Model Predictive Control for Deferrable Loads Scheduling

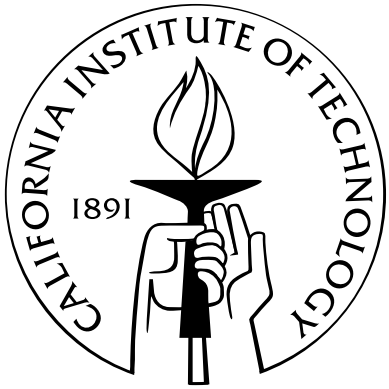
Thesis by

Niangjun Chen

In Partial Fulfillment of the Requirements

for the Degree of

Master of Science



California Institute of Technology

Pasadena, California

2014

(Submitted June 2, 2014)

Acknowledgements

I would like to express my gratitude towards my advisers Prof. Adam Wierman and Prof. Steven Low. I feel very fortunate and privileged receive enormous guidance and help from them. Their wisdom and encouragement help me overcome many obstacles during the past two years. I really learned a lot from both of them.

I am also grateful to my collaborators, Lachlan Andrew and Lingwen Gan. They have always provided insightful discussions and constructive suggestions. The atmosphere and environment of the RSRG, Computer Science department and Caltech are amazing and helpful. It is really a pleasure to work and study here.

Last but not least, I would like to thank my family. Their unconditional love and support for me over the years is what made this thesis possible.

Abstract

Real-time demand response is essential for handling the uncertainties of renewable generation. Traditionally, demand response has been focused on large industrial and commercial loads, however it is expected that a large number of small residential loads such as air conditioners, dish washers, and electric vehicles will also participate in the coming years. The electricity consumption of these smaller loads, which we call deferrable loads, can be shifted over time, and thus be used (in aggregate) to compensate for the random fluctuations in renewable generation.

In this thesis, we propose a real-time distributed deferrable load control algorithm to reduce the variance of aggregate load (load minus renewable generation) by shifting the power consumption of deferrable loads to periods with high renewable generation. The algorithm is model predictive in nature, i.e., at every time step, the algorithm minimizes the expected variance to go with updated predictions. We prove that suboptimality of this model predictive algorithm vanishes as time horizon expands in the average case analysis. Further, we prove strong concentration results on the distribution of the load variance obtained by model predictive deferrable load control. These concentration results highlight that the typical performance of model predictive deferrable load control is tightly concentrated around the average-case performance. Finally, we evaluate the algorithm via trace-based simulations.

Contents

Chapter 1

Introduction

The electricity grid is expected to change dramatically over the coming decades. Conventional coal and nuclear generation is being rapidly substituted by renewable generation such as wind and solar [9]. However, these renewables are difficult to predict. For example, wind generation prediction has a root-mean-square error of around 18% of the nameplate capacity looking 24 hours ahead [23]. Such high uncertainty in generation calls the traditional control strategy of "generation follows demand" into question.

Real-time demand response programs seek to induce dynamic demand management of customers' electricity load in response to power supply conditions, e.g., by reducing or deferring power consumption in response to requests from the utility. Such programs have the potential to compensate for the uncertainties in renewables in real-time so as to ease the incorporation of renewable energy into the grid, and so are recognized as priority areas for the future smart grid by both the National Institute of Standards and Technology [37] and the Department of Energy [16].

The success of demand response depends on the willingness and ability of consumers' electrical loads to be deferred over time. Such *deferrable loads* are expected to take many forms, e.g., plug-in electric vehicles, dryers, air conditioners, etc. The penetration of deferrable loads is expected to grow significantly in the coming years as a result of increasing penetration of electric vehicles and smart appliances [17]. This expected increase high-

lights the potential for scheduling deferrable loads in order to compensate for the random fluctuations of renewable energy.

However, realizing the potential of deferrable loads requires the coordination of a large number of distributed loads. Current approaches for achieving such coordination include 1) direct load control (DLC) by load serving entities (LSE) [27, 34, 19, 20], and 2) time-of-use pricing and other complex pricing structures [4, 11, 31]. DLC is the focus of this thesis since the LSE has full control over the loads. Specifically, this thesis focuses on decentralized DLC algorithms. The motivation for this approach is that, as the penetration of deferrable loads grows, the scale of the task of controlling deferrable loads will prevent centralized control and so distributed, decentralized coordination will become necessary.

Related work There is a growing body of work on decentralized direct load control algorithms. This literature focuses on both evaluating algorithms in simulation-based evaluations [3, 36, 28] and on deriving theoretical performance guarantees [34, 19]. For example, [34] proposes a decentralized charging strategy for electric vehicles (EV) that is optimal if all EVs are identical, and [19] provides an algorithm for the setting when EVs are not necessarily identical.

Typically, the algorithms proposed in the literature, e.g., [3, 36, 28, 34, 19], have not considered uncertainties in renewable generation and deferrable load arrivals. However, of course, only predictions of these quantities are known ahead of time in practice, and the impact of prediction errors can be dramatic, e.g., see Figure 5.2.

Only very recently have algorithms that consider the uncertainties in renewable generation and deferrable load arrivals been proposed. Most of these works focus on simulation-based studies, e.g., [14, 10, 15]; however some work does derive analytic performance guarantees [39, 13, 32, 8]. For example, reference [13] proposes an algorithm that achieves the optimal competitive ratio in the case where renewable generation is precisely known (and constant) and [32] proposes an algorithm with some worst-case performance guarantees. Note that, while the algorithms proposed in [13, 32] are analyzed with a "worst-case"

perspective, this thesis first focuses on the "average-case" perspective, then we show via distributional analysis that the "average case" is indeed the representative case.

Summary of contributions We provide a model predictive algorithm for decentralized deferrable load control in the context of uncertain predictions about both future loads and future renewable generation. More specifically, in this thesis we propose a novel extension of the "optimal deferrable load control problem" studied in [19]. This extension incorporates uncertainty about both deferrable and non-deferrable loads, in addition to inexact predictions of renewable generation; and then uses this problem to derive a new algorithm for deferrable load control. Further, we perform both analytic and trace-based performance analysis of the algorithm in order to quantify the impact of prediction uncertainties on deferrable load control. In particular, the contributions of the work are threefold.

First, we model renewable generation prediction as a Wiener filtering process [41] (Section 2.1), that is able to model any zero mean, stationary prediction evolutions. Additionally, the formulation includes a very general model for deferable loads that allows for heterogeneous deadlines and maximum charging rates, as well as stochastic arrivals.

Second, in the context of this model, we introduce a model predictive algorithm for deferrable load control with uncertainty (Section 3.2). The model predictive algorithm essentially solves a series of optimal control problems whose horizon lengths shrink with time. At any time, the algorithm uses only the information that is available, i.e., specifications of deferrable loads that have already arrived and predictions on future loads and renewable generation. In this sense, the algorithm we propose is a (non-trivial) extension of the algorithm proposed in [19], which applies only in the case of exact knowledge of loads and renewables. A key technique introduced by the algorithm is the concept of a "pseudo deferrable load," which is simulated at the utility to represent future deferrable load arrivals.

Third, we perform a detailed performance analysis of our proposed algorithm. The performance analysis uses both analytic results and trace-based experiments to study (i) the

reduction in expected load variance achieved via deferrable load control, and (ii) the value of using model predictive control via our algorithm when compared with static (open-loop) control. *The theorems in Section 4.1 characterize the impact of prediction inaccuracy on deferrable load control.* These analytic results highlight that as the time horizon expands, the expected load variance obtained by our proposed algorithm approaches the optimal value (Corollary 3). Also, as the time horizon expands, the algorithm obtains an increasing variance reduction over the optimal static algorithm (Corollary 5, 6). Furthermore, in Section 5 we provide trace-based experiments using data from Southern California Edison and Alberta Electric System Operator to validate the analytic results. These experiments highlight that our proposed algorithm obtains a small suboptimality under high uncertainties of renewable generation, and has significant performance improvement over the optimal static control.

The model predictive algorithm for controlling deferrable load as well as the average case analysis of the performance is presented in [21], while the worst case analysis and distributional analysis of the performance of this algorithm can be found in [12].

Chapter 2

Real-time Deferrable Load Control

2.1 Model overview and Notation

this thesis studies the design and analysis of real-time control algorithms for scheduling deferrable loads to compensate for the random fluctuations in renewable generation. In the following we present a model of this scenario that serves as the basis for our algorithm design and performance evaluation. The model includes renewable generation, non-deferrable loads, and deferrable loads, which are described in turn. The key differentiation of this model from that of [19] is the inclusion of uncertainties (prediction errors) on future renewable generation and loads.

Throughout, we consider a discrete-time model over a finite time horizon. The time horizon is divided into T time slots of equal length and labeled $1, \ldots, T$. In practice, the time horizon could be one day and the length of a time slot could be 10 minutes.

2.2 Renewable generation and non-deferrable load

Renewable generation like wind is stochastic and difficult to predict. Similarly, non-deferrable load including lights are hard to predict at a low aggregation level.

Since the focus is on scheduling deferrable loads, we aggregate renewable generation and non-deferrable load into one process termed the *base load*, $b = \{b(\tau)\}_{\tau=1}^{T}$, which is

defined as the difference between non-deferrable load and renewable generation, and is a stochastic process.

To model the uncertainty of base load, we use a causal filter based model described as follows, and illustrated in Figure 2.1. In particular, the base load at time τ is modeled as a random deviation $\delta b = \{\delta b(\tau)\}_{\tau=1}^{T}$ around its expectation $\bar{b} = \{\bar{b}(\tau)\}_{\tau=1}^{T}$. The process $\bar{b}$ is specified externally to the model, e.g., from historical data and weather report, and the process $\delta b(\tau)$ is further modeled as an uncorrelated sequence of identically distributed random variables $e = \{e(\tau)\}_{\tau=1}^{T}$ with mean 0 and variance σ^2, passing through a causal filter. Specifically, let $f = \{f(\tau)\}_{\tau=-\infty}^{\infty}$ denote the impulse response of this causal filter and assume that $f(0) = 1$, then $f(\tau) = 0$ for $\tau < 0$ and

$$\delta b(\tau) = \sum_{s=1}^{T} e(s) f(\tau - s), \quad \tau = 1, \ldots, T.$$

At time $t = 1, \ldots, T$, a prediction algorithm can observe the sequence $e(s)$ for $s = 1, \ldots, t$, and predicts b as[1]

$$b_t(\tau) = \bar{b}(\tau) + \sum_{s=1}^{t} e(s) f(\tau - s), \quad \tau = 1, \ldots, T. \tag{2.1}$$

Note that $b_t(\tau) = b(\tau)$ for $\tau = 1, \ldots, t$ since f is causal.

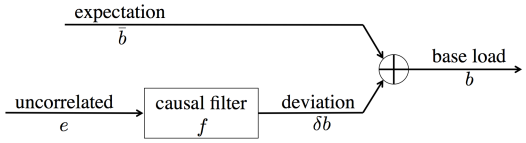


Figure 2.1: Diagram of the notation and structure of the model for base load, i.e., non-deferrable load minus renewable generation.

This model allows for non-stationary base load through the specification of $\bar{b}$ and a broad class of models for uncertainty via f and e. In particular, two specific filters f that

[1] This prediction algorithm is a Wiener filter [41].

we consider in detail later in the paper are:

1. A filter with finite but flat impulse response, i.e., there exists $\Delta > 0$ such that

$$f(t) = \begin{cases} 1 & \text{if } 0 \leq t < \Delta \\ 0 & \text{otherwise;} \end{cases}$$

2. A filter with an infinite and exponentially decaying impulse response, i.e., there exists $a \in (0, 1)$ such that

$$f(t) = \begin{cases} a^t & \text{if } t \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

These two filters provide simple but informative examples for our discussion in Section 4.1.

2.3 Deferrable load

To model deferrable loads we consider a setting where N deferrable loads arrive over the time horizon, each requiring a certain amount of electricity by a given deadline. Further, a real-time algorithm has imperfect information about the arrival times and sizes of these deferrable loads.

More specifically, we assume a total of N deferrable loads and label them in increasing order of their arrival times by $1, \ldots, N$, i.e., load n arrives no later than load $n + 1$ for $n = 1, \ldots, N - 1$. Further, we define $N(t)$ as the number of loads that arrive before (or at) time t for $t = 1, \ldots, T$ and fix $N(0) := 0$. Thus, load $1, \ldots, N(t)$ arrive before or at time t for $t = 1, \ldots, T$ and $N(T) = N$.

For each deferrable load, its arrival time and deadline, as well as other constraints on its power consumption, are captured via upper and lower bounds on its possible power consumption during each time. Specifically, the power consumption of deferrable load n at

time t, $p_n(t)$, must be between given lower and upper bounds $\underline{p}_n(t)$ and $\overline{p}_n(t)$, i.e.,

$$\underline{p}_n(t) \leq p_n(t) \leq \overline{p}_n(t), \quad n = 1, \ldots, N, \ t = 1, \ldots, T. \tag{2.2}$$

These are specified externally to the model. For example, if an electric vehicle plugs in with Level II charging, then its power consumption must be within $[0, 3.3]$kW. However, if it is not plugged in (has either not arrived yet or has already departed) then its power consumption is 0kW, i.e., within $[0, 0]$kW. Further, we assume that a deferrable load n must withdraw a fixed amount of energy P_n by its deadline, i.e.,

$$\sum_{t=1}^{T} p_n(t) = P_n, \quad n = 1, \ldots, N. \tag{2.3}$$

Finally, the N deferrable loads arrive randomly throughout the time horizon. Define

$$a(t) := \sum_{n=N(t-1)+1}^{N(t)} P_n \tag{2.4}$$

as the total energy request of all deferrable loads that arrive at time t for $t = 1, \ldots, T$. We assume that $\{a(t)\}_{t=1}^{T}$ is a sequence of independent identically distributed random variables with mean λ and variance s^2. Further, define

$$A(t) := \sum_{\tau=t+1}^{T} a(\tau) \tag{2.5}$$

as the total energy requested after time t for $t = 1, \ldots, T$.

In summary, at time $t = 1, \ldots, T$, a real-time algorithm has full information about the deferrable loads that have arrived, i.e., $\underline{p}_n$, $\overline{p}_n$, and P_n for $n = 1, \ldots, N(t)$, and knows the expectation of future deferrable load total energy request $\mathbb{E}(A(t))$. However, a real-time algorithm has no other knowledge about deferrable loads that arrive after time t.

2.4 The deferrable load control problem

We can now formally state the deferrable load control problem that is the focus of this thesis. Recall that the objective of real-time deferrable load control is to compensate for the random fluctuations in renewable generation and non-deferrable load in order to "flatten" the *aggregate load* $d = \{d(t)\}_{t=1}^{T}$, which is defined as

$$d(t) = b(t) + \sum_{n=1}^{N} p_n(t), \quad t = 1, \ldots, T. \tag{2.6}$$

In this thesis, we focus on minimizing the *sample path variance* of the aggregate load d, $V(d)$, as a measure of "flatness", that is defined as

$$V(d) = \frac{1}{T} \sum_{t=1}^{T} \left(d(t) - \frac{1}{T} \sum_{\tau=1}^{T} d(\tau) \right)^2. \tag{2.7}$$

We can now formally specify the optimal deferrable load control (ODLC) problem that we seek to solve:

$$\begin{aligned} \textbf{ODLC:} \ \min \quad & \frac{1}{T} \sum_{t=1}^{T} \left(d(t) - \frac{1}{T} \sum_{\tau=1}^{T} d(\tau) \right)^2 \\ \text{over} \quad & p_n(t), d(t), \quad \forall n, t \\ \text{s.t.} \quad & d(t) = b(t) + \sum_{n=1}^{N} p_n(t), \quad \forall t; \\ & \underline{p}_n(t) \le p_n(t) \le \overline{p}_n(t), \quad \forall n, t; \\ & \sum_{t=1}^{T} p_n(t) = P_n, \quad \forall n. \end{aligned} \tag{2.8}$$

In the above ODLC, the objective is simply the sample path variance of the aggregate load, $V(d)$, and the constraints correspond to equations (2.6), (2.2), and (2.3), respectively. We chose $V(d)$ as the objective for ODLC because of its significance for microgrid operators [26]. However, additionally, [19] has proven that the optimal solution does not change if

the objective function $V(d)$ is replaced by $f(d) = \sum_{t=1}^{T} U(d(t))$ where $U : \mathbb{R} \to \mathbb{R}$ is strictly convex. Hence, we can use $V(d)$ without loss of generality.

Chapter 3

Model Predictive Algorithm

Given the optimal deferrable load control (ODLC) problem defined in (2.8), the first contribution of this thesis is to design an algorithm that solves ODLC in real-time, given uncertain predictions of base and deferrable loads.

There are two key challenges for the algorithm design. First, the algorithm has access only to uncertain predictions at any given time, i.e., at time t the algorithm only knows deferrable loads 1 to $N(t)$ rather than 1 to N, and only knows the prediction b_t instead of b itself. Second, even if there was no uncertainty in predictions, solving the ODLC problem requires significant computational effort when there are a large number of deferrable loads.

Motivated by these challenges, in this section we design a decentralized algorithm with strong performance guarantees even when there is uncertainty in the predictions. The algorithm builds on the work of [19], which provides a decentralized algorithm for the case without uncertainty in predictions. We present the details of the algorithm from [19] in Section 3.1 and then present a modification of the algorithm to handle uncertain predictions in Section 3.2.

3.1 Load control without uncertainty

We start with the case where the algorithm has complete knowledge (no uncertainty) about base load and deferrable loads. In this context, the key algorithmic challenge is to solve

the ODLC problem in (2.8) via a decentralized algorithm. Such a decentralized algorithm was proposed in [19], and we summarize the algorithm and its analysis here.

Algorithm definition: The algorithm from [19] is given in Algorithm 1. It is iterative and the superscripts in brackets denote the round of iteration. In each iteration $k \geq 0$, there are two key steps: Step (ii) and (iii). In Step (ii), the utility calculates the average load $g^{(k)}$ and broadcasts it to all deferrable loads. Note that the utility only needs to know the reported schedule $p_n^{(k)}$, the base load b, and the number of deferrable loads N. It does not need to know the constraints of the deferable loads. In Step (iii), each deferrable load n updates $p_n^{(k+1)}$ by solving a convex optimization. The objective function has two terms. The first term can be interpreted as the electricity bill if the electricity price was set to $g^{(k)}$. The second term vanishes as iterations continue.

Algorithm convergence results: Importantly, though Algorithm 1 is iterative, it converges very fast. In fact, the simulations in [19] stop the iterations after 15 rounds (i.e., K=15) in all cases because convergence is already achieved. Further, Algorithm 1 provably solves the ODLC problem given in (2.8) when there is no uncertainty, i.e., when $N(t) = N$ and $b_t = b$ for $t = 1, \ldots, T$ [19]. More precisely, let $\mathcal{O}$ denote the set of optimal solutions to (2.8), and define $d(p, \mathcal{O}) := \min_{\hat{p} \in \mathcal{O}} \|p - \hat{p}\|$ as the distance from a deferrable load schedule p to optimal deferrable load schedules $\mathcal{O}$.

Proposition 1 ([19])**.** *When there is no uncertainty, i.e., $N(t) = N$ and $b_t = b$ for $t = 1, \ldots, T$, the deferrable load schedules $p^{(k)}$ obtained by Algorithm 1 converge to optimal schedules to ODLC, i.e., $d(p^{(k)}, \mathcal{O}) \to 0$ as $k \to \infty$.*

A particular class of optimal solutions will be of interest to us later in the paper, so we define them here. Specifically, we call a feasible deferrable load schedule $p = (p_1, \ldots, p_N)$ *valley-filling*, if there exists some constant $C \in \mathbb{R}$ such that $\sum_{n=1}^{N} p_n(t) = (C - b(t))^+$ for $t = 1, \ldots, T$.

Proposition 2 ([19])**.** *If a valley-filling deferrable load schedule exists, then it solves*

ODLC. Further, in such cases, all optimal schedules to ODLC have the same aggregate load.

Note that valley-filling schedules tend to exist if there is a large number of deferrable loads, in such settings optimal solutions to ODLC are valley-filling.

3.2 Load control with uncertainty

Algorithm 1 provides a decentralized approach for solving the ODLC problem; however it assumes exact knowledge (certainty) about base load and deferrable loads. In this section, we adapt Algorithm 1 to the setting where there is uncertainty in base load and deferrable load predictions, while maintaining strong performance guarantees. In particular, in this section we assume that at time t, only the prediction b_t is known, not b itself, and only information about deferrable loads 1 to $N(t)$ and the expectation of future energy requests $\mathbb{E}(A(t))$ are known.

Algorithm definition: To adapt Algorithm 1 to deal with uncertainty, the first step is straightforward. In particular, it is natural to replace the base load b by its prediction b_t in Algorithm 1 to deal with the unavailability of b.

However, dealing with unavailable future deferrable load information is trickier. To do this we use a pseudo deferrable load, which is simulated at the utility, to represent future deferrable loads. More specifically, let $q = \{q(\tau)\}_{\tau=t}^{T}$ with $q(t) = 0$ denote the power consumption of the pseudo load, and assume that it requests $\mathbb{E}(A(t))$ amount of energy, i.e.,

$$\sum_{\tau=t}^{T} q(\tau) = \mathbb{E}(A(t)). \tag{3.1}$$

We also assume that q is point-wise upper and lower bounded by some upper and lower bounds $\overline{q}$ and $\underline{q}$, i.e.,

$$\underline{q}(\tau) \leq q(\tau) \leq \overline{q}(\tau), \quad \tau = t, \ldots, T. \tag{3.2}$$

Note that $\underline{q}(t) = \overline{q}(t) = 0$. The bounds $\underline{q}$ and $\overline{q}$ should be set according to historical data. Here, for simplicity, we consider them to be $\underline{q}(\tau) = 0$ and $\overline{q}(\tau) = \infty$ for $\tau = t+1, \ldots, T$.

Given the above setup, the utility solves the following problem at every time slot $t = 1, \ldots, T$, to accommodate the availability of only partial information.

$$
\begin{aligned}
\textbf{ODLC-t:} \quad \min \quad & \sum_{\tau=t}^{T} \left(d(\tau) - \frac{1}{T-t+1} \sum_{s=t}^{T} d(s) \right)^2 \\
\text{over} \quad & p_n(\tau), q(\tau), d(\tau), \quad n \le N(t), \tau \ge t \\
\text{s.t.} \quad & d(\tau) = b_t(\tau) + \sum_{n=1}^{N(t)} p_n(\tau) + q(\tau), \quad \tau \ge t; \\
& \underline{p}_n(\tau) \le p_n(\tau) \le \overline{p}_n(\tau), \quad n \le N(t), \ \tau \ge t; \\
& \sum_{\tau=t}^{T} p_n(\tau) = P_n(t), \quad n \le N(t); \\
& \underline{q}(\tau) \le q(\tau) \le \overline{q}(\tau), \quad \tau \ge t; \\
& \sum_{\tau=t}^{T} q(\tau) = \mathbb{E}(A(t))
\end{aligned}
\tag{3.3}
$$

where $P_n(t) = P_n - \sum_{\tau=1}^{t-1} p_n(\tau)$ is the energy to be consumed at or after time t, for $n = 1, \ldots, N(t)$ and $t = 1, \ldots, T$.

Now, adjusting Algorithm 1 to solve ODLC-t gives Algorithm 2, which is real-time and shrinking-horizon. Note that if base load prediction is exact (i.e., $b_t = b$ for $t = 1, \ldots, T$) and all deferrable loads arrive at the beginning of the time horizon (i.e., $N(t) = N$ for $t = 1, \ldots, T$), then ODLC-1 reduces to ODLC and Algorithm 2 reduces to Algorithm 1.

Algorithm convergence results: We provide analytic guarantees on the convergence and optimality of Algorithm 2. In particular, we prove that Algorithm 2 solves ODLC-t at every time slot t. Specifically, let $\mathcal{O}(t)$ denote the set of optimal schedules to ODLC-t, and define $d(p, \mathcal{O}(t)) := \min_{(\hat{p},\hat{q}) \in \mathcal{O}(t)} \|p - \hat{p}\|$ as the distance from a schedule p to optimal schedules $\mathcal{O}(t)$ at time t, for $t = 1, \ldots, T$.

Theorem 1. *At time $t = 1, \ldots, T$, the deferrable load schedules $p^{(k)}$ obtained by Algorithm*

2 converge to optimal schedules to ODLC-t, i.e., $d(p^{(k)}, \mathcal{O}(t)) \to 0$ *as* $k \to \infty$.

The theorem is proved in Appendix A.1. Though iterative, Algorithm 2 converges fast, similar to Algorithm 1. In the simulations, setting $K = 15$ is enough for all test cases.

Similar to Proposition 2, "t-valley-filling" provides a simple characterization of the solutions to ODLC-t. Specifically, at time $t = 1, \ldots, T$, a feasible schedule (p, q) is called *t-valley-filling*, if there exists some constant $C(t) \in \mathbb{R}$ such that

$$q(\tau) + \sum_{n=1}^{N(t)} p_n(\tau) = (C(t) - b_t(\tau))^+, \quad \tau = t, \ldots, T. \tag{3.4}$$

Given this definition of t-valley-filling, the following corollary follows immediately from Proposition 2.

Corollary 1. *At time* $t = 1, \ldots, T$*, a t-valley-filling deferrable load schedule, if exists, solves ODLC-t. Furthermore, in such cases, all optimal schedules to ODLC-t have the same aggregate load.*

This corollary serves as the basis for the performance analysis we perform in Section 4.1. Remember that t-valley-filling schedules tend to exist in cases where there are a large numbers of deferrable loads.

Algorithm 1 Deferrable load control without uncertainty

Input: The utility knows the base load b and the number N of deferrable loads. Each load $n \in \{1, \ldots, N\}$ knows its energy request P_n and power consumption bounds $\overline{p}_n$ and $\underline{p}_n$. The utility sets K, the number of iterations.

Output: Deferrable load schedule $p = (p_1, \ldots, p_N)$.

(i) Set $k \leftarrow 0$ and intitialize the schedule $p^{(k)}$ as

$$p_n^{(k)}(t) \leftarrow 0, \quad t = 1, \ldots, T, \;\; n = 1, \ldots, N.$$

(ii) The utility calculates the average load $g^{(k)} = d^{(k)}/N$,

$$g^{(k)}(t) \leftarrow \frac{1}{N}\left(b(t) + \sum_{n=1}^{N} p_n^{(k)}(t)\right), \quad t = 1, \ldots, T,$$

and broadcasts $g^{(k)}$ to all deferrable loads.

(iii) Each load n updates a new schedule $p_n^{(k+1)}$ by solving

$$\begin{aligned} \min \quad & \sum_{\tau=1}^{T} g^{(k)}(\tau) p_n(\tau) + \frac{1}{2}\left(p_n(\tau) - p_n^{(k)}(\tau)\right)^2 \\ \text{over} \quad & p_n(1), \ldots, p_n(T) \\ \text{s.t.} \quad & \underline{p}_n(\tau) \le p_n(\tau) \le \overline{p}_n(\tau), \quad \forall \tau; \\ & \sum_{\tau=1}^{T} p_n(\tau) = P_n, \end{aligned}$$

and reports $p_n^{(k+1)}$ to the utility.

(iv) Set $k \leftarrow k + 1$. If $k < K$, go to Step (ii).

Algorithm 2 Deferrable load control with uncertainty

Input: At time t, the utility knows the prediction b_t of base load and the number $N(t)$ of deferrable loads. Each deferrable load $n \in \{1, \ldots, N(t)\}$ knows its future energy request $P_n(t)$ and power consumption bounds $\overline{p}_n$ and $\underline{p}_n$. The utility sets K, the number iterations.

Output: At time t, output the power consumption $p_n(t)$ for deferrable loads $1, \ldots, N(t)$.

At time slot $t = 1, \ldots, T$:

(i) Set $k \leftarrow 0$. Each deferrable load $n \in \{1, \ldots, N(t)\}$ initializes its schedule $\{p_n^{(0)}(\tau)\}_{\tau=t}^{T}$ as

$$p_n^{(0)}(\tau) \leftarrow \begin{cases} p_n^{(K)}(\tau) & \text{if } n \leq N(t-1) \\ 0 & \text{if } n > N(t-1) \end{cases}, \quad \tau = t, \ldots, T$$

where $p_n^{(K)}$ is the schedule of load n in iteration K of the previous time slot $t-1$.

(ii) The utility solves

$$\begin{aligned} \min \quad & \sum_{\tau=t}^{T} \left(b_t(\tau) + \sum_{n=1}^{N(t)} p_n^{(k)}(\tau) + q(\tau) \right)^2 \\ \text{over} \quad & q(t), \ldots, q(T) \\ \text{s.t.} \quad & \underline{q}(\tau) \leq q(\tau) \leq \overline{q}(\tau), \quad \tau \geq t; \\ & \sum_{\tau=t}^{T} q(\tau) = \mathbb{E}(A(t)) \end{aligned}$$

to obtain a pseudo schedule $\{q^{(k)}(\tau)\}_{\tau=t}^{T}$. The utility then calculates the average aggregate load per deferrable load $g^{(k)}$ as

$$g^{(k)}(\tau) \leftarrow \frac{1}{N(t)} \left(b_t(\tau) + \sum_{n=1}^{N(t)} p_n^{(k)}(\tau) + q^{(k)}(\tau) \right)$$

for $\tau = t, \ldots, T$, and broadcasts $\{g^{(k)}(\tau)\}_{\tau=t}^{T}$ to deferrable loads $n = 1, \ldots, N(t)$.

(iii) Each deferrable load $n = 1, \ldots, N(t)$ solves

$$\begin{aligned} \min \quad & \sum_{\tau=t}^{T} g^{(k)}(\tau) p_n(\tau) + \frac{1}{2} \left(p_n(\tau) - p_n^{(k)}(\tau) \right)^2 \\ \text{over} \quad & p_n(t), \ldots, p_n(T) \\ \text{s.t.} \quad & \underline{p}_n(\tau) \leq p_n(\tau) \leq \overline{p}_n(\tau), \quad \tau \geq t; \\ & \sum_{\tau=t}^{T} p_n(\tau) = P_n(t), \end{aligned}$$

to obtain a new schedule $\{p_n^{(k+1)}(\tau)\}_{\tau=t}^{T}$, and reports $\{p_n^{k+1}(\tau)\}_{\tau=t}^{T}$ to the utility.

(iv) Set $k \leftarrow k+1$. If $k < K$, go to Step (ii).

(v) Deferrable load $n \in \{1, \ldots, N(t)\}$ sets $p_n(t) \leftarrow p_n^K(t)$ and $P_n(t+1) \leftarrow P_n(t) - p_n(t)$.

Chapter 4

Performance Analysis

4.1 Average-case Analysis

To this point, we have shown that Algorithm 2 makes "optimal" decisions with the information available at every time slot, i.e., it solves ODLC-t at time t for $t = 1, \ldots, T$. However, these decisions are still suboptimal compared to what could be achieved if exact information was available. In this section, our goal is to understand the impact of uncertainty on the performance. In particular, we study two questions:

(i) How do the uncertainties about base load and deferrable loads impact the expected sample path load variance obtained by Algorithm 2?

(ii) What is the improvement of using the real-time control provided by Algorithm 2 over using the optimal static control?

Our answers to these questions are below. Throughout, we focus on the special, but practically relevant, case when a t-valley-filling schedule exists at every time $t = 1, \ldots, T$. As we have mentioned previously, when the number of deferrable loads is large this is a natural assumption that holds for practical load profiles. The reason for making this assumption is that it allows us to use the characterization of optimal schedules given in (3.4). In fact, without loss of generality, we further assume $C(t) \geq b_t(\tau)$ for $\tau = t, \ldots, T$,

under which (3.4) implies

$$d(t) = C(t) = \frac{1}{T-t+1}\left(\sum_{\tau=t}^{T} b_t(\tau) + \mathbb{E}(A(t)) + \sum_{n=1}^{N(t)} P_n(t)\right) \tag{4.1}$$

for $t = 1, \ldots, T$. Thus, equation (4.1) defines the model we use for the performance analysis of Algorithm 2.

4.1.1 The expected load variance of Algorithm 2

We start by calculating the expected load variance, $\mathbb{E}(V)$, of Algorithm 2. The goal is to understand how uncertainty about base load and deferrable loads impacts the load variance. Note that, if there are no base load prediction errors and deferrable loads arrive at the beginning of the time horizon, then Algorithm 2 obtains optimal schedules that have zero load variance. In contrast, when there are base load prediction errors and stochastic deferrable load arrivals, the expected load variance is given by the following theorem.

To state the result, recall that $\{f(t)\}_{t=-\infty}^{\infty}$ is the causal filter modeling the correlation of base load and define $F(t) := \sum_{s=0}^{t} f(s)$ for $t = 0, \ldots, T$.

Theorem 2. *The expected load variance* $\mathbb{E}(V)$ *obtained by Algorithm 2 is*

$$\mathbb{E}(V) = \frac{s^2}{T}\sum_{t=2}^{T}\frac{1}{t} + \frac{\sigma^2}{T^2}\sum_{t=0}^{T-1} F^2(t)\frac{T-t-1}{t+1}. \tag{4.2}$$

The theorem is proved in Appendix A.4.

Theorem 2 explicitly states the interaction of the variability of base load prediction (σ) and deferrable load prediction (s) with the horizon length T. Besides, it highlights the correlation of base load prediction error through F. More specifically, the expected load variance $\mathbb{E}(V)$ tends to 0 as the uncertainties in base load and deferrable loads vanish, i.e., $\sigma \to 0$ and $s \to 0$.

Corollary 2. *The expected load variance* $\mathbb{E}(V) \to 0$ *as* $\sigma \to 0$ *and* $s \to 0$.

Another remark about Theorem 2 is that the two terms in (4.2) correspond to the impact of the uncertainties in deferrable loads and base load respectively. In particular, Theorem 2 is proved in Section A.4 by analyzing these two cases separately and then combining the results. Specifically, the following two lemmas are the key pieces in the proof of Theorem 2, but are also of interest in their own right.

Lemma 1. *If there is no base load prediction error, i.e., $b_t = b$ for $t = 1, \ldots, T$, then the expected load variance obtained by Algorithm 2 is*

$$\mathbb{E}(V) = s^2 \frac{\sum_{t=2}^{T} \frac{1}{t}}{T} \approx s^2 \frac{\ln T}{T}.$$

The lemma is proved in Appendix A.2.

Lemma 2. *If there are no deferrable load arrivals after time 1, i.e., $N(t) = N$ for $t = 1, \ldots, T$, then the expected load variance obtained by Algorithm 2 is*

$$\mathbb{E}(V) = \frac{\sigma^2}{T^2} \sum_{t=0}^{T-1} F^2(t) \frac{T-t-1}{t+1}.$$

The lemma is proved in Appendix A.3.

Lemma 1 highlights that the more uncertainty in deferrable load arrival, i.e., the larger s, the larger the expected load variance $\mathbb{E}(V)$. On the other hand, the longer the time horizon T, the smaller the expected load variance $\mathbb{E}(V)$.

Similarly, Lemma 2 highlights that a larger base load prediction error, i.e., a larger σ, results in a larger expected load variance $\mathbb{E}(V)$. However, if the impulse response $\{f(t)\}_{t=-\infty}^{\infty}$ of the modeling filter of the base load decays fast enough with t, then the following corollary highlights that the expected load variance actually tends to 0 as time horizon T increases despite the uncertainty of base load predictions.

Corollary 3. *If there are no deferrable load arrivals after time 1, i.e., $N(t) = N$ for $t = 1, \ldots, T$, and $|f(t)| \sim O(t^{-1/2-\alpha})$ for some $\alpha > 0$, then the expected load variance*

obtained by Algorithm 2 satisfies $\mathbb{E}(V) \to 0$ *as* $T \to \infty$.

The corollary is proved in Appendix A.5.

4.1.2 Improvement over static control

The goal of this section is to quantify the improvement of real-time control via Algorithm 2 over the optimal static (open-loop) control. To be more specific, we compare the expected load variance $\mathbb{E}(V)$ obtained by the real-time control Algorithm 2, with the expected load variance $\mathbb{E}(V')$ obtained by the optimal static control, which only uses base load prediction at the beginning of the time horizon (i.e., $\bar{b}$) to compute deferrable load schedules. We assume $N(t) = N$ for $t = 1, \ldots, T$ in this section since otherwise any static control cannot obtain a schedule for all deferrable loads. Thus, the interpretation of the results that follow is as a quantification of the value of incorporating updated base load predictions into the deferrable load controller.

To begin the analysis, note that $\mathbb{E}(V)$ for this setting is given in Lemma 2. Further, it can be verified that the optimal static control is to solve ODLC with b replaced by $\bar{b}$, and the corresponding expected load variance $\mathbb{E}(V')$ is given by the following lemma.

Lemma 3. *If there is no stochastic load arrival, i.e.,* $N(t) = N$ *for* $t = 1, \ldots, T$*, then the expected load variance* $\mathbb{E}(V')$ *obtained by the optimal static control is*

$$\mathbb{E}(V') = \frac{\sigma^2}{T^2} \sum_{t=0}^{T-1} \left(T(T-t) f^2(t) - F^2(t) \right).$$

The lemma is proved in Appendix A.6.

Next, comparing $\mathbb{E}(V)$ and $\mathbb{E}(V')$ given in Lemma 2 and 3 shows that Algorithm 2 always obtains a smaller expected load variance than the optimal static control. Specifically,

Corollary 4. *If there is no deferrable load arrival after time* 1*, i.e.,* $N(t) = N$ *for* $t =$

$1, \ldots, T$, *then*

$$\mathbb{E}(V') - \mathbb{E}(V) = \frac{\sigma^2}{T} \sum_{t=1}^{T} \frac{1}{2t} \sum_{m=0}^{t-1} \sum_{n=0}^{t-1} (f(m) - f(n))^2 \geq 0.$$

The corollary is proved in the extended version [22].

Corollary 4 highlights that Algorithm 2 is guaranteed to obtain a smaller expected load variance than the optimal static control. The next step is to quantify how much smaller $\mathbb{E}(V)$ is in comparison with $\mathbb{E}(V')$.

To do this we compute the ratio $\mathbb{E}(V')/\mathbb{E}(V)$. Unfortunately, the general expression for the ratio is too complex to provide insight, so we consider two representative cases for the impulse response $f(t)$ of the causal filter in order to obtain insights. Specifically, we consider examples (i) and (ii) from Section 2.2. Briefly, in (i) $f(t)$ is finite and in (ii) $f(t)$ is infinite but decays exponentially in t. For these two cases, the ratio $\mathbb{E}(V')/\mathbb{E}(V)$ is summarized in the following corollaries.

Corollary 5. *If there is no deferrable load arrival after time* 1*, i.e.,* $N(t) = N$ *for* $t = 1, \ldots, T$*, and there exists* $\Delta > 0$ *such that*

$$f(t) = \begin{cases} 1 & \text{if } 0 \leq t < \Delta \\ 0 & \text{otherwise,} \end{cases}$$

then

$$\frac{\mathbb{E}(V')}{\mathbb{E}(V)} = \frac{T/\Delta}{\ln(T/\Delta)} \left(1 + O\left(\frac{1}{\ln(T/\Delta)}\right)\right).$$

The corollary is proved in the extended version [22].

Corollary 6. *If there is no deferrable load arrival after time* 1*, i.e.,* $N(t) = N$ *for* $t =$

$1, \ldots, T$, *and there exists* $a \in (0, 1)$ *such that*

$$f(t) = \begin{cases} a^t & \text{if } t \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

then

$$\frac{\mathbb{E}(V')}{\mathbb{E}(V)} = \frac{1-a}{1+a} \frac{T}{\ln T} \left(1 + O\left(\frac{\ln \ln T}{\ln T}\right)\right).$$

The corollary is proved in the extended version [22].

Corollary 5 highlights that, in the case where f is finite, if we define $\lambda = T/\Delta$ as the ratio of time horizon to filter length, then the load reduction roughly scales as $\lambda / \ln(\lambda)$. Thus, the longer the time horizon is in comparison to the filter length, the larger expected load variance reduction we obtain from using Algorithm 2 as compared with the optimal static control.

Similarly, Corollary 6 highlights that, in the case where f is infinite and exponentially decaying, the expected load variance reduction scales with T as $T / \ln T$ with coefficient $(1-a)/(1+a)$. Thus, the smaller a is, which means the faster f dies out, the more load variance reduction we obtain by using real-time control. This is similar to having a smaller Δ in the previous case.

4.2 Worst-case analysis

The results surveyed above highlight that Algorithm 2 performs well on average; however, it is often important to guarantee more than good average case performance. For that reason, many results in the literature focus on worst case analysis, e.g., [30, 33, 6]. While no existing results apply directly to the setting of this thesis, it is straightforward to see that the worst-case performance of Algorithm 2 is quite bad.

To see this, let us consider a setting where the prediction error for generation, e, and deferrable load, a, have bounded deviations from their means (0 and λ respectively).

Definition 1. *We say that* ***prediction errors are bounded*** *if there exist* ϵ_1 *and* ϵ_2 *such that, at any time* $t = 1, \ldots, T$*,*

$$|a(t) - \lambda| \leq \epsilon_1, \ |e(t)| \leq \epsilon_2. \tag{4.3}$$

In this situation, it is straightforward to see that the worst case performance of Algorithm 2 can potentially be quite bad. For two real numbers $a, b \in \mathbb{R}$, define $a \vee b := \max\{a, b\}$.

Proposition 3. *If a* t*-valley-filling solution exists for* $t = 1, 2, \ldots, T$*, and prediction errors are bounded by* ϵ_1 *and* ϵ_2 *as in* (4.3)*, then the worst-case load variance* $\sup_{a,e} V$ *achieved by Algorithm 2 is*

$$\begin{aligned} \sup_{a,e} \ V \ &= \ \epsilon_1^2 \left(1 - \frac{1}{T} \sum_{k=1}^{T} \frac{1}{k}\right) \\ &\quad + \frac{\epsilon_2^2}{T^2} \sum_{\tau=0}^{T-1} \sum_{s=0}^{T-1} \left(\frac{T}{\tau \vee s + 1} - 1\right) |F(\tau) F(s)|. \end{aligned}$$

The worst-case performance is achieved when all prediction errors on the load arrivals are equal to ϵ_1 while all prediction errors on the generation are equal to ϵ_2 in magnitude with the appropriate signs—the case where $a(t) = \lambda + \epsilon_1$ and $e(t) = \epsilon_2 \cdot \operatorname{sgn}(F(T - t))$ for all t.

Corollary 7. *If a* t*-valley-filling solution exists for* $t = 1, 2, \ldots, T$*, and prediction errors are bounded by* ϵ_1 *and* ϵ_2 *as in* (4.3)*, then the worst-case load variance* $\sup_{a,e} V$ *achieved by Algorithm 2 is lower bounded as*

$$\sup_{a,e} \ V \ \geq \ \epsilon_1^2 \left(1 - \frac{1}{T} \sum_{k=1}^{T} \frac{1}{k}\right) \ \approx \ \epsilon_1^2 \left(1 - \frac{\ln T}{T}\right).$$

Interestingly, the form of Corollary 7 implies that, in the worst-case, Algorithm 2 can be as bad as having no control at all: the time averaged load variance behaves like the worst one step load variance. Meanwhile, recall from Proposition **??** that the average performance $\mathbb{E}(V) \to 0$ as $T \to \infty$. Hence, while the the load variance V has a small mean $\mathbb{E}(V)$, it can be quite large in the worst case.

4.3 Distributional analysis

The contrast between the worst-case analysis (Proposition 3) and average-case analysis (Proposition **??**) motivates the main goal of this thesis – to understand how often the "bad cases," where V takes large values, happen. That is, we want to understand what typical variations of V under Algorithm 2 look like.

4.3.1 Concentration bounds

We start with analyzing the tail probability of V. Concretely, our focus is on

$$V_\eta := \min\{c \in \mathbb{R} \mid V \leq c \text{ with probability } \eta\},$$

which denotes the minimum value c such that $V \leq c$ with probability η for $\eta \in [0, 1]$. Our main result provides upper bounds on V_η, for large values of η, for arbitrary of prediction error distributions.

More specifically, we prove that *with high probability*, the load variance of Algorithm 2 does not deviate much from its average-case performance, i.e., we prove a concentration result for model predictive deferable load control.

Theorem 3. *Suppose a t-valley filling solution exists for $t = 1, 2, \ldots, T$, and prediction errors bounded by ϵ_1 and ϵ_2 as in* (4.3)*. Then the distribution of the load variance V*

obtained by Algorithm 2 satisfies a Bernstein type concentration, i.e.,

$$\mathbb{P}(V - \mathbb{E}V > t) \leq \exp\left(\frac{-t^2}{16\epsilon^2\lambda_1(2\mathbb{E}V + t)}\right) \tag{4.4}$$

where $\epsilon = \max(\epsilon_1, \epsilon_2)$ *and*

$$\lambda_1 = \frac{\ln T}{T} + \frac{1}{T^2}\sum_{t=0}^{T-1} F^2(t)\frac{T-t+1}{t+1}.$$

The theorem is proven in Appendix B.2.

To build intuition, the tail probability bound of V in (4.4) can be simplified for two different regimes of t as

$$\mathbb{P}(V - \mathbb{E}V > t) \leq \begin{cases} \exp\left(\frac{-t^2}{48\epsilon^2\lambda_1\mathbb{E}V}\right), & t < \mathbb{E}V \\ \exp\left(\frac{-t}{48\epsilon^2\lambda_1}\right), & t \geq \mathbb{E}V. \end{cases} \tag{4.5}$$

Though looser than that in (4.4), the tail bound in (4.5) highlights that V has a Gaussian tail probability bound when $t < \mathbb{E}V$ and an Exponential tail probability bound when $t \geq \mathbb{E}V$.

Theorem 3 relates the tail behavior of V with the maximum prediction error ϵ and the error correlation F over time. It implies that the actual performance of Algorithm 2 does not deviate much from its mean. To illustrate this, consider the following example where the prediction on baseload is precise, since the parameter λ_1 has a simple expression in this scenario.

Example 1. *Suppose that the baseload prediction is precise, i.e.,* $\epsilon_2 = 0$*. Then the average load variance is*

$$\mathbb{E}[V] = \frac{s^2}{T}\sum_{t=2}^{T}\frac{1}{t} \approx s^2 \ln T/T$$

and the tail bound in Theorem 3 can be simplified as

$$\mathbb{P}(V - \mathbb{E}V > c\mathbb{E}V) \leq \exp\left(-\frac{c^2}{2+c}\frac{s^2}{16\epsilon^2}\right).$$

Recall that constant s is the variance of a and constant ϵ is the maximum deviation of a from its mean. The above expression shows that, with high probability, V is at most a constant $c+1$ times of its mean $\mathbb{E}V$.

More generally, the quantity λ_1 controls the decaying speed of the tail bound in (4.4): the smaller λ_1, the faster the tail bound $\mathbb{P}(V - \mathbb{E}V > t)$ decays in t, and the load variance V achieved by Algorithm 2 concentrates sharper around its mean $\mathbb{E}V$. The following corollary highlights that λ_1 tends to 0 as T increases, provided that the error correlation $f(t)$ decays fast enough in t. Note that the condition on f is the same for Corollary 8 and Proposition **??**.

Corollary 8. *Under the assumptions of Theorem 3, if the error correlation $f \sim O(t^{-\frac{1}{2}-\alpha})$ for some $\alpha > 0$, then $\lambda_1 \to 0$ as $T \to \infty$.*

A detailed proof of Theorem 3 is included in the Appendix; however it is useful to provide some informal intuition for the argument used.

In general, tail probability bounds can be obtained by controlling the moments of a random variable. For example, the Markov inequality gives inverse linear tail probability bound using the first moment, and the Chebyshev inequality provides inverse quadratic tail probability bound using the second moment. However, the bound we obtained in Theorem 3 approaches 0 much faster for large t than the aforementioned Markov and Chebyshev bounds. This is done by controlling the moment generating function of V using the convex Log-Sobolev inequality.

A challenge in controlling the moment generating function of V is that, the most commonly used approach—the Martingale bounded difference approach [35]—only obtains very loose tail probability bounds in our case. This is because V can change dramatically when

one of the sources $a(t)$ or $e(t)$ of the randomness changes. Instead, we exploit the fact that the gradient of V is bounded by a linear function of itself (similar but slightly different from the "self-bounding" property defined in [7]). Using this property together with Log-Sobolev inequality in the product measure gives us a nice way to bound the entropy of V. After this we apply the Herbst's argument [29] to compute a good estimate on the concentration of V.

4.3.2 Bounds on the variance

To further understand the scale of typical load variance V under Algorithm 2, it is useful to also study the variance. In addition, the form of the variance highlights the impact of the tight concentration shown in Theorem 3.

Theorem 4. *Suppose a t-valley-filling solution exists for $t = 1, 2, \ldots, T$, and prediction errors are bounded by ϵ_1 and ϵ_2 as in* (4.3)*. Then the variance* $\mathrm{var}(V)$ *of V obtained by Algorithm 2 is bounded above by*

$$\mathrm{var}(V) \leq \left(\frac{4\epsilon_1 s \ln T}{T}\right)^2 + \left(\frac{4\epsilon_2 \sigma}{T^2} \sum_{t=0}^{T-1} F^2(t) \frac{T-t+1}{t+1}\right)^2. \tag{4.6}$$

To interpret this result, let $\overline{\mathrm{var}(V)}$ denote the upper bound on $\mathrm{var}(V)$ provided in (4.6). Theorem 4 implies that $\mathbb{E}V$ and $\sqrt{\overline{\mathrm{var}(V)}}$ scale similarly with T. In particular, the first term $\frac{s^2}{T}\sum_{t=2}^{T}\frac{1}{t}$ in $\mathbb{E}V$ scales with T as $\Omega(\ln T/T)$ while the first term $(4\epsilon_1 s \ln T/T)^2$ in $\overline{\mathrm{var}(V)}$ scales with T as $\Omega\left((\ln T/T)^2\right)$, and the second terms in $\mathbb{E}V$ and $\overline{\mathrm{var}(V)}$ have the same relationship. Hence, the standard deviation $\sqrt{\mathrm{var}(V)}$, which is upper bounded by $\sqrt{\overline{\mathrm{var}(V)}}$, is at most on the same scale as $\mathbb{E}V$ as T expands. It immediately follows from the Chebyshev inequality that V can only deviate significantly from $\mathbb{E}(V)$ with a small probability.

Corollary 9. *Under the assumptions in Theorem 4, for $t > 0$,*

$$\begin{aligned}&\mathbb{P}(|V - \mathbb{E}V| > t)\\ &\leq \frac{1}{t^2}\left[\left(\frac{4\epsilon_1 s \ln T}{T}\right)^2 + \left(\frac{4\epsilon_2\sigma}{T^2}\sum_{\tau=0}^{T-1} F^2(\tau)\frac{T-\tau+1}{\tau+1}\right)^2\right].\end{aligned} \tag{4.7}$$

While the tail bound (4.4) in Theorem 3 scales at least exponentially in t, the Chebyshev inequality only provides a tail bound (4.7) that scales inverse quadratically in t. Hence for large t, (4.4) provides a much tighter tail bound. However for small values of t, the tail bound (4.7) is usually tighter since the variance $\mathrm{var}(V)$ is well estimated in (4.6).

Furthermore, the variance $\mathrm{var}(V)$ vanishes as T expands, provided that $f(t)$ decays sufficiently fast as t grows, as formally stated in the following corollary.

Corollary 10. *Under the assumptions of Theorem 4, if the error correlation $f \sim O(t^{-\frac{1}{2}-\alpha})$ for some $\alpha > 0$, then $\mathrm{var}(V) \to 0$ as $T \to \infty$.*

Note that the condition on f parallels that in Proposition **??**.

Chapter 5

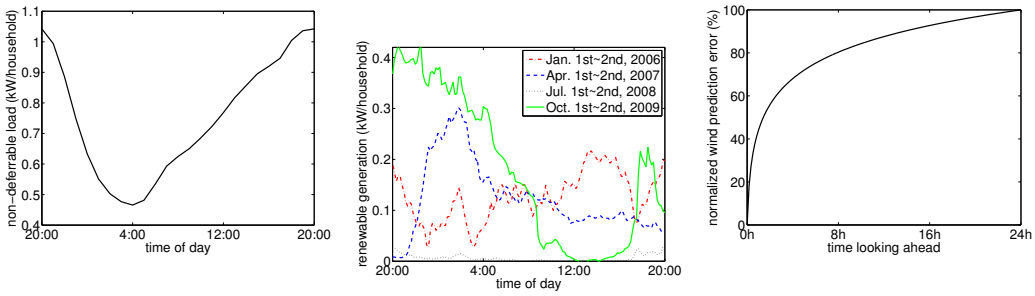
Simulation

In this Chapter we use trace-based experiments to explore the generality of the analytic results in the previous section. In particular, the results in the previous section characterize the expected load variance obtained by Algorithm 2 as a function of prediction uncertainties, and quantify the improvement of Algorithm 2 over the optimal static (open-loop) controller. However, the analytic results make simplifying assumptions on the form of uncertainties and solution schedules (equation (4.1)). Therefore, it is important to assess the performance of the algorithm using real-world data.

5.1 Experimental setup

The numerical experiments we perform use a time horizon of 24 hours, from 20:00 to 20:00 on the following day. The time slot length is 10 minutes, which is the granularity of the data we have obtained about renewable generation.

Base load Recall that base load is a combination of non-deferrable load and renewable generation. The non-deferrable load traces used in the experiments come from the average residential load in the service area of Southern California Edison in 2012 [38]. In the simulations, we assume that non-deferrable load is precisely known so that uncertainties in the base load only come from renewable generation. In particular, non-deferrable load



(a) non-deferrable load (b) wind generation (c) prediction error over time

Figure 5.1: Illustration of the traces used in the experiments. (a) shows the average residential load in the service area of Southern California Edison in 2012. (b) shows the total wind power generation of the Alberta Electric System Operator scaled to represent 20% penetration. (c) shows the normalized root-mean-square wind prediction error as a function of the time looking ahead for the model used in the experiments.

over the time horizon of a day is taken to be the average over the 366 days in 2012 as in Figure 5.1a, and assumed to be known to the utility at the beginning of the time horizon. In practice, non-deferrable load at the substation feeder level can be predicted within 1–3% root-mean-square error looking 24 hours ahead [18].

The renewable generation traces we use come from the 10-minute historical data for total wind power generation of the Alberta Electric System Operator from 2004 to 2009 [5]. In the simulations, we scale the wind power generation so that its average over the 6 years corresponds to a number of penetration levels in the range between 5% and 30%, and pick the wind power generation of a randomly chosen day as the renewable generation during each run. Figure 5.1b shows the wind power generation for four representative days, one for each season, after scaling to 20% penetration.

We assume that the renewable generation is not precisely known until it is realized, but that a prediction of the generation, which improves over time, is available to the utility. The modeling of prediction evolution over time is according to a martingale forecasting process [25, 24], which is a standard model for an unbiased prediction process that improves over time.

Specifically, the prediction model is as follows: For wind generation $w(\tau)$ at time τ,

the prediction error $w_t(\tau) - w(\tau)$ at time $t < \tau$ is the sum of a sequence of independent random variables $n_s(\tau)$ as

$$w_t(\tau) = w(\tau) + \sum_{s=t+1}^{\tau} n_s(\tau), \quad 0 \le t < \tau \le T.$$

Here $w_0(\tau)$ is the wind prediction without any observation, i.e., the expected wind generation $\bar{w}(\tau)$ at the beginning of the time horizon (used by static control).

The random variables $n_s(\tau)$ are assumed to be Gaussian with mean 0. Their variances are chosen as

$$\mathbb{E}(n_s^2(\tau)) = \frac{\sigma^2}{\tau - s + 1}, \quad 1 \le s \le \tau \le T$$

where $\sigma > 0$ is such that the root-mean-square prediction error $\sqrt{\mathbb{E}(w_0(T) - w(T))^2}$ looking T time slots (i.e., 24 hours) ahead is 0%–22.5% of the nameplate wind generation capacity.[1] According to this choice of the variances of $n_s(\tau)$, root-mean-square prediction error only depends on how far ahead the prediction is, in particular as in Figure 5.1c. This choice is motivated by [23].

Deferrable loads For simplicity, we consider the hypothetical case where all deferrable loads are electric vehicles. Since historical data for electric vehicle usage is not available, we are forced to use synthetic traces for this component of the experiments. Specifically, in the simulations the electric vehicles are considered to be identical, each requests 10kWh electricity by a deadline 8 hours after it arrives, and each must consume power at a rate within $[0, 3.3]$kW after it arrives and before its deadline.

In the simulations, the arrival process starts at 20:00 and ends at 12:00 the next day so that the deadlines of all electric vehicles lie within the time horizon of 24 hours. In each time slot during the arrival process, we assume that the number of arriving electric vehicles is uniformly distributed in $[0.8\lambda, 1.2\lambda]$, where λ is chosen so that electric vehicles

[1] Average wind generation is 15% of the nameplate capacity, so the root-mean-square prediction error looking T time slots ahead is 0%–150% the average wind generation.

(on average) account for 5%–30% of the non-deferrable loads. While this synthetic workload is simplistic, the results we report are representative of more complex setups as well.

Uncertainty about deferrable load arrivals is captured as follows. The prediction $\mathbb{E}(A(t))$ of future deferrable load total energy request is simply the arrival rate λ times the length of the rest of the arrival process $T' - t$ where T' is the end of the arrival process (12:00), i.e.,

$$\mathbb{E}(A(t)) = \lambda(T' - t), \quad t = 1, \ldots, T'.$$

If $t > T'$, i.e., the deferrable load arrival process has ended, then $\mathbb{E}(A(t)) = 0$.

Baselines for comparison Our goal in the simulations is to contrast the performance of Algorithm 2 with a number of common benchmarks to tease apart the impact of real-time control and the impact of different forms of uncertainty. To this end, we consider four controllers in our experiments:

(i) *Offline optimal control:* The controller has full knowledge about the base load and deferrable loads, and solves the ODLC problem offline. It is not realistic in practice, but serves as a benchmark for the other controllers since offline optimal control obtains the smallest possible load variance.

(ii) *Static control with exact deferrable load arrival information*: The controller has full knowledge about deferrable loads (including those that have not arrived), but uses only the prediction of base load that is available at the beginning of the time horizon to compute a deferrable load schedule that minimizes the expected load variance. This static control is still unrealistic since a deferrable load is known only after it arrives. But, this controller corresponds to what is considered in prior works, e.g., [34, 19, 20].

(iii) *Real-time control with exact deferrable load arrival information.* The controller has full knowledge about deferrable loads (including those that have not arrived), and

uses the prediction of base load that is available at the current time slot to update the deferrable load schedule by minimizing the expected load variance to go, i.e., Algorithm 2 with $N(t) = N$ for $t = 1, \ldots, T$. The control is unrealistic since a deferrable load is known only after it arrives; however it provides the natural comparison for case (ii) above.

(iv) *Real-time control without exact deferrable load arrival information, i.e., Algorithm 2.* This corresponds to the realistic scenario where only predictions are available about future deferrable loads and base load. The comparison with case (iii) highlights the impact of deferrable load arrival uncertainties.

The performance measure that we show in all plots is the "suboptimality" of the controllers, which we define as

$$\eta := \frac{V - V^{\text{opt}}}{V^{\text{opt}}},$$

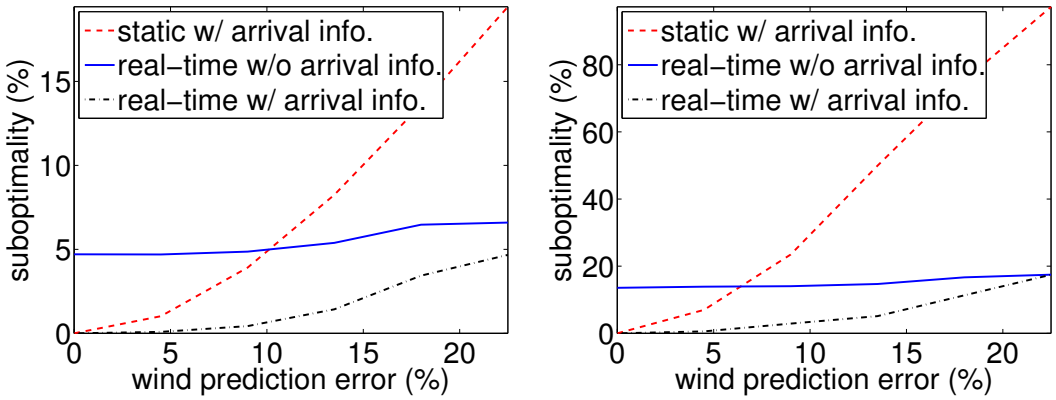
where V is the load variance obtained by the controller and V^{opt} is the load variance obtained by the offline optimal, i.e., case (i) above. Thus, the lines in the figures correspond to cases (ii)-(iv).

5.2 Experimental results

Our experimental results focus on two main goals: (i) understanding the impact of prediction accuracy on the expected load variance obtained by deferrable load control algorithms, and (ii) contrasting the real-time (closed-loop) control of Algorithm 2 with the optimal static (open-loop) controller. We focus on the impact of three key factors: wind prediction error, the penetration of deferrable load, and the penetration of renewable energy.

The impact of prediction error To study the impact of prediction error, we fix the penetration of both renewable generation (wind) and deferrable loads at 10% of non-deferrable load, and simulate the load variance obtained under different levels of root-

mean-square wind prediction errors (0%–22.5% of the nameplate capacity looking 24 hours ahead). The results are summarized in Figure 5.2a. It is not surprising that suboptimality of both the static and the real-time controllers that have exact information about deferrable load arrivals is zero when the wind prediction error is 0, since there is no uncertainty for these controllers in this case.



(a) Wind and deferrable load penetration are both 10%.

(b) Wind and deferrable load penetration are both 20%.

Figure 5.2: Illustration of the impact of wind prediction error on suboptimality of load variance.

As prediction error increases, the suboptimality of both the static and the real-time control increases. However, notably, the suboptimality of real-time control grows much more slowly than that of static control, and remains small (¡4.7%) if deferrable load arrivals are known, over the whole range 0%–22.5% of wind prediction error. At 22.5% prediction error, the suboptimality of static control is 4.2 times that of real-time control. This highlights that real-time control mitigates the influence of imprecise base load prediction over time.

Moving to the scenario where deferrable load arrivals are not known precisely, we see that the impact of this inexact information is less than 6.6% of the optimal variance. However, real-time control yields a load variance that is surprisingly resilient to the growth of wind prediction error, and eventually beats the optimal static control at around 10% wind prediction error, even though the optimal static control has exact knowledge of deferrable

loads and the adaptive control does not.

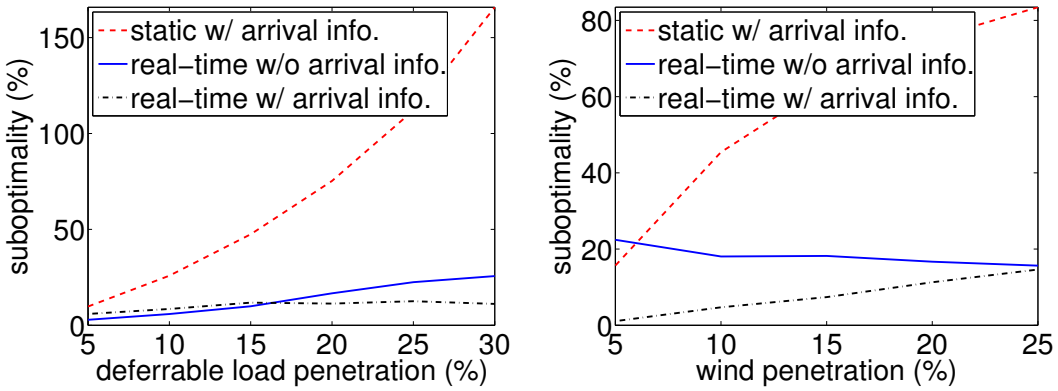
As prediction error increases, the suboptimality of the real-time control with or without deferrable load arrival information gets close, i.e., the benefit of knowing additional information on future deferrable load arrivals vanishes as base load uncertainty increases. This is because the additional information is used to overfit the base load prediction error.

The same comparison is shown in Figure 5.2b for the case where renewable and deferrable load penetration are both 20%. Qualitatively the conclusions are the same, however at this higher penetration the contrast between the resilience of adaptive control and static control is magnified, while the benefit of knowing deferrable load arrival information is minified. In particular, real-time control without arrival information beats static control with arrival information, at a lower (around 7%) wind prediction error, and knowing deferrable load arrival information does not reduce suboptimality of real-time control with 22.5% wind prediction error.

The impact of deferrable load penetration Next, we look at the impact of deferrable load penetration on the performance of the various controllers. To do this, we fix the wind penetration level to be 20% and wind prediction error looking 24 hours ahead to be 18%, and simulate the load variance obtained under different deferrable load penetration levels (5%–30%). The results are summarized in Figure 5.3a.

Not surprisingly, if future deferrable loads are known and uncertainty only comes from base load prediction error, then the suboptimality of real-time control is very small (¡11.2%) over the whole range 5%–30% of deferrable load penetration, while the suboptimality of static control increases with deferrable load penetration, up to as high as 166% (14.9 times that of real-time control) at 30% deferrable load penetration.

However, without knowing future deferrable loads, the suboptimality of real-time control increases with the deferrable load penetration. This is because larger amount of deferrable loads introduces larger uncertainties in deferrable load arrivals. But the suboptimality remains smaller than that of static control over the whole range 5%–30% of



(a) Impact of deferrable load penetration (b) Impact of wind penetration

Figure 5.3: Suboptimality of load variance as a function of (a) deferrable load penetration and (b) wind penetration. In (a) the wind penetration is 20% and in (b) the deferrable load penetration is 20%. In both, the wind prediction error looking 24 hours ahead is 18%.

deferrable load penetration. The highest suboptimality 25.7% occurs at 30% deferrable load penetration, and is less than 1/6 of the suboptimality of static control, which assumes exact deferrable load arrival information.

The impact of renewable penetration Finally, we study the impact of renewable penetration. To do this we fix the deferrable load penetration level to be 20% and the wind prediction error looking 24 hours ahead to be 18%, and simulate the load variance obtained by the 4 test cases under different wind penetration levels (5%–25%). The results are summarized in Figure 5.3b.

A key observation is that if future deferrable loads are known and uncertainty only comes from base load prediction error, then the suboptimality of real-time control grows much slower than that of static control, as wind penetration level increases. As explained before, this highlights that real-time control mitigates the impact of base load prediction error over time. In fact, the suboptimality of real-time control is small (¡15%) over the whole range 5%–25% of wind penetration levels. Of course, without knowledge of future

deferrable loads, the suboptimality of real-time control becomes bigger. However, it still eventually outperforms the optimal static controller at around 6% wind penetration, despite the fact that the optimal static controller is using exact information about deferrable loads.

5.3 A case study

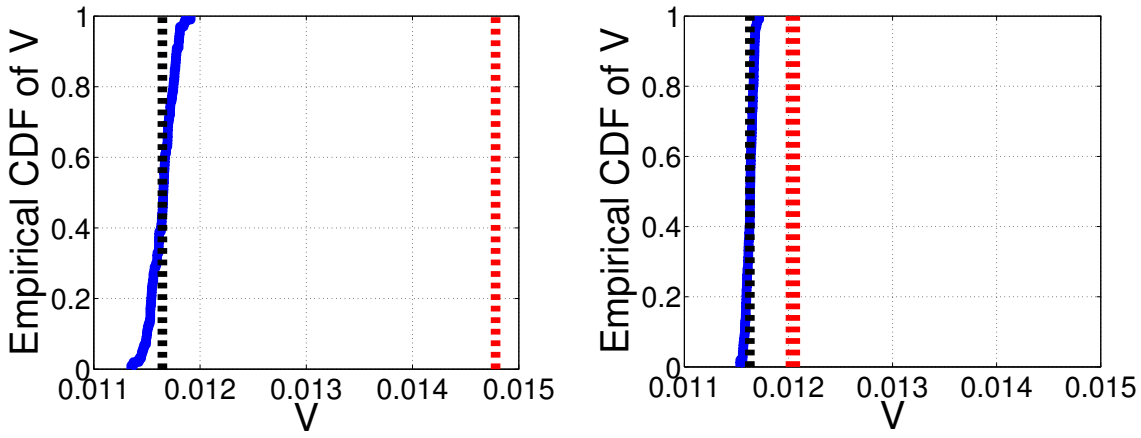
Theorems 3 and 4 provide theoretical guarantees that the load variance V obtained by Algorithm 2 concentrates around its mean, if prediction errors are bounded as in (4.3) and error correlation decays sufficiently fast (c.f. Corollary 2). Thus, they give the intuition that the expected performance of Algorithm 2 is a useful metric to focus on, and does indeed give an indication of the "typical" performance of the algorithm.

However, our analysis is based on the assumption that a t-valley-filling solution exists, which relies on the penetration of deferrable load being high enough. This is a necessary technical assumption for our analysis, and has been used by the previous analysis of Algorithm 2 as well, e.g., [21].

Given this assumption in the analytic results, it is important to understand the robustness of the results to this assumption. To that end, here we provide a case study to demonstrate that this intuition is robust to the t-valley-filling assumption.

In our case study, we mimic the setting of [21], where an average-case analysis of Algorithm 2 is performed. In particular, we use 24 hour residential load trace in the Southern California Edison (SCE) service area averaged over the year 2012 and 2013 [2] as the non-deferrable load, and wind power generation data from the Alberta Electric System Operator from 2004 to 2012 [1]. The wind power generation data is scaled so that its average over 9 years corresponds to 30% penetration level, and pick the wind generation of a random day as renewable during each run. We generate random prediction error in baseload and arrival of deferrable load similar to [21].

Given this setting, we simulate 100 instances in each scenario and compare the results with the Theorems 3. The results are shown in Fig. 5.4 where we plot the cumulative



(a) 30% prediction error

(b) 10% prediction error

Figure 5.4: The empirical cumulative distribution function of the load variance under Algorithm 2 over 24 hour control horizon using real data. The red line represents the analytic bound on the 90% confidence interval computed from Theorem 3, and the black line shows the empirical mean.

distribution (CDF) of the load variance produced by Algorithm 2 under two different scenarios. Specifically, in Fig. 5.4a, we assume the prediction error in wind power generation is 30%, and in Fig. 5.4b, we assume the prediction error is 10%. We plot the CDF on the same scale in both plots and additionally show an analytic bound on the 90% confidence interval computed from Theorem 3. For both cases, the results highlight a strong concentration around the mean, and the analytic bound from Theorem 3 is valid despite the fact that the t-valley-filling assumption is not satisfied. Further, note that the analytic bound is much tighter when prediction error is small, which coincides the statement of Theorem 3.

Chapter 6

Concluding Remarks

We have proposed a model predictive algorithm for decentralized deferrable load control that can schedule a large number of deferrable loads to compensate for the random fluctuations in renewable generation. At any time, the algorithm incorporates updated predictions about deferrable loads and renewable generation to minimize the expected load variance to go. Further, by modeling the base load prediction updates as a Wiener filtering process, we have conducted performance analysis to our algorithm in average case analysis and distributional analysis. We derived an explicit expression for the aggregate load variance obtained by the average case performance of the algorithm, which quantitatively showed the improvement of model predictive control over static control. Interestingly, the sub-optimality of static control is $O(T/\ln T)$ times that of real-time control in two representative cases of base load prediction updates. Besides average case analysis, we have provided a distributional analysis of the algorithm and shown that the load variance is tightly concentrated around its mean. Thus, our results highlight that the typical performance one should expect to see under model predictive deferrable load control is not-too-different from the average-case analysis. Importantly, the proof technique we develop may be useful for the analysis of model predictive control in more general settings as well. The qualitative insights from the analytic results were validated using trace-based simulations, which confirm that the algorithm has significantly smaller sub-optimality than the optimal static control.

The main limitation in our analysis (which is also true for the prior stochastic analysis of model predictive deferrable load control) is the assumption that a t-valley-filling solution exists. Practically, one can expect this to be satisfied if the penetration of deferrable loads is high; however, relaxing the need for this technical assumption remains an interesting and important challenge. Interestingly, the numerical results we report here highlight that one should also expect a tight concentration in the case where a t-valley-filling solution does not exist.

There remain many open questions on deferrable load control. For example, is it possible to reduce the communication and computation requirements of the proposed algorithm by assuming achievability of t-valley-filling? How to extend the algorithm to a receding horizon implementation? Additionally, how to apply the technique used here to incorporate prediction evolution for other demand response settings.

Bibliography

[1] Alberta eelctric system operator. wind power and alberta internal load data. `http://www.aeso.ca/gridoperations/20544.html`, 2012.

[2] Southern california edison dynamic load profiles. `https://www.sce.com/wps/portal/home/regulatory/load-profiles`, 2013.

[3] S. Acha, T. C. Green, and N. Shah. Effects of optimised plug-in hybrid vehicle charging strategies on electric distribution network losses. In *IEEE PES Transmission and Distribution Conference and Exposition*, pages 1–6, 2010.

[4] D. J. Aigner and J. G. Hirschberg. Commercial/industrial customer response to time-of-use electricity prices: Some experimental results. *The RAND Journal of Economics*, 16(3):341–355, 1985.

[5] Alberta Electric System Operator. Wind power / ail data, 2009. `http://www.aeso.ca/gridoperations/20544.html`.

[6] A. Bemporad and M. Morari. Robust model predictive control: A survey. In *Robustness in identification and control*, pages 207–226. Springer, 1999.

[7] S. Boucheron, G. Lugosi, P. Massart, et al. On concentration of self-bounding functions. *Electronic Journal of Probability*, 14(64):1884–1899, 2009.

[8] J.-Y. L. Boudec and D.-C. Tomozei. Satisfiability of elastic demand in the smart grid. *arXiv preprint arXiv:1011.5606*, 2010.

[9] California Public Utilities Commission. Zero net energy action plan, 2008. http://www.cpuc.ca.gov/NR/rdonlyres/6C2310FE-AFE0-48E4-AF03-530A99D28FCE/0/ZNEActionPlanFINAL83110.pdf.

[10] M. Caramanis and J. Foster. Management of electric vehicle charging to mitigate renewable generation intermittency and distribution network congestion. In *IEEE CDC*, pages 4717–4722, 2009.

[11] L. Chen, N. Li, S. H. Low, and J. C. Doyle. Two market models for demand response in power networks. In *IEEE SmartGridComm*, pages 397–402, 2010.

[12] N. Chen, L. Gan, S. H. Low, and A. Wierman. Distributional Analysis for Model Predictive Deferrable Load Control. *ArXiv e-prints*, Mar. 2014.

[13] S. Chen and L. Tong. iems for large scale charging of electric vehicles: architecture and optimal online scheduling. In *IEEE SmartGridComm*, pages 629–634, 2012.

[14] A. Conejo, J. Morales, and L. Baringo. Real-time demand response model. *IEEE Transactions on Smart Grid*, 1(3):236–242, 2010.

[15] S. Deilami, A. Masoum, P. Moses, and M. Masoum. Real-time coordination of plug-in electric vehicle charging in smart grids to minimize power losses and improve voltage profile. *IEEE Transactions on Smart Grid*, 2(3):456–467, 2011.

[16] Department of Energy. The smart grid: an introduction, 2008. http://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/DOE_SG_Book_Single_Pages%281%29.pdf.

[17] Department of Energy. One million electric vehicles by 2015, 2011. http://www1.eere.energy.gov/vehiclesandfuels/pdfs/1_million_electric_vehicles_rpt.pdf.

[18] E. A. Feinberg and D. Genethliou. Load forecasting. In *Applied Mathematics for Restructured Electric Power Systems*, Power Electronics and Power Systems, pages 269–285. Springer US, 2005.

[19] L. Gan, U. Topcu, and S. H. Low. Optimal decentralized protocol for electric vehicle charging. In *IEEE CDC*, pages 5798–5804, 2011.

[20] L. Gan, U. Topcu, and S. H. Low. Stochastic distributed protocol for electric vehicle charging with discrete charging rate. In *IEEE PES General Meeting*, pages 1–8, 2012.

[21] L. Gan, A. Wierman, U. Topcu, N. Chen, and S. H. Low. Real-time deferrable load control: handling the uncertainties of renewable generation. In *Proceedings of the fourth international conference on Future energy systems*, pages 113–124. ACM, 2013.

[22] L. Gan, A. Wierman, U. Topcu, N. Chen, and S. H. Low. Real-time deferrable load control: handling the uncertainties of renewable generation, 2013. Technical report, available at `http://www.its.caltech.edu/~lgan/index.html`.

[23] G. Giebel, R. Brownsword, G. Kariniotakis, M. Denhard, and C. Draxl. *The State-Of-The-Art in Short-Term Prediction of Wind Power*. ANEMOS.plus, 2011.

[24] S. C. Graves, D. B. Kletter, and W. B. Hetzel. A dynamic model for requirements planning with application to supply chain optimization. *Manufacturing & Service Operation Management*, 1(1):50–61, 1998.

[25] S. C. Graves, H. C. Meal, S. Dasu, and Y. Qiu. Two-stage production planning in a dynamic environment, 1986. `http://web.mit.edu/sgraves/www/papers/GravesMealDasuQiu.pdf`.

[26] N. Hatziargyriou, H. Asano, R. Iravani, and C. Marnay. Microgrids. *IEEE Power and Energy Magazine*, 5(4):78–94, 2007.

[27] Y.-Y. Hsu and C.-C. Su. Dispatch of direct load control using dynamic programming. *IEEE Transactions on Power Systems*, 6(3):1056–1061, 1991.

[28] M. Ilic, J. Black, and J. Watz. Potential benefits of implementing load control. In *IEEE PES Winter Meeting*, volume 1, pages 177–182, 2002.

[29] M. Ledoux. Concentration of measure and logarithmic sobolev inequalities. In *Seminaire de probabilites XXXIII*, pages 120–216. Springer, 1999.

[30] J. a. Lee and Z. Yu. Worst-case formulations of model predictive control for systems with bounded parameters. *Automatica*, 33(5):763–781, 1997.

[31] N. Li, L. Chen, and S. H. Low. Optimal demand response based on utility maximization in power networks. In *IEEE PES General Meeting*, pages 1–8, 2011.

[32] Q. Li, T. Cui, R. Negi, F. Franchetti, and M. D. Ilic. On-line decentralized charging of plug-in electric vehicles in power systems. *arXiv:1106.5063*, 2011.

[33] M. Lin, Z. Liu, A. Wierman, and L. L. Andrew. Online algorithms for geographical load balancing. In *Green Computing Conference (IGCC), 2012 International*, pages 1–10. IEEE, 2012.

[34] Z. Ma, D. Callaway, and I. Hiskens. Decentralized charging control for large populations of plug-in electric vehicles. In *IEEE CDC*, pages 206–212, 2010.

[35] C. McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[36] K. Mets, T. Verschueren, W. Haerick, C. Develder, and F. De Turck. Optimizing smart energy control strategies for plug-in hybrid electric vehicle charging. In *IEEE/IFIP NOMS Wksps*, pages 293–299, 2010.

[37] National Institute of Standards and Technology. Nist framework and roadmap for smart grid interoperability standards, 2010. http://www.nist.gov/public_affairs/releases/upload/smartgrid_interoperability_final.pdf.

[38] Southern California Edison. 2012 static load profiles, 2012. http://www.sce.com/005_regul_info/eca/DOMSM12.DLP.

[39] A. Subramanian, M. Garcia, A. Dominguez-Garcia, D. Callaway, K. Poolla, and P. Varaiya. Real-time scheduling of deferrable electric loads. In *ACC*, pages 3643–3650, 2012.

[40] Wikipedia. Krasovskii-lasalle principle. http://en.wikipedia.org/wiki/Krasovskii-LaSalle_principle.

[41] Wikipedia. Wiener filter. http://en.wikipedia.org/wiki/Wiener_filter.

Appendix A

Proofs of average case results

In this section, we only include proofs of the main results due to space restrictions. The remainder of the proofs can be found in the extended version [22].

A.1 Proof of Theorem 1

For brevity and without loss of generality, we prove Theorem 1 for $t = 1$ only. Thus, we can abbreviate b_t and $N(t)$ by b and N respectively without introducing confusion.

For feasible p, q to ODLC-t and $p = (p_1, \ldots, p_N)$, define

$$L(p, q) = \sum_{\tau=1}^{T} \left(b(\tau) + \sum_{n=1}^{N} p_n(\tau) + q(\tau) \right)^2 .$$

Since the sum of the aggregate load $\sum_{\tau=1}^{T} d(\tau)$ is a constant, minimizing the ℓ_2 norm of the aggregate load is equivalent to minimizing its variance. Hence, if subject to the same constraints, the minimizer of L is also the solution to ODLC-t. According to the proof of Proposition 1 in [19], we have

$$L(p^{(k+1)}, q^{(k)}) \leq L(p^{(k)}, q^{(k)})$$

for $k \geq 0$, and the equality is attained if and only if $p^{(k+1)} = p^{(k)}$ and $p^{(k)}$ minimizes

$L(p, q^{(k)})$ over all feasible p, i.e., (the first order optimality condition)

$$\left\langle b + \sum_{n=1}^{N} p_n^{(k)} + q^{(k)},\ p'_n - p_n^{(k)} \right\rangle \geq 0$$

for $n = 1, \ldots, N$ and all feasible p'_n. According to Step (ii) of Algorithm 2, it is straightforward that

$$L(p^{(k+1)}, q^{(k+1)}) \leq L(p^{(k+1)}, q^{(k)})$$

for $k \geq 0$, and the equality is attained if and only if $q^{(k+1)} = q^{(k)}$ and $q^{(k)}$ minimizes $L(p^{(k+1)}, q)$ over all feasible q, i.e., (the first order optimality condition)

$$\left\langle b + \sum_{n=1}^{N} p_n^{(k+1)} + q^{(k)},\ q' - q^{(k)} \right\rangle \geq 0$$

for all feasible q'. It then follows that

$$L(p^{(k+1)}, q^{(k+1)}) \leq L(p^{(k)}, q^{(k)})$$

and the equality if attained if and only if $(p^{(k+1)}, q^{(k+1)}) = (p^{(k)}, q^{(k)})$, and

$$\begin{aligned} \left\langle b + \sum_{n=1}^{N} p_n^{(k)} + q^{(k)},\ p'_n - p_n^{(k)} \right\rangle &\geq 0, \\ \left\langle b + \sum_{n=1}^{N} p_n^{(k)} + q^{(k)},\ q' - q^{(k)} \right\rangle &\geq 0 \end{aligned}$$

for all feasible p and q, i.e., $(p^{(k)}, q^{(k)})$ minimizes $L(p, q)$. Then by Lasalle's Theorem [40], we have $d(p^{(k)}, \mathcal{O}(t)) \to 0$ as $k \to \infty$. ■

A.2 Proof of Lemma 1

When $b_t = b$ and $\mathbb{E}(a(t)) = \lambda$ for $t = 1, \ldots, T$, the model (4.1) for Algorithm 2 reduces to

$$d(t) = \frac{1}{T-t+1}\left(\sum_{\tau=t}^{T} b(\tau) + \lambda(T-t) + \sum_{n=1}^{N(t)} P_n(t)\right) \tag{A.1}$$

for $t = 1, \ldots, T$. Then

$$(T-t+1)d(t) = \sum_{\tau=t}^{T} b(\tau) + \lambda(T-t) + \sum_{n=1}^{N(t)} P_n(t)$$

$$(T-t+2)d(t-1) = \sum_{\tau=t-1}^{T} b(\tau) + \lambda(T-t+1) + \sum_{n=1}^{N(t-1)} P_n(t-1)$$

for $t = 2, \ldots, T$. Subtract the two equations and simplify using the fact that $b(t-1) + \sum_{n=1}^{N(t-1)} (P_n(t-1) - P_n(t)) = b(t-1) + \sum_{n=1}^{N(t-1)} p_n(t-1) = d(t-1)$ and the definition of $a(t)$ to obtain

$$d(t) - d(t-1) \quad = \quad \frac{1}{T-t+1}\,(a(t) - \lambda)$$

for $t = 2, \ldots, T$. Substituting $t = 1$ into (A.1), it can be verified that $d(1) = \lambda + \sum_{\tau=1}^{T} b(\tau)/T + (a(1) - \lambda)/T$, therefore

$$d(t) = \lambda + \frac{1}{T}\sum_{\tau=1}^{T} b(\tau) + \sum_{\tau=1}^{t} \frac{1}{T-\tau+1}\,(a(\tau) - \lambda)$$

for $t = 1, \ldots, T$. The average aggregate load is

$$u = \frac{1}{T}\sum_{t=1}^{T} d(t) = \lambda + \frac{1}{T}\left(\sum_{\tau=1}^{T} b(\tau) + \sum_{\tau=1}^{T} (a(\tau) - \lambda)\right).$$

Hence,

$$\begin{aligned}
& \mathbb{E}(d(t)-u)^2 \\
= \;& \mathbb{E}\left(\sum_{\tau=1}^{t}\frac{1}{T-\tau+1}\left(a(\tau)-\lambda\right)-\frac{1}{T}\sum_{\tau=1}^{T}(a(\tau)-\lambda)\right)^2 \\
= \;& \mathbb{E}\left(\sum_{\tau=1}^{t}\frac{\tau-1}{T(T-\tau+1)}\left(a(\tau)-\lambda\right)-\frac{1}{T}\sum_{\tau=t+1}^{T}\left(a(\tau)-\lambda\right)\right)^2 \\
= \;& \frac{s^2}{T^2}\left(\sum_{\tau=1}^{t}\frac{(\tau-1)^2}{(T-\tau+1)^2}+T-t\right)
\end{aligned}$$

for $t = 1, \ldots, T$. The last equality holds because $(a(\tau) - \lambda)$ are independent for all τ and each of them have mean zero and variance s^2. It follows that

$$\begin{aligned}
\mathbb{E}(V) &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(d(t)-u)^2 \\
&= \frac{s^2}{T^3}\left(\sum_{t=1}^{T}\sum_{\tau=1}^{t}\frac{(\tau-1)^2}{(T-\tau+1)^2}+\sum_{t=1}^{T}(T-t)\right) \\
&= \frac{s^2}{T^3}\left(\sum_{\tau=1}^{T}\frac{(\tau-1)^2}{T-\tau+1}+\sum_{t=1}^{T}(T-t)\right) \\
&= \frac{s^2}{T^3}\left(\sum_{t=1}^{T}\frac{(T-t)^2}{t}+\sum_{t=1}^{T}\frac{(T-t)t}{t}\right) \\
&= s^2\frac{\sum_{t=2}^{T}\frac{1}{t}}{T} \sim s^2\frac{\ln T}{T}. \qquad \blacksquare
\end{aligned}$$

A.3 Proof of Lemma 2

In the case where no deferrable arrival after $t = 1$, i.e., $N(t) = N$ for $t = 1, \ldots, T$, the model (4.1) for Algorithm 2 reduces to

$$(T-t+1)d(t) = \sum_{\tau=t}^{T} b_t(\tau) + \sum_{n=1}^{N} P_n(t) \tag{A.2}$$

for $t = 1, \ldots, T$. Substitute t by $t-1$ to obtain

$$(T-t+2)d(t-1) = \sum_{\tau=t-1}^{T} b_{t-1}(\tau) + \sum_{n=1}^{N} P_n(t-1)$$

for $t = 2, \ldots, T$. Subtract the two equations to obtain

$$\begin{aligned} & (T-t+1)d(t) - (T-t+2)d(t-1) \\ = & \sum_{\tau=t}^{T} e(t)f(\tau-t) - b(t-1) - \sum_{n=1}^{N} p_n(t-1) \\ = & e(t)F(T-t) - d(t-1), \end{aligned}$$

which implies

$$d(t) - d(t-1) = \frac{1}{T-t+1} e(t)F(T-t)$$

for $t = 2, \ldots, T$. Substituting $t = 1$ into (A.2) and recalling the definition of b_t in (2.1), it can be verified that

$$d(1) = \frac{1}{T}\left(\sum_{n=1}^{N} P_n + \sum_{\tau=1}^{T} \bar{b}(\tau)\right) + \frac{1}{T} e(1)F(T-1).$$

Therefore,

$$d(t) = \frac{1}{T}\left(\sum_{n=1}^{N} P_n + \sum_{\tau=1}^{T} \bar{b}(\tau)\right) + \sum_{\tau=1}^{t} \frac{1}{T-\tau+1} e(\tau)F(T-\tau)$$

for $t = 1, \ldots, T$. The average aggregate load is

$$u = \frac{1}{T}\left(\sum_{n=1}^{N} P_n + \sum_{t=1}^{T} \bar{b}(t)\right) + \frac{1}{T}\sum_{\tau=1}^{T} e(\tau)F(T-\tau).$$

Hence,

$$\begin{aligned}
& \mathbb{E}(d(t)-u)^2 \\
= \; & \mathbb{E}\left(\sum_{\tau=1}^{t}\frac{1}{T-\tau+1}e(\tau)F(T-\tau)-\sum_{\tau=1}^{T}\frac{1}{T}e(\tau)F(T-\tau)\right)^2 \\
= \; & \mathbb{E}\left(\sum_{\tau=1}^{t}\frac{\tau-1}{T(T-\tau+1)}e(\tau)F(T-\tau)\right. \\
& \qquad \left.-\sum_{\tau=t+1}^{T}\frac{1}{T}e(\tau)F(T-\tau)\right)^2 \\
= \; & \frac{\sigma^2}{T^2}\left(\sum_{\tau=1}^{t}\frac{(\tau-1)^2}{(T-\tau+1)^2}F^2(T-\tau)+\sum_{\tau=t+1}^{T}F^2(T-\tau)\right)
\end{aligned}$$

for $t=1,\ldots,T$. The last equality holds because $e(\tau)$ are uncorrelated random variables with mean zero and variance σ^2. It follows that

$$\begin{aligned}
& \mathbb{E}(V)=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(d(t)-u)^2 \\
= \; & \frac{\sigma^2}{T^3}\sum_{t=1}^{T}\left(\sum_{\tau=1}^{t}\frac{(\tau-1)^2}{(T-\tau+1)^2}F^2(T-\tau)+\sum_{\tau=t+1}^{T}F^2(T-\tau)\right) \\
= \; & \frac{\sigma^2}{T^3}\sum_{\tau=1}^{T}F^2(T-\tau)\frac{(\tau-1)^2}{T-\tau+1}+\frac{\sigma^2}{T^3}\sum_{\tau=2}^{T}(\tau-1)F^2(T-\tau) \\
= \; & \frac{\sigma^2}{T^2}\sum_{\tau=1}^{T}F^2(T-\tau)\frac{\tau-1}{T-\tau+1}=\frac{\sigma^2}{T^2}\sum_{t=0}^{T-1}F^2(t)\frac{T-t-1}{t+1}.\blacksquare
\end{aligned}$$

A.4 Proof of Theorem 2

Similar to the proof of Lemma 1 and 2, use the model (4.1) to obtain

$$d(t)=\lambda+\frac{1}{T}\sum_{\tau=1}^{T}\bar{b}(\tau)+\sum_{\tau=1}^{t}\frac{1}{T-\tau+1}\left(e(\tau)F(T-\tau)+a(\tau)-\lambda\right)$$

for $t = 1, \ldots, T$ and

$$u = \lambda + \frac{1}{T}\sum_{\tau=1}^{T} \bar{b}(\tau) + \sum_{\tau=1}^{T} \frac{1}{T}\left(e(\tau)F(T-\tau) + a(\tau) - \lambda\right).$$

Hence,

$$\begin{aligned} & \mathbb{E}(d(t)-u)^2 \\ = \; & \mathbb{E}\left(\sum_{\tau=1}^{t} \frac{1}{T-\tau+1} e(\tau)F(T-\tau) - \sum_{\tau=1}^{T} \frac{1}{T} e(\tau)F(T-\tau)\right)^2 \\ & + \mathbb{E}\left(\sum_{\tau=1}^{t} \frac{1}{T-\tau+1}\left(a(\tau)-\lambda\right) - \sum_{\tau=1}^{T} \frac{1}{T}\left(a(\tau)-\lambda\right)\right)^2. \end{aligned}$$

The first term is exactly that in Lemma 2, and the second term is exactly that in Lemma 1. Hence, the expected load variance is

$$\mathbb{E}(V) = \frac{\sigma^2}{T^2}\sum_{t=0}^{T-1} F^2(t)\frac{T-t-1}{t+1} + \frac{s^2}{T}\sum_{t=2}^{T}\frac{1}{t}. \qquad \blacksquare$$

A.5 Proof of Corollary 3

If $|f(t)| \sim O(t^{-1/2-\alpha})$ for some $\alpha > 0$, then $|f(t)| \leq Ct^{-1/2-\alpha}$ for some $C > 0$ and all $t \geq 1$. Without loss of generality, assume that $0 < \alpha < 1/2$ and $C \geq (1-2\alpha)/(1+2\alpha)$. Then $F(0) = 1$ and

$$|F(t)| = \left|\sum_{\tau=0}^{t} f(\tau)\right| \leq 1 + \sum_{\tau=1}^{t} C\tau^{-1/2-\alpha} \leq \frac{2C}{1-2\alpha}t^{1/2-\alpha}$$

for $t = 1, \ldots, T$. The last inequality holds because $C \geq (1-2\alpha)/(1+2\alpha)$. Therefore it follows from Lemma 2 that

$$\begin{aligned}
\mathbb{E}(V) &\leq \frac{\sigma^2}{T}\sum_{s=0}^{T-1} F^2(s)\frac{1}{s+1} \\
&\leq \frac{\sigma^2}{T} + \frac{\sigma^2}{T}\sum_{s=1}^{T-1}\frac{4C^2}{(1-2\alpha)^2}s^{1-2\alpha}\frac{1}{s+1} \\
&\leq \frac{\sigma^2}{T} + \frac{\sigma^2}{T}\frac{4C^2}{(1-2\alpha)^2}\sum_{s=1}^{T-1}\frac{1}{s^{2\alpha}} \\
&\leq \frac{\sigma^2}{T} + \frac{4\sigma^2C^2}{(1-2\alpha)^2T} + \frac{4\sigma^2C^2}{(1-2\alpha)^3T^{2\alpha}}.
\end{aligned}$$

Hence, $\mathbb{E}(V) \to 0$ as $T \to \infty$. ■

A.6 Proof of Lemma 3

The aggregate load d obtained by the optimal static algorithm is

$$\begin{aligned}
d(t) &= \frac{1}{T}\left(\sum_{n=1}^{N} P_n + \sum_{\tau=1}^{T}\bar{b}(\tau)\right) - \bar{b}(t) + b(t) \\
&= \frac{1}{T}\left(\sum_{n=1}^{N} P_n + \sum_{\tau=1}^{T}\bar{b}(\tau)\right) + \sum_{\tau=1}^{T} e(\tau)f(t-\tau)
\end{aligned}$$

for $t = 1, \ldots, T$. Hence,

$$\begin{aligned}
&\mathbb{E}(d(t)-u)^2 \\
=\ &\mathbb{E}\left(\sum_{\tau=1}^{T} e(\tau)\left(f(t-\tau) - \frac{1}{T}F(T-\tau)\right)\right)^2 \\
=\ &\frac{\sigma^2}{T^2}\sum_{\tau=1}^{T} T^2f^2(t-\tau) - 2Tf(t-\tau)F(T-\tau) + F^2(T-\tau)
\end{aligned}$$

for $t = 1, \ldots, T$. It follows that

$$\begin{aligned}
& \mathbb{E}(V') = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(d(t)-u)^2 \\
= \quad & \frac{\sigma^2}{T}\sum_{t=1}^{T}\sum_{\tau=1}^{T} f^2(t-\tau) - \frac{2\sigma^2}{T^2}\sum_{\tau=1}^{T} F(T-\tau)\sum_{t=1}^{T} f(t-\tau) \\
& + \frac{\sigma^2}{T^2}\sum_{\tau=1}^{T} F^2(T-\tau) \\
= \quad & \frac{\sigma^2}{T}\sum_{t=1}^{T}\sum_{\tau=0}^{t-1} f^2(\tau) - \frac{\sigma^2}{T^2}\sum_{\tau=1}^{T} F^2(T-\tau) \\
= \quad & \frac{\sigma^2}{T}\sum_{\tau=0}^{T-1}(T-\tau) f^2(\tau) - \frac{\sigma^2}{T^2}\sum_{\tau=0}^{T-1} F^2(\tau) \\
= \quad & \frac{\sigma^2}{T^2}\sum_{t=0}^{T-1}\left(T(T-t)f^2(t) - F^2(t)\right). \qquad \blacksquare
\end{aligned}$$

Appendix B

Proofs of distributional results

B.1 Proof of Proposition 3

It has been computed in [21] that the load variance V obtained by Algorithm 2 is composed of two parts:

$$V = V_1 + V_2$$

where

$$V_1 := \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{\tau=1}^{t}\frac{\tau-1}{T(T-\tau+1)}(a(\tau)-\lambda) - \sum_{\tau=t+1}^{T}\frac{1}{T}(a(\tau)-\lambda)\right]^2$$

is the variance due to the prediction error on deferrable load and

$$V_2 := \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{\tau=1}^{t}\frac{\tau-1}{T(T-\tau+1)}e(\tau)F(T-\tau) - \sum_{\tau=t+1}^{T}\frac{1}{T}e(\tau)F(T-\tau)\right]^2$$

is the variance due to the prediction error on baseload. Now we compute the worst-case V_1 and V_2 under the bounded prediction error assumption (4.3).

We start with computing the worst-case V_1. Let $x(\tau) := a(\tau) - \lambda$ for $\tau = 1, 2, \ldots, T$, then

$$\begin{aligned}
V_1 &= \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{\tau=1}^{t}\frac{\tau-1}{T(T-\tau+1)}x(\tau) - \sum_{\tau=t+1}^{T}\frac{1}{T}x(\tau)\right]^2 \\
&= \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{\tau=1}^{t}\frac{1}{T-\tau+1}x(\tau) - \sum_{\tau=1}^{T}\frac{1}{T}x(\tau)\right]^2 \\
&= \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{\tau=1}^{t}\frac{1}{T-\tau+1}x(\tau)\right]^2 + \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{\tau=1}^{T}\frac{1}{T}x(\tau)\right]^2 \\
&\qquad - \frac{2}{T}\sum_{t=1}^{T}\sum_{\tau=1}^{t}\frac{1}{T-\tau+1}x(\tau)\sum_{s=1}^{T}\frac{1}{T}x(s) \\
&= \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{\tau=1}^{t}\frac{1}{T-\tau+1}x(\tau)\right]^2 + \left[\sum_{\tau=1}^{T}\frac{1}{T}x(\tau)\right]^2 \\
&\qquad - \frac{2}{T^2}\sum_{s=1}^{T}x(s)\sum_{\tau=1}^{T}\sum_{t=\tau}^{T}\frac{1}{T-\tau+1}x(\tau) \\
&= \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{\tau=1}^{t}\frac{1}{T-\tau+1}x(\tau)\right]^2 + \frac{1}{T^2}\left[\sum_{\tau=1}^{T}x(\tau)\right]^2 \\
&\qquad - \frac{2}{T^2}\sum_{s=1}^{T}x(s)\sum_{\tau=1}^{T}x(\tau) \\
&= \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{\tau=1}^{t}\frac{1}{T-\tau+1}x(\tau)\right]^2 - \frac{1}{T^2}\left[\sum_{\tau=1}^{T}x(\tau)\right]^2.
\end{aligned}$$

The first term

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\left[\sum_{\tau=1}^{t}\frac{1}{T-\tau+1}x(\tau)\right]^2 \\
&= \frac{1}{T}\sum_{t=1}^{T}\sum_{\tau=1}^{t}\left[\frac{1}{T-\tau+1}x(\tau)\right]^2 \\
&\quad + \frac{2}{T}\sum_{t=1}^{T}\sum_{\tau=1}^{t}\frac{1}{T-\tau+1}x(\tau)\sum_{s=\tau+1}^{t}\frac{1}{T-s+1}x(s) \\
&= \frac{1}{T}\sum_{\tau=1}^{T}\sum_{t=\tau}^{T}\frac{1}{(T-\tau+1)^2}x^2(\tau) \\
&\quad + \frac{2}{T}\sum_{\tau=1}^{T}\sum_{s=\tau+1}^{T}\sum_{t=s}^{T}\frac{1}{T-\tau+1}\frac{1}{T-s+1}x(\tau)x(s) \\
&= \frac{1}{T}\sum_{\tau=1}^{T}\frac{1}{T-\tau+1}x^2(\tau) \\
&\quad + \frac{2}{T}\sum_{\tau=1}^{T}\sum_{s=\tau+1}^{T}\frac{1}{T-\tau+1}x(\tau)x(s) \\
&= \frac{1}{T}\sum_{\tau=1}^{T}\sum_{s=1}^{T}\frac{1}{T-\tau\wedge s+1}x(\tau)x(s)
\end{aligned}$$

where $a \wedge b := \min\{a, b\}$ for $a, b \in \mathbb{R}$. Let the matrix $A \in \mathbb{R}^{T\times T}$ be given by

$$A_{\tau s} := \frac{T}{T-\tau\wedge s+1}$$

for $\tau, s = 1, 2, \ldots, T$, i.e.,

$$A = \begin{bmatrix}
\frac{T}{T} & \frac{T}{T} & \frac{T}{T} & \cdots & \frac{T}{T} \\
\frac{T}{T} & \frac{T}{T-1} & \frac{T}{T-1} & \cdots & \frac{T}{T-1} \\
\frac{T}{T} & \frac{T}{T-1} & \frac{T}{T-2} & \cdots & \frac{T}{T-2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{T}{T} & \frac{T}{T-1} & \frac{T}{T-2} & \cdots & \frac{T}{1},
\end{bmatrix}$$

then

$$V_1 = \frac{1}{T^2} x^T \left(A - \mathbf{1}\mathbf{1}^T \right) x$$

where the vector $x := (x(1), x(2), \ldots, x(T))^T$. When prediction error is bounded as in (4.3), one has $|x(t)| \leq \epsilon_1$ for all t, and therefore

$$\begin{aligned} V_1 &= \frac{1}{T^2} \sum_{\tau=1}^{T} \sum_{s=1}^{T} \left(A_{\tau s} - 1 \right) x(\tau) x(s) \\ &\leq \frac{1}{T^2} \sum_{\tau=1}^{T} \sum_{s=1}^{T} \frac{\tau \wedge s - 1}{T - \tau \wedge s + 1} \epsilon_1^2 \end{aligned}$$

and the equality is attained if and only if $x(t) = \epsilon_1$ for all t, or $x(t) = -\epsilon_1$ for all t. Finally, we simplify the worst-case expression of V_1 as follows:

$$\begin{aligned} \sup_a V_1 &= \frac{1}{T^2} \sum_{\tau=1}^{T} \sum_{s=1}^{T} \frac{\tau \wedge s - 1}{T - \tau \wedge s + 1} \epsilon_1^2 \\ &= \frac{\epsilon_1^2}{T^2} \sum_{k=1}^{T} \frac{k-1}{T-k+1} (2T + 1 - 2k) \\ &= \epsilon_1^2 \left(1 - \frac{1}{T} \sum_{k=1}^{T} \frac{1}{k} \right) \approx \epsilon_1^2 \left(1 - \frac{\ln T}{T} \right). \end{aligned}$$

We proceed to compute the worst-case V_2. Using the same derivation, it can be computed that

$$V_2 = \frac{1}{T^2} y^T \left(A - \mathbf{1}\mathbf{1}^T \right) y$$

where

$$\begin{aligned} y &:= (y(1), y(2), \ldots, y(T))^T, \\ y(t) &:= e(t) F(T-t), \quad t = 1, 2, \ldots, T. \end{aligned}$$

It follows that

$$\begin{aligned} V_2 &= \frac{1}{T^2}\sum_{\tau=1}^{T}\sum_{s=1}^{T}(A_{\tau s}-1)\,y(\tau)y(s) \\ &\le \frac{1}{T^2}\sum_{\tau=1}^{T}\sum_{s=1}^{T}\frac{\tau\wedge s-1}{T-\tau\wedge s+1}\epsilon_2^2|F(T-\tau)F(T-s)| \end{aligned}$$

and that the equality is attained if and only if $e(t) = \epsilon_2 \cdot \mathrm{sgn}(F(T-t))$ for all t, or $e(t) = -\epsilon_2 \cdot \mathrm{sgn}(F(T-t))$ for all t. Finally, we simplify the worst-case expression of V_2 as follows:

$$\begin{aligned} \sup_e V_2 &= \frac{1}{T^2}\sum_{\tau=1}^{T}\sum_{s=1}^{T}\frac{\tau\wedge s-1}{T-\tau\wedge s+1}\epsilon_2^2|F(T-\tau)F(T-s)| \\ &= \frac{\epsilon_2^2}{T^2}\sum_{\tau=0}^{T-1}\sum_{s=0}^{T-1}\left(\frac{T}{\tau\vee s+1}-1\right)|F(\tau)F(s)| \end{aligned}$$

To summarize, the worst-case load variance V obtained by Algorithm 2 is

$$\begin{aligned} \sup_{a,e} V &= \epsilon_1^2\left(1-\frac{1}{T}\sum_{k=1}^{T}\frac{1}{k}\right) \\ &\quad + \frac{\epsilon_2^2}{T^2}\sum_{\tau=0}^{T-1}\sum_{s=0}^{T-1}\left(\frac{T}{\tau\vee s+1}-1\right)|F(\tau)F(s)|. \end{aligned}$$

The lower bound in the lemma can be obtained from the case where all prediction errors

of the load arrival is equal to $d_1/2$, then

$$\begin{aligned}
\sup_a V &\geq \frac{d_1^2}{4T}\sum_{t=1}^{T}\left(\sum_{\tau=1}^{t}\frac{\tau-1}{T(T-\tau+1)}-\sum_{\tau=t+1}^{T}\frac{1}{T}\right)^2 \\
&= \frac{d_1^2}{4T^3}\sum_{t=1}^{T}\left(\sum_{\tau=1}^{t}\frac{T}{T-\tau+1}-T\right)^2 \\
&= \frac{d_1^2}{4T}\sum_{t=1}^{T}\left(\sum_{\tau=1}^{t}\frac{1}{T-\tau+1}-1\right)^2 \\
&= \frac{d_1^2}{4T}\left(\sum_{t=1}^{T}(\sum_{\tau=T-t+1}^{T}\frac{1}{\tau})^2-T\right) \\
&\geq \frac{d_1^2}{4T}\left(\sum_{t=1}^{T}(\int_{T-t+1}^{T}\frac{1}{u}du)^2-T\right) \\
&= \frac{d_1^2}{4T}\left(\sum_{k=1}^{T}(\ln(\frac{T}{k}))^2-T\right)
\end{aligned}$$

B.2 Proof of Theorem 3

The theorem relies on a variant of the Log-Sobolev inequality provided in the following lemma.

Lemma 4 (Theorem 3.2, [29]). *Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be convex and X be supported on $[-d/2, d/2]^n$, then*

$$\begin{aligned}
&\mathbb{E}[\exp(f(X))f(X)] - \mathbb{E}[\exp(f(X))]\log\mathbb{E}[\exp(f(X))] \\
\leq\ &\frac{d^2}{2}\mathbb{E}[\exp(f(X))||\nabla f(X)||^2]. \qquad \text{(B.1)}
\end{aligned}$$

If f is further "self-bounded", then its tail probability can be bounded as in the following lemma.

Lemma 5. *Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be convex and X be supported on $[-d/2, d/2]^n$. If $\mathbb{E}[f(X)] = 0$ and f satisfies the following self-bounding property*

$$||\nabla f||^2 \leq af + b, \tag{B.2}$$

then the tail probability of $f(X)$ can be bound as

$$\mathbb{P}\left\{f(X) > t\right\} \leq \exp\left(\frac{-t^2}{2b + at}\right). \tag{B.3}$$

Proof. Denote the moment generating function of $f(X)$ by

$$m(\theta) := \mathbb{E}e^{\theta f(X)}, \qquad \theta > 0.$$

The function $\theta f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex, and therefore it follows from Lemma 4 that

$$\begin{aligned}
&\mathbb{E}\left[e^{\theta f}\theta f\right] - \mathbb{E}\left[e^{\theta f}\right]\ln\mathbb{E}\left[e^{\theta f}\right] \leq \frac{d^2}{2}\mathbb{E}\left[e^{\theta f}||\theta\nabla f||^2\right], \\
&\theta m'(\theta) - m(\theta)\ln m(\theta) \leq \frac{1}{2}\theta^2 d^2 \mathbb{E}[e^{\theta f}||\nabla f||^2].
\end{aligned}$$

According to the self-bounding property (B.2), one has

$$\begin{aligned}
\theta m'(\theta) - m(\theta)\ln m(\theta) &\leq \frac{1}{2}\theta^2 d^2 \mathbb{E}[e^{\theta f}(af + b)] \\
&= \frac{1}{2}\theta^2 d^2 \left[am'(\theta) + bm(\theta)\right].
\end{aligned}$$

Divide both sides by $\theta^2 m(\theta)$ to get

$$\frac{d}{d\theta}\left[\left(\frac{1}{\theta} - \frac{ad^2}{2}\right)\ln m(\theta)\right] \leq \frac{bd^2}{2}.$$

Integrate both sides from 0 to s to get

$$\left(\frac{1}{\theta} - \frac{ad^2}{2}\right) \ln m(\theta) \Big|_{\theta=0}^{s} \leq \frac{1}{2} bd^2 s$$

for $s \geq 0$. Noting that $m(0) = 1$ and $m'(0) = \mathbb{E}f = 0$, one has

$$\lim_{\theta \to 0^+} \left(\frac{1}{\theta} - \frac{ad^2}{2}\right) \ln m(\theta) \;=\; 0,$$

and therefore

$$\left(\frac{1}{s} - \frac{ad^2}{2}\right) \ln m(s) \leq \frac{1}{2} bd^2 s \tag{B.4}$$

for $s \geq 0$. We can bound the tail probability $\mathbb{P}\{f > t\}$ with the control (B.4) over the moment generating function $m(s)$.

In particular, one has

$$\begin{aligned}
\mathbb{P}\{f > t\} \;&=\; \mathbb{P}\left\{e^{sf} > e^{st}\right\} \;\leq\; e^{-st}\mathbb{E}\left[e^{sf}\right] \\
&=\; \exp[-st + \ln m(s)] \\
&\leq\; \exp\left[-st + \frac{bd^2 s^2}{2 - asd^2}\right]
\end{aligned}$$

for $s \geq 0$. Choose $s = t/(bd^2 + ad^2 t/2)$ to get

$$\mathbb{P}\{f > t\} \leq \exp\left(\frac{-t^2}{d^2(2b + at)}\right).$$

□

Proof of Theorem 3. It has been computed in [21] that the load variance V obtained by Algorithm 2 is composed of two parts:

$$V = V_1 + V_2$$

where

$$V_1 := \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{\tau=1}^{t}\frac{\tau-1}{T(T-\tau+1)}(a(\tau)-\lambda) - \sum_{\tau=t+1}^{T}\frac{1}{T}(a(\tau)-\lambda)\right]^2$$

is the variance due to the prediction error on deferrable load and

$$V_2 := \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{\tau=1}^{t}\frac{\tau-1}{T(T-\tau+1)}e(\tau)F(T-\tau) - \sum_{\tau=t+1}^{T}\frac{1}{T}e(\tau)F(T-\tau)\right]^2$$

is the variance due to the prediction error on baseload.

Let $x(\tau) := a(\tau) - \lambda$ for $\tau = 1, 2, \ldots, T$, then

$$\begin{aligned} V_1 &= \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{\tau=1}^{t}\frac{\tau-1}{T(T-\tau+1)}x(\tau) - \sum_{\tau=t+1}^{T}\frac{1}{T}x(\tau)\right]^2 \\ &= \frac{1}{T}||Bx||_2^2 \end{aligned}$$

where the $T \times T$ matrix B is given by

$$B_{t\tau} := \begin{cases} \frac{\tau-1}{T(T-\tau+1)} & \tau \le t \\ -\frac{1}{T} & \tau > t \end{cases}, \qquad 1 \le t, \tau \le T.$$

Similarly, the variance V_2 due to the prediction error on baseload can be written as

$$V_2 = g(e) = \frac{1}{T}||Ce||_2^2$$

where the $T \times T$ matrix C is given by

$$C_{t\tau} := \begin{cases} \frac{\tau-1}{T(T-\tau+1)} F(T-\tau), & \tau \leq t \\ -\frac{1}{T} F(T-\tau), & \tau > t \end{cases}$$

for $1 \leq t, \tau \leq T$. Therefore, the load variance

$$V = V_1 + V_2 = \frac{1}{T} \|Ay\|_2^2$$

where

$$A = \begin{bmatrix} B & 0 \\ 0 & C \end{bmatrix}, \quad y = \begin{bmatrix} x \\ e \end{bmatrix}.$$

Define a centered random variable

$$Z := h(y) := V - \mathbb{E}V = \frac{1}{T} ||Ay||^2 - \mathbb{E}V$$

and note that the function h is convex. Let $\lambda_{\max}$ be the maximum eigenvalue of AA^T/T, then

$$\begin{aligned} ||\nabla h(y)||^2 &= \frac{4}{T^2} ||A^T Ay||^2 = \frac{4}{T} (Ay)^T \left(\frac{AA^T}{T} \right) (Ay) \\ &\leq \frac{4\lambda_{\max}}{T} (Ay)^T (Ay) = 4\lambda_{\max} [h(y) + \mathbb{E}V]. \end{aligned}$$

According to the bounded prediction error assumption (4.3), one has $|y| \leq \epsilon$ componentwise. Then, apply Lemma 5 to the random variable Z to obtain

$$\mathbb{P}\{Z > t\} \leq \exp\left(-\frac{t^2}{16\lambda_{\max}\epsilon^2(2\mathbb{E}V + t)} \right)$$

for $t > 0$, i.e.,

$$\mathbb{P}\{V - \mathbb{E}V > t\} \leq \exp\left(-\frac{t^2}{16\lambda_{\max}\epsilon^2(2\mathbb{E}V + t)} \right)$$

for $t > 0$. Finally, the largest eigenvalue $\lambda_{\max}$ of AA^T/T can be bounded above as

$$\begin{aligned}\lambda_{\max} &\leq \mathrm{tr}\left(\frac{AA^T}{T}\right) = \mathrm{tr}\left(\frac{BB^T}{T}\right) + \mathrm{tr}\left(\frac{CC^T}{T}\right) \\ &= \frac{1}{T}\left(\sum_{t=2}^{T}\frac{1}{t}\right) + \frac{1}{T^2}\sum_{t=0}^{T-1}F^2(t)\frac{T-t-1}{t+1} \\ &\leq \frac{\ln T}{T} + \frac{1}{T^2}\sum_{t=0}^{T-1}F^2(t)\frac{T-t-1}{t+1} \;=:\; \lambda_1,\end{aligned}$$

which completes the proof of Theorem 3.

□

B.3 Proof of Theorem 4

The derivation of the theorem is based on the following two lemma, which separates the cases when there is only one type of prediction error.

Lemma 6. *If there is no prediction error in the base load, then the variance of the performance of Algorithm 2 is bounded by*

$$\mathrm{Var}(V) \leq 4d_1^2 s^2 \left(\frac{\ln T}{T}\right)^2. \tag{B.5}$$

Lemma 7. *If there is no prediction error in the deferrable load, then the variance of the performance of Algorithm 2 is bounded by*

$$\mathrm{Var}(V) \leq 4d_2^2\sigma^2 \left(\frac{1}{T^2}\sum_{t=0}^{T-1}F^2(t)\frac{T-t+1}{t+1}\right)^2. \tag{B.6}$$

Firstly we will prove Lemma 6, where we only consider prediction error in deferrable

load.

Proof of Lemma 6. Let $x(\tau) = a(\tau) - \lambda$, then $x(\tau)$ is centered, with variance s^2. Let $x = (x(1), \ldots, x(T))$. From the results in [21] Lemma 1, we have

$$V = \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{\tau=1}^{t}\frac{\tau-1}{T(T-\tau+1)}x(\tau) - \sum_{\tau=t+1}^{T}\frac{1}{T}x(\tau)\right)^2$$

Define an auxilary matrix B such that

$$B_{t\tau} = \begin{cases} \frac{\tau-1}{T(T-\tau+1)} & \tau \leq t \\ -\frac{1}{T} & \tau > t. \end{cases}$$

Then we have

$$V_1 = f(x(1), x(2), \ldots, x(T)) = \frac{1}{T}||Bx||_2^2.$$

Hence $V_1 = f(x)$ is a convex function, by convex Poincaré inequality, we have

$$\mathrm{Var}(V) \leq d_1^2 \mathbb{E}[||\nabla f(x)||^2]. \tag{B.7}$$

Whereas

$$\begin{aligned}
\mathbb{E}\left[||\nabla f(x)||^2\right] &= \frac{4}{T^2}\mathbb{E}\left[||B^TBx||^2\right] \\
&\leq \frac{4}{T^2}\lambda_{\max}(B^TB)\mathbb{E}\left[||Bx||^2\right] \\
&\leq 4\mathrm{tr}\left(\frac{1}{T}B^TB\right)\mathbb{E}\left[\frac{1}{T}||Bx||^2\right] \\
&= 4s^2\left[\mathrm{tr}\left(\frac{1}{T}B^TB\right)\right]^2 \\
&\leq 4s^2\left(\frac{\ln T}{T}\right)^2
\end{aligned}$$

The last inequality is because

$$\begin{aligned}
\mathrm{tr}\left(B^T B\right) &= \frac{1}{T}\sum_{i=1}^{T}(B^T B)_{ii} \\
&= \sum_{i=1}^{T}\sum_{k=1}^{T}(B_{ki})^2 \\
&= \frac{1}{T^2}\sum_{i=1}^{T}\left(\sum_{k=1}^{i}\frac{(k-1)^2}{(T-k+1)^2}+(T-i)\right) \\
&= \frac{1}{T^2}\sum_{k=1}^{T}(\frac{(k-1)^2}{(T-k+1)}+\sum_{i=1}^{T}(T-i) \\
&= \frac{1}{T^2}\sum_{k=1}^{T}\frac{(T-k)^2}{k}+\sum_{k=1}^{T}\frac{(T-k)k}{k} \\
&= \sum_{k=2}^{T}\frac{1}{k} \leq \ln T. \qquad \square
\end{aligned}$$

Next we proof lemma 7 the case where we only consider the prediction error in the base load.

Proof of Lemma 7. Let $e = (e(1), \ldots, e(T))$, when there is no prediction error in the deferrable load arrival, we have

$$\begin{aligned}
V =& \frac{1}{T}\sum_{t=1}^{T}(\sum_{\tau=1}^{t}\frac{\tau-1}{T(T-\tau+1)}F(T-\tau)e(\tau) \\
&- \sum_{\tau=t+1}^{T}\frac{1}{T}F(T-\tau)e(\tau))^2.
\end{aligned}$$

If we define an auxilary matrix C such that

$$C_{t\tau} = \begin{cases} \frac{\tau-1}{T(T-\tau+1)}F(T-\tau), & \tau \leq t \\ -\frac{1}{T}F(T-\tau), & \tau > t \end{cases}$$

Then we have

$$V = g(e(1), e(2), \ldots, e(T)) = \frac{1}{T}||Ce||_2^2.$$

Hence $V = g(e)$ is a convex function in e. By similar argument as Lemma 6

$$\mathrm{Var}(V) \leq d_2^2 \mathbb{E}[||\nabla g(e)||^2]. \tag{B.8}$$

Whereas

$$\begin{aligned}
\mathbb{E}\left[||\nabla g(e)||^2\right] &= \frac{4}{T^2}\mathbb{E}\left[||C^T Ce||^2\right] \\
&\leq \frac{4}{T^2}\lambda_{\max}(C^T C)\mathbb{E}\left[||Ce||^2\right] \\
&\leq 4\mathrm{tr}\left(\frac{1}{T}C^T C\right)\mathbb{E}\left[\frac{1}{T}||Ce||^2\right] \\
&= 4\sigma^2\left[\mathrm{tr}\left(\frac{1}{T}C^T C\right)\right]^2 \\
&= 4\sigma^2\left(\frac{1}{T^2}\sum_{t=0}^{T-1} F^2(t)\frac{T-t+1}{t+1}\right)^2.
\end{aligned}$$

The last equality is because

$$\begin{aligned}
&\mathrm{tr}(C^T C) \\
=&\sum_{i=1}^{T}\left(\sum_{k=1}^{T} C_{ki}^2\right) \\
=&\frac{1}{T^2}\sum_{i=1}^{T}\left(\sum_{k=1}^{i}\frac{(k-1)^2}{(T-k+1)^2}F^2(T-k) + \sum_{k=i+1}^{T}F^2(T-k)\right) \\
=&\frac{1}{T^2}\left(\sum_{k=2}^{T}\frac{(k-1)^2}{T-k+1}F^2(T-k) + \sum_{k=2}^{T}(k-1)F^2(T-k)\right) \\
=&\frac{1}{T}\sum_{k=2}^{T}F^2(T-k)\frac{k-1}{T-k+1}.
\end{aligned}$$

□

Next, we bring the two results together to get a proof of Theorem 4.

Proof of Theorem 2. Let V_1 be the load variance without prediction error in base load and V_2 be the load variance without prediction error in the deferrable load.

$$V = V_1 + V_2.$$

By independence of x and e, the variance of V is bounded by

$$\begin{aligned}
&\mathrm{Var}(V) = \mathrm{Var}(V_1) + \mathrm{Var}(V_2) \\
&\leq \left(\frac{2d_1 s \ln T}{T}\right)^2 + \left(\frac{2d_2\sigma}{T^2}\sum_{t=0}^{T-1} F^2(t)\frac{T-t+1}{t+1}\right)^2.
\end{aligned}$$

□