Distributional Analysis for Model Predictive Deferrable Load Control

Niangjun Chen, Lingwen Gan, Steven H. Low, Adam Wierman

***Abstract*— Deferrable load control is essential for handling the uncertainties associated with the increasing penetration of renewable generation. Model predictive control has emerged as an effective approach for deferrable load control, and has received considerable attention. In particular, previous work has analyzed the average-case performance of model predictive deferrable load control. However, to this point, distributional analysis of model predictive deferrable load control has been elusive. In this paper, 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.**

I. Introduction

The electricity grid is at the brink of change. On the generation side, the penetration of wind and solar in the energy portfolio is on the rise due to environmental concerns and, on the demand side, many smart appliances and devices that can adjust its power consumption level to minimize cost are entering the market. The combination of these two changes make generation less controllable and load less predictable, which makes the traditional "generation follows load" model of control much more difficult.

Fortunately, while smart devices make demand forecasting more challenging, they also provide an opportunity to mitigate the intermittency of wind and solar generation from the load side by allowing for demand response. There are two major categories of demand response, direct load control (DLC) and price-based demand response. See [1] for a discussion of the contrasts between these approaches.

In this paper we focus on *direct load control* with the goal of using demand response to *reduce variations of the aggregate load*. This objective has been studied frequently in the literature, e.g., [2], [3], because reducing the variations of the aggregate load corresponds to minimizing the generation cost of the utilities. In particular, large generators with the smallest marginal costs, e.g., nuclear generators and hydro generators, have limited ramp rates, i.e., their power output cannot be adjusted too quickly. So, if load varies frequently, then it must balanced by more expensive generators (i.e., "peakers") that have fast ramp rate. Thus, if the load variance is reduced, then the utility can use the least expensive sources of power generation to satisfy the electricity demand.

A. Model predictive deferrable load control

There are a variety of algorithmic challenges inherent in designing a direct load control algorithm for minimizing the load variance. Two of the most fundamental challenges are the following. First, the algorithm needs to scale well with the number of controllable devices, since the expectation is that the number of loads participating in such demand response programs will be large. Second, the algorithm needs to perform well in the face of uncertain predictions about future loads, renewable generation, etc.

There is a growing body of work on direct load control algorithms, which includes both simulation-based evaluations [4]–[6] and papers that focus on deriving theoretical performance guarantees [7], [8]. The most commonly proposed framework for algorithm design from this literature is, perhaps, model predictive control.

Model predictive control is a classical control algorithm, e.g., see [9] for a survey. In the context of direct load control, many variations have been proposed. At this point, there exist model predictive deferrable load control algorithms that are distributed with guaranteed convergence to optimal deferrable load schedules, e.g., [10].

However, to this point, the evaluation of model predictive deferrable load control has focused primarily on average-case analysis, e.g., [11], [12], or worst-case analysis, e.g., [13], [14]. While such analysis provides important insights, there is still much to learn about the performance of model predictive deferrable load control.

For example, though the average-case performance of model predictive deferrable load control is good, this alone says little about the "typical" performance of the algorithm. If the variability is high or the distribution is skewed, then bad events may still be quite likely. It is this sort of thinking that has motivated the study of worst-case analysis in this context; however worst-case analysis is often quite loose when compared with the typical performance. Instead, what is really needed is a distributional analysis, which can say, e.g., that the load variance will be less than the desired level 95 percent of the time. But, to this point, no results on the distribution of the load variation under model predictive deferrable load control exist.

B. Contributions of this paper

The main contribution of this paper is to provide a distributional analysis of the load variance under model predictive deferrable load control. More specifically, we prove sharp concentration results for the load variance arising from model predictive distributed load control. These results can, e.g., be used to bound the probability that the load variance exceeds a certain value and to understand what "typical" load variances will look like.

Our results are derived in the context of a standard formulation of the so-called "optimal deferrable load control" (OLDC) problem, where predictions of future loads are modeled using a Weiner filter [10]. The model is introduced in Section II. We also adopt the model predictive deferrable

load control mechanism in [10] since it has provable optimality guarantees: average-case analysis suggests that it performs well in environments with uncertain predictions. Details of the model predictive deferrable load control algorithm are introduced in Section III.

Given this context, the main result of the paper is Theorem 1, which proves a Bernstein-type concentration for the load variance under model predictive deferrable load control. This result highlights that the load variance is concentrated around its mean, and therefore the typical performance is as tightly concentrated around the average performance. Additionally, the result provides useful performance bounds on, e.g., the 95th percentile.

Note that it is useful to contrast the distributional analysis in Theorem 1 with worst-case bounds. Since worst-case bounds have not previously been derived for the exact setting we consider, we provide a new worst-case analysis in Proposition 5. In contrast with the insight derived from the distributional analysis, the worst-case analysis simply states that model predictive deferrable control can be as bad as having no control at all if predictions are adversarial.

Finally, in addition to the usefulness of Theorem 1 in the context of deferrable load control, it is important to note that the proof technique we develop may also be useful for understanding the distributional performance of model predictive control in other settings. In particular, to obtain tight bounds we use a Log-Sobolev inequality to bound the moment generating function of the load variance. However, a challenge is that the most commonly used approach – the martingale bounded difference approach, e.g., [15] – applies only if the function does not change much when one of the random variable is substituted by its identical copy. Unfortunately, this is not the case in our context, so we develop a new approach that, instead of bounding the target by step-wise changes, bounds its entropy using its gradient. This allows us to exploit more structure of the problem and obtain tighter bounds. Note that this approach should be applicable to the analysis of model predictive control in many other contexts too.

II. Model

In this paper we consider a standard model for deferrable load control introduced by [16] and then studied in, e.g., [6], [7], [17]. It is a discrete-time model where the time-slot length matches the timescale at which the power grid system operator makes control decisions.

The goal is to flatten the aggregate load over the control horizon $t \in \{1, ..., T\}$. In practice, the control horizon could be a day and a time slot could be on the order of minutes. To formalize the objective of flattening the aggregate load, previous work has tended to focus on minimizing the load variance

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

where $d = (d(1), d(2), \ldots, d(T))$ is the aggregate load profile at each time slot.

Importantly, the aggregate load consists of two types. The first type, which is termed *baseload*, includes loads like lighting and heating and is stochastic and non-controllable. Note that renewable generation like wind and solar can be considered as a negative stochastic and non-controllable load. Denote the baseload by $b = (b(1), b(2), \ldots, b(T))$, and note that b can be interpreted as the difference between non-deferrable load and renewable generation during each time period.

The second type of load, which is called *deferrable load*, consists of devices whose power consumption can be controlled by the utility, e.g., pool pumps, dryers, and electric vehicles taking part in direct load control programs [3], [18]. It is the control of these devices that can be used to minimize (1), provided that energy constraints and charging rate constraints are satisfied. To model deferrable load we consider N devices indexed $1, 2, \ldots, N$, and let $p_n(t)$ denote the power consumption of device n at time t for $n = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$. Further, each device has associated constraints on the power consumption as follows

$$\underline{p}_n(t) \le p_n(t) \le \bar{p}_n(t), \tag{2a}$$

$$\sum_{t=1}^{T} p_n(t) = P_n. \tag{2b}$$

Note that, using the above, arrival and deadline constraints can be specified by setting $\underline{p}_n(t) = \bar{p}_n(t) = 0$ for t before arrival and after deadline.

Given the previous notation, we can now formally specify the optimal deferrable load control (ODLC) problem that is the focus of this paper. Define $[k] := \{1, 2, \ldots, k\}$ for $k \in \mathbb{Z}^+$.

$$\begin{aligned}
\textbf{ODLC:}\ \min\ & \frac{1}{T}\sum_{t=1}^{T}\left(d(t) - \frac{1}{T}\sum_{\tau=1}^{T} d(\tau)\right)^2 \\
\text{over}\ & p_n(t), d(t), \quad \forall n, t \\
\text{s.t.}\ & d(t) = b(t) + \sum_{n=1}^{N} p_n(t), \quad t \in [T]; \\
& \underline{p}_n(t) \le p_n(t) \le \bar{p}_n(t), \quad n \in [N], t \in [T]; \\
& \sum_{t=1}^{T} p_n(t) = P_n, \quad n \in [N].
\end{aligned} \tag{3}$$

An important point about the formulation is that ODLC is a convex optimization problem, but cannot be solved in real time since the optimal decision at time t depends on future information about the baseload and future information about the arrivals of deferrable load. This information is not known exactly, but commonly there do exist predictions of future baseload and deferrable load arrivals. So, in practice such predictions are used for real time control.

Thus, the final component of the model is to specify a model for the predictions. Crucially, prediction errors should grow as prediction is made further into the future. Further, it is likely that errors are correlated, e.g., an underestimate for time slot $t+1$ likely leads to an underestimate for time slot $t+2$. To capture these issues, [10] has suggested a model based on Weiner filters, and we adopt the same assumptions here.

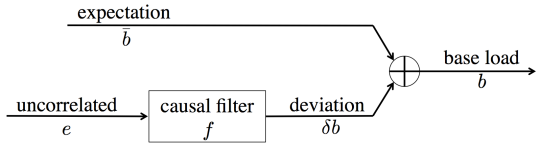


Fig. 1: *Diagram of the structure of the baseload model.*

Specifically, baseload b is modeled as a random deviation δb around its expectation $\bar{b}$ as illustrated in Fig. 1. The process δb is modeled as a sequence of independent random variables $e(1), \ldots, e(T)$, each with mean 0 and variance σ^2, passing through a causal filter with impulse response f ($f(\tau) = 0$ for $\tau < 0$), i.e.,

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

Using the current information, one can update the prediction at time t by

$$b_t(\tau) = \bar{b}(\tau) + \sum_{s=1}^{t} e(s) f(\tau - s), \qquad \tau = 1, \ldots, T. \quad (4)$$

Further, deferrable loads are modeled as random arrivals over time. Let $N(t)$ be the number of loads that arrive before (or at) time t for $t = 1, ..., T$. Define

$$a(t) := \sum_{n=N(t-1)+1}^{N(t)} P_n, \qquad t = 1, \ldots, T$$

as the energy request of deferrable loads that arrive at time t. Then $\{a(t)\}_{t=1}^{T}$ is assumed to be a sequence of independent random variables with mean λ and variance s^2. Further, let $A(t) := \sum_{\tau=t+1}^{T} a(\tau)$ denote the total energy requested after time t for $t = 1, 2, \ldots, T$.

In summary, when attempting to solve ODLC an algorithm has, at time t, the following information is available: (i) the present deferrable loads, i.e., $\underline{p}_n$, $\overline{p}_n$, and P_n for $n \leq N(t)$; (ii) the expectation $\mathbb{E}(A(t))$ of future energy requests; and (iii) the prediction b_t of b.

III. Model predictive deferrable load control

A natural approach for solving the optimal deferrable load control (ODLC) problem described in the previous section is model predictive control, which has been applied in many settings, e.g., see [9] for a survey. In the context of deferrable load control, model predictive control is perhaps the most commonly suggested framework for algorithm design.

In general, model predictive control uses the available predictions about future to solve an optimization problem over the remainder of the control horizon, and implements only the first step in the solution obtained. In the context of the ODLC problem, at each time t, such an approach uses the updated prediction of baseload b_t and the updated prediction of future energy request $\mathbb{E}[A(t)]$ to solve an optimization problem over the remainder of the control horizon, and obtains deferrable load profiles $(p_n(t), p_n(t+1), \ldots, p_n(T))$ for the remainder $\{t, t+1, \ldots, T\}$ of the control horizon. Only $p_n(t)$ will be implemented at time t, and $p_n(t+1), \ldots, p_n(T)$ will be recomputed in the future with more updated predictions.

Interestingly, previous work has found that the optimization problem that is solved should not simply be a truncated version of the ODLC problem. Instead, [10] suggests introducing a pseudo load q to account for the future arrival of deferrable load. The introduction of this term allows for strong analytic guarantees on performance [10]. Hence, this is the version of model predictive control we consider in this paper.

Specifically, we consider the model predictive deferable load control algorithm described in Algorithm 1, where at each time t the following optimization problem is solved

Algorithm 1 Model Predictive Deferrable Load Control

Initialize $P_n(1) \leftarrow P_n$ for $n = 1, 2, \ldots, N$;
At time step $t = 1, \ldots, T$,

1: Update predictions b_t and $A(t)$;
2: Solve **ODLC-t**$\left(b_t, A(t), \left[P_n(t), \overline{p}_n, \underline{p}_n\right]_{n \in [N(t)]}\right)$ to obtain time-t power consumptions $p_n(t)$ for deferrable loads $n \leq N(t)$ that have already arrived;
3: Update $P_n(t+1) \leftarrow P_n(t) - p_n(t)$ for $n \leq N(t)$;

$$\begin{aligned}
&\textbf{ODLC-t}\left(b_t, A(t), \left[P_n(t), \overline{p}_n, \underline{p}_n\right]_{n \in [N(t)]}\right) \\
\min \quad & \sum_{\tau=t}^{T} \left(\sum_{n=1}^{N(t)} p_n(\tau) + q(\tau) + b_t(\tau) \right)^2 \\
\text{over} \quad & p_n(\tau), q(\tau), \quad n \leq N(t), \tau \geq t \\
\text{s.t.} \quad & \underline{p}_n(\tau) \leq p_n(\tau) \leq \overline{p}_n(\tau), \quad n \leq N(t),\ \tau \geq t; \\
& \sum_{\tau=t}^{T} p_n(\tau) = P_n(t), \quad n \leq N(t); \\
& \underline{q}(\tau) \leq q(\tau) \leq \overline{q}(\tau), \quad \tau \geq t; \\
& \sum_{\tau=t}^{T} q(\tau) = \mathbb{E}(A(t)),
\end{aligned}$$

In this formulation, $P_n(t) = P_n - \sum_{\tau=1}^{t-1} p_n(\tau)$ is the energy to be consumed at or after time t, for all n and all t, and $\underline{q}$, $\overline{q}$ are given constants (from historical data) with $\underline{q}(t) = \overline{\overline{q}}(t) = 0$.

Importantly, if predictions are exact then Algorithm 1 solves ODLC exactly. Further, prior papers have shown that Algorithm 1 can be run in an completely distributed manner and still ensure (fast) convergence to an optimal solution [17].

Proposition 1 ([17]). *When there is no uncertainty, i.e., $N(t) = N$ and $b_t = b$ for $t = 1, 2, \ldots, T$, then the sequence of deferrable load schedules $(p^{(1)}, p^{(2)}, \ldots, p^{(k)}, \ldots)$ obtained by Algorithm 1 converge to optimal schedules of ODLC.*

For our purposes, the most relevant part of previous studies of Algorithm 1 is that there exists simple characterizations

of the solutions to ODLC-t, which prove quite useful when analyzing the performance of the algorithm.

Specifically, in cases where there are a large number of deferrable loads, the solutions to ODLC-t satisfy a property that is referred to as t-valley-filling.

Definition 1. *For any time $t = 1, \dots, T$, a feasible schedule (p, q) is called t-**valley-filling**, if there exists $C(t) \in \mathbb{R}$ such that*

$$\sum_{n=1}^{N(t)} p_n(\tau) + q(\tau) + b_t(\tau) = C(t), \qquad \tau = t, \dots, T. \quad (5)$$

Proposition 2. *At time $t = 1, \dots, T$, a t-valley-filling deferrable load schedule, if it exists, solves ODLC-t.*

This characterization provides a strong basis for the performance analysis of Algorithm 1. To see this, note that if there exists a t-valley-filling solution then, besides being optimal, it ensures that the aggregate load satisfies

$$d(t) = \frac{1}{T-t+1}\left(\sum_{n=1}^{N(t)} P_n(t) + \mathbb{E}(A(t)) + \sum_{\tau=t}^{T} b_t(\tau)\right) \quad (6)$$

for $t = 1, 2, \dots, T$. This property is used to analyze the load variance obtained by Algorithm 1 throughout this paper.

IV. Performance analysis

The main focus of this paper is the performance analysis of model predictive deferrable load control (Algorithm 1). As discussed, the algorithm has been introduced in [10] where a decentralized version of the algorithm was proven to optimally solve the ODLC problem (3) if predictions are exact. Then, [10] analyzed the *average-case* performance in a context with uncertain predictions, paralleling the current paper. The goal of this paper is to perform a *distributional analysis*, rather than simply average-case analysis. However, to provide context we first introduce the previous average-case analysis and contrast it with a (novel) worst-case analysis.

Recall that, throughout, we measure performance via the load variance V obtained by the algorithm, i.e., (1) and we focus on the case where there are enough deferrable loads in order to ensure that a t-valley-filling solution exists.

A. Average-case analysis (previous work)

An average-case analysis of Algorithm 1 was performed in [10]. The following is the main result from that paper.

Proposition 3 ([10])**.** *If a t-valley-filling solution exists for $t = 1, 2, \dots, T$, then the expected load variance obtained by Algorithm 1 is*

$$\mathbb{E}(V) = \frac{s^2}{T}\sum_{t=2}^{T}\frac{1}{t} + \frac{\sigma^2}{T^2}\sum_{t=0}^{T-1} F^2(t)\frac{T-t-1}{t+1}. \quad (7)$$

where $F(t) := \sum_{s=0}^{t} f(s)$ for $t = 0, \dots, T$.

This expression is quite informative, and has been explored in depth in [10]. Briefly, note that Proposition 3 explicitly states how the generation prediction error (σ) and deferrable load prediction error (s) affect the expected load variance $\mathbb{E}(V)$. Further, it highlights that $\mathbb{E}(V) \to 0$ as the predictions get precise, i.e., $\sigma \to 0$ and $s \to 0$. More importantly, it follows from Proposition 3 that $\mathbb{E}(V)$ tends to 0 as time horizon T increases, provided that the error correlation $f(t)$ decays sufficiently fast with t.

Proposition 4 ([10])**.** *If* (6) *holds, and the error correlation $f \sim O(t^{-\frac{1}{2}-\alpha})$ for some $\alpha > 0$, then $\mathbb{E}(V) \to 0$ as $T \to \infty$.*

We highlight the above proposition because the condition on f reappears in our distributional analysis of V. Though technical, this condition is practically relevant since the error correlation $f(t)$ usually decays fast with t and the time horizon T is usually long, which implies that Algorithm 1 should typically have good average case performance.

B. Worst-case analysis

The results surveyed above highlight that Algorithm 1 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., [19]–[21]. While no existing results apply directly to the setting of this paper, it is straightforward to see that the worst-case performance of Algorithm 1 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 2. *We say that **prediction errors are bounded** if there exist ϵ_1 and ϵ_2 such that, at any time $t = 1, \dots, T$,*

$$|a(t) - \lambda| \le \epsilon_1, \ |e(t)| \le \epsilon_2. \quad (8)$$

In this situation, it is straightforward to see that the worst case performance of Algorithm 1 can potentially be quite bad. For two real numbers $a, b \in \mathbb{R}$, define $a \vee b := \max\{a, b\}$.

Proposition 5. *If a t-valley-filling solution exists for $t = 1, 2, \dots, T$, and prediction errors are bounded by ϵ_1 and ϵ_2 as in* (8)*, then the worst-case load variance $\sup_{a,e} V$ achieved by Algorithm 1 is*

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

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 1. *If a t-valley-filling solution exists for $t = 1, 2, \dots, T$, and prediction errors are bounded by ϵ_1 and ϵ_2 as in* (8)*, then the worst-case load variance $\sup_{a,e} V$ achieved by Algorithm 1 is lower bounded as*

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

Interestingly, the form of Corollary 1 implies that, in the worst-case, Algorithm 1 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 4 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.

V. Distributional analysis

The contrast between the worst-case analysis (Proposition 5) and average-case analysis (Proposition 3) motivates the main goal of this paper – 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 1 look like.

A. Concentration bounds

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

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

which denotes the minimum value c such that $V \le c$ 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 1 does not deviate much from its average-case performance, i.e., we prove a concentration result for model predictive deferable load control.

Theorem 1. *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* (8)*. Then the distribution of the load variance V obtained by Algorithm 1 satisfies a Bernstein type concentration, i.e.,*

$$\mathbb{P}(V - \mathbb{E}V > t) \le \exp\left(\frac{-t^2}{16\epsilon^2\lambda_1(2\mathbb{E}V + t)}\right) \tag{9}$$

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 A.

To build intuition, the tail probability bound of V in (9) can be simplified for two different regimes of t as

$$\mathbb{P}(V - \mathbb{E}V > t) \le \begin{cases} \exp\left(\frac{-t^2}{48\epsilon^2\lambda_1\mathbb{E}V}\right), & t < \mathbb{E}V \\ \exp\left(\frac{-t}{48\epsilon^2\lambda_1}\right), & t \ge \mathbb{E}V. \end{cases} \tag{10}$$

Though looser than that in (9), the tail bound in (10) highlights that V has a Gaussian tail probability bound when $t < \mathbb{E}V$ and an Exponential tail probability bound when $t \ge \mathbb{E}V$.

Theorem 1 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 1 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 1 can be simplified as

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

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 (9): 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 1 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 2 and Proposition 4.

Corollary 2. *Under the assumptions of Theorem 1, 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 1 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 1 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 [15]—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 [22]). 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 [23] to compute a good estimate on the concentration of V.

B. Bounds on the variance

To further understand the scale of typical load variance V under Algorithm 1, 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 1.

Theorem 2. *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 (8). Then the variance $\mathrm{var}(V)$ of V obtained by Algorithm 1 is bounded above by*

$$\mathrm{var}(V) \leq \left(\frac{4\epsilon_1 s \ln T}{T}\right)^2 + \left(\frac{4\epsilon_2 \sigma}{T^2} \sum_{t=0}^{T-1} F^2(t) \frac{T-t+1}{t+1}\right)^2. \tag{11}$$

To interpret this result, let $\overline{\mathrm{var}(V)}$ denote the upper bound on $\mathrm{var}(V)$ provided in (11). Theorem 2 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 $\left(4\epsilon_1 s \ln T/T\right)^2$ 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 3. *Under the assumptions in Theorem 2, for $t > 0$,*

$$\mathbb{P}(|V - \mathbb{E}V| > t) \leq \frac{1}{t^2}\left[\left(\frac{4\epsilon_1 s \ln T}{T}\right)^2 + \left(\frac{4\epsilon_2 \sigma}{T^2} \sum_{\tau=0}^{T-1} F^2(\tau) \frac{T-\tau+1}{\tau+1}\right)^2\right]. \tag{12}$$

While the tail bound (9) in Theorem 1 scales at least exponentially in t, the Chebyshev inequality only provides a tail bound (12) that scales inverse quadratically in t. Hence for large t, (9) provides a much tighter tail bound. However for small values of t, the tail bound (12) is usually tighter since the variance $\mathrm{var}(V)$ is well estimated in (11).

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.

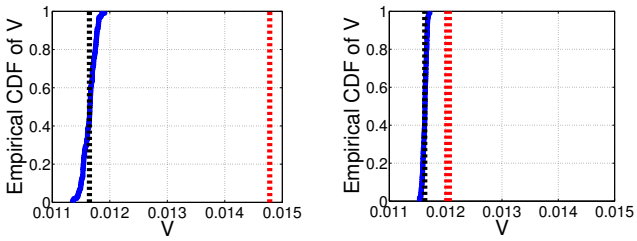
Corollary 4. *Under the assumptions of Theorem 2, if the error correlation $f \sim O(t^{-\frac{1}{2}-\alpha})$ for some $\alpha > 0$, then $\mathrm{var}(V) \to 0$ as $T \to \infty$.*

Note that the condition on f parallels that in Proposition 4.

C. A case study

Theorems 1 and 2 provide theoretical guarantees that the load variance V obtained by Algorithm 1 concentrates around its mean, if prediction errors are bounded as in (8) and error correlation decays sufficiently fast (c.f. Corollary 2). Thus, they give the intuition that the expected performance of Algorithm 1 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 1 as well, e.g., [10].



(a) 30% prediction error (b) 10% prediction error

Fig. 2: *The empirical cumulative distribution function of the load variance under Algorithm 1 over 24 hour control horizon using real data. The red line represents the analytic bound on the 90% confidence interval computed from Theorem 1, and the black line shows the empirical mean.*

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 [10], where an average-case analysis of Algorithm 1 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 [24] as the non-deferrable load, and wind power generation data from the Alberta Electric System Operator from 2004 to 2012 [25]. 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 [10].

Given this setting, we simulate 100 instances in each scenario and compare the results with the Theorems 1. The results are shown in Fig. 2 where we plot the cumulative distribution (CDF) of the load variance produced by Algorithm 1 under two different scenarios. Specifically, in Fig. 2a, we assume the prediction error in wind power generation is 30%, and in Fig. 2b, 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 1. For both cases, the results highlight a strong concentration around the mean, and the analytic bound from Theorem 1 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 1.

VI. Conclusion

In this paper we have studied a promising algorithm for direct control demand response: model predictive deferrable load control. We have, for the first time, 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 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.

REFERENCES

[1] M. H. Albadi and E. El-Saadany, "Demand response in electricity markets: An overview," in *Power Engineering Society General Meeting, 2007. IEEE*, June 2007, pp. 1–5.

[2] E. Sortomme, M. Hindi, S. MacPherson, and S. Venkata, "Coordinated charging of plug-in hybrid electric vehicles to minimize distribution system losses," *Smart Grid, IEEE Transactions on*, vol. 2, no. 1, pp. 198–205, March 2011.

[3] K. Clement-Nyns, E. Haesen, and J. Driesen, "The impact of charging plug-in hybrid electric vehicles on a residential distribution grid," *Power Systems, IEEE Transactions on*, vol. 25, no. 1, pp. 371–380, Feb 2010.

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

[5] 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 *Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP*. IEEE, 2010, pp. 293–299.

[6] M. Ilic, J. W. Black, and J. L. Watz, "Potential benefits of implementing load control," in *Power Engineering Society Winter Meeting, 2002. IEEE*, vol. 1. IEEE, 2002, pp. 177–182.

[7] Z. Ma, D. Callaway, and I. Hiskens, "Decentralized charging control for large populations of plug-in electric vehicles," in *Decision and Control (CDC), 2010 49th IEEE Conference on*. IEEE, 2010, pp. 206–212.

[8] L. Gan, U. Topcu, and S. H. Low, "Stochastic distributed protocol for electric vehicle charging with discrete charging rate," in *Power and Energy Society General Meeting, 2012 IEEE*. IEEE, 2012, pp. 1–8.

[9] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control engineering practice*, vol. 11, no. 7, pp. 733–764, 2003.

[10] 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*. ACM, 2013, pp. 113–124.

[11] A. J. Conejo, J. M. Morales, and L. Baringo, "Real-time demand response model," *Smart Grid, IEEE Transactions on*, vol. 1, no. 3, pp. 236–242, 2010.

[12] J. Roos and I. Lane, "Industrial power demand response analysis for one-part real-time pricing," *Power Systems, IEEE Transactions on*, vol. 13, no. 1, pp. 159–164, 1998.

[13] S. Chen and L. Tong, "iems for large scale charging of electric vehicles: Architecture and optimal online scheduling," in *Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference on*. IEEE, 2012, pp. 629–634.

[14] 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 preprint arXiv:1106.5063*, 2011.

[15] C. McDiarmid, "On the method of bounded differences," *Surveys in combinatorics*, vol. 141, no. 1, pp. 148–188, 1989.

[16] K.-H. Ng and G. B. Sheble, "Direct load control-a profit-based load management using linear programming," *Power Systems, IEEE Transactions on*, vol. 13, no. 2, pp. 688–694, 1998.

[17] L. Gan, U. Topcu, and S. Low, "Optimal decentralized protocol for electric vehicle charging," in *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*. IEEE, 2011, pp. 5798–5804.

[18] M. Pedrasa, T. Spooner, and I. MacGill, "Coordinated scheduling of residential distributed energy resources to optimize smart home energy services," *Smart Grid, IEEE Transactions on*, vol. 1, no. 2, pp. 134–143, Sept 2010.

[19] J. a. Lee and Z. Yu, "Worst-case formulations of model predictive control for systems with bounded parameters," *Automatica*, vol. 33, no. 5, pp. 763–781, 1997.

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

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

[22] S. Boucheron, G. Lugosi, P. Massart *et al.*, "On concentration of self-bounding functions," *Electronic Journal of Probability*, vol. 14, no. 64, pp. 1884–1899, 2009.

[23] M. Ledoux, "Concentration of measure and logarithmic sobolev inequalities," in *Seminaire de probabilites XXXIII*. Springer, 1999, pp. 120–216.

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

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

APPENDIX

A. Proof of Theorem 1

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

Lemma 1 (Theorem 3.2, [23])**.** *Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be convex and X be supported on $[-d/2, d/2]^n$, then*

$$\mathbb{E}[\exp(f(X))f(X)] - \mathbb{E}[\exp(f(X))]\log\mathbb{E}[\exp(f(X))] \leq \frac{d^2}{2}\mathbb{E}[\exp(f(X))||\nabla f(X)||^2]. \quad (13)$$

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

Lemma 2. *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*

$$||\nabla f||^2 \leq af + b, \quad (14)$$

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

$$\mathbb{P}\{f(X) > t\} \leq \exp\left(\frac{-t^2}{2b + at}\right). \quad (15)$$

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 1 that

$$\mathbb{E}\left[e^{\theta f}\theta f\right] - \mathbb{E}\left[e^{\theta f}\right]\ln\mathbb{E}\left[e^{\theta f}\right] \leq \frac{d^2}{2}\mathbb{E}\left[e^{\theta f}||\theta\nabla f||^2\right],$$

$$\theta m'(\theta) - m(\theta)\ln m(\theta) \leq \frac{1}{2}\theta^2 d^2\mathbb{E}[e^{\theta f}||\nabla f||^2].$$

According to the self-bounding property (14), one has

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

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

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

Integrate both sides from 0 to s to get

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

for $s \ge 0$. 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

$$\left(\frac{1}{s} - \frac{ad^2}{2}\right)\ln m(s) \le \frac{1}{2}bd^2 s \tag{16}$$

for $s \ge 0$. We can bound the tail probability $\mathbb{P}\{f > t\}$ with the control (16) over the moment generating function $m(s)$.

In particular, one has

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

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

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

□

Proof of Theorem 1. It has been computed in [10] that the load variance V obtained by Algorithm 1 is composed of two parts:

$$V = V_1 + V_2$$

where

$$\begin{aligned}V_1 \;:=\; \frac{1}{T}\sum_{t=1}^{T}\Bigg[\sum_{\tau=1}^{t}\frac{\tau-1}{T(T-\tau+1)}(a(\tau)-\lambda) \\ - \sum_{\tau=t+1}^{T}\frac{1}{T}(a(\tau)-\lambda)\Bigg]^2\end{aligned}$$

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

$$\begin{aligned}V_2 \;:=\; \frac{1}{T}\sum_{t=1}^{T}\Bigg[\sum_{\tau=1}^{t}\frac{\tau-1}{T(T-\tau+1)}e(\tau)F(T-\tau) \\ - \sum_{\tau=t+1}^{T}\frac{1}{T}e(\tau)F(T-\tau)\Bigg]^2\end{aligned}$$

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

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

for $1 \le t, \tau \le T$. 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

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

According to the bounded prediction error assumption (8), one has $|y| \le \epsilon$ componentwise. Then, apply Lemma 2 to the random variable Z to obtain

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

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

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

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

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

which completes the proof of Theorem 1.

□