\sigma_3}\mathbb{1}(\sigma_3>s_t>0)|\mathcal{F}_{t-1}\right]\\ &\leq 2\mathbb{E}\left[\frac{s_t}{\sigma_3}\mathbb{1}(\sigma_3>s_t>0)|\mathcal{F}_{t-1}\right]+\bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})\\ &\leq 2\mathbb{E}\left[\mathbb{1}(\sigma_3>s_t>0)|\mathcal{F}_{t-1}\right]+\bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})\\ &\leq 2\mathbb{E}\left[\mathbb{1}(s_t>0)|\mathcal{F}_{t-1}\right]+\bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2}).\end{aligned}$$

Similarly, using (B.14) on the right-hand side of (B.12) we have

$$\begin{aligned}\frac{\frac{\bar{\sigma}_w^2}{2}-4\bar{\sigma}_\nu^2(1+\frac{\sigma_3^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})}{\sigma_3^2} &\leq \mathbb{E}\left[\frac{|s_t|}{\sigma_3}\mathbb{1}(-\sigma_3<s_t<0)|\mathcal{F}_{t-1}\right]+\mathbb{E}\left[\frac{s_t}{\sigma_3}\mathbb{1}(\sigma_3>s_t>0)|\mathcal{F}_{t-1}\right]\\ &\leq 2\mathbb{E}\left[\mathbb{1}(s_t<0)|\mathcal{F}_{t-1}\right]+\bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2}).\end{aligned}$$

Therefore, it results in the two following lower bounds,

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}(s_t<0)|\mathcal{F}_{t-1}\right] &\geq \frac{\frac{\bar{\sigma}_w^2}{2}-4\bar{\sigma}_\nu^2(1+\frac{\sigma_3^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})}{2\sigma_3^2}-0.5\bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})\\ \mathbb{E}\left[\mathbb{1}(s_t>0)|\mathcal{F}_{t-1}\right] &\geq \frac{\frac{\bar{\sigma}_w^2}{2}-4\bar{\sigma}_\nu^2(1+\frac{\sigma_3^2}{2\bar{\sigma}_\nu^2})\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2})}{2\sigma_3^2}-0.5\bar{\sigma}_\nu^2\exp(\frac{-\sigma_3^2}{2\bar{\sigma}_\nu^2}). \quad \text{(B.15)}\end{aligned}$$

Choosing σ_3 sufficiently large results in the right-hand sides in inequalities (B.15) to be positive and bounded away from zero. Let $c_p'' > 0$ denote the right-hand sides in the (B.15). We use this fact to analyze $\bar{s}_t$ when $|v^\top \mathbb{E}\left[z_t|\mathcal{F}_{t-1}\right]| \geq \sigma_1/2$.

When $v^\top \mathbb{E}\left[z_t|\mathcal{F}_{t-1}\right] \geq \sigma_1/2$, since probability c_p'', s_t is positive, therefore, $|\bar{s}_t| \geq \sigma_1/2$ with probability c_p''. When $v^\top \mathbb{E}\left[z_t|\mathcal{F}_{t-1}\right] \leq -\sigma_1/2$, since probability c_p'', s_t is negative, therefore, $|\bar{s}_t| \geq \sigma_1/2$ with probability c_p''. Therefore, overall, with probability $c_p' := \min\{c_p, c_p''\}$, we have that $|\bar{s}_t| \geq \sigma_1/2$, resulting in the statement of the lemma. □

Theorem B.1 (Precise version of Theorem 3.1, Persistence of Excitation During the Extra Exploration)**.** *If the duration of the adaptive control with improved exploration* $T_w \geq \frac{6n}{c_p'}\log(12/\delta)$*, then with probability at least* $1-\delta$*, for* $\sigma_\star^2 = \frac{c_p'\sigma_1^2}{16}$*, StabL has*

$$\lambda_{\min}(V_{T_w}) \geq \sigma_\star^2 T_w.$$

Proof. Let $U_t = \bar{E}_t - \mathbb{E}_t\left[\bar{E}_t|\mathcal{F}_{t-1}\right]$. Then U_t is a martingale difference sequence with $|U_t| \leq 1$. Applying Azuma's inequality, we have that with probability at least $1-\delta$

$$\sum_{t=1}^{T_w} U_t \geq -\sqrt{2T_w \log\frac{1}{\delta}}.$$

Using the Lemma B.2, we have

$$\begin{aligned}\sum_{t}^{T_w} \bar{E}_t &\geq \sum_{t}^{T_w} \mathbb{E}_t\left[\bar{E}_t|\mathcal{F}_{t-n}\right] - \sqrt{2T_w \log\frac{1}{\delta}} \\ &\geq c'_p T_w - \sqrt{2T_w \log\frac{1}{\delta}},\end{aligned}$$

where for $T_w \geq 8\log(1/\delta)/c'^2_p$, we have $\sum_t^{T_w} \bar{E}_t \geq \frac{c'_p}{2} T_w$. Now, for any unit vector v, define $\bar{s}_t = v^\top z_t$, therefore from the definition of $\bar{E}_t$ we have,

$$v^\top V_{T_w} v = \sum_{t}^{T_w} \bar{s}_t^2 \geq \bar{E}_t \sigma_1^2/4 \geq \frac{c'_p \sigma_1^2}{8} T_w$$

This inequality hold for a given v. In the following we show a similar inequality for all v together. Similar to the Theorem 20 in [62], consider a 1/4-net of $\mathbb{S}^{n+d-1}$, $\mathbb{N}(1/4)$ and set $M_{T_w} := \{V_{T_w}^{-1/2} v/\|V_{T_w}^{-1/2} v\| : v \in \mathbb{N}(1/4)\}$. These two sets have at most 12^{n+d-1} members. Using union bound over members of this set, when $T_w \geq \frac{20}{c'^2_p}((n+d) + \log(1/\delta))$, we have that $v^\top V_{T_w} v \geq \frac{c'_p \sigma_1^2}{8} T_w$ for all $v \in M_{T_w}$ with a probability at least $1-\delta$. Using the definition of members in M_{T_w}, for each $v \in \mathbb{N}(1/4)$, we have $v^\top V_{T_w}^{-1} v \leq \frac{8}{T_w c'_p \sigma_1^2}$. Let v_n denote the eigenvector of the largest eigenvalue of $V_{T_w}^{-1}$, and a vector $v' \in \mathbb{N}(1/4)$ such that $\|v_n - v'\| \leq 1/4$. Then we have

$$\begin{aligned}\|V_{T_w}^{-1}\| = v_n^\top V_{T_w}^{-1} v_n &= v'^\top V_{T_w}^{-1} v' + (v_n - v')^\top V_{T_w}^{-1}(z_n + v') \\ &\leq \frac{8}{T_w c'_p \sigma_1^2} + \|v_n - v'\|\|V_{T_w}^{-1}\|\|z_n + v'\| \leq \frac{8}{T_w c'_p \sigma_1^2} + \|V_{T_w}^{-1}\|/2.\end{aligned}$$

Rearranging, we get that $\|V_{T_w}^{-1}\| \leq \frac{16}{T_w c'_p \sigma_1^2}$. Therefore, the advertised bound holds for $T_w \geq \frac{20}{c'^2_p}((n+d) + \log(1/\delta))$ with probability at least $1-\delta$. □

B.1.2 System Identification & Confidence Set Construction, Proof of Lemma 3.2

To have completeness, for the proof of Lemma 3.2, we first provide the proof for confidence set construction borrowed from Abbasi-Yadkori and Szepesvári [2], since Lemma 3.2 builds upon this confidence set construction. First, let

$$\kappa_e = \left(\frac{\sigma_w}{\sigma_\star}\sqrt{n(n+d)\log\left(1 + \frac{cT(1+\kappa^2)(n+d)^{2(n+d)}}{\mu(n+d)}\right) + 2n\log\frac{1}{\delta}} + \sqrt{\mu}S\right). \tag{B.16}$$

Proof. Define $\Theta_*^\top = [A, B]$ and $z_t = \left[x_t^\top u_t^\top\right]^\top$. The system in (3.1) can be characterized equivalently as

$$x_{t+1} = \Theta_*^\top z_t + w_t,$$

Given a single input-output trajectory $\{x_t, u_t\}_{t=1}^T$, one can rewrite the input-output relationship as,

$$X_T = Z_T \Theta_* + W_T, \tag{B.17}$$

for $X_T^\top = [x_1, \ldots, x_T] \in \mathbb{R}^{n\times T}$, $Z_T^\top = [z_1, \ldots, z_T] \in \mathbb{R}^{(n+d)\times T}$, and $W_T^\top = [w_1, \ldots, w_T] \in \mathbb{R}^{n\times T}$. Then, we estimate Θ_* by solving the following least square problem,

$$\begin{aligned}
\hat{\Theta}_T &= \arg\min_X ||X_T - Z_T X||_F^2 + \mu||X||_F^2 \\
&= (Z_T^\top Z_T + \mu I)^{-1} Z_T^\top X_T \\
&= (Z_T^\top Z_T + \mu I)^{-1} Z_T^\top W_T + \Theta_* - \mu(Z_T^\top Z_T + \mu I)^{-1}\Theta_*.
\end{aligned}$$

The confidence set is obtained using the expression for $\hat{\Theta}_T$ and sub-Gaussianity of the w_t,

$$\begin{aligned}
|\operatorname{Tr}((\hat{\Theta}_T - \Theta_*)^\top X)| &= |\operatorname{Tr}(W_T^\top Z_T (Z_T^\top Z_T + \mu I)^{-1} X) - \mu \operatorname{Tr}(\Theta_*^\top (Z_T^\top Z_T + \mu I)^{-1} X)| \\
&\le |\operatorname{Tr}(W_T^\top Z_T (Z_T^\top Z_T + \mu I)^{-1} X)| + \mu |\operatorname{Tr}(\Theta_*^\top (Z_T^\top Z_T + \mu I)^{-1} X)| \\
&\le \sqrt{\operatorname{Tr}(X^\top (Z_T^\top Z_T + \mu I)^{-1} X) \operatorname{Tr}(W_T^\top Z_T (Z_T^\top Z_T + \mu I)^{-1} Z_T^\top W_T)} \\
&\quad + \mu\sqrt{\operatorname{Tr}(X^\top (Z_T^\top Z_T + \mu I)^{-1} X) \operatorname{Tr}(\Theta_*^\top (Z_T^\top Z_T + \mu I)^{-1}\Theta_*)}, \qquad \text{(B.18)} \\
&= \sqrt{\operatorname{Tr}(X^\top (Z_T^\top Z_T + \mu I)^{-1} X)}\left[\sqrt{\operatorname{Tr}(W_T^\top Z_T (Z_T^\top Z_T + \mu I)^{-1} Z_T^\top W_T)} + \mu\sqrt{\operatorname{Tr}(\Theta_*^\top (Z_T^\top Z_T + \mu I)^{-1}\Theta_*)}\right]
\end{aligned}$$

where the result follows from $|\operatorname{Tr}(A^\top B C)| \le \sqrt{\operatorname{Tr}(A^\top B A)\operatorname{Tr}(C^\top B C)}$ for square positive definite B due to Cauchy Schwarz (weighted inner-product). For $X = (Z_T^\top Z_T + \mu I)(\hat{\Theta}_T - \Theta_*)$, we get

$$\sqrt{\operatorname{Tr}((\hat{\Theta}_T - \Theta_*)^\top (Z_T^\top Z_T + \mu I)(\hat{\Theta}_T - \Theta_*))} \le \sqrt{\operatorname{Tr}(W_T^\top Z_T (Z_T^\top Z_T + \mu I)^{-1} Z_T^\top W_T)} + \sqrt{\mu}\sqrt{\operatorname{Tr}(\Theta_*^\top \Theta_*)}.$$

Let $\mathcal{S}_T = Z_T^\top W_T \in \mathbb{R}^{(n+d)\times n}$ and s_i denote the columns of it. Also, let $V_T = (Z_T^\top Z_T + \mu I)$. Thus,

$$\operatorname{Tr}(W_T^\top Z_T (Z_T^\top Z_T + \mu I)^{-1} Z_T^\top W_T) = \operatorname{Tr}(\mathcal{S}_T^\top V_T^{-1} \mathcal{S}_T) = \sum_{i=1}^n s_i^\top V_T^{-1} s_i = \sum_{i=1}^n \|s_i\|_{V_T^{-1}}^2. \tag{B.19}$$

Notice that $s_i = \sum_{j=1}^{T} w_{j,i} z_j$ where $w_{j,i}$ is the i'th element of w_j. From Assumption 3.1, we have that $w_{j,i}$ is σ_w-sub-Gaussian, thus we can use Theorem 1 of [3], which gives a self-normalized bound for vector-valued martingales and show that,

$$\mathrm{Tr}(W_T^\top Z_T (Z_T^\top Z_T + \mu I)^{-1} Z_T^\top W_T) \leq 2n\sigma_w^2 \log\left(\frac{\det(V_T)^{1/2}\det(\mu I)^{-1/2}}{\delta}\right). \quad \text{(B.20)}$$

with probability $1-\delta$. From Assumption 3.2, we also have that $\sqrt{\mathrm{Tr}(\Theta_*^\top \Theta_*)} \leq S$. Combining these gives the self-normalized confidence set or the model estimate:

$$\mathrm{Tr}((\hat{\Theta}_T - \Theta_*)^\top V_T (\hat{\Theta}_T - \Theta_*)) \leq \left(\sigma_w \sqrt{2n \log\left(\frac{\det(V_T)^{1/2}\det(\mu I)^{-1/2}}{\delta}\right)} + \sqrt{\mu} S\right)^2. \quad \text{(B.21)}$$

Notice that we have $\mathrm{Tr}((\hat{\Theta}_T - \Theta_*)^\top V_T (\hat{\Theta}_T - \Theta_*)) \geq \lambda_{\min}(V_T)\|\hat{\Theta}_T - \Theta_*\|_F^2$. Therefore,

$$\|\hat{\Theta}_T - \Theta_*\|_2 \leq \frac{1}{\sqrt{\lambda_{\min}(V_T)}}\left(\sigma_w \sqrt{2n \log\left(\frac{\det(V_T)^{1/2}\det(\mu I)^{-1/2}}{\delta}\right)} + \sqrt{\mu} S\right). \quad \text{(B.22)}$$

To complete the proof, we need a lower bound on $\lambda_{\min}(V_{T_w})$. Using Lemma 3.1, we obtain the following with probability at least $1-2\delta$:

$$\|\hat{\Theta}_{T_w} - \Theta_*\|_2 \leq \frac{\beta_t(\delta)}{\sigma_\star \sqrt{T_w}}.$$

From Lemma 3.5, for $t \leq T_w$, we have that $\|z_t\| \leq c(n+d)^{n+d}$ with probability at least $1-2\delta$, for some constant c. Combining this with Lemma 11 of [3], we get

$$\|\hat{\Theta}_{T_w} - \Theta_*\|_2 \leq \frac{\kappa_e}{\sqrt{T_w}}. \quad \text{(B.23)}$$

□

B.1.3 Boundedness of State, Proof of Lemma 3.5

In this section, we will provide the proof of Lemma 3.5, *i.e.* bounds on states for the adaptive control with improved exploration and stabilizing adaptive control phases. First, define the following. Let

$$T_w = \frac{\kappa_e^2}{\min\{\bar{\sigma}_w^2 nD/C_0, \epsilon^2\}} \quad \text{(B.24)}$$

such that for $T > T_w$, we have $\|\hat{\Theta}_T - \Theta_*\|_2 \leq \min\{\sqrt{\bar{\sigma}_w^2 nD/C_0}, \epsilon\}$ with probability at least $1-2\delta$. Notice that due to Lemma 3.3 and as shown in the following, these

guarantee the stability of the closed-loop dynamics for deploying an optimistic controller for the remaining part of StabL. Choose an error probability, $\delta > 0$. Consider the following events, in the probability space Ω:

- The event that the confidence sets hold for $s = 0, \ldots, T$,

$$\mathcal{E}_t = \{\omega \in \Omega : \forall s \leq T, \quad \Theta_* \in C_s(\delta)\}$$

- The event that the state vector stays "small" for $s = 0, \ldots, T_w$,

$$\mathcal{F}_t = \{\omega \in \Omega : \forall s \leq T_w, \quad \|x_s\| \leq \bar{\alpha}_t\}$$

where

$$\bar{\alpha}_t = \frac{18\kappa^3}{\gamma(8\kappa - 1)}\bar{\eta}^{n+d}\left[GZ_t^{\frac{n+d}{n+d+1}}\beta_t(\delta)^{\frac{1}{2(n+d+1)}} + (\|B_*\|\sigma_\nu + \sigma_w)\sqrt{2n\log\frac{nt}{\delta}}\right],$$

for

$$\bar{\eta} \geq \sup_{\Theta\in\mathcal{S}} \|A_* + B_*K(\Theta)\|, \qquad Z_T = \max_{1\leq t\leq T}\|z_t\|$$

$$G = 2\left(\frac{2S(n+d)^{n+d+1/2}}{\sqrt{U}}\right)^{1/(n+d+1)}, \quad U = \frac{U_0}{H}, \quad U_0 = \frac{1}{16^{n+d-2}\max\left(1,\ S^{2(n+d-2)}\right)},$$

and H is any number satisfying

$$H > \max\left(16,\ \frac{4S^2M^2}{(n+d)U_0}\right), \quad \text{where} \quad M = \sup_{Y\geq 1}\frac{\left(\sigma_w\sqrt{n(n+d)\log\left(\frac{1+TY/\mu}{\delta}\right)} + \mu^{1/2}S\right)}{Y}.$$

Notice that $\mathcal{E}_1 \supseteq \mathcal{E}_2 \supseteq \ldots \supseteq \mathcal{E}_T$ and $\mathcal{F}_1 \supseteq \mathcal{F}_2 \supseteq \ldots \supseteq \mathcal{F}_{T_s}$. This means considering the probability of the last event is sufficient in lower bounding all events happening simultaneously. In Abbasi-Yadkori and Szepesvári [2], an argument regarding projection onto subspaces is constructed to show that the norm of the state is well-controlled except $n + d$ times at most in any horizon T. The set of time steps that is not well-controlled is denoted as $\mathcal{T}_t$. The given lemma shows how well controlled $\|(\Theta_* - \hat{\Theta}_t)^\top z_t\|$ is besides $\mathcal{T}_t$.

Lemma B.3 (Lemma 18 of Abbasi-Yadkori and Szepesvári [2])**.** *We have that for any* $0 \leq t \leq T$*,*

$$\max_{s\leq t, s\notin T_t}\left\|(\Theta_* - \hat{\Theta}_s)^\top z_s\right\| \leq GZ_t^{\frac{n+d}{n+d+1}}\beta_t(\delta/4)^{\frac{1}{2(n+d+1)}}.$$

Notice that Lemma B.3 does not depend on controllability or the stabilizability of the system. Thus, we will use Lemma B.3 for $t \leq T_w$ for the adaptive control with improved exploration phase of StabL. Then we consider the effect of stabilizing controllers for the remaining time steps.

State Bound for the Adaptive Control with Improved Exploration Phase

One can write the state update as $x_{t+1} = \Gamma_t x_t + r_t$, where

$$\Gamma_t = \begin{cases} \tilde{A}_{t-1} + \tilde{B}_{t-1} K(\tilde{\Theta}_{t-1}) & t \notin \mathcal{T}_T \\ A_* + B_* K(\tilde{\Theta}_{t-1}) & t \in \mathcal{T}_T \end{cases} \text{ and } r_t = \begin{cases} (\Theta_* - \tilde{\Theta}_{t-1})^\top z_t + B_* \nu_t + w_t & t \notin \mathcal{T}_T \\ B_* \nu_t + w_t & t \in \mathcal{T}_T \end{cases}. \tag{B.25}$$

Thus, using the fact that $x_0 = 0$, we can obtain the following roll out for x_t,

$$\begin{aligned} x_t &= \Gamma_{t-1} x_{t-1} + r_{t-1} = \Gamma_{t-1} \left(\Gamma_{t-2} x_{t-2} + r_{t-2} \right) + r_t \\ &= \Gamma_{t-1} \Gamma_{t-2} \Gamma_{t-3} x_{t-3} + \Gamma_{t-1} \Gamma_{t-2} r_{t-2} + \Gamma_{t-1} r_{t-1} + r_t \\ &= \Gamma_{t-1} \Gamma_{t-2} \dots \Gamma_{t-(t-1)} r_1 + \dots + \Gamma_{t-1} \Gamma_{t-2} r_{t-2} + \Gamma_{t-1} r_{t-1} + r_t \\ &= \sum_{k=1}^{t} \left(\prod_{s=k}^{t-1} \Gamma_s \right) r_k . \end{aligned} \tag{B.26}$$

Recall that the controller is optimistically designed from the set of parameters are (κ, γ)-strongly stabilizable by their optimal controllers. Therefore, we have

$$1 - \gamma \geq \max_{t \leq T} \rho \left(\tilde{A}_t + \tilde{B}_t K(\tilde{\Theta}_t) \right). \tag{B.27}$$

Therefore, multiplication of closed-loop system matrices, $\tilde{A}_t + \tilde{B}_t K(\tilde{\Theta}_t)$, is not guaranteed to be contractive. In Abbasi-Yadkori and Szepesvári [2], the authors assume these matrices are contractive under controllability assumption. In order to bound the state similarly, we need to satisfy that the epochs that we use a particular optimistic controller is long enough that the state doesn't scale too badly during the exploration and produces bounded state. Thus, by choosing $H_0 = 2\gamma^{-1} \log(2\kappa\sqrt{2})$ and adopting Lemma 39 of Cassel et al. [48], we have that

$$\|x_t\| \leq \frac{18 \kappa^3 \bar{\eta}^{n+d}}{\gamma(8\kappa - 1)} \left(\max_{1 \leq k \leq t} \|r_k\| \right). \tag{B.28}$$

Furthermore, we have that $\|r_k\| \leq \left\| (\Theta_* - \tilde{\Theta}_{k-1})^\top z_k \right\| + \|B_* \nu_k + w_k\|$ when $k \notin \mathcal{T}_T$, and $\|r_k\| = \|B_* \nu_k + w_k\|$, otherwise. Hence,

$$\max_{k\le t}\|r_k\| \le \max_{k\le t, k\notin\mathcal{T}_t}\left\|(\Theta_* - \tilde{\Theta}_{k-1})^\top z_k\right\| + \max_{k\le t}\|B_*\nu_k + w_k\|\,.$$

The first term is bounded by the Lemma B.3. The second term involves summation of independent $\|B_*\|\sigma_\nu$ and σ_w subgaussian vectors. Using standard sub-Gaussian vector norm upper bound with a union bound argument, for all $k \le t$, we have $\|B_*\nu_k + w_k\| \le (\|B_*\|\sigma_\nu + \sigma_w)\sqrt{2n\log\frac{nt}{\delta}}$ with probability at least $1-\delta$. Therefore, on the event of $\mathcal{E}$,

$$\|x_t\| \le \frac{18\kappa^3\bar{\eta}^{n+d}}{\gamma(8\kappa-1)}\left[GZ_t^{\frac{n+d}{n+d+1}}\beta_t(\delta)^{\frac{1}{2(n+d+1)}} + (\|B_*\|\sigma_\nu+\sigma_w)\sqrt{2n\log\frac{nt}{\delta}}\right] \tag{B.29}$$

for $t \le T_w$. Using union bound, we can deduce that $\mathcal{E}_T \cap \mathcal{F}_{T_s}$ holds with probability at least $1-2\delta$. Notice that this bound depends on Z_t and $\beta_t(\delta)$ which in turn depends on x_t. Using Lemma 5 of [2], one can obtain the following bound

$$\|x_t\| \le c'(n+d)^{n+d}, \tag{B.30}$$

for some large enough constant c'. The adaptive control with improved exploration phase of StabL has this exponential dimension dependent state bound for all $t \le T_w$. In the following section, we show that during the stabilizing adaptive control phase, the bound on the state has a polynomial dependency on the dimensions.

State Bound in Stabilizing Adaptive Control phase

In the stabilizing adaptive control phase, StabL stops using the additive isotropic exploration component ν_t, the state follows the dynamics of

$$x_{t+1} = (A_* + B_*K(\tilde{\Theta}_{t-1}))x_t + w_t. \tag{B.31}$$

Denote $\mathbf{M_t} = A_* + B_*K(\tilde{\Theta}_{t-1})$ as the closed loop dynamics of the system. From the choice of T_w for the stabilizable systems, we have that $\mathbf{M_t}$ is $(\kappa\sqrt{2}, \gamma/2)$-strongly stable. Thus, we have $\rho(\mathbf{M_t}) \le 1-\gamma/2$ for all $t > T_s$ and $\|H_t\|\|H_t^{-1}\| \le \kappa\sqrt{2}$ for $H_t \succ 0$, such that $\|L_t\| \le 1-\gamma/2$ for $\mathbf{M_t} = H_tL_tH_t^{-1}$. Then for $T > t > T_w$, if the same policy, $\mathbf{M}$ is applied starting from state x_{T_w}, we have

$$\|x_t\| = \left\|\prod_{i=T_w+1}^{t}\mathbf{M}x_{T_w} + \sum_{i=T_w+1}^{t}\left(\prod_{s=i}^{t-1}\mathbf{M}\right)w_i\right\| \tag{B.32}$$

$$\le \kappa\sqrt{2}(1-\gamma/2)^{t-T_w}\|x_{T_w}\| + \max_{T_w<i\le T}\|w_i\|\left(\sum_{i=T_w+1}^{t}\kappa\sqrt{2}(1-\gamma/2)^{t-i+1}\right) \tag{B.33}$$

$$\le \kappa\sqrt{2}(1-\gamma/2)^{t-T_w}\|x_{T_w}\| + \frac{2\kappa\sigma_w\sqrt{2}}{\gamma}\sqrt{2n\log(n(t-T_w)/\delta)}. \tag{B.34}$$

Note that $H_0 = 2\gamma^{-1}\log(2\kappa\sqrt{2})$. This gives that $\kappa\sqrt{2}(1-\gamma/2)^{H_0} \leq 1/2$. Therefore, at the end of each controller period, the effect of the previous state is halved. Using this fact, at the ith policy change after T_w, we get

$$\begin{aligned}\|x_{t_i}\| &\leq 2^{-i}\|x_{T_w}\| + \sum_{j=0}^{i-1} 2^{-j}\frac{2\kappa\sigma_w\sqrt{2}}{\gamma}\sqrt{2n\log(n(t-T_w)/\delta)} \\ &\leq 2^{-i}\|x_{T_w}\| + \frac{4\kappa\sigma_w\sqrt{2}}{\gamma}\sqrt{2n\log(n(t-T_w)/\delta)}.\end{aligned}$$

For all $i > (n+d)\log(n+d) - \log(\frac{2\kappa\sigma_w\sqrt{2}}{\gamma}\sqrt{2n\log(n(t-T_w)/\delta)})$, at policy change i, we get

$$\|x_{t_i}\| \leq \frac{6\kappa\sigma_w\sqrt{2}}{\gamma}\sqrt{2n\log(n(t-T_w)/\delta)}.$$

Moreover, due to the stability of the synthesized controller, the worst possible controller update scheme is to update the controller every H_0 time-steps, *i.e.,* invoking the condition of $t-\tau > H_0$ in the update rule. Notice that this update rule considers the worst effect of similarity transformation on the growth of the state, since otherwise applying the same controller for longer periods would have a further reduction on the state due to the contraction that the stabilizing controller brings. Thus, from (B.34) we have that

$$\|x_t\| \leq \frac{(12\kappa^2 + 2\kappa\sqrt{2})\sigma_w}{\gamma}\sqrt{2n\log(n(t-T_w)/\delta)}, \tag{B.35}$$

for all $t > T_r \coloneqq T_w + T_{base}$ where $T_{base} = ((n+d)\log(n+d))\,H_0$.

B.1.4 Regret Decomposition

The regret decomposition leverages the OFU principle. Since during the adaptive control with improved exploration period StabL applies independent isotropic perturbations through the controller but still designs the optimistic controller, one can consider the external perturbation as a component of the underlying system. With this way, we consider the regret obtained by using the improved exploration separately. First noted that based on the definition ofOFUprinciple, StabL solves $J(\tilde{\Theta}_t) \leq \inf_{\Theta\in C_t(\delta)\cap\mathcal{S}} J(\Theta) + 1/\sqrt{t}$ to find the optimistic parameter. This search is done over only $C_t(\delta)$ in the stabilizing adaptive control phase. Denote the system evolution noise at time t as ζ_t. For $t \leq T_w$, system evolution noise can be considered as $\zeta_t = B_*\nu_t + w_t$ and for $t > T_w$, $\zeta_t = w_t$. Denote the optimal average cost of system

$\tilde{\Theta}$ under ζ_t as $J_*(\tilde{\Theta}, \zeta_t)$. The regret of the StabL can be decomposed as

$$\sum_{t=0}^{T} x_t^\top Q x_t + u_t^\top R u_t + 2\nu_t^\top R u_t + \nu_t^\top R \nu_t - J_*(\Theta_*, w_t), \tag{B.36}$$

where u_t is the optimal controller input for the optimistic system $\tilde{\Theta}_{t-1}$, ν_t is the noise injected and x_t is the state of the system $\tilde{\Theta}_{t-1}$ with the system evolution noise of ζ_t. From Bellman optimality equation for LQR, [28], we can write the following for the optimistic system, $\tilde{\Theta}_{t-1}$,

$$\begin{aligned} &J_*(\tilde{\Theta}_{t-1}, \zeta_t) + x_t^\top \tilde{P}_{t-1} x_t = x_t^\top Q x_t + u_t^\top R u_t \\ &+ \mathbb{E}\left[(\tilde{A}_{t-1} x_t + \tilde{B}_{t-1} u_t + \zeta_t)^\top \tilde{P}_{t-1} (\tilde{A}_{t-1} x_t + \tilde{B}_{t-1} u_t + \zeta_t) \middle| \mathcal{F}_{t-1}\right], \end{aligned}$$

where $\tilde{P}_{t-1}$ is the solution of DARE for $\tilde{\Theta}_{t-1}$. Following the decomposition used in without additional exploration [2], we get,

$$\begin{aligned} &J_*(\tilde{\Theta}_{t-1}, \zeta_t) + x_t^\top \tilde{P}_{t-1} x_t - (x_t^\top Q x_t + u_t^\top R u_t) \\ &= (\tilde{A}_{t-1} x_t + \tilde{B}_{t-1} u_t)^\top \tilde{P}_{t-1} (\tilde{A}_{t-1} x_t + \tilde{B}_{t-1} u_t) \\ &\quad + \mathbb{E}\left[x_{t+1}^\top \tilde{P}_{t-1} x_{t+1} \middle| \mathcal{F}_{t-1}\right] - (A_* x_t + B_* u_t)^\top \tilde{P}_{t-1} (A_* x_t + B_* u_t) \end{aligned}$$

where we use the fact that $x_{t+1} = A_* x_t + B_* u_t + \zeta_t$, the martingale property of the noise and the conditioning on the filtration $\mathcal{F}_{t-1}$. Hence, summing up over time, we get

$$\sum_{t=0}^{T} \left(x_t^\top Q x_t + u_t^\top R u_t\right) = \sum_{t=0}^{T} J_*(\tilde{\Theta}_{t-1}, \zeta_t) + R_1^\zeta - R_2^\zeta - R_3^\zeta$$

for

$$R_1^\zeta = \sum_{t=0}^{T} \left\{x_t^\top \tilde{P}_{t-1} x_t - \mathbb{E}\left[x_{t+1}^\top \tilde{P}_t x_{t+1} \middle| \mathcal{F}_{t-1}\right]\right\} \tag{B.37}$$

$$R_2^\zeta = \sum_{t=0}^{T} \mathbb{E}\left[x_{t+1}^\top \left(\tilde{P}_{t-1} - \tilde{P}_t\right) x_{t+1} \middle| \mathcal{F}_{t-1}\right] \tag{B.38}$$

$$R_3^\zeta = \sum_{t=0}^{T} \bar{x}_{t+1,\tilde{\Theta}_{t-1}}^\top \tilde{P}_{t-1} \bar{x}_{t+1,\tilde{\Theta}_{t-1}} - \bar{x}_{t+1,\Theta_*}^\top \tilde{P}_{t-1} \bar{x}_{t+1,\Theta_*} \tag{B.39}$$

where $\bar{x}_{t+1,\tilde{\Theta}_{t-1}} = \tilde{A}_{t-1} x_t + \tilde{B}_{t-1} u_t$ and $\bar{x}_{t+1,\Theta_*} = A_* x_t + B_* u_t$.

Therefore, when we jointly have that $\Theta_* \in C_t(\delta)$ for all time steps t and the state is bounded as shown in Lemma 3.5,

$$\sum_{t=0}^{T} (x_t^\top Q x_t + u_t^\top R u_t) = \sum_{t=0}^{T_w} \sigma_\nu^2 \operatorname{Tr}(\tilde{P}_{t-1} B_* B_*^\top) + \sum_{t=0}^{T} \bar{\sigma}_w^2 \operatorname{Tr}(\tilde{P}_{t-1}) + R_1^\zeta - R_2^\zeta - R_3^\zeta$$

where the equality follows from the fact that, $J_*(\tilde{\Theta}_{t-1}, \zeta_t) = \mathrm{Tr}(\tilde{P}_{t-1}W)$ where $W = \mathbb{E}[\zeta_t\zeta_t^\top|\mathcal{F}_{t-1}]$ for a corresponding filtration $\mathcal{F}_t$. The optimistic choice of $\tilde{\Theta}_t$ provides that

$$\bar{\sigma}_w^2 \,\mathrm{Tr}(\tilde{P}_{t-1}) = J_*(\tilde{\Theta}_{t-1}, w_t) \leq J_*(\Theta_*, w_t) + 1/\sqrt{t} = \bar{\sigma}_w^2 \,\mathrm{Tr}(P_*) + 1/\sqrt{t}.$$

Combining this with (B.36) and Assumption 3.2, we obtain the following expression for the regret of StabL :

$$\mathrm{R}(T) \leq \sigma_v^2 T_w D\|B_*\|_F^2 + R_1^\zeta - R_2^\zeta - R_3^\zeta + \sum_{t=0}^{T_w} 2\nu_t^\top R u_t + \nu_t^\top R \nu_t. \tag{B.40}$$

B.1.5 Regret Analysis, Proof of Theorem 3.2

In this section, we provide the bounds on each term in the regret decomposition separately. We show that the regret suffered from the improved exploration is tolerable in the upcoming stages via the guaranteed stabilizing controller, yielding polynomial dimension dependency in regret.

Direct Effect of Improved Exploration

The following gives an upper bound on the regret attained due to isotropic perturbations in the adaptive control with improved exploration phase of StabL.

Lemma B.4 (Direct Effect of Improved Exploration on Regret)**.** *If $\mathcal{E}_T \cap \mathcal{F}_{T_w}$ holds then with probability at least $1-\delta$,*

$$\sum_{t=0}^{T_w} \left(2\nu_t^\top R u_t + \nu_t^\top R \nu_t\right) \leq d\sigma_\nu \sqrt{B_\delta} + d\|R\|\sigma_\nu^2 \left(T_w + \sqrt{T_w}\log\frac{4dT_w}{\delta}\sqrt{\log\frac{4}{\delta}}\right), \tag{B.41}$$

where

$$B_\delta = 8\left(1 + T_w\kappa^2\|R\|^2(n+d)^{2(n+d)}\right)\log\left(\frac{4d}{\delta}\left(1 + T_w\kappa^2\|R\|^2(n+d)^{2(n+d)}\right)^{1/2}\right).$$

Proof. Let $q_t^\top = u_t^\top R$. The first term can be written as

$$2\sum_{t=0}^{T_w}\sum_{i=1}^{d} q_{t,i}\nu_{t,i} = 2\sum_{i=1}^{d}\sum_{t=0}^{T_w} q_{t,i}\nu_{t,i}.$$

Let $M_{t,i} = \sum_{k=0}^{t} q_{k,i}\nu_{k,i}$. From Theorem 1 of [3] on some event $G_{\delta,i}$ that holds with probability at least $1-\delta/(2d)$, for any $t \geq 0$,

$$M_{t,i}^2 \leq 2\sigma_\nu^2\left(1 + \sum_{k=0}^{t} q_{k,i}^2\right)\log\left(\frac{2d}{\delta}\left(1 + \sum_{k=0}^{t} q_{k,i}^2\right)^{1/2}\right).$$

On $\mathcal{E}_T \cap \mathcal{F}_{T_w}$, $\|q_k\| \leq \kappa\|R\|(n+d)^{n+d}$, thus $q_{k,i} \leq \kappa\|R\|(n+d)^{n+d}$. Using union bound we get, for probability at least $1 - \frac{\delta}{2}$,

$$\sum_{t=0}^{T_w} 2\nu_t^\top R u_t \leq$$
$$d\sqrt{8\sigma_\nu^2 \left(1 + T_w\kappa^2\|R\|^2(n+d)^{2(n+d)}\right) \log\left(\frac{4d}{\delta}\left(1 + T_w\kappa^2\|R\|^2(n+d)^{2(n+d)}\right)^{1/2}\right)}. \tag{B.42}$$

Let $W = \sigma_\nu\sqrt{2d\log\frac{4dT_w}{\delta}}$. Define $\Psi_t = \nu_t^\top R\nu_t - \mathbb{E}\left[\nu_t^\top R\nu_t|\mathcal{F}_{t-1}\right]$ and its truncated version $\tilde{\Psi}_t = \Psi_t\mathbb{I}_{\{\Psi_t \leq 2DW^2\}}$.

$$\mathbb{P}\left(\sum_{t=1}^{T_w}\Psi_t > 2\|R\|W^2\sqrt{2T_w\log\frac{4}{\delta}}\right) \leq$$
$$\mathbb{P}\left(\max_{1\leq t\leq T_w}\Psi_t > 2\|R\|W^2\right) + \mathbb{P}\left(\sum_{t=1}^{T_w}\tilde{\Psi}_t > 2\|R\|W^2\sqrt{2T_w\log\frac{4}{\delta}}\right).$$

Using standard sub-Gaussian vector norm bound with union bound and Azuma's inequality, the summation of terms on the right-hand side is bounded by $\delta/2$. Thus, with probability at least $1 - \delta/2$,

$$\sum_{t=0}^{T_w}\nu_t^\top R\nu_t \leq dT_w\sigma_\nu^2\|R\| + 2\|R\|W^2\sqrt{2T_w\log\frac{4}{\delta}}. \tag{B.43}$$

Combining (B.42) and (B.43) gives the statement of lemma for the regret of external exploration noise. □

Bounding R_1^ζ

In this section, we state an upper bound on R_1^ζ given in (B.37). We first provide a high probability bound on the system noise.

Lemma B.5 (Bounding sub-Gaussian vector)**.** *With probability* $1 - \frac{\delta}{8}$, $\|\zeta_k\| \leq (\sigma_w + \|B_*\|\sigma_\nu)\sqrt{2n\log\frac{8nT}{\delta}}$ *for* $k \leq T_w$ *and* $\|\zeta_k\| \leq \sigma_w\sqrt{2n\log\frac{8nT}{\delta}}$ *for* $T_w < k \leq T$.

Proof. From the sub-Gaussianity assumption, we have that for any index $1 \leq i \leq n$ and any time k, $|w_{k,i}| \leq \sigma_w\sqrt{2\log\frac{8}{\delta}}$ and $|(B_*\nu_k)_i| < \|B_*\|\sigma_\nu\sqrt{2\log\frac{8}{\delta}}$ with probability $1 - \frac{\delta}{8}$. Using the union bound, we get the statement of lemma. □

Using this we state the bound on R_1^ζ for stabilizable systems.

Lemma B.6 (Bounding R_1^ζ for StabL)**.** *Let R_1^ζ be as defined by (B.37). Under the event of $\mathcal{E}_T \cap \mathcal{F}_{T_w}$, with probability at least $1 - \delta/2$, using StabL for $t > T_r$, we have*

$$\begin{aligned} R_1 \le\ & k_{s,1}(n+d)^{n+d}(\sigma_w + \|B_*\|\sigma_v)n\sqrt{T_r}\log((n+d)T_r/\delta) \\ & + \frac{k_{s,2}(12\kappa^2 + 2\kappa\sqrt{2})}{\gamma}\sigma_w^2 n\sqrt{n}\sqrt{T - T_w}\log(n(t-T_w)/\delta) \\ & + k_{s,3}n\sigma_w^2\sqrt{T-T_w}\log(nT/\delta) + k_{s,4}n(\sigma_w + \|B_*\|\sigma_v)^2\sqrt{T_w}\log(nT/\delta), \end{aligned}$$

for some problem dependent coefficients $k_{s,1}, k_{s,2}, k_{s,3}, k_{s,4}$.

Proof. Assume that the event $\mathcal{E}_T \cap \mathcal{F}_{T_w}$ holds. Let $f_t = A_* x_t + B_* u_t$. One can decompose R_1 as

$$R_1 = x_0^\top P(\tilde{\Theta}_0)x_0 - x_{T+1}^\top P(\tilde{\Theta}_{T+1})x_{T+1} + \sum_{t=1}^{T} x_t^\top P(\tilde{\Theta}_t)x_t - \mathbb{E}\left[x_t^\top P(\tilde{\Theta}_t)x_t \middle| \mathcal{F}_{t-2}\right].$$

Since $P(\tilde{\Theta}_0)$ is positive semidefinite and $x_0 = 0$, the first two terms are bounded above by zero. The second term is decomposed as follows

$$\begin{aligned} & \sum_{t=1}^{T} x_t^\top P(\tilde{\Theta}_t)x_t - \mathbb{E}\left[x_t^\top P(\tilde{\Theta}_t)x_t \middle| \mathcal{F}_{t-2}\right] \\ & = \sum_{t=1}^{T} f_{t-1}^\top P(\tilde{\Theta}_t)\zeta_{t-1} + \sum_{t=1}^{T}\left(\zeta_{t-1}^\top P(\tilde{\Theta}_t)\zeta_{t-1} - \mathbb{E}\left[\zeta_{t-1}^\top P(\tilde{\Theta}_t)\zeta_{t-1}\middle|\mathcal{F}_{t-2}\right]\right). \end{aligned}$$

Let $R_{1,1} = \sum_{t=1}^{T} f_{t-1}^\top P(\tilde{\Theta}_t)\zeta_{t-1}$ and $R_{1,2} = \sum_{t=1}^{T}\left(\zeta_{t-1}^\top P(\tilde{\Theta}_t)\zeta_{t-1} - \mathbb{E}\left[\zeta_{t-1}^\top \tilde{P}_t \zeta_{t-1}\middle|\mathcal{F}_{t-2}\right]\right)$. Let $v_{t-1}^\top = f_{t-1}^\top P(\tilde{\Theta}_t)$. $R_{1,1}$ can be written as

$$R_{1,1} = \sum_{t=1}^{T}\sum_{i=1}^{n} v_{t-1,i}\zeta_{t-1,i} = \sum_{i=1}^{n}\sum_{t=1}^{T} v_{t-1,i}\zeta_{t-1,i}.$$

Let $M_{t,i} = \sum_{k=1}^{t} v_{k-1,i}\zeta_{k-1,i}$. By Theorem 1 of [3] on some event $G_{\delta,i}$ that holds with probability at least $1 - \delta/(4n)$, for any $t \ge 0$,

$$\begin{aligned} M_{t,i}^2 \le\ & 2(\sigma_w^2 + \|B_*\|^2\sigma_v^2)\left(1 + \sum_{k=1}^{T_r} v_{k-1,i}^2\right)\log\left(\frac{4n}{\delta}\left(1 + \sum_{k=1}^{T_r} v_{k-1,i}^2\right)^{1/2}\right) \\ & + 2\sigma_w^2\left(1 + \sum_{k=T_r+1}^{t} v_{k-1,i}^2\right)\log\left(\frac{4n}{\delta}\left(1 + \sum_{k=T_r+1}^{t} v_{k-1,i}^2\right)^{1/2}\right) \quad \text{for } t > T_r. \end{aligned}$$

Notice that StabL stops additional isotropic perturbation after $t = T_w$, and the state starts decaying until $t = T_r$. For simplicity of presentation we treat the time between T_w and T_r as exploration sacrificing the tightness of the result. On $\mathcal{E}_T \cap \mathcal{F}_{T_w}$, $\|v_k\| \leq DS(n+d)^{n+d}\sqrt{1+\kappa^2}$ for $k \leq T_r$ and $\|v_k\| \leq \frac{(12\kappa^2+2\kappa\sqrt{2})DS\sigma_w\sqrt{1+\kappa^2}}{\gamma}\sqrt{2n\log(n(t-T_w)/\delta)}$ for $k > T_r$. Thus, we have the same bounds for the individual components of v_k, i.e., $v_{k,i}$. Using union bound we get, for probability at least $1 - \frac{\delta}{4}$, for $t > T_r$,

$$R_{1,1} \leq n\sqrt{2(\sigma_w^2 + \|B_*\|^2\sigma_v^2)\left(1 + T_r D^2 S^2 (n+d)^{2(n+d)}(1+\kappa^2)\right)}\times \sqrt{\log\left(\frac{4n}{\delta}\left(1 + T_r D^2 S^2 (n+d)^{2(n+d)}(1+\kappa^2)\right)^{1/2}\right)} + n\sqrt{2\sigma_w^2\left(1 + \frac{2(t-T_r)(12\kappa^2+2\kappa\sqrt{2})^2 D^2 S^2 n\sigma_w^2(1+\kappa^2)}{\gamma^2}\log(n(T-T_w)/\delta)\right)}\times \sqrt{\log\left(\frac{4n}{\delta}\left(1 + \frac{2(t-T_r)(12\kappa^2+2\kappa\sqrt{2})^2 D^2 S^2 n\sigma_w^2(1+\kappa^2)}{\gamma^2}\log(n(T-T_w)/\delta)\right)\right)}.$$

Let $\mathcal{W}_{exp} = (\sigma_w + \|B_*\|\sigma_v)\sqrt{2n\log\frac{8nT}{\delta}}$ and $\mathcal{W}_{noexp} = \sigma_w\sqrt{2n\log\frac{8nT}{\delta}}$. Define $\Psi_t = \zeta_{t-1}^\top P(\tilde{\Theta}_t)\zeta_{t-1} - \mathbb{E}\left[\zeta_{t-1}^\top P(\tilde{\Theta}_t)\zeta_{t-1}|\mathcal{F}_{t-2}\right]$ and its truncated version $\tilde{\Psi}_t = \Psi_t \mathbb{I}_{\{\Psi_t \leq 2DW_{exp}^2\}}$ for $t \leq T_w$ and $\tilde{\Psi}_t = \Psi_t \mathbb{I}_{\{\Psi_t \leq 2DW_{noexp}^2\}}$ for $t > T_w$. Notice that $R_{1,2} = \sum_{t=1}^T \Psi_t$.

$$\mathbb{P}\left(\sum_{t=1}^{T_w}\Psi_t > 2DW_{exp}^2\sqrt{2T_w\log\frac{8}{\delta}}\right) + \mathbb{P}\left(\sum_{t=T_w+1}^{T}\Psi_t > 2DW_{noexp}^2\sqrt{2(T-T_w)\log\frac{8}{\delta}}\right)$$
$$\leq \mathbb{P}\left(\max_{1\leq t\leq T_w}\Psi_t > 2DW_{exp}^2\right) + \mathbb{P}\left(\max_{T_w+1\leq t\leq T}\Psi_t > 2DW_{noexp}^2\right)$$
$$+ \mathbb{P}\left(\sum_{t=1}^{T_w}\tilde{\Psi}_t > 2DW_{exp}^2\sqrt{2T_w\log\frac{8}{\delta}}\right) + \mathbb{P}\left(\sum_{t=T_w+1}^{T}\tilde{\Psi}_t > 2DW_{noexp}^2\sqrt{2(T-T_w)\log\frac{8}{\delta}}\right).$$

Using sub-Gaussian vector norm bound with union bound and Azuma's inequality, the summation of terms on the right-hand side is bounded by $\delta/4$. Thus, with probability at least $1 - \delta/4$, for $t > T_w$,

$$R_{1,2} \leq 4nD\sigma_w^2\sqrt{2(t-T_w)\log\frac{8}{\delta}}\log\frac{8nT}{\delta} + 4nD(\sigma_w + \|B_*\|\sigma_v)^2\sqrt{2T_w\log\frac{8}{\delta}}\log\frac{8nT}{\delta}.$$

Combining $R_{1,1}$ and $R_{1,2}$ gives the statement. □

Bounding R_2^ζ

In this section, we will bound $|R_2^\zeta|$ given in (B.38). We first provide a bound on the maximum number of policy changes.

Lemma B.7 (Number of Policy Changes for StabL)**.** *On the event of $\mathcal{E}_T \cap \mathcal{F}_{T_w}$, StabL changes the policy at most*

$$\min\left\{T/H_0, (n+d)\log_2\left(1+\frac{\mu+T_r(n+d)^{2(n+d)}(1+\kappa^2)+(T-T_r)(1+\kappa^2)X_s^2}{\mu}\right)\right\}, \tag{B.44}$$

where $X_s = \frac{(12\kappa^2+2\kappa\sqrt{2})\sigma_w}{\gamma}\sqrt{2n\log(n(T-T_w)/\delta)}$.

Proof. Changing policy K times up to time T_w requires $\det(V_T) \geq \mu^{n+d}2^K$. We also have that

$$\lambda_{\max}(V_T) \leq \mu + \sum_{t=0}^{T}\|z_t\|^2 \leq \mu + T_r(n+d)^{2(n+d)}(1+\kappa^2) + (T-T_r)(1+\kappa^2)X_s^2.$$

Thus, $\mu^{n+d}2^K \leq \left(\mu + T_r(n+d)^{2(n+d)}(1+\kappa^2) + (T-T_r)(1+\kappa^2)X_s^2\right)^{n+d}$. Solving for K gives

$$K \leq (n+d)\log_2\left(1+\frac{T_r(n+d)^{2(n+d)}(1+\kappa^2)+(T-T_r)(1+\kappa^2)X_s^2}{\mu}\right).$$

Moreover, the number of policy changes is also controlled by the lower bound H_0 on the duration of each controller. This policy update method would give at most T/H_0 policy changes. Since for the policy update of StabL requires both conditions to be met, the upper bound on the number of policy changes is minimum of these. □

Notice that besides the policy change instances, all the terms in R_2^ζ are 0. Therefore, we have the following results for stabilizable systems.

Lemma B.8 (Bounding R_2^ζ for StabL)**.** *Let R_2^ζ be as defined by* (B.38)*. Under the event of $\mathcal{E}_T \cap \mathcal{F}_{T_w}$, using StabL, we have*

$$\begin{aligned}|R_2^\zeta| \leq\ & 2D(n+d)^{2(n+d)+1}\log_2\left(1+\frac{T_r(n+d)^{2(n+d)}(1+\kappa^2)}{\mu}\right)\\ &+ 2DX_s^2(n+d)\log_2\left(1+\frac{T_r(n+d)^{2(n+d)}(1+\kappa^2)+(T-T_r)(1+\kappa^2)X_s^2}{\mu}\right)\end{aligned}$$

where $X_s = \frac{(12\kappa^2+2\kappa\sqrt{2})\sigma_w}{\gamma}\sqrt{2n\log(n(T-T_w)/\delta)}$.

Proof. On the event $\mathcal{E}_T \cap \mathcal{F}_{T_w}$, we know the maximum number of policy changes up to T_r and T using Lemma B.7. Using the fact that $\|x_t\| \leq (n+d)^{n+d}$ for $t \leq T_r$ and $\|x_t\| \leq \frac{(12\kappa^2+2\kappa\sqrt{2})\sigma_w}{\gamma}\sqrt{2n\log(n(t-T_w)/\delta)}$, we obtain the statement of the lemma. □

Bounding R_3^ζ

Before bounding R_3^ζ, first, consider the following for stabilizable LQRs.

Lemma B.9. *On the event of $\mathcal{E}_T \cap \mathcal{F}_{T_w}$, using StabL in a stabilizable LQR, the following holds,*

$$\sum_{t=0}^{T} \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2 \leq \frac{8(1+\kappa^2)\beta_T^2(\delta)}{\mu} \times \left((n+d)^{2(n+d)} \max\left\{2, \left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\mu}\right)^{H_0}\right\} \log\frac{\det(V_{T_r})}{\det(\mu I)} + X_s^2 \max\left\{2, \left(1 + \frac{(1+\kappa^2)X_s^2}{\mu}\right)^{H_0}\right\} \log\frac{\det(V_T)}{\det(V_{T_r})}\right)$$

where $X_s = \frac{(12\kappa^2+2\kappa\sqrt{2})\sigma_w}{\gamma}\sqrt{2n\log(n(t-T_w)/\delta)}$.

Proof. Let $s_t = (\Theta_* - \tilde{\Theta}_t)^\top z_t$ and $\tau \leq t$ be the time step that the last policy change happened. We have the following using triangle inequality,

$$\|s_t\| \leq \|(\Theta_* - \hat{\Theta}_t)^\top z_t\| + \|(\hat{\Theta}_t - \tilde{\Theta}_t)^\top z_t\|.$$

For all $\Theta \in C_\tau(\delta)$, for $\tau \leq T_r$, we have

$$\|(\Theta - \hat{\Theta}_t)^\top z_t\| \leq \|V_t^{1/2}(\Theta - \hat{\Theta}_t)\|\|z_t\|_{V_t^{-1}} \tag{B.45}$$

$$\leq \|V_\tau^{1/2}(\Theta - \hat{\Theta}_t)\|\sqrt{\frac{\det(V_t)}{\det(V_\tau)}}\|z_t\|_{V_t^{-1}} \tag{B.46}$$

$$\leq \max\left\{\sqrt{2}, \sqrt{\left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\mu}\right)^{H_0}}\right\}\|V_\tau^{1/2}(\Theta - \hat{\Theta}_t)\|\|z_t\|_{V_t^{-1}} \tag{B.47}$$

$$\leq \max\left\{\sqrt{2}, \sqrt{\left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\mu}\right)^{H_0}}\right\}\beta_\tau(\delta)\|z_t\|_{V_t^{-1}}. \tag{B.48}$$

Similarly, for for all $\Theta \in C_\tau(\delta)$, for $\tau > T_r$, we have

$$\|(\Theta - \hat{\Theta}_t)^\top z_t\| \leq \max\left\{\sqrt{2}, \sqrt{\left(1 + \frac{(1+\kappa^2)X_s^2}{\mu}\right)^{H_0}}\right\}\beta_\tau(\delta)\|z_t\|_{V_t^{-1}}.$$

Using these results, we obtain,

$$\sum_{t=0}^{T}\|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2$$
$$\leq 8\max\left\{2, \left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\mu}\right)^{H_0}\right\} \frac{\beta_T^2(\delta)(1+\kappa^2)(n+d)^{2(n+d)}}{\mu} \log\left(\frac{\det(V_{T_r})}{\det(\mu I)}\right)$$
$$+ 8\max\left\{2, \left(1 + \frac{(1+\kappa^2)X_s^2}{\mu}\right)^{H_0}\right\} \frac{\beta_T^2(\delta)(1+\kappa^2)X_s^2}{\mu} \log\left(\frac{\det(V_T)}{\det(V_{T_r})}\right),$$

where we use Lemma 11 of [3]. □

Using Lemma B.9, we bound R_3^ζ as follows.

Lemma B.10 (Bounding R_3^ζ for StabL)**.** *Let R_3^ζ be as defined by* (B.39)*. Under the event of $\mathcal{E}_T \cap \mathcal{F}_{T_w}$, using StabL with the choice of $\mu = (1+\kappa^2)X_s^2$, we have*

$$|R_3^\zeta| = O\left((n+d)^{(H_0+2)(n+d)+2}\sqrt{n}\sqrt{T_r} + (n+d)n\sqrt{T-T_r}\right).$$

Proof. Let $Y_1 = \frac{8(1+\kappa^2)\beta_T^2(\delta)}{\mu}(n+d)^{2(n+d)}\max\left\{2, \left(1 + \frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\mu}\right)^{H_0}\right\}\log\frac{\det(V_{T_r})}{\det(\mu I)}$ and $Y_2 = \frac{8(1+\kappa^2)\beta_T^2(\delta)}{\mu}X_s^2\max\left\{2, \left(1 + \frac{(1+\kappa^2)X_s^2}{\mu}\right)^{H_0}\right\}\log\frac{\det(V_T)}{\det(V_{T_r})}$, where we define X_s as $X_s = \frac{(12\kappa^2+2\kappa\sqrt{2})\sigma_w}{\gamma}\sqrt{2n\log(n(t-T_w)/\delta)}$. The following uses triangle inequality

and Cauchy Schwarz inequality and again triangle inequality to give:

$$
\begin{aligned}
\left|R_3^{\zeta}\right| &\leq \sum_{t=0}^{T}\left|\left\|P(\tilde{\Theta}_t)^{1/2}\tilde{\Theta}_t^{\top}z_t\right\|^2-\left\|P(\tilde{\Theta}_t)^{1/2}\Theta_*^{\top}z_t\right\|^2\right| \\
&=\sum_{t=0}^{T_r}\left|\left\|P(\tilde{\Theta}_t)^{1/2}\tilde{\Theta}_t^{\top}z_t\right\|^2-\left\|P(\tilde{\Theta}_t)^{1/2}\Theta_*^{\top}z_t\right\|^2\right|+\sum_{t=T_r}^{T}\left|\left\|P(\tilde{\Theta}_t)^{1/2}\tilde{\Theta}_t^{\top}z_t\right\|^2-\left\|P(\tilde{\Theta}_t)^{1/2}\Theta_*^{\top}z_t\right\|^2\right| \\
&\leq\left(\sum_{t=0}^{T_r}\left(\left\|P(\tilde{\Theta}_t)^{1/2}\tilde{\Theta}_t^{\top}z_t\right\|-\left\|P(\tilde{\Theta}_t)^{1/2}\Theta_*^{\top}z_t\right\|\right)^2\right)^{1/2}\left(\sum_{t=0}^{T_r}\left(\left\|P(\tilde{\Theta}_t)^{1/2}\tilde{\Theta}_t^{\top}z_t\right\|+\left\|P(\tilde{\Theta}_t)^{1/2}\Theta_*^{\top}z_t\right\|\right)^2\right)^{1/2} \\
&+\left(\sum_{t=T_r}^{T}\left(\left\|P(\tilde{\Theta}_t)^{1/2}\tilde{\Theta}_t^{\top}z_t\right\|-\left\|P(\tilde{\Theta}_t)^{1/2}\Theta_*^{\top}z_t\right\|\right)^2\right)^{1/2}\left(\sum_{t=T_r}^{T}\left(\left\|P(\tilde{\Theta}_t)^{1/2}\tilde{\Theta}_t^{\top}z_t\right\|+\left\|P(\tilde{\Theta}_t)^{1/2}\Theta_*^{\top}z_t\right\|\right)^2\right)^{1/2} \\
&\leq\left(\sum_{t=0}^{T_r}\left\|P(\tilde{\Theta}_t)^{1/2}\left(\tilde{\Theta}_t-\Theta_*\right)^{\top}z_t\right\|^2\right)^{1/2}\left(\sum_{t=0}^{T_r}\left(\left\|P(\tilde{\Theta}_t)^{1/2}\tilde{\Theta}_t^{\top}z_t\right\|+\left\|P(\tilde{\Theta}_t)^{1/2}\Theta_*^{\top}z_t\right\|\right)^2\right)^{1/2} \\
&\quad+\left(\sum_{t=T_r}^{T}\left\|P(\tilde{\Theta}_t)^{1/2}\left(\tilde{\Theta}_t-\Theta_*\right)^{\top}z_t\right\|^2\right)^{1/2}\left(\sum_{t=T_r}^{T}\left(\left\|P(\tilde{\Theta}_t)^{1/2}\tilde{\Theta}_t^{\top}z_t\right\|+\left\|P(\tilde{\Theta}_t)^{1/2}\Theta_*^{\top}z_t\right\|\right)^2\right)^{1/2} \\
&\leq\sqrt{Y_1}\sqrt{4T_rD(1+\kappa^2)S^2(n+d)^{2(n+d)}}+\sqrt{Y_2}\sqrt{4(T-T_r)D(1+\kappa^2)S^2X_s^2} \\
&\leq\frac{\max\left\{8,4\sqrt{2}\left(1+\frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\mu}\right)^{H_0/2}\right\}DS(1+\kappa^2)\beta_T(\delta)(n+d)^{2(n+d)}}{\sqrt{\mu}}\times \\
&\qquad\sqrt{T_r(n+d)\log\left(1+\frac{T_r(1+\kappa^2)(n+d)^{2(n+d)}}{\mu(n+d)}\right)} \\
&+\frac{\max\left\{8,4\sqrt{2}\left(1+\frac{(1+\kappa^2)X_s^2}{\mu}\right)^{H_0/2}\right\}DS(1+\kappa^2)\beta_T(\delta)}{\sqrt{\mu}}X_s^2\times \\
&\qquad\sqrt{(T-T_r)(n+d)\log\left(1+\frac{T_r(1+\kappa^2)(n+d)^{2(n+d)}+(T-T_r)X_s^2}{\mu(n+d)}\right)}.
\end{aligned}
$$

Examining the first term, it has the dimension dependency of $(n+d)^{(n+d)H_0}\times\sqrt{n(n+d)}\times(n+d)^{2(n+d)}\times\sqrt{n+d}$ where $\sqrt{n(n+d)}$ is due to $\beta_T(\delta)$. For the second term, with the choice of $\mu=(1+\kappa^2)X_s^2$, the exponential dependency on the dimension with H_0 can be converted to a scalar multiplier, i.e., $\left(1+\frac{(1+\kappa^2)X_s^2}{\mu}\right)^{H_0/2}=\sqrt{2}^{H_0}$ and $(1+\kappa^2)X_s^2/\sqrt{\mu}=\sqrt{(1+\kappa^2)}X_s$. Therefore, for the second term, we have the dimension dependency of $\sqrt{n(n+d)}\times\sqrt{n}\times\sqrt{n+d}$ which gives the advertised bound. □

Combining Terms for Final Regret Upper Bound

Proof of Theorem 3.2: Recall that

$$\text{REGRET}(T) \leq \sigma_v^2 T_w D \|B_*\|_F^2 + \sum_{t=0}^{T_w} \left(2v_t^\top R u_t + v_t^\top R v_t\right) + R_1^\zeta - R_2^\zeta - R_3^\zeta.$$

Combining Lemma B.4 for $\sum_{t=0}^{T_w} \left(2v_t^\top R u_t + v_t^\top R v_t\right)$, Lemma B.6 for R_1^ζ, Lemma B.8 for $|R_2^\zeta|$ and Lemma B.10 for $|R_3^\zeta|$, we get the advertised regret bound. □

B.2 Proofs of Section 3.3

In Appendix B.2.1, we provide the system identification and stabilization guarantees of TSAC. In particular, we give the proof of Lemma 3.6 and give the precise duration of the TS with improved exploration phase T_w. In Appendix B.2.2, we provide the complete proof of Theorem 3.4, as well as the intermediate results discussed in the main text. In Appendix B.2.3, we provide the precise regret decomposition and discuss the individual terms in the regret upper bound. Appendix B.2.4 comprises the analysis of individual terms in the regret decomposition. Before proceeding the next section, we define the following high probability events which are standard in TS-based algorithms. First recall the RLS confidence ellipsoid given in (3.8):

$$\mathcal{E}_t^{\text{RLS}}(\delta) = \{\Theta : \|\Theta - \hat{\Theta}_t\|_{V_t} \leq \beta_t(\delta)\},$$

for $\beta_t(\delta) = \sigma_w \sqrt{2n \log((T \det(V_t)^{1/2})/(\delta \det(\mu I)^{1/2}))} + \sqrt{\mu} S$. Further define

$$\mathcal{E}_t^{\text{TS}}(\delta) = \{\Theta : \|\Theta - \hat{\Theta}_t\|_{V_t} \leq \upsilon_t(\delta)\},$$

for $\upsilon_t(\delta) = \beta_t(\delta) n \sqrt{(n+d)\log(n(n+d)/\delta)}$. Define the events

$$\hat{E}_t = \{\forall s \leq t, \Theta_* \in \mathcal{E}_t^{\text{RLS}}(\delta)\} \tag{B.49}$$

$$\tilde{E}_t = \{\forall s \leq t, \tilde{\Theta}_s \in \mathcal{E}_t^{\text{TS}}(\delta)\}. \tag{B.50}$$

As described in Section 3.3.2, $\hat{E}_t$ defines the event that RLS estimates $\hat{\Theta}_t$ concentrate around Θ_* and $\tilde{E}_t$ defines the event that the sampled model parameter concentrates around $\hat{\Theta}_t$. From standard Gaussian tail bound and the self-normalized estimation error, we have that $\hat{E} \cap \tilde{E}$ for all $t \leq T$, with probability at least $1 - 2\delta$. Here the time dependency dropped since $\hat{E} := \hat{E}_T \subset \ldots \subset \hat{E}_1$ and $\tilde{E} := \tilde{E}_T \subset \ldots \subset \tilde{E}_1$. These events will be key in providing all the technical results starting from stabilization guarantees to final regret upper bound.

B.2.1 System Identification and Stabilization Guarantees, Proof of Lemma 3.6

In this section, we show that improved exploration of TSAC provides persistently exciting inputs, which will be used to enable reaching a stabilizing neighborhood around Θ_*. Note that in Appendix B.1.1, we studied the same problem for a more general sub-Gaussian process noise setting. Here we provide the special case for Gaussian process noise. From the Gaussian process noise assumption, we have that $\mathbb{E}[x_{t+1}x_{t+1}^\top \mid \mathcal{F}_t] \succcurlyeq \sigma_w^2 I$. Thus, with the input $u_t = K(\tilde{\Theta}_t)x_t + \nu_t$ for $\nu_t \sim \mathcal{N}(0, 2\kappa^2\sigma_w^2 I)$, we have that $\mathbb{E}[z_{t+1}z_{t+1}^\top \mid \mathcal{F}_t] \succcurlyeq \frac{\sigma_w^2}{2} I$. Using Theorem 20 of Cohen et al. [62], we have that $V_t \succcurlyeq t\frac{\sigma_w^2}{40} I$ for $t \geq 200(n+d)\log\frac{12}{\delta}$ with probability at least $1-\delta$. Using the RLS estimate error bound given in (3.8), i.e., under the event of $\hat{E}_t$ we have

$$\|\hat{\Theta}_t - \Theta_*\|_2 \leq \frac{\beta_t}{\sqrt{\lambda_{\min}(V_t)}}, \tag{B.51}$$

with probability at least $1-\delta$. Plugging in the $\lambda_{\min}(V_t)$ in its place yields the first result of Lemma 3.6. For the second result, we use Lemma 3.3. Recall that $D = \overline{\alpha}\gamma^{-1}\kappa^2(1+\kappa^2)$. Lemma 3.3 states that for any (κ,γ)-stabilizable system Θ_* and for any $\varepsilon \leq \min\{\sqrt{(\sigma_w^2 n)/(142D^7)}, 1/(54D^5\}$, such that $\|\Theta' - \Theta_*\| \leq \varepsilon$, $K(\Theta')$ produces $(\kappa\sqrt{2}, \gamma/2)$-stable closed-loop dynamics on Θ_* such that there exists L and $H \succ 0$ such that $A_* + B_* K(\Theta') = H'LH'^{-1}$, with $\|L\| \leq 1-\gamma/2$ and $\|H'\|\|H'^{-1}\| \leq \kappa\sqrt{2}$. Under the event of $\hat{E} \cap \tilde{E}$, we have $\|\tilde{\Theta}_t - \Theta_*\|_2 \leq \frac{\beta_t(\delta)+\upsilon_t(\delta)}{\sqrt{\lambda_{\min}(V_T)}}$. Under the event of $\hat{E}\cap\tilde{E}$, this yields $\|\tilde{\Theta}_t - \Theta_*\|_2 \leq \frac{7(\beta_t(\delta)+\upsilon_t(\delta))}{\sigma_w\sqrt{t}}$ with probability $1-\delta$. Combining this result with the required ε for finding the stabilizing neighborhood, for TS with the exploration duration of

$$T_w \geq T_0 := \frac{49(\beta_T(\delta)+\upsilon_T(\delta))^2}{\sigma_w \min\{(\sigma_w^2 n)/(142D^7), 1/(54^2 D^{10}\}}, \tag{B.52}$$

TSAC achieves $(\kappa\sqrt{2}, \gamma/2)$-stable closed-loop dynamics on Θ_*, with probability at least $1-3\delta$.

B.2.2 Probability of Sampling Optimistic Models, Proof of Theorem 3.4

In this section, we give proof of the main technical contribution for TSAC, showing that TS samples optimistic model parameters with constant probability (Theorem 3.4). The proof follows the outline provided in Section 3.3.3. We first provide the proofs of each lemma in Section 3.3.3. Finally, we combine these results to prove Theorem 3.4.

Proof of Lemma 3.8

Given a stabilizable system $\Theta = (A, B)^\top$, and a stabilizing linear feedback controller K, we can find the LQR cost as follows

$$J(\Theta, K) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} x_t^\top Q x_t + u_t^\top R u_t\right], \tag{B.53}$$

$$= \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} \operatorname{tr}\left((Q + K^\top R K) x_t x_t^\top\right)\right], \tag{B.54}$$

$$= \lim_{T\to\infty} \operatorname{tr}\left((Q + K^\top R K) \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[x_t x_t^\top\right]\right), \tag{B.55}$$

$$= \operatorname{tr}\left((Q + K^\top R K) \Sigma(\Theta, K)\right), \tag{B.56}$$

where $\Sigma(\Theta, K) \coloneqq \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[x_t x_t^\top\right]$ is the stationary state covariance of the closed-loop system. In (B.54), we used the feedback control policy relation $u_t = K x_t$ and trace trick for inner products of vectors. Note that the closed-loop system evolves as

$$x_{t+1} = (A + BK) x_t + w_t. \tag{B.57}$$

The covariance of the state at time t can be written as a recursive relation

$$\mathbb{E}\left[x_{t+1} x_{t+1}^\top\right] = \mathbb{E}\left[((A + BK) x_t + w_t)((A + BK) x_t + w_t)^\top\right] \tag{B.58}$$

$$= (A + BK)\, \mathbb{E}\left[x_t x_t^\top\right] (A + BK)^\top + \mathbb{E}\left[w_t w_t^\top\right] \tag{B.59}$$

$$= (A + BK)\, \mathbb{E}\left[x_t x_t^\top\right] (A + BK)^\top + \sigma_w^2 I, \tag{B.60}$$

where (B.59) is because $\mathbb{E}[w_t] = 0$ and w_t and x_t are independent. Since $\rho(A + BK) < 1$, the above iteration converges to a finite fixed point. Furthermore, we have the following relation

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[x_{t+1} x_{t+1}^\top\right] = (A + BK) \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[x_t x_t^\top\right] (A + BK)^\top + \sigma_w^2 I. \tag{B.61}$$

Denoting by $\Sigma_T(\Theta, K) \coloneqq \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[x_t x_t^\top\right]$ the finite averaged state covariance, we have the following

$$\Sigma_T(\Theta, K) + \frac{\mathbb{E}\left[x_{T+1} x_{T+1}^\top\right] - \mathbb{E}\left[x_1 x_1^\top\right]}{T} = (A + BK) \Sigma_T(\Theta, K) (A + BK)^\top + \sigma_w^2 I.$$

Taking the limit of both sides as $T \to \infty$ and noting that $\mathbb{E}\left[x_{T+1} x_{T+1}^\top\right]$ has a finite value at the limit, we obtain the following Lyapunov equation

$$\Sigma(\Theta, K) = (A + BK) \Sigma(\Theta, K) (A + BK)^\top + \sigma_w^2 I,$$

whose solution is given by the following convergent infinite sum

$$\Sigma(\Theta, K) = \sum_{t=0}^{\infty} (A + BK)^t \sigma_w^2 I \left((A + BK)^\top\right)^t .$$

It is well known that the optimal control policy of infinite-horizon LQR systems can be achieved by stationary linear feedback controllers [28]. Therefore, we can find the optimal LQR cost of a stabilizable system by minimizing its closed-loop cost among all stabilizing stationary linear feedback controllers.

Suppose $\Theta \in \mathcal{S}^{\text{surr}}$, i.e., $J(\Theta, K(\Theta_*)) \leq J(\Theta_*, K(\Theta_*))$. Then, the optimal LQR cost of Θ is given as

$$J(\Theta) = J(\Theta, K(\Theta)) = \min_{K \in \mathbb{R}^{d \times n}} J(\Theta, K) \tag{B.62}$$

$$\leq J(\Theta, K(\Theta_*)) \overset{(a)}{\leq} J(\Theta_*, K(\Theta_*)) = J(\Theta_*), \tag{B.63}$$

where (a) is due to $\Theta \in \mathcal{S}^{\text{surr}}$. Thus, $\Theta \in \mathcal{S}^{\text{opt}}$. □

Proof of Lemma 3.9

The following lemma will be used as the backbone for Lemma 3.9.

Lemma B.11. *Let $V_1, V_2 \in \mathbb{R}^{n \times n}$ be symmetric positive semi-definite matrices. Define two ellipsoids as*

$$\mathcal{E}_1 \coloneqq \left\{\Theta \in \mathbb{R}^{n \times m} \;\middle|\; \operatorname{tr}\left(\Theta^\top V_1 \Theta\right) \leq 1\right\} \quad \text{and} \quad \mathcal{E}_2 \coloneqq \left\{\Theta \in \mathbb{R}^{n \times m} \;\middle|\; \operatorname{tr}\left(\Theta^\top V_2 \Theta\right) \leq 1\right\}. \tag{B.64}$$

Then, $\mathcal{E}_1 \subseteq \mathcal{E}_2$ if and only if $V_1 \succcurlyeq V_2$.

Proof. For the forward direction, assume $V_1 - V_2$ has a negative eigenvalue, i.e., there exist $\lambda < 0$ and a unit vector $\theta \in \mathbb{R}^n / \{0\}$ such that $(V_1 - V_2)\theta = \lambda\theta$. Construct $\Theta = [\theta, \theta, \ldots, \theta] \in \mathbb{R}^{n \times m}$. Observe that $\operatorname{tr}(\Theta^\top V_1 \Theta) = m\theta^\top V_1 \theta$ and $\operatorname{tr}(\Theta^\top V_2 \Theta) = m\theta^\top V_2 \theta$. Therefore, we have the relationship $\operatorname{tr}(\Theta^\top V_2 \Theta) = \operatorname{tr}(\Theta^\top V_1 \Theta) - m\lambda$.

If $V_1\theta = 0$, then $\operatorname{tr}(\Theta^\top V_1 \Theta) = 0 \leq 1$ and therefore for any scalar $\alpha > 0$, $\alpha\Theta \in \mathcal{E}_1$. On the other hand, $\operatorname{tr}(\Theta^\top V_2 \Theta) = -m\lambda > 0$ and therefore, one can find a scalar $\alpha > 0$ such that $\operatorname{tr}((\alpha\Theta)^\top V_2 (\alpha\Theta)) = -m\lambda\alpha^2 > 1$, i.e., $\alpha\Theta \notin \mathcal{E}_2$. If $V_1\theta \neq 0$, then define $\Theta' = \frac{1}{\sqrt{m\theta^\top V_1 \theta}}\Theta$ and observe that $\operatorname{tr}\left(\Theta'^\top V_1 \Theta'\right) = 1$, i.e., $\Theta' \in \mathcal{E}_1$. On the other hand, $\operatorname{tr}\left(\Theta'^\top V_2 \Theta'\right) = 1 - \frac{\lambda}{\theta^\top V_1 \theta} > 1$, i.e., $\Theta' \notin \mathcal{E}_2$. Therefore, we have that if $\mathcal{E}_1 \subseteq \mathcal{E}_2$ then $V_1 \succcurlyeq V_2$.

For the reverse direction, assume that $V_1 \succcurlyeq V_2$ and $\Theta \in \mathcal{E}_1$. Then, $\operatorname{tr}(\Theta^\top (V_1 - V_2)\Theta) \geq 0$ and $\operatorname{tr}(\Theta^\top V_2 \Theta) \leq \operatorname{tr}(\Theta^\top V_1 \Theta) \leq 1$. Therefore, $\Theta \in \mathcal{E}_2$. □

Proof of Lemma 3.9. Let us rewrite the ellipsoids. For the time being, we will drop δ dependence for simplicity.

$$\mathcal{E}_t^{\mathrm{RLS}} = \left\{ \hat{\Theta} \in \mathbb{R}^{(n+d)\times n} \;\middle|\; \operatorname{tr}\left((\hat{\Theta} - \Theta_*)^\top \beta_t^{-1} V_t (\hat{\Theta} - \Theta_*) \right) \leq 1 \right\}, \tag{B.65}$$

$$\mathcal{E}_t^{\mathrm{cl}} = \left\{ \hat{\Theta} \in \mathbb{R}^{(n+d)\times n} \;\middle|\; \operatorname{tr}\left((\hat{\Theta} - \Theta_*)^\top H_* F_t^{-1} H_*^\top (\hat{\Theta} - \Theta_*) \right) \leq 1 \right\}. \tag{B.66}$$

To prove the lemma, it is necessary and sufficient to show $\beta_t^{-1} V_t \succcurlyeq H_* F_t^{-1} H_*^\top$ by Lemma B.11. Eliminating b_t terms from both sides and multiplying by $V_t^{-\frac{1}{2}}$ from left and right, we obtain the equivalent condition,

$$I \succcurlyeq V_t^{-\frac{1}{2}} H_* (H_*^\top V_t^{-1} H_*)^{-1} H_*^\top V_t^{-\frac{1}{2}} = V_t^{-\frac{1}{2}} H_* (H_*^\top V_t^{-1} H_*)^{-\frac{1}{2}} (H_*^\top V_t^{-1} H_*)^{-\frac{1}{2}} H_*^\top V_t^{-\frac{1}{2}}.$$

In other words, we have that $\mathcal{E}_t^{\mathrm{RLS}} \subseteq \mathcal{E}_t^{\mathrm{cl}}$ if and only if $\| (H_*^\top V_t^{-1} H_*)^{-\frac{1}{2}} H_*^\top V_t^{-\frac{1}{2}} \|_2 \leq 1$. Notice that

$$\| (H_*^\top V_t^{-1} H_*)^{-\frac{1}{2}} H_*^\top V_t^{-\frac{1}{2}} \|_2^2 = \sigma_1 \left((H_*^\top V_t^{-1} H_*)^{-\frac{1}{2}} H_*^\top V_t^{-\frac{1}{2}} \right)^2, \tag{B.67}$$

$$= \lambda_{\max} \left(V_t^{-\frac{1}{2}} H_* (H_*^\top V_t^{-1} H_*)^{-1} H_*^\top V_t^{-\frac{1}{2}} \right), \tag{B.68}$$

$$= \lambda_{\max} \left((H_*^\top V_t^{-1} H_*)^{-\frac{1}{2}} H_*^\top V_t^{-1} H_* (H_*^\top V_t^{-1} H_*)^{-\frac{1}{2}} \right), \tag{B.69}$$

$$= \lambda_{\max}(I) = 1, \tag{B.70}$$

where we used the fact that $\sigma_1(A) = \sqrt{\lambda_{\max}(A^\top A)} = \sqrt{\lambda_{\max}(A A^\top)}$. This is true for any time t and δ and therefore completes the proof. □

Proof of Lemma 3.10

The following lemma guarantees the existence of a stable neighborhood around any stable matrix.

Lemma B.12. *Let $A_c \in \mathcal{M}_{Schur}$, i.e., $\rho(A_c) < 1$. Then, there exists $\epsilon > 0$ such that for any $\Delta \in \mathcal{M}_n$ with $\|\Delta\|_F \leq 1$, we have that $A_c + \epsilon\Delta \in \mathcal{M}_{Schur}$, i.e., $\rho(A_c + \epsilon\Delta) < 1$.*

Proof. Per Gelfand's formula, we have that for any $\delta > 0$, there exists $N_\delta \in \mathbb{N}$ such that

$$\rho(A_c) \leq \|A_c^k\|_F^{1/k} < \rho(A_c) + \delta, \tag{B.71}$$

for any $k \geq N_\delta$. Since the mapping $A_c \mapsto \|A_c^k\|_F^{1/k}$ is smooth for any $k \in \mathbb{N}$, we can write the following expansion by Taylor's theorem for any $t \in \mathbb{R}$,

$$\|(A_c + t\Delta)^k\|_F^{1/k} = \|A_c^k\|_F^{1/k} + t\frac{\mathrm{d}}{\mathrm{d}t}\|(A_c + t\Delta)^k\|_F^{1/k}\Big|_{\lambda t}, \tag{B.72}$$

where $\lambda \in [0, 1]$. For a given $t \in \mathbb{R}$, there exists a constant $M_{k,t} > 0$ such that for any $\|\Delta\|_F \leq 1$, we have that $\left|\frac{\mathrm{d}}{\mathrm{d}t}\|(A_c + t\Delta)^k\|_F^{1/k}\right| \leq M_{k,t}$ by Taylor's theorem. Then, we can write the following upper bound

$$\|(A_c + t\Delta)^k\|_F^{1/k} \leq \|A_c^k\|_F^{1/k} + |t|\, M_{t,k}. \tag{B.73}$$

Using the relation (B.71) and the upper bound (B.73), we have that for any $\delta > 0$, $t > 0$, and $\|\Delta\|_F \leq 1$, there exists $N_\delta \in \mathbb{N}$ and $M_{t,N_\delta} > 0$ such that

$$\rho(A_c + t\Delta) \leq \|(A_c + t\Delta)^{N_\delta}\|_F^{1/N_\delta} \leq \|A_c^{N_\delta}\|_F^{1/N_\delta} + tM_{t,N_\delta} \tag{B.74}$$

$$< \rho(A_c) + \delta + tM_{t,N_\delta} \tag{B.75}$$

Fix a $\delta > 0$ such that $\rho(A_c) + \delta < 1$ and fix a $t > 0$. Then, we can find $0 < \epsilon \leq t$ such that $\rho(A_c) + \delta + \epsilon M_{t,N_\delta} < 1$ and thus

$$\rho(A_c + \epsilon\Delta) < \rho(A_c) + \delta + \epsilon M_{t,N_\delta} < 1 \tag{B.76}$$

for any $\|\Delta\|_F \leq 1$ by (B.75). □

Proof of Lemma 3.10. For any $A_c \in \mathcal{M}_{\text{Schur}}$, there exists a constant $\epsilon > 0$, such that for any $\|\Delta\|_F \leq 1$, we have that $A_c + \epsilon\Delta \in \mathcal{M}_{\text{Schur}}$ by Lemma B.12. To see smoothness of L, we write $A_t \coloneqq A_c + t\Delta$ and $L(A_t) = \operatorname{tr}(Q_*\Sigma_t)$ for any $|t| \leq \epsilon$ and $\|\Delta\|_F \leq 1$ where Σ_t solves the following Lyapunov equation:

$$\Sigma_t - A_t\Sigma_t A_t^\top = \sigma_w^2 I \text{ and } \Sigma_0 - A_c\Sigma_0 A_c^\top = \sigma_w^2 I. \tag{B.77}$$

Note that, $\rho(A_t) < 1$ for any $|t| \leq \epsilon$ and therefore both equations in (B.77) have unique solutions for any $|t| \leq \epsilon$. The Jacobian $\nabla L(A_c) \in \mathcal{M}_n$ satisfies $\nabla L(A_c) \bullet \Delta = \frac{\mathrm{d}}{\mathrm{d}t}L(A_t)\big|_{t=0} = \operatorname{tr}(Q_*\dot{\Sigma}_0)$ for any $\|\Delta\|_F \leq 1$ where $\dot{\Sigma}_t$ is the derivative of Σ_t and satisfies the following Lyapunov equation:

$$\dot{\Sigma}_t - A_t\dot{\Sigma}_t A_t^\top = \Delta\Sigma_t A_t^\top + A_t\Sigma_t\Delta^\top \text{ and } \dot{\Sigma}_0 - A_c\dot{\Sigma}_0 A_c^\top = \Delta\Sigma_0 A_c^\top + A_c\Sigma_0\Delta^\top. \tag{B.78}$$

Similarly, both equations in (B.78) have unique solutions for any $|t| \leq \epsilon$ and therefore $\nabla L(A_c)$ exists for any A_c. To find the Jacobian, we have that $\dot{\Sigma}_0 =$

$\sum_{k=0}^{\infty} A_c^k \left(\Delta\Sigma_0 A_c^\top + A_c\Sigma_0\Delta^\top\right)(A_c^\top)^k$ and

$$\operatorname{tr}(Q_*\dot{\Sigma}_0) = \operatorname{tr}\left(Q_* \sum_{k=0}^{\infty} A_c^k \left(\Delta\Sigma_0 A_c^\top + A_c\Sigma_0\Delta^\top\right)(A_c^\top)^k\right) \tag{B.79}$$

$$= 2\operatorname{tr}\left(\sum_{k=0}^{\infty}(A_c^\top)^k Q_* A_c^k A_c\Sigma_0\Delta^\top\right) = 2\sum_{k=0}^{\infty}(A_c^\top)^k Q_* A_c^k A_c\Sigma_0 \bullet \Delta. \tag{B.80}$$

Therefore, $\nabla L(A_c) = 2\sum_{k=0}^{\infty}(A_c^\top)^k Q_* A_c^k A_c\Sigma_0$. In particular, in the case of $A_{c,*}$, we have that $\sum_{k=0}^{\infty}(A_{c,*}^\top)^k Q_* A_{c,*}^k = P_*$, the solution to the Riccati equation, and thus $\nabla L(A_{c,*}) = 2P_* A_{c,*}\Sigma_*$. Repeating the same process, one can see that $L(A_t)$ is infinitely differentiable and thus we conclude L is a smooth function.

Denote by $\mathcal{B}_\epsilon := \{A \in \mathcal{M}_n \mid \|A - A_c\|_F \le \epsilon\} \subset \mathcal{M}_{\text{Schur}}$ the ball of radius $\epsilon > 0$ around $A_c \in \mathcal{M}_{\text{Schur}}$. Consider the function L restricted to the domain $\mathcal{B}_\epsilon$. Since $\mathcal{B}_\epsilon$ is a convex set, we can apply Taylor's theorem to L around A_c in this domain to obtain

$$L(A_c + \epsilon\Delta) = L(A_c) + \nabla L(A_c) \bullet \epsilon\Delta + \frac{1}{2}\epsilon\Delta \bullet \mathcal{H}_{A_c+s\Delta}(\epsilon\Delta), \tag{B.81}$$

for $\|\Delta\|_F \le 1$ and for some $s \in [0, \epsilon]$. Here, $\mathcal{H}_{A_c} : \mathcal{M}_n \to \mathcal{M}_n$ is the Hessian operator evaluated at a point $A_c \in \mathcal{M}_{\text{Schur}}$ and satisfies the following relationship

$$\Delta \bullet \mathcal{H}_{A_c}(\Delta) = \frac{\mathrm{d}^2}{\mathrm{d}t^2} L(A_c + t\Delta)\Big|_{t=0}, \tag{B.82}$$

for any $\|\Delta\|_F \le 1$. Finally, there exists a constant $r > 0$, such that for any $G \in \mathcal{M}_n$, we have that $\left|G \bullet \mathcal{H}_{A_c+s\Delta}(G)\right| \le r\|G\|_F^2$ for any $s \in [0, \epsilon]$ and $\|\Delta\|_F \le 1$ by Taylor's theorem. □

Proof of Lemma 3.11

First, we need to show the boundedness of z_t.

Lemma B.13. *Define the terms*

$$Z'_{T_w} := (1+\kappa)c'(n+d)^{n+d} + \kappa\sigma_w\sqrt{4d\log(dT_w/\delta)}, \tag{B.83}$$

$$Z''_T := (1+\kappa)(12\kappa^2 + 2\kappa\sqrt{2})\gamma^{-1}\sigma_w\sqrt{2n\log(n(T-T_w)/\delta)}. \tag{B.84}$$

Then, the following holds w.p. at least $1 - 4\delta$*,*

$$\|z_t\| \le \begin{cases} Z'_{T_w}, & \text{for } t \le T_r \\ Z''_T, & \text{for } T_r < t \le T \end{cases}. \tag{B.85}$$

Proof. From Lemma 3.7, we know that $\|x_t\| \leq c'(n+d)^{n+d}$ with $c' > 0$ a constant for $t \leq T_r$ and $\|x_t\| \leq (12\kappa^2 + 2\kappa\sqrt{2})\gamma^{-1}\sigma_w\sqrt{2n\log(n(t-T_w)/\delta)}$ for all $T_r < s \leq T$ w.p. at least $1-4\delta$. Furthermore, under the event of E_t, we have that $\|u_t\| \leq \kappa\|x_t\| + \|v_t\| \leq \kappa\|x_t\| + \kappa\sigma_w\sqrt{4d\log(dT_w/\delta)}$ for all $0 \leq t \leq T_w$. Observing that $\|z_t\| = \sqrt{\|x_t\|^2 + \|u_t\|^2} \leq \|x_t\| + \|u_t\|$, one can reach the desired result by substituting the appropriate bounds on $\|x_t\|$ and $\|u_t\|$ and considering the maximal case achieved when $t = T$. □

The following lemma will be used to bound V_t.

Lemma B.14. *Let $V_t = \mu I + \sum_{s=0}^{t-1} z_s z_s^\top$. On the event of $E_T = \hat{E} \cap \tilde{E} \cap \bar{E}$, we have*

$$\lambda_{\max}(V_t) \leq \begin{cases} \mu + tZ'^2_{T_w}, & \text{for } t \leq T_r \\ \mu + T_r Z'^2_{T_w} + (t-T_r)Z''^2_T, & \text{for } T_r < t \leq T \end{cases} \tag{B.86}$$

$$\text{and} \quad \lambda_{\min}(V_t) \geq \begin{cases} \mu + t\frac{\sigma_w^2}{40}, & \text{for } 200(n+d)\log\frac{12}{\delta} \leq t \leq T_w \\ \mu + T_w\frac{\sigma_w^2}{40}, & \text{for } T_w < t \leq T \end{cases}. \tag{B.87}$$

Proof. Recall that on the event E_T, the RLS estimates, TS sampled systems are concentrated and the state is bounded, i.e., Lemma 3.7. Conditioned on this event, we will start with bounding $\lambda_{\max}(V_t)$. For any time $0 \leq t \leq T$, triangle inequality gives $\lambda_{\max}(V_t) = \|\mu I + \sum_{s=0}^{t-1} z_s z_s^\top\|_2 \leq \mu + \sum_{s=0}^{t-1}\|z_s\|^2$. Using the bounds on $\|z_t\|$ given in Lemma B.13, we can write $\lambda_{\max}(V_t) \leq \mu + tZ'^2_{T_w}$ for $t \leq T_r$ and $\lambda_{\max}(V_t) \leq \mu + T_r Z'^2_{T_w} + (t-T_r)Z''^2_T$ for $T_r < t \leq T$. For the lower bound, note that we have that $\mathbb{E}[z_{t+1}z_{t+1}^\top \mid \mathcal{F}_t] \succcurlyeq \frac{\sigma_w^2}{2} I$. Using Lemma 3.6, on the event E_T, we have that $V_t \succcurlyeq \mu I + t\frac{\sigma_w^2}{40} I$ for $200(n+d)\log\frac{12}{\delta} \leq t \leq T_w$. Since $V_{t+1} = V_t + z_t z_t^\top$, we have that $V_t \succcurlyeq V_{T_w} \succcurlyeq \mu I + T_w\frac{\sigma_w^2}{40} I$ for $T_w < t \leq T$. □

Finally, we will use the following lemma to bound $\beta_t(\delta) = \sigma_w\sqrt{2n\log\left(\frac{\det(V_t)^{1/2}}{\delta\det(\mu I)^{1/2}}\right)} + \sqrt{\mu}S$.

Lemma B.15. *On the event of E_T, we have the following upper bound on $\beta_T(\delta)$:*

$$\beta_T(\delta) \leq 4\sigma_w^2 n\log\left(\frac{1}{\delta}\right) + 2\sigma_w^2 n(n+d)\log\left(1 + \frac{T_r Z'^2_{T_w} + (T-T_r)Z''^2_T}{(n+d)\mu}\right) + 2\mu S^2. \tag{B.88}$$

Proof. Following a similar approach pursued in Lemma 10 of [2], we can bound the log-determinant of V_t as

$$\log \frac{\det(V_T)}{\det(\mu I)} \leq (n+d) \log \left(1 + \frac{T_r Z'^2_{T_w} + (T - T_r) Z''^2_T}{(n+d)\mu}\right),$$

by Lemma B.14. This leads to the following upper bound on $\beta_t(\delta)$

$$\begin{aligned}\beta_T(\delta)^2 &\leq \left(\sigma_w \sqrt{2n \log\left(\frac{1}{\delta}\right) + n(n+d) \log\left(1 + \frac{T_r Z'^2_{T_w} + (T - T_r) Z''^2_T}{(n+d)\mu}\right)} + \sqrt{\mu} S\right)^2 \\ &\leq 4\sigma_w^2 n \log\left(\frac{1}{\delta}\right) + 2\sigma_w^2 n(n+d) \log\left(1 + \frac{T_r Z'^2_{T_w} + (T - T_r) Z''^2_T}{(n+d)\mu}\right) + 2\mu S^2.\end{aligned}$$

□

Proof of Lemma 3.11. We will first show the desired bounds on $\lambda_{\min}(F_t)$ and $\lambda_{\max}(F_t)$. Recall that the event E_T holds with probability at least $1 - 4\delta$. Noting that $H_*^\top H_* = I + K_*^\top K_*$, it is clear that $F_t \succcurlyeq \beta_t^2 \lambda_{\min}(V_t^{-1}) H_*^\top H_* \succcurlyeq \frac{\beta_t^2}{\lambda_{\max}(V_t)} I$. Thus, from Lemma B.14, for $T_r < t \leq T$, we have that $\lambda_{\min,t} \geq \frac{\beta_t^2}{\lambda_{\max}(V_t)} \geq \frac{\beta_t^2}{\mu + T_r Z'^2_{T_w} + (t - T_r) Z''^2_T}$.

On the other hand, $F_t \preccurlyeq \beta_t^2 \lambda_{\max}(V_t^{-1}) H_*^\top H_* \preccurlyeq \frac{\beta_t^2 (1+\kappa^2)}{\lambda_{\min}(V_t)} I$. Again using Lemma B.14, for $T_r < t \leq T$, we have that $\lambda_{\max,t} \leq \frac{(1+\kappa^2)\beta_t^2}{\lambda_{\min}(V_t)} \leq \frac{(1+\kappa^2)\beta_t^2}{\mu + T_w \frac{\sigma_w^2}{40}}$. Since $t \mapsto \beta_t$ is increasing, $t \mapsto \lambda_{\max,t}$ is increasing as well. The condition number $\kappa_t \coloneqq \frac{\lambda_{\max,t}}{\lambda_{\min,t}} \leq \frac{\mu + T_r Z'^2_{T_w} + (t - T_r) Z''^2_T}{(1+\kappa^2)^{-1}(\mu + T_w \frac{\sigma_w^2}{40})}$ is increasing for $T_r < t \leq T$.

If $T_w = O(\sqrt{T}^{1+o(1)})$, then we have that $\lambda_{\max}(V_T) \leq O(\mathrm{poly}(n, d, \log(1/\delta)) T \log T)$ and $\beta_T(\delta) \leq O(\mathrm{poly}(n, d, \log(1/\delta)) \log T)$. Thus, there are positive constants $C = \mathrm{poly}(n, d, \log(1/\delta))$ and $c = \mathrm{poly}(n, d, \log(1/\delta))$ such that $\lambda_{\max,T} \leq C \frac{\log T}{T_w}$ and $\kappa_t = \frac{\lambda_{\max,T}}{\lambda_{\min,t}} \leq c \frac{T \log T}{T_w}$ for $T_r < t \leq T$ for large enough T. Choosing the larger between C and c yields the desired result. □

Proof of Theorem 3.4

Defining by $p_t^{\text{opt}} := \mathbb{P}\left\{\tilde{\Theta}_t \in \mathcal{S}^{\text{opt}} \,\middle|\, \mathcal{F}_t^{\text{cnt}}, \hat{E}_t\right\}$ the optimistic probability, and by $\mathbb{P}_t\{\cdot\} := \mathbb{P}\{\cdot \mid \mathcal{F}_t^{\text{cnt}}\}$ conditional probability measure, we can write

$$p_t^{\text{opt}} \geq \mathbb{P}\left\{\tilde{\Theta}_t \in \mathcal{S}^{\text{surr}} \,\middle|\, \mathcal{F}_t^{\text{cnt}}, \hat{E}_t\right\}, \tag{B.89}$$

$$= \mathbb{P}\left\{L(\tilde{\Theta}_t^\top H_*) \leq L(\Theta_*^\top H_*) \,\middle|\, \mathcal{F}_t^{\text{cnt}}, \hat{E}_t\right\}, \tag{B.90}$$

$$\geq \min_{\hat{\Theta} \in \mathcal{E}_t^{\text{RLS}}} \mathbb{P}_t\{L(\hat{\Theta}^\top H_* + \eta^\top \beta_t V_t^{-\frac{1}{2}} H_*) \leq L(\Theta_*^\top H_*)\}, \tag{B.91}$$

$$= \min_{\hat{\Theta} \in \mathcal{E}_t^{\text{RLS}}} \mathbb{P}_t\{L(\hat{\Theta}^\top H_* + \Xi\sqrt{F_t}) \leq L(\Theta_*^\top H_*)\}, \tag{B.92}$$

where (B.89) is by Lemma 3.8, (B.91) is a worst-case estimation bound within high-probability confidence region, and (B.92) is because $\eta^\top \beta_t V_t^{-\frac{1}{2}} H_*$ and $\Xi\sqrt{F_t}$ have the same distributions with $\eta \in \mathbb{R}^{(n+d)\times n}$ and $\Xi \in \mathbb{R}^{n\times n}$ being i.i.d. standard normal random matrices.

The bound in (B.92) can be further lower bounded by minimizing over a larger confidence set as

$$p_t^{\text{opt}} \geq \min_{\hat{\Theta} \in \mathcal{E}_t^{\text{cl}}} \mathbb{P}_t\{L(\hat{\Theta}^\top H_* + \Xi\sqrt{F_t}) \leq L(A_{c,*})\}, \tag{B.93}$$

$$= \min_{\hat{\Upsilon} \,:\, \|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t\{L(A_{c,*} + (\Xi + \hat{\Upsilon})\sqrt{F_t}) \leq L(A_{c,*})\}, \tag{B.94}$$

where (B.93) is by Lemma (3.9) and (B.94) is because H_* is full column rank and therefore we can minimize over closed-loop matrices instead of open-loop system parameters.

Denoting by $G_t = (\Xi + \hat{\Upsilon})\sqrt{F_t}$ the perturbation due to estimation and sampling, Lemma 3.10 suggests that there exists constants $\epsilon_* > 0$ and $r_* > 0$ such that

$$L(A_{c,*} + G_t) = L(A_{c,*}) + \nabla L_* \bullet G_t + \frac{1}{2} G_t \bullet \mathcal{H}_{A_{c,*}+sG_t}(G_t), \tag{B.95}$$

$$\leq L(A_{c,*}) + \nabla L_* \bullet G_t + \frac{r_*}{2}\|G_t\|_F^2, \tag{B.96}$$

whenever $\|G_t\|_F \leq \epsilon_*$. Substituting (B.96) into (B.94) leads to the following lower bound

$$p_t^{\text{opt}} \geq \min_{\hat{\Upsilon} \,:\, \|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t\{L(A_{c,*} + G_t) \leq L(A_{c,*})\} \tag{B.97}$$

$$\geq \min_{\hat{\Upsilon} \,:\, \|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t \left\{ \begin{array}{c} L(A_{c,*}) + \nabla L_* \bullet G_t + \frac{r_*}{2}\|G_t\|_F^2 \leq L(A_{c,*}), \\ \text{and} \quad \|G_t\|_F \leq \epsilon_* \end{array} \right\} \tag{B.98}$$

$$= \min_{\hat{\Upsilon} \,:\, \|\hat{\Upsilon}\|_F \leq 1} \mathbb{P}_t \left\{ \begin{array}{c} \frac{r_*}{2}\|(\Xi + \hat{\Upsilon})\sqrt{F_t}\|_F^2 + \nabla L_* \bullet (\Xi + \hat{\Upsilon})\sqrt{F_t} \leq 0, \\ \text{and} \quad \|(\Xi + \hat{\Upsilon})\sqrt{F_t}\|_F \leq \epsilon_* \end{array} \right\}. \tag{B.99}$$

Noting that $\|(\Xi+\hat{\Upsilon})\sqrt{F_t}\|_F \le \sqrt{\lambda_{\max,t}}\|\Xi+\hat{\Upsilon}\|_F$ where $\lambda_{\max,t} \coloneqq \lambda_{\max}(F_t)$, we can further relax the lower bound (B.99) as

$$p_t^{\text{opt}} \ge \min_{\hat{\Upsilon}\,:\,\|\hat{\Upsilon}\|_F\le 1} \mathbb{P}_t \left\{ \begin{array}{c} \frac{\lambda_{\max,t} r_*}{2}\|\Xi+\hat{\Upsilon}\|_F^2 + (\nabla L_* \sqrt{F_t}) \bullet (\Xi+\hat{\Upsilon}) \le 0, \\ \text{and} \quad \sqrt{\lambda_{\max,t}}\|\Xi+\hat{\Upsilon}\|_F \le \epsilon_* \end{array} \right\}, \tag{B.100}$$

$$= \min_{\hat{\Upsilon}\,:\,\|\hat{\Upsilon}\|_F\le 1} \mathbb{P}_t \left\{ \begin{array}{c} \left\|\Xi+\hat{\Upsilon}+\frac{\nabla L_*\sqrt{F_t}}{\lambda_{\max,t} r_*}\right\|_F^2 \le \left\|\frac{\nabla L_*\sqrt{F_t}}{\lambda_{\max,t} r_*}\right\|_F^2, \\ \text{and} \quad \|\Xi+\hat{\Upsilon}\|_F^2 \le \frac{\epsilon_*^2}{\lambda_{\max,t}} \end{array} \right\}, \tag{B.101}$$

where (B.101) is obtained by the completion of squares. Let $\mathcal{U} : \mathcal{M}_n \to \mathcal{M}_n$ be an orthogonal transformation such that $\mathcal{U}\left(\hat{\Upsilon}+\frac{\nabla L_*\sqrt{F_t}}{\lambda_{\max,t} r_*}\right) = \left\|\hat{\Upsilon}+\frac{\nabla L_*\sqrt{F_t}}{\lambda_{\max,t} r_*}\right\|_F E_{11}$ where $E_{11} \in \mathcal{M}_n$ has 1 in its $(1,1)$ entry and zeros elsewhere. Since Frobenius norm and the probability density of Ξ are invariant under orthogonal transformations, (B.101) can be rewritten as

$$p_t^{\text{opt}} \ge \min_{\hat{\Upsilon}\,:\,\|\hat{\Upsilon}\|_F\le 1} \mathbb{P}_t \left\{ \begin{array}{c} \left\|\mathcal{U}\left(\Xi+\hat{\Upsilon}+\frac{\nabla L_*\sqrt{F_t}}{\lambda_{\max,t} r_*}\right)\right\|_F^2 \le \left\|\frac{\nabla L_*\sqrt{F_t}}{\lambda_{\max,t} r_*}\right\|_F^2, \\ \text{and} \quad \|\mathcal{U}(\Xi+\hat{\Upsilon})\|_F^2 \le \frac{\epsilon_*^2}{\lambda_{\max,t}} \end{array} \right\} \tag{B.102}$$

$$= \min_{\hat{\Upsilon}\,:\,\|\hat{\Upsilon}\|_F\le 1} \mathbb{P}_t \left\{ \begin{array}{c} \left\|\Xi+\left\|\hat{\Upsilon}+\frac{\nabla L_*\sqrt{F_t}}{\lambda_{\max,t} r_*}\right\|_F E_{11}\right\|_F^2 \le \left\|\frac{\nabla L_*\sqrt{F_t}}{\lambda_{\max,t} r_*}\right\|_F^2, \\ \text{and} \quad \|\Xi+\mathcal{U}(\hat{\Upsilon})\|_F^2 \le \frac{\epsilon_*^2}{\lambda_{\max,t}} \end{array} \right\} \tag{B.103}$$

$$= \min_{\hat{\Upsilon}\,:\,\|\hat{\Upsilon}\|_F\le 1} \mathbb{P}_t \left\{ \begin{array}{c} \left(\Xi_{11}+\left\|\hat{\Upsilon}+\frac{\nabla L_*\sqrt{F_t}}{\lambda_{\max,t} r_*}\right\|_F\right)^2 + \sum_{i,j\neq 1,1}\Xi_{ij}^2 \le \left\|\frac{\nabla L_*\sqrt{F_t}}{\lambda_{\max,t} r_*}\right\|_F^2, \\ \text{and} \quad \|\Xi+\mathcal{U}(\hat{\Upsilon})\|_F^2 \le \frac{\epsilon_*^2}{\lambda_{\max,t}} \end{array} \right\}, \tag{B.104}$$

Notice that the probability in (B.104) is described by the intersection of two balls whose centers are far apart by $\frac{\|\nabla L_*\sqrt{F_t}\|_F}{\lambda_{\max,t} r_*}$ and hence the intersection has a fixed shape. Choosing $\hat{\Upsilon}$ along the direction of $\|\nabla L_*\sqrt{F_t}\|_F$ moves the center of the first ball furthest possible from the origin which leads to the intersection of the balls to move furthest away from the origin as well. Therefore, the probability in (B.104) attains its minimum at $\hat{\Upsilon}_\# \coloneqq \frac{\nabla L_*\sqrt{F_t}}{\|\nabla L_*\sqrt{F_t}\|_F}$ and (B.104) can be equivalently expressed by

$$p_t^{\text{opt}} \ge \mathbb{P}_t \left\{ \begin{array}{c} \left(\Xi_{11}+1+\frac{\|\nabla L_*\sqrt{F_t}\|_F}{\lambda_{\max,t} r_*}\right)^2 + \sum_{i,j\neq 1,1}\Xi_{ij}^2 \le \frac{\|\nabla L_*\sqrt{F_t}\|_F^2}{\lambda_{\max,t}^2 r_*^2}, \\ \text{and} \quad \|\Xi+E_{11}\|_F^2 \le \frac{\epsilon_*^2}{\lambda_{\max,t}} \end{array} \right\} \tag{B.105}$$

$$= \mathbb{P}_t \left\{ \begin{array}{c} \left(\xi+1+\frac{\|\nabla L_*\sqrt{F_t}\|_F}{\lambda_{\max,t} r_*}\right)^2 + X \le \frac{\|\nabla L_*\sqrt{F_t}\|_F^2}{\lambda_{\max,t}^2 r_*^2}, \\ \text{and} \quad (\xi+1)^2 + X \le \frac{\epsilon_*^2}{\lambda_{\max,t}} \end{array} \right\}, \tag{B.106}$$

where $\xi \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_{n^2-1}$ are independent normal and chi-squared random variables, respectively. Denoting by $a_t \coloneqq \frac{\|\nabla L_* \sqrt{F_t}\|_F}{\lambda_{\max,t} r_*}$ and $b_t = \frac{\epsilon_*}{\sqrt{\lambda_{\max,t}}}$ the radii of the balls, we can rewrite (B.106) as

$$p_t^{\text{opt}} \geq \mathbb{P}_t \left\{ (\xi + 1 + a_t)^2 + X \leq a_t^2, \text{ and } (\xi+1)^2 + X \leq b_t^2 \right\} \tag{B.107}$$

$$= \mathbb{P}_t \left\{ \begin{array}{c} |\xi + 1 + a_t| \leq \sqrt{a_t^2 - X}, \text{ and } |\xi + 1| \leq \sqrt{b_t^2 - X}, \\ \text{and } X \leq \min(a_t^2, b_t^2) \end{array} \right\} \tag{B.108}$$

$$= \int_0^{\min(a_t^2, b_t^2)} \mathbb{P}_t \left\{ |\xi + 1 + a_t| \leq \sqrt{a_t^2 - x}, \text{ and } |\xi + 1| \leq \sqrt{b_t^2 - x} \right\} f_{n^2-1}(x) \mathrm{d}x \tag{B.109}$$

$$= \int_0^{\min(a_t^2, b_t^2)} \mathbb{P}_t \left\{ \begin{array}{c} 1 + a_t - \sqrt{a_t^2 - x} \leq \xi \leq 1 + a_t + \sqrt{a_t^2 - x}, \\ \text{and } 1 - \sqrt{b_t^2 - x} \leq \xi \leq 1 + \sqrt{b_t^2 - x}, \end{array} \right\} f_{n^2-1}(x) \mathrm{d}x, \tag{B.110}$$

where $f_k(x) \coloneqq \left(2^{\frac{k}{2}} \Gamma(\frac{k}{2})\right)^{-1} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$ is the probability density function of the chi-squared distribution with $k \in \mathbb{N}$ degrees of freedom. (B.109) is derived from the law of total probability. Notice that the probability inside the integral in (B.110) is determined by the intersection of two intervals. This probability will have a non-zero value only for a fixed interval of x depending on the relation between a_t and b_t. We will investigate three cases:

i. $\mathbf{0 \leq b_t \leq \sqrt{2} a_t}$: There is a non-empty intersection if and only if $0 \leq x \leq b_t^2 \left(1 - \frac{b_t^2}{4a_t^2}\right)$ and the integral (B.110) becomes

$$p_t^{\text{opt}} \geq \int_0^{b_t^2\left(1 - \frac{b_t^2}{4a_t^2}\right)} \mathbb{P}_t \left\{ 1 + a_t - \sqrt{a_t^2 - x} \leq \xi \leq 1 + \sqrt{b_t^2 - x} \right\} f_{n^2-1}(x) \mathrm{d}x \tag{B.111}$$

$$= \int_0^{b_t^2\left(1 - \frac{b_t^2}{4a_t^2}\right)} \left[Q\left(1 + a_t - \sqrt{a_t^2 - x}\right) - Q\left(1 + \sqrt{b_t^2 - x}\right) \right] f_{n^2-1}(x) \mathrm{d}x, \tag{B.112}$$

where Q is the Gaussian Q-function. Notice that for fixed values of b_t, (B.112) is monotonically increasing with respect to a_t and vice versa.

ii. $\sqrt{2}\mathbf{a_t} \leq \mathbf{b_t} \leq \mathbf{2a_t}$: There is a non-empty intersection if and only if $0 \leq x \leq a_t^2$ and the integral (B.110) becomes

$$p_t^{\text{opt}} \geq \int_0^{a_t^2} \mathbb{P}_t \left\{ 1 + a_t - \sqrt{a_t^2 - x} \leq \xi \leq 1 + \sqrt{b_t^2 - x} \right\} f_{n^2-1}(x) \mathrm{d}x \tag{B.113}$$

$$= \int_0^{a_t^2} \left[Q\left(1 + a_t - \sqrt{a_t^2 - x}\right) - Q\left(1 + \sqrt{b_t^2 - x}\right) \right] f_{n^2-1}(x) \mathrm{d}.x \tag{B.114}$$

Notice that for fixed values of b_t, (B.114) is monotonically increasing with respect to a_t and vice versa.

iii. $\mathbf{2a_t} \leq \mathbf{b_t}$: There is a non-empty intersection if and only if $0 \leq x \leq a_t^2$ and the integral (B.110) becomes

$$p_t^{\text{opt}} \geq \int_0^{a_t^2} \mathbb{P}_t \left\{ 1 + a_t - \sqrt{a_t^2 - x} \leq \xi \leq 1 + a_t + \sqrt{a_t^2 - x} \right\} f_{n^2-1}(x) \mathrm{d}x \tag{B.115}$$

$$= \int_0^{a_t^2} \left[Q\left(1 + a_t - \sqrt{a_t^2 - x}\right) - Q\left(1 + a_t + \sqrt{a_t^2 - x}\right) \right] f_{n^2-1}(x) \mathrm{d}x \tag{B.116}$$

Notice that for fixed values of b_t, (B.116) is monotonically increasing with respect to a_t and vice versa.

As seen from all three case, the integral in (B.110) is monotonically increasing with respect to both a_t, and b_t regardless of their relative relation. Therefore, we will consider tight lower bounds of $a_t = \frac{\|\nabla L_* \sqrt{F_t}\|_F}{\lambda_{\max,t} r_*}$ so that the relation $b_t \geq 2a_t$ holds for large enough $t \geq 0$. Noting that $\nabla L_* = 2P_* A_{c,*} \Sigma_*$ by Lemma 3.10 and $P_* \succ 0$, $\Sigma_* \succ 0$, we will consider two cases.

1. **Singular $\mathbf{A_{c,*}}$** : In this case, the Jacobian matrix ∇L_* becomes singular as well. Then, we can bound a_t from below as $a_t = \frac{\|\nabla L_* \sqrt{F_t}\|_F}{\lambda_{\max,t} r_*} \geq \sqrt{\lambda_{\min,t}} \frac{\|\nabla L_*\|_F}{\lambda_{\max,t} r_*} = \sqrt{\frac{\lambda_{\min,t}}{\lambda_{\max,t}}} \frac{\|r_*^{-1} \nabla L_*\|_F}{\sqrt{\lambda_{\max,t}}}$. Furthermore, choosing $T_w = O((\sqrt{T})^{1+o(1)})$, we can use upper bounds for $\frac{\lambda_{\max,t}}{\lambda_{\min,t}}$ and $\lambda_{\max,t}$ from Lemma 3.11 to write down, $a_t \geq \frac{T_w}{\sqrt{T}\log T} \frac{\|\nabla L_*\|_F}{C r_*} =: a_{1,T}$ and $b_t \geq \sqrt{\frac{T_w}{\log T}} \frac{\epsilon_*}{\sqrt{C}} =: b_{1,T}$ for all $T_r < t \leq T$ under the event E_T for large enough T. Therefore, replacing a_t and b_t with $a_{1,T}$ and $b_{1,T}$ in (B.110) gives a lower bound to (B.110). Noting that the ratio $\frac{b_{1,T}}{a_{1,T}} = \sqrt{\frac{T \log T}{T_w}} \frac{\epsilon_* r_* \sqrt{C}}{\|\nabla L_*\|_F}$ can be made to be greater than or equal to 2 by an appropriate choice of T_w leading to the case (*iii*) bound

$$p_t^{\text{opt}} \geq \int_0^{a_{1,T}^2} \left[Q\left(1 + a_{1,T} - \sqrt{a_{1,T}^2 - x}\right) - Q\left(1 + a_{1,T} + \sqrt{a_{1,T}^2 - x}\right) \right] f_{n^2-1}(x) \mathrm{d}x \tag{B.117}$$

for all $T_r < t \leq T$ for large enough T.

2. Nonsingular $\mathbf{A}_{\mathbf{c},*}$: In this case, the Jacobian matrix ∇L_* becomes nonsingular as well. Then, we can bound a_t from below as $a_t = \frac{\|\nabla L_* \sqrt{F_t}\|_F}{\lambda_{\max,t} r_*} \geq \sigma_{\min,*} \frac{\|\sqrt{F_t}\|_F}{\lambda_{\max,t} r_*} \geq \frac{\sigma_{\min,*}}{r_* \sqrt{\lambda_{\max,t}}}$. Choosing $T_w = O((\sqrt{T})^{1+o(1)})$, we can use the upper bound for $\lambda_{\max,t}$ from Lemma 3.11 to write the lower bound, $a_t \geq \sqrt{\frac{T_w}{\log T}} \frac{\min(\sigma_{\min,*}, \epsilon_* r_*/2)}{\sqrt{C} r_*} =: a_{2,T}$ and $b_t \geq \sqrt{\frac{T_w}{\log T}} \frac{\epsilon_*}{\sqrt{C}} =: b_{2,T}$ for all $T_r < t \leq T$ under the event E_T for large enough T. Therefore, replacing a_t and b_t with $a_{2,T}$ and $b_{2,T}$ in (B.110) gives a lower bound to (B.110) for $T_r < t \leq T$. Noting that the ratio $\frac{\beta_{2,T}}{a_{2,T}} = \frac{\epsilon_* r_*}{\min(\sigma_{\min,*}, \epsilon_* r_*/2)} = \max\left(\frac{\epsilon_* r_*}{\sigma_{\min,*}}, 2\right) \geq 2$, we can use the case (iii) bound

$$p_t^{\text{opt}} \geq \int_0^{a_{2,T}^2} \left[Q\left(1 + a_{2,T} - \sqrt{a_{2,T}^2 - x}\right) - Q\left(1 + a_{2,T} + \sqrt{a_{2,T}^2 - x}\right) \right] f_{n^2-1}(x) \mathrm{d}x \tag{B.118}$$

for all $T_r < t \leq T$ for large enough T.

In both cases, our focus will be on the following probability with a parameters $a > 0$, and $k \in \mathbb{N}$

$$p_k(a) \coloneqq \int_0^{a^2} \left[Q(1 + a - \sqrt{a^2 - x}) - Q(1 + a + \sqrt{a^2 - x}) \right] f_k(x) \mathrm{d}x \tag{B.119}$$

The following lemma summarizes some of the important properties of the function $a \mapsto p_k(a)$.

Lemma B.16. *The non-negative real valued function $a \mapsto p_k(a)$ is monotonically increasing with respect to $a \geq 0$. Furthermore, we have that $\frac{1}{p_k(a)} \leq \frac{1}{Q(1)} \left(1 + \frac{Ck}{a^{1/2}}\right)$ for $a \geq ck$ for problem independent constants $c, C > 0$.*

Proof. Notice that for a fixed value of $0 \leq x \leq a^2$, the functions $a \mapsto 1 + a - \sqrt{a^2 - x}$ and $a \mapsto 1 + a + \sqrt{a^2 - x}$ are monotonically decreasing and monotonically increasing, respectively. As Q-function is monotonically decreasing, the function $a \mapsto Q(1 + a - \sqrt{a^2 - x}) - Q(1 + a + \sqrt{a^2 - x})$ is monotonically increasing for fixed $0 \leq x \leq a^2$. Therefore, the function $a \mapsto p_k(a)$ is also monotonically increasing.

In order to obtain the desired asymptotic bound, let $\epsilon \in (0,1)$ and we can write

$$\begin{aligned}
p_k(a) &= \int_0^{a^2} \left[Q(1+a-\sqrt{a^2-x}) - Q(1+a+\sqrt{a^2-x}) \right] f_k(x)\mathrm{d}x \\
&\geq \int_0^{\epsilon a^2} \left[Q(1+a-\sqrt{a^2-x}) - Q(1+a+\sqrt{a^2-x}) \right] f_k(x)\mathrm{d}x \\
&\geq \int_0^{\epsilon a^2} \min_{0\leq x' \leq \epsilon a^2} \left[Q(1+a-\sqrt{a^2-x'}) - Q(1+a+\sqrt{a^2-x'}) \right] f_k(x)\mathrm{d}x \\
&= \left[Q(1+a(1-\sqrt{1-\epsilon})) - Q(1+a(1+\sqrt{1-\epsilon})) \right] F_k(\epsilon a^2)
\end{aligned}$$

where $F_k(x) \coloneqq 1 - \frac{\Gamma(k/2, x/2)}{\Gamma(k/2)}$ is the cumulative distribution function of chi-square distribution and $(s,x) \mapsto \Gamma(s,x) \coloneqq \int_x^\infty t^{s-1}e^{-t}\mathrm{d}t$ and $s \mapsto \Gamma(s) \coloneqq \int_o^\infty t^{s-1}e^{-t}\mathrm{d}t$ are upper incomplete Gamma and ordinary Gamma functions respectively. Notice that the functions $(s,x) \mapsto \Gamma(s,x)$ and $x \mapsto Q(x)$ are monotonically decreasing with increasing $x > 0$. Therefore, for large enough $\epsilon a^2 \gg 1$ and large enough $a \gg 1$, we can claim that $\Gamma(k/2, \epsilon a^2/2) \ll 1$ and $Q(1+a) \ll 1$ are small enough. Furthermore, for small enough $\epsilon \ll 1$, we can use Taylor expansion to see that $1 - \sqrt{1-\epsilon} = \frac{\epsilon}{2} \sum_{k=0}^\infty \frac{\epsilon^k}{2^k}(2k-1)! \leq c_1\epsilon$ for a problem-independent constant $c_1 > 0$. Then, for small enough $\epsilon \ll 1$, we have that

$$\begin{aligned}
p_k(a) &\geq \left[Q(1+a(1-\sqrt{1-\epsilon})) - Q(1+a(1+\sqrt{1-\epsilon})) \right] \left(1 - \frac{\Gamma(k/2, \epsilon a^2/2)}{\Gamma(k/2)} \right) \\
&\geq \left[Q(1+c_1\epsilon a) - Q(1+a) \right] \left(1 - \frac{\Gamma(k/2, \epsilon a^2/2)}{\Gamma(k/2)} \right)
\end{aligned}$$

Furthermore, for small enough $\epsilon a \ll 1$, we have that $Q(1+c_1\epsilon a) \geq Q(1) - c_2\epsilon a$ by Taylor's theorem where c_2 is a problem-independent constant. Using these bounds, we can bound the inverse of $p_k(a)$ from above for small enough $\epsilon \ll 1$, small enough $\epsilon a \ll 1$, large enough $a \gg 1$ and large enough $\epsilon a^2 \gg 1$ as

$$\begin{aligned}
\frac{1}{p_k(a)} &\leq \frac{1}{Q(1) - c_2\epsilon a - Q(1+a)} \frac{1}{1 - \frac{\Gamma(k/2, \epsilon a^2/2)}{\Gamma(k/2)}} \\
&= \frac{1}{Q(1)} \frac{1}{(1 - c_2\epsilon a - Q(1+a)) \left(1 - \frac{\Gamma(k/2, \epsilon a^2/2)}{\Gamma(k/2)} \right)} \\
&\leq \frac{1}{Q(1)} \left[1 + 2C \left(c_2\epsilon a + Q(1+a) + \frac{\Gamma(k/2, \epsilon a^2/2)}{\Gamma(k/2)} \right) \right] \qquad \text{(B.120)}
\end{aligned}$$

where we used the Taylor expansion $\frac{1}{1-x} = \sum_{k=0}^\infty x^k \leq 1 + Cx$ for small enough $x \ll 1$ with $C > 0$ being a problem-independent constant.

The assumption $\epsilon a^2 \gg 1$ can be used to write the asymptotic expansion of incomplete Gamma function $\Gamma(k/2,\, \epsilon a^2/2) = (\epsilon a^2/2)^{k/2-1} e^{-\epsilon a^2/2} \left[1 + O\left((\epsilon a^2/2)^{-1}\right)\right]$. Noting that the Q function is always bounded as $Q(1+a) \leq \frac{e^{-\frac{(1+a)^2}{2}}}{\sqrt{2\pi}(1+a)}$, we claim that choosing $\epsilon = \frac{k}{2ea^{1+1/2}}$, for $\alpha \geq c''k$ with a constant $c'' > 0$ guarantees that $\epsilon a = \frac{k}{2e} a^{-1/2} \ll 1$ and $\epsilon a^2 = \frac{k}{2e} a^{1-1/2} \gg 1$. Therefore, the upper bound (B.120) is valid for $\alpha \geq c''k$. Furthermore, the term ϵa decays slower than both $Q(1+a)$ and $\frac{\Gamma(k/2, \epsilon a^2/2)}{\Gamma(k/2)}$ and thus ϵa dominates as

$$\frac{1}{p_k(a)} \leq \frac{1}{Q(1)} \left(1 + \frac{Ck}{2e} a^{-1/2}\right),$$

for a problem-independent constant $C > 0$. □

Based on Lemma B.16, the integrals in (B.117) and (B.118) are asymptotically constant if both $a_{1,T}$ and $a_{2,T}$ are asymptotically large enough. This can be achieved if $a_{1,T} = \frac{T_w}{\sqrt{T}\log T} \frac{\|\nabla L_*\|_F}{C r_*} = \omega(1)$ for singular $A_{c,*}$ and $a_{2,T} = \sqrt{\frac{T_w}{\log T}} \frac{\min(\sigma_{\min,*},\, \epsilon_* r_*/2)}{\sqrt{C} r_*} = \omega(1)$ for non-singular $A_{c,*}$. In other words, choosing $T_w = n^2 \omega(\sqrt{T}\log T)$ for singular $A_{c,*}$ and $T_w = n^2 \omega(\log T)$ for non-singular $A_{c,*}$ yields the desired bound

$$p_t^{\text{opt}} \geq \frac{Q(1)}{1 + o(1)},$$

for $T_r < t \leq T$ for large enough T. Combined with the upper $T_w = O((\sqrt{T})^{1+o(1)})$, the proposed choices of T_w satisfy the asymptotic conditions. □

B.2.3 Regret Decomposition

Denote the optimal expected average cost of an LQR system Θ with process noise covariance W by $J_*(\Theta, W) = \text{tr}(P(\Theta)W)$. Note that during the initial exploration period, we have that $u_t = \bar{u}_t + \nu_t$ for $t \leq T_w$ and after the initial exploration, we have that $u_t = \bar{u}_t$ for $t > T_w$ where we denote by $\bar{u}_t \coloneqq K(\tilde{\Theta}_t) x_t$ the optimal control action assuming the system $\tilde{\Theta}_t$. Since initial exploration period injects independent random perturbations through the optimal control input, $\bar{u}_t$, for sampled system, $\tilde{\Theta}_t$, the state dynamics can be reformulated in order to take the external perturbations into account by adding it to the process noise:

$$x_{t+1} = A_* x_t + B_* \bar{u}_t + \zeta_t, \tag{B.121}$$

where $\bar{u}_t = K(\tilde{\Theta}_t) x_t$, $\zeta_t = B_* \nu_t + w_t$ for $t \leq T_w$, and $\zeta_t = w_t$ for $t > T_w$. We can write the regret explicitly as

$$R_T = \sum\nolimits_{t=0}^{T} \left\{x_t^\top Q x_t + u_t^\top R u_t - J_*(\Theta_*, \sigma_w^2 I)\right\} = R_{T_w}^{\text{exp}} + R_T^{\text{noexp}}, \tag{B.122}$$

where

$$R_{T_w}^{\text{exp}} := \sum_{t=0}^{T_w} \left(2\bar{u}_t^\top R\nu_t + \nu_t^\top R\nu_t\right), R_T^{\text{noexp}} := \sum_{t=0}^{T} \left\{x_t^\top Q x_t + \bar{u}_t^\top R\bar{u}_t - J_*(\Theta_*, \sigma_w^2 I)\right\}$$

Since $E_s \subset E_t$ for any $0 \le s \le t$, we have that

$$\begin{aligned} R_T^{\text{noexp}} \mathbb{1}_{E_T} &= \sum_{t=0}^{T} \left\{x_t^\top Q x_t + \bar{u}_t^\top R\bar{u}_t - J_*(\Theta_*, \sigma_w^2 I)\right\} \mathbb{1}_{E_T} \\ &\le \sum_{t=0}^{T} \left\{x_t^\top Q x_t + \bar{u}_t^\top R\bar{u}_t - J_*(\Theta_*, \sigma_w^2 I)\right\} \mathbb{1}_{E_t}, \end{aligned} \tag{B.123}$$

$$R_{T_w}^{\text{exp}} \mathbb{1}_{E_T} = \sum_{t=0}^{T_w} \left(2\bar{u}_t^\top R\nu_t + \nu_t^\top R\nu_t\right) \mathbb{1}_{E_T} \le \sum_{t=0}^{T_w} \left(2\bar{u}_t^\top R\nu_t + \nu_t^\top R\nu_t\right) \mathbb{1}_{E_t}. \tag{B.124}$$

From Bellman optimality equations [28], we obtain

$$\begin{aligned} &J_*(\tilde{\Theta}_t, \text{Cov}[\zeta_t]) + x_t^\top P(\tilde{\Theta}_t) x_t \\ &\quad = x_t^\top Q x_t + \bar{u}_t^\top R\bar{u}_t + \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \,\middle|\, \mathcal{F}_t\right] + \bar{z}_t^\top \tilde{\Theta}_t P(\tilde{\Theta}_t) \tilde{\Theta}_t^\top \bar{z}_t - \bar{z}_t^\top \Theta_* P(\tilde{\Theta}_t) \Theta_*^\top \bar{z}_t, \end{aligned}$$

where $\bar{z}_t^\top = [x_t^\top, \bar{u}_t^\top]$. Rearranging the terms and subtracting the optimal expected average cost of the true system, we obtain the following for each term in (B.123),

$$\begin{aligned} &\left\{x_t^\top Q x_t + \bar{u}_t^\top R\bar{u}_t - J_*(\Theta_*, \sigma_w^2 I)\right\} \mathbb{1}_{E_t} \\ &= \left\{J_*(\tilde{\Theta}_t, \text{Cov}[\zeta_t]) - J_*(\Theta_*, \sigma_w^2 I)\right\} \mathbb{1}_{E_t} + \left\{\bar{z}_t^\top \Theta_* P(\tilde{\Theta}_t) \Theta_*^\top \bar{z}_t - \bar{z}_t^\top \tilde{\Theta}_t P(\tilde{\Theta}_t) \tilde{\Theta}_t^\top \bar{z}_t\right\} \mathbb{1}_{E_t}, \\ &\quad + x_t^\top P(\tilde{\Theta}_t) x_t \mathbb{1}_{E_t} - \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mathbb{1}_{E_t} \,\middle|\, \mathcal{F}_t\right]. \end{aligned}$$

Note that, $\mathbb{1}_{E_t} \mathbb{1}_{E_{t+1}} = \mathbb{1}_{E_{t+1}}$ since $E_{t+1} \subset E_t$. Since $P(\tilde{\Theta}_t) \succ 0$, we obtain

$$\begin{aligned} &\mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mathbb{1}_{E_t} \,\middle|\, \mathcal{F}_t\right] \\ &= \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mathbb{1}_{E_t} \left(\mathbb{1}_{E_{t+1}} + \mathbb{1}_{E_{t+1}^c}\right) \,\middle|\, \mathcal{F}_t\right], \\ &= \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mathbb{1}_{E_{t+1}} \,\middle|\, \mathcal{F}_t\right] + \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mathbb{1}_{E_t} \mathbb{1}_{E_{t+1}^c} \,\middle|\, \mathcal{F}_t\right], \\ &\ge \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_t) x_{t+1} \mathbb{1}_{E_{t+1}} \,\middle|\, \mathcal{F}_t\right], \\ &= \mathbb{E}\left[x_{t+1}^\top \left(P(\tilde{\Theta}_t) - P(\tilde{\Theta}_{t+1})\right) x_{t+1} \mathbb{1}_{E_{t+1}} \,\middle|\, \mathcal{F}_t\right] + \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_{t+1}) x_{t+1} \mathbb{1}_{E_{t+1}} \,\middle|\, \mathcal{F}_t\right]. \end{aligned}$$

Therefore,

$$\begin{aligned} \left\{x_t^\top Q x_t + \bar{u}_t^\top R\bar{u}_t - J_*(\Theta_*, \sigma_w^2 I)\right\} \mathbb{1}_{E_t} \le\ & \left\{J_*(\tilde{\Theta}_t, \text{Cov}[\zeta_t]) - J_*(\Theta_*, \sigma_w^2 I)\right\} \mathbb{1}_{E_t}, \\ &+ \left\{\bar{z}_t^\top \Theta_* P(\tilde{\Theta}_t) \Theta_*^\top \bar{z}_t - \bar{z}_t^\top \tilde{\Theta}_t P(\tilde{\Theta}_t) \tilde{\Theta}_t^\top \bar{z}_t\right\} \mathbb{1}_{E_t}, \\ &+ \left\{x_t^\top P(\tilde{\Theta}_t) x_t \mathbb{1}_{E_t} - \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_{t+1}) x_{t+1} \mathbb{1}_{E_{t+1}} \,\middle|\, \mathcal{F}_t\right]\right\}, \\ &+ \mathbb{E}\left[x_{t+1}^\top \left(P(\tilde{\Theta}_{t+1}) - P(\tilde{\Theta}_t)\right) x_{t+1} \mathbb{1}_{E_{t+1}} \,\middle|\, \mathcal{F}_t\right]. \end{aligned} \tag{B.125}$$

Notice that $\mathrm{Cov}[\zeta_t] = \sigma_v^2 B_* B_*^\top + \sigma_w^2 I$ for $t \le T_w$ and $\mathrm{Cov}[\zeta_t] = \sigma_w^2 I$ for $t > T_w$ and therefore

$$J_*(\tilde{\Theta}_t, \mathrm{Cov}[\zeta_t]) = \mathrm{Tr}(P(\tilde{\Theta}_t)\,\mathrm{Cov}[\zeta_t]) = \begin{cases} \sigma_v^2\,\mathrm{Tr}(P(\tilde{\Theta}_t)B_*B_*^\top) + \sigma_w^2\,\mathrm{Tr}(P(\tilde{\Theta}_t)) & t \le T_w \\ \sigma_w^2\,\mathrm{Tr}(P(\tilde{\Theta}_t)) & t > T_w \end{cases}. \tag{B.126}$$

Summing the terms in (B.125) upto time T and adding the $R_{T_w}^{\mathrm{exp}}$ term, we obtain

$$R_T \mathbb{1}_{E_T} = R_{T_w}^{\mathrm{exp}} \mathbb{1}_{E_T} + R_T^{\mathrm{noexp}} \mathbb{1}_{E_T} \le R_{T_w}^{\mathrm{exp},1} + R_{T_w}^{\mathrm{exp},2} + R_T^{\mathrm{TS}} + R_T^{\mathrm{RLS}} + R_T^{\mathrm{mart}} + R_T^{\mathrm{gap}} \tag{B.127}$$

where

$$R_{T_w}^{\mathrm{exp},1} = \sum\nolimits_{t=0}^{T_w} \left(2\bar{u}_t^\top R \nu_t + \nu_t^\top R \nu_t\right) \mathbb{1}_{E_t}, \tag{B.128}$$

$$R_{T_w}^{\mathrm{exp},2} = \sum\nolimits_{t=0}^{T_w} \sigma_v^2\,\mathrm{Tr}(P(\tilde{\Theta}_t)B_*B_*^\top)\mathbb{1}_{E_t}, \tag{B.129}$$

$$R_T^{\mathrm{TS}} = \sum\nolimits_{t=0}^{T} \left\{J_*(\tilde{\Theta}_t, \sigma_w^2 I) - J_*(\Theta_*, \sigma_w^2 I)\right\} \mathbb{1}_{E_t}, \tag{B.130}$$

$$R_T^{\mathrm{RLS}} = \sum\nolimits_{t=0}^{T} \left\{\bar{z}_t^\top \Theta_* P(\tilde{\Theta}_t)\Theta_*^\top \bar{z}_t - \bar{z}_t^\top \tilde{\Theta}_t P(\tilde{\Theta}_t)\tilde{\Theta}_t^\top \bar{z}_t\right\} \mathbb{1}_{E_t}, \tag{B.131}$$

$$R_T^{\mathrm{mart}} = \sum\nolimits_{t=0}^{T} \left\{x_t^\top P(\tilde{\Theta}_t) x_t \mathbb{1}_{E_t} - \mathbb{E}\left[x_{t+1}^\top P(\tilde{\Theta}_{t+1}) x_{t+1} \mathbb{1}_{E_{t+1}} \,\middle|\, \mathcal{F}_t\right]\right\}, \tag{B.132}$$

$$R_T^{\mathrm{gap}} = \sum\nolimits_{t=0}^{T} \mathbb{E}\left[x_{t+1}^\top \left(P(\tilde{\Theta}_{t+1}) - P(\tilde{\Theta}_t)\right) x_{t+1} \mathbb{1}_{E_{t+1}} \,\middle|\, \mathcal{F}_t\right]. \tag{B.133}$$

B.2.4 Regret Analysis

Bounding $R_{T_w}^{\mathrm{exp},1}$ and $R_{T_w}^{\mathrm{exp},2}$

The following gives an upper bound on the regret attained due to isotropic perturbations in the TS with improved exploration phase of TSAC.

Lemma B.17 (Direct Effect of Improved Exploration on Regret)**.** *The following holds with probability at least* $1 - \delta$*,*

$$R_{T_w}^{exp,1} = \sum_{t=0}^{T_w} \left\{2\bar{u}_t^\top R \nu_t + \nu_t^\top R \nu_t\right\} \mathbb{1}_{E_t} \le d\sigma_\nu \sqrt{B_\delta} + d\|R\|\sigma_\nu^2 \left(T_w + \sqrt{T_w} \log \frac{4dT_w}{\delta} \sqrt{\log \frac{4}{\delta}}\right)$$

where

$$B_\delta = 8\left(1 + T_w \kappa^2 \|R\|^2 (n+d)^{2(n+d)}\right) \log\left(\frac{4d}{\delta}\left(1 + T_w \kappa^2 \|R\|^2 (n+d)^{2(n+d)}\right)^{1/2}\right).$$

Furthermore, we have $R_{T_w}^{exp,2} \le \sigma_v^2 D\,\|B_*\|_F^2 T_w$.

Proof. The proof follows directly from Lemma B.17 and Assumption 3.2. □

Bounding $R_T^{\mathbf{RLS}}$

Bounding this term is achieved by manipulating the similar bounds in Abbasi-Yadkori and Szepesvári [2], Abeille and Lazaric [6] to our setting and TS algorithm. In particular, this regret term corresponds to R_3^{ζ} of StabL studied in Lemma B.10. We first have the following result from regularized least squares estimate.

Lemma B.18. *On the event of E_T, for $X_s = \frac{(12\kappa^2+2\kappa\sqrt{2})\sigma_w}{\gamma}\sqrt{2n\log(n(T-T_w)/\delta)}$, we have,*

$$\sum_{t=0}^{T}\|(\Theta_*-\tilde{\Theta}_t)^\top z_t\|^2 \le 2(\beta_T(\delta)+\upsilon_T(\delta))^2\Bigg(\left(1+\frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\mu}\right)^{\tau_0+1}\log\frac{\det(V_{T_r})}{\det(\mu I)} + \left(1+\frac{(1+\kappa^2)X_s^2}{\mu}\right)^{\tau_0+1}\log\frac{\det(V_T)}{\det(V_{T_r})}\Bigg).$$

Proof. Let $\tau \le t$ be the last time step before t, when the policy was updated. Using Cauchy-Schwarz inequality, we have:

$$\sum_{t=0}^{T}\|(\Theta_*-\tilde{\Theta}_t)^\top z_t\|^2 \le \sum_{t=0}^{T}\|V_t^{\frac{1}{2}}(\tilde{\Theta}_t-\Theta_*)\|^2\|z_t\|^2_{V_t^{-1}} \le \sum_{t=0}^{T}\frac{\det(V_t)}{\det(V_\tau)}\|V_\tau^{\frac{1}{2}}(\tilde{\Theta}_\tau-\Theta_*)\|^2\|z_t\|^2_{V_t^{-1}}. \tag{B.134}$$

Note that $t-\tau \le \tau_0$ due to policy update rule. Moreover, we have

$$\det(V_t) = \det(V_\tau)\prod_{i=0}^{t-\tau}(1+\|z_t\|^2_{V_{t-i}^{-1}}) \le \det(V_\tau)\left(1+\frac{\|z_t\|^2}{\mu}\right)^{\tau_0}.$$

Combining this with (B.134), on the event of E_T, for $t \le T_r$, we have:

$$\sum_{t=0}^{T_r}\|(\Theta_*-\tilde{\Theta}_t)^\top z_t\|^2 \le \sum_{t=0}^{T_r}\left(1+\frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\mu}\right)^{\tau_0}\|V_\tau^{1/2}(\tilde{\Theta}_\tau-\Theta_*)\|^2\|z_t\|^2_{V_t^{-1}} \tag{B.135}$$

$$\le \sum_{t=0}^{T_r}\left(1+\frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\mu}\right)^{\tau_0}(\beta_T(\delta)+\upsilon_T(\delta))^2\|z_t\|^2_{V_t^{-1}}, \tag{B.136}$$

$$\le \frac{2(1+\kappa^2)(n+d)^{2(n+d)}}{\mu}\left(1+\frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\mu}\right)^{\tau_0}(\beta_T(\delta)+\upsilon_T(\delta))^2\log\left(\frac{\det(V_{T_r})}{\det(\mu I)}\right), \tag{B.137}$$

where in (B.136) we used the fact that on the event of E_T, using triangle inequality, we have $\|\tilde{\Theta}_\tau-\Theta_*\|_{V_\tau} \le \|\tilde{\Theta}_\tau-\hat{\Theta}_\tau\|_{V_\tau}+\|\hat{\Theta}_\tau-\Theta_*\|_{V_\tau} \le \upsilon_\tau(\delta)+\beta_\tau(\delta) \le \upsilon_T(\delta)+\beta_T(\delta)$

and in (B.137) we used used the upper bound of $\|z_t\|_{V_t^{-1}}$ to utilize Lemma 10 of Abbasi-Yadkori and Szepesvári [2]. Similarly, on the even of E_t, for $t > T_r$, we get:

$$\sum_{t=T_r+1}^{T} \|(\Theta_* - \tilde{\Theta}_t)^\top z_t\|^2 \leq \frac{2(1+\kappa^2)X_s^2}{\mu}\left(1 + \frac{(1+\kappa^2)X_s^2}{\mu}\right)^{\tau_0}(\beta_T(\delta) + \upsilon_T(\delta))^2 \log\left(\frac{\det(V_T)}{\det(V_{T_r})}\right).$$

□

Lemma B.19 (Bounding R_T^{RLS} for TSAC)**.** *Let R_T^{RLS} be as defined by* (B.131)*. Under the event of E_T, setting $\mu = (1+\kappa^2)X_s^2$, we have*

$$\left|R_T^{RLS}\right| = \tilde{O}\left((n+d)^{(\tau_0+2)(n+d)+1.5}\sqrt{n}\sqrt{T_r} + (n+d)n\sqrt{T-T_r}\right).$$

Proof.

$$\left|R_T^{\mathrm{RLS}}\right| \leq \sum_{t=0}^{T}\left|\left\|P(\tilde{\Theta}_t)^{\frac{1}{2}}\tilde{\Theta}_t^\top z_t\right\|^2 - \left\|P(\tilde{\Theta}_t)^{\frac{1}{2}}\Theta_*^\top z_t\right\|^2\right| \tag{B.138}$$

$$\begin{aligned}
&= \sum_{t=0}^{T_r}\left|\left\|P(\tilde{\Theta}_t)^{\frac{1}{2}}\tilde{\Theta}_t^\top z_t\right\|^2 - \left\|P(\tilde{\Theta}_t)^{\frac{1}{2}}\Theta_*^\top z_t\right\|^2\right| + \sum_{t=T_r}^{T}\left|\left\|P(\tilde{\Theta}_t)^{\frac{1}{2}}\tilde{\Theta}_t^\top z_t\right\|^2 - \left\|P(\tilde{\Theta}_t)^{\frac{1}{2}}\Theta_*^\top z_t\right\|^2\right| \\
&\leq \left(\sum_{t=0}^{T_r}\left\|P(\tilde{\Theta}_t)^{\frac{1}{2}}\left(\tilde{\Theta}_t - \Theta_*\right)^\top z_t\right\|^2\right)^{\frac{1}{2}}\left(\sum_{t=0}^{T_r}\left(\left\|P(\tilde{\Theta}_t)^{\frac{1}{2}}\tilde{\Theta}_t^\top z_t\right\| + \left\|P(\tilde{\Theta}_t)^{\frac{1}{2}}\Theta_*^\top z_t\right\|\right)^2\right)^{\frac{1}{2}} \\
&\quad + \left(\sum_{t=T_r}^{T}\left\|P(\tilde{\Theta}_t)^{\frac{1}{2}}\left(\tilde{\Theta}_t - \Theta_*\right)^\top z_t\right\|^2\right)^{\frac{1}{2}}\left(\sum_{t=T_r}^{T}\left(\left\|P(\tilde{\Theta}_t)^{\frac{1}{2}}\tilde{\Theta}_t^\top z_t\right\| + \left\|P(\tilde{\Theta}_t)^{\frac{1}{2}}\Theta_*^\top z_t\right\|\right)^2\right)^{\frac{1}{2}}
\end{aligned} \tag{B.139}$$

where (B.138) and (B.139) follow from triangle inequality. Note that for $t \leq T_r$, we have $\|z_t\|^2 \leq (1+\kappa^2)(n+d)^{2(n+d)}$ and for $t > T_r$ we have $\|z_t\|^2 \leq (1+\kappa^2)X_s^2$. Moreover, since $\tilde{\Theta}$ belongs to $\mathcal{S}$ by construction of the rejection sampling, we get

$$\begin{aligned}
\left|R_T^{\mathrm{RLS}}\right| &\leq \left(D\sum_{t=0}^{T_r}\left\|\left(\tilde{\Theta}_t - \Theta_*\right)^\top z_t\right\|^2\right)^{\frac{1}{2}}\sqrt{4T_r D(1+\kappa^2)S^2(n+d)^{2(n+d)}} \\
&\quad + \left(D\sum_{t=T_r}^{T}\left\|\left(\tilde{\Theta}_t - \Theta_*\right)^\top z_t\right\|^2\right)^{\frac{1}{2}}\sqrt{4(T-T_r)D(1+\kappa^2)S^2X_s^2} \\
&\leq \frac{\sqrt{8T_r}DS(1+\kappa^2)(n+d)^{2(n+d)}(\beta_T(\delta)+\upsilon_T(\delta))}{\sqrt{\mu}}\left(1+\frac{(1+\kappa^2)(n+d)^{2(n+d)}}{\mu}\right)^{\frac{\tau_0}{2}}\sqrt{\log\left(\frac{\det(V_{T_r})}{\det(\mu I)}\right)} \\
&\quad + \frac{\sqrt{8(T-T_r)}DS(1+\kappa^2)X_s^2\left(\beta_T(\delta)+\upsilon_T(\delta)\right)}{\sqrt{\mu}}\left(1+\frac{(1+\kappa^2)X_s^2}{\mu}\right)^{\frac{\tau_0}{2}}\sqrt{\log\left(\frac{\det(V_T)}{\det(V_{T_r})}\right)}
\end{aligned} \tag{B.140}$$

From Lemma 10 of Abbasi-Yadkori and Szepesvári [2], we have that $\log(\frac{\det(V_{T_r})}{\det(\mu I)}) \leq (n+d)\log(1+\frac{T_r(1+\kappa^2)(n+d)^{2(n+d)}}{\mu(n+d)})$ and $\log(\frac{\det(V_T)}{\det(V_{T_r})}) \leq (n+d)\log(1+\frac{T_r(1+\kappa^2)(n+d)^{2(n+d)}+(T-T_r)X_s^2}{\mu(n+d)})$.

After inserting these quantities into (B.140), we have the dimension dependency of $(n+d)^{2(n+d)} \times \sqrt{n(n+d)} \times (n+d)^{(n+d)\tau_0} \times (n+d)$ on the first term where $\sqrt{n(n+d)}$ is due to $\beta_T(\delta) + \upsilon_T(\delta)$. For the second term, for large enough T, we have the dimension dependency of $n \times \sqrt{n(n+d)} \times n^{(\tau_0/2)} \times \sqrt{n+d}$, where n comes from X_s^2. Thus, we achieve the following bound for $\left|R_T^{\mathrm{RLS}}\right|$:

$$\left|R_T^{\mathrm{RLS}}\right| = \tilde{O}\left((n+d)^{(\tau_0+2)(n+d)+1.5}\sqrt{n}\sqrt{T_r} + (n+d)n^{1.5+\tau_0/2}\sqrt{T-T_r}\right).$$

With the choice of $\mu = (1+\kappa^2)X_s^2$, the dependency of $n^{(\tau_0/2)}$ on the second term can be converted to a scalar multiplier of $\sqrt{2}^{\tau_0}$ and reduces the dependency of X_s^2 to X_s, which gives the advertised bound. □

Bounding $R_T^{\mathbf{mart}}$

Notice that this term is the same as R_1^{ζ} of StabL studied in Lemma B.6. Therefore, the same bound translates to R_T^{mart}.

Bounding $R_T^{\mathbf{TS}}$

Lemma B.20 (Bounding R_T^{TS} for TSAC)**.** *Let R_T^{TS} be as defined by* (B.130)*. Under the event of E_T, we have that*

$$\left|R_T^{TS}\right| \leq \tilde{O}\left(\sqrt{n}T_w + \mathrm{poly}(n, d, \log(1/\delta))\sqrt{T-T_w}\right),$$

with probability at least $1-2\delta$ if $T_w = \omega(\sqrt{T}\log T)$ for singular $A_{c,*}$ and $T_w = \omega(\log T)$ for non-singular $A_{c,*}$.

Proof. We decompose R_T^{TS} into two pieces as

$$R_T^{\mathrm{TS}} = \underbrace{\sum_{t=0}^{T_w}\left\{J_*(\tilde{\Theta}_t, \sigma_w^2 I) - J_*(\Theta_*, \sigma_w^2 I)\right\}\mathbb{1}_{E_t}}_{R_{T_w}^{\mathrm{TS,exp}}} + \underbrace{\sum_{t=T_w+1}^{T}\left\{J_*(\tilde{\Theta}_t, \sigma_w^2 I) - J_*(\Theta_*, \sigma_w^2 I)\right\}\mathbb{1}_{E_t}}_{R_T^{\mathrm{TS,noexp}}}$$

Since every sampled system is in set $\mathcal{S}$, we have that $\|P(\tilde{\Theta}_t)\|_F \leq D$ and therefore

$$\begin{aligned} R_{T_w}^{\text{TS,exp}} &\leq \sum_{t=0}^{T_w} \left| J_*(\tilde{\Theta}_t, \sigma_w^2 I) - J_*(\Theta_*, \sigma_w^2 I) \right| \mathbb{1}_{E_t} \\ &\leq \sum_{t=0}^{T_w} \left(\left| J_*(\tilde{\Theta}_t, \sigma_w^2 I) \right| + \left| J_*(\Theta_*, \sigma_w^2 I) \right| \right) \qquad \text{(B.141)} \\ &\leq \sqrt{n}\sigma_w^2 \sum_{t=0}^{T_w} \left(\|P(\tilde{\Theta}_t)\|_F + \|P(\Theta_*)\|_F \right) \leq 2\sqrt{n}\sigma_w^2 D T_w, \qquad \text{(B.142)} \end{aligned}$$

where we used the relation $\text{tr}(P) \leq \sqrt{n}\|P\|_F$ in (B.141). Considering the number of times a new TS sample is drawn, the second term in R_T^{TS} can be written as

$$R_K^{\text{TS,noexp}} = \sum_{k=0}^{K} \tau_0 \left\{ J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) - J_*(\Theta_*, \sigma_w^2 I) \right\} \mathbb{1}_{E_{t_k}},$$

where $t_k = T_w + 1 + k\tau_0$ and $K = \left\lceil \frac{T - T_w}{\tau_0} \right\rceil$. Denoting the information available to the controller up to time $t \geq 0$ via $\mathcal{F}_t^{\text{cnt}} := \sigma\left(\mathcal{F}_{t-1}, x_t\right)$, $R_K^{\text{TS,noexp}}$ can be further decomposed into two pieces as

$$\begin{aligned} R_K^{\text{TS,noexp}} = &\underbrace{\sum_{k=0}^{K} \tau_0 \left\{ J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) - \mathbb{E}\left[J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] \right\} \mathbb{1}_{E_t}}_{R_K^{\text{TS,1}}}, \\ &+ \underbrace{\sum_{k=0}^{K} \tau_0 \left\{ \mathbb{E}\left[J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) \,\middle|\, \mathcal{F}_t^{\text{cnt}}, E_{t_k} \right] - J_*(\Theta_*, \sigma_w^2 I) \right\} \mathbb{1}_{E_{t_k}}}_{R_K^{\text{TS,2}}}. \end{aligned}$$

We will investigate each term in order under the event of E_T.

Bounding $\mathbf{R_K^{TS,1}}$. Notice that $\{R_K^{\text{TS,1}}\}_{K \geq 0}$ is a martingale sequence with $\left| R_K^{\text{TS,1}} - R_{K-1}^{\text{TS,1}} \right| \leq 2\tau_0 \sigma_w^2 \sqrt{n} D$. Therefore it can be bounded by Azuma's inequality w.p. at least $1 - \delta$ as

$$R_K^{\text{TS,1}} \leq \sigma_w^2 D \sqrt{8n\tau_0^2 K \log(2/\delta)} \leq \sigma_w^2 D \sqrt{8n\tau_0 (T - T_w) \log(2/\delta)}. \qquad \text{(B.143)}$$

Bounding $\mathbf{R_K^{TS,2}}$. Denoting by $\mathcal{S}^{\text{opt}} := \left\{ \Theta \in \mathbb{R}^{(n+d)\times n} \,\middle|\, J_*(\Theta, \sigma_w^2 I) \leq J_*(\Theta_*, \sigma_w^2 I) \right\}$ the set of optimistic parameters and defining

$$R_k^{TS,2} := \left\{ \mathbb{E}\left[J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k} \right] - J_*(\Theta_*, \sigma_w^2 I) \right\} \mathbb{1}_{E_{t_k}}.$$

Notice that, for any $\Theta \in \mathcal{S}^{\text{opt}}$, we can write

$$\begin{aligned} R_k^{TS,2} &\leq \left\{\mathbb{E}\left[J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k}\right] - J_*(\Theta, \sigma_w^2 I)\right\} \mathbb{1}_{E_{t_k}} \\ &\leq \left|J_*(\Theta, \sigma_w^2 I) - \mathbb{E}\left[J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k}\right]\right| \mathbb{1}_{E_{t_k}}. \end{aligned}$$

As the above bound holds for any $\Theta \in \mathcal{S}^{\text{opt}}$, we can replace the right-hand side with an expectation over the optimistic set $\mathcal{S}^{\text{opt}}$. Specifically, we choose an i.i.d. copy of $\tilde{\Theta}_{t_k}$, that is, we choose a random variable $\tilde{\Theta}'_{t_k}$ which has the same distribution as $\tilde{\Theta}_{t_k}$ and independent from it. Then, we have that

$$\begin{aligned} R_k^{TS,2} &\leq \mathbb{E}\left[\left|J_*(\tilde{\Theta}'_{t_k}, \sigma_w^2 I) - \mathbb{E}\left[J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k}\right]\right| \mathbb{1}_{E_{t_k}} \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k}, \tilde{\Theta}'_{t_k} \in \mathcal{S}^{\text{opt}}\right] \\ &= \frac{\mathbb{E}\left[\left|J_*(\tilde{\Theta}'_{t_k}, \sigma_w^2 I) - \mathbb{E}\left[J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k}\right]\right| \mathbb{1}_{E_{t_k}} \mathbb{1}_{\tilde{\Theta}'_{t_k} \in \mathcal{S}^{\text{opt}}} \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k}\right]}{\mathbb{P}\left(\tilde{\Theta}'_{t_k} \in \mathcal{S}^{\text{opt}} \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, \hat{E}_{t_k}\right)} \end{aligned}$$

Denoting by $p_t^{\text{opt}} = \mathbb{P}\left(\tilde{\Theta}'_t \in \mathcal{S}^{\text{opt}} \,\middle|\, \mathcal{F}_t^{\text{cnt}}, \hat{E}_t\right)$ the probability of drawing cost optimistic TS samples, we can write further bounds on $R_k^{TS,2}$ as

$$\begin{aligned} R_k^{TS,2} &\leq \frac{1}{p_{t_k}^{\text{opt}}} \mathbb{E}\left[\left|J_*(\tilde{\Theta}'_{t_k}, \sigma_w^2 I) - \mathbb{E}\left[J_*(\tilde{\Theta}_{t_k}, \sigma_w^2 I) \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k}\right]\right| \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k}\right] \\ &= \frac{\sigma_w^2}{p_{t_k}^{\text{opt}}} \mathbb{E}\left[\left|\text{Tr}\left(P(\tilde{\Theta}'_{t_k}) - \mathbb{E}\left[P(\tilde{\Theta}_{t_k}) \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k}\right]\right)\right| \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k}\right] \\ &\leq \frac{n\sigma_w^2}{p_{t_k}^{\text{opt}}} \mathbb{E}\left[\left\|P(\tilde{\Theta}'_{t_k}) - \mathbb{E}\left[P(\tilde{\Theta}_{t_k}) \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k}\right]\right\|_2 \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k}\right] \end{aligned} \tag{B.144}$$

where we used the relation $|\text{tr}(A)| \leq n\|A\|_2$. Denoting $P_k \coloneqq \mathbb{E}\left[P(\tilde{\Theta}_{t_k}) \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k}\right]$, the following definition will be used in the rest of the section to understand the behavior of $R_k^{TS,2}$

$$\Delta_k \coloneqq \mathbb{E}\left[\|P(\tilde{\Theta}_{t_k}) - P_k\|_2 \,\middle|\, \mathcal{F}_{t_k}^{\text{cnt}}, E_{t_k}\right] \tag{B.145}$$

The following lemma will be used to bound Δ_k from above.

Lemma B.21. *For any $\Theta \in \mathcal{S}$, any positive definite matrix $V \in \mathbb{R}^{(n+d)\times(n+d)}$, and for any $i, j \in [n]$,*

$$\|\nabla P_{ij}(\Theta)\|_V \leq \Gamma \|H(\Theta)\|_V,$$

where $\Gamma \geq 0$ is a problem dependent constant.

Proof. Let $\delta P(\Theta, \delta\Theta)$ be the differential of $P(\Theta)$ in the direction $\delta\Theta$. Then, we have that

$$\begin{aligned}\delta P(\Theta, \delta\Theta) = A_c(\Theta)^\top \delta P(\Theta, \delta\Theta) A_c(\Theta) \\ + A_c(\Theta)^\top P(\Theta)\delta\Theta^\top H(\Theta) + H(\Theta)^\top \delta\Theta P(\Theta) A_c(\Theta)\end{aligned} \tag{B.146}$$

where $A_c(\Theta) = \Theta^\top H(\Theta)$ is the closed-loop matrix. We know that $P(\Theta)$ satisfies the Riccati equation as

$$P - A_c^\top P A_c = Q + K^\top R K \succ 0 \implies \left(P^{\frac{1}{2}} A_c P^{-\frac{1}{2}}\right)^\top P^{\frac{1}{2}} A_c P^{-\frac{1}{2}} \prec I$$

where we dropped Θ dependence for simplicity. Therefore, similarity transformation of the closed-loop matrix $\bar{A}_c := P^{\frac{1}{2}} A_c P^{-\frac{1}{2}}$ is a contraction, i.e., $\|P^{\frac{1}{2}} A_c P^{-\frac{1}{2}}\|_2 =: \sigma_\Theta < 1$. Multiplying both sides of (B.146) by $P^{-\frac{1}{2}}$ we obtain

$$\delta\bar{P}(\delta\Theta) = \bar{A}_c^\top \delta\bar{P}(\delta\Theta)\bar{A}_c + \bar{A}_c^\top P^{\frac{1}{2}}\delta\Theta^\top H P^{-\frac{1}{2}} + P^{-\frac{1}{2}} H^\top \delta\Theta P^{\frac{1}{2}} \bar{A}_c$$

where $\delta\bar{P}(\delta\Theta) = P^{-\frac{1}{2}}\delta P(\delta\Theta) P^{-\frac{1}{2}}$. Taking the spectral norm of both sides and using sub-multiplicativity of spectral norm as well as equivalence of matrix norms, we have that

$$\begin{aligned}\|\delta\bar{P}(\delta\Theta)\|_2 &\le \|A_c\|_2^2 \|\delta\bar{P}(\delta\Theta)\|_2 + 2\|\bar{A}_c\|_2 \|P^{\frac{1}{2}}\delta\Theta^\top H P^{-\frac{1}{2}}\|_2 \\ &\le \|A_c\|_2^2 \|\delta\bar{P}(\delta\Theta)\|_2 + 2\|\bar{A}_c\|_2 \|P^{\frac{1}{2}}\delta\Theta^\top H P^{-\frac{1}{2}}\|_F \\ &= \sigma_\Theta^2 \|\delta\bar{P}(\delta\Theta)\|_2 + 2\sigma_\Theta \|\delta\Theta^\top H\|_F\end{aligned}$$

By rearranging the inequality and using the property $\|\delta\Theta^\top H\|_F \le \|\delta\Theta\|_{V^{-1}} \|H\|_V$, we obtain

$$\|\delta\bar{P}(\delta\Theta)\|_2 \le \frac{2\sigma_\Theta}{1 - \sigma_\Theta^2} \|\delta\Theta\|_{V^{-1}} \|H\|_V$$

Observing that $\|\delta P(\delta\Theta)\|_2 = \|P^{\frac{1}{2}}\delta\bar{P}(\delta\Theta)P^{\frac{1}{2}}\|_2 \le \|P\|_2 \|\delta\bar{P}(\delta\Theta)\|_2 \le D\|\delta\bar{P}(\delta\Theta)\|_2$ and noting that $\|\nabla P_{ij}(\Theta)\|_V = \sup_{\|\delta\Theta\|_{V^{-1}}=1} \left|\delta P_{ij}(\delta\Theta)\right| \le \sup_{\|\delta\Theta\|_{V^{-1}}=1} \|\delta P(\delta\Theta)\|_2$, one can get

$$\|\nabla P_{ij}(\Theta)\|_V \le \frac{2D\sigma_\Theta}{1 - \sigma_\Theta^2} \|H(\Theta)\|_V$$

Observing that the function $\sigma_\Theta : \mathcal{S} \to \mathbb{R}_+$ is continuous on $\mathcal{S}$ and $\sigma_* := \max_{\Theta\in\mathcal{S}} \sigma_\Theta < 1$ as $\mathcal{S}$ is compact, we can further bound the scalar from above by Θ independent constant $\Gamma = \frac{2D\sigma_*}{1-\sigma_*^2} > 0$. □

The following lemma gives a useful upper bound on Δ_k.

Lemma B.22. *Let Δ_k be defined as in* (B.145)*. Then, for all $k \geq 0$, we have that*

$$\Delta_k \leq 2n^2 \upsilon_{t_k} \Gamma \, \mathbb{E}\left[\|H(\tilde{\Theta}_{t_k})\|_{V_{t_k}^{-1}} \,\middle|\, \mathcal{F}_{t_k}^{cnt}, E_{t_k} \right].$$

Proof. The proof follows directly from applying the bound in Lemma B.21 to Equation 11 in [7]. □

Finally, we are ready to give a bound on the summation of Δ_k terms

Lemma B.23. *Let Δ_k be defined as in* (B.145) *for any $k \geq 0$. Then, the following bound holds with probability at least $1 - \delta$*

$$\begin{aligned}\sum_{k=0}^{K} \Delta_k &\leq \frac{16n^2 \alpha \upsilon_T \Gamma}{1 + \frac{1}{\beta_T}} \left(\sum_{t=T_w+1}^{T} \|z_t\|_{V_t^{-1}} + 2\alpha \sqrt{2 \frac{T - T_w}{\tau_0} \frac{1+\kappa^2}{\mu} \log\left(\frac{2}{\delta}\right)} \right) \\ &\leq \tilde{O}(\mathrm{poly}(n, d, \log(1/\delta)) \sqrt{T - T_w})\end{aligned}$$

where $\alpha = (1 + 1/\beta_0^2)(\sqrt{2n \log(3n)} + \upsilon_T + (1+\kappa) S X_s)$.

Proof. Define $\bar{\Theta}_{t_k} = \hat{\Theta}_{t_k} + \beta_{t_k} V_{t_k}^{-\frac{1}{2}} \eta_{t_k}$. Using Proposition 9 in [7], we have that

$$\begin{aligned}\|H(\bar{\Theta}_{t_k})\|_{V_{t_k}^{-1}} &\leq \frac{8}{1 + \frac{1}{\beta_{t_k}}} \left\| H(\bar{\Theta}_{t_k}) \, \mathbb{E}\left[x_{t_k} x_{t_k}^\top \mathbb{1}_{\|x_{t_k}\| \leq \alpha} \,\middle|\, \mathcal{F}_{t_k - 1}, E_{t_k - 1}, \bar{\Theta}_{t_k} \right] \right\|_{V_{t_k}^{-1}} \\ &\leq \frac{8}{1 + \frac{1}{\beta_{t_k}}} \left\| \mathbb{E}\left[H(\bar{\Theta}_{t_k}) x_{t_k} x_{t_k}^\top \mathbb{1}_{\|x_{t_k}\| \leq \alpha} \,\middle|\, \mathcal{F}_{t_k - 1}, E_{t_k - 1}, \bar{\Theta}_{t_k} \right] \right\|_{V_{t_k}^{-1}} \\ &\leq \frac{8\alpha}{1 + \frac{1}{\beta_{t_k}}} \mathbb{E}\left[\left\| H(\bar{\Theta}_{t_k}) x_{t_k} \right\|_{V_{t_k}^{-1}} \mathbb{1}_{\|x_{t_k}\| \leq \alpha} \,\middle|\, \mathcal{F}_{t_k - 1}, E_{t_k - 1}, \bar{\Theta}_{t_k} \right]\end{aligned}$$

By Lemma B.22 and the preceding bound, we can write

$$\begin{aligned}\Delta_k &\leq 2n^2 \upsilon_{t_k} \Gamma \, \mathbb{E}\left[\|H(\tilde{\Theta}_{t_k})\|_{V_{t_k}^{-1}} \,\middle|\, \mathcal{F}_{t_k}^{\mathrm{cnt}}, E_{t_k} \right] = 2n^2 \upsilon_{t_k} \Gamma \frac{\mathbb{E}\left[\|H(\bar{\Theta}_{t_k})\|_{V_{t_k}^{-1}} \mathbb{1}_{\bar{\Theta}_{t_k} \in \mathcal{S}} \,\middle|\, \mathcal{F}_{t_k}^{\mathrm{cnt}}, E_{t_k} \right]}{\mathbb{P}\left\{ \bar{\Theta}_{t_k} \in \mathcal{S} \,\middle|\, \mathcal{F}_{t_k}^{\mathrm{cnt}}, E_{t_k} \right\}} \\ &\leq \frac{16n^2 \alpha \upsilon_{t_k} \Gamma}{1 + \frac{1}{\beta_{t_k}}} \frac{\mathbb{E}\left[\mathbb{E}\left[\|H(\bar{\Theta}_{t_k}) x_{t_k}\|_{V_{t_k}^{-1}} \mathbb{1}_{\|x_{t_k}\| \leq \alpha} \,\middle|\, \mathcal{F}_{t_k - 1}, E_{t_k - 1}, \bar{\Theta}_{t_k} \right] \mathbb{1}_{\bar{\Theta}_{t_k} \in \mathcal{S}} \,\middle|\, \mathcal{F}_{t_k}^{\mathrm{cnt}}, E_{t_k} \right]}{\mathbb{P}\left\{ \bar{\Theta}_{t_k} \in \mathcal{S} \,\middle|\, \mathcal{F}_{t_k}^{\mathrm{cnt}}, E_{t_k} \right\}} \\ &= \frac{16n^2 \alpha \upsilon_{t_k} \Gamma}{1 + \frac{1}{\beta_{t_k}}} \underbrace{\mathbb{E}[\mathbb{E}[\|\underbrace{H(\tilde{\Theta}_{t_k}) x_{t_k}}_{z_{t_k}}\|_{V_{t_k}^{-1}} \mathbb{1}_{\|x_{t_k}\| \leq \alpha} \,|\, \mathcal{F}_{t_k - 1}, E_{t_k - 1}, \tilde{\Theta}_{t_k}] \,|\, \mathcal{F}_{t_k}^{\mathrm{cnt}}, E_{t_k}]}_{=: Y_k}.\end{aligned}$$

Notice that $\mathbb{E}\left[Y_k \middle| \mathcal{F}_{t_{k-1}}\right] = \mathbb{E}\left[\|z_{t_k}\|_{V_{t_k}^{-1}} \mathbb{1}_{\|x_{t_k}\| \le \alpha} \middle| \mathcal{F}_{t_{k-1}}\right]$ by law of iterated expectations and $\|z_{t_k}\|_{V_{t_k}^{-1}} \mathbb{1}_{\|x_{t_k}\| \le \alpha} \le \frac{1}{\sqrt{\mu}} \|H(\tilde{\Theta}_{t_k}) x_{t_k}\| \mathbb{1}_{\|x_{t_k}\| \le \alpha} \le \sqrt{\frac{1+\kappa^2}{\mu}}\alpha$.

Therefore, the sequence $\left\{Y_k - \|z_{t_k}\|_{V_{t_k}^{-1}} \mathbb{1}_{\|x_{t_k}\| \le \alpha}\right\}_{k \ge 0}$ is a bounded martingale difference sequence. By Azuma's inequality, we have that with probability at least $1 - \delta$,

$$\sum_{k=0}^{K} \left(Y_k - \|z_{t_k}\|_{V_{t_k}^{-1}} \mathbb{1}_{\|x_{t_k}\| \le \alpha}\right) \le 2\alpha \sqrt{2 \frac{T - T_w}{\tau_0} \frac{1 + \kappa^2}{\mu} \log\left(\frac{2}{\delta}\right)}.$$

We can bound the sum of $\|z_{t_k}\|_{V_{t_k}^{-1}}$ terms using Lemma 10 of [3] and Hölder's inequality as

$$\begin{aligned}
\sum_{k=0}^{K} \|z_{t_k}\|_{V_{t_k}^{-1}} &\le \sum_{k=0}^{K} \|z_{t_k}\|_{V_{t_k}^{-1}} + \sum_{k=0}^{K} \sum_{t=t_k+1}^{t_{k+1}-1} \|z_t\|_{V_t^{-1}} \\
&= \sum_{t=T_w+1}^{T} \|z_t\|_{V_t^{-1}} \le \sqrt{T - T_w} \log \frac{\det(V_T)}{\det(V_{T_w})}
\end{aligned}$$

Combining these results, we obtain the desired bound

$$\sum_{k=0}^{K} \Delta_k \le \frac{16 n^2 \alpha \upsilon_T \Gamma}{1 + \frac{1}{\beta_T}} \left(\sum_{t=T_w+1}^{T} \|z_t\|_{V_t^{-1}} + 2\alpha \sqrt{2 \frac{T - T_w}{\tau_0} \frac{1 + \kappa^2}{\mu} \log\left(\frac{2}{\delta}\right)} \right).$$

□

Now, we are ready to bound $R_K^{TS,2}$. Under the event E_T Theorem 3.4 suggests that $1/p_{t^{\text{opt}}} \le O(1)$ if $T_w = \omega(\sqrt{T} \log T)$ for singular $A_{c,*}$ and $T_w = \omega(\log T)$ for non-singular $A_{c,*}$. Using this result together with Lemma B.23, we have that

$$R_K^{TS,2} = \sum_{k=0}^{K} \tau_0 R_k^{TS,2} \le n \sigma_w^2 \tau_0 \sum_{k=0}^{K} \frac{\Delta_k}{p_{t_k^{\text{opt}}}} \le \tilde{O}(\text{poly}(n, d) \sqrt{(T - T_w) \log(1/\delta)}) \tag{B.147}$$

with probability at least $1 - \delta$. Combining the above with (B.142) and (B.143), we obtain the desired bound. □

Bounding R_T^{gap}

Lemma B.24 (Bounding R_T^{gap} for TSAC)**.** *Let R_T^{gap} be as defined by* (B.133)*. Under the event of E_T, we have that*

$$\left|R_T^{gap}\right| = \tilde{O}\left(\text{poly}(n, d) \sqrt{T \log(1/\delta)}\right),$$

with probability at least $1 - 2\delta$ for large enough T.

Proof.

$$R_T^{\text{gap}} = \sum_{t=0}^{T} \mathbb{E}\left[x_{t+1}^\top \left(P(\tilde{\Theta}_{t+1}) - P(\tilde{\Theta}_t)\right) x_{t+1} \mathbb{1}_{E_{t+1}} \,\middle|\, \mathcal{F}_t\right] \tag{B.148}$$

$$= \sum_{t=0}^{K} \mathbb{E}\left[x_{t_k+1}^\top \left(P(\tilde{\Theta}_{t_k+1}) - P(\tilde{\Theta}_{t_k})\right) x_{t_k+1} \mathbb{1}_{E_{t_k+1}} \,\middle|\, \mathcal{F}_{t_k}\right] \tag{B.149}$$

Separating the duration of TSAC into two parts at $t = T_r$, we obtain two same term achieved in [7]. Note that in Abeille and Lazaric [7], the authors follow frequent update rule and TSAC updates every τ_0 time-steps. The proof of these terms similarly follow Section 5.2 in [7] and using Lemma B.23 we obtain

$$O((n+d)^{n+d}\sqrt{T_r} + \text{poly}(n,d)\sqrt{T - T_r}).$$

Note that there is an additional τ_0 factor in these bounds, due to the "relatively slower" update of TSAC. For large enough T such that the second term dominates the overall upper bound, we obtain the advertised guarantee. □

B.2.5 Proof of Theorem 3.3

Collecting the regret terms derived in subsections of Appendix B.2.4, for large enough T, under the event E_T, we have that

$$\begin{aligned}
R_{T_w}^{\text{exp}} &= \tilde{O}\left((n+d)^{n+d} T_w\right), \quad \text{w.p. } 1-\delta \\
R_T^{\text{RLS}} &= \tilde{O}\left((n+d)^{n+d}\sqrt{T_r} + \text{poly}(n,d,\log(1/\delta))\sqrt{T - T_r}\right), \\
R_T^{\text{mart}} &= \tilde{O}\left((n+d)^{n+d}\sqrt{T_r} + \text{poly}(n,d,\log(1/\delta))\sqrt{T - T_w}\right), \quad \text{w.p. } 1-\delta \\
R_T^{\text{gap}} &= \tilde{O}\left((n+d)^{n+d}\sqrt{T_r} + \text{poly}(n,d,\log(1/\delta))\sqrt{T - T_r}\right), \quad \text{w.p. } 1-2\delta
\end{aligned}$$

and choosing $T_w = \omega(\sqrt{T}\log T)$ for singular $A_{c,*}$ and $T_w = \omega(\log T)$ for non-singular $A_{c,*}$ gives

$$R_T^{\text{TS}} = \tilde{O}\left(\text{poly}(n,d) T_w + \text{poly}(n,d,\log(1/\delta))\sqrt{T - T_r}\right), \quad \text{w.p. } 1-2\delta.$$

Recall that the event E_T is true with probability at least $1 - 4\delta$. Combining all these bounds, we have the overall regret bound as

$$R_T = \tilde{O}\left((n+d)^{n+d} T_w + \text{poly}(n,d,\log(1/\delta))\sqrt{T - T_w}\right), \quad \text{w.p. } 1-10\delta. \tag{B.150}$$

Notice that R_T is linear in the initial exploration time T_w with an exponential dimension dependency. Also note that $T_w \geq T_0 \coloneqq \text{poly}(\log(1/\delta), \sigma_w^{-1}, n, d, \bar{\alpha}, \gamma^{-1}, \kappa)$ guarantees a stabilizing controller by Lemma 3.6. In order to control the growth of R_T by $\tilde{O}(\sqrt{T})$, the initial exploration time can maximally be in the order of $(\sqrt{T})^{1+o(1)}$ where $T^{o(1)}$ hides all multiplicative sub-polynomial growths, i.e., $T_w = O\left((\sqrt{T})^{1+o(1)}\right) = \tilde{O}(\sqrt{T})$.

On the other hand, Theorem 3.4 puts strict lower bounds on the growth of T_w in order to maintain asymptotically constant optimistic probability. In particular, for singular $A_{c,*}$, this condition is stated as $T_w = \omega(\sqrt{T}\log T)$. Combined with the required upper bound $O\left((\sqrt{T})^{1+o(1)}\right)$, it must be that $T_w = \max\left(T_0,\ c(\sqrt{T}\log T)^{1+o(1)}\right)$ for a constant $c > 0$ for large enough T. Inserting this result in (B.150) gives us

$$R_T = \tilde{O}\left((n+d)^{n+d}\sqrt{T}\right), \quad \text{w.p. } 1 - 10\delta$$

for large enough T. Observe that exponential dimension dependence is unavoidable in this case as the system is excited with isotropic noise in every direction long enough to dominate with exponential dimension.

For non-singular $A_{c,*}$, the lower bound is stated as $T_w = \omega(\log T)$. For large enough T, choosing $T_w = \max\left(T_0,\ c(\log T)^{1+o(1)}\right)$ for a constant $c > 0$ is sufficient to satisfy both the upper and lower bounds on T_w. Inserting this result in (B.150) gives us

$$R_T = \tilde{O}\left(\text{poly}(n, d, \log(1/\delta))\sqrt{T}\right), \quad \text{w.p. } 1 - 10\delta$$

for large enough T. Observe that the exponential dimension dependence is not dominant anymore since logarithmically large T_w is sufficient to guarantee asymptotically constant optimistic probability.

Appendix C

FURTHER PROOFS FOR CHAPTER 5

C.1 Proofs of Section 5.3

C.1.1 Proof of Theorem 5.3

Proof. For a single input-output trajectory $\{y_t, u_t\}_{t=1}^{\tau}$, where $\tau \leq T$, using the representation in (5.18), we can write the following for the given system,

$$Y_\tau = \Phi_\tau \mathcal{G}_{\mathbf{yu}}{}^\top + \underbrace{E_\tau + N_\tau}_{\text{Noise}} \qquad \text{where} \tag{C.1}$$

$$\begin{aligned}
\mathcal{G}_{\mathbf{yu}} &= \begin{bmatrix} CF, & C\bar{A}F, & \ldots, & C\bar{A}^{H-1}F, & CB, & C\bar{A}B, & \ldots, & C\bar{A}^{H-1}B \end{bmatrix} \in \mathbb{R}^{m\times(m+p)H} \\
Y_\tau &= [y_H,\ y_{H+1},\ \ldots,\ y_\tau]^\top \in \mathbb{R}^{(\tau-H)\times m} \\
\Phi_\tau &= [\phi_H,\ \phi_{H+1},\ \ldots,\ \phi_\tau]^\top \in \mathbb{R}^{(\tau-H)\times(m+p)H} \\
E_\tau &= [e_H,\ e_{H+1},\ \ldots,\ e_\tau]^\top \in \mathbb{R}^{(\tau-H)\times m} \\
N_\tau &= \begin{bmatrix} C\bar{A}^H x_0,\ C\bar{A}^H x_1, \ldots, C\bar{A}^H x_{\tau-H} \end{bmatrix}^\top \in \mathbb{R}^{(\tau-H)\times m}.
\end{aligned}$$

$\widehat{\mathcal{G}}_{\mathbf{y}}$ is the solution to (5.21), *i.e.,* $\min_X \|Y_\tau - \Phi_\tau X^\top\|_F^2 + \lambda\|X\|_F^2$. Hence, we get $\widehat{\mathcal{G}}_{\mathbf{y}}^\top = (\Phi_\tau^\top\Phi_\tau + \lambda I)^{-1}\Phi_\tau^\top Y_\tau$.

$$\begin{aligned}
\widehat{\mathcal{G}}_{\mathbf{y}} &= \left[(\Phi_\tau^\top\Phi_\tau + \lambda I)^{-1}\Phi_\tau^\top(\Phi_\tau \mathcal{G}_{\mathbf{yu}}{}^\top + E_\tau + N_\tau)\right]^\top \\
&= \Big[(\Phi_\tau^\top\Phi_\tau + \lambda I)^{-1}\Phi_\tau^\top\,(E_\tau + N_\tau) + (\Phi_\tau^\top\Phi_\tau + \lambda I)^{-1}\Phi_\tau^\top\Phi_\tau \mathcal{G}_{\mathbf{yu}}{}^\top \\
&\qquad + \lambda(\Phi_\tau^\top\Phi_\tau + \lambda I)^{-1}\mathcal{G}_{\mathbf{yu}}{}^\top - \lambda(\Phi_\tau^\top\Phi_\tau + \lambda I)^{-1}\mathcal{G}_{\mathbf{yu}}{}^\top\Big]^\top \\
&= \left[(\Phi_\tau^\top\Phi_\tau + \lambda I)^{-1}\Phi_\tau^\top E_\tau + (\Phi_\tau^\top\Phi_\tau + \lambda I)^{-1}\Phi_\tau^\top N_\tau + \mathcal{G}_{\mathbf{yu}}{}^\top - \lambda(\Phi_\tau^\top\Phi_\tau + \lambda I)^{-1}\mathcal{G}_{\mathbf{yu}}{}^\top\right]^\top
\end{aligned}$$

Using $\widehat{\mathcal{G}}_{\mathbf{y}}$, we get

$$
\begin{aligned}
&|\operatorname{Tr}(X(\widehat{\mathcal{G}}_{\mathbf{y}}-\mathcal{G}_{\mathbf{yu}})^\top)| && \text{(C.2)}\\
&= |\operatorname{Tr}(X(\Phi_\tau^\top\Phi_\tau+\lambda I)^{-1}\Phi_\tau^\top E_\tau)+\operatorname{Tr}(X(\Phi_\tau^\top\Phi_\tau+\lambda I)^{-1}\Phi_\tau^\top N_\tau)-\lambda\operatorname{Tr}(X(\Phi_\tau^\top\Phi_\tau+\lambda I)^{-1}\mathcal{G}_{\mathbf{yu}}{}^\top)|\\
&\le |\operatorname{Tr}(X(\Phi_\tau^\top\Phi_\tau+\lambda I)^{-1}\Phi_\tau^\top E_\tau)|+|\operatorname{Tr}(X(\Phi_\tau^\top\Phi_\tau+\lambda I)^{-1}\Phi_\tau^\top N_\tau)|+\lambda|\operatorname{Tr}(X(\Phi_\tau^\top\Phi_\tau+\lambda I)^{-1}\mathcal{G}_{\mathbf{yu}}{}^\top)|\\
&\le \sqrt{\operatorname{Tr}(X(\Phi_\tau^\top\Phi_\tau+\lambda I)^{-1}X^\top)\operatorname{Tr}(E_\tau^\top\Phi_\tau(\Phi_\tau^\top\Phi_\tau+\lambda I)^{-1}\Phi_\tau^\top E_\tau)} && \text{(C.3)}\\
&\quad+\sqrt{\operatorname{Tr}(X(\Phi_\tau^\top\Phi_\tau+\lambda I)^{-1}X^\top)\operatorname{Tr}(N_\tau^\top\Phi_\tau(\Phi_\tau^\top\Phi_\tau+\lambda I)^{-1}\Phi_\tau^\top N_\tau)}\\
&\quad+\lambda\sqrt{\operatorname{Tr}(X(\Phi_\tau^\top\Phi_\tau+\lambda I)^{-1}X^\top)\operatorname{Tr}(\mathcal{G}_{\mathbf{yu}}(\Phi_\tau^\top\Phi_\tau+\lambda I)^{-1}\mathcal{G}_{\mathbf{yu}}{}^\top)}\\
&= \sqrt{\operatorname{Tr}(X(\Phi_\tau^\top\Phi_\tau+\lambda I)^{-1}X^\top)}\ \times\\
&\left[\sqrt{\operatorname{Tr}(E_\tau^\top\Phi_\tau(\Phi_\tau^\top\Phi_\tau+\lambda I)^{-1}\Phi_\tau^\top E_\tau)}+\sqrt{\operatorname{Tr}(N_\tau^\top\Phi_\tau(\Phi_\tau^\top\Phi_\tau+\lambda I)^{-1}\Phi_\tau^\top N_\tau)}+\lambda\sqrt{\operatorname{Tr}(\mathcal{G}_{\mathbf{yu}}(\Phi_\tau^\top\Phi_\tau+\lambda I)^{-1}\mathcal{G}_{\mathbf{yu}}{}^\top)}\right]
\end{aligned}
$$

where (C.3) follows from $|\operatorname{Tr}(ABC^\top)| \le \sqrt{\operatorname{Tr}(ABA^\top)\operatorname{Tr}(CBC^\top)}$ for positive definite B due to Cauchy Schwarz (weighted inner-product). For $X = (\widehat{\mathcal{G}}_{\mathbf{y}}-\mathcal{G}_{\mathbf{yu}})(\Phi_\tau^\top\Phi_\tau+\lambda I)$, we get

$$
\sqrt{\operatorname{Tr}((\widehat{\mathcal{G}}_{\mathbf{y}}-\mathcal{G}_{\mathbf{yu}})V_\tau(\widehat{\mathcal{G}}_{\mathbf{y}}-\mathcal{G}_{\mathbf{yu}})^\top)} \le \sqrt{\operatorname{Tr}(E_\tau^\top\Phi_\tau V_\tau^{-1}\Phi_\tau^\top E_\tau)}+\sqrt{\operatorname{Tr}(N_\tau^\top\Phi_\tau V_\tau^{-1}\Phi_\tau^\top N_\tau)}+\sqrt{\lambda}\|\mathcal{G}_{\mathbf{yu}}\|_F \tag{C.4}
$$

where V_τ is the regularized design matrix at time τ. Let $\max_{i\le\tau}\|\phi_i\| \le \Upsilon\sqrt{H}$ and $\max_{H\le i\le\tau}\|x_i\| \le \mathcal{X}$, *i.e.,* in data collection bounded inputs are used. The first term on the right hand side of (C.4) can be bounded using Theorem 1 of [2] since e_t is $\|C\Sigma C^\top+\sigma_z^2 I\|$-sub-Gaussian vector. Therefore, for $\delta\in(0,1)$, with probability at least $1-\delta$,

$$
\sqrt{\operatorname{Tr}(E_\tau^\top\Phi_t V_\tau^{-1}\Phi_\tau^\top E_\tau)} \le \sqrt{m\|C\Sigma C^\top+\sigma_z^2 I\|\log\left(\frac{\det(V_\tau)^{1/2}}{\delta\det(\lambda I)^{1/2}}\right)} \tag{C.5}
$$

For the second term,

$$
\begin{aligned}
\sqrt{\operatorname{Tr}(N_\tau^\top\Phi_\tau V_\tau^{-1}\Phi_\tau^\top N_\tau)} \le \frac{1}{\sqrt{\lambda}}\|N_\tau^\top\Phi_\tau\|_F &\le \sqrt{\frac{m}{\lambda}}\left\|\sum_{i=H}^{\tau}\phi_i(C\bar{A}^H x_{i-H})^\top\right\|\\
&\le \tau\sqrt{\frac{m}{\lambda}}\max_{i\le\tau}\left\|\phi_i(C\bar{A}^H x_{i-H})^\top\right\|\\
&\le \tau\sqrt{\frac{m}{\lambda}}\|C\|\upsilon^H\max_{i\le\tau}\|\phi_i\|\|x_{i-H}\|\\
&\le \tau\sqrt{\frac{m}{\lambda}}\|C\|\upsilon^H\Upsilon\sqrt{H}\mathcal{X}.
\end{aligned}
$$

Picking $H = \frac{2\log(T)+\log(\Upsilon X)+0.5\log(m/\lambda)+\log(\|C\|)}{\log(1/\upsilon)}$ gives

$$\sqrt{\mathrm{Tr}(N_\tau^\top \Phi_\tau V_\tau^{-1} \Phi_\tau^\top N_\tau)} \leq \frac{\tau}{T^2}\sqrt{H}. \tag{C.6}$$

Combining (C.5) and (C.6) gives the self-normalized estimation error bound state in the theorem. □

C.2 Proofs of Section 5.4

C.2.1 Proof of Lemma 5.6:

The proof of Lemma 5.6 follows similar arguments with the proof of Lemma 4.2 of [161]. The main difference is that in LqgOpt, the system estimations are refined during the adaptive control period, thus the control policy is refined. Also, since the behavior of a system and its similarity transformation is the same, without loss of generality we assume that similarity transformation $\mathbf{T} = I$. In the following, we show the boundedness for the contractible systems which can be extended to stabilizable systems applying the same policy for a long enough duration to cancel out the effects of similarity transformations that makes the closed-loop system contractible i.e., Lemma 3.5. Let $\rho = \max\{1-\gamma_1, 1-\gamma_2, 1-\gamma_3\}$ Assume that $\Theta \in (C_A(t) \times C_B(t) \times C_C(t) \times C_L(t))$ for all $t \geq T_w$, which is holds with probability $1-\delta$. We can write the decomposition for $\hat{x}_{t|t,\tilde{\Theta}}$ as follows,

$$\begin{aligned}
\hat{x}_{t|t,\tilde{\Theta}} &= \hat{x}_{t|t-1,\tilde{\Theta}} + \tilde{L}_t(y_t - \tilde{C}_t\hat{x}_{t|t-1,\tilde{\Theta}}) \\
&= \tilde{A}_{t-1}\hat{x}_{t-1|t-1,\tilde{\Theta}} - \tilde{B}_{t-1}\tilde{K}_{t-1}\hat{x}_{t-1|t-1,\tilde{\Theta}} + \tilde{L}_t(y_t - \tilde{C}_t(\tilde{A}_{t-1}\hat{x}_{t-1|t-1,\tilde{\Theta}} - \tilde{B}_{t-1}\tilde{K}_{t-1}\hat{x}_{t-1|t-1,\tilde{\Theta}})) \\
&= (I - \tilde{L}_t\tilde{C}_t)(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1})\hat{x}_{t-1|t-1,\tilde{\Theta}} + \tilde{L}_t y_t \\
&= (I - \tilde{L}_t\tilde{C}_t)(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1})\hat{x}_{t-1|t-1,\tilde{\Theta}} \\
&\qquad + \tilde{L}_t\left(Cx_t - C\hat{x}_{t|t-1,\tilde{\Theta}} + C\hat{x}_{t|t-1,\tilde{\Theta}} + z_t\right) \\
&= (I - \tilde{L}_t\tilde{C}_t)(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1})\hat{x}_{t-1|t-1,\tilde{\Theta}} \\
&\qquad + \tilde{L}_t\left(Cx_t - C\hat{x}_{t|t-1,\tilde{\Theta}} + C(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1})\hat{x}_{t-1|t-1,\tilde{\Theta}} + z_t\right) \\
&= \left(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1} - \tilde{L}_t\left(\tilde{C}_t\tilde{A}_{t-1} - \tilde{C}_t\tilde{B}_{t-1}\tilde{K}_{t-1} - C\tilde{A}_{t-1} + C\tilde{B}_{t-1}\tilde{K}_{t-1}\right)\right)\hat{x}_{t-1|t-1,\tilde{\Theta}} \\
&\qquad + \tilde{L}_t C(x_t - \hat{x}_{t|t-1,\Theta} + \hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\tilde{\Theta}}) + \tilde{L}_t z_t \\
&= \left(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1} - \tilde{L}_t\left(\tilde{C}_t\tilde{A}_{t-1} - \tilde{C}_t\tilde{B}_{t-1}\tilde{K}_{t-1} - C\tilde{A}_{t-1} + C\tilde{B}_{t-1}\tilde{K}_{t-1}\right)\right)\hat{x}_{t-1|t-1,\tilde{\Theta}} \\
&\qquad + \tilde{L}_t C(x_t - \hat{x}_{t|t-1,\Theta}) + \tilde{L}_t C(\hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\tilde{\Theta}}) + \tilde{L}_t z_t.
\end{aligned} \tag{C.7}$$

Thus, the dynamics of $\hat{x}_{t|t,\tilde{\Theta}}$ is governed by

$$\mathbf{N}_t = \tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1} - \tilde{L}_t\left(\tilde{C}_t\tilde{A}_{t-1} - \tilde{C}_t\tilde{B}_{t-1}\tilde{K}_{t-1} - C\tilde{A}_{t-1} + C\tilde{B}_{t-1}\tilde{K}_{t-1}\right)$$

and it is driven by the process of $\tilde{L}_t C(x_t - \hat{x}_{t|t-1,\Theta}) + \tilde{L}_t C(\hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\tilde{\Theta}}) + \tilde{L}_t z_t$. Let $T_u = T_B \left(\frac{2\psi\rho}{1-\rho}\right)^2$. With the Assumption 5.1, and for $T_w \geq T_u$, we have that $\|\tilde{C}_t - C\| \leq \frac{1-\rho}{2\psi\rho}$ which gives $\|\mathbf{N}_t\| \leq \frac{1+\rho}{2} < 1$ for all $t \geq T_w$. Similar to the proof of Lemma 4.2 in [161], we have that $\tilde{L}_t C(x_t - \hat{x}_{t|t-1,\Theta}) + \tilde{L}_t z_t$ is $\psi(\|C\|\|\Sigma\|^{1/2} + \sigma_z)$-sub-Gaussian, thus it's ℓ_2-norm can be bounded using Lemma C.3.12:

$$\|\tilde{L}_t C(x_t - \hat{x}_{t|t-1,\Theta}) + \tilde{L}_t z_t\| \leq \psi(\|C\|\|\Sigma\|^{1/2} + \sigma_z)\sqrt{2n \log(2nT/\delta)}$$

for all $t \geq T_w$ with probability at least $1 - \delta$. A special care is needed for $\hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\tilde{\Theta}}$. Denote $\Delta_t = \hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\tilde{\Theta}}$. Consider the decomposition given in equation (51) in Lale et al. [161]. In this setting, since at each time step after the warm-up, the estimation errors are monotonically decreasing, therefore we can upper bound the norm of each term in the decomposition by the norm of the term at the time of end of warm-up. Let

$$\begin{aligned} T_\alpha &= T_B \left(\frac{\kappa_2 \left(1 + \psi(1 + \|C\|)\right)}{\rho/2}\right)^2, \qquad T_\gamma = T_A \frac{\sigma_n^2(\bar{A})}{4} \left(\frac{1 + \kappa_2(1 + \psi\|B\|)}{\rho/2}\right)^2, \\ T_\beta &= T_A \frac{\sigma_n^2(\bar{A})}{4} \left(\frac{\kappa_2\|B\|(1+\psi+\psi\|C\|)(\kappa_1\psi+(1+\kappa_2)(1+\psi))}{(1-\rho)^2}\right)^2. \end{aligned} \tag{C.8}$$

Thus, using the arguments in [161], we can show that after a warm-up period of $T_w \geq \max\{T_\alpha, T_\gamma\}$, we have that for all $t \geq T_w$, $\max\{\|(A+(\tilde{A}_t - A - \tilde{B}_t\tilde{K}_t + B\tilde{K}_t))(I - \tilde{L}_t\tilde{C}_t)\|, \|A - B\tilde{K}_t + B\tilde{K}_t\tilde{L}_t(\tilde{C}_t - C)\|\} \leq \sigma < 1$. Using the inductive argument given in [161], we can show that for all $t \geq T_w \geq T_\beta$, $\|\Delta_t\| \leq \bar{\Delta}$ with probability $1 - \delta$. Notice that the definition of $\bar{\Delta}$ still includes the same terms given in equation (54) of Lale et al. [161] but $\beta_A, \beta_B, \beta_C$ is replaced with $\beta_A(T_w), \beta_B(T_w), \beta_C(T_w)$ and ΔL is replaced by $2\beta_L(T_w)$ due to new estimation method, i.e.,

$$\bar{\Delta} = 10 \left(\frac{\bar{\kappa}}{1-\rho} + \frac{\bar{\beta}\bar{\xi}}{(1-\rho)^2}\right) \left(\|C\|\|\Sigma\|^{1/2} + \sigma_z\right) \sqrt{2m \log(2mT/\delta)}$$

for $\bar{\kappa} = 2\kappa_1\beta_L(T_w) + 2\psi(\beta_A(T_w) + \kappa_2\beta_B(T_w))$, $\bar{\beta} = 2\psi\beta_C(T_w)(\kappa_1 + 2(\beta_A(T_w) + \kappa_2\beta_B(T_w))) + 2(\beta_A(T_w) + \kappa_2\beta_B(T_w))$ and $\bar{\xi} = \psi(\rho + 2(\beta_A(T_w) + \kappa_2\beta_B(T_w))) +$

$2\|B\|\kappa_2\beta_L(T_w)$. Thus, we get

$$\|\hat{x}_{t|t,\tilde{\Theta}}\| = \left\|\sum_{i=1}^{t} \mathbf{N}^{t-i}\left(\tilde{L}_i C(x_{i-1} - \hat{x}_{i|i-1,\Theta}) + \tilde{L}_i C(\hat{x}_{i|i-1,\Theta} - \hat{x}_{i|i-1,\tilde{\Theta}}) + \tilde{L}_i z_i\right)\right\| \quad \text{(C.9)}$$

$$\leq \max_{1\leq i\leq t}\left\|\tilde{L}_i C(x_{i-1} - \hat{x}_{i|i-1,\Theta}) + \tilde{L}_i C(\hat{x}_{i|i-1,\Theta} - \hat{x}_{i|i-1,\tilde{\Theta}}) + \tilde{L}_i z_i\right\|\left(\sum_{i=1}^{t}\|\mathbf{M}\|^{t-i}\right) \quad \text{(C.10)}$$

$$\leq \frac{2}{1-\rho}\max_{1\leq i\leq t}\left\|\tilde{L}_i C(x_{i-1} - \hat{x}_{i|i-1,\Theta}) + \tilde{L}_i C(\hat{x}_{i|i-1,\Theta} - \hat{x}_{i|i-1,\tilde{\Theta}}) + \tilde{L}_i z_i\right\| \quad \text{(C.11)}$$

$$\leq \tilde{\mathcal{X}} := \frac{2\psi\left(\|C\|\bar{\Delta} + \left(\|C\|\|\Sigma\|^{1/2} + \sigma_z\right)\sqrt{2n\log(2nT/\delta)}\right)}{1-\rho}. \quad \text{(C.12)}$$

with probability $1 - 2\delta$. For y_t, we have the following decomposition,

$$\begin{aligned} y_t &= C\hat{x}_{t|t-1,\tilde{\Theta}} + C(x_t - \hat{x}_{t|t-1,\tilde{\Theta}}) + z_t \\ &= C\hat{x}_{t|t-1,\tilde{\Theta}} + C(x_t - \hat{x}_{t|t-1,\Theta}) + C(\hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\tilde{\Theta}}) + z_t \\ &= C(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1})\hat{x}_{t-1|t-1,\tilde{\Theta}} + C(x_t - \hat{x}_{t|t-1,\Theta}) + C(\hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\tilde{\Theta}}) + z_t \end{aligned}$$

Using similar analysis with $\hat{x}_{t|t,\tilde{\Theta}}$, we get the following bound for y_t for all $t \geq T_w$:

$$\|y_t\| \leq \rho\|C\|\tilde{\mathcal{X}} + (\|C\|\|\Sigma\|^{1/2} + \sigma_z)\sqrt{2m\log(2mT/\delta)} + \|C\|\bar{\Delta}$$

with probability $1 - 2\delta$. Thus, all statements of Lemma 5.6 hold with probability at least $1 - 3\delta$.

C.2.2 Upper bound on $\|\tilde{\Sigma} - \mathbf{S}^{-1}\Sigma\mathbf{S}\|$, Proof of Lemma 5.10

In this section, we provide the concentration results on $\|\tilde{\Sigma} - \mathbf{S}^{-1}\Sigma\mathbf{S}\|$. $\mathbf{S} \in \mathbb{R}^{n\times n}$ is a similarity transformation that is composed of two similarity transformations. The first one takes the system Θ and transforms it to $\bar{\Theta}$, the output of SysId algorithm. The second similarity transformation is the unitary matrix that is proven to exist in Theorem 5.4. We deploy the fixed point argument from [191] to bound $\|\tilde{\Sigma} - \mathbf{S}^{-1}\Sigma\mathbf{S}\|$.

Proof. For the simplicity of the presentation of the proof, without loss of generality, let $A = \mathbf{S}^{-1}A\mathbf{S}$, $C = C\mathbf{S}$, *i.e.* assume that $\mathbf{S} = I$. Given parameters $(A, C, \sigma_w^2 I, \sigma_z^2 I)$, define $F(X, A, C)$ such that

$$\begin{aligned} F(X, A, C) &= X - AXA^\top + AXC^\top\left(CXC^\top + \sigma_z^2 I\right)^{-1}CXA^\top - \sigma_w^2 I \\ &= X - A(I + \sigma_z^{-2}XC^\top C)^{-1}XA^\top - \sigma_w^2 I \end{aligned}$$

where last equality follows from matrix inversion lemma. Moreover, notice that solving algebraic Riccati equation for steady state error covariance matrix of state estimation for $(A, C, \sigma_w^2 I, \sigma_z^2 I)$ is equivalent to finding the unique positive definite solution to X such that $F(X, A, C) = 0$. The solution for the underlying system Θ, $F(X, A, C) = 0$, is denoted as Σ and the solution for the optimistic system $\hat{\Theta}$ chosen from the set $(C_A \times C_B \times C_C) \cap \mathcal{S}$, $F(X, \tilde{A}, \tilde{C}) = 0$, is denoted as $\tilde{\Sigma}$. Denote $D_\Sigma = \tilde{\Sigma} - \Sigma$ and $M = A(I - LC)$. Recall that $L = \Sigma C^\top \left(C\Sigma C^\top + \sigma_z^2 I\right)^{-1}$. For any matrix X such that $I + (\Sigma + X)(\sigma_z^{-2} C^\top C)$ is invertible we have

$$F(\Sigma + X, A, C) = X - MXM^\top + MX(\sigma_z^{-2} C^\top C)[I + (\Sigma + X)(\sigma_z^{-2} C^\top C)]^{-1} XM^\top. \tag{C.13}$$

One can verify the identity by adding $F(\Sigma, A, C) = 0$ to the right hand side of (C.13) and use the identity that $M = A(I - LC) = A(I + \sigma_z^{-2}\Sigma C^\top C)^{-1} = A(I - \Sigma C^\top (C\Sigma C^\top + \sigma_z^2 I)^{-1} C)$. Define two operators $\mathcal{T}(X)$, $\mathcal{H}(X)$ such that $\mathcal{T}(X) = X - MXM^\top$ and $\mathcal{H}(X) = MX(\sigma_z^{-2} C^\top C)[I + (\Sigma + X)(\sigma_z^{-2} C^\top C)]^{-1} XM^\top$. Thus,

$$F(\Sigma + X, A, C) = \mathcal{T}(X) + \mathcal{H}(X).$$

Notice that since (C.13) is satisfied for any X such that $I + (\Sigma + X)(\sigma_z^{-2} C^\top C)$ is invertible,

$$F(\Sigma + X, A, C) - F(\Sigma + X, \tilde{A}, \tilde{C}) = \mathcal{T}(X) + \mathcal{H}(X) \tag{C.14}$$

has a unique solution $X = D_\Sigma$ where $\Sigma + D_\Sigma \succeq 0$.

Recall that M is stable. Therefore, the linear map $\mathcal{T} : X \mapsto X - MXM^\top$ has non-zero eigenvalues, *i.e* $\mathcal{T}$ is invertible. Using this, define the following operator,

$$\Psi(X) = \mathcal{T}^{-1}\left(F(\Sigma + X, A, C) - F(\Sigma + X, \tilde{A}, \tilde{C}) - \mathcal{H}(X)\right).$$

Notice that solving for X in (C.14) is equivalent to solving for X that satisfies $\Sigma + X \succeq 0$ and $\Psi(X) = X$. This shows that $\Psi(X)$ has a unique fixed point X that is D_Σ. Consider the set

$$\mathcal{S}_{\Sigma,\beta} = \{X : \|X\| \leq \beta, X = X^\top, \Sigma + X \succeq 0\}. \tag{C.15}$$

Let $X \in \mathcal{S}_{\Sigma,\beta}$ for $\beta < \sigma_n(\Sigma)/2$. First of all, recalling Assumption 5.1, notice that operator norm of $\mathcal{T}^{-1}$ is upper bounded as $\|\mathcal{T}^{-1}\| \leq \frac{1}{1-\upsilon^2}$. We also have $\|\mathcal{H}(X)\| \leq$

$\sigma_z^{-2}\upsilon^2\|X\|^2\|C\|^2 \leq \sigma_z^{-2}\upsilon^2\beta^2\|C\|^2$. Now consider $F(\Sigma+X, A, C) - F(\Sigma+X, \tilde{A}, \tilde{C})$:

$$F(\Sigma + X, \tilde{A}, \tilde{C}) - F(\Sigma + X, A, C) \tag{C.16}$$
$$= A(I + \sigma_z^{-2}(\Sigma + X)C^\top C)^{-1}(\Sigma + X)A^\top - \tilde{A}(I + \sigma_z^{-2}(\Sigma + X)\tilde{C}^\top\tilde{C})^{-1}(\Sigma + X)\tilde{A}^\top$$
$$= A(I + \sigma_z^{-2}(\Sigma + X)\tilde{C}^\top\tilde{C})^{-1}(\Sigma + X)\sigma_z^{-2}(C^\top C - \tilde{C}^\top\tilde{C})(I+\sigma_z^{-2}(\Sigma+X)C^\top C)^{-1}(\Sigma+X)A^\top$$
$$- (\tilde{A}-A)(I+\sigma_z^{-2}(\Sigma+X)\tilde{C}^\top\tilde{C})^{-1}(\Sigma+X)A^\top - A(I + \sigma_z^{-2}(\Sigma+X)\tilde{C}^\top\tilde{C})^{-1}(\Sigma+X)(\tilde{A}-A)^\top$$
$$- (\tilde{A} - A)(I + \sigma_z^{-2}(\Sigma + X)\tilde{C}^\top\tilde{C})^{-1}(\Sigma + X)(\tilde{A} - A)^\top \tag{C.17}$$

Note that for two PSD matrices of the same dimension M and N, we have $\|N(I + MN)^{-1}\| \leq \|N\|$. Using this result and the fact that $X \in \mathcal{S}_{\Sigma,\beta}$,

$$\|F(\Sigma+X, \tilde{A}, \tilde{C}) - F(\Sigma+X, A, C)\| \tag{C.18}$$
$$\leq \sigma_z^{-2}\kappa_1^2\|\Sigma + X\|^2\|C^\top C - \tilde{C}^\top\tilde{C}\| + 2\kappa_1\|\Sigma+X\|\|\tilde{A} - A\| + \|\Sigma + X\|\|\tilde{A} - A\|^2$$
$$\leq \sigma_z^{-2}\kappa_1^2(\beta+\|\Sigma\|)^2(2\|C\|\|\tilde{C}-C\| + \|\tilde{C}-C\|^2) + (\beta+\|\Sigma\|)(2\kappa_1\|\tilde{A}-A\| + \|\tilde{A}-A\|^2) \tag{C.19}$$

This gives us the following,

$$\|\Psi(X)\| \leq \frac{\sigma_z^{-2}\kappa_1^2\left(\beta + \|\Sigma\|\right)^2\left(2\|C\|\|\tilde{C} - C\| + \|\tilde{C}-C\|^2\right)}{1 - \upsilon^2} + \frac{\left(\beta + \|\Sigma\|\right)\left(2\kappa_1\|\tilde{A} - A\| + \|\tilde{A}-A\|^2\right) + \sigma_z^{-2}\upsilon^2\beta^2\|C\|^2}{1 - \upsilon^2}$$

Again using the fact that $\|N(I + MN)^{-1}\| \leq \|N\|$ for two PSD matrices and the definition of $\mathcal{H}(X)$, for $X_1, X_2 \in \mathcal{S}_{\Sigma,\beta}$

$$\|\mathcal{H}(X_1) - \mathcal{H}(X_2)\| \leq \upsilon^2\left((\sigma_z^{-2}\|C\|^2\beta)^2 + 2(\sigma_z^{-2}\|C\|^2\beta)\right)\|X_1 - X_2\|$$

Next, we bound

$$\|\mathcal{D}(X_1, X_2)\| = \|F(\Sigma+X_1, \tilde{A}, \tilde{C}) - F(\Sigma+X_1, A, C) - F(\Sigma+X_2, \tilde{A}, \tilde{C}) + F(\Sigma+X_2, A, C)\|.$$

Notice that we have $\|(I + \sigma_z^{-2}(\Sigma + X)\tilde{C}^\top\tilde{C})^{-1}\|, \|(I + \sigma_z^{-2}(\Sigma + X)C^\top C)^{-1}\| \leq \frac{2(\|\Sigma\|+\beta)}{\sigma_n(\Sigma)}$ from the choice of β. Let $V_1 = (I + \sigma_z^{-2}(\Sigma + X_1)C^\top C)^{-1}(\Sigma + X_1)$ and $\tilde{V}_1 = (I + \sigma_z^{-2}(\Sigma + X_1)\tilde{C}^\top\tilde{C})^{-1}(\Sigma + X_1)$. Define similarly V_2 and $\tilde{V}_2$. Note that

$\|V_1\|, \|V_2\|, \|\tilde{V}_1\|, \|\tilde{V}_2\| \leq \|\Sigma\| + \beta$. Using these, we bound $\|\mathcal{D}(X_1, X_2)\|$ as follows

$$
\begin{aligned}
&\left\|\mathcal{D}(X_1, X_2)\right\| \\
&= \Big\|A\tilde{V}_1\sigma_z^{-2}(C^\top C - \tilde{C}^\top\tilde{C})V_1A^\top - A\tilde{V}_2\sigma_z^{-2}(C^\top C - \tilde{C}^\top\tilde{C})V_2A^\top - (\tilde{A}-A)\tilde{V}_1A^\top + (\tilde{A}-A)\tilde{V}_2A^\top \\
&\qquad - A\tilde{V}_1(\tilde{A}-A)^\top + A\tilde{V}_2(\tilde{A}-A)^\top - (\tilde{A}-A)\tilde{V}_1(\tilde{A}-A)^\top + (\tilde{A}-A)\tilde{V}_2(\tilde{A}-A)^\top\Big\| \\
&\leq \kappa_1^2\|(\tilde{V}_1 - \tilde{V}_2)\sigma_z^{-2}(C^\top C - \tilde{C}^\top\tilde{C})V_1\| + \kappa_1^2\|\tilde{V}_2\sigma_z^{-2}(C^\top C - \tilde{C}^\top\tilde{C})(V_1 - V_2)\| \\
&\qquad + \|\tilde{V}_1 - \tilde{V}_2\|\left(2\kappa_1\|\tilde{A} - A\| + \|\tilde{A} - A\|^2\right) \\
&\leq \sigma_z^{-2}\kappa_1^2(2\|C\|\|\tilde{C}-C\| + \|\tilde{C}-C\|^2)\left(\|\tilde{V}_1 - \tilde{V}_2\|\|V_1\| + \|\tilde{V}_2\|\|V_1 - V_2\|\right) \\
&\qquad + \|\tilde{V}_1 - \tilde{V}_2\|\left(2\kappa_1\|\tilde{A} - A\| + \|\tilde{A} - A\|^2\right) \qquad \text{(C.20)}
\end{aligned}
$$

We need to consider $\|\tilde{V}_1 - \tilde{V}_2\|$ and $\|V_1 - V_2\|$:

$$
\begin{aligned}
\|\tilde{V}_1 - \tilde{V}_2\| &\leq \|(I + \sigma_z^{-2}(\Sigma + X_1)\tilde{C}^\top\tilde{C})^{-1}(X_1 - X_2)\| \\
&\qquad + \left\|\left((I + \sigma_z^{-2}(\Sigma + X_1)\tilde{C}^\top\tilde{C})^{-1} - (I + \sigma_z^{-2}(\Sigma + X_2)\tilde{C}^\top\tilde{C})^{-1}\right)(\Sigma + X_2)\right\| \\
&\leq \|X_1 - X_2\|\frac{2(\|\Sigma\| + \beta)}{\sigma_n(\Sigma)} + \sigma_z^{-2}\frac{4(\|\Sigma\| + \beta)^3}{\sigma_n^2(\Sigma)}(\|C\| + \|\tilde{C} - C\|)^2\|X_1 - X_2\| \\
\|V_1 - V_2\| &\leq \|X_1 - X_2\|\frac{2(\|\Sigma\| + \beta)}{\sigma_n(\Sigma)} + \sigma_z^{-2}\frac{4(\|\Sigma\| + \beta)^3}{\sigma_n^2(\Sigma)}\|C\|^2\|X_1 - X_2\|
\end{aligned}
$$

Combining these with (C.20), we get

$$
\begin{aligned}
&\left\|\mathcal{D}(X_1, X_2)\right\| \\
&\leq \left[(2\|C\|\|\tilde{C}-C\| + \|\tilde{C}-C\|^2)\kappa_1^2\left(\frac{4\sigma_z^{-2}(\|\Sigma\|+\beta)^2}{\sigma_n(\Sigma)} + \frac{8\sigma_z^{-4}(\|\Sigma\|+\beta)^4}{\sigma_n^2(\Sigma)}((\|C\| + \|\tilde{C} - C\|)^2 + \|C\|^2)\right)\right. \\
&\left. + \left(2\kappa_1\|\tilde{A} - A\| + \|\tilde{A} - A\|^2\right)\left(\frac{2(\|\Sigma\| + \beta)}{\sigma_n(\Sigma)} + \frac{4\sigma_z^{-2}(\|\Sigma\| + \beta)^3}{\sigma_n^2(\Sigma)}(\|C\| + \|\tilde{C} - C\|)^2\right)\right]\|X_1 - X_2\|
\end{aligned}
$$

Therefore we have the following inequality for $\Psi(X_1) - \Psi(X_2)$:

$$
\begin{aligned}
&\|\Psi(X_1) - \Psi(X_2)\| \\
&\leq \left[(2\|C\|\|\tilde{C}-C\| + \|\tilde{C}-C\|^2)\kappa_1^2\left(\frac{4\sigma_z^{-2}(\|\Sigma\|+\beta)^2}{\sigma_n(\Sigma)} + \frac{8\sigma_z^{-4}(\|\Sigma\|+\beta)^4}{\sigma_n^2(\Sigma)}((\|C\| + \|\tilde{C} - C\|)^2 + \|C\|^2)\right)\right. \\
&+ \left(2\kappa_1\|\tilde{A} - A\| + \|\tilde{A} - A\|^2\right)\left(\frac{2(\|\Sigma\| + \beta)}{\sigma_n(\Sigma)} + \sigma_z^{-2}\frac{4(\|\Sigma\| + \beta)^3}{\sigma_n^2(\Sigma)}(\|C\| + \|\tilde{C} - C\|)^2\right) \\
&\left. + \upsilon^2\left((\sigma_z^{-2}\|C\|^2\beta)^2 + 2(\sigma_z^{-2}\|C\|^2\beta)\right)\right]\frac{\|X_1 - X_2\|}{1 - \upsilon^2} \qquad \text{(C.21)}
\end{aligned}
$$

Denote ϵ such that $\epsilon := \max\{\|C - \tilde{C}\|, \|A - \tilde{A}\|\}$. The choice of T_w (due to T_A, T_B) guarantees that $\epsilon < 1$. In order to show that D_Σ is the unique fixed point of Ψ in

$\mathcal{S}_{\Sigma,\beta}$, one needs to show that Ψ maps $\mathcal{S}_{\Sigma,\beta}$ to itself and it's contraction. To this end, we need to have ϵ and β that gives $\|\Psi(X)\| \leq \beta$ and $\|\Psi(X_1)-\Psi(X_2)\| < \|X_1 - X_2\|$. Let $\beta = 2k^*\epsilon < \frac{\sigma_n(\Sigma)}{2}$ where

$$k^* = \frac{\sigma_z^{-2}\kappa_1^2(2\|C\|+1)\|\Sigma\|^2 + (2\kappa_1+1)\|\Sigma\|}{1-\upsilon^2}$$

One can verify that this gives $\|\Psi(X)\| \leq \beta$. In order to get contraction, the coefficient of $\|X_1 - X_2\|$ in (C.21) must be less than 1. This requires

$$\epsilon < \frac{1-\upsilon^2}{(2\|C\|+1)\kappa_1^2 c_1 + (2\kappa_1+1)c_2 + 6k^* c_3},$$

for

$$c_1 = \left(\frac{4\sigma_z^{-2}(\|\Sigma\|+\sigma_n(\Sigma)/2)^2}{\sigma_n(\Sigma)} + \frac{8\sigma_z^{-4}(\|\Sigma\|+\sigma_n(\Sigma)/2)^4}{\sigma_n^2(\Sigma)}(2\|C\|^2 + 2\|C\| + 1)\right)$$

$$c_2 = \left(\frac{2(\|\Sigma\| + \sigma_n(\Sigma)/2)}{\sigma_n(\Sigma)} + \frac{4\sigma_z^{-2}(\|\Sigma\| + \sigma_n(\Sigma)/2)^3}{\sigma_n^2(\Sigma)}(\|C\|^2 + 2\|C\| + 1)\right)$$

$$c_3 = \upsilon^2\sigma_z^{-2}\|C\|^2.$$

From the choice of T_w (Due to T_L), ϵ satisfies the stated bound. Thus, Ψ has a unique fixed point in $\mathcal{S}_{\Sigma,2k\epsilon}$, *i.e.* $\|\tilde{\Sigma} - \Sigma\| \leq 2k^* \max\{\|C - \tilde{C}\|, \|A - \tilde{A}\|\}$. Bringing back the similarity transformations, this gives us the following bound

$$\begin{aligned}\|\tilde{\Sigma} - \mathbf{S}^{-1}\Sigma\mathbf{S}\| &\leq 2k^* \max\{\|\tilde{C} - C\mathbf{S}\|, \|\tilde{A} - \mathbf{S}^{-1}A\mathbf{S}\|\} \\ &\leq 4k^* \max\{\beta_A, \beta_C\} \coloneqq \Delta\Sigma\end{aligned}$$

since $\|\tilde{A} - \mathbf{S}^{-1}A\mathbf{S}\| \leq 2\beta_A$ and $\|\tilde{C} - C\mathbf{S}\| \leq 2\beta_C$. □

C.2.3 Proof of Lemma 5.9

From Lemma 5.3, we have the following with probability $1-\delta/2$, for all $1 \leq t \leq T_w$,

$$\|x_t\| \leq X_{exp} \coloneqq \frac{(\sigma_w + \sigma_u\|B\|)\kappa_1(1-\gamma_1)}{\sqrt{1-(1-\gamma_1)^2}}\sqrt{2n\log(12nT_w/\delta)}, \tag{C.22}$$

$$\|z_t\| \leq Z \coloneqq \sigma_z\sqrt{2m\log(12mT_w/\delta)}, \tag{C.23}$$

$$\|u_t\| \leq U_{exp} \coloneqq \sigma_u\sqrt{2p\log(12pT_w/\delta)}. \tag{C.24}$$

Let $\Omega = 2(\|C^\top QC\|X_{exp}^2 + \|Q\|Z^2 + \|R\|U_{exp}^2)$. Define $\mathcal{X}_t = x_t^\top C^\top QCx_t + z_t^\top Qz_t + u_t^\top Ru_t - \mathbb{E}[x_t^\top C^\top QCx_t + z_t^\top Qz_t + u_t^\top Ru_t]$ and its truncated version $\tilde{\mathcal{X}}_t = \mathbb{1}_{\mathcal{X}_t \leq \Omega}\mathcal{X}_t$. Define $S = \sum_{t=1}^{T_w}\mathcal{X}_t$ and $\tilde{S} = \sum_{t=1}^{T_w}\tilde{\mathcal{X}}_t$. By Lemma I.4 of [161],

$$\mathbb{P}\left(S > \Omega\sqrt{2T_w\log\frac{2}{\delta}}\right) \leq \mathbb{P}\left(\max_{1\leq t\leq T_w}\mathcal{X}_t \geq \Omega\right) + \mathbb{P}\left(\tilde{S} > \Omega\sqrt{2T_w\log\frac{2}{\delta}}\right). \tag{C.25}$$

From (C.22)-(C.24) and Azuma Inequality, each term on the right-hand side is bounded by $\delta/2$. Thus, with probability $1-\delta$,

$$\sum_{t=1}^{T_w} y_t^\top Q y_t + u_t^\top R u_t - \mathbb{E}[y_t^\top Q y_t + u_t^\top R u_t] \leq \Omega\sqrt{2T_w \log\frac{2}{\delta}} \tag{C.26}$$

$$\sum_{t=1}^{T_w} y_t^\top Q y_t + u_t^\top R u_t \leq T_w \left((\sigma_w^2 + \sigma_u^2\|B\|^2)\frac{\kappa_1^2(1-\gamma_1)^2}{1-(1-\gamma_1)^2}\operatorname{Tr}(C^\top Q C) + \sigma_u^2\operatorname{Tr}(R) + \sigma_z^2\operatorname{Tr}(Q)\right) \tag{C.27}$$

$$+ 2\left(\|C^\top Q C\|X_{exp}^2 + \|Q\|Z^2 + \|R\|U_{exp}^2\right)\sqrt{2T_w\log\frac{2}{\delta}} \tag{C.28}$$

Recall that cost obtained in T_w by the optimal controller of Θ is

$$T_w\left(\operatorname{Tr}(C^\top Q C\bar{\Sigma}) + \operatorname{Tr}(P(\Sigma-\bar{\Sigma})) + \sigma_z^2\operatorname{Tr}(Q)\right).$$

Thus the regret obtained from T_w length exploration is upper bounded as described in the statement of lemma.

C.3 Proofs of Section 5.6

C.3.1 Proof of Lemma 5.15

Proof. Let $B'_{\pi,w} := [I_{n\times n}^\top \;\; 0_{s\times n}^\top]^\top$, and $B'_{\pi,z} := [D_\pi^\top B^\top \;\; B_\pi^\top]^\top$, the columns of B'_π applied on process noise, and measurement noise respectively. Similarly $C'_{\pi,y} := \begin{bmatrix} C & 0_{s\times d}\end{bmatrix}$ and $C'_{\pi,u} := \begin{bmatrix} D_\pi C & C_\pi\end{bmatrix}$ are rows of C'_π generating the observation and action.

Rolling out the dynamical system defining a policy π in (5.82), we can restate the action u_t^π as follows,

$$\begin{aligned}
u_t^\pi &= D_\pi z_t + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} z_{t-i} + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,w} w_{t-i} \\
&= D_\pi z_t + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} z_{t-i} + C'_{\pi,u} B'_{\pi,w} w_{t-1} + \sum_{i=2}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,w} w_{t-i} \\
&= D_\pi z_t + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} z_{t-i} + D_\pi C w_{t-1} + \sum_{i=2}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,w} w_{t-i}
\end{aligned}$$

Note that $A'_\pi B'_{\pi,w}$ is equal to $\begin{bmatrix} A + BD_\pi C \\ B_\pi C \end{bmatrix}$. Based on the definition of A'_π in (5.82), we restate A'_π as follows,

$$A'_\pi = \begin{bmatrix} A + BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix} = \begin{bmatrix} BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix} + \begin{bmatrix} A & 0_{n\times s} \\ 0_{s\times n} & 0_{s\times s} \end{bmatrix}$$

For any given bounded matrices A'_π and A, and any integer $i > 0$, we have

$$\begin{aligned}
{A'_\pi}^i &= \begin{bmatrix} A + BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix}^i \\
&= \begin{bmatrix} A + BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix}^{i-1} \begin{bmatrix} BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix} + \begin{bmatrix} A + BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix}^{i-1} \begin{bmatrix} A & 0_{n\times s} \\ 0_{s\times n} & 0_{s\times s} \end{bmatrix} \\
&= \begin{bmatrix} A + BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix}^{i-1} \begin{bmatrix} BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix} \\
&\quad + \begin{bmatrix} A + BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix}^{i-2} \begin{bmatrix} BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix} \begin{bmatrix} A & 0_{n\times s} \\ 0_{s\times n} & 0_{s\times s} \end{bmatrix} \\
&\quad + \begin{bmatrix} A + BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix}^{i-2} \begin{bmatrix} A^2 & 0_{n\times s} \\ 0_{s\times n} & 0_{s\times s} \end{bmatrix} \\
&\ \ \vdots \\
&= \begin{bmatrix} A^i & 0_{n\times s} \\ 0_{s\times n} & 0_{s\times s} \end{bmatrix} + \sum_{j=1}^{i} {A'_\pi}^{j-1} \begin{bmatrix} BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix} \begin{bmatrix} A & 0_{n\times s} \\ 0_{s\times n} & 0_{s\times s} \end{bmatrix}^{i-j}
\end{aligned}$$

We use this decomposition to relate u_t^π and $u_t^{\mathbf{M}}$. Now considering ${A'_\pi}^{i-1} B'_{\pi,w}$, for $i - 1 > 0$ we have

$${A'_\pi}^{i-1} B'_{\pi,w} = \begin{bmatrix} A^{i-1} \\ 0_{s\times n} \end{bmatrix} + \sum_{j=1}^{i-1} {A'_\pi}^{j-1} \begin{bmatrix} BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix} \begin{bmatrix} A^{i-1-j} \\ 0_{s\times n} \end{bmatrix} = \begin{bmatrix} A^{i-1} \\ 0_{s\times n} \end{bmatrix} + \sum_{j=1}^{i-1} {A'_\pi}^{j-1} B'_{\pi,z} C A^{i-1-j}$$

Using this equality in the derivation of u_t^π we derive,

$$u_t^\pi = D_\pi z_t + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} z_{t-i} + D_\pi C w_{t-1}$$
$$+ \sum_{i=2}^{t-1} \begin{bmatrix} D_\pi C & C_\pi \end{bmatrix} \begin{bmatrix} A^{i-1} \\ 0_{s\times n} \end{bmatrix} w_{t-i} + \sum_{i=2}^{t-1} C'_{\pi,u} \left(\sum_{j=1}^{i-1} {A'_\pi}^{j-1} B'_{\pi,z} C A^{i-1-j} \right) w_{t-i}$$
$$= D_\pi z_t + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} z_{t-i} + \sum_{i=1}^{t-1} D_\pi C A^{i-1} w_{t-i} + \sum_{i=2}^{t-1} \sum_{j=1}^{i-1} C'_{\pi,u} {A'_\pi}^{j-1} B'_{\pi,z} C A^{i-1-j} w_{t-i}$$

Note that $b_t(\mathbf{G}) = z_t + \sum_{i=1}^{t-1} CA^{t-i-1} w_i = z_t + \sum_{i=1}^{t-1} CA^{i-1} w_{t-i}$. Inspired by this expression, we rearrange the previous sum as follows:

$$u_t^\pi = D_\pi \left(z_t + \sum_{i=1}^{t-1} CA^{i-1} w_{t-i} \right) + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} z_{t-i} + \sum_{i=2}^{t-1} \sum_{j=1}^{i-1} C'_{\pi,u} {A'_\pi}^{j-1} B'_{\pi,z} C A^{i-1-j} w_{t-i}$$
$$= D_\pi \left(z_t + \sum_{i=1}^{t-1} CA^{i-1} w_{t-i} \right) + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} z_{t-i} + \sum_{j=1}^{t-2} \sum_{i=j+1}^{t-1} C'_{\pi,u} {A'_\pi}^{j-1} B'_{\pi,z} C A^{i-1-j} w_{t-i}$$
$$= D_\pi \left(z_t + \sum_{i=1}^{t-1} CA^{i-1} w_{t-i} \right) + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} z_{t-i} + \sum_{j=1}^{t-2} C'_{\pi,u} {A'_\pi}^{j-1} B'_{\pi,z} \sum_{i=1}^{t-j-1} CA^{t-j-i-1} w_i$$
$$= D_\pi b_t + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} b_{t-i}$$

Now setting $M^{[0]} = D_\pi$, and $M^{[i]} = C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z}$ for all $0 < i < H'$, we conclude that for any LDC policy $\pi \in \Pi$, there exists at least one length H' DFC policy $\mathbf{M}(H')$ such that

$$u_t^\pi - u_t^{\mathbf{M}} = \sum_{i=H'}^{t} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} b_{t-i}$$

Using Cauchy Schwarz inequality we have

$$\|u_t^\pi - u_t^{\mathbf{M}}\| \leq \left\| \sum_{i=H'}^{t} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} b_{t-i} \right\| \leq \psi(H') \kappa_b$$

which states the first half of the Lemma.

Using the definition of y_t^π in (5.82), we have

$$y_t^\pi = z_t + \sum_{i=1}^{t-1} CA^{t-i-1} w_i + \sum_{i}^{t-1} G^{[i]} u_{t-i}^\pi.$$

Similarly for $y_t^{\mathbf{M}}$ we have,

$$y_t^{\mathbf{M}} = z_t + \sum_{i=1}^{t-1} CA^{t-i-1}w_i + \sum_{i}^{t-1} G^{[i]}u_{t-i}^{\mathbf{M}}.$$

Subtracting these two equations, we derive,

$$y_t^{\pi} - y_t^{\mathbf{M}} = \sum_{i}^{t-1} G^{[i]}u_{t-i}^{\pi} - \sum_{i}^{t-1} G^{[i]}u_{t-i}^{\mathbf{M}} = \sum_{i}^{t-1} G^{[i]}(u_{t-i}^{\pi} - u_{t-i}^{\mathbf{M}})$$

resulting in

$$\|y_t^{\pi} - y_t^{\mathbf{M}}\| \leq \psi(H')\kappa_{\mathbf{G}}\kappa_b$$

which states the second half of the Lemma. □

C.3.2 Bound on the Markov Parameters Estimation Errors

Finally, we will consider the Markov parameter estimates that is constructed by using the parameter estimates. From Theorem 5.4, for some unitary matrix $\mathbf{T}$, we denote $\Delta A := \|\widehat{A}_t - \mathbf{T}^{\top}A\mathbf{T}\|$, $\Delta B := \|\widehat{B}_t - \mathbf{T}^{\top}B\| = \|\widehat{C}_t - C\mathbf{T}\|$. Let $T_A = T_{\mathcal{G}_{\mathbf{yu}}}\frac{4c_1^2\left(\frac{\sqrt{nH}(\|\mathcal{H}\|+\sigma_n(\mathcal{H}))}{\sigma_n^2(\mathcal{H})}\right)^2}{(1-(1-\gamma_1))^2}$. For $t > \max\{T_A, T_B\}$, $\Delta A \leq \frac{1-(1-\gamma_1)}{2}$ and $\Delta B \leq 1$. Using this fact, we have

$$\begin{aligned}
&\sum_{j\geq 1}^{H} \|\widehat{C}_t\widehat{A}_t^{j-1}\widehat{B}_t - CA^{j-1}B\| \\
&\leq \Delta B(\|B\|+\|C\|+1)+\sum_{i=1}^{H-1}\kappa_1\rho^i(A)\Delta B(\|B\|+\|C\|+1) + \|\widehat{A}_t^i-\mathbf{T}^{\top}A^i\mathbf{T}\|(\|C\|\|B\|+\|B\|+\|C\|+1) \\
&\leq \left(1+\frac{\kappa_1}{1-(1-\gamma_1)}\right)\Delta B(\|B\|+\|C\|+1) + \Delta A(\|C\|\|B\|+\|B\|+\|C\|+1)\sum_{i=1}^{H-1}\sum_{j=0}^{i-1}\binom{i}{j}\|A^j\|(\Delta A)^{i-1-j} \\
&\leq \left(1+\frac{\kappa_1}{1-(1-\gamma_1)}\right)\Delta B(\|B\|+\|C\|+1) \\
&\qquad + \Delta A\kappa_1(\|C\|\|B\|+\|B\|+\|C\|+1)\sum_{i=1}^{H-1}\sum_{j=0}^{i-1}\binom{i}{j}\rho^j(A)\left(\frac{1-(1-\gamma_1)}{2}\right)^{i-1-j} \\
&\leq \left(1+\frac{\kappa_1}{1-(1-\gamma_1)}\right)\Delta B(\|B\|+\|C\|+1) + \frac{2\Delta A\kappa_1}{1-(1-\gamma_1)}(\|C\|\|B\|+\|B\|+\|C\|+1)\sum_{i=1}^{H-1}\left[\left(\frac{1+\rho}{2}\right)^i - \rho^i\right] \\
&\leq \Delta B\left(1+\frac{\kappa_1}{1-(1-\gamma_1)}\right)(\|B\|+\|C\|+1) + \frac{2\Delta A\kappa_1}{(1-(1-\gamma_1))^2}(\|C\|\|B\|+\|B\|+\|C\|+1)
\end{aligned}$$

$$\gamma_{\mathbf{G}} = (\|B\|+\|C\|+1)\left(1+\frac{\kappa_1}{1-(1-\gamma_1)}+\frac{2\kappa_1}{(1-(1-\gamma_1))^2}\right)+\frac{2\kappa_1}{(1-(1-\gamma_1))^2}\|C\|\|B\|$$

Assuming that $\|F\|+\|C\| > 1$ for simplicity, from the exact expressions of Theorem 5.4, we have $\Delta A > \Delta B$. For the given $\gamma_{\mathbf{G}}$ and $\gamma_{\mathcal{H}}$, we can upper bound the last expression above as follow,

$$\sum_{j\geq 1}^{H} \|\widehat{C}_t \widehat{A}_t^{j-1} \widehat{B}_t - CA^{j-1}B\| \leq \gamma_{\mathbf{G}} \Delta A \leq \frac{c_1 \gamma_{\mathbf{G}} \gamma_{\mathcal{H}} \kappa_e}{\sigma_\star \sqrt{t}}, \tag{C.29}$$

for

$$\gamma_{\mathbf{G}} \coloneqq (\|B\| + \|C\| + 1)\left(1+\frac{\kappa_1}{1-(1-\gamma_1)}+\frac{2\kappa_1}{(1-(1-\gamma_1))^2}\right) + \frac{2\kappa_1}{(1-(1-\gamma_1))^2}\|C\|\|B\|, \tag{C.30}$$

$$\kappa_e \coloneqq \sqrt{m\|C\Sigma C^\top + \sigma_z^2 I\|\left(\log(1/\delta) + \frac{H(m+p)}{2}\log\left(\frac{\lambda(m+p)+T\Upsilon^2}{\lambda(m+p)}\right)\right)} + S\sqrt{\lambda} + \frac{\sqrt{H}}{T}, \tag{C.31}$$

$$\gamma_{\mathcal{H}} \coloneqq \frac{\sqrt{nH}(\|\mathcal{H}\| + \sigma_n(\mathcal{H}))}{\sigma_n^2(\mathcal{H})}. \tag{C.32}$$

The proof of Theorem 5.2 is completed by noticing that $\|\widehat{\mathbf{G}}(H) - \mathbf{G}(H)\| = \|[\widehat{G}^{[1]}\ \widehat{G}^{[2]}\ \dots\ \widehat{G}^{[H]}] - [G^{[1]}\ G^{[2]}\ \dots\ G^{[H]}] \leq \sqrt{\sum_{i=1}^{H}\|\widehat{G}^{[i]} - G^{[i]}\|^2}$.

C.3.3 Boundedness Lemmas

Lemma C.3.1 (Bounded Nature's y)**.** . *For all $t \in [T]$, the following holds with probability at least $1-\delta$,*

$$\|b_t(\mathbf{G})\| \leq \kappa_b \coloneqq \overline{\sigma}_z\sqrt{2m\log\frac{6mT}{\delta}} + \|C\|\kappa_1\sqrt{2n}\left((1-\gamma_1)^t\sqrt{\|\Sigma\|\log\frac{6n}{\delta}} + \frac{\overline{\sigma}_w\sqrt{\log\frac{4nT}{\delta}}}{\gamma_1}\right).$$

Proof. Using standard sub-Gaussian tail bound, the following hold for all $t \in [T]$, with probability at least $1-\delta$,

$$\|w_t\| \leq \overline{\sigma}_w\sqrt{2n\log\frac{6nT}{\delta}}, \qquad \|z_t\| \leq \overline{\sigma}_z\sqrt{2m\log\frac{6mT}{\delta}}, \qquad \|x_0\| \leq \sqrt{2n\|\Sigma\|\log\frac{6n}{\delta}}. \tag{C.33}$$

Thus we have,

$$\|b_t(\mathbf{G})\| = \|z_t + CA^t x_0 + \sum_{i=0}^{t-1} CA^{t-i-1} w_i\| \le \|z_t\| + \|C\| \left(\|A^t\| \|x_0\| + \left(\max_{1 \le t \le T} \|w_t\| \right) \sum_{i=0}^{\infty} \|A^i\| \right)$$

$$\le \overline{\sigma}_z \sqrt{2m \log \frac{6mT}{\delta}} + \|C\| \kappa_1 \sqrt{2n} \left((1-\gamma_1)^t \sqrt{\|\Sigma\| \log \frac{6n}{\delta}} + \frac{\overline{\sigma}_w \sqrt{\log \frac{4nT}{\delta}}}{\gamma_1} \right). \tag{C.34}$$

□

Lemma C.3.2 (Boundedness Lemma)**.** *Let $\delta \in (0,1)$, $T > T_w \ge T_{\max}$ and $\psi_{\mathbf{G}}(H+1) \le 1/10T$. For* AdaptOn*, we have the boundedness of the following with probability at least $1-2\delta$:*
Nature's y *: $\|b_t(\mathbf{G})\| \le \kappa_b$, $\forall t$,*
Inputs: *$\|u_t\| \le \kappa_u := 2\max\{\kappa_{u_b}, \kappa_{\mathcal{M}}\kappa_b\}, \forall t$,* ***Outputs:*** *$\|y_t\| \le \kappa_y := \kappa_b + \kappa_{\mathbf{G}}\kappa_u$, $\forall t$, and*
Nature's y estimates: *$\|b_t(\widehat{\mathbf{G}})\| \le 2\kappa_b$, for all $t > T_w$.*

Proof of this lemma follows similarly from the proof of Lemma 6.1 in Simchowitz et al. [245].

Additional Bound on the Markov Parameter Estimates

Define α, such that $\alpha \le \underline{\alpha}_{loss}\left(\sigma_z^2 + \sigma_w^2 \left(\frac{\sigma_{\min}(C)}{1+\|A\|^2}\right)^2\right)$, where right-hand side is the effective strong convexity parameter. Define $T_{cx} := T_{\mathcal{G}_{\mathbf{yu}}} \frac{16 c_1^2 \kappa_b^2 \kappa_{\mathcal{M}}^2 \kappa_{\mathbf{G}}^2 H' \gamma_{\mathbf{G}}^2 \gamma_{\mathcal{H}}^2 \underline{\alpha}_{loss}}{\alpha}$, $T_{\epsilon_{\mathbf{G}}} := 4c_1^2 \kappa_{\mathcal{M}}^2 \kappa_{\mathbf{G}}^2 \gamma_{\mathbf{G}}^2 \gamma_{\mathcal{H}}^2 T_{\mathcal{G}_{\mathbf{yu}}}$ and $T_r = c_1^2 \gamma_{\mathbf{G}}^2 \gamma_{\mathcal{H}}^2 \kappa_\psi^2 T_{\mathcal{G}_{\mathbf{yu}}}/r^2$.

Lemma C.3.3 (Additional Boundedness of Markov Parameter Estimation Error)**.** *Let $T_w > T_{\max}$, i.e. $T_w > \max\{T_{cx}, T_{\epsilon_{\mathbf{G}}}, T_r\}$ and $\psi_{\mathbf{G}}(H+1) \le 1/10T$. Then*

$$\|\sum_{j \ge 1} \widehat{G}_i^{[j]} - G^{[j]}\| \le \epsilon_{\mathbf{G}}(i,\delta) \le \min\left\{ \frac{1}{4\kappa_b \kappa_{\mathcal{M}} \kappa_{\mathbf{G}}} \sqrt{\frac{\alpha}{H' \underline{\alpha}_{loss}}}, \frac{1}{2\kappa_{\mathcal{M}} \kappa_{\mathbf{G}}}, \frac{r}{\kappa_\psi} \right\}$$

with probability at least $1-4\delta$, where $\epsilon_{\mathbf{G}}(i,\delta) = \frac{2c_1 \gamma_{\mathbf{G}} \gamma_{\mathcal{H}} \kappa_e}{\sigma_\star \sqrt{2^{i-1} T_w}}$.

Proof. At the beginning of epoch i, using persistence of excitation with high probability in (C.29), we get

$$\sum_{j \ge 1}^{H} \|\widehat{C}_i \widehat{A}_i^{j-1} \widehat{B}_i - CA^{j-1}B\| \le \epsilon_{\mathbf{G}}(i,\delta)/2 = \frac{c_1 \gamma_{\mathbf{G}} \gamma_{\mathcal{H}} \kappa_e}{\sigma_\star \sqrt{2^{i-1} T_w}}. \tag{C.35}$$

From the assumption that $\psi_{\mathbf{G}}(H+1) \leq 1/10T$, we have that $\sum_{j \geq H+1} \|\widehat{G}_1^{[j]} - G^{[j]}\| \leq \epsilon_{\mathbf{G}}(1, \delta)/2$. The second inequality follows from the choice of $T_{\epsilon_{\mathbf{G}}}$,$T_{cx}$ and T_r. □

C.3.4 Proofs for Regret Bound

In order to prove Theorem 5.10, we follow the proof steps of Theorem 5 of Simchowitz et al. [245]. The main difference is that AdaptOn updates the Markov parameter estimates in epochs throughout the adaptive control period which provides a decrease in the gradient error in each epoch. These updates allow AdaptOn to remove $O(\sqrt{T})$ term in the regret expression of Theorem 5.

Proof. Consider the hypothetical "true prediction" y's, y_t^{pred} and losses, $f_t^{pred}(M)$ defined in Definition 8.1 of Simchowitz et al. [245]. Up to truncation by H, they describe the true counterfactual output of the system for AdaptOn inputs during the adaptive control period and the corresponding counterfactual loss functions. Lemma C.3.7, shows that at all epoch i, at any time step $t \in [t_i, \ \ldots, \ t_{i+1} - 1]$, the gradient $f_t^{pred}(M)$ is close to the gradient of the loss function of AdaptOn:

$$\left\| \nabla f_t \left(\mathbf{M}, \widehat{\mathbf{G}}_i, b_1(\widehat{\mathbf{G}}_i), \ldots, b_t(\widehat{\mathbf{G}}_i) \right) - \nabla f_t^{\text{pred}} (\mathbf{M}) \right\|_{\text{F}} \leq C_{\text{approx}} \epsilon_{\mathbf{G}}(i, \delta), \qquad \text{(C.36)}$$

where $C_{\text{approx}} := \sqrt{H'} \kappa_{\mathbf{G}} \kappa_{\mathcal{M}} \kappa_b^2 \, (16\overline{\alpha}_{loss} + 24L)$. For a comparing controller $\mathbf{M}_{comp} \in \mathcal{M}(H', \kappa_{\mathcal{M}})$ and the competing set $\mathcal{M}_\psi(H'_0, \kappa_\psi)$, where $\kappa_{\mathcal{M}} = (1 + r)\kappa_\psi$ and

$H_0' = \lfloor \frac{H'}{2} \rfloor - H$, we have the following regret decomposition:

$$\begin{aligned}
\textsc{Regret}(T) \leq & \underbrace{\left(\sum_{t=1}^{T_w} \ell_t\left(y_t, u_t\right)\right)}_{\text{warm-up regret}} + \underbrace{\left(\sum_{t=T_w+1}^{T} \ell_t\left(y_t, u_t\right) - \sum_{t=T_w+1}^{T} F_t^{\text{pred}}\left[\mathbf{M}_{t:t-H}\right]\right)}_{\text{algorithm truncation error}} \\
& + \underbrace{\left(\sum_{t=T_w+1}^{T} F_t^{\text{pred}}\left[\mathbf{M}_{t:t-H}\right] - \sum_{t=T_w+1}^{T} f_t^{\text{pred}}\left(\mathbf{M}_{comp}\right)\right)}_{f^{\text{pred}} \text{ policy regret}} \\
& + \underbrace{\left(\sum_{t=T_w+1}^{T} f_t^{\text{pred}}\left(\mathbf{M}_{comp}\right) - \inf_{\mathbf{M}\in\mathcal{M}_\psi} \sum_{t=T_w+1}^{T} f_t\left(\mathbf{M}, \mathbf{G}, b_1(\mathbf{G}), \ldots, b_t(\mathbf{G})\right)\right)}_{\text{comparator approximation error}} \\
& + \underbrace{\left(\inf_{\mathbf{M}\in\mathcal{M}_\psi} \sum_{t=T_w+1}^{T} f_t\left(\mathbf{M}, \mathbf{G}, b_1(\mathbf{G}), \ldots, b_t(\mathbf{G})\right) - \inf_{\mathbf{M}\in\mathcal{M}_\psi} \sum_{t=T_w+1}^{T} \ell_t\left(y_t^{\mathbf{M}}, u_t^{\mathbf{M}}\right)\right)}_{\text{comparator truncation error}} \\
& + \underbrace{\left(\inf_{\mathbf{M}\in\mathcal{M}_\psi} \sum_{t=1}^{T} \ell_t\left(y_t^{\mathbf{M}}, u_t^{\mathbf{M}}\right) - \sum_{t=0}^{T} \ell(y^{\pi\star}, u^{\pi\star})\right)}_{\text{policy approximation error}} \qquad \text{(C.37)}
\end{aligned}$$

Notice that the last term is only required to extend the Theorem 5.10 to Corollary 5.10.1. The result of Theorem 5.10 does not require the last term. We will consider each term separately.

Warm-up Regret: From (5.81) and Lemma C.3.2, we get $\sum_{t=1}^{T_w} \ell_t\left(y_t, u_t\right) \leq T_w L \kappa_y^2$.

Algorithm Truncation Error: From (5.81), we get

$$\begin{aligned}
\sum_{t=T_w+1}^{T} \ell_t\left(y_t, u_t\right) - \sum_{t=T_w+1}^{T} F_t^{\text{pred}}\left[\mathbf{M}_{t:t-H}\right] &\leq \sum_{t=T_w+1}^{T} \left|\ell_t\left(y_t, u_t\right) - \ell_t\left(b_t(\mathbf{G}) + \sum_{i=1}^{H} G^{[i]} u_{t-i}, u_t\right)\right| \\
&\leq \sum_{t=T_w+1}^{T} L\kappa_y \left\|y_t - b_t(\mathbf{G}) + \sum_{i=1}^{H} G^{[i]} u_{t-i}\right\| \\
&\leq \sum_{t=T_w+1}^{T} L\kappa_y \left\|\sum_{i=H+1} G^{[i]} u_{t-i}\right\| \\
&\leq TL\kappa_y\kappa_u\psi_{\mathbf{G}}(H+1)
\end{aligned}$$

Since $\psi_{\mathbf{G}}(H+1) \leq 1/10T$, we get $\sum_{t=T_w+1}^{T} \ell_t\left(y_t, u_t\right) - \sum_{t=T_w+1}^{T} F_t^{\text{pred}}\left[\mathbf{M}_{t:t-H}\right] \leq L\kappa_y\kappa_u/10$.

Comparator Truncation Error: Similar to algorithm truncation error above,

$$\begin{aligned}\inf_{\mathbf{M}\in\mathcal{M}_\psi} \sum_{t=T_w+1}^{T} f_t\left(\mathbf{M}, \mathbf{G}, b_1(\mathbf{G}), \ldots, b_t(\mathbf{G})\right) - \inf_{\mathbf{M}\in\mathcal{M}_\psi} \sum_{t=T_w+1}^{T} \ell_t\left(y_t^{\mathbf{M}}, u_t^{\mathbf{M}}\right) &\leq TL\kappa_{\mathbf{G}}\kappa_{\mathcal{M}}^2\kappa_b^2\psi_{\mathbf{G}}(H+1) \\ &\leq L\kappa_{\mathbf{G}}\kappa_{\mathcal{M}}^2\kappa_b^2/10\end{aligned}$$

Policy Approximation Error: By the assumption that $M_\star$ lives in the given convex set $\mathcal{M}_\psi$ and (5.81), using Lemma 5.15, we get

$$\begin{aligned}\inf_{\mathbf{M}\in\mathcal{M}_\psi} \sum_{t=1}^{T} \ell_t\left(y_t^{\mathbf{M}}, u_t^{\mathbf{M}}\right) - \sum_{t=1}^{T} \ell_t(y_t^{\pi_\star}, u_t^{\pi_\star}) &\leq \sum_{t=1}^{T} \ell_t\left(y_t^{\mathbf{M}_\star}, u_t^{\mathbf{M}_\star}\right) - \ell_t(y_t^{\pi_\star}, u_t^{\pi_\star}) \\ &\leq TL\kappa_y\left(\psi(H_0')\kappa_b + \psi(H_0')\kappa_{\mathbf{G}}\kappa_b\right) \\ &\leq 2TL\kappa_y\kappa_{\mathbf{G}}\kappa_b\psi(H_0')\end{aligned}$$

Since $\psi(H_0') \leq \kappa_{\mathcal{M}}/T$, we get $\inf_{\mathbf{M}\in\mathcal{M}_0} \sum_{t=1}^{T} \ell_t\left(y_t^{\mathbf{M}}, u_t^{\mathbf{M}}\right) - \sum_{t=1}^{T} \ell_t(y_t^{\pi_\star}, u_t^{\pi_\star}) \leq 2L\kappa_{\mathcal{M}}\kappa_y\kappa_{\mathbf{G}}\kappa_b$.

$\mathbf{f}^{\text{pred}}$ Policy Regret : In order to utilize Theorem C.3.6, we need the strong convexity, Lipschitzness and smoothness properties stated in the theorem. Due to Lemma C.3.3, Lemmas C.3.8-C.3.10 provide those conditions. Combining these with (C.36), we obtain the following adaptation of Theorem C.3.6:

Lemma C.3.4. *For step size $\eta = \frac{12}{\alpha t}$, the following bound holds with probability $1-\delta$:*

$$\begin{aligned}&\mathbf{f}^{\text{pred}}\textit{ policy regret } + \frac{\alpha}{48}\sum_{t=T_w+1}^{T}\|\mathbf{M}_t - \mathbf{M}_{comp}\|_F^2 \\ &\lesssim \frac{L^2H'^3\min\{m,p\}\kappa_b^4\kappa_{\mathbf{G}}^4\kappa_{\mathcal{M}}^2}{\min\{\alpha, L\kappa_b^2\kappa_{\mathbf{G}}^2\}}\left(1+\frac{\overline{\alpha}_{loss}}{\min\{m,p\}L\kappa_{\mathcal{M}}}\right)\log\left(\frac{T}{\delta}\right) + \frac{1}{\alpha}\sum_{t=T_w+1}^{T} C_{approx}^2\epsilon_{\mathbf{G}}^2\left(\left\lceil\log_2\left(\frac{t}{T_w}\right)\right\rceil, \delta\right).\end{aligned}$$

Proof. Let $d = \min\{m,p\}$. We can upper bound the right-hand side of Theorem

C.3.6 via following proof steps of Theorem 4 of Simchowitz et al. [245]:

$$\mathbf{f}^{\text{pred}}\mathbf{p.r.}-\left(\frac{6}{\alpha}\sum_{t=k+1}^{T}\|\boldsymbol{\epsilon}_t\|_2^2-\frac{\alpha}{48}\sum_{t=1}^{T}\left\|\mathbf{M}_t-\mathbf{M}_{comp}\right\|_F^2\right)\lesssim\frac{L^2H'^3d\kappa_b^4\kappa_{\mathbf{G}}^4\kappa_{\mathcal{M}}^2}{\min\{\alpha,L\kappa_b^2\kappa_{\mathbf{G}}^2\}}\left(1+\frac{\overline{\alpha}_{loss}}{dL\kappa_{\mathcal{M}}}\right)\log\left(\frac{T}{\delta}\right)$$

$$\mathbf{f}^{\text{pred}}\mathbf{p.r.}+\frac{\alpha}{48}\sum_{t=1}^{T}\left\|\mathbf{M}_t-\mathbf{M}_{comp}\right\|_F^2\lesssim\frac{L^2H'^3d\kappa_b^4\kappa_{\mathbf{G}}^4\kappa_{\mathcal{M}}^2}{\min\{\alpha,L\kappa_b^2\kappa_{\mathbf{G}}^2\}}\left(1+\frac{\overline{\alpha}_{loss}}{dL\kappa_{\mathcal{M}}}\right)\log\left(\frac{T}{\delta}\right)$$
$$+\frac{1}{\alpha}\sum_{t=T_w+1}^{T}C_{\text{approx}}^2\epsilon_{\mathbf{G}}^2\left(\left\lceil\log_2\left(\frac{t}{T_w}\right)\right\rceil,\delta\right),\tag{C.38}$$

where (C.38) follows from (C.36). □

Comparator Approximation Error:

Lemma C.3.5. *Suppose that* $H'\geq 2H_0'-1+H$, $\psi_{\mathbf{G}}(H+1)\leq 1/10T$. *Then for all* $\tau>0$,

Comp. app. err. $\leq 4L\kappa_y\kappa_u\kappa_{\mathcal{M}}$

$$+\sum_{t=T_w+1}^{T}\left[\tau\left\|\mathbf{M}_t-\mathbf{M}_{\text{comp}}\right\|_F^2+8\kappa_y^2\kappa_b^2\kappa_{\mathcal{M}}^2(H+H')\max\left\{L,\frac{L^2}{\tau}\right\}\epsilon_{\mathbf{G}}^2\left(\left\lceil\log_2\left(\frac{t}{T_w}\right)\right\rceil,\delta\right)\right]$$

Proof. The lemma can be proven using the proof of Proposition 8.2 of Simchowitz et al. [245]. Using Lemma E.3 and adapting Lemma E.4 in Simchowitz et al. [245] such that $\mathbf{M}_{comp}^{[i]}=M_*^{[i]}I_{i\leq H_0'-1}+\sum_{a=0}^{H_0'-1}\sum_{b=0}^{H}\sum_{c=0}^{H_0'-1}M_*^{[a]}(\widehat{G}_1^{[b]}-G^{[b]})M_*^{[c]}\mathbb{I}_{a+b+c=i}$ for $\mathbf{M}_*=\text{argmin}_{\mathbf{M}\in\mathcal{M}_\psi}\sum_{t=T_w+1}^{T}\ell_t(y_t^{\mathbf{M}},u_t^{\mathbf{M}})$ and due to Lemma C.3.3 we have $\mathbf{M}_{comp}\in\mathcal{M}$:

$$\sum_{t=T_w+1}^{T}f_t^{\text{pred}}\left(\mathbf{M}_{comp}\right)-\inf_{\mathbf{M}\in\mathcal{M}_0}\sum_{t=T_w+1}^{T}f_t\left(\mathbf{M},\mathbf{G},b_1(\mathbf{G}),\ldots,b_t(\mathbf{G})\right)$$

$$\leq 4L\kappa_y\sum_{t=T_w+1}^{T}\epsilon_{\mathbf{G}}^2\left(\left\lceil\log_2\left(\frac{t}{T_w}\right)\right\rceil,\delta\right)\kappa_{\mathcal{M}}^2\kappa_b\left(\kappa_{\mathcal{M}}+\frac{\kappa_b}{4\tau}\right)+\kappa_u\kappa_{\mathcal{M}}\psi_{\mathbf{G}}(H+1)+(H+H')\tau\left\|\mathbf{M}_t-\mathbf{M}_{\text{comp}}\right\|_F^2$$

$$\leq\sum_{t=T_w+1}^{T}\left[\tau\left\|\mathbf{M}_t-\mathbf{M}_{\text{comp}}\right\|_F^2+8\kappa_y^2\kappa_b^2\kappa_{\mathcal{M}}^2(H+H')\max\left\{L,\frac{L^2}{\tau}\right\}\epsilon_{\mathbf{G}}^2\left(\left\lceil\log_2\left(\frac{t}{T_w}\right)\right\rceil,\delta\right)\right]$$

$$+4TL\kappa_y\kappa_u\kappa_{\mathcal{M}}\psi_{\mathbf{G}}(H+1)$$

$$\leq 4L\kappa_y\kappa_u\kappa_{\mathcal{M}}+\sum_{t=T_w+1}^{T}\left[\tau\left\|\mathbf{M}_t-\mathbf{M}_{\text{comp}}\right\|_F^2+8\kappa_y^2\kappa_b^2\kappa_{\mathcal{M}}^2(H+H')\max\left\{L,\frac{L^2}{\tau}\right\}\epsilon_{\mathbf{G}}^2\left(\left\lceil\log_2\left(\frac{t}{T_w}\right)\right\rceil,\delta\right)\right]$$

□

Combining all the terms bounded above, with $\tau = \frac{\alpha}{48}$ gives

$$
\begin{aligned}
&\textsc{Regret}(T) \\
&\lesssim T_w L\kappa_y^2 + L\kappa_y\kappa_u/10 + L\kappa_{\mathbf{G}}\kappa_{\mathcal{M}}^2\kappa_b^2/10 + 2L\kappa_{\mathcal{M}}\kappa_y\kappa_{\mathbf{G}}\kappa_b + 4L\kappa_y\kappa_u\kappa_{\mathcal{M}} \\
&+ \frac{L^2H'^3\min\{m,p\}\kappa_b^4\kappa_{\mathbf{G}}^4\kappa_{\mathcal{M}}^2}{\min\{\alpha, L\kappa_b^2\kappa_{\mathbf{G}}^2\}}\left(1+\frac{\overline{\alpha}_{loss}}{\min\{m,p\}L\kappa_{\mathcal{M}}}\right)\log\left(\frac{T}{\delta}\right) + \frac{1}{\alpha}\sum_{t=T_w+1}^{T} C_{\text{approx}}^2\epsilon_{\mathbf{G}}^2\left(\left\lceil\log_2\left(\frac{t}{T_w}\right)\right\rceil,\delta\right) \\
&+ \sum_{t=T_w+1}^{T} 8\kappa_y^2\kappa_b^2\kappa_{\mathcal{M}}^2(H+H')\max\left\{L, \frac{48L^2}{\alpha}\right\}\epsilon_{\mathbf{G}}^2\left(\left\lceil\log_2\left(\frac{t}{T_w}\right)\right\rceil,\delta\right) \\
&\lesssim T_w L\kappa_y^2 \\
&+ \frac{L^2H'^3\min\{m,p\}\kappa_b^4\kappa_{\mathbf{G}}^4\kappa_{\mathcal{M}}^2}{\min\{\alpha, L\kappa_b^2\kappa_{\mathbf{G}}^2\}}\left(1+\frac{\overline{\alpha}_{loss}}{\min\{m,p\}L\kappa_{\mathcal{M}}}\right)\log\left(\frac{T}{\delta}\right) \\
&+ \sum_{t=T_w+1}^{T}\epsilon_{\mathbf{G}}^2\left(\left\lceil\log_2\left(\frac{t}{T_w}\right)\right\rceil,\delta\right)\left\{\frac{H'\kappa_{\mathbf{G}}^2\kappa_{\mathcal{M}}^2\kappa_b^4\left(\overline{\alpha}_{loss}+L\right)^2}{\alpha} + \kappa_y^2\kappa_b^2\kappa_{\mathcal{M}}^2(H+H')\max\left\{L,\frac{48L^2}{\alpha}\right\}\right\}
\end{aligned}
$$

□

Following the doubling update rule of AdaptOn for the epoch lengths, after T time steps of agent-environment interaction, the number of epochs is $O(\log T)$. From Lemma C.3.3, at any time step t during the i^{th} epoch, i.e., $t \in [t_i, \ldots, t_i - 1]$, $\epsilon_{\mathbf{G}}^2(i,\delta) = O(polylog(T)/2^{i-1}T_w)$. Therefore, update rule of AdaptOn yields,

$$
\sum_{t=T_w+1}^{T}\epsilon_{\mathbf{G}}^2\left(\left\lceil\log_2\left(\frac{t}{T_w}\right)\right\rceil,\delta\right) = \sum_{i=1}^{O(\log T)} 2^{i-1}T_w\epsilon_{\mathbf{G}}^2(i,\delta) \leq O\left(polylog(T)\right) \tag{C.39}
$$

Using the result of (C.39), we can bound the third term of the regret upper bound in Theorem 5.10 with a $polylog(T)$ bound which gives the advertised result and using the policy approximation error term we obtain Corollary 5.10.1.

□

C.3.5 Technical Lemmas and Theorems

Theorem C.3.6 (Theorem 8 of Simchowitz et al. [245])**.** *Suppose that $\mathcal{K} \subset \mathbb{R}^d$ and $h \geq 1$. Let $F_t := \mathcal{K}^{h+1} \to \mathbb{R}$ be a sequence of L_c coordinatewise-Lipschitz functions with the induced unary functions $f_t(x) := F_t(x, \ldots, x)$ which are L_{f}-Lipschitz and β-smooth. Let $f_{t;k}(x) := \mathbb{E}\left[f_t(x)|\mathcal{F}_{t-k}\right]$ be α-strongly convex on $\mathcal{K}$ for a filtration $(\mathcal{F}_t)_{t\geq 1}$. Suppose that $z_{t+1} = \Pi_{\mathcal{K}}\left(z_t - \eta \boldsymbol{g}_t\right)$, where $\boldsymbol{g}_t = \nabla f_t\left(z_t\right) + \epsilon_t$ for $\left\|\boldsymbol{g}_t\right\|_2 \leq L_{\mathbf{g}}$,*

and $\mathrm{Diam}(\mathcal{K}) \leq D$. *Let the gradient descent iterates be applied for* $t \geq t_0$ *for some* $t_0 \leq k$, *with* $z_0 = z_1 = \cdots = z_{t_0} \in \mathcal{K}$ *for* $k \geq 1$. *Then with step size* $\eta_t = \frac{3}{\alpha t}$, *the following bound holds with probability* $1-\delta$ *for all comparators* $z_\star \in \mathcal{K}$ *simultaneously:*

$$\sum_{t=k+1}^{T} f_t(z_t) - f_t(z_\star) - \left(\frac{6}{\alpha}\sum_{t=k+1}^{T}\|\boldsymbol{\epsilon}_t\|_2^2 - \frac{\alpha}{12}\sum_{t=1}^{T}\|z_t - z_\star\|_2^2\right)$$
$$\lesssim \alpha k D^2 + \frac{\left(kL_{\mathrm{f}} + h^2 L_{\mathrm{c}}\right)L_{\mathrm{g}} + kdL_{\mathrm{f}}^2 + k\beta L_{\mathrm{g}}}{\alpha}\log(T) + \frac{kL_{\mathrm{f}}^2}{\alpha}\log\left(\frac{1+\log\left(e+\alpha D^2\right)}{\delta}\right)$$

Lemma C.3.7 (Lemma 8.1 of Simchowitz et al. [245])**.** *For any* $\mathbf{M} \in \mathcal{M}$, *let* $f_t^{pred}(\mathbf{M})$ *denote the unary counterfactual loss function induced by true truncated counterfactuals (Definition 8.1 of Simchowitz et al. [245]). During the i'th epoch of adaptive control period, at any time step* $t \in [t_i, \ \ldots, \ t_{i+1}-1]$, *for all i, we have that*

$$\left\|\nabla f_t\left(\mathbf{M}, \widehat{\mathbf{G}}_i, b_1(\widehat{\mathbf{G}}_i), \ldots, b_t(\widehat{\mathbf{G}}_i)\right) - \nabla f_t^{pred}(\mathbf{M})\right\|_{\mathrm{F}} \leq C_{approx}\,\epsilon_{\mathbf{G}}(i,\delta),$$

where $C_{approx} := \sqrt{H'}\kappa_{\mathbf{G}}\kappa_{\mathcal{M}}\kappa_b^2\left(16\overline{\alpha}_{loss} + 24L\right)$.

Lemma C.3.8 (Lemma 8.2 of Simchowitz et al. [245])**.** *For any* $\mathbf{M} \in \mathcal{M}$, $f_t^{pred}(\mathbf{M})$ *is* β*-smooth, where* $\beta = 16H'\kappa_b^2\kappa_{\mathbf{G}}^2\overline{\alpha}_{loss}$.

Lemma C.3.9 (Lemma 8.3 of Simchowitz et al. [245])**.** *For any* $\mathbf{M} \in \mathcal{M}$, *given* $\epsilon_{\mathbf{G}}(i,\delta) \leq \frac{1}{4\kappa_b\kappa_{\mathcal{M}}\kappa_{\mathbf{G}}}\sqrt{\frac{\alpha}{H'\underline{\alpha}_{loss}}}$, *conditional unary counterfactual loss function induced by true counterfactuals are* $\alpha/4$ *strongly convex.*

Lemma C.3.10 (Lemma 8.4 of Simchowitz et al. [245])**.** *Let* $L_f = 4L\sqrt{H'}\kappa_b^2\kappa_{\mathbf{G}}^2\kappa_{\mathcal{M}}$. *For any* $\mathbf{M} \in \mathcal{M}$ *and for* $T_w \geq T_{\max}$, $f_t^{pred}(\mathbf{M})$ *is* $4L_f$*-Lipschitz,* $f_t^{pred}[\mathbf{M}_{t:t-H}]$ *is* $4L_f$ *coordinate Lipschitz. Moreover,* $\max_{\mathbf{M}\in\mathcal{M}}\left\|\nabla f_t\left(\mathbf{M}, \widehat{\mathbf{G}}_i, b_1(\widehat{\mathbf{G}}_i), \ldots, b_t(\widehat{\mathbf{G}}_i)\right)\right\|_2 \leq 4L_f$.

Lemma C.3.11 (Doubling Trick [125])**.** *For any sequence of numbers* $z_1, \ldots, z_n$ *with* $0 \leq z_k \leq Z_{k-1} := \max\left\{1, \sum_{i=1}^{k-1} z_i\right\}$

$$\sum_{k=1}^{n}\frac{z_k}{\sqrt{Z_{k-1}}} \leq (\sqrt{2}+1)\sqrt{Z_n}$$

Lemma C.3.12 (Norm of a Sub-Gaussian Vector)**.** *Let* $v \in \mathbb{R}^d$ *be a entry-wise R-subgaussian random variable. Then with probability* $1-\delta$, $\|v\| \leq R\sqrt{2d\log(2d/\delta)}$.

Appendix D

FURTHER PROOFS FOR CHAPTER 6

D.1 Proofs of Section 6.3

In this section, we present the technical results and the proofs. First, we discuss the related assumptions in the main text. Then we provide proof of the results and discuss them. In particular, in Appendix D.1.1, we provide the proof for Proposition 6.1, i.e., model learning guarantees, in Appendix D.1.2, we provide the stability and boundedness results for the warm-up and adaptive control periods, and in Appendix D.1.3 we give the proofs for main regret guarantees, namely Theorem 6.4 and Corollary 6.4.1.

Assumptions

Notice that Assumption 6.2 provides that for bounded inputs, the dynamics of the system stay bounded. In particular, the Lipschitz assumption avoids arbitrary changes in the output due to the noise term and the exponential stability assumption avoids output blow-ups due to unmodeled dynamics. Moreover, the exponential open-loop stability assumption can be replaced by the existence of a controller that keeps the state bounded. Note that in many nonlinear control systems, these conditions are already satisfied due to physical laws, e.g., the state of the system cannot be unbounded. This behavior in particular demonstrated in dissipative dynamical systems [120, 253]. As explained in the main text, the assumption that the periodic extension of the ith mapping of F from s_t to y_t, $F_i(\cdot)$, lives in Sobolev space of periodic functions, $W_p^{m,2}([0, 2\pi]^{h(d_y+d_u)})$, allows one to use Fourier Series as a learning basis. Note that the theoretical guarantees of FALCON hold for all systems that have up-to mth order (mixed) derivative in all indices in the weak sense such that they are all in L^2 space, and there exists linear transformation (scaling and/or shifting) and periodic extension that brings them in $W_p^{m,2}([0, 2\pi]^{h(d_y+d_u)})$. This class of systems constitutes a wide range of dynamical systems.

Assumption 6.1 says that the difference in cost of two observation and input pairs that are close to each other are quadratic in the differences of outputs and inputs. This assumption makes the regret minimization problem meaningful since learning the model dynamics would result in a smaller instantaneous cost difference.

This assumption is satisfied for bounded input-output systems with quadratic cost functions.

Assumption 6.8 is on the MPC policy and it relates to both the underlying system and the cost functions. It guarantees that the cost functions are designed in a meaningful way (running or terminal costs) such that minimizing these objectives in return stabilizes the system dynamics and this property is valid under small modeling errors. This assumption is valid in practice, since while deploying MPC in real-world control systems, engineers design the policies based on possible modeling errors, and Assumption 6.8 merely quantifies this. As described in the main text, this assumption holds for linearized systems. The persistence of excitation condition means that the inputs excite the system uniformly, i.e., the smallest eigenvalue of the design matrix $\Phi_t \Phi_t^\top$ scales linearly over time. This assumption is fairly standard in a system identification setting and it guarantees the consistent and reliable estimation of the underlying system. The combination of the unmodelled dynamics and the system noise, as well as the sampling-based MPC policy design, can provide the required randomness to satisfy this assumption, yielding a mild assumption. Notice that if this assumption does not hold, one can show that FALCON can still attain sublinear regret rate, *i.e.* $O(T^{2/3})$, by setting a horizon-dependent warm-up period. Finally, the local Lipschitz condition for the neighborhood of models is valid for many dynamical systems and follows intuitively by the Lipschitz assumption of the underlying system. This is particularly important that the outputs of the MPC policy, i.e., the designed control actions u_t, do not vary arbitrarily. This is again a mild assumption since the control actions that stabilize the system dynamics are not expected to vary significantly in most of the dynamical systems.

D.1.1 Model Learning Guarantees

We first provide the following system identification guarantee for learning the unknown optimal Fourier Series coefficients Θ_* using a finite number of data points. It follows standard least-squares estimation error analysis under sub-Gaussian perturbations [2, 164, 166].

Proof of Proposition 6.1

Proof. Recall that the dynamics can be written as

$$\begin{bmatrix} y_{t,1} \\ \vdots \\ y_{t,m} \end{bmatrix} = \begin{bmatrix} F_1(s_t) \\ \vdots \\ F_{d_y}(s_t) \end{bmatrix} + e_t. \tag{D.1}$$

From Assumption 6.4, we know that $F_i(\cdot) : \mathbb{R}^{h(d_y+d_u)} \to \mathbb{R}$, that lives in $W_p^{k,2}(\Omega)$ for $\Omega = [0, 2\pi]^{h(d_y+d_u)}$ and for $1 \leq i \leq d_y$. The Fourier basis that FALCON learns the underlying system $F(\cdot)$ is given as

$$\phi(s) = [\cos(\boldsymbol{\omega}_\mathbf{1}^\mathsf{T} s), \sin(\boldsymbol{\omega}_\mathbf{1}^\mathsf{T} s), ..., \cos(\boldsymbol{\omega}_{\frac{\boldsymbol{D}}{\mathbf{2}}}^\mathsf{T} s), \sin(\boldsymbol{\omega}_{\frac{\boldsymbol{D}}{\mathbf{2}}}^\mathsf{T} s)]^\top,$$

where $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_d]$, $\omega_j \in \{0, 1, \ldots, n\}$, $1 \leq j \leq d$, and $s \in \mathbb{R}^{h(d_y+d_u)}$. Moreover, due to Assumption 6.2, we have bounded inputs and outputs such that $\|s_t\| \leq S$ for all $t \leq T_w$, see Lemma D.1. After the warm-up period FALCON approximates the underlying dynamics as $\hat{\Theta}_1^\top \phi(s)$ where $\hat{\Theta}_1 = [\hat{\theta}_1^{[1]}, \hat{\theta}_1^{[2]}, \ldots, \hat{\theta}_1^{[d_y]}] \in \mathbb{R}^{D \times d_y}$ with ith column denoted as $\hat{\theta}_1^{[i]}$. The absolute approximation error of any function F_i can be decomposed as follows:

$$\begin{aligned}
\sup_{\|s\| \leq S} |F_i(s) - \hat{F}_i(s)| &= \sup_{\|s\| \leq S} \left| \sum_{\boldsymbol{\omega}} \left[a_{\boldsymbol{\omega}} \cos\left(\boldsymbol{\omega}^\mathsf{T} s\right) + b_{\boldsymbol{\omega}} \sin\left(\boldsymbol{\omega}^\mathsf{T} s\right) \right] - \hat{\theta}^{[i]\top} \phi(s) \right| \\
&\leq \underbrace{\sup_{\|s\| \leq S} \left| \sum_{\boldsymbol{\omega}} \left[a_{\boldsymbol{\omega}} \cos\left(\boldsymbol{\omega}^\mathsf{T} s\right) + b_{\boldsymbol{\omega}} \sin\left(\boldsymbol{\omega}^\mathsf{T} s\right) \right] - \theta_*^{[i]\top} \phi(s) \right|}_{\text{Finite number of Fourier basis approximation}} + \underbrace{\left| \theta_*^{[i]\top} \phi(s) - \hat{\theta}^{[i]\top} \phi(s) \right|}_{\text{Finite sample approximation}}
\end{aligned}$$

for $\theta_*^{[i]} \in \mathbb{R}^D$, where $\theta_*^{[i]\top} \phi(s)$ denotes the best Fourier series approximation of F_i, using the nth order Fourier expansion, i.e. D-dimensional Fourier basis. The first term can be bounded by a multivariate analog of Jackson's theorem for trigonometric polynomial approximation [124]. In particular, Theorem 4.3 of [237] states that for large enough n,

$$\sup_{\|s\| \leq S} \left| F_i - \theta_*^{[i]\top} \phi(s) \right| \leq C n^{-m} \|\partial^m F_i(\cdot)\|_{L^\infty}.$$

The exact values of required n and C can be collected from Sections 2 and 3 of [237]. In particular, they depend on S, exponentially on d_y, and combinatorically on n. Using Theorem 6.1 to bound the second term with the fact that $\|\phi(s)\| \leq \sqrt{D}$

and $D = n^d$, we get $\sup_{\|s\|\leq S} |F_i(s) - \hat{F}_i(s)| \leq O(\frac{n^d}{\sqrt{T}} + \frac{\|\partial^m F_i(\cdot)\|_{L^\infty}}{n^m})$ with probability $1 - 2\delta$. Since the dynamics are written as (D.1), via union bound over all columns, we have

$$\sup_{\|s\|\leq S} \|F(s) - \hat{F}(s)\|_\infty = O\left(\frac{n^d}{\sqrt{T}} + \frac{\sup_i \|\partial^m F_i(\cdot)\|_{L^\infty}}{n^m}\right) \tag{D.2}$$

with probability at least $1 - 2\delta$. Note that union bound provides additional logarithmic factor depending on δ/d_y which is hidden under $O(\cdot)$ notation. For $T_w > \left(\frac{n^d}{1-\alpha_m/n^m}\right)^2$, we have that $\sup_{\|s\|\leq S} \|F(s) - \hat{F}_1(s)\|_\infty = O(T_w^{\varepsilon-0.5})$ for $0.5 > \varepsilon \geq 0$. Note that we achieve $\varepsilon = 0$ as m, *i.e.* smoothness of the underlying system, goes to infinity.

□

D.1.2 Stability / Bounded Output Guarantees

In this subsection, we provide the stability and boundedness guarantees for FALCON. In particular, Lemma D.1 shows that the bounded exploratory inputs of warm-up yield bounded outputs, and Lemma D.1.1 shows that for long enough warm-up duration T_w and a sufficiently large order of Fourier basis n, FALCON stabilizes the underlying system. First, let $\|u_t\| < B_{u_{exp}}$ for $t \leq T_w$, i.e., the persistently exciting warm-up control inputs are bounded. Moreover, using Lemma C.3.12, we have that $\|e_t\| \leq B_e := \sigma_e\sqrt{2d_y \log(2d_yT/\delta)}$ with probability $1 - \delta$ for all $t \leq T$. Without loss of generality, we assume that $y_0 = 0$. Using these definitions we have the following result.

Lemma D.1 (Bounded Output During Warm-Up)**.** *Suppose Assumption 6.2 holds and FALCON uses exploratory inputs such that $\|u_t\| \leq B_{u_{exp}}$. Starting from zero initial condition $y_0 = 0$, after the warm-up period we have $\|s_t\| \leq S$ such that $S = \sqrt{h}\left((K+1)B_{u_{exp}} + B_e\right)$, with probability $1 - \delta$.*

Proof. Using the exponential input-to-output stability assumption of the underlying open-loop system, we have $\|y_t\| \leq KB_{u_{exp}} + B_e$, with probability $1 - \delta$. Using this result for the order-h NARX state $s_i = [y_{i-1}^\top, \ldots, y_{i-h}^\top, u_{i-1}^\top, \ldots, u_{i-h}^\top]^\top$, we get $\|s_t\| \leq S$ as given in the lemma. □

Lemma D.1.1 (Stability and Bounded Output During Adaptive Control)**.** *Suppose Assumption 6.8 holds. Given the order of the Sobolev space m and*

$$\alpha_m := \sup_{i,\|s\|\leq S} \|\partial^m F_i(s)\|_{L^\infty},$$

with probability $1-2\delta$*, FALCON with* $n \geq \left(\frac{2\alpha_m}{\epsilon}\right)^{1/m}$ *as the order of Fourier basis designs stabilizing controllers for the underlying system after*

$$T_w \geq \max\left\{\left(\frac{n^d}{1-\alpha_m/n^m}\right)^2, 4(4\alpha_m)^{\frac{d}{m}}\epsilon^{-2(\frac{d}{m}+1)}\right\}$$

time-steps of warm-up period and keeps s_t *bounded, i.e.,* $\|s_t\| \leq S$ *for all* t*.*

Proof. Using Assumption 6.8, FALCON needs to make sure that the estimation error of the model dynamics is well-controlled. To achieve this, FALCON is required to use high enough order of Fourier Series basis and maintain a long enough warm-up duration. In particular, given the order of the Sobolev space m and $\alpha_m = \sup_{i,\|s\|\leq S}\|\partial^m F_i(s)\|_{L^\infty}$, FALCON requires $n = \left(\frac{2\alpha_m}{\epsilon}\right)^{1/m}$ as the order of Fourier basis. This corresponds to $D = \left(\frac{2\alpha_m}{\epsilon}\right)^{d/m}$ number of Fourier Series basis vectors. Thus, for $T_w = O\left((4\alpha_m)^{\frac{d}{m}}\epsilon^{-2(\frac{d}{m}+1)}\right)$, with probability $1-2\delta$, we have that $\sup_{\|s\|\leq S}\|F(s)-\hat{F}(s)\|_\infty \leq \epsilon$. Thus, after T_w time-steps of warm-up period FALCON stabilizes the underlying system, and the outputs decay over time due to exponential input to output stability. □

D.1.3 Regret Analysis

In this section, we provide the proof of Theorem 6.4, the regret upper bound of FALCON. We first present the precise theorem statement for Theorem 6.4 and provide the proof. The proof follows similarly to [164] in terms of regret decomposition but the improved system identification result via Fourier series basis given in Proposition 6.1 yields an improved regret.

Theorem D.1.2. *Let Assumptions 6.2, 6.1, 6.4, and 6.8 hold. Given the order of the Sobolev space* m*,* $\alpha_m = \sup_{i,\|s\|\leq S}\|\partial^m F_i(s)\|_{L^\infty}$*, and* $d = h(d_y + d_u)$*, suppose FALCON uses* $n \geq \left(\frac{2\alpha_m}{\epsilon}\right)^{1/m}$ *as the order of Fourier basis. Then, with probability at least* $1-2\delta$*, after a warm-up period of*

$$T_w \geq \max\left\{\left(\tfrac{n^d}{1-\alpha_m/n^m}\right)^2, C_{m,d}\epsilon^{-2(\frac{d}{m}+1)}, C_{m,d}(\Gamma/L_o)^{-2(\frac{d}{m}+1)}, C_{m,d}(\Gamma/(LL_o))^{-2(\frac{d}{m}+1)}\right\}$$

for $C_{m,d} = 4(4\alpha_m)^{\frac{d}{m}}$*, FALCON with doubling epochs,* $t_{ep} = T_w 2^{i-1}$*, attains regret of* $\textsc{Regret}(T) = O(\sqrt{T} + \epsilon' T)$*, where* $\epsilon' = \frac{\alpha_m^2}{n^{2m}}$*. For sufficiently smooth systems, i.e.* $m = O(\log(T))$*, FALCON achieves* $\textsc{Regret}(T) = \tilde{O}(\sqrt{T})$*.*

Proof of Theorem 6.4 and Corollary 6.4.1

Recall that from Lemma D.1.1, for the given warm-up duration, we have

$$\sup_{\|s\|\leq S} \|F(s)-\hat{F}(s)\|_\infty \leq \bar{\epsilon}_1 < \epsilon,$$

which means that the closed-loop dynamics are stabilized. In the beginning of the adaptive control period, for simplicity, assume that both FALCON and policy $\pi_\star$ have the same initial condition $s' = \{y_{t:t-h+1}, u_{t-1:t-h+1}\}$. Note that this is for simplicity and due to exponential stability it only results in small deviation which could be made arbitrarily small, see Lemma D.1.1. Let $u_t^{\hat{F}}$ denote the control input designed by the MPC policy using the estimated model dynamics $\hat{F}(\cdot)$ in the planning, and similarly u_t^F for using $F(\cdot)$ in the planning. Furthermore, denote the next observation of the system once $u_t^{\hat{F}}$ and u_t^F are deployed as $y_{t+1}^{\hat{F}}$ and y_{t+1}^F respectively. Using the Lipschitz assumption of the underlying system dynamics and the MPC policy in planning models, for the same initial condition s', we have

$$\|u_t^F - u_t^{\hat{F}}\| \leq L_o\bar{\epsilon}, \qquad \|y_{t+1}^F - y_{t+1}^{\hat{F}}\| \leq LL_o\bar{\epsilon}. \tag{D.3}$$

Since the MPC policy provides exponential stability for both F and $\hat{F}$, we have that the output of the system is decaying. Therefore, for each time-step in the epoch one can show that differences given in (D.3) decays over time. Therefore, for $\bar{\epsilon} \leq \min\{\Gamma/L_o, \Gamma/(LL_o)\}$, we have that Assumption 6.1 holds. Due to the choice of T_w, FALCON satisfies this condition. By choosing $t_{ep} = T_w \times 2^{i-1}$ for ith epoch, we get the following regret decomposition for FALCON:

$$\textsc{Regret}(T) = \sum_{t=1}^{T_w}(C_t(y_t,u_t) - C_t(y_t^{\pi_\star}, u_t^{\pi_\star})) + \sum_{t=T_w+1}^{T}(C_t(y_t,u_t) - C_t(y_t^{\pi_\star}, u_t^{\pi_\star})) \tag{D.4}$$

$$\leq T_w(KB_{u_{exp}} + B_e)^2 + \sum_{i=1}^{\log_2(T-T_w)} 2^{i-1}T_wR(1+L)^2L_o^2\bar{\epsilon}_i^2, \tag{D.5}$$

where (D.5) follows from Assumption 6.1 and (D.3) for the upper bound on the difference in the inputs and outputs of FALCON and $\pi_\star$. Since the inputs are persistently exciting during the adaptive control phase, from Proposition 6.1, we have that $\bar{\epsilon}_i = O\left(\frac{n^d}{\sqrt{2^{i-1}T_w}} + \frac{\alpha_m}{n^m}\right)$. Let $\lesssim$ denote the upper bound up to system-

dependent constants. Inserting $\bar{\epsilon}_i$ to (D.5), we get

$$\textsc{Regret}(T) \lesssim T_w + \sum_{i=1}^{\log_2(T-T_w)} 2^{i-1}T_w \left(\frac{n^{2d}}{2^{i-1}T_w} + \frac{2n^{d-m}\alpha_m}{\sqrt{2^{i-1}T_w}} + \frac{\alpha_m^2}{n^{2m}} \right) \tag{D.6}$$

$$\lesssim T_w + n^{2d}\log_2(T - T_w) + n^{d-m}\alpha_m\sqrt{T} + T\frac{\alpha_m^2}{n^{2m}} \tag{D.7}$$

$$= O\left(n^{d-m}\alpha_m\sqrt{T} + T\frac{\alpha_m^2}{n^{2m}} \right). \tag{D.8}$$

Here (D.6) hides the constants in each term coming from (D.5), namely $(KB_{u_{exp}} + B_{y_{exp}} + B_e)^2$ for the first term and $R(1+L)^2L_o^2$ for the second term. (D.7) uses Lemma C.3.11 and (D.8) hides the logarithmic factors and constants that do not depend on the Fourier series and smoothness of the underlying function $F(\cdot)$. Note that (D.8) recovers the first statement of Theorem 6.4 for $\epsilon' = \frac{\alpha_m^2}{n^{2m}}$, where ϵ' directly depends on the smoothness of the underlying system. For sufficiently smooth systems, *i.e.* for a given horizon T if $m = O(\log(T))$, then FALCON attains $O(\sqrt{T})$ regret, proving the second statement in Theorem 6.4. Finally, for infinitely smooth systems, i.e., $m = \infty$, we see that last two terms of (D.7) vanish for $n > 1$, yielding regret that scales polylogarithmically in time, as stated in Corollary 6.4.1.