Statistical Foundations of Operator Learning

Thesis by
Nicholas H. Nelsen

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

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2024
Defended April 29, 2024

Nicholas H. Nelsen
ORCID: 0000-0002-8328-1199

ACKNOWLEDGEMENTS

First, I am extremely grateful to my advisor Andrew Stuart for his guidance, mentorship, kindness, and patience over the course of my Ph.D. It is a great privilege to say that I am one of Andrew's students. Thank you, Andrew, for taking me on as a student, supporting my professional development through collaborations and conference travel, and providing me with the freedom to grow as a researcher and independently explore my interests. I have learned so much from you about teaching, analysis, and how to be a successful academic.

I also thank Kaushik Bhattacharya, Tim Colonius, and Houman Owhadi for serving on my thesis committee. Several discussions with Kaushik served to further underscore the importance of the research in the penultimate chapter of this thesis. I remember Tim telling me how interesting the subject of statistical inference is when I first arrived at Caltech to start my Ph.D.; I am glad I pursued this line of work in my thesis. Furthermore, I always learn something new from every research meeting with Houman.

The teaching and lecturing quality at Caltech is outstanding. In particular, I learned everything I know about probability, matrix analysis, and mathematical writing from Joel Tropp in his excellent graduate courses. I also appreciated Joel's advice when I needed it from time to time. Andrew Stuart's courses such as functional analysis and inverse problems were equally stellar.

Thank you to my amazing collaborators: Maarten de Hoop, Oliver Dunbar, Bamdad Hosseini, Daniel Zhengyu Huang, Nikola Kovachki, Samuel Lanthaler, Matti Lassas, Oscar Leong, Zachary Morrow, Houman Owhadi, Andrew Stuart, Margaret Trautner, and Xianjin Yang. I have learned so much from you all!

Next, I thank Maarten de Hoop, Omar Ghattas, Youssef Marzouk, Houman Owhadi, Philippe Rigollet, Andrew Stuart, Alex Townsend, Karen Willcox, and Yunan Yang for their support of my career at such an early stage.

I wish to thank the staff in CMS and at Caltech more broadly for making the non-research aspects of my time as a graduate student go so smoothly, especially Jolene Brink, Holly Golcher, Sydney Garstang, Diana Bohler, and Jenni Campbell. Jolene *always* does a flawless job and I commend her for it. The Hixon Writing Center at Caltech was a helpful resource during the job application season. I also appreciated advice from Eviatar Bach, Mateo Díaz,

Bamdad Hosseini, Daniel Z. Huang, Matthew Levine, Krithika Manohar, Eliza O'Reilly, Andrew Stuart, Joel Tropp, and Robert Webber during this time.

I am grateful for the research visits, seminars, and/or interesting discussions at conferences around the world with Simone Brugiapaglia, Maarten de Hoop, Ernesto De Vito, Omar Ghattas, Tapio Helin, Bamdad Hosseini, Nikola Kovachki, Samuel Lanthaler, Matti Lassas, Youssef Marzouk, Richard Nickl, Houman Owhadi, Daniel Sanz-Alonso, Hayden Schaeffer, Claudia Schillings, Christoph Schwab, Alex Townsend, Joel Tropp, Rachel Ward, and Yunan Yang. These interactions have helped me shape my taste for research problems.

I am also thankful for my friends and academic colleagues from all over the globe: Giovanni Alberti, Eviatar Bach, Ricardo Baptista, Francesca Bartolucci, Pau Batlle, Nicolas Boullé, Tan Bui-Thanh, Edoardo Calvello, Lianghao Cao, Elizabeth Carlson, Haoxuan Chen, Peng Chen, Yifan Chen, Matthieu Darcy, Nick Dexter, Mateo Díaz, Oliver Dunbar, Tomi Esho, Ethan Epperly, Paz Fink-Shustin, Nicola Rares Franco, Alfredo Garbuno Iñigo, Scott Habermehl, Joey Hart, Franca Hoffmann, Vesa Kaarnioja, Brendan Keith, Daniel Leibovici, Oscar Leong, Eitan Levin, Matthew Levine, Matt Li, Zongyi Li, Lu Lu, Dingcheng Luo, Yury Korolev, Rob Macedo, Andreas Mang, Romit Maulik, Aimee Maurais, Roberto Molinaro, Kevin Miller, Tram Nguyen, Thomas O'Leary-Roseberry, Eliza O'Reilly, Assad Oberai, Elizabeth Qian, Luca Ratti, Deep Ray, Matteo Santacesaria, Florian Schäfer, Tapio Schneider, George Stepaniants, Ben Stevens, Cody Sutton, So Takao, Nathaniel Trask, Margaret Trautner, Richard Tsai, Roy Wang, Sven Wang, Jacob Waugh, Robert Webber, Jin-Long Wu, Xianjin Yang, Jakob Zech, and Yuhua Zhu.

The SIAM organization has played a big role in my career so far, from amazing conferences to outreach opportunities and even having my picture on the 2024 "Join SIAM" flyer. I was lucky to work with Edoardo Calvello, Jiajie Chen, Mateo Díaz, Daniel Leibovici, Eitan Levin, Eliza O'Reilly, Roy Wang, and Robert Webber on building up the Caltech SIAM Student Chapter and the CMX Student/Postdoc Seminar in CMS during the post-pandemic time.

I am fortunate to have my research supported by several funding sources including the National Science Foundation, Amazon through the Amazon/Caltech AI4Science Fellowship, the Office of Naval Research, the Department of Defense, the Air Force Office of Scientific Research, the Army Research Office, and the Resnick High Performance Computing Center at Caltech.

I greatly enjoyed my summer research internship in the Center for Computing Research at Sandia National Laboratories in 2018 and thank Peter Bosler for this opportunity. I developed many strong connections with staff scientists there and an early taste for independent computational mathematics research.

I would not be where I am today without my excellent educators and mentors at Oklahoma State University during my undergrad: Chris Francisco, Bus Jaco, Jamey Jacob, Jiří Lebl, Lisa Mantini, Omer San, Arvind Santhanakrishnan, Henry Segerman, and Jiahong Wu. Thank you for helping set me on the path.

Last but not least, I want to thank my family and Lindee for supporting me and my ambitions throughout this journey.

Nicholas H. Nelsen
Pasadena, CA
April 2024

ABSTRACT

This thesis studies operator learning from a statistical perspective. Operator learning uses observed data to estimate mappings between infinite-dimensional spaces. It does so at the conceptually continuum level, leading to discretization-independent machine learning methods when implemented in practice. Although this framework shows promise for physical model acceleration and discovery, the mathematical theory of operator learning lags behind its empirical success. Motivated by scientific computing and inverse problems where the available data are often scarce, this thesis develops scalable algorithms for operator learning and theoretical insights into their data efficiency.

The thesis begins by introducing a convergent operator learning algorithm that is implementable on a computer with controlled complexity. The method is based on linear combinations of function-valued random features, enjoys efficient training via convex optimization, and accurately approximates nonlinear solution operators of parametric partial differential equations. A statistical analysis derives state-of-the-art error bounds for the method and establishes its robustness to errors stemming from noisy observations and model misspecification. Next, the thesis tackles fundamental statistical questions about how problem structure, data quality, and prior information influence learning accuracy. Specializing to a linear setting, a sharp Bayesian nonparametric analysis shows that continuum linear operators, such as the integration or differentiation of spatially varying functions, are provably learnable from noisy input-output pairs. The theory reveals that smoothing operators are easier to learn than unbounded ones and that training with rough or high-frequency input data improves sample complexity. When only specific linear functionals of the operator's output are the primary quantities of interest, the final part of the thesis proves that the smoothness of the functionals determines whether learning directly from these finite-dimensional observations carries a statistical advantage over plug-in estimators based on learning the entire operator. To validate the findings beyond linear problems, the thesis develops practical deep operator learning architectures for nonlinear mappings that send functions to vectors, or vice versa, and shows their corresponding universal approximation properties. Altogether, this thesis advances the reliability and efficiency of operator learning for continuum problems in the physical and data sciences.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] Nicholas H. Nelsen and Andrew M. Stuart. "The random feature model for input-output maps between Banach spaces". *SIAM Journal on Scientific Computing* 43.5 (2021), A3212–A3243. DOI: 10.1137/20M133957X.
N.H.N. was the main contributor to this work, and A.M.S. had an advisory role. N.H.N. led the conceptualization of the project, designed and implemented the operator random features algorithm, wrote all code, generated all data, performed all numerical experiments, proved all theoretical results except for Result 2.8, and led the writing of the manuscript. This content is adapted for Chapter 2.

[2] Nicholas H. Nelsen and Andrew M. Stuart. "Operator learning using random features: a tool for scientific computing". *SIAM Review,* accepted for SIGEST section (2024).
This SIGEST article was originally published by the authors in *SIAM J. Sci. Comput.*, Vol. 43, No. 5, pp. A3212–A3243, 2021. N.H.N. was the main contributor of the revised material and wrote the revision, while A.M.S. provided input about the revision structure. Excerpts of this content are adapted for Chapters 1 and 2.

[3] Samuel Lanthaler and Nicholas H. Nelsen. "Error bounds for learning with vector-valued random features". In: *Advances in Neural Information Processing Systems* (spotlight paper). Ed. by A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 71834–71861. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/e34d908241aef40440e61d2a27715424-Abstract-Conference.html.
S.L. and N.H.N. contributed equally to this work. S.L. initiated the project and proof approach, proved approximation results for the noise-free and well-specified setting, proved the strong consistency and discretization error results, and removed the logarithmic factor by improving N.H.N.'s noise proof. N.H.N. participated in the conception of the project, developed the theoretical analysis for noise and misspecification errors (including Theorems 3.4 and 3.12 and the self-bounding argument in Appendix D.1), refined the rest of the generalization error analysis and proofs, wrote the code, prepared the data, performed the numerical experiment, and wrote the majority of the manuscript (except for Section 1). This content is adapted for Chapter 3.

[4] Maarten V. de Hoop, Nikola B. Kovachki, Nicholas H. Nelsen, and Andrew M. Stuart. "Convergence rates for learning linear operators from noisy data". *SIAM/ASA Journal on Uncertainty Quantification* 11.2 (2023), pp. 480–513. DOI: 10.1137/21M1442942.
N.H.N. was the main contributor to this work. A.M.S. had an advisory role and initiated the project and proof approach. N.B.K., N.H.N., and A.M.S. participated in the conceptualization of the project. N.H.N. proved all theoretical results (with suggestions from N.B.K. regarding the proofs of Theorems 3.3–3.4 and Lemma

B.6), designed and implemented all numerical experiments, generated all data, wrote all code (with help from N.B.K. on efficient GPU implementation), and wrote the manuscript. N.H.N. led the major revision of the article with input from A.M.S.. All authors edited the manuscript. This content is adapted for Chapter 4.

[5] Daniel Zhengyu Huang, Nicholas H. Nelsen, and Margaret Trautner. "An operator learning perspective on parameter-to-observable maps". *preprint arXiv:2402.06031 cs.LG* (2024). DOI: 10.48550/arXiv.2402.06031. N.H.N. was the main contributor to this work and conceptualized the project, designed the FNM architectures, proved all theoretical results, wrote the public codebase used in all numerical experiments, generated the data for and applied the new method to the advection–diffusion example, and led the writing of the manuscript. D.Z.H. performed the airfoil experiment. M.T. performed the homogenization experiment and helped write drafts of Sections 1 and 2. This content is adapted for Chapter 5.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Number *Page*

LIST OF TABLES

Chapter 1

THESIS INTRODUCTION

This thesis develops scalable data-driven methods for solving continuum problems, establishes theoretical guarantees on the reliability and robustness of these methods, and applies the methods in the physical and data sciences. It does so by combining mathematical analysis with domain-specific insight in a statistical *operator learning* framework. Operator learning lifts machine learning principles originally built for the task of function estimation from finite-dimensional data to the task of operator estimation from infinite-dimensional data. It is beginning to influence fields such as scientific computing, engineering, and imaging, where continuum objects play a central role. For instance, operator learning tools are starting to be adopted as fast surrogates for high fidelity partial differential equation solvers (e.g., for seismic waves [275]) or as data-driven components for larger and more complex systems (e.g., weather models [215]). More generally, it displays potential for accelerating and discovering complex physical models and solving previously intractable problems.

The foundations of operator learning are typically motivated from an approximation theory [153] or numerical linear algebra [41] perspective. These perspectives are founded on the principles of numerical analysis, which studies continuum physical models and their discretized approximations when simulated on computers. At the infinite-dimensional level, such problems are formulated using the language of functional analysis. The restriction of the underlying function spaces that contain the phenomena of interest to appropriate finite-dimensional approximation subspaces informs the design of accurate numerical methods such as the classical finite element method. These classical methods are problem-specific, yet are interpretable due to the physical structure encoded by the underlying mathematical model. They assume noise-free initial data, boundary conditions, and source terms. Theoretical results study convergence as the number of degrees of freedom of the approximation tends to infinity. Much of the research on operator learning approximations of continuum physical models parallels this numerical analysis landscape. For instance, the training data are noiseless and generated from a well-posed mathematical model. The theory often focuses on convergence guarantees as the number of

degrees of freedom of the operator learning architecture—i.e., the number of free parameters—increases.

On the other hand, statistical learning theory works under the assumption that the observed data are samples from an unknown joint probability distribution. In the supervised learning setting, the finite samples of data consist of inputs and corresponding labeled outputs. The goal of statistical learning is to build a model from these data that accurately describes the underlying input-output relationship on average with respect to the unknown joint distribution. Theoretical work in the field centers on rates of estimation as the sample size, i.e., the number of data points, grows. The difficulty of such problems is often exacerbated by the high or infinite dimensionality of the data. Unlike in numerical analysis, there may not exist a ground truth mathematical model relating inputs to outputs in the statistical learning setting. Even if there is such a true model, it could be corrupted by complicated noise processes. Both schools of thought ultimately seek to make accurate predictions about interesting phenomena in computationally efficient ways given certain information about the problem.

This thesis belongs somewhere in between the statistical learning and numerical analysis viewpoints—although it is more closely aligned with statistical learning theory. As in numerical analysis, the present work always assumes the existence of an underlying continuum map that relates the infinite-dimensional input space to the infinite-dimensional output space. It also prescribes a random noise model on the observed outputs to represent measurement errors, as in statistical learning. This mathematical setup has close similarities to *Bayesian inverse problems* [253]. These inverse problems take the form

$$y = \mathcal{G}(u) + \eta\,, \tag{1.1}$$

where $\mathcal{G}$ is the true forward operator, u is an infinite-dimensional input parameter, and η represents noise. The Bayesian approach to statistical inverse problems models both u and η as random variables. The goal is to find the distribution of u given the observation y. In contrast, operator learning solves the completely different problem of finding $\mathcal{G}$ from many realizations of the pair (u, y) in (1.1). Nonetheless, one of the key technical strategies of this thesis is to reformulate such supervised operator learning problems into Bayesian inverse problems of the form (1.1) (with a different forward map to account for

the task of estimating $\mathcal{G}$). This enables the thesis to exploit the rich history of advances in statistical and Bayesian inverse problems and quantify the uncertainty inherent in the inference procedure.

As in statistical learning, the primary theoretical focus of the thesis is on *sample complexity*, which is the amount of training data required to achieve a desired accuracy level. Deep understanding of sample complexity has enormous practical consequences. For example, due to high computational or experimental burden, scientific data is not as abundant as data in the information sciences. Data generation is often too expensive or too time consuming. To make the most out of the limited quantities of data available, the study of sample complexity is essential. The thesis uncovers powerful theoretical insights into how to improve sample complexity for certain operator learning algorithms and also develops new algorithms that are provably sample-efficient.

Model misspecification is another recurring theme of the thesis. Since the ground truth operator $\mathcal{G}$ is usually partially or fully unknown, it is unlikely that $\mathcal{G}$ will belong to the chosen data-driven model class used to approximate it. This effect is typically quantified by a regularity or smoothness mismatch between the true $\mathcal{G}$ and elements of the model class. Hence, it is desirable to design and work with operator learning methods that are robust and stable to such model misspecification errors. The thesis studies this question for both linear and nonlinear problems.

This introductory chapter continues with a discussion on the need for continuum learning algorithms in Section 1.1 and how operator learning addresses this need in Section 1.2. Section 1.3 details the major contributions of the present work, while Section 1.4 outlines the organization of the thesis.

1.1 Continuum Algorithms for Continuum Problems

The challenge of inferring or approximating infinite-dimensional objects from finite amounts of information is ubiquitous in science, engineering, medicine, and beyond. Classically, this challenge is usually due to prediction or downstream tasks that arise from a well-defined mathematical model. Such tasks, which include inversion, control, optimization, and uncertainty propagation, are then tackled with mature numerical methods that have been developed over several decades. These methods are tailored to the specific type of underlying continuum model, typically an ordinary, partial, or stochastic differential

equation, and require domain-specific expertise to fully exploit their power.

In contrast, due to the ongoing revolution in the data and information sciences, there is increasing interest in using machine learning methods for these complex scientific tasks that overcome deficiencies stemming from imperfect knowledge of the mathematical model or from the existing numerical methods themselves (e.g., high dimensionality and high computational cost). For example, data-driven techniques are being applied in problems that range from discovering new closure relations in climate models [273] to accelerating the design of novel materials [24]. Black-box machine learning tools are extremely flexible and come with fully fledged software libraries, which supports their user-friendliness and spurs widespread adoption. However, the character of scientific data differs substantially from data in the computer sciences, which is the area where machine learning has thus far exhibited the most success. Indeed, the physical world is naturally modeled with infinite-dimensional continuum quantities, for example, temperature or pressure fields of a fluid. Such objects are spatially and temporally varying functions that have intrinsic smoothness properties, long-range correlations, and span multiple scales. The mathematical models that underpin natural phenomena are highly structured and interpretable. Continuum data may also be heterogeneous, partially observed, and represented in several different forms (e.g., in finite element, Fourier, or wavelet bases). However, by using off-the-shelf machine learning tools that treat discretizations of these data at face-value—purely as finite-dimensional vectors—researchers are inadvertently throwing away continuum information that is crucial for making accurate inferences. In response, this thesis tackles the design and analysis of machine learning algorithms that are tailor-made for scientific data and, more generally, other types of continuum data and models.

The general philosophy of designing algorithms at the continuum level has been successful across mathematical disciplines. In optimization problems constrained by partial differential equations, there is the "optimize-then-discretize" principle [127] which uses ideas from variational analysis to perform optimization in infinite dimensions and only discretizes the result at implementation time. In applied probability, there are continuum Markov chain Monte Carlo (MCMC) algorithms for sampling probability distributions supported on infinite-dimensional function spaces [68]. The speed of convergence of these continuum MCMC methods does not degenerate under mesh refinement [121]. This

makes them more practical for large-scale problems. The Bayesian formulation of inverse problems on Banach spaces provides another example of the philosophy [253]. Here, an infinite-dimensional version of Bayes' rule leads to well-defined posterior measures that solve the infinite-dimensional inverse problem and can be sampled in a computationally scalable manner with the continuum MCMC algorithms previously discussed. There is also work along similar lines that extends numerical linear algebra routines for finite-dimensional vectors and matrices to new ones for infinite-dimensional functions and linear operators [262, 261]. For example, the work of Townsend and Trefethen [262] generalizes standard QR, LU, and Cholesky factorizations of matrices to analogous factorizations for compact linear operators acting on function spaces. Special care is taken so that the discretization of these continuum algorithms does not pollute the expected infinite-dimensional behavior [67]. All such methods inherit particular dimension-independent properties that make them more robust, more computationally efficient, and possibly more accurate. Operator learning, described in the next section, brings this powerful way of thinking to the realm of machine learning. The work in the present thesis may be understood within this general "machine-learn-then-discretize" framework.

1.2 Supervised Operator Learning

Recognizing the need for new mathematical development of learning algorithms that are tailor-made for continuum problems, researchers established the operator learning paradigm to build data-driven models that map between infinite-dimensional input and output spaces. An *operator* is an input-output relationship such that each input and corresponding output are infinite-dimensional objects. For example, the mapping from the current temperature in a room to the temperature one hour later is an operator. This is because temperature at a fixed time is a function characterized by its values at an uncountably infinite number of spatial locations. A more concrete example of an operator is the mapping $\mathcal{G}\colon (a, f) \mapsto u$ from coefficient function a and source term f to solution function u governed by the elliptic partial differential equation

$$-\nabla \cdot (a \nabla u) = f \tag{1.2}$$

equipped with appropriate boundary conditions. The thesis returns to this example in Chapter 2.

Although operator learning represents a great conceptual advance in data

science, infinite-dimensional quantities must always be discretized when represented on a computer or in real experiments. What distinguishes operator learning from traditional machine learning architectures that operate on high-dimensional discretized vectors is that in the continuum limit of infinite resolution, operator learning architectures have a well-defined and consistent meaning. They capture the fundamental continuum structure of the problem and not artifacts due to the particular choice of discretization. For a fixed set of trainable parameters, operator learning methods by design produce consistent results given any finite-dimensional discretization of the formally infinite-dimensional data. That is, they are inherently dimension- and discretization-independent. In practice, this means that when trained at one resolution or discretization, the learned operator can be transferred to any other resolution or discretization without the need for re-training. This property endows such methods with significant potential for discovering new scientific laws from diverse sources of real-world data.

The present work focuses primarily on *supervised operator learning*, which concerns the training of models to fit infinite-dimensional input-output pairs of labeled data. However, just as supervised learning is a subset of the whole field of machine learning, the operator learning framework extends to many more classes of problems, including some that are discussed in Chapter 6. Moreover, although the theory in Chapters 3, 4, and 5 is formulated within the setting of quite general infinite-dimensional spaces, the thesis is primarily motivated by operators of interest that map between spaces of spatially varying functions. Indeed, all of the numerical experiments in this work come from continuum science and engineering problems that are of this form.

To this end, supervised operator learning assumes that N random pairs of functional data $\{(u_n, y_n)\}_{n=1}^N$ are available. The inputs $\{u_n\}_{n=1}^N \subset \mathcal{U}$ from input space $\mathcal{U}$ might be functions $x \mapsto u_n(x)$ of some spatial variable $x \in \mathcal{D}$. The observed outputs

$$y_n = \mathcal{G}(u_n) + \eta_n \tag{1.3}$$

are possibly noisy evaluations of some unknown ground truth operator $\mathcal{G}$ at the input function u_n. These could be functions $x' \mapsto y_n(x')$, on a possibly different spatial domain $\mathcal{D}'$, belonging to output function space $\mathcal{Y}$. Given these data, the goal is to estimate the operator $\mathcal{G}\colon \mathcal{U} \to \mathcal{Y}$. Although various reconstruction

procedures exist for this task, by far the most common involves solving the empirical risk minimization problem

$$\min_{\Psi \in \mathscr{H}} \left\{ \frac{1}{N} \sum_{n=1}^{N} \ell\big(y_n, \Psi(u_n)\big) + \mathscr{R}(\Psi) \right\} \tag{1.4}$$

for some loss function $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that quantifies the data misfit. Here $\mathscr{H}$ is a user-defined hypothesis space of operators mapping $\mathcal{U}$ to $\mathcal{Y}$ and $\mathscr{R} \colon \mathscr{H} \to \mathbb{R}$ is a regularization functional that ameliorates the ill-posedness of learning from a finite number samples.[1] For the function-valued regression problems considered in this thesis, common choices for the loss function $\ell(\cdot, \cdot)$ include those derived from the squared $L^2(\mathcal{D}')$ norm, the $L^2(\mathcal{D}')$ norm itself, the squared $H^1(\mathcal{D}')$ norm, and relative versions of these choices. The $L^1(\mathcal{D}')$ norm is also a good option if the output functions are discontinuous in some way.

A recurring theme in this thesis is the tradeoff between *bias* and *variance* that occurs when adjusting the size of the set $\mathscr{H}$ (equivalently, adjusting the approximation power of a parametric model by changing the number of learnable parameters). Making $\mathscr{H}$ bigger reduces the bias error, but in general increases the variance. One must carefully balance the size of $\mathscr{H}$, the strength of regularization $\mathscr{R}$, and other errors—such as those due to discretization—in order to obtain optimal statistical guarantees on learning accuracy.

The accuracy of the trained model that solves (1.4) can be assessed in many different ways. The most common metric is the expected loss (i.e., expected risk) over the training distribution. Under an independent and identically distributed statistical model for the training data $\{(u_n, y_n)\}_{n=1}^N$, this involves replacing the N-term average in (1.4) with the full expectation over the joint law of the pair (u_1, y_1). Other notions of error include the worst-case error over compact sets of inputs, excess risk, and out-of-distribution expected loss. By working in a statistical learning framework, this thesis develops average-case convergence guarantees that are valid for the latter two accuracy metrics.

This section concludes by contextualizing the present thesis with other work in the literature. Several practically implementable operator learning architectures were developed concurrently [4, 34, 173, 181, 203, 207, 274]. These include the DeepONet [181], which generalizes and makes practical the main idea

[1] A precise definition of all the involved quantities $\mathcal{U}$, $\mathcal{Y}$, $\mathcal{G}$, $\mathscr{H}$, $\mathscr{R}$, and ℓ requires a technical functional-analytic setup that is deferred until Chapter 2 and beyond.

of Chen and Chen [59], PCA-Net [34], and the function-valued random features method that is developed in this thesis [203]. These models were followed by neural operators [154, 173], which lift the structure of finite-dimensional neural networks to the infinite-dimensional function space setting, and in particular the Fourier Neural Operator [172]. A related line of research aiming to more closely align model reduction with operator learning is the work on deep learning-based reduced order models [44, 100, 101, 102]; some of these studies also derive approximation guarantees. Details for and comparisons among these architectures are given in [153, Section 3]. At a high-level, each architecture represents a different choice of the hypothesis set $\mathscr{H}$ in (1.4). There are fewer studies on the effect of changing the regularization functional $\mathscr{R}$ or the loss function ℓ, but these quantities are just as important as $\mathscr{H}$.

Apart from the random features method, what the preceding architectures—collectively called "neural operators" in this section—share is a deep learning backbone. The approximation theory of such neural operators is fairly well developed by now [125, 130, 150, 152, 154, 158, 159, 160, 163]. It includes qualitative universal approximation, i.e., density, results as well as quantitative parameter complexity bounds, that is, the number of model parameters required to achieve accuracy ε. The paper [163] reveals a "curse of parametric complexity" in which the parameter complexity required to approximate general Lipschitz continuous operators is shown to be exponential in powers of ε^{-1}. This exponentially large parameter complexity aligns with the findings of older work [191] and suggests that efficient neural operator learning is impossible without further assumptions. If enough regularity is assumed, however, the curse is lifted. For example, this thesis shows that operators belonging to reproducing kernel Hilbert spaces are efficient to approximate. It is also known that linear or holomorphic target operators enjoy efficient algebraic approximation rates [3, 125]. However, what rates are possible for sets of operators "in between" holomorphic and Lipschitz ones is still an open question.

Some of the simplest operators are linear ones. There is a substantial body of work in this setting ranging from the learning of general linear operators [76, 135, 196, 251] to estimating Koopman operators [151], conditional mean embeddings of probability distributions [118, 142, 247, 259], and Green's functions of specific linear PDEs [40, 42, 111, 239]. The linear setting allows for thorough and sharp statistical analysis that leads to fundamental insights about the data

efficiency of operator learning in terms of problem structure and the quality of continuum data [40, 76, 130]. For this reason, a large fraction of the present thesis is devoted to learning linear operators and linear functionals. Although these mappings form a relatively small subset of all possible operators, they nonetheless cover a wide range of practical applications and physical phenomena. The linear setting is especially interesting from a theoretical viewpoint because it enables tight error bounds and convergence rates. As a byproduct, theorems of this type are usually valid for *infinitely many problems* specified by certain regularity assumptions on the linear maps. In contrast, theoretical analysis of nonlinear problems is often undertaken on a case-by-case basis [204, 205]. The insights gained from such specialized study are impressive, but generally are not as widely applicable as insights obtained from linear analysis, such as those in this thesis.

Regardless, there are some noteworthy results for nonlinear functionals and operators. In terms of sample complexity, which provides the training dataset size required to obtain ε accuracy, most of the existing theory depends on kernel methods, either in a reproducing kernel Hilbert space framework [52, 162] or via local averaging (e.g., kernel smoothers) [97, 210]. Indeed, the paper [209] performs nonlinear operator learning in the encoder–decoder paradigm, where the input and output spaces are represented by truncated orthonormal bases and the finite-dimensional coefficient-to-coefficient mapping is performed with a kernel smoother. The kernel smoother is then approximated with random Fourier features [221]. A similar idea is undertaken from a Gaussian process perspective [25], building upon more classical work on operator-valued kernels [52, 136]. Turning toward deep learning, error bounds are obtained for encoder–decoder neural operators such as DeepONet and PCA-Net in [178]. These results imply a "curse of sample complexity", i.e., exponentially large sample sizes, for learning general Lipschitz operators. Similar to the parameter complexity case, when enough regularity is assumed on the nonlinear operators of interest, as expressed through weighted tensor product structure, operator holomorphy, or analyticity, for example, minimax lower bounds return to much better algebraic rates with respect to sample size [5, 52, 133, 132]. Moreover, there exist both constructive and nonconstructive estimators that achieve these algebraic convergence rates for operator learning [3, 52, 162].

The mathematically oriented review articles by Boullé and Townsend [41] and

Kovachki, Lanthaler, and Stuart [153] contain more exhaustive literature reviews on the subject of operator learning. Additionally, the individual chapters in the present thesis provide more targeted exposition of closely related work. It should now be clear that growing evidence from the literature suggests that operator learning is emerging as a powerful tool to accelerate model-centric tasks in science and engineering or to discover unknown physical laws from experimental data. Nonetheless, the mathematical theory of operator learning is still relatively incomplete, which limits its impact. This thesis fills in some of the remaining gaps.

1.3 Thesis Contributions

By blending ideas from Bayesian inference, statistical inverse problems, and high-dimensional probability, this thesis develops new infinite-dimensional machine learning algorithms and analysis for continuum problems and datasets. Of central interest is the subtle interplay between the underlying problem structure, prior knowledge of such structure, and the amount of training data required to learn an accurate model. The next four subsections describe the major contributions of the work.

1.3.1 Regression With Function-Valued Random Features

Originally published in *SIAM J. Sci. Comput.*, Vol. 43, No. 5, pp. A3212–A3243 (2021), Chapter 2 proposes a randomized algorithm for learning nonlinear operators mapping between infinite-dimensional spaces of functions. The model consists of a linear combination of M random operators (i.e., the random features). The proposed *function-valued random features algorithm* involves learning the coefficients of this linear expansion. For a suitable training objective function, this is a finite-dimensional convex, quadratic optimization problem that is scalable to high data dimensions and large sample sizes. This contrasts with more complicated deep learning methods that are plagued by nonconvex training routines. Moreover, Result 2.8, summarized in the next display, provides further insight.

Equivalence to a Finite-Rank Operator-Valued Kernel Method

The supervised training procedure for function-valued random features is equivalent to ridge regression over a reproducing kernel Hilbert space of operators spanned by the M random features.

Equivalently, this result implies that the method may be interpreted as approximating the Gaussian process prior distribution (i.e., operator-valued covariance kernel) in a function-valued Gaussian process regression method [52, 228].

Function-valued random feature regression departs from traditional function-valued (also known as vector-valued) Gaussian process regression in two fundamental ways. First, function-valued random features significantly improve the computational complexity (time and memory) of full vector-valued Gaussian process regression. This is accomplished by the low-rank (i.e., rank at most M) random feature approximation of the infinite-rank prior covariance and simple linear algebra identities. The result is a relatively small $M \times M$ matrix inversion for random features, while the full Gaussian process kernel method requires inversion of an enormous $Nd_{\text{out}} \times Nd_{\text{out}}$ matrix in practice, where d_{out} is the dimension of the discretized output function space (which conceptually should be infinite) and N is the sample size. The latter inversion is basically impossible (at least computationally) in a true function space setting unless the kernel has some sort of trivial structure.

For the second point of departure, function-valued random features explicitly model (spatial) correlations in the output function space. This is of fundamental importance in the operator learning context because any random infinite-dimensional output function must necessarily have nonzero cross correlations to be a well-defined element of the ambient function space (which is always a Hilbert space). Mathematically, this is equivalent to the requirement that the prior covariance operator be trace-class. Full vector-valued Gaussian process regression typically assumes isotropic covariance kernels that are a scalar multiple of the identity operator. However, the identity operator is not trace-class in infinite dimensions. This undesirable choice is often made because unlike in the scalar kernel regression setting, there are no canonical operator-valued kernels. Until now, they have been defined on a case-by-case basis and usually involve some diagonal (e.g., identity) or rank-one Kronecker structure to make exact posterior inference possible. The proposed methodology is able to alleviate these issues with randomization and efficient convex optimization.

The final major contribution of Chapter 2 is the design of practically implementable random feature maps for parametric partial differential equation problems. These maps include the so-called *Fourier Space Random Features* (2.34), which define random operators by way of Fourier series coefficients and helps

to inspire the now widely adopted Fourier Neural Operator architecture. The chapter also introduces the now widely used viscous Burgers' equation benchmark problem and demonstrates the excellent performance of Fourier Space Random Features on this operator learning benchmark.

1.3.2 Complete Error Analysis of Function-Valued Random Features

Chapter 3 develops the theoretical foundations of the function-valued random features method. Originally published in *Adv. Neural Inf. Process. Syst.*, Vol. 36, pp. 71834–71861 (2023), the chapter proves strong statistical consistency and quantitative convergence rates for the algorithm. In the well-specified model setting, Theorem 3.9 gives the sharpest parameter complexity error bound for random features to date, even in the scalar regression setting (also see Table 3.1). The next display states this result informally.

State-of-the-Art Parameter Complexity for Random Features

For a dataset of size N, only $M \simeq \sqrt{N}$ random features suffice to guarantee that the squared generalization error of the trained function-valued random features method is $O(N^{-1/2})$ with high probability.

The theoretical analysis—which goes far beyond mere existence proofs—stands out from related work in the literature because it unifies all sources of error from the trained random feature model: approximation (number of features), estimation (finite sample size), optimization, noisy measurements, model misspecification, and discretization of the continuum data. As a result, *the function-valued random features method is the first provably convergent operator learning algorithm for nonlinear problems that is actually implementable on a computer with controlled complexity.*

The main technical contributions of Chapter 3 revolve around a novel empirical risk (i.e., training error) bound for the algorithm. In particular, the chapter obtains an $O(1)$ high probability bound on the norm $\|\widehat{\alpha}\|_M$ of the trained random feature model's coefficients $\widehat{\alpha} \in \mathbb{R}^M$ (Corollary 3.17). To do this, the analysis develops a novel recursive self-bounding argument that avoids logarithmic factors appearing in previous works. The rest of the proofs use Bernstein's inequality in Hilbert spaces (with careful truncations) and empirical process theory to control the generalization error. This approach avoids the suboptimal matrix concentration inequalities that all other papers use to derive

statistical guarantees for random feature regression and may prove useful in the analysis of other learning algorithms beyond random features.

1.3.3 Sample Complexity Analysis of Linear Operator Learning

Before the results in this thesis were established, several basic questions remained open: can unbounded operators such as differentiation of spatially varying functions be rigorously learned, and if so, how much data is required? Can robustness to data distribution shifts be guaranteed? How should a trained operator's accuracy be assessed? These problems are technically challenging, even in the linear setting. Resolving them is paramount before even considering the additional complexity that nonlinear operators bring. Chapter 4 attempts to close these major theoretical gaps in the setting of learning a self-adjoint linear operator on an infinite-dimensional Hilbert space.

Originally published in *SIAM/ASA J. Uncertain. Quantif.*, Vol. 11, No. 2, pp. 480–513 (2023), the chapter performs a theoretical analysis of Bayesian nonparametric posterior estimators of unknown linear operators. The theory leads to three fundamental principles that reveal the types of linear operators, types of training data, and types of distribution shift that improve sample complexity (Subsection 4.1.2). The next display summarizes these findings.

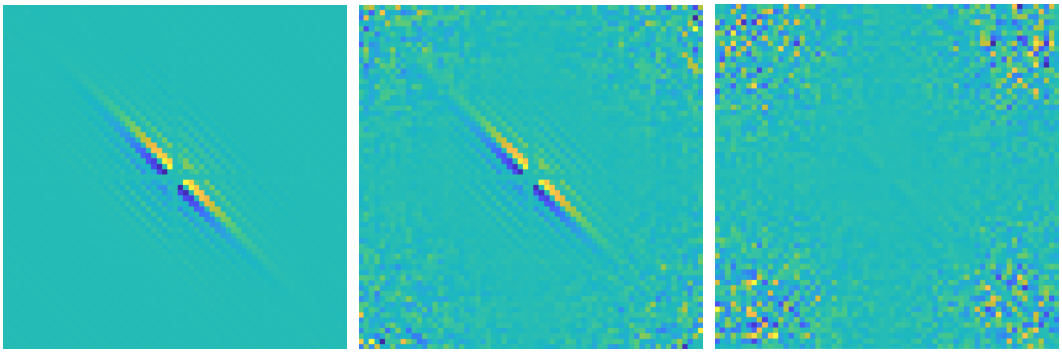
Fundamental Principles of Linear Operator Learning

(i) Smoothing operators are easier to learn than nonsmoothing ones.

(ii) Rougher training input functions improve data efficiency.

(iii) Test error improves whenever the input test distribution is supported on smoother input functions.

These principles have interesting implications. For instance, Item (i) suggests that compact operators that smooth input functions—such as integration operators—require less data to accurately learn than do unbounded operators that amplify perturbations in input functions—such as differential operators. Figure 1.1 illustrates the effect of Item (ii) in the context of a data completion problem from electrical impedance tomography (Subsection 4.1.3). It shows that training a linear estimator on input functions with a lot of high-frequency content (equivalently, rougher functions with less regularity) is more accurate



(a) True linear operator (b) Estimate from rough data (c) Estimate from smooth data

Figure 1.1: Bayesian linear regression of the imaginary part of the Dirichlet-to-Neumann map in Fourier basis coordinates arising from electrical impedance tomography on the disk (Subsection 4.1.3). The true linear operator is shown in Figure 1.1a. The reconstruction from noisy input-output data with a Bayesian posterior mean estimator is more accurate when the training input functions are rough (Figure 1.1b) than when the input functions are smooth (Figure 1.1c).

than training it on input functions that are too smooth. This is a statement about the quality of the continuum training data. Similar insights are also revealed in the work of Boullé and Townsend [42], albeit in a less explicit form. Moreover, the fact that test error improves with smoother covariate shifts, as stated in Item (iii), suggests that claims about the generalization accuracy of operator learning algorithms should always be qualified with the properties of the test distribution. Altogether, these insights provide useful practical guidance to users of operator learning, especially in data-scarce regimes or scenarios where the data generation procedure is controlled by the user. They further give qualitative theoretical validity to equation discovery methods [45]. The principles, verified in the linear case by the theory and numerics in Chapter 4, have also been observed in nonlinear problems [75, 282].

The proof approach in Chapter 4 relies on a reduction to an infinite sequence statistical regression model under white noise. This reduction is accomplished by assuming prior knowledge of the unknown operator's eigenvectors, so that only its eigenvalue sequence must be estimated. Thus, the use of problem structure here reduces the complexity of the inference task. The chapter's corresponding Bayesian inverse problems-style analysis of the random sequence space model is likely of independent interest to the nonparametric statistics community. Working with Gaussian process priors, the chapter proves convergence rates in the infinite data limit for the Bayesian posterior eigenvalue estimator under

multiple data distribution assumptions, as well as individual lower bounds. Figure 4.3 and Corollary 4.18 verify the sharpness of the results both numerically and theoretically, respectively. The sharp analysis also allows for a precise characterization of smoothness misspecification with respect to the Gaussian prior. Indeed, Figures 4.1c and 4.3 reveal that undersmoothing priors may lead to estimators that do not converge at all, while oversmoothing priors always lead to convergence, albeit at a possibly arbitrarily slow rate.

In addition to the main statistical guarantees from Section 4.3, the key technical contributions of Chapter 4 also include new upper and lower bounds on the generalization gap between test and train errors (leading to slow rates that appear to be optimal; see Theorem 4.24 and Figure 4.3) and probabilistic techniques to control infinite sums of dependent subexponential random variables (Lemmas C.5 and C.6) and ways to lower bound such sums (see the proofs of Theorems 4.17 and 4.24). These results may be of independent interest as well.

1.3.4 Operator Learning for Parameter-to-Observable Maps

Available as the preprint *arXiv:2402.06031 cs.LG* (2024), Chapter 5 of this thesis addresses the challenging setting in which the underlying continuum operator might only be accessible from finite-dimensional and possibly indirect quantity of interest measurements of its output. To facilitate a precise theoretical development, the chapter formulates this problem as the data-driven estimation of a factorized linear functional

$$f = q \circ L\,, \tag{1.5}$$

where q is another linear functional that represents a scalar quantity of interest and L is a linear operator between Hilbert spaces. The objective is to determine the best training procedure for estimating f. To this end, the analysis in the chapter establishes sample complexity bounds for two different Bayesian estimators of f, each corresponding to a distinct training data access model.

The first estimator, end-to-end (EE), is based on direct supervised learning of f itself. Here, the training data consist of input functions and noisy *scalar-valued labels* corresponding to f evaluated at these input functions. This setting is natural whenever the observed data is acquired from real experiments in a laboratory, for example. The second estimator, full-field (FF), is based on Bayesian linear regression of the operator L as in Chapter 4 and use of the compositional structure of the target functional f (1.5). This method also

requires prior knowledge of the action of q on any new input function. Here, the training data consist of input functions and noisy *function-valued labels* corresponding to L evaluated at the inputs. Once an auxiliary estimator $\widehat{L}$ for L is built from these data, the final (FF) plug-in estimator for f is given by the composition $q \circ \widehat{L}$. The required full-field continuum training dataset may be difficult (or even impossible) to acquire for some problems. However, once the auxiliary estimator $\widehat{L}$ is obtained, it is much more versatile than the (EE) estimator because it can be used to predict quantities of interest different from q or be deployed in other downstream tasks.

The main theoretical result of the chapter (Corollary 5.15) shows that the regularity of the quantity of interest map q determines the regimes in which one of the two estimators has a statistical advantage over the other with respect to sample complexity. This insight is summarized by the next display.

Use of Domain Knowledge Can Be Statistically Beneficial

If the quantity of interest map q is sufficiently smooth, then the (FF) plug-in estimator of f based on continuum data and domain-specific knowledge is more data-efficient than the purely data-driven (EE) estimator.

This result suggests that prior domain-specific knowledge of problem structure (i.e., the compositional form of f and the exact form of q) can quantitatively improve the data efficiency of operator learning algorithms. If the map q is instead unbounded with regularity below a certain threshold, then the opposite conclusion holds—end-to-end learning is statistically advantageous. How much of a statistical advantage one estimator has over the other intricately depends on the regularity properties of the full operator L and the data distributions (Figure 5.2). Subsection 5.4.3.2 provides more detailed discussion.

To obtain the preceding results, the chapter develops new theoretical analysis of Bayesian nonparametric regression of linear functionals that may be of independent interest. The main novelty is the bias error bound in Proposition D.12, which dominates the total test error and relies on a delicate conditioning argument along with clever matrix perturbation identities. This bound is especially novel because it accommodates smoothness misspecification from the Gaussian prior. Results in related work are not as robust and typically require some notion of well specification to hold. Due to the kernel trick, the main (EE)

result for learning *linear* functionals, Theorem D.1, implies new convergence rates for Gaussian process regression of scalar-valued *nonlinear* functions as a special case. The other technical contributions of the chapter involve building upon the stability results for infinite sums of dependent subexponential random variables from Chapter 4 and using these in the more challenging setting of linear functional regression, which involves non-commuting and non-diagonal empirical covariance operators.

The other major contributions of Chapter 5 revolve around the design and implementation of practical operator learning architectures that have a finite-dimensional Euclidean input space, output space, or both. Existing neural operators only accommodate function-to-function maps. To go beyond this limitation, the chapter proposes linear functional and linear decoder layers that map functions to vectors and vectors to functions, respectively. Appending these new layers near the beginning or end of standard neural operators leads to expressive nonlinear *neural mappings* and *Fourier Neural Mappings* that can learn function-to-vector, vector-to-function, and vector-to-vector *parameter-to-observable maps* in a unified way while provably preserving universal approximation properties. These architectures are especially well suited for applications that involve a finite number of input parameters or observed outputs, such as design optimization or inverse problems, respectively. Composite linear functionals of the form $q \circ L$ arising from the chapter's main theoretical contributions are a special case of the general nonlinear function-to-vector setting here. Three numerical experiments involving environmental science, aerodynamics, and materials modeling applications demonstrate that the new Fourier Neural Mappings architectures qualitatively validate intuition from the linear theory and empirically outperform standard finite-dimensional neural networks that do not access any continuum data or continuum problem structure.

1.4 Thesis Outline

The remainder of this thesis is organized as follows. Chapter 2 proposes the function-valued random features method and demonstrates its performance on two parametric partial differential equation benchmark problems. A theoretical analysis develops statistical guarantees for the function-valued random features method in Chapter 3. The results include the strong consistency of the methodology as well as convergence rates. Chapter 4 shifts the focus of the thesis to linear problems. It develops the fundamental principles of learning

linear operators from noisy data. Chapter 5 goes further by additionally studying quantity of interest functionals composed with linear operators and statistical tradeoffs that arise in this more realistic setting. This chapter also develops and implements new universal neural operator architectures that can accommodate finite-dimensional vector inputs or outputs, or both, which greatly expands the applicability of operator learning. The thesis concludes in Chapter 6 with a research summary and an outlook toward future developments.

The chapters in this thesis are adapted from research papers that target several different audiences across applied and computational mathematics, statistics, machine learning, and engineering. Thus, each chapter is self-contained and equipped with its own notation.

Chapter 2

OPERATOR LEARNING WITH FUNCTION-VALUED RANDOM FEATURES

This chapter is adapted from the following publications:

[1] Nicholas H. Nelsen and Andrew M. Stuart. "Operator learning using random features: a tool for scientific computing". *SIAM Review,* accepted for SIGEST section (2024).

[2] Nicholas H. Nelsen and Andrew M. Stuart. "The random feature model for input-output maps between Banach spaces". *SIAM Journal on Scientific Computing* 43.5 (2021), A3212–A3243. DOI: 10.1137/20M133957X.

Operator learning is the subject that centers on the data-driven approximation of maps between infinite-dimensional spaces. It is emerging as a powerful tool to complement traditional scientific computing, which is often concerned with operators mapping between spaces of functions. Building on classical random features for scalar-valued regression, this chapter introduces the function-valued random features method as an operator learning architecture that is practical for nonlinear problems yet is structured enough to facilitate efficient training. At its core, the proposed approach builds a linear combination of random operators. This turns out to be a low-rank approximation of an operator-valued kernel ridge regression algorithm, and hence the method also has strong connections to Gaussian process regression. The chapter designs function-valued random features that are tailored to the structure of two nonlinear operator learning benchmark problems arising from parametric partial differential equations. Numerical results demonstrate the scalability, discretization invariance, and transferability of the function-valued random features method.

2.1 Introduction

The *random feature model*, an architecture for the data-driven approximation of maps between finite-dimensional spaces, was formalized in [221, 222, 223], building on earlier precursors in [21, 200, 271]. The goal of this chapter is to extend the random feature model to a methodology for the data-driven approximation of maps between infinite-dimensional spaces. Canonical examples

of such maps include the semigroup generated by a time-dependent partial differential equation (PDE) mapping the initial condition (an input parameter) to the solution at a later time and the operator mapping a coefficient function (an input parameter) appearing in a PDE to its solution. Obtaining efficient and potentially low-dimensional representations of PDE solution maps is not only conceptually interesting, but also practically useful. Many applications in science and engineering require repeated evaluations of a complex and expensive forward model for different configurations of a system parameter. The model often represents a discretized PDE and the parameter, serving as input to the model, often represents a high-dimensional discretized quantity such as an initial condition or uncertain coefficient field. These *outer loop* applications commonly arise in inverse problems or uncertainty quantification tasks that involve control, optimization, or inference [218]. Full order forward models do not perform well in such many-query contexts, either due to excessive computational cost (requiring the most powerful high performance computing architectures) or slow evaluation time (unacceptable in real-time contexts such as on-the-fly optimal control). In contrast to that of the *big data* regime that dominates computer vision and other technological fields, only a relatively small amount of high-resolution data can be generated from computer simulations or physical experiments in scientific applications. Fast approximate solvers built from this limited available data that can efficiently and accurately emulate the full order model would be highly advantageous.

In this work, we demonstrate that the random feature model holds considerable potential for such a purpose. Resembling [181, 274] and the contemporaneous work in [34, 150, 173, 207], we present a methodology for true function space learning of black-box input-output maps between a Banach space and separable Hilbert space. We formulate the approximation problem as supervised learning in infinite dimensions and show that the natural hypothesis space is a reproducing kernel Hilbert space associated with an operator-valued kernel. For a suitable loss functional, training the random feature model is equivalent to solving a finite-dimensional convex optimization problem. As a consequence of our careful construction of the method as mapping between Banach spaces, the resulting emulator naturally scales favorably with respect to the high input and output dimensions arising in practical, discretized applications; furthermore, it is shown to achieve small relative test error for two model problems arising from approximation of a semigroup and of the solution map corresponding to

an elliptic PDE exhibiting parametric dependence on a coefficient function.

2.1.1 Related Work

In recent years, two different lines of research have emerged that address PDE approximation problems with machine learning techniques. The first perspective takes a more traditional approach akin to point collocation methods from the field of numerical analysis. Here, the goal is to use a deep neural network (NN) to solve a prescribed initial boundary value problem with as high accuracy as possible. Given a point cloud in a spatio-temporal domain $\mathcal{D}$ as input data, the prevailing approach first directly parametrizes the PDE solution field as a NN and then optimizes the NN parameters by minimizing the PDE residual with respect to (w.r.t.) some loss functional (see [225, 245, 92] and the references therein). To clarify, the object approximated with this novel method is a *low-dimensional* input-output map $\mathcal{D} \to \mathbb{R}$, i.e., the real-valued function that solves the PDE. This approach is mesh-free by definition but highly intrusive as it requires full knowledge of the specified PDE. Any change to the original formulation of the initial boundary value problem or related PDE problem parameters necessitates an (expensive) re-training of the NN solution. We do not explore this first approach any further in this chapter.

The second direction is arguably more ambitious: use a NN as an emulator for the infinite-dimensional mapping between an input parameter and the PDE solution itself or a functional of the solution, i.e., a quantity of interest; the latter is widely prevalent in uncertainty quantification problems. We emphasize that the object approximated in this setting, unlike in the aforementioned first approach, is an input-output map $\mathcal{X} \to \mathcal{Y}$, i.e., the PDE solution operator, where $\mathcal{X}$ and $\mathcal{Y}$ are infinite-dimensional Banach spaces; this map is generally nonlinear. For an approximation-theoretic treatment of parametric PDEs in general, we refer the reader to the article of Cohen and DeVore [63]. In applications, the solution operator is represented by a discretized forward model $\mathbb{R}^K \to \mathbb{R}^K$, where K is the mesh size, and hence represents a *high-dimensional* object. It is this second line of research that inspires our work.

Of course, there are many approaches to forward model reduction that do not explicitly involve machine learning ideas. The reduced basis method (see [20, 29, 82] and the references therein) is a classical idea based on constructing an empirical basis from data snapshots and solving a cheaper variational problem;

it is still widely used in practice due to computationally efficient offline-online decompositions that eliminate dependence on the full order degrees of freedom. Recently, machine learning extensions to the reduced basis methodology, of both intrusive (e.g., projection-based reduced order models) and nonintrusive (e.g., model-free data only) type, have further improved the applicability of these methods [61, 103, 126, 165, 72]. However, the input-output maps considered in these works involve high dimension in only one of the input or output space, not both. Other popular surrogate modeling techniques include Gaussian processes [228], polynomial chaos expansions [248], and radial basis functions [270]; yet, these are only practically suitable for problems with input space of low to moderate dimension. Classical numerical methods for PDEs may also represent the forward model $\mathbb{R}^K \to \mathbb{R}^K$, albeit implicitly in the form a computer code (e.g., finite element, finite difference, finite volume methods). However, the approximation error is sensitive to K and repeated evaluations of this forward model often becomes cost prohibitive due to poor scaling with input dimension K.

Instead, deep NNs have been identified as strong candidate surrogate models for parametric PDE problems due to their empirical ability to emulate high-dimensional nonlinear functions with minimal evaluation cost once trained. Early work in the use of NNs to learn the solution operator, or vector field, defining ODEs and time-dependent PDEs, may be found in the 1990s [59, 115, 233]. There are now more theoretical justifications for NNs breaking the *curse of dimensionality* [150, 157, 91], leading to increased interest in PDE applications [4, 105, 211, 241]. A suite of work on data-driven discretizations of PDEs has surfaced that allow for identification of the governing model [19, 36, 179, 214, 252, 264]; however, we note that only the operators appearing in the equation itself are approximated with these approaches, not the solution operator of the PDE. More in line with our focus in this chapter, architectures based on deep convolutional NNs have proven quite successful for learning elliptic PDE solution maps (for example, see [265, 272, 283], which take an image-to-image regression approach). Other NNs have been used in similar elliptic problems for quantity of interest prediction [140], error estimation [58], or unsupervised learning [168]. Yet in all the approaches above, the architectures and resulting error are dependent on the mesh resolution. To circumvent this issue, the surrogate map must be well-defined on function space and independent of any finite-dimensional realization of the map that arises from discretization.

This is not a new idea (see [59, 234] or for functional data analysis, [136, 192]). The aforementioned reduced basis method is an example, as is the method of Chkifa, Cohen, DeVore, and Schwab [62] and Cohen and DeVore [63], which approximates the solution map with sparse Taylor polynomials and is proved to achieve optimal convergence rates in idealized settings. However, it is only recently that machine learning methods have been explicitly designed to operate in an infinite-dimensional setting, and there is little work in this direction [34, 173]. Here we propose the function-valued random features method as another such model.

The random feature model (RFM) [221, 222, 223], detailed in Section 2.2.3, is in some sense the simplest possible machine learning model; it may be viewed as an ensemble average of randomly parametrized functions: an expansion in a randomized basis. These *random features* could be defined, for example, by randomizing the internal parameters of a NN. Compared to NN emulators with enormous learnable parameter counts (e.g., $O(10^5)$ to $O(10^7)$, see [94, 96, 168]) and methods that are intrusive or lead to nontrivial implementations [62, 165, 72], the RFM is one of the simplest models to formulate and train (often $O(10^4)$ parameters, or fewer, suffice). The theory of the RFM for real-valued outputs is well developed, partly due to its close connection to kernel methods [17, 51, 134, 221, 270] and Gaussian processes [200, 271], and includes generalization rates and dimension-free estimates [91, 222, 255]. A quadrature viewpoint on the RFM provides further insight and leads to Monte Carlo sampling ideas [17]; we remark on this further in Section 2.2.3. As in modern deep learning practice, the RFM has also been shown to perform best when the model is over-parametrized [27]. In a similar high-dimensional setting of relevance in this chapter, Griebel and Rieger [117] and Kempf, Wendland, and Rieger [139] theoretically investigated nonparametric kernel regression for parametric PDEs with real-valued solution map outputs. The specific random Fourier feature approach of Rahimi and Recht [221] was generalized by Brault, Heinonen, and Buc [43] to the finite-dimensional matrix-valued kernel setting with vector-valued random Fourier features. However, most of these works require explicit knowledge of the kernel itself. Here our viewpoint is to work directly with random features as the basis for a standalone method, choosing them for their properties and noting that they implicitly define a kernel, but not working directly with this kernel; furthermore, our work considers both infinite-dimensional input *and* output spaces, not just one or the other.

A key idea underlying our approach is to formulate the proposed random features algorithm on infinite-dimensional space and only then discretize. This philosophy in algorithm development has been instructive in a number of areas in scientific computing, such as optimization [127] and the development of Markov chain Monte Carlo methodology [68]. It has recently been promoted as a way of designing and analyzing algorithms within machine learning [120, 89, 236, 87, 88], and our work may be understood within this general framework.

2.1.2 Contributions

Our primary contributions in this chapter are now listed.

1. We develop the random feature model, directly formulated on the function space level, for learning input-output maps between Banach spaces purely from data. As a method for parametric PDEs, the methodology is nonintrusive but also has the additional advantage that it may be used in settings where only data is available and no model is known.

2. We show that our proposed method is more computationally tractable to both train and evaluate than standard kernel methods in infinite dimensions. Furthermore, we show that the method is equivalent to operator-valued kernel ridge regression performed in a finite-dimensional space spanned by random features.

3. We apply our methodology to learn the semigroup defined by the solution operator for viscous Burgers' equation and the coefficient-to-solution operator for the Darcy flow equation.

4. We demonstrate in several numerical experiments two mesh-independent approximation properties that are built into the proposed methodology: invariance of relative error to mesh resolution and evaluation ability on any mesh resolution.

The remainder of this chapter is structured as follows. In Section 2.2, we communicate the mathematical framework required to work with the random feature model in infinite dimensions, identify an appropriate approximation space, and explain the training procedure. We introduce two instantiations of random feature maps that target physical science applications in Section 2.3 and

detail the corresponding numerical results for these applications in Section 2.4. We conclude in Section 2.5 with discussion and future work.

2.2 Methodology

In this work, the overarching problem of interest is the approximation of a map $F^\dagger \colon \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are infinite-dimensional spaces of real-valued functions defined on some bounded open subset of $\mathbb{R}^d$, and $F^\dagger$ is defined by $a \mapsto F^\dagger(a) \coloneqq u$, where u is the solution of a (possibly time-dependent) PDE and a is an input function required to make the problem well-posed. Our proposed approach for this approximation, constructing a surrogate map F for the true map $F^\dagger$, is data-driven, nonintrusive, and based on least squares. Least-squares-based methods are integral to the random feature methodology as proposed in low dimensions [221, 222] and generalized here to the infinite-dimensional setting; they have also been shown to work well in other algorithms for high-dimensional numerical approximation [33, 65, 83]. Within the broader scope of reduced order modeling techniques [29], the approach we adopt in this chapter falls within the class of data-fit emulators. In its essence, our method interpolates the solution manifold

$$\mathcal{M} = \{u \in \mathcal{Y} \colon u = F^\dagger(a) \quad \text{and} \quad a \in \mathcal{X}\}\,. \tag{2.1}$$

The solution map $F^\dagger$, as the inverse of a differential operator, is often smoothing and admits a notion of compactness, i.e., the output space compactly embeds into the input space. Then, the idea is that $\mathcal{M}$ should have some compact, low-dimensional structure (intrinsic dimension). However, actually finding a model F that exploits this structure despite the high dimensionality of the truth map $F^\dagger$ is quite difficult. Further, the effectiveness of many model reduction techniques, such as those based on the reduced basis method, are dependent on inherent properties of the map $F^\dagger$ itself (e.g., analyticity), which in turn may influence the decay rate of the Kolmogorov width of the manifold $\mathcal{M}$ [63]. While such subtleties of approximation theory are crucial to developing rigorous theory and provably convergent algorithms, we choose to work in the nonintrusive setting where knowledge of the map $F^\dagger$ and its associated PDE are only obtained through measurement data, and hence detailed characterizations such as those aforementioned are essentially unavailable. Thus, we emphasize that our proposed operator learning methodology is applicable to general continuum problems with function space data, not just to PDEs.

The remainder of this section introduces the mathematical preliminaries for our methodology. With the goal of operator approximation in mind, in Section 2.2.1 we formulate a supervised learning problem in an infinite-dimensional setting. We provide the necessary background on reproducing kernel Hilbert spaces in Section 2.2.2 and then define the RFM in Section 2.2.3. In Section 2.2.4, we describe the optimization principle which leads to algorithms for the RFM and an example problem in which $\mathcal{X}$ and $\mathcal{Y}$ are one-dimensional vector spaces.

2.2.1 Problem Formulation

Let $\mathcal{X}$ and $\mathcal{Y}$ be real Banach spaces and $F^\dagger \colon \mathcal{X} \to \mathcal{Y}$ be a (possibly nonlinear) map. It is natural to frame the approximation of $F^\dagger$ as a supervised learning problem. Suppose we are given training data in the form of input-output pairs $\{(a_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$, where $a_i \sim \nu$ i.i.d., ν is a probability measure supported on $\mathcal{X}$, and $y_i = F^\dagger(a_i) \sim F^\dagger_\sharp \nu$ with, potentially, noise added to the evaluations of $F^\dagger(\cdot)$. In the examples in this chapter, the noise is viewed as resulting from model error (the PDE does not perfectly represent the physics) or from discretization error (in approximating the PDE); situations in which the data acquisition process is inherently noisy can also be envisioned but are not studied here. We aim to build a parametric reconstruction of the true map $F^\dagger$ from the data, that is, construct a model $F \colon \mathcal{X} \times \mathcal{P} \to \mathcal{Y}$ and find $\alpha^\dagger \in \mathcal{P} \subseteq \mathbb{R}^m$ such that $F(\cdot, \alpha^\dagger) \approx F^\dagger$ are close as maps from $\mathcal{X}$ to $\mathcal{Y}$ in some suitable sense. The natural number m here denotes the total number of model parameters. The standard approach to determine parameters in supervised learning is to first define a loss functional $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ and then minimize the expected risk,

$$\min_{\alpha \in \mathcal{P}} \mathbb{E}^{a \sim \nu} \left[\ell\big(F^\dagger(a), F(a, \alpha)\big) \right]. \tag{2.2}$$

With only the data $\{(a_i, y_i)\}_{i=1}^n$ at our disposal, we approximate problem (2.2) by replacing ν with the empirical measure $\nu^{(n)} := \frac{1}{n} \sum_{j=1}^n \delta_{a_j}$, which leads to the empirical risk minimization problem

$$\min_{\alpha \in \mathcal{P}} \frac{1}{n} \sum_{j=1}^n \ell\big(y_j, F(a_j, \alpha)\big). \tag{2.3}$$

The hope is that given minimizer $\alpha^{(n)}$ of (2.3) and $\alpha^\dagger$ of (2.2), $F(\cdot, \alpha^{(n)})$ well approximates $F(\cdot, \alpha^\dagger)$, that is, the learned model *generalizes* well; these ideas may be made rigorous with results from statistical learning theory [123]. Solving the problem (2.3) is called *training* the model F. Once trained, the model is

then validated on a new set of i.i.d. input-output pairs previously unseen during the training process. This *testing* phase indicates how well F approximates $F^\dagger$. From here on out, we assume that $(\mathcal{Y}, \langle\cdot,\cdot\rangle_\mathcal{Y}, \|\cdot\|_\mathcal{Y})$ is a real separable Hilbert space and focus on the squared loss

$$\ell(y, y') \coloneqq \frac{1}{2}\|y - y'\|_\mathcal{Y}^2 . \tag{2.4}$$

We stress that our entire formulation is in an infinite-dimensional setting and we will remain in this setting throughout the chapter; as such, the random feature methodology we propose will inherit desirable discretization-invariant properties, to be observed in the numerical experiments of Section 2.4.

Notation 2.1 (expectation)**.** For a Borel measurable map $G\colon \mathcal{U} \to \mathcal{V}$ between two Banach spaces $\mathcal{U}$ and $\mathcal{V}$ and a probability measure π supported on $\mathcal{U}$, we denote the expectation of G under π by

$$\mathbb{E}^{u\sim\pi}\big[G(u)\big] = \int_\mathcal{U} G(u)\pi(du) \in \mathcal{V} \tag{2.5}$$

in the sense of Bochner integration [74, Section A.2]. We will drop the domain of integration in situations where no confusion is caused by doing so.

2.2.2 Operator-Valued Reproducing Kernels

The RFM is naturally formulated in a reproducing kernel Hilbert space (RKHS) setting, as our exposition will demonstrate in Section 2.2.3. However, the usual RKHS theory is concerned with real-valued functions [14, 31, 70, 270]. Our setting, with the output space $\mathcal{Y}$ a separable Hilbert space, requires several ideas that generalize the real-valued case. We now outline these ideas with a review of operator-valued kernels; parts of the presentation that follow may be found in the references [17, 55, 192, 201].

We first consider the special case $\mathcal{Y} \coloneqq \mathbb{R}$ for ease of exposition. A real RKHS is a Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_\mathcal{H}, \|\cdot\|_\mathcal{H})$ comprised of real-valued functions $f\colon \mathcal{X} \to \mathbb{R}$ such that the pointwise evaluation functional $f \mapsto f(a)$ is bounded for every $a \in \mathcal{X}$. It then follows that there exists a unique, symmetric, positive definite kernel function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for every $a \in \mathcal{X}$, $k(\cdot, a) \in \mathcal{H}$ and the *reproducing kernel property* $f(a) = \langle k(\cdot, a), f\rangle_\mathcal{H}$ holds. These two properties are often taken as the definition of a RKHS. The converse direction is also true: every symmetric, positive definite kernel defines a unique RKHS [14].

We now introduce the needed generalization of the reproducing property to the case of arbitrary real Hilbert spaces $\mathcal{Y}$, as this result will motivate the construction of the RFM. Kernels in this setting are now operator-valued.

Definition 2.2 (operator-valued kernel)**.** Let $\mathcal{X}$ be a real Banach space and $\mathcal{Y}$ a real separable Hilbert space. An *operator-valued kernel* is a map

$$k\colon \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})\,, \tag{2.6}$$

where $\mathcal{L}(\mathcal{Y})$ denotes the Banach space of all bounded linear operators on $\mathcal{Y}$, such that its adjoint satisfies $k(a,a')^* = k(a',a)$ for all a, $a' \in \mathcal{X}$ and for every $N \in \mathbb{N}$,

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\langle y_i, k(a_i,a_j)y_j\rangle_{\mathcal{Y}} \geq 0 \tag{2.7}$$

for all pairs $\{(a_i,y_i)\}_{i=1}^{N} \subset \mathcal{X}\times\mathcal{Y}$.

Paralleling the development for the real-valued case, an operator-valued kernel k also uniquely (up to isomorphism) determines an associated real RKHS $\mathcal{H}_k = \mathcal{H}_k(\mathcal{X};\mathcal{Y})$. Now, choosing a probability measure ν supported on $\mathcal{X}$, we define a kernel integral operator T_k associated to k by

$$\begin{aligned} T_k\colon L^2_\nu(\mathcal{X};\mathcal{Y}) &\to L^2_\nu(\mathcal{X};\mathcal{Y}) \\ F &\mapsto T_kF \coloneqq \int k(\cdot,a')F(a')\nu(da')\,, \end{aligned} \tag{2.8}$$

which is nonnegative, self-adjoint, and compact (provided $k(a,a) \in \mathcal{L}(\mathcal{Y})$ is compact for all $a \in \mathcal{X}$ [55]). Let us further assume that all conditions needed for $T_k^{1/2}$ to be an isometry from L^2_ν into $\mathcal{H}_k$ are satisfied, i.e., $\mathcal{H}_k = \mathrm{Im}(T_k^{1/2})$. Generalizing the standard Mercer theory (see, e.g., [17, 31]), we may write the RKHS inner product as

$$\langle F,G\rangle_{\mathcal{H}_k} = \langle F, T_k^{-1}G\rangle_{L^2_\nu} \quad \text{for all} \quad F,G \in \mathcal{H}_k\,. \tag{2.9}$$

Note that while (2.9) appears to depend on the measure ν on $\mathcal{X}$, the RKHS $\mathcal{H}_k$ is itself determined by the kernel without any reference to a measure [70, Chapter 3, Theorem 4]. With the inner product now explicit, we may directly deduce a reproducing property. A fully rigorous justification of the methodology is outside the scope of this chapter; however, we perform formal computations

which provide intuition underpinning the methodology. To this end we fix $a \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then

$$\begin{aligned}\langle k(\cdot, a)y, T_k^{-1}F\rangle_{L^2_\nu} &= \int \left\langle k(a', a)y, (T_k^{-1}F)(a')\right\rangle_{\mathcal{Y}} \nu(da') \\ &= \int \left\langle y, k(a, a')(T_k^{-1}F)(a')\right\rangle_{\mathcal{Y}} \nu(da') \\ &= \left\langle y, \int k(a, a')(T_k^{-1}F)(a')\, \nu(da')\right\rangle_{\mathcal{Y}} \\ &= \langle y, F(a)\rangle_{\mathcal{Y}}\,,\end{aligned}$$

by using Definition 2.2 of operator-valued kernel and the fact that $k(\cdot, a)y \in \mathcal{H}_k$ [55]. So, we deduce the following.

Result 2.3 (reproducing property for operator-valued kernels)**.** *Let $F \in \mathcal{H}_k$ be given. Then for every $a \in \mathcal{X}$ and $y \in \mathcal{Y}$, it holds that*

$$\langle y, F(a)\rangle_{\mathcal{Y}} = \langle k(\cdot, a)y, F\rangle_{\mathcal{H}_k}\,. \tag{2.10}$$

This identity, paired with a special choice of k, is the basis of the RFM in our abstract infinite-dimensional setting.

2.2.3 Random Feature Model

One could approach the approximation of target map $F^\dagger\colon \mathcal{X} \to \mathcal{Y}$ from the perspective of kernel methods. However, it is generally a difficult task to explicitly design operator-valued kernels of the form (2.6) since the spaces $\mathcal{X}$ and $\mathcal{Y}$ may be of different regularity, for example. Example constructions of operator-valued kernels studied in the literature include those taking value as diagonal operators, multiplication operators, or composition operators [136, 192], but these all involve some simple generalization of scalar-valued kernels. Instead, the RFM allows one to implicitly work with operator-valued kernels through the use of a *random feature map* $\varphi\colon \mathcal{X} \times \Theta \to \mathcal{Y}$ and a probability measure μ supported on Banach space Θ. The map φ is assumed to be square integrable w.r.t. the product measure $\nu \times \mu$, i.e., $\varphi \in L^2_{\nu\times\mu}(\mathcal{X} \times \Theta; \mathcal{Y})$, where ν is the (sometimes a modeling choice at our discretion, sometimes unknown) data distribution on $\mathcal{X}$. Together, (φ, μ) form a *random feature pair*. With this setup in place, we now describe the connection between random features and kernels; to this end, recall the following standard notation.

Notation 2.4 (outer product)**.** Given a Hilbert space $(H, \langle\cdot,\cdot\rangle, \|\cdot\|)$, the *outer product* $a \otimes b \in \mathcal{L}(H, H)$ is defined by $(a \otimes b)c = \langle b, c\rangle a$ for any $a, b, c \in H$.

Given the pair (φ, μ), consider maps $k_\mu \colon \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ of the form

$$k_\mu(a, a') \coloneqq \int \varphi(a;\theta) \otimes \varphi(a';\theta)\mu(d\theta)\,. \tag{2.11}$$

Such representations need not be unique; different pairs (φ, μ) may induce the same kernel $k = k_\mu$ in (2.11). Since k_μ may readily be shown to be an operator-valued kernel via definition 2.2, it defines a unique real RKHS $\mathcal{H}_{k_\mu} \subset L^2_\nu(\mathcal{X};\mathcal{Y})$. Our approximation theory will be based on this space or finite-dimensional approximations thereof. We now perform a purely formal but instructive calculation, following from application of the reproducing property (2.10) to operator-valued kernels of the form (2.11). Doing so leads to an integral representation of any $F \in \mathcal{H}_{k_\mu}$: for all $a \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$\begin{aligned}
\langle y, F(a)\rangle_{\mathcal{Y}} = \langle k_\mu(\cdot, a)y, F\rangle_{\mathcal{H}_{k_\mu}} &= \left\langle \int \langle \varphi(a;\theta), y\rangle_{\mathcal{Y}}\, \varphi(\cdot;\theta)\, \mu(d\theta), F \right\rangle_{\mathcal{H}_{k_\mu}} \\
&= \int \langle \varphi(a;\theta), y\rangle_{\mathcal{Y}} \langle \varphi(\cdot;\theta), F\rangle_{\mathcal{H}_{k_\mu}} \mu(d\theta) \\
&= \int c_F(\theta) \langle y, \varphi(a;\theta)\rangle_{\mathcal{Y}}\, \mu(d\theta) \\
&= \left\langle y, \int c_F(\theta)\varphi(a;\theta)\mu(d\theta) \right\rangle_{\mathcal{Y}},
\end{aligned}$$

where the coefficient function $c_F \colon \Theta \to \mathbb{R}$ is defined by

$$c_F(\theta) \coloneqq \langle \varphi(\cdot;\theta), F\rangle_{\mathcal{H}_{k_\mu}}\,. \tag{2.12}$$

Since $\mathcal{Y}$ is a Hilbert space, the above holding for all $y \in \mathcal{Y}$ implies the integral representation

$$F = \int c_F(\theta)\varphi(\cdot;\theta)\mu(d\theta)\,. \tag{2.13}$$

The formal expression (2.12) for $c_F(\theta)$ needs careful interpretation (provided in Appendix A.2). For instance, if $\varphi(\cdot;\theta)$ is a realization of a Gaussian process as in Example 2.9, then $\varphi(\cdot;\theta) \notin \mathcal{H}_{k_\mu}$ with probability one; indeed, in this case c_F is defined only as an $L^2_\mu(\Theta;\mathbb{R})$ limit. Nonetheless, the RKHS may be completely characterized by this integral representation. Define the map

$$\begin{aligned}
&\mathcal{A} \colon L^2_\mu(\Theta;\mathbb{R}) \to L^2_\nu(\mathcal{X};\mathcal{Y}) \\
&\qquad c \mapsto \mathcal{A}c \coloneqq \int c(\theta)\varphi(\cdot;\theta)\mu(d\theta)\,.
\end{aligned} \tag{2.14}$$

The map $\mathcal{A}$ may be shown to be a bounded linear operator that is a particular square root of T_{k_μ} (Appendix A.2). We have the following result whose proof, provided in Appendix A.1, is a straightforward generalization of the real-valued case given by Bach [17, Section 2.2].

Result 2.5 (infinite-dimensional RKHS)**.** *Under the assumption that* $\varphi \in L^2_{\nu\times\mu}(\mathcal{X}\times\Theta;\mathcal{Y})$*, the RKHS defined by the kernel* k_μ *in* (2.11) *is precisely*

$$\mathcal{H}_{k_\mu} = \mathrm{Im}(\mathcal{A}) = \left\{ \int c(\theta)\varphi(\cdot;\theta)\mu(d\theta) \colon c \in L^2_\mu(\Theta;\mathbb{R}) \right\}. \tag{2.15}$$

We stress that the integral representation of mappings in RKHS (2.15) is not unique since $\mathcal{A}$ is not injective in general. However, the particular choice $c = c_F$ (2.12) in representation (2.13) does enjoy a sense of uniqueness as described in Appendix A.2. In particular, the $L^2_\mu(\Theta;\mathbb{R})$ norm of c_F equals the $\mathcal{H}_{k_\mu}$ norm of F. The formula (2.15) suggests that $\mathcal{H}_{k_\mu}$, which is built from (φ,μ) and completely determined by coefficient functionals $c \in L^2_\mu(\Theta;\mathbb{R})$, is a natural nonparametric class of operators to perform approximation with. However, the actual implementation of estimators based on the model class $\mathcal{H}_{k_\mu}$ is known to incur an enormous computational cost without further assumptions on the structure of (φ,μ), as we discuss later in this section. Instead, we next adopt a parametric approximation to this full RKHS approach.

A central role in what follows is the approximation of measure μ by the empirical measure

$$\mu^{(m)} := \frac{1}{m}\sum_{j=1}^m \delta_{\theta_j}\,, \quad \text{where} \quad \theta_j \overset{\text{i.i.d.}}{\sim} \mu\,. \tag{2.16}$$

Given this, define $k^{(m)} := k_{\mu^{(m)}}$ to be the empirical approximation to k_μ:

$$k^{(m)}(a,a') = \mathbb{E}^{\theta\sim\mu^{(m)}}\left[\varphi(a;\theta)\otimes\varphi(a';\theta)\right] = \frac{1}{m}\sum_{j=1}^m \varphi(a;\theta_j)\otimes\varphi(a';\theta_j)\,. \tag{2.17}$$

Then we let $\mathcal{H}_{k^{(m)}}$ be the unique RKHS induced by the kernel $k^{(m)}$; note that $k^{(m)}$ and hence $\mathcal{H}_{k^{(m)}}$ are themselves random variables. The following characterization of $\mathcal{H}_{k^{(m)}}$ is proved in Appendix A.1.

Result 2.6 (finite-dimensional RKHS)**.** *Assume that* $\varphi \in L^2_{\nu\times\mu}(\mathcal{X}\times\Theta;\mathcal{Y})$ *and that the random features* $\{\varphi(\cdot;\theta_j)\}_{j=1}^m$ *are linearly independent in* $L^2_\nu(\mathcal{X};\mathcal{Y})$*. Then the RKHS* $\mathcal{H}_{k^{(m)}}$ *is equal to the linear span of* $\{\varphi_j := \varphi(\cdot;\theta_j)\}_{j=1}^m$*.*

Applying a simple Monte Carlo sampling approach to elements in RKHS (2.15) by replacing probability measure μ by empirical measure $\mu^{(m)}$ gives, for $c \in L^2_\mu$,

$$\frac{1}{m}\sum_{j=1}^m c(\theta_j)\varphi(\cdot;\theta_j) \approx \int c(\theta)\varphi(\cdot;\theta)\mu(d\theta)\,. \tag{2.18}$$

This approximation achieves the Monte Carlo rate $O(m^{-1/2})$ in expectation and, by virtue of Result 2.6, is in $\mathcal{H}_{k^{(m)}}$. However, in the setting of this work, the Monte Carlo approach does not give rise to a practical method for learning a target map $F^\dagger \in \mathcal{H}_{k_\mu}$ because $F^\dagger$, k_μ, and $\mathcal{H}_{k_\mu}$ are all unknown; only the random feature pair (φ, μ) is assumed to be given. Hence one cannot apply (2.12) (or (A.7)) to evaluate $c = c_{F^\dagger}$ in (2.18). Furthermore, in realistic settings it may be that $F^\dagger \notin \mathcal{H}_{k_\mu}$, which leads to an additional smoothness misspecification gap not accounted for by the Monte Carlo method. To sidestep these difficulties, the RFM adopts a data-driven optimization approach to determine a different approximation to $F^\dagger$, also from the space $\mathcal{H}_{k^{(m)}}$. We now define the RFM.

Definition 2.7 (RFM)**.** Given probability spaces $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \nu)$ and $(\Theta, \mathcal{B}(\Theta), \mu)$ with $\mathcal{X}$ and Θ being real finite- or infinite-dimensional Banach spaces, real separable Hilbert space $\mathcal{Y}$, and $\varphi \in L^2_{\nu\times\mu}(\mathcal{X} \times \Theta; \mathcal{Y})$, the *random feature model* is the parametric map

$$\begin{aligned} F_m &\colon \mathcal{X} \times \mathbb{R}^m \to \mathcal{Y} \\ &(a; \alpha) \mapsto F_m(a; \alpha) \coloneqq \frac{1}{m} \sum_{j=1}^{m} \alpha_j \varphi(a; \theta_j)\,, \quad \text{where} \quad \theta_j \overset{\text{i.i.d.}}{\sim} \mu\,. \end{aligned} \tag{2.19}$$

We use the Borel σ-algebras $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\Theta)$ to define the probability spaces in the preceding definition. Our goal with the RFM is to choose parameters $\alpha \in \mathbb{R}^m$ so as to approximate mappings $F^\dagger \in \mathcal{H}_{k_\mu}$ (in the ideal setting) by mappings $F_m(\cdot; \alpha) \in \mathcal{H}_{k^{(m)}}$. The RFM is itself a random variable and may be viewed as a *spectral method* since the randomized basis $\varphi(\cdot; \theta)$ in the linear expansion (2.19) is defined on all of $\mathcal{X}$ ν-a.e. Determining the coefficient vector α from data obviates the difficulties associated with the oracle Monte Carlo approach since the method only requires knowledge of the pair (φ, μ) and knowledge of sample input-output pairs from target operator $F^\dagger$.

As written, (2.19) is incredibly simple. The operator F_m is nonlinear in its input a but linear in its coefficient parameters α. In practice, the linearity w.r.t. the RFM parameters is broken by also learning *hyperparameters* that appear in the pair (φ, μ) [85]. Moreover, similar to operator learning architectures such as neural operators [154] and Fourier neural operators [172], the RFM is a *nonlinear approximation*. This means that the output $F_m(a; \alpha)$ of the RFM belongs to a nonlinear manifold in $\mathcal{Y}$ (cp. Equation 2.1) instead of a

fixed linear subspace of $\mathcal{Y}$. In contrast, methods such as PCA-Net [34] and DeepONet [181] are restricted to such fixed linear spaces, which may limit their approximation power for specific classes of problems. More theory is required to quantitatively separate these two classes of approximation methods.

Overall, it is clear that the choice of random feature map and measure pair (φ, μ) will determine the quality of approximation. Most papers deploying these methods, including [43, 221, 222], take a kernel-oriented perspective by first choosing a kernel and then finding a random feature map to estimate this kernel. Our perspective, more aligned with [223, 255], is the opposite in that we allow the choice of random feature map φ to implicitly *define* the kernel via the formula (2.11) instead of picking the kernel first. This methodology also has implications for numerics: the kernel never explicitly appears in any computations, which leads to memory and other cost savings. It does, however, leave open the question of characterizing the universality [255] of such kernels and the RKHS $\mathcal{H}_{k_\mu}$ of mappings from $\mathcal{X}$ to $\mathcal{Y}$ that underlies the approximation method; this is an important avenue for future work.

The close connection to kernels explains the origins of the RFM in the machine learning literature. Moreover, the RFM may also be interpreted in the context of neural networks [200, 255, 271]. To see this explicitly, consider the setting where $\mathcal{X}$ and $\mathcal{Y}$ are both equal to the Euclidean space $\mathbb{R}$ and choose φ to be a family of hidden neurons $\varphi_{\text{NN}}(a;\theta) \coloneqq \sigma(\theta^{(1)} \cdot a + \theta^{(2)})$. A single hidden layer NN would seek to find $\{(\alpha_j, \theta_j)\}_{j=1}^m$ in $\mathbb{R} \times \mathbb{R}^2$ so that

$$\frac{1}{m} \sum_{j=1}^{m} \alpha_j \varphi_{\text{NN}}(\cdot;\theta_j) \tag{2.20}$$

matches the given training data $\{(a_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$. More generally, and in arbitrary Euclidean spaces, one may allow $\varphi_{\text{NN}}(\cdot;\theta)$ to be any deep NN. However, while the RFM has the same *form* as (2.20), there is a difference in the *training*: the θ_j are drawn i.i.d. from a probability measure and then fixed, and only the α_j are chosen to fit the training data. This connection is quite profound: given any deep NN with randomly initialized parameters θ, studies of the lazy training regime and neural tangent kernel [51, 134] suggest that adopting a RFM approach and optimizing over only α is quite natural, as it is observed that in this regime the internal NN parameters do not stray far from their random initialization during gradient descent while the last layer of parameters $\{\alpha_j\}_{j=1}^m$ adapt considerably.

Once the feature parameters $\{\theta_j\}_{j=1}^m$ are chosen at random and fixed, training the RFM F_m only requires optimizing over $\alpha \in \mathbb{R}^m$ which, due to linearity of F_m in α, is a straightforward task to which we now turn our attention.

2.2.4 Optimization

One of the most attractive characteristics of the RFM is its training procedure. With the L^2-type loss (2.4) as in standard regression settings, optimizing the coefficients of the RFM with respect to the empirical risk (2.3) is a convex optimization problem, requiring only the solution of a finite-dimensional system of linear equations; the convexity also suggests the possibility of appending convex constraints (such as linear inequalities), although we do not pursue this here. Further, the kernels k_μ or $k^{(m)}$ are not required anywhere in the algorithm. We emphasize the simplicity of the underlying optimization tasks as they suggest the possibility of numerical implementation of the RFM into complicated black-box computer codes. This is in contrast with most other supervised operator learning methods, which are trained with variants of stochastic gradient descent. Such a training strategy leads to nonconvexity that is notoriously difficult to study both computationally and theoretically.

We now proceed to show that a regularized version of the optimization problem (2.3)–(2.4) arises naturally from approximation of a nonparametric regression problem defined over the RKHS $\mathcal{H}_{k_\mu}$. To this end, recall the supervised learning formulation in Section 2.2.1. Given n i.i.d. input-output pairs $\{(a_i, y_i = F^\dagger(a_i))\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ as data, with the a_i drawn from (possibly unknown) probability measure ν on $\mathcal{X}$, the objective is to find an approximation $\widehat{F}$ to the map $F^\dagger$. Let $\mathcal{H}_{k_\mu}$ be the hypothesis space and k_μ its operator-valued reproducing kernel of the form (2.11). The most straightforward learning algorithm in this RKHS setting is kernel ridge regression, also known as penalized least squares. This method produces a nonparametric model by finding a minimizer $\widehat{F}$ of

$$\min_{F \in \mathcal{H}_{k_\mu}} \left\{ \sum_{j=1}^n \frac{1}{2} \big\| y_j - F(a_j) \big\|_{\mathcal{Y}}^2 + \frac{\lambda}{2} \big\| F \big\|_{\mathcal{H}_{k_\mu}}^2 \right\}, \tag{2.21}$$

where $\lambda \geq 0$ is a penalty parameter. By the representer theorem for operator-valued kernels [192, Theorems 2 and 4], the minimizer has the form

$$\widehat{F} = \sum_{j=1}^n k_\mu(\cdot, a_j)\beta_j \tag{2.22}$$

for some functions $\{\beta_j\}_{j=1}^n \subset \mathcal{Y}$. In practice, finding these n functions in the output space requires solving a block linear operator equation. For the high-dimensional PDE problems we consider in this work, solving such an equation may become prohibitively expensive from both operation count and memory required. A few workarounds were proposed in [136] such as certain diagonalizations, but these rely on simplifying assumptions that are somewhat limiting. More fundamentally, the representation of the solution in (2.22) requires knowledge of the kernel k_μ; in our setting we assume access only to the random feature pair (φ, μ) which defines k_μ and not k_μ itself.

We thus explain how to make progress with this problem given knowledge only of random features. Recall the empirical kernel given by (2.17), the RKHS $\mathcal{H}_{k^{(m)}}$, and Result 2.6. The following result, proved in Appendix A.1, shows that a RFM hypothesis class with a penalized least squares empirical loss function in optimization problem (2.3)–(2.4) is equivalent to kernel ridge regression (2.21) restricted to $\mathcal{H}_{k^{(m)}}$.

Result 2.8 (random feature ridge regression is equivalent to a kernel method)**.** *Assume that $\varphi \in L^2_{\nu\times\mu}(\mathcal{X}\times\Theta;\mathcal{Y})$ and that the random features $\{\varphi(\cdot;\theta_j)\}_{j=1}^m$ are linearly independent in $L^2_\nu(\mathcal{X};\mathcal{Y})$. Fix $\lambda \geq 0$. Let $\widehat{\alpha} \in \mathbb{R}^m$ be the unique minimum norm solution of the following problem:*

$$\min_{\alpha\in\mathbb{R}^m}\left\{\sum_{j=1}^n \frac{1}{2}\left\| y_j - \frac{1}{m}\sum_{\ell=1}^m \alpha_\ell \varphi(a_j;\theta_\ell)\right\|_{\mathcal{Y}}^2 + \frac{\lambda}{2m}\|\alpha\|_2^2\right\}. \tag{2.23}$$

Then the RFM defined by this choice $\alpha = \widehat{\alpha}$ satisfies

$$F_m(\cdot;\widehat{\alpha}) = \operatorname*{arg\,min}_{F\in\mathcal{H}_{k^{(m)}}}\left\{\sum_{j=1}^n \frac{1}{2}\big\|y_j - F(a_j)\big\|_{\mathcal{Y}}^2 + \frac{\lambda}{2}\big\|F\big\|_{\mathcal{H}_{k^{(m)}}}^2\right\}. \tag{2.24}$$

Solving the convex, quadratic problem (2.23) trains the RFM. The first-order condition for a global minimizer leads to the normal equations

$$\sum_{j=1}^m\left(\frac{1}{m}\sum_{i=1}^n \big\langle\varphi(a_i;\theta_l),\varphi(a_i;\theta_j)\big\rangle_{\mathcal{Y}} + \lambda\delta_{lj}\right)\widehat{\alpha}_j = \sum_{i=1}^n \big\langle\varphi(a_i;\theta_l), y_i\big\rangle_{\mathcal{Y}} \tag{2.25}$$

for each $l \in \{1,\ldots,m\}$, where $\delta_{lj} = 1$ if $l = j$ and equals zero otherwise. This is an m-by-m linear system of equations for $\widehat{\alpha} \in \mathbb{R}^m$ that is standard to solve. In the case $\lambda = 0$, the minimum norm solution may be written in terms of a pseudoinverse operator [182, Section 6.11].

Equation (2.25) reveals that the trained RFM $F_m(\cdot;\widehat{\alpha})$ is a linear function of the labeled output data $\{y_i\}_{i=1}^n$. This property is undesirable from the perspective of statistical optimality. Indeed, it is known that any estimator that is linear in the output training data is minimax *suboptimal* for certain classes of problems [256, Theorem 1, Section 4.1, p. 6]. However, any adaptation of the feature pair (φ,μ) to the training data will break this property and potentially restore optimality. For example, choosing λ or hyperparameters appearing in (φ,μ) based on a cross validation procedure would make the RF pair data-dependent as desired [85]. This is typically done in practice.

Example 2.9 (Brownian bridge)**.** We now provide a one-dimensional instantiation of the RFM to illustrate the methodology. Take the input space as $\mathcal{X} \coloneqq (0,1)$, output space $\mathcal{Y} \coloneqq \mathbb{R}$, input space measure $\nu \coloneqq \mathsf{Unif}(0,1)$, and random parameter space $\Theta \coloneqq \mathbb{R}^\infty$. Denote the input by $a = x \in \mathcal{X}$. Then, consider the random feature map $\varphi\colon (0,1)\times\mathbb{R}^\infty \to \mathbb{R}$ defined by the *Brownian bridge*

$$\varphi(x;\theta) \coloneqq \sum_{j\in\mathbb{N}} \theta^{(j)}(j\pi)^{-1}\sqrt{2}\sin(j\pi x)\,, \quad \text{where} \quad \theta^{(j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)\,, \tag{2.26}$$

$\theta \coloneqq \{\theta^{(j)}\}_{j\in\mathbb{N}}$, and $\mu \coloneqq \mathcal{N}(0,1)\times\mathcal{N}(0,1)\times\cdots$. For any realization of $\theta\sim\mu$, the function $\varphi(\cdot;\theta)$ is a Brownian motion constrained to zero at $x=0$ and $x=1$. The induced kernel $k_\mu\colon (0,1)\times(0,1)\to\mathbb{R}$ is then simply the covariance function of this stochastic process:

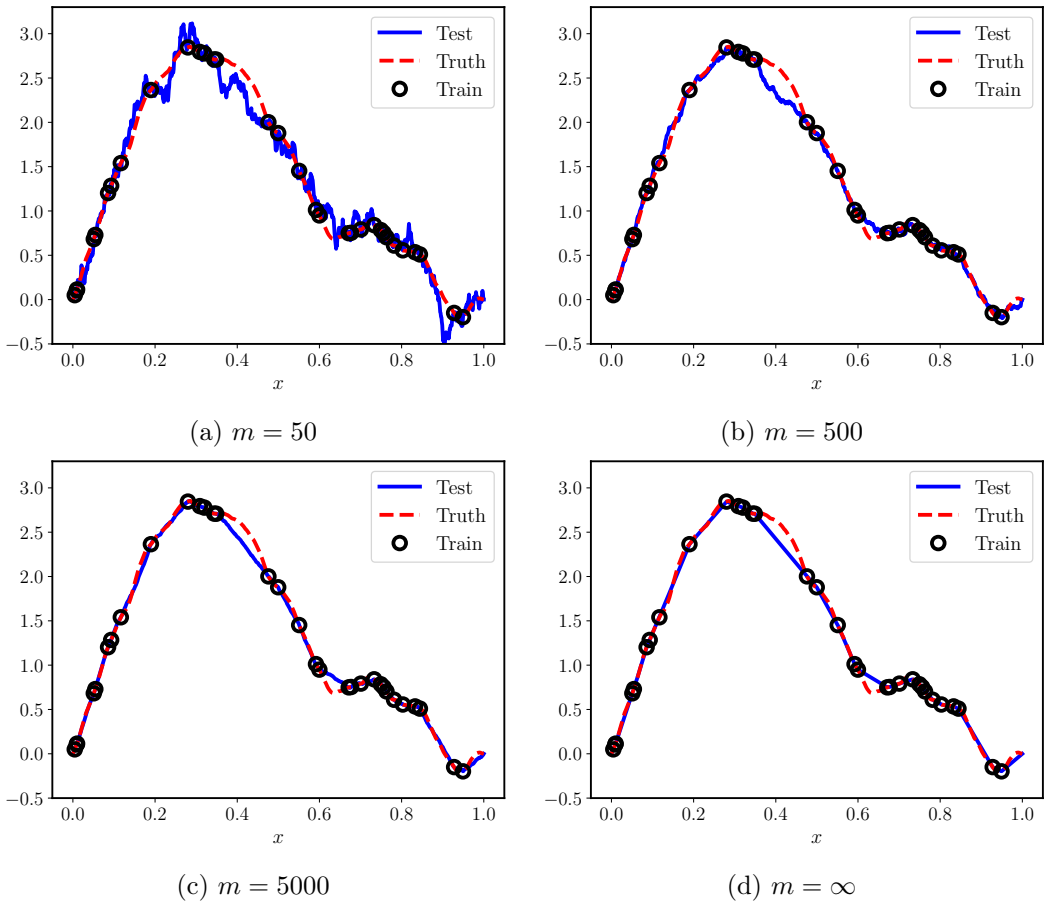
$$k_\mu(x,x') = \mathbb{E}^{\theta\sim\mu}\big[\varphi(x;\theta)\varphi(x';\theta)\big] = \min\{x,x'\} - xx'\,. \tag{2.27}$$

Note that k_μ is the Green's function for the negative Laplacian on $(0,1)$ with Dirichlet boundary conditions. Using this fact, we may explicitly characterize the associated RKHS $\mathcal{H}_{k_\mu}$ as follows. First, we have

$$T_{k_\mu} f = \int_0^1 k_\mu(\cdot,y) f(y)\,dy = \left(-\frac{d^2}{dx^2}\right)^{-1} f\,, \tag{2.28}$$

where the negative Laplacian has domain $H^2((0,1);\mathbb{R})\cap H_0^1((0,1);\mathbb{R})$. Viewing T_{k_μ} as an operator from $L^2((0,1);\mathbb{R})$ into itself, from (2.9) we conclude, upon integration by parts, that

$$\langle f,g\rangle_{\mathcal{H}_{k_\mu}} = \langle f, T_{k_\mu}^{-1} g\rangle_{L^2} = \left\langle \frac{df}{dx}, \frac{dg}{dx}\right\rangle_{L^2} = \langle f,g\rangle_{H_0^1} \quad \text{for all} \quad f,g\in\mathcal{H}_{k_\mu}\,. \tag{2.29}$$



(a) $m = 50$ (b) $m = 500$

(c) $m = 5000$ (d) $m = \infty$

Figure 2.1: Brownian bridge RFM for one-dimensional input-output spaces with $n = 32$ training points fixed and $\lambda = 0$ (Example 2.9): as $m \to \infty$, the RFM approaches the nonparametric interpolant given by the representer theorem (Figure 2.1d), which in this case is a piecewise linear approximation of the true function (an element of RKHS $\mathcal{H}_{k_\mu} = H_0^1$, shown in red). Blue lines denote the trained model evaluated on test data points and black circles denote evaluation at training points.

Note that the last identity does indeed define an inner product on H_0^1. By this formal argument we identify the RKHS $\mathcal{H}_{k_\mu}$ as the Sobolev space $H_0^1((0,1);\mathbb{R})$. Furthermore, Brownian bridge may be viewed as the Gaussian measure $\mathcal{N}(0, T_{k_\mu})$. Approximation using the RFM with the Brownian bridge random features is illustrated in Figure 2.1. Since $k_\mu(\cdot, x)$ is a piecewise linear function, a kernel interpolation or regression method will produce a piecewise linear approximation. Indeed, the figure indicates that the RFM with n training points fixed approaches the optimal piecewise linear kernel interpolant as $m \to \infty$ (see [91] for a related theoretical result).

The Brownian bridge example 2.9 illuminates a more fundamental idea. For

this low-dimensional problem, an expansion in a deterministic Fourier sine basis would be more natural. But if we do not have a natural, computable orthonormal basis, then randomness provides a useful alternative representation; notice that the random features each include random combinations of the deterministic Fourier sine basis in this example. For the more complex problems that we study numerically in the next two sections, we lack knowledge of good, computable bases for general maps in infinite dimensions. The RFM approach exploits randomness to explore, implicitly discover the structure of, and represent, such maps. Thus we now turn away from this example of real-valued maps defined on a subset of the real line and instead consider the use of random features to represent maps between spaces of functions.

2.3 Application to PDE Solution Operators

In this section, we design the random feature maps $\varphi\colon \mathcal{X}\times\Theta \to \mathcal{Y}$ and measures μ for the RFM approximation of two particular PDE parameter-to-solution maps: the evolution semigroup of viscous Burgers' equation in Section 2.3.1 and the coefficient-to-solution operator for the Darcy problem in Section 2.3.2. It is well known to kernel method practitioners that the choice of kernel (which in this work follows from the choice of (φ, μ)) plays a central role in the quality of the function reconstruction. While our method is purely data-driven and requires no knowledge of the governing PDE, we take the view that any prior knowledge can, and should, be introduced into the design of (φ, μ). However, the question of how to automatically determine good random feature pairs for a particular problem or dataset, inducing data-adapted kernels, is open. A preliminary strategy is provided by Dunbar, Mutic, and Nelsen [85]. The maps φ that we choose to employ are nonlinear in both arguments. We also detail the probability measure ν on the input space $\mathcal{X}$ for each of the two PDE applications; this choice is crucial because while we desire the trained RFM to transfer to arbitrary out-of-distribution inputs from $\mathcal{X}$, we can in general only expect the learned map to perform well when restricted to inputs statistically similar to those sampled from ν.

2.3.1 Burgers' Equation: Formulation

Viscous Burgers' equation in one spatial dimension is representative of the advection-dominated PDE problem class in some regimes; these time-dependent equations are not conservation laws due to the presence of small dissipative

terms, but nonlinear transport still plays a central role in the evolution of solutions. The initial value problem we consider is

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial}{\partial x}\left(\dfrac{u^2}{2}\right) - \varepsilon \dfrac{\partial^2 u}{\partial x^2} = f & \text{in } (0,\infty)\times(0,1)\,, \\ u(\cdot,0) = u(\cdot,1)\,, \quad \dfrac{\partial u}{\partial x}(\cdot,0) = \dfrac{\partial u}{\partial x}(\cdot,1) & \text{in } (0,\infty)\,, \\ u(0,\cdot) = a & \text{in } (0,1)\,, \end{cases} \tag{2.30}$$

where $\varepsilon > 0$ is the viscosity (i.e., diffusion coefficient) and we have imposed periodic boundary conditions. The initial condition a serves as the input and is drawn according to a Gaussian measure defined by

$$a \sim \nu \coloneqq \mathcal{N}(0, C) \tag{2.31}$$

with Matérn-like covariance operator [86, 187]

$$C \coloneqq \tau^{2\alpha-d}(-\Delta + \tau^2 \,\mathrm{Id})^{-\alpha}\,, \tag{2.32}$$

where $d = 1$ and the negative Laplacian $-\Delta$ is defined over $\mathbb{T}^1 = [0,1]_{\text{per}}$ and restricted to functions which integrate to zero over $\mathbb{T}^1$. The hyperparameter $\tau \geq 0$ is an inverse length scale and $\alpha > 1/2$ controls the regularity of the draw. Such a are almost surely Hölder and Sobolev regular with exponent up to $\alpha - 1/2$ [74, Theorem 12, p. 338], so in particular $a \in \mathcal{X} \coloneqq L^2(\mathbb{T}^1;\mathbb{R})$. Then for all $\varepsilon > 0$, the unique global solution $u(t,\cdot)$ to (2.30) is real analytic for all $t > 0$ [141, Theorem 1.1]. Hence, setting the output space to be $\mathcal{Y} \coloneqq H^s(\mathbb{T}^1;\mathbb{R})$ for any $s > 0$, we may define the solution map

$$\begin{aligned} F^\dagger \colon L^2 &\to H^s \\ a &\mapsto F^\dagger(a) \coloneqq \Psi_T(a) = u(T,\cdot)\,, \end{aligned} \tag{2.33}$$

where $\{\Psi_t\}_{t>0}$ forms the solution operator semigroup for (2.30) and we fix the final time $t = T > 0$. The map $F^\dagger$ is smoothing and nonlinear.

We now describe a random feature map for use in the RFM (2.19) that we call *Fourier space random features*. Let $\mathcal{F}$ denote the Fourier transform over spatial domain $\mathbb{T}^1$ and define $\varphi \colon \mathcal{X} \times \Theta \to \mathcal{Y}$ by

$$\varphi(a;\theta) \coloneqq \sigma\big(\mathcal{F}^{-1}(\chi \mathcal{F} a \mathcal{F} \theta)\big)\,, \tag{2.34}$$

where $\sigma(\cdot)$, the ELU function defined below, is defined as a mapping on $\mathbb{R}$ and applied pointwise to functions. Viewing $\Theta \subseteq L^2(\mathbb{T}^1;\mathbb{R})$, the randomness

enters through $\theta \sim \mu \coloneqq \mathcal{N}(0, C')$ with C' the same covariance operator as in (2.32) but with potentially different inverse length scale and regularity, and the *wavenumber filter function* $\chi : \mathbb{Z} \to \mathbb{R}_{\geq 0}$ is

$$\chi(k) \coloneqq \sigma_\chi(2\pi|k|\delta)\,, \quad \text{where} \quad \sigma_\chi(r) \coloneqq \max\big\{0, \min\{2r, (r+1/2)^{-\beta}\}\big\}\,, \tag{2.35}$$

where δ, $\beta > 0$. The map $\varphi(\cdot;\theta)$ essentially performs a filtered random convolution with the initial condition. Figure 2.2a illustrates a sample input and output from φ. Although simply hand-tuned for performance and not optimized, the filter χ is designed to shuffle energy in low to medium wavenumbers and cut off high wavenumbers (see Figure 2.2b), reflecting our prior knowledge of solutions to (2.30).

We choose the activation function σ in (2.34) to be the exponential linear unit

$$\mathrm{ELU}(r) \coloneqq \begin{cases} r\,, & r \geq 0 \\ e^r - 1\,, & r < 0\,. \end{cases} \tag{2.36}$$

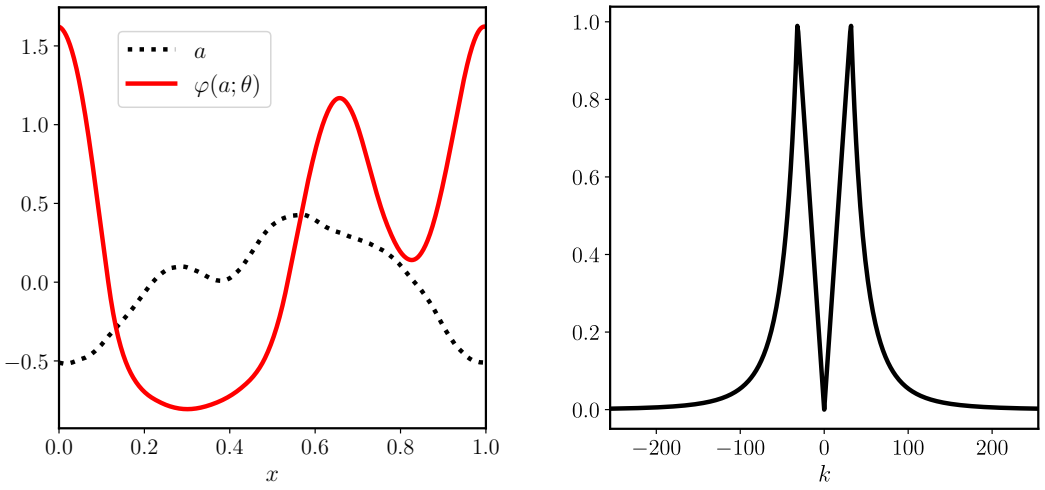
The ELU function has successfully been used as activation in other machine learning frameworks for related nonlinear PDE problems [165, 213, 214]. We also find ELU to perform better in the RFM framework over several other choices including $\mathrm{ReLU}(\cdot)$, $\tanh(\cdot)$, $\mathrm{sigmoid}(\cdot)$, $\sin(\cdot)$, $\mathrm{SELU}(\cdot)$, and $\mathrm{softplus}(\cdot)$. Note that the pointwise evaluation of ELU in (2.34) will be well defined, by Sobolev embedding, for $s > 1/2$ sufficiently large in the definition of $\mathcal{Y} = H^s$. Since the solution operator maps into H^s for any $s > 0$, this does not constrain the method.

2.3.2 Darcy Flow: Formulation

Divergence form elliptic equations [109] arise in a variety of applications, in particular, the groundwater flow in a porous medium governed by Darcy's law [26]. This linear elliptic boundary value problem reads

$$\begin{cases} -\nabla \cdot (a \nabla u) = f & \text{in } D\,, \\ u = 0 & \text{on } \partial D\,, \end{cases} \tag{2.37}$$

where D is a bounded open subset in $\mathbb{R}^d$, f represents sources and sinks of fluid, a the permeability of the porous medium, and u the piezometric head; all three functions map D into $\mathbb{R}$ and, in addition, a is strictly positive almost everywhere in D. We work in a setting where f is fixed and consider the



(a) (b)

Figure 2.2: Random feature map construction for Burgers' equation: Figure 2.2a displays a representative input-output pair for the random feature $\varphi(\cdot;\theta)$, $\theta \sim \mu$ (2.34), while Figure 2.2b shows the filter $k \mapsto \chi(k)$ for $\delta = 0.0025$ and $\beta = 4$ (2.35).

input-output map defined by $a \mapsto u$. The measure ν on a is a high contrast level set prior constructed as the pushforward of a Gaussian measure:

$$a \sim \nu \coloneqq \psi_\sharp \mathcal{N}(0, C)\,. \tag{2.38}$$

Here $\psi : \mathbb{R} \to \mathbb{R}$ is a threshold function defined by

$$\psi(r) \coloneqq a^+ \mathbb{1}_{(0,\infty)}(r) + a^- \mathbb{1}_{(-\infty,0)}(r)\,, \quad \text{where} \quad 0 < a^- \le a^+ < \infty\,, \tag{2.39}$$

applied pointwise to functions, and the covariance operator C is given in (2.32) with $d = 2$ and homogeneous Neumann boundary conditions on $-\Delta$. That is, the resulting coefficient a almost surely takes only two values (a^+ or a^-) and, as the zero level set of a Gaussian random field, exhibits random geometry in the physical domain D. It follows that $a \in L^\infty(D; \mathbb{R}_{\ge 0})$ almost surely. Further, the size of the contrast ratio a^+/a^- measures the scale separation of this elliptic problem and hence controls the difficulty of reconstruction [32]. See Figure 2.3a for a representative sample.

Given $f \in L^2(D;\mathbb{R})$, the standard Lax-Milgram theory may be applied to show that for coefficient $a \in \mathcal{X} \coloneqq L^\infty(D;\mathbb{R}_{\ge 0})$, there exists a unique weak solution $u \in \mathcal{Y} \coloneqq H_0^1(D;\mathbb{R})$ for (2.37) (see, e.g., Evans [93]). Thus, we define the ground truth solution map

$$\begin{aligned} F^\dagger \colon L^\infty &\to H_0^1 \\ a &\mapsto F^\dagger(a) \coloneqq u\,. \end{aligned} \tag{2.40}$$

Although the PDE (2.37) is linear, the solution map $F^\dagger$ is nonlinear.

We now describe the chosen random feature map for this problem, which we call *predictor-corrector random features*. Define $\varphi\colon \mathcal{X} \times \Theta \to \mathcal{Y}$ by $\varphi(a;\theta) \coloneqq p_1$ such that

$$-\Delta p_0 = \frac{f}{a} + \sigma_\gamma(\theta_1) \quad \text{and} \tag{2.41a}$$

$$-\Delta p_1 = \frac{f}{a} + \sigma_\gamma(\theta_2) + \nabla(\log a) \cdot \nabla p_0\,, \tag{2.41b}$$

where the boundary conditions are homogeneous Dirichlet, $\theta = (\theta_1, \theta_2) \sim \mu \coloneqq \mu' \times \mu'$ are two Gaussian random fields each drawn from $\mu' \coloneqq \mathcal{N}(0, C')$, f is the source term in (2.37), and $\gamma = (s^+, s^-, \delta)$ are parameters for a thresholded sigmoid $\sigma_\gamma : \mathbb{R} \to \mathbb{R}$ given by

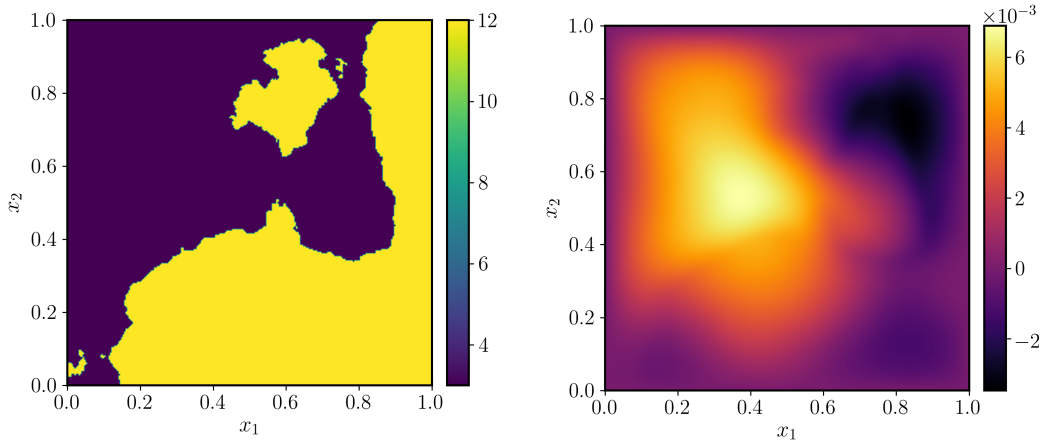
$$r \mapsto \sigma_\gamma(r) \coloneqq \frac{s^+ - s^-}{1 + e^{-r/\delta}} + s^- \tag{2.42}$$

and extended as a Nemytskii operator when applied to $\theta_1(\cdot)$ or $\theta_2(\cdot)$. We view $\Theta \subseteq L^2(D;\mathbb{R}) \times L^2(D;\mathbb{R})$. In practice, since ∇a is not well-defined when drawn from the level set measure, we replace a with a_ε, where $a_\varepsilon \coloneqq v(1)$ is a smoothed version of a obtained by evolving the following linear heat equation for one time unit:

$$\begin{cases} \dfrac{\partial v}{\partial t} = \eta \Delta v & \text{in } (0,1) \times D\,, \\ \mathsf{n} \cdot \nabla v = 0 & \text{on } (0,1) \times \partial D\,, \\ v(0) = a & \text{in } D\,, \end{cases} \tag{2.43}$$

where n is the outward unit normal vector to ∂D. An example of the response $\varphi(a;\theta)$ to a piecewise constant input $a \sim \nu$ is shown in Figure 2.3 for some $\theta \sim \mu$.

We remark that by removing the two random terms involving θ_1 and θ_2 in (2.41), we obtain a remarkably accurate surrogate model for the PDE. This observation is representative of a more general iterative method, a predictor-corrector type iteration, for solving the Darcy equation (2.37), whose convergence depends on the size of a. The map φ is essentially a random perturbation of a single step of this iterative method: Equation (2.41a) makes a coarse prediction of the output, then (2.41b) improves this prediction with a correction term derived from expanding the original PDE. This choice of φ falls within an ensemble viewpoint that the RFM may be used to improve pre-existing surrogate models



(a) $a \sim \nu$ (b) $\varphi(a;\theta)$, where $\theta \sim \mu$

Figure 2.3: Random feature map construction for Darcy flow: Figure 2.3a displays a representative input draw a with $\tau = 3$, $\alpha = 2$ and $a^+ = 12$, $a^- = 3$; Figure 2.3b shows the output random feature $\varphi(a;\theta)$ (Equation 2.41) taking the coefficient a as input. Here, $f \equiv 1$, $\tau' = 7.5$, $\alpha' = 2$, $s^+ = 1/a^+$, $s^- = -1/a^-$, and $\delta = 0.15$.

by taking $\varphi(\cdot;\theta)$ to be an existing emulator, but randomized in a principled way through $\theta \sim \mu$.

For this particular example, we are cognizant of the facts that the random feature map φ requires full knowledge of the Darcy equation and a naïve evaluation of φ may be as expensive as solving the original PDE, which is itself a linear PDE; however, we believe that the ideas underlying the random features used here are intuitive and suggestive of what is possible in other applications areas. For example, RFMs may be applied on larger domains with simple geometries, viewed as supersets of the physical domain of interest, enabling the use of efficient algorithms such as the fast Fourier transform (FFT) even though these may not be available on the original problem, either because the operator to be inverted is spatially inhomogeneous or because of the complicated geometry of the physical domain.

2.4 Numerical Experiments

We now assess the performance of our proposed methodology on the approximation of operators $F^\dagger \colon \mathcal{X} \to \mathcal{Y}$ presented in Section 2.3. Practical implementation of the approach on a computer necessitates discretization of the input-output function spaces $\mathcal{X}$ and $\mathcal{Y}$. Hence in the numerical experiments that follow, all infinite-dimensional objects such as the training data, evaluations of random feature maps, and random fields are discretized on an equispaced mesh with

K grid points to take advantage of the $O(K \log K)$ computational speed of the FFT. The simple choice of equispaced points does not limit the proposed approach, as our formulation of the RFM on function space allows the method to be implemented numerically with any choice of spatial discretization. Such a numerical discretization procedure leads to the problem of high- but finite-dimensional approximation of discretized target operators mapping $\mathbb{R}^K$ to $\mathbb{R}^K$ by similarly discretized RFMs. However, we emphasize the fact that K is allowed to vary, and we study the properties of the discretized RFM as K varies, noting that since the RFM is defined conceptually on function space in Section 2.2 without reference to discretization, its discretized numerical realization has approximation quality consistent with the infinite-dimensional limit $K \to \infty$. This implies that the same trained model can be deployed across the entire hierarchy of finite-dimensional spaces $\mathbb{R}^K$ parametrized by $K \in \mathbb{N}$ without the need to be re-trained, provided that K is sufficiently large. Thus in this section, our notation does not make explicit the dependence of the discretized RFM or target operators on mesh size K. We demonstrate these claimed properties numerically.[1]

The input functions and our chosen random feature maps (2.34) and (2.41) require i.i.d. draws of Gaussian random fields to be fully defined. We efficiently sample these fields by truncating a Karhunen–Loéve expansion and employing fast summation of the eigenfunctions with FFT. More precisely, on a mesh of size K, denote by $g(\cdot)$ a numerical approximation of a Gaussian random field on domain $D = (0,1)^d$, $d = 1,\, 2$:

$$g = \sum_{k \in Z_K} \xi_k \sqrt{\lambda_k} \phi_k \approx \sum_{k' \in \mathbb{Z}^d_{\geq 0}} \xi_{k'} \sqrt{\lambda_{k'}} \phi_{k'} \sim \mathcal{N}(0, C)\,, \tag{2.44}$$

where $\{\xi_j\} \sim \mathcal{N}(0,1)$ i.i.d. and $Z_K \subset \mathbb{Z}_{\geq 0}$ is a truncated one-dimensional lattice of cardinality K ordered such that $\{\lambda_j\}$ is nonincreasing. The pairs $(\lambda_{k'}, \phi_{k'})$ are found by solving the eigenvalue problem $C\phi_{k'} = \lambda_{k'}\phi_{k'}$ for non-negative, symmetric, trace-class operator C (2.32). Concretely, these solutions

[1] The datasets are available at doi.org/10.22002/55tdh-hda68. The code used to produce the numerical results and figures in this chapter is available at

https://github.com/nickhnelsen/random-features-banach .

are given by

$$\phi_{k'}(x) = \begin{cases} \sqrt{2}\cos(k_1'\pi x_1)\cos(k_2'\pi x_2), & k_1' \text{ or } k_2' = 0, \\ 2\cos(k_1'\pi x_1)\cos(k_2'\pi x_2), & \text{otherwise,} \end{cases}, \quad \text{and} \tag{2.45a}$$

$$\lambda_{k'} = \tau^{2\alpha-2}(\pi^2|k'|^2 + \tau^2)^{-\alpha} \tag{2.45b}$$

for homogeneous Neumann boundary conditions when $d = 2$, $k' = (k_1', k_2') \in \mathbb{Z}^2_{\geq 0}\backslash\{0\}$, $x = (x_1, x_2) \in (0,1)^2$, and given by

$$\phi_{2j}(x) = \sqrt{2}\cos(2\pi jx)\,, \quad \phi_{2j-1}(x) = \sqrt{2}\sin(2\pi jx)\,, \quad \phi_0(x) = 1\,, \tag{2.46a}$$

$$\lambda_{2j} = \lambda_{2j-1} = \tau^{2\alpha-1}(4\pi^2 j^2 + \tau^2)^{-\alpha}\,, \quad \lambda_0 = \tau^{-1} \tag{2.46b}$$

for periodic boundary conditions when $d = 1$, $j \in \mathbb{Z}_{>0}$, and $x \in (0,1)$. In both cases, we enforce that g integrate to zero over D by manually setting to zero the Fourier coefficient corresponding to multi-index $k' = 0$. We use such g in all experiments that follow. Additionally, the k and k' used in this section to denote wavenumber indices should not be confused with our previous notation for kernels.

With the discretization and data generation setup now well-defined, and the pairs (φ, μ) given in Section 2.3, the last algorithmic step is to train the RFM by solving (2.25) and then test its performance. For a fixed number of random features m, we only train and test a single realization of the RFM, viewed as a random variable itself. In each instance m is varied in the experiments that follow, the draws $\{\theta_j\}_{j=1}^m$ are re-sampled i.i.d. from μ. To measure the distance between the trained RFM $F_m(\cdot;\widehat{\alpha})$ and the ground truth map $F^\dagger$, we employ the *approximate expected relative test error*

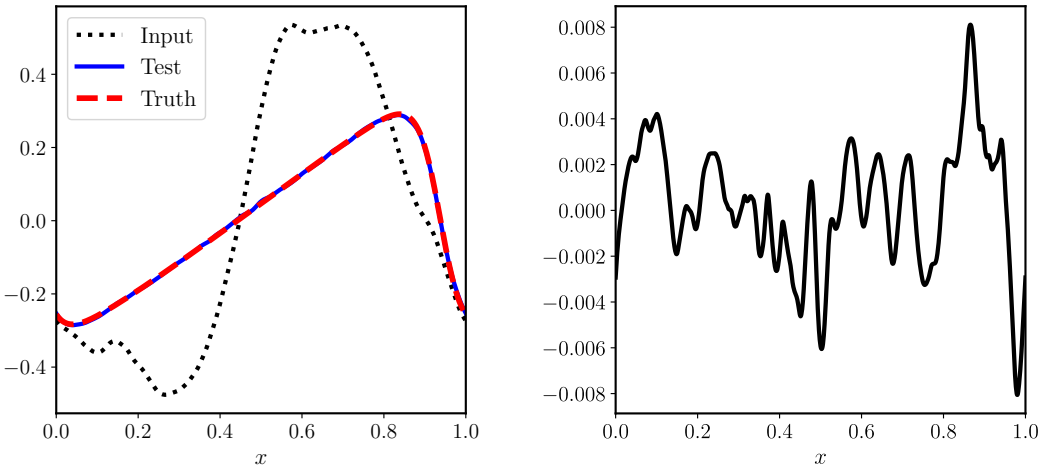
$$e_{n',m} \coloneqq \frac{1}{n'}\sum_{j=1}^{n'} \frac{\|F^\dagger(a_j') - F_m(a_j';\widehat{\alpha})\|_{L^2}}{\|F^\dagger(a_j')\|_{L^2}} \approx \mathbb{E}^{a'\sim\nu}\left[\frac{\|F^\dagger(a') - F_m(a';\widehat{\alpha})\|_{L^2}}{\|F^\dagger(a')\|_{L^2}}\right], \tag{2.47}$$

where the $\{a_j'\}_{j=1}^{n'}$ are drawn i.i.d. from ν and n' denotes the number of input-output pairs used for testing. All $L^2(D;\mathbb{R})$ norms on the physical domain are numerically approximated by composite trapezoid rule quadrature. Since $\mathcal{Y} \subset L^2$ for both the PDE solution operators (2.33) and (2.40), we also perform all required inner products during training in L^2 rather than in $\mathcal{Y}$; this results in smaller relative test error $e_{n',m}$.

2.4.1 Burgers' Equation: Experiment

We generate a high-resolution dataset of input-output pairs by solving Burgers' equation (2.30) on an equispaced periodic mesh of size $K = 1025$ (identifying the first mesh point with the last) with random initial conditions sampled from $\nu = \mathcal{N}(0, C)$ using (2.44), where C is given by (2.32) with parameter choices $\tau = 7$ and $\alpha = 2.5$. The full order solver is a FFT-based pseudospectral method for spatial discretization [99] and a fourth-order Runge-Kutta integrating factor time-stepping scheme for time discretization [138]. All data represented on mesh sizes $K < 1025$ used in both training and testing phases are subsampled from this original dataset, and hence we consider numerical realizations of $F^\dagger$ (2.33) up to $\mathbb{R}^{1025} \to \mathbb{R}^{1025}$. We fix $n = 512$ training and $n' = 4000$ testing pairs unless otherwise noted, and also fix the viscosity to $\varepsilon = 10^{-2}$ in all experiments. Lowering ε leads to smaller length scale solutions and more difficult reconstruction; more data (higher n) and features (higher m) or a more expressive choice of (φ, μ) would be required to achieve comparable error levels due to the slow decaying Kolmogorov width of the solution map. For simplicity, we set the forcing $f \equiv 0$, although nonzero forcing could lead to other interesting solution maps such as $f \mapsto u(T, \cdot)$. It is easy to check that the solution will have zero mean for all time and a steady state of zero. Hence, we choose $T \leq 2$ to ensure that the solution is far enough away from steady state. For the random feature map (2.34), we fix the hyperparameters $\alpha' = 2$, $\tau' = 5$, $\delta = 0.0025$, and $\beta = 4$. The map itself is evaluated efficiently with the FFT and requires no other tools to be discretized. RFM hyperparameters were hand-tuned but not optimized. We find that regularization during training had a negligible effect for this problem, so the RFM is trained with $\lambda = 0$ by solving the normal equations (2.25) with the pseudoinverse to deliver the minimum norm least squares solution; we use the truncated singular value decomposition (SVD) implementation in Python's `scipy.linalg.pinv2` for this purpose.

Our experiments study the RFM approximation to the viscous Burgers' equation evolution operator semigroup (2.33). As a visual aid for the high-dimensional problem at hand, Figure 2.4 shows a representative sample input and output along with a trained RFM test prediction. To determine whether the RFM has actually learned the correct evolution operator, we test the semigroup property of the map; [274] pursues closely related work also in a Fourier space setting. Denote the $(j-1)$-fold composition of a function G with itself by G^j. Then,



(a) (b)

Figure 2.4: Representative input-output test sample for the Burgers' equation solution map $F^\dagger := \Psi_1$: Here, $n = 512$, $m = 1024$, and $K = 1025$. Figure 2.4a shows a sample input, output (truth), and trained RFM prediction (test), while Figure 2.4b displays the pointwise error. The relative L^2 error for this single prediction is 0.0146.

Table 2.1: Expected relative error $e_{n',m}$ for time upscaling with the learned RFM operator semigroup for Burgers' equation: Here, $n' = 4000$, $m = 1024$, $n = 512$, and $K = 129$. The RFM is trained on data from the evolution operator $\Psi_{T=0.5}$, and then tested on input-output samples generated from Ψ_{jT}, where $j = 2, 3, 4$, by repeated composition of the learned model. The increase in error is small even after three compositions, reflecting excellent out-of-distribution performance.

Train on: $T = 0.5$	Test on: $2T = 1.0$	$3T = 1.5$	$4T = 2.0$
0.0360	0.0407	0.0528	0.0788

with $u(0, \cdot) = a$, we have

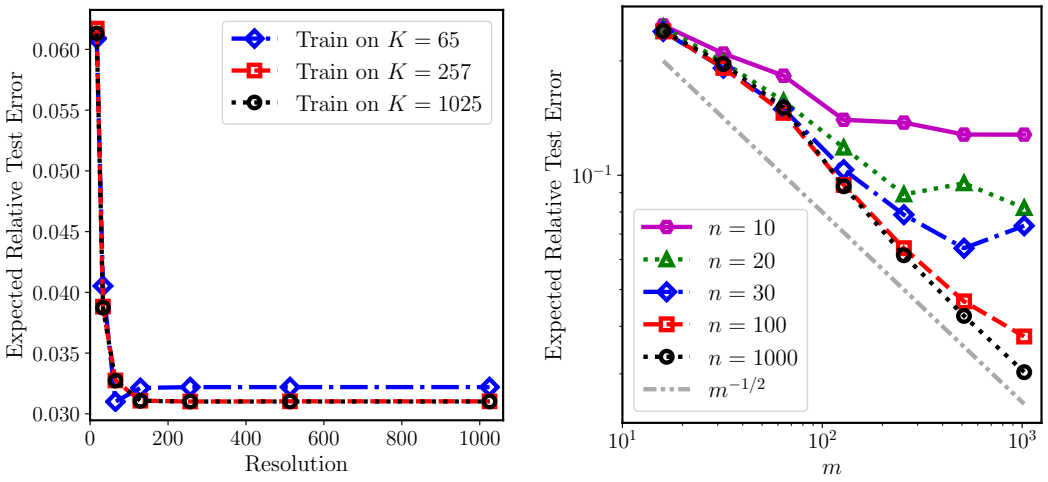
$$(\Psi_T \circ \cdots \circ \Psi_T)(a) = \Psi_T^j(a) = \Psi_{jT}(a) = u(jT, \cdot) \tag{2.48}$$

by definition. We train the RFM on input-output pairs from the map Ψ_T with $T := 0.5$ to obtain $\widehat{F} := F_m(\cdot; \widehat{\alpha})$. Then, it should follow from (2.48) that $\widehat{F}^j \approx \Psi_{jT}$, that is, each application of $\widehat{F}$ should evolve the solution T time units. We test this semigroup approximation by learning the map $\widehat{F}$ and then comparing $\widehat{F}^j$ on $n' = 4000$ fixed inputs to outputs from each of the operators Ψ_{jT}, with $j \in \{1, 2, 3, 4\}$ (the solutions at time T, $2T$, $3T$, $4T$). The results are presented in Table 2.1 for a fixed mesh size $K = 129$. We observe that the composed RFM map $\widehat{F}^j$ accurately captures Ψ_{jT}, though this accuracy deteriorates as j increases due to error propagation in time as is common with any traditional integrator. However, even after three

compositions corresponding to 1.5 time units past the training time $T = 0.5$, the relative error only increases by around 0.04. It is remarkable that the RFM learns time evolution without explicitly time-stepping the PDE (2.30) itself. Such a procedure is coined *time upscaling* in the PDE context and in some sense breaks the CFL stability barrier [80]. Table 2.1 is evidence that the RFM has excellent out-of-distribution performance: although only trained on inputs $a \sim \nu$, the model outputs accurate predictions given new input samples $\Psi_{jT}(a) \sim (\Psi_{jT})_\sharp \nu$.

We next study the ability of the RFM to transfer its learned coefficients $\widehat{\alpha}$ obtained from training on mesh size K to different mesh resolutions K' in fig. 2.5a. We fix $T \coloneqq 1$ from here on and observe that the lowest test error occurs when $K = K'$, that is, when the train and test resolutions are identical; this behavior was also observed in the contemporaneous work [173]. At very low resolutions, such as $K = 17$ here, the test error is dominated by discretization error which can become quite large; for example, resolving conceptually infinite-dimensional objects such as the Fourier space-based feature map in (2.34) or the L^2 norms in (2.47) with only 17 grid points gives bad accuracy. But outside this regime, the errors are essentially constant across resolution regardless of the training resolution K, indicating that the RFM learns its optimal coefficients independently of the resolution and hence generalizes well to any desired mesh size. In fact, the trained model could be deployed on different discretizations of the domain D (e.g., various choices of finite elements, graph-based/particle methods), not just with different mesh sizes. Practically speaking, this means that high-resolution training sets can be subsampled to smaller mesh sizes K (yet still large enough to avoid large discretization error) for faster training, leading to a trained model with nearly the same accuracy at all higher resolutions.

The smallest expected relative test error achieved by the RFM is 0.0303 for the configuration detailed in Figure 2.5b. This excellent performance is encouraging because the error we report is of the same order of magnitude as that reported by Li et al. [172, Section 5.1] for the same Burgers' solution operator that we study, but with slightly different problem parameter choices. We emphasize that the neural operator methods in that work are based on deep learning, which involves training neural networks by solving a nonconvex optimization problem with stochastic gradient descent, while our random feature methods have orders of magnitude fewer trainable parameters that are easily optimized

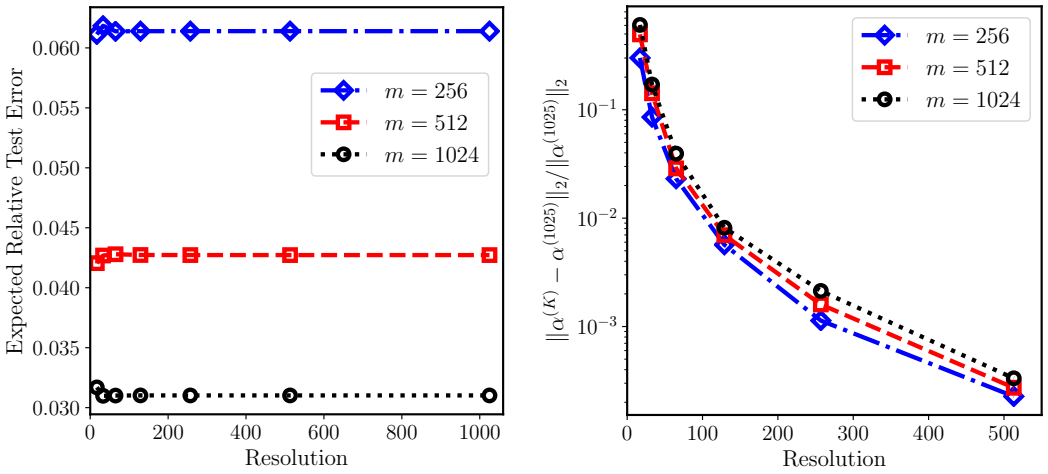


(a) (b)

Figure 2.5: Expected relative test error of a trained RFM for the Burgers' evolution operator $F^\dagger = \Psi_1$ with $n' = 4000$ test pairs: Figure 2.5a displays the invariance of test error w.r.t. training and testing on different resolutions for $m = 1024$ and $n = 512$ fixed; the RFM can train and test on different mesh sizes without loss of accuracy. Figure 2.5b shows the decay of the test error for resolution $K = 129$ fixed as a function of m and n; the error follows the $O(m^{-1/2})$ Monte Carlo rate remarkably well. The smallest error achieved is 0.0303 for $n = 1000$ and $m = 1024$.

through convex optimization. In Figure 2.5b, we see that for large enough n, the error empirically follows the $O(m^{-1/2})$ parameter complexity bound that is suggested by Monte Carlo intuition, as discussed in Section 2.2.3. The figure also indicates a delicate dependence of m as a function of n, in particular, n must increase with m as is expected from parametric estimation. A detailed account of the dependence of m on n required to achieve a certain error tolerance for the RFM is given in the next chapter [162]. We also refer the interested reader to [52] for a sharp statistical analysis in a related setting.

Finally, Figure 2.6 demonstrates the invariance of the expected relative test error to the mesh resolution used for training and testing. This result is a consequence of framing the RFM on function space; other machine-learning-based surrogate methods defined in finite dimensions exhibit an *increase* in test error as mesh resolution is increased (see [34, Section 4] for a numerical account of this phenomenon). The first panel, Figure 2.6a, shows the error as a function of mesh resolution for three values of m. For very low resolution, the error varies slightly but then flattens out to a constant value as $K \to \infty$. The second panel, Figure 2.6b, indicates that the learned coefficient $\alpha^{(K)}$ for each K converges to some $\alpha^{(\infty)}$ as $K \to \infty$, again reflecting the design of the



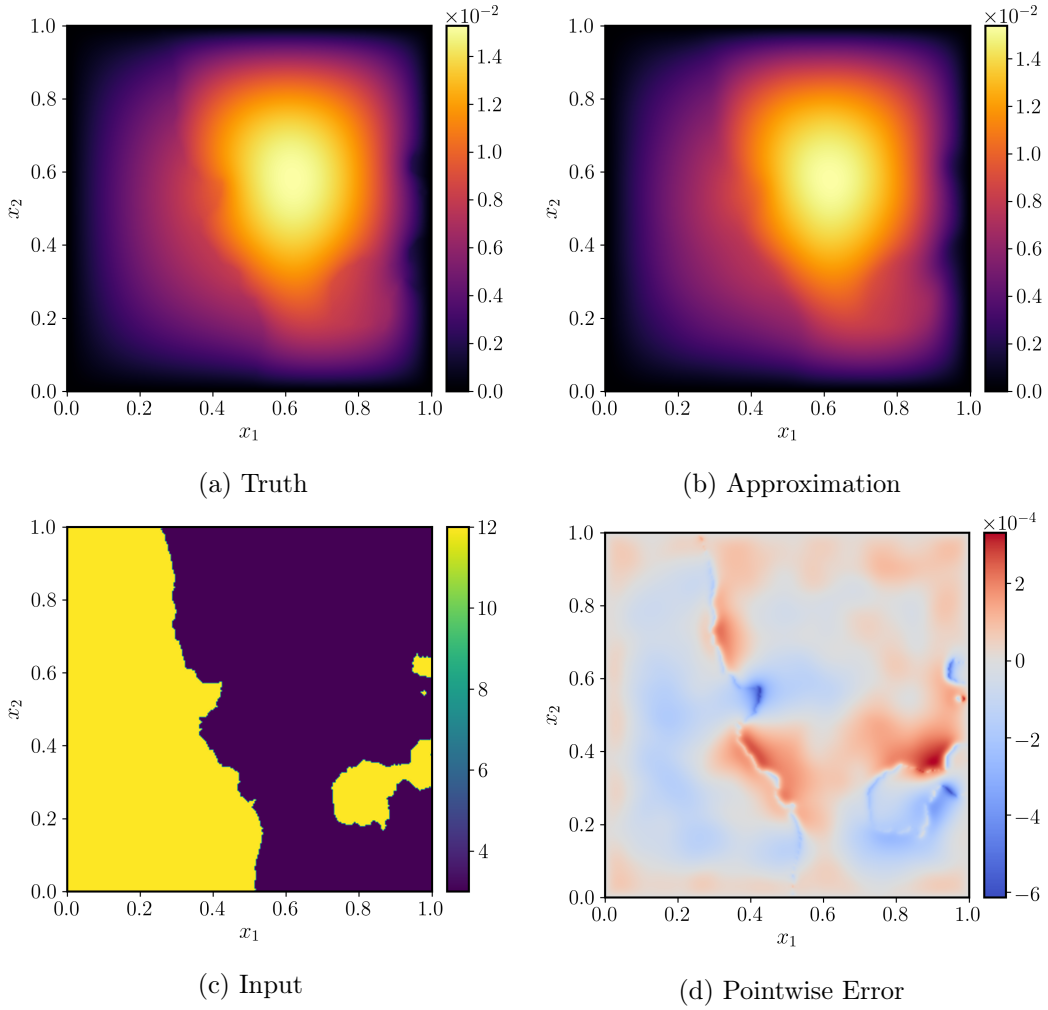
(a) (b)

Figure 2.6: Results of a trained RFM for the Burgers' equation evolution operator $F^\dagger = \Psi_1$: Here, $n = 512$ training and $n' = 4000$ testing pairs were used. Figure 2.6a shows resolution-invariant test error for various m. Figure 2.6b displays the relative error of the learned coefficient α w.r.t. the coefficient learned on the highest mesh size ($K = 1025$).

RFM as a mapping between infinite-dimensional spaces.

2.4.2 Darcy Flow: Experiment

In this section, we consider Darcy flow on the physical domain $D \coloneqq (0,1)^2$, the unit square. We generate a high-resolution dataset of input-output pairs for $F^\dagger$ (2.40) by solving (2.37) on an equispaced 257×257 mesh (size $K = 257^2$) using a second-order finite difference scheme. All mesh sizes $K < 257^2$ are subsampled from this original dataset and hence we consider numerical realizations of $F^\dagger$ up to $\mathbb{R}^{66049} \to \mathbb{R}^{66049}$. We denote *resolution* by r such that $K = r^2$. We fix $n = 128$ training and $n' = 1000$ testing pairs unless otherwise noted. The input data are drawn from the level set measure ν (2.38) with $\tau = 3$ and $\alpha = 2$ fixed. We choose $a^+ = 12$ and $a^- = 3$ in all experiments that follow and hence the contrast ratio $a^+/a^- = 4$ is fixed. The source is fixed to $f \equiv 1$, the constant function. We evaluate the predictor-corrector random features φ (2.41) using an FFT-based fast Poisson solver corresponding to an underlying second-order finite difference stencil at a cost of $O(K \log K)$ per solve. The smoothed coefficient a_ε in the definition of φ is obtained by solving (2.43) with time step 0.03 and diffusion constant $\eta = 10^{-4}$; with centered second-order finite differences, this incurs 34 time steps and hence a cost $O(34K)$. We fix the hyperparameters $\alpha' = 2$, $\tau' = 7.5$, $s^+ = 1/12$, $s^- = -1/3$, and $\delta = 0.15$



(a) Truth

(b) Approximation

(c) Input

(d) Pointwise Error

Figure 2.7: Representative input-output test sample for the Darcy flow solution map: Here, $n = 256$, $m = 350$, and $K = 257^2$. Figure 2.7c shows a sample input, Figure 2.7a the resulting output (truth), Figure 2.7b a trained RFM prediction, and Figure 2.7d the pointwise error. The relative L^2 error for this single prediction is 0.0122.

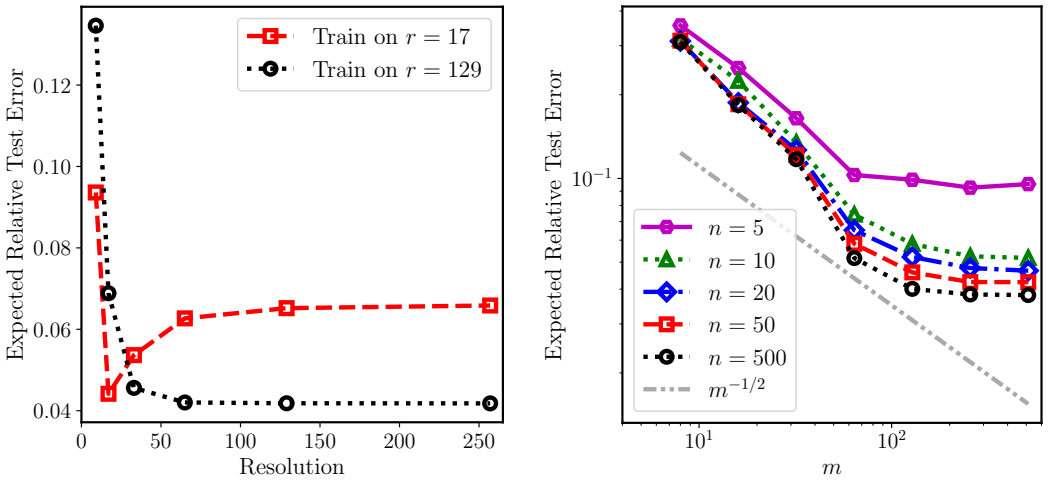
for the map φ. Unlike in Section 2.4.1, we find via grid search on λ that regularization during training does improve the reconstruction of the Darcy flow solution operator and hence we train with $\lambda := 10^{-8}$ fixed. We remark that, for simplicity, the above hyperparameters were not systematically and jointly optimized; as a consequence the RFM performance has the capacity to improve beyond the results in this section.

Darcy flow is characterized by the geometry of the high contrast coefficients $a \sim \nu$. As seen in Figure 2.7, the solution inherits the steep interfaces of the input. However, we see that a trained RFM with predictor-corrector random

features (2.41) captures these interfaces well, albeit with slight smoothing; the error concentrates on the location of the interface. The effect of increasing m and n on the test error is shown in Figure 2.8b. Here, the error appears to saturate more than was observed for the Burgers' equation problem (Figure 2.5b) and does not follow the $O(m^{-1/2})$ rate. This is likely due to fixing λ to be constant instead of scaling it with m. It is also possible that the Darcy flow solution map does not belong to the RKHS $\mathcal{H}_{k_\mu}$, leading to an additional misspecification error. However, the smallest test error achieved for the best performing RFM configuration is 0.0381, which is on the same scale as the error reported in competing neural operator methods [34, 173] for the same Darcy flow setup.

The RFM is able to be successfully trained and tested on different resolutions for Darcy flow. Figure 2.8a shows that, again, for low resolutions, the smallest relative test error is achieved when the train and test resolutions are identical (here, for $r = 17$). However, when the resolution is increased away from this low resolution regime, the relative test error slightly increases then approaches a constant value, reflecting the function space design of the method. Training the RFM on a high-resolution mesh poses no issues when transferring to lower or higher resolutions for model evaluation, and it achieves consistent error for test resolutions sufficiently large (i.e., $r \geq 33$, the regime where discretization error starts to become negligible). Additionally, the RFM basis functions $\{\varphi(\cdot;\theta_j)\}_{j=1}^m$ are defined without any dependence on the training data unlike in other competing approaches based on similar shallow linear approximations, such as the reduced basis method or the PCA-Net method in [34]. Consequently, our RFM may be directly evaluated on any desired mesh resolution once trained ("super-resolution"), whereas those aforementioned approaches require some form of interpolation to transfer between different mesh sizes [34, Section 4.3].

In Figure 2.9, we again confirm that our method is invariant to the refinement of the mesh and improves with more random features. While the difference at low resolutions is more pronounced than that observed for Burgers' equation, our results for Darcy flow still suggest that the expected relative test error converges to a constant value as resolution increases; an estimate of this rate of convergence is seen in Figure 2.9b, where we plot the relative error of the learned parameter $\alpha^{(r)}$ at resolution r w.r.t. the parameter learned at the highest resolution trained, which was $r = 129$. Although we do not observe the limiting error following the Monte Carlo rate in m, which suggests that the



(a) (b)

Figure 2.8: Expected relative test error of a trained RFM for Darcy flow with $n' = 1000$ test pairs: Figure 2.8a displays the invariance of test error w.r.t. training and testing on different resolutions for $m = 512$ and $n = 256$ fixed; the RFM can train and test on different mesh sizes without significant loss of accuracy. Figure 2.8b shows the decay of the test error for resolution $r = 33$ fixed as a function of m and n; the smallest error achieved is 0.0381 for $n = 500$ and $m = 512$.

RKHS $\mathcal{H}_{k_\mu}$ induced by the choice of φ may not be expressive enough (e.g., not universal [255]), the numerical results make clear that our method nonetheless performs well as an operator approximator.

2.5 Conclusion

This chapter introduces a random feature methodology for the data-driven estimation of operators mapping between infinite-dimensional Banach spaces. It may be interpreted as a low-rank approximation to operator-valued kernel ridge regression. Training the function-valued random features only requires solving a quadratic optimization problem for an m-dimensional coefficient vector. The conceptually infinite-dimensional algorithm is nonintrusive and results in a scalable method that is consistent with the continuum limit, robust to discretization, and highly flexible in practical use cases. Numerical experiments confirm these benefits in scientific machine learning applications involving two nonlinear forward operators arising from PDEs. Backed by tractable training routines and theoretical guarantees, operator learning with the function-valued random features method displays considerable potential for accelerating many-query computational tasks and for discovering new models from high-dimensional experimental data in science and engineering.

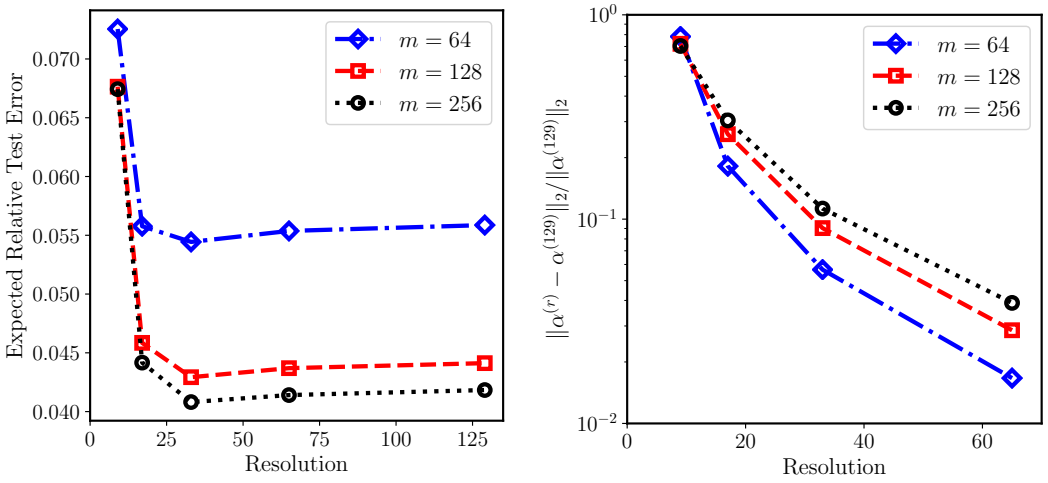


Figure 2.9: Results of a trained RFM for Darcy flow: Here, $n = 128$ training and $n' = 1000$ testing pairs were used. Figure 2.9a demonstrates resolution-invariant test error for various m, while Figure 2.9b displays the relative error of the learned coefficient $\alpha^{(r)}$ at resolution r w.r.t. the coefficient learned on the highest resolution ($r = 129$).

Going beyond this chapter, several directions for future research remain open. It is of interest to characterize the quality of the operator RKHS spaces induced by random feature pairs and whether practical problem classes actually belong to these spaces. This would be both mathematically interesting and highly desirable as it would help guide algorithmic development. Also of importance is the question of how to automatically adapt function-valued random features to data instead of manually constructing them. Some possibilities along this line of work include the Bayesian estimation of hyperparameters [85], as is frequently used in Gaussian process regression, or more general hierarchical learning of the random feature pair (φ, μ) itself. In tandem, there is a need for a mature function-valued random features software library that includes efficient linear solvers and GPU implementations, benchmark problems, and robust hyperparameter optimizers. These advances will further enable the random features method to learn from real-world data and solve challenging forward and inverse problems from the physical sciences, such as climate modeling and material modeling, with controlled computational complexity.

Chapter 3

ERROR BOUNDS FOR FUNCTION-VALUED RANDOM FEATURES

This chapter is adapted from the following publication:

[1] Samuel Lanthaler and Nicholas H. Nelsen. "Error bounds for learning with vector-valued random features". In: *Advances in Neural Information Processing Systems* (spotlight paper). Ed. by A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 71834–71861. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/e34d908241aef40440e61d2a27715424-Abstract-Conference.html.

This chapter provides a comprehensive error analysis of learning with vector-valued random features (RF). The theory is developed for RF ridge regression in a fully general infinite-dimensional input-output setting—which in particular includes function-valued RFs—but nonetheless applies to and improves existing finite-dimensional analyses. In contrast to comparable work in the literature, the approach proposed here relies on a direct analysis of the underlying risk functional and completely avoids the explicit RF ridge regression solution formula in terms of random matrices. This removes the need for suboptimal matrix concentration inequalities. The main results established in this chapter include strong consistency of vector-valued RF estimators under model misspecification and minimax optimal convergence rates in the well-specified setting. The parameter complexity (number of random features) and sample complexity (number of labeled data) required to achieve such rates are comparable with Monte Carlo intuition and free from logarithmic factors for the first time.

3.1 Introduction

Supervised learning of an unknown mapping $\mathcal{G}\colon \mathcal{X} \to \mathcal{Y}$ is a core task in machine learning. The *random feature model* (RFM), proposed in [221, 223], combines randomization with optimization to accomplish this task. The RFM is based on a linear expansion with respect to a randomized basis, the *random features*. The coefficients in this RF expansion are optimized to fit the given data of input-output pairs. For popular loss functions, such as the square

loss, the RFM leads to a convex optimization problem which can be solved efficiently and reliably.

The RFM provides a scalable approximation of an underlying kernel method [221, 223]. While the former is based on an expansion in M random features $\varphi(\,\cdot\,;\theta_1),\dots,\varphi(\,\cdot\,;\theta_M)$, the corresponding kernel method relies on an expansion in values of a positive definite kernel function $K(\,\cdot\,,u_1),\dots,K(\,\cdot\,,u_N)$ on a dataset of size N. Kernel methods are conceptually appealing, theoretically sound, and have attracted considerable interest [14, 52, 240]. However, they require the storage, manipulation, and often inversion of the kernel matrix K with entries $K(u_i,u_j)$. The size of K scales quadratically in the number of samples N, which can be prohibitive for large datasets. When the underlying input-output map is vector-valued with $\dim(\mathcal{Y})=p$, the often significant computational cost of kernel methods is further exacerbated by the fact that each entry $K(u_i,u_j)$ of K is, in general, a p-by-p matrix. Hence, the size of K scales quadratically in both N and p. This severely limits the applicability of kernel methods to problems with high-dimensional, or indeed infinite-dimensional, output space. In contrast, learning with RF only requires storage of RF matrices whose size is quadratic in the number of features M. When $M \ll Np$, this implies substantial computational savings, with the most extreme case being the infinite-dimensional setting in which $p=\infty$.

In the context of operator learning, the underlying target mapping is an operator $\mathcal{G}\colon \mathcal{X}\to\mathcal{Y}$ with infinite-dimensional input and output spaces. Such operators appear naturally in scientific computing and often arise as solution maps of an underlying partial differential equation. Operator learning has attracted considerable interest, e.g., [34, 111, 126, 172, 181], and in this context, the RFM serves as an alternative to neural network-based methods with considerable potential for a sound theoretical basis. Indeed, an extension of the RFM to this infinite-dimensional setting is proposed and implemented in the previous chapter [202, 203]. Although the results show promise, a mathematical analysis of this approach including error bounds and rates of convergence has so far been outstanding.

Related Work. Several papers have derived error bounds for learning with RF. Early work on the RFM [223] proceeded by direct inspection of the risk functional, demonstrating that $M \simeq N$ random features suffice to achieve

a squared error $O(1/\sqrt{N})$ for RF ridge regression (RR). This result was considerably improved in [235], where $\sqrt{N}\log N$ random features were shown to be sufficient to achieve the same squared error. This improvement in parameter complexity is based on the explicit RF RR solution formula, combined with extensive use of matrix analysis and matrix concentration inequalities. Similar analysis in [170] sharpens the parameter complexity to $\sqrt{N}\log d_{\mathsf{K}}^{\lambda}$ random features. Here d_{K}^{λ} is the *number of effective degrees of freedom* [17, 52], with λ the RR regularization parameter and K the kernel matrix. In this context, we also mention related analysis in [17]. In all these works, the squared error in terms of sample size N match the minimax optimal rates for kernel RR derived in [52]. Going beyond the above error bounds, [17, 170, 235] also derive fast rates under additional assumptions on the underlying data distribution and/or with improved RF sampling schemes.

Many works also study the interpolatory ($M \simeq N$) or overparametrized ($M \gg N$) regimes in the scalar output setting [60, 91, 107, 122, 190]. However, when $p = \dim(\mathcal{Y}) \gg 1$ or $p = \infty$, such regimes may no longer be relevant. This is because the kernel matrix K now has size Np-by-Np, and it is possible that the number of random features M satisfies $M \ll Np$ even though $M \gg N$. In this case, high-dimensional vector-valued learning naturally operates in the underparametrized regime.

In the area of operator learning for scientific problems, approximation results are common [64, 81, 125, 114, 152, 160, 78, 241] but statistical guarantees are lacking, the main exceptions being [42, 76, 135, 196, 239, 251] in the linear operator setting and [52] in the nonlinear setting. The RFM also has potential for such nonlinear problems. Indeed, vector-valued random Fourier features have been studied before [43, 193]. However, theory is only provided for kernel approximation, not generalization guarantees.

To summarize, while previous analyses have provided considerable insight into the generalization properties of the RFM, they have almost exclusively focused on the scalar-valued setting. Given the paucity of theoretical work beyond this setting, it is *a priori* unclear whether similar estimates continue to hold when the RFM is applied to infinite-dimensional vector-valued mappings. This includes the function-valued random features method from Chapter 2.

Table 3.1: A summary of available error estimates for the RFM, with regularization parameter λ, output space $\mathcal{Y}$, and number of random features M. (†): the truth is assumed to be written as $\mathcal{G}(u) = \mathbb{E}_\theta[\alpha^\star(\theta)\varphi(u;\theta)]$ with restrictive almost sure bound $|\alpha^\star(\theta)| \leq R$ to avoid explicit regularization.

Paper	Approach	λ	$\dim(\mathcal{Y})$	M	Squared Error
Rahimi & Recht [223]	"kitchen sinks"	n/a (†)	1	N	$R/\sqrt{N}$
Rudi & Rosasco [235]	matrix concen.	$1/\sqrt{N}$	1	$\sqrt{N}\log(N)$	$1/\sqrt{N}$
Li et al. [170]	matrix concen.	$1/\sqrt{N}$	1	$\sqrt{N}\log(d_{\mathsf{K}}^\lambda)$	$1/\sqrt{N}$
This work	**"kitchen sinks"**	$\mathbf{1/\sqrt{N}}$	$\boldsymbol{\infty}$	$\boldsymbol{\sqrt{N}}$	$\mathbf{1/\sqrt{N}}$

Contributions. The primary purpose of the present chapter is to extend earlier results on learning with random features to the vector-valued setting. The theory developed in this work unifies sources of error stemming from approximation, generalization, misspecification, and noisy observations. We focus on training via ridge regression with the square loss. Our results differ from existing work not only in the scope of applicability, but also in the strategy employed to derive our results. Similar to [223], we do not rely on the explicit random feature ridge regression solution (RF-RR) formula, which is specific to the square loss. One main benefit of this approach is that it entirely avoids the use of matrix concentration inequalities, thereby making the extension to an infinite-dimensional vector-valued setting straightforward. Our main contributions are now listed (see also Table 3.1).

(C1) Given N training samples, we prove that $M \simeq \sqrt{N}$ random features and regularization strength $\lambda \simeq 1/\sqrt{N}$ is enough to guarantee that the squared error is $O(1/\sqrt{N})$, provided that the target operator belongs to a specific reproducing kernel Hilbert space (Theorem 3.9);

(C2) we establish that the vector-valued RF-RR estimator is strongly consistent (Theorem 3.12);

(C3) under additional regularity assumptions, we derive rates of convergence even when the target operator does not belong to the specific reproducing kernel Hilbert space (Theorem 3.14);

(C4) we demonstrate that the approach of Rahimi and Recht [223] can be used to derive state-of-the-art rates for the RFM which, for the first time, are free from logarithmic factors.

Outline. The remainder of this chapter is organized as follows. We set up the ridge regression problem in Section 3.2. The main results are stated in Section 3.3 and their proofs are sketched in Section 3.4. Section 3.5 provides a simulation study and Section 3.6 gives concluding remarks. Detailed proofs are deferred to Appendix B.

3.2 Preliminaries

We now set up our vector-valued learning framework by introducing notational conventions, reviewing random features, and formulating the ridge regression problem.

Notation. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a sufficiently rich probability space on which all random variables in this chapter are defined. Let $\mathcal{X}$ be the input space, $\mathcal{Y}$ the output space, and Θ a set. We consistently use u to denote elements of $\mathcal{X}$ and θ for RF parameters in Θ. The set of probability measures supported on a set $\mathcal{Q}$ is denoted by $\mathscr{P}(\mathcal{Q})$. We write expectation (in the sense of Bochner integration) with respect to $u \sim \nu \in \mathscr{P}(\mathcal{X})$ as $\mathbb{E}_u[\,\cdot\,]$ and similarly for $\theta \sim \mu \in \mathscr{P}(\Theta)$. Independent and identically distributed (i.i.d.) samples $u_1, \dots, u_N$ from ν will be denoted by $\{u_n\} \sim \nu^{\otimes N}$ and similarly for $\{\theta_m\} \sim \mu^{\otimes M}$. We write $a \simeq b$ to mean that there exists a constant $C \geq 1$ such that $C^{-1}b \leq a \leq Cb$ and similarly for the one-sided inequalities $a \lesssim b$ and $a \gtrsim b$.

Random Features and Reproducing Kernel Hilbert Spaces. Random features are defined by a pair (φ, μ), where $\varphi\colon \mathcal{X} \times \Theta \to \mathcal{Y}$ and $\mu \in \mathscr{P}(\Theta)$. Fixing $\theta \sim \mu$ defines a map $\varphi(\,\cdot\,;\theta)\colon \mathcal{X} \to \mathcal{Y}$. Considering linear combinations of such maps leads to the following definition.

Definition 3.1 (random feature model)**.** The map $\Phi(\,\cdot\,;\alpha, \{\theta_m\})$, which we denote by $\Phi(\,\cdot\,;\alpha)\colon \mathcal{X} \to \mathcal{Y}$, given by

$$u \mapsto \Phi(u;\alpha) \coloneqq \frac{1}{M} \sum_{m=1}^{M} \alpha_m \varphi(u;\theta_m) \tag{3.1}$$

is a *random feature model* (RFM) with coefficients $\alpha \in \mathbb{R}^M$ and fixed realizations $\{\theta_m\} \sim \mu^{\otimes M}$.

Associated to the pair (φ, μ) is a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of maps from $\mathcal{X}$ to $\mathcal{Y}$ [203, Section 2.3]. Under mild assumptions (see

Appendix B.2) assumed in our main results, it holds that

$$\mathcal{H} = \left\{ \mathcal{G} \in L^2_\nu(\mathcal{X}; \mathcal{Y}) \,\middle|\, \mathcal{G} = \mathbb{E}_{\theta\sim\mu}[\alpha(\theta)\varphi(\,\cdot\,;\theta)] \text{ and } \alpha \in L^2_\mu(\Theta;\mathbb{R}) \right\} \tag{3.2}$$

with RKHS norm $\|\mathcal{G}\|_{\mathcal{H}} = \min_\alpha \|\alpha\|_{L^2_\mu}$, where α ranges over all decompositions of $\mathcal{G}$ of the form in (3.2). A minimizer $\alpha_{\mathcal{H}}$ of this problem always exists [17, Section 2.2]. We use this fact to identify any $\mathcal{G} \in \mathcal{H}$ with its minimizing $\alpha_{\mathcal{H}} \in L^2_\mu(\Theta;\mathbb{R})$ without further comment.

Random Feature Ridge Regression. Let $\mathcal{P} \in \mathscr{P}(\mathcal{X} \times \mathcal{Y})$ be the *joint data distribution*. The goal of RF-RR is to estimate an underlying operator $\mathcal{G}\colon \mathcal{X} \to \mathcal{Y}$ from finitely many i.i.d. input-output pairs $\{(u_n, y_n)\}_{n=1}^N \sim \mathcal{P}^{\otimes N}$, where typically the y_n are noisy transformations of the point values $\mathcal{G}(u_n)$. To describe RF-RR, we first make some definitions.

Definition 3.2 (empirical risk)**.** Writing $Y = \{y_n\}$ for the collection of observed output data and fixing a regularization parameter $\lambda > 0$, the *regularized Y-empirical risk* of $\alpha \in \mathbb{R}^M$ is given by[1]

$$\mathscr{R}_N^\lambda(\alpha; Y) \coloneqq \frac{1}{N}\sum_{n=1}^N \|y_n - \Phi(u_n;\alpha)\|_{\mathcal{Y}}^2 + \lambda\|\alpha\|_M^2\,, \quad \text{where} \tag{3.3a}$$

$$\|\alpha\|_M^2 \coloneqq \frac{1}{M}\sum_{m=1}^M |\alpha_m|^2 \tag{3.3b}$$

is a scaled Euclidean norm on $\mathbb{R}^M$. The *regularized $\mathcal{G}$-empirical risk*, $\mathscr{R}_N^\lambda(\alpha;\mathcal{G})$, is defined analogously with $\mathcal{G}(u_n)$ in place of y_n. In the absence of regularization, i.e., $\lambda = 0$, these expressions define the *Y-empirical risk* and *$\mathcal{G}$-empirical risk*, denoted by $\mathscr{R}_N(\alpha; Y)$ and $\mathscr{R}_N(\alpha;\mathcal{G})$, respectively.

RF-RR is the minimization problem $\min_{\alpha\in\mathbb{R}^M} \mathscr{R}_N^\lambda(\alpha; Y)$. The minimizer, which we denote by $\widehat{\alpha}$, is referred to as *trained coefficients* and $\Phi(\,\cdot\,;\widehat{\alpha})$ the *trained RFM*. For M and N sufficiently large and $\lambda > 0$ sufficiently small, we expect the trained RFM to well approximate $\mathcal{G}$. This intuition is made precise by quantitative error bounds and statistical performance guarantees in the next section.

[1] Note that λ in (3.3) is equal to N^{-1} times the regularization parameter from Chapter 2, which is denoted by the same symbol λ.

3.3 Main Results

The main result of this chapter is an abstract bound on the population squared error (Section 3.3.2). From this widely applicable theorem, several more specialized results are deduced. These include consistency (Section 3.3.3) and convergence rates (Section 3.3.4) of the RF-RR estimator trained on noisy data. The assumptions under which this theory is developed are provided next in Section 3.3.1.

3.3.1 Assumptions

Throughout this chapter, we assume that the input space $\mathcal{X}$ is a Polish space and the output space $\mathcal{Y}$ is a real separable Hilbert space. These are common assumptions in learning theory [52]. We view $\mathcal{X}$ and $\mathcal{Y}$ as measurable spaces equipped with their respective Borel σ-algebras.

Next, we make the following minimal assumptions on the random feature pair (φ, μ).

Assumption 3.3 (random feature regularity)**.** *Let $\nu \in \mathscr{P}(\mathcal{X})$ be the input distribution and (Θ, Σ, μ) be a probability space. The random feature map $\varphi\colon \mathcal{X} \times \Theta \to \mathcal{Y}$ and the probability measure $\mu \in \mathscr{P}(\Theta)$ are such that (i) φ is measurable; (ii) φ is uniformly bounded; in fact, $\|\varphi\|_{L^\infty} \coloneqq \operatorname{ess\,sup}_{(u,\theta)\sim\nu\otimes\mu} \|\varphi(u;\theta)\|_{\mathcal{Y}} \leq 1$; and (iii) the RKHS $\mathcal{H}$ corresponding to (φ, μ) is separable.*

The boundedness assumption on φ is shared in general theoretical analyses of RF [170, 223, 235]; the unit bound can always be ensured by a simple rescaling. We work in a general misspecified setting.

Assumption 3.4 (misspecification)**.** *There exist $\rho \in L^\infty_\nu(\mathcal{X};\mathcal{Y})$ and $\mathcal{G}_\mathcal{H} \in \mathcal{H}$ such that the operator $\mathcal{G}\colon \mathcal{X} \to \mathcal{Y}$ satisfies the decomposition $\mathcal{G} = \rho + \mathcal{G}_\mathcal{H}$.*

Since Assumption 3.3 implies that $\mathcal{H} \subset L^\infty_\nu(\mathcal{X};\mathcal{Y})$, any $\mathcal{G} = \mathcal{G}_\mathcal{H} + \rho$ as in Assumption 3.4 is automatically bounded in the sense that $\mathcal{G} \in L^\infty_\nu(\mathcal{X};\mathcal{Y})$. Conversely, any $\mathcal{G} \in L^\infty_\nu(\mathcal{X};\mathcal{Y})$ allows such a decomposition. We interpret ρ as a residual from the operator $\mathcal{G}_\mathcal{H}$ belonging to the RKHS. It may be prescribed by the problem, as we will see later in the context of discretization errors in operator learning (Example 3.11), or be arbitrary, as is customary in learning theory when the only information about $\mathcal{G}$ is that it is bounded.

Our main goal is to recover $\mathcal{G}$ from i.i.d. data $\{(u_n, y_n)\}$ arising from the following statistical model.

Assumption 3.5 (joint data distribution)**.** *The joint distribution $\mathcal{P} \in \mathscr{P}(\mathcal{X} \times \mathcal{Y})$ of the random variable $(u, y) \sim \mathcal{P}$ is given by $u \sim \nu$ with $\nu \in \mathscr{P}(\mathcal{X})$ and $y = \mathcal{G}(u) + \eta$. Here, $\mathcal{G}$ satisfies Assumption 3.4. The additive noise η is a random variable in $\mathcal{Y}$ that is conditionally centered, $\mathbb{E}[\eta \,|\, u] = 0$, and is subexponential: $\|\eta\|_{\psi_1(\mathcal{Y})} < \infty$; see (B.7) for the definition of $\|\cdot\|_{\psi_1(\mathcal{Y})}$.*

Assumption 3.5 implies that $\mathcal{G}(u) = \mathbb{E}[y \,|\, u]$. In contrast to related work [17, 170, 223], we allow for unbounded input-dependent noise. In particular, our results also hold for bounded or subgaussian noise, as well as *multiplicative noise* (e.g., $\eta = \xi\mathcal{G}(u)$ with $\mathbb{E}[\xi \,|\, u] = 0$ and $\|\xi\|_{\psi_1(\mathbb{R})} < \infty$).

3.3.2 General Error Bound

For any $\mathcal{G}$, define the *$\mathcal{G}$-population risk functional* or *$\mathcal{G}$-population squared error* by

$$\mathscr{R}(\alpha; \mathcal{G}) \coloneqq \mathbb{E}_{u\sim\nu}\|\mathcal{G}(u) - \Phi(u; \alpha, \{\theta_m\})\|_{\mathcal{Y}}^2 \quad \text{for} \quad \alpha \in \mathbb{R}^M . \tag{3.4}$$

The main result of this chapter establishes an upper bound for this quantity that holds with high probability, provided that the number of random features and number of data pairs are large enough.

Theorem 3.6 ($\mathcal{G}$-population squared error bound)**.** *Suppose that $\mathcal{G} = \rho + \mathcal{G}_{\mathcal{H}}$ satisfies Assumption 3.4. Fix a failure probability $\delta \in (0, 1)$, regularization strength $\lambda \in (0, 1)$, and sample size N. Let $\{\theta_m\} \sim \mu^{\otimes M}$ be the M random feature parameters and $\{(u_n, y_n)\} \sim \mathcal{P}^{\otimes N}$ be the data according to Assumption 3.5. For Φ the RFM (3.1) satisfying Assumption 3.3, let $\widehat{\alpha} \in \mathbb{R}^M$ be the minimizer of the regularized Y-empirical risk $\mathscr{R}_N^\lambda(\,\cdot\,; Y)$ given by (3.3). If $M \geq \lambda^{-1}\log(32/\delta)$ and $N \geq \lambda^{-2}\log(16/\delta)$, then*

$$\mathbb{E}_{u\sim\nu}\|\mathcal{G}(u) - \Phi(u; \widehat{\alpha}, \{\theta_m\})\|_{\mathcal{Y}}^2 \leq 79e^{3/2}\big(\|\mathcal{G}\|_{L^\infty_\nu}^2 + 2\beta(\rho, \lambda, \mathcal{G}_{\mathcal{H}}, \eta)\big)\lambda \tag{3.5}$$

with probability at least $1 - \delta$, where

$$\beta(\rho, \lambda, \mathcal{G}_{\mathcal{H}}, \eta) \coloneqq 328\|\mathcal{G}_{\mathcal{H}}\|_{\mathcal{H}}^2 + 2023e^3\|\eta\|_{\psi_1(\mathcal{Y})}^2 + 8\lambda^{-1}\,\mathbb{E}_{u\sim\nu}\|\rho(u)\|_{\mathcal{Y}}^2 + 18\lambda\|\rho\|_{L^\infty_\nu}^2 \tag{3.6}$$

is a function of ρ, λ, $\mathcal{G}_{\mathcal{H}}$, and the law of the noise variable η.

The main elements of the proof of Theorem 3.6 will be explained in Section 3.4.

Remark 3.7 (excess risk)**.** *We note that other work [170, 223, 235] often focuses on bounding the* excess risk $\widehat{\mathscr{E}} \coloneqq \mathscr{E}(\Phi(\,\cdot\,;\widehat{\alpha})) - \inf_{\mathcal{G}_\mathcal{H}\in\mathcal{H}} \mathscr{E}(\mathcal{G}_\mathcal{H})$*, where* $\mathscr{E}(F) \coloneqq \mathbb{E}\,\|y - F(u)\|_\mathcal{Y}^2 = \mathbb{E}_{u\sim\nu}\,\|\mathcal{G}(u) - F(u)\|_\mathcal{Y}^2 + \mathbb{E}\,\|\eta\|_\mathcal{Y}^2$*. In particular, this bias-variance decomposition implies that* $\widehat{\mathscr{E}} \leq \mathbb{E}_{u\sim\nu}\|\mathcal{G}(u) - \Phi(u;\widehat{\alpha})\|_\mathcal{Y}^2$*. Thus, Theorem 3.6 also gives a corresponding bound on the excess risk* $\widehat{\mathscr{E}}$*.*

Remark 3.8 (the factor β)**.** *In the well-specified setting, that is,* $\mathcal{G}-\mathcal{G}_\mathcal{H} = \rho \equiv 0$*, the factor* β *in Theorem (3.6) satisfies the uniform bound*

$$\beta(\rho, \lambda, \mathcal{G}_\mathcal{H}, \eta) \leq B \coloneqq 328\|\mathcal{G}\|_\mathcal{H}^2 + 2023e^3\|\eta\|_{\psi_1(\mathcal{Y})}^2\,. \tag{3.7}$$

In particular, the constant B *does not depend on* λ *in this case. Otherwise,* β *in general depends on* λ*. We can characterize this dependence precisely if it is known that* $\mathcal{G} \in L_\nu^\infty(\mathcal{X};\mathcal{Y})$*. In this case, Assumption 3.4 is satisfied with* $\rho \coloneqq \mathcal{G} - \mathcal{G}_\mathcal{H}$ for any $\mathcal{G}_\mathcal{H} \in \mathcal{H}$*. Choosing* $\mathcal{G}_\mathcal{H} = \mathcal{G}_\vartheta|_{\vartheta=\lambda}$ *as in Appendix B.2 (which is optimal in the sense described there) and a short calculation deliver the bound*

$$\beta(\rho, \lambda, \mathcal{G}_\mathcal{H}, \eta) \lesssim \lambda^{-1}\lambda^{\min(r,1)} = \lambda^{-(1-r)_+} \tag{3.8}$$

if $\mathcal{G}$ *additionally satisfies a particular* r*-th-order regularity condition (see Lemma B.8 for the details). Here,* $a_+ \coloneqq \max(a, 0)$ *for any* $a \in \mathbb{R}$*. Thus,* β *is uniformly bounded if* $\mathcal{G} \in \mathcal{H}$ $(r \geq 1)$ *and grows algebraically as a power of* λ^{-1} *otherwise* $(0 \leq r < 1)$*.*

Consequences. The general error bound (3.5) in Theorem 3.6 has several implications for vector-valued learning with the RFM. First, it immediately implies a rate of convergence if $\mathcal{G} \in \mathcal{H}$.

Theorem 3.9 (well-specified)**.** *Instantiate the hypotheses and notation of Theorem 3.6. Suppose that* $\rho \equiv 0$ *so that* $\mathcal{G} \in \mathcal{H}$ *(3.2). If* $M \geq \lambda^{-1}\log(32/\delta)$ *and* $N \geq \lambda^{-2}\log(16/\delta)$*, then*

$$\mathbb{E}_{u\sim\nu}\|\mathcal{G}(u) - \Phi(u;\widehat{\alpha})\|_\mathcal{Y}^2 \leq 79e^{3/2}\big(\|\mathcal{G}\|_{L_\nu^\infty}^2 + 2B\big)\lambda \lesssim \lambda \tag{3.9}$$

with probability at least $1 - \delta$*, where the constant* $B \geq 0$ *is defined by (3.7).*

Given a number of samples N, Theorem 3.9 shows that RF-RR with regularization $\lambda \simeq 1/\sqrt{N}$ and number of features $M \gtrsim \sqrt{N}$ leads to a population squared

error of size $1/\sqrt{N}$ with high probability. This result should be compared to the previous state-of-the-art convergence rates in the literature for RF-RR with i.i.d. sampled features [17, 170, 223, 235]. See Table 3.1, which indicates that our analysis gives the lowest parameter complexity to date. We emphasize that such a convergence rate rests on the assumption that $\mathcal{G} \in \mathcal{H}$. This corresponds to a *compatibility condition* between $\mathcal{G}$ and the pair (φ, μ), i.e., the random feature map φ and the probability measure μ, which determine the RKHS $\mathcal{H}$. Designing suitable φ and μ for a given operator $\mathcal{G}$ remains an open problem. For an explanation of the poor parameter complexity in Rahimi and Recht's original paper [223], see the work of Sun, Gilbert, and Tewari [254, Appendix E].

Theorem 3.6 also implies convergence of $\mathscr{R}(\widehat{\alpha}; \mathcal{G})$ when $\mathcal{G} \notin \mathcal{H}$, as we will see in Sections 3.3.3 and 3.3.4. But first, the next corollary shows that the same general bound (3.5) also holds for the $\mathcal{G}_{\mathcal{H}}$-population squared error $\mathscr{R}(\widehat{\alpha}; \mathcal{G}_{\mathcal{H}})$, up to enlarged constant factors. The proof is given in Appendix B.3.

Corollary 3.10 ($\mathcal{G}_{\mathcal{H}}$-population squared error bound)**.** *Instantiate the hypotheses and notation of Theorem 3.6. If $M \geq \lambda^{-1} \log(32/\delta)$ and $N \geq \lambda^{-2} \log(16/\delta)$, then there exists an absolute constant $C > 1$ such that with probability at least $1 - \delta$, it holds that*

$$\mathbb{E}_{u \sim \nu} \|\mathcal{G}_{\mathcal{H}}(u) - \Phi(u; \widehat{\alpha})\|_{\mathcal{Y}}^2 \leq C\big(\|\mathcal{G}\|_{L^\infty_\nu}^2 + 2\beta(\rho, \lambda, \mathcal{G}_{\mathcal{H}}, \eta)\big)\lambda\,. \tag{3.10}$$

Although our main goal is to learn $\mathcal{G}$ from noisy data, there are settings instead in which the learning of $\mathcal{G}_{\mathcal{H}} \in \mathcal{H}$ as in Corollary 3.10 is of primary interest, but only values of some approximation $\mathcal{G} \in L^\infty_\nu(\mathcal{X}; \mathcal{Y})$ are available. The following example illustrates this.

Example 3.11 (numerical discretization error)**.** One practically relevant setting to which Corollary 3.10 applies arises when training a RFM from functional data generated by a numerical approximation $\mathcal{G} = \mathcal{G}_\Delta$ of some underlying operator $\mathcal{G}_{\mathcal{H}} \in \mathcal{H}$. Here $\Delta > 0$ represents a numerical parameter, such as the grid resolution when approximating the solution operator of a partial differential equation. In this setting, $\rho = \mathcal{G}_\Delta - \mathcal{G}_{\mathcal{H}}$ is nonzero and it is crucial to include the discretization error in the analysis, which we define as $\varepsilon_\Delta := \|\rho\|_{L^\infty_\nu}^2 = \|\mathcal{G}_\Delta - \mathcal{G}_{\mathcal{H}}\|_{L^\infty_\nu}^2$. Assume $\eta \equiv 0$, so that $\widehat{\alpha}$ minimizes $\mathscr{R}_N^\lambda(\,\cdot\,; Y) = \mathscr{R}_N^\lambda(\,\cdot\,; \mathcal{G}_\Delta)$. Using Corollary 3.10, it follows that for M and N sufficiently large,

$$\mathbb{E}_{u \sim \nu} \|\mathcal{G}_{\mathcal{H}}(u) - \Phi(u; \widehat{\alpha})\|_{\mathcal{Y}}^2 \lesssim \lambda \|\mathcal{G}_{\mathcal{H}}\|_{\mathcal{H}}^2 + \varepsilon_\Delta \tag{3.11}$$

with high probability. Thus, as suggested by intuition, in addition to the error contribution that is present when training on perfect data (the first term on the right-hand side), there is an additional discretization error of size ε_Δ. We also see that the performance of RF-RR is stable with respect to such discretization errors stemming from the training data. Actually obtaining a rate of convergence would require problem-specific information about the particular numerical solver and discretization scheme that are used.

3.3.3 Statistical Consistency

We now return to the objective of recovering $\mathcal{G}$ from data. In particular, suppose that $\mathcal{G} \notin \mathcal{H}$; the RKHS, viewed as a hypothesis class, is misspecified. Our analysis demonstrates that statistical guarantees for RF-RR are still possible in this setting.

To this end, assume that $\mathcal{G} \in L^\infty_\nu(\mathcal{X};\mathcal{Y})$. It follows that Assumption 3.4 is satisfied with $\rho := \mathcal{G} - \mathcal{G}_\mathcal{H}$ and $\mathcal{G}_\mathcal{H} \in \mathcal{H}$ being *any* element of the RKHS. Applying Theorem 3.6 and minimizing over $\mathcal{G}_\mathcal{H}$ yields

$$\mathbb{E}_{u\sim\nu}\|\mathcal{G}(u) - \Phi(u;\widehat{\alpha})\|^2_\mathcal{Y} \lesssim \lambda + \inf_{\mathcal{G}_\mathcal{H}\in\mathcal{H}}\left\{\mathbb{E}_{u\sim\nu}\|\mathcal{G}(u) - \mathcal{G}_\mathcal{H}(u)\|^2_\mathcal{Y} + \lambda\|\mathcal{G}_\mathcal{H}\|^2_\mathcal{H}\right\} \tag{3.12}$$

with probability at least $1-\delta$ if $M \gtrsim \lambda^{-1}\log(2/\delta)$ and $N \gtrsim \lambda^{-2}\log(2/\delta)$. To obtain (3.12), we enlarged constants and used the bound $\|\rho\|^2_{L^\infty_\nu} \lesssim \|\mathcal{G}\|^2_{L^\infty_\nu} + \|\mathcal{G}_\mathcal{H}\|^2_{L^\infty_\nu}$ in (3.6).

If $\mathcal{G}$ is in the L^2_ν-closure of $\mathcal{H}$, then with high probability, the population squared error on the left-hand side of (3.12) converges to zero as $\lambda \to 0$ (by application of Lemma B.7 to the second term on the right). This is a statement about the (weak) *statistical consistency* of the trained RF-RR estimator; it can be upgraded to an almost sure statement, as expressed precisely in the next main result.

Theorem 3.12 (strong consistency)**.** *Suppose that $\mathcal{G} \in L^\infty_\nu(\mathcal{X};\mathcal{Y})$ belongs to the $L^2_\nu(\mathcal{X};\mathcal{Y})$-closure of $\mathcal{H}$ (3.2). Let $\{\lambda_k\}_{k\in\mathbb{N}} \subset (0,1)$ be a sequence of positive regularization parameters such that $\sum_{k\in\mathbb{N}}\lambda_k < \infty$. For Φ the RFM (3.1) satisfying Assumption 3.3 and for each k, let $\widehat{\alpha}^{(k)} \in \mathbb{R}^{M_k}$ be the trained RFM coefficients that minimize the regularized Y-empirical risk $\mathscr{R}^{\lambda_k}_{N_k}(\,\cdot\,;Y)$ given by (3.3) with $M_k \simeq \lambda_k^{-1}\log(2/\lambda_k)$ i.i.d. random features and $N_k \simeq \lambda_k^{-2}\log(2/\lambda_k)$ i.i.d. data pairs under Assumption 3.5. It holds true that*

$$\lim_{k\to\infty}\mathbb{E}_{u\sim\nu}\|\mathcal{G}(u) - \Phi(u;\widehat{\alpha}^{(k)})\|^2_\mathcal{Y} = 0 \quad \textit{with probability one.} \tag{3.13}$$

The proof relies on a standard Borel–Cantelli argument. See Appendix B.3 for the details.

Remark 3.13 (universal RKHS)**.** *The assumption that $\mathcal{G}$ belongs to the L^2_ν-closure of the RKHS $\mathcal{H}$ is automatically satisfied if $\mathcal{H}$ is dense in $L^2_\nu(\mathcal{X};\mathcal{Y})$. This is equivalent to its kernel being* universal *[53, 56]. In this case, the trained RFM is a strongly consistent estimator of any $\mathcal{G} \in L^\infty_\nu$. However, we are unaware of any practical characterizations of universality of the kernel in terms of its corresponding random feature pair (φ, μ) for the vector-valued setting studied here.*

3.3.4 Convergence Rates

The previous subsection establishes convergence guarantees without any rates. We now establish quantitative bounds. Throughout what follows, we denote by $\mathcal{K}\colon L^2_\nu(\mathcal{X};\mathcal{Y}) \to L^2_\nu(\mathcal{X};\mathcal{Y})$ the integral operator (B.16) corresponding to the operator-valued kernel function of the RKHS $\mathcal{H}$ (see Appendix B.2).

Theorem 3.14 (slow rates under misspecification)**.** *Suppose that $\mathcal{G} \in L^\infty_\nu(\mathcal{X};\mathcal{Y})$ and that Assumption 3.5 holds. Additionally, assume that $\mathcal{G} \in \operatorname{Im}(\mathcal{K}^{r/2})$ for some $r > 0$, where $\mathcal{K}$ is the integral operator corresponding to the kernel of RKHS $\mathcal{H}$ (3.2). Fix $\delta \in (0,1)$ and $\lambda \in (0,1)$. For Φ the RFM (3.1) satisfying Assumption 3.3, let $\widehat{\alpha} \in \mathbb{R}^M$ minimize $\mathscr{R}^\lambda_N(\,\cdot\,;Y)$ given by (3.3). If $M \geq \lambda^{-1}\log(32/\delta)$ and $N \geq \lambda^{-2}\log(16/\delta)$, then with probability at least $1-\delta$ it holds that*

$$\mathbb{E}_{u\sim\nu}\|\mathcal{G}(u) - \Phi(u;\widehat{\alpha})\|^2_{\mathcal{Y}} \lesssim \lambda^{\min(r,1)}\,. \tag{3.14}$$

The implied constant in (3.14) depends only on $\|\mathcal{G}\|_{L^\infty_\nu}$ and $\|\eta\|_{\psi_1(\mathcal{Y})}$.

Theorem 3.14 provides a quantitative convergence rate as $\lambda \to 0$. For $r \geq 1$, i.e., when $\mathcal{G} \in \mathcal{H}$, we recover the linear convergence rate of order λ from Theorem 3.9. The assumption that $\mathcal{G} \in \operatorname{Im}(\mathcal{K}^{r/2})$ can be viewed as a "fractional regularity" assumption on the underlying operator; indeed, in specific settings it corresponds to a fractional (Sobolev) regularity of the underlying function. In general, it appears difficult to check this condition in practice, which is one limitation of our result.

A quantitative analog to the almost sure statement of Theorem 3.12 also holds.

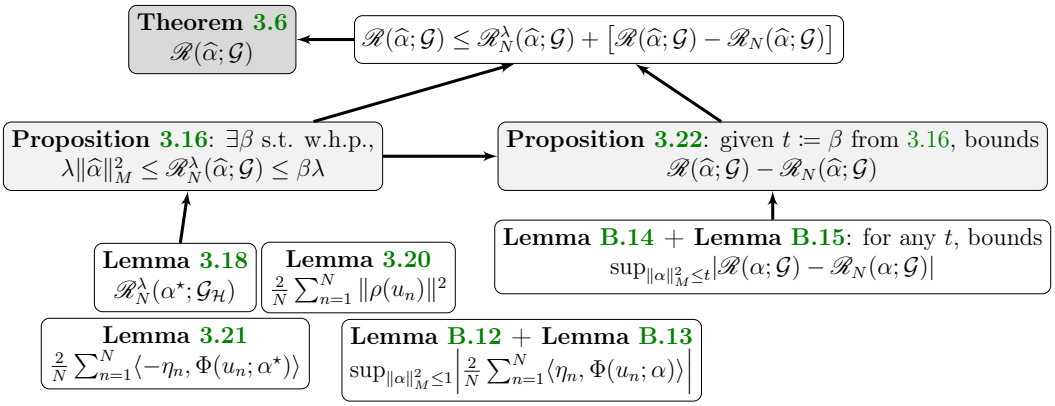


Figure 3.1: Flow chart illustrating the proof of Theorem 3.6.

Corollary 3.15 (strong convergence rate)**.** *Instantiate the hypotheses and notation of Theorem 3.12. Assume in addition that* $\mathcal{G} \in \mathrm{Im}(\mathcal{K}^{r/2})$ *for some* $r > 0$*. Let* $\{\lambda_k\}_{k\in\mathbb{N}} \subset (0,1)$ *be a sequence of positive regularization parameters such that* $\sum_{k\in\mathbb{N}} \lambda_k < \infty$*. For each* k*, let* $\widehat{\alpha}^{(k)} \in \mathbb{R}^{M_k}$ *be the trained RFM coefficients with* $M_k \simeq \lambda_k^{-1}\log(2/\lambda_k)$ *and* $N_k \simeq \lambda_k^{-2}\log(2/\lambda_k)$*. It holds true that*

$$\limsup_{k\to\infty} \left(\frac{\mathbb{E}_{u\sim\nu}\|\mathcal{G}(u) - \Phi(u;\widehat{\alpha}^{(k)})\|_\mathcal{Y}^2}{\lambda_k^{\min(r,1)}} \right) < \infty \quad \textit{with probability one.} \tag{3.15}$$

Short proofs of both Theorem 3.14 and Corollary 3.15 may be found in Appendix B.3.

3.4 Proof Outline for the Main Theorem

Our main results are all derived from Theorem 3.6, whose proof, schematically illustrated in Figure 3.1, we now outline. Following [223], we break the proof into several steps that arise from the error decomposition

$$\mathscr{R}(\widehat{\alpha};\mathcal{G}) = \mathscr{R}_N(\widehat{\alpha};\mathcal{G}) + \left[\mathscr{R}(\widehat{\alpha};\mathcal{G}) - \mathscr{R}_N(\widehat{\alpha};\mathcal{G})\right]. \tag{3.16}$$

Section 3.4.1 estimates the first term on the right-hand side of (3.16) while Section 3.4.2 estimates the second.

3.4.1 Bounding the Regularized Empirical Risk

The main technical contribution of this work is a tight bound on the $\mathcal{G}$-empirical risk $\mathscr{R}_N(\widehat{\alpha};\mathcal{G})$ for the trained RFM. The analysis involves controlling several sources of error and careful truncation arguments to avoid unnecessarily strong assumptions on the problem. The result is the following.

Proposition 3.16 (regularized $\mathcal{G}$-empirical risk bound)**.** *Let Assumptions 3.3 and 3.5 hold. Suppose that $\mathcal{G} = \rho + \mathcal{G}_{\mathcal{H}}$ satisfies Assumption 3.4. Fix $\delta \in (0,1)$, $\lambda \in (0,1)$, $M \in \mathbb{N}$, and $N \in \mathbb{N}$. Let $\widehat{\alpha} \in \mathbb{R}^M$ be the minimizer of the regularized Y-empirical risk $\mathscr{R}_N^\lambda(\,\cdot\,; Y)$ given by (3.3). If $M \geq \lambda^{-1}\log(16/\delta)$ and $N \geq \lambda^{-2}\log(8/\delta)$, then*

$$\mathscr{R}_N^\lambda(\widehat{\alpha}; \mathcal{G}) \leq \lambda\beta(\rho, \lambda, \mathcal{G}_{\mathcal{H}}, \eta) \tag{3.17}$$

with probability at least $1-\delta$, where the multiplicative factor $\beta(\rho, \lambda, \mathcal{G}_{\mathcal{H}}, \eta)$ is given by (3.6).

Since $\lambda\|\widehat{\alpha}\|_M^2 \leq \mathscr{R}_N^\lambda(\widehat{\alpha}; \mathcal{G})$, the next corollary controlling the norm (3.3) of $\widehat{\alpha}$ is immediate. It plays a crucial role in developing an upper bound for the second term on the right side of (3.16).

Corollary 3.17 (trained RFM norm bound)**.** *Instantiate the hypotheses and notation of Proposition 3.16. Fix $\delta \in (0,1)$ and $\lambda \in (0,1)$. If $M \geq \lambda^{-1}\log(16/\delta)$ and $N \geq \lambda^{-2}\log(8/\delta)$, then*

$$\widehat{\alpha} \in \mathcal{A}_\beta \coloneqq \left\{\alpha \in \mathbb{R}^M \,\middle|\, \|\alpha\|_M^2 \leq \beta\right\} \tag{3.18}$$

with probability at least $1-\delta$. The radius $\beta \coloneqq \beta(\rho, \lambda, \mathcal{G}_{\mathcal{H}}, \eta)$ of the norm bound is given by (3.6).

The core elements of the proof of Proposition 3.16 are provided in the next few subsections, with the full argument in Appendix B.4.1. The main idea is to upper bound the $\mathcal{G}$-empirical risk by its regularized counterpart and then decompose the latter into several (coupled) error contributions.

To do this, first fix any $\alpha \in \mathbb{R}^M$. It holds that

$$\mathscr{R}_N^\lambda(\alpha; Y) = \mathscr{R}_N^\lambda(\alpha; \mathcal{G}) + \frac{2}{N}\sum_{n=1}^N \langle \eta_n, \mathcal{G}(u_n) - \Phi(u_n; \alpha)\rangle_{\mathcal{Y}} + \frac{1}{N}\sum_{n=1}^N \|\eta_n\|_{\mathcal{Y}}^2 \tag{3.19}$$

because $\mathcal{Y}$ is a Hilbert space and $y_n = \mathcal{G}(u_n) + \eta_n$. Using this, a short calculation

shows that

$$\begin{aligned}\mathscr{R}_N^\lambda(\widehat{\alpha};\mathcal{G}) &= \big[\mathscr{R}_N^\lambda(\widehat{\alpha};Y) - \mathscr{R}_N^\lambda(\alpha;Y)\big] + \mathscr{R}_N^\lambda(\alpha;\mathcal{G}) \\ &\qquad + \frac{2}{N}\sum_{n=1}^N \langle \eta_n, \Phi(u_n;\widehat{\alpha}) - \Phi(u_n;\alpha)\rangle_{\mathcal{Y}} \\ &\leq \mathscr{R}_N^\lambda(\alpha;\mathcal{G}) + \frac{2}{N}\sum_{n=1}^N \langle -\eta_n, \Phi(u_n;\alpha)\rangle_{\mathcal{Y}} + \frac{2}{N}\sum_{n=1}^N \langle \eta_n, \Phi(u_n;\widehat{\alpha})\rangle_{\mathcal{Y}}\,. \end{aligned} \tag{3.20}$$

In the final line, we used the fact that $\widehat{\alpha}$ minimizes $\mathscr{R}_N^\lambda(\,\cdot\,;Y)$. Since $\alpha \in \mathbb{R}^M$ is arbitrary, we have the freedom to choose α so that the first term is small (see Section 3.4.1.1 and 3.4.1.2). With α fixed, the second term averages to zero by our assumptions on the noise, and hence, we expect it to be small with high probability (see Section 3.4.1.3).

The third term in (3.20) exhibits high correlation between the noise η_n and the trained RFM coefficients $\widehat{\alpha}$, making it more difficult to estimate. To control this last term, we first note that it is homogeneous in $\|\widehat{\alpha}\|_M$, which can be used to derive an upper bound in terms of a supremum over the unit ball with respect to $\|\cdot\|_M$. The resulting expression is then bounded with empirical process techniques (see Appendix B.4.1.3). For the complete details of the required argument we refer the reader to Appendix B.4.1.

In the remainder of this subsection, we estimate the first two terms on the right-hand side of (3.20). Using the fact that $\mathcal{G} = \rho + \mathcal{G}_{\mathcal{H}}$, the first term can be split into two contributions,

$$\mathscr{R}_N^\lambda(\alpha;\mathcal{G}) \leq 2\mathscr{R}_N^\lambda(\alpha;\mathcal{G}_{\mathcal{H}}) + \frac{2}{N}\sum_{n=1}^N \|\rho(u_n)\|_{\mathcal{Y}}^2\,. \tag{3.21}$$

These contributions to the first term in (3.20) are bounded in Section 3.4.1.1 and 3.4.1.2. The second term in (3.20) is controlled in Section 3.4.1.3.

3.4.1.1 Bounding the Approximation Error

We begin with the term $\mathscr{R}_N^\lambda(\alpha;\mathcal{G}_{\mathcal{H}})$, which may be viewed as an empirical *approximation error* due to α being arbitrary. Its only dependence on the data is through $\{u_n\}$ in (3.3). Intuitively, this term should behave like its population counterpart. It is then natural to choose a Monte Carlo approximation $\alpha_m =$

$\alpha_{\mathcal{H}}(\theta_m)$ for α, where $\alpha_{\mathcal{H}} \in L^2_\mu(\Theta;\mathbb{R})$ is identified with $\mathcal{G}_{\mathcal{H}}$ as in (3.2). However, our intuition that $\lambda\|\alpha\|_M^2$ should concentrate around $\lambda\|\alpha_{\mathcal{H}}\|^2_{L^2_\mu}$ fails because it is generally not possible to control the tail of the random variable $|\alpha_{\mathcal{H}}(\theta)|^2$. We next show that this problem can be overcome by a carefully tuned truncation argument combined with Bernstein's inequality.

Lemma 3.18 (construction of approximator)**.** *Suppose that $\mathcal{G}_{\mathcal{H}} \coloneqq \mathcal{G} \in \mathcal{H}$. Fix $\delta \in (0,1)$, $\lambda > 0$, $N \in \mathbb{N}$, and $M \in \mathbb{N}$. Let $\{\theta_m\} \sim \mu^{\otimes M}$ and $\{u_n\} \sim \nu^{\otimes N}$. Define $\alpha^\star \in \mathbb{R}^M$ componentwise by*

$$\alpha^\star_m \coloneqq \alpha_{\mathcal{H}}(\theta_m)\mathbb{1}_{\{|\alpha_{\mathcal{H}}(\theta_m)|\le T\}}, \quad \text{where} \quad T \coloneqq \sqrt{\lambda^{-1}\,\mathbb{E}_{\theta\sim\mu}\,|\alpha_{\mathcal{H}}(\theta)|^2} \tag{3.22}$$

and $\mathcal{G}_{\mathcal{H}} = \mathbb{E}_{\theta\sim\mu}[\alpha_{\mathcal{H}}(\theta)\varphi(\,\cdot\,;\theta)]$ with $\|\mathcal{G}_{\mathcal{H}}\|^2_{\mathcal{H}} = \mathbb{E}_{\theta\sim\mu}|\alpha_{\mathcal{H}}(\theta)|^2$. If $M \ge \lambda^{-1}\log(4/\delta)$, then with probability at least $1-\delta$ in the random feature parameters $\theta_1,\dots,\theta_M$, it holds that

$$\mathscr{R}^\lambda_N(\alpha^\star;\mathcal{G}_{\mathcal{H}}) \le 81\lambda\|\mathcal{G}_{\mathcal{H}}\|^2_{\mathcal{H}}. \tag{3.23}$$

Appendix B.4.1.1 provides the proof.

Remark 3.19 (well-specified and noise-free)**.** *Lemma 3.18 gives a $O(\lambda)$ bound on the regularized $\mathcal{G}_{\mathcal{H}}$-empirical risk of a RFM trained on well-specified and noise-free i.i.d. data $\{u_n, \mathcal{G}_{\mathcal{H}}(u_n)\}$.*

3.4.1.2 Bounding the Misspecification Error

The second contribution to (3.21) is easily bounded by Bernstein's inequality because $\rho \in L^\infty_\nu$. We refer the reader to Appendix B.4.1.2 for the detailed proof.

Lemma 3.20 (concentration of misspecification error)**.** *Let ρ be as in Assumption 3.4. Fix $\delta \in (0,1)$. With probability at least $1-\delta$, it holds that*

$$\frac{2}{N}\sum_{n=1}^{N}\|\rho(u_n)\|^2_{\mathcal{Y}} \le 4\,\mathbb{E}_{u\sim\nu}\|\rho(u)\|^2_{\mathcal{Y}} + \frac{9\|\rho\|^2_{L^\infty_\nu}\log(2/\delta)}{N}. \tag{3.24}$$

3.4.1.3 Bounding the Noise Error

The second term on the right-hand side of (3.20) is a zero-mean error contribution due to the noise corrupting the output training data. By the fact that η is subexponential (Assumption 3.5), Bernstein's inequality delivers exponential concentration. The proof details are in Appendix B.4.1.3.

Lemma 3.21 (concentration of noise error cross term)**.** *Let Assumptions 3.3 and 3.5 hold. Fix $\alpha \in \mathbb{R}^M$, $\{\theta_m\} \sim \mu^{\otimes M}$, and $\delta \in (0,1)$. With probability at least $1-\delta$, it holds that*

$$\frac{2}{N}\sum_{n=1}^{N}\langle -\eta_n, \Phi(u_n;\alpha)\rangle_{\mathcal{Y}} \leq 16e^{3/2}\|\eta_1\|_{\psi_1(\mathcal{Y})}\|\alpha\|_M\sqrt{\frac{\log(2/\delta)}{N}}\,. \tag{3.25}$$

Appendix B.4.1.3 also details the techniques used to upper bound the third and final term in (3.20).

3.4.2 Bounding the Generalization Gap

Having bounded the empirical risk with approximation arguments, it remains to control the estimation error $\mathscr{R}(\widehat{\alpha};\mathcal{G}) - \mathscr{R}_N(\widehat{\alpha};\mathcal{G})$ due to finite data in (3.16). We call this the *generalization gap*: the difference between the population test error and its empirical version. If $\widehat{\alpha}$ satisfies $\|\widehat{\alpha}\|_M^2 \leq t$ for some $t > 0$, then one can upper bound the generalization gap by its supremum over this set. The main challenge is to show existence of a (sufficiently small) t that satisfies this inequality. This is handled by Corollary 3.17. As summarized in the following proposition, the resulting supremum of the empirical process defined by the generalization gap is shown to be of size $N^{-1/2}$ with high probability.

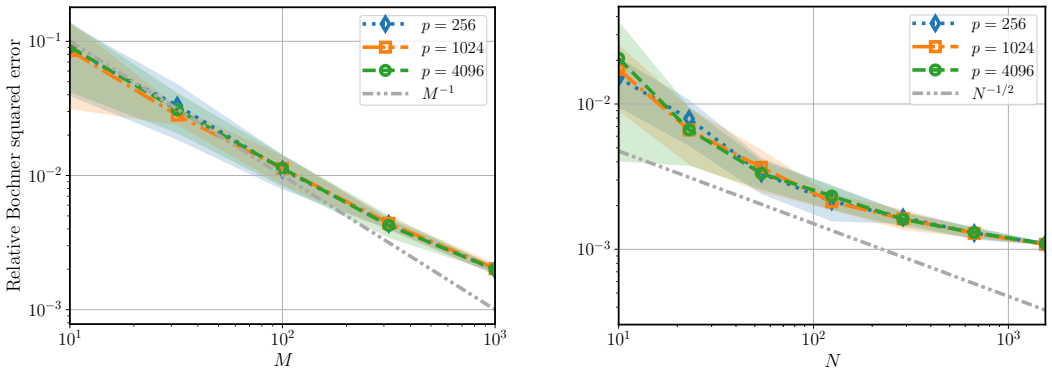
Proposition 3.22 (uniform bound on the generalization gap)**.** *Let Assumption 3.3 hold. Suppose $\mathcal{G}$ satisfies Assumption 3.4. Let $\{\theta_m\} \sim \mu^{\otimes M}$ for the RFM Φ given by (3.1). Fix $\delta \in (0,1)$. For i.i.d. input samples $\{u_n\} \sim \nu^{\otimes N}$, define the random variable*

$$\mathcal{E}_\beta\big(\{u_n\},\{\theta_m\}\big) \coloneqq \sup_{\alpha\in\mathcal{A}_\beta}\left|\frac{1}{N}\sum_{n=1}^{N}\|\mathcal{G}(u_n)-\Phi(u_n;\alpha)\|_{\mathcal{Y}}^2 - \mathbb{E}_u\|\mathcal{G}(u)-\Phi(u;\alpha)\|_{\mathcal{Y}}^2\right|, \tag{3.26}$$

where $\mathcal{A}_\beta \coloneqq \{\alpha' \in \mathbb{R}^M \,|\, \|\alpha'\|_M^2 \leq \beta\}$ and the deterministic radius $\beta = \beta(\rho,\lambda,\mathcal{G}_{\mathcal{H}},\eta)$ is given in (3.6) with $\mathcal{G}$ as above. If $N \geq \log(1/\delta)$, then with probability at least $1-\delta$ it holds that

$$\mathcal{E}_\beta\big(\{u_n\},\{\theta_m\}\big) \leq 32e^{3/2}\big(\|\mathcal{G}\|_{L^\infty_\nu}^2 + \beta(\rho,\lambda,\mathcal{G}_{\mathcal{H}},\eta)\big)\sqrt{\frac{6\log(2/\delta)}{N}}\,. \tag{3.27}$$

The proof of Proposition 3.22 is given in Appendix B.4.2. The argument is composed of two steps. The first is to show that $\mathcal{E}_\beta \,|\, \{\theta_m\}$ concentrates around its (conditional) expectation (Lemma B.14). This follows easily using



(a) Varying p, M, λ for fixed $N = 1548$.

(b) Varying p, N, λ for fixed $M = 10^4$.

Figure 3.2: Squared test error of trained RFM for learning the Burgers' equation solution operator. All shaded bands denote two empirical standard deviations from the empirical mean of the error computed over 10 different models, each with i.i.d. sampling of the features and training data indices.

the boundedness of the summands. The second step is to upper bound the expectation of $\mathcal{E}_\beta \,|\, \{\theta_m\}$ (Lemma B.15). This is achieved by exploiting the Hilbert space structure of the $\mathcal{Y}$-square loss and the linearity of the RFM with respect to its coefficients.

3.4.3 Combining the Bounds

Since we now have control over the $\mathcal{G}$-empirical risk and the generalization gap, the $\mathcal{G}$-population risk is also under control by (3.16). The proof of Theorem 3.6 follows by putting together the pieces (Appendix B.3).

3.5 Numerical Experiment

To study how our theory holds up in practice, we numerically implement the vector-valued RF-RR algorithm on a benchmark operator learning dataset[2]. The data $\{(u_n, \mathcal{G}(u_n))\}_{n=1}^N$ is noise-free, the $\{u_n\}$ are i.i.d. Gaussian random fields, and $\mathcal{G}\colon L^2(\mathbb{T};\mathbb{R}) \to L^2(\mathbb{T};\mathbb{R})$ is a nonlinear operator defined as the time one flow map of the viscous Burgers' equation on the torus $\mathbb{T}$. Appendix B.5 provides more details about the problem setting and the choice of random feature pair (φ, μ).

Figure 3.2a shows the decay of the relative squared test error as M increases

[2]The code is available at

https://github.com/nickhnelsen/error-bounds-for-vvRF .

(with $\lambda \simeq 1/M$) for fixed N. The error closely follows the rate $O(M^{-1})$ until it begins to saturate at larger M. This is due to either $\mathcal{G}$ not belonging to the RKHS of (φ, μ) or the finite data error dominating. As implied by our theory, the error does not depend on the discretized output dimension $p < \infty$. Figure 3.2b displays similar behavior as N is varied (now with $\lambda \simeq 1/\sqrt{N}$ and fixed M). Overall, the observed parameter and sample complexity reasonably validate our theoretical insights.

3.6 Conclusion

This chapter establishes several fundamental results for learning with infinite-dimensional vector-valued random features; these include strong consistency and explicit convergence rates. When the underlying mapping belongs to the RKHS, the rates obtained in this work match minimax optimal rates in the number of samples N, requiring only a number of random features $M \simeq \sqrt{N}$. Despite being derived in a very general setting, to the best of our knowledge, this provides the sharpest parameter complexity in M, which is free from logarithmic factors for the first time.

There are several interesting directions for future work. These include deriving fast rates for function-valued random features, relaxing the boundedness assumption on the features and the true mapping, and accommodating heavier-tailed or white noise distributions. Obtaining fast rates would require a sharpening of several estimates, and in particular, replacing the global Rademacher complexity-type estimate, implicit in our work, by its local counterpart. As our approach does not make use of an explicit solution formula, which is only available for a square loss, this might pave the way for improved rates for other loss functions, such as a general L^p-loss. We leave such potential extensions of the present approach for future research.

Chapter 4

LEARNING LINEAR OPERATORS FROM NOISY DATA

This chapter is adapted from the following publication:

[1] Maarten V. de Hoop, Nikola B. Kovachki, Nicholas H. Nelsen, and Andrew M. Stuart. "Convergence rates for learning linear operators from noisy data". *SIAM/ASA Journal on Uncertainty Quantification* 11.2 (2023), pp. 480–513. DOI: 10.1137/21M1442942.

This chapter studies the learning of linear operators between infinite-dimensional Hilbert spaces. The training data comprises pairs of random input vectors in a Hilbert space and their noisy images under an unknown self-adjoint linear operator. Assuming that the operator is diagonalizable in a known basis, this work solves the equivalent inverse problem of estimating the operator's eigenvalues given the data. Adopting a Bayesian approach, the theoretical analysis establishes posterior contraction rates in the infinite data limit with Gaussian priors that are not directly linked to the forward map of the inverse problem. The main results also include learning-theoretic generalization error guarantees for a wide range of distribution shifts. These convergence rates quantify the effects of data smoothness and true eigenvalue decay or growth, for compact or unbounded operators, respectively, on sample complexity. Numerical evidence supports the theory in diagonal and non-diagonal settings.

4.1 Introduction

The supervised learning of operators between Hilbert spaces provides a natural framework for the acceleration of scientific computation and discovery. This framework can lead to fast surrogate models that approximate expensive existing models or to the discovery of new models that are consistent with observed data when no first principles model exists. To develop some of the fundamental principles of operator learning, this chapter concerns (Bayesian) nonparametric linear regression under random design. Although the previous chapter obtains quite general error bounds for a class of nonlinear operator learning problems, the restriction to linear problems in the present chapter allows for a much sharper theoretical analysis.

To this end, let H be a real infinite-dimensional Hilbert space and L be an unknown—possibly unbounded and in general densely defined on H—self-adjoint linear operator from its domain in H into H itself. We study the following linear operator learning problem.

Main Problem. *Let $\{x_n\} \subset H$ be random design vectors and $\{\xi_n\}$ be noise vectors. Given the training data pairs $\{(x_n, y_n)\}_{n=1}^N$ with sample size $N \in \mathbb{N}$, where*

$$y_n = Lx_n + \gamma\xi_n \quad \textit{for} \quad n \in \{1, \dots, N\} \quad \textit{and} \quad \gamma > 0\,, \tag{4.1}$$

find an estimator $L^{(N)}$ of L that is accurate when evaluated outside of the samples $\{x_n\}$.

The estimation of L from the data (4.1) is generally an ill-posed linear inverse problem [79]. In principle, the chosen reconstruction procedure should be consistent: the estimator $L^{(N)}$ converges to the true L as $N \to \infty$. The rate of this convergence is equivalent to the *sample complexity* of the estimator, which determines the efficiency of statistical estimation. The sample complexity $N(\varepsilon) \in \mathbb{N}$ is the number of samples required for the estimator to achieve an error less than a fixed tolerance $\varepsilon > 0$. It quantifies the difficulty of Main Problem.

In modern scientific machine learning problems where operator learning is used, the demand on data from different operator learning architectures often outpaces the availability of computational or experimental resources needed to generate the data. Ideally, theoretical analysis of sample complexity should reveal guidelines for how to reduce the requisite data volume. To that end, the broad purpose of this chapter is to provide an answer to the question:

What factors can reduce sample size requirements for linear operator learning?

Our goal is not to develop a practical procedure to regress linear operators between infinite-dimensional vector spaces. Various methods already exist for that purpose, including those based on (functional) principal component analysis (PCA) [34, 69, 128]. Instead, we aim to strengthen the rather sparse but slowly growing theoretical foundations of operator learning.

We overview our approach to solve Main Problem in Section 4.1.1. We summarize one of our main convergence results in Section 4.1.2. In Section 4.1.3,

we illustrate examples to which our theory applies. Section 4.1.4 surveys work related to ours. The primary contributions of this chapter and its organization are given in Sections 4.1.5 and 4.1.6, respectively.

4.1.1 Key Ideas

In this subsection, we communicate the key ideas of our methodology at an informal level and distinguish our approach from similar ones in the literature.

4.1.1.1 Operator Learning as an Inverse Problem

We cast Main Problem as a Bayesian inverse problem with a *linear operator as the unknown object to be inferred from data*. Suppose the input training data $\{x_n\}$ from (4.1) are independent and identically distributed (i.i.d.) according to a (potentially unknown) centered probability measure ν on H with finite second moment. Let $\Lambda\colon H \to H$ be the covariance operator of ν with orthonormal eigenbasis $\{\phi_k\}$. Let the $\{\xi_n\}$ be i.i.d. $\mathcal{N}(0, \mathrm{Id}_H)$ Gaussian white noise processes independent of $\{x_n\}$. Writing $Y = (y_1, \ldots, y_N)$, $X = (x_1, \ldots, x_N)$, and $\Xi = (\xi_1, \ldots, \xi_N)$ yields the concatenated data model

$$Y = K_X L + \gamma \Xi\,. \tag{4.2}$$

The forward operator of this inverse problem is $K_X\colon T \mapsto (Tx_1, \ldots, Tx_N)$. Under a Gaussian prior $L \sim \mathcal{N}(0, \Sigma)$, the solution is the Gaussian posterior $L \,|\, (X, Y)$. For a fixed orthonormal basis $\{\varphi_j\}$ of H, it will be convenient to identify (4.2) with the countable inverse problem

$$y_{jn} = \sum_{k=1}^{\infty} x_{kn} \mathsf{L}_{jk} \; + \gamma \xi_{jn} \quad \text{for} \quad j \in \mathbb{N} \quad \text{and} \quad n \in \{1, \ldots, N\}\,, \tag{4.3}$$

where $\xi_{jn} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $x_{kn} = \langle \phi_k, x_n \rangle_H$, and $\mathsf{L}_{jk} = \langle \varphi_j, L\phi_k \rangle_H$. See Section 4.2.2.2 for details.

4.1.1.2 Comparison to Nonparametric Inverse Problems

In contrast, most theoretical studies of Bayesian inverse problems concern the *estimation of a vector* $f \in H_1$ from data

$$Y' = Kf + N^{-1/2}\xi\,, \quad \text{where} \quad \xi \sim \mathcal{N}(0, \mathrm{Id}_{H_2}) \tag{4.4}$$

and $K\colon H_1 \to H_2$ is a known bounded linear operator between Hilbert spaces H_1 and H_2. This is a signal in white noise model. Under a prior on f,

the asymptotic behavior of the posterior $f \,|\, Y'$ as the noise tends to zero ($N \to \infty$) is of primary interest. Many analyses of (4.4) consider the *singular value decomposition* (SVD) of K [10, 11, 57, 147, 230]. Projecting f into its coordinates $\{f_k\}$ in the basis of right singular vectors $\{\phi'_k\}$ of K and writing $\{Y'_j\}$ for observations of the stochastic process Y' on the basis of left singular vectors of K yields

$$Y'_j = \kappa_j f_j + N^{-1/2}\xi_j \quad \text{for} \quad j \in \mathbb{N}\,, \tag{4.5}$$

where the $\{\xi_j\}$ are i.i.d. $\mathcal{N}(0,1)$ and $\{\kappa_j\}$ are the singular values of K. Obtaining a sequence space model of this form is always possible if K is a compact operator [57, Section 1.2].

Some notable differences between the traditional inverse problem (4.4) and the operator learning inverse problem (4.2) are evident. Equation (4.2) is directly tied to (functional) regression, while (4.4) is not. The unknown f is a vector while L is an unknown operator. A more major distinction is that K in (4.4) is deterministic and arbitrary, while K_X in (4.2) is a *random forward map* defined by point evaluations. Their sequence space representations also differ. Equation (4.5) is diagonal with a singly-indexed unknown $\{f_j\}$, while (4.3) is non-diagonal (because the SVD of K_X was not invoked) with a doubly-indexed unknown $\{\mathsf{L}_{jk}\}$. Thus, our work deviates significantly from existing studies.

4.1.1.3 Diagonalization Leads to Eigenvalue Learning

The technical core of this chapter concerns the sequence space representation (4.3) of Main Problem in the ideal setting that a diagonalization of L is known.

Assumption 4.1 (diagonalizing eigenbasis given for L)**.** *The unknown linear operator L from Main Problem is diagonalized in the known orthonormal basis $\{\varphi_j\}_{j\in\mathbb{N}} \subset H$.*

Under this assumption and denoting the eigenvalues of L by $\{l_j\}$, Equation (4.3) simplifies to

$$y_{jn} = \langle \varphi_j, x_n \rangle_H \, l_j \; + \gamma \xi_{jn} \quad \text{for} \quad j \in \mathbb{N} \quad \text{and} \quad n \in \{1, \ldots, N\}\,. \tag{4.6}$$

In general, the random coefficient $\langle \varphi_j, x_n \rangle_H$ depends on every $\{x_{kn}\}_{k\in\mathbb{N}}$ from (4.3). To summarize, under Assumption 4.1 we obtain a white noise sequence

space regression model with *correlated random coefficients*. Inference of the full operator is reduced to only that of its eigenvalue sequence. Equation (4.6) is at the heart of our analysis of linear operator learning. The convergence results we establish for this model may also be of independent interest.

Our proof techniques in this diagonal setting closely follow those in the paper [147], which studies posterior contraction for (4.5) in a simultaneously diagonalizable Bayesian setting. However, our work exhibits some crucial differences with [147] which we now summarize.

(D1) (forward map) The coefficients $\{\langle \varphi_j, x_n \rangle_H\}$ in our problem (4.6) are random variables (r.v.s), while in [147] the singular values $\{\kappa_j\}$ in the problem (4.5) are fixed by K. Also, the law of $\{\langle \varphi_j, x_n \rangle_H\}$ may not be known in practice; only the samples $\{x_n\}$ may be given.

(D2) (link condition) Unlike in [147], our prior covariance operator Σ *is not linked to the SVD* of the forward map K_X. That is, we do not assume simultaneous diagonalizability.

(D3) (prior support) The Gaussian prior we induce on $\{l_j\}$ is supported on a (potentially) much larger sequence space in the scale $\mathcal{H}^s$ (relative to $\{\varphi_j\}$, with $s \in \mathbb{R}$),[1] instead of just the space $\ell^2(\mathbb{N}; \mathbb{R})$ (relative to $\{\phi'_k\}$) charged by the prior on $\{f_j\}$ in [147].

(D4) (reconstruction norm) Solution convergence for (4.6) is in $\mathcal{H}^{-s}$ norms relative to $\{\varphi_j\}$, while only the $\ell^2(\mathbb{N}; \mathbb{R})$ norm relative to $\{\phi'_k\}$ (i.e., H_1 norm) is considered in [147].

These differences deserve further elaboration.

Item (D1). If $x_n \in H$ almost surely (a.s.), then $\langle \varphi_j, x_n \rangle_H \to 0$ a.s. as $j \to \infty$ in (4.6), just as $\kappa_j \to 0$ if K in (4.4) is compact. However, we later observe that our K_X *is not compact*.

[1] The Sobolev-like sequence Hilbert spaces $\mathcal{H}^s = \mathcal{H}^s(\mathbb{N}; \mathbb{R})$ are defined for $s \in \mathbb{R}$ by

$$\mathcal{H}^s(\mathbb{N}; \mathbb{R}) := \left\{ v \colon \mathbb{N} \to \mathbb{R} \,\middle|\, \sum_{j=1}^{\infty} j^{2s} |v_j|^2 < \infty \right\}.$$

They are equipped with the natural $\{j^s\}$-weighted $\ell^2(\mathbb{N}; \mathbb{R})$ inner product and norm. We will usually interpret these spaces as defining a smoothness scale [119, Section 2] of vectors relative to the orthonormal basis $\{\varphi_j\}$ of H.

Item (D2). The authors in [147] assume that the eigenbasis of the prior covariance of f is precisely $\{\phi'_k\}$, the right singular vectors of K in (4.4). This direct link condition between the prior and K ensures that the implied prior (and posterior) on $\{f_j\}$ is an infinite product measure. Our analysis of (4.6) still induces an infinite product prior on $\{l_j\}$ *without using the SVD of the forward operator* K_X. Instead, we make mild assumptions that only weakly link K_X to the prior covariance operator Σ. See (4.7) for a relevant smoothness condition.

Item (D3). The reason we work with a sequence prior having support on sets larger than ℓ^2 is to include *unbounded operators* (with eigenvalues $|l_j| \to \infty$ as $j \to \infty$) in the analysis.[2]

Item (D4). Only the H_1 estimation error is considered in [147] because the unknown quantity is a vector $f \in H_1$. Since our unknown is an operator, we also consider the *prediction error* [48] on new test inputs (see Section 4.2.2.5). This relates to the $\mathcal{H}^{-s}$ norms in (D4).

4.1.2 Main Result

Here and in the sequel, we assume that there is some fixed ground truth operator that underlies the observed output data.

Assumption 4.2 (true linear operator)**.** *The data Y, observed as $\{y_{jn}\}$ in (4.6), is generated according to (4.2) for a fixed self-adjoint linear operator $L = L^\dagger$ with eigenvalues $\{l_j^\dagger\}$.*

Under (4.6), we study the performance of the posterior $\{l_j\} \,|\, (X, Y)$ (and related point estimators) for estimating the true $\{l_j^\dagger\}$ in the limit of infinite data. The following concrete theorem is representative of more general convergence results established later in the chapter.

Theorem 4.3 (asymptotic convergence rate with Gaussian design)**.** *Suppose that Assumptions 4.1 and 4.2 hold with $\{l_j^\dagger\} \in \mathcal{H}^s$ for some $s \in \mathbb{R}$. Let $\nu = \mathcal{N}(0, \Lambda)$ be a Gaussian measure satisfying*

$$c_1^{-1} j^{-2\alpha} \leq \langle \varphi_j, \Lambda \varphi_j \rangle_H \leq c_1 j^{-2\alpha} \quad \text{for all sufficiently large } j \in \mathbb{N} \tag{4.7}$$

[2] Note, however, that unbounded operators with continuous spectra [67] are beyond the scope of this chapter.

for some $c_1 \geq 1$ *and* $\alpha > 1/2$. *Let* $\bigotimes_{j=1}^{\infty} \mathcal{N}(0, \sigma_j^2)$ *be the prior on* $\{l_j\}$ *in* (4.6) *with variances* $\{\sigma_j^2\}$ *satisfying* $c_2^{-1} j^{-2p} \leq \sigma_j^2 \leq c_2 j^{-2p}$ *for all sufficiently large* $j \in \mathbb{N}$ *for some* $c_2 \geq 1$ *and* $p \in \mathbb{R}$. *Denote by* P^{D_N} *the posterior distribution for* $\{l_j\}$ *arising from the observed data* $D_N \coloneqq (X, Y)$. *Fix* $\alpha' \in [0, \alpha + 1/2)$. *If* $\min(\alpha, \alpha') + \min(p - 1/2, s) > 0$, *then there exists a constant* $C > 0$, *independent of the sample size* N, *such that*

$$\mathbb{E}^{D_N} \mathbb{E}^{\{l_i^{(N)}\}_{i=1}^{\infty} \sim P^{D_N}} \sum_{j=1}^{\infty} j^{-2\alpha'} \left| l_j^{\dagger} - l_j^{(N)} \right|^2 \leq C N^{-\left(\frac{\alpha' + \min(p - 1/2, s)}{\alpha + p} \right)} \tag{4.8}$$

for all sufficiently large N. *The first expectation in* (4.8) *is over the joint law of* D_N.

Equation (4.8) shows that, on average, posterior sample eigenvalue estimates converge in $\mathcal{H}^{-\alpha'}$ to the true eigenvalues of $L^{\dagger}$ in the infinite data limit. The hypothesis (4.7), which controls the regularity of the data $\{x_n\}$, is immediately satisfied if, e.g., Λ is a Matérn-like covariance operator with eigenvectors $\{\varphi_j\}$. Theorem 4.3, whose proof is in Appendix C.1, is a consequence of Theorem 4.16, which is valid for a much larger class of input data measures.

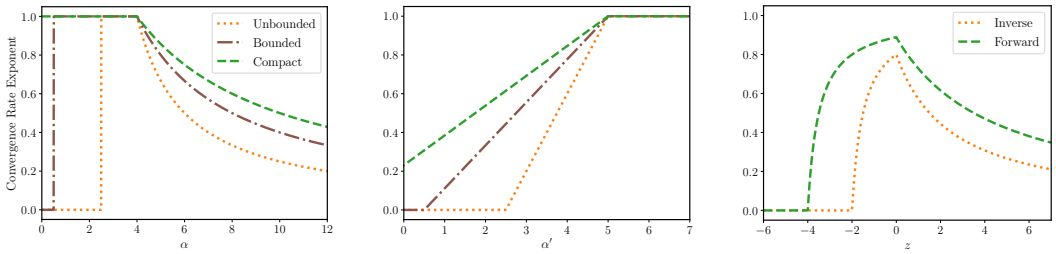
Nonetheless, Theorem 4.3 nearly tells the whole story. The convergence rate exponent in (4.8) shows that the regularity of the ground truth, data, and prior each have an influence on sample complexity. Figure 4.1 summarizes this complex relationship. The figure, and this chapter more generally, reveals three fundamental principles of (linear) operator learning:

(P1) (smoothness of outputs) The ground truth operator becomes statistically more efficient to learn whenever the smoothness of its (noise-free) outputs *increases*. Moreover, as the degree of smoothing of the operator *increases*, sample complexity improves.

(P2) (smoothness of inputs) *Decreasing* the smoothness of input training data improves sample complexity (in norms that do not depend on the training distribution itself).[3]

(P3) (distribution shift) As the smoothness of samples from the input test distribution *increases*, average out-of-distribution prediction error improves.

[3] If the norm used to measure error depends on the training data distribution, this may no longer be true. For example, in-distribution error (train and test on the same distribution) would correspond in Theorem 4.3 to setting $\alpha' = \alpha$ (see Section 4.2.2.5). In this case, *increasing* α would improve sample complexity.



(a) Varying α and s (b) Varying α' and s (c) Varying p, α, and s

Figure 4.1: Fundamental principles of linear operator learning. The theoretical convergence rate exponents (from Theorem 4.18) corresponding to unbounded ($-\Delta$, $s < -2.5$), bounded (Id, $s < -1/2$), and compact ($(-\Delta)^{-1}$, $s < 1.5$) true operators are displayed (see Principle (P1) and Section 4.4.1). With $p = s + 1/2$, Figure 4.1a ($\alpha' = 4.5$) and Figure 4.1b ($\alpha = 4.5$) illustrate the effects that varying input training data and test data smoothness have on convergence rates, respectively (Principles (P2) and (P3)). Figure 4.1c shows that learning the unbounded "inverse map" $-\Delta$ (with $\alpha = \alpha' = 4.5$) is always harder than learning the compact "forward map" $(-\Delta)^{-1}$ (with $\alpha = \alpha' = 2.5$) as the shift $z = p - s - 1/2$ in prior regularity is varied (Section 4.1.4).

Below, we discuss how the Principles (P1) to (P3) manifest in Theorem 4.3 and Figure 4.1.

Item (P1). In Theorem 4.3, the left side of (4.8) is equivalent to the expected prediction error over some input test distribution (see Section 4.2.2.5 for details). Increasing α' increases the regularity of test samples. Assuming for simplicity that $s = p - 1/2$, the convergence rate in (4.8) is $N^{-(\alpha'+s)/(\alpha+s+1/2)}$ as $N \to \infty$. Thus, besides large α', it is beneficial to have large regularity exponents $\alpha' + s$ of the operator's evaluation on sampled test inputs or large regularity exponents s of the true operator's eigenvalues. Indeed, Figures 4.1a to 4.1c suggest that unbounded operators (whose eigenvalues grow without bound) are more difficult to learn than bounded (eigenvalues remain bounded) or compact ones (eigenvalues decay to zero).

Item (P2). *Training inputs* with low smoothness are favorable. This is quantified in Theorem 4.3 by *decreasing* α, which means that the $\{x_n\}$ become "rougher" (Figure 4.1a).

Item (P3). Figure 4.1b illustrates that *increasing* α' in Theorem 4.3 improves the error.

We reinforce items (P1) to (P3) throughout the rest of the chapter.

4.1.3 Examples

Although quite a strong assumption, the known diagonalization from Assumption 4.1 is still realizable in practice. For instance, there may be prior knowledge that the data covariance operator commutes with the true operator (and hence shares the same eigenbasis) or that the true operator obeys known physical principles (e.g., commutes with translation or rotation operators). Regarding the latter, in [220] the authors infer the eigenvalues of a differential operator closure for an advection–diffusion model from indirect observations. As in [263], the operator could be known up to some uncertain parameter. This is the case for several smoothing forward operators that define commonly studied linear inverse problems, including the severely ill-posed inverse boundary problem for the Helmholtz equation with unknown wavenumber parameter [11, Section 5] or the inverse heat equation with unknown scalar diffusivity parameter [263, Section 6.1]. In both references, the eigenbases are already known. Thus, our learning theory applies to these uncertain operators: taking s and p large enough in (4.8) yields prediction error rates of convergence as close to N^{-1} as desired.

More concretely, the theory in this chapter may be applied directly to the following examples.

4.1.3.1 Blind Deconvolution

Periodic deconvolution on the d-dimensional torus $\mathbb{T}^d$ is a linear inverse problem that arises frequently in the imaging sciences. The goal is to recover a periodic signal $f\colon \mathbb{T}^d \to \mathbb{C}$ from noisy measurements

$$y = \mu * f + \eta\,, \quad \text{where} \quad \mu * f \coloneqq \int_{\mathbb{T}^d} f(\cdot - t)\,\mu(dt) \quad \text{and } \eta \text{ is noise,}$$

of its convolution with a filter μ. The filter may be identified with a periodic signal or more generally with a signed measure [263, Section 6.2]. However, μ is sometimes unknown; this leads to *blind* or *semi-blind deconvolution*. One path forward is to first estimate the smoothing operator $K_\mu\colon f \mapsto \mu * f$ from many random (f, y) pairs under the given model. By the known translation-invariance of the problem, K_μ is diagonalized in the complex Fourier basis. Inference is then reduced to estimating the Fourier coefficients $\{\mu_j\}$ of μ, which are the eigenvalues of K_μ. Since $\{\mu_j\} \in \mathcal{H}^s$ for some $s \in \mathbb{R}$, Theorem 4.3 provides a convergence rate.

4.1.3.2 Radial Electrical Impedance Tomography

Electrical impedance tomography (EIT) is a non-invasive imaging procedure that is used in medical, industrial, and geophysical applications [199]. Abstractly, EIT concerns the following severely ill-posed nonlinear inverse problem. Let $\mathbb{D} \subset \mathbb{R}^2$ be the unit disk and let $\sigma \colon \mathbb{D} \to \mathbb{R}_{>0}$ be the strictly positive electrical conductivity of a medium. With electric potential $u \colon \mathbb{D} \to \mathbb{R}$ governed by the elliptic partial differential equation (PDE)

$$-\nabla \cdot (\sigma \nabla u) = 0 \quad \text{in} \quad \mathbb{D}\,,$$

the goal is to reconstruct the unknown conductivity σ in $\mathbb{D}$ from voltage and current boundary measurements of u. These are modeled (to infinite precision) by the linear operators

$$\Lambda_\sigma \colon u|_{\partial\mathbb{D}} \mapsto \sigma \frac{\partial u}{\partial \mathsf{n}}\Big|_{\partial\mathbb{D}} \quad \text{or} \quad \mathscr{R}_\sigma \colon \sigma \frac{\partial u}{\partial \mathsf{n}}\Big|_{\partial\mathbb{D}} \mapsto u|_{\partial\mathbb{D}}\,,$$

where $\partial/\partial \mathsf{n}$ is the outward normal derivative. In practical EIT, either Λ_σ or $\mathscr{R}_\sigma$ must be recovered from finite data. One way to solve this *data completion* step [47] involves making random boundary measurements and employing operator learning (4.1). If σ is *radial*, then Λ_σ and $\mathscr{R}_\sigma$ are diagonalized in the complex Fourier basis over $\partial\mathbb{D} = \mathbb{T}^1$ [199, Section 13.1]. In this case, the theory in this chapter immediately applies to learn the eigenvalues of both operators.

4.1.4 Related Work

A natural setting to apply operator learning is one in which the ambient Hilbert space H comprises real-valued functions over a domain $\mathsf{D} \subset \mathbb{R}^d$. For example, there is an emerging body of work focused on learning surrogates for forward, typically nonlinear, solution operators of PDEs [4, 34, 150, 154, 181, 203, 207, 241]. In the context of dynamical systems, there is literature focused on learning the Koopman operator or its generator, both linear operators, from time series data [45, 108, 143, 144, 214]. There also is interest in speeding up (Bayesian) inversion techniques with forward surrogates [154] and in directly learning regularizers for inversion [12, 15] (or even entire regularized inverse solution operators [16, 66, 77]). However, more theory is needed to quantify the difficulty of learning forward versus inverse operators that arise in these contexts. Some sharp theory already exists for nonlinear operator learning. For example, the authors of [52, 229] establish optimal convergence rates for direct and inverse least squares regression problems with both infinite-dimensional input *and*

output spaces under the condition that point evaluation is a Hilbert–Schmidt operator. However, this condition never holds when H is infinite-dimensional in our linear operator setting (4.1).

We now highlight three subfields that are closely linked to our statistical framework.

Linear Operator Learning. The study of linear function-to-function models within functional data analysis (FDA) [227] is well established [69, 128, 231, 268]. Much of this work concerns the setting $H = L^2((0,1);\mathbb{R})$ and linear models based on kernel integral operators under colored noise. Operator estimation is then reduced to learning the kernel, usually in a reproducing kernel Hilbert space (RKHS) framework. Linear operator learning has also been considered in machine learning [1], particularly in the context of conditional expectation operators [195] and conditional mean embeddings [118, 142, 247]. The authors of [128, 231] study functional linear regression with a spectral operator estimator. This allows them to obtain consistency of the prediction error assuming only boundedness of the true operator [128], rather than compactness as assumed in much of the FDA literature. Convergence rates are established in [231]. While unbounded operators are not considered in these two works, their approaches could likely be modified to handle them. Relatedly, the authors of [257] and [42] share our motivations. The former establishes sample complexities for learning Schatten-class compact operators (motivated by inverse problem solution operators) while the latter for learning compact operators associated to Green's functions of elliptic PDEs (motivated by PDE discovery). Our theory also treats these types of operators but goes further by proving sample complexities for the direct learning of *unbounded operators*, which are of primary interest in these papers (the inverse operator in the former and the partial differential operator in the latter).

Inverse Operator Learning. The direct learning of solution operators of inverse problems is currently a popular research area, catalyzed by the success of deep neural networks [15, 46, 77, 94]. However, theoretical analysis in this area is lacking. One difficulty is the interplay between the ill-posedness of the learning and ill-posedness of the inverse problem itself. For a compact operator T, our diagonal theory suggests that learning $L = T$ under model (4.1) is easier than learning the unbounded inverse operator $L = T^{-1}$ under the same model

(Figure 4.1c). Although less common than the former, the latter setting could arise from noisy differentiation of time series in PDE system identification, for example. One limitation of our theory is that it does not account for *errors-in-covariates* that distinguishes true inverse operator learning, where (a regularized version of) T^{-1} must be estimated only from noisy forward map samples (4.1) with $L = T$. Total least squares [113] is one solution approach in finite dimensions. The infinite-dimensional setting was considered in [38] but with non-Bayesian methods. Regardless, inverse operator learning in this challenging setting is an important area for future research.

Bayesian Nonparametric Statistics. Although the theoretical analysis of inverse problems with linear operator unknowns is largely absent from the Bayesian nonparametrics literature (see Sections 4.1.1.2 and 4.1.1.3), this literature still has some similarities with Equations (4.1) and (4.2). Many works go beyond [147] by deriving posterior contraction rates for problem (4.4) without assuming simultaneous diagonalizability of the prior covariance and the forward operator. In [230], the author studies linear inverse problems in a non-conjugate setting. However, knowledge of the forward map's SVD is used heavily in the analysis even though the prior is (in one case) represented in a non-SVD basis (one comprised of finite linear combinations of singular vectors). These ideas are generalized in [119] to priors linked to smoothness scales instead of the SVD. For Gaussian priors not linked to the SVD, new methods were introduced in [198] that yield optimal posterior performance for X-ray transform inverse problems. These techniques were refined for general linear inverse problems in [112]. However, the previous two papers focus on semiparametric inference (i.e., linear functionals) instead of full nonparametric reconstruction (our main interest). The closest work to ours is [263]. There, the author studies a linear inverse problem in which the forward map is only known up to an uncertain parameter θ. Given a noisy observation of θ in addition to data of the form (4.4), the author analyzes a Bayesian joint reconstruction procedure. Other papers that use Gaussian priors not linked to the SVD include [8, 9, 145]. While notable, all of these works mentioned *do not help us extend the results in this chapter for* (4.6) *to non-diagonal linear operator learning* (4.3) *because our framework already avoids the SVD from the start*; see Item (D2) in Section 4.1.1.3. Removing Assumption 4.1 while preserving sharp rates will likely require new ideas; see Section 4.5. Last, although the

three papers [2, 39, 197] develop powerful new methods, these methods are specific to the particular nonlinear inverse problem studied. In contrast, the aim of this chapter is to develop widely applicable theoretical insights into operator learning. We view [2, 39, 197] as being more relevant to follow-up work in the area of nonlinear inverse operator learning.

4.1.5 Contributions

This chapter provides a unified framework for the supervised learning of compact, bounded, and unbounded linear operators. The analysis is performed in the ideal situation that the eigenvectors of the true operator are known. Thus, much like the work in [147] on Bayesian posterior contraction for linear inverse problems, our results give a theoretical roadmap for linear operator learning. Although we do not explicitly learn solution maps of inverse problems from noisy data, our theory provides insight into the difficulty of learning operators defined by both forward and inverse problems. Our primary contributions are now listed:

(C1) we formulate linear operator learning as a nonparametric Bayesian inverse problem with a linear operator as the unknown quantity, generalizing the work of Knapik, van der Vaart, and van Zanten [147] to operators;

(C2) under a known eigenbasis assumption, in the large sample limit we prove convergence of the full posterior eigenvalues to the truth by deriving in-expectation and high probability upper and lower bounds for the generalization error under distribution shift;

(C3) we establish analogous convergence rate guarantees for the posterior mean eigenvalues with respect to learning-theoretic notions of excess risk and generalization gap;

(C4) we present numerical results for learning compact, bounded, and unbounded operators arising from canonical linear PDEs in a diagonal setting, which directly support the theory, and in a non-diagonal setting, which support conjecture that our theoretical insights remain valid beyond the confines of the theory.

A consequence of these contributions are the theoretical principles (P1) to (P3) (visualized in Figure 4.1). Although only proved for linear operators, these may still inform state-of-the-art nonlinear operator learning techniques used in practice [34, 154, 181, 207]. Indeed, the influence of output space smoothness on sample complexity, reflecting (P1), has been observed in neural operators [75, 152, 160]. Item (P2) implies that training on Gaussian random field data with the commonly chosen squared exponential covariance (leading to infinitely smooth samples) is actually statistically *disadvantageous*. Regarding robustness of models under distribution shift, (P2) and (P3) suggest that it may be misleading to only report prediction errors on test data with the same smoothness as the training data. Further exploration of these and related issues is crucial to guide the development of operator learning as an emerging field.

4.1.6 Outline

The remainder of the chapter is organized as follows. Contribution (C1) (summarized in Section 4.1.1) is described in Section 4.2, where we give a full functional-analytic problem setup and characterize the posterior. Our main theoretical results, items (C2) and (C3), are presented and discussed in Section 4.3. Numerical experiments (C4) that illustrate, support, and extend beyond the theory are provided in Section 4.4. Concluding remarks follow in Section 4.5. Appendix C.1 is devoted to proofs of the main results, with supporting lemmas in Appendix C.2. Remaining proofs of auxiliary results are located in Appendix C.3.

4.2 Setup

After overviewing some notation in Section 4.2.1, we detail our Bayesian inverse problems approach to (4.1) in Section 4.2.2. Section 4.2.3 gives an optimization perspective and defines expected risk and generalization gap in the infinite-dimensional setting.

4.2.1 Preliminaries

We now detail the conventions used in this chapter.

Linear Spaces. Let $(H, \langle\cdot,\cdot\rangle, \|\cdot\|)$ from Section 4.1 be a real, separable, infinite-dimensional Hilbert space. For any self-adjoint positive definite linear operator A on H, we define $A^{-1/2}$ by functional calculus, $\langle\cdot,\cdot\rangle_A := \langle A^{-1/2}\cdot, A^{-1/2}\cdot\rangle$, and $\|\cdot\|_A := \|A^{-1/2}\cdot\|$. The set $\mathcal{L}(H_1; H_2)$ is the space of bounded linear op-

erators mapping Hilbert spaces H_1 into H_2, and when $H_1 = H_2 = H$, we write $\mathcal{L}(H)$. The separable Hilbert space of Hilbert–Schmidt operators from H_1 to H_2 is denoted by $\mathrm{HS}(H_1; H_2)$ with inner product $\langle \cdot, \cdot \rangle_{\mathrm{HS}(H_1;H_2)}$. When $H_1 = H_2 = H$, we write $(\mathrm{HS}(H), \langle \cdot, \cdot \rangle_{\mathrm{HS}}, \|\cdot\|_{\mathrm{HS}})$. For any $a \in H_2$ and $b \in H_1$, the map $a \otimes_{H_1} b \in \mathrm{HS}(H_1; H_2)$ denotes the outer product $(a \otimes_{H_1} b)c \coloneqq \langle b, c \rangle_{H_1} a$ for any $c \in H_1$. We use the shorthand $a \otimes b \in \mathrm{HS}(H)$ when $H_1 = H_2 = H$. For a possibly unbounded linear operator T on H, we denote its domain by the subspace $\mathcal{D}(T) \subseteq H$. The identity map on H is written as $\mathrm{Id} \in \mathcal{L}(H)$.

Probability. We primarily consider centered Borel probability measures Π on H with finite second moment $\mathbb{E}^{h\sim\Pi}\|h\|^2 < \infty$. Such a Π has a covariance operator $\mathrm{Cov}[\Pi] \coloneqq \mathbb{E}^{h\sim\Pi}[h \otimes h]$ in $\mathcal{L}(H)$ that is symmetric, nonnegative, and trace-class. This leads to the Karhunen–Loève (KL) expansion $h = \sum_{j=1}^{\infty} \theta_j \xi_j \psi_j \sim \Pi$ [250]. The $\{\xi_j\}$ are zero mean, unit variance, pairwise uncorrelated real r.v.s on a complete probability space denoted by $(\Omega, \mathcal{F}, \mathbb{P})$. The $\{\psi_j\}$ are the eigenvectors of $\mathrm{Cov}[\Pi]$, extended to form an orthonormal basis of H, and $\{\theta_j^2\}$ are its nonnegative eigenvalues. If Π is a Gaussian measure, then the $\{\xi_j\}$ are i.i.d. $\mathcal{N}(0,1)$ [253]. When appropriate, expectations are taken in the sense of Bochner integration. We use $\mathbb{E}$ with no additional scripts to denote an average over all sources of randomness. We implicitly justify the exchange of expectation and infinite summation with the Fubini–Tonelli theorem.

Notation. For real p and q, we write $p \wedge q \coloneqq \min(p, q)$ and $p \vee q \coloneqq \max(p, q)$. For two nonnegative real sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \simeq b_n$ if $\{a_n/b_n\}$ is bounded away from zero and infinity and $a_n \lesssim b_n$ if there exists $C > 0$ such that $a_n/b_n \leq C$ for all n. We use computer science asymptotic notation. This means that we write $a_n = O(b_n)$ as $n \to \infty$ if $\limsup_{n\to\infty} a_n/b_n < \infty$, $a_n = \Omega(b_n)$ as $n \to \infty$ if $b_n = O(a_n)$, $a_n = \Theta(b_n)$ as $n \to \infty$ if both $a_n = O(b_n)$ and $a_n = \Omega(b_n)$, and $a_n = o(b_n)$ as $n \to \infty$ if $\lim_{n\to\infty} a_n/b_n = 0$. We sometimes use $a_n \asymp b_n$ as convenient shorthand for $a_n = \Theta(b_n)$ and $a_n \ll b_n$ for $a_n = o(b_n)$.

4.2.2 Bayesian Inference

In this subsection, we continue the development of operator learning as an inverse problem. We adopt the following conventions. Define D_N to be the collection of all the data, $D_N \coloneqq (X, Y)$. We equip the N-fold product space H^N

with the inner product $\langle U, V\rangle_{H^N} = \frac{1}{N}\sum_{n=1}^{N}\langle u_n, v_n\rangle$ for any $U = (u_1, \ldots, u_N)$ and $V = (v_1, \ldots, v_N) \in H^N$. This makes H^N a Hilbert space. For any symmetric positive definite $\mathcal{C} \in \mathcal{L}(H)$, define $H_{\mathcal{C}} \coloneqq \mathrm{Im}(\mathcal{C}^{1/2}) \subseteq H$. Equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{C}}$, the set $H_{\mathcal{C}}$ is a Hilbert space.

4.2.2.1 Weighted Hilbert–Schmidt Operators

Thus far we have not specified the space to which the self-adjoint operator $L\colon \mathcal{D}(L) \subseteq H \to H$ in Main Problem belongs. Since L may not be bounded on H, the ideal Hilbert space $\mathrm{HS}(H)$ is not sufficient. Instead, we consider particular Lebesgue–Bochner spaces. Let ν' be a centered Borel probability measure on a sufficiently large space containing H with bounded covariance $\Lambda' \coloneqq \mathrm{Cov}[\nu'] \in \mathcal{L}(H)$. Then $L^2_{\nu'}(H; H)$ is defined as the set of all Borel measurable maps $F\colon H \to H$ such that $\|F\|_{L^2_{\nu'}(H;H)} \coloneqq (\mathbb{E}^{x\sim\nu'}\|F(x)\|^2)^{1/2}$ is finite. Linearity gives additional structure. For any linear $T\colon \mathcal{D}(T) \subseteq H \to H$, the identity $\langle v, Tu\rangle = \mathrm{tr}(Tu \otimes v)$ for all $u \in \mathcal{D}(T)$ and $v \in H$ yields

$$\mathbb{E}^{x\sim\nu'}\|Tx\|^2 = \mathrm{tr}\big(T\Lambda'^{1/2}\big(T\Lambda'^{1/2}\big)^*\big) = \|T\Lambda'^{1/2}\|^2_{\mathrm{HS}} = \|T\|^2_{\mathrm{HS}(H_{\Lambda'};H)}\,. \tag{4.9}$$

By (4.9), linear maps with finite $L^2_{\nu'}$ Bochner norm can be identified with weighted Hilbert–Schmidt operators. This is useful, as the next fact (proved in Appendix C.3) demonstrates.

Fact 4.4 (weighted Hilbert–Schmidt spaces)**.** *Suppose there is a symmetric positive definite linear operator $\mathcal{K} \in \mathcal{L}(H)$ that satisfies $\mathcal{K}^{-1/2} \in \mathrm{HS}(H_{\Lambda'}; H)$, where $\Lambda' = \mathrm{Cov}[\nu']$. Then $\nu'(H_{\mathcal{K}}) = 1$. Additionally, if $T \in \mathrm{HS}(H_{\mathcal{K}}; H)$, then $\mathbb{E}^{x\sim\nu'}\|Tx\|^2 < \infty$.*

For $\mathcal{K}$ satisfying the hypotheses of Fact 4.4, the fact suggests that $\mathrm{HS}(H_{\mathcal{K}}; H)$ is a natural Hilbert space for L to belong to. Defining $\mathcal{D}(L) \coloneqq \{h \in H\colon Lh \in H\}$ (the usual domain for many self-adjoint operators), Fact 4.4 also implies that $\nu'(\mathcal{D}(L)) = 1$. Identifying such a valid $\mathcal{K}$ requires some *a priori* knowledge about the unknown L. For example, later in Subsection 3.1 we show how to choose a $\mathcal{K}$ "smoothing enough" so that $L \in \mathrm{HS}(H_{\mathcal{K}}; H)$. For now, to make sense of the remainder of Section 4.2 we *assume* that the following condition holds.

Condition 4.5 (existence of $\mathcal{K}$)**.** *There exists a symmetric positive definite linear operator $\mathcal{K} \in \mathcal{L}(H)$ such that $\{\mathcal{K}^{-1/2}\Lambda^{1/2}, \mathcal{K}^{-1/2}\Lambda'^{1/2}\} \subset \mathrm{HS}(H)$ and $L^\dagger \in \mathrm{HS}(H_\mathcal{K}; H)$.*

Our use of weighted Hilbert–Schmidt spaces is closely related to the notion of Π measurable linear operators for a probability measure Π, which is a common way to work with unbounded operators; see [104, 186] and [147, Sections 3–4]. If $\mathcal{K}$ is compact, the weighted norm is weak in the sense that $\mathrm{HS}(H_\mathcal{K}; H) \supset \mathcal{L}(H) \supset \mathrm{HS}(H)$ [104, Section 2.2]. However, if L is already Hilbert–Schmidt on H, then the choice $\mathcal{K} = \mathrm{Id} \in \mathcal{L}(H)$ in Condition 4.5 is valid (if Λ' is trace-class).

4.2.2.2 Data Model

Recall the statistical model $Y = K_X L + \gamma\Xi$ (4.2) from Section 4.1.1.1. We now give further details about each component in this data model.

Forward Map. The input data $X \sim \nu^{\otimes N}$ in H^N defines the linear forward map K_X. We enforce that the Borel probability measure ν has finite second moment. Hence, its covariance $\Lambda \in \mathcal{L}(H)$ is symmetric, nonnegative, and trace-class on H. We take Λ to be strictly positive definite for simplicity. For $\mathcal{K}$ as in Condition 4.5 and for any $Z \in H_\mathcal{K}^N$ (the N-fold product of $H_\mathcal{K}$), we define the forward map $K_Z \in \mathcal{L}(\mathrm{HS}(H_\mathcal{K}; H); H^N)$ by $T \mapsto K_Z T \coloneqq (Tz_1, \dots, Tz_N)$. Fact 4.6, proved in Appendix C.3, addresses the compactness of this map.

Fact 4.6 (non-compact)**.** *If $Z \in H_\mathcal{K}^N \setminus \{0\}$, then $K_Z \in \mathcal{L}(\mathrm{HS}(H_\mathcal{K}; H); H^N)$ is not compact.*

Noise. Define $\pi \coloneqq \mathcal{N}(0, \mathrm{Id})$. Since $\mathrm{Id} \in \mathcal{L}(H)$ is not trace-class on H, the white noise $\xi \sim \pi$ is not a proper random element in H. It is instead defined as the H-indexed centered Gaussian process $\xi \coloneqq \{\xi_h \colon h \in H\}$ with covariance function $(h, h') \mapsto \mathbb{E}[\xi_h \xi_{h'}] = \langle h, h' \rangle$ [147, Section 2]. For $\gamma > 0$, the noise is then $\gamma\Xi$, where $\Xi \sim \pi^{\otimes N}$ is assumed independent of X and L. Finally, we interpret Y in (4.2) as N independent stochastic processes $Y_n \coloneqq \{\langle y_n, h \rangle \colon h \in H\}$ for $n \in \{1, \dots, N\}$, such that for $h \in H$, it holds that $\langle y_n, h \rangle \mid X, L \sim \mathcal{N}(\langle Lx_n, h \rangle, \gamma^2 \|h\|^2)$. Observing Y entrywise on the indices $\{\varphi_j\}$ leads to (4.3). The case of general $\mathrm{Cov}[\pi] = \Gamma \in \mathcal{L}(H)$ may be handled by pre-whitening the data (4.2) [10, Section 1]; see also the related Corollary 4.19.

Prior. We assume that $L \sim \mu$ is *a priori* Gaussian, where $\mu \coloneqq \mathcal{N}(0, \Sigma)$ is conjugate to the likelihood, and independent of X and Ξ. Since we view the r.v. $L \colon \mathcal{D}(L) \subseteq H \to H$ as a densely defined operator, the sense in which μ is a proper Gaussian measure requires some care. Specifically, we take $\Sigma \in \mathcal{L}(\mathrm{HS}(H_{\mathcal{K}}; H))$ to be symmetric, positive definite, and trace-class on $\mathrm{HS}(H_{\mathcal{K}}; H) \supseteq \mathrm{HS}(H)$, but not necessarily trace-class on $\mathrm{HS}(H)$. Here $\mathcal{K}$ ensures that the support of μ is large enough to encompass unbounded operators on H.

Posterior. Recall that the realized data Y is given by (4.2) with $L = L^{\dagger}$ under Assumption 4.2. Since Ξ, X, and L are *a priori* independent, the posterior for L given Y and X, denoted by μ^{D_N}, is the same as that obtained when L is conditioned on Y with X fixed, a.s.; see [74, Theorems 32, 13, and 37] for more justification. The Bayesian inverse problem (4.2) is linear and Gaussian. Thus, the posterior is also a Gaussian on $\mathrm{HS}(H_{\mathcal{K}}; H)$ and is denoted by

$$\mu^{D_N} = \mathcal{N}(\bar{L}^{(N)}, \Sigma^{(N)}) \,. \tag{4.10}$$

The posterior mean is $\bar{L}^{(N)} = \mathbb{E}^{L \sim \mu^{D_N}} L \in \mathrm{HS}(H_{\mathcal{K}}; H)$. The posterior covariance operator is $\Sigma^{(N)} \in \mathcal{L}(\mathrm{HS}(H_{\mathcal{K}}; H))$. Explicit formulas for both are known even in this infinite-dimensional setting [147, 166, 186]. We link (4.10) to our diagonal formulation in the next three subsections.

4.2.2.3 Diagonalization

Recall the scalar sequence space model $y_{jn} = \langle \varphi_j, x_n \rangle l_j + \gamma \xi_{jn}$ for $j \in \mathbb{N}$ and $n \in \{1, \dots, N\}$ (4.6).[4] This model arises from the matrix sequence problem (4.3) under Assumption 4.1 by noting that $\mathsf{L}_{jk} = \langle \varphi_j, L\phi_k \rangle = l_j \langle \varphi_j, \phi_k \rangle$ because L is self-adjoint. The $\{\phi_k\}$ are the orthonormal eigenvectors of $\Lambda = \mathrm{Cov}[\nu]$. For each n, the $\{x_{kn} = \langle \phi_k, x_n \rangle\}_{k \in \mathbb{N}}$ are pairwise uncorrelated r.v.s by KL expansion. If L and Λ commute, then $\{\phi_k = \varphi_k\}$ can be taken as the eigenbasis for L. For each n, the scalar model's coefficients $\{\langle \varphi_j, x_n \rangle\}$ are pairwise uncorrelated in this case. However, in general L and Λ do not commute, so the coefficients are

[4] In the absence of noise $\{\gamma \xi_{jn}\}$, determination of $\{l_j = l_j^{\dagger}\}$ is trivial: the diagonalizable structure arising from Assumption 4.1 means that $\{l_j^{\dagger}\}$ may be recovered from a *single* input-output pair, say $(x_1, L^{\dagger} x_1)$. However, our non-diagonal simulation studies in Section 4.4.2 will demonstrate the relevance of our theory beyond Assumption 4.1. In this setting, determination of $\{l_j^{\dagger}\}$ is no longer trivial in the noise-free case.

correlated. For $n \in \{1, \dots, N\}$, it is useful to write these as

$$g_{jn} \coloneqq \langle \varphi_j, x_n \rangle = \sum_{k=1}^{\infty} \langle \varphi_j, \phi_k \rangle x_{kn} \quad \text{and} \quad \vartheta_j^2 \coloneqq \mathrm{Var}[g_{j1}] = \langle \varphi_j, \Lambda \varphi_j \rangle \tag{4.11}$$

for $j \in \mathbb{N}$. Our proofs use some independence-agnostic methods to deal with the dependent, correlated family $\{g_{jn}\}_{j \in \mathbb{N}}$. Nonetheless, $\{g_{jn}\}_{n=1}^{N}$ is still i.i.d. for fixed j and $\mathbb{E}[g_{jn} g_{jn'}] = 0$ for $n \neq n'$.

4.2.2.4 Posterior Characterization

For two sequences $\{a_{jn}\}$ and $\{b_{jn}\}$, we henceforth use the averaging notation $\overline{a_j b_j}^{(N)} \coloneqq \frac{1}{N} \sum_{n=1}^{N} a_{jn} b_{jn}$. For (4.6), we assume a prior $\{l_j\} \sim \mu_{\mathrm{seq}} \coloneqq \bigotimes_{j=1}^{\infty} \mathcal{N}(0, \sigma_j^2)$. We will identify $L \sim \mu$ with $l \coloneqq \{l_j\} \sim \mu_{\mathrm{seq}}$ in Section 4.2.2.5. Under this product prior, (4.6) decouples (i.e., $\{l_j\} \,|\, D_N = \{l_j \,|\, D_N\}$) into an infinite number of random scalar Bayesian inverse problems that are equivalent to the full infinite-dimensional problem (4.2). By completing the square [253, Example. 6.23], we obtain the following Gaussian posterior.

Fact 4.7 (posterior)**.** *The law of* $\{l_j\} \,|\, D_N$ *is* $\mu_{\mathrm{seq}}^{D_N} = \bigotimes_{j=1}^{\infty} \mathcal{N}(\bar{l}_j^{(N)}, (\sigma_j^{(N)})^2)$*, where*

$$\bar{l}_j^{(N)} = \frac{N \gamma^{-2} \sigma_j^2 \overline{y_j g_j}^{(N)}}{1 + N \gamma^{-2} \sigma_j^2 \overline{g_j g_j}^{(N)}} \quad \textit{and} \quad \left(\sigma_j^{(N)}\right)^2 = \frac{\sigma_j^2}{1 + N \gamma^{-2} \sigma_j^2 \overline{g_j g_j}^{(N)}} \tag{4.12}$$

for each $j \in \mathbb{N}$.

4.2.2.5 Bayesian Test Error

The true $L^\dagger$ is naturally approximated by the *posterior mean estimator* $\bar{l}^{(N)} \coloneqq \{\bar{l}_j^{(N)}\}$ and the *posterior sample estimator* $l^{(N)} \coloneqq \{l_j^{(N)}\} \sim \mu_{\mathrm{seq}}^{D_N}$. Defining the linear bijection $B \colon \{l_j\} \mapsto \sum_{j=1}^{\infty} l_j \varphi_j \otimes \varphi_j$, it follows that the actual posterior μ^{D_N} (4.10) on L is the pushforward of $\mu_{\mathrm{seq}}^{D_N}$ under B, that is, $L^{(N)} \sim \mu^{D_N} = B_\sharp \mu_{\mathrm{seq}}^{D_N} = \mathcal{N}(\bar{L}^{(N)}, \Sigma^{(N)})$.

Recall the measure ν' from Section 4.2.2.1 that has bounded covariance $\Lambda' \in \mathcal{L}(H)$ (e.g., $\Lambda' = \mathrm{Id}$ is allowed). Assume Λ' has an orthonormal eigenbasis $\{\phi_k'\}$ of H. We now view ν' as an arbitrary *test data distribution* that we are interested in predictions on. A useful representation of the weighted norm (4.9) is $T \mapsto \mathbb{E}^{x \sim \nu'} \|Tx\|^2 = \sum_{j,k} \lambda_k(\Lambda') \langle \varphi_j, T \phi_k' \rangle^2$, where $\{\lambda_k(\Lambda')\}$ denotes the

eigenvalues of Λ'. In our setting, L is diagonal in $\{\varphi_j\}$ which leads to

$$\|L\|^2_{L^2_{\nu'}(H;H)} = \sum_{i=1}^{\infty} \vartheta_i'^2 l_i^2\,, \quad \text{where} \quad \vartheta_j'^2 := \sum_{k=1}^{\infty} \lambda_k(\Lambda')\langle \varphi_j, \phi_k'\rangle^2 = \langle \varphi_j, \Lambda' \varphi_j\rangle \tag{4.13}$$

for $j \in \mathbb{N}$. We can now define a notion of test error (i.e., prediction or "generalization" error).

Definition 4.8 (test error: posterior)**.** The *test error of the posterior sample estimator* is

$$\mathbb{E}^{D_N}\, \mathbb{E}^{L^{(N)}\sim\mu^{D_N}} \left\| L^\dagger - L^{(N)} \right\|^2_{L^2_{\nu'}(H;H)} = \mathbb{E}^{D_N}\, \mathbb{E}^{l^{(N)}\sim\mu_{\text{seq}}^{D_N}} \sum_{j=1}^{\infty} \vartheta_j'^2 \left| l_j^\dagger - l_j^{(N)} \right|^2. \tag{4.14}$$

The outer expectation is with respect to the data, and the inner expectation is with respect to the Bayesian posterior. The definition of test error for the posterior mean is similar.

Definition 4.9 (test error: mean)**.** The *test error of the posterior mean estimator* is

$$\mathbb{E}^{D_N} \left\| L^\dagger - \bar{L}^{(N)} \right\|^2_{L^2_{\nu'}(H;H)} = \mathbb{E}^{D_N} \sum_{j=1}^{\infty} \vartheta_j'^2 \left| l_j^\dagger - \bar{l}_j^{(N)} \right|^2. \tag{4.15}$$

We say that (4.14) or (4.15) tests *in-distribution* if $\nu' = \nu$ and *out-of-distribution* or *under distribution shift* otherwise. If $\Lambda' = \mathrm{Id}$, then the $L^2_{\nu'}$ Bochner norm equals the familiar unweighted $\mathrm{HS}(H)$ norm. In Section 4.3, we study the $N \to \infty$ asymptotics of Equations (4.14) and (4.15).

4.2.3 Statistical Learning

We briefly adopt a statistical learning theory perspective to complement the Bayesian approach of Section 4.2.2. Let $\mathcal{P}$ denote the joint distribution on (x, y) implied by $y = L^\dagger x + \gamma\xi$, where $x \sim \nu$ and $\xi \sim \pi = \mathcal{N}(0, \mathrm{Id})$ independently. The data in (4.1) is then $(x_n, y_n) \sim \mathcal{P}$ i.i.d., $n \in \{1, \dots, N\}$. Since regression is our focus, it is natural to work with the square loss function on H. Then $\mathbb{E}^{(x,y)\sim\mathcal{P}} \frac{1}{2}\|y - Lx\|^2$ and $\frac{1}{N}\sum_{n=1}^N \frac{1}{2}\|y_n - Lx_n\|^2$ define the expected risk and empirical risk for L, respectively. However, these expressions are not well-defined because infinite-dimensional H implies $\|y\| = \|\xi\| = \infty$ a.s. [253, Remark 3.8]. Inspired by the negative log likelihood of μ^{D_N} as in [12, 205], we re-define the risks as follows.

Definition 4.10 (expected risk)**.** Given L, the *expected risk (*or *population risk)* is

$$\mathcal{R}_\infty(L) \coloneqq \mathbb{E}^{(x,y)\sim\mathcal{P}}\left[\tfrac{1}{2}\|Lx\|^2 - \langle y, Lx\rangle\right]. \tag{4.16}$$

Definition 4.11 (empirical risk)**.** Given L, the *empirical risk* is

$$\mathcal{R}_N(L) \coloneqq \frac{1}{N}\sum_{n=1}^{N}\left[\tfrac{1}{2}\|Lx_n\|^2 - \langle y_n, Lx_n\rangle\right] = \tfrac{1}{2}\|K_X L\|^2_{H^N} - \langle Y, K_X L\rangle_{H^N}, \tag{4.17}$$

and the *regularized empirical risk* is

$$\mathcal{R}_{N,W}(L) \coloneqq \mathcal{R}_N(L) + \tfrac{1}{2N}\|W^{-1/2}L\|^2_{\mathrm{HS}(H_\mathcal{K};H)}, \tag{4.18}$$

where $W \in \mathcal{L}(\mathrm{HS}(H_\mathcal{K}; H))$ is symmetric positive definite and $\mathcal{K}$ is as in Condition 4.5.

Equations (4.16) and (4.17) *are well-defined* because the "infinite constants" $\frac{1}{2}\|y\|^2$ and $\frac{1}{2}\|y_n\|^2$ from the original risk expressions are subtracted away and the linear cross terms $\langle y, Lx\rangle$ and $\langle y_n, Lx_n\rangle$, viewed as actions under stochastic processes (see Section 4.2.2.2), are finite a.s..

The role of risk is to quantify the accuracy of a hypothesis L. By the independence of x and ξ plus the stochastic process definition of π in Section 4.2.2.2, $\mathbb{E}^{(x,y)\sim\mathcal{P}}\langle y, Lx\rangle = \mathbb{E}^{x\sim\nu}\langle L^\dagger x, Lx\rangle$ so that $\mathcal{R}_\infty(L) = \frac{1}{2}\mathbb{E}^{x\sim\nu}\|L^\dagger x - Lx\|^2 - \frac{1}{2}\mathbb{E}^{x\sim\nu}\|L^\dagger x\|^2$. Thus, the infimum of $\mathcal{R}_\infty$ is achieved at the *regression function* [52] $\mathbb{E}[y \,|\, x = \cdot] = L^\dagger \in \mathrm{HS}(H_\mathcal{K}; H)$. Minimizers of the empirical risk over the RKHS *hypothesis class* $\mathscr{L} = \mathrm{Im}(W^{1/2})$ are point estimates of the true $L^\dagger$ (but we do not require $L^\dagger \in \mathscr{L}$). Our focus is the minimizer $\widehat{L}^{(N,W)}$ of the convex functional (4.18) over $\mathscr{L}$. It may be identified as the posterior mean $\bar{L}^{(N)}$ from (4.10) whenever $\gamma^2 W$ equals the prior covariance Σ [73]. We enforce this and write $\widehat{L}^{(N,W)} \equiv \bar{L}^{(N)}$. To quantify the performance of $\bar{L}^{(N)}$, we employ the following notions of error from statistical learning.

Definition 4.12 (excess risk)**.** The *excess risk* of the posterior mean is defined by

$$\mathcal{E}_N \coloneqq 2\mathcal{R}_\infty(\bar{L}^{(N)}) - 2\mathcal{R}_\infty(L^\dagger) = \mathbb{E}^{x\sim\nu}\|L^\dagger x - \bar{L}^{(N)}x\|^2. \tag{4.19}$$

The excess risk is always nonnegative and provides a notion of consistency for $\bar{L}^{(N)}$. In Section 4.3.5, we control (4.19) either in expectation, $\mathbb{E}^{D_N}\mathcal{E}_N$,

or with high probability over the input training samples, $\mathbb{E}^{Y|X}\mathcal{E}_N$. The last expectation is over the noise only, under (4.2).

Next, we define the generalization gap. It can take any sign and, as the difference between test and training errors, controls the amount of "overfitting" that $\bar{L}^{(N)}$ can exhibit.

Definition 4.13 (generalization gap)**.** The *generalization gap* of the posterior mean is

$$\mathcal{G}_N \coloneqq \mathcal{R}_\infty(\bar{L}^{(N)}) - \mathcal{R}_N(\bar{L}^{(N)})\,. \tag{4.20}$$

Equation (4.20) may be written in terms of $L^\dagger$ instead of y (see Equation (C.4) in the proof of Theorem 4.24). In Section 4.3.6, we bound the *expected generalization gap* $\mathbb{E}^{D_N}|\mathcal{G}_N|$.

4.3 Convergence Rates

We are now ready to study the sample complexity of the posterior estimator (4.12) with respect to the notions of error defined in Sections 4.2.2.5 and 4.2.3. In Section 4.3.1, we list and interpret our main assumptions. In Section 4.3.2, under fourth moment conditions we establish asymptotic convergence rates of both the posterior sample and mean estimators and related lower bounds. Posterior contraction is discussed in Section 4.3.3. Analogous high probability results are developed in Section 4.3.4 for subgaussian design. Last, both upper and lower bounds are established in expectation for the excess risk and generalization gap in Sections 4.3.5 and 4.3.6. We collect all of the proofs in Appendix C.1.

4.3.1 Main Assumptions

In the setting of the sequence model (4.6), our convergence theory for diagonal linear operator learning is primarily developed under five assumptions.

Assumption 4.14 (eigenvalue learning assumptions)**.** *The following conditions hold true.*

(A1) (diagonal true operator) *Assumptions 4.1 and 4.2 hold, so that* $L^\dagger = \sum_{j=1}^\infty l_j^\dagger \varphi_j \otimes \varphi_j$.

(A2) (smoothness of true operator) *The true eigenvalues satisfy* $l^\dagger \coloneqq \{l_j^\dagger\} \in \mathcal{H}^s$ *for some* $s \in \mathbb{R}$.

(A3) (smoothness of prior) *The prior variance sequence* $\{\sigma_j^2\}$ *that appears in* $\mu_{\text{seq}} = \bigotimes_{j=1}^{\infty} \mathcal{N}(0, \sigma_j^2)$ *satisfies*

$$\sigma_j^2 = \Theta(j^{-2p}) \quad \text{as} \quad j \to \infty \quad \text{for some} \quad p \in \mathbb{R}\,. \tag{4.21}$$

(A4) (smoothness of data) *The trace-class covariance operator* $\Lambda \in \mathcal{L}(H)$ *of the input training data distribution* ν *satisfies*

$$\vartheta_j^2 = \langle \varphi_j, \Lambda \varphi_j \rangle = \Theta(j^{-2\alpha}) \quad \text{as} \quad j \to \infty \quad \text{for some} \quad \alpha > 1/2\,. \tag{4.22}$$

The input test data distribution ν' *is a centered Borel probability measure with a bounded covariance operator* $\Lambda' \in \mathcal{L}(H)$ *that satisfies*

$$\vartheta_j'^2 = \langle \varphi_j, \Lambda' \varphi_j \rangle = \Theta(j^{-2\alpha'}) \quad \text{as} \quad j \to \infty \quad \text{for some} \quad \alpha' \geq 0\,. \tag{4.23}$$

(A5) (smoothness range) *It holds that* $(\alpha \wedge \alpha') + s > 0$ *and* $(\alpha \wedge \alpha') + (p - 1/2) > 0$.

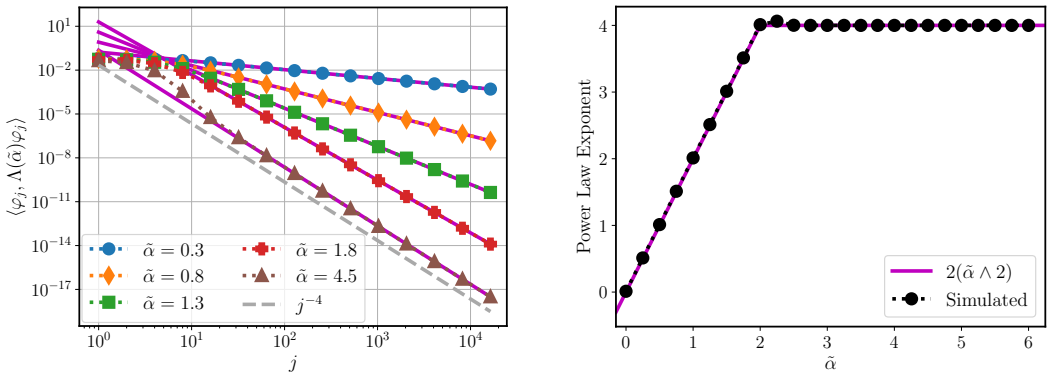
These assumptions are interpreted as follows.

Item (A1). The diagonalization allows us to identify $L^\dagger$ with its eigenvalues $l^\dagger$. The domain $\mathcal{D}(L^\dagger) := \{h \in H \colon \|L^\dagger h\|^2 = \sum_{j=1}^{\infty} |l_j^\dagger|^2 \langle \varphi_j, h \rangle^2 < \infty\}$ ensures that $L^\dagger$ is self-adjoint on H.

Item (A2). The regularity condition $l^\dagger \in \mathcal{H}^s$ implicitly determines the sense in which the series expansion for $L^\dagger$ in (A1) converges. If $s \geq 0$, then $L^\dagger \in \mathrm{HS}(H)$. Otherwise, there exists $\mathcal{K} \in \mathcal{L}(H)$ such that $L^\dagger \in \mathrm{HS}(H_{\mathcal{K}}; H)$. For example, define $\mathcal{K}_{s'} := \sum_j \kappa_j^2 \varphi_j \otimes \varphi_j$ with $\kappa_j^2 = j^{2s'}$. Then $\|L^\dagger\|_{\mathrm{HS}(H_{\mathcal{K}_{s'}}; H)} = \|l^\dagger\|_{\mathcal{H}^{s'}}$, so $L^\dagger$ converges in $\mathrm{HS}(H_{\mathcal{K}_{s'}}; H)$ for any $s' \leq s < 0$.

Item (A3). The exponent $p \in \mathbb{R}$ in (4.21) adjusts the regularity of prior draws $l \sim \mu_{\text{seq}}$: $l \in \mathcal{H}^{s'}$ a.s. for every $s' < p - 1/2$. The choice $p = s + 1/2$ thus gives the closest match to the true regularity of $l^\dagger \in \mathcal{H}^s$. Relating back to Section 4.2.2.2, the full prior is $\mu = B_\sharp \mu_{\text{seq}} = \mathcal{N}(0, \Sigma)$. With, e.g., $\mathcal{K} = \mathcal{K}_{s'}$ as above, Σ then satisfies $\Sigma \varphi_i \otimes \varphi_j = \kappa_j^2 \sigma_j^2 \delta_{ij} \varphi_i \otimes \varphi_j$ for all i and j.

Item (A4). Equations (4.22) and (4.23) reflect algebraic spectral decay of the input data covariance operators with respect to the eigenbasis $\{\varphi_j\}$ of $L^\dagger$. This provides a weak link between the data distributions and the prior;



(a) Decay of $\langle\varphi_j, \Lambda(\tilde{\alpha})\varphi_j\rangle$ as $j \to \infty$

(b) Fit to power law as $\tilde{\alpha}$ varies

Figure 4.2: Example of exact power law spectral decay (4.22) when Λ is not diagonalized by $\{\varphi_j\}$; see the discussion in Item (A4).

see Item (D2). Although sharp bounds such as Equations (4.22) and (4.23) may be difficult to verify when Λ or Λ' is not diagonalized in $\{\varphi_j\}$, Figure 4.2 provides strong numerical evidence that exact power law decay can still exist in this setting. Indeed, in this figure we choose $\Lambda = \Lambda(\tilde{\alpha})$ such that its eigenpairs $\{(\lambda_k^2, \phi_k)\}$ satisfy $\lambda_k^2 = 15^{2\tilde{\alpha}-1}((k\pi)^2 + 225)^{-\tilde{\alpha}} = \Theta(k^{-2\tilde{\alpha}})$ with $\tilde{\alpha} \in \mathbb{R}$ and $z \mapsto \phi_k(z) = \sqrt{2}\sin(k\pi z)$. We choose output basis $z \mapsto \varphi_j(z) = \sqrt{2}\cos((j - \frac{1}{2})\pi z)$. Both $\{\phi_k\}$ and $\{\varphi_j\}$ are orthonormal bases of $H = L^2((0,1);\mathbb{R})$. One can show that $\vartheta_j^2 = \langle\varphi_j, \Lambda(\tilde{\alpha})\varphi_j\rangle = \sum_{k=1}^{\infty} 64\pi^{-2}\lambda_k^2 k^2 (4(j(j-1) - k^2) + 1)^{-2}$. For select $j \leq 2^{21}$, we sum the first 2^{21} terms of this series to approximately compute the $\{\vartheta_j^2\}$. Figure 4.2a shows that ϑ_j^2 decays asymptotically as a power law (with magenta lines being linear least square fits) for various $\tilde{\alpha}$ (saturating near $2\tilde{\alpha} = 4$). Figure 4.2b suggests that Λ satisfies Assumption (A4) with $\alpha = \tilde{\alpha} \wedge 2$.

Item (A5). The first inequality in (A5) ensures that $L^\dagger$ has finite L^2_ν and $L^2_{\nu'}$ Bochner norms (4.13). In particular, $\nu(\mathcal{D}(L^\dagger)) = \nu'(\mathcal{D}(L^\dagger)) = 1$.[5] The second inequality ensures that the prior covariance Σ is trace-class on both $\mathrm{HS}(H_\Lambda; H)$ and $\mathrm{HS}(H_{\Lambda'}; H)$. This means $L \sim \mu = \mathcal{N}(0, \Sigma)$ has finite L^2_ν and $L^2_{\nu'}$ Bochner norms a.s.. It follows that the latter two assertions also hold for the posterior $\mu^{D_N} = \mathcal{N}(\bar{L}^{(N)}, \Sigma^{(N)})$, a.s. with respect to D_N.

[5]Notice that we *do not* invoke the $\mathcal{K}$-weighted Hilbert–Schmidt formulation from Sections 4.2.2.1 and 4.2.2.2 in Assumption 4.14. Such abstraction is unnecessary for our straightforward diagonal approach (Item (A1)). In particular, the scalar sequence space model (4.6) is well-defined without reference to any $\mathcal{K}$. However, work going beyond diagonal operators may need to use $\mathrm{HS}(H_\mathcal{K}; H)$ spaces, with $\mathcal{K}$ satisfying Condition 4.5.

4.3.2 Expectation Bounds

To develop error bounds in expectation, we only require mild polynomial moment conditions on the input training data measure ν.

Assumption 4.15 (expectation: training data)**.** *The training data distribution ν is a centered Borel probability measure on H with KL expansion $x = \sum_{k=1}^{\infty} \lambda_k \zeta_k \phi_k \sim \nu$. The eigenvalues $\{\lambda_k^2\}$ of $\mathrm{Cov}[\nu] = \Lambda$ are ordered to be nonincreasing, and the zero mean and unit variance r.v.s $\{\zeta_k\}$ are independent, have finite fourth moments, and satisfy $\mathbb{E}[\zeta_j^4] = O(1)$ as $j \to \infty$. In particular, $\mathbb{E}^{x\sim\nu}\|x\|^4 < \infty$. Last, the r.v.s $\{\overline{g_j g_j}^{(N)}\}_{j,N\in\mathbb{N}}$, defined in Section 4.2.2.4 as $\overline{g_j g_j}^{(N)} = \frac{1}{N}\sum_{n=1}^{N}\langle\varphi_j, x_n\rangle^2$, satisfy $\limsup_{N\to\infty}\mathbb{E}[(\overline{g_j g_j}^{(N)})^{-4}] \lesssim \langle\varphi_j, \Lambda\varphi_j\rangle^{-4}$ for all $j \in \mathbb{N}$.*

Henceforth, it is useful to define the parametrized sequences $\{J_N\}_{N\in\mathbb{N}}$ and $\{\rho_N\}_{N\in\mathbb{N}}$ by

$$J_N(\alpha, p) \coloneqq \left\lfloor N^{\frac{1}{2(\alpha+p)}} \right\rfloor \quad \text{and}$$

$$\rho_N(\alpha, \alpha', p) \coloneqq \begin{cases} N^{-\left(1-\frac{\alpha+1/2-\alpha'}{\alpha+p}\right)}, & \text{if } \alpha' < \alpha + 1/2\,, \\ N^{-1}\log N\,, & \text{if } \alpha' = \alpha + 1/2\,, \\ N^{-1}\,, & \text{if } \alpha' > \alpha + 1/2\,, \end{cases} \tag{4.24}$$

respectively, for $N \in \mathbb{N}$. Notice that $J_N \to \infty$ (if $\alpha + p > 0$) and $\rho_N = \Omega(N^{-1})$ as $N \to \infty$. Our main result gives asymptotic convergence rates of the test errors from Section 4.2.2.5.

Theorem 4.16 (expectation: upper bound)**.** *Let the ground truth $L^\dagger$, prior μ on L, training data distribution ν, and test data distribution ν' satisfy Assumptions 4.14 and 4.15. Let $\rho_N = \rho_N(\alpha, \alpha', p)$ in (4.24) with α, α', and p as in Assumption 4.14. Denote by μ^{D_N} the posterior distribution (4.10) for L arising from the observed data $D_N = (X, Y)$ in (4.2). Then*

$$\mathbb{E}^{D_N}\,\mathbb{E}^{L^{(N)}\sim\mu^{D_N}}\left\|L^\dagger - L^{(N)}\right\|^2_{L^2_{\nu'}(H;H)} = O\big(\rho_N\big) + o\big(N^{-\left(\frac{\alpha'+s}{\alpha+p}\right)}\big) \quad \text{as} \quad N \to \infty\,, \tag{4.25}$$

where the constants in this upper bound depend on $L^\dagger$ or, equivalently, on $l^\dagger$. Furthermore,

$$\sup_{\|l^\dagger\|_{\mathcal{H}^s}\lesssim 1} \mathbb{E}^{D_N}\,\mathbb{E}^{L^{(N)}\sim\mu^{D_N}}\left\|L^\dagger - L^{(N)}\right\|^2_{L^2_{\nu'}(H;H)} = O\big(\rho_N + N^{-\left(\frac{\alpha'+s}{\alpha+p}\right)}\big) \quad \text{as} \quad N \to \infty\,. \tag{4.26}$$

Both assertions also hold for the test error (4.15) *of the posterior mean* $\bar{L}^{(N)}$.

Theorem 4.16 has the same implications as Theorem 4.3, namely, Principles (P1) to (P3). The effect of distribution shift (P3) in (4.25) is apparent: increasing α' always improves the sample complexity (until the rate N^{-1} is achieved). We note that the three smoothness cases in the ρ_N term from (4.25) are similar to those in functional linear regression [48]. The "matching" prior smoothness choice $p = s + 1/2$ leads to asymptotically balanced contributions from both error terms in (4.26). The rate is then $N^{-(2\alpha'+2s)/(1+2\alpha+2s)}$ if $\alpha' < \alpha + 1/2$ (which for $\alpha' = \alpha$ is minimax optimal [57, 145, 147]) or N^{-1} (up to logarithms) if $\alpha' \geq \alpha + 1/2$. Principles (P1) and (P2) are evident: as s *decreases* ($L^\dagger$ becomes "less compact" and possibly unbounded) and α *increases* (the $\{x_n\}$ become smoother), the rates *degrade*. Figure 4.1 visualizes these rates in various settings. Last, we note that the rate of convergence in (4.25) can be strictly faster when $L^\dagger$ is fixed as opposed to when $L^\dagger$ is varying for the worst-case error (4.26). Our interest is mainly in individual bounds (i.e., fixed $L^\dagger$) because these are more useful in practice.

Next, we provide a lower bound corresponding to a given $L^\dagger$, equivalently, $l^\dagger$.

Theorem 4.17 (expectation: lower bound)**.** *Let the hypotheses of Theorem 4.16 be satisfied. Let* $J_N = J_N(\alpha, p)$ *in* (4.24) *with* α *and* p *as in Assumption 4.14. Then for any positive sequence* $\{\tau_n\}$ *such that* $\tau_n \to 0$ *and* $n\tau_n \to \infty$ *as* $n \to \infty$*, the posterior mean test error satisfies*

$$\mathbb{E}^{D_N} \left\| L^\dagger - \bar{L}^{(N)} \right\|^2_{L^2_{\nu'}(H;H)} = \Omega\left(\tau_N \rho_N + \sum_{j > J_N} j^{-2\alpha'} |l_j^\dagger|^2 \right) \quad \text{as} \quad N \to \infty\,. \tag{4.27}$$

The same assertion holds for the test error (4.14) *of the full posterior* μ^{D_N}*, but without* $\{\tau_n\}$.

The tail series term in (4.27) is closely related to the lower bound in [160, Theorem 3.4] for nonlinear operator learning because both involve the spectral decay of the covariance operator of the pushforward measure $L^\dagger_\sharp \nu'$. Since this tail term is order $N^{-(\alpha'+s)/(\alpha+p)} o(1)$ by (C.6) in Lemma C.1, the lower bound (4.27) "matches" the corresponding terms in the individual upper bound (4.25) up to $o(1)$ factors. But without further conditions on $l^\dagger$ (and hence knowledge about the $o(1)$ factors), the bounds are not guaranteed to be sharp in the over-smoothing prior regime $p > s + 1/2$. The rates do match (up to τ_N

in (4.27) for $\bar{L}^{(N)}$, but τ_N is under control) for under-smoothing priors with $p \leq s + 1/2$ because the ρ_N terms in both Theorems 4.16 and 4.17 dominate. The $\{\tau_n\}$ factor is likely an artifact of our proof technique. By choosing $l^\dagger$ such that $|l_j^\dagger| = J_N^{-s}\delta_{j-1,J_N}$ in (4.27), Theorem 4.17 also implies that (4.26) is truly sharp.

So far, we have assumed that $l^\dagger \in \mathcal{H}^s$ for some s. However, this does not preclude the possibility that $l^\dagger \in \mathcal{H}^{s'}$ for another $s' > s$. For example, the previous theorems account for operators with analytic spectral smoothness (see Section 4.1.3): $l_j^\dagger \asymp \exp(-c_1 j^{c_2})$ for c_1 and $c_2 > 0$ (here $l^\dagger \in \mathcal{H}^s$ for every $s \in \mathbb{R}$). Nevertheless, many scientific problems are naturally distinguished by *regularly varying* eigenvalues. These behave like a power law up to a slowly varying function $S\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ [37] (this means that $S(\lambda x)/S(x) \to 1$ as $x \to \infty$ for every $\lambda > 0$; examples include logarithms or functions with positive limit). The following sharp convergence result concerns posterior sample estimates of regularly varying true eigenvalues. Similar may be proved for the posterior mean, but the upper and lower bounds must be considered separately as in Theorems 4.16 and 4.17. The implications are the same.

Theorem 4.18 (asymptotically sharp bound for regularly varying eigenvalues)**.** *Let the hypotheses of Theorem 4.16 be satisfied, but instead of (A2), let $L^\dagger$ be such that $|l_j^\dagger| = \Theta(j^{-1/2-s}S(j))$ as $j \to \infty$ for some slowly varying function S at infinity. Let $J_N = J_N(\alpha, p)$ in (4.24). Then*

$$\mathbb{E}^{D_N}\,\mathbb{E}^{L^{(N)}\sim\mu^{D_N}}\left\|L^\dagger - L^{(N)}\right\|^2_{L^2_{\nu'}(H;H)} = \Theta\left(\rho_N + N^{-\left(\frac{\alpha'+s}{\alpha+p}\right)}S^2(J_N)\right) \quad \text{as} \quad N \to \infty\,. \tag{4.28}$$

Although we have thus far restricted our attention to the Gaussian white noise model (4.2), the next corollary shows that our theory remains valid for *smoother* Gaussian noise.

Corollary 4.19 (colored noise)**.** *Suppose that the Gaussian distribution of the $\{\xi_n\}$ determining the data Y in (4.2) is not necessarily white, but is instead given by $\pi = \mathcal{N}(0, \Gamma)$, where $\Gamma \in \mathcal{L}(H)$ is symmetric positive definite with eigenbasis $\{\varphi_j\}$ shared with $L^\dagger$ and eigenvalues $\lambda_j(\Gamma) = \Theta(j^{-2\beta})$ as $j \to \infty$ for some $\beta \geq 0$. Let $\mu_{\text{seq}}^{D_N}$ be given by (4.12) except with each γ^2 replaced by $\gamma^2\lambda_j(\Gamma)$. Let the hypotheses of Theorems 4.16 to 4.18 hold, respectively, except let $\rho_N = \rho_N(\alpha - \beta, \alpha', p)$, $J_N = J_N(\alpha - \beta, p)$, and instead of (A5), let*

$\min(\alpha-\beta,\alpha')+\min(p-1/2,s)>0$. *Then the results of Theorems 4.16 to 4.18 remain valid, respectively, if in each display of Equations (4.25) to (4.28) every instance of α is replaced by $\alpha-\beta$.*

The corollary follows from the hypothesis that Γ and $L^\dagger$ commute. Indeed, pre-whitening the new output data gives $\Gamma^{-1/2}y_n = L^\dagger\Gamma^{-1/2}x_n + \gamma\mathcal{N}(0,\mathrm{Id})$, which our existing theory can handle. This result implies that larger β (smoother noise) improves convergence rates because the input data smoothness has effectively been reduced from α to $\alpha-\beta$ (see Principle (P2)).

4.3.3 Posterior Contraction

The performance of Bayesian procedures is often quantified by the rate of contraction of the posterior around the true data-generating parameter as $N\to\infty$. In the setting of operator learning, we follow [8, 11, 145, 147] and consider finding a positive sequence $\varepsilon_N\to 0$ such that for any positive sequence $M_N\to\infty$, it holds that

$$\mathbb{E}^{D_N}\,\mu^{D_N}\big(\{L\colon \|L^\dagger-L\|_{L^2_{\nu'}(H;H)}\geq M_N\varepsilon_N\}\big)\to 0 \quad\text{as}\quad N\to\infty\,. \tag{4.29}$$

We say that ε_N is a *contraction rate* of the posterior μ^{D_N} with respect to the $L^2_{\nu'}(H;H)$ Bochner norm. By Chebyshev's inequality, $(M_n\varepsilon_N)^{-2}$ times the posterior test error (4.14) is an upper bound for the left-hand side of (4.29). Thus, the limit in (4.29) holds true if (4.14) is $O(\varepsilon_N^2)$ as $N\to\infty$. The next corollary is then a consequence of Theorem 4.16.

Corollary 4.20 (posterior contraction)**.** *Let the hypotheses of Theorem 4.16 be satisfied. Then any sequence $\{\varepsilon_N\}_{N\in\mathbb{N}}$ such that ε_N^2 is of the order of the right-hand side of (4.25) as $N\to\infty$ is a contraction rate of μ^{D_N} with respect to the $L^2_{\nu'}(H;H)$ Bochner norm.*

We deduce that the inverse problem (4.2) for linear operator learning is *moderately ill-posed* in the sense of [10, Section 4] under Assumptions 4.14 and 4.15 because ε_N^2 follows a power law (4.25). Since μ^{D_N} is Gaussian, (4.14) admits a decomposition into three terms: the squared estimation bias, estimation variance, and posterior spread (i.e., the trace of $\Sigma^{(N)}$) [10, Section 1.1]. Inspection of the proof of Theorem 4.16 shows that the second term on the right of (4.25) is the contribution from the squared estimation bias, while the first is from both the estimation variance and posterior spread (C.1). Interpretations are similar for the remaining theorems.

4.3.4 High Probability Bounds

A stronger assumption on the input data distribution is needed to obtain concentration bounds. It includes Gaussian measures as a special case.

Assumption 4.21 (high probability: training data)**.** *The training data distribution ν is a centered Borel probability measure on H with KL expansion $x = \sum_{k=1}^{\infty} \lambda_k \zeta_k \phi_k \sim \nu$. The eigenvalues $\{\lambda_k^2\}$ of $\mathrm{Cov}[\nu] = \Lambda$ are ordered to be nonincreasing, and the zero mean and unit variance r.v.s $\{\zeta_k\}$ are independent σ_ν^2-subgaussian for some absolute constant $\sigma_\nu \geq 1$. In particular, ν is a strict subgaussian measure with trace-class covariance operator proxy Λ.*[6]

The next result holds with exceptionally high probability over the subgaussian design $X \sim \nu^{\otimes N}$.

Theorem 4.22 (high probability: upper and lower bounds)**.** *Let the ground truth $L^\dagger$, prior μ on L, training data distribution ν, and test data distribution ν' satisfy Assumptions 4.14 and 4.21. Let $J_N = J_N(\alpha, p)$ and $\rho_N = \rho_N(\alpha, \alpha', p)$ in (4.24) with α, α', and p as in Assumption 4.14. Denote by μ^{D_N} the posterior distribution (4.10) for L arising from the observed data $D_N = (X, Y)$ in (4.2). Then there exist two constants $c_1 > 0$ and $c_2 \in (0, 1/2)$ such that, as $N \to \infty$, it holds that*

$$\mathbb{E}^{Y \mid X} \mathbb{E}^{L^{(N)} \sim \mu^{D_N}} \left\| L^\dagger - L^{(N)} \right\|_{L^2_{\nu'}(H;H)}^2 = O\big(\rho_N\big) + o\big(N^{-\left(\frac{\alpha'+s}{\alpha+p}\right)}\big) \tag{4.30}$$

with probability at least $1 - c_1 \exp(-c_2 N)$ over $X \sim \nu^{\otimes N}$ and

$$\mathbb{E}^{Y \mid X} \mathbb{E}^{L^{(N)} \sim \mu^{D_N}} \left\| L^\dagger - L^{(N)} \right\|_{L^2_{\nu'}(H;H)}^2 = \Omega\Big(\rho_N + \sum_{j > J_N} j^{-2\alpha'} |l_j^\dagger|^2\Big) \tag{4.31}$$

with probability at least $1 - c_1 \exp(-c_2 N)$ over $X \sim \nu^{\otimes N}$. Both assertions remain valid if the inner expectations are removed and $L^{(N)}$ is replaced by the posterior mean $\bar{L}^{(N)}$.

We explicitly see that the probability of failure for Theorem 4.22 is exponentially small in the sample size. The implications of this theorem are the same as those of Theorem 4.16. The corresponding lower bounds are analogous to the in-expectation results from Theorem 4.17.

[6] A centered real-valued r.v. Z is σ^2-subgaussian, denoted by $Z \in \mathrm{SG}(\sigma^2)$, if $\mathbb{E} \exp(tZ) \leq \exp(\sigma^2 t^2/2)$ for all $t \in \mathbb{R}$ [267]. On a Hilbert space $(H, \langle \cdot, \cdot \rangle)$, a centered H-valued r.v. x is subgaussian with respect to trace-class covariance operator proxy $Q \in \mathcal{L}(H)$, denoted by $x \in \mathrm{SG}(Q)$, if there exists $q \geq 0$ such that $\mathbb{E} \exp(\langle h, x \rangle) \leq \exp(q^2 \langle h, Qh \rangle / 2)$ for all $h \in H$ [13]. It is strictly subgaussian if $Q \preccurlyeq c\, \mathbb{E}[x \otimes x]$ for some $c > 0$.

4.3.5 Excess Risk

In the previous two subsections, we bounded the test error (4.14) from above and below. It follows that corresponding bounds for the excess risk $\mathcal{E}_N$ (4.19) may be obtained by specializing to the in-distribution case $\nu' = \nu$ for the posterior mean (so $\alpha' = \alpha$).

Corollary 4.23 (expected excess risk: upper and lower bounds)**.** *Let the hypotheses of Theorems 4.16 and 4.17 be satisfied. Then the expected excess risk $\mathbb{E}^{D_N}\mathcal{E}_N$ satisfies the bounds*

$$\mathbb{E}^{D_N}\,\mathbb{E}^{x\sim\nu}\|L^\dagger x - \bar{L}^{(N)}x\|^2 = O\big(N^{-\left(\frac{\alpha+p-1/2}{\alpha+p}\right)}\big) + o\big(N^{-\left(\frac{\alpha+s}{\alpha+p}\right)}\big) \quad \text{as} \quad N\to\infty\,, \tag{4.32}$$

and for any positive sequence $\{\tau_n\}$ such that $\tau_n \to 0$ and $n\tau_n \to \infty$ as $n\to\infty$, it holds that

$$\mathbb{E}^{D_N}\,\mathbb{E}^{x\sim\nu}\|L^\dagger x - \bar{L}^{(N)}x\|^2 = \Omega\Big(\tau_N N^{-\left(\frac{\alpha+p-1/2}{\alpha+p}\right)} + \sum_{j>J_N} j^{-2\alpha}|l_j^\dagger|^2\Big) \quad \text{as} \quad N\to\infty\,. \tag{4.33}$$

Corollary 4.23 is proved as a consequence of Theorems 4.16 and 4.17. A similar result may be established for $\mathbb{E}^{Y\,|\,X}\mathcal{E}_N$ by using Theorem 4.22. We omit the details for brevity. It is also interesting that *fast rates* for the excess risk (i.e., faster than $N^{-1/2}$ [188]) are attained by the posterior mean eigenvalue estimator in certain regimes. The usual statistical learning techniques based on bounding suprema of empirical processes typically yield slow $N^{-1/2}$ rates or worse [257]. Our results are sharper because we use explicit diagonal calculations.

4.3.6 Generalization Gap

Last, we estimate the generalization gap (4.20) in $L^1_{\mathbb{P}}(\Omega;\mathbb{R})$.

Theorem 4.24 (expected generalization gap: upper and lower bounds)**.** *Let the hypotheses of Theorem 4.16 be satisfied. Then for $\mathcal{G}_N$ as in (4.20), it holds that*

$$\mathbb{E}^{D_N}|\mathcal{G}_N| = O\big(N^{-\left(\frac{1}{2}\wedge\frac{\alpha+p-1/2}{\alpha+p}\right)}\big) \quad \text{as} \quad N\to\infty\,. \tag{4.34}$$

Additionally, for any positive sequence $\{\tau_n\}$ such that $\tau_n\to 0$ and $n^{1/2}\tau_n\to\infty$ as $n\to\infty$,

$$\mathbb{E}^{D_N}|\mathcal{G}_N| = \Omega\big(\tau_N N^{-\left(\frac{\alpha+p-1/2}{\alpha+p}\right)}\big) \quad \text{as} \quad N\to\infty \tag{4.35}$$

if $(\alpha+s)/(\alpha+p)\geq 2$. Otherwise, the previous assertion (4.35) remains valid provided that $p<1+\alpha+2s$ and $\tau_n \gg n^{-1/2}\vee n^{-(1+\alpha+2s-p)/(2\alpha+2p)}$ as $n\to\infty$.

We see that the expected generalization gap decays at least as fast as the standard Monte Carlo parametric rate $N^{-1/2}$ if $\alpha + p \geq 1$. Otherwise, it decays at a slower rate that is arbitrarily slow as $\alpha + p$ approaches $1/2$ from above. The lower bound only matches the latter contribution.

4.4 Numerical Experiments

We now instantiate our operator learning framework numerically, both according to the theory (Section 4.4.1) and beyond (Section 4.4.2). For clarity, we only implement the posterior mean estimator $\bar{L}^{(N)}$. Our conceptually infinite-dimensional problem must be carefully discretized to avoid obscuring the theoretical infinite-dimensional behavior [6, Section 1.2]. We use spectral truncation [6, 9] to finite-dimensionalize infinite sequence spaces. For $v = \{v_j\} \in \mathbb{R}^\infty$, its truncation is $v^{(J)} \coloneqq \{v_j\}_{j \leq J} \in \mathbb{R}^J$ for $J \in \mathbb{N}$ "Fourier" modes. We use the relative expected squared $L^2_{\nu'}$ Bochner norm as a numerical error metric, i.e.,

$$\frac{\mathbb{E}^{D_N} \mathbb{E}^{x \sim \nu'} \|L^\dagger x - \bar{L}^{(N)} x\|^2}{\mathbb{E}^{x \sim \nu'} \|L^\dagger x\|^2} = \frac{\mathbb{E}^{D_N} \sum_{j=1}^\infty \vartheta_j'^2 \left| l_j^\dagger - \bar{l}_j^{(N)} \right|^2}{\sum_{k=1}^\infty \vartheta_k'^2 |l_k^\dagger|^2}. \tag{4.36}$$

4.4.1 Within the Theory

We now confirm the theoretical results of this chapter with simulation studies. Define $A \colon \mathcal{D}(A) \subset H \to H$ by $h \mapsto Ah \coloneqq -\Delta h$ with domain $\mathcal{D}(A) \coloneqq H_0^1(I; \mathbb{R}) \cap H^2(I; \mathbb{R})$, where $I \coloneqq (0,1)$, $H \coloneqq L^2(I; \mathbb{R})$, and Δ is the Laplacian (i.e., second derivative). We consider truths $L^\dagger = A, \mathrm{Id}$, and A^{-1} corresponding to unbounded, bounded, and compact self-adjoint operators on H, respectively. The map A is diagonalized in the orthonormal basis $\{\varphi_j\}$ of H given by $z \mapsto \varphi_j(z) = \sqrt{2} \sin(j\pi z)$. This is the output space basis used henceforth. Then $L^\dagger = A, \mathrm{Id}$, and A^{-1} have eigenvalue sequences $l^\dagger = \{(j\pi)^2\}, \{1\}$, and $\{(j\pi)^{-2}\} \in \mathcal{H}^s$ for any $s < s^\star$, where $s^\star = -5/2, -1/2$, and $3/2$, respectively. These eigenvalues are regularly varying (with $S \equiv 1$) as in Theorem 4.18.

We work in the Gaussian setting of Theorem 4.3. We choose Matérn-like covariances

$$\Lambda = \tau_1^{2\alpha - 1} (A + \tau_1^2 \,\mathrm{Id})^{-\alpha} \quad \text{and} \quad \Lambda' = \tau_2^{2\alpha' - 1} (A + \tau_2^2 \,\mathrm{Id})^{-\alpha'} \tag{4.37}$$

for ν and ν'. Here $\{\tau_i\}_{i=1,2}$ are inverse length scales. Draws from ν (resp. ν') are in $\mathcal{H}^{s'}$ for all $s' < \alpha - 1/2$ (resp. $s' < \alpha' - 1/2$). Notice that $L^\dagger$, Λ, and Λ' are simultaneously diagonalizable in $\{\phi_j \equiv \varphi_j\}$. The eigenvalues are $\lambda_j(\Lambda) =$

Table 4.1: Matching test measure. Theoretical vs. experimental (in parentheses) convergence rate exponents r in $O(N^{-r})$ of the relative expected squared $L^2_\nu(H;H)$ in-distribution error (i.e., the scaled excess risk $\mathbb{E}^{D_N}\mathcal{E}_N$).

$L^\dagger$	{*Operator Class*}	Rough Prior	Matching Prior	Smooth Prior
A	{*Unbounded*}	0.714 (0.714)	0.800 (0.809)	0.615 (0.616)
Id	{*Bounded*}	0.867 (0.865)	0.889 (0.889)	0.762 (0.762)
A^{-1}	{*Compact*}	0.913 (0.913)	0.923 (0.920)	0.828 (0.830)

$\vartheta_j^2 = \tau_1^{2\alpha-1}((j\pi)^2+\tau_1^2)^{-\alpha} \asymp j^{-2\alpha}$ and similarly for $\lambda_j(\Lambda') = \vartheta_j'^2 \asymp j^{-2\alpha'}$. These satisfy Assumption (A4). We directly define the prior covariance Σ in sequence space according to Assumption (A3), choosing $\sigma_j^2 := \tau_3^{2p-1}((j\pi)^2+\tau_3^2)^{-p} \asymp j^{-2p}$ for $\tau_3 > 0$. We enforce Assumption (A5) for the values of α, α', and p.

An independent random dataset D_N (as in (4.6)) is generated for each sample size $N \in \mathbb{N}$ to construct $\bar{l}^{(N)}$. For each N, this is repeated 250, 500, or 1000 times for $L^\dagger = A, \mathrm{Id}$, and A^{-1}, respectively, to approximate the outer expectation in (4.36) by sample averages. Convergence rates are produced by linear least square fits to the logarithm of computed errors. We fix the noise scale to be $\gamma = 10^{-1}, 10^{-3}$, and 10^{-5} for $L^\dagger = A, \mathrm{Id}$, and A^{-1}, respectively.

4.4.1.1 In-Distribution

We set $\alpha = \alpha' = 4.5$ (in-distribution), $\tau_1 = \tau_2 = 15$, $\tau_3 = 1$, and define the prior smoothness $p = p(L^\dagger) = 1/2 + s^\star(L^\dagger) + z$, where $z = -0.75, 0$, or 0.75 is a fixed shift to replicate rough, matching, or smooth priors, respectively. Sequences are discretized by keeping up to $J = 2^{16} = 65,536$ Fourier modes. The sample size is $N \in \{2^4, 2^5, \ldots, 2^{14}\}$. Table 4.1 empirically verifies our sharp theoretical predictions from Theorem 4.18 for $\mathbb{E}^{D_N}\mathcal{E}_N$. The convergence as N increases is visualized in Figure 4.4d for the smooth prior case.

Moving on to study the rates of convergence of $\mathbb{E}^{D_N}\mathcal{E}_N$ and $\mathbb{E}^{D_N}|\mathcal{G}_N|$ for unbounded $L^\dagger = A$ in more detail, we now use *N-dependent* spectral truncation. For each N, we only take Fourier modes from the set $\{j \in \mathbb{N} \colon j \le cJ_N\}$, where $c > 0$ is a tunable constant and $J_N := N^{1/(2\alpha+2p)} \ll N$. This approach is justified because it is more stable numerically and the results in Section 4.3 remain valid with this N-dependent truncation. Contributions from the tail set $\{j \in \mathbb{N} \colon j > cJ_N\}$ are of equal order or negligible, asymptotically, relative to those from the truncated set (Appendix C.1). Figure 4.3 shows results with N

Table 4.2: Distribution shift. Theoretical vs. experimental (in parentheses) convergence rate exponents r in $O(N^{-r})$ of the relative expected squared $L^2_\nu(H;H)$ out-of-distribution error (4.36) for rougher and smoother test measures.

	Rougher Test Measure: $\alpha' = 4 < \alpha = 4.5$			*Smoother Test Measure*: $\alpha' = 5.25 > \alpha = 4.5$		
$L^\dagger$	Rough Prior	Matching Prior	Smooth Prior	Rough Prior	Matching Prior	Smooth Prior
A	0.429 (0.428)	0.600 (0.607)	0.462 (0.462)	1.000 (0.992)	1.000 (0.996)	0.846 (0.849)
Id	0.733 (0.734)	0.778 (0.788)	0.667 (0.667)	1.000 (0.986)	1.000 (0.979)	0.905 (0.905)
A^{-1}	0.826 (0.837)	0.846 (0.861)	0.759 (0.764)	1.000 (0.981)	1.000 (0.975)	0.931 (0.926)

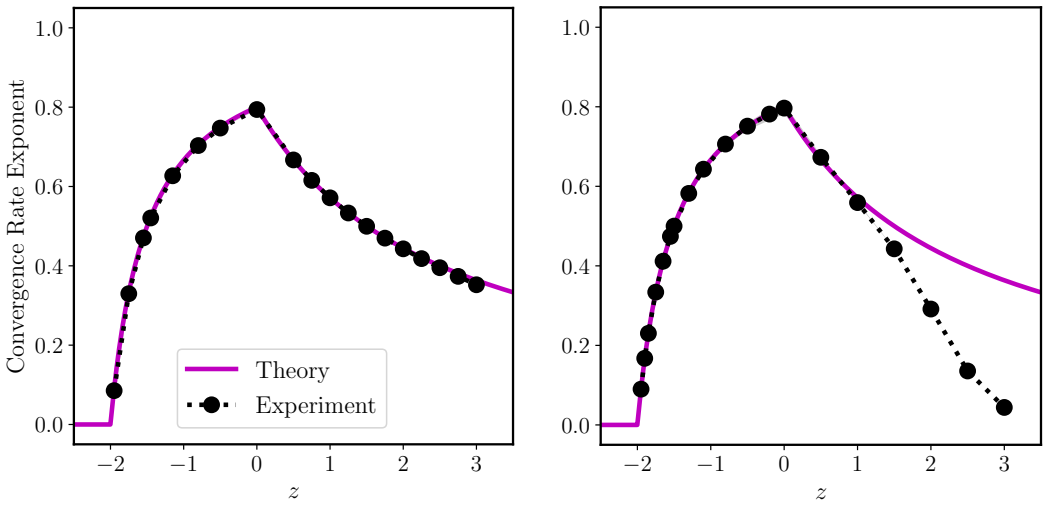
up to 2^{21} and c such that $cJ_{2^{21}} \approx 2^{14}$ (maximal truncation level). The influence of discretization manifests itself through γ. For $\mathbb{E}^{D_N} \mathcal{E}_N$, the under-smoothing prior region ($z < 0$) is relatively insensitive to γ and the rate exponents closely match (4.32). But in the over-smoothing prior region $z > 0$ for large γ, the rates begin to deviate from the theory because large constants mask the theoretical asymptotic behavior in this finite sample regime. Similarly, for finite N, the noise scale can alter the correct behavior of the competing terms in the bound (4.34) for $\mathbb{E}^{D_N}|\mathcal{G}_N|$. For small γ, terms $O(N^{-1/2})$ have large hidden constants that obscure terms $\gg N^{-1/2}$ for small $z < 0$ (Figure 4.3c). For large γ, this behavior is reversed (Figure 4.3d).

4.4.1.2 Out-of-Distribution

We now vary α' to simulate distribution shift. With $J = 2^{16}$ and $\tau_2 = 15$, our results in Table 4.2 show near perfect agreement with Theorem 4.18 for out-of-distribution regimes on both sides of the boundary case $\alpha' = \alpha + 1/2$. In the matching prior setting ($z = 0$), Figures 4.4a to 4.4c show the decay of the test error (4.36) with N. The magenta lines are least square fits and the shaded regions denote one standard deviation from the mean with respect to resampling D_N. The excellent numerical fits verify our assertions.

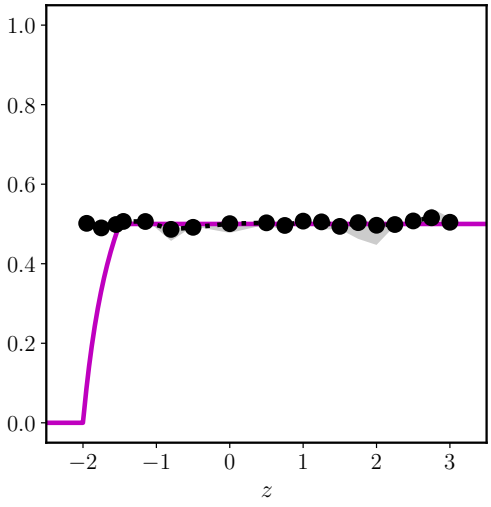
4.4.2 Beyond the Theory

In this subsection, we consider truths $L^\dagger \in \mathrm{HS}(H_\mathcal{K}; H)$ (with $\mathcal{K}$ satisfying Condition 4.5) that are *not necessarily diagonalized* by $\{\varphi_j\}$. So, the infinite matrix $\mathsf{L}^\dagger := \{\mathsf{L}^\dagger_{jk}\}$ from (4.3) must be estimated instead of $l^\dagger$. Recall that Λ has eigenpairs $\{(\lambda_k^2, \phi_k)\}$. By Fact 4.4, $L^\dagger \in \mathrm{HS}(H_\Lambda; H)$ so the expansion $L^\dagger = \sum_{i,j}(\lambda_j \mathsf{L}^\dagger_{ij})\varphi_i \otimes_{H_\Lambda} (\lambda_j \phi_j) = \sum_{i,j} \mathsf{L}^\dagger_{ij} \varphi_i \otimes \phi_j$ always exists and is unique. Yet, we have no theory for posterior estimators of $\mathsf{L}^\dagger$. To derive the posterior mean, we notice that the inverse problem for $\mathsf{L} \,|\, D_N$ decouples along rows of

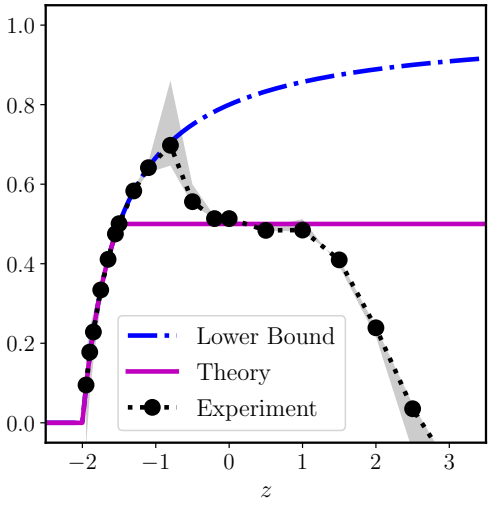


(a) $\mathbb{E}^{D_N}\,\mathcal{E}_N$, $\gamma^2 = 10^{-6}$

(b) $\mathbb{E}^{D_N}\,\mathcal{E}_N$, $\gamma^2 = 25^2$

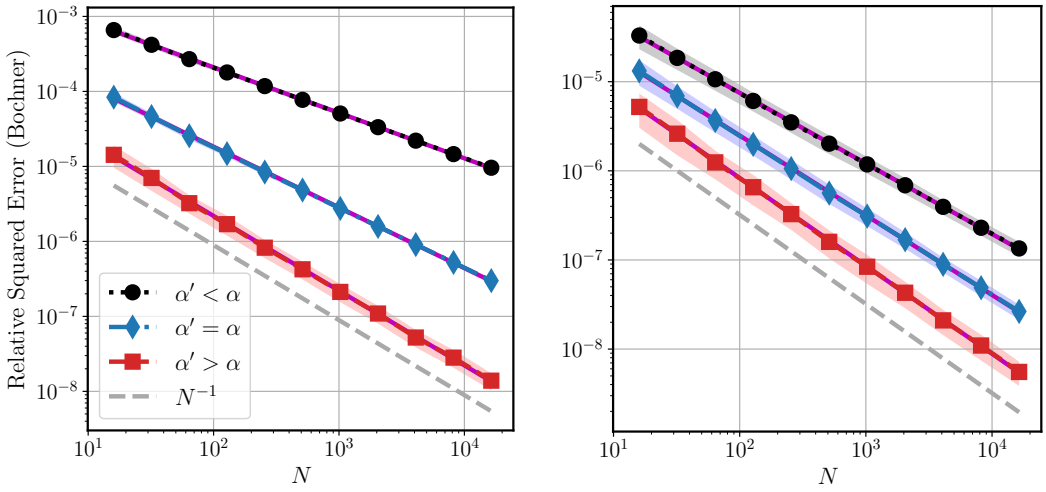


(c) $\mathbb{E}^{D_N}|\mathcal{G}_N|$, $\gamma^2 = 10^{-6}$



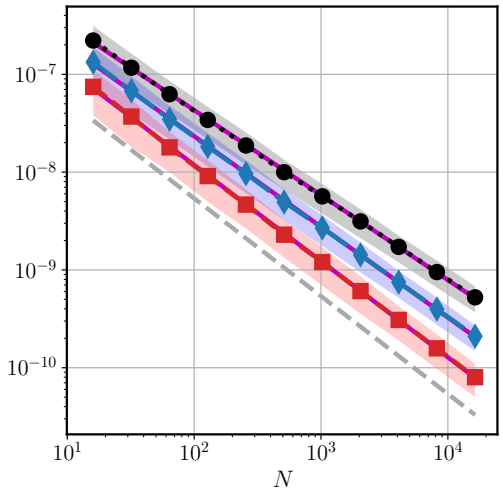
(d) $\mathbb{E}^{D_N}|\mathcal{G}_N|$, $\gamma^2 = 25^2$

Figure 4.3: The numerical influence of data noise variance γ^2 for $L^\dagger = A$. For two distinct γ^2 values, Figures 4.3a and 4.3b show convergence rate exponents for $\mathbb{E}^{D_N}\,\mathcal{E}_N$ vs. z, with $z = p + 2$ being the prior smoothness shift parameter, while Figures 4.3c and 4.3d display rates for $\mathbb{E}^{D_N}|\mathcal{G}_N|$ vs. z. Throughout, the solid magenta "Theory" curves denote the theoretical upper bound rate exponents, and the shaded regions denote one standard deviation from the mean rate exponent computed from 250 repetitions of the numerical experiment.

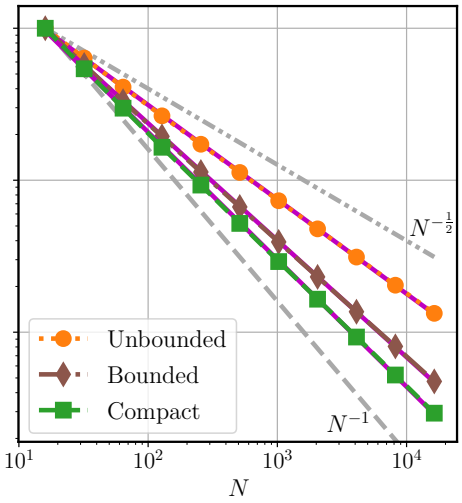


(a) $L^\dagger = A$, $\gamma = 10^{-1}$

(b) $L^\dagger = \mathrm{Id}$, $\gamma = 10^{-3}$



(c) $L^\dagger = A^{-1}$, $\gamma = 10^{-5}$



(d) $\mathbb{E}^{D_N} \mathcal{E}_N$ vs. N

Figure 4.4: Within the theory. Figures 4.4a to 4.4c are such that $z = 0$ (matching $p = s^\star + 1/2$) and the test measures ν' are either equal to ($\alpha' = \alpha$), rougher than ($\alpha' < \alpha$), or smoother than ($\alpha' > \alpha$) the training measure ν. For fixed $L^\dagger$, the same $\bar{L}^{(N)}$ achieves smaller relative error (4.36) as α' increases, that is, when testing against smoother input functions. In all cases, the observed rates closely match the theoretical ones (see Tables 4.1 and 4.2). Figure 4.4d (corresponding to Table 4.1 column four) shows that convergence improves with increased operator smoothing (the logarithmic vertical axis is rescaled to ease comparison of the slopes).

$\mathsf{L} = \{\mathsf{L}_{jk}\}$, which are denoted by $\mathsf{L}_{j:}$ for $j \in \mathbb{N}$. We assume a Gaussian prior $\mathsf{L}_{j:} \sim \mathcal{N}(0, \Sigma_j)$, where $\Sigma_j = \operatorname{diag}(\{\sigma_{jk}^2\}_{k\in\mathbb{N}})$ is diagonal for simplicity. Thus $(\mathsf{L}_{j:})_k = \mathsf{L}_{jk} \sim \mathcal{N}(0, \sigma_{jk}^2)$. By deriving the normal equations, we obtain for j, k, and $\ell \in \mathbb{N}$ the posterior mean

$$\bar{\mathsf{L}}_{j:}^{(N)} = \big(\mathsf{A}^{(N)} + \tfrac{\gamma^2}{N}\Sigma_j^{-1}\big)^{-1}\mathsf{b}_j^{(N)}, \quad \text{where} \quad \mathsf{A}_{\ell k}^{(N)} \coloneqq \overline{x_\ell x_k}^{(N)} \quad \text{and} \quad \big(\mathsf{b}_j^{(N)}\big)_\ell \coloneqq \overline{y_j x_\ell}^{(N)}. \tag{4.38}$$

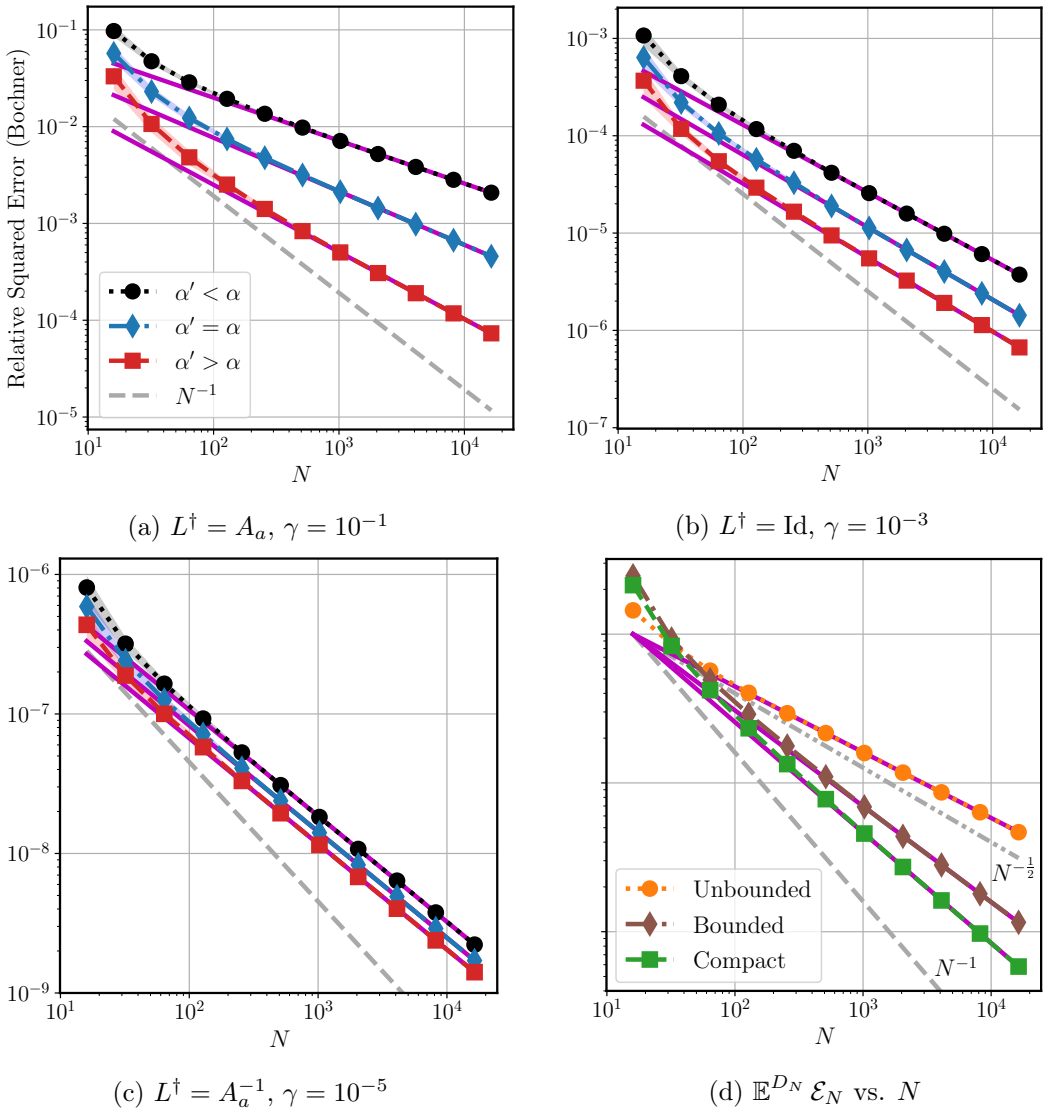
We use the same covariances (4.37) diagonalized in the Fourier sine input basis $\{\phi_j\}$, but now use the Volterra cosine output basis $\{\varphi_j\}$ as in Figure 4.2, where $z \mapsto \varphi_j(z) \coloneqq \sqrt{2}\cos((j-\frac{1}{2})\pi z)$. Define the *divergence form elliptic operator* $A_a \colon \mathcal{D}(A_a) \subset H \to H$ by $h \mapsto A_a h \coloneqq -\nabla\cdot(a\nabla h)$, where $\mathcal{D}(A_a) = \mathcal{D}(A)$ is as before and $z \mapsto a(z) \coloneqq \exp(-3z)$ is a smooth coefficient function. We learn (via $\bar{\mathsf{L}}^{(N)}$) unbounded, bounded, and compact self-adjoint operators $L^\dagger = A_a, \mathrm{Id}$, and A_a^{-1}, respectively. For each of the three $L^\dagger$, we pick prior variance sequences $\sigma_{jk}^2 = \sigma_{jk}^2(L^\dagger)$ given by

$$\sigma_{jk}^2(L^\dagger) \coloneqq \begin{cases} (jk)^{-(z-2)}\big(\frac{1+(k/j)^2}{1+(j-k)^2}\big)^2, & \text{if } L^\dagger = A_a\,, \\ (jk)^{-z}\big(\frac{k+k/j}{1+j+(j-k)^2}\big)^2, & \text{if } L^\dagger = \mathrm{Id}\,, \\ (jk)^{-(z+2)}\big(\frac{1+j/k}{1+(j-k)^2}\big)^2, & \text{if } L^\dagger = A_a^{-1}. \end{cases} \tag{4.39}$$

These priors ensure that L matches the exact asymptotic behavior (as $j \to \infty$, $k \to \infty$, and $j = k \to \infty$) of $\mathsf{L}^\dagger$ when $z = 0$. Our simulation setup follows Section 4.4.1, except now with $J = 2^{12}$, N up to 2^{14}, and only 100 Monte Carlo repetitions. Although A_a is not diagonal in $\varphi_j \neq \phi_j$ (each $\mathsf{L}^\dagger$ is dense) and the posterior mean estimator is now a doubly-indexed sequence, our results in Figure 4.5 support the same conclusions previously asserted.

4.5 Conclusion

This chapter concerns the supervised learning of linear operators between Hilbert spaces. Learning is framed as a Bayesian inverse problem with a linear operator as the unknown quantity. Working in the best-case scenario of known eigenvectors, the analysis establishes convergence rates in the infinite data limit. The main results reveal useful theoretical insights about operator learning, including what types of operators are harder to learn than others, what types of training data lead to reduced sample complexity, and how distribution shift affects error. The work opens up the following directions for future research.



(a) $L^\dagger = A_a$, $\gamma = 10^{-1}$

(b) $L^\dagger = \mathrm{Id}$, $\gamma = 10^{-3}$

(c) $L^\dagger = A_a^{-1}$, $\gamma = 10^{-5}$

(d) $\mathbb{E}^{D_N} \mathcal{E}_N$ vs. N

Figure 4.5: Beyond the theory. Analogous to Figure 4.4 except with the non-diagonal elliptic operator A_a.

Extensions in Diagonal Setting. One immediate extension of our diagonal approach involves generalizing it from self-adjoint operators with known eigenvectors to non-self-adjoint operators with known singular vectors. Another involves taking the simultaneous large data and small noise limit under both well-specified and misspecified likelihoods. Although our approach requires Gaussian conjugacy, Gaussian priors are not suitable for all problems. Recent work using non-conjugate priors may prove useful in our setting [119, 145, 230, 263]. To exploit the Bayesian posterior beyond just theoretical contraction performance, exploration of uncertainty quantification via credible sets is also of interest.

Beyond Diagonal Operators. In the linear setting, it is desirable to remove the known eigenbasis assumption but retain rates of convergence. The proof of Fact 4.6 in Appendix C.3 implies that the SVD of the random forward map K_X in (4.2) is determined by functional PCA of X. Thus, the SVD approach in Section 4.1.1.2 and [147] could be used to recover the doubly-indexed infinite matrix coordinates of the true operator in the (random) SVD basis. Another approach is to directly study the non-diagonal problem (4.3) as in Section 4.4.2. Nonlinear operators also deserve attention, as the experimental results in [75] demonstrate. Central to their statistical analysis will be the modern architectures (beyond kernel methods [52, 229]) that parametrize the unknown operators and their inherent problem-dependent structure.

Chapter 5

AN OPERATOR LEARNING PERSPECTIVE ON PARAMETER-TO-OBSERVABLE MAPS

This chapter is adapted from the following preprint:

[1] Daniel Zhengyu Huang, Nicholas H. Nelsen, and Margaret Trautner. "An operator learning perspective on parameter-to-observable maps". *preprint arXiv:2402.06031 cs.LG* (2024). DOI: 10.48550/arXiv.2402.06031.

Computationally efficient surrogates for parametrized physical models play a crucial role in science and engineering. Operator learning provides data-driven surrogates for such models that map between function spaces. However, instead of full-field measurements, often the available data are only finite-dimensional parametrizations of model inputs or finite observables of model outputs. Building on Fourier Neural Operators, this chapter introduces the Fourier Neural Mappings (FNMs) framework that is able to accommodate such finite-dimensional vector inputs or outputs. The work develops universal approximation theorems for the method. Moreover, in many applications the underlying parameter-to-observable (PtO) map is defined implicitly through an infinite-dimensional operator, such as the solution operator of a partial differential equation. A natural question is whether it is more data-efficient to learn the PtO map end-to-end or to first learn the solution operator and subsequently compute the observable from the full-field solution. A theoretical analysis of Bayesian nonparametric regression of linear functionals, which is of independent interest, suggests that the end-to-end approach can actually have worse sample complexity in some regimes. Going beyond the theory, numerical results for the FNM approximation of three nonlinear PtO maps demonstrate the benefits of the operator learning perspective that this chapter adopts.

5.1 Introduction

Operator learning has emerged as a methodology that enables the machine learning of maps between spaces of functions. Many surrogate modeling tasks in areas such as uncertainty quantification, inverse problems, and design optimization involve a map between function spaces, such as the solution operator

of a partial differential equation (PDE). However, the primary quantities of interest (QoI) in these tasks are usually just a finite number of design parameters or output observables. This may be because full-field data, such as initial conditions, boundary conditions, and solutions of PDEs, are not accessible from measurements or are too expensive to acquire. The prevailing approach then involves emulating the parameter-to-observable (PtO) map instead of the underlying solution map between function spaces. Yet, it is natural to wonder if the success of operator learning in the function-to-function setting can be brought to bear in this more realistic setting where inputs or outputs may necessarily be finite-dimensional vectors. To this end, the present chapter introduces Fourier Neural Mappings (FNMs) as a way to extend operator learning architectures such as the Fourier Neural Operator (FNO) to finite-dimensional input and output spaces in a manner that is compatible with the underlying operator between infinite-dimensional spaces. The admissible types of FNM models considered in this work are visualized in Figure 5.1.

Nevertheless, it is possible to accommodate finite-dimensional inputs or outputs through other means. For instance, one could lift a finite-dimensional input vector to a function by expanding in predetermined basis functions, apply traditional operator learning architectures to the full-field function space data, and then directly compute a known finite-dimensional QoI from the output function. In contrast, the end-to-end FNM approach in this work is fully data-driven and operates directly on finite-dimensional vector data without the need for pre- and postprocessing. A natural question is whether one of these two approaches achieves better accuracy than the other when the goal is to predict certain QoIs. In the present chapter, we address this important question both from a theoretical and a numerical perspective. Indeed, it has been empirically observed in various nonlinear problems ranging from electronic structure calculations [260] to metamaterial design [24] that data-driven methods that predict the full-field response of a system are superior to end-to-end approaches for the same downstream tasks or QoIs. The analysis in this chapter takes the first steps toward a rigorous theoretical understanding of these empirical findings. The theory adapts the techniques developed in the previous chapter to this more challenging setting.

Throughout the chapter, we refer to learning a function-valued map as *full-field learning*. Given such a learned map, various known QoIs may be directly

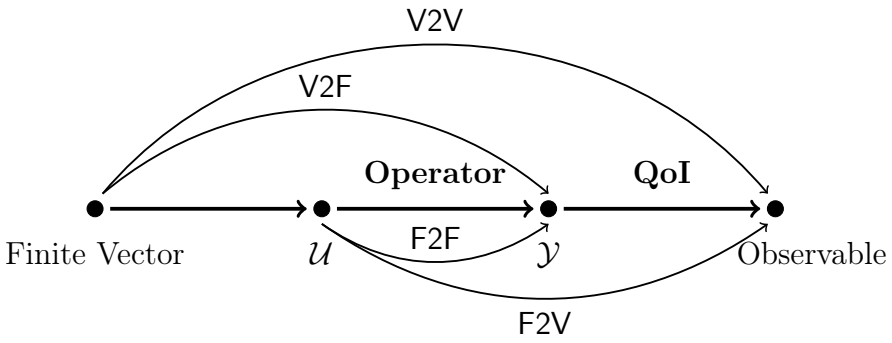


Figure 5.1: Illustration of the factorization of an underlying PtO map into a QoI and an operator between function spaces. Also shown are the four variants of input and output representations considered in this work. Here, $\mathcal{U}$ is an input function space and $\mathcal{Y}$ is an intermediate function space.

computed from the output of the map. On the other hand, we refer to the direct estimation of the map from an input to the observed QoI as *end-to-end learning*. This terminology distinguishes between output spaces. When either the input or output is finite- and the other is infinite-dimensional, we label it as "vector-to-function" (V2F) or "function-to-vector" (F2V), respectively, to avoid ambiguity. The abbreviations V2V and F2F for "vector-to-vector" and "function-to-function" are analogous.

5.1.1 Contributions

In this chapter, we make the following contributions.

(C1) We introduce FNMs as a function space architecture that is able to accommodate finite-dimensional vector inputs, outputs, or both.

(C2) We prove universal approximation theorems for FNMs.

(C3) We establish convergence rates for Bayesian nonparametric regression of linear functionals under smoothness misspecification; as a byproduct of this analysis, we prove that full-field learning of linear functionals that are factorized into the composition of a linear QoI and a linear operator enjoys better sample complexity than end-to-end estimators in certain regimes.

(C4) We perform numerical experiments with FNMs in three examples—an advection–diffusion equation, flow over an airfoil, and an elliptic homogenization problem—that show empirical evidence that the theoretical

linear intuition from Contribution (C3) remains valid for nonlinear maps.

Next, we provide an overview of related work in the literature in Subsection 5.1.2. Subsection 5.1.3 contains relevant notation, and Subsection 5.1.4 gives an outline of the remainder of the chapter.

5.1.2 Related Work

Several works have established neural operators as a viable tool for scientific machine learning. The general neural operator formalism is described in [154] and contains several subclasses including DeepONet [181], graph neural operator [173, 174], and FNO [172]. These architectures allow for function data evaluated at different grid points or resolutions to be used with the same model. In particular, the FNO is primarily parametrized in Fourier space. It exploits the fact that the Fourier basis spans L^2 on the torus and uses the efficient Fast Fourier Transform (FFT) algorithm for computations. The idea of parametrizing operators in Fourier space is explored in earlier works as well [203, 214]. The FNO has been shown to be applicable both to domains other than the torus and to nonuniform meshes [171, 176]. These neural operators have been used in various areas of application, including climate modeling [156], fracture mechanics [116], and catheter design [281]. In several of these applications, neural operators have been implemented with finite-dimensional vector inputs or outputs by using constant functions as replacements for finite vectors, which is theoretically justified by statements of universal approximation [35], or by using other hand-designed maps. However, learning a constant function as a representation for a constant is arguably unnatural and computationally wasteful; it is desirable to substitute a more suitable architecture. The present chapter develops FNMs that extend neural operators to this important setting while retaining desirable universal approximation properties.

The theory of neural operators—and scientific machine learning more broadly—generally falls into three tiers. In the first tier, universal approximation results [59, 71, 129, 90] use classical approximation theory to guarantee that the architecture is capable of representing maps from within a class of interest to any desired accuracy. Some of the proofs of these results contain constructive arguments, but the corresponding architectures are usually not as empirically effective as those that solely come with existence results. Examples of con-

structive arguments for operator approximation are contained in [124], which constructs ReLU neural networks, and [40], which uses randomized numerical linear algebra to sketch Green's functions for linear elliptic PDEs. Each of these works also comes equipped with convergence rates with respect to model size and data size, respectively; these rates form the second tier of operator learning theory. Many papers in this tier prove bounds on the required model size, i.e., parameter complexity [81, 125, 152, 158, 161, 163, 178, 180, 242]. Some are able to obtain sample complexity bounds, although most results are restricted to linear or kernelized settings [52, 76, 135, 162, 169, 196, 239, 251]. The third tier of theory describes the likelihood of actually obtaining an accurate approximation through optimization. While some results along these lines exist for linear models, linear maps, and constructive operators [162, 164], they are absent for the class of neural operators optimized through variants of stochastic gradient descent (SGD). This is the class that has proven empirically most effective in applications thus far and is the class used in this work.

To provide theoretical intuition for nonlinear settings, this chapter establishes convergence rates in the tractable setting of learning a linear functional—the PtO map—from noisy data. The setting of functional linear regression has a long history in statistics [48, 54, 149, 231, 278]. A zoo of different estimators exist, e.g., those based on reproducing kernel Hilbert space (RKHS) methods, principal component analysis, and wavelets. We target the frequentist convergence properties of a linear Gaussian posterior estimator that arises from reinterpreting the regression task as a Bayesian inverse problem. Apart from the previous chapter [76], the closest work to ours is [175]. There, the authors derive posterior contraction rates for Bayesian functional linear regression with Gaussian priors. However, their main results are only sharp if the prior is correctly specified to match the regularity of the true linear functional. Our work goes beyond this by proving sharp high-probability error bounds for out-of-distribution prediction error under very general smoothness misspecification. However, to do so we require that the prior and data covariance operators commute, which is a limitation of our theory. Another relevant work is [49], in which the authors make minimal assumptions on the prior and data covariances but still make the well-specified assumption that the truth belongs to the RKHS of the prior. They also require a data-dependent scaling of the prior. Similarly, [216] obtains convergence guarantees for Bayesian nonparametric regression of functions with Gaussian process priors in the misspecified setting.

However, their work only considers a squared exponential covariance structure and requires a careful rescaling of the prior to deal with the smoothness misspecification. Our work holds for the larger class of misspecified Matérn-like Gaussian priors and delivers rates without rescaling the prior distribution.

Beyond linear functionals, recent work proposes and analyzes a kernelized deep learning method for nonlinear functionals [243]. The idea of such neural functionals, a subclass of the FNMs proposed in this work, is not new. One appearance is in the context of a function space discriminator for generative adversarial networks [224]. However, that work uses only a single bounded linear functional that is appended to the output of a FNO and is parametrized by a standard neural network. This is a special case of our FNMs for F2V maps. Another paper that shares similar ideas to the present chapter is [280]. There, the authors also formulate a V2V neural network approach that maps through a latent 1D function space. However, their encoder and decoder maps are prescribed by hand-picked basis functions, while for FNMs the encoder and decoder maps are learned from data.

In this chapter, three illustrative applications are highlighted. The first application is an advection–diffusion model where the input is a velocity field and the output is the state at a fixed future time. This problem is considered a benchmark for scientific machine learning [258]. Some theoretical approximation rates for it have been developed for DeepONet in the F2F setting [81]. The second application centers on the compressible flow over an airfoil, i.e., an airplane wing cross section. This experiment is explored for FNO in [171] and used as a shape optimization example in the F2F setting for DeepONet in [244] and for reduced basis networks in [206]. Several other related works devise V2V-based neural network approaches and novel training strategies for this aerodynamics problem [183, 184, 185, 194]. The third application involves learning the homogenized elasticity coefficient for a multiscale elliptic PDE. This example is explored in detail for FNO in [35] and for other constitutive laws in [177]. For the Darcy flow—or scalar coefficient—setting of this equation, other work has adopted the F2F setting to efficiently compute QoIs [277]. For each of these applications, we compare the generalization error performance of all four F2F, F2V, V2F, and V2V variants of FNMs as well as standard fully-connected neural networks.

5.1.3 Notation

The set of continuous linear operators from a Banach space $\mathcal{U}$ to a Banach space $\mathcal{V}$ is denoted by $\mathcal{L}(\mathcal{U};\mathcal{V})$, and if $\mathcal{U} = \mathcal{V}$, then we write $\mathcal{L}(\mathcal{V})$. For a separable Hilbert space $(H, \langle\cdot,\cdot\rangle, \|\cdot\|)$, the outer product operator $a \otimes b \in \mathcal{L}(H)$ is defined by $(a \otimes b)c \coloneqq \langle b, c\rangle a$ for any elements a, b, and c of H. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space that is sufficiently rich to support all random variables that appear in this chapter. All expectations are interpreted as Bochner integrals. For a probability measure Π supported on H with two finite moments, we denote by $\mathrm{Cov}(\Pi) = \mathbb{E}^{u\sim\Pi}[(u - \mathbb{E}\,u) \otimes (u - \mathbb{E}\,u)] \in \mathcal{L}(H)$ its covariance operator. Independent and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots, X_n$ from Π are denoted by $\{X_i\}_{i=1}^n \sim \Pi^{\otimes n}$. For two random variables X and Z, the conditional expectation notation $\mathbb{E}^{Z\,|\,X}[\,\cdot\,]$ denotes integration with respect to the law of $Z\,|\,X$. For $N \in \mathbb{N}$, we write $[N] \coloneqq \{1, 2, \ldots, N\}$. For two nonnegative real sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, we write $a_n \lesssim b_n$ if there exists $c > 0$ such that $a_n \le c b_n$ for all $n \in \mathbb{N}$ and $a_n \simeq b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$. To denote asymptotic equivalence, we write $a_n \asymp b_n$ as $n \to \infty$ if there exist $c \ge 1$ and $n_0 \in \mathbb{N}$ such that $c^{-1} b_n \le a_n \le c b_n$ for all $n \ge n_0$. The Sobolev-like sequence Hilbert spaces $\mathcal{H}^s = \mathcal{H}^s(\mathbb{N}, \mathbb{R})$ are defined for $s \in \mathbb{R}$ by $\mathcal{H}^s \coloneqq \{v\colon \mathbb{N} \to \mathbb{R} \mid \sum_{j=1}^\infty j^{2s}|v_j|^2 < \infty\}$. We write $\mathbb{T}^d$ for the d-dimensional unit torus $[0,1]^d_{\mathrm{per}}$.

5.1.4 Outline

The remainder of this article is organized as follows. We define the architecture of FNMs as a slight adjustment of FNOs in Section 5.2 (Contribution (C1)) and confirm that FNMs retain desirable properties of FNOs such as universal approximation in Section 5.3 (Contribution (C2)). In Section 5.4, we analyze end-to-end and full-field learning of linear functionals to establish a theoretical foundation that underlies the data volume requirements of the two approaches (Contribution (C3)). Going beyond the theory, Section 5.5 provides numerical experiments that compare end-to-end and full-field learning with FNMs with both finite- and infinite-dimensional input space representations for predicting QoIs in several nonlinear PDE problems (Contribution (C4)). Concluding remarks are given in Section 5.6. Appendix D.1 contains additional theorems related to Section 5.4. All proofs are provided in Appendices D.2 and D.3.

5.2 Neural Mappings for Finite-Dimensional Vector Data

In this section, we recall the FNO architecture (Subsection 5.2.1) and describe modifications of it to form FNMs (Subsection 5.2.2).

5.2.1 A Review of Neural Operators

Let $\mathcal{U} = \mathcal{U}(\mathcal{D}; \mathbb{R}^{d_u})$ and $\mathcal{Y} = \mathcal{Y}(\mathcal{D}; \mathbb{R}^{d_y})$ be Banach function spaces over Euclidean domain $\mathcal{D} \subset \mathbb{R}^d$. Finite-dimensional fully-connected neural networks are repeated compositions of affine mappings alternating with pointwise nonlinearities. To extend this framework to the infinite-dimensional function space setting, depth T neural operators from $\mathcal{U}$ to $\mathcal{Y}$ take the form

$$\Psi^{(\mathrm{NO})}(u) \coloneqq \big(\mathcal{Q} \circ \mathscr{L}_T \circ \mathscr{L}_{T-1} \circ \cdots \circ \mathscr{L}_1 \circ \mathcal{S}\big)(u) \quad \text{for all} \quad u \in \mathcal{U}\,, \tag{5.1}$$

where $\mathcal{S}$ is a pointwise-defined local lifting operator, $\mathcal{Q}$ is a pointwise-defined local projection operator, and for each $t \in [T]$, the layer $\mathscr{L}_t$ is a nonlinear map between function spaces that is the composition of a local (and usually nonlinear) operator with a nonlocal affine kernel integral operator [154].

The Fourier Neural Operator (FNO) is a specific instance of the class of neural operators (5.1) where, for $t \in [T]$, the form of the layer $\mathscr{L}_t \colon \{v \colon \mathbb{T}^d \to \mathbb{R}^{d_{t-1}}\} \to \{v \colon \mathbb{T}^d \to \mathbb{R}^{d_t}\}$ is given by

$$v \mapsto \mathscr{L}_t(v) = \Big\{\sigma_t\big(W_t v(x) + (\mathcal{K}_t v)(x) + b_t(x)\big)\Big\}_{x \in \mathbb{T}^d}\,. \tag{5.2}$$

In (5.2), $W_t \in \mathbb{R}^{d_t \times d_{t-1}}$ is a weight matrix, $b_t \colon \mathbb{T}^d \to \mathbb{R}^{d_t}$ is a bias function, and $\mathcal{K}_t$ is a convolution operator given, for $v \colon \mathbb{T}^d \to \mathbb{R}^{d_{t-1}}$ and any $x \in \mathbb{T}^d$, by the expression

$$(\mathcal{K}_t v)(x) = \left\{\sum_{k \in \mathbb{Z}^d} \left(\sum_{j=1}^{d_{t-1}} (P_t^{(k)})_{\ell j} \langle \psi_k, v_j \rangle_{L^2(\mathbb{T}^d; \mathbb{C})}\right) \psi_k(x)\right\}_{\ell \in [d_t]} \in \mathbb{R}^{d_t}\,. \tag{5.3}$$

In the preceding display, the $\psi_k = e^{2\pi \mathrm{i} \langle k, \cdot \rangle_{\mathbb{R}^d}}$ are the complex Fourier basis elements of $L^2(\mathbb{T}^d; \mathbb{C})$ and $P_t^{(k)} \in \mathbb{C}^{d_t \times d_{t-1}}$ are the learnable parameters of the integral operator $\mathcal{K}_t$ for each $k \in \mathbb{Z}^d$. The functions $\sigma_t \colon \mathbb{R} \to \mathbb{R}$ are nonlinear activations that act pointwise when applied to vectors. Additional details of more general versions and computational implementations of the FNO may be found in [152, 154, 171].

Though the internal FNO layers $\{\mathscr{L}_t\}$ in (5.2) and (5.3) are defined on the periodic domain $\mathbb{T}^d$, it is possible to apply the FNO to other domains $\mathcal{D} \subset \mathbb{R}^d$.

To do this, introduce an operator $\mathcal{E}\colon \{h\colon \mathcal{D} \to \mathbb{R}^{d_0}\} \to \{h\colon \mathbb{T}^d \to \mathbb{R}^{d_0}\}$ that maps to functions on $\mathbb{T}^d$ and replace $\mathcal{S}$ in (5.1) with $\mathcal{E} \circ \mathcal{S}$. Similarly, let $\mathcal{R}\colon \{h\colon \mathbb{T}^d \to \mathbb{R}^{d_T}\} \to \{h\colon \mathcal{D} \to \mathbb{R}^{d_T}\}$ be an operator that maps back to functions on the desired domain $\mathcal{D}$ and replace $\mathcal{Q}$ in (5.1) with $\mathcal{Q} \circ \mathcal{R}$. These modifications to the lifting and projecting components yield the final FNO architecture

$$\Psi^{(\mathrm{FNO})} = \mathcal{Q} \circ \mathcal{R} \circ \mathscr{L}_T \circ \mathscr{L}_{T-1} \circ \cdots \circ \mathscr{L}_1 \circ \mathcal{E} \circ \mathcal{S}\,. \tag{5.4}$$

In practice, the map $\mathcal{E}$ is usually represented by zero padding the input domain and $\mathcal{R}$ by restricting to the output domain of interest.

5.2.2 The Neural Mappings Framework

The neural operator architecture described in Section 5.2.1 only accepts inputs, outputs, and intermediate states that are elements of function spaces. Finite-dimensional vector inputs, outputs, and states are not directly compatible with neural operators. We propose *neural mappings*, which lift this restriction through two fundamental building blocks. The first, linear functional layers, map from function space to finite dimensions. The second, linear decoder layers, map from finite dimensions to function space. We combine these two building blocks with standard iterative neural operator layers to form several classes of nonlinear and function space consistent architectures.

Instating the neural operator notation from Section 5.2.1, we define a *linear functional layer* $\mathscr{G}\colon \{h\colon \mathcal{D} \to \mathbb{R}^{d_{T-1}}\} \to \mathbb{R}^{d_T}$ and a *linear decoder layer* $\mathscr{D}\colon \mathbb{R}^{d_0} \to \{h\colon \mathcal{D} \to \mathbb{R}^{d_1}\}$ to be maps of the form

$$\begin{aligned} h \mapsto \mathscr{G}h &\coloneqq \int_{\mathcal{D}} \kappa(x)h(x)\,dx\,, \quad \text{where} \quad \kappa\colon \mathcal{D} \to \mathbb{R}^{d_T \times d_{T-1}}\,, \quad \text{and} \\ z \mapsto \mathscr{D}z &\coloneqq \kappa(\cdot)z\,, \quad \text{where} \quad \kappa\colon \mathcal{D} \to \mathbb{R}^{d_1 \times d_0}\,, \end{aligned} \tag{5.5}$$

respectively. The linear functional layer $\mathscr{G}$ takes a vector-valued function h and integrates it against a fixed matrix-valued function κ to produce a finite vector output. In duality to $\mathscr{G}$, the linear decoder layer $\mathscr{D}$ takes as input a finite vector z and multiplies it by a fixed matrix-valued function κ to produce an output function. The functions κ are the sole learnable parameters of these two layers.

Although $\mathscr{G}$ and $\mathscr{D}$ may be incorporated into general neural operators (5.1), we will specialize our method to the FNO. In anticipation of this periodic setting,

we view $\mathscr{G}$ as a Fourier linear functional layer by replacing $\mathcal{D}$ in (5.5) by the torus $\mathbb{T}^d$ and using Fourier series to expand $\mathscr{G}$ as

$$h \mapsto \mathscr{G}h = \left\{ \sum_{k\in\mathbb{Z}^d} \left(\sum_{j=1}^{d_{T-1}} P_{\ell j}^{(k)} \langle \psi_k, h_j \rangle_{L^2(\mathbb{T}^d;\mathbb{C})} \right) \right\}_{\ell\in[d_T]} \in \mathbb{R}^{d_T}, \tag{5.6}$$

where we recall that $\{\psi_k\}$ is the Fourier basis of $L^2(\mathbb{T}^d;\mathbb{C})$. In (5.6), the entries of the matrices $\{P^{(k)}\} \subset \mathbb{C}^{d_T \times d_{T-1}}$ correspond to the Fourier coefficients of the function κ in (5.5). Similar calculations show that, on the torus $\mathbb{T}^d$, the linear decoder $\mathscr{D}$ takes the form

$$z \mapsto \mathscr{D}z = \left\{ \sum_{k\in\mathbb{Z}^d} \left(P^{(k)} z\right)_j \psi_k \right\}_{j\in[d_1]}, \quad \text{where} \quad P^{(k)} \in \mathbb{C}^{d_1 \times d_0}. \tag{5.7}$$

Just like for the FNO kernel integral layers (5.3), the expressions (5.6) and (5.7) are efficiently implemented and learned in Fourier space.

We are now able to define the general FNMs architecture.

Definition 5.1 (Fourier Neural Mappings)**.** Let $Q\colon \mathbb{R}^{d_T} \to \mathbb{R}^{d_y}$ and $S\colon \mathbb{R}^{d_u} \to \mathbb{R}^{d_0}$ be finite-dimensional maps. For $\{\mathscr{L}_t\}$ defined as in (5.4) and $\mathscr{G}$ and $\mathscr{D}$ defined as in (5.6) and (5.7), let

$$\Psi^{(\mathrm{FNM})} := Q \circ \mathscr{G} \circ \mathscr{L}_{T-1} \circ \cdots \circ \mathscr{L}_2 \circ \mathscr{D} \circ S. \tag{5.8}$$

be the base level map. The *Fourier Neural Mappings* architecture is comprised of the following four main models that are obtained by modifying the base map:

(M-V2V) vector-to-vector (V2V): $\Psi^{(\mathrm{FNM})}$ in (5.8) as written, thus mapping finite vector inputs to finite vector outputs;

(M-V2F) vector-to-function (V2F): $\Psi^{(\mathrm{FNM})}$ with operator $\mathscr{G}$ in (5.8) replaced by $\mathcal{R} \circ \mathscr{L}_T$, where $\mathcal{R}$ and $\mathscr{L}_T$ are as in (5.4) and (5.2), respectively, and Q in (5.8) now viewed as a pointwise-defined operator acting on vector-valued functions;

(M-F2V) function-to-vector (F2V): $\Psi^{(\mathrm{FNM})}$ with operator $\mathscr{D}$ in (5.8) replaced by $\mathscr{L}_1 \circ \mathcal{E}$, where $\mathscr{L}_1$ and $\mathcal{E}$ are as in (5.2) and (5.4), respectively, and S in (5.8) now viewed as a pointwise-defined operator acting on vector-valued functions;

(M-F2F) function-to-function (F2F): $\Psi^{(\mathrm{FNM})}$ with modifications (M-V2F) and (M-F2V), thus the resulting architecture is the standard FNO $\Psi^{(\mathrm{FNO})}$ (5.4).[1]

When the (M-F2V) FNM is of primary interest, we sometimes call this architecture *Fourier Neural Functionals*. Similarly, we may also call the (M-V2F) FNM a *Fourier Neural Decoder*.

5.3 Universal Approximation Theory for Fourier Neural Mappings

In this section, we establish universal approximation theorems for FNMs; this is a confirmation that the architectures maintain this desirable property of neural operators. The results are stated for the cases of the F2V and V2F architectures; the case of V2V trivially follows. Similar results also hold for general neural mappings by invoking the appropriate universal approximation theorems for general neural operators from [152, Section 9.3] and for the topology induced by Lebesgue–Bochner norms, i.e., average error with respect to a probability measure supported on the input space. For more details regarding these extensions, see [154, Theorems 11–14, Section 9.3, pp. 55–57] and [152, pp. 12–14 and Theorem 18]. Our proofs, which are collected in Appendix D.2, use arguments based on constant functions that are similar to those used to prove universal approximation theorems at the level of operators.

The approximation theory in this section relies on the following assumption.

Assumption 5.2 (activation function)**.** *All nonlinear layers* $\{\mathscr{L}_t\}_{t=1}^T$ *from* (5.2) *have the same non-polynomial and globally Lipschitz activation function* $\sigma \in C^\infty(\mathbb{R};\mathbb{R})$.

We note that in practice, the final Fourier layer activation function is often set to be the identity. Moreover, the bias functions b_t in $\mathscr{L}_t$ are typically chosen to be constant functions. The universal approximation theory does not distinguish these differences. Additionally, to align with the existing theory developed in [152], our existence proofs rely on a reduction to the setting that

(i) the channel dimension d_t is constant across all layers, say $d_t = d_v \in \mathbb{N}$ for all $t \in [T]$, and

(ii) the maps S and Q in (5.8) are linear and act pointwise on functions.

[1] Notice that yet another function-to-function FNM architecture is possible by exchanging the roles of $\mathscr{G}$ and $\mathscr{D}$ in (5.8); this is a nonlinear Fourier neural autoencoder.

These conditions are certainly special cases of nonconstant channel dimension and nonlinear lifting and projection maps, respectively. Hence, the forthcoming universality properties still hold for more sophisticated architectures that deviate from conditions (i) and (ii), such as those used in Section 5.5 in this chapter.

Our first result delivers a universal approximation result for Fourier Neural Functionals, i.e., the F2V setting. Appendix D.2.2 contains the proof.

Theorem 5.3 (universal approximation: function-to-vector mappings)**.** *Let $s \geq 0$, $\mathcal{D} \subset \mathbb{R}^d$ be an open Lipschitz domain such that $\overline{\mathcal{D}} \subset (0,1)^d$, and $\mathcal{U} = H^s(\mathcal{D}; \mathbb{R}^{d_u})$. Let $\Psi^\dagger \colon \mathcal{U} \to \mathbb{R}^{d_y}$ be a continuous mapping. Let $K \subset \mathcal{U}$ be compact in $\mathcal{U}$. Under Assumption 5.2, for any $\varepsilon > 0$, there exist Fourier Neural Functionals $\Psi \colon \mathcal{U} \to \mathbb{R}^{d_y}$ of the form (5.8) with modification (M-F2V) such that*

$$\sup_{u \in K} \left\| \Psi^\dagger(u) - \Psi(u) \right\|_{\mathbb{R}^{d_y}} < \varepsilon \,. \tag{5.9}$$

The approximation theorem for the Fourier Neural Decoder, i.e., the V2F case, is analogous.

Theorem 5.4 (universal approximation: vector-to-function mappings)**.** *Let $t \geq 0$, $\mathcal{D} \subset \mathbb{R}^d$ be an open Lipschitz domain such that $\overline{\mathcal{D}} \subset (0,1)^d$, and $\mathcal{Y} = H^t(\mathcal{D}; \mathbb{R}^{d_y})$. Let $\Psi^\dagger \colon \mathbb{R}^{d_u} \to \mathcal{Y}$ be a continuous mapping. Let $\mathcal{Z} \subset \mathbb{R}^{d_u}$ be compact. Under Assumption 5.2, for any $\varepsilon > 0$, there exists a Fourier Neural Decoder $\Psi \colon \mathbb{R}^{d_u} \to \mathcal{Y}$ of the form (5.8) with modification (M-V2F) such that*

$$\sup_{z \in \mathcal{Z}} \left\| \Psi^\dagger(z) - \Psi(z) \right\|_{\mathcal{Y}} < \varepsilon \,. \tag{5.10}$$

The proof may also be found in Appendix D.2.2. While perhaps not surprising, the results in Theorems 5.3 and 5.4 nonetheless show that the proposed FNM architectures are sensible for the tasks of approximating continuous function-to-vector or vector-to-function mappings.

5.4 Statistical Theory for Regression of Linear Functionals

The previous two sections propose and justify a general nonlinear framework for approximating PtO maps with finite-dimensional input or output spaces. Although universality properties of the proposed architectures are established,

the efficiency of statistically estimating the underlying PtO map from a finite dataset remains to be addressed. Such sample complexity results are crucial to understand the expected performance of learning algorithms in scenarios where experimental or computational resources for data generation are limited. However, it is an open challenge to develop such a theory in the general *nonlinear* setting previously considered.

To still shed some light on this issue, we provide a detailed theoretical analysis of learning a *linear* PtO map f that admits a factorization into a linear functional q, the QoI, composed with a self-adjoint linear operator L, the forward map. To set the stage, let $(H, \langle\cdot,\cdot\rangle, \|\cdot\|)$ be an infinite-dimensional real separable Hilbert space. We view the PtO map $f = q \circ L \colon H \to \mathbb{R}$ as a linear functional on H. Hence, the *input space is always infinite-dimensional* in this section. Let ν be the data-generating Borel probability measure on the input space H. Here and in the sequel, suppose that

$$\mathbb{E}^{u\sim\nu}\, u = 0 \quad \text{and} \quad \Sigma \coloneqq \mathrm{Cov}(\nu) = \sum_{j=1}^{\infty} \sigma_j \varphi_j \otimes \varphi_j \tag{5.11}$$

for some orthonormal basis $\{\varphi_j\}_{j\in\mathbb{N}}$ of H and eigenvalue sequence $\{\sigma_j\}_{j\in\mathbb{N}} \subset \mathbb{R}_{\geq 0}$, and that N i.i.d. input data samples $\{u_n\}_{n=1}^N \sim \nu^{\otimes N}$ are available. We consider two different supervised learning approaches (visualized in Figure 5.1 as F2V and F2F) that correspond to the given labeled output data $\{y_n\}_{n=1}^N$ being either noise-perturbed versions of

(EE) (end-to-end learning) *the entire PtO map f applied to the input data*, or

(FF) (full-field learning) *the forward map L applied to the input data. Moreover, in this latter case the linear functional q is assumed to be known.*

In both cases, the primary goal is to estimate f given certain input-output data pairs. Notice that in the (EE) approach, the responses are scalar-valued, while for the (FF) approach the responses are function-valued.[2] In some cases, the problem itself specifies the approach that can be used. For example, in physical experiments, it is often impossible to experimentally acquire the full output of L. Instead, only a finite number of possibly indirect measurements are

[2]Although the Hilbert space H is general and not necessarily comprised of functions, we still refer to its elements as "functions" to avoid confusion caused by attempts to distinguish between finite-dimensional Euclidean vectors and general infinite-dimensional vectors.

available, which represent the QoI q in the (EE) framework. Even if the data generation procedure is algorithmic, the cost of generating full-field data may be much higher than simply measuring finite-dimensional QoIs. Nevertheless, for a large class of forward maps L and *continuous* QoIs q, we demonstrate that (FF) is more data-efficient than (EE) in this linear setup; see insights (I1) and (I2) and Figure 5.2 for more details.

To this end, Subsection 5.4.1 sets up the framework for statistical error analysis of learning general linear functionals on H with a particular Bayesian posterior estimator; this is the end-to-end setting. Subsection 5.4.2 then continues the setup for the setting of factorized linear functionals in the full-field learning setting. Subsection 5.4.3 contains the main results. Here, two theorems provide convergence rates for the end-to-end and full-field settings, respectively. In particular, the end-to-end result may be of independent interest to the functional data analysis and Bayesian nonparametric statistics communities. The theorems are followed by a discussion and a corollary that rigorously compares the data efficiency of end-to-end versus full-field learning of factorized linear functionals.

5.4.1 End-to-End Learning

Consider a general linear functional $f\colon H \to \mathbb{R}$. Working in a nonparametric functional regression framework, we adopt a linear Gaussian posterior mean estimator that is obtained by conditioning a Gaussian process prior on the training dataset under the access model (EE). Our primary concern is the development of large sample convergence rates for the average squared prediction error of the estimator with respect to some test distribution ν'. We allow ν' to be different from the input training distribution ν; that is, our error bounds hold *out-of-distribution* or under *covariate shift*. We describe our statistical model and Bayesian inference approach in Subsection 5.4.1.1. In Subsection 5.4.1.2, we list and interpret the main assumptions that underlie the theory.

5.4.1.1 Setup and Estimator

We adopt a Bayesian inverse problems perspective on the linear functional regression task. To simplify the analysis, suppose that $f \in H^*$ is continuous

on H.[3] As a mild abuse of notation, we use the same symbol f for both the linear functional itself and its Riesz representer. We thus view f as an element of H in all that follows.

Next, for noise level $\gamma > 0$ and for $n \in [N]$, we consider the statistical model

$$y_n = \langle f, u_n \rangle + \gamma \xi_n \,, \quad \text{where} \quad u_n \overset{\text{i.i.d.}}{\sim} \nu \quad \text{and} \quad \xi_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \,. \tag{5.12}$$

We get to observe y_n and u_n, but not the noise ξ_n. Concatenate the data and noise as $U = \{u_n\}_{n=1}^N$, $Y = \{y_n\}_{n=1}^N$, and $\Xi = \{\xi_n\}_{n=1}^N$. This allows (5.12) to be recast as the linear inverse problem of finding f from inputs U and outputs

$$Y = S_N f + \gamma \Xi \,, \tag{5.13}$$

where $S_N \colon H \to \mathbb{R}^N$ is the (random) sampling operator $h \mapsto \{\langle h, u_n \rangle\}_{n=1}^N$.

Proceeding with the Bayesian approach, we endow f with a Gaussian prior distribution $f \sim \mathcal{N}(0, \Lambda)$. Here $\Lambda \in \mathcal{L}(H)$ is a trace-class covariance operator on H. The advantage of working with the preceding linear Gaussian model is that the posterior distribution for f is also a Gaussian measure supported on H with closed form expressions for its mean and covariance. To this end, let

$$\widehat{\Sigma} \coloneqq \frac{S_N^* S_N}{N} = \frac{1}{N} \sum_{n=1}^N u_n \otimes u_n \tag{5.14}$$

denote the empirical covariance operator of the input distribution ν. Additionally, suppose that U, Ξ, and f are independent as random variables. Although the randomness of the operator S_N is a slight deviation from the usual Bayesian setting, application of [147, Proposition 3.1, pp. 2630–2631] and [74, Theorems 32, 13, and 37] still imply that the posterior distribution obtained by conditioning f on the training dataset (U, Y) is given by

$$f \,|\, (U, Y) \sim \mathcal{N}(\bar{f}^{(N)}, \Lambda^{(N)}), \quad \text{where} \quad \bar{f}^{(N)} = A_N Y \tag{5.15}$$

and the operator $A_N \colon \mathbb{R}^N \to H$ and posterior covariance $\Lambda^{(N)} \in \mathcal{L}(H)$ satisfy

$$A_N \coloneqq \gamma^{-2} \Lambda^{(N)} S_N^* \quad \text{and} \quad \Lambda^{(N)} = \frac{\gamma^2}{N} \Lambda^{1/2} \left(\Lambda^{1/2} \widehat{\Sigma} \Lambda^{1/2} + \frac{\gamma^2}{N} \operatorname{Id}_H \right)^{-1} \Lambda^{1/2} \,. \tag{5.16}$$

[3] It is possible to handle *unbounded* linear functionals f using the weighted Hilbert–Schmidt approach from [76, Section 2.2] or the framework of measurable linear functionals from [147, Section 3 and 5], possibly at the expense of stronger assumptions on the data and prior covariances.

We take a frequentist consistency perspective by assuming there exists a fixed ground truth $f^\dagger \in H$ that generates the data Y in (5.13). Abusing notation by using the same symbol for Y, this means that we observe the output response data[4]

$$Y = S_N f^\dagger + \gamma \Xi \,. \tag{5.17}$$

Our estimator of $f^\dagger$ is then taken to be the posterior mean $\bar{f}^{(N)} = A_N Y$ from (5.15) with Y as in (5.17). However, this particular estimator is chosen to simplify the exposition. As explained in Remark 5.14, the analysis remains valid if the full posterior distribution $\mathcal{N}(\bar{f}^{(N)}, \Lambda^{(N)})$ from (5.15) is used to estimate $f^\dagger$ instead of just its mean. Posterior contraction rates also follow as a consequence.

Instead of measuring the quality of our estimate of $f^\dagger$ in the H-norm, we consider a weaker weighted norm induced by the average squared prediction error of the estimator. This is more common in statistical learning than in statistical inverse problems. To this end, let ν' be a centered Borel probability measure supported on a sufficiently large space containing H. We are interested in the out-of-distribution test squared error of $\bar{f}^{(N)}$ with respect to ν', which is given by

$$\mathbb{E}^{u' \sim \nu'} \left| \langle f^\dagger, u' \rangle - \langle \bar{f}^{(N)}, u' \rangle \right|^2 = \left\| \mathrm{Cov}(\nu')^{1/2} (f^\dagger - \bar{f}^{(N)}) \right\|^2 . \tag{5.18}$$

The derivation of this identity is explained in Appendix D.3.2. Equation (5.18) is equivalent to the squared $L^2_{\nu'}(H; \mathbb{R})$ Bochner norm error between the functionals.

Finally, we explain how our posterior mean estimator relates to standard regularized least squares minimizers from machine learning.

Remark 5.5 (equivalence to regularized empirical risk minimization)**.** *Under the setting described in this section, the posterior mean is equivalent to a generalized-Tikhonov regularized least squares estimator. See [76, Section 2.3] and [147, pp. 3631–2632] for more details and relations to RKHS methods. This connects the Bayesian approach taken here back to traditional supervised learning frameworks.*

[4]The actual noise process Ξ in the observed data (5.17) need not match the assumed Gaussian likelihood model implied by (5.12). Indeed, the main results in Subsection 5.4.3 remain valid for any centered square integrable random vector with isotropic covariance.

5.4.1.2 Assumptions

We work under three primary assumptions that involve the data, the prior, and the truth. The first assumption concerns the covariance operators that characterize the end-to-end learning framework (EE).

Assumption 5.6 (end-to-end learning: data and prior)**.** *Instate the setup and hypotheses developed in Subsection 5.4.1.1. The following hold true.*

(A1) (simultaneous diagonalization) *The covariance operator* $\Sigma = \mathrm{Cov}(\nu)$ *of input training data distribution* ν *satisfies* (5.11)*. The covariance operators* $\Sigma' := \mathrm{Cov}(\nu')$ *of the input test data distribution* ν' *and* Λ *of the prior are diagonalized in the eigenbasis* $\{\varphi_j\}_{j\in\mathbb{N}}$ *of* Σ *and have the representations*

$$\Sigma' = \sum_{j=1}^{\infty} \sigma'_j \varphi_j \otimes \varphi_j \quad \textit{and} \quad \Lambda = \sum_{j=1}^{\infty} \lambda_j \varphi_j \otimes \varphi_j \,. \tag{5.19}$$

(A2) (data decay) *For some* $\alpha > \frac{1}{2}$ *and* $\alpha' \geq 0$*, the eigenvalues of* Σ *and* Σ' *satisfy*

$$\sigma_j \asymp j^{-2\alpha} \quad \textit{as} \quad j \to \infty \quad \textit{and} \quad \sigma'_j \lesssim j^{-2\alpha'} \quad \textit{as} \quad j \to \infty \,. \tag{5.20}$$

(A3) (prior decay) *For some* $p > \frac{1}{2}$*, the eigenvalues of* Λ *satisfy*

$$\lambda_j \asymp j^{-2p} \quad \textit{as} \quad j \to \infty \,. \tag{5.21}$$

Although in traditional linear inverse problems the simultaneous diagonalizability (A1) of the normal operator and the prior covariance is considered a strong assumption [8, 230], in the linear functional regression setting here we interpret this condition more mildly. Indeed, Σ and Λ only commute in the infinite data limit. The actual normal operator corresponding to the inverse problem (5.17) is $S_N^* S_N / N = \widehat{\Sigma}$ which *does not* commute with Λ in general for any finite N. Moreover, it is often the case—especially for operator learning-based surrogate models—that the data generation procedure is controlled by the practitioner. In this case, it is simple to choose the data and prior covariance operators to have the same eigenfunctions. For example, Matérn-like Gaussian measures with different regularity exponents and lengthscales are a common choice for the data measure in operator learning and for the prior measure in Bayesian inverse problems; it is natural to assume their covariances share the same

eigenbasis. However, the assumption that Σ' and Λ have the same eigenbasis is stronger because the test distribution ν' may be outside of the user's control and differ substantially from the training distribution in some applications. Going beyond this assumption is an important future direction. Finally, the power law eigenvalue decay conditions (A2) and (A3) are standard in learning theory and help to facilitate explicit convergence rates.

The previous assumption provides fine-grained control of the second moments of the data and prior distributions. Next, we impose a slightly strong assumption about the tails of the Karhunen–Loève (KL) expansion of the training data distribution ν (5.11) in order to obtain high probability error bounds.

Assumption 5.7 (strongly-subgaussian training data)**.** *The input training data distribution ν is a centered Borel probability measure on H with KL expansion*

$$u = \sum_{j=1}^{\infty} \sqrt{\sigma_j} z_j \varphi_j \sim \nu \,, \tag{5.22}$$

where the eigenvalues $\{\sigma_j\}_{j\in\mathbb{N}}$ of $\operatorname{Cov}(\nu) = \Sigma$ are nonincreasing and the $\{z_j\}_{j\in\mathbb{N}}$ are zero mean, unit variance, independent random variables that satisfy[5]

$$m \coloneqq \sup_{j\in\mathbb{N}} \|z_j\|_{\psi_2} = \sup_{j\in\mathbb{N}} \left(\sup_{\ell\geq 1} \ell^{-1/2} \big(\mathbb{E}|z_j|^{\ell}\big)^{1/\ell} \right) < \infty \,. \tag{5.23}$$

Equation (5.23) in Assumption 5.7 implies that the training data distribution ν (5.11) is subgaussian in a relatively strong sense. This enables the use of exponential concentration inequalities in the proofs of forthcoming results. We further insist that the KL expansion coefficients $\{z_j\}_{j\in\mathbb{N}}$ (5.22) are an independent family in order to align with a similar assumption made for full-field learning in the next subsection. However, the following remark explains how this undesirable independence condition can be eliminated for end-to-end learning at the expense of worse tail bounds.

Remark 5.8 (independence of the KL coefficients)**.** *The requirement of independence of the KL expansion coefficients $\{z_j\}_{j\in\mathbb{N}}$ (5.22), while commonly found in the literature [23, 131], is undesirable because it excludes many interesting statistical models. This condition is only used in Lemma D.25, which shows*

[5] See Appendix D.3.1 for the definition of subgaussian random variables and their norms $\|\cdot\|_{\psi_2}$.

that a particular event has high probability. This lemma requires a strong notion of subgaussianity (D.56)*; the assumed independence of the KL coefficients suffices to satisfy this condition. However, we show in Lemma D.26 that the requirement of independence can be replaced by the condition that the* $\{z_j\}_{j\in\mathbb{N}}$ *are pairwise uncorrelated. At the cost of a lower probability for the event of interest, this improvement extends the applicability of our end-to-end learning theory. For example, using the kernel trick, the main theoretical results of this chapter imply convergence rates for Gaussian process regression of scalar-valued* nonlinear *functions as a special case.*

Last, we require a regularity assumption on the true linear functional $f^\dagger$ in order to derive convergence rates.

Assumption 5.9 (regularity of ground truth linear functional)**.** *For each* $j \in \mathbb{N}$*, denote by* $f_j^\dagger = \langle f^\dagger, \varphi_j \rangle$ *the coefficients of* $f^\dagger$*. For some* $s \geq 0$*, it holds that*

$$\left\|f^\dagger\right\|_{\mathcal{H}^s}^2 := \sum_{j=1}^{\infty} j^{2s} |f_j^\dagger|^2 < \infty\,. \tag{5.24}$$

Additionally, $\alpha + s > 1$*, where* α *is as in* (5.20)*.*

Notice that Assumption 5.9 *implies* that $f^\dagger \in H^*$ is a continuous linear functional on H because $s \geq 0$ and hence the coefficients of $f^\dagger$ belong to $\ell^2(\mathbb{N};\mathbb{R})$. The regularity constraint linking α with s is a technical condition that ensures a certain event has vanishing probability in the large sample limit (see Lemma D.20); it may be possible to weaken this constraint with alternative proof techniques. Regardless, this condition is not hard to satisfy.

5.4.2 Full-Field Learning

In this subsection, we provide the additional assumptions and framework required for the setting in which the true linear PtO map $f^\dagger$ is factorized as $f^\dagger = q^\dagger \circ L^\dagger$, where the QoI $q^\dagger$ is a linear functional on H and $L^\dagger$ is a self-adjoint linear operator on H. We allow $q^\dagger$ or $L^\dagger$ to potentially be unbounded with respect to the topology of H. With the full-field learning data access model (FF), we fully observe noisy versions of the *function-valued* output of $L^\dagger$ at the training input functions. We adopt a Bayesian posterior mean estimator $\bar{L}^{(N)}$ for $L^\dagger$ based on these data. The final estimator of the true

linear functional $f^\dagger$ is obtained by composing $q^\dagger$ with the learned operator $\bar{L}^{(N)}$. Subsection 5.4.2.1 contains the setup, while Subsection 5.4.2.2 contains the assumptions we invoke and gives examples of some QoIs that satisfy these assumptions.

5.4.2.1 Setup and Estimator

Our full-field training data observations are given by

$$Y_n = L^\dagger u_n + \eta_n\,, \quad \text{where} \quad u_n \overset{\text{i.i.d.}}{\sim} \nu \quad \text{and} \quad \eta_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathrm{Id}_H) \tag{5.25}$$

for $n \in [N]$. Equation (5.25) should be interpreted in a weak sense, i.e., as H-indexed stochastic processes, because $\eta_n \notin H$ almost surely [76, Section 2.2.2., p. 11]. Without loss of generality, we assume a unit noise level because this does not affect the asymptotic results. To make the analysis tractable, we work in the setting that $L^\dagger$ is diagonalized in the eigenbasis $\{\varphi_j\}_{j\in\mathbb{N}}$ of Σ from (5.11). Thus, we write

$$L^\dagger = \sum_{j=1}^{\infty} l_j^\dagger \varphi_j \otimes \varphi_j \tag{5.26}$$

and develop an estimator for the eigenvalue sequence $l^\dagger = \{l_j^\dagger\}_{j\in\mathbb{N}}$. Details about the domain of $L^\dagger$ and the topology in which (5.26) converges may be found in the previous chapter [76].

Our Bayesian approach follows [76] by modeling the eigenvalue sequence $l^\dagger$ with an independent Gaussian prior $l_j \sim \mathcal{N}(0, \mu_j)$ on each eigenvalue. Write $\Upsilon = \{Y_n\}_{n=1}^N$ and $U = \{u_n\}_{n=1}^N$ and assume that $l = \{l_j\}_{n\in\mathbb{N}}$, Υ, and U are independent. Then [76, Fact 2.4, p. 12] furnishes the posterior distribution

$$l \,|\, (U, \Upsilon) \sim \bigotimes_{j=1}^{\infty} \mathcal{N}\left(\bar{l}_j^{(N)}, c_j^{(N)}\right). \tag{5.27}$$

In (5.27), $\{\bar{l}_j^{(N)}\}_{j\in\mathbb{N}}$ are the posterior mean eigenvalues and $\{c_j^{(N)}\}_{j\in\mathbb{N}}$ are the posterior variances. The plug-in estimator for $f^\dagger = q^\dagger \circ L^\dagger$ is then given by

$$q^\dagger \circ \bar{L}^{(N)}\,, \quad \text{where} \quad \bar{L}^{(N)} \coloneqq \sum_{j=1}^{\infty} \bar{l}_j^{(N)} \varphi_j \otimes \varphi_j\,. \tag{5.28}$$

The precise formulas for the mean and variance in (5.27) are given in [76, Equation (2.4), p. 12]. As in Subsection 5.4.1, we are interested in the out-of-distribution test squared error of the estimator $q^\dagger \circ \bar{L}^{(N)}$ with respect to input test measure ν'.

5.4.2.2 Assumptions

We now collect the main assumptions for the full-field learning approach; these are primarily drawn from the previous chapter [76, Assumption 3.1, pp. 14–15].

Assumption 5.10 (full-field learning: main assumptions)**.** *Instate the setup and hypotheses of Subsection 5.4.2.1. The following hold true.*

(A-I) *The input distributions ν and ν' and their covariance operators Σ and Σ' satisfy Assumptions (A1) and (A2).*[6] *Moreover, ν satisfies Assumption 5.7.*

(A-II) *The eigenvalues $l^\dagger$ of the operator $L^\dagger$ in (5.26) satisfy $l^\dagger \in \mathcal{H}^\beta$ for some $\beta \in \mathbb{R}$.*

(A-III) *The prior $l_j \sim \mathcal{N}(0, \mu_j)$ has variances satisfying $\mu_j \asymp j^{-2\beta-1}$ as $j \to \infty$.*

(A-IV) *The QoI $q^\dagger$ satisfies $|q^\dagger(\varphi_j)|^2 \lesssim j^{-2r-1}$ as $j \to \infty$ for some $r \in \mathbb{R}$ such that $\min(\alpha, \alpha' + r + 1/2) + \beta > 0$, where α and α' are as in (A-I).*

(A-V) *The PtO map $q^\dagger \circ L^\dagger$ is continuous, i.e., $\sum_{j=1}^\infty |q^\dagger(\varphi_j)|^2 |l_j^\dagger|^2 < \infty$.*

The conditions in Assumption 5.10 have similar interpretations to those in Subsection 5.4.1.2 and [76, Assumption 3.1, pp. 14–15]. For simplicity, in (A-III) we have already chosen the optimal prior smoothness exponent $\beta + 1/2$. The condition $\alpha' + r + 1/2 + \beta > 0$ in (A-IV) ensures that the PtO map $q^\dagger \circ L^\dagger$ has finite $L^2_{\nu'}(H; \mathbb{R})$ Bochner norm; thus, the test error is well-defined. The continuity of the PtO map enforced by (A-V) aligns with the (EE) setting. The power law decay of the coefficients of $q^\dagger$ in (A-IV) allows for a sharp convergence analysis. Several common, concrete linear QoIs satisfy this condition, as the next remark demonstrates.

Remark 5.11 (examples of linear QoIs)**.** *Several simple QoIs $q^\dagger$ satisfy the power law decay in (A-IV). Let $H = L^2((0,1); \mathbb{R})$, which has orthonormal basis $x \mapsto \varphi_j(x) = \sqrt{2}\sin(j\pi x)$ for each $j \in \mathbb{N}$. For convenience, denote $q_j^\dagger \coloneqq q^\dagger(\varphi_j)$.*

[6] Inspection of the proof of [76, Theorem 3.9, pp. 18–19] shows that the high probability upper bound (Equation 3.10) there remains valid if it is assumed that σ_j' is only bounded above asymptotically by $j^{-2\alpha'}$ and not necessarily from below.

- (mean on an interval) *The map* $q^\dagger\colon h \mapsto \int_0^1 h(x)\,dx = \langle \mathbf{1}, h\rangle_{L^2((0,1))}$ *is continuous and has Riesz representer* $\mathbf{1}\colon x \mapsto 1$. *The coefficients of* $q^\dagger$ *satisfy*

$$|q_j^\dagger|^2 = \left|\frac{\sqrt{2}(1-\cos(j\pi))}{j\pi}\right|^2 = \frac{8\mathbb{1}_{\{j \text{ odd}\}}}{j^2\pi^2} \lesssim j^{-2}.$$

Hence, $r = 1/2$ *is a valid decay exponent in Assumption (A-IV).*

- (point evaluation) *The map* $q^\dagger\colon h \mapsto h(x_0)$ *for a fixed* $x_0 \in (0,1)$ *is not continuous on* H. *It holds that*

$$|q_j^\dagger|^2 = 2|\sin(2\pi j x_0)|^2 \lesssim 1$$

and hence $r = -1/2$ *is a valid decay exponent in (A-IV).*

- (point evaluation of derivative) *The map* $q^\dagger\colon h \mapsto (dh/dx)(x_0)$ *for a fixed* $x_0 \in (0,1)$ *is not continuous on* H. *Its coefficients satisfy*

$$|q_j^\dagger|^2 = 8\pi^2 j^2 |\cos(2\pi j x_0)|^2 \lesssim j^2$$

and hence $r = -3/2$ *is a valid decay exponent. This QoI is not smooth.*

5.4.3 Main Results

Building upon the setup from the previous two subsections, this subsection establishes convergence rates for end-to-end learning in Theorem 5.12 and full-field learning in Theorem 5.13. Both results are stated for the expectation of the out-of-distribution test error (5.18) conditioned on the input data U. Intuitively, this averages out the noise in the data. The two theorems are interpreted in Subsection 5.4.3.1. This discussion is followed by Subsection 5.4.3.2, which directly compares the end-to-end and full-field methods in Corollary 5.15.

The first theorem describes convergence rates in the end-to-end learning setting.

Theorem 5.12 (end-to-end learning: optimized convergence rate)**.** *Let the input training data distribution* ν, *the test data distribution* ν', *and the Gaussian prior* $\mathcal{N}(0,\Lambda)$ *satisfy Assumptions 5.6 and 5.7. Let the ground truth linear functional* $f^\dagger \in \mathcal{H}^s$ *satisfy Assumption 5.9 with* $s > 0$. *Let* α *and* α' *be as in* (5.20) *and* $p = s + 1/2$ *be as in* (5.21). *Then there exists* $c \in (0, 1/4)$ *and* $N_0 \geq 1$ *such that for any* $N \geq N_0$, *the mean* $\bar{f}^{(N)}$ *of the Gaussian posterior distribution* (5.15) *arising from the* N *pairs of observed training data* (U, Y) *in* (5.17) *satisfies the error bound*

$$\mathbb{E}^{Y\,|\,U}\,\mathbb{E}^{u'\sim\nu'}\left|\langle f^\dagger, u'\rangle - \langle \bar{f}^{(N)}, u'\rangle\right|^2 \lesssim \left(1 + \|f^\dagger\|_{\mathcal{H}^s}^2\right)\varepsilon_N^2 \tag{5.29}$$

with probability at least $1 - 2\exp(-cN^{\min(1, \frac{\alpha+s-1}{\alpha+s+1/2})})$ *over* $U \sim \nu^{\otimes N}$*, where*

$$\varepsilon_N^2 = \begin{cases} N^{-\left(\frac{2\alpha'+2s}{1+2\alpha+2s}\right)}, & \text{if } \alpha' < \alpha + 1/2, \\ N^{-1}\log 2N, & \text{if } \alpha' = \alpha + 1/2, \\ N^{-1}, & \text{if } \alpha' > \alpha + 1/2. \end{cases} \tag{5.30}$$

The constants c*,* N_0*, and the implied constant in* (5.29) *do not depend on* N *or* $f^\dagger$*.*

Appendix D.1 contains a more general version of the preceding theorem that is valid *for any* $p > 1/2$ (Theorem D.1). The assertion of this version is optimized for the choice $p = s + 1/2$ made in Theorem 5.12. The proof of Theorem D.1, from which Theorem 5.12 follows immediately, may be found in Appendix D.3.2. Appendix D.1 also includes an expectation bound instead of the high probability bound in Theorem 5.12; the consequences are the same.

The second main theorem in this subsection describes the expected squared error in the QoI after learning an approximate forward map from full-field data.

Theorem 5.13 (full-field learning: convergence rate for power law QoI)**.** *Let the input training data distribution* ν*, the test data distribution* ν'*, the true forward map* $L^\dagger$*, and the QoI* $q^\dagger$ *satisfy Assumption 5.10. Let* α *and* α' *be as in* (5.20) *and* β *and* r *be as in (A-II) and (A-IV). Then there exist constants* $c > 0$ *and* $C > 0$ *such that for all sufficiently large* N*, the plug-in estimator* $q^\dagger \circ \bar{L}^{(N)}$ *in* (5.28) *based on the Gaussian posterior distribution* (5.27) *arising from the* N *pairs of observed full-field training data* (U, Υ) *in* (5.25) *satisfies the error bound*

$$\mathbb{E}^{\Upsilon \,|\, U}\, \mathbb{E}^{u' \sim \nu'} \left| q^\dagger(L^\dagger u') - q^\dagger(\bar{L}^{(N)} u') \right|^2 \lesssim \varepsilon_N^2 \tag{5.31}$$

with probability at least $1 - Ce^{-cN}$ *over* $U \sim \nu^{\otimes N}$*, where*

$$\varepsilon_N^2 \coloneqq \begin{cases} N^{-\left(\frac{1+2\alpha'+2\beta+2r}{1+2\alpha+2\beta}\right)}, & \text{if } \alpha' + r < \alpha, \\ N^{-1}\log N, & \text{if } \alpha' + r = \alpha, \\ N^{-1}, & \text{if } \alpha' + r > \alpha. \end{cases} \tag{5.32}$$

The constants c*,* C*, and the implied constant in* (5.31) *do not depend on* N*.*

Appendix D.1 also contains a similar convergence result for QoIs with an assumed Sobolev-like regularity instead of power law regularity. Proofs of both this result and Theorem 5.13 are collected in Appendix D.3.3.

5.4.3.1 Discussion

The proof of Theorem 5.12 is based on a bias–variance decomposition argument that is detailed in Appendix D.3.2. Notice that the error bound in the theorem is uniform over $\mathcal{H}^s$–balls because the implied constant in the inequality (5.29) does not depend on $f^\dagger$. The probability that the bound fails to hold decays to zero faster than any power law as a function of the sample size N because $\alpha + s - 1 > 0$ by hypothesis. The convergence rate (5.29) depends on the three smoothness exponents α, α', and s. The consequences are the same as those identified in [76, Section 3] because the error bound takes the same form. Indeed, rougher training data points $\{u_n\}_{n=1}^N$ (smaller α), smoother test data points $u' \sim \nu'$ (larger α'), and smoother target functionals $f^\dagger$ (larger s) all serve to reduce the test error (up to saturation).

Moreover, the optimal choice of p made in Theorem 5.12 depends on the regularity exponent s of the ground truth $f^\dagger$, which is unknown. However, there exist more sophisticated estimators that adapt to the unknown regularity and achieve the optimal convergence rate [7, 146]. Nonetheless, the fact that we are even able to choose $p = s + 1/2$ in the first place is one of the novelties of our result. Most existing theoretical work on functional linear regression requires some constraint linking the regularity p of the prior to the regularity s of $f^\dagger$. The most common assumption corresponds to the *well-specified* setting [49, 175, 278, 279], which means that $f^\dagger$ belongs to the RKHS $\mathrm{Im}(\Lambda^{1/2}) \subset H$ of the prior $\mathcal{N}(0, \Lambda)$. In terms of coefficients, this is equivalent to assuming that $\sum_{j=1}^\infty \lambda_j^{-1} |f_j^\dagger|^2 \lesssim \|f^\dagger\|_{\mathcal{H}^p}^2 < \infty$ because $\lambda_j \asymp j^{-2p}$. However, we only have that $f^\dagger \in \mathcal{H}^s$. With the optimal choice $p = s + 1/2$, it is possible that $f^\dagger \notin \mathcal{H}^p = \mathcal{H}^{s+1/2}$. Our theory allows for such RKHS misspecification. In the full-field learning setting, similar notions of robustness to misspecification are guaranteed by the error bounds in [76].

The consequences of Theorem 5.13 for full-field learning are similar to those of Theorem 5.12 for end-to-end learning with regard to the smoothness exponents that define the estimation problem. However, Theorem 5.13 is only valid for QoIs $q^\dagger$ with asymptotic power law decay of the form $|q^\dagger(\varphi_j)|^2 \lesssim j^{-2r-1}$ as $j \to \infty$. While many QoIs in practice satisfy this condition, it still corresponds to a relatively small set within the class of all linear functionals. For example, if the asymptotic power law decay condition holds, then $\{q^\dagger(\varphi_j)\}_{j \in \mathbb{N}} \in \mathcal{H}^{r-\varepsilon}$ for every $\varepsilon > 0$. It is natural to wonder whether Theorem 5.13 remains valid

if $q^\dagger$ is only assumed to satisfy such a Sobolev-like regularity condition. To this end, Theorem D.3 in Appendix D.1 generalizes Theorem 5.13 to all of $\mathcal{H}^r$, but at the expense of a worse convergence rate where r in (5.32) is replaced by $r - 1/2$.

To conclude the discussion, we remark on more general estimators based on full posterior distributions.

Remark 5.14 (posterior contraction rates)**.** *First consider the (EE) setting. With minor modifications, the more general Theorem D.1, and hence also Theorem 5.12, remains valid for the* posterior sample estimator $f^{(N)} \sim \mathcal{N}(\bar{f}^{(N)}, \Lambda^{(N)})$ *instead of its mean. To see this, note that the KL expansion of the posterior* (5.15) *yields*

$$\mathbb{E}^{f^{(N)}\sim\mathcal{N}(\bar{f}^{(N)},\Lambda^{(N)})}\big\|(\Sigma')^{1/2}(f^\dagger - f^{(N)})\big\|^2 = \big\|(\Sigma')^{1/2}(f^\dagger - \bar{f}^{(N)})\big\|^2 + \operatorname{tr}\big((\Sigma')^{1/2}\Lambda^{(N)}(\Sigma')^{1/2}\big)\,. \tag{5.33}$$

Theorem D.1 bounds the conditional expectation of the first term on the right-hand side of the preceding equality. We see that the only new error term that the full posterior introduces is the second term, the posterior spread*. But the end of Subsection D.3.2.1 explains that the posterior spread may be upper bounded by a constant times the rate ε_N^2 from Theorem D.1. Thus, the end-to-end error bounds* (D.1) *and* (5.29) *remain valid for the posterior sample estimator $f^{(N)}$ at the expense of enlarged constant factors. Posterior contraction rates then follow from a standard Chebyshev inequality argument [76, Section 3.3, p. 18]. Similar results may be deduced for the full-field setting (FF) because the error analysis for the forward map in [76, Section 3.4] already takes into account the full posterior distribution* (5.27)*.*

5.4.3.2 Sample Complexity Comparison

To conclude Subsection 5.4.3, we provide a detailed comparison of the end-to-end (EE) and full-field (FF) PtO map learning approaches to provide intuition about their statistical performance. Focusing on the specific setting of QoIs with power law coefficient decay and *in-distribution* test error, the following corollary is a consequence of Theorem 5.12 and Theorem 5.13. A short proof is provided in Appendix D.3.4.

Corollary 5.15 (sample complexity comparison)**.** *Instate the notation and assertions in Assumptions 5.6, 5.7, and (A-III). Suppose that the training and test distribution covariances have equivalent smoothness, i.e., $\alpha' = \alpha$. Let the underlying true PtO map $f^\dagger$ have the factorization $f^\dagger = q^\dagger \circ L^\dagger$, where $|q^\dagger(\varphi_j)|^2 \lesssim j^{-2r-1}$ as $j \to \infty$ and $L^\dagger$ is as in (5.26) with eigenvalues $l^\dagger \in \mathcal{H}^\beta$. If $\beta + r + 1/2 > 0$, $\alpha + \beta + r > 1/2$, and $\alpha + \beta > 0$, then there exist constants $c > 0$ and $C > 0$ such that for all sufficiently large N, the following holds on an event with probability at least $1 - C\exp(-cN^{\min(1, \frac{\alpha+\beta+r-1/2}{1+\alpha+\beta+r})})$ over $U = \{u_n\}_{n=1}^N \sim \nu^{\otimes N}$. The (EE) posterior mean estimator $\bar{f}^{(N)}$ in (5.15) (with $p \coloneqq \beta + r + 1$ in (A3)) trained on end-to-end data (U, Y) satisfies*

$$\mathbb{E}^{Y\,|\,U}\,\mathbb{E}^{u\sim\nu}\big|q^\dagger(L^\dagger u) - \langle \bar{f}^{(N)}, u\rangle\big|^2 \lesssim N^{-\left(1-\frac{1}{2+2\alpha+2\beta+2r}\right)}\,. \tag{5.34}$$

On the other hand, the (FF) plug-in estimator $q^\dagger \circ \bar{L}^{(N)}$ in (5.28) trained on full-field data (U, Υ) satisfies

$$\mathbb{E}^{\Upsilon\,|\,U}\,\mathbb{E}^{u\sim\nu}\big|q^\dagger(L^\dagger u) - q^\dagger(\bar{L}^{(N)}u)\big|^2 \lesssim \begin{cases} N^{-\left(1-\frac{-2r}{1+2\alpha+2\beta}\right)}\,, & \text{if } r < 0\,, \\ N^{-1}\log N\,, & \text{if } r = 0\,, \\ N^{-1}\,, & \text{if } r > 0\,. \end{cases} \tag{5.35}$$

Several interesting insights may be deduced from the convergence rates (5.34) and (5.35). Since a common set of assumptions have been identified in the statement of Corollary 5.15, these rates may be directly compared to assess whether the (EE) or (FF) approach is more accurate or, equivalently, more data-efficient, than the other (up to the sharpness of the upper bounds). First, we note that the squared generalization error of (EE) in (5.34) has a nonparametric convergence rate $N^{-(1-\delta)}$ that is always slower than the parametric estimation rate N^{-1} by a polynomial factor N^δ (where $\delta > 0$). On the other hand, if $r \geq 0$, then (FF) achieves the fast parametric rate N^{-1} (up to a log factor if $r = 0$); this always beats (EE) in the regime $r \geq 0$. This regime has an interesting interpretation because *the QoI is continuous if* $r > 0$. These types of QoIs appear naturally in scientific applications.

To study the regime $r < 0$, let

$$\rho_{\mathrm{EE}}(r) \coloneqq 1 - \frac{1}{2 + 2\alpha + 2\beta + 2r} \quad \text{and} \quad \rho_{\mathrm{FF}}(r) \coloneqq 1 - \frac{2\max(-r, 0)}{1 + 2\alpha + 2\beta} \tag{5.36}$$

for $r \neq 0$ denote the convergence rate exponents of the end-to-end and full-field estimators, respectively (ignoring the log factor when $r = 0$ in (5.35)). A larger

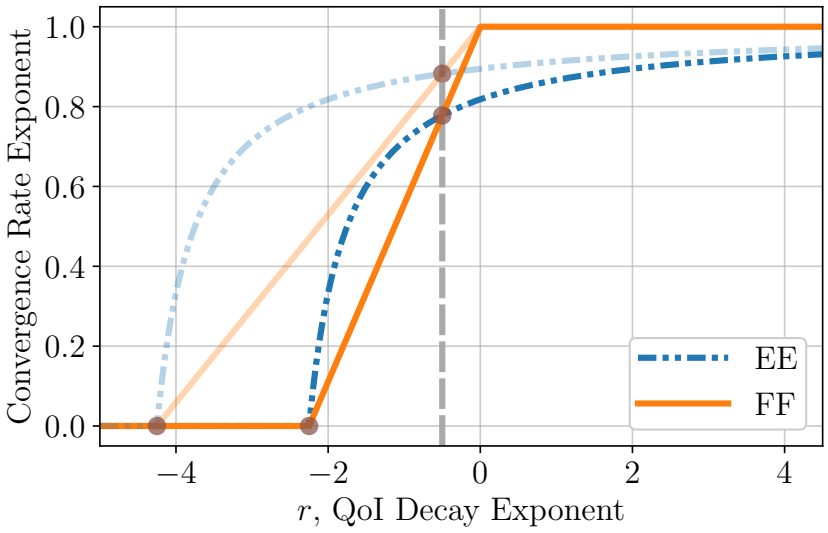


Figure 5.2: (EE) vs. (FF) convergence rate exponents (5.36) as a function of QoI decay exponent r. Larger exponents imply faster convergence rates. As the curves gets lighter, $\alpha + \beta$, an indicator of the smoothness of the problem, increases. The vertical dashed line corresponds to $r = -1/2$, which is the transition point where (EE) and (FF) have the same rate and the onset of power law decay for the QoI coefficients begins.

exponent implies faster convergence and better sample complexity. A simple algebraic factorization shows that $\rho_{\mathrm{EE}}(r) = \rho_{\mathrm{FF}}(r) = \rho$ at the points

$$(r_0, \rho_0) = \left(-\frac{1 + 2\alpha + 2\beta}{2}, 0\right) \quad \text{and} \quad (r_1, \rho_1) = \left(-\frac{1}{2}, \frac{2\alpha + 2\beta}{1 + 2\alpha + 2\beta}\right).$$

By the concavity of $r \mapsto \rho_{\mathrm{EE}}(r)$ in the range $[r_0, r_1]$ and affine structure of $r \mapsto \rho_{\mathrm{FF}}(r)$, we deduce the following two insights:

(I1) ((EE) is better for rough QoIs) $\rho_{\mathrm{EE}}(r) > \rho_{\mathrm{FF}}(r)$ for $r_0 < r < -1/2$ and

(I2) ((FF) is better for smooth QoIs) $\rho_{\mathrm{EE}}(r) < \rho_{\mathrm{FF}}(r)$ for $r > -1/2$.

The inequalities in (I1) and (I2) suggest that the (FF) approach is advantageous when the QoI is smooth and the (EE) approach is advantageous when the QoI is rough. However, definitive conclusions would require lower bounds. An example of a rough QoI is $q^\dagger \colon h \mapsto (dh/dx)(x_0)$, which returns a point evaluation of the first derivative of a univariate function (see the last item in Remark 5.11). Direct application of such a rough QoI to a function is an ill-posed operation (e.g., amplifies perturbations in the function). This may

partially explain why (EE) is preferable in this case, as the (EE) estimator does not require the evaluation of $q^\dagger$ while (FF) does. We plot the functions ρ_{EE} and ρ_{FF} in Figure 5.2, which visualizes the main insights (I1) and (I2) from the preceding discussion.

5.5 Numerical Experiments

We now perform numerical experiments with the proposed FNM architectures.[7] These experiments have two main purposes. The first is to numerically implement and compare the various FNM models on several PtO maps of practical interest; the second is to qualitatively validate the theory developed in the chapter. We focus on nontrivial nonlinear problems with finite-dimensional observables that define the QoI maps. Although our linear theory from Section 5.4 does not apply to such nonlinear problems, we still observe qualitative validation of the main implications of the linear analysis. That is, for smooth enough QoIs, full-field learning is at least as data-efficient as end-to-end learning. Unlike the theory, however, our numerical results distinguish the two approaches only by constant factors and not by the actual convergence rates.

The continuum FNM architectures from Section 5.2 are implemented numerically by replacing all forward and inverse Fourier series calculations with their Discrete Fourier Transform counterparts. This enables fast summation of the series (5.3), (5.6), and (5.7) with the FFT. The inner products in these formulas are also computed with the FFT. In particular, the FFT performs Fourier space operations in the set $\{k \in \mathbb{Z}^d \colon \|k\|_{\ell^\infty([d];\mathbb{Z})} \leq K\}$ rather than over all $k \in \mathbb{Z}^d$.[8] In this case, we say that the FNM architecture has K modes. This is analogous to the mode truncation used in standard FNO layers (see, e.g., [172]). Additionally, since we work with real vector-valued functions, conjugate symmetry of the Fourier coefficients may be exploited to write the Fourier linear functional (5.6) and decoder (5.7) layers only in terms of the real part of the coefficients appearing in the summands. We also make a minor modification to the F2V and V2V FNMs. Since the discrete implementation of $\mathscr{G}$ in (5.6) requires discarding the higher frequencies in the input function, we define an

[7] The datasets are available at doi.org/10.22002/r5ga1-55d06. The code used to produce the numerical results and figures in this chapter is available at

`https://github.com/nickhnelsen/fourier-neural-mappings` .

[8] In all numerical experiments to follow, $d = 1$ or $d = 2$.

auxiliary map $\mathscr{W}\colon h \mapsto \int_{\mathbb{T}^d} \mathsf{NN}(h(x))\,dx$ that makes use of all frequencies. Here $\mathsf{NN}(\,\cdot\,)$ is a one hidden layer fully-connected neural network (NN). Then we replace $\mathscr{G}$ in Definition 5.1 by the concatenated operator $(\mathscr{G}, \mathscr{W})^\top$.

Given a dataset of input-output pairs $\{(u_n, \tilde{y}_n)\}_{n=1}^N$, we train a FNM Ψ_θ taking one of the forms given in Definition 5.1 (with the modifications from the preceding discussion) in a supervised manner by minimizing the average relative error

$$\frac{1}{N}\sum_{n=1}^{N} \frac{\|\tilde{y}_n - \Psi_\theta(u_n)\|}{\|\tilde{y}_n\|} \tag{5.37}$$

or the average absolute squared error

$$\frac{1}{N}\sum_{n=1}^{N} \|\tilde{y}_n - \Psi_\theta(u_n)\|^2 \tag{5.38}$$

over the FNM's tunable parameters θ using mini-batch SGD with the ADAM optimizer. The choice of the loss function is dependent on the underlying problem. Moreover, the norm in the preceding displays are inferred from the space that the $\tilde{y}_n$ takes values in (i.e., finite-dimensional or infinite-dimensional output spaces). To avoid numerical instability in our actual computations, we add 10^{-6} to the denominator of the ratio in (5.37).

The numerical experiments are organized as follows. In Subsection 5.5.1, we extract the first four polynomial moments from the solution of a velocity-parametrized 2D advection–diffusion equation. Next, Subsection 5.5.2 considers the flow over an airfoil modeled by the steady compressible Euler equation. The PtO map sends the shape of the airfoil to the resultant drag and lift force vector. Last, we study an elliptic homogenization problem parametrized by material microstructure in Subsection 5.5.3. Here, the QoI returns the effective tensor of the material.

5.5.1 Moments of an Advection–Diffusion Model

Our first model problem concerns a canonical advection–diffusion PDE in two spatial dimensions. This equation often arises in the environmental sciences and is useful for modeling the spread of passive tracers (e.g., contaminants, pollutants, aerosols), especially when the driving velocity field is coupled to another PDE such as the Navier–Stokes equation. Our setup is as follows. Let $\mathcal{D} = (0,1)^2$ be the spatial domain and n denote the unit inward normal vector to $\mathcal{D}$. For a prescribed time-independent velocity field $v\colon \mathcal{D} \to \mathbb{R}^2$, the state

$\phi\colon \mathcal{D} \times \mathbb{R}_{>0} \to \mathbb{R}$ solves

$$\begin{aligned} \partial_t \phi + \nabla \cdot (v\phi) - 0.05\Delta\phi &= g \quad \text{in} \quad \mathcal{D} \times \mathbb{R}_{>0}\,, \\ \mathsf{n} \cdot \nabla\phi &= 0 \quad \text{on} \quad \partial\mathcal{D} \times \mathbb{R}_{>0}\,, \\ \phi &= 0 \quad \text{on} \quad \mathcal{D} \times \{0\}\,. \end{aligned} \tag{5.39}$$

The time-independent source term g is a smoothed impulse located at $x_0 \coloneqq (0.2, 0.5)^\top$ and is defined for $x \in \mathcal{D}$ by

$$g(x) \coloneqq \frac{5}{2\pi(50)^{-2}} \exp\left(-\frac{\|x - x_0\|_{\mathbb{R}^2}^2}{2(50)^{-2}}\right).$$

We associate our input parameter with the velocity field v appearing in (5.39). Our parametrization takes the form

$$v = (u, 0)^\top\,, \quad \text{where} \quad u(x_1, x_2) = 3 + \sum_{j=1}^{d_{\mathrm{KL}}} \sqrt{\tau_j} z_j e_j(x_1) \tag{5.40}$$

for all $x = (x_1, x_2) \in \mathcal{D}$. Note that u is constant in the vertical x_2 direction. The eigenvalues $\{\tau_j\}_{j\in\mathbb{N}}$ and eigenfunctions $\{e_j\}_{j\in\mathbb{N}}$ correspond to the Mercer decomposition of a kernel obtained by restricting a Matérn covariance function over $\mathbb{R}$ to $(0, 1) \subset \mathbb{R}$. The covariance function has smoothness exponent 1.5 and lengthscale 0.25 [228]. We choose

$$z_j \overset{\text{i.i.d.}}{\sim} \mathsf{Uniform}([-1, 1]) \quad \text{for all} \quad j \in [d_{\mathrm{KL}}]\,.$$

Thus, up to normalization constants, the velocity field (5.40) is the (truncated) KL expansion of a subgaussian stochastic process. We take the input to either be the full x_1-velocity field $u\colon \mathcal{D} \to \mathbb{R}$ or the i.i.d. realizations $z \coloneqq (z_1, \ldots, z_{d_{\mathrm{KL}}})^\top$ of the random variables that affinely parametrize u.

Define the *nonlinear* QoI map $q^\dagger\colon L^4(\mathcal{D}; \mathbb{R}) \to \mathbb{R}^4$ as follows. First, for any $h \in L^2(\mathcal{D}; \mathbb{R})$, let

$$\bar{m}(h) \coloneqq \int_{\mathcal{D}} h(x)\,dx \quad \text{and} \quad \bar{s}(h) \coloneqq \left(\int_{\mathcal{D}} |h(x) - \bar{m}(h)|^2\,dx\right)^{1/2} \tag{5.41}$$

denote the mean and variance of the pushforward of the uniform distribution on $\mathcal{D} = (0, 1)^2$ under h, respectively. Then $q^\dagger = (q_1^\dagger, q_2^\dagger, q_3^\dagger, q_4^\dagger)^\top$ is given by

$$h \mapsto q^\dagger(h) \coloneqq \begin{pmatrix} \bar{m}(h) \\ \bar{s}(h) \\ \bar{s}(h)^{-3} \int_{\mathcal{D}} \big(h(x) - \bar{m}(h)\big)^3\,dx \\ -3 + \bar{s}(h)^{-4} \int_{\mathcal{D}} \big|h(x) - \bar{m}(h)\big|^4\,dx \end{pmatrix}. \tag{5.42}$$

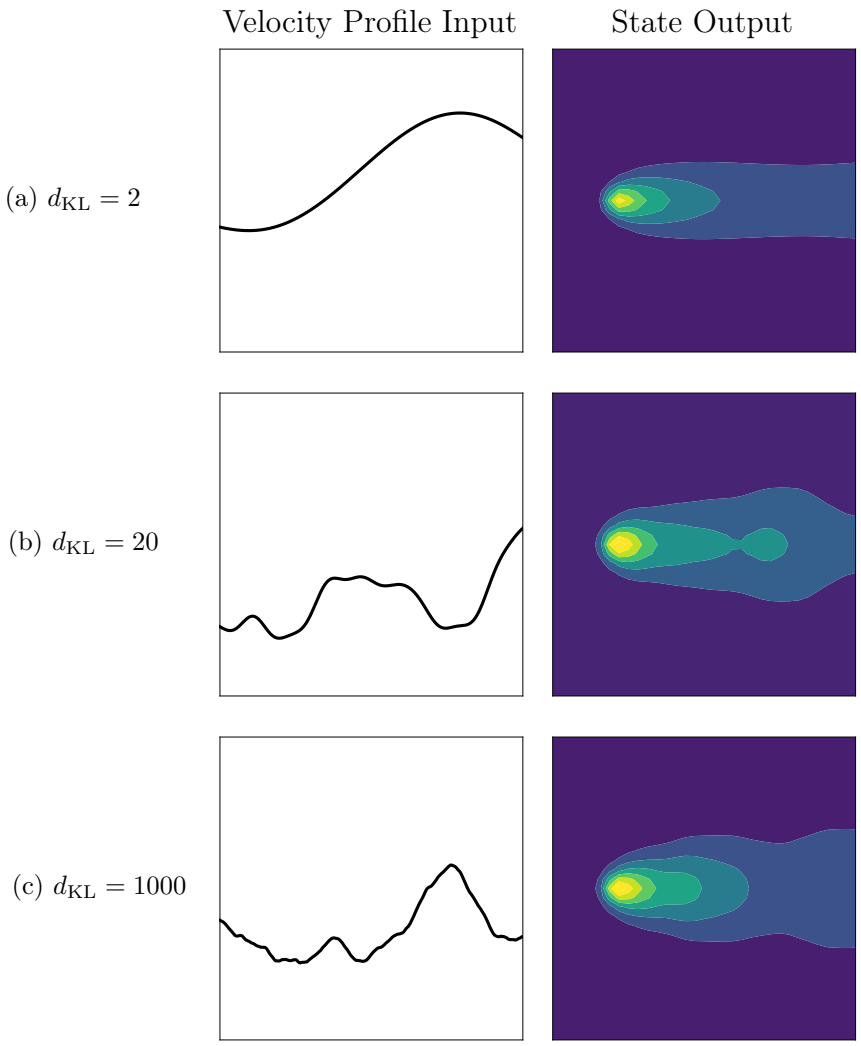


Figure 5.3: Visualization of the velocity-to-state map for the advection–diffusion model. Rows denote the dimension of the KL expansion of the velocity profile and columns display representative input and output fields.

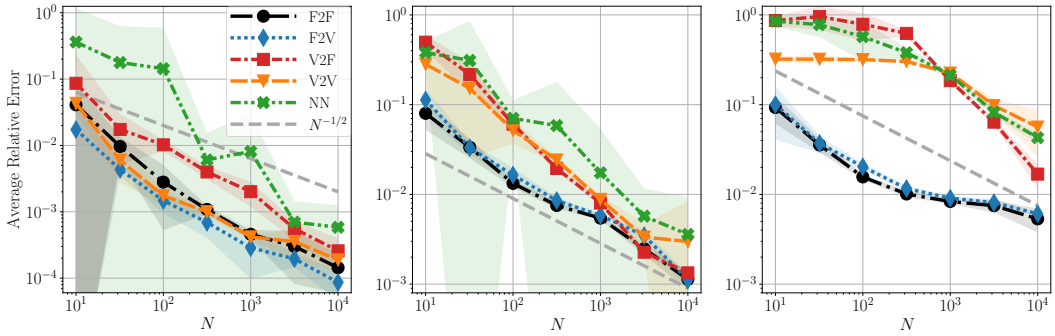
Hence, $q_1^\dagger$ is the mean, $q_2^\dagger$ the standard deviation, $q_3^\dagger$ the skewness, and $q_4^\dagger$ the excess kurtosis. Our goal is to build FNM surrogates for the PtO map that sends the input representation (either the full velocity field or its finite number of i.i.d. coefficients) to the QoI values of the state ϕ at final time $t = 3/4$ (see Figure 5.3). Therefore, we train FNMs to approximate each of the following

ground truth maps:

$$\begin{aligned}
&\Psi^{\dagger}_{\mathrm{F2F}} \colon u \mapsto \phi\big|_{t=3/4}\,, \\
&\Psi^{\dagger}_{\mathrm{F2V}} \colon u \mapsto q^{\dagger}\Big(\phi\big|_{t=3/4}\Big)\,, \\
&\Psi^{\dagger}_{\mathrm{V2F}} \colon z \mapsto \phi\big|_{t=3/4}\,, \quad \text{and} \\
&\Psi^{\dagger}_{\mathrm{V2V}} \colon z \mapsto q^{\dagger}\Big(\phi\big|_{t=3/4}\Big)\,.
\end{aligned}$$

The training data is obtained by solving (5.39) with a second-order Lagrange finite element method on a mesh of size 32×32 and Euler time step 0.01. For each $d_{\mathrm{KL}} \in \{2, 20, 1000\}$, we generate 10^4 i.i.d. data pairs for training, 1500 pairs for computing the test error (which is (5.37) over the 1500 test pairs instead of over the N training pairs), and 500 pairs for validation. All FNM models with 2D spatial input or output functions use 12 modes per dimension and a channel width of 32. For the V2V-FNM, we use a 1D latent function space with 12 modes and channel width of 96. We compare all FNM models to a standard fully-connected NN with three layers and constant hidden width 2048. These architecture settings were selected based on a hyperparameter search over the validation dataset for $d_{\mathrm{KL}} = 1000$ that mimics the parameter complexity experiments in [35, 159]. The models are trained on the relative loss (5.37) for 500 epochs in L^2 output space norm for functions and Euclidean norm for vectors. The optimizer settings include a minibatch size of 20, weight decay of 10^{-4}, and an initial learning rate of 10^{-3} which is halved every 100 epochs. We train 5 i.i.d. realizations of the models for various values of N and d_{KL} and report the results in Figure 5.4.

Figure 5.4 reveals several interesting trends. In general, training models to emulate the advection–diffusion PtO map with finite-dimensional vectors as input is more difficult than adopting function space input variants of the problem. The difficulty is further exacerbated as the dimension of the input vector (here, d_{KL}) increases. We hypothesis that this gap in performance would reduce if the vector input models received the weighted KL coefficients $\{\sqrt{\tau_j} z_j\}$ as input instead of the i.i.d. sequence $\{z_j\}$. This way the model would have access to decay information and hence an ordering of the coefficients. The standard finite-dimensional NN performs poorly across all KL expansion dimensions. The training of the NN is also quite erratic, as evidenced by the large green shaded regions indicating large variance over multiple training runs.



(a) $d_{\mathrm{KL}} = 2$ (b) $d_{\mathrm{KL}} = 20$ (c) $d_{\mathrm{KL}} = 1000$

Figure 5.4: Empirical sample complexity of FNM and NN architectures for the advection–diffusion PtO map (note that Figure 5.4a has a different vertical axis range). The shaded regions denote two standard deviations away from the mean of the test error over 5 realizations of the random training dataset indices, batch indices during SGD, and model parameter initializations.

The output space seems to play less of a role than the input space. Indeed, the F2F and F2V FNMs with function space inputs generally achieve the lowest test error regardless of N and d_{KL}. The full-field F2F method slightly out performs the end-to-end F2V method by a small constant factor (except for when $d_{\mathrm{KL}} = 2$). Since $q^\dagger$ is a smoothing QoI due to its integral definition, this observation aligns with the theoretical insights from Subsection 5.4.3.2. The fast convergence of some of the FNM models, especially for the low-dimensional cases $d_{\mathrm{KL}} = 2$ and $d_{\mathrm{KL}} = 20$, could potentially be explained by the lack of noise in the data, the smoothness of the QoIs, and the nonconvexity of the training procedure. When $d_{\mathrm{KL}} = 1000$, the problem is essentially infinite-dimensional. The function space input FNMs (F2F and F2V) exhibit a nonparametric decay of test error as expected.

5.5.2 Aerodynamic Force Exerted on an Airfoil

Consider the following steady compressible Euler equation applied to an airfoil problem (see Figure 5.5), as introduced in [171]:

$$\begin{aligned} \nabla \cdot (\rho v) &= 0\,, \\ \nabla \cdot (\rho v v^\top + p\,\mathrm{Id}_{\mathbb{R}^2}) &= 0\,, \\ \nabla \cdot \big((E+p)v\big) &= 0\,. \end{aligned} \tag{5.43}$$

Here ρ is the fluid density, v is the velocity vector, p is the pressure, and E is the total energy. Equation (5.43) is equipped with the following far-field boundary conditions: $\rho_\infty = 1$, $p_\infty = 1$, $M_\infty = 0.8$, and $\mathsf{AoA} = 0$, where M_∞ is the Mach

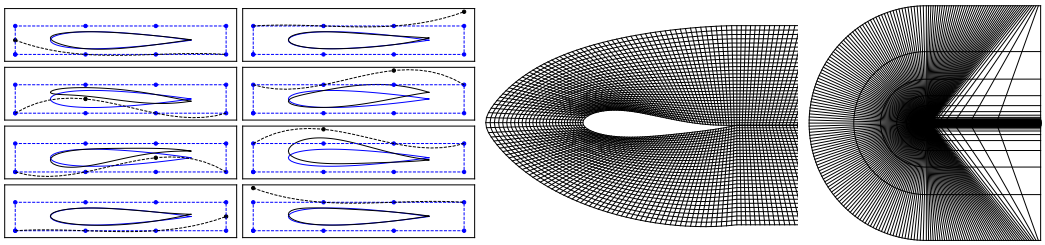


Figure 5.5: Flow over an airfoil. From left to right: visualization of the cubic design element and different airfoil configurations, guided by the displacement field of the control nodes; a close-up view of the C-grid surrounding the airfoil; the physical domain discretized by the C-grid.

number and AoA is the angle of attack. This setup indicates that the flow condition is in the transonic regime. Additionally, the no-penetration condition $v \cdot \mathsf{n} = 0$ is imposed at the airfoil, where n represents the inward-pointing normal vector to the airfoil. Additional mathematical details about the setup may be found in [183, 184, 185, 194].

In this context, we are interested in solving the aforementioned 2D Euler equation to predict the drag and lift performance of different airfoil shapes. Building fast yet accurate surrogates for this task facilities aerodynamic shape optimization [206, 244] for various design goals, such as maximizing the lift to drag ratio [171]. The drag and lift QoIs, which only depend on the pressure on the airfoil, are given by the force vector

$$(\mathsf{Drag},\, \mathsf{Lift})^\top = \oint_{\mathcal{A}} p\mathsf{n}\, ds \in \mathbb{R}^2\,. \tag{5.44}$$

Here $\mathcal{A}$ denotes the closed curve defined by the union of the upper and lower surfaces of the airfoil. Different airfoil shapes are generated following the design element approach [95] (Figure 5.5). The initial NACA-0012 shape is embedded into a "cubic" design element featuring eight control nodes, and the initial shape is morphed to a different one following the displacement field of the control nodes of the design element. The displacements of control nodes are restricted to the vertical direction only. Consequently, the intrinsic dimension of the input is seven, as displacing all nodes in the vertical direction by a constant value does not change the shape of the airfoil.

To generate the training data, we used the traditional second-order finite volume method with the implicit backward Euler time integrator. The process begins by generating a new airfoil shape. Subsequently, a C-grid mesh [249]

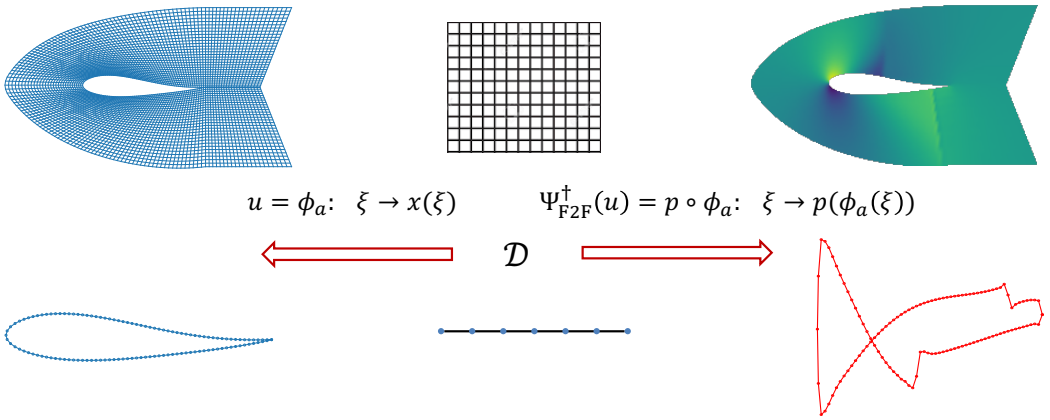


Figure 5.6: Flow over an airfoil. The 1D (bottom) and 2D (top) latent spaces are illustrated at the center; the input functions $u(\,\cdot\,)$ encoding the irregular physical domains, are shown on the left; and the output functions $p \circ u$ representing the pressure field on the irregular physical domains, are depicted on the right.

consisting of 221×51 quadrilateral elements is created around the airfoil with adaptation near the airfoil. In total, we generated 2000 training data and 400 test data with the vertical displacements of each control node being sampled from a uniform distribution $\mathsf{Uniform}([-0.05, 0.05])$.

Next, we will define the operator learning problem (see Figure 5.6). In the 2D setting, we aim to learn the entire pressure field. Let $\mathcal{D}_a$ represent the irregular physical domain parametrized by a, indicating the shape of the airfoil. The domain $\mathcal{D}_a$ is discretized by a structured C-grid [249]. We introduce a latent space $\mathcal{D} = [0, 1]^2$ and the deformation map $\phi_a \colon \xi \to x(\xi)$ between $\mathcal{D}$ and $\mathcal{D}_a$. Here the deformation map has an analytical format and maps the uniform grid in $\mathcal{D}$ to the C-grid in $\mathcal{D}_a$. Subsequently, we formulate the operator learning problem in the latent space as

$$\Psi^\dagger_{\mathrm{F2F}} \colon \phi_a \to p \circ \phi_a \,. \tag{5.45}$$

In the preceding display, the deformation map ϕ_a is a function defined in $\mathcal{D}$, and $p \circ \phi_a$ represents the pressure function defined in $\mathcal{D}$. As mentioned previously, both lift and drag depend solely on the pressure distribution over the airfoil. Hence, we can alternatively formulate the learning problem in a 1D setting by focusing solely on learning the pressure distribution over the airfoil. We construct a one-dimensional latent space $\mathcal{D} = [0, 1]$ and also denote the deformation map as $\phi_a \colon \xi \to x(\xi)$ mapping from $\mathcal{D}$ to the shape of the airfoil. The corresponding operator learning problem in this 1D setting has the same

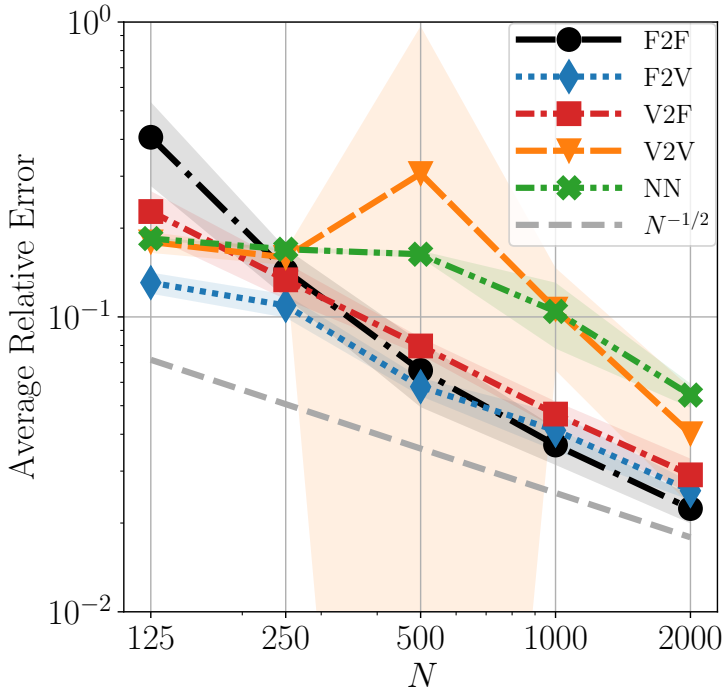


Figure 5.7: Flow over an airfoil. Comparative analysis of data size versus relative test error for the FNM and NN approaches. The shaded regions denote two standard deviations away from the mean of the test error over 5 realizations of the batch indices during SGD and model parameter initializations.

form as (5.45). The ground truth maps $\Psi^\dagger_{\text{F2V}}$, $\Psi^\dagger_{\text{V2F}}$, and $\Psi^\dagger_{\text{V2V}}$ are defined similarly, mapping either the deformation function ϕ_a or the 7-dimensional control node vector input to the pressure function or the QoI (5.44) itself. We use all four variants of the FNM architectures and a finite-dimensional NN to approximate these maps from data.

For each sample size N, five i.i.d. realizations of the models are trained on the relative loss (5.37) for 2000 epochs in L^2 output space norm for functions and Euclidean norm for vectors. All FNM models use 4 hidden layers, 12 modes per dimension, and a channel width of 128. We compare these models to a standard fully-connected NN with four layers and a hidden width of 128. In the case of FNM models, we observe that learning in the 1D setting consistently outperforms the 2D setting across all sizes of training data. Therefore, we only present results for the 1D setting. Moreover, this set of architectural hyperparameters with a large channel width of 128 in general outperforms other hyperparameter settings. Figure 5.7 reveals several trends. As the data volume N increases, all error curves decay at an algebraic rate that is slightly

faster than $N^{-1/2}$. This may be due to the small sample sizes considered (under 2000 data pairs) or, especially since the training data is noise-free, could be evidence of a data-driven "superconvergence" effect similar to that observed for QoI computations in adjoint methods for PDEs [110]. Overall, emulating PtO maps by training models with finite-dimensional vectors as both input and output (V2V and NN) is more challenging for this problem than adopting function space variants (F2F, F2V, V2F). The standard finite-dimensional NN performs similarly to V2V.

5.5.3 Effective Tensor for a Multiscale Elliptic Equation

This example considers an equation that arises in elasticity in computational solid mechanics and relates the material properties on small scales to the effective property on a larger scale: *homogenization*. Consider the linear multiscale elliptic equation on a bounded domain $\mathcal{D} \subset \mathbb{R}^2$ given by

$$\begin{aligned} -\nabla \cdot (A^\epsilon \nabla u^\epsilon) &= g \quad \text{in} \quad \mathcal{D}\,, \\ u^\epsilon &= 0 \quad \text{on} \quad \partial\mathcal{D}\,. \end{aligned} \tag{5.46}$$

Here A^ϵ is given by $x \mapsto A^\epsilon(x) = A\left(\frac{x}{\epsilon}\right)$ for some $A\colon \mathbb{T}^2 \to \mathbb{R}^{2\times 2}_{\text{sym},\succ 0}$ which is 1-periodic and positive definite. The source term is g. This equation contains fine-scale dependence through A^ϵ, which may be computationally expensive to evaluate without taking advantage of periodicity. The method of homogenization allows for elimination of the small scales in this manner and yields the homogenized equation

$$\begin{aligned} -\nabla \cdot \left(\overline{A} \nabla u\right) &= g \quad \text{in} \quad \mathcal{D}\,, \\ u &= 0 \quad \text{on} \quad \partial\mathcal{D}\,, \end{aligned} \tag{5.47}$$

where $\overline{A}$ is given by

$$\overline{A} = \int_{\mathbb{T}^2} \left(A(y) + A(y)\nabla\chi(y)^\top\right) dy \tag{5.48}$$

and $\chi\colon \mathbb{T}^2 \to \mathbb{R}^2$ solves the cell problem

$$-\nabla \cdot \left((\nabla\chi)A\right) = \nabla \cdot A \quad \text{in} \quad \mathbb{T}^2\,, \tag{5.49}$$

$$\int_{\mathbb{T}^2} \chi(y)\, dy = 0 \quad \text{and} \quad \chi \text{ is 1-periodic}\,. \tag{5.50}$$

For $0 < \epsilon \ll 1$, the solution u^ϵ of (5.46) is approximated by the solution u of (5.47). The error between the solutions converges to zero as $\epsilon \to 0$ [30, 217].

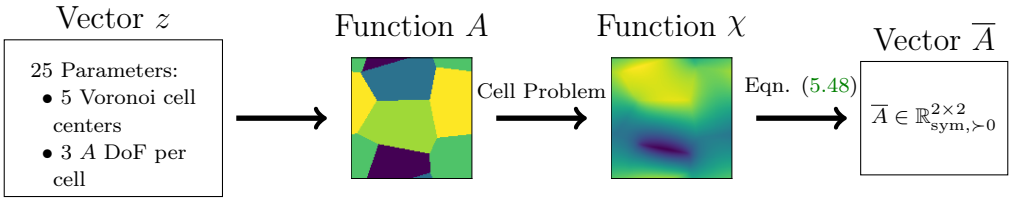


Figure 5.8: Diagram showing the homogenization experiment ground truth maps. The function A is parametrized by a finite vector z. The quantity of interest $\overline{A}$ (5.48) is computed from both the material function A and the solution χ to the cell problem (5.49). Note that both A and χ are functions on the torus $\mathbb{T}^2$.

The bottleneck step in obtaining the effective tensor $\overline{A}$, which is our QoI, is solving the cell problem (5.49). Learning the solution map $A \mapsto \chi$ in (5.49) corresponds to the F2F setting and is explored in detail in [35]. Alternately, one could learn the effective tensor $\overline{A}$ directly using the F2V-FNM architecture to approximate $A \mapsto \overline{A}$. Furthermore, though A is a function from $\mathbb{T}^2$ to $\mathbb{R}^{2\times 2}_{\mathrm{sym},\succ 0}$, in certain cases it may have an exact finite vector parametrization. One example of this case is finite piecewise-constant Voronoi tessellations; A takes constant values on a fixed number of cells, and the cell centers uniquely determine the Voronoi geometry. Denoting these parameters as $z \in \mathbb{R}^{d_u}$ for appropriate $d_u \in \mathbb{N}$, one could also learn the V2F map $z \to \chi$ or the V2V map $z \to \overline{A}$. In this experiment, we compare the error in the QoI $\overline{A}$ using all four methods. A visualization of the possible maps is shown in Figure 5.8. Since our example is defined in two spatial dimensions, the five Voronoi cell centers have two components each. The symmetry of A yields three degrees of freedom (DoF) on each Voronoi cell. Altogether, this yields 25 parameters that comprise the finite-dimensional vector input.

For training, we use the absolute squared loss in (5.38) with the H^1 norm for function output and Frobenius norm for vector output. Test error is also evaluated using these metrics. Data are generated with a finite element solver using the method described in [35]; both A and χ are interpolated to a 128×128 grid, and the Voronoi geometry is randomly generated for each sample. The test set size is 500. Each map uses hyperparameters obtained via a grid search. For F2F, F2V, V2F, and V2V, the number of modes are 18, 12, 12, and 18, and the channel widths are 64, 96, 96, and 64, respectively. The fully-connected NN used as a comparison has a channel width of 576 and 2 hidden layers. As a consequence, all methods have a fixed model size of modes times width equaling 1152.

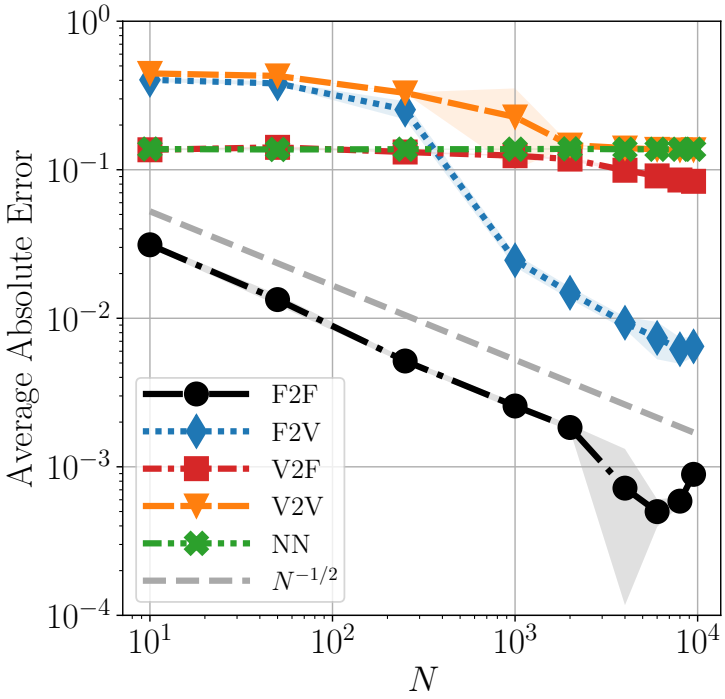


Figure 5.9: Elliptic homogenization problem. Absolute $\overline{A}$ error in the Frobenius norm versus data size for the FNM and NN architectures. The shaded regions denote two standard deviations away from the mean of the test error over 5 realizations of batch indices during SGD and model parameter initializations.

The results for the homogenization experiment in Figure 5.9 reinforce the theoretical intuition from Section 5.4 that learning with finite-dimensional vector data results in higher error than learning with functional data. Both the F2F and the F2V models approximately track the $N^{-1/2}$ rate, where N is the number of training data. On the other hand, the V2V model and NN model fail to attain this rate and saturate at the same level of roughly 10% error. The V2F map does achieve a slightly faster error decay rate than the V2V architecture for large enough sample sizes N, but it does not approach the $N^{-1/2}$ rate obtained by the F2F and F2V models. These convergence rate differences occur when there is a difference in input dimension. On the other hand, for a difference in output dimension, while both the F2F and F2V models reach roughly the same convergence rate, the F2V error remains an order of magnitude higher than the F2F error. We remark that when measuring performance with relative test error instead, the qualitative behavior of Figure 5.9 remains the same.

5.6 Conclusion

This chapter proposes the Fourier Neural Mappings (FNMs) framework as an operator learning method for approximating parameter-to-observable (PtO) maps with finite-dimensional vector inputs or outputs, or both. Universal approximation theorems demonstrate that FNMs are well suited for this task. Of central interest is the setting in which the PtO map factorizes into a vector-valued quantity of interest (QoI) map composed with a forward operator mapping between two function spaces. For this setting, the chapter introduces the end-to-end (EE) and full-field (FF) learning approaches. The (EE) approach directly estimates the PtO map from its own input-output pairs, while the (FF) approach estimates the forward map first and then plugs this estimator into the QoI. The main theoretical results of the chapter establish sample complexity bounds for Bayesian nonparametric regression of linear functionals with the (EE) and (FF) methods. The analysis reveals useful insights into how the smoothness of the QoI influences data efficiency. In particular, (FF) is superior to (EE) for smooth QoIs in this setting. The situation reverses for QoIs of low regularity. Finally, the chapter implements the FNM architectures for three nonlinear problems arising from environmental science, aerodynamics, and materials modeling. The numerical results support the linear theory and extend beyond it by revealing the supremacy of function space representations of the input space over analogous finite-dimensional vector parametrizations.

Several avenues for future work remain open. One way to understand the data efficiency of the (EE) and (FF) learning approaches beyond the specific Bayesian linear estimators that this chapter analyzes would involve the development of fundamental lower bounds. Besides a few recent works [5, 135, 158, 163, 169], there has been little attention on minimax lower bounds and (statistical) optimality for operator learning. The statistical theory in the present chapter fixes an infinite-dimensional input space and studies the influence of the output space (being either one-dimensional or infinite-dimensional). The derivation of similar insights to those in Subsection 5.4.3.2 for PtO map learning with vector-to-function estimators could help explain the strong influence of the input space observed in the numerical experiments from Section 5.5. The scalar input and infinite-dimensional output case is partially addressed by [232]. Furthermore, it remains to be seen whether the theory developed in the present chapter for the linear functional setting can extend to certain classes of nonlinear maps and QoIs, perhaps by linearizing the maps in an appropriate manner. For a

specific nonlinear map, relevant work in [238] studies the approximation of nonsmooth QoIs. On the practical side, it is of interest to further explore architectural improvements for the various FNMs and in particular whether the latent function space introduced by the vector-to-vector FNM (M-V2V) can actually lead to improved performance over standard finite-dimensional neural networks.

Chapter 6

THESIS CONCLUSION

Operator learning concerns the training of data-driven models that map between infinite-dimensional input and output spaces instead of spaces of finite-dimensional vectors. As a result, these continuum architectures are not sensitive to the discretization chosen at implementation time. An emerging paradigm, operator learning has the potential to enhance and accelerate scientific computation and discovery. This thesis studies statistical aspects of supervised learning for operators mapping between spaces of spatially-varying functions.

In Chapter 2, the thesis proposes the function-valued random features method for scalable nonlinear operator learning. Numerical experiments verify the computational efficiency and discretization invariance of the methodology in parametric partial differential equation examples. Chapter 3 provides an analysis of function-valued random feature regression with square loss. The theoretical results include statistical consistency guarantees and some of the sharpest convergence rates for random features to date. Going beyond universal approximation existence theorems, the theory developed in this chapter is appealing because it holds for a continuum algorithm that can actually be trained and implemented on a computer, as done in Chapter 2. Next, Chapter 4 frames the supervised learning of a linear operator between Hilbert spaces as a Bayesian inverse problem with a random forward map. The resulting analysis of this inverse problem establishes posterior contraction rates and generalization error bounds in the large data limit. These results provide practical insights into how to improve the sample complexity of linear operator learning. For the task of predicting finite-dimensional observables from the output state of some continuum system operator, Chapter 5 theoretically explores the relative difficulty of full-field linear operator learning versus end-to-end learning of the desired quantities of interest. Additionally, this chapter devises function space-consistent neural operator architectures for universally approximating nonlinear function-to-vector and vector-to-function mappings. Throughout the thesis, numerical evidence supports the theoretical findings and demonstrates the practical utility of the proposed operator learning algorithms.

The present chapter concludes this thesis by summarizing the main contributions in Section 6.1 and providing fruitful directions for future developments in the field of operator learning in Section 6.2.

6.1 Summary of Contributions

This thesis tackles the challenge of theoretically uncovering how problem structure, data quality, and prior information affect the generalization error of operator learning algorithms. The thesis also develops new algorithms that both realize the theory and go beyond it. This section reviews both the algorithmic and analytical contributions of the thesis.

6.1.1 Algorithms

This thesis develops theoretically justified operator learning algorithms that scalably and accurately approximate nonlinear mappings.

Chapter 2 proposes the use of function-valued random features to learn nonlinear operators mapping between infinite-dimensional Banach spaces of functions. It shows that the method is a parametric, finite-rank approximation of a full-rank operator-valued kernel or Gaussian process regression method. Due to efficient convex optimization, the function-valued random features method has a manageable offline training cost that is substantially cheaper than the offline cost for full-rank function-valued kernel regression. The proposed algorithm further enables the modeling of fully general output space correlations, while existing kernel methods are still unable to do so. When implemented on a computer, the conceptually infinite-dimensional random features algorithm is consistent with the continuum limit, robust to the particular choice of discretization, and highly transferrable in practice. The chapter further designs problem-adapted Fourier Space Random Feature maps with cheap online evaluation cost, which makes the algorithm well suited for the task of speeding up otherwise prohibitively expensive many-query problems.

Chapter 5 identifies the need for operator learning architectures that are able to handle finite-dimensional vectors as inputs or outputs. This need arises when input functions are parametrized in a low-dimensional way, such as in design optimization problems, or when state measurements of an underlying continuum system are limited in resolution or involve indirect quantities. The chapter introduces the neural mappings framework for supervised learning in such scenarios, with a focus on the particular Fourier Neural Mappings class

that builds upon the existing Fourier Neural Operator architecture. The proposed methodology is a mathematically principled, fully data-driven approach that does not require explicit knowledge about the finite-dimensional input parametrization, output quantity of interest map, or the underlying continuum system operator. New functional encoder layers map functions to vectors in a discretization-invariant way. Similarly, linear decoder layers consistently map vectors to functions. Fourier Neural Mappings inherit all of the benefits of the original Fourier Neural Operator, including universal approximation properties and fast forward passes through the network. Numerical results support the linear insights developed later in Chapter 5 and extend beyond them by revealing that continuum representations of the input space are superior to finite-dimensional vector representations, especially for a material homogenization problem.

6.1.2 Analysis

The theoretical analysis in this thesis is united by how smoothness, model misspecification, and prior domain knowledge influence the data efficiency of operator learning. Additionally, both Chapters 4 and 5 contribute to the development of principled uncertainty quantification for linear operator learning through Bayesian inference methodology and analysis.

Chapter 3 develops state-of-the-art parameter complexity bounds for random feature ridge regression. It focuses on model error due to smoothness misspecification, leading to strong asymptotic consistency theorems for the methodology and non-asymptotic error bounds robust to misspecification that hold with high probability. Although the high-level proof approach based on decomposing the squared test error into the training error plus the generalization gap is a classical idea, the analysis develops several novel techniques that sharpen the argument and lead to tighter rates. These include the self-bounding argument leading to a high probability bound on the norm of the trained random feature model's coefficients and empirical process bounds that do not require vector contraction results for Rademacher complexity.

Chapter 4 uncovers novel theoretical principles about how the smoothness of the problem, smoothness of the training data, and smoothness of the test data affect the sample complexity of linear operator learning. These principles have implications for the optimal acquisition of training data, the robustness of

learned models under data distribution shifts, and how the accuracy of such models should be evaluated. The theory is applied to answer several basic questions in the field. For example, the chapter shows that self-adjoint linear operators involving differentiation or integration of functions can be learned from noisy data pairs and bounds how much data is required to do so. The core of the analysis centers on new statistical guarantees for a white noise sequence space regression model with correlated random coefficients and smoothness misspecification from the prior distribution. Along the way, the chapter proves technical lemmas that are applicable to sums of dependent subexponential random variables. These independently interesting results lead to concentration bounds for random series and are of wide applicability, as demonstrated by their use in both Chapters 4 and 5. The lower bounds in expectation are also of interest both in terms of the proof technique and their potential applicability to other statistical problems.

Chapter 5 provides a new Bayesian functional linear regression analysis that remains valid when the nonparametric model class is misspecified, i.e., the smoothness implied by the prior distribution and the smoothness of the ground truth functional do not match. This result is of independent interest, owing to a refined bias bound in the underlying bias–variance decomposition. Going further, the theory establishes a sample complexity comparison between end-to-end and full-field learning of composite linear parameter-to-observable maps of the form $f = q \circ L$. Specifically, the chapter analyzes the sample complexity of (i) directly regressing f from paired data, and (ii) a plug-in estimator based on exact knowledge of q plus regressing only the operator L. The full-field approach (ii) delivers a more accurate estimate of f than approach (i) does, provided that the quantity of interest functional q is smooth enough. In this case, prior knowledge of continuum problem structure gives a quantitative statistical advantage over purely data-driven end-to-end learning. This result gives theoretical validity to empirical observations made about the benefits of full-field learning in various fields. The chapter also proves new universal approximation theorems for the Fourier Neural Mappings architectures, giving the nonlinear numerical experiments in Chapter 5 some theoretical grounding.

6.2 Outlook

Although the new field of operator learning has witnessed rapid growth in recent times, the bulk of the research has centered on surrogate modeling of

partial differential equation solution operators in one, two, or three spatial dimensions and possibly one time dimension. This is understandable, as such continuum physical models are ubiquitous in science and engineering. However, more challenging problems include those with much higher state space dimensions. These include the high-dimensional settings of quantum many-body problems, reinforcement learning, mathematical finance, and filtering and data assimilation. Also, solution operators of high- or infinite-dimensional partial differential equations appear in physics, e.g., Schrödinger or quantum field equations [18], and in control theory, e.g., Hamilton–Jacobi equations or mean field games [50]. Numerical solvers for these problems are still in their infancy. It is of interest to probe the limits of operator learning for such applications. Beyond ordinary, partial, and stochastic differential equation models, there are entirely different classes of problems in other fields that are underexplored and may benefit from an operator learning perspective. For example, there are diverse datasets in biology [276], medicine [237], finance [269], and the social sciences [22] that deserve a proper infinite-dimensional treatment.

The present thesis studies supervised operator learning problems. However, the framework of performing machine learning in infinite dimensions is much more general than this. One can envision unsupervised [12], semi-supervised [137], active [167], and online operator learning [226] from potentially diverse sources of noisy and incomplete data. For instance, in the unsupervised setting, only unlabeled data are available. One canonical problem of this type is covariance estimation [106]. Here, the unlabeled dataset consists of centered random functions $x \mapsto u_n(x)$ sampled from an unknown probability measure μ, that is,

$$u_n \sim \mu \quad \text{for} \quad n = 1, \dots, N\,. \tag{6.1}$$

Given these data, the goal is to estimate the unknown covariance function

$$(x, x') \mapsto \mathbb{E}^{u\sim\mu}\big[u(x)u(x')\big] \tag{6.2}$$

or its operator representation on a suitable function space. One may also be interested in estimating specific functionals of the covariance operator and not the entire operator itself [148]. This problem is typically addressed from a nonparametric statistics perspective. It would be interesting to apply (deep) operator learning ideas here too.

Alternatively, the observed data might only consist of noisy, indirect, and sparse measurements of a continuum system, as is common in inverse problems.

Here, the only accessible data are the labeled outputs. The inputs that give rise to the labels are unknown. The inverse problem is to infer the (possible range of) inputs that produce the observed output quantity. Some inverse problems are even more challenging, such as blind deconvolution [28]. Given a blurry image $y\colon \mathcal{D} \to \mathbb{R}$, the goal of blind deconvolution is to recover both the underlying clean image function $u\colon \mathcal{D} \to \mathbb{R}$ and the convolution filter kernel function $\kappa\colon \mathcal{D} \to \mathbb{R}$ from y subject to the assumed statistical model

$$y = \int_{\mathcal{D}} \kappa(\cdot - x)u(x)\,dx + \eta\,, \tag{6.3}$$

where η represents a random noise process. That is, the solution to blind deconvolution is the convolution operator itself as well as the unknown clean image u. Sometimes an optional collection of labels $\{y_n\}_{n=1}^N$ is also provided in addition to y, where each label y_n corresponds to a true image u_n according to (6.3). This data access model suggests that an indirect form of operator learning could be applicable. It is clear that inverse problems pose new opportunities for operator learning methods because traditional algorithms for solving inverse problems are not always reliable or accurate due to inherent ill-posedness issues. Most existing machine learning research in this area studies the learning of regularization operators for inverse problems [15]. However, this line of work does not learn the entire inversion map itself. An operator learning approach may prove to be useful in this challenging setting.

One of the main theoretical findings in the present thesis is that the continuum qualities of the infinite-dimensional training data quantitatively impact statistical performance of operator learning algorithms. Building on this insight, future development of active learning or optimal data acquisition strategies applicable in infinite dimensions would serve to boost the effectiveness of operator learning in challenging limited data scenarios. Moreover, a collective community shift away from random training data generation and toward greedy or more principled data generation strategies that exploit problem-specific structure is highly desirable and sorely needed to address difficult problems currently considered to be out of reach. Altogether, the methodological and theoretical contributions of this thesis serve to improve the robustness, reliability, and efficiency of continuum machine learning algorithms and help lay the foundation for future advances in the emerging field of operator learning.

BIBLIOGRAPHY

[1] Jacob Abernethy, Francis Bach, Theodoros Evgeniou, and Jean-Philippe Vert. "A new approach to collaborative filtering: Operator estimation with spectral regularization". *Journal of Machine Learning Research* 10.3 (2009), pp. 803–826 (cited on page 84).

[2] Kweku Abraham and Richard Nickl. "On statistical Calderón problems". *Mathematical Statistics and Learning* 2.2 (2019), pp. 165–216 (cited on page 86).

[3] Ben Adcock, Simone Brugiapaglia, Nick Dexter, and Sebastian Moraga. "Near-optimal learning of Banach-valued, high-dimensional functions via deep neural networks". *preprint arXiv:2211.12633* (2022) (cited on pages 8, 9).

[4] Ben Adcock, Simone Brugiapaglia, Nick Dexter, and Sebastian Morage. "Deep neural networks are effective at learning high-dimensional Hilbert-valued functions from limited data". In: *Mathematical and Scientific Machine Learning*. PMLR. 2022, pp. 1–36 (cited on pages 7, 22, 83).

[5] Ben Adcock, Nick Dexter, and Sebastian Moraga. "Optimal approximation of infinite-dimensional holomorphic functions". *Calcolo* 61.1 (2024), p. 12 (cited on pages 9, 151).

[6] Sergios Agapiou, Johnathan M. Bardsley, Omiros Papaspiliopoulos, and Andrew M. Stuart. "Analysis of the Gibbs sampler for hierarchical inverse problems". *SIAM/ASA Journal on Uncertainty Quantification* 2.1 (2014), pp. 511–544 (cited on page 104).

[7] Sergios Agapiou and Ismaël Castillo. "Heavy-tailed Bayesian nonparametric adaptation". *preprint arXiv:2308.04916* (2023) (cited on page 135).

[8] Sergios Agapiou, Stig Larsson, and Andrew M Stuart. "Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems". *Stochastic Processes and Their Applications* 123.10 (2013), pp. 3828–3860 (cited on pages 85, 101, 128).

[9] Sergios Agapiou and Peter Mathé. "Designing truncated priors for direct and inverse Bayesian problems". *Electronic Journal of Statistics* 16.1 (2022), pp. 158–200 (cited on pages 85, 104).

[10] Sergios Agapiou and Peter Mathé. "Posterior contraction in Bayesian inverse problems under Gaussian priors". In: *New Trends in Parameter Identification for Mathematical Models*. Ed. by B. Hofmann, A. Leitão, and J. Zubelli. Trends in Mathematics. Birkhäuser, Cham, 2018, pp. 1–29 (cited on pages 77, 90, 101).

[11] Sergios Agapiou, Andrew M Stuart, and Yuan-Xiang Zhang. "Bayesian posterior contraction rates for linear severely ill-posed inverse problems". *Journal of Inverse and Ill-Posed Problems* 22.3 (2014), pp. 297–321 (cited on pages 77, 82, 101, 234).

[12] Giovanni S. Alberti, Ernesto De Vito, Matti Lassas, Luca Ratti, and Matteo Santacesaria. "Learning the optimal Tikhonov regularizer for inverse problems". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 25205–25216 (cited on pages 83, 93, 157).

[13] Rita Giuliano Antonini. "Subgaussian random variables in Hilbert spaces". *Rendiconti del Seminario Matematico della Università di Padova* 98 (1997), pp. 89–99 (cited on page 102).

[14] Nachman Aronszajn. "Theory of reproducing kernels". *Transactions of the American Mathematical Society* 68.3 (1950), pp. 337–404 (cited on pages 27, 56).

[15] Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. "Solving inverse problems using data-driven models". *Acta Numerica* 28 (2019), pp. 1–174 (cited on pages 83, 84, 158).

[16] Andrea Aspri, Yury Korolev, and Otmar Scherzer. "Data driven regularization by projection". *Inverse Problems* 36, 125009 (2020) (cited on page 83).

[17] Francis Bach. "On the equivalence between kernel quadrature rules and random feature expansions". *Journal of Machine Learning Research* 18.1 (2017), pp. 714–751 (cited on pages 23, 27, 28, 30, 57, 60, 62, 64, 187).

[18] C. Bagnuls and C. Bervillier. "Exact renormalization group equations: an introductory review". *Physics Reports* 348.1-2 (2001), pp. 91–157 (cited on page 157).

[19] Yohai Bar-Sinai, Stephan Hoyer, Jason Hickey, and Michael P. Brenner. "Learning data-driven discretizations for partial differential equations". *Proceedings of the National Academy of Sciences* 116.31 (2019), pp. 15344–15349 (cited on page 22).

[20] Maxime Barrault, Yvon Maday, Ngoc Cuong Nguyen, and Anthony T. Patera. "An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations". *Comptes Rendus Mathematique* 339.9 (2004), pp. 667–672 (cited on page 21).

[21] Andrew R. Barron. "Universal approximation bounds for superpositions of a sigmoidal function". *IEEE Transactions on Information theory* 39.3 (1993), pp. 930–945 (cited on page 19).

[22] David J. Bartholomew, Fiona Steele, and Irini Moustaki. *Analysis of multivariate social science data*. CRC Press, 2008 (cited on page 157).

[23] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. "Benign overfitting in linear regression". *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30063–30070 (cited on page 129).

[24] Jan-Hendrik Bastek and Dennis M. Kochmann. "Inverse-design of nonlinear mechanical metamaterials via video denoising diffusion models". *Nature Machine Intelligence* 5 (2023), pp. 1466–1475 (cited on pages 4, 113).

[25] Pau Batlle, Matthieu Darcy, Bamdad Hosseini, and Houman Owhadi. "Kernel methods are competitive for operator learning". *Journal of Computational Physics* 496, 112549 (2024) (cited on page 9).

[26] Jacob Bear and M. Yavuz Corapcioglu. *Fundamentals of transport phenomena in porous media*. Vol. 82. Springer Science & Business Media, 2012 (cited on page 40).

[27] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. "Reconciling modern machine-learning practice and the classical bias–variance trade-off". *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854 (cited on page 23).

[28] Anthony J. Bell and Terrence J. Sejnowski. "An information-maximization approach to blind separation and blind deconvolution". *Neural Computation* 7.6 (1995), pp. 1129–1159 (cited on page 158).

[29] Peter Benner, Albert Cohen, Mario Ohlberger, and Karen Willcox. *Model reduction and approximation: theory and algorithms*. Vol. 15. SIAM, 2017 (cited on pages 21, 25).

[30] Alain Bensoussan, Jacques-Louis Lions, and George Papanicolaou. *Asymptotic analysis for periodic structures*. Vol. 374. American Mathematical Society, 2011 (cited on page 148).

[31] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011 (cited on pages 27, 28).

[32] Christine Bernardi and Rüdiger Verfürth. "Adaptive finite element methods for elliptic equations with non-smooth coefficients". *Numerische Mathematik* 85.4 (2000), pp. 579–608 (cited on page 41).

[33] Gregory Beylkin and Martin J. Mohlenkamp. "Algorithms for numerical analysis in high dimensions". *SIAM Journal on Scientific Computing* 26.6 (2005), pp. 2133–2159 (cited on page 25).

[34] Kaushik Bhattacharya, Bamdad Hosseini, Nikola B. Kovachki, and Andrew M. Stuart. "Model reduction and neural networks for parametric PDEs". *The SMAI Journal of Computational Mathematics* 7 (2021), pp. 121–157 (cited on pages 7, 8, 20, 23, 33, 49, 52, 56, 75, 83, 87).

[35] Kaushik Bhattacharya, Nikola B. Kovachki, Aakila Rajan, Andrew M. Stuart, and Margaret Trautner. "Learning homogenization for elliptic operators". *preprint arXiv:2306.12006* (2023) (cited on pages 115, 117, 143, 149).

[36] Daniele Bigoni, Yuming Chen, Nicolas Garcia Trillos, Youssef Marzouk, and Daniel Sanz-Alonso. "Data-driven forward discretizations for Bayesian inversion". *Inverse Problems* 36.10, 105008 (2020) (cited on page 22).

[37] Nicholas H. Bingham, Charles M. Goldie, Jozef L. Teugels, and J. L. Teugels. *Regular Variation*. Vol. 27. Cambridge University Press, 1989 (cited on page 100).

[38] Ismael Rodrigo Bleyer and Ronny Ramlau. "A double regularization approach for inverse problems with noisy data and inexact operator". *Inverse Problems* 29.2, 025004 (2013) (cited on page 85).

[39] Jan Bohr. "Stability of the non-abelian X-ray transform in dimension ≥ 3". *Journal of Geometric Analysis* 31.11 (2021), pp. 11226–11269 (cited on page 86).

[40] Nicolas Boullé, Diana Halikias, and Alex Townsend. "Elliptic PDE learning is provably data-efficient". *Proceedings of the National Academy of Sciences* 120.39, e2303904120 (2023) (cited on pages 8, 9, 116).

[41] Nicolas Boullé and Alex Townsend. "A Mathematical Guide to Operator Learning". *preprint arXiv:2312.14688* (2023) (cited on pages 1, 9).

[42] Nicolas Boullé and Alex Townsend. "Learning elliptic partial differential equations with randomized linear algebra". *Foundations of Computational Mathematics* 23.2 (2023), pp. 709–739 (cited on pages 8, 14, 57, 84).

[43] Romain Brault, Markus Heinonen, and Florence Buc. "Random Fourier features for operator-valued kernels". In: *Asian Conference on Machine Learning*. 2016, pp. 110–125 (cited on pages 23, 33, 57).

[44] Simone Brivio, Stefania Fresca, Nicola Rares Franco, and Andrea Manzoni. "Error estimates for POD-DL-ROMs: a deep learning framework for reduced order modeling of nonlinear parametrized PDEs enhanced by proper orthogonal decomposition". *Advances in Computational Mathematics* 50.3 (2024), p. 33 (cited on page 8).

[45] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. "Discovering governing equations from data by sparse identification of nonlinear dynamical systems". *Proceedings of the National Academy of Sciences* 113.15 (2016), pp. 3932–3937 (cited on pages 14, 83).

[46] Tatiana A. Bubba, Mathilde Galinier, Matti Lassas, Marco Prato, Luca Ratti, and Samuli Siltanen. "Deep neural networks for inverse problems with pseudodifferential operators: An application to limited-angle tomography". *SIAM Journal on Imaging Sciences* 14.2 (2021), pp. 470–505 (cited on page 84).

[47] Tan Bui-Thanh, Qin Li, and Leonardo Zepeda-Núñez. "Bridging and improving theoretical and computational electrical impedance tomography via data completion". *SIAM Journal on Scientific Computing* 44.3 (2022), B668–B693 (cited on page 83).

[48] T. Tony Cai and Peter Hall. "Prediction in functional linear regression". *Annals of Statistics* 34.5 (2006), pp. 2159–2179 (cited on pages 79, 99, 116).

[49] T. Tony Cai and Ming Yuan. "Minimax and adaptive prediction for functional linear regression". *Journal of the American Statistical Association* 107.499 (2012), pp. 1201–1216 (cited on pages 116, 135).

[50] Piermarco Cannarsa and Maria Elisabetta Tessitore. "Infinite-dimensional Hamilton–Jacobi equations and Dirichlet boundary control problems of parabolic type". *SIAM Journal on Control and Optimization* 34.6 (1996), pp. 1831–1847 (cited on page 157).

[51] Yuan Cao and Quanquan Gu. "Generalization bounds of stochastic gradient descent for wide and deep neural networks". In: *Advances in Neural Information Processing Systems*. 2019, pp. 10835–10845 (cited on pages 23, 33).

[52] Andrea Caponnetto and Ernesto De Vito. "Optimal rates for the regularized least-squares algorithm". *Foundations of Computational Mathematics* 7 (2007), pp. 331–368 (cited on pages 9, 11, 49, 56, 57, 61, 83, 94, 111, 116, 189, 266).

[53] Andrea Caponnetto, Charles A. Micchelli, Massimiliano Pontil, and Yiming Ying. "Universal multi-task kernels". *The Journal of Machine Learning Research* 9 (2008), pp. 1615–1646 (cited on page 66).

[54] Hervé Cardot, André Mas, and Pascal Sarda. "CLT in functional linear regression models". *Probability Theory and Related Fields* 138 (2007), pp. 325–361 (cited on pages 116, 234).

[55] Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. "Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem". *Analysis and Applications* 4.4 (2006), pp. 377–408 (cited on pages 27–29, 193, 194).

[56] Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. "Vector valued reproducing kernel Hilbert spaces and universality". *Analysis and Applications* 8.1 (2010), pp. 19–61 (cited on page 66).

[57] Laurent Cavalier. "Nonparametric statistical inverse problems". *Inverse Problems* 24.3, 034004 (2008) (cited on pages 77, 99).

[58] Guodong Chen and Krzysztof Fidkowski. "Output-based error estimation and mesh adaptation using convolutional neural networks: application to a scalar advection-diffusion problem". In: *AIAA Scitech 2020 Forum*. 2020, 1143 (cited on page 22).

[59] Tianping Chen and Hong Chen. "Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems". *IEEE Transactions on Neural Networks* 6.4 (1995), pp. 911–917 (cited on pages 8, 22, 23, 115).

[60] Zhijun Chen, Hayden Schaeffer, and Rachel Ward. "Concentration of random feature matrices in high-dimensions". In: *Mathematical and Scientific Machine Learning*. PMLR. 2022, pp. 287–302 (cited on page 57).

[61] Mulin Cheng, Thomas Y. Hou, Mike Yan, and Zhiwen Zhang. "A data-driven stochastic method for elliptic PDEs with random coefficients". *SIAM/ASA Journal on Uncertainty Quantification* 1.1 (2013), pp. 452–493 (cited on page 22).

[62] Abdellah Chkifa, Albert Cohen, Ronald DeVore, and Christoph Schwab. "Sparse adaptive Taylor approximation algorithms for parametric and stochastic elliptic PDEs". *ESAIM: Mathematical Modelling and Numerical Analysis* 47.1 (2013), pp. 253–280 (cited on page 23).

[63] Albert Cohen and Ronald DeVore. "Approximation of high-dimensional parametric PDEs". *Acta Numerica* 24 (2015), pp. 1–159 (cited on pages 21, 23, 25).

[64] Albert Cohen, Ronald Devore, and Christoph Schwab. "Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDEs". *Analysis and Applications* 9.1 (2011), pp. 11–47 (cited on page 57).

[65] Albert Cohen and Giovanni Migliorati. "Optimal weighted least-squares methods". *The SMAI Journal of Computational Mathematics* 3 (2017), pp. 181–203 (cited on page 25).

[66] Matthew J. Colbrook, Vegard Antun, and Anders C. Hansen. "The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem". *Proceedings of the National Academy of Sciences* 119.12, e2107151119 (2022) (cited on page 83).

[67] Matthew J. Colbrook, Andrew Horning, and Alex Townsend. "Computing spectral measures of self-adjoint operators". *SIAM Review* 63.3 (2021), pp. 489–524 (cited on pages 5, 79).

[68] Simon L. Cotter, Gareth O. Roberts, Andrew M. Stuart, and David White. "MCMC methods for functions: modifying old algorithms to make them faster". *Statistical Science* (2013), pp. 424–446 (cited on pages 4, 24).

[69] Christophe Crambes and André Mas. "Asymptotics of prediction in functional linear regression with functional outputs". *Bernoulli* 19.5B (2013), pp. 2627–2651 (cited on pages 75, 84).

[70] Felipe Cucker and Steve Smale. "On the mathematical foundations of learning". *Bulletin of the American Mathematical Society* 39.1 (2002), pp. 1–49 (cited on pages 27, 28, 186).

[71] George Cybenko. "Approximation by superpositions of a sigmoidal function". *Mathematics of Control, Signals and Systems* 2.4 (1989), pp. 303–314 (cited on page 115).

[72] Niccolò Dal Santo, Simone Deparis, and Luca Pegolotti. "Data driven approximation of parametrized PDEs by reduced basis and neural networks". *Journal of Computational Physics* 416, 109550 (2020) (cited on pages 22, 23).

[73] Masoumeh Dashti, Kody J. H. Law, Andrew M. Stuart, and Jochen Voss. "MAP estimators and their consistency in Bayesian nonparametric inverse problems". *Inverse Problems* 29.9, 095017 (2013) (cited on page 94).

[74] Masoumeh Dashti and Andrew M. Stuart. "The Bayesian approach to inverse problems". In: *Handbook of Uncertainty Quantification*. Ed. by R. Ghanem, D. Higdon, and H. Owhadi. Springer, Cham, 2017, pp. 311–428. ISBN: 978-3-319-12385-1 (cited on pages 27, 39, 91, 126).

[75] Maarten V. de Hoop, Daniel Zhengyu Huang, Elizabeth Qian, and Andrew M. Stuart. "The cost-accuracy trade-off in operator learning with neural networks". *Journal of Machine Learning* 1.3 (2022), pp. 299–341 (cited on pages 14, 87, 111).

[76] Maarten V. de Hoop, Nikola B. Kovachki, Nicholas H. Nelsen, and Andrew M. Stuart. "Convergence rates for learning linear operators from noisy data". *SIAM/ASA Journal on Uncertainty Quantification* 11.2 (2023), pp. 480–513 (cited on pages 8, 9, 57, 116, 126, 127, 131, 132, 135, 136, 234, 240, 246, 247, 256–258).

[77] Maarten V. de Hoop, Matti Lassas, and Christopher A. Wong. "Deep learning architectures for nonlinear operator functions and nonlinear inverse problems". *Mathematical Statistics and Learning* 4.1 (2022), pp. 1–86 (cited on pages 83, 84).

[78] Tim De Ryck and Siddhartha Mishra. "Generic bounds on the approximation error for physics-informed (and) operator learning". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 10945–10958 (cited on page 57).

[79] Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Umberto De Giovannini, and Francesca Odone. "Learning from examples as an inverse problem". *Journal of Machine Learning Research* 6 (2005), pp. 883–904 (cited on page 75).

[80] Laurent Demanet. "Curvelets, wave atoms, and wave equations". PhD thesis. California Institute of Technology, 2006 (cited on page 48).

[81] Beichuan Deng, Yeonjong Shin, Lu Lu, Zhongqiang Zhang, and George Em Karniadakis. "Approximation rates of DeepONets for learning operators arising from advection–diffusion equations". *Neural Networks* 153 (2022), pp. 411–426 (cited on pages 57, 116, 117).

[82] Ronald A. DeVore. "The theoretical foundation of reduced basis methods". *Model Reduction and Approximation: Theory and Algorithms* 15 (2017), pp. 137–168 (cited on page 21).

[83] Alireza Doostan and Gianluca Iaccarino. "A least-squares approximation of partial differential equations with high-dimensional random inputs". *Journal of Computational Physics* 228.12 (2009), pp. 4332–4345 (cited on page 25).

[84] James Dugundji. "An extension of Tietze's theorem". *Pacific Journal of Mathematics* 1.3 (1951), pp. 353–367 (cited on page 230).

[85] Oliver R. A. Dunbar, Maya Mutic, and Nicholas H. Nelsen. "Hyperparameter optimization for randomized algorithms: a case study for random features". *In preparation* (2024) (cited on pages 32, 36, 38, 54).

[86] Matthew M. Dunlop, Marco A. Iglesias, and Andrew M. Stuart. "Hierarchical Bayesian level set inversion". *Statistics and Computing* 27.6 (2017), pp. 1555–1584 (cited on page 39).

[87] Weinan E. "A proposal on machine learning via dynamical systems". *Communications in Mathematics and Statistics* 5.1 (2017), pp. 1–11 (cited on page 24).

[88] Weinan E, Jiequn Han, and Qianxiao Li. "A mean-field optimal control formulation of deep learning". *Research in the Mathematical Sciences* 6.1 (2019), p. 10 (cited on page 24).

[89] Weinan E, Chao Ma, and Lei Wu. "Machine learning from a continuous viewpoint, I". *Science China Mathematics* 63.11 (2020), pp. 2233–2266 (cited on page 24).

[90] Weinan E, Chao Ma, and Lei Wu. "The Barron space and the flow-induced function spaces for neural network models". *Constructive Approximation* 55.1 (2022), pp. 369–406 (cited on page 115).

[91] Weinan E, Chao Ma, and Lei Wu. "The generalization error of the minimum-norm solutions for over-parameterized neural networks". *Pure and Applied Functional Analysis* 5.6 (2020), pp. 1445–1460 (cited on pages 22, 23, 37, 57).

[92] Weinan E and Bing Yu. "The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems". *Communications in Mathematics and Statistics* 6.1 (2018), pp. 1–12 (cited on page 21).

[93] Lawrence C. Evans. *Partial Differential Equations*. Vol. 19. American Mathematical Society, 2010 (cited on page 41).

[94] Yuwei Fan and Lexing Ying. "Solving electrical impedance tomography with deep learning". *Journal of Computational Physics* 404 (2020), pp. 109–119 (cited on pages 23, 84).

[95] Gerald Farin. *Curves and surfaces for computer-aided geometric design: a practical guide*. Elsevier, 2014 (cited on page 145).

[96] Jordi Feliu-Faba, Yuwei Fan, and Lexing Ying. "Meta-learning pseudo-differential operators with deep neural networks". *Journal of Computational Physics* 408, 109309 (2020) (cited on page 23).

[97] Frédéric Ferraty, André Mas, and Philippe Vieu. "Nonparametric regression on functional data: inference and practical aspects". *Australian & New Zealand Journal of Statistics* 49.3 (2007), pp. 267–286 (cited on page 9).

[98] Simon Fischer and Ingo Steinwart. "Sobolev norm learning rates for regularized least-squares algorithms". *Journal of Machine Learning Research* 21.1 (2020), pp. 8464–8501 (cited on page 266).

[99] Bengt Fornberg. *A practical guide to pseudospectral methods*. Vol. 1. Cambridge University Press, 1998 (cited on page 46).

[100] Nicola Rares Franco and Simone Brugiapaglia. "A practical existence theorem for reduced order models based on convolutional autoencoders". *preprint arXiv:2402.00435* (2024) (cited on page 8).

[101] Nicola Rares Franco, Daniel Fraulin, Andrea Manzoni, and Paolo Zunino. "On the latent dimension of deep autoencoders for reduced order modeling of PDEs parametrized by random fields". *preprint arXiv:2310.12095* (2023) (cited on page 8).

[102] Nicola Rares Franco, Stefania Fresca, Andrea Manzoni, and Paolo Zunino. "Approximation bounds for convolutional neural networks in operator learning". *Neural Networks* 161 (2023), pp. 129–141 (cited on page 8).

[103] Han Gao, Jian-Xun Wang, and Matthew J. Zahr. "Non-intrusive model reduction of large-scale, nonlinear dynamical systems using deep learning". *Physica D: Nonlinear Phenomena* 412, 132614 (2020) (cited on page 22).

[104] Leszek Gawarecki and Vidyadhar Mandrekar. *Stochastic differential equations in infinite dimensions: with applications to stochastic partial differential equations*. Springer Berlin Heidelberg, 2010 (cited on page 90).

[105] Moritz Geist, Philipp Petersen, Mones Raslan, Reinhold Schneider, and Gitta Kutyniok. "Numerical solution of the parametric diffusion equation by deep neural networks". *Journal of Scientific Computing* 88.1 (2021), p. 22 (cited on page 22).

[106] Omar Al-Ghattas, Jiaheng Chen, Daniel Sanz-Alonso, and Nathan Waniorek. "Covariance operator estimation: sparsity, lengthscale, and ensemble Kalman filters". *preprint arXiv:2310.16933* (2023) (cited on page 157).

[107] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "Linearized two-layers neural networks in high dimension". *Annals of Statistics* 49 (2021), pp. 1029–1054 (cited on page 57).

[108] Dimitrios Giannakis. "Data-driven spectral decomposition and forecasting of ergodic dynamical systems". *Applied and Computational Harmonic Analysis* 47.2 (2019), pp. 338–396 (cited on page 83).

[109] David Gilbarg and Neil S. Trudinger. *Elliptic partial differential equations of second order*. springer, 2015 (cited on page 40).

[110] Michael B. Giles and Endre Süli. "Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality". *Acta Numerica* 11 (2002), pp. 145–236 (cited on page 148).

[111] Craig R. Gin, Daniel E. Shea, Steven L. Brunton, and J. Nathan Kutz. "DeepGreen: deep learning of Green's functions for nonlinear boundary value problems". *Scientific Reports* 11.1, 21614 (2021) (cited on pages 8, 56).

[112] Matteo Giordano and Hanne Kekkonen. "Bernstein–von Mises theorems and uncertainty quantification for linear inverse problems". *SIAM/ASA Journal on Uncertainty Quantification* 8.1 (2020), pp. 342–373 (cited on page 85).

[113] Gene H. Golub and Charles F. Van Loan. "An analysis of the total least squares problem". *SIAM Journal on Numerical Analysis* 17.6 (1980), pp. 883–893 (cited on page 85).

[114] Lukas Gonon, Lyudmila Grigoryeva, and Juan-Pablo Ortega. "Approximation bounds for random neural networks and reservoir systems". *The Annals of Applied Probability* 33.1 (2023), pp. 28–69 (cited on page 57).

[115] R. Gonzalez-Garcia, R. Rico-Martinez, and I. G. Kevrekidis. "Identification of distributed parameter systems: A neural net based approach". *Computers & Chemical Engineering* 22 (1998), S965–S968 (cited on page 22).

[116] Somdatta Goswami, Minglang Yin, Yue Yu, and George Em Karniadakis. "A physics-informed variational DeepONet for predicting crack path in quasi-brittle materials". *Computer Methods in Applied Mechanics and Engineering* 391, 114587 (2022) (cited on page 115).

[117] Michael Griebel and Christian Rieger. "Reproducing kernel Hilbert spaces for parametric partial differential equations". *SIAM/ASA Journal on Uncertainty Quantification* 5.1 (2017), pp. 111–137 (cited on page 23).

[118] Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil. "Conditional mean embeddings as regressors". *preprint arXiv:1205.4656* (2012) (cited on pages 8, 84).

[119] Shota Gugushvili, Aad van der Vaart, and Dong Yan. "Bayesian linear inverse problems in regularity scales". *Annales de l'Institut Henri Poincaré Probability and Statistics* 56.3 (2020), pp. 2081–2107 (cited on pages 78, 85, 110).

[120] Eldad Haber and Lars Ruthotto. "Stable architectures for deep neural networks". *Inverse Problems* 34.1, 014004 (2017) (cited on page 24).

[121] Martin Hairer, Andrew M. Stuart, and Sebastian Vollmer. "Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions". *The Annals of Applied Probability* 24.6 (2014), pp. 2455–2490 (cited on page 4).

[122] Abolfazl Hashemi, Hayden Schaeffer, Robert Shi, Ufuk Topcu, Giang Tran, and Rachel Ward. "Generalization bounds for sparse random feature expansions". *Applied and Computational Harmonic Analysis* 62 (2023), pp. 310–330 (cited on page 57).

[123] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009 (cited on page 26).

[124] Lukas Herrmann, Joost A. A. Opschoor, and Christoph Schwab. "Constructive deep ReLU neural network approximation". *Journal of Scientific Computing* 90.2, 75 (2022) (cited on page 116).

[125] Lukas Herrmann, Christoph Schwab, and Jakob Zech. "Neural and GPC operator surrogates: construction and expression rate bounds". *preprint arXiv:2207.04950* (2022) (cited on pages 8, 57, 116).

[126] Jan S. Hesthaven and Stefano Ubbiali. "Non-intrusive reduced order modeling of nonlinear problems using neural networks". *Journal of Computational Physics* 363 (2018), pp. 55–78 (cited on pages 22, 56).

[127] Michael Hinze, René Pinnau, Michael Ulbrich, and Stefan Ulbrich. *Optimization with PDE constraints*. Vol. 23. Springer Science & Business Media, 2008 (cited on pages 4, 24).

[128] Siegfried Hörmann and Łukasz Kidziński. "A note on estimation in Hilbertian linear models". *Scandinavian Journal of Statistics* 42.1 (2015), pp. 43–62 (cited on pages 75, 84).

[129] Kurt Hornik. "Approximation capabilities of multilayer feedforward networks". *Neural Networks* 4.2 (1991), pp. 251–257 (cited on page 115).

[130] Daniel Zhengyu Huang, Nicholas H. Nelsen, and Margaret Trautner. "An operator learning perspective on parameter-to-observable maps". *preprint arXiv:2402.06031* (2024) (cited on pages 8, 9).

[131] Laura Hucker and Martin Wahl. "A note on the prediction error of principal component regression in high dimensions". *Theory of Probability and Mathematical Statistics* 109 (2023), pp. 37–53 (cited on pages 129, 262).

[132] Yuri Ingster and Natalia Stepanova. "Estimation and detection of functions from anisotropic Sobolev classes". *Electronic Journal of Statistics* 5 (2011), pp. 484–506 (cited on page 9).

[133] Yuri Ingster and Natalia Stepanova. "Estimation and detection of functions from weighted tensor product spaces". *Mathematical Methods of Statistics* 18 (2009), pp. 310–340 (cited on page 9).

[134] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in Neural Information Processing Systems*. 2018, pp. 8571–8580 (cited on pages 23, 33).

[135] Jikai Jin, Yiping Lu, Jose Blanchet, and Lexing Ying. "Minimax optimal kernel operator learning via multilevel training". In: *The Eleventh International Conference on Learning Representations*. 2022 (cited on pages 8, 57, 116, 151).

[136] Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren. "Operator-valued kernels for learning from functional response data". *The Journal of Machine Learning Research* 17.1 (2016), pp. 613–666 (cited on pages 9, 23, 29, 35, 186).

[137] Sebastian Kaltenbach, Paris Perdikaris, and Phaedon-Stelios Koutsourelakis. "Semi-supervised invertible neural operators for Bayesian inverse problems". *Computational Mechanics* 72.3 (2023), pp. 451–470 (cited on page 157).

[138] Aly-Khan Kassam and Lloyd N. Trefethen. "Fourth-order time-stepping for stiff PDEs". *SIAM Journal on Scientific Computing* 26.4 (2005), pp. 1214–1233 (cited on page 46).

[139] Rüdiger Kempf, Holger Wendland, and Christian Rieger. "Kernel-based Reconstructions for Parametric PDEs". In: *International Workshop on Meshfree Methods for Partial Differential Equations*. Springer. 2017, pp. 53–71 (cited on page 23).

[140] Yuehaw Khoo, Jianfeng Lu, and Lexing Ying. "Solving parametric PDE problems with artificial neural networks". *European Journal of Applied Mathematics* 32.3 (2021), pp. 421–435 (cited on page 22).

[141] Alexander Kiselev, Fedor Nazarov, and Roman Shterenberg. "Blow up and regularity for fractal Burgers equation". *Dynamics of Partial Differential Equations* 5.3 (2008), pp. 211–240 (cited on page 39).

[142] Ilja Klebanov, Ingmar Schuster, and T. J. Sullivan. "A rigorous theory of conditional mean embeddings". *SIAM Journal on Mathematics of Data Science* 2.3 (2020), pp. 583–606 (cited on pages 8, 84).

[143] Stefan Klus, Feliks Nüske, Sebastian Peitz, Jan-Hendrik Niemann, Cecilia Clementi, and Christof Schütte. "Data-driven approximation of the Koopman generator: Model reduction, system identification, and control". *Physica D: Nonlinear Phenomena* 406, 132416 (2020) (cited on page 83).

[144] Stefan Klus, Ingmar Schuster, and Krikamol Muandet. "Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces". *Journal of Nonlinear Science* 30.1 (2020), pp. 283–315 (cited on page 83).

[145] Bartek T. Knapik and Jean-Bernard Salomond. "A general approach to posterior contraction in nonparametric inverse problems". *Bernoulli* 24.3 (2018), pp. 2091–2121 (cited on pages 85, 99, 101, 110).

[146] Bartek T. Knapik, Botond T. Szabó, Aad W. Van Der Vaart, and J. Harry van Zanten. "Bayes procedures for adaptive inference in inverse problems for the white noise model". *Probability Theory and Related Fields* 164 (2016), pp. 771–813 (cited on page 135).

[147] Bartek T. Knapik, Aad W. van der Vaart, and J. Harry van Zanten. "Bayesian inverse problems with Gaussian priors". *Annals of Statistics* 39.5 (2011), pp. 2626–2657 (cited on pages 77–79, 85, 86, 90, 91, 99, 101, 111, 126, 127, 215, 216, 221, 222, 240, 259).

[148] Vladimir Koltchinskii. "Asymptotically efficient estimation of smooth functionals of covariance operators". *Journal of the European Mathematical Society* 23.3 (2021), pp. 765–843 (cited on page 157).

[149] Vladimir Koltchinskii. "Estimation of smooth functionals in high-dimensional models: bootstrap chains and Gaussian approximation". *Annals of Statistics* 50.4 (2022), pp. 2386–2415 (cited on page 116).

[150] Yury Korolev. "Two-layer neural networks with values in a Banach space". *SIAM Journal on Mathematical Analysis* 54.6 (2022), pp. 6358–6389 (cited on pages 8, 20, 22, 83).

[151] Vladimir Kostic, Pietro Novelli, Andreas Maurer, Carlo Ciliberto, Lorenzo Rosasco, and Massimiliano Pontil. "Learning dynamical systems via Koopman operator regression in reproducing kernel Hilbert spaces". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 4017–4031 (cited on page 8).

[152] Nikola B. Kovachki, Samuel Lanthaler, and Siddhartha Mishra. "On universal approximation and error bounds for Fourier neural operators". *Journal of Machine Learning Research* 22.290 (2021), pp. 1–76 (cited on pages 8, 57, 87, 116, 119, 122, 227).

[153] Nikola B. Kovachki, Samuel Lanthaler, and Andrew M. Stuart. "Operator learning: Algorithms and analysis". *preprint arXiv:2402.15715* (2024) (cited on pages 1, 8, 10).

[154] Nikola B. Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. "Neural operator: Learning maps between function spaces with applications to PDEs". *Journal of Machine Learning Research* 24.89 (2023), pp. 1–97 (cited on pages 8, 32, 83, 87, 115, 119, 122, 211).

[155] Carlos Kubrusly and Joao Zanni. "A note on compactness of tensor products". *Acta Mathematica Universitatis Comenianae* 84.1 (2015), pp. 59–62 (cited on page 224).

[156] Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. "FourCastNet: Accelerating global high-resolution weather forecasting using adaptive Fourier neural operators". In: *Proceedings of the Platform for Advanced Scientific Computing Conference*. 2023, pp. 1–11 (cited on page 115).

[157] Gitta Kutyniok, Philipp Petersen, Mones Raslan, and Reinhold Schneider. "A theoretical analysis of deep neural networks and parametric PDEs". *Constructive Approximation* 55.1 (2022), pp. 73–125 (cited on page 22).

[158] Samuel Lanthaler. "Operator learning with PCA-Net: upper and lower complexity bounds". *Journal of Machine Learning Research* 24.318 (2023), pp. 1–67 (cited on pages 8, 116, 151).

[159] Samuel Lanthaler, Zongyi Li, and Andrew M. Stuart. "The nonlocal neural operator: Universal approximation". *preprint arXiv:2304.13221* (2023) (cited on pages 8, 143).

[160] Samuel Lanthaler, Siddhartha Mishra, and George Em Karniadakis. "Error estimates for DeepONets: A deep learning framework in infinite dimensions". *Transactions of Mathematics and Its Applications* 6.1 (2022), pp. 1–141 (cited on pages 8, 57, 87, 99).

[161] Samuel Lanthaler, Roberto Molinaro, Patrik Hadorn, and Siddhartha Mishra. "Nonlinear reconstruction for operator learning of PDEs with discontinuities". In: *The Eleventh International Conference on Learning Representations*. 2022 (cited on page 116).

[162] Samuel Lanthaler and Nicholas H. Nelsen. "Error bounds for learning with vector-valued random features". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 71834–71861 (cited on pages 9, 49, 116, 265).

[163] Samuel Lanthaler and Andrew M. Stuart. "The parametric complexity of operator learning". *preprint arXiv:2306.15924* (2023) (cited on pages 8, 116, 151).

[164] Thomas Laurent and James Brecht. "Deep linear networks with arbitrary loss: All local minima are global". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2902–2907 (cited on page 116).

[165] Kookjin Lee and Kevin T. Carlberg. "Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders". *Journal of Computational Physics* 404, 108973 (2020) (cited on pages 22, 23, 40).

[166] Markku S. Lehtinen, Lassi Paivarinta, and Erkki Somersalo. "Linear inverse problems for generalised random variables". *Inverse Problems* 5.4, 599 (1989) (cited on page 91).

[167] Shibo Li, Xin Yu, Wei Xing, Mike Kirby, Akil Narayan, and Shandian Zhe. "Multi-resolution active learning of Fourier neural operators". *preprint arXiv:2309.16971* (2023) (cited on page 157).

[168] Yingzhou Li, Jianfeng Lu, and Anqi Mao. "Variational training of neural network approximations of solution maps for physical models". *Journal of Computational Physics* 409, 109338 (2020) (cited on pages 22, 23).

[169] Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. "Towards optimal Sobolev norm rates for the vector-valued regularized least-squares algorithm". *preprint arXiv:2312.07186* (2023) (cited on pages 116, 151).

[170] Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. "Towards a unified analysis of random Fourier features". *Journal of Machine Learning Research* 22.108 (2021), pp. 1–51 (cited on pages 57, 58, 61–64).

[171] Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. "Fourier neural operator with learned deformations for PDEs on general geometries". *Journal of Machine Learning Research* 24.388 (2023), pp. 1–26 (cited on pages 115, 117, 119, 144, 145).

[172] Zongyi Li, Nikola B. Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. "Fourier neural operator for parametric partial differential equations". *International Conference on Learning Representations* (2021) (cited on pages 8, 32, 48, 56, 115, 139).

[173] Zongyi Li, Nikola B. Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. "Neural operator: Graph kernel network for partial differential equations". *preprint arXiv:2003.03485* (2020) (cited on pages 7, 8, 20, 23, 48, 52, 115).

[174] Zongyi Li, Nikola B. Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Andrew M. Stuart, Kaushik Bhattacharya, and Anima Anandkumar. "Multipole graph neural operator for parametric partial differential equations". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 6755–6766 (cited on page 115).

[175] Heng Lian, Taeryon Choi, Jie Meng, and Seongil Jo. "Posterior convergence for Bayesian functional linear regression". *Journal of Multivariate Analysis* 150 (2016), pp. 27–41 (cited on pages 116, 135).

[176] Levi Lingsch, Mike Michelis, Sirani M. Perera, Robert K. Katzschmann, and Siddartha Mishra. "A structured matrix method for nonequispaced neural operators". *preprint arXiv:2305.19663* (2023) (cited on page 115).

[177] Burigede Liu, Nikola B. Kovachki, Zongyi Li, Kamyar Azizzadenesheli, Anima Anandkumar, Andrew M. Stuart, and Kaushik Bhattacharya. "A learning-based multiscale method and its application to inelastic impact problems". *Journal of the Mechanics and Physics of Solids* 158, 104668 (2022) (cited on page 117).

[178] Hao Liu, Haizhao Yang, Minshuo Chen, Tuo Zhao, and Wenjing Liao. "Deep nonparametric estimation of operators between infinite dimensional spaces". *Journal of Machine Learning Research* 25.24 (2024), pp. 1–67 (cited on pages 9, 116).

[179] Zichao Long, Yiping Lu, Xianzhong Ma, and Bin Dong. "PDE-Net: Learning PDEs from data". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 3208–3216 (cited on page 22).

[180] Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. "Deep network approximation for smooth functions". *SIAM Journal on Mathematical Analysis* 53.5 (2021), pp. 5465–5506 (cited on page 116).

[181] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. "Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators". *Nature Machine Intelligence* 3.3 (2021), pp. 218–229 (cited on pages 7, 20, 33, 56, 83, 87, 115).

[182] David G. Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997 (cited on page 35).

[183] Kjetil O. Lye, Siddhartha Mishra, and Roberto Molinaro. "A multilevel procedure for enhancing accuracy of machine learning algorithms". *European Journal of Applied Mathematics* 32.3 (2021), pp. 436–469 (cited on pages 117, 145).

[184] Kjetil O. Lye, Siddhartha Mishra, and Deep Ray. "Deep learning observables in computational fluid dynamics". *Journal of Computational Physics* 410, 109339 (2020) (cited on pages 117, 145).

[185] Kjetil O. Lye, Siddhartha Mishra, Deep Ray, and Praveen Chandrashekar. "Iterative surrogate model optimization (ISMO): An active learning algorithm for PDE constrained optimization with deep neural networks". *Computer Methods in Applied Mechanics and Engineering* 374, 113575 (2021) (cited on pages 117, 145).

[186] Avi Mandelbaum. "Linear estimators and measurable linear transformations on a Hilbert space". *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 65.3 (1984), pp. 385–397 (cited on pages 90, 91).

[187] Bertil Matérn. *Spatial variation*. Vol. 36. Springer Science & Business Media, 2013 (cited on page 39).

[188] Timothée Mathieu and Stanislav Minsker. "Excess risk bounds in robust empirical risk minimization". *Information and Inference: A Journal of the IMA* 10.4 (2021), pp. 1423–1490 (cited on page 103).

[189] Andreas Maurer and Massimiliano Pontil. "Concentration inequalities under sub-Gaussian and sub-exponential conditions". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 7588–7597 (cited on pages 189, 193).

[190] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration". *Applied and Computational Harmonic Analysis* 59 (2022), pp. 3–84 (cited on page 57).

[191] Hrushikesh Narhar Mhaskar and Nahmwoo Hahm. "Neural networks for functional approximation and system identification". *Neural Computation* 9.1 (1997), pp. 143–159 (cited on page 8).

[192] Charles A. Micchelli and Massimiliano Pontil. "On learning vector-valued functions". *Neural Computation* 17.1 (2005), pp. 177–204 (cited on pages 23, 27, 29, 34).

[193] Ha Quang Minh. "Operator-valued Bochner theorem, Fourier feature maps for operator-valued kernels, and vector-valued learning". *preprint arXiv:1608.05639* (2016) (cited on page 57).

[194] Siddhartha Mishra and T. Konstantin Rusch. "Enhancing accuracy of deep learning algorithms by training with low-discrepancy sequences". *SIAM Journal on Numerical Analysis* 59.3 (2021), pp. 1811–1834 (cited on pages 117, 145).

[195] Mattes Mollenhauer and Péter Koltai. "Nonparametric approximation of conditional expectation operators". *preprint arXiv:2012.12917* (2020) (cited on page 84).

[196] Mattes Mollenhauer, Nicole Mücke, and T. J. Sullivan. "Learning linear operators: Infinite-dimensional regression as a well-behaved non-compact inverse problem". *preprint arXiv:2211.08875* (2022) (cited on pages 8, 57, 116, 191).

[197] François Monard, Richard Nickl, and Gabriel P. Paternain. "Consistent inversion of noisy non-abelian X-ray transforms". *Communications on Pure and Applied Mathematics* 74.5 (2021), pp. 1045–1099 (cited on page 86).

[198] François Monard, Richard Nickl, and Gabriel P. Paternain. "Efficient nonparametric Bayesian inference for X-ray transforms". *Annals of Statistics* 47.2 (2019), pp. 1113–1147 (cited on page 85).

[199] Jennifer L. Mueller and Samuli Siltanen. *Linear and Nonlinear Inverse Problems with Practical Applications*. Computational Science & Engineering. SIAM, 2012 (cited on page 83).

[200] Radford M. Neal. "Priors for infinite networks". In: *Bayesian Learning for Neural Networks*. Springer, 1996, pp. 29–53 (cited on pages 19, 23, 33).

[201] Nicholas H Nelsen. "Operator-Valued Kernels". In: *ACM 204: Matrix Analysis*. Ed. by Joel A Tropp. Pasadena, CA: California Institute of Technology CMS Lecture Notes, 2022, pp. 286–297. DOI: 10.7907/m421-yb89 (cited on page 27).

[202] Nicholas H. Nelsen and Andrew M. Stuart. "Operator learning using random features: a tool for scientific computing". *SIAM Review* 66.3 (2024) (cited on page 56).

[203] Nicholas H. Nelsen and Andrew M. Stuart. "The random feature model for input-output maps between Banach spaces". *SIAM Journal on Scientific Computing* 43.5 (2021), A3212–A3243 (cited on pages 7, 8, 56, 59, 83, 115, 193, 211).

[204] Richard Nickl. *Bayesian non-linear statistical inverse problems*. EMS Press, 2023 (cited on page 9).

[205] Richard Nickl, Sara van de Geer, and Sven Wang. "Convergence rates for penalized least squares estimators in PDE constrained regression problems". *SIAM/ASA Journal on Uncertainty Quantification* 8.1 (2020), pp. 374–413 (cited on pages 9, 93).

[206] Thomas O'Leary-Roseberry, Xiaosong Du, Anirban Chaudhuri, Joaquim R. R. A. Martins, Karen Willcox, and Omar Ghattas. "Learning high-dimensional parametric maps via reduced basis adaptive residual networks". *Computer Methods in Applied Mechanics and Engineering* 402, 115730 (2022) (cited on pages 117, 145).

[207] Thomas O'Leary-Roseberry, Umberto Villa, Peng Chen, and Omar Ghattas. "Derivative-informed projected neural networks for high-dimensional parametric maps governed by PDEs". *Computer Methods in Applied Mechanics and Engineering* 388, 114199 (2022) (cited on pages 7, 20, 83, 87).

[208] Hidemitsu Ogawa. "An operator pseudo-inversion lemma". *SIAM Journal on Applied Mathematics* 48.6 (1988), pp. 1527–1531 (cited on page 194).

[209] Junier Oliva, William Neiswanger, Barnabás Póczos, Eric Xing, Hy Trac, Shirley Ho, and Jeff Schneider. "Fast function to function regression". In: *Artificial Intelligence and Statistics*. PMLR. 2015, pp. 717–725 (cited on page 9).

[210] Junier Oliva, Barnabás Póczos, and Jeff Schneider. "Distribution to distribution regression". In: *International Conference on Machine Learning*. PMLR. 2013, pp. 1049–1057 (cited on page 9).

[211] Joost A. A. Opschoor, Christoph Schwab, and Jakob Zech. "Deep learning in high dimension: ReLU network expression rates for Bayesian PDE inversion". *SAM Research Report* (2020) (cited on page 22).

[212] Houman Owhadi and Clint Scovel. "Separability of reproducing kernel spaces". *Proceedings of the American Mathematical Society* 145.5 (2017), pp. 2131–2138 (cited on page 193).

[213] Ravi G. Patel and Olivier Desjardins. "Nonlinear integro-differential operator regression with neural networks". *preprint arXiv:1810.08552* (2018) (cited on page 40).

[214] Ravi G. Patel, Nathaniel A. Trask, Mitchell A. Wood, and Eric C. Cyr. "A physics-informed operator regression framework for extracting data-driven continuum models". *Computer Methods in Applied Mechanics and Engineering* 373, 113500 (2021) (cited on pages 22, 40, 83, 115).

[215] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Anima Anandkumar. "FourCastNet: a global data-driven high-resolution weather model using adaptive Fourier neural operators". *preprint arXiv:2202.11214* (2022) (cited on page 1).

[216] Debdeep Pati, Anirban Bhattacharya, and Guang Cheng. "Optimal Bayesian estimation in random covariate design with a rescaled Gaussian process prior". *Journal of Machine Learning Research* 16.87 (2015), pp. 2837–2851 (cited on page 116).

[217] Grigoris Pavliotis and Andrew M. Stuart. *Multiscale methods: averaging and homogenization*. Springer Science & Business Media, 2008 (cited on page 148).

[218] Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. "Survey of multifidelity methods in uncertainty propagation, inference, and optimization". *SIAM Review* 60.3 (2018), pp. 550–591 (cited on page 20).

[219] Iosif F. Pinelis and Aleksandr Ivanovich Sakhanenko. "Remarks on inequalities for large deviation probabilities". *Theory of Probability & Its Applications* 30.1 (1986), pp. 143–148 (cited on pages 189–191, 243).

[220] Teresa Portone and Robert D. Moser. "Bayesian inference of an uncertain generalized diffusion operator". *SIAM/ASA Journal on Uncertainty Quantification* 10.1 (2022), pp. 151–178 (cited on page 82).

[221] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". *Advances in Neural Information Processing Systems* 20 (2007), pp. 1177–1184 (cited on pages 9, 19, 23, 25, 33, 55, 56).

[222] Ali Rahimi and Benjamin Recht. "Uniform approximation of functions with random bases". In: *2008 46th Annual Allerton Conference on Communication, Control, and Computing*. IEEE. 2008, pp. 555–561 (cited on pages 19, 23, 25, 33).

[223] Ali Rahimi and Benjamin Recht. "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning". *Advances in Neural Information Processing Systems* 21 (2008), pp. 1313–1320 (cited on pages 19, 23, 33, 55, 56, 58, 61–64, 67).

[224] Md Ashiqur Rahman, Manuel A. Florez, Anima Anandkumar, Zachary E. Ross, and Kamyar Azizzadenesheli. "Generative adversarial neural operators". *Transactions on Machine Learning Research* (2022) (cited on page 117).

[225] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations". *Journal of Computational Physics* 378 (2019), pp. 686–707 (cited on page 21).

[226] Vinod Raman, Unique Subedi, and Ambuj Tewari. "Online infinite-dimensional regression: learning linear operators". *preprint arXiv:2309.06548* (2023) (cited on page 157).

[227] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. 2nd ed. Springer Series in Statistics. Springer, New York, 2005 (cited on page 84).

[228] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Vol. 1. Springer, 2006 (cited on pages 11, 22, 141).

[229] Abhishake Rastogi, Gilles Blanchard, and Peter Mathé. "Convergence analysis of Tikhonov regularization for non-linear statistical inverse problems". *Electronic Journal of Statistics* 14.2 (2020), pp. 2798–2841 (cited on pages 83, 111).

[230] Kolyan Ray. "Bayesian inverse problems with non-conjugate priors". *Electronic Journal of Statistics* 7 (2013), pp. 2516–2549 (cited on pages 77, 85, 110, 128).

[231] Matthew Reimherr. "Functional regression with repeated eigenvalues". *Stat & Probability Letters* 107 (2015), pp. 62–70 (cited on pages 84, 116).

[232] Matthew Reimherr, Bharath Sriperumbudur, and Hyun Bin Kang. "Optimal function-on-scalar regression over complex domains". *Electronic Journal of Statistics* 17.1 (2023), pp. 156–197 (cited on page 151).

[233] R. Rico-Martinez, K. Krischer, I. G. Kevrekidis, M. C. Kube, and J. L. Hudson. "Discrete-vs. continuous-time nonlinear signal processing of Cu electrodissolution data". *Chemical Engineering Communications* 118.1 (1992), pp. 25–48 (cited on page 22).

[234] Fabrice Rossi and Brieuc Conan-Guez. "Functional multi-layer perceptron: a non-linear tool for functional data analysis". *Neural Networks* 18.1 (2005), pp. 45–60 (cited on page 23).

[235] Alessandro Rudi and Lorenzo Rosasco. "Generalization properties of learning with random features". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017 (cited on pages 57, 58, 61, 63, 64, 189).

[236] Lars Ruthotto and Eldad Haber. "Deep neural networks motivated by partial differential equations". *Journal of Mathematical Imaging and Vision* (2019), pp. 1–13 (cited on page 24).

[237] Milad Samaee, Nicholas H. Nelsen, Manikantam G. Gaddam, and Arvind Santhanakrishnan. "Diastolic vortex alterations with reducing left ventricular volume: an in vitro study". *Journal of Biomechanical Engineering* 142.12, 121006 (2020) (cited on page 157).

[238] Laura Scarabosio. "Deep neural network surrogates for nonsmooth quantities of interest in shape uncertainty quantification". *SIAM/ASA Journal on Uncertainty Quantification* 10.3 (2022), pp. 975–1011 (cited on page 152).

[239] Florian Schäfer and Houman Owhadi. "Sparse recovery of elliptic solvers from matrix-vector products". *SIAM Journal on Scientific Computing* 46.2 (2024), A998–A1025 (cited on pages 8, 57, 116).

[240] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2018 (cited on page 56).

[241] Christoph Schwab and Jakob Zech. "Deep learning in high dimension: Neural network expression rates for generalized polynomial chaos expansions in UQ". *Analysis and Applications* 17.1 (2019), pp. 19–55 (cited on pages 22, 57, 83).

[242] Zuowei Shen, Haizhao Yang, and Shijun Zhang. "Deep network approximation characterized by number of neurons". *Communications in Computational Physics* 28.5 (2020), pp. 1768–1811 (cited on page 116).

[243] Zhongjie Shi, Jun Fan, Linhao Song, Ding-Xuan Zhou, and Johan A. K. Suykens. "Nonlinear functional regression by functional deep neural network with kernel embedding". *preprint arXiv:2401.02890* (2024) (cited on page 117).

[244] Khemraj Shukla, Vivek Oommen, Ahmad Peyvan, Michael Penwarden, Nicholas Plewacki, Luis Bravo, Anindya Ghoshal, Robert M. Kirby, and George Em Karniadakis. "Deep neural operators as accurate surrogates for shape optimization". *Engineering Applications of Artificial Intelligence* 129, 107615 (2024) (cited on pages 117, 145).

[245] Justin Sirignano and Konstantinos Spiliopoulos. "DGM: A deep learning algorithm for solving partial differential equations". *Journal of Computational Physics* 375 (2018), pp. 1339–1364 (cited on page 21).

[246] Steve Smale and Ding-Xuan Zhou. "Shannon sampling II: Connections to learning theory". *Applied and Computational Harmonic Analysis* 19.3 (2005), pp. 285–302 (cited on pages 193, 195).

[247] Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. "Hilbert space embeddings of conditional distributions with applications to dynamical systems". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. Association for Computing Machinery, 2009, pp. 961–968 (cited on pages 8, 84).

[248] Pol D. Spanos and Roger Ghanem. "Stochastic finite element expansion for random media". *Journal of Engineering Mechanics* 115.5 (1989), pp. 1035–1053 (cited on page 22).

[249] Joseph L. Steger and Denny S. Chaussee. "Generation of body-fitted coordinates using hyperbolic partial differential equations". *SIAM Journal on Scientific and Statistical Computing* 1.4 (1980), pp. 431–437 (cited on pages 145, 146).

[250] Ingo Steinwart. "Convergence types and rates in generic Karhunen-Loève expansions with applications to sample path properties". *Potential Analysis* 51.3 (2019), pp. 361–395 (cited on page 88).

[251] George Stepaniants. "Learning partial differential equations in reproducing kernel Hilbert spaces". *Journal of Machine Learning Research* 24.86 (2023), pp. 1–72 (cited on pages 8, 57, 116).

[252] Ben Stevens and Tim Colonius. "FiniteNet: A fully convolutional LSTM network architecture for time-dependent partial differential equations". *preprint arXiv:2002.03014* (2020) (cited on page 22).

[253] Andrew M. Stuart. "Inverse problems: a Bayesian perspective". *Acta Numerica* 19 (2010), pp. 451–559 (cited on pages 2, 5, 88, 92, 93).

[254] Yitong Sun, Anna Gilbert, and Ambuj Tewari. "But how does it work in theory? Linear SVM with random features". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018 (cited on page 64).

[255] Yitong Sun, Anna Gilbert, and Ambuj Tewari. "On the approximation properties of random ReLU features". *preprint arXiv:1810.04374* (2019) (cited on pages 23, 33, 53).

[256] Taiji Suzuki and Shunta Akiyama. "Benefit of deep learning with non-convex noisy gradient descent: Provable excess risk bound and superiority to kernel methods". In: *International Conference on Learning Representations*. 2020 (cited on page 36).

[257] Puoya Tabaghi, Maarten V. de Hoop, and Ivan Dokmanić. "Learning Schatten–von Neumann operators". *preprint arXiv:1901.10076* (2019) (cited on pages 84, 103).

[258] Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. "PDEBench: An extensive benchmark for scientific machine learning". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 1596–1611 (cited on page 117).

[259] Prem Talwai, Ali Shameli, and David Simchi-Levi. "Sobolev norm learning rates for conditional mean embeddings". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 10422–10447 (cited on page 8).

[260] Ying Shi Teh, Swarnava Ghosh, and Kaushik Bhattacharya. "Machine-learned prediction of the electronic fields in a crystal". *Mechanics of Materials* 163, 104070 (2021) (cited on page 113).

[261] Alex Townsend and Lloyd N. Trefethen. "An extension of Chebfun to two dimensions". *SIAM Journal on Scientific Computing* 35.6 (2013), pp. C495–C518 (cited on page 5).

[262] Alex Townsend and Lloyd N. Trefethen. "Continuous analogues of matrix factorizations". *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 471.2173, 20140585 (2015) (cited on page 5).

[263] Mathias Trabs. "Bayesian inverse problems with unknown operators". *Inverse Problems* 34.8, 085001 (2018) (cited on pages 82, 85, 110).

[264] Nathaniel Trask, Ravi G. Patel, Ben J. Gross, and Paul J. Atzberger. "GMLS-Nets: A framework for learning from unstructured data". *preprint arXiv:1909.05371* (2019) (cited on page 22).

[265] Rohit K. Tripathy and Ilias Bilionis. "Deep UQ: Learning deep neural network surrogate models for high dimensional uncertainty quantification". *Journal of Computational Physics* 375 (2018), pp. 565–588 (cited on page 22).

[266] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018 (cited on pages 191, 217, 232, 263).

[267] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019 (cited on pages 102, 208, 217, 222, 244, 248).

[268] Daren Wang, Zifeng Zhao, Yi Yu, and Rebecca Willett. "Functional linear regression with mixed predictors". *Journal of Machine Learning Research* 23.266 (2022), pp. 1–94 (cited on page 84).

[269] Shanshan Wang, Wolfgang Jank, and Galit Shmueli. "Explaining and forecasting online auction prices and their dynamics using functional data analysis". *Journal of Business & Economic Statistics* 26.2 (2008), pp. 144–160 (cited on page 157).

[270] Holger Wendland. *Scattered data approximation*. Vol. 17. Cambridge University Press, 2004 (cited on pages 22, 23, 27).

[271] Christopher K. I. Williams. "Computing with infinite networks". In: *Advances in Neural Information Processing Systems*. 1997, pp. 295–301 (cited on pages 19, 23, 33).

[272] Nick Winovich, Karthik Ramani, and Guang Lin. "ConvPDE-UQ: Convolutional neural networks with quantified uncertainty for heterogeneous elliptic partial differential equations on varied domains". *Journal of Computational Physics* 394 (2019), pp. 263–279 (cited on page 22).

[273] Jin-Long Wu, Matthew E. Levine, Tapio Schneider, and Andrew M. Stuart. "Learning about structural errors in models of complex dynamical systems". *preprint arXiv:2401.00035* (2023) (cited on page 4).

[274] Kailiang Wu and Dongbin Xiu. "Data-driven deep learning of partial differential equations in modal space". *Journal of Computational Physics* 408, 109307 (2020) (cited on pages 7, 20, 46).

[275] Yan Yang, Angela F. Gao, Jorge C. Castellanos, Zachary E. Ross, Kamyar Azizzadenesheli, and Robert W. Clayton. "Seismic wave propagation and inversion with neural operators". *The Seismic Record* 1.3 (2021), pp. 126–134 (cited on page 1).

[276] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. "Functional data analysis for sparse longitudinal data". *Journal of the American Statistical Association* 100.470 (2005), pp. 577–590 (cited on page 157).

[277] Huaiqian You, Quinn Zhang, Colton J. Ross, Chung-Hao Lee, and Yue Yu. "Learning deep implicit Fourier neural operators (IFNOs) with applications to heterogeneous material modeling". *Computer Methods in Applied Mechanics and Engineering* 398, 115296 (2022) (cited on page 117).

[278] Ming Yuan and T. Tony Cai. "A reproducing kernel Hilbert space approach to functional linear regression". *Annals of Statistics* (2010), pp. 3412–3444 (cited on pages 116, 135).

[279] Fode Zhang, Weiping Zhang, Rui Li, and Heng Lian. "Faster convergence rate for functional linear regression in reproducing kernel Hilbert spaces". *Statistics* 54.1 (2020), pp. 167–181 (cited on page 135).

[280] Zezhong Zhang, Feng Bao, and Guannan Zhang. "Improving the expressive power of deep neural networks through integral activation transform". *preprint arXiv:2312.12578* (2023) (cited on page 117).

[281] Tingtao Zhou, Xuan Wan, Daniel Zhengyu Huang, Zongyi Li, Zhiwei Peng, Anima Anandkumar, John F. Brady, Paul W. Sternberg, and Chiara Daraio. "AI-aided geometric design of anti-infection catheters". *Science Advances* 10.1, eadj1741 (2024) (cited on page 115).

[282] Min Zhu, Handi Zhang, Anran Jiao, George Em Karniadakis, and Lu Lu. "Reliable extrapolation of deep neural operators informed by physics or sparse observations". *Computer Methods in Applied Mechanics and Engineering* 412, 116064 (2023) (cited on page 14).

[283] Yinhao Zhu and Nicholas Zabaras. "Bayesian deep convolutional encoder–decoder networks for surrogate modeling and uncertainty quantification". *Journal of Computational Physics* 366 (2018), pp. 415–447 (cited on page 22).

Appendix A

APPENDIX TO CHAPTER 2

This appendix collects the proofs of the results appearing in the main body of Chapter 2: Operator Learning With Function-Valued Random Features (Appendix A.1). In Appendix A.2, it also provides further insight into the uniqueness of the integral representation of operators in the reproducing kernel Hilbert space induced by random features.

A.1 Proofs of Results

We begin by proving the integral characterization of the operator RKHS.

Proof of Result 2.5. Fix $a \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then, we note that

$$k_\mu(\cdot, a)y = \int \langle \varphi(a;\theta), y\rangle_{\mathcal{Y}} \varphi(\cdot;\theta)\mu(d\theta) = \mathcal{A}\langle \varphi(a;\cdot), y\rangle_{\mathcal{Y}} \in \mathrm{Im}(\mathcal{A}) \tag{A.1}$$

since $\langle \varphi(a;\cdot), y\rangle_{\mathcal{Y}} \in L^2_\mu(\Theta;\mathbb{R})$ by the Cauchy–Schwarz inequality.

Now we show that $\mathrm{Im}(\mathcal{A})$ admits a reproducing property of the form (2.10). First, note that $\mathcal{A}$ can be viewed as a bijection between its coimage and image spaces, and we denote this bijection by

$$\tilde{\mathcal{A}}\colon\ \ker(\mathcal{A})^\perp \to \mathrm{Im}(\mathcal{A})\,. \tag{A.2}$$

For any F and $G \in \mathrm{Im}(\mathcal{A})$, define the candidate RKHS inner product $\langle \cdot, \cdot \rangle$ by

$$\langle F, G\rangle \coloneqq \big\langle \tilde{\mathcal{A}}^{-1}F, \tilde{\mathcal{A}}^{-1}G\big\rangle_{L^2_\mu(\Theta;\mathbb{R})}\,. \tag{A.3}$$

This is indeed a valid inner product since $\tilde{\mathcal{A}}$ is invertible. Note that for any $q \in \ker(\mathcal{A})$, it holds that

$$\begin{aligned}
\big\langle q, \langle \varphi(a;\cdot), y\rangle_{\mathcal{Y}}\big\rangle_{L^2_\mu(\Theta;\mathbb{R})} &= \int q(\theta)\langle \varphi(a;\theta), y\rangle_{\mathcal{Y}}\,\mu(d\theta) \\
&= \Big\langle \int q(\theta)\varphi(a;\theta)\mu(d\theta), y\Big\rangle_{\mathcal{Y}} \\
&= 0
\end{aligned}$$

so that $\langle \varphi(a;\cdot), y\rangle_{\mathcal{Y}} \in \ker(\mathcal{A})^{\perp}$. Then for any $F \in \mathrm{Im}(\mathcal{A})$, we compute

$$\begin{aligned}
\langle k_\mu(\cdot,a)y, F\rangle &= \left\langle \langle \varphi(a;\cdot), y\rangle_{\mathcal{Y}}, \tilde{\mathcal{A}}^{-1}F \right\rangle_{L^2_\mu(\Theta;\mathbb{R})} \\
&= \int \langle \varphi(a;\theta), y\rangle_{\mathcal{Y}} (\tilde{\mathcal{A}}^{-1}F)(\theta)\mu(d\theta) \\
&= \left\langle \int (\tilde{\mathcal{A}}^{-1}F)(\theta)\varphi(a;\theta)\mu(d\theta), y \right\rangle_{\mathcal{Y}} \\
&= \left\langle y, (\mathcal{A}\tilde{\mathcal{A}}^{-1}F)(a) \right\rangle_{\mathcal{Y}} \\
&= \langle y, F(a)\rangle_{\mathcal{Y}}\,,
\end{aligned}$$

which gives exactly (2.10) if our candidate inner product is defined to be the RKHS inner product. Since $F \in \mathrm{Im}(\mathcal{A})$ is arbitrary, this and (A.1) together imply that $\mathrm{Im}(\mathcal{A}) = \mathcal{H}_{k_\mu}$ is the RKHS induced by k_μ as shown in [70, 136]. □

The characterization of the finite-dimensional operator RKHS follows as a corollary.

Proof of Result 2.6. Since $L^2_{\mu^{(m)}}(\Theta;\mathbb{R})$ is isomorphic to $\mathbb{R}^m$, we can consider the map $\mathcal{A}\colon \mathbb{R}^m \to L^2_\nu(\mathcal{X};\mathcal{Y})$ defined in (2.14) and use Result 2.5 to conclude that

$$\mathcal{H}_{k^{(m)}} = \mathrm{Im}(\mathcal{A}) = \left\{ \frac{1}{m}\sum_{j=1}^{m} c_j \varphi(\cdot;\theta_j) \colon c \in \mathbb{R}^m \right\} = \mathrm{span}\{\varphi_j\}_{j=1}^m \tag{A.4}$$

since the $\{\varphi_j\}_{j=1}^m$ are assumed linearly independent. □

Finally, we prove that function-valued random feature ridge regression is equivalent to an operator-valued kernel method (equivalently, function-valued Gaussian process regression).

Proof of Result 2.8. Recall from Result 2.6 that the RKHS $\mathcal{H}_{k^{(m)}}$ comprises the linear span of the $\{\varphi_j \coloneqq \varphi(\cdot;\theta_j)\}_{j=1}^m$. Hence $\varphi_j \in \mathcal{H}_{k^{(m)}}$, and note that by the reproducing kernel property (2.10), for any $F \in \mathcal{H}_{k^{(m)}}$, $a \in \mathcal{X}$, and $y \in \mathcal{Y}$,

$$\begin{aligned}
\langle y, F(a)\rangle_{\mathcal{Y}} &= \left\langle k^{(m)}(\cdot,a)y, F \right\rangle_{\mathcal{H}_{k^{(m)}}} \\
&= \frac{1}{m}\sum_{j=1}^{m} \langle \varphi_j(a), y\rangle_{\mathcal{Y}} \langle \varphi_j, F\rangle_{\mathcal{H}_{k^{(m)}}} \\
&= \left\langle y, \frac{1}{m}\sum_{j=1}^{m} \langle \varphi_j, F\rangle_{\mathcal{H}_{k^{(m)}}} \varphi_j(a) \right\rangle_{\mathcal{Y}}.
\end{aligned}$$

Since this is true for all $y \in \mathcal{Y}$, we deduce that

$$F = \frac{1}{m} \sum_{j=1}^{m} \alpha_j \varphi_j \,, \quad \text{where} \quad \alpha_j = \langle \varphi_j, F \rangle_{\mathcal{H}_{k^{(m)}}} \,. \tag{A.5}$$

As the $\{\varphi_j\}_{j=1}^m$ are assumed linearly independent, we deduce that the representation (A.5) is unique.

Finally, we calculate the RKHS norm of any such F in terms of α:

$$\begin{aligned} \|F\|^2_{\mathcal{H}_{k^{(m)}}} = \langle F, F \rangle_{\mathcal{H}_{k^{(m)}}} &= \left\langle \frac{1}{m} \sum_{j=1}^{m} \alpha_j \varphi_j, F \right\rangle_{\mathcal{H}_{k^{(m)}}} \\ &= \frac{1}{m} \sum_{j=1}^{m} \alpha_j \langle \varphi_j, F \rangle_{\mathcal{H}_{k^{(m)}}} \\ &= \frac{1}{m} \sum_{j=1}^{m} \alpha_j^2 \,. \end{aligned}$$

Substituting this into (2.24), we obtain the desired equivalence with (2.23). □

A.2 Further Remarks on the Integral Representation of RKHS

We recall the linear operator $\mathcal{A}$ (2.14) from Section 2.2.3. In this appendix, we clarify the meaning of Equation (2.12) and show that $\mathcal{A}$ is a square root of T_{k_μ}. Similar discussion is provided by Bach [17, Sec. 2] for the special case $\mathcal{Y} = \mathbb{R}$.

By the assumption $\varphi \in L^2_{\nu \times \mu}(\mathcal{X} \times \Theta; \mathcal{Y})$ and Cauchy–Schwarz inequality, it holds that

$$\mathcal{A} \in \mathcal{L}\big(L^2_\mu(\Theta; \mathbb{R}); L^2_\nu(\mathcal{X}; \mathcal{Y})\big) \,. \tag{A.6}$$

Now let $F \in \mathrm{Im}(\mathcal{A}) = \mathcal{H}_{k_\mu}$. We have $F = \mathcal{A}c$ for some $c \in L^2_\mu$. But since $\ker(\mathcal{A})$ is closed, $L^2_\mu = \ker(\mathcal{A}) \oplus \ker(\mathcal{A})^\perp$. Hence, there exist unique $q_F \in \ker(\mathcal{A})$ and $c_F \in \ker(\mathcal{A})^\perp$ such that $c = q_F + c_F$. Using the notation in Equation (A.2), we have $c_F = \tilde{\mathcal{A}}^{-1} F$ by definition of $\tilde{\mathcal{A}}$. The reproducing property in the proof A.1 produced the representation $F = \mathcal{A}c_F$; in fact, the similar calculation leading to (2.12) in Section 2.2.3 also identified the unique c_F, there defined formally by $\theta \mapsto c_F(\theta) = \langle \varphi(\cdot; \theta), F \rangle_{\mathcal{H}_{k_\mu}}$. Indeed,

$$\begin{aligned} \langle c_F, q \rangle_{L^2_\mu(\Theta;\mathbb{R})} &= \int \langle \varphi(\cdot; \theta), F \rangle_{\mathcal{H}_{k_\mu}} q(\theta) \mu(d\theta) \\ &= \left\langle \int q(\theta) \varphi(\cdot; \theta) \mu(d\theta), F \right\rangle_{\mathcal{H}_{k_\mu}} \\ &= 0 \end{aligned}$$

for any $q \in \ker(\mathcal{A})$. Hence $c_F \in \ker(\mathcal{A})^\perp$, and we interpret (2.12) as formal notation for the unique element $\tilde{\mathcal{A}}^{-1}F \in \ker(\mathcal{A})^\perp$. Using the formula (A.3) and orthogonality, we also obtain the following useful characterization of the RKHS norm:

$$\|F\|^2_{\mathcal{H}_{k_\mu}} = \big\|\tilde{\mathcal{A}}^{-1}F\big\|^2_{L^2_\mu} = \|c_F\|^2_{L^2_\mu} = \min_{c\in\mathcal{C}_F}\|c\|^2_{L^2_\mu}\,, \tag{A.7}$$

where $\mathcal{C}_F \coloneqq \{c \in L^2_\mu(\Theta;\mathbb{R})\colon \mathcal{A}c = F\}$.

Finally, we show that $\mathcal{A}\mathcal{A}^* = T_{k_\mu}$. This means that the RKHS is equal to the image of two different square roots of integral operator T_{k_μ}, namely, $\mathcal{H}_{k_\mu} = \mathrm{Im}(T^{1/2}_{k_\mu}) = \mathrm{Im}(\mathcal{A})$. First, for any $F \in L^2_\nu(\mathcal{X};\mathcal{Y})$ and $c \in L^2_\mu(\Theta;\mathbb{R})$, it holds that

$$\begin{aligned}
\langle F, \mathcal{A}c\rangle_{L^2_\nu} &= \Big\langle F, \int c(\theta)\varphi(\cdot;\theta)\mu(d\theta)\Big\rangle_{L^2_\nu} \\
&= \int c(\theta)\langle F, \varphi(\cdot;\theta)\rangle_{L^2_\nu}\,\mu(d\theta) \\
&= \Big\langle \int \langle F(a'), \varphi(a';\cdot)\rangle_{\mathcal{Y}}\,\nu(da'), c\Big\rangle_{L^2_\mu}
\end{aligned}$$

by the Fubini–Tonelli theorem. So, we deduce that the adjoint of $\mathcal{A}$ is

$$\begin{aligned}
\mathcal{A}^*\colon\ L^2_\nu(\mathcal{X};\mathcal{Y}) &\to L^2_\mu(\Theta;\mathbb{R}) \\
F &\mapsto \mathcal{A}^*F \coloneqq \int \langle F(a'), \varphi(a';\cdot)\rangle_{\mathcal{Y}}\,\nu(da')\,,
\end{aligned} \tag{A.8}$$

which is bounded because $\mathcal{A}$ is bounded. For any $F \in L^2_\nu(\mathcal{X};\mathcal{Y})$, we compute

$$\begin{aligned}
\mathcal{A}\mathcal{A}^*F &= \int_\Theta (\mathcal{A}^*F)(\theta)\varphi(\cdot;\theta)\mu(d\theta) \\
&= \int_\Theta \int_{\mathcal{X}} \langle F(a'), \varphi(a';\theta)\rangle_{\mathcal{Y}}\,\varphi(\cdot;\theta)\nu(da')\mu(d\theta) \\
&= \int_{\mathcal{X}} \bigg(\int_\Theta \varphi(\cdot;\theta)\otimes\varphi(a';\theta)\mu(d\theta)\bigg)F(a')\nu(da') \\
&= T_{k_\mu}F\,,
\end{aligned}$$

again by Fubini–Tonelli, as desired. This concludes the appendix.

Appendix B

APPENDIX TO CHAPTER 3

This appendix is the companion to Chapter 3: Error Bounds for Function-Valued Random Features and is organized as follows. Appendix B.1 collects several useful concentration inequalities and facts about subexponential random variables. Appendix B.2 provides a reproducing kernel Hilbert space approach to misspecification error in regression problems. Appendix B.3 collects the proofs of the mains results in Chapter 3, while Appendix B.4 gives the detailed proofs of all the required technical results. Finally, Appendix B.5 describes the numerical experiment from Section 3.5 in more detail.

B.1 Concentration of Measure

In this appendix, we recall two classical results from [219] that estimate the difference between empirical averages and true averages of random vectors taking values in Banach spaces. These are then specialized to the setting of subexponential random variables, which play a major role in this chapter. To set the notation, we use Pr to denote probability with respect to the underlying probability space.

The first result is a vector-valued Bernstein concentration inequality with various applications to problems posed in infinite-dimensional Hilbert spaces [52, 189, 235]. It is used throughout this work.

Theorem B.1 (vector-valued Bernstein inequality in Hilbert space)**.** *Let Z be an H-valued random variable, where $(H, \langle\cdot,\cdot\rangle, \|\cdot\|)$ is a separable Hilbert space. Suppose there exist positive numbers $b > 0$ and $\sigma > 0$ such that*

$$\mathbb{E}\|Z - \mathbb{E}\,Z\|^p \leq \frac{1}{2}p!\sigma^2 b^{p-2} \quad \text{for all} \quad p \geq 2\,. \tag{B.1}$$

For any $\delta \in (0,1)$ and $N \in \mathbb{N}$, denoting by $\{Z_n\}_{n=1}^N$ a sequence of N i.i.d. copies of Z, it holds that

$$\Pr\left\{\left\|\frac{1}{N}\sum_{n=1}^N Z_n - \mathbb{E}\,Z\right\| \leq \frac{2b\log(2/\delta)}{N} + \sqrt{\frac{2\sigma^2\log(2/\delta)}{N}}\right\} \geq 1-\delta\,. \tag{B.2}$$

Proof. The result is a direct consequence of [219, Corollary 1, p. 144] in the i.i.d. Hilbert space setting. In this case, it holds for any $t > 0$ that

$$\Pr\{\|S_N - \mathbb{E}\, S_N\| \geq t\} \leq 2\exp\left(-\frac{N^2 t^2}{2N\sigma^2 + 2Ntb}\right) = 2\exp\left(-\frac{Nt^2}{2\sigma^2 + 2bt}\right),$$

where $S_n \coloneqq \frac{1}{N}\sum_{n=1}^{N} Z_n$. Setting the right-hand side equal to δ, solving a quadratic equation for $t = t(\delta)$, and using the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ to bound $t(\delta)$ from above leads to (B.2). □

The most common use case of Bernstein's inequality is in the following bounded setting.

Lemma B.2. *Let Z be a (potentially) uncentered random variable such that*

$$\|Z\| \leq c \quad \textit{almost surely} \qquad \textit{and} \qquad \mathbb{E}\|Z - \mathbb{E}\, Z\|^2 \leq v^2 \tag{B.3}$$

for some $c > 0$ and $v > 0$. Then Z satisfies Bernstein's moment condition (B.1) with $b = 2c$ and $\sigma = v$. If $\mathbb{E}\, Z = 0$, then taking $b = c$ suffices.

Proof. It holds that $\|Z - \mathbb{E}\, Z\| \leq \|Z\| + \mathbb{E}\|Z\| \leq 2c$ almost surely. We compute

$$\begin{aligned}\mathbb{E}\|Z - \mathbb{E}\, Z\|^p &= \mathbb{E}[\|Z - \mathbb{E}\, Z\|^2 \|Z - \mathbb{E}\, Z\|^{p-2}] \leq \mathbb{E}\|Z - \mathbb{E}\, Z\|^2 (2c)^{p-2} \\ &\leq v^2 (2c)^{p-2} \leq \frac{1}{2} p! v^2 (2c)^{p-2}\end{aligned}$$

because $1 \leq p!/2$ for all $p \geq 2$. The centered improvement is trivial. □

The second classic result we present is a one-sided Bernstein-type tail bound in a general Banach space. We invoke this theorem to control the tails of suprema of empirical processes.

Theorem B.3 (vector-valued Bernstein inequality in Banach space)**.** *Let Z be a $\mathcal{Z}$-valued random variable, where $(\mathcal{Z}, \|\cdot\|)$ is a separable Banach space. Suppose there exist positive numbers $b > 0$ and $\sigma > 0$ such that*

$$\mathbb{E}\|Z - \mathbb{E}\, Z\|^p \leq \frac{1}{2} p! \sigma^2 b^{p-2} \quad \textit{for all} \quad p \geq 2\,. \tag{B.4}$$

For any $\delta \in (0,1)$ and $N \in \mathbb{N}$, denoting by $\{Z_n\}_{n=1}^N$ a sequence of N i.i.d. copies of Z, it holds that

$$\Pr\left\{\left\|\frac{1}{N}\sum_{n=1}^{N} Z_n\right\| - \mathbb{E}\left\|\frac{1}{N}\sum_{n=1}^{N} Z_n\right\| \leq \frac{2b\log(1/\delta)}{N} + \sqrt{\frac{2\sigma^2 \log(1/\delta)}{N}}\right\} \geq 1 - \delta\,. \tag{B.5}$$

Proof. The assertion is [219, Corollary 1, p. 144] in the i.i.d. Banach space setting (hence only convergence of norms in (B.5) instead of strong convergence). It is proved the same way as Theorem B.1. □

In Theorem B.3, a random variable Z satisfying the *Bernstein moment condition* (B.4) is subexponential in the sense that $\|Z - \mathbb{E}\, Z\|$ is subexponential on $\mathbb{R}$, i.e., exhibits exponential tail decay. Recall that for a real-valued random variable Z, its subexponential norm may be defined by

$$\|Z\|_{\psi_1} \coloneqq \sup_{p \in [1,\infty)} \frac{(\mathbb{E}|Z|^p)^{1/p}}{p}\,. \tag{B.6}$$

See [266, Section 2.7] for equivalent definitions. We say that Z is subexponential if its subexponential norm is finite. Following [196, Section 4.3, pp. 19–20], for a random variable Z with values in a Banach space $(\mathcal{Z}, \|\cdot\|_{\mathcal{Z}})$ we define

$$\|Z\|_{\psi_1(\mathcal{Z})} \coloneqq \big\|\|Z\|_{\mathcal{Z}}\big\|_{\psi_1} = \sup_{p \in [1,\infty)} \frac{(\mathbb{E}\|Z\|_{\mathcal{Z}}^p)^{1/p}}{p} \tag{B.7}$$

as its subexponential norm. It is known that a random variable has finite subexponential norm if and only if it satisfies the Bernstein moment condition (B.4) [see, e.g., 196, Appendix A.2]. Next, we give explicit constants in the Bernstein moment condition that depend on the subexponential norm.

Proposition B.4 (subexponential implies Bernstein moment condition)**.** *Let Z be a $(\mathcal{Z}, \|\cdot\|)$-valued subexponential random variable, that is, $\|Z\|_{\psi_1(\mathcal{Z})} < \infty$. Then Z satisfies*

$$\mathbb{E}\|Z - \mathbb{E}\, Z\|^p \le \frac{1}{2} p! \sigma^2 b^{p-2} \quad \textit{for all} \quad p \ge 2\,, \tag{B.8}$$

where

$$\sigma^2 \coloneqq 4e\sqrt{\mathbb{E}\|Z - \mathbb{E}\, Z\|^2}\|Z\|_{\psi_1(\mathcal{Z})} \quad \textit{and} \quad b \coloneqq 4e\|Z\|_{\psi_1(\mathcal{Z})}\,. \tag{B.9}$$

Proof. By the Cauchy–Schwarz inequality,

$$\mathbb{E}\|Z - \mathbb{E}\, Z\|^p = \mathbb{E}[\|Z - \mathbb{E}\, Z\|\|Z - \mathbb{E}\, Z\|^{p-1}] \le \|Z - \mathbb{E}\, Z\|_{L^2_{\mathbb{P}}} (\mathbb{E}\|Z - \mathbb{E}\, Z\|^{2p-2})^{1/2}\,.$$

The inequality $|a+b|^q \le 2^{q-1}(|a|^q + |b|^q)$, Jensen's inequality, and $\|Z\|_{\psi_1(\mathcal{Z})} < \infty$ show that

$$\mathbb{E}\|Z - \mathbb{E}\, Z\|^{2p-2} \le 2^{2p-2}\, \mathbb{E}\|Z\|^{2p-2} \le 2^{2p-2}(2p-2)^{2p-2}\|Z\|^{2p-2}_{\psi_1(\mathcal{Z})}\,.$$

Next, Stirling's approximation $(q/e)^q \le q!$ and $q! = q(q-1)! \ge 2(q-1)!$ for $q \ge 2$ yields

$$\begin{aligned}\mathbb{E}\|Z - \mathbb{E}\,Z\|^p &\le 2^{2p-2}\|Z - \mathbb{E}\,Z\|_{L^2_{\mathbb{P}}}^2 (p-1)^{p-1}\|Z\|_{\psi_1(\mathcal{Z})}^{p-1} \\ &\le (p!/2)\|Z - \mathbb{E}\,Z\|_{L^2_{\mathbb{P}}}^2 2^{2p-2} e^{p-1}\|Z\|_{\psi_1(\mathcal{Z})}^{p-1}.\end{aligned}$$

Rearranging the exponents to fit the Bernstein moment condition form completes the proof. □

This leads to the following corollary, which is useful in the setting that the variance $\mathbb{E}\|Z - \mathbb{E}\,Z\|_{\mathcal{Z}}^2 = \|Z - \mathbb{E}\,Z\|_{L^2_{\mathbb{P}}(\Omega;\mathcal{Z})}^2$ of random variable Z is not small or hard to compute.

Corollary B.5 (subexponential tail bound in Banach space)**.** *Fix $N \in \mathbb{N}$. Let $\{Z_n\}_{n=1}^N$ be i.i.d. random variables with values in a separable Banach space $(\mathcal{Z}, \|\cdot\|)$. Suppose that $\|Z_1\|_{\psi_1(\mathcal{Z})} < \infty$. Let $S_N := \frac{1}{N}\sum_{n=1}^N Z_n$. Fix $\delta \in (0,1)$. With probability at least $1-\delta$, it holds that*

$$\begin{aligned}\big\|S_N\big\| - \mathbb{E}\big\|S_N\big\| \le{}& \frac{8e\|Z_1\|_{\psi_1(\mathcal{Z})}\log(1/\delta)}{N} \\ &+ \sqrt{\frac{8e\|Z_1 - \mathbb{E}\,Z_1\|_{L^2_{\mathbb{P}}(\Omega;\mathcal{Z})}\|Z_1\|_{\psi_1(\mathcal{Z})}\log(1/\delta)}{N}}.\end{aligned} \tag{B.10}$$

In particular, if $N \ge \log(1/\delta)$, then with probability at least $1-\delta$ it holds that

$$\big\|S_N\big\| - \mathbb{E}\big\|S_N\big\| \le \sqrt{\frac{64e^3\|Z_1\|_{\psi_1(\mathcal{Z})}^2\log(1/\delta)}{N}}. \tag{B.11}$$

Proof. Apply Proposition B.4 to Theorem B.3 to obtain the first assertion. For the second assertion, first note that $\mathbb{E}\|Z_1 - \mathbb{E}\,Z_1\|^2 \le 4\,\mathbb{E}\|Z_1\|^2$ (by triangle inequality and using $(a+b)^2 \le 2(a^2+b^2)$) and $\|Z_1\|_{\psi_1(\mathcal{Z})}^2 \ge \mathbb{E}\|Z_1\|^2/4$ (by (B.7)). Since $N \ge \log(1/\delta)$, we have $\log(1/\delta)/N \le \sqrt{\log(1/\delta)/N}$. Combining these facts, it follows that the right-hand side of (B.10) is bounded above by

$$\begin{aligned}&\sqrt{\frac{64e^2\|Z_1\|_{\psi_1(\mathcal{Z})}^2\log(1/\delta)}{N}} + \sqrt{\frac{32e\|Z_1\|_{\psi_1(\mathcal{Z})}^2\log(1/\delta)}{N}} \\ &\qquad\le \sqrt{\frac{64(2e^2+e)\|Z_1\|_{\psi_1(\mathcal{Z})}^2\log(1/\delta)}{N}}.\end{aligned}$$

We used $\sqrt{a}+\sqrt{b} \le \sqrt{2(a+b)}$ on the right. Noting that $2e^2+e \le e^3$ completes the proof. □

Comparing this result to a similar result [189, Proposition 7(i), pp. 4–5, the i.i.d. case], we note that (B.10) in Corollary B.5 is sharper in the sense that the constant in the $N^{-1/2}$ term depends on the variance of the summands instead of just its subexponential norm (which can be much larger than the variance).

B.2 Misspecification Error With RKHS Methods

Let $K\colon \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ be an operator-valued kernel [203] corresponding to a separable[1] RKHS $\mathcal{H}$ of functions mapping $\mathcal{X}$ to $\mathcal{Y}$. Here, $\mathcal{L}(\mathcal{Y})$ denotes the set of bounded linear operators from $\mathcal{Y}$ into itself. In this appendix, we present a general analysis based on [246] for the approximation of elements in $L^2_\nu(\mathcal{X};\mathcal{Y})$ by elements in the RKHS $\mathcal{H}$. This problem is relevant to our learning theory framework whenever the true data-generating map does not belong to the RKHS (3.2) associated to the given random feature pair (φ, μ). Specializing to this setting, suppose that (φ, μ) satisfy Assumption 3.3. Let $K\colon (u, u') \mapsto \mathbb{E}_{\theta\sim\mu}[\varphi(u;\theta) \otimes_{\mathcal{Y}} \varphi(u';\theta)]$ be the corresponding limit random feature kernel. It holds that $K(u,u)$ is trace-class for each $u \in \mathcal{X}$ ν-almost surely because

$$\operatorname{tr}(K(u,u)) = \mathbb{E}_\theta \operatorname{tr}(\varphi(u;\theta) \otimes_{\mathcal{Y}} \varphi(u;\theta)) = \mathbb{E}_\theta \|\varphi(u;\theta)\|^2_{\mathcal{Y}} \leq \|\varphi\|^2_{L^\infty} < \infty \tag{B.12}$$

by the Fubini–Tonelli theorem. It follows from [55, Proposition 4.8, p. 394] that $\mathcal{H}$ is compactly embedded into $L^2_\nu(\mathcal{X};\mathcal{Y})$ because

$$\mathbb{E}_{u\sim\nu}\|K(u,u)\|_{\mathcal{L}(\mathcal{Y})} \leq \mathbb{E}_{u\sim\nu} \operatorname{tr}(K(u,u)) \leq \|\varphi\|^2_{L^\infty} < \infty\,. \tag{B.13}$$

We denote the canonical embedding by $\iota\colon \mathcal{H} \hookrightarrow L^2_\nu(\mathcal{X};\mathcal{Y})$. Now let $\mathcal{G} \in L^2_\nu(\mathcal{X};\mathcal{Y})$ be an arbitrary operator. We consider an approximation $\mathcal{G}_\vartheta$ to $\mathcal{G}$ defined by

$$\mathcal{G}_\vartheta \coloneqq \operatorname*{arg\,min}_{F\in\mathcal{H}} \Big\{ \|\mathcal{G} - \iota F\|^2_{L^2_\nu} + \vartheta \|F\|^2_{\mathcal{H}} \Big\}\,. \tag{B.14}$$

This operator has a simple representation.

Lemma B.6. *There exists a unique solution $\mathcal{G}_\vartheta$ to (B.14) given by*

$$\mathcal{G}_\vartheta = (\iota^*\iota + \vartheta \operatorname{Id})^{-1} \iota^* \mathcal{G}\,. \tag{B.15}$$

[1] The *assumption* of separability of the RKHS could be removed if additional conditions are placed on its kernel that would *imply* separability, the most relevant being continuity-type assumptions. In the case $\mathcal{Y} = \mathbb{R}$, it is known that existence of a Borel measurable feature map for the kernel suffices for separability of its RKHS [212], which is much weaker than continuity. However, we are unaware of similar results for general $\mathcal{Y}$.

Proof. The result is a consequence of convex optimization on Hilbert spaces and the fact that ι is a bounded linear operator. □

The adjoint of the inclusion map, $\iota^*\colon L^2_\nu(\mathcal{X};\mathcal{Y}) \to \mathcal{H}$, is given by the vector-valued reproducing kernel property as the (Bochner) integral operator

$$F \mapsto \iota^* F = \mathbb{E}_{u\sim\nu}[K(\,\cdot\,,u)F(u)]\,. \tag{B.16}$$

Since ι is compact, so is $\mathcal{K} \coloneqq \iota\iota^* \in \mathcal{L}(L^2_\nu(\mathcal{X};\mathcal{Y}))$. The action of $\mathcal{K}$ is the same as that of the integral operator ι^* above. Since $\mathcal{K}$ is also symmetric, the spectral theorem yields the operator norm convergent expansion $\mathcal{K} = \sum_j \lambda_j e_j \otimes_{L^2(\nu)} e_j$. The sequence $\{\lambda_j\} \subset \mathbb{R}_{\geq 0}$ is a nonincreasing rearrangement of the eigenvalues of $\mathcal{K}$ and $\{e_j\}$ is its corresponding eigenbasis.

Now define the regularized RKHS approximation error

$$\mathscr{A}_{\mathcal{G}}(\vartheta) \coloneqq \inf_{F\in\mathcal{H}} \Big\{ \|\mathcal{G} - \iota F\|^2_{L^2_\nu} + \vartheta \|F\|^2_{\mathcal{H}} \Big\} \tag{B.17}$$

which is parametrized by $\mathcal{G} \in L^2_\nu(\mathcal{X};\mathcal{Y})$. We have the following convergence result for this error.

Lemma B.7 (convergence of regularized RKHS approximation error)**.** *Suppose that $\mathcal{G}$ is in the L^2_ν-closure of $\mathcal{H}$. Under the prevailing assumptions of this appendix, it holds that*

$$\lim_{\vartheta\to 0} \mathscr{A}_{\mathcal{G}}(\vartheta) = 0\,. \tag{B.18}$$

Proof. By Lemma B.6 and the Woodbury identity [208, Theorem 1], it holds that

$$\iota\mathcal{G}_\vartheta = \iota(\iota^*\iota \ + \vartheta \operatorname{Id}_{\mathcal{H}})^{-1}\iota^*\mathcal{G} = \mathcal{K}(\mathcal{K} + \vartheta \operatorname{Id}_{L^2_\nu})^{-1}\mathcal{G} = (\mathcal{K} + \vartheta \operatorname{Id}_{L^2_\nu})^{-1}\mathcal{K}\mathcal{G}\,.$$

The second equality holds by simultaneous diagonalization. Writing $\mathcal{G}$ in the eigenbasis of $\mathcal{K}$ yields

$$\mathcal{G} - \iota\mathcal{G}_\vartheta = \big[\operatorname{Id}_{L^2_\nu} - (\mathcal{K} + \vartheta \operatorname{Id}_{L^2_\nu})^{-1}\mathcal{K}\big]\mathcal{G} = \sum_{j\in\mathbb{N}} \frac{\vartheta}{\lambda_j + \vartheta} \langle e_j, \mathcal{G}\rangle_{L^2_\nu}\, e_j\,.$$

Similarly, using the norm isometry between L^2_ν and the RKHS [see, e.g., 55, pp. 403–404] we obtain

$$\|\mathcal{G}_\vartheta\|^2_{\mathcal{H}} = \|\mathcal{K}^{-1/2}\iota\mathcal{G}_\vartheta\|^2_{L^2_\nu} = \|\mathcal{K}^{1/2}(\mathcal{K}+\vartheta \operatorname{Id}_{L^2_\nu})^{-1}\mathcal{G}\|^2_{L^2_\nu} = \sum_{j\in\mathbb{N}} \left(\frac{\sqrt{\lambda_j}}{\lambda_j + \vartheta}\right)^2 \langle e_j, \mathcal{G}\rangle^2_{L^2_\nu}\,.$$

Since the infimum in (B.17) is attained at $\mathcal{G}_\vartheta$ (B.14), we deduce that

$$\begin{aligned}\mathscr{A}_{\mathcal{G}}(\vartheta) &= \|\mathcal{G} - \iota\mathcal{G}_\vartheta\|_{L^2_\nu}^2 + \vartheta\|\mathcal{G}_\vartheta\|_{\mathcal{H}}^2 \\ &= \sum_{j\in\mathbb{N}} \frac{\vartheta^2}{(\lambda_j+\vartheta)^2}\langle e_j, \mathcal{G}\rangle_{L^2_\nu}^2 + \sum_{j\in\mathbb{N}} \frac{\vartheta\lambda_j}{(\lambda_j+\vartheta)^2}\langle e_j, \mathcal{G}\rangle_{L^2_\nu}^2 \\ &= \sum_{j\in\mathbb{N}} \Big(\frac{\vartheta}{\lambda_j+\vartheta}\Big)\langle e_j, \mathcal{G}\rangle_{L^2_\nu}^2 .\end{aligned}$$

Using $\vartheta/(\lambda_j+\vartheta) \leq 1$ for each $j \in \mathbb{N}$ and $\mathcal{G} \in \overline{\mathcal{H}}^{L^2_\nu}$ (the L^2_ν-closure of $\mathcal{H}$), it follows that

$$\mathscr{A}_{\mathcal{G}}(\vartheta) = \sum_{\{j\in\mathbb{N}\,|\,\lambda_j\neq 0\}} \Big(\frac{\vartheta}{\lambda_j+\vartheta}\Big)\langle e_j, \mathcal{G}\rangle_{L^2_\nu}^2 \to 0 \quad \text{as} \quad \vartheta \to 0 \tag{B.19}$$

by dominated convergence. □

The *rate* of convergence of $\mathscr{A}_{\mathcal{G}}$ to zero can be quantified under an additional regularity assumption.

Lemma B.8 (convergence rate under Hölder source condition)**.** *Suppose $\mathcal{G} \in \mathrm{Im}(\mathcal{K}^{r/2})$ for some $r \geq 0$. Then for any $\vartheta > 0$, it holds under the prevailing assumptions of this appendix that*

$$\mathscr{A}_{\mathcal{G}}(\vartheta) \leq \|\mathcal{K}^{-r/2}\mathcal{G}\|_{L^2_\nu(\mathcal{X};\mathcal{Y})}^2 \times \begin{cases} \vartheta^r, & \text{if } \ r \in [0,1]\,, \\ \vartheta\|\mathcal{K}\|_{\mathcal{L}(L^2_\nu)}^{r-1}, & \text{if } \ r > 1\,. \end{cases} \tag{B.20}$$

Proof. The proof closely follows the argument of Smale and Zhou [246, Theorem 4, p. 295]. By hypothesis, there exists $F_{\mathcal{G}} \in L^2_\nu(\mathcal{X};\mathcal{Y})$ such that $\mathcal{G} = \mathcal{K}^{r/2}F_{\mathcal{G}}$. Then

$$\mathscr{A}_{\mathcal{G}}(\vartheta) = \sum_{j\in\mathbb{N}} \frac{\vartheta\lambda_j^r}{\lambda_j+\vartheta}\langle e_j, F_{\mathcal{G}}\rangle_{L^2_\nu}^2 \leq \Big(\sup_{j\in\mathbb{N}} \frac{\vartheta\lambda_j^r}{\lambda_j+\vartheta}\Big)\|F_{\mathcal{G}}\|_{L^2_\nu}^2 .$$

For any $j \in \mathbb{N}$, the argument of the supremum equals

$$\frac{\vartheta\lambda_j^r}{\lambda_j+\vartheta} = \Big(\frac{\lambda_j}{\lambda_j+\vartheta}\Big)\Big(\frac{\lambda_j}{\lambda_j+\vartheta}\Big)^{r-1}\frac{\vartheta}{(\lambda_j+\vartheta)^{1-r}} = \vartheta^r\Big(\frac{\lambda_j}{\lambda_j+\vartheta}\Big)^r\Big(\frac{\vartheta}{\lambda_j+\vartheta}\Big)^{1-r}.$$

This is bounded above by ϑ^r for all $j \in \mathbb{N}$ if $r \geq 0$ and $r \leq 1$. Otherwise, $r > 1$ and

$$\frac{\vartheta\lambda_j^r}{\lambda_j+\vartheta} = \vartheta\lambda_j^{r-1}\Big(\frac{\lambda_j}{\lambda_j+\vartheta}\Big) \leq \vartheta\lambda_j^{r-1}\,.$$

Taking the supremum over $j \in \mathbb{N}$ completes the proof. □

In Lemma B.8, $\mathcal{G}$ satisfies $\mathcal{G} \in \mathcal{H}$ if $r \geq 1$, and the rate of convergence of $\mathscr{A}_{\mathcal{G}}(\vartheta)$ is at least as fast as $O(\vartheta)$ as $\vartheta \to 0$. When $r \in [0,1)$, then $\mathcal{G} \notin \mathcal{H}$ and the rate becomes slower than linear in ϑ.

B.3 Proofs for Section 3.3

In this appendix, we prove Theorem 3.6 and its main consequences.

Proof of Theorem 3.6. Under the hypotheses, (3.17) in Proposition 3.16 holds with probability at least $1-\delta$ provided that $M \geq \lambda^{-1}\log(16/\delta)$ and $N \geq \lambda^{-2}\log(8/\delta)$. That is, $\mathscr{R}_N(\widehat{\alpha};\mathcal{G}) \leq \mathscr{R}_N^{\lambda}(\widehat{\alpha};\mathcal{G}) \leq \beta\lambda$, where $\beta = \beta(\rho,\lambda,\mathcal{G}_{\mathcal{H}},\eta)$ is given by (3.6). In particular, $\widehat{\alpha} \in \mathcal{A}_{\beta}$ (3.18) on the same event (by Corollary 3.17). It follows that, on this event,

$$\mathscr{R}(\widehat{\alpha};\mathcal{G}) - \mathscr{R}_N(\widehat{\alpha};\mathcal{G}) \leq \sup_{\alpha\in\mathcal{A}_{\beta}} \left|\mathscr{R}_N(\alpha;\mathcal{G}) - \mathscr{R}(\alpha;\mathcal{G})\right|.$$

Application of Proposition 3.22 shows that the right-hand side of the above display is bounded above by

$$32\sqrt{6}e^{3/2}\big(\|\mathcal{G}\|_{L^{\infty}_{\nu}}^2 + \|\varphi\|_{L^{\infty}}^2\beta(\rho,\lambda,\mathcal{G}_{\mathcal{H}},\eta)\big)\lambda \leq 79e^{3/2}\big(\|\mathcal{G}\|_{L^{\infty}_{\nu}}^2 + \|\varphi\|_{L^{\infty}}^2\beta(\rho,\lambda,\mathcal{G}_{\mathcal{H}},\eta)\big)\lambda$$

with probability at least $1-\delta$ because $N \geq \lambda^{-2}\log(8/\delta) \geq \log(2/\delta)$. Recalling (3.16), using $1 \leq 79e^{3/2}$, and applying a union bound completes the proof. □

Proof of Corollary 3.10. Since $\mathcal{G} = \rho + \mathcal{G}_{\mathcal{H}}$, we compute

$$\begin{aligned}\mathscr{R}(\widehat{\alpha};\mathcal{G}_{\mathcal{H}}) = \big\|\mathcal{G}_{\mathcal{H}} - \Phi(\,\cdot\,;\widehat{\alpha})\big\|_{L^2_{\nu}}^2 &= \big\|-\rho + [\mathcal{G} - \Phi(\,\cdot\,;\widehat{\alpha})]\big\|_{L^2_{\nu}}^2 \\ &\leq 2\|\rho\|_{L^2_{\nu}}^2 + 2\|\mathcal{G} - \Phi(\,\cdot\,;\widehat{\alpha})\|_{L^2_{\nu}}^2 \\ &= 2\,\mathbb{E}_{u\sim\nu}\|\rho(u)\|_{\mathcal{Y}}^2 + 2\mathscr{R}(\widehat{\alpha};\mathcal{G})\,.\end{aligned}$$

By (3.5) and (3.6), we see that the term $\mathbb{E}\|\rho(u)\|_{\mathcal{Y}}^2$ also appears in the upper bound for $\mathscr{R}(\widehat{\alpha};\mathcal{G})$. Collecting like terms and enlarging constants proves the assertion. □

Proof of Theorem 3.12. For $k \in \mathbb{N}$ and $\delta \in (0,1)$, the trained RFM satisfies $\|\mathcal{G} - \Phi(\,\cdot\,;\widehat{\alpha}^{(k)})\|_{L^2_{\nu}} \leq cs_k$ for some deterministic constant $c > 0$, where the sequence $s_k \to 0$ as $k \to \infty$ is given by the right-hand side of (3.12) with $\lambda = \lambda_k$ for each $k \in \mathbb{N}$ (by Lemma B.7). This inequality holds with probability

at least $1-\delta$ if $M \gtrsim \lambda_k^{-1}\log(2/\delta)$ and $N \gtrsim \lambda_k^{-2}\log(2/\delta)$ by Theorem 3.6. Now choose $\delta = \lambda_k$. Then

$$\sum_{k\in\mathbb{N}} \Pr\bigl\{\|\mathcal{G} - \Phi(\,\cdot\,;\widehat{\alpha}^{(k)})\|_{L^2_\nu} > cs_k\bigr\} \leq \sum_{k\in\mathbb{N}} \lambda_k < \infty\,.$$

The first Borel–Cantelli lemma establishes that there exists an $\mathbb{N}$-valued random variable k_0 such that $\|\mathcal{G} - \Phi(\,\cdot\,;\widehat{\alpha}^{(k)})\|_{L^2_\nu} \leq cs_k$ for all $k > k_0$ almost surely. This implies the desired result. □

Proof of Theorem 3.14. Application of Theorem 3.6 leads to the high probability bound (3.12) by the same argument from Section 3.3.3. Using that $\lambda \lesssim 1$ and Lemma B.8 proves the assertion. □

Proof of Corollary 3.15. Apply Theorem 3.14 to get a high probability bound. Choose $\lambda = \lambda_k = \delta$. Then the proof follows that of Theorem 3.12 except with $\{s_k\}$ replaced by $\{\lambda_k^{\min(r,1)/2}\}$. □

B.4 Proofs for Section 3.4

This appendix provides proofs of the error bounds for the regularized empirical risk (Appendix B.4.1) and the generalization gap (Appendix B.4.2).

B.4.1 Proofs for Subsection 3.4.1: Bounding the Regularized Empirical Risk

We now prove Proposition 3.16. Supporting results used in the proof appear after the argument in the subsequent subsections (Appendix B.4.1.1, B.4.1.2, and B.4.1.3).

Proof of Proposition 3.16. Our starting point is (3.20). We first note by linearity that

$$\begin{aligned}\frac{2}{N}\sum_{n=1}^{N}\langle \eta_n, \Phi(u_n;\widehat{\alpha})\rangle_{\mathcal{Y}} &= \|\widehat{\alpha}\|_M \left(\frac{2}{N}\sum_{n=1}^{N}\bigl\langle \eta_n, \Phi(u_n;\widehat{\alpha}/\|\widehat{\alpha}\|_M)\bigr\rangle_{\mathcal{Y}}\right)\\ &\leq \|\widehat{\alpha}\|_M \left(2\sup_{\alpha'\in\mathcal{A}_1}\left|\frac{1}{N}\sum_{n=1}^{N}\langle \eta_n, \Phi(u_n;\alpha')\rangle_{\mathcal{Y}}\right|\right),\end{aligned}$$

where $\mathcal{A}_1 = \{\alpha' \in \mathbb{R}^M \,|\, \|\alpha'\|_M^2 \le 1\}$, provided that $\|\widehat{\alpha}\|_M > 0$. Otherwise, the inequality in the above display holds trivially. Next, we define

$$t := \frac{1}{M}\sum_{m=1}^{M} |\widehat{\alpha}_m|^2 = \|\widehat{\alpha}\|_M^2\,, \tag{B.21}$$

$$A_{N,M}^{\lambda} := \mathscr{R}_N^{\lambda}(\alpha;\mathcal{G}) + \frac{2}{N}\sum_{n=1}^{N}\langle -\eta_n, \Phi(u_n;\alpha)\rangle_{\mathcal{Y}}\,, \quad \text{and} \tag{B.22}$$

$$c_{N,M} := \left(2\sup_{\alpha'\in\mathcal{A}_1}\left|\frac{1}{N}\sum_{n=1}^{N}\langle \eta_n, \Phi(u_n;\alpha')\rangle_{\mathcal{Y}}\right|\right)^2. \tag{B.23}$$

The inequality in (3.20) and the arithmetic-mean–geometric-mean inequality imply that

$$\lambda t \le \mathscr{R}_N^{\lambda}(\widehat{\alpha};\mathcal{G}) \le A_{N,M}^{\lambda} + \sqrt{c_{N,M}t} \le A_{N,M}^{\lambda} + \frac{1}{2}\lambda^{-1}c_{N,M} + \frac{1}{2}\lambda t\,. \tag{B.24}$$

Subtracting $\lambda t/2$ from both sides and multiplying through by $2\lambda^{-1}$ yields

$$t \le 2\lambda^{-1}A_{N,M}^{\lambda} + \lambda^{-2}c_{N,M}\,. \tag{B.25}$$

Substituting (B.25) back into the rightmost side of (B.24) gives

$$\mathscr{R}_N^{\lambda}(\widehat{\alpha};\mathcal{G}) \le 2A_{N,M}^{\lambda} + \lambda^{-1}c_{N,M}\,. \tag{B.26}$$

All of the above calculations hold with probability one. To complete our estimate of $\mathscr{R}_N^{\lambda}(\widehat{\alpha};\mathcal{G})$, it remains to upper bound the $A_{N,M}^{\lambda}$ (B.22) and $c_{N,M}$ (B.23) terms. We begin with the latter.

Lemmas B.12 and B.13 (with $t = 1$) and (B.7) show that

$$\begin{aligned}\sqrt{c_{N,M}} &\le \frac{4\|\eta\|_{\psi_1(\mathcal{Y})}\|\varphi\|_{L^\infty}}{\sqrt{N}} + 16e^{3/2}\|\eta\|_{\psi_1(\mathcal{Y})}\|\varphi\|_{L^\infty}\sqrt{\frac{\log(1/\delta)}{N}} \\ &\le 16e^{3/2}\|\eta\|_{\psi_1(\mathcal{Y})}\|\varphi\|_{L^\infty}\sqrt{\frac{6\log(2/\delta)}{N}}\end{aligned} \tag{B.27}$$

with probability at least $1-\delta$ if $N \ge \log(2/\delta) \ge \log(1/\delta)$. We used the inequalities $4 \le 16e^{3/2}$, $\sqrt{a}+\sqrt{b} \le \sqrt{2(a+b)}$, and $1 \le 2\log(2/\delta)$ to get to the last line.

Continuing, we bound $A_{N,M}^{\lambda}$. It has several error contributions originating from two terms. The first term in (B.22) is $\mathscr{R}_N^{\lambda}(\alpha;\mathcal{G})$, where $\alpha \in \mathbb{R}^M$ is arbitrary.

By Assumption 3.4, $\mathcal{G} = \rho + \mathcal{G}_{\mathcal{H}}$ so that

$$\begin{aligned}\mathscr{R}_N^{\lambda}(\alpha;\mathcal{G}) &\le \lambda\|\alpha\|_M^2 + 2\mathscr{R}_N(\alpha;\mathcal{G}_{\mathcal{H}}) + \frac{2}{N}\sum_{n=1}^{N}\|\rho(u_n)\|_{\mathcal{Y}}^2 \\ &\le 2\mathscr{R}_N^{\lambda}(\alpha;\mathcal{G}_{\mathcal{H}}) + \frac{2}{N}\sum_{n=1}^{N}\|\rho(u_n)\|_{\mathcal{Y}}^2 .\end{aligned} \tag{B.28}$$

By Lemma 3.20, it holds with probability at least $1-\delta$ that

$$\frac{2}{N}\sum_{n=1}^{N}\|\rho(u_n)\|_{\mathcal{Y}}^2 \le 4\,\mathbb{E}_u\|\rho(u)\|_{\mathcal{Y}}^2 + \frac{9\|\rho\|_{L^\infty_\nu}^2\log(2/\delta)}{N} . \tag{B.29}$$

Next, we bound the first term on the right-hand side in (B.28). Since $\mathcal{G}_{\mathcal{H}} \in \mathcal{H}$, there exists $\alpha_{\mathcal{H}} \in L^2_\mu(\Theta;\mathbb{R})$ such that

$$\mathcal{G}_{\mathcal{H}} = \mathbb{E}_{\theta\sim\mu}\big[\alpha_{\mathcal{H}}(\theta)\varphi(\,\cdot\,;\theta)\big] \quad \text{and} \quad \|\mathcal{G}_{\mathcal{H}}\|_{\mathcal{H}}^2 = \mathbb{E}_{\theta\sim\mu}|\alpha_{\mathcal{H}}(\theta)|^2 .$$

With $\alpha_{\mathcal{H}}$ as in the above display, choose once and for all $\alpha \equiv \alpha^\star \in \mathbb{R}^M$ as in (3.22). By Lemma 3.18,

$$2\mathscr{R}_N^{\lambda}(\alpha^\star;\mathcal{G}_{\mathcal{H}}) \le 162\lambda\|\mathcal{G}_{\mathcal{H}}\|_{\mathcal{H}}^2 \tag{B.30}$$

with probability at least $1-\delta$ if $M \ge \lambda^{-1}\log(4/\delta)$. On the same event,

$$\begin{aligned}\|\alpha^\star\|_M^2 &\le \mathbb{E}_\theta\,|\alpha_{\mathcal{H}}(\theta)|^2\left(1 + \frac{2\log(2/\delta)}{\lambda M} + \sqrt{\frac{2\log(2/\delta)}{\lambda M}}\right) \\ &\le 5\,\mathbb{E}_\theta\,|\alpha_{\mathcal{H}}(\theta)|^2 \\ &= 5\|\mathcal{G}_{\mathcal{H}}\|_{\mathcal{H}}^2\end{aligned}$$

by Lemma B.11. This fact and Lemma 3.21 (with $\alpha \equiv \alpha^\star$ as in Equation 3.22) shows that the second and final term in $A^{\lambda}_{N,M}$ (B.22) satisfies, with probability at least $1-\delta$, the upper bound

$$\frac{2}{N}\sum_{n=1}^{N}\langle -\eta_n, \Phi(u_n;\alpha^\star)\rangle_{\mathcal{Y}} \le 40e^{3/2}\|\mathcal{G}_{\mathcal{H}}\|_{\mathcal{H}}\|\eta\|_{\psi_1(\mathcal{Y})}\|\varphi\|_{L^\infty}\sqrt{\frac{\log(2/\delta)}{N}} . \tag{B.31}$$

Combining the estimates (B.27), (B.28), (B.29), (B.30), and (B.31), recalling (B.26), and invoking the union bound, we deduce that if $N \ge \lambda^{-2}\log(2/\delta)$, then

$$\mathscr{R}_N^{\lambda}(\widehat{\alpha};\mathcal{G}) \le C_0\lambda + 8\,\mathbb{E}_{u\sim\nu}\|\rho(u)\|_{\mathcal{Y}}^2 + 18\|\rho\|_{L^\infty_\nu}^2\lambda^2$$

with probability at least $1 - 4\delta$, where

$$\begin{aligned} C_0 &\coloneqq 324\|\mathcal{G}_{\mathcal{H}}\|_{\mathcal{H}}^2 + 80e^{3/2}\|\eta\|_{\psi_1(\mathcal{Y})}\|\varphi\|_{L^\infty}\|\mathcal{G}_{\mathcal{H}}\|_{\mathcal{H}} + 1536e^3\|\eta\|_{\psi_1(\mathcal{Y})}^2\|\varphi\|_{L^\infty}^2 \\ &\leq (324+4)\|\mathcal{G}_{\mathcal{H}}\|_{\mathcal{H}}^2 + 1936e^3\|\eta\|_{\psi_1(\mathcal{Y})}^2\|\varphi\|_{L^\infty}^2 \,. \end{aligned}$$

In the last line, we used Young's inequality with $\varepsilon = 1/8$, that is, $ab \leq \varepsilon a^2/2 + b^2/(2\varepsilon)$ with $a = 80e^{3/2}\|\eta\|_{\psi_1(\mathcal{Y})}\|\varphi\|_{L^\infty}$ and $b = \|\mathcal{G}_{\mathcal{H}}\|_{\mathcal{H}}$. This proves the asserted upper bound. □

B.4.1.1 Proofs for Subsection 3.4.1.1: Bounding the Approximation Error

Given a function $\alpha \in L^2_\mu(\Theta;\mathbb{R})$, we denote its cut-off at level $T > 0$ by

$$\theta \mapsto \alpha_{\leq T}(\theta) \coloneqq \alpha(\theta)\mathbb{1}_{\{|\alpha(\theta)|\leq T\}} \,. \tag{B.32}$$

This subsection is devoted to the proof of Lemma 3.18, which is based on the following three lemmas.

Lemma B.9. *Suppose $\mathcal{G} \in \mathcal{H}$ belongs to the RKHS $\mathcal{H}$. Let $\alpha \in L^2_\mu(\Theta;\mathbb{R})$ be such that $\mathcal{G} = \mathbb{E}_\theta[\alpha(\theta)\varphi(\,\cdot\,;\theta)]$. Let $u_1, \ldots, u_N \sim \nu$ be i.i.d. samples and $\nu_N = \frac{1}{N}\sum_{n=1}^N \delta_{u_n}$ be the corresponding empirical measure. Then almost surely,*

$$\big\|\mathcal{G} - \mathbb{E}_\theta[\alpha_{\leq T}(\theta)\varphi(\,\cdot\,;\theta)]\big\|_{L^2_{\nu_N}}^2 \leq \frac{\|\varphi\|_{L^\infty}^2(\mathbb{E}_\theta\,|\alpha(\theta)|^2)^2}{T^2} \quad \textit{for all} \quad T > 0\,. \tag{B.33}$$

Proof. Fix $u \in \mathcal{X}$ and define $\alpha_{>T} \coloneqq \alpha - \alpha_{\leq T}$. The claim follows from

$$\|\mathcal{G}(u) - \mathbb{E}_\theta[\alpha_{\leq T}(\theta)\varphi(u;\theta)]\|_{\mathcal{Y}}^2 = \|\,\mathbb{E}_\theta[\alpha_{>T}(\theta)\varphi(u;\theta)]\|_{\mathcal{Y}}^2 \leq \|\varphi\|_{L^\infty}^2(\mathbb{E}_\theta\,|\alpha_{>T}(\theta)|)^2$$

and the observation that $\mathbb{E}_\theta\,|\alpha_{>T}(\theta)| \leq \mathbb{E}_\theta\,|\alpha(\theta)|^2/T$. □

The previous lemma controls the error incurred by truncating the coefficient function of elements in the RKHS. The next lemma provides a bound on sample average approximations of these truncations.

Lemma B.10. *Let $u_1, \ldots, u_N \sim \nu$ be i.i.d. samples and let $\nu_N = \frac{1}{N}\sum_{n=1}^N \delta_{u_n}$ denote the corresponding empirical measure. For $\alpha \in L^2_\mu(\Theta;\mathbb{R})$, let $Z = Z(\theta)$ be the $L^2_{\nu_N}(\mathcal{X};\mathcal{Y})$-valued random variable defined for $\theta \sim \mu$ by*

$$Z = \alpha_{\leq T}(\theta)\varphi(\,\cdot\,;\theta)\,. \tag{B.34}$$

If $Z_1, \dots, Z_M$ are M i.i.d. copies of Z, then it holds with probability at least $1-\delta$ that

$$\left\| \frac{1}{M} \sum_{m=1}^{M} Z_m - \mathbb{E}\, Z \right\|_{L^2_{\nu_N}}^2 \leq \frac{32T^2 \|\varphi\|_{L^\infty}^2 \log^2(2/\delta)}{M^2} + \frac{4\|\varphi\|_{L^\infty}^2 \log(2/\delta)\, \mathbb{E}_\theta\, |\alpha(\theta)|^2}{M}\,. \tag{B.35}$$

Proof. By boundedness of $|\alpha_{\leq T}| \leq T$, we have the trivial uniform upper bound $\|Z_m\|_{L^2_{\nu_N}} \leq T\|\varphi\|_{L^\infty}$ for each m. The variance is bounded above as

$$\sigma^2 \coloneqq \mathbb{E}\, \|Z - \mathbb{E}\, Z\|_{L^2_{\nu_N}}^2 \leq \mathbb{E}\, \|Z\|_{L^2_{\nu_N}}^2 \leq \|\varphi\|_{L^\infty}^2\, \mathbb{E}_\theta\, |\alpha(\theta)|^2\,.$$

By Lemma B.2 and Theorem B.1, it holds with probability at least $1-\delta$ that

$$\left\| \frac{1}{M} \sum_{m=1}^{M} Z_m - \mathbb{E}\, Z \right\|_{L^2_{\nu_N}} \leq \frac{4T\|\varphi\|_{L^\infty} \log(2/\delta)}{M} + \sqrt{\frac{2\sigma^2 \log(2/\delta)}{M}}\,.$$

Squaring both sides and substitution of the above bound on σ^2 yields the claimed estimate. □

The third lemma below develops a high probability bound on the empirical approximation of the RKHS norm of truncated elements in the RKHS.

Lemma B.11. *Let $\alpha \in L^2_\mu(\Theta; \mathbb{R})$ and $\{\theta_m\} \sim \mu^{\otimes M}$. With probability at least $1-\delta$, it holds that*

$$\frac{1}{M} \sum_{m=1}^{M} |\alpha_{\leq T}(\theta_m)|^2 \leq \mathbb{E}_\theta\, |\alpha(\theta)|^2 + \frac{4T^2 \log(2/\delta)}{M} + \sqrt{\frac{2T^2\, \mathbb{E}_\theta\, |\alpha(\theta)|^2 \log(2/\delta)}{M}}\,. \tag{B.36}$$

Proof. We apply Bernstein's inequality (B.2) to the random variable $Z(\theta) \coloneqq |\alpha_{\leq T}(\theta)|^2$ with $\theta \sim \mu$ and $M \in \mathbb{N}$ i.i.d. copies $Z_1, \dots, Z_M$ of Z defined by $Z_m = Z(\theta_m)$ for each m. We note that $|Z| \leq T^2$ by definition. The variance of Z satisfies the upper bound

$$\sigma^2 \coloneqq \mathbb{E}|Z - \mathbb{E}\, Z|^2 \leq \mathbb{E}|Z|^2 = \mathbb{E}_\theta\, |\alpha_{\leq T}(\theta)|^4 \leq T^2\, \mathbb{E}_\theta\, |\alpha(\theta)|^2\,.$$

It follows from Lemma B.2 and Theorem B.1 that

$$\frac{1}{m} \sum_{m=1}^{M} Z_m \leq \mathbb{E}\, Z + \frac{4T^2 \log(2/\delta)}{M} + \sqrt{\frac{2\sigma^2 \log(2/\delta)}{M}}$$

with probability at least $1-\delta$. This in turn implies that

$$\frac{1}{M}\sum_{m=1}^{M}|\alpha_{\le T}(\theta_m)|^2 \le \mathbb{E}_\theta\,|\alpha(\theta)|^2] + \frac{4T^2\log(2/\delta)}{M} + \sqrt{\frac{2T^2\,\mathbb{E}\,|\alpha(\theta)|^2\log(2/\delta)}{M}}$$

with at least the same probability. This is the claim. □

We are now in a position to prove Lemma 3.18.

Proof of Lemma 3.18. Write $T \coloneqq T(\lambda)$ and $\alpha \coloneqq \alpha_{\mathcal{H}}$. Next, define $\alpha_{\le T} \in L^2_\mu(\Theta;\mathbb{R})$ by

$$\theta \mapsto \alpha_{\le T}(\theta) \coloneqq \alpha(\theta)\mathbb{1}_{\{|\alpha(\theta)|\le T\}} = \begin{cases}\alpha(\theta), & \text{if } |\alpha(\theta)|\le T\,,\\ 0, & \text{otherwise}\,.\end{cases}$$

We define $\alpha_{>T} \coloneqq \alpha\mathbb{1}_{\{|\alpha|>T\}}$ similarly, so that $\alpha \equiv \alpha_{\le T} + \alpha_{>T}$ holds true. Then for $\theta_1,\dots,\theta_M$, we have $\alpha^\star \in \mathbb{R}^M$ given by $\alpha^\star_m = \alpha_{\le T}(\theta_m)$ for each $m \in \{1,\dots,M\}$. We claim that $\mathscr{R}^\lambda_N(\alpha^\star;\mathcal{G}) \le (74\|\varphi\|^2_{L^\infty}+7)\lambda\,\mathbb{E}_\theta\,|\alpha_{\mathcal{H}}(\theta)|^2$ with high probability, which implies the asserted bound (3.23). To see this, we make the error decomposition

$$\mathscr{R}^\lambda_N(\alpha^\star;\mathcal{G}) = \frac{1}{N}\sum_{n=1}^{N}\left\|\mathcal{G}(u_n) - \frac{1}{M}\sum_{m=1}^{M}\alpha_{\le T}(\theta_m)\varphi(u_n;\theta_m)\right\|^2_{\mathcal{Y}} + \frac{\lambda}{M}\sum_{m=1}^{M}|\alpha_{\le T}(\theta_m)|^2 \tag{B.37}$$

$$\le \frac{2}{N}\sum_{n=1}^{N}\|\mathcal{G}(u_n) - \mathbb{E}_\theta[\alpha_{\le T}(\theta)\varphi(u_n;\theta)]\|^2_{\mathcal{Y}} \tag{I}$$

$$+ \frac{2}{N}\sum_{n=1}^{N}\left\|\frac{1}{M}\sum_{m=1}^{M}\alpha_{\le T}(\theta_m)\varphi(u_n;\theta_m) - \mathbb{E}_\theta[\alpha_{\le T}(\theta)\varphi(u_n;\theta)]\right\|^2_{\mathcal{Y}} \tag{II}$$

$$+ \frac{\lambda}{M}\sum_{m=1}^{M}|\alpha_{\le T}(\theta_m)|^2\,. \tag{III}$$

Each of the three terms (I)–(III) is estimated as follows.

By Lemma B.9, we can bound

$$\text{(I)} \le \frac{2\|\varphi\|^2_{L^\infty}(\mathbb{E}_\theta\,|\alpha(\theta)|^2)^2}{T^2} = 2\lambda\|\varphi\|^2_{L^\infty}\,\mathbb{E}_\theta\,|\alpha(\theta)|^2\,.$$

Lemma B.10 delivers the bound

$$\begin{aligned}(\mathrm{II}) &\le \frac{64T^2\|\varphi\|_{L^\infty}^2\log(2/\delta)^2}{M^2} + \frac{8\|\varphi\|_{L^\infty}^2\log(2/\delta)\,\mathbb{E}_\theta\,|\alpha(\theta)|^2}{M} \\ &= \lambda\,\mathbb{E}_\theta\,|\alpha(\theta)|^2\left(\frac{64\|\varphi\|_{L^\infty}^2\log(2/\delta)^2}{\lambda^2M^2} + \frac{8\|\varphi\|_{L^\infty}^2\log(2/\delta)}{\lambda M}\right)\end{aligned}$$

with probability at least $1-\delta$.

Last, Lemma B.11 yields

$$\begin{aligned}(\mathrm{III}) &\le \lambda\,\mathbb{E}\,|\alpha(\theta)|^2 + \frac{4\lambda T^2\log(2/\delta)}{M} + \lambda\sqrt{\frac{2T^2\,\mathbb{E}\,|\alpha(\theta)|^2\log(2/\delta)}{M}} \\ &= \lambda\,\mathbb{E}_\theta\,|\alpha(\theta)|^2\left(1 + \frac{4\log(2/\delta)}{\lambda M} + \sqrt{\frac{2\log(2/\delta)}{\lambda M}}\right)\end{aligned}$$

with probability at least $1-\delta$.

Combining the three estimates, it follows that if

$$\frac{\log(2/\delta)}{\lambda M} \le 1\,,$$

then

$$\mathscr{R}_N^\lambda(\alpha^\star;\mathcal{G}_\mathcal{H}) = \mathscr{R}_N^\lambda(\alpha^\star;\mathcal{G}) \le (74\|\varphi\|_{L^\infty}^2 + 7)\lambda\,\mathbb{E}_\theta\,|\alpha_\mathcal{H}(\theta)|^2$$

with probability at least $1-2\delta$. We used the fact that $\sqrt{2}\le 2$. This is the claimed upper bound. □

B.4.1.2 Proof for Subsection 3.4.1.2: Bounding the Misspecification Error

Recall that $\rho \in L^\infty_\nu$ under Assumption 3.4. We now prove Lemma 3.20.

Proof of Lemma 3.20. Let $Z_1 = \|\rho(u_1)\|_\mathcal{Y}^2$, which is uncentered. Almost surely, $Z_1 \le \|\rho\|_{L^\infty_\nu}^2$ and

$$\mathbb{E}|Z_1 - \mathbb{E}\,Z_1|^2 \le \mathbb{E}\,Z_1^2 = \mathbb{E}_{u\sim\nu}\|\rho(u)\|_\mathcal{Y}^4 \le \|\rho\|_{L^\infty_\nu}^2\,\mathbb{E}_{u\sim\nu}\|\rho(u)\|_\mathcal{Y}^2\,.$$

Thus with probability at least $1-\delta$, Corollary B.2 and Theorem B.1 provide the bound

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^N\|\rho(u_n)\|_\mathcal{Y}^2 &\le \mathbb{E}_u\|\rho(u)\|_\mathcal{Y}^2 + \frac{4\|\rho\|_{L^\infty_\nu}^2\log(\frac{2}{\delta})}{N} + \sqrt{\frac{2\,\mathbb{E}_u\|\rho(u)\|_\mathcal{Y}^2\|\rho\|_{L^\infty_\nu}^2\log(\frac{2}{\delta})}{N}} \\ &\le 2\,\mathbb{E}_u\|\rho(u)\|_\mathcal{Y}^2 + \frac{\frac{9}{2}\|\rho\|_{L^\infty_\nu}^2\log(2/\delta)}{N}\,.\end{aligned}$$

To get the last inequality, we used the arithmetic-mean–geometric-mean inequality $\sqrt{ab} \le (a+b)/2$ to obtain

$$\sqrt{\left(2\,\mathbb{E}_u\|\rho(u)\|_{\mathcal{Y}}^2\right)\left(\|\rho\|_{L^\infty_\nu}^2 \log(2/\delta)/N\right)} \le \mathbb{E}_u\|\rho(u)\|_{\mathcal{Y}}^2 + \frac{\frac{1}{2}\|\rho\|_{L^\infty_\nu}^2 \log(2/\delta)}{N}\,.$$

Multiplying the penultimate chain of inequalities through by two completes the proof. □

B.4.1.3 Proofs for Subsection 3.4.1.3: Bounding the Noise Error

This subsection provides proofs for the lemmas used to control the error stemming from i.i.d. noise corrupting the output data as in Assumption 3.5. The estimates themselves could be improved by using (B.10) instead of (B.11) and by tracking the noise variance $\mathbb{E}\|\eta\|_{\mathcal{Y}}^2$ instead of bounding it above by $4\|\eta\|_{\psi_1(\mathcal{Y})}^2$. This would be relevant in settings where the noise is small or tends to zero with the sample size. We defer such considerations to future work.

Proof of Lemma 3.21. Define $Z_n(\alpha) \coloneqq \langle -\eta_n, \Phi(u_n;\alpha)\rangle_{\mathcal{Y}}$ for each n. Conditioned on $\{\theta_m\}$, it holds that Z_n is an i.i.d. copy of Z_1. By the assumption $\mathbb{E}[\eta_1 \,|\, u_1] = 0$, we have

$$\begin{aligned}
\mathbb{E}\, Z_1(\alpha) &= \mathbb{E}_{(u_1,\eta_1)}[\langle -\eta_1, \Phi(u_1;\alpha)\rangle_{\mathcal{Y}}] \\
&= \mathbb{E}_{u_1\sim\nu}[\mathbb{E}[\langle -\eta_1, \Phi(u_1;\alpha)\rangle_{\mathcal{Y}} \,|\, u_1]] \\
&= \mathbb{E}_{u_1\sim\nu}[\langle -\mathbb{E}[\eta_1 \,|\, u_1], \Phi(u_1;\alpha)\rangle_{\mathcal{Y}}] \\
&= 0\,.
\end{aligned} \tag{B.38}$$

Next, we compute that

$$|Z_1(\alpha)| \le \|\eta_1\|_{\mathcal{Y}}\|\Phi(u_1;\alpha)\|_{\mathcal{Y}} \le \|\eta_1\|_{\mathcal{Y}}\|\alpha\|_M\|\varphi\|_{L^\infty}$$

by two applications of the Cauchy–Schwarz inequality, one in $\mathcal{Y}$ and the other in $\mathbb{R}^M$. We deduce that $\|Z_1(\alpha)\|_{\psi_1} \le \|\eta_1\|_{\psi_1(\mathcal{Y})}\|\alpha\|_M\|\varphi\|_{L^\infty}$, conditioned on $\{\theta_m\}$. Proposition B.4, Theorem B.1 (Bernstein's inequality), and a similar argument to that in the proof of Corollary B.5 deliver the asserted bound. □

The next two lemmas are used in the proof of Proposition 3.16 to control the third and final term in (3.20).

Lemma B.12 (linear empirical process: concentration)**.** *Fix $t > 0$ and $\delta \in (0,1)$. Define*

$$Z_t \coloneqq \sup_{\alpha \in \mathcal{A}_t} \left| \frac{1}{N} \sum_{n=1}^{N} \langle \eta_n, \Phi(u_n;\alpha) \rangle_{\mathcal{Y}} \right|, \quad \text{where} \quad \mathcal{A}_t \coloneqq \left\{ \alpha \in \mathbb{R}^M \,\middle|\, \|\alpha\|_M^2 \le t \right\}. \tag{B.39}$$

If $N \ge \log(1/\delta)$, then conditioned on the realizations $\{\theta_m\}$ in the family Φ it holds that

$$Z_t \le \mathbb{E}_{\{(u_n,\eta_n)\}}[Z_t] + 8e^{3/2} \|\eta_1\|_{\psi_1(\mathcal{Y})} \|\varphi\|_{L^\infty} \sqrt{\frac{\log(1/\delta)}{N}} \sqrt{t} \tag{B.40}$$

with probability at least $1-\delta$.

Proof. For any $\alpha \in \mathcal{A}_t$, we compute

$$\begin{aligned} |\langle \eta_1, \Phi(u_1;\alpha) \rangle_{\mathcal{Y}}| &= \left| \frac{1}{M} \sum_{m=1}^{M} \alpha_m \langle \eta_1, \varphi(u_1;\theta_m) \rangle_{\mathcal{Y}} \right| \\ &\le \left(\frac{1}{M} \sum_{m=1}^{M} |\alpha_m|^2 \right)^{1/2} \left(\frac{1}{M} \sum_{m=1}^{M} \langle \eta_1, \varphi(u_1;\theta_m) \rangle_{\mathcal{Y}}^2 \right)^{1/2} \\ &\le \sqrt{t} \left(\|\eta_1\|_{\mathcal{Y}}^2 \frac{1}{M} \sum_{m=1}^{M} \|\varphi(u_1;\theta_m)\|_{\mathcal{Y}}^2 \right)^{1/2}. \end{aligned}$$

We used the Cauchy–Schwarz inequality twice. By the boundedness of φ, the above display gives

$$\left\| \langle \eta_1, \Phi(u_1;\,\cdot\,) \rangle \right\|_{\psi_1(C(\mathcal{A}_t;\mathbb{R}))} = \left\| \sup_{\alpha \in \mathcal{A}_t} \left| \langle \eta_1, \Phi(u_1;\alpha) \rangle_{\mathcal{Y}} \right| \right\|_{\psi_1} \le \|\eta_1\|_{\psi_1(\mathcal{Y})} \|\varphi\|_{L^\infty} \sqrt{t}\,.$$

The i.i.d. random variables $\langle \eta_n, \Phi(u_n;\,\cdot\,) \rangle_{\mathcal{Y}} \colon \alpha \mapsto \langle \eta_n, \Phi(u_n;\alpha) \rangle_{\mathcal{Y}}$ (conditional on $\{\theta_m\}$) are linear and hence continuous. Application of (B.11) in Corollary B.5 to $\langle \eta_n, \Phi(u_n;\,\cdot\,) \rangle_{\mathcal{Y}}$ taking value in the separable Banach space $C(\mathcal{A}_t;\mathbb{R})$ of continuous functions from compact set $\mathcal{A}_t \subset \mathbb{R}^M$ into $\mathbb{R}$, equipped with the supremum norm, completes the proof. □

The previous lemma gives a concentration bound for the linear empirical process and the next lemma estimates its expectation.

Lemma B.13 (linear empirical process: expectation)**.** *Fix $t > 0$. Define $\mathcal{A}_t$ as in Lemma B.12. Then*

$$\mathbb{E}_{\{(u_n,\eta_n)\}} \sup_{\alpha \in \mathcal{A}_t} \left| \frac{1}{N} \sum_{n=1}^{N} \langle \eta_n, \Phi(u_n;\alpha) \rangle_{\mathcal{Y}} \right| \le \frac{\|\eta_1\|_{L^2_{\mathbb{P}}(\Omega;\mathcal{Y})} \|\varphi\|_{L^\infty}}{\sqrt{N}} \sqrt{t}\,. \tag{B.41}$$

Proof. For any $\alpha \in \mathcal{A}_t$, the Cauchy–Schwarz inequality in $\mathbb{R}^M$ delivers the bound

$$\begin{aligned}\left|\frac{1}{N}\sum_{n=1}^{N}\langle \eta_n, \Phi(u_n;\alpha)\rangle_{\mathcal{Y}}\right| &= \left|\frac{1}{M}\sum_{m=1}^{M}\alpha_m \frac{1}{N}\sum_{n=1}^{N}\langle \eta_n, \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}\right| \\ &\le \left\|\{\alpha_m\}_{m=1}^{M}\right\|_M \left(\frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N}\sum_{n=1}^{N}\langle \eta_n, \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}\right]^2\right)^{1/2}.\end{aligned}$$

Let the left-hand side of (B.41) be denoted by Ξ_t. We next note that

$$\begin{aligned}\mathbb{E}_{\{(u_n,\eta_n)\}}\left[\langle \eta_n, \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}\right] &= \mathbb{E}_{u_n\sim\nu}\left[\mathbb{E}\left[\langle \eta_n, \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}} \,|\, u_n\right]\right] \\ &= \mathbb{E}_{u_n\sim\nu}[\langle \mathbb{E}[\eta_n \,|\, u_n], \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}] \\ &= 0\,.\end{aligned}$$

Using the independence of (u_n, η_n) and $(u_{n'}, \eta_{n'})$ for any two indices $n \neq n'$, together with the above observation, we thus obtain

$$\begin{aligned}&\mathbb{E}_{\{(u_n,\eta_n)\}}\left[\langle \eta_n, \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}\,\langle \eta_{n'}, \varphi(u_{n'};\theta_m)\rangle_{\mathcal{Y}}\right] \\ &\qquad = \mathbb{E}_{(u_n,\eta_n)}\left[\langle \eta_n, \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}\right]\,\mathbb{E}_{(u_{n'},\eta_{n'})}\left[\langle \eta_{n'}, \varphi(u_{n'};\theta_m)\rangle_{\mathcal{Y}}\right] \\ &\qquad = 0\,.\end{aligned}$$

This implies that

$$\begin{aligned}\Xi_t &\le \frac{\sqrt{t}}{N}\,\mathbb{E}_{\{(u_n,\eta_n)\}}\sqrt{\frac{1}{M}\sum_{m=1}^{M}\sum_{n,n'=1}^{N}\langle \eta_n, \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}\,\langle \eta_{n'}, \varphi(u_{n'};\theta_m)\rangle_{\mathcal{Y}}} \\ &\le \frac{\sqrt{t}}{N}\sqrt{\frac{1}{M}\sum_{m=1}^{M}\sum_{n,n'=1}^{N}\mathbb{E}_{\{(u_n,\eta_n)\}}\left[\langle \eta_n, \varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}\,\langle \eta_{n'}, \varphi(u_{n'};\theta_m)\rangle_{\mathcal{Y}}\right]} \\ &= \frac{\sqrt{t}}{\sqrt{N}}\sqrt{\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{(u_1,\eta_1)}\left\langle \eta_1, \varphi(u_1;\theta_m)\right\rangle_{\mathcal{Y}}^2} \\ &\le \frac{\sqrt{t}}{\sqrt{N}}\|\eta_1\|_{L^2_{\mathbb{P}}(\Omega;\mathcal{Y})}\|\varphi\|_{L^\infty}\,.\end{aligned}$$

We used Jensen's inequality in the second line, independence and the zero-mean property of the summands in the third line, and the Cauchy–Schwarz inequality in $\mathcal{Y}$ in the final line. □

B.4.2 Proofs for Subsection 3.4.2: Bounding the Generalization Gap

This subsection upper bounds the generalization gap with suprema techniques. We begin with the following empirical process concentration inequality. It gives uniform control on the difference between the empirical and population risk functionals. The process, as a function of its index α, is quadratic because the RFM $\Phi(\,\cdot\,;\alpha)$ is linear in α.

Lemma B.14 (quadratic empirical process: concentration)**.** *Fix $t > 0$ and $\delta \in (0,1)$. Define*

$$Z_t \coloneqq \sup_{\alpha \in \mathcal{A}_t} \left| \mathscr{R}_N(\alpha; \mathcal{G}) - \mathscr{R}(\alpha; \mathcal{G}) \right| \tag{B.42}$$

$$= \sup_{\alpha \in \mathcal{A}_t} \left| \frac{1}{N} \sum_{n=1}^{N} \|\mathcal{G}(u_n) - \Phi(u_n;\alpha)\|_{\mathcal{Y}}^2 - \mathbb{E}_{u\sim\nu} \|\mathcal{G}(u) - \Phi(u;\alpha)\|_{\mathcal{Y}}^2 \right|, \tag{B.43}$$

where

$$\mathcal{A}_t \coloneqq \left\{ \alpha \in \mathbb{R}^M \;\middle|\; \|\alpha\|_M^2 \le t \right\}. \tag{B.44}$$

If $N \ge \log(1/\delta)$, then conditioned on the realizations $\{\theta_m\}$ in the family Φ, it holds that

$$Z_t \le \mathbb{E}_{\{u_n\}}[Z_t] + 32 e^{3/2} (\|\mathcal{G}\|_{L_\nu^\infty}^2 + \|\varphi\|_{L^\infty}^2 t) \sqrt{\frac{\log(1/\delta)}{N}} \tag{B.45}$$

with probability at least $1 - \delta$.

Proof. For any $\alpha \in \mathcal{A}_t$ and $n \in \{1, \dots, N\}$, let

$$X_n(t,\alpha) \coloneqq \|\mathcal{G}(u_n) - \Phi(u_n;\alpha)\|_{\mathcal{Y}}^2 - \mathbb{E}_{u\sim\nu} \|\mathcal{G}(u) - \Phi(u;\alpha)\|_{\mathcal{Y}}^2 .$$

We compute

$$\begin{aligned} |X_1(t,\alpha)| &\le 2\|\mathcal{G}(u_1)\|_{\mathcal{Y}}^2 + 2\,\mathbb{E}_{u\sim\nu}\|\mathcal{G}(u)\|_{\mathcal{Y}}^2 + 2\|\Phi(u_1;\alpha)\|_{\mathcal{Y}}^2 + 2\,\mathbb{E}_{u\sim\nu}\|\Phi(u;\alpha)\|_{\mathcal{Y}}^2 \\ &\le 4\|\mathcal{G}\|_{L_\nu^\infty}^2 + 4\|\varphi\|_{L^\infty}^2 t . \end{aligned}$$

We used the fact that for any $u \in \mathcal{X}$ ν-almost surely, $\|\Phi(u;\alpha)\|_{\mathcal{Y}}^2 \le t\|\varphi\|_{L^\infty}^2$ on the set $\mathcal{A}_t$ (by the Cauchy–Schwarz inequality). This implies that

$$\big\|X_1(t,\,\cdot\,)\big\|_{\psi_1(C(\mathcal{A}_t;\mathbb{R}))} = \Big\| \sup_{\alpha\in\mathcal{A}_t} |X_1(t,\alpha)| \Big\|_{\psi_1} \le 4\|\mathcal{G}\|_{L_\nu^\infty}^2 + 4\|\varphi\|_{L^\infty}^2 t .$$

The $X_n(t,\,\cdot\,)$ do indeed belong to $C(\mathcal{A}_t;\mathbb{R})$ almost surely, as they can be written as a sum of affine and quadratic forms on $\mathbb{R}^M$ in the α variable. Application of (B.11) in Corollary B.5 (taking the separable Banach space to be $C(\mathcal{A}_t;\mathbb{R})$ equipped with the supremum norm) completes the proof. □

Since the supremum concentrates around its mean, it remains to show that its mean is small as a function of the sample size. The next lemma does this with Rademacher symmetrization.

Lemma B.15 (quadratic empirical process: expectation)**.** *Fix $t > 0$. Define Z_t as in Lemma B.14. Conditioned on the realizations $\{\theta_m\}$ in the family Φ, it holds that*

$$\mathbb{E}_{\{u_n\}}[Z_t] \le \frac{4\|\mathcal{G}\|_{L^\infty_\nu}^2}{\sqrt{N}} + \frac{4\|\varphi\|_{L^\infty}^2}{\sqrt{N}} t\,. \tag{B.46}$$

Proof. By Giné–Zinn symmetrization [see, e.g., 267, Section 4.2, Proposition 4.11, pp. 107–108],

$$\mathbb{E}_{\{u_n\}}[Z_t] \le 2\,\mathbb{E} \sup_{\alpha\in\mathcal{A}_t} \left| \frac{1}{N} \sum_{n=1}^N \varepsilon_n \|\mathcal{G}(u_n) - \Phi(u_n;\alpha)\|_{\mathcal{Y}}^2 \right|, \quad \text{where}$$
$$\varepsilon_n \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\big(\{+1,-1\}\big)\,,$$

because the original summands (conditioned on $\{\theta_m\}$) are independent. The expectation on the right is interpreted as the conditional expectation given $\{\theta_m\}$ (i.e., $\mathbb{E}_{\{u_n\},\{\varepsilon_n\}}$ over the data and Rademacher variables only). The right-hand side is the Rademacher complexity of the RF model class composed with the square loss. Expanding the square, it is bounded above by

$$2\,\mathbb{E}_{\{u_n\},\{\varepsilon_n\}} \left| \frac{1}{N} \sum_{n=1}^N \varepsilon_n \|\mathcal{G}(u_n)\|_{\mathcal{Y}}^2 \right| \tag{I}$$
$$+\,4\,\mathbb{E}_{\{u_n\},\{\varepsilon_n\}} \sup_{\alpha\in\mathcal{A}_t} \left| \frac{1}{N} \sum_{n=1}^N \varepsilon_n \langle \mathcal{G}(u_n), \Phi(u_n;\alpha)\rangle_{\mathcal{Y}} \right| \tag{II}$$
$$+\,2\,\mathbb{E}_{\{u_n\},\{\varepsilon_n\}} \sup_{\alpha\in\mathcal{A}_t} \left| \frac{1}{N} \sum_{n=1}^N \varepsilon_n \|\Phi(u_n;\alpha)\|_{\mathcal{Y}}^2 \right|. \tag{III}$$

We now estimate each term. The first term (I) satisfies the standard Monte Carlo bound

$$\text{(I)} \le 2\left(\mathbb{E}\left| \frac{1}{N} \sum_{n=1}^N \varepsilon_n \|\mathcal{G}(u_n)\|_{\mathcal{Y}}^2 \right|^2 \right)^{1/2} = \frac{2}{\sqrt{N}} \left(\frac{1}{N} \sum_{n=1}^N \mathbb{E}\|\mathcal{G}(u_n)\|_{\mathcal{Y}}^4 \right)^{1/2} \le \frac{2\|\mathcal{G}\|_{L^\infty_\nu}^2}{\sqrt{N}}\,.$$

For the second term (II), we begin by estimating the empirical average on the set $\mathcal{A}_t$ as

$$\begin{aligned}\left|\frac{1}{N}\sum_{n=1}^{N}\varepsilon_n\langle\mathcal{G}(u_n),\Phi(u_n;\alpha)\rangle_{\mathcal{Y}}\right| &= \left|\frac{1}{M}\sum_{m=1}^{M}\alpha_m\left(\frac{1}{N}\sum_{n=1}^{N}\varepsilon_n\langle\mathcal{G}(u_n),\varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}\right)\right| \\ &\le \sqrt{t}\left(\frac{1}{M}\sum_{m=1}^{M}\left|\frac{1}{N}\sum_{n=1}^{N}\varepsilon_n\langle\mathcal{G}(u_n),\varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}\right|^2\right)^{1/2}\end{aligned}$$

by the Cauchy–Schwarz inequality in $\mathbb{R}^M$. We deduce by Jensen's inequality and independence that (II) is less than or equal to

$$\begin{aligned}&\frac{4\sqrt{t}}{N}\left(\frac{1}{M}\sum_{m=1}^{M}\sum_{n,n'=1}^{N}\mathbb{E}[\varepsilon_n\varepsilon_{n'}]\,\mathbb{E}_{\{u_n\}}\left[\langle\mathcal{G}(u_n),\varphi(u_n;\theta_m)\rangle_{\mathcal{Y}}\langle\mathcal{G}(u_{n'}),\varphi(u_{n'};\theta_m)\rangle_{\mathcal{Y}}\right]\right)^{1/2} \\ &= \frac{4\sqrt{t}}{N}\left(\frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N}\mathbb{E}_{u\sim\nu}\langle\mathcal{G}(u),\varphi(u;\theta_m)\rangle_{\mathcal{Y}}^2\right)^{1/2}.\end{aligned}$$

A final application of the Cauchy–Schwarz inequality in $\mathcal{Y}$ in the last line shows that the second term (II) is bounded above by $4\sqrt{t}\|\mathcal{G}\|_{L^\infty_\nu}\|\varphi\|_{L^\infty}/\sqrt{N}$. By Young's inequality $ab \le a^2/2 + b^2/2$, we further bound

$$\frac{4\|\mathcal{G}\|_{L^\infty_\nu}\|\varphi\|_{L^\infty}\sqrt{t}}{\sqrt{N}} = \left(\frac{2\|\mathcal{G}\|_{L^\infty_\nu}}{N^{1/4}}\right)\left(\frac{2\|\varphi\|_{L^\infty}\sqrt{t}}{N^{1/4}}\right) \le \frac{2\|\mathcal{G}\|^2_{L^\infty_\nu}}{\sqrt{N}} + \frac{2\|\varphi\|^2_{L^\infty}t}{\sqrt{N}}\,.$$

The third term (III) is estimated in a similar manner. Expanding the empirical average on $\mathcal{A}_t$ yields

$$\begin{aligned}\left|\frac{1}{N}\sum_{n=1}^{N}\varepsilon_n\|\Phi(u_n;\alpha)\|^2_{\mathcal{Y}}\right| &= \left|\frac{1}{M}\sum_{m=1}^{M}\alpha_m\left(\frac{1}{M}\sum_{m'=1}^{M}\alpha_{m'}\beta^{(N)}_{m,m'}\right)\right|, \quad \text{where} \\ \beta^{(N)}_{m,m'} &= \frac{1}{N}\sum_{n=1}^{N}\varepsilon_n\langle\varphi(u_n;\theta_m),\varphi(u_n;\theta_{m'})\rangle_{\mathcal{Y}}\,.\end{aligned}$$

The first equality in the above display satisfies the upper bound

$$\begin{aligned}\sqrt{t}\left(\frac{1}{M}\sum_{m=1}^{M}\left|\frac{1}{M}\sum_{m'=1}^{M}\alpha_{m'}\beta^{(N)}_{m,m'}\right|^2\right)^{1/2} &\le \sqrt{t}\left(\frac{1}{M}\sum_{m=1}^{M}t\left[\frac{1}{M}\sum_{m'=1}^{M}\left|\beta^{(N)}_{m,m'}\right|^2\right]\right)^{1/2} \\ &= \frac{t}{N}\sqrt{\frac{1}{M^2}\sum_{m,m'=1}^{M}\left|\sum_{n=1}^{N}\varepsilon_n\langle\varphi(u_n;\theta_m),\varphi(u_n;\theta_{m'})\rangle_{\mathcal{Y}}\right|^2}\end{aligned}$$

by two applications of the Cauchy–Schwarz inequality in $\mathbb{R}^M$. Finally, we deduce that

$$\begin{aligned}(\text{III}) &\leq \frac{2t}{N}\sqrt{\frac{1}{M^2}\sum_{m,m'=1}^{M}\sum_{n=1}^{N}\mathbb{E}_{\{u_n\}}\big\langle\varphi(u_n;\theta_m),\varphi(u_n;\theta_{m'})\big\rangle_{\mathcal{Y}}^2} \\ &\leq \frac{2t}{\sqrt{N}}\sqrt{\frac{1}{M^2}\sum_{m,m'=1}^{M}\mathbb{E}_u\Big[\|\varphi(u;\theta_m)\|_{\mathcal{Y}}^2\|\varphi(u;\theta_{m'})\|_{\mathcal{Y}}^2\Big]} \\ &\leq \frac{2t\|\varphi\|_{L^\infty}^2}{\sqrt{N}}\end{aligned}$$

by Jensen's inequality, the fact that $\mathbb{E}[\varepsilon_n\varepsilon_{n'}] = \delta_{n,n'}$, and the Cauchy–Schwarz inequality in $\mathcal{Y}$.

Combining the three estimates completes the proof. □

The proof of the main generalization gap bound (3.27) is now immediate.

Proof of Proposition 3.22. Lemma B.14 and B.15 (applied with $t = \beta$) show that

$$\mathcal{E}_\beta(\{u_n\},\{\theta_m\}) \leq \frac{4(\|\mathcal{G}\|_{L^\infty_\nu}^2 + \|\varphi\|_{L^\infty}^2\beta)}{\sqrt{N}} + 32e^{3/2}(\|\mathcal{G}\|_{L^\infty_\nu}^2 + \|\varphi\|_{L^\infty}^2\beta)\sqrt{\frac{\log(1/\delta)}{N}} \tag{B.47}$$

with conditional probability (over $\{\theta_m\}$) at least $1-\delta$ if $N \geq \log(1/\delta)$. Since δ does not depend on $\{\theta_m\}$, we deduce by the tower rule of conditional expectation that the event implied by (B.47) has $\mathbb{P}$-probability at least $1-\delta$ as well. Using the inequalities $4 \leq 32e^{3/2}$ and $\sqrt{a}+\sqrt{b} \leq \sqrt{2(a+b)}$ shows that the expression in (B.47) is bounded above by

$$32e^{3/2}\big(\|\mathcal{G}\|_{L^\infty_\nu}^2 + \|\varphi\|_{L^\infty}^2\beta\big)\sqrt{\frac{2(1+\log(1/\delta))}{N}}\,.$$

Application of the inequality $1 \leq 2\log(2/\delta)$, valid for $\delta \in (0,1)$, implies (3.27) as asserted. □

B.5 Numerical Experiment Details

In this appendix, we detail the setup of the numerical experiment from Section 3.5 and provide additional visualization of the function-valued RFM's discretization-independence in Figure B.1. All code used to produce the numerical results and figures in this chapter are available at

https://github.com/nickhnelsen/error-bounds-for-vvRF .

A RFM with M features is trained on N input-output pairs $\{(u_n, \mathcal{G}(u_n))\}_{n=1}^N$ according to the vector-valued RF-RR algorithm. The ground truth map $\mathcal{G}\colon L^2(\mathbb{T};\mathbb{R}) \to L^2(\mathbb{T};\mathbb{R})$ is a nonlinear operator defined by $u^{(0)}(\cdot) \mapsto u(\cdot, 1)$, where $u = \{u(x,t)\}_{x,t}$ solves the partial differential equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = 10^{-1}\frac{\partial^2 u}{\partial x^2}\,, \quad (x,t) \in \mathbb{T} \times (0,\infty)\,,$$

with initial condition $u(\cdot, 0) = u^{(0)} \in L^2(\mathbb{T};\mathbb{R})$. Here, $\mathbb{T} \simeq (0, 2\pi)_{\text{per}}$ is the 1D torus which comes with periodic boundary conditions. The initial conditions $u_n \sim \nu$ are sampled i.i.d. from a centered Matérn-like Gaussian process according to [154, Section 6.3, p. 32].

The random features are defined in a similar way to the Fourier Space RFs introduced by Nelsen and Stuart [203, Section 3.1, p. 15]:

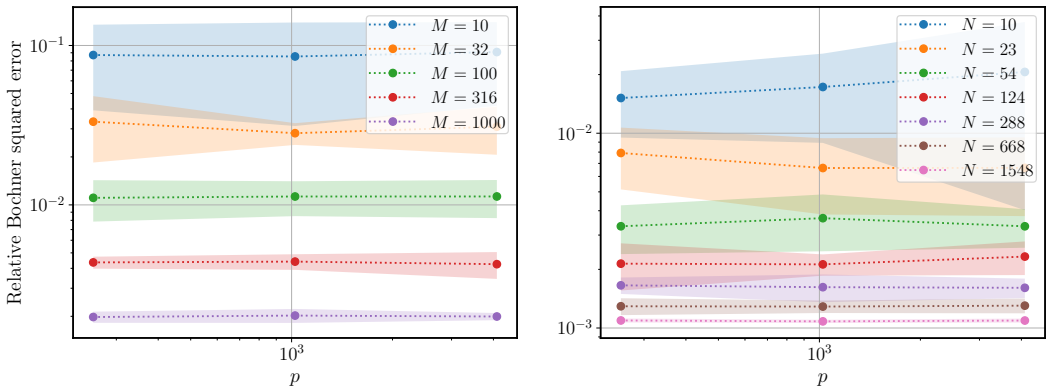
$$\varphi(u^{(0)};\theta) = 2.6 \cdot \mathrm{ELU}\big(\mathcal{F}^{-1}\{1_{(|k|\le k_{\max})}\chi_k \cdot (\mathcal{F}u^{(0)})_k \cdot (\mathcal{F}\theta)_k\}_{k\in\mathbb{Z}}\big) \quad \text{and} \quad \theta \sim \mu\,,$$

where μ is also a centered Matérn Gaussian measure with covariance operator $1.8^2(-\frac{d^2}{dx^2} + 15^2\,\mathrm{Id})^{-3}$. In the above display, $\mathcal{F}$ maps a function to its Fourier series coefficients, and $\mathcal{F}^{-1}$ expresses a Fourier coefficient sequence as a function expanded in the Fourier basis. The filter χ is given by [203, Eqn. 3.6, p. 15] with $\delta = 0.32$ and $\beta = 0.1$. We take $k_{\max} = 64$. The feature map φ lifts the notion of hidden neuron in neural network architectures to function space.

In Figures 3.2 and B.1, the quantity represented on the vertical axis is an empirical approximation to the relative Bochner squared error:

$$\begin{aligned}\frac{\frac{1}{N'}\sum_{n=1}^{N'}\|\mathcal{G}(u'_n) - \Phi(u'_n;\widehat{\alpha},\{\theta_m\})\|_{L^2}^2}{\frac{1}{N'}\sum_{n=1}^{N'}\|\mathcal{G}(u'_n)\|_{L^2}^2} &\approx \frac{\mathbb{E}_{u\sim\nu}\|\mathcal{G}(u) - \Phi(u;\widehat{\alpha},\{\theta_m\})\|_{L^2}^2}{\mathbb{E}_{u\sim\nu}\|\mathcal{G}(u)\|_{L^2}^2} \\ &= \frac{\mathscr{R}(\widehat{\alpha};\mathcal{G})}{\mathscr{R}(0;\mathcal{G})}\,.\end{aligned} \tag{B.48}$$

In (B.48), $N' = 500$ is the size of the test set $\{(u'_n, \mathcal{G}(u'_n))\}_{n=1}^{N'}$, where $\{u'_n\}_{n=1}^{N'} \sim \nu^{\otimes N'}$ is disjoint from the training input set $\{u_n\}_{n=1}^N$. The input and output spaces are discretized on the same p-point equally spaced grid in $(0, 2\pi)$. Thus, the discretized version of any input or output function belonging to $\mathcal{X} = \mathcal{Y} = L^2(\mathbb{T};\mathbb{R})$ may be identified with an element of $\mathbb{R}^p$. In Figures 3.2a and B.1a,



(a) Varying M, λ, p for fixed $N = 1548$.

(b) Varying N, λ, p for fixed $M = 10^4$.

Figure B.1: Squared test error—which empirically approximates the population risk $\mathscr{R}(\widehat{\alpha}; \mathcal{G})$—versus discretized output space dimension p, where $\mathcal{G}$ is the Burgers' equation solution operator (Appendix B.5).

the regularization factor is chosen as $\lambda = 7 \cdot 10^{-4}/M$ and as $\lambda = 3 \cdot 10^{-6}/\sqrt{N}$ in Figures 3.2b and B.1b.

A priori, it is not clear whether this operator learning benchmark satisfies our theoretical assumptions because we cannot verify that the Burgers' solution operator belongs to the RKHS of (φ, μ) (or the range of some power of the RKHS kernel integral operator). At a more technical level, the feature map φ uses an unbounded activation function (ELU, the exponential linear unit), while our theory is only developed for bounded RFs (Assumption 3.3). Nevertheless, the empirically obtained parameter and sample complexity in Figure 3.2 reasonably fit the main result of our well-specified theory (Theorem 3.9).

Appendix C

APPENDIX TO CHAPTER 4

This appendix collects all the proofs for Chapter 4: Learning Linear Operators From Noisy Data. The proofs for the main theorems appearing in Section 4.3 are given in Appendix C.1. Appendix C.2 states and proves key technical results used in the arguments. Finally, Appendix C.3 proves auxiliary results appearing in the main exposition of Chapter 4.

C.1 Proofs of Main Results

In this appendix, we provide proofs of the theorems from the main body of the chapter, in order of appearance. We begin with Theorem 4.3.

Proof of Theorem 4.3. Theorem 4.3 is a special case of Theorem 4.16 in the case $\alpha' < \alpha + 1/2$ in $\rho_N(\alpha, \alpha', p)$ (4.24). It remains to show that the Gaussian measure $\nu = \mathcal{N}(0, \Lambda)$ satisfies Assumption 4.15. The KL expansion coefficients certainly satisfy the fourth moment condition. The final condition on $\{\overline{g_j g_j}^{(N)}\}$ is verified by Lemma C.7 because $\{g_{jn}\}_{n=1}^N \sim \mathcal{N}(0, \vartheta_j^2)^{\otimes N}$. □

C.1.1 Proofs for Subsection 4.3.2

Under Assumptions 4.14 and 4.15, we calculate from Equations (4.12) and (4.13) that $\mathbb{E}^{Y\,|\,X}\,\mathbb{E}^{L^{(N)}\sim\mu^{D_N}}\|L^\dagger - L^{(N)}\|^2_{L^2_{\nu'}(H;H)} = \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3$ for $N \in \mathbb{N}$, where

$$\mathcal{I}_1 = \sum_{j=1}^{\infty} \frac{\vartheta_j'^2 |l_j^\dagger|^2}{(1 + N\gamma^{-2}\sigma_j^2 \overline{g_j g_j}^{(N)})^2}\,, \quad \mathcal{I}_2 = \sum_{j=1}^{\infty} \frac{N\vartheta_j'^2 \gamma^{-2}\sigma_j^4 \overline{g_j g_j}^{(N)}}{(1 + N\gamma^{-2}\sigma_j^2 \overline{g_j g_j}^{(N)})^2}\,, \tag{C.1a}$$

$$\mathcal{I}_3 = \sum_{j=1}^{\infty} \frac{\vartheta_j'^2 \sigma_j^2}{1 + N\gamma^{-2}\sigma_j^2 \overline{g_j g_j}^{(N)}}\,. \tag{C.1b}$$

This is the test error averaged only over the posterior and noise distributions, keeping the random design X fixed. The posterior mean test error (4.15) is given by $\mathbb{E}[\mathcal{I}_1 + \mathcal{I}_2]$ only. Recall from Assumption 4.14 that $\vartheta_j'^2$ (4.23) decays as $j^{-2\alpha'}$ (determining the test distribution ν') and σ_j^2 (4.21) decays (or grows) as j^{-2p} (determining the prior on L). The three series depend on $X = \{x_n\}$ through the correlated r.v.s $\{g_{jn} = \langle \varphi_j, x_n \rangle\}$ (4.11). These are mean zero with variance ϑ_j^2 (4.22) decaying as $j^{-2\alpha}$. The truth is $l^\dagger \in \mathcal{H}^s$, as in item (A2). All

of the following proofs involve estimating the three random series (C.1), which converge $\mathbb{P}$-a.s. by Item (A5) in Assumption 4.14 (and by Lemma C.3 for $\mathcal{I}_2$). For convenience, we set $u \coloneqq 2(\alpha+p) > 1$ and write $\overline{g_j g_j}^{(N)} =: \vartheta_j^2 Z_j^{(N)}$. Thus, $\mathbb{E}\, Z_j^{(N)} = 1$. We also set $\gamma \equiv 1$ without loss of generality.

Proof of Theorem 4.16. We split each of the three series (C.1) into sums over two disjoint index sets $\{j \in \mathbb{N} \colon j \leq N^{1/u}\}$ and $\{j \in \mathbb{N} \colon j > N^{1/u}\}$. We denote such sums by $\mathcal{I}_i^{\leq}$ and $\mathcal{I}_i^{>}$, respectively, for each $i \in \{1,2,3\}$. We must estimate their expectations over $X \sim \nu^{\otimes N}$ to prove the assertion (4.25). Notice that $Nj^{-u} \simeq 1 + Nj^{-u}$ whenever $j \leq N^{1/u}$.

Beginning with $\mathbb{E}\,\mathcal{I}_2$, its partial sum $\mathbb{E}\,\mathcal{I}_2^{\leq}$ satisfies

$$\begin{aligned}\mathbb{E} \sum_{j \leq N^{1/u}} \frac{N\vartheta_j'^2 \sigma_j^4 \overline{g_j g_j}^{(N)}}{(1 + N\sigma_j^2 \overline{g_j g_j}^{(N)})^2} &\leq \sum_{j \leq N^{1/u}} \frac{\vartheta_j'^2\, \mathbb{E}\big[(\overline{g_j g_j}^{(N)})^{-1}\big]}{N} \\ &\lesssim \sum_{j \leq N^{1/u}} \frac{\vartheta_j'^2 \vartheta_j^{-2}}{N} \asymp \sum_{j \leq N^{1/u}} \frac{j^{-2(\alpha'+p)}}{1 + Nj^{-u}}\end{aligned}$$

as $N \to \infty$. We used Assumption 4.15 and Lyapunov's inequality to bound the negative moment. By applying (C.8b) in Lemma C.2 (with $t = 2(\alpha'+p)$, $v = 1$, and condition $t > 1$ satisfied by item (A5)) to the last sum, we deduce that $\mathbb{E}\,\mathcal{I}_2^{\leq} = O(\rho_N)$. The tail series satisfies

$$\mathbb{E}\,\mathcal{I}_2^{>} \leq \sum_{j > N^{1/u}} N\vartheta_j'^2 \sigma_j^4\, \mathbb{E}\big[\vartheta_j^2 Z_j^{(N)}\big] \asymp N \sum_{j > N^{1/u}} j^{-2(\alpha'+\alpha+2p)} \asymp N^{-\left(1 - \frac{\alpha+1/2-\alpha'}{\alpha+p}\right)}$$

as $N \to \infty$ by (C.8a) in Lemma C.2 (applied with $t = 2(\alpha'+\alpha+2p) > 1$ by Item (A5)). This is always the same order as, or negligible compared to, the upper bound on $\mathbb{E}\,\mathcal{I}_2^{\leq}$.

By the same argument used for $\mathbb{E}\,\mathcal{I}_2^{\leq}$ (bounding its denominator by one and using Assumption 4.15 plus Lyapunov's inequality), we deduce that $\mathbb{E}\,\mathcal{I}_3^{\leq} = O(\rho_N)$ also. The tail $\mathbb{E}\,\mathcal{I}_3^{>}$ is bounded above by $\sum_{j > N^{1/u}} \vartheta_j'^2 \sigma_j^2 \asymp \sum_{j > N^{1/u}} j^{-2(\alpha'+p)}$. This sum is the same order as the bound on $\mathbb{E}\,\mathcal{I}_2^{>}$ by (C.8a) in Lemma C.2 (with $t = 2(\alpha'+p) > 1$ by Item (A5)).

Last, again by Assumption 4.15 and Lyapunov's inequality, $\mathbb{E}\,\mathcal{I}_1^{\leq}$ is bounded above by

$$\sum_{j \leq N^{\frac{1}{u}}} \frac{\vartheta_j'^2 |l_j^{\dagger}|^2\, \mathbb{E}\big[(\overline{g_j g_j}^{(N)})^{-2}\big]}{(N\sigma_j^2)^2} \lesssim \sum_{j \leq N^{\frac{1}{u}}} \frac{\vartheta_j'^2 |l_j^{\dagger}|^2 (\vartheta_j^2)^{-2}}{(N\sigma_j^2)^2} \asymp \sum_{j \leq N^{\frac{1}{u}}} \frac{j^{-2\alpha'} |l_j^{\dagger}|^2}{(1 + Nj^{-u})^2} \tag{C.2}$$

as $N \to \infty$. Application of (C.7) in Lemma C.1 (with $\xi = l^\dagger$, $t = 2\alpha'$, $q = s$, $v = 2$, and $t \geq -2q$ satisfied by Item (A5)) shows that this last sum is $o(N^{-(\alpha'+s)/(\alpha+p)})$ if $(\alpha'+s)/(\alpha+p) < 2$ or $\Theta(N^{-2})$ otherwise. The tail sum matches this bound in the first case and is strictly smaller otherwise because $\mathbb{E}\,\mathcal{I}_1^> \leq \sum_{j>N^{1/u}} \vartheta_j'^2 |l_j^\dagger|^2 \asymp \sum_{j>N^{1/u}} j^{-2\alpha'} |l_j^\dagger|^2$ (apply (C.6) in Lemma C.1 with $\xi = l^\dagger$, $t = 2\alpha'$ and $q = s$). All together, we deduce that $\mathbb{E}\,\mathcal{I}_2$ and $\mathbb{E}\,\mathcal{I}_3$ have the same upper bound $\rho_N \gg N^{-2}$. This implies (4.25). The uniform bound over $\|l^\dagger\|_{\mathcal{H}^s} \lesssim 1$ follows from the first assertion in [147, Lemma 8.1] (this turns the little-o into a big-O as claimed). The final assertion follows because the posterior mean test error only corresponds to $\mathcal{I}_1$ and $\mathcal{I}_2$. □

Proof of Theorem 4.17. The proof proceeds by developing lower bounds on each of the three series (C.1), using the same disjoint index sets approach in the proof of Theorem 4.16. For $\mathbb{E}\,\mathcal{I}_3$, since $r \mapsto (1+ar)^{-1}$ is convex on $[0,\infty)$ for all $a \geq 0$, Jensen's inequality yields

$$\mathbb{E} \sum_{j=1}^{\infty} \frac{\vartheta_j'^2 \sigma_j^2}{1 + N\sigma_j^2 \overline{g_j g_j}^{(N)}} \geq \sum_{j=1}^{\infty} \frac{\vartheta_j'^2 \sigma_j^2}{1 + N\sigma_j^2\, \mathbb{E}[\vartheta_j^2 Z_j^{(N)}]} \asymp \sum_{j=1}^{\infty} \frac{j^{-2(\alpha'+p)}}{1 + Nj^{-u}} \tag{C.3}$$

as $N \to \infty$. The last sum is $\Theta(\rho_N)$ by (C.8b) in Lemma C.2 (with $t = 2(\alpha'+p) > 1$ by (A5) and $v = 1$).

Next, $\mathbb{E}\,\mathcal{I}_2 \geq \mathbb{E}\,\mathcal{I}_2^\leq$ by nonnegativity. For any positive $\tau_N \to 0$, define the events $A_j^{(N)} \coloneqq \{\omega \in \Omega \colon Z_j^{(N)}(\omega) \geq \tau_N\}$ for every j and $N \in \mathbb{N}$. The law of total expectation yields

$$\begin{aligned}
\mathbb{E}\,\mathcal{I}_2^\leq = &\sum_{j \leq N^{1/u}} N\vartheta_j'^2 \sigma_j^4 \vartheta_j^2\, \mathbb{E}\left[\frac{Z_j^{(N)}}{(1 + N\sigma_j^2 \vartheta_j^2 Z_j^{(N)})^2} \,\middle|\, A_j^{(N)} \right] \mathbb{P}\big(A_j^{(N)}\big) \\
&+ \sum_{j \leq N^{1/u}} N\vartheta_j'^2 \sigma_j^4 \vartheta_j^2\, \mathbb{E}\left[\frac{Z_j^{(N)}}{(1 + N\sigma_j^2 \vartheta_j^2 Z_j^{(N)})^2} \,\middle|\, \big(A_j^{(N)}\big)^\mathsf{c} \right] \mathbb{P}\big(A_j^{(N)}\big)^\mathsf{c}.
\end{aligned}$$

The second term in the above display is nonnegative, so we obtain

$$\begin{aligned}
\mathbb{E}\,\mathcal{I}_2^\leq &\geq \sum_{j \leq N^{1/u}} \tau_N N\vartheta_j'^2 \sigma_j^4 \vartheta_j^2\, \mathbb{E}\big[(1 + N\sigma_j^2 \vartheta_j^2 Z_j^{(N)})^{-2} \,\big|\, A_j^{(N)}\big] \mathbb{P}\big(A_j^{(N)}\big) \\
&\geq \sum_{j \leq N^{1/u}} \frac{\tau_N N\vartheta_j'^2 \sigma_j^4 \vartheta_j^2\, \mathbb{P}\big(A_j^{(N)}\big)}{\big(1 + N\sigma_j^2 \vartheta_j^2\, \mathbb{E}\big[Z_j^{(N)} \,\big|\, A_j^{(N)}\big]\big)^2} \\
&= \sum_{j \leq N^{1/u}} \frac{\tau_N N\vartheta_j'^2 \sigma_j^4 \vartheta_j^2\, \mathbb{P}\big(A_j^{(N)}\big)^3}{\big(\mathbb{P}\big(A_j^{(N)}\big) + N\sigma_j^2 \vartheta_j^2\, \mathbb{E}\big[\mathbb{1}_{A_j^{(N)}} Z_j^{(N)}\big]\big)^2}.
\end{aligned}$$

We applied conditional Jensen's inequality to yield the second inequality because $r \mapsto (1+ar)^{-2}$ is convex on $[0,\infty)$ for any $a \geq 0$. Next, Markov's inequality plus Assumption 4.15 gives

$$\begin{aligned}\sup_{j\geq 1} \mathbb{P}\big(A_j^{(N)}\big)^{\mathsf{c}} &= \sup_{j\geq 1} \mathbb{P}\big\{(Z_j^{(N)})^{-1} > \tau_N^{-1}\big\} \\ &\leq \sup_{j\geq 1} \tau_N \, \mathbb{E}\big[(Z_j^{(N)})^{-1}\big] \to 0 \quad \text{as} \quad N \to \infty \,.\end{aligned}$$

This implies $\inf_{j\geq 1} \mathbb{P}(A_j^{(N)}) \to 1$ as $N \to \infty$. Using this and the facts $\mathbb{E}[\mathbb{1}_A Z_j^{(N)}] \leq \mathbb{E}[Z_j^{(N)}] = 1$ and $\mathbb{P}(A) \leq 1$ for any $A \in \mathcal{F}$ and applying $1 + Nj^{-u} \simeq Nj^{-u}$ for $j \leq N^{1/u}$ twice yields

$$\mathbb{E}\,\mathcal{I}_2^{\leq} \gtrsim \sum_{j \leq N^{1/u}} \frac{\tau_N N \vartheta_j'^2 \sigma_j^4 \vartheta_j^2 \, \mathbb{P}\big(A_j^{(N)}\big)^3}{(1+N\sigma_j^2\vartheta_j^2)^2} \gtrsim \tau_N \sum_{j \leq N^{1/u}} \frac{j^{-2(\alpha'+p)}}{1+Nj^{-u}} \quad \text{as} \quad N \to \infty \,.$$

Comparing to (C.3), we deduce $\mathbb{E}\,\mathcal{I}_2 = \Omega(\tau_N \rho_N)$. This is negligible relative to $\mathbb{E}\,\mathcal{I}_3 = \Omega(\rho_N)$.

Last, by Jensen's inequality, we lower bound $\mathbb{E}\,\mathcal{I}_1$ by the rightmost sum in (C.2), which is always $\Omega(N^{-2})$ (if $l^\dagger \neq 0$), plus the tail of the same sum, which gives the second term in (4.27) by (C.6) in Lemma C.1 (with $\xi = l^\dagger, t = 2\alpha', q = s$, and $v = 2$). The $\Omega(N^{-2})$ contribution from the first sum is dominated by both $\rho_N = \Omega(N^{-1})$ and $\tau_N \rho_N$ if $\tau_N \gg N^{-1}$. Therefore, the posterior sample test error (4.14) enjoys the asserted rate, while the posterior mean test error $\mathbb{E}[\mathcal{I}_1 + \mathcal{I}_2]$ only admits the bound (4.27) with the τ_N factor as claimed. □

Proof of Theorem 4.18. The ρ_N term (corresponding to $\mathcal{I}_2$ and $\mathcal{I}_3$ in (C.1)) in the assertion (4.28) follows from Theorems 4.16 and 4.17 for the posterior sample estimator. It remains to obtain the second term in (4.28). Following the argument from the proof of Theorem 4.16, $\mathbb{E}\,\mathcal{I}_1^{\leq}$ is asymptotically bounded above by the last sum in (C.2). Now given $|l_j^\dagger| \asymp j^{-1/2-s} S(j)$, by the full version of [147, Lemma 8.2] (applied with $\xi = l^\dagger$, $t = -2\alpha'$, $v = 2$, $q = s$, $\mathcal{S} = S$, and $t > -2q$ by item (A5)), this sum has exact order the second term in (4.28) if $(\alpha' + s)/(\alpha + p) < 2$ and is negligible relative to ρ_N otherwise. The tail $\mathbb{E}\,\mathcal{I}_1^{>}$ is always bounded above by the second term in (4.28) by the proof of [147, Lemma 8.2]. After an application of Jensen's inequality, the argument leading to a matching lower bound for $\mathbb{E}\,\mathcal{I}_1$ is the same as the one above. □

C.1.2 Proof for Subsection 4.3.4

We follow Appendix C.1.1 by letting $u \coloneqq 2(\alpha + p) > 1$, $\overline{g_j g_j}^{(N)} =: \vartheta_j^2 Z_j^{(N)}$, and $\gamma \equiv 1$, but instead of Assumption 4.15 we now enforce Assumption 4.21, which defines our Λ-subgaussian data. This yields $g_{jn} = \langle \varphi_j, x_n \rangle \in \mathrm{SG}(\sigma_\nu^2 \vartheta_j^2)$. Henceforth, let $\mathrm{SE}(v^2, a)$ denote the set of real subexponential (SE) r.v.s with parameters $(v, a) \in \mathbb{R}^2_{\geq 0}$. The inclusion $X \in \mathrm{SE}(v^2, a)$ is characterized by the moment generating function (MGF) bound $\mathbb{E} \exp(\theta(X - \mathbb{E}\, X)) \leq \exp(v^2\theta^2/2)$ for all $|\theta| < 1/a$ [267]. Using [266, Lemma 2.7.6] gives $g_{jn}^2/\vartheta_j^2 \in \mathrm{SE}(c^2\sigma_\nu^4, c\sigma_\nu^2)$ for an absolute constant $c > 0$. By independence, $Z_j^{(N)} \in \mathrm{SE}(c^2\sigma_\nu^4/N, c\sigma_\nu^2/N)$ [267, Section 2.1.3]. The following proof relies on SE concentration from Appendix C.2.

Proof of Theorem 4.22. We prove the upper and lower concentration bounds separately.

Upper Bound. Fix $\delta \in (0, 1 \wedge c\sigma_\nu^2)$ and define $N_{\delta^-} \coloneqq (1 - \delta)N$. We follow the disjoint index sets approach from Theorem 4.16, except now we sum over $\{j \in \mathbb{N} \colon j \leq N_{\delta^-}^{1/u}\}$ and $\{j \in \mathbb{N} \colon j > N_{\delta^-}^{1/u}\}$. Denote these sums by $\mathcal{I}_i^{\leq,\delta}$ and $\mathcal{I}_i^{>,\delta}$, respectively (C.1). We first bound

$$\mathcal{I}_1^{\leq,\delta} \leq \sum_{j \leq N_{\delta^-}^{1/u}} \frac{\vartheta_j'^2 |l_j^\dagger|^2}{(1 + N_{\delta^-}\sigma_j^2\vartheta_j^2)^2} \asymp \sum_{j \leq N_{\delta^-}^{1/u}} \frac{j^{-2\alpha'}|l_j^\dagger|^2}{(1 + N_{\delta^-} j^{-u})^2} \quad \text{as} \quad N \to \infty$$

with probability (w.p.) at least $1 - N_{\delta^-}^{1/u} \exp(-N\delta^2/(2c^2\sigma_\nu^4))$ by Lemma C.4 (with $n = N$, $X_j^{(N)} = Z_j^{(N)}$, $v = a = c\sigma_\nu^2$, $J = \lfloor N_{\delta^-}^{1/u} \rfloor$, and the lower tail only). The remaining bounds for $\mathcal{I}_1$ (including the almost sure bound for $\mathcal{I}_1^{>,\delta}$) are the same as those in the proof of Theorem 4.16, except with N replaced by N_{δ^-}. This gives the second term in (4.30).

Following the arguments in the proof of Theorem 4.16 for $\mathbb{E}\mathcal{I}_2^{\leq}$ and by a similar application of Lemma C.4, we deduce that $\mathcal{I}_2^{\leq,\delta} = O(\rho_{N_{\delta^-}})$ w.p. at least $1 - N_{\delta^-}^{1/u} \exp(-N\delta^2/(2c^2\sigma_\nu^4))$. For the infinite tail series $\mathcal{I}_2^{>,\delta}$, bounding its denominator by one yields

$$\mathcal{I}_2^{>,\delta} \leq \sum_{j > N_{\delta^-}^{1/u}} N\vartheta_j'^2\vartheta_j^2\sigma_j^4 Z_j^{(N)} \lesssim N(1 + \delta) \sum_{j > N_{\delta^-}^{1/u}} j^{-2(\alpha' + \alpha + 2p)} \quad \text{as} \quad N \to \infty$$

w.p. at least $1 - \exp(-N\delta^2/(2c^2\sigma_\nu^4))$. The second inequality is from Lemma C.6 (with $n = N$, $X_j^{(N)} = Z_j^{(N)}$, $v = a = c\sigma_\nu^2$, $\{w_j = \vartheta_j'^2\vartheta_j^2\sigma_j^4\}$, and the upper tail

only), where $\{\vartheta_j'^2\vartheta_j^2\sigma_j^4\}$ is in ℓ^1 because $\alpha'+\alpha+2p>1$ by item (A5). We deduce $\mathcal{I}_2^{>,\delta}=O((1+\delta)/(1-\delta)N_{\delta^-}^{-(1-(\alpha+1/2-\alpha')/(\alpha+p))})$ by the same argument used for $\mathbb{E}\,\mathcal{I}_2^{>}$ in the proof of Theorem 4.16.

Along similar lines as the proof of Theorem 4.16, the posterior covariance term $\mathcal{I}_3^{\leq,\delta}$ has the same order as $\mathcal{I}_2^{\leq,\delta}$ with the same probability (by Lemma C.4). The tail $\mathcal{I}_3^{>,\delta}$ is bounded above a.s. by the first case in $\rho_{N_{\delta^-}}$ (4.24) as $N\to\infty$. Since $a+b(1+\delta)/(1-\delta)\lesssim(1+\delta)/(1-\delta)$ for any $a,b>0$, we deduce $\mathcal{I}_2+\mathcal{I}_3$ has order the first term in (4.30) after choosing δ sufficiently small. The asserted total probability follows by combining the individual event probabilities with the union bound and the fact that there exists $c_1(\delta)>0$ and $0<c_3<c'\coloneqq 1/(2c^2\sigma_\nu^4)$ with $c_2=c_3\delta^2$ such that $\sup_{n\gtrsim 1}n^{1/u}\exp(-(c'-c_3)n\delta^2)<c_1(\delta)$. The assertion about the upper bound for $\bar{L}^{(N)}$ follows by ignoring $\mathcal{I}_3$.

Lower Bound. Since $1+\delta\in(1,2)$ is bounded, we do not track this factor in what follows. The proof proceeds by splitting all series at the critical index $J_N=\lfloor N^{1/u}\rfloor$ (since $(1+\delta)N\simeq N$) as in Theorem 4.16. By nonnegativity, we lower bound the error (C.1) by $\mathcal{I}_1^{>}+\mathcal{I}_2^{\leq}+\mathcal{I}_3^{\leq}$. The tail term $\mathcal{I}_1^{>}$ is bounded below by the second term in (4.31) with high probability by Lemma C.4 and Equation (C.6) in Lemma C.1. The remaining calculations showing that $\mathcal{I}_2^{\leq}$ and $\mathcal{I}_3^{\leq}$ are $\Omega(\rho_N)$ with high probability follow directly from Lemma C.4 and Equation (C.8b) in Lemma C.2 and are omitted. For $\bar{L}^{(N)}$, the only variance contribution is from $\mathcal{I}_2^{\leq}$; its lower bound $(1-\delta)\rho_N$ has a small pre-factor $1-\delta$. Combining the individual event probabilities as was done for the upper bound completes the proof of Theorem 4.22. □

C.1.3 Proof for Subsection 4.3.6

This subsection proves Theorem 4.24 by bounding the generalization gap $\mathcal{G}_N$ (4.20), which only involves in-distribution notions of error. We work in the setting of Appendix C.1.1, letting $u\coloneqq 2(\alpha+p)>1$ and $\gamma\equiv 1$ and enforcing Assumptions 4.14 and 4.15. Then, the $L^1_{\mathbb{P}}(\Omega;\mathbb{R})$ norm of $\mathcal{G}_N$ satisfies $\mathbb{E}^{D_N}|\mathcal{G}_N|=\mathbb{E}^{D_N}|\mathcal{J}_1+\mathcal{J}_2+\mathcal{J}_3|$, where

$$\frac{1}{2}\sum_{j=1}^{\infty}(\vartheta_j^2-\overline{g_jg_j}^{(N)})\left|\bar{l}_j^{(N)}-l_j^{\dagger}\right|^2,\quad \frac{1}{2}\sum_{j=1}^{\infty}(\overline{g_jg_j}^{(N)}-\vartheta_j^2)|l_j^{\dagger}|^2,\quad \sum_{j=1}^{\infty}\overline{g_j\xi_j}^{(N)}\bar{l}_j^{(N)} \tag{C.4}$$

define $\mathcal{J}_1$, $\mathcal{J}_2$, and $\mathcal{J}_3$, respectively. In (C.4), the r.v.s $\{\xi_{jn}\}$ from (4.6) are i.i.d. $\mathcal{N}(0,1)$. Using the explicit form (4.12) of the posterior mean $\{\bar{l}_j^{(N)}\}$, we find

that $\mathbb{E}^{D_N}|\mathcal{G}_N|$ equals

$$\mathbb{E}\left|\frac{1}{2}\sum_{j=1}^{\infty}(\vartheta_j^2-\overline{g_jg_j}^{(N)})\frac{|l_j^\dagger|^2+N\sigma_j^4\overline{g_jg_j}^{(N)}}{(1+N\sigma_j^2\overline{g_jg_j}^{(N)})^2}+\mathcal{J}_2+\sum_{j=1}^{\infty}\frac{(\overline{g_j\xi_j}^{(N)})^2+l_j^\dagger\overline{g_jg_j}^{(N)}\overline{g_j\xi_j}^{(N)}}{N^{-1}\sigma_j^{-2}+\overline{g_jg_j}^{(N)}}\right|. \tag{C.5}$$

The following proof and Lemma C.3 imply the convergence of (C.4) $\mathbb{P}$-a.s. and (C.5).

Proof of Theorem 4.24. We prove the upper and lower bounds on $\mathbb{E}^{D_N}|\mathcal{G}_N|$ separately.

Upper Bound. By the triangle inequality, (C.5) is bounded above by five terms G_i for $i \in \{1,\dots,5\}$. Here, $\{G_1, G_2\}$ corresponds to $\mathbb{E}|\mathcal{J}_1|$, G_3 to $\mathbb{E}|\mathcal{J}_2|$, and $\{G_4, G_5\}$ to $\mathbb{E}|\mathcal{J}_3|$.

By triangle and Jensen's inequality, $G_3 = \mathbb{E}|\mathcal{J}_2| \le \frac{1}{2}\sum_{j=1}^{\infty}|l_j^\dagger|^2(\mathrm{Var}[\overline{g_jg_j}^{(N)}])^{1/2}$. Independence of $\{x_n\}$ yields $\mathrm{Var}[\overline{g_jg_j}^{(N)}] \le \frac{1}{N}\mathbb{E}^{x\sim\nu}\langle\varphi_j,x\rangle^4$. Using Equation (4.11) and Assumption 4.15 ($\{\zeta_j\}$ are zero mean, unit variance, and independent), $\mathbb{E}^{x\sim\nu}\langle\varphi_j,x\rangle^4 \simeq \sum_k c_{jk}^4\,\mathbb{E}\,\zeta_k^4 + \sum_{k'\neq k} c_{jk}^2c_{jk'}^2$, where $c_{jk} \coloneqq \langle\Lambda^{1/2}\varphi_j,\phi_k\rangle$. The second term is bounded above by a constant times $(\sum_k c_{jk}^2)^2 = \vartheta_j^4$ and so is the first term (using $\limsup_{j\to\infty}\mathbb{E}\,\zeta_j^4 < \infty$ and $\ell^2 \subset \ell^4$). Thus, $G_3 \lesssim \|L^\dagger\|_{L^2_\nu(H;H)}^2\,N^{-1/2}$.

Using the disjoint index sets approach from the proof of Theorem 4.16, $G_1^{\le}$ is bounded above by $\frac{1}{2}\sum_{j\le N^{1/u}}(N\sigma_j^2)^{-2}|l_j^\dagger|^2\,\mathbb{E}[|\vartheta_j^2-\overline{g_jg_j}^{(N)}|(\overline{g_jg_j}^{(N)})^{-2}]$. By the Cauchy–Schwarz inequality and Assumption 4.15, the expectation on the right is bounded above by $(\mathrm{Var}[\overline{g_jg_j}^{(N)}])^{1/2}\vartheta_j^{-4}$ for sufficiently large N. It follows that $G_1^{\le}$ is of the order $N^{-1/2}$ times the rightmost sum in (C.2) with $\alpha' = \alpha$, which all together is $o(N^{-1/2})$. This contribution is negligible relative to G_3. A similar argument shows that the tail sum $G_1^{>}$ is never bigger than $G_1^{\le}$.

The other term associated with $\mathcal{J}_1$, which is G_2, satisfies

$$G_2^{\le} \le \frac{1}{2}\sum_{j\le N^{1/u}} N^{-1}\,\mathbb{E}[|\vartheta_j^2-\overline{g_jg_j}^{(N)}|(\overline{g_jg_j}^{(N)})^{-1}] = O\Big(N^{-\frac{1}{2}}N^{-\left(1-\frac{1}{u}\right)}\Big) = o(N^{-1/2})$$

(since $u > 1$) by an argument similar to the one used for G_1. The Cauchy–Schwarz inequality and the variance bound used for G_1 yields

$$G_2^{>} \le \frac{1}{2}\sum_{j>N^{1/u}} N\sigma_j^4\,\mathbb{E}[|\vartheta_j^2-\overline{g_jg_j}^{(N)}|\,\overline{g_jg_j}^{(N)}] \lesssim N^{-1/2}\sum_{j>N^{1/u}} N\sigma_j^4\vartheta_j^4\,.$$

The last sum is asymptotic to $N^{-1/2}\sum_{j>N^{1/u}} Nj^{-2u}$ as $N\to\infty$, which is the same order as $G_2^{\leq}$ by (C.8a) in Lemma C.2 (with $t=2u>1$). Thus, G_2 is also negligible relative to G_3.

Moving on to G_4 from $\mathbb{E}|\mathcal{J}_3|$, we first average out the noise $\{\xi_{jn}\}$ to obtain

$$G_4 = \mathbb{E}\sum_{j=1}^{\infty}\frac{(\overline{g_j\xi_j}^{(N)})^2}{N^{-1}\sigma_j^{-2}+\overline{g_jg_j}^{(N)}} = \mathbb{E}^X\sum_{j=1}^{\infty}\frac{N\sigma_j^2\,\mathbb{E}^{Y\,|\,X}\big[(\overline{g_j\xi_j}^{(N)})^2\big]}{1+N\sigma_j^2\overline{g_jg_j}^{(N)}}$$
$$= \mathbb{E}^X\sum_{j=1}^{\infty}\frac{\sigma_j^2\overline{g_jg_j}^{(N)}}{1+N\sigma_j^2\overline{g_jg_j}^{(N)}}\,.$$

Since the map $r\mapsto r(1+ar)^{-1}$ is concave on $[0,\infty)$ for all $a\geq 0$, Jensen's inequality yields $G_4\lesssim\sum_{j=1}^{\infty}j^{-u}/(1+Nj^{-u}) = O(N^{-(\alpha+p-1/2)/(\alpha+p)})$ as $N\to\infty$ by (C.8b) in Lemma C.2 (with $t=u>1$ and $v=1$, satisfying the first case).

Last, Jensen's inequality applied to the entire series G_5 from $\mathbb{E}|\mathcal{J}_3|$ yields

$$G_5 \leq \left(\mathbb{E}^X\,\mathbb{E}^{Y\,|\,X}\left|\sum_{j=1}^{\infty}\frac{N\sigma_j^2 l_j^{\dagger}\overline{g_jg_j}^{(N)}\overline{g_j\xi_j}^{(N)}}{1+N\sigma_j^2\overline{g_jg_j}^{(N)}}\right|^2\right)^{1/2}$$
$$= \left(\mathbb{E}^X\sum_{j=1}^{\infty}\frac{N|l_j^{\dagger}|^2\sigma_j^4(\overline{g_jg_j}^{(N)})^3}{(1+N\sigma_j^2\overline{g_jg_j}^{(N)})^2}\right)^{1/2}$$

because

$$\mathbb{E}^{Y\,|\,X}[(\overline{g_j\xi_j}^{(N)})(\overline{g_{j'}\xi_{j'}}^{(N)})] = \frac{1}{N^2}\sum_{n,n'\leq N}g_{jn}g_{j'n'}\,\mathbb{E}[\xi_{jn}\xi_{j'n'}] = \frac{\delta_{jj'}}{N}\left(\frac{1}{N}\sum_{n=1}^{N}g_{jn}g_{j'n}\right)$$

for any j and $j'\in\mathbb{N}$. Thus,

$$G_5 \leq \sqrt{\sum_{j=1}^{\infty}N^{-1}|l_j^{\dagger}|^2\vartheta_j^2} = \|L^{\dagger}\|_{L^2_{\nu}(H;H)}^2\,N^{-1/2}\,.$$

Comparing each $\{G_i\}_{i=1,\ldots,5}$, we conclude that $\mathbb{E}^{D_N}|\mathcal{G}_N| = O(N^{-1/2}+G_4)$ as $N\to\infty$ as asserted.

Lower Bound. By the triangle inequality,

$$\mathbb{E}^{D_N}|\mathcal{G}_N| \geq \mathbb{E}|\mathcal{J}_3+\mathcal{J}_2| - \mathbb{E}|\mathcal{J}_1| \geq |\mathbb{E}\,\mathcal{J}_3+\mathbb{E}\,\mathcal{J}_2| - \mathbb{E}|\mathcal{J}_1| = |\mathbb{E}\,\mathcal{J}_3| - \mathbb{E}|\mathcal{J}_1|\,.$$

We first develop a lower bound on $|\mathbb{E}\,\mathcal{J}_3|$, which equals G_4 by the zero mean property of the $\{\xi_{jn}\}$. By an argument similar to the one used to lower bound

$\mathbb{E}\,\mathcal{I}_2$ in the proof of Theorem 4.17,

$$|\mathbb{E}\,\mathcal{J}_3| \gtrsim \tau_N \sum_{j \le N^{1/u}} \frac{j^{-u}}{1 + Nj^{-u}} = \Omega\big(\tau_N N^{-(1-1/u)}\big) \quad \text{as} \quad N \to \infty$$

for any positive $\tau_N \to 0$. This is the asserted lower bound in (4.35). To conclude the proof, we claim that the upper bounds previously developed for $\mathbb{E}|\mathcal{J}_1|$ (i.e., for G_1 and G_2) are asymptotically negligible relative to $\tau_N N^{-(1-1/u)}$ under the hypotheses. Enforcing $\tau_N \gg N^{-1/2}$ ensures that this is true for the G_2 bound. By (C.2), if $(\alpha + s)/(\alpha + p) \ge 2$, then the G_1 contribution is $N^{-1/2}N^{-2} \ll \tau_N N^{-(1-1/u)}$. Otherwise, G_1 is strictly smaller than $N^{-1/2}N^{-(\alpha+s)/(\alpha+p)}$. This term is negligible relative to the $|\mathbb{E}\,\mathcal{J}_3|$ contribution if $\tau_N \gg N^{-(1+\alpha+2s-p)/(2\alpha+2p)} \to 0$, which requires $p < 1 + \alpha + 2s$ as assumed in the hypotheses. □

C.2 Supporting Lemmas

This appendix provides technical lemmas that are used repeatedly in the chapter. The first two results, which are variations of [147, Lemmas 8.1–8.2], develop sharp asymptotics for certain series that arise from $\mu_{\mathrm{seq}}^{D_N}$ in (4.12).

Lemma C.1 (series asymptotics: Sobolev regularity)**.** *Let $q \in \mathbb{R}$, $t \ge -2q$, $u > 0$, and $v \ge 0$. Then for every $\xi \in \mathcal{H}^q(\mathbb{N};\mathbb{R})$, it holds that*

$$\sum_{j > N^{1/u}} \frac{j^{-t}\xi_j^2}{(1 + Nj^{-u})^v} \simeq \sum_{j > N^{1/u}} j^{-t}\xi_j^2 \le N^{-\left(\frac{t+2q}{u}\right)} \Big(\sum_{j > N^{1/u}} j^{2q}\xi_j^2 \Big) \tag{C.6}$$

for all $N \in \mathbb{N}$. Additionally, for every fixed $\xi \in \mathcal{H}^q(\mathbb{N};\mathbb{R})$, it holds that

$$\sum_{j \le N^{1/u}} \frac{j^{-t}\xi_j^2}{(1 + Nj^{-u})^v} = \begin{cases} o\big(N^{-\left(\frac{t+2q}{u}\right)}\big), & \text{if } (t+2q)/u < v, \\ N^{-v}\,\|\xi\|^2_{\mathcal{H}^{(uv-t)/2}}\,\big(1 + o(1)\big), & \text{if } (t+2q)/u \ge v \end{cases} \tag{C.7}$$

as $N \to \infty$. The previous assertion (C.7) remains valid for the full infinite series.

Proof. The claims follow from [147, Lemma 8.1, p. 2653] and its proof therein. □

Lemma C.2 (series asymptotics: sharp)**.** *Let $t > 1$, $u > 0$, and $v \geq 0$. Then as $N \to \infty$,*

$$\sum_{j>N^{1/u}} \frac{j^{-t}}{(1+Nj^{-u})^v} \simeq \sum_{j>N^{1/u}} j^{-t} = \Theta\big(N^{-\left(\frac{t-1}{u}\right)}\big) \quad \textit{and} \tag{C.8a}$$

$$\sum_{j=1}^{\infty} \frac{j^{-t}}{(1+Nj^{-u})^v} \asymp \sum_{j\leq N^{1/u}} \frac{j^{-t}}{(1+Nj^{-u})^v} = \begin{cases} \Theta(N^{-\left(\frac{t-1}{u}\right)}), & \textit{if } (t-1)/u < v\,, \\ \Theta(N^{-v}\log N)\,, & \textit{if } (t-1)/u = v\,, \\ \Theta(N^{-v})\,, & \textit{if } (t-1)/u > v\,. \end{cases} \tag{C.8b}$$

Proof. The claims follow from [147, pp. 2654–2655]. Choose the slowly varying function used there to be identically constant, $q = -1/2$, and use the fact that $\sum_{j=1}^{J} 1/j \asymp \log J$ as $J \to \infty$. □

The next lemma justifies the a.s. convergence of various random series in our proofs.

Lemma C.3 (almost sure convergence of series)**.** *Let $\{X_j\}_{j\geq 1}$ be a sequence of (possibly dependent) real r.v.s. If $\sum_{j=1}^{\infty} \mathbb{E}|X_j| < \infty$, then*

$$\sum_{j=1}^{J} X_j \xrightarrow{\text{a.s.}} \sum_{j=1}^{\infty} X_j \quad \textit{as} \quad J \to \infty\,. \tag{C.9}$$

Proof. An application of the monotone convergence theorem shows that the random series $\sum_j |X_j|$ converges almost surely. □

We now turn to some useful concentration inequalities for subexponential r.v.s.

Lemma C.4 (subexponential: union)**.** *For $n \in \mathbb{N}$, let $\{X_j^{(n)}\}_{j\geq 1}$ be a (possibly dependent) family of unit mean $\mathrm{SE}(v^2/n, a/n)$ r.v.s. Fix $\delta \in (0, \min(1, v^2/a))$ and $J \in \mathbb{N}$. Then with probability at least $1 - 2J\exp(-n\delta^2/(2v^2))$, it holds that $(1-\delta) \leq X_j^{(n)} \leq (1+\delta)$ for all $j \leq J$.*

Proof. The result follows from application of the union bound to [267, Proposition 2.9]. □

To develop tighter concentration for subexponential series, we need the next two lemmas. The first result may be of independent interest.

Lemma C.5 (subexponential: closure under addition)**.** *Let $J \in \mathbb{N}$. If $\{X_j\}_{j=1,\dots,J}$ are (possibly dependent) real-valued r.v.s such that $X_j \in \mathrm{SE}(v_j^2, a_j)$ for every $j \in \{1,\dots,J\}$, then*

$$\sum_{j=1}^{J} X_j \in \mathrm{SE}\left(\left(\sum_{j=1}^{J} v_j\right)^2, \left(\sum_{j=1}^{J} v_j\right) \max_{1\le i\le J} \frac{a_i}{v_i}\right). \tag{C.10}$$

Proof. Defining the centered r.v. $Y_j \coloneqq X_j - \mathbb{E}\, X_j$ for each $j \in \{1,\dots,J\}$, we estimate

$$\begin{aligned}
\mathbb{E}\exp\left(\theta \sum_{j=1}^{J} Y_j\right) = \mathbb{E}\prod_{j=1}^{J} \exp(\theta Y_j) &\le \prod_{j=1}^{J} \big(\mathbb{E}\exp(\theta Y_j p_j)\big)^{1/p_j} \\
&\le \prod_{j=1}^{J} \big(\exp(v_j^2\theta^2 p_j^2/2)\big)^{1/p_j} \\
&= \exp\left(\left(\sum_{j=1}^{J} v_j\right)^2 \theta^2/2\right).
\end{aligned}$$

We used the generalized Hölder's inequality to yield the first inequality with $\sum_{i=1}^{J} 1/p_i = 1$ and $p_i \coloneqq v_i^{-1}\sum_{j=1}^{J} v_j$. The SE MGF bound applied for each $j \in \{1,\dots,J\}$ yields the second inequality, which is valid for all $|\theta| < \min_{i\le J}(p_i a_i)^{-1} = (\max_{i\le J} p_i a_i)^{-1}$ as asserted. □

Lemma C.6 (subexponential: series)**.** *For $n \in \mathbb{N}$, let $\{X_j^{(n)}\}_{j\ge 1}$ be a (possibly dependent) family of nonnegative unit mean $\mathrm{SE}(v^2/n, a/n)$ r.v.s. Let $w \in \ell^1(\mathbb{N};\mathbb{R})$ be nonnegative. Fix $\delta \in (0, \min(1, v^2/a))$. Then with probability at least $1 - 2\exp(-n\delta^2/(2v^2))$, it holds that*

$$(1-\delta)\sum_{j=1}^{\infty} w_j \le \sum_{j=1}^{\infty} w_j X_j^{(n)} \le (1+\delta)\sum_{j=1}^{\infty} w_j\,. \tag{C.11}$$

Proof. For any $J \in \mathbb{N}$, define $Y_J \coloneqq \sum_{j\le J} w_j X_j^{(n)}$. It follows from Lemma C.5 that $Y_J \in \mathrm{SE}(\frac{v^2}{n}\|\{w_j\}_{j\le J}\|_1^2, \frac{a}{n}\|\{w_j\}_{j\le J}\|_1)$. Since $\sum_{j=1}^{\infty}\mathbb{E}|w_j X_j^{(n)}| = \sum_{j=1}^{\infty} w_j < \infty$ holds by hypothesis, we deduce that $Y_J \to Y_\infty$ as $J \to \infty$ $\mathbb{P}$-a.s. by monotone convergence (Lemma C.3). Fatou's lemma applied to the Y_J SE MGF bound yields $Y_\infty \in \mathrm{SE}(\frac{v^2}{n}\|w\|_{\ell^1}^2, \frac{a}{n}\|w\|_{\ell^1})$. Thus, the fact that $\mathbb{E}\, Y_\infty = \|w\|_{\ell^1}$ and the SE tail bound (Lemma C.4) establish that $\mathbb{P}\big\{|Y_\infty - \mathbb{E}\, Y_\infty| \le \mathbb{E}\, Y_\infty\, \delta\big\} \ge 1 - 2\exp(-n\delta^2/(2v^2))$ for all $\delta \in (0, \min(1, v^2/a))$ as asserted. □

Our last result, specific to Gaussian design, is used in the proof of Theorem 4.3.

Lemma C.7 (chi-square moments)**.** *Let $W \sim \chi^2(n)$ be a chi-square r.v. with $n \in \mathbb{N}$ degrees of freedom. Then for any $q > -n/2$, $\mathbb{E}[W^q] = 2^q \frac{\Gamma(q+n/2)}{\Gamma(n/2)}$, where Γ is Euler's gamma function.*

Proof. A direct calculation with the PDF of $\chi^2(n)$ yields the q-th noncentral moment in closed form. □

C.3 Proofs of Auxiliary Results

In this appendix, we prove the facts asserted in Section 4.2.2.

Proof of Fact 4.4. By (4.9), $\mathcal{K}^{-1/2} \in \mathrm{HS}(H_{\Lambda'}; H)$ if and only if $\mathbb{E}^{x\sim\nu'}\|x\|_{\mathcal{K}}^2 < \infty$. Hence, $\nu'(H_{\mathcal{K}}) = 1$ as claimed. For the second claim, for any orthonormal basis $\{\psi_j\}$ of H we compute

$$\begin{aligned}\|T\Lambda'^{1/2}\|_{\mathrm{HS}}^2 &= \sum_{i,j}\langle\psi_i, T(\mathcal{K}^{1/2}\mathcal{K}^{-1/2})\Lambda'^{1/2}\psi_j\rangle^2 \\ &= \sum_{i,j}\langle(T\mathcal{K}^{1/2})^*\psi_i, \mathcal{K}^{-1/2}\Lambda'^{1/2}\psi_j\rangle^2 .\end{aligned}$$

Applying the Cauchy–Schwarz inequality to the rightmost equality yields the upper bound $\|(T\mathcal{K}^{1/2})^*\|_{\mathrm{HS}}^2\|\mathcal{K}^{-1/2}\Lambda'^{1/2}\|_{\mathrm{HS}}^2$. This is finite by hypothesis. So, we deduce $\mathbb{E}^{x\sim\nu'}\|Tx\|^2 < \infty$. □

Proof of Fact 4.6. For $N \in \mathbb{N}$, let $Z = (z_1, \dots, z_N) \in H_{\mathcal{K}}^N \setminus \{0\}$. By definition of K_Z, the map $K_Z^* K_Z \in \mathcal{L}(\mathrm{HS}(H_{\mathcal{K}}; H))$ acts as the right multiplication operator $T \mapsto T\mathcal{C}_{\mathcal{K}}^{(N)}$, where $\mathcal{C}_{\mathcal{K}}^{(N)} = \frac{1}{N}\sum_{n=1}^N z_n \otimes_{H_{\mathcal{K}}} z_n \in \mathcal{L}(H_{\mathcal{K}}) \setminus \{0\}$ is the empirical covariance of Z on $H_{\mathcal{K}}$. Thus, $K_Z^* K_Z = \mathrm{Id}_H \otimes \mathcal{C}_{\mathcal{K}}^{(N)}$ is a tensor product operator on $H \otimes H_{\mathcal{K}}$. But $\mathrm{Id}_H \in \mathcal{L}(H)$ is not compact on H. By [155, Corollary 1], $K_Z^* K_Z$ is not compact. Thus, K_Z is not compact either. □

Appendix D

APPENDIX TO CHAPTER 5

This appendix is the companion to Chapter 5: Operator Learning for Parameter-to-Observable Maps. Appendix D.1 states additional sample complexity results that complement those in the main body of the chapter. Detailed proofs for the approximation theory and statistical theory established by Sections 5.3 and 5.4 are provided in Appendices D.2 and D.3, respectively.

D.1 Additional Variants of the Sample Complexity Theorems

This appendix contains two additional theorems that are more general than their counterparts in Section 5.4 of the main text. While the main theorems as presented in Subsection 5.4.3 are sufficient to convey the primary message of that part of this chapter, we present the extra theorems here for completeness, as the results may be of independent interest.

First, we state the general convergence theorem for the end-to-end posterior mean estimator $\bar{f}^{(N)}$ from (5.15). The result is valid for any choice of prior covariance operator eigenvalue decay exponent $p > 1/2$. The proof is in Appendix D.3.2.

Theorem D.1 (end-to-end learning: general convergence rate)**.** *Let the input training data distribution ν, the test data distribution ν', and the Gaussian prior $\mathcal{N}(0, \Lambda)$ satisfy Assumptions 5.6 and 5.7. Let the ground truth linear functional $f^\dagger \in \mathcal{H}^s$ satisfy Assumption 5.9. Let α, α', and p be as in (5.20) and (5.21). Then there exists $c \in (0, 1/4)$ and $N_0 \geq 1$ such that for any $N \geq N_0$, the mean $\bar{f}^{(N)}$ of the Gaussian posterior distribution (5.15) arising from the N pairs of observed training data (U, Y) in (5.17) satisfies the error bound*

$$\mathbb{E}^{Y\,|\,U}\,\mathbb{E}^{u'\sim\nu'}\left|\langle f^\dagger, u'\rangle - \langle \bar{f}^{(N)}, u'\rangle\right|^2 \lesssim \left(1 + \|f^\dagger\|_{\mathcal{H}^s}^2\right)\varepsilon_N^2 \tag{D.1}$$

with probability at least $1 - 2\exp(-cN^{\min(1, \frac{\alpha+s-1}{\alpha+p})})$ over $U \sim \nu^{\otimes N}$, where

$$\varepsilon_N^2 := \begin{cases} N^{-\min\left(\frac{\alpha'+s}{\alpha+p},\, 1-\frac{\alpha+1/2-\alpha'}{\alpha+p}\right)}, & \text{if } \alpha' < \alpha + 1/2\,, \\ \max\left(N^{-\left(\frac{\alpha'+s}{\alpha+p}\right)},\, N^{-1}\log 2N\right), & \text{if } \alpha' = \alpha + 1/2\,, \\ N^{-\min\left(\frac{\alpha'+s}{\alpha+p},\, 1\right)}, & \text{if } \alpha' > \alpha + 1/2\,. \end{cases} \tag{D.2}$$

The constants c, N_0, and the implied constant in (D.1) do not depend on N or $f^\dagger$.

Theorem D.1 immediately implies an expectation bound for the test error, which we state now as a corollary and prove later in Appendix D.3.2.4.

Corollary D.2 (end-to-end learning: expectation bound)**.** *Instate the hypotheses and notation of Theorem D.1. Then there exists $N_\star \geq 1$ such that for any $N \geq N_\star$, the mean $\bar{f}^{(N)}$ of the Gaussian posterior distribution (5.15) arising from the N pairs of observed training data (U, Y) in (5.17) satisfies the expected error bound*

$$\mathbb{E}\Big[\mathbb{E}^{u'\sim\nu'}\big|\langle f^\dagger, u'\rangle - \langle \bar{f}^{(N)}, u'\rangle\big|^2\Big] \lesssim \big(1 + \|f^\dagger\|_{\mathcal{H}^s}^2\big)\varepsilon_N^2\,, \tag{D.3}$$

where ε_N^2 is as in (D.2). The constant $N_\star$ and the implied constant in (D.3) do not depend on N or $f^\dagger$.

The second extra theorem develops convergence rates for the full-field learning plug-in estimator under a Sobolev regularity condition on the underlying QoI.

Theorem D.3 (full-field learning: convergence rate for Sobolev QoI)**.** *Let the input training data distribution ν, the test data distribution ν', the true forward map $L^\dagger$, and the QoI $q^\dagger$ satisfy Assumption 5.10, but instead of (A-IV), suppose that $\{q^\dagger(\varphi_j)\}_{j\in\mathbb{N}} \in \mathcal{H}^r$ for some $r \in \mathbb{R}$. Let α and α' be as in (5.20) and β be as in (A-II). If $\min(\alpha, \alpha' + r) + \beta > 0$, then there exist constants $c > 0$ and $C > 0$ such that for all sufficiently large N, the plug-in estimator $q^\dagger \circ \bar{L}^{(N)}$ in (5.28) based on the Gaussian posterior distribution (5.27) arising from the N pairs of observed full-field training data (U, Υ) in (5.25) satisfies the error bound*

$$\mathbb{E}^{\Upsilon\,|\,U}\,\mathbb{E}^{u'\sim\nu'}\big|q^\dagger(L^\dagger u') - q^\dagger(\bar{L}^{(N)}u')\big|^2 \lesssim \varepsilon_N^2 \tag{D.4}$$

with probability at least $1 - Ce^{-cN}$ over $U \sim \nu^{\otimes N}$, where

$$\varepsilon_N^2 \coloneqq \begin{cases} N^{-\left(\frac{2\alpha'+2\beta+2r}{1+2\alpha+2\beta}\right)}\,, & \text{if } \alpha' + r < \alpha + 1/2\,, \\ N^{-1}\log N\,, & \text{if } \alpha' + r = \alpha + 1/2\,, \\ N^{-1}\,, & \text{if } \alpha' + r > \alpha + 1/2\,. \end{cases} \tag{D.5}$$

The constants c, C, and the implied constant in (D.4) do not depend on N.

The proof is found in Appendix D.3.3. The convergence rate (D.5) should be compared to the rate (5.32) from Theorem 5.13. Theorem D.3 covers a larger class of QoIs than does Theorem 5.13 due to the generality of the Sobolev regularity condition. Nonetheless, Theorem D.3 is not sharp when $q^\dagger$ decays asymptotically like a power law, as discussed in Subsection 5.4.3.1. It is straightforward to derive comparison results like Corollary 5.15 for the Sobolev QoI setting here, both for in-distribution and out-of-distribution test errors. Similar to Corollary D.2, expectation bounds are also readily derived from Theorem D.3. We refrain from doing so for the sake of brevity.

D.2 Proofs for Section 5.3

This appendix begins with some universal approximation results for neural operators before establishing similar universal approximation results for neural mappings (i.e., neural functionals and decoders).

D.2.1 Supporting Approximation Results for Neural Operators

We need the following two lemmas that are simple generalizations of the universal approximation theorem for FNOs [152, Theorem 9, p. 9] to the setting where only one of the input or output domain is the torus. These results may be extracted from the proof of [152, Theorem 9, p. 9].

Lemma D.4 (universal approximation for FNO: periodic output domain)**.** *Let Assumption 5.2 hold. Let $s \geq 0$ and $s' \geq 0$, $\mathcal{D} \subset \mathbb{R}^d$ be an open Lipschitz domain such that $\overline{\mathcal{D}} \subset (0,1)^d$, and $\mathcal{U} = H^s(\mathcal{D}; \mathbb{R}^{d_u})$. Let $\mathcal{Y} = H^{s'}(\mathbb{T}^d; \mathbb{R}^{d_y})$ and $\mathcal{G} \colon \mathcal{U} \to \mathcal{Y}$ be a continuous operator. There exists a continuous linear extension operator $E \colon \mathcal{U} \to H^s(\mathbb{T}^d; \mathbb{R}^{d_u})$ such that $(Eu)|_{\mathcal{D}} = u$ for all $u \in \mathcal{U}$. Moreover, let $K \subset \mathcal{U}$ be compact in $\mathcal{U}$. For any $\varepsilon > 0$, there exists a Fourier Neural Operator $\Psi \colon H^s(\mathbb{T}^d; \mathbb{R}^{d_u}) \to \mathcal{Y}$ of the form (5.4) (with $\mathcal{E} = \mathrm{Id}$, $\mathcal{R} = \mathrm{Id}$, and items (i) and (ii) both holding true) such that*

$$\sup_{u \in K} \|\mathcal{G}(u) - \Psi(Eu)\|_{\mathcal{Y}} < \varepsilon \,. \tag{D.6}$$

The next lemma is analogous to the previous one and deals with periodic input domains.

Lemma D.5 (universal approximation for FNO: periodic input domain)**.** *Let Assumption 5.2 hold. Let $s \geq 0$ and $s' \geq 0$, $\mathcal{D} \subset \mathbb{R}^d$ be an open Lipschitz domain such that $\overline{\mathcal{D}} \subset (0,1)^d$, and $\mathcal{U} = H^s(\mathbb{T}^d; \mathbb{R}^{d_u})$. Let $\mathcal{Y} = H^{s'}(\mathcal{D}; \mathbb{R}^{d_y})$ and*

$\mathcal{G}\colon \mathcal{U} \to \mathcal{Y}$ be a continuous operator. Denote by $R \in \mathcal{L}(H^{s'}(\mathbb{T}^d;\mathbb{R}^{d_y});\mathcal{Y})$ the restriction operator $y \mapsto y|_{\mathcal{D}}$. Let $K \subset \mathcal{U}$ be compact in $\mathcal{U}$. For any $\varepsilon > 0$, there exists a Fourier Neural Operator $\Psi\colon \mathcal{U} \to H^{s'}(\mathbb{T}^d;\mathbb{R}^{d_y})$ of the form (5.4) (with $\mathcal{E} = \mathrm{Id}$, $\mathcal{R} = \mathrm{Id}$, and items (i) and (ii) both holding true) such that

$$\sup_{u\in K} \|\mathcal{G}(u) - R\Psi(u)\|_{\mathcal{Y}} < \varepsilon\,. \tag{D.7}$$

D.2.2 Universal Approximation Proofs

The remainder of this appendix provides proofs of the main universal approximation theorems found in Section 5.3 for the proposed FNM family of architectures. We begin with the F2V FNF architecture.

Theorem 5.3 (universal approximation: function-to-vector mappings)**.** *Let $s \geq 0$, $\mathcal{D} \subset \mathbb{R}^d$ be an open Lipschitz domain such that $\overline{\mathcal{D}} \subset (0,1)^d$, and $\mathcal{U} = H^s(\mathcal{D};\mathbb{R}^{d_u})$. Let $\Psi^\dagger\colon \mathcal{U} \to \mathbb{R}^{d_y}$ be a continuous mapping. Let $K \subset \mathcal{U}$ be compact in $\mathcal{U}$. Under Assumption 5.2, for any $\varepsilon > 0$, there exist Fourier Neural Functionals $\Psi\colon \mathcal{U} \to \mathbb{R}^{d_y}$ of the form (5.8) with modification (M-F2V) such that*

$$\sup_{u\in K} \left\|\Psi^\dagger(u) - \Psi(u)\right\|_{\mathbb{R}^{d_y}} < \varepsilon\,. \tag{5.9}$$

Proof. Let $\mathcal{Y} \coloneqq L^2(\mathbb{T}^d;\mathbb{R}^{d_y})$ and $\mathbf{1}\colon x \mapsto 1$ be the constant function on $\mathbb{T}^d$. We first convert the function-to-vector mapping $\Psi^\dagger$ to the function-to-function operator $\mathcal{G}^\dagger\colon \mathcal{U} \to \mathcal{Y}$ defined by $u \mapsto \Psi^\dagger(u)\mathbf{1}$. We then establish the existence of a FNO that approximates $\mathcal{G}^\dagger$. Finally, from this FNO we construct a FNF that approximates $\Psi^\dagger$. To this end, fix $\varepsilon' > 0$. By the continuity of $\Psi^\dagger$, there exists $\delta > 0$ such that $\|u_1 - u_2\|_{\mathcal{U}} < \delta$ implies $\|\Psi^\dagger(u_1) - \Psi^\dagger(u_2)\|_{\mathbb{R}^{d_y}} < \varepsilon'$. Then

$$\begin{aligned}
\|\mathcal{G}^\dagger(u_1) - \mathcal{G}^\dagger(u_2)\|_{\mathcal{Y}}^2 &= \int_{\mathbb{T}^d} \|\Psi^\dagger(u_1)\mathbf{1}(x) - \Psi^\dagger(u_2)\mathbf{1}(x)\|_{\mathbb{R}^{d_y}}^2\,dx \\
&= |\mathbb{T}^d|\,\|\Psi^\dagger(u_1) - \Psi^\dagger(u_2)\|_{\mathbb{R}^{d_y}}^2 \\
&< (\varepsilon')^2\,.
\end{aligned}$$

We used the fact that $|\mathbb{T}^d| = 1$ for the identification $\mathbb{T}^d \equiv (0,1)^d_{\mathrm{per}}$. This shows the continuity of $\mathcal{G}^\dagger\colon \mathcal{U} \to \mathcal{Y}$. By the universal approximation theorem for FNOs (Lemma D.4, applied with $s = s$, $s' = 0$, $d_y = d_y$, and $\mathcal{G} = \mathcal{G}^\dagger$), there exists a continuous linear operator $E\colon \mathcal{U} \to H^s(\mathbb{T}^d;\mathbb{R}^{d_u})$ and a FNO $\mathcal{G}\colon H^s(\mathbb{T}^d;\mathbb{R}^{d_u}) \to \mathcal{Y}$ of the form (5.4) (with $\mathcal{R} = \mathrm{Id}$, $\mathcal{E} = \mathrm{Id}$, and items (i) and (ii) both holding

true) such that

$$\sup_{u\in K}\|\mathcal{G}^\dagger(u)-\mathcal{G}(Eu)\|_{\mathcal{Y}}<\varepsilon\,.$$

To complete the proof, we construct a FNF by appending a specific linear layer to the output of $\mathcal{G}\circ E$. To this end, let $\overline{P}\colon \mathcal{Y}\to\mathbb{R}^{d_y}$ be the averaging operator

$$u\mapsto \overline{P}u \coloneqq \int_{\mathbb{T}^d} u(x)\,dx\,.$$

Clearly $\overline{P}$ is linear. It is continuous on $\mathcal{Y}$ because

$$\|\overline{P}u\|_{\mathbb{R}^{d_y}}\le\int_{\mathbb{T}^d}\|u(x)\|_{\mathbb{R}^{d_y}}\mathbf{1}(x)\,dx\le\|u\|_{\mathcal{Y}}$$

by the triangle and Cauchy–Schwarz inequalities. Now define $\Psi\coloneqq(\overline{P}\circ\mathcal{G}\circ E)\colon\mathcal{U}\to\mathbb{R}^{d_y}$. This map has the representation

$$\Psi=\overline{P}\circ\tilde{\mathcal{Q}}\circ F\circ\tilde{\mathcal{S}}\circ E$$

for some local linear operators $\tilde{\mathcal{Q}}$ (identified with $\tilde{Q}\in\mathbb{R}^{d_y\times d_v}$ for channel dimension d_v) and $\tilde{\mathcal{S}}$ (identified with $\tilde{S}\in\mathbb{R}^{d_v\times d_u}$), and where F denotes the repeated composition of all nonlinear FNO layers of the form $\mathscr{L}_t$ as in (5.2). We claim that Ψ belongs to the FNF class, i.e., (5.8) with modification (M-F2V). To see this, choose $Q=I_{\mathbb{R}^{d_y}}\in\mathbb{R}^{d_y\times d_y}$ and $S=I_{\mathbb{R}^{d_u}}\in\mathbb{R}^{d_u\times d_u}$ (which we identify with $\mathrm{Id}_{\mathcal{U}}\in\mathcal{L}(\mathcal{U})$). Let $\mathcal{E}\coloneqq(\tilde{\mathcal{S}}\circ E)\colon\mathcal{U}\to H^s(\mathbb{T}^d;\mathbb{R}^{d_v})$. Define the linear functional layer $\mathscr{G}\coloneqq(\overline{P}\circ\tilde{\mathcal{Q}})\colon L^2(\mathbb{T}^d;\mathbb{R}^{d_v})\to\mathbb{R}^{d_y}$ which has the kernel linear functional representation

$$u\mapsto\mathscr{G}u=\int_{\mathbb{T}^d}\kappa(x)u(x)\,dx\,,\quad\text{where}\quad x\mapsto\kappa(x)\coloneqq\mathbf{1}(x)\tilde{Q}\in\mathbb{R}^{d_y\times d_v}$$

as in (5.5). Thus,

$$\begin{aligned}\Psi&=\overline{P}\circ\tilde{\mathcal{Q}}\circ F\circ\tilde{\mathcal{S}}\circ E\\&=I_{\mathbb{R}^{d_y}}\circ(\overline{P}\circ\tilde{\mathcal{Q}})\circ F\circ(\tilde{\mathcal{S}}\circ E)\circ\mathrm{Id}_{\mathcal{U}}\\&=Q\circ\mathscr{G}\circ F\circ\mathcal{E}\circ S\end{aligned}$$

as claimed. Finally, using the fact that $\overline{P}(z\mathbf{1})=z$ for any $z\in\mathbb{R}^{d_y}$, it holds that

$$\sup_{u\in K}\|\Psi^\dagger(u)-\Psi(u)\|_{\mathbb{R}^{d_y}}=\sup_{u\in K}\|\overline{P}\mathcal{G}^\dagger(u)-\overline{P}\mathcal{G}(Eu)\|_{\mathbb{R}^{d_y}}\le\sup_{u\in K}\|\mathcal{G}^\dagger(u)-\mathcal{G}(Eu)\|_{\mathcal{Y}}\,.$$

The rightmost expression is less than ε and hence (5.9) holds. □

The universality proof for the vector-to-function Fourier Neural Decoder (FND) architecture follows similar arguments.

Theorem 5.4 (universal approximation: vector-to-function mappings)**.** *Let $t \geq 0$, $\mathcal{D} \subset \mathbb{R}^d$ be an open Lipschitz domain such that $\overline{\mathcal{D}} \subset (0,1)^d$, and $\mathcal{Y} = H^t(\mathcal{D}; \mathbb{R}^{d_y})$. Let $\Psi^\dagger \colon \mathbb{R}^{d_u} \to \mathcal{Y}$ be a continuous mapping. Let $\mathcal{Z} \subset \mathbb{R}^{d_u}$ be compact. Under Assumption 5.2, for any $\varepsilon > 0$, there exists a Fourier Neural Decoder $\Psi \colon \mathbb{R}^{d_u} \to \mathcal{Y}$ of the form (5.8) with modification (M-V2F) such that*

$$\sup_{z \in \mathcal{Z}} \left\| \Psi^\dagger(z) - \Psi(z) \right\|_{\mathcal{Y}} < \varepsilon \,. \tag{5.10}$$

Proof. Let $\mathcal{U} \coloneqq L^2(\mathbb{T}^d; \mathbb{R}^{d_u})$ and $\mathbf{1} \colon \mathbb{T}^d \to \mathbb{R}$ be the constant function $x \mapsto 1$. Define the map $\mathsf{L} \colon \mathbb{R}^{d_u} \to \mathcal{U}$ by $z \mapsto z\mathbf{1}$. Clearly L is linear. To see that it is continuous, we compute

$$\|\mathsf{L}z\|_{\mathcal{U}}^2 = \int_{\mathbb{T}^d} \|z\mathbf{1}(x)\|_{\mathbb{R}^{d_u}}^2 \, dx = |\mathbb{T}^d| \|z\|_{\mathbb{R}^{d_u}}^2 = \|z\|_{\mathbb{R}^{d_u}}^2 \,. \tag{D.8}$$

Thus, L is injective with $\|\mathsf{L}\|_{\mathcal{L}(\mathbb{R}^{d_u}; \mathcal{U})} = 1$. Choose $K \coloneqq \mathsf{L}\mathcal{Z} = \{\mathsf{L}z \colon z \in \mathcal{Z}\} \subset \mathcal{U}$, which is compact in $\mathcal{U}$ because continuous functions map compact sets to compact sets. Define $\mathcal{G}^\dagger \colon K \to \mathcal{Y}$ by $\mathsf{L}z \mapsto \Psi^\dagger(z)$. First, we show that $\mathcal{G}^\dagger$ is continuous. Fix $\varepsilon' > 0$. By the continuity of $\Psi^\dagger$, there exists $\delta > 0$ such that if $\|\mathsf{L}z_1 - \mathsf{L}z_2\|_{\mathcal{U}} = \|z_1 - z_2\|_{\mathbb{R}^{d_u}} < \delta$, then $\|\Psi^\dagger(z_1) - \Psi^\dagger(z_2)\|_{\mathcal{Y}} < \varepsilon'$. Thus for any $u_1 = \mathsf{L}z_1 \in K$ and $u_2 = \mathsf{L}z_2 \in K$ with $\|u_1 - u_2\|_{\mathcal{U}} < \delta$, we have

$$\|\mathcal{G}^\dagger(u_1) - \mathcal{G}^\dagger(u_2)\|_{\mathcal{Y}} = \|\Psi^\dagger(z_1) - \Psi^\dagger(z_2)\|_{\mathcal{Y}} < \varepsilon' \,.$$

It follows that $\mathcal{G}^\dagger \colon K \to \mathcal{Y}$ is continuous. By the Dugundji extension theorem [84], there exists a continuous operator $\tilde{\mathcal{G}}^\dagger \colon \mathcal{U} \to \mathcal{Y}$ such that $\tilde{\mathcal{G}}^\dagger(u) = \mathcal{G}^\dagger(u)$ for every $u \in K$. By the universal approximation theorem for FNOs (Lemma D.5, applied with $s = 0$, $s' = t$, $d_u = d_u$, and $\mathcal{G} = \tilde{\mathcal{G}}^\dagger$), there exists a FNO $\mathcal{G} \colon \mathcal{U} \to H^t(\mathbb{T}^d; \mathbb{R}^{d_y})$ of the form (5.4) (with $\mathcal{R} = \mathrm{Id}$, $\mathcal{E} = \mathrm{Id}$, and items (*i*) and (*ii*) both holding true) such that

$$\sup_{u \in K} \|\tilde{\mathcal{G}}^\dagger(u) - R\mathcal{G}(u)\|_{\mathcal{Y}} = \sup_{u \in K} \|\mathcal{G}^\dagger(u) - R\mathcal{G}(u)\|_{\mathcal{Y}} < \varepsilon \,.$$

In the preceding display, $R \in \mathcal{L}(H^t(\mathbb{T}^d; \mathbb{R}^{d_y}); \mathcal{Y})$ denotes the restriction operator $y \mapsto y|_D$. Now define the map $\Psi \coloneqq (R \circ \mathcal{G} \circ \mathsf{L}) \colon \mathbb{R}^{d_u} \to \mathcal{Y}$. This map has the representation

$$\Psi = R \circ \tilde{\mathcal{Q}} \circ F \circ \tilde{\mathcal{S}} \circ \mathsf{L}$$

for some local linear operators $\tilde{\mathcal{Q}}$ (identified with $\tilde{Q} \in \mathbb{R}^{d_y \times d_v}$ for channel dimension d_v) and $\tilde{\mathcal{S}}$ (identified with $\tilde{S} \in \mathbb{R}^{d_v \times d_u}$), and where F denotes the repeated composition of all nonlinear FNO layers of the form $\mathscr{L}_t$ as in (5.2). We claim that Ψ is of the FND form, i.e., (5.8) with modification (M-V2F). To see this, choose $Q = I_{\mathbb{R}^{d_y}} \in \mathbb{R}^{d_y \times d_y}$ (which we identify with $\mathrm{Id}_{\mathcal{Y}} \in \mathcal{L}(\mathcal{Y})$) and $S = I_{\mathbb{R}^{d_u}} \in \mathbb{R}^{d_u \times d_u}$. Let $\mathcal{R} \coloneqq (R \circ \tilde{\mathcal{Q}})\colon H^t(\mathbb{T}^d; \mathbb{R}^{d_v}) \to \mathcal{Y}$. Define the linear decoder layer $\mathscr{D} \coloneqq (\tilde{\mathcal{S}} \circ \mathsf{L})\colon \mathbb{R}^{d_u} \to L^2(\mathbb{T}^d; \mathbb{R}^{d_v})$ which has the kernel function product representation

$$z \mapsto \mathscr{D}z = \kappa(\cdot)z\,, \quad \text{where} \quad x \mapsto \kappa(x) \coloneqq \mathbf{1}(x)\tilde{S} \in \mathbb{R}^{d_v \times d_u}$$

as in (5.5). Thus,

$$\begin{aligned}\psi &= R \circ \tilde{\mathcal{Q}} \circ F \circ \tilde{\mathcal{S}} \circ \mathsf{L} \\ &= \mathrm{Id}_{\mathcal{Y}} \circ (R \circ \tilde{\mathcal{Q}}) \circ F \circ (\tilde{\mathcal{S}} \circ \mathsf{L}) \circ I_{\mathbb{R}^{d_u}} \\ &= Q \circ \mathcal{R} \circ F \circ \mathscr{D} \circ S\end{aligned}$$

as claimed. Finally, by the injectivity of L implied by (D.8), any $u' \in K$ has the representation $u' = \mathsf{L}z'$ for some unique $z' \in \mathcal{Z} \subset \mathbb{R}^{d_u}$. It follows that

$$\sup_{u \in K} \|\mathcal{G}^\dagger(u) - R\mathcal{G}(u)\|_{\mathcal{Y}} \geq \|\mathcal{G}^\dagger(u') - R\mathcal{G}(u')\|_{\mathcal{Y}} = \|\Psi^\dagger(z') - \Psi(z')\|_{\mathcal{Y}}\,.$$

This implies the asserted result (5.10). □

D.3 Proofs for Section 5.4

This appendix contains the lengthy arguments that underlie the statistical learning theory for regression of linear functionals from Section 5.4. We begin by recalling convenient properties of subgaussian and subexponential probability distributions in Appendix D.3.1. Appendix D.3.2 contains proofs of the main (EE) results from Section 5.4.3. In particular, it develops a new bias–variance analysis of the linear functional regression problem in a Bayesian nonparametric setting that may be of independent interest. Proofs for the full-field learning approach (Subsection 5.4.2) to factorized linear functionals are provided in Appendix D.3.3. The sample complexity comparison corollary is proved in Appendix D.3.4. Technical lemmas used throughout the analysis are collected in Appendix D.3.5.

D.3.1 Subgaussian and Subexponential Distributions

This appendix reviews the concept of subgaussian and subexponential random variables. These play a central role in the analysis leading to the high probability error bounds in Section 5.4.3.

Definition D.6 (subgaussian)**.** A real-valued random variable X is said to be *subgaussian* [266, Section 2.5] if for some $\sigma > 0$ it satisfies the moment generating function bound

$$\mathbb{E}\, e^{\lambda(X - \mathbb{E}\, X)} \leq e^{\lambda^2 \sigma^2 / 2} \quad \text{for all} \quad \lambda \in \mathbb{R}\,. \tag{D.9}$$

We write $X \in \mathrm{SG}(\sigma^2)$ when (D.9) holds and define the subgaussian norm of X by

$$\|X\|_{\psi_2} \coloneqq \sup_{p \in [1,\infty)} \frac{\big(\mathbb{E}|X|^p\big)^{1/p}}{\sqrt{p}}\,. \tag{D.10}$$

It is known that X is subgaussian if and only if $\|X\|_{\psi_2} < \infty$. However, we often require random variables with heavier tails.

Definition D.7 (subexponential)**.** A real-valued random variable Z is said to be *subexponential* [266, Section 2.7] if for some $v > 0$ and $b > 0$ it satisfies the moment generating function bound

$$\mathbb{E}\, e^{\lambda(Z - \mathbb{E}\, Z)} \leq e^{\lambda^2 v^2 / 2} \quad \text{for all} \quad |\lambda| \leq \frac{1}{b}\,. \tag{D.11}$$

In contrast to the subgaussian case, the moment generating function of a subexponential random variable need only exist in a neighborhood of the origin instead of everywhere on the real line. We write $Z \in \mathrm{SE}(v^2, \alpha)$ when (D.11) holds and define subexponential norm by

$$\|Z\|_{\psi_1} \coloneqq \sup_{p \in [1,\infty)} \frac{\big(\mathbb{E}|Z|^p\big)^{1/p}}{p}\,. \tag{D.12}$$

It is known that X is subgaussian if and only if $Z = X^2$ is subexponential. In fact, we have the following estimate relating the two norms.

Lemma D.8 (squared subgaussian)**.** *Let X be a real-valued random variable. Then*

$$\big\|X\big\|_{\psi_2}^2 \leq \big\|X^2\big\|_{\psi_1} \leq 2\big\|X\big\|_{\psi_2}^2\,. \tag{D.13}$$

Proof. We compute

$$\begin{aligned}\left\|X^2\right\|_{\psi_1} = \sup_{p\in[1,\infty)} \frac{\left(\mathbb{E}|X|^{2p}\right)^{1/p}}{p} &= 2 \sup_{p\in[1,\infty)} \frac{\left(\mathbb{E}|X|^{2p}\right)^{1/2p}\left(\mathbb{E}|X|^{2p}\right)^{1/2p}}{\sqrt{2p}\sqrt{2p}} \\ &= 2\left(\sup_{p\in[1,\infty)} \frac{\left(\mathbb{E}|X|^{2p}\right)^{1/2p}}{\sqrt{2p}}\right)^2 \\ &\le 2\left(\sup_{2p\in[1,\infty)} \frac{\left(\mathbb{E}|X|^{2p}\right)^{1/2p}}{\sqrt{2p}}\right)^2.\end{aligned}$$

The final term inside parentheses on the right-hand side equals the subgaussian norm (upon replacing $2p$ with p). This is the asserted upper bound. The lower bound follows from the inequality $(\mathbb{E}|X|^{2p})^{1/2} \ge \mathbb{E}|X|^p$. □

D.3.2 Proofs for End-to-End Learning of General Linear Functionals

The goal of this appendix is to prove Theorem D.1 (which implies Theorem 5.12). This theorem provides a high probability convergence rate in terms of the sample size N for the *out-of-distribution test error*

$$\mathbb{E}^{u'\sim\nu'}\left|\langle f^\dagger, u'\rangle - \langle \bar{f}^{(N)}, u'\rangle\right|^2 = \left\|(\Sigma')^{1/2}(f^\dagger - \bar{f}^{(N)})\right\|^2 \tag{D.14}$$

conditioned on the covariates U. The equality in (D.14) is due to linearity and the fact that $|\langle f^\dagger, u'\rangle - \langle \bar{f}^{(N)}, u'\rangle|^2 = \langle f^\dagger - \bar{f}^{(N)}, (u'\otimes u')(f^\dagger - \bar{f}^{(N)})\rangle$. For notational convenience in the proofs, we write

$$\mathscr{R}_N \coloneqq \mathbb{E}^{Y\,|\,U}\left\|(\Sigma')^{1/2}(f^\dagger - \bar{f}^{(N)})\right\|^2 = \mathbb{E}\left[\left\|(\Sigma')^{1/2}(f^\dagger - \bar{f}^{(N)})\right\|^2 \,\middle|\, u_1,\dots,u_N\right]. \tag{D.15}$$

The argument behind the proof of Theorem D.1 follows a classical bias–variance decomposition. We now state this decomposition in the following lemma.

Lemma D.9 (bias–variance decomposition)**.** *Instate the setting of Subsection 5.4.1.1. The test error* (D.15) *satisfies the decomposition* $\mathscr{R}_N = B_N + V_N$*, where*

$$B_N \coloneqq \left\|(\Sigma')^{1/2}(\mathrm{Id}_H - A_N S_N) f^\dagger\right\|^2 \quad \textit{and} \tag{D.16a}$$

$$V_N \coloneqq \gamma^2\,\mathbb{E}\left[\left\|(\Sigma')^{1/2} A_N \Xi\right\|^2 \,\middle|\, U\right]. \tag{D.16b}$$

Proof. Denote $\|\cdot\|_{\nu'} \coloneqq \|(\Sigma')^{1/2}\cdot\|$. Let $m_N \coloneqq \mathbb{E}^{Y\,|\,U}[\bar{f}^{(N)}]$. Expanding $\mathbb{E}^{Y\,|\,U}\|f^\dagger - \bar{f}^{(N)}\|_{\nu'}^2 = \mathbb{E}^{Y\,|\,U}\|(f^\dagger - m_N) + (m_N - \bar{f}^{(N)})\|_{\nu'}^2$ shows that $\mathscr{R}_N$ equals

$$\|f^\dagger - m_N\|_{\nu'}^2 + 2\,\mathbb{E}^{Y\,|\,U}\left\langle f^\dagger - m_N, \Sigma'(m_N - \bar{f}^{(N)})\right\rangle + \mathbb{E}^{Y\,|\,U}\|m_N - \bar{f}^{(N)}\|_{\nu'}^2.$$

By linearity, the middle term equals zero. Recalling that $\bar{f}^{(N)} = A_N Y$ from (5.15) and $Y = S_N f^\dagger + \gamma\Xi$ from (5.17), the fact Ξ is centered implies that $m_N = A_N S_N f^\dagger$. Thus, we recover B_N (D.16a) as the first term in the preceding display. Similarly, $m_N - \bar{f}^{(N)} = -A_N(\gamma\Xi)$ so that V_N (D.16b) equals the last term. □

The main proof novelty lies in the bound for the bias B_N, which relies on a careful conditioning argument and clever matrix identities. The analysis begins by estimating the variance V_N in Appendix D.3.2.1. The bias is studied in Appendix D.3.2.2. Finally, the bounds are combined to prove Theorem D.1 in Appendix D.3.2.3 and Corollary D.2 in Appendix D.3.2.4.

D.3.2.1 Bounding the Variance

We focus on controlling the variance term (D.16b) first because it is easier to estimate than the bias. Our goal is to prove the following.

Proposition D.10 (variance upper bound)**.** *Under Assumptions 5.6 and 5.7, there exists $c \in (0,1)$ and $N_0 \geq 1$ (depending only on ν, ν', Λ, and γ^2) such that for all $N \geq N_0$, it holds that the variance term V_N in (D.16b) satisfies the estimate*

$$V_N \leq 2\sum_{j=1}^{\infty} \frac{\sigma_j' \lambda_j}{1 + N\gamma^{-2}\sigma_j\lambda_j} \lesssim \begin{cases} N^{-\left(1-\frac{\alpha+1/2-\alpha'}{\alpha+p}\right)}, & \text{if } \alpha' < \alpha + 1/2\,, \\ N^{-1}\log 2N\,, & \text{if } \alpha' = \alpha + 1/2\,, \\ N^{-1}\,, & \text{if } \alpha' > \alpha + 1/2 \end{cases} \tag{D.17}$$

with probability at least $1 - e^{-cN}$.

The expression for the variance upper bound in Proposition D.10 is in precisely the same form as that found in [76, Equation (A.1b), p. 24]. We derive quantitative convergence rates for V_N under the assumption of power law decay of the eigenvalues of the data and prior covariance operators. However, other types such as exponential decay [11] or convex eigenvalues [54] are also possible and interesting.

To prove the proposition, we require some preparatory results. First, notice that

$$\mathbb{E}\Big[\big\|(\Sigma')^{1/2} A_N \Xi\big\|^2 \,\Big|\, U\Big] = \mathbb{E}^{g\sim\mathcal{N}(0,\mathrm{Id}_{\mathbb{R}^N})}\big\|(\Sigma')^{1/2} A_N g\big\|^2 = \mathrm{tr}\big((\Sigma')^{1/2} A_N A_N^* (\Sigma')^{1/2}\big)\,.$$

Now using the fact that $\widehat{\Sigma} = S_N^* S_N / N$ from (5.14) and defining

$$\widehat{\mathcal{C}} \coloneqq \Lambda^{1/2} \widehat{\Sigma} \Lambda^{1/2} , \tag{D.18}$$

we see that

$$A_N A_N^* = \frac{1}{N} \left[\Lambda^{1/2} \left(\widehat{\mathcal{C}} + \frac{\gamma^2}{N} \operatorname{Id}_H \right)^{-1} \widehat{\mathcal{C}} \left(\widehat{\mathcal{C}} + \frac{\gamma^2}{N} \operatorname{Id}_H \right)^{-1} \Lambda^{1/2} \right] .$$

Next, define

$$\mu \coloneqq \gamma^2 / N > 0 \quad \text{and} \quad \widehat{\mathcal{C}}_\mu \coloneqq \widehat{\mathcal{C}} + \mu \operatorname{Id}_H \ . \tag{D.19}$$

By the cyclic property of the trace,

$$\begin{aligned} V_N &= \mu \operatorname{tr}\big(\Lambda^{1/2} \Sigma' \Lambda^{1/2} \widehat{\mathcal{C}}_\mu^{-1} \widehat{\mathcal{C}} \widehat{\mathcal{C}}_\mu^{-1} \big) \\ &= \mu \operatorname{tr}\big([\widehat{\mathcal{C}}_\mu^{-1/2} \Lambda^{1/2} \Sigma' \Lambda^{1/2} \widehat{\mathcal{C}}_\mu^{-1/2}] [\widehat{\mathcal{C}}_\mu^{-1/2} \widehat{\mathcal{C}} \widehat{\mathcal{C}}_\mu^{-1/2}] \big) \\ &\leq \mu \operatorname{tr}\big(\widehat{\mathcal{C}}_\mu^{-1/2} \Lambda^{1/2} \Sigma' \Lambda^{1/2} \widehat{\mathcal{C}}_\mu^{-1/2} \big) \big\| \widehat{\mathcal{C}}_\mu^{-1/2} \widehat{\mathcal{C}} \widehat{\mathcal{C}}_\mu^{-1/2} \big\|_{\mathcal{L}(H)} \\ &\leq \mu \operatorname{tr}\big(\widehat{\mathcal{C}}_\mu^{-1/2} \Lambda^{1/2} \Sigma' \Lambda^{1/2} \widehat{\mathcal{C}}_\mu^{-1/2} \big) . \end{aligned}$$

The first inequality is due to $\operatorname{tr}(AB) \leq \operatorname{tr}(A) \|B\|_{\mathcal{L}(H)}$, which holds for any symmetric positive-semidefinite trace-class A and any bounded B; this follows from the von Neumann trace inequality. The second inequality follows from the simultaneous diagonalizability of the factors in the triple product inside the operator norm and the fact that $\lambda/(\lambda + \mu) \leq 1$ for any eigenvalue λ of $\widehat{\mathcal{C}}$. Now define

$$\mathcal{C}' \coloneqq \Lambda^{1/2} \Sigma' \Lambda^{1/2} , \quad \mathcal{C} \coloneqq \Lambda^{1/2} \Sigma \Lambda^{1/2} , \quad \text{and} \quad \mathcal{C}_\mu \coloneqq \mathcal{C} + \mu \operatorname{Id}_H \ . \tag{D.20}$$

Lemma D.27 (with $A = \widehat{\mathcal{C}}$, $B = \mathcal{C}$, and $\lambda = \mu$) shows that V_N is bounded above by

$$\begin{aligned} \mu \operatorname{tr}\big(\mathcal{C}' \widehat{\mathcal{C}}_\mu^{-1} \big) &= \mu \operatorname{tr}\Big(\mathcal{C}' \mathcal{C}_\mu^{-1/2} \big(\operatorname{Id}_H - \mathcal{C}_\mu^{-1/2} (\mathcal{C} - \widehat{\mathcal{C}}) \mathcal{C}_\mu^{-1/2} \big)^{-1} \mathcal{C}_\mu^{-1/2} \Big) \\ &= \mu \operatorname{tr}\Big(\big[\mathcal{C}_\mu^{-1/2} \mathcal{C}' \mathcal{C}_\mu^{-1/2} \big] \big(\operatorname{Id}_H - \mathcal{C}_\mu^{-1/2} (\mathcal{C} - \widehat{\mathcal{C}}) \mathcal{C}_\mu^{-1/2} \big)^{-1} \Big) \\ &\leq \mu \operatorname{tr}\big(\mathcal{C}_\mu^{-1/2} \mathcal{C}' \mathcal{C}_\mu^{-1/2} \big) \big\| \big(\operatorname{Id}_H - \mathcal{C}_\mu^{-1/2} (\mathcal{C} - \widehat{\mathcal{C}}) \mathcal{C}_\mu^{-1/2} \big)^{-1} \big\|_{\mathcal{L}(H)} . \end{aligned} \tag{D.21}$$

The final inequality is again due to von Neumann's trace inequality. To bound the operator norm in the preceding display, we apply a Neumann series argument.

Lemma D.11 (Neumann series bound)**.** *Let $\mu = \gamma^2/N$. There exists $c \in (0,1)$ and $N_0 \geq 1$ such that for any $N \geq N_0$, it holds that*

$$\left\| \left(\mathrm{Id}_H - \mathcal{C}_\mu^{-1/2} (\mathcal{C} - \widehat{\mathcal{C}}) \mathcal{C}_\mu^{-1/2} \right)^{-1} \right\|_{\mathcal{L}(H)} \leq 2 \tag{D.22}$$

with probability at least $1 - e^{-cN}$.

Proof. By Lemma D.25, the event (D.55) holds with probability at least $1 - e^{-cN}$ for any $N \geq N_0$. On this event, we can invoke the Neumann series expansion

$$\left(\mathrm{Id}_H - \mathcal{C}_\mu^{-1/2} (\mathcal{C} - \widehat{\mathcal{C}}) \mathcal{C}_\mu^{-1/2} \right)^{-1} = \sum_{k=0}^{\infty} \left(\mathcal{C}_\mu^{-1/2} (\mathcal{C} - \widehat{\mathcal{C}}) \mathcal{C}_\mu^{-1/2} \right)^k .$$

This delivers the operator norm bound

$$\begin{aligned} \left\| \left(\mathrm{Id}_H - \mathcal{C}_\mu^{-1/2} (\mathcal{C} - \widehat{\mathcal{C}}) \mathcal{C}_\mu^{-1/2} \right)^{-1} \right\|_{\mathcal{L}(H)} &\leq \sum_{k=0}^{\infty} \left\| \left(\mathcal{C}_\mu^{-1/2} (\widehat{\mathcal{C}} - \mathcal{C}) \mathcal{C}_\mu^{-1/2} \right)^k \right\|_{\mathcal{L}(H)} \\ &\leq \sum_{k=0}^{\infty} \left\| \mathcal{C}_\mu^{-1/2} (\widehat{\mathcal{C}} - \mathcal{C}) \mathcal{C}_\mu^{-1/2} \right\|_{\mathcal{L}(H)}^k \\ &\leq \sum_{k=0}^{\infty} \left(\frac{1}{2} \right)^k \end{aligned}$$

by (D.55). The fact that $\sum_{k=0}^{\infty} (1/2)^k = (1 - 1/2)^{-1} = 2$ completes the proof. □

We may now prove Proposition D.10.

Proof of Proposition D.10. Combining (D.21) and Lemma D.11 (with c and N_0 as in the hypotheses there) shows that

$$V_N \leq 2\mu \operatorname{tr}\left(\mathcal{C}_\mu^{-1/2} \mathcal{C}' \mathcal{C}_\mu^{-1/2} \right) = 2 \sum_{j=1}^{\infty} \frac{\mu \sigma_j' \lambda_j}{\mu + \sigma_j \lambda_j} = 2 \sum_{j=1}^{\infty} \frac{\sigma_j' \lambda_j}{1 + N\gamma^{-2} \sigma_j \lambda_j} \tag{D.23}$$

with probability at least $1 - e^{-cN}$ if $N \geq N_0$. We used the assumed simultaneous diagonalizability of the prior and data covariance operators. Since $\sigma_j' \lesssim j^{-2\alpha'}$, $\sigma_j \asymp j^{-2\alpha}$, and $\lambda_j \asymp j^{-2p}$, all as $j \to \infty$, under Assumption 5.6, there exists $j_0 \in \mathbb{N}$ (independent of N) such that the rightmost expression in (D.23) is bounded above by

$$\begin{aligned} \sum_{j \leq j_0} \frac{\sigma_j' \lambda_j}{1 + N\gamma^{-2} \sigma_j \lambda_j} + \sum_{j > j_0} \frac{j^{-2(\alpha'+p)}}{1 + N j^{-2(\alpha+p)}} &\lesssim \frac{\gamma^2}{N} \sum_{j \leq j_0} \frac{\sigma_j'}{\sigma_j} + \sum_{j=1}^{\infty} \frac{j^{-2(\alpha'+p)}}{1 + N j^{-2(\alpha+p)}} \\ &\lesssim \sum_{j=1}^{\infty} \frac{j^{-2(\alpha'+p)}}{1 + N j^{-2(\alpha+p)}} . \end{aligned}$$

The last inequality in the preceding display follows from the fact that $1 + N \leq 2N$ and an argument similar to the one used in the proof of Lemma D.24. The proof is complete after an application of Lemma D.23 (with $t = 2(\alpha' + p) > 1$, $u = 2(\alpha + p) > 0$, and $v = 1$) yields the rightmost expression in (D.17). □

The proof of Proposition D.10 also justifies the claim made in Remark 5.14. Indeed, the second term on the right-hand side of the equality (5.33) is the posterior spread with respect to the weighted norm (5.18). Equation (D.21) upper bounds the posterior spread, which in turn upper bounds the variance (D.16b) in the bias–variance decomposition of (5.18). Thus, the variance and the posterior spread have the same upper bound.

D.3.2.2 Bounding the Bias

Recall from (D.16a) that the bias term is given by

$$B_N = \left\| (\Sigma')^{1/2} (\mathrm{Id}_H - A_N S_N) f^\dagger \right\|^2 .$$

In this appendix, we establish the following upper bound on B_N.

Proposition D.12 (bias upper bound)**.** *Let Assumptions 5.6, 5.7, and 5.9 hold. Let the bias B_N be as in (D.16a). There exists $c_0 > 8$, $c \in (0, 1/4)$, and $N_0 \geq 1$ (all independent of N and $f^\dagger$) such that for any $N \geq N_0$, it holds that*

$$B_N \leq 2 \sum_{j=1}^{\infty} \frac{\sigma'_j |f_j^\dagger|^2}{\left(1 + N\gamma^{-2}\sigma_j\lambda_j\right)^2} + c_0 \left\| f^\dagger \right\|_{\mathcal{H}^s}^2 \sum_{j=1}^{\infty} \frac{\sigma'_j \lambda_j}{1 + N\gamma^{-2}\sigma_j\lambda_j} \tag{D.24}$$

with probability at least $1 - 2\exp(-cN^{\min(1, \frac{\alpha+s-1}{\alpha+p})})$. On the same event, the variance bound (D.17) also holds true.

This bias bound is interesting because the second term in (D.24) is the same as the upper bound on the variance V_N in (D.17) (up to constant factors depending on $\|f^\dagger\|_{\mathcal{H}^s}$). Thus, as long as $f^\dagger$ is nonzero and not too small in norm, the total test error of the posterior mean estimator (i.e., the sum of bias and variance) is essentially dominated by the bias. Moreover, the hypotheses of Proposition D.12 do not require the true linear functional $f^\dagger$ to belong to the reproducing kernel Hilbert space of the prior $\mathcal{N}(0, \Lambda)$ (i.e., we allow for $\sum_{j=1}^{\infty} \lambda_j^{-1} |f_j^\dagger|^2 = \infty$). This is a significant advantage of our approach over related work; see Subsection 5.4.3.1 for related discussion.

The proof of Proposition D.12 is very lengthy. We break up the argument into several lemmas and steps. To set the stage, we instate the notation and definitions from Appendix D.3.2.1, in particular, the objects $\widehat{\mathcal{C}}$ from (D.18), μ and $\widehat{\mathcal{C}}_\mu$ from (D.19), and $\mathcal{C}'$, $\mathcal{C}$, and $\mathcal{C}_\mu$ from (D.20). We also use the shorthand notation

$$\widehat{T} \coloneqq \mathcal{C}_\mu^{-1/2}(\mathcal{C} - \widehat{\mathcal{C}})\mathcal{C}_\mu^{-1/2} \quad \text{and} \quad \widehat{M} \coloneqq (\mathrm{Id}_H - \widehat{T})^{-1} \tag{D.25}$$

for two random operators that appear frequently in the sequel.

We begin our analysis with a useful random series representation of the bias.

Lemma D.13 (bias: series)**.** *Under Assumption 5.6, B_N satisfies the identity*

$$B_N = \mu^2 \sum_{k=1}^{\infty} \frac{\sigma_k' \lambda_k}{\sigma_k \lambda_k + \mu} \left| \sum_{j=1}^{\infty} \frac{f_j^\dagger \lambda_j^{-1/2}}{(\sigma_j \lambda_j + \mu)^{1/2}} \langle \varphi_k, \widehat{M} \varphi_j \rangle \right|^2 . \tag{D.26}$$

Proof. By (5.14) and (5.16), we observe that

$$A_N S_N = \Lambda^{1/2} \widehat{\mathcal{C}}_\mu^{-1} \Lambda^{1/2} \widehat{\Sigma}$$

and hence

$$A_N S_N f^\dagger = \sum_{j=1}^{\infty} \Lambda^{1/2} \widehat{\mathcal{C}}_\mu^{-1} \Lambda^{1/2} \widehat{\Sigma} f_j^\dagger \varphi_j = \sum_{j=1}^{\infty} \Lambda^{1/2} \widehat{\mathcal{C}}_\mu^{-1} \widehat{\mathcal{C}} f_j^\dagger \lambda_j^{-1/2} \varphi_j \, .$$

We used the diagonalization of $\Lambda = \sum_j \lambda_j \varphi_j \otimes \varphi_j$ in the last equality. Noticing that

$$\mathrm{Id}_H - \widehat{\mathcal{C}}_\mu^{-1} \widehat{\mathcal{C}} = \widehat{\mathcal{C}}_\mu^{-1} \widehat{\mathcal{C}}_\mu - \widehat{\mathcal{C}}_\mu^{-1} \widehat{\mathcal{C}} = \mu \widehat{\mathcal{C}}_\mu^{-1} \, ,$$

we have the chain of equalities

$$\begin{aligned} f^\dagger - A_N S_N f^\dagger &= \sum_{j=1}^{\infty} \Big(f_j^\dagger \lambda_j^{-1/2} \Lambda^{1/2} \varphi_j - f_j^\dagger \lambda_j^{-1/2} \Lambda^{1/2} \widehat{\mathcal{C}}_\mu^{-1} \widehat{\mathcal{C}} \varphi_j \Big) \\ &= \sum_{j=1}^{\infty} f_j^\dagger \lambda_j^{-1/2} \Lambda^{1/2} (\mathrm{Id}_H - \widehat{\mathcal{C}}_\mu^{-1} \widehat{\mathcal{C}}) \varphi_j \\ &= \mu \sum_{j=1}^{\infty} f_j^\dagger \lambda_j^{-1/2} \Lambda^{1/2} \widehat{\mathcal{C}}_\mu^{-1} \varphi_j \, . \end{aligned}$$

Recalling $\widehat{M}$ from (D.25) and using the identity (D.58) from Lemma D.27 gives

$$\widehat{\mathcal{C}}_\mu^{-1} = \mathcal{C}_\mu^{-1/2} \big(\mathrm{Id}_H - \mathcal{C}_\mu^{-1/2} (\mathcal{C} - \widehat{\mathcal{C}}) \mathcal{C}_\mu^{-1/2} \big)^{-1} \mathcal{C}_\mu^{-1/2} = \mathcal{C}_\mu^{-1/2} \widehat{M} \mathcal{C}_\mu^{-1/2} \, .$$

Next, we expand in the shared orthonormal eigenbasis $\{\varphi_j\}$ of Λ and Σ to obtain

$$\begin{aligned}(\Sigma')^{1/2}(\mathrm{Id}_H - A_N S_N) f^\dagger &= \mu \sum_{j=1}^{\infty} \frac{f_j^\dagger \lambda_j^{-1/2}}{(\sigma_j \lambda_j + \mu)^{1/2}} (\Sigma')^{1/2} \Lambda^{1/2} \mathcal{C}_\mu^{-1/2} \widehat{M} \varphi_j \\ &= \mu \sum_{j=1}^{\infty} \frac{f_j^\dagger \lambda_j^{-1/2}}{(\sigma_j \lambda_j + \mu)^{1/2}} \sum_{k=1}^{\infty} \langle \varphi_k, (\Sigma')^{1/2} \Lambda^{1/2} \mathcal{C}_\mu^{-1/2} \widehat{M} \varphi_j \rangle \varphi_k \\ &= \mu \sum_{j=1}^{\infty} \frac{f_j^\dagger \lambda_j^{-1/2}}{(\sigma_j \lambda_j + \mu)^{1/2}} \sum_{k=1}^{\infty} \frac{(\sigma_k')^{1/2} \lambda_k^{1/2}}{(\sigma_k \lambda_k + \mu)^{1/2}} \langle \varphi_k, \widehat{M} \varphi_j \rangle \varphi_k \,.\end{aligned}$$

By continuity of the inner product,

$$\begin{aligned}\left\langle (\Sigma')^{1/2}(\mathrm{Id}_H - A_N S_N) f^\dagger, \varphi_i \right\rangle &= \mu \sum_{j=1}^{\infty} \frac{f_j^\dagger \lambda_j^{-1/2}}{(\sigma_j \lambda_j + \mu)^{1/2}} \frac{(\sigma_i')^{1/2} \lambda_i^{1/2}}{(\sigma_i \lambda_i + \mu)^{1/2}} \langle \varphi_i, \widehat{M} \varphi_j \rangle \\ &= \mu \frac{(\sigma_i')^{1/2} \lambda_i^{1/2}}{(\sigma_i \lambda_i + \mu)^{1/2}} \sum_{j=1}^{\infty} \frac{f_j^\dagger \lambda_j^{-1/2}}{(\sigma_j \lambda_j + \mu)^{1/2}} \langle \varphi_i, \widehat{M} \varphi_j \rangle \,.\end{aligned}$$

Summing the square of the preceding display over all $i \in \mathbb{N}$ completes the proof. □

Next, we note by direct calculation that $\widehat{M}$ from (D.25) satisfies the key identity

$$\widehat{M} = \mathrm{Id}_H + \widehat{M}\widehat{T} \,.$$

Thus, the right-hand side of the display (D.26) in Lemma D.13 is bounded above by

$$\begin{aligned}&2\mu^2 \sum_{k=1}^{\infty} \frac{\sigma_k' \lambda_k}{\sigma_k \lambda_k + \mu} \left| \sum_{j=1}^{\infty} \frac{|f_j^\dagger| \lambda_j^{-1/2}}{(\sigma_j \lambda_j + \mu)^{1/2}} \left| \langle \varphi_k, \varphi_j \rangle \right| \right|^2 \\ &\qquad + 2\mu^2 \sum_{k=1}^{\infty} \frac{\sigma_k' \lambda_k}{\sigma_k \lambda_k + \mu} \left| \sum_{j=1}^{\infty} \frac{|f_j^\dagger| \lambda_j^{-1/2}}{(\sigma_j \lambda_j + \mu)^{1/2}} \left| \langle \varphi_k, \widehat{M}\widehat{T} \varphi_j \rangle \right| \right|^2 \\ &= \underbrace{2\mu^2 \sum_{k=1}^{\infty} \frac{\sigma_k' |f_k^\dagger|^2}{(\sigma_k \lambda_k + \mu)^2}}_{=: \overline{B}_N} + \underbrace{2\mu^2 \sum_{k=1}^{\infty} \frac{\sigma_k' \lambda_k}{\sigma_k \lambda_k + \mu} \left| \sum_{j=1}^{\infty} \frac{|f_j^\dagger| \lambda_j^{-1/2}}{(\sigma_j \lambda_j + \mu)^{1/2}} \left| \langle \varphi_k, \widehat{M}\widehat{T} \varphi_j \rangle \right| \right|^2}_{=: \widehat{B}_N} .\end{aligned} \tag{D.27}$$

We used the fact that the $\{\varphi_j\}$ are orthonormal to obtain the equality. The first term $\overline{B}_N$ is the standard bias term one would expect from a simultaneously

diagonalizable linear inverse problem [76, 147]. The second term $\widehat{B}_N$ is a residual due to finite data. This is the term that we focus on estimating.

To this end, let E be the event from (D.55):

$$\mathsf{E} = \left\{ \left\| \widehat{T} \right\|_{\mathcal{L}(H)} \leq 1/2 \right\}. \tag{D.28}$$

Fix $\varepsilon > 0$ to be determined later. Define another event

$$\mathsf{E}_0 \coloneqq \left\{ \widehat{B}_N \leq \varepsilon \right\}. \tag{D.29}$$

Let the intersection $\mathsf{E}_0 \cap \mathsf{E}$ be our "good" event. On this event, our variance and bias bounds will hold simultaneously. For our results to be meaningful, we must show that $\mathsf{E}_0 \cap \mathsf{E}$ has high probability. Since $\mathbb{P}(\mathsf{E}_0 \,|\, \mathsf{E}) = 1 - \mathbb{P}(\mathsf{E}_0^{\mathsf{c}} \,|\, \mathsf{E})$, we have

$$\begin{aligned} \mathbb{P}(\mathsf{E}_0 \cap \mathsf{E}) &= \mathbb{P}(\mathsf{E}_0 \,|\, \mathsf{E})\, \mathbb{P}(\mathsf{E}) \\ &= \mathbb{P}(\mathsf{E}) - \mathbb{P}(\mathsf{E}_0^{\mathsf{c}} \,|\, \mathsf{E})\, \mathbb{P}(\mathsf{E}) \\ &= \mathbb{P}(\mathsf{E}) - \mathbb{P}(\mathsf{E}_0^{\mathsf{c}} \cap \mathsf{E}). \end{aligned} \tag{D.30}$$

Thus, to lower bound the probability of $\mathsf{E}_0 \cap \mathsf{E}$, it suffices to upper bound

$$\mathbb{P}(\mathsf{E}_0^{\mathsf{c}} \cap \mathsf{E}) = \mathbb{P}\big(\{\widehat{B}_N > \varepsilon\} \cap \mathsf{E}\big).$$

On the event E, it holds that $\|\widehat{M}\|_{\mathcal{L}(H)} \leq 2$ by (D.22). This, the symmetry of $\widehat{M}$, and the Cauchy–Schwarz inequality imply that

$$\begin{aligned} \widehat{B}_N &\leq 2\mu^2 \sum_{k=1}^{\infty} \frac{\sigma_k' \lambda_k}{\sigma_k \lambda_k + \mu} \left| \sum_{j=1}^{\infty} \frac{|f_j^\dagger| \lambda_j^{-1/2}}{(\sigma_j \lambda_j + \mu)^{1/2}} \big\| \widehat{M} \varphi_k \big\| \big\| \widehat{T} \varphi_j \big\| \right|^2 \\ &\leq 8 \underbrace{\left| \sum_{j=1}^{\infty} \frac{\mu^{1/2} |f_j^\dagger| \lambda_j^{-1/2}}{(\sigma_j \lambda_j + \mu)^{1/2}} \big\| \widehat{T} \varphi_j \big\| \right|^2}_{=: (\mathcal{I}_N)^2} \sum_{k=1}^{\infty} \frac{\mu \sigma_k' \lambda_k}{\sigma_k \lambda_k + \mu} \end{aligned}$$

on the event E. Notice that in the last line of the preceding display, the factor $(\mathcal{I}_N)^2$ is multiplying our high probability upper bound (D.17) on V_N. Hence, for the contribution from $\widehat{B}_N$ to be negligible relative to the upper bound on V_N, it suffices to show that $\mathcal{I}_N \lesssim 1$ for all sufficiently large N with high probability. Indeed, for some $\tau > 0$ to be determined later, choose

$$\varepsilon \coloneqq 8\tau^2 \sum_{k=1}^{\infty} \frac{\mu \sigma_k' \lambda_k}{\sigma_k \lambda_k + \mu} > 0 \tag{D.31}$$

in the definition (D.29) of E_0. Then the monotonicity of probability measure (i.e., if $\mathsf{A}_1 \subseteq \mathsf{A}_2$, then $\mathbb{P}(\mathsf{A}_1) \le \mathbb{P}(\mathsf{A}_2)$) implies that

$$\mathbb{P}(\mathsf{E}_0^{\mathsf{c}} \cap \mathsf{E}) \le \mathbb{P}(\{\mathcal{I}_N > \tau\} \cap \mathsf{E}) \le \mathbb{P}\{\mathcal{I}_N > \tau\}\,. \tag{D.32}$$

In the rest of the argument, we develop an upper tail bound for the random series $\mathcal{I}_N$ to control the rightmost expression in (D.32). To ease the notation, we write

$$\mathcal{I}_N = \sum_{j=1}^{\infty} s_j \big\|\widehat{T}\varphi_j\big\|\,, \quad \text{where} \quad s_j \coloneqq \frac{\mu^{1/2}|f_j^\dagger|\lambda_j^{-1/2}}{(\sigma_j\lambda_j+\mu)^{1/2}}\,. \tag{D.33}$$

Our strategy is to show that:[1]

Step 1. *the individual summands of $\mathcal{I}_N$ are subexponential random variables,*

Step 2. *the entire random series $\mathcal{I}_N$ is subexponential, and*

Step 3. *the entire random series $\mathcal{I}_N$ has a fast tail decay.*

We now proceed with this three step proof procedure.

Step 1. Recalling the definition of $\widehat{T}$ from (D.25), we see by the symmetry of $\mathcal{C}_\mu^{-1/2}$ and $\Lambda^{1/2}$ that

$$\begin{aligned}
-\widehat{T} &= \mathcal{C}_\mu^{-1/2}(\widehat{\mathcal{C}}-\mathcal{C})\mathcal{C}_\mu^{-1/2} \\
&= \mathcal{C}_\mu^{-1/2}\frac{1}{N}\sum_{n=1}^{N}\Big(\Lambda^{1/2}u_n \otimes \Lambda^{1/2}u_n - \mathbb{E}\big[\Lambda^{1/2}u_1 \otimes \Lambda^{1/2}u_1\big]\Big)\mathcal{C}_\mu^{-1/2} \\
&= \frac{1}{N}\sum_{n=1}^{N}\big(v_n \otimes v_n - \mathbb{E}[v_1 \otimes v_1]\big)\,, \quad \text{where} \quad v_n \coloneqq \mathcal{C}_\mu^{-1/2}\Lambda^{1/2}u_n
\end{aligned}$$

and $\{u_n\}_{n=1}^N \sim \nu^{\otimes N}$. Thus, it holds that

$$-\widehat{T}\varphi_j = \frac{1}{N}\sum_{n=1}^{N}\zeta_{jn}\,, \quad \text{where} \quad \zeta_{jn} \coloneqq \langle v_n, \varphi_j\rangle v_n - \mathbb{E}\big[\langle v_1,\varphi_j\rangle v_1\big]\,. \tag{D.34}$$

That is, $\widehat{T}\varphi_j$ is a sum of i.i.d. random vectors in the Hilbert space H. To show that the scalar random variables $\|\widehat{T}\varphi_j\|$ are subexponential, we first need to control the subexponential norm of the $\|\zeta_{jn}\|$. The next lemma accomplishes this task.

[1] Appendix D.3.1 reviews subgaussian and subexponential random variables.

Lemma D.14 (moments)**.** *Under Assumption 5.6 and 5.7, for every j it holds that*

$$\mathbb{E}\|\zeta_{j1}\|^\ell \le \left(4em^2\rho_j\sqrt{\operatorname{tr}\big(\mathcal{C}_\mu^{-1}\mathcal{C}\big)}\right)^\ell \ell! \quad \textit{for all} \quad \ell \in \{2,3,\dots\}\,, \tag{D.35}$$

where $m \ge 0$ is as in (5.23) and

$$\rho_j \coloneqq \sqrt{\frac{\sigma_j\lambda_j}{\mu+\sigma_j\lambda_j}}\,. \tag{D.36}$$

Proof. Fix an integer $\ell \ge 2$. The inequality $|a+b|^\ell \le 2^{\ell-1}(|a|^\ell + |b|^\ell)$ shows that

$$\begin{aligned}\mathbb{E}\|\zeta_{j1}\|^\ell &\le 2^{\ell-1}\big(\mathbb{E}\|\langle v_1,\varphi_j\rangle v_1\|^\ell + \|\mathbb{E}[\langle v_1,\varphi_j\rangle v_1]\|^\ell\big)\\ &\le 2^\ell\,\mathbb{E}\|\langle v_1,\varphi_j\rangle v_1\|^\ell\,.\end{aligned}$$

The second line is due to Jensen's inequality. Let u be an i.i.d. copy of $u_1 \sim \nu$, so that $v \coloneqq \mathcal{C}_\mu^{-1/2}\Lambda^{1/2}u$ is an i.i.d. copy of v_1. By Assumption 5.7 and the assumption (A1) that Λ and Σ share the orthonormal eigenbasis $\{\varphi_j\}$, it holds that

$$v = \sum_{j=1}^\infty \rho_j z_j \varphi_j\,, \quad \text{where} \quad \rho_j = \sqrt{\frac{\sigma_j\lambda_j}{\mu+\sigma_j\lambda_j}} \ge 0\,.$$

Thus, $\langle v,\varphi_j\rangle = \rho_j z_j$ and $\mathbb{E}\|\langle v_1,\varphi_j\rangle v_1\|^\ell = \mathbb{E}\|\langle v,\varphi_j\rangle v\|^\ell$ equals

$$\begin{aligned}\rho_j^\ell\,\mathbb{E}\big[|z_j|^\ell\|v\|^\ell\big] &= \rho_j^\ell\,\mathbb{E}\big[|z_j|^\ell\big(\|v\|^2\big)^{\ell/2}\big]\\ &= \rho_j^\ell\,\mathbb{E}\left[|z_j|^\ell\left(\sum_{k=1}^\infty \rho_k^2 z_k^2\right)^{\ell/2}\right]\\ &= \rho_j^\ell\,\mathbb{E}\left|\sum_{k=1}^\infty \rho_k^2 z_j^2 z_k^2\right|^{\ell/2}.\end{aligned}$$

The triangle inequality in the Banach space $L_{\mathbb{P}}^{\ell/2}(\Omega;\mathbb{R})$ (since $\ell \ge 2$) and the Cauchy–Schwarz inequality yields

$$\begin{aligned}\left\|\sum_{k=1}^\infty \rho_k^2 z_j^2 z_k^2\right\|_{L_{\mathbb{P}}^{\ell/2}} &\le \sum_{k=1}^\infty \rho_k^2 \big\|z_j^2 z_k^2\big\|_{L_{\mathbb{P}}^{\ell/2}}\\ &= \sum_{k=1}^\infty \rho_k^2\Big(\mathbb{E}\big[|z_j|^\ell|z_k|^\ell\big]\Big)^{2/\ell}\\ &\le \sum_{k=1}^\infty \rho_k^2\big(\mathbb{E}|z_j|^{2\ell}\big)^{1/\ell}\big(\mathbb{E}|z_k|^{2\ell}\big)^{1/\ell}.\end{aligned}$$

By the definition (D.12) of the subexponential norm, Lemma D.8, and (5.23), we have

$$\begin{aligned}\sum_{k=1}^{\infty} \rho_k^2 \big(\mathbb{E}|z_j|^{2\ell}\big)^{1/\ell} \big(\mathbb{E}|z_k|^{2\ell}\big)^{1/\ell} &\le \sum_{k=1}^{\infty} \rho_k^2 \big(\ell \|z_j^2\|_{\psi_1}\big) \big(\ell \|z_k^2\|_{\psi_1}\big) \\ &\le \sum_{k=1}^{\infty} \rho_k^2 \big(2\ell \|z_j\|_{\psi_2}^2\big) \big(2\ell \|z_k\|_{\psi_2}^2\big) \\ &\le 4\ell^2 m^4 \sum_{k=1}^{\infty} \frac{\sigma_k \lambda_k}{\mu + \sigma_k \lambda_k} \,.\end{aligned}$$

Taking the $\ell/2$-th power and putting together the pieces, we deduce that

$$\mathbb{E}\|\zeta_{j1}\|^{\ell} \le (2\rho_j)^{\ell} \left(2m^2 \sqrt{\operatorname{tr}\big(\mathcal{C}_{\mu}^{-1}\mathcal{C}\big)}\right)^{\ell} \ell^{\ell} \,.$$

Recalling from Stirling's formula that $(\ell/e)^{\ell} \le \ell!$ completes the proof. □

We need the following proposition that relies heavily on [219, Theorem 1, p. 144].

Proposition D.15 (Hilbert space norm of independent sums)**.** *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of independent centered random vectors with values in a separable Hilbert space $(\mathcal{X}, \langle\cdot,\cdot\rangle, \|\cdot\|)$. Let $N \in \mathbb{N}$ be arbitrary. If the Bernstein moment condition*

$$\sum_{n=1}^{N} \mathbb{E}\|X_n\|^{\ell} \le \frac{1}{2} \ell! \sigma^2 b^{\ell-2} \quad \text{for all} \quad \ell \in \{2, 3, \ldots\} \tag{D.37}$$

holds for some $\sigma > 0$ and $b > 0$, then the partial sums $\mathcal{S}_N := \sum_{n=1}^{N} X_n$ satisfy the subexponential condition $\|\mathcal{S}_N\| \in \mathrm{SE}(2\sigma^2, 2b)$, that is,

$$\mathbb{E}\, e^{\lambda(\|\mathcal{S}_N\| - \mathbb{E}\|\mathcal{S}_N\|)} \le e^{\lambda^2\sigma^2} \quad \text{for all} \quad |\lambda| \le \frac{1}{2b} \,. \tag{D.38}$$

Proof. Since $\mathcal{X}$ is a separable Banach space, [219, Theorem 1, p. 144] shows that

$$\mathbb{E}\, e^{|\lambda|(\|\mathcal{S}_N\| - \mathbb{E}\|\mathcal{S}_N\|)} \le \exp\left(\sum_{n=1}^{N} \mathbb{E}\Big[e^{|\lambda|\|X_n - \mathbb{E}\,X_n\|} - 1 - |\lambda|\|X_n - \mathbb{E}\,X_n\|\Big]\right)$$

for all $\lambda \in \mathbb{R}$. Under the Bernstein moment condition (D.37), we compute using the Taylor expansion of the exponential function that

$$\begin{aligned}\sum_{n=1}^{N} \mathbb{E}\Big[e^{|\lambda|\|X_n - \mathbb{E}\,X_n\|} - 1 - |\lambda|\|X_n - \mathbb{E}\,X_n\|\Big] &= \sum_{\ell=2}^{\infty}\sum_{n=1}^{N} \frac{|\lambda|^\ell\,\mathbb{E}\|X_n - \mathbb{E}\,X_n\|^\ell}{\ell!} \\ &\le \frac{1}{2}\lambda^2\sigma^2 \sum_{\ell=2}^{\infty} \big(|\lambda| b\big)^{\ell-2} \\ &= \frac{\lambda^2\sigma^2}{2(1-|\lambda| b)} \\ &\le \lambda^2\sigma^2\end{aligned}$$

provided that $|\lambda| b \le 1/2$. Thus, it holds that

$$\mathbb{E}\, e^{|\lambda|(\|\mathcal{S}_N\| - \mathbb{E}\|\mathcal{S}_N\|)} \le e^{\lambda^2\sigma^2} \quad \text{for all} \quad |\lambda| \le 1/(2b)\,.$$

For $\|\mathcal{S}_N\|$ to be subexponential, by the definition (D.11) we also need to show that

$$\mathbb{E}\, e^{-|\lambda|(\|\mathcal{S}_N\| - \mathbb{E}\|\mathcal{S}_N\|)} \le e^{\lambda^2\sigma^2} \quad \text{for all} \quad |\lambda| \le 1/(2b)\,.$$

But since $\|\mathcal{S}_N\| \ge 0$ a.s., the one-sided Bernstein moment generating function bound [267, Proposition 2.14, Equation (2.22a), p. 31] applied to $-\|\mathcal{S}_N\|$ yields

$$\mathbb{E}\, e^{-|\lambda|(\|\mathcal{S}_N\| - \mathbb{E}\|\mathcal{S}_N\|)} \le e^{\lambda^2\,\mathbb{E}\|\mathcal{S}_N\|^2/2} \quad \text{for all} \quad |\lambda| < \infty\,.$$

It remains to bound $\mathbb{E}\|\mathcal{S}_N\|^2$ in terms of σ^2. Using the facts that $\mathbb{E}\,X_n = 0$ and $\mathcal{X}$ is a Hilbert space yield

$$\mathbb{E}\|\mathcal{S}_N\|^2 = \sum_{n=1}^{N}\sum_{n'=1}^{N} \mathbb{E}\langle X_n, X_{n'}\rangle = \sum_{n=1}^{N} \mathbb{E}\|X_n\|^2 \le \sigma^2\,.$$

We used the Bernstein moment condition (D.37) with $\ell = 2$ to obtain the final inequality. Noting that $|\lambda| < \infty$ implies $|\lambda| \le 1/(2b)$ and that $\sigma^2/2 \le \sigma^2$ completes the proof. □

We are now in a position to prove the following lemma about the empirical sums.

Lemma D.16 (norm of empirical sum is subexponential)**.** *Fix $j \in \mathbb{N}$. Let*

$$\varsigma_j \coloneqq 8em^2\rho_j\sqrt{\operatorname{tr}\big(\mathcal{C}_\mu^{-1}\mathcal{C}\big)} \tag{D.39}$$

with ρ_j as in (D.36). Under Assumptions 5.6 and 5.7, it holds that

$$\big\|\widehat{T}\varphi_j\big\| \in \mathrm{SE}\bigg(\frac{2\varsigma_j^2}{N}, \frac{2\varsigma_j}{N}\bigg)\,. \tag{D.40}$$

Proof. For fixed j, the independence of the $\{\zeta_{jn}\}_{n=1}^N$ and Lemma D.14 imply that

$$\begin{aligned}\sum_{n=1}^N \mathbb{E}\|\zeta_{jn}\|^\ell = N\,\mathbb{E}\|\zeta_{j1}\|^\ell &\le 2\left(\frac{N}{2}\right)\left(4em^2\rho_j\sqrt{\operatorname{tr}\big(\mathcal{C}_\mu^{-1}\mathcal{C}\big)}\right)^\ell \ell! \\ &\le \frac{N}{2}\ell!\varsigma_j^\ell \\ &= \frac{1}{2}\ell!\big(N\varsigma_j^2\big)\varsigma_j^{\ell-2}\end{aligned}$$

for any $\ell \in \{2,3,\ldots\}$ (using $2 \le 2^\ell$ to get to the second line). Recalling from (D.34) that $\mathbb{E}\,\zeta_{j1} = 0$, Proposition D.15 applied with $\sigma^2 = N\varsigma_j^2$ and $b = \varsigma_j$ in (D.37) yields

$$\mathbb{E}\,e^{N\lambda(\|\widehat{T}\varphi_j\| - \mathbb{E}\|\widehat{T}\varphi_j\|)} \le e^{\lambda^2 N\varsigma_j^2} \quad \text{for all} \quad |\lambda| \le \frac{1}{2\varsigma_j}\,.$$

Replacing λ with λ/N and recalling Definition D.7 gives the asserted result. □

Step 2. The $\ell^1(\mathbb{N}) \coloneqq \ell^1(\mathbb{N};\mathbb{R})$ norm of the nonnegative sequence $\{s_j\varsigma_j\}_{j\in\mathbb{N}}$ plays a central role in the analysis to follow. To this end, denote

$$w^{(N)} \coloneqq \{w_j^{(N)}\}_{j\in\mathbb{N}}\,, \quad \text{where} \quad w_j^{(N)} \coloneqq s_j\varsigma_j \quad \text{for all} \quad j \in \mathbb{N}\,. \tag{D.41}$$

We now upper bound the deterministic series $\|w^{(N)}\|_{\ell^1(\mathbb{N})}$.

Lemma D.17 (deterministic series convergence rate)**.** *Under Assumption 5.6 and 5.9, it holds that*

$$\frac{1}{N}\big\|w^{(N)}\big\|_{\ell^1(\mathbb{N})}^2 \lesssim \big\|f^\dagger\big\|_{\mathcal{H}^s}^2 \times \begin{cases} N^{-\left(\frac{\alpha+s-1}{\alpha+p}\right)}\,, & \text{if } \frac{\alpha+s-1/2}{\alpha+p} < 2\,, \\ N^{-\left(1+\frac{\alpha+p-1/2}{\alpha+p}\right)}\log 2N\,, & \text{if } \frac{\alpha+s-1/2}{\alpha+p} = 2\,, \\ N^{-\left(1+\frac{\alpha+p-1/2}{\alpha+p}\right)}\,, & \text{if } \frac{\alpha+s-1/2}{\alpha+p} > 2 \end{cases} \tag{D.42}$$

for all $N \in \mathbb{N}$.

Proof. Recalling the definitions of $\{s_j\}_{j\in\mathbb{N}}$ (D.33) and $\{\varsigma_j\}_{j\in\mathbb{N}}$ (D.39), we have

$$\begin{aligned}\big\|w^{(N)}\big\|_{\ell^1(\mathbb{N})} = \sum_{j=1}^\infty s_j\varsigma_j &= 8em^2\sqrt{\operatorname{tr}\big(\mathcal{C}_\mu^{-1}\mathcal{C}\big)}\sum_{j=1}^\infty \frac{\mu^{1/2}|f_j^\dagger|\sigma_j^{1/2}}{\mu+\sigma_j\lambda_j} \\ &\simeq N^{1/2}\sqrt{\operatorname{tr}\big(\mathcal{C}_\mu^{-1}\mathcal{C}\big)}\sum_{j=1}^\infty \frac{|f_j^\dagger|\sigma_j^{1/2}}{1+N\sigma_j\lambda_j}\,.\end{aligned}$$

The preceding display, the asymptotics of $\{\sigma_j\}_{j\in\mathbb{N}}$ and $\{\lambda_j\}_{j\in\mathbb{N}}$ from Assumption 5.6, and the Cauchy–Schwarz inequality imply that

$$\frac{1}{N}\big\|w^{(N)}\big\|_{\ell^1(\mathbb{N})}^2 \lesssim \big\|f^\dagger\big\|_{\mathcal{H}^s}^2 \operatorname{tr}\big(\mathcal{C}_\mu^{-1}\mathcal{C}\big)\frac{1}{N^2} + \operatorname{tr}\big(\mathcal{C}_\mu^{-1}\mathcal{C}\big)\left|\sum_{j=1}^{\infty}\frac{|f_j^\dagger| j^{-\alpha}}{1+Nj^{-2(\alpha+p)}}\right|^2 .$$

By Lemma D.24, it holds that $\operatorname{tr}(\mathcal{C}_\mu^{-1}\mathcal{C}) \simeq N^{1/(2(\alpha+p))}$ because $\mu = \gamma^2/N$ (D.19). The series factor in the second term in the preceding display satisfies the estimate

$$\left|\sum_{j=1}^{\infty}\frac{|f_j^\dagger| j^{-\alpha}}{1+Nj^{-2(\alpha+p)}}\right|^2 \le \big\|f^\dagger\big\|_{\mathcal{H}^s}^2 \sum_{j=1}^{\infty}\frac{j^{-2(\alpha+s)}}{(1+Nj^{-2(\alpha+p)})^2}$$

by the Cauchy–Schwarz inequality. The rightmost series converges (because $2(\alpha+s) > 2 > 1$ by the last assertion of Assumption 5.9) and is bounded above by a constant (independent of N) times

$$\begin{cases} N^{-\left(\frac{\alpha+s-1/2}{\alpha+p}\right)}, & \text{if } \frac{\alpha+s-1/2}{\alpha+p} < 2\,, \\ N^{-2}\log 2N\,, & \text{if } \frac{\alpha+s-1/2}{\alpha+p} = 2\,, \\ N^{-2}\,, & \text{if } \frac{\alpha+s-1/2}{\alpha+p} > 2 \end{cases}$$

by Lemma D.23 (applied with $t = 2(\alpha+s) > 1$, $u = 2(\alpha+p) > 1 > 0$, and $v = 2$). Bounding N^{-2} above by the preceding display and multiplying this bound by the $N^{1/(2(\alpha+p))}$ trace bound completes the proof. □

Combining the previous results with those of [76, Appendix B, pp. 30–31] establishes that the entire series $\mathcal{I}_N$ (D.33) is a real-valued subexponential random variable.

Lemma D.18 (random series: subexponential)**.** *Let Assumptions 5.6, 5.7, and 5.9 be satisfied and $\mathcal{I}_N$ be defined as in (D.33). It holds that*

$$\mathcal{I}_N \in \mathrm{SE}\left(\frac{2}{N}\big\|w^{(N)}\big\|_{\ell^1(\mathbb{N})}^2, \frac{2}{N}\big\|w^{(N)}\big\|_{\ell^1(\mathbb{N})}\right) \quad \textit{for all} \quad N\in\mathbb{N}\,. \tag{D.43}$$

Proof. Fix $N \in \mathbb{N}$. A change of variables in Definition D.7 of subexponential and Lemma D.16 imply that $s_j\|\widehat{T}\varphi_j\| \in \mathrm{SE}(2s_j^2\varsigma_j^2/N, 2s_j\varsigma_j/N)$ for any j. For any $J \in \mathbb{N}$, let $\mathcal{I}_N^{(J)} := \sum_{j=1}^{J} s_j\|\widehat{T}\varphi_j\|$. Even though the summands

$\{s_j\|\widehat{T}\varphi_j\|\}_{j\in\mathbb{N}}$ are a dependent sequence of random variables, [76, Lemma B.5, p. 30] shows that

$$\mathcal{I}_N^{(J)} \in \mathrm{SE}\left(\frac{2}{N}\left|\sum_{j=1}^{J} s_j\varsigma_j\right|^2, \frac{2}{N}\sum_{j=1}^{J} s_j\varsigma_j\right).$$

Next, Jensen's inequality yields the bound

$$\sum_{j=1}^{\infty}\mathbb{E}\big[s_j\|\widehat{T}\varphi_j\|\big] \le \sum_{j=1}^{\infty} s_j\sqrt{\mathbb{E}\|\widehat{T}\varphi_j\|^2}\,.$$

For fixed $j\in\mathbb{N}$, we compute

$$\begin{aligned}\mathbb{E}\big\|\widehat{T}\varphi_j\big\|^2 = \frac{1}{N^2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\mathbb{E}\langle\zeta_{jn},\zeta_{jn'}\rangle &= \frac{1}{N}\,\mathbb{E}\|\zeta_{j1}\|^2\\ &\le \frac{2}{N}\Big(4em^2\rho_j\sqrt{\mathrm{tr}(\mathcal{C}_\mu^{-1}\mathcal{C})}\Big)^2\\ &= \frac{\varsigma_j^2}{2N}\,.\end{aligned}$$

In the preceding display, we used orthogonality, independence, and the Bernstein moment condition (D.35) (Lemma D.14 applied with $\ell=2$). Thus,

$$\sum_{j=1}^{\infty}\mathbb{E}\Big[s_j\big\|\widehat{T}\varphi_j\big\|\Big] \le \frac{1}{\sqrt{2N}}\sum_{j=1}^{\infty} s_j\varsigma_j\,. \tag{D.44}$$

Since the right-hand side of (D.44) is finite by Lemma D.17, a monotone convergence argument [76, Lemma B.3, p. 30] shows that

$$\mathbb{P}\Big\{\lim_{J\to\infty}\mathcal{I}_N^{(J)} = \mathcal{I}_N\Big\} = 1\,.$$

Using this almost sure convergence and again recalling Definition D.7, it holds that

$$\begin{aligned}\mathbb{E}\exp\big(\lambda(\mathcal{I}_N - \mathbb{E}\,\mathcal{I}_N)\big) &= \mathbb{E}\lim_{J\to\infty}\exp\Big(\lambda\big(\mathcal{I}_N^{(J)} - \mathbb{E}\,\mathcal{I}_N^{(J)}\big)\Big)\\ &\le \liminf_{J\to\infty}\mathbb{E}\exp\Big(\lambda\big(\mathcal{I}_N^{(J)} - \mathbb{E}\,\mathcal{I}_N^{(J)}\big)\Big)\\ &\le \liminf_{J\to\infty}\exp\left(\frac{\lambda^2}{2}\frac{2}{N}\left|\sum_{j=1}^{J} s_j\varsigma_j\right|^2\right)\\ &= \exp\Big(\frac{\lambda^2}{2}\frac{2}{N}\big\|w^{(N)}\big\|_{\ell^1(\mathbb{N})}^2\Big)\end{aligned}$$

for all $|\lambda| \le (\frac{2}{N}\|w^{(N)}\|_{\ell^1})^{-1}$ because $\frac{2}{N}\sum_{j=1}^{J} w_j^{(N)} \le \frac{2}{N}\|w^{(N)}\|_{\ell^1}$ for any $J\in\mathbb{N}$. In the preceding display, the first line is due to the identity $\mathbb{E}\lim_J \mathcal{I}_N^{(J)} =$

$\lim_J \mathbb{E}\,\mathcal{I}_N^{(J)}$ (which follows from monotone convergence), the second due to Fatou's lemma, and the third due to the fact that $\mathcal{I}_N^{(J)}$ is subexponential (hence (D.11) holds). Therefore, the entire random series $\mathcal{I}_N$ satisfies Definition D.7 and the proof is complete. □

Step 3. A consequence of the previous lemma is a strong tail decay bound for $\mathcal{I}_N$.

Lemma D.19 (random series: tail bound)**.** *Let Assumptions 5.6, 5.7, and 5.9 be satisfied and $\mathcal{I}_N$ be defined as in (D.33). For any $N \in \mathbb{N}$, it holds that*

$$\mathbb{P}\left\{\mathcal{I}_N \geq \frac{\|w^{(N)}\|_{\ell^1(\mathbb{N})}}{\sqrt{2N}} + t\right\} \leq \exp\left(-\frac{Nt^2}{4\|w^{(N)}\|_{\ell^1(\mathbb{N})}^2}\right) \tag{D.45}$$

for all $0 \leq t \leq \|w^{(N)}\|_{\ell^1(\mathbb{N})}$.

Proof. By Lemma D.18 and [267, Proposition 2.9, p. 26], it holds that

$$\mathbb{P}\{\mathcal{I}_N \geq \mathbb{E}\,\mathcal{I}_N + t\} \leq \exp\left(-\frac{Nt^2}{4\|w^{(N)}\|_{\ell^1(\mathbb{N})}^2}\right) \quad \text{for all} \quad 0 \leq t \leq \left\|w^{(N)}\right\|_{\ell^1(\mathbb{N})}.$$

By the Fubini–Tonelli theorem and (D.44), we obtain

$$\mathbb{E}\,\mathcal{I}_N = \sum_{j=1}^{\infty} \mathbb{E}\left[s_j \left\|\widehat{T}\varphi_j\right\|\right] \leq \frac{\|w^{(N)}\|_{\ell^1(\mathbb{N})}}{\sqrt{2N}}.$$

The assertion (D.45) follows from the monotonicity of probability measure. □

The previous lemma implies a high probability uniform upper bound on $\mathcal{I}_N$.

Lemma D.20 (random series: uniform bound)**.** *Let Assumptions 5.6, 5.7, and 5.9 be satisfied and $\mathcal{I}_N$ be defined as in (D.33). There exists $c_0 > 1$ and $c \in (0, 1/4)$, both independent of N and $f^\dagger$, such that*

$$\mathbb{P}\left\{\mathcal{I}_N \geq c_0 \|f^\dagger\|_{\mathcal{H}^s}\right\} \leq \exp\left(-cN^{\min\left(1, \frac{\alpha+s-1}{\alpha+p}\right)}\right) \quad \textit{for all} \quad N \in \mathbb{N}. \tag{D.46}$$

Proof. Let $t := \min(\|f^\dagger\|_{\mathcal{H}^s}, \|w^{(N)}\|_{\ell^1})$. Clearly $0 \leq t \leq \|w^{(N)}\|_{\ell^1}$. Also, $t \leq \|f^\dagger\|_{\mathcal{H}^s}$ so that monotonicity of probability measure yields

$$\begin{aligned}
\mathbb{P}\left\{\mathcal{I}_N \geq \frac{\|w^{(N)}\|_{\ell^1}}{\sqrt{2N}} + \|f^\dagger\|_{\mathcal{H}^s}\right\} &\leq \mathbb{P}\left\{\mathcal{I}_N \geq \frac{\|w^{(N)}\|_{\ell^1}}{\sqrt{2N}} + t\right\} \\
&\leq \exp\left(-\frac{N \min\left(\|f^\dagger\|_{\mathcal{H}^s}^2, \|w^{(N)}\|_{\ell^1}^2\right)}{4\|w^{(N)}\|_{\ell^1}^2}\right) \\
&= \begin{cases} e^{-N\|f^\dagger\|_{\mathcal{H}^s}^2/(4\|w^{(N)}\|_{\ell^1}^2)}, & \text{if } \|w^{(N)}\|_{\ell^1} \geq \|f^\dagger\|_{\mathcal{H}^s}, \\ e^{-N/4}, & \text{if } \|w^{(N)}\|_{\ell^1} < \|f^\dagger\|_{\mathcal{H}^s}. \end{cases}
\end{aligned}$$

The second inequality is due to Lemma D.19. Next, we upper bound the probability in the first case $\|w^{(N)}\|_{\ell^1} \geq \|f^\dagger\|_{\mathcal{H}^s}$ by bounding $N^{-1}\|w^{(N)}\|_{\ell^1}^2$ from above. To do so, let $\delta \coloneqq (\alpha + p - \frac{1}{2})/(\alpha + p)$. Then $\delta > 0$ because $\alpha + p > 1/2$ by Assumption 5.6. Clearly $N^{-\delta} \leq 1$. Since $x \mapsto \log 2x$ is slowly varying, $\lim_{N\to\infty} N^{-\delta} \log 2N = 0$. Hence, $\sup_{N\in\mathbb{N}} N^{-\delta} \log 2N < \infty$. Lemma D.17 then yields

$$\begin{aligned}
\frac{1}{N}\big\|w^{(N)}\big\|_{\ell^1}^2 &\lesssim \|f^\dagger\|_{\mathcal{H}^s}^2 \times \begin{cases} N^{-\left(\frac{\alpha+s-1}{\alpha+p}\right)}, & \text{if } \frac{\alpha+s-1/2}{\alpha+p} < 2, \\ N^{-1}\big(N^{-\delta}\log 2N\big), & \text{if } \frac{\alpha+s-1/2}{\alpha+p} = 2, \\ N^{-1}N^{-\delta}, & \text{if } \frac{\alpha+s-1/2}{\alpha+p} > 2 \end{cases} \\
&\lesssim \|f^\dagger\|_{\mathcal{H}^s}^2 \times \begin{cases} N^{-\left(\frac{\alpha+s-1}{\alpha+p}\right)}, & \text{if } \frac{\alpha+s-1/2}{\alpha+p} < 2, \\ N^{-1}, & \text{if } \frac{\alpha+s-1/2}{\alpha+p} \geq 2 \end{cases} \\
&\leq \|f^\dagger\|_{\mathcal{H}^s}^2 N^{-\min\left(1, \frac{\alpha+s-1}{\alpha+p}\right)}
\end{aligned}$$

for all $N \in \mathbb{N}$. It follows that there exists $c' > 0$ such that

$$\begin{aligned}
\exp\left(-\frac{N\|f^\dagger\|_{\mathcal{H}^s}^2}{4\|w^{(N)}\|_{\ell^1}^2}\right) &= \exp\left(-\frac{\|f^\dagger\|_{\mathcal{H}^s}^2}{4\|w^{(N)}\|_{\ell^1}^2/N}\right) \\
&\leq \exp\left(-c' N^{\min\left(1, \frac{\alpha+s-1}{\alpha+p}\right)}/4\right).
\end{aligned}$$

For the second case $\|w^{(N)}\|_{\ell^1} < \|f^\dagger\|_{\mathcal{H}^s}$, taking the minimum yields the bound

$$e^{-N/4} \leq \exp\left(-N^{\min\left(1, \frac{\alpha+s-1}{\alpha+p}\right)}/4\right).$$

Writing $c \coloneqq \min(1, c')/4 \in (0, 1/4)$, it follows that

$$\max\left(e^{-N\|f^\dagger\|_{\mathcal{H}^s}^2/(4\|w^{(N)}\|_{\ell^1}^2)}, e^{-N/4}\right) \leq \exp\left(-cN^{\min\left(1, \frac{\alpha+s-1}{\alpha+p}\right)}\right)$$

for all $N \in \mathbb{N}$. The right-hand side of the preceding display is thus an upper bound to $\mathbb{P}\{\mathcal{I}_N \geq \|w^{(N)}\|_{\ell^1}/\sqrt{2N} + \|f^\dagger\|_{\mathcal{H}^s}\}$. To complete the proof, notice that $\|w^{(N)}\|_{\ell^1(\mathbb{N})}/\sqrt{2N} \leq (c'/2)\|f^\dagger\|_{\mathcal{H}^s} N^{-\min(1,(\alpha+s-1)/(\alpha+p))/2} \leq (c'/2)\|f^\dagger\|_{\mathcal{H}^s}$ because $\alpha + s > 1$ by Assumption 5.9. By monotonicity of probability measure, this implies the asserted result (D.46) with $c_0 \coloneqq 1 + c'/2$. □

This completes **Step 1.**, **Step 2.**, and **Step 3.** With a uniform upper bound on $\mathcal{I}_N$ in hand from the previous lemma, the claimed bound on the bias in Proposition D.12 follows easily. The details are provided in the following proof.

Proof of Proposition D.12. Recalling the event E_0 from (D.29) with ε as in (D.31), choose $\tau = c'\|f^\dagger\|_{\mathcal{H}^s}$ with $c' > 1$ equal to the constant c_0 in the hypotheses of Lemma D.20. On the good event $\mathsf{E}_0 \cap \mathsf{E}$ from (D.29) and (D.28), it holds by (D.27) that $B_N \leq \overline{B}_N + \varepsilon$ which is precisely the claimed upper bound (D.24) with $c_0 \coloneqq 8(c')^2$. It remains to lower bound the probability of $\mathsf{E}_0 \cap \mathsf{E}$. By (D.30) and (D.32), it holds that

$$\mathbb{P}(\mathsf{E}_0 \cap \mathsf{E}) = \mathbb{P}(\mathsf{E}) - \mathbb{P}(\mathsf{E}_0^{\mathsf{c}} \cap \mathsf{E}) \geq \mathbb{P}(\mathsf{E}) - \mathbb{P}(\mathcal{I}_N > c'\|f^\dagger\|_{\mathcal{H}^s}) \,.$$

By hypothesis, the assertion of Lemma D.25 (i.e., that $\mathbb{P}(\mathsf{E}) \geq 1 - e^{-c_1 N}$ for some $c_1 \in (0,1)$) holds true provided that $N \geq N_0$ with $N_0 \geq 1$ defined in (D.54). Combining this with Lemma D.20 shows that, for all $N \geq N_0$, the good set satisfies

$$\begin{aligned}\mathbb{P}(\mathsf{E}_0 \cap \mathsf{E}) &\geq 1 - e^{-c_1 N} - \exp\Big(-c_2 N^{\min\left(1, \frac{\alpha+s-1}{\alpha+p}\right)}\Big) \\ &\geq 1 - 2\exp\Big(-c N^{\min\left(1, \frac{\alpha+s-1}{\alpha+p}\right)}\Big)\end{aligned}$$

for some $c_2 \in (0, 1/4)$. We defined $c \coloneqq \min(c_1, c_2) \in (0, 1/4)$ in the last line of the preceding display. To complete the proof, notice that if $\mathsf{E}_0 \cap \mathsf{E}$ occurs, then E also occurs. But the variance bound (D.17) also holds true on E (as shown in the proof of Proposition D.10). This proves the final assertion of the proposition. □

D.3.2.3 Proof of Theorem D.1

Combining the bias and variance bounds leads to the main theoretical result for end-to-end learning, Theorem D.1, which we now prove.

Theorem D.1 (end-to-end learning: general convergence rate)**.** *Let the input training data distribution ν, the test data distribution ν', and the Gaussian prior $\mathcal{N}(0, \Lambda)$ satisfy Assumptions 5.6 and 5.7. Let the ground truth linear functional $f^\dagger \in \mathcal{H}^s$ satisfy Assumption 5.9. Let α, α', and p be as in (5.20) and (5.21). Then there exists $c \in (0, 1/4)$ and $N_0 \geq 1$ such that for any $N \geq N_0$, the mean $\bar{f}^{(N)}$ of the Gaussian posterior distribution (5.15) arising from the N pairs of observed training data (U, Y) in (5.17) satisfies the error bound*

$$\mathbb{E}^{Y\,|\,U}\,\mathbb{E}^{u' \sim \nu'}\big|\langle f^\dagger, u'\rangle - \langle \bar{f}^{(N)}, u'\rangle\big|^2 \lesssim \big(1 + \|f^\dagger\|_{\mathcal{H}^s}^2\big)\varepsilon_N^2 \tag{D.1}$$

with probability at least $1 - 2\exp(-cN^{\min(1, \frac{\alpha+s-1}{\alpha+p})})$ *over* $U \sim \nu^{\otimes N}$*, where*

$$\varepsilon_N^2 \coloneqq \begin{cases} N^{-\min\left(\frac{\alpha'+s}{\alpha+p}, 1-\frac{\alpha+1/2-\alpha'}{\alpha+p}\right)}, & \text{if } \alpha' < \alpha + 1/2\,, \\ \max\left(N^{-\left(\frac{\alpha'+s}{\alpha+p}\right)}, N^{-1}\log 2N\right), & \text{if } \alpha' = \alpha + 1/2\,, \\ N^{-\min\left(\frac{\alpha'+s}{\alpha+p}, 1\right)}, & \text{if } \alpha' > \alpha + 1/2\,. \end{cases} \tag{D.2}$$

The constants c*,* N_0*, and the implied constant in* (D.1) *do not depend on* N *or* $f^\dagger$*.*

Proof. Combining Propositions D.10 and D.12 shows that, for all $N \geq N_0$, the out-of-distribution test error (D.15) satisfies the upper bound

$$\mathscr{R}_N \leq 2\sum_{j=1}^{\infty} \frac{\sigma_j' |f_j^\dagger|^2}{\left(1 + N\gamma^{-2}\sigma_j\lambda_j\right)^2} + \left(2 + c_0 \left\|f^\dagger\right\|_{\mathcal{H}^s}^2\right) \sum_{j=1}^{\infty} \frac{\sigma_j'\lambda_j}{1 + N\gamma^{-2}\sigma_j\lambda_j} \tag{D.47}$$

with the asserted probability. The second term converges at the rate specified by the rightmost expression in (D.17). We focus on controlling the first term in the upper bound (D.47). Using the asymptotics of $\{\sigma_j'\}_{j\in\mathbb{N}}$, $\{\sigma_j\}_{j\in\mathbb{N}}$, and $\{\lambda_j\}_{j\in\mathbb{N}}$ from Assumption 5.6, there exists $j_0 \in \mathbb{N}$ (independent of N and $f^\dagger$) such that

$$\begin{aligned} \sum_{j=1}^{\infty} \frac{\sigma_j' |f_j^\dagger|^2}{\left(1 + N\gamma^{-2}\sigma_j\lambda_j\right)^2} &\lesssim \sum_{j\leq j_0} \frac{\sigma_j' |f_j^\dagger|^2}{\left(1 + N\gamma^{-2}\sigma_j\lambda_j\right)^2} + \sum_{j>j_0} \frac{j^{-2\alpha'}|f_j^\dagger|^2}{\left(1 + Nj^{-2(\alpha+p)}\right)^2} \\ &\lesssim \frac{\gamma^4}{N^2} \sum_{j\leq j_0} \left(j^{-2s}\sigma_j'\sigma_j^{-2}\lambda_j^{-2}\right) j^{2s}|f_j^\dagger|^2 + \sum_{j=1}^{\infty} \frac{j^{-2\alpha'}|f_j^\dagger|^2}{\left(1 + Nj^{-2(\alpha+p)}\right)^2} \\ &\lesssim N^{-2}\left\|f^\dagger\right\|_{\mathcal{H}^s}^2 + \sum_{j=1}^{\infty} \frac{j^{-2\alpha'}|f_j^\dagger|^2}{\left(1 + Nj^{-2(\alpha+p)}\right)^2}\,. \end{aligned}$$

To obtain the last line, we took the maximum over $j \leq j_0$ of the factor in parentheses in the first term appearing in the second line. Lemma D.22 (applied with $t = 2\alpha' \geq -2s$, $u = 2(\alpha+p) > 1 > 0$, and $v = 2$) shows that the rightmost series in the preceding display is bounded above by

$$N^{-\min\left(2, \frac{\alpha'+s}{\alpha+p}\right)} \left\|f^\dagger\right\|_{\mathcal{H}^s}^2 \leq N^{-2}\left\|f^\dagger\right\|_{\mathcal{H}^s}^2 + N^{-\left(\frac{\alpha'+s}{\alpha+p}\right)} \left\|f^\dagger\right\|_{\mathcal{H}^s}^2\,.$$

However, the N^{-2} contribution is bounded above by the variance contribution in (D.17). Putting together the pieces by enlarging constants and bounding the sum of two nonnegative terms by twice their maximum yields (D.1) and (D.2) as required. □

D.3.2.4 Proof of Corollary D.2

We now prove Corollary D.2, which shows that our (conditional on the design U) high probability bounds imply full expectation bounds for the end-to-end learning linear functional estimator.

Corollary D.2 (end-to-end learning: expectation bound)**.** *Instate the hypotheses and notation of Theorem D.1. Then there exists $N_\star \geq 1$ such that for any $N \geq N_\star$, the mean $\bar{f}^{(N)}$ of the Gaussian posterior distribution (5.15) arising from the N pairs of observed training data (U, Y) in (5.17) satisfies the expected error bound*

$$\mathbb{E}\Big[\mathbb{E}^{u'\sim\nu'}\big|\langle f^\dagger, u'\rangle - \langle \bar{f}^{(N)}, u'\rangle\big|^2\Big] \lesssim \big(1 + \|f^\dagger\|_{\mathcal{H}^s}^2\big)\varepsilon_N^2\,, \tag{D.3}$$

where ε_N^2 is as in (D.2). The constant $N_\star$ and the implied constant in (D.3) do not depend on N or $f^\dagger$.

Proof. Recall $\mathscr{R}_N$ from (D.15). Our goal is to bound $\mathbb{E}\,\mathscr{R}_N$. Recalling the bias–variance decomposition $\mathscr{R}_N = B_N + V_N$ from Lemma D.9, our task reduces to separately bounding $\mathbb{E}\,B_N$ and $\mathbb{E}\,V_N$. We begin with $\mathbb{E}\,V_N$.

Let $\mathsf{E}_\star$ be the event that the high probability variance bound from Proposition D.10 occurs. So, $\mathbb{P}(\mathsf{E}_\star) \geq 1 - e^{-cN}$. In particular, we see from (D.23) that

$$V_N \leq 2\mu\,\mathrm{tr}\big(\mathcal{C}_\mu^{-1/2}\mathcal{C}'\mathcal{C}_\mu^{-1/2}\big)$$

holds true on $\mathsf{E}_\star$ for sufficiently large N. Next, we decompose

$$\begin{aligned} V_N &= V_N \mathbb{1}_{\mathsf{E}_\star} + V_N \mathbb{1}_{\mathsf{E}_\star^\mathsf{c}} \\ &\leq 2\mu\,\mathrm{tr}\big(\mathcal{C}_\mu^{-1/2}\mathcal{C}'\mathcal{C}_\mu^{-1/2}\big) + V_N \mathbb{1}_{\mathsf{E}_\star^\mathsf{c}}\,. \end{aligned}$$

The first inequality holds because $\mathbb{1}_A \leq 1$ for any event A. Thus,

$$\mathbb{E}\,V_N \leq 2\mu\,\mathrm{tr}\big(\mathcal{C}_\mu^{-1/2}\mathcal{C}'\mathcal{C}_\mu^{-1/2}\big) + \mathbb{E}\big[V_N \mathbb{1}_{\mathsf{E}_\star^\mathsf{c}}\big]\,.$$

It remains to bound the second term. We recall from the proof of Proposition D.10 that the bound

$$V_N \leq \mu\,\mathrm{tr}\big(\widehat{\mathcal{C}}_\mu^{-1/2}\mathcal{C}'\widehat{\mathcal{C}}_\mu^{-1/2}\big) = \mu\,\mathrm{tr}\big(\widehat{\mathcal{C}}_\mu^{-1}\mathcal{C}'\big)$$

holds almost surely. Applying von Neumann's trace inequality gives

$$\begin{aligned}\operatorname{tr}\big(\widehat{\mathcal{C}}_\mu^{-1}\mathcal{C}'\big) &\leq \big\|\widehat{\mathcal{C}}_\mu^{-1}\big\|_{\mathcal{L}(H)} \operatorname{tr}(\mathcal{C}') \\ &\leq \mu^{-1} \operatorname{tr}(\mathcal{C}')\,.\end{aligned}$$

The last inequality follows by operator monotonicity of the spectral norm because

$$\widehat{\mathcal{C}}_\mu^{-1} = (\widehat{\mathcal{C}} + \mu \operatorname{Id}_H)^{-1} \preccurlyeq \mu^{-1} \operatorname{Id}_H\ .$$

We deduce that $V_N \leq \operatorname{tr}(\mathcal{C}')$ almost surely so that

$$\begin{aligned}\mathbb{E}\, V_N &\leq 2\mu \operatorname{tr}\big(\mathcal{C}_\mu^{-1/2}\mathcal{C}'\mathcal{C}_\mu^{-1/2}\big) + \mathbb{E}\big[V_N \mathbb{1}_{\mathsf{E}_\star^\mathsf{c}}\big] \\ &\leq 2\mu \operatorname{tr}\big(\mathcal{C}_\mu^{-1/2}\mathcal{C}'\mathcal{C}_\mu^{-1/2}\big) + \operatorname{tr}(\mathcal{C}')\,\mathbb{P}(\mathsf{E}_\star^\mathsf{c}) \\ &\leq 2\mu \operatorname{tr}\big(\mathcal{C}_\mu^{-1/2}\mathcal{C}'\mathcal{C}_\mu^{-1/2}\big) + \operatorname{tr}(\mathcal{C}')e^{-cN} \\ &\leq 2\sum_{j=1}^{\infty} \frac{\sigma_j' \lambda_j}{1 + N\gamma^{-2}\sigma_j\lambda_j} + \operatorname{tr}(\mathcal{C}')e^{-cN} \\ &\lesssim \begin{cases} N^{-\left(1-\frac{\alpha+1/2-\alpha'}{\alpha+p}\right)}\,, & \text{if } \alpha' < \alpha + 1/2\,, \\ N^{-1}\log 2N\,, & \text{if } \alpha' = \alpha + 1/2\,, \\ N^{-1}\,, & \text{if } \alpha' > \alpha + 1/2\,. \end{cases}\end{aligned}$$

The final inequality holds for sufficiently large N because the exponential decay is always bounded above by power law decay. Thus, apart from constant factors, the expectation bound for the variance does not change from the high probability bound.

To complete the proof, we establish an upper bound on the expected squared bias $\mathbb{E}\, B_N$, where B_N is as in (D.16a). To proceed, we recall the high probability bias bound from Proposition D.12. Let A be the event that this bound holds. Denote by b_N the right-hand side of (D.24). Just as we did for the variance, we decompose the expected bias as

$$\begin{aligned}\mathbb{E}\, B_N &= \mathbb{E}\big[B_N \mathbb{1}_\mathsf{A}\big] + \mathbb{E}\big[B_N \mathbb{1}_{\mathsf{A}^\mathsf{c}}\big] \\ &\leq b_N + \mathbb{E}\big[B_N \mathbb{1}_{\mathsf{A}^\mathsf{c}}\big]\,.\end{aligned}$$

Next, we develop a bound for the second term in the preceding display. We estimate

$$\begin{aligned}B_N &= \big\|(\Sigma')^{1/2}(\operatorname{Id}_H - A_N S_N) f^\dagger\big\|^2 \\ &\leq 2\big\|(\Sigma')^{1/2} f^\dagger\big\|^2 + 2\big\|(\Sigma')^{1/2} A_N S_N f^\dagger\big\|^2 \\ &= 2\big\|(\Sigma')^{1/2} f^\dagger\big\|^2 + 2\big\langle f^\dagger, ((\Sigma')^{1/2} A_N S_N)^* (\Sigma')^{1/2} A_N S_N f^\dagger\big\rangle\,.\end{aligned}$$

Next, we compute that the second term in the last line equals

$$\begin{aligned}
&2\operatorname{tr}\Big(((\Sigma')^{1/2}A_NS_N)^*(\Sigma')^{1/2}A_NS_Nf^\dagger\otimes f^\dagger\Big)\\
&\le 2\|f^\dagger\|^2\operatorname{tr}\Big(((\Sigma')^{1/2}A_NS_N)^*(\Sigma')^{1/2}A_NS_N\Big)\\
&= 2\|f^\dagger\|^2\operatorname{tr}\big((A_NS_N)^*\Sigma'A_NS_N\big)\,.
\end{aligned}$$

Recall from the proof of Lemma D.13 that

$$A_NS_N=\Lambda^{1/2}\widehat{\mathcal{C}}_\mu^{-1}\Lambda^{1/2}\widehat{\Sigma}\,.$$

Thus, we find that

$$\begin{aligned}
\operatorname{tr}\big((A_NS_N)^*\Sigma'A_NS_N\big)&=\operatorname{tr}\big(\widehat{\Sigma}\Lambda^{1/2}\widehat{\mathcal{C}}_\mu^{-1}\mathcal{C}'\widehat{\mathcal{C}}_\mu^{-1}\Lambda^{1/2}\widehat{\Sigma}\big)\\
&\le\|\mathcal{C}'\|\operatorname{tr}\big(\widehat{\mathcal{C}}_\mu^{-1}\Lambda^{1/2}\widehat{\Sigma}\widehat{\Sigma}\Lambda^{1/2}\widehat{\mathcal{C}}_\mu^{-1}\big)\\
&=\|\mathcal{C}'\|\operatorname{tr}\big(\Lambda^{1/2}\widehat{\mathcal{C}}_\mu^{-2}\Lambda^{1/2}\widehat{\Sigma}^2\big)\\
&\le\|\mathcal{C}'\|\|\Lambda^{1/2}\widehat{\mathcal{C}}_\mu^{-2}\Lambda^{1/2}\|\operatorname{tr}(\widehat{\Sigma}^2)\\
&\le\mu^{-2}\|\mathcal{C}'\|\|\Lambda\|\operatorname{tr}(\widehat{\Sigma}^2)\,.
\end{aligned}$$

To summarize, we have shown that

$$B_N\le 2\big\|(\Sigma')^{1/2}f^\dagger\big\|^2+2\mu^{-2}\|f^\dagger\|^2\|\mathcal{C}'\|\|\Lambda\|\operatorname{tr}(\widehat{\Sigma}^2)$$

almost surely and hence

$$\begin{aligned}
\mathbb{E}\big[B_N\mathbb{1}_{\mathsf{A}^\mathsf{c}}\big]&\le 2\big\|(\Sigma')^{1/2}f^\dagger\big\|^2\,\mathbb{P}(\mathsf{A}^\mathsf{c})+2\mu^{-2}\|f^\dagger\|^2\|\mathcal{C}'\|\|\Lambda\|\,\mathbb{E}\big[\operatorname{tr}(\widehat{\Sigma}^2)\mathbb{1}_{\mathsf{A}^\mathsf{c}}\big]\\
&\le 2\big\|(\Sigma')^{1/2}f^\dagger\big\|^2\,\mathbb{P}(\mathsf{A}^\mathsf{c})+2\mu^{-2}\|f^\dagger\|^2\|\mathcal{C}'\|\|\Lambda\|\big(\mathbb{E}\operatorname{tr}(\widehat{\Sigma}^2)^2\big)^{1/2}\,\mathbb{P}(\mathsf{A}^\mathsf{c})^{1/2}
\end{aligned}$$

by the Cauchy–Schwarz inequality. We now estimate

$$\begin{aligned}
\mathbb{E}\Big[\big(\operatorname{tr}(\widehat{\Sigma}^2)\big)^2\Big]&=\mathbb{E}\big\|\widehat{\Sigma}\big\|_{\mathrm{HS}}^4\\
&=\mathbb{E}\bigg(\frac{1}{N^2}\sum_{1\le i,j\le N}\langle u_i,u_j\rangle^2\bigg)^2\\
&=\frac{1}{N^4}\sum_{1\le i,j,k,l\le N}\mathbb{E}\big[\langle u_i,u_j\rangle^2\langle u_k,u_l\rangle^2\big]\\
&\le\frac{1}{N^4}\sum_{1\le i,j,k,l\le N}\big(\mathbb{E}\langle u_i,u_j\rangle^4\big)^{1/2}\big(\mathbb{E}\langle u_k,u_l\rangle^4\big)^{1/2}\\
&\le\frac{1}{N^4}\sum_{1\le i,j,k,l\le N}\big(\mathbb{E}\|u_i\|^4\|u_j\|^4\big)^{1/2}\big(\mathbb{E}\|u_k\|^4\|u_l\|^4\big)^{1/2}\\
&\le\mathbb{E}^{u\sim\nu}\|u\|^8\,.
\end{aligned}$$

The chain of inequalities follow from repeated application of Cauchy–Schwarz. The eighth moment in the last line of the preceding display is finite because ν is subgaussian by hypothesis.

Putting together the pieces, we deduce that

$$\begin{aligned}\mathbb{E}\, B_N &\leq b_N + 2\big\|(\Sigma')^{1/2} f^\dagger\big\|^2\, \mathbb{P}(\mathsf{A}^\mathsf{c}) + 2\mu^{-2}\|f^\dagger\|^2 \|\mathcal{C}'\|\|\Lambda\| \big(\mathbb{E}^{u\sim\nu}\|u\|^8\big)^{1/2}\, \mathbb{P}(\mathsf{A}^\mathsf{c})^{1/2} \\ &\lesssim b_N + \big\|(\Sigma')^{1/2} f^\dagger\big\|^2\, \mathbb{P}(\mathsf{A}^\mathsf{c}) + N^2\|f^\dagger\|^2\, \mathbb{P}(\mathsf{A}^\mathsf{c})^{1/2} \\ &\lesssim b_N + N^2\|f^\dagger\|^2\, \mathbb{P}(\mathsf{A}^\mathsf{c})^{1/2} \\ &\leq b_N + N^2 \exp\big(-cN^{\min(\frac{1}{2}, \frac{\alpha+s-1}{2(\alpha+p)})}\big)\|f^\dagger\|^2\,.\end{aligned}$$

The last term decays faster than any power of N^{-1} for sufficiently large N. Thus, $\mathbb{E}\, B_N \lesssim b_N$ and the proof is complete. □

D.3.3 Proofs for Full-Field Learning of Factorized Linear Functionals

This appendix proves the main results from Appendix D.1 and Section 5.4.3 for the (FF) approach. We begin with a lemma that gives a convenient expression for the $L^2_{\nu'}(H;\mathbb{R})$ Bochner norm of a factorized linear PtO map.

Lemma D.21 (squared Bochner norm of linear PtO map)**.** *Let ν' satisfy Assumption (A-I). Let q be a linear functional and L be a linear operator such that $L = \sum_{j=1}^\infty l_j \varphi_j \otimes \varphi_j$ for eigenbasis $\{\varphi_j\}_{j\in\mathbb{N}}$ of $\Sigma' = \mathrm{Cov}(\nu')$ and eigenvalue sequence $\{l_j\}_{j\in\mathbb{N}} \subset \mathbb{R}$. Write $\{\sigma'_j\}_{j\in\mathbb{N}}$ for the eigenvalues of Σ'. If $q \circ L$ is continuous, then*

$$\mathbb{E}^{u'\sim\nu'}|q(Lu')|^2 = \sum_{j=1}^\infty \sigma'_j |q(\varphi_j)|^2 l_j^2\,. \tag{D.48}$$

Proof. Using linearity and the fact that $qL = q\circ L$ is scalar-valued, we compute

$$\begin{aligned}\mathbb{E}^{u'\sim\nu'}|qLu'|^2 = \mathbb{E}^{u'\sim\nu'}[(qLu')(qLu')] &= \mathbb{E}^{u'\sim\nu'}[(qLu')(qLu')^*] \\ &= \mathbb{E}^{u'\sim\nu'}[(qL)u' \otimes u'(qL)^*] \\ &= (qL)\Sigma'(qL)^* \\ &= \mathrm{tr}\big(qL(\Sigma')^{1/2}(qL(\Sigma')^{1/2})^*\big)\,.\end{aligned}$$

The adjoint $(qL)^* \in H$ is well-defined due to the continuity of qL. The definition

of Hilbert–Schmidt norm and the fact that L and Σ' share an eigenbasis yield

$$\operatorname{tr}\big(qL(\Sigma')^{1/2}(qL(\Sigma')^{1/2})^*\big) = \big\|qL(\Sigma')^{1/2}\big\|_{\mathrm{HS}(H;\mathbb{R})}^2 = \sum_{j=1}^{\infty}\big|qL(\Sigma')^{1/2}\varphi_j\big|^2 = \sum_{j=1}^{\infty}\sigma_j' l_j^2 |q(\varphi_j)|^2$$

as asserted. □

Lemma D.21 will be used in the proofs of following two theorems. The arguments rely on [76, Theorem 3.9, pp. 18–19].

Theorem D.3 (full-field learning: convergence rate for Sobolev QoI)**.** *Let the input training data distribution ν, the test data distribution ν', the true forward map $L^\dagger$, and the QoI $q^\dagger$ satisfy Assumption 5.10, but instead of (A-IV), suppose that $\{q^\dagger(\varphi_j)\}_{j\in\mathbb{N}} \in \mathcal{H}^r$ for some $r \in \mathbb{R}$. Let α and α' be as in (5.20) and β be as in (A-II). If $\min(\alpha, \alpha' + r) + \beta > 0$, then there exist constants $c > 0$ and $C > 0$ such that for all sufficiently large N, the plug-in estimator $q^\dagger \circ \bar{L}^{(N)}$ in (5.28) based on the Gaussian posterior distribution (5.27) arising from the N pairs of observed full-field training data (U, Υ) in (5.25) satisfies the error bound*

$$\mathbb{E}^{\Upsilon\,|\,U}\,\mathbb{E}^{u'\sim\nu'}\big|q^\dagger(L^\dagger u') - q^\dagger(\bar{L}^{(N)}u')\big|^2 \lesssim \varepsilon_N^2 \tag{D.4}$$

with probability at least $1 - Ce^{-cN}$ over $U \sim \nu^{\otimes N}$, where

$$\varepsilon_N^2 \coloneqq \begin{cases} N^{-\left(\frac{2\alpha'+2\beta+2r}{1+2\alpha+2\beta}\right)}, & \text{if } \alpha' + r < \alpha + 1/2\,, \\ N^{-1}\log N\,, & \text{if } \alpha' + r = \alpha + 1/2\,, \\ N^{-1}\,, & \text{if } \alpha' + r > \alpha + 1/2\,. \end{cases} \tag{D.5}$$

The constants c, C, and the implied constant in (D.4) do not depend on N.

Proof. Write $q_j^\dagger = q^\dagger(\varphi_j)$ for each $j \in \mathbb{N}$. Lemma D.21 shows that

$$\mathbb{E}^{u'\sim\nu'}\big|q^\dagger(L^\dagger u') - q^\dagger(\bar{L}^{(N)}u')\big|^2 = \sum_{j=1}^{\infty}\sigma_j'|q_j^\dagger|^2\big|l_j^\dagger - \bar{l}_j^{(N)}\big|^2 \tag{D.49}$$

because $L^\dagger$ and $\bar{L}^{(N)}$ share an eigenbasis and $q^\dagger \circ L^\dagger$ is continuous by hypothesis. The assumed asymptotics $\sigma_j' \lesssim j^{-2\alpha'}$ as $j \to \infty$ deliver an index $j_0 \in \mathbb{N}$ such

that the preceding display is bounded above by

$$\begin{aligned}
&\sum_{j\le j_0}\left(j^{2\alpha'}\sigma'_j j^{2r}|q_j^\dagger|^2\right)j^{-2(\alpha'+r)}\left|l_j^\dagger-\bar{l}_j^{(N)}\right|^2+\sum_{j>j_0}\left(j^{2r}|q_j^\dagger|^2\right)j^{-2(\alpha'+r)}\left|l_j^\dagger-\bar{l}_j^{(N)}\right|^2\\
&\qquad\le\left(\max_{k\le j_0}k^{2\alpha'}\sigma'_k k^{2r}|q_k^\dagger|^2\right)\sum_{j\le j_0}j^{-2(\alpha'+r)}\left|l_j^\dagger-\bar{l}_j^{(N)}\right|^2\\
&\qquad\qquad+\left(\sup_{k\in\mathbb{N}}k^{2r}|q_k^\dagger|^2\right)\sum_{j>j_0}j^{-2(\alpha'+r)}\left|l_j^\dagger-\bar{l}_j^{(N)}\right|^2\\
&\qquad\lesssim\sum_{j=1}^{\infty}j^{-2(\alpha'+r)}\left|l_j^\dagger-\bar{l}_j^{(N)}\right|^2.
\end{aligned}$$

The supremum is finite because its argument is summable due to $q^\dagger\in\mathcal{H}^r$. The asserted result follows from [76, Theorem 3.9, pp. 18–19] by using the result for the posterior mean, choosing $|\vartheta'_j|^2=j^{-2(\alpha'+r)}$ (i.e., replacing α' with $\alpha'+r$), and choosing δ to be a sufficiently small constant. □

The theorem for power law QoI decay has a proof similar to the previous one.

Theorem 5.13 (full-field learning: convergence rate for power law QoI)**.** *Let the input training data distribution ν, the test data distribution ν', the true forward map $L^\dagger$, and the QoI $q^\dagger$ satisfy Assumption 5.10. Let α and α' be as in (5.20) and β and r be as in (A-II) and (A-IV). Then there exist constants $c>0$ and $C>0$ such that for all sufficiently large N, the plug-in estimator $q^\dagger\circ\bar{L}^{(N)}$ in (5.28) based on the Gaussian posterior distribution (5.27) arising from the N pairs of observed full-field training data (U,Υ) in (5.25) satisfies the error bound*

$$\mathbb{E}^{\Upsilon\,|\,U}\,\mathbb{E}^{u'\sim\nu'}\left|q^\dagger(L^\dagger u')-q^\dagger(\bar{L}^{(N)}u')\right|^2\lesssim\varepsilon_N^2 \tag{5.31}$$

with probability at least $1-Ce^{-cN}$ over $U\sim\nu^{\otimes N}$, where

$$\varepsilon_N^2:=\begin{cases}N^{-\left(\frac{1+2\alpha'+2\beta+2r}{1+2\alpha+2\beta}\right)}, & \text{if } \alpha'+r<\alpha,\\ N^{-1}\log N, & \text{if } \alpha'+r=\alpha,\\ N^{-1}, & \text{if } \alpha'+r>\alpha.\end{cases} \tag{5.32}$$

The constants c, C, and the implied constant in (5.31) do not depend on N.

Proof. The proof mimics that of Theorem D.3. After enlarging j_0 to accommodate the asymptotics of $q_j^\dagger$, the only difference is that (D.49) is now bounded

above by

$$\max_{k\le j_0} k^{2\alpha'}\sigma_k' k^{2r+1}|q_k^\dagger|^2 \sum_{j\le j_0} j^{-2(\alpha'+r+1/2)}\big|l_j^\dagger - \bar{l}_j^{(N)}\big|^2 + \sum_{j>j_0} j^{-2(\alpha'+r+1/2)}\big|l_j^\dagger - \bar{l}_j^{(N)}\big|^2$$
$$\lesssim \sum_{j=1}^{\infty} j^{-2(\alpha'+r+1/2)}\big|l_j^\dagger - \bar{l}_j^{(N)}\big|^2 .$$

Replacing α' with $\alpha' + r + 1/2$ in [76, Theorem 3.9, pp. 18–19] completes the proof. □

D.3.4 Proof of Sample Complexity Comparison

We now prove Corollary 5.15.

Corollary 5.15 (sample complexity comparison)**.** *Instate the notation and assertions in Assumptions 5.6, 5.7, and (A-III). Suppose that the training and test distribution covariances have equivalent smoothness, i.e., $\alpha' = \alpha$. Let the underlying true PtO map $f^\dagger$ have the factorization $f^\dagger = q^\dagger \circ L^\dagger$, where $|q^\dagger(\varphi_j)|^2 \lesssim j^{-2r-1}$ as $j \to \infty$ and $L^\dagger$ is as in (5.26) with eigenvalues $l^\dagger \in \mathcal{H}^\beta$. If $\beta + r + 1/2 > 0$, $\alpha + \beta + r > 1/2$, and $\alpha + \beta > 0$, then there exist constants $c > 0$ and $C > 0$ such that for all sufficiently large N, the following holds on an event with probability at least $1 - C\exp(-cN^{\min(1,\frac{\alpha+\beta+r-1/2}{1+\alpha+\beta+r})})$ over $U = \{u_n\}_{n=1}^N \sim \nu^{\otimes N}$. The (EE) posterior mean estimator $\bar{f}^{(N)}$ in (5.15) (with $p \coloneqq \beta + r + 1$ in (A3)) trained on end-to-end data (U, Y) satisfies*

$$\mathbb{E}^{Y\,|\,U}\,\mathbb{E}^{u\sim\nu}\big|q^\dagger(L^\dagger u) - \langle \bar{f}^{(N)}, u\rangle\big|^2 \lesssim N^{-\left(1-\frac{1}{2+2\alpha+2\beta+2r}\right)} . \tag{5.34}$$

On the other hand, the (FF) plug-in estimator $q^\dagger \circ \bar{L}^{(N)}$ in (5.28) trained on full-field data (U, Υ) satisfies

$$\mathbb{E}^{\Upsilon\,|\,U}\,\mathbb{E}^{u\sim\nu}\big|q^\dagger(L^\dagger u) - q^\dagger(\bar{L}^{(N)}u)\big|^2 \lesssim \begin{cases} N^{-\left(1-\frac{-2r}{1+2\alpha+2\beta}\right)}, & \text{if } r < 0, \\ N^{-1}\log N, & \text{if } r = 0, \\ N^{-1}, & \text{if } r > 0. \end{cases} \tag{5.35}$$

Proof. First, we claim that $f^\dagger = q^\dagger \circ L^\dagger \in \mathcal{H}^s$ for any $s \le \beta + r + 1/2$ under the hypotheses. Since $f_j^\dagger \coloneqq f^\dagger(\varphi_j) = l_j^\dagger q^\dagger(\varphi_j) =: l_j^\dagger q_j^\dagger$, we compute

$$\sum_{j=1}^{\infty} j^{2s}|f_j^\dagger|^2 = \sum_{j=1}^{\infty} j^{2s}|l_j^\dagger|^2|q_j^\dagger|^2 \lesssim 1 + \sum_{j=1}^{\infty}\big(j^{2s-2r-1-2\beta}\big)j^{2\beta}|l_j^\dagger|^2 \lesssim 1 + \|l^\dagger\|_{\mathcal{H}^\beta}^2 < \infty$$

if $2s - 2r - 1 - 2\beta \leq 0$. This proves the claim. For the end-to-end bound, choose the maximal $s = \beta + r + 1/2$ in Theorem 5.12. With optimal $p = s + 1/2$, the assumption $p > 1/2$ gives $s > 0$ so that $\beta + r + 1/2 > 0$ as hypothesized. This condition also satisfies the continuity requirement (A-V) by a similar calculation. The condition $\alpha + s > 1$ from Assumption 5.9 is the same as $\alpha + \beta + r > 1/2$. Finally, to satisfy the condition $\min(\alpha, \alpha + r + 1/2) + \beta > 0$ from Theorem 5.13, it suffices for $\alpha + \beta > 0$ because the other case has $\alpha + r + 1/2 + \beta > \max(\alpha, 1) > 0$. With a common set of hypotheses identified, the convergence rates may now simply be read off from Theorem 5.12 and Theorem 5.13 after plugging in $\alpha' = \alpha$ and $s = \beta + r + 1/2$. The fact that these bounds hold simultaneously on an event with the asserted probability follows by a union bound and enlarging constants. □

D.3.5 Technical Lemmas

We conclude Appendix D.3 with several supporting lemmas. The following two technical results emphasize the nonasymptotic nature of analogous asymptotic bounds on parametrized series from [147, Lemmas 8.1–8.2, pp. 2653–2655]. The first result is useful for controlling the bias error term arising in Appendix D.3.2.2.

Lemma D.22 (series decay: Sobolev regularity)**.** *Let $q \in \mathbb{R}$, $t \geq -2q$, $u > 0$, and $v \geq 0$. Let $N \in \mathbb{N}$ be arbitrary. For every $\xi \in \mathcal{H}^q(\mathbb{N}; \mathbb{R})$, it holds that*

$$\sum_{j=1}^{\infty} \frac{j^{-t} \xi_j^2}{(1 + N j^{-u})^v} \lesssim N^{-\min\left(v, \frac{t+2q}{u}\right)} \|\xi\|_{\mathcal{H}^q}^2 . \tag{D.50}$$

Proof. The asserted result may be extracted from the proof of [147, Lemma 8.1]. □

The next lemma is analogous to the previous one and is useful for controlling the variance error term from Appendix D.3.2.1.

Lemma D.23 (series decay: power law regularity)**.** *Let $t > 1$, $u > 0$, and $v \geq 0$. For all $N \in \mathbb{N}$, it holds that*

$$\sum_{j=1}^{\infty} \frac{j^{-t}}{(1 + N j^{-u})^v} \lesssim \begin{cases} N^{-\left(\frac{t-1}{u}\right)}, & \text{if } (t-1)/u < v, \\ N^{-v} \log 2N, & \text{if } (t-1)/u = v, \\ N^{-v}, & \text{if } (t-1)/u > v. \end{cases} \tag{D.51}$$

Proof. We split the series into two parts by summing over disjoint index sets defined by the critical index $N^{1/u}$. If $j \le N^{1/u}$, then $Nj^{-u} \le 1+Nj^{-u} \le 2Nj^{-u}$. Otherwise $j > N^{1/u}$ and $1 \le 1 + Nj^{-u} \le 2$. Hence, for the bulk part of the series,

$$\sum_{j \le N^{1/u}} \frac{j^{-t}}{(1+Nj^{-u})^v} \simeq \frac{1}{N^v} \sum_{j \le N^{1/u}} j^{uv-t}\,. \tag{D.52}$$

If $(t-1)/u > v$, then $uv - t < -1$ so that the right-hand side of (D.52) is bounded above by $N^{-v}\sum_{j=1}^{\infty} j^{uv-t} \lesssim N^{-v}$. Next, if $(t-1)/u = v$, then the right-hand side equals $N^{-v}\sum_{j \le N^{1/u}} j^{-1}$. The J-th harmonic number satisfies $\sum_{j=1}^{J} j^{-1} \le 1 + \log J$ (by an integral comparison). Since $\log(\,\cdot\,)$ is increasing, $1 + \log N^{1/u} \le (\log 2)^{-1}\log(2N) + (\log 2N)/u \lesssim \log 2N$ and the second case in (D.51) follows. Finally, if $(t-1)/u < v$, then we consider the regimes $-1 < uv - t < 0$ and $uv - t \ge 0$ separately. In the first regime, $h\colon x \mapsto x^{uv-t}$ is nonincreasing. Thus, if $j - 1 \le x \le j$, then $h(j) \le h(x)$ and hence $h(j) \le \int_{j-1}^{j} h(x)\,dx$. Summing this inequality leads to

$$\sum_{j \le N^{1/u}} j^{uv-t} \le \int_0^{N^{1/u}} x^{uv-t}\,dx = \frac{N^{v-(t-1)/u}}{1+uv-t}$$

because $0 < 1 + uv - t < 1$. Multiplying by N^{-v} shows that (D.52) is bounded above by a constant times $N^{-(t-1)/u}$. In the second regime, $uv - t \ge 0$ so h is now nondecreasing. A similar argument to the preceding one yields

$$\sum_{j \le N^{1/u}} j^{uv-t} \le \int_1^{1+N^{1/u}} x^{uv-t}\,dx \le \int_0^{1+N^{1/u}} x^{uv-t}\,dx = \frac{(1+N^{1/u})^{1+uv-t}}{1+uv-t}\,.$$

Since $1 + uv - t \ge 1$ and $N^{1/u} \ge 1$, the right-hand side is bounded above by $(2N^{1/u})^{1+uv-t} \lesssim N^{v-(t-1)/u}$. This proves that the inequality (D.51) remains valid if the infinite series is replaced by the bulk partial sum (D.52). It remains to estimate the tail part of the series. By an analogous integral comparison,

$$\sum_{j > N^{1/u}} \frac{j^{-t}}{(1+Nj^{-u})^v} \simeq \sum_{j > N^{1/u}} j^{-t} \le \int_{\lceil N^{1/u}\rceil - 1}^{\infty} x^{-t}\,dx \le \int_{\max(1, N^{1/u}-1)}^{\infty} x^{-t}\,dx\,.$$

The rightmost integral converges because $t > 1$ and evaluates to

$$\frac{\max(1, N^{1/u}-1)^{-(t-1)}}{t-1} \le \frac{2^{t-1}}{t-1} N^{-\left(\frac{t-1}{u}\right)}\,.$$

We used $\max(a,b) \geq (a+b)/2$ for nonnegative a and b to obtain the inequality. This shows that the tail series has the same upper bound $N^{-(t-1)/u}$ as the bulk sum if $(t-1)/u < v$. Otherwise, $N^{-(t-1)/u} \lesssim N^{-v} \log 2N$ if $(t-1)/u = v$ or $N^{-(t-1)/u} \lesssim N^{-v}$ if $(t-1)/u > v$. Thus, the assertion (D.51) follows. □

Application of the previous lemma gives an exact estimate for the *effective dimension* corresponding to the prior-normalized training data covariance operator $\mathcal{C} = \Lambda^{1/2}\Sigma\Lambda^{1/2}$ under the assumption of asymptotically exact power law decay of its eigenvalues. It plays a role in both the bias and variance bounds.

Lemma D.24 (effective dimension)**.** *Under Assumption 5.6, it holds that*

$$\operatorname{tr}\big(\mathcal{C}_\mu^{-1}\mathcal{C}\big) \simeq \mu^{-\frac{1}{2(\alpha+p)}} \quad \textit{for all} \quad 0 < \mu \lesssim 1\,. \tag{D.53}$$

Proof. Let $u \coloneqq 2(\alpha+p)$. Then $u > 1$ by hypothesis. By the simultaneous diagonalization from Assumption 5.6, the eigenvalues of $\mathcal{C} = \Lambda^{1/2}\Sigma\Lambda^{1/2}$ are $\{\sigma_j\lambda_j\}_{j\in\mathbb{N}}$. Then since $\sigma_j \asymp j^{-2\alpha}$ and $\lambda_j \asymp j^{-2p}$ as $j \to \infty$, there exists $j_0 \in \mathbb{N}$ such that

$$\begin{aligned}\mu \operatorname{tr}\big(\mathcal{C}_\mu^{-1}\mathcal{C}\big) = \sum_{j=1}^{\infty} \frac{\mu\sigma_j\lambda_j}{\mu+\sigma_j\lambda_j} &\simeq \sum_{j\leq j_0} \frac{\sigma_j\lambda_j}{1+\mu^{-1}\sigma_j\lambda_j} + \sum_{j>j_0} \frac{j^{-u}}{1+\mu^{-1}j^{-u}} \\ &\leq \mu j_0 + \sum_{j=1}^{\infty} \frac{j^{-u}}{1+\mu^{-1}j^{-u}}\,.\end{aligned}$$

Since $1+\mu^{-1} \lesssim 2\mu^{-1}$ follows from the hypothesis $\mu \lesssim 1$, it holds that

$$\mu \lesssim \frac{2}{1+\mu^{-1}} \leq 2\sum_{j=1}^{\infty} \frac{j^{-u}}{1+\mu^{-1}j^{-u}}\,.$$

Application of Lemma D.23 (applied in the first case with $t=u$, $u=u$, $v=1$, and $N=\mu^{-1}$ because $(u-1)/u < 1$) shows that the series in the preceding display is bounded above by a constant times $\mu^{1-1/u}$. This implies the upper bound in (D.53).

Now let $J_\mu \coloneqq \max(j_0, \mu^{-1/u})$. For the lower bound, we compute

$$\mu \operatorname{tr}\big(\mathcal{C}_\mu^{-1}\mathcal{C}\big) \geq \sum_{j>j_0} \frac{\mu\sigma_j\lambda_j}{\mu+\sigma_j\lambda_j} \simeq \sum_{j>j_0} \frac{j^{-u}}{1+\mu^{-1}j^{-u}} \geq \sum_{j>J_\mu} \frac{j^{-u}}{1+\mu^{-1}j^{-u}}\,.$$

Since $j > J_\mu \geq \mu^{-1/u}$, it holds that $1 \leq 1 + \mu^{-1} j^{-u} \leq 2$. Hence, the right-hand side of the preceding display is bounded above and below by a constant times $\sum_{j>J_\mu} j^{-u}$. By comparison to an integral as in the proof of Lemma D.23, we obtain

$$\begin{aligned}\sum_{j>J_\mu} j^{-u} \geq \int_{\lceil J_\mu \rceil}^{\infty} x^{-u}\, dx &\geq \frac{(J_\mu + 1)^{-(u-1)}}{u-1} \\ &\geq \frac{(\mu^{-1/u} + j_0 + 1)^{-(u-1)}}{u-1} \\ &\geq \frac{\big(\mu^{-1/u} + \mu^{-1/u}(j_0+1)\big)^{-(u-1)}}{u-1}\,.\end{aligned}$$

We used $J_\mu \leq j_0 + \mu^{-1/u}$ in the second line and $1 \lesssim \mu^{-1/u}$ in the third line. Since the third line evaluates to $((j_0+2)^{-(u-1)}/(u-1))\mu^{1-1/u}$, it follows that $\operatorname{tr}(\mathcal{C}_\mu^{-1}\mathcal{C}) \gtrsim \mu^{-1/u}$ as asserted. □

The next lemma, whose proof requires the previous effective dimension estimate, defines a high probability event on which the operator norm of a certain normalized and centered empirical covariance is uniformly bounded in the sample size.

Lemma D.25 (good set)**.** *Let $\mu = \gamma^2/N$ and $\mathcal{C}$, $\mathcal{C}_\mu$, and $\widehat{\mathcal{C}}$ be as in (D.18) and (D.20). Under Assumption 5.6 and 5.7, there exists a constant $c \in (0,1)$ (depending only on ν, Λ, and γ^2) such that if*

$$N \geq N_0 \coloneqq c^{-1} \mathbb{1}_{\{\gamma^{-2}\operatorname{tr}(\mathcal{C}) > c\}} + 1\,, \tag{D.54}$$

then the event

$$\mathsf{E} \coloneqq \Big\{ \big\| \mathcal{C}_\mu^{-1/2} (\widehat{\mathcal{C}} - \mathcal{C}) \mathcal{C}_\mu^{-1/2} \big\|_{\mathcal{L}(H)} \leq 1/2 \Big\} \tag{D.55}$$

satisfies $\mathbb{P}(\mathsf{E}) \geq 1 - e^{-cN}$.

Proof. By [131, Lemma 5, p. 13], there exists $c_0 \in (0,1)$ such that if $\operatorname{tr}(\mathcal{C}_\mu^{-1}\mathcal{C}) \leq c_0 N$ and the pushforward $(\Lambda^{1/2})_\sharp \nu$ is strongly-subgaussian, then $\mathbb{P}(\mathsf{E}) \geq 1 - e^{-c_0 N}$. An H-valued random variable Z (equivalently, its law) is *strongly-subgaussian* if

$$\big\| \langle Z, h \rangle \big\|_{\psi_2} = \sup_{p \geq 1} \frac{(\mathbb{E}|\langle Z, h\rangle|^p)^{1/p}}{\sqrt{p}} \lesssim \sqrt{\mathbb{E}\langle Z, h\rangle^2} \quad \text{for all} \quad h \in H\,. \tag{D.56}$$

To complete the proof, we will verify the trace condition (for sufficiently large N) and the subgaussian condition. For the latter, it is sufficient to show that

ν itself is strongly-subgaussian because then the random variable $\Lambda^{1/2}u$ with $u \sim \nu$ satisfies

$$\big\|\langle \Lambda^{1/2}u, h\rangle\big\|_{\psi_2} = \sup_{p\geq 1} \frac{(\mathbb{E}|\langle u, \Lambda^{1/2}h\rangle|^p)^{1/p}}{\sqrt{p}} \lesssim \sqrt{\mathbb{E}\langle u, \Lambda^{1/2}h\rangle^2} = \sqrt{\langle h, \mathcal{C}h\rangle}$$

for every $h \in H$ as desired. By Assumption 5.7, $u \sim \nu$ has the series expansion $u = \sum_j \sqrt{\sigma_j} z_j \varphi_j$ with the $\{z_j\}_{j\in\mathbb{N}}$ centered and independent. Fix $h \in H$ and define $h_j \coloneqq \langle h, \varphi_j\rangle$ for each $j \in \mathbb{N}$. Estimating the moment generating function of $\langle u, h\rangle$ with an argument similar to the one used in the proof of Lemma D.18 leads to

$$\begin{aligned}
\mathbb{E}\, e^{\lambda\langle u,h\rangle} = \mathbb{E}\prod_{j=1}^{\infty} e^{\lambda\sqrt{\sigma_j} z_j h_j} &= \prod_{j=1}^{\infty} \mathbb{E}\, e^{\lambda\sqrt{\sigma_j} z_j h_j} \\
&\leq \prod_{j=1}^{\infty} e^{c'\lambda^2\sigma_j h_j^2 \|z_j\|_{\psi_2}^2} \\
&\leq \exp\left(c' m^2 \lambda^2 \sum_{j=1}^{\infty} \sigma_j h_j^2\right) \\
&= \exp\big(c' m^2 \lambda^2 \langle h, \Sigma h\rangle\big)
\end{aligned}$$

for some absolute constant $c' > 0$ [266, p. 28] and any $\lambda \in \mathbb{R}$. We used independence of the $\{z_j\}_{j\in\mathbb{N}}$ to obtain the second equality in the preceding display and subgaussianity to obtain the first inequality. Again by [266, p. 28], $m\langle h, \Sigma h\rangle^{1/2}$ is equivalent to the subgaussian norm of $\langle u, h\rangle$. Thus, ν is strongly-subgaussian as required.

Next, we show that the trace condition holds. Denote the eigenvalues of $\mathcal{C} = \Lambda^{1/2}\Sigma\Lambda^{1/2}$ by $\{\lambda_j(\mathcal{C})\}_{j\in\mathbb{N}}$. Since $\lambda_j(\mathcal{C}) \asymp j^{-2(\alpha+p)}$ and $2(\alpha + p) > 1$ by Assumption 5.6, the operator $\mathcal{C}$ is trace-class and

$$\operatorname{tr}\big(\mathcal{C}_\mu^{-1}\mathcal{C}\big) = \sum_{j=1}^{\infty} \frac{\lambda_j(\mathcal{C})}{\lambda_j(\mathcal{C}) + \gamma^2/N} \leq N\gamma^{-2} \sum_{j=1}^{\infty} \lambda_j(\mathcal{C}) = N\gamma^{-2} \operatorname{tr}(\mathcal{C})\,.$$

Thus, $\operatorname{tr}(\mathcal{C}_\mu^{-1}\mathcal{C}) \leq c_0 N$ if $\gamma^{-2}\operatorname{tr}(\mathcal{C}) \leq c_0$. Otherwise, application of Lemma D.24 shows that there exists a constant $c_1 > 0$ such that $\operatorname{tr}(\mathcal{C}_\mu^{-1}\mathcal{C}) \leq c_1 N^{1/(2(\alpha+p))}$. Hence, $\operatorname{tr}(\mathcal{C}_\mu^{-1}\mathcal{C}) \leq c_0 N$ holds if $N \geq c_2 \coloneqq \max(1, (c_1/c_0)^{(\alpha+p)/(\alpha+p-1/2)})$. Finally, let $c = \min(c_0, c_2^{-1})$. We conclude by showing that the constant N_0 in (D.54) suffices to verify the trace condition. If $\gamma^{-2}\operatorname{tr}(\mathcal{C}) \leq c$, then $\gamma^{-2}\operatorname{tr}(\mathcal{C}) \leq c_0$ and $N_0 \geq 1$ suffices. Otherwise, $N_0 \geq c^{-1} + 1 \geq \max(c_0^{-1}, c_2) + 1 \geq c_2$ as required. The fact that $1 - e^{-c_0 N} \geq 1 - e^{-cN}$ completes the proof. □

As discussed in Remark 5.8, we are able to weaken the strongly-subgaussian requirement on the input training distribution ν (5.11) by only requiring its KL expansion coefficients (5.22) to be pairwise uncorrelated instead of statistically independent. However, the following result shows that this improvement to Assumption 5.7 leads to a strictly worse failure probability for the good event E. Indeed, Lemma D.25 gives $\mathbb{P}(\mathsf{E}^{\mathsf{c}}) \leq e^{-cN}$, while the next lemma yields $\mathbb{P}(\mathsf{E}^{\mathsf{c}}) \leq 2e^{-cN^r}$ with $r = (\alpha + p - 1)/(\alpha + p) < 1$ strictly smaller than one.

Lemma D.26 (good set without independent KL coefficients)**.** *Let $\mu = \gamma^2/N$ and $\mathcal{C}$, $\mathcal{C}_\mu$, and $\widehat{\mathcal{C}}$ be as in (D.18) and (D.20). Let Assumption 5.6 hold. Suppose that the hypotheses of Assumption 5.7 hold, but instead of the requirement that the normalized KL expansion coefficients $\{z_j\}_{j\in\mathbb{N}}$ (5.22) of $u \sim \nu$ are independent, assume now that the $\{z_j\}_{j\in\mathbb{N}}$ are only pairwise uncorrelated. Then there exist constants $c \in (0,1)$ and $N_0 \geq 1$ (depending only on ν, Λ, and γ^2) such that if $N \geq N_0$, then the event*

$$\mathsf{E} \coloneqq \left\{ \left\| \mathcal{C}_\mu^{-1/2} (\widehat{\mathcal{C}} - \mathcal{C}) \mathcal{C}_\mu^{-1/2} \right\|_{\mathcal{L}(H)} \leq 1/2 \right\} \tag{D.57}$$

satisfies $\mathbb{P}(\mathsf{E}) \geq 1 - 2\exp(-cN^{\frac{\alpha+p-1}{\alpha+p}})$, where α and p are as in (A2) and (A3).

Proof. The proof closely mimics the argument in the proof of Lemma D.14. Recalling the definition of $\widehat{T} = \mathcal{C}_\mu^{-1/2}(\mathcal{C} - \widehat{\mathcal{C}})\mathcal{C}_\mu^{-1/2}$ from (D.25), it holds that

$$-\widehat{T} = \frac{1}{N} \sum_{n=1}^{N} \big(v_n \otimes v_n - \mathbb{E}[v_1 \otimes v_1] \big), \quad \text{where} \quad v_n \coloneqq \mathcal{C}_\mu^{-1/2} \Lambda^{1/2} u_n$$

and $\{u_n\}_{n=1}^N \sim \nu^{\otimes N}$. Writing $(\mathrm{HS}(H), \langle \cdot, \cdot \rangle_{\mathrm{HS}(H)}, \| \cdot \|_{\mathrm{HS}(H)})$ for the Hilbert space of Hilbert–Schmidt operators mapping H into itself, we control all moments of $\|\widehat{T}\|_{\mathrm{HS}(H)}$ and then apply a Bernstein inequality. To this end, let

$$Z_n \coloneqq \big(v_n \otimes v_n - \mathbb{E}[v_1 \otimes v_1] \big)$$

and fix an integer $\ell \geq 2$. The inequality $|a+b|^\ell \leq 2^{\ell-1}(|a|^\ell + |b|^\ell)$ shows that

$$\begin{aligned} \mathbb{E}\|Z_1\|_{\mathrm{HS}(H)}^\ell &\leq 2^{\ell-1} \big(\mathbb{E}\|v_1 \otimes v_1\|_{\mathrm{HS}(H)}^\ell + \|\mathbb{E}[v_1 \otimes v_1]\|_{\mathrm{HS}(H)}^\ell \big) \\ &\leq 2^\ell \, \mathbb{E}\|v_1\|^{2\ell} . \end{aligned}$$

The second line is due to Jensen's inequality and the identity $\|a \otimes b\|_{\mathrm{HS}(H)} = \|a\|\|b\|$. By the KL expansion hypothesis and the assumption (A1) that Λ and

Σ share the orthonormal eigenbasis $\{\varphi_j\}$, it holds that

$$v_1 = \sum_{j=1}^{\infty} \rho_j z_j \varphi_j \,, \quad \text{where} \quad \rho_j = \sqrt{\frac{\sigma_j \lambda_j}{\mu + \sigma_j \lambda_j}} \geq 0$$

and the $\{z_j\}_{j\in\mathbb{N}}$ (5.22) are pairwise uncorrelated. Thus,

$$\mathbb{E}\|v_1\|^{2\ell} = \mathbb{E}\big[\big(\|v_1\|^2\big)^{\ell}\big] = \mathbb{E}\left[\left(\sum_{k=1}^{\infty} \rho_k^2 z_k^2\right)^{\ell}\right].$$

The triangle inequality in the Banach space $L^{\ell}_{\mathbb{P}}(\Omega;\mathbb{R})$ yields

$$\left\|\sum_{k=1}^{\infty} \rho_k^2 z_k^2\right\|_{L^{\ell}_{\mathbb{P}}} \leq \sum_{k=1}^{\infty} \rho_k^2 \big\|z_k^2\big\|_{L^{\ell}_{\mathbb{P}}} = \sum_{k=1}^{\infty} \rho_k^2 \Big(\mathbb{E}\big[|z_k|^{2\ell}\big]\Big)^{1/\ell}.$$

By the definition (D.12) of the subexponential norm, Lemma D.8, and hypothesis (5.23),

$$\begin{aligned}
\sum_{k=1}^{\infty} \rho_k^2 \Big(\mathbb{E}\big[|z_k|^{2\ell}\big]\Big)^{1/\ell} &\leq \sum_{k=1}^{\infty} \rho_k^2 \big(\ell \|z_k^2\|_{\psi_1}\big) \\
&\leq \sum_{k=1}^{\infty} \rho_k^2 \big(2\ell \|z_k\|_{\psi_2}^2\big) \\
&\leq 2\ell m^2 \sum_{k=1}^{\infty} \frac{\sigma_k \lambda_k}{\mu + \sigma_k \lambda_k}\,.
\end{aligned}$$

Taking the ℓ-th power and putting together the pieces, we deduce that

$$\mathbb{E}\|Z_1\|^{\ell}_{\mathrm{HS}(H)} \leq 2^{\ell}\big(2m^2 \operatorname{tr}\big(\mathcal{C}_{\mu}^{-1}\mathcal{C}\big)\big)^{\ell} \ell^{\ell} \leq \frac{1}{2}\ell!\big(8em^2 \operatorname{tr}\big(\mathcal{C}_{\mu}^{-1}\mathcal{C}\big)\big)^{\ell} =: \frac{1}{2}\ell!\sigma^2 b^{\ell-2}$$

by Stirling's formula $(\ell/e)^{\ell} \leq \ell!$. Here

$$\sigma = b \equiv b_N = 8em^2 \operatorname{tr}(\mathcal{C}_{\mu}^{-1}\mathcal{C}) \simeq N^{\frac{1}{2(\alpha+p)}}\,,$$

where the last equivalence follows from Lemma D.24. By the Pinelis–Sakhanenko Bernstein inequality for Hilbert spaces—where the version we use is [162, Theorem A.1, p. 14]—it holds that

$$\mathbb{P}\left\{\big\|\widehat{T}\big\|_{\mathrm{HS}(H)} \leq \frac{2b_N t}{N} + \sqrt{\frac{2b_N^2 t}{N}}\right\} \geq 1 - 2e^{-t} \quad \text{for all} \quad t > \log 2\,.$$

For $\|\widehat{T}\|_{\mathrm{HS}(H)} \leq 1/2$ to hold, it suffices that $2b_N t/N \leq 1/4$ and $2b_N^2 t/N \leq 1/8$. There exists $c \in (0,1)$ and $N_0 \geq 1$ such that the choice $t \coloneqq cN^{\frac{\alpha+p-1}{\alpha+p}}$ makes both inequalities true as long as $N \geq N_0$. The desired result (D.57) follows from the inequality $\|\widehat{T}\|_{\mathcal{L}(H)} \leq \|\widehat{T}\|_{\mathrm{HS}(H)}$ and the monotonicity of probability measure. □

Finally, following [52, Section 5.2, p. 350] and [98, Section 6.2, p. 26], we recall the following identity for regularized inverses of linear operators. This result is especially useful in conjunction with the cyclic property of the trace for deriving trace bounds.

Lemma D.27 (identity for regularized inverses)**.** *Let $\mathcal{H}$ be a separable Hilbert space. For any $\lambda > 0$ and any symmetric positive-semidefinite bounded linear operators $A \in \mathcal{L}(\mathcal{H})$ and $B \in \mathcal{L}(\mathcal{H})$, let $A_\lambda \coloneqq A + \lambda \operatorname{Id}_\mathcal{H}$ and $B_\lambda \coloneqq B + \lambda \operatorname{Id}_\mathcal{H}$. It holds that*

$$A_\lambda^{-1} = B_\lambda^{-1/2}\big(\operatorname{Id}_\mathcal{H} - B_\lambda^{-1/2}(B - A)B_\lambda^{-1/2}\big)^{-1}B_\lambda^{-1/2}. \tag{D.58}$$

Proof. Write $\operatorname{Id} \coloneqq \operatorname{Id}_\mathcal{H}$. A direct calculation shows that

$$\begin{aligned} A + \lambda \operatorname{Id} &= A - B + B + \lambda \operatorname{Id} \\ &= (B + \lambda \operatorname{Id})^{1/2}(B + \lambda \operatorname{Id})^{1/2} - (B - A) \\ &= B_\lambda^{1/2}\big(\operatorname{Id}\big)B_\lambda^{1/2} - B_\lambda^{1/2}\big(B_\lambda^{-1/2}(B - A)B_\lambda^{-1/2}\big)B_\lambda^{1/2} \\ &= B_\lambda^{1/2}\big(\operatorname{Id} - B_\lambda^{-1/2}(B - A)B_\lambda^{-1/2}\big)B_\lambda^{1/2}. \end{aligned}$$

Inverting the both sides of the equality yields the assertion. □