# Online Convex Optimization Using Predictions*

Niangjun Chen†, Anish Agarwal†, Adam Wierman†,
Siddharth Barman†, Lachlan L. H. Andrew‡
†California Institute of Technology, Pasadena, CA, USA,
E-mail: {ncchen,aagarwal,adamw,sid.barman} @caltech.edu
‡Monash University, Australia, E-mail: lachlan.andrew@monash.edu

## ABSTRACT

Making use of predictions is a crucial, but under-explored, area of online algorithms. This paper studies a class of online optimization problems where we have external noisy predictions available. We propose a stochastic prediction error model that generalizes prior models in the learning and stochastic control communities, incorporates correlation among prediction errors, and captures the fact that predictions improve as time passes. We prove that achieving sublinear regret and constant competitive ratio for online algorithms requires the use of an unbounded prediction window in adversarial settings, but that under more realistic stochastic prediction error models it is possible to use Averaging Fixed Horizon Control (AFHC) to simultaneously achieve sublinear regret and constant competitive ratio in expectation using only a constant-sized prediction window. Furthermore, we show that the performance of AFHC is tightly concentrated around its mean.

## Categories and Subject Descriptors

F.2.0 [**Analysis of Algorithms and Problem Complexity**]: General

## General Terms

Algorithms, Performance, Theory

## 1. INTRODUCTION

Making use of predictions about the future is a crucial, but under-explored, area of online algorithms. In this paper, we use online convex optimization to illustrate the insights that can be gained from incorporating a general, realistic model of prediction noise into the analysis of online algorithms.

**Online convex optimization.** In an online convex optimization (OCO) problem, a learner interacts with an environment in a sequence of rounds. In round $t$ the learner chooses an action $x_t$ from a convex decision/action space $G$, and then the environment reveals a convex cost function $c_t$ and the learner pays cost $c_t(x_t)$. An algorithm's goal is to minimize total cost over a (long) horizon $T$.

OCO has a rich theory and a wide range of important applications. In computer science, it is most associated with the so-called $k$-experts problem, an online learning problem where in each round $t$ the algorithm chooses one of $k$ possible actions, viewed as following the advice of one of $k$ "experts".

However, OCO is being increasingly broadly applied and, recently has become prominent in networking and cloud computing applications, including the design of dynamic capacity planning, load shifting and demand response for data centers [25, 29, 30, 31, 35], geographical load balancing of internet-scale systems [28, 42], electrical vehicle charging [16, 25], video streaming [21, 22] and thermal management of systems-on-chip [46, 47].

In typical applications of online convex optimization in networking and cloud computing there is an additional cost in each round, termed a "switching cost", that captures the cost of changing actions during a round. Specifically, the cost is $c_t(x_t) + \| x_t - x_{t-1} \|$, where $\| \cdot \|$ is a norm (often the one-norm). This additional term makes the online problem more challenging since the optimal choice in a round then depends on future cost functions. These "smoothed" online convex optimization problems have received considerable attention in the context of networking and cloud computing applications, e.g., [28, 29, 30, 31, 32], and are also relevant for many more traditional online convex optimization applications where, in reality, there is a cost associated with a change in action, e.g., portfolio management. *We focus on smoothed online convex optimization problems.*

**A mismatch between theory and practice.** As OCO algorithms make their way into networking and cloud computing applications, it is increasingly clear that there is a mismatch between the pessimistic results provided by the theoretical analysis (which is typically adversarial) and the near-optimal performance observed in practice.

Concretely, two main performance metrics have been studied in the literature: *regret*, defined as the difference between the cost of the algorithm and the cost of the offline optimal static solution, and the *competitive ratio*, defined as the maximum ratio between the cost of the algorithm and the cost of the offline optimal (dynamic) solution.

Within the *machine learning community*, regret has been heavily studied [20, 45, 49] and there are many simple algorithms that provide provably sublinear regret (also called "no regret"). For example, online gradient descent achieves $O(\sqrt{T})$-regret [49], even when there are switching costs [4]. In contrast, the *online algorithms community* considers a more general class of problems called "metrical task systems" (MTS) and focuses on competitive ratio [7, 8, 29]. Most results in this literature are "negative", e.g., when $c_t$ are ar-

*This work is partially supported by the NSF through CNS-1319820, EPAS-1307794, CNS-0846025, CCF-1101470 and the ARC through DP130101378.


*SIGMETRICS'15,* June 15–19, 2015, Portland, OR, USA.

http://dx.doi.org/10.1145/2745844.2745854.

bitrary, the competitive ratio grows without bound as the number of states in the decision space grows [8]. Exceptions to such negative results come only when structure is imposed on either the cost functions or the decision space, e.g., when the decision space is *one-dimensional* it is possible for an online algorithm to have a constant competitive ratio, e.g., [29]. However, *even in this simple setting no algorithms performs well for both competitive ratio and regret.* No online algorithm can have sublinear regret and a constant competitive ratio, even if the decision space is one-dimensional and cost functions are linear [4].

In contrast to the pessimism of the analytic work, applications in networking and cloud computing have shown that OCO algorithms can significantly outperform the static optimum while nearly matching the performance of the dynamic optimal, i.e., simultaneously do well for regret and the competitive ratio. Examples include dynamic capacity management of a single data center [2, 29] and geographical load balancing across multiple data centers [28, 35, 36].

It is tempting to attribute this discrepancy to the fact that practical workloads are not adversarial. However, a more important factor is that, in reality, *algorithms can exploit relatively accurate predictions about the future*, such as diurnal variations [5, 19, 31]. But a more important contrast between the theory and application is simply that, in reality, *predictions about the future are available and accurate, and thus play a crucial role in the algorithms.*

**Incorporating predictions.** It is no surprise that predictions are crucial to online algorithms in practice. In OCO, knowledge about future cost functions is valuable, even when noisy. However, despite the importance of predictions, we do not understand how prediction noise affects the performance (and design) of online convex optimization algorithms.

This is not due to a lack of effort. Most papers that apply OCO algorithms to networking and cloud computing applications study the impact of prediction noise, e.g., [1, 3, 31, 35]. Typically, these consider numerical simulations where i.i.d. noise terms with different levels of variability are added to the parameter being predicted, e.g., [18, 40]. While this is a valuable first step, it does not provide any *guarantees* about the performance of the algorithm with realistic prediction errors (which tend to be correlated, since an overestimate in one period is likely followed by another overestimate) and further does not help inform the *design* of algorithms that can effectively use predictions.

Though most work on predictions has been simulation based, there has also been significant work done seeking analytic guarantees. This literature can be categorized into:

(i) *Worst-case models* of prediction error typically assume that there exists a lookahead window $\omega$ such that within that window, prediction is near-perfect (*too optimistic*), and outside that window the workload is adversarial (*too pessimistic*), e.g., [8, 29, 28, 32, 12].

(ii) *Simple stochastic models* of prediction error typically consider i.i.d. errors, e.g., [10, 31, 30]. Although this is analytically appealing, it ignores important features of prediction errors, as described in the next section.

(iii) *Detailed stochastic models* of specific predictors applied for specific signal models, such as [48, 38, 39, 23]. This leads to less pessimistic results, but the guarantees, and the algorithms themselves, become *too fragile* to assumptions on the system evolution.

**Contributions of this paper.** First, *we introduce a general colored noise model for studying prediction errors in online convex optimization problems.* The model captures three important features of real predictors: (i) it allows for arbitrary correlations in prediction errors (e.g., both short and long range correlations); (ii) the quality of predictions decreases the further in the future we try to look ahead; and (iii) predictions about the future are updated as time passes. Further, it strikes a middle ground between the worst-case and stochastic approaches. In particular, it does not make any assumptions about an underlying stochastic process or the design of the predictor. Instead, it only makes (weak) assumptions about the stochastic form of the error of the predictor; these assumptions are satisfied by many common stochastic models, e.g., the prediction error of standard Weiner filters [44] and Kalman filters [24]. Importantly, by being agnostic to the underlying stochastic process, the model allows worst-case analysis with respect to the realization of the underlying cost functions.

Second, using this model, *we show that a simple algorithm, Averaging Fixed Horizon Control (AFHC) [28], simultaneously achieves sublinear regret and a constant competitive ratio in expectation using very limited prediction*, i.e., a prediction window of size $O(1)$, in nearly all situations when it is feasible for an online algorithm to do so (Theorem 2). Further, we show that the performance of AFHC is tightly concentrated around its mean (Theorem 10). Thus, AFHC extracts the asymptotically optimal value from predictions. Additionally, our results inform the choice of the optimal prediction window size. (For ease of presentation, both Theorems 9 and 10 are stated and proven for the specific case of online LASSO – see Section 2 – but the proof technique can be generalized in a straightforward way.)

Importantly, Theorem 5 highlights that the dominant factor impacting whether the prediction window should be long or short in AFHC is not the variance of the noise, but rather the correlation structure of the noise. For example, if prediction errors are i.i.d. then it is optimal for AFHC to look ahead as far as possible (i.e., $T$) regardless of the variance, but if prediction errors have strong short-range dependencies then the optimal prediction window is constant sized regardless of the variance.

Previously, AFHC had only been analyzed in the adversarial model [29], and our results are in stark contrast to the pessimism of prior work. To highlight this, we prove that in the "easiest" adversarial model (where predictions are exact within the prediction window), no online algorithm can achieve sublinear regret *and* a constant competitive ratio when using a prediction window of constant size (Theorem 1). This contrast emphasizes the value of moving to a more realistic stochastic model of prediction error.

# 2. ONLINE CONVEX OPTIMIZATION WITH SWITCHING COSTS

Throughout this paper we consider online convex optimization problems with switching costs, i.e., "smoothed" online convex optimization (SOCO) problems.

## 2.1 Problem Formulation

The standard formulation of an online optimization problem with switching costs considers a convex decision/action space $G \subset \mathbb{R}^n$ and a sequence of cost functions $\{c_1, c_2, \ldots\}$, where each $c_t : G \to \mathbb{R}^+$ is convex. At time $t$, the online algorithm first chooses an action, which is a vector $x_t \in G$, the environment chooses a cost function $c_t$ from a set $\mathcal{C}$, and the algorithm pays a stage cost $c_t(x_t)$ and a switching cost $\beta||x_t - x_{t-1}||$ where $\beta \in (\mathbb{R}^+)$. Thus, the total cost of the online algorithm is defined to be

$$\text{cost}(ALG) = \mathbb{E}_x \left[ \sum_{t=1}^{T} c_t(x_t) + \beta||x_t - x_{t-1}|| \right], \qquad (1)$$

where $x_1, \dots, x_T$ are the actions chosen by the algorithm, $ALG$. Without loss of generality, assume the initial action $x_0 = 0$, the expectation is over any randomness used by the algorithm, and $||\cdot||$ is a seminorm on $\mathbb{R}^n$.

Typically, a number of assumptions about the action space $G$ and the cost functions $c_t$ are made to allow positive results to be derived. In particular, the action set $G$ is often assumed to be closed, nonempty, and bounded, where by bounded we mean that there exists $D \in \mathbb{R}$ such that for all $x, y \in G, ||x - y|| \leq D$. Further, the cost functions $c_t$ are assumed to have a uniformly bounded subgradient, i.e., there exists $N \in \mathbb{R}^+$ such that, for all $x \in G$, $||\nabla c_t(x)|| \leq N$.

Since our focus in this paper is on predictions, we consider a variation of the above with *parameterized* cost functions $c_t(x_t; y_t)$, where the parameter $y_t$ is the focus of prediction. Further, except when considering worst-case predictions, we adopt a specific form of $c_t$ for concreteness. We focus on a tracking problem where the online algorithm is trying to do a "smooth" tracking of $y_t$ and pays a least square penalty each round.

$$\text{cost}(ALG) = \mathbb{E}_x \left[ \sum_{t=1}^{T} \frac{1}{2}||y_t - Kx_t||_2^2 + \beta||x_t - x_{t-1}||_1. \right], \tag{2}$$

where the target $y_t \in \mathbb{R}^m$, and $K \in \mathbb{R}^{m \times n}$ is a (known) linear map that transforms the control variable into the space of the tracking target. Let $K^\dagger$ be the Moore-Penrose pseudoinverse of $K$.

We focus on this form because it represents an online version of the LASSO (Least Absolute Shrinkage and Selection Operator) formulation, which is widely studied in a variety of contexts, e.g., see [13, 14, 15, 41] and the references therein. Typically in LASSO the one-norm regularizer is used to induce sparsity in the solution. In our case, this corresponds to specifying that a good solution does not change too much, i.e., $x_t - x_{t-1} \neq 0$ is infrequent. Importantly, the focus on LASSO, i.e., the two-norm loss function and one-norm regularizer, is simply for concreteness and ease of presentation. Our proof technique generalizes (at the expense of length and complexity).

We assume that $K^T K$ is invertible and that the static optimal solution to (2) is positive. Neither of these is particularly restrictive. If $K$ has full column rank then $K^T K$ is invertible. This is a reasonable, for example, when the dimensionality of the action space $G$ is small relative to the output space. Note that typically $K$ is designed, and so it can be chosen to ensure these assumptions are satisfied. Additionally if K is invertible, then it no longer appears in the results provided.

Finally, it is important to highlight a few contrasts between the cost function in (2) and the typical assumptions in the online convex optimization literature. First, note that the feasible action set $G = \mathbb{R}^n$ is unbounded. Second, note that gradient of $c_t$ can be arbitrarily large when $y_t$ and $Kx_t$ are far apart. Thus, both of these are relaxed compared to what is typically studied in the online convex optimization literature. We show in Section 5 that, we can have sublinear regret even in this relaxed setting.

## 2.2 Performance Metrics

The performance of online algorithms for SOCO problems is typically evaluated via two performance metrics: *regret* and the *competitive ratio*. Regret is the dominant choice in the machine learning community and competitive ratio is the dominant choice in the online algorithms community. The key difference between these measures is whether they compare the performance of the online algorithm to the offline optimal static solution or the offline optimal dynamic solution. Specifically, the optimal offline *static* solution, is[1]

$$STA = \operatorname*{argmin}_{x \in G} \sum_{t=1}^{T} c_t(x) + \beta||x||, \tag{3}$$

and the optimal *dynamic* solution is

$$OPT = \operatorname*{argmin}_{(x_1,\dots,x_T) \in G^T} \sum_{t=1}^{T} c_t(x_t) + \beta||(x_t - x_{t-1})||. \tag{4}$$

DEFINITION 1. *The **regret** of an online algorithm, ALG, is less than $\rho(T)$ if the following holds:*

$$\sup_{(c_1,\dots,c_T) \in \mathcal{C}^T} \text{cost}(ALG) - \text{cost}(STA) \leq \rho(T). \tag{5}$$

DEFINITION 2. *An online algorithm ALG is said to be $\rho(T)$-**competitive** if the following holds:*

$$\sup_{(c^1,\dots,c^T) \in \mathcal{C}^T} \frac{\text{cost}(ALG)}{\text{cost}(OPT)} \leq \rho(T) \tag{6}$$

The goals are typically to find algorithms with a (small) constant competitive ratio ("constant-competitive") and to find online algorithms with sublinear regret, i.e., an algorithm $ALG$ that has regret $\rho(T)$ bounded above by some $\hat{\rho}(T) \in o(T)$; note that $\rho(T)$ may be negative if the concept we seek to learn varies dynamically. Sublinear regret is also called "no-regret", since the time-average loss of the online algorithm goes to zero as $T$ grows.

## 2.3 Background

To this point, there are large literatures studying both the designs of no-regret algorithms and the design of constant-competitive algorithms. However, in general, these results tell a pessimistic story.

In particular, on a positive note, it is possible to design simple, no-regret algorithms, e.g., *online gradient descent* (OGD) based algorithms [49, 20] and *Online Newton Step and Follow the Approximate Leader* algorithms [20]. (Note that the classical setting does not consider switching costs; however, [4] shows that similar regret bounds can be obtained when switching costs are considered.)

However, when one considers the competitive ratio, results are much less optimistic. Historically, results about competitive ratio have considered weaker assumptions, i.e., the cost functions $c_t$ and the action set $G$ can be nonconvex, and the switching cost is an arbitrary metric $d(x_t, x_{t-1})$ rather than a seminorm $||x_t - x_{t-1}||$. The weakened assumptions, together with the tougher offline target for comparison, leads to the fact that most results are "negative". For example, [8] has shown that any deterministic algorithm must be $\Omega(n)$-competitive given metric decision space of size $n$. Furthermore, [7] has shown that any randomized algorithm must be $\Omega(\sqrt{\log n / \log\log n})$-competitive. To this point, positive results are only known in very special cases. For example, [29] shows that, when $G$ is a one dimensional normed space, there exists a deterministic online algorithm that is 3-competitive.

Results become even more pessimistic when one asks for algorithms that perform well for both competitive ratio and regret. Note that performing well for both measures is particularly desirable for many networking and cloud computing

[1]One switching cost is incurred due to the fact that we enforce $x_0 = 0$.

applications where it is necessary to both argue that a dynamic control algorithm provides benefits over a static control algorithm (sublinear regret) and is near optimal (competitive ratio). However, a recent result in [4] highlights that such as goal is impossible: even when the setting is restricted to a one dimensional normed space with linear cost functions no online algorithm can simultaneously achieve sublinear regret and constant competitive ratio [2].

# 3. MODELING PREDICTION ERROR

The adversarial model underlying most prior work on online convex optimization has led to results that tend to be pessimistic; however, in reality, *algorithms can often use predictions about future cost functions in order to perform well.*

Knowing information about future cost functions is clearly valuable for smoothed online convex optimization problems, since it allows you to better justify whether it is worth it to incur a switching cost during the current stage. Thus, it is not surprising that predictions have proven valuable in practice for such problems.

Given the value of predictions in practice, it is not surprising that there have been numerous attempts to incorporate models of prediction error into the analysis of online algorithms. We briefly expand upon the worst-case and stochastic approaches described in the introduction to motivate our approach, which is an integration of the two.

**Worst-case models.** Worst-case models of prediction error tend to assume that there exists a lookahead window, $w$, such that within that window, a perfect (or near-perfect, e.g., error bounded by $\varepsilon$) prediction is available. Then, outside of that window the workload is adversarial. A specific example is that, for any $t$ the online algorithm knows $y_t, \dots, y_{t+w}$ precisely, while $y_{t+w+1}, \dots$ are adversarial.

Clearly, such models are both *too optimistic* about the the predictions used and *too pessimistic* about what is outside the prediction window. The result is that algorithms designed using such models tend to be too trusting of short term predictions and too wary of unknown fluctuations outside of the prediction window. Further, such models tend to underestimate the value of predictions for algorithm performance. To illustrate this, we establish the following theorem.

THEOREM 1. *For any constant $\gamma > 0$ and any online algorithm $A$ (deterministic or randomized) with constant lookahead $w$, either the competitive ratio of the algorithm is at least $\gamma$ or its regret, is $\Omega(T)$. Here $T$ is the number of cost functions in an instance.*

The above theorem focuses on the "easiest" worst-case model, i.e., where the algorithm is allowed perfect lookahead for $w$ steps. Even in this case, an online algorithm must have super-constant lookahead in order to simultaneously have sublinear regret and a constant competitive ratio. Further, the proof (given in Appendix A.1) highlights that this holds even in the scalar setting with linear cost functions. Thus, worst-case models are *overly pessimistic* about the value of prediction.

**Stochastic models.** Stochastic models tend to come in two forms: (i) i.i.d. models or (ii) detailed models of stochastic processes and specific predictors for those processes.

In the first case, for reasons of tractability, prediction errors are simply assumed to be i.i.d. mean zero random variables. While such an assumption is clearly analytically appealing, it is also quite simplistic and ignores many important features of prediction errors. For example, in reality, predictions have increasing error the further in time we look ahead due to correlation of predictions errors in nearby points in time. Further, predictions tend to be updated or refined as time passes. These fundamental characteristics of predictions cannot be captured by the i.i.d. model.

In the second case, which is common in control theory, a specific stochastic model for the underlying process is assumed and then an optimal predictor (filter) is derived. Examples here include the derivation of Weiner filters and Kalaman filters for the prediction of wide-sense stationary processes and linear dynamical systems respectively, see [23]. While such approaches avoid the pessimism of the worst-case viewpoint, they instead tend to be fragile to the underlying modeling assumptions. In particular, an online algorithm designed to use a particular filter based on a particular stochastic model lacks the robustness to be used in settings where the underlying assumptions are not valid.

## 3.1 A General Prediction Model

A key contribution of this paper is the development of a model for studying predictions that provides a middle ground between the worst-case and the stochastic viewpoints. The model we propose below seeks a middle ground by not making any assumption on the underlying stochastic process or the design of the predictor, but instead making assumptions only on the form of the error of the predictor. Thus, it is agnostic to the predictor and can be used in worst-case analysis with respect to the realization of the underlying cost functions.

Further, the model captures three important features of real predictors: (i) it allows for correlations in prediction errors (both short range and long range); (ii) the quality of predictions decreases the further in the future we try to look ahead; and (iii) predictions about the future are refined as time passes.

Concretely, throughout this paper we model prediction error via the following equation:

$$y_t = y_{t|\tau} + \sum_{s=\tau+1}^{t} f(t-s)e(s). \tag{7}$$

Here, $y_{t|\tau}$ is the prediction of $y_t$ made at time $\tau < t$. Thus, $y_t - y_{t|\tau}$ is the prediction error, and is specified by the summation in (7). In particular, the prediction error is modeled as a weighted linear combination of per-step noise terms, $e(s)$ with weights $f(t-s)$ for some deterministic impulse function $f$. The key assumptions of the model are that $e(s)$ are i.i.d. with mean zero and positive definite covariance $R_e$; and that $f$ satisfies $f(0) = I$ and $f(t) = 0$ for $t < 0$. Note that, as the examples below illustrate, it is common for the impulse function to decay as $f(s) \sim 1/s^{\alpha}$. As we will see later, this simple model is flexible enough to capture the prediction error that arise from classical filters on time series, and it can represent all forms of stationary prediction error by using appropriate forms of $f$.

Some intuition for the form of the model can be obtained by expanding the summation in (7). In particular, note that for $\tau = t - 1$ we have

$$y_t - y_{t|t-1} = f(0)e(t) = e(t), \tag{8}$$

which highlights why we refer to $e(t)$ as the per-step noise.

Further, expanding the summation further gives

$$y_t - y_{t|\tau} = f(0)e(t) + f(1)e(t-1) + \ldots + f(t-\tau-1)e(\tau+1). \tag{9}$$

[2] Note that this impossibility is not the result of the regret being additive and the competitive ratio being multiplicative, as [4] proves the parallel result for competitive difference, which is an additive comparison to the dynamic optimal.

Note that the first term is the one-step prediction error $y_t - y_{t|t-1}$; the first two terms make up the two-step prediction error $y_t - y_{t|t-2}$; and so on. This highlights that predictions in the model have increasing noise as one looks further ahead in time and that predictions are refined as time goes forward.

Additionally, note that the form of (9) highlights that the impulse function $f$ captures the degree of short-term/long-term correlation in prediction errors. Specifically, the form of $f(t)$ determines how important the error $t$ steps in the past is for the prediction. Since we assume no structural form for $f$, complex correlation structures are possible.

Naturally, the form of the correlation structure plays a crucial role in the performance results we prove. But, the detailed structure is not important, only its effect on the aggregate variance. Specifically, the impact of the correlation structure on performance is captured through the following two definitions, which play a prominent role in our analysis. First, for any $w > 0$, let $||f_w||^2$ be the two norm of prediction error covariance over $(w+1)$ steps of prediction, i.e.,

$$||f_w||^2 = \text{tr}(\mathbb{E}[\delta y_w \delta y_w^T]) = \text{tr}(R_e \sum_{s=0}^{w} f(s)^T f(s)), \tag{10}$$

where $\delta y_w^T = y_{t+w} - y_{t+w|t-1} = \sum_{s=t}^{t+w} f(t+w-s)e(s)$. The derivation of (10) is found in the proof of Theorem 5.

Second, let $F(w)$ be the two norm square of the projected cumulative prediction error covariance, i.e.,

$$F(w) = \sum_{t=0}^{w} \mathbb{E}||KK^\dagger \delta y_w||^2 = \text{tr}(R_e \sum_{s=0}^{w} (w-s+1) f(s)^T KK^\dagger f(s)). \tag{11}$$

Note that $KK^\dagger$ is the orthogonal projector onto the range space of $K$. Hence it is natural that the definitions are over the induced norm of $KK^\dagger$ since any action chosen from the space $F$ can only be mapped to the range space of $K$ i.e. no algorithm, online or offline, can track the portion of $y$ that falls in the null space of $K$.

Finally, unraveling the summation all the way to time zero highlights that the process $y_t$ can be viewed as a random deviation around the predictions made at time zero, $y_{t|0} := \hat{y}_t$, which are specified externally to the model:

$$y_t = \hat{y}_t + \sum_{s=1}^{t} f(t-s)e(s). \tag{12}$$

This highlights that an instance of the online convex optimization problem can be specified via either the process $y_t$ or via the initial predictions $\hat{y}_t$, and then the random noise from the model determines the other. We discuss this more when defining the notions of regret and competitive ratio we study in this paper in Section 3.3.

## 3.2 Examples

While the form of the prediction error in the model may seem mysterious, it is quite general, and includes many of the traditional models as special cases. For example, to recover the worst-case prediction model one can set, $\forall t, e(t) = 0$ and $\hat{y}_{t'}$ as unknown $\forall t' > t + w$ and then take the worst case over $\hat{y}$. Similarly, a common approach in robust control is to set $f(t) = \begin{cases} I, & t = 0; \\ 0, & t \neq 0 \end{cases}$, $|e(s)| < D, \forall s$ and then consider the worst case over $e$.

Additionally, strong motivation for it can be obtained by studying the predictors for common stochastic processes. In particular, the form of (7) matches the prediction error of standard Weiner filters [44] and Kalman filters [24], etc. To highlight this, we include a few brief examples below.

Example 1 (Wiener Filter). *Let $\{y_t\}_{t=0}^T$ be a wide-sense stationary stochastic process with $\mathbb{E}[y_t] = \hat{y}_t$, and covariance $\mathbb{E}[(y_i - \hat{y}_i)(y_j - \hat{y}_j)^T] = R_y(i-j)$, i.e., the covariance matrix $R_y > 0$ of $y = [y_1 \; y_2 \; \dots y_T]^T$ is a Toeplitz matrix. The corresponding $e(s)$ in the Wiener filter for the process is called the "innovation process" and can be computed via the Wiener-Hopf technique [23]. Using the innovation process $e(s)$, the optimal causal linear prediction is*

$$y_{t|\tau} = \hat{y}_t + \sum_{s=1}^{\tau} \langle y_t, e(s) \rangle ||e(s)||^{-2} e(s),$$

*and so the correlation function $f(s)$ as defined in (7) is*

$$f(s) = \langle y_s, e(0) \rangle ||e(0)||^{-2} = R_y(s) R_e^{-1}, \tag{13}$$

*which yields*

$$||f_w||^2 = \frac{1}{R_e} \sum_{s=0}^{w} R_y(s)^2 \text{ and } F(w) = \frac{1}{R_e} \sum_{s=0}^{w} (w-s+1) R_y(s)^2.$$

Example 2 (Kalman Filter). *Consider a stationary dynamical system described by the hidden state space model*

$$x'_{t+1} = Ax'_t + Bu_t, \quad y_t = Cx'_t + v_t.$$

*where the $\{u_t, v_t, x_0\}$ are $m \times 1, p \times 1$, and $n \times 1$-dimensional random variables such that*

$$\left\langle \begin{bmatrix} u_i \\ v_j \\ x_0 \end{bmatrix}, \begin{bmatrix} u_i \\ v_j \\ x_0 \\ 1 \end{bmatrix} \right\rangle = \begin{bmatrix} Q\delta_{ij} & S\delta_{ij} & 0 & 0 \\ S^*\delta_{ij} & R\delta_{ij} & 0 & 0 \\ 0 & 0 & \Pi_0 & 0 \end{bmatrix}.$$

*The Kalman filter for this process yields the optimal causal linear estimator $y_{t|\tau} = K[y_1^T, \dots, y_\tau^T]^T$ such that $y_{t|\tau} = \arg\min \mathbb{E}_\tau ||y_t - K'[y_1^T, \dots, y_\tau^T]^T||^2$. When $t$ is large and the system reaches steady state, the optimal prediction is given in the following recursive form [23]:*

$$x'_{t+1|t} = Ax'_{t|t-1} + K_p e(t), \; y_{0|-1} = 0, \; e(0) = y_0,$$
$$e(t) = y_t - Cx'_{t|t-1},$$

*where $K_p = (APC^* + BS)R_e^{-1}$, is the Kalman gain, and $R_e = R + CPC^*$ is the covariance of the innovation process $e_t$, and $P$ solves*

$$P = APA^* + CQC^* - K_p R_e K_p^*.$$

*This yields the predictions*

$$y_{t|\tau} = \sum_{s=1}^{\tau} \langle y_t, e(s) \rangle R_e^{-1} e(s)$$
$$= \sum_{s=1}^{\tau} CA^{t-s-1}(APC^* + BS)R_e^{-1} e(s).$$

*Thus, for stationary Kalman filter, the prediction error correlation function is*

$$f(s) = CA^{s-1}(APC^* + BS)R_e^{-1} = CA^{s-1}K_p, \tag{14}$$

*which yields*

$$||f_w||^2 = \sum_{s=0}^{w} \text{tr}(R_e (CA^{s-1}K_p)^T KK^\dagger (CA^{s-1}K_p)) \text{ and}$$
$$F(w) = \sum_{s=0}^{w} (w-s+1) \text{tr}(R_e (CA^{s-1}K_p)^T KK^\dagger (CA^{s-1}K_p)).$$

### 3.3 Performance Metrics

A key feature of the prediction model described above is that it provides a general stochastic model for prediction errors while not imposing any particular underlying stochastic process or predictor. Thus, it generalizes a variety of stochastic models while allowing worst-case analysis.

More specifically, when studying online algorithms using the prediction model above, one could either specify the instance via $y_t$ and then use the form of (7) to give random predictions about the instance to the algorithm or, one could specify the instance using $\hat{y} := y_{t|0}$ and then let the $y_t$ be randomly revealed using the form of (12). Note that, of the two interpretations, the second is preferable for analysis, and thus we state our theorems using it.

In particular, our setup can be interpreted as allowing an adversary to specify the instance via the initial (time 0) predictions $\hat{y}$, and then using the prediction error model to determine the instance $y_t$. We then take the worst-case over $\hat{y}$. This corresponds to having an adversary with a "shaky hand" or, alternatively, letting the adversary specify the instance but forcing them to also provide unbiased initial predictions.

In this context, we study the following notions of (expected) regret and (expected) competitive ratio, where the expectation is over the realization of the prediction noise $e$ and the measures consider the worst-case specification of the instance $\hat{y}$.

DEFINITION 3. *We say an online algorithm ALG, has (expected)* ***regret*** *at most $\rho(T)$ if*

$$\sup_{\hat{y}} \mathbb{E}_e[\text{cost}(ALG) - \text{cost}(STA)] \leq \rho(T). \tag{15}$$

DEFINITION 4. *We say an online algorithm ALG is $\rho(T)$-****competitive*** *(in expectation) if*

$$\sup_{\hat{y}} \frac{\mathbb{E}_e[\text{cost}(ALG)]}{\mathbb{E}_e[\text{cost}(OPT)]} \leq \rho(T). \tag{16}$$

Our proofs bound the competitive ratio through an analysis of the competitive difference, which is defined as follows.

DEFINITION 5. *We say an online algorithm ALG has (expected)* ***competitive difference*** *at most $\rho(T)$ if*

$$\sup_{\hat{y}} \mathbb{E}_e\left[\text{cost}(ALG) - \text{cost}(OPT)\right] \leq \rho(T). \tag{17}$$

Note that these expectations are with respect to the prediction noise, $(e(t))_{t=1}^{T}$, and so cost$(OPT)$ is also random. Note also that when cost$(OPT) \in \Omega(\rho(T))$ and ALG has competitive difference at most $\rho(T)$, then the algorithm has a constant (bounded) competitive ratio.

## 4. AVERAGING FIXED HORIZON CONTROL

A wide variety of algorithms have been proposed for online convex optimization problems. Given the focus of this paper on predictions, the most natural choice of an algorithm to consider is Receding Horizon Control (RHC), a.k.a., Model Predictive Control (MPC).

There is a large literature in control theory that studies RHC/MPC algorithms, e.g., [17, 33] and the references therein; and thus RHC is a popular choice for online optimization problems when predictions are available, e.g., [6, 43, 26, 11]. However, recent results have highlighted that while RHC can perform well for one-dimensional smoothed online optimization problems, it does not perform well (in the worst case) outside of the one-dimension case. Specifically, the competitive ratio of RHC with perfect lookahead $w$ is $1+O(1/w)$ in the one-dimensional setting, but is $1+\Omega(1)$ outside of this setting, i.e., the competitive ratio does not decrease to 1 as the prediction window $w$ increases [28].

In contrast, a promising new algorithm, Averaging Fixed Horizon Control (AFHC) proposed by [28] in the context of geographical load balancing maintains good performance in high-dimensional settings, i.e., maintains a competitive ratio of $1 + O(1/w)$[3]. Thus, in this paper, we focus on AFHC. Our results highlight that AFHC extracts the asymptotically optimal value from predictions, and so validates this choice.

As the name implies, AFHC averages the choices made by Fixed Horizon Control (FHC) algorithms. In particular, AFHC with prediction window size $(w+1)$ averages the actions of $(w+1)$ FHC algorithms.

ALGORITHM 1 (FIXED HORIZON CONTROL). *Let $\Omega_k = \{i : i \equiv k \mod (w+1)\} \cap [-w, T]$ for $k = 0, \ldots, w$. Then $FHC^{(k)}(w+1)$, the kth FHC algorithm is defined in the following manner. At timeslot $\tau \in \Omega_k$ (i.e., before $c_\tau$ is revealed), choose actions $x^{(k)}_{FHC,t}$ for $t = \tau, \ldots \tau + w$ as follows:*

*If $t \leq 0$, $x^{(k)}_{FHC,t} = 0$. Otherwise, let $x_{\tau-1} = x^{(k)}_{FHC,\tau-1}$, and let $(x^{(k)}_{FHC,t})_{t=\tau}^{\tau+w}$ be the vector that solves*

$$\min_{x_\tau,\ldots,x_{\tau+w}} \sum_{t=\tau}^{\tau+w} \hat{c}_t(x_t) + \beta||(x_t - x_{t-1})||$$

*where $\hat{c}_t(\cdot)$ is the prediction of the future cost $c_t(\cdot)$ for $t = \tau, \ldots, \tau + w$.*

Note that in the classical OCO with $(w+1)$-lookahead setting, $\hat{c}_t(\cdot)$ is exactly equal to the true cost $c(\cdot)$. Each FHC$^{(k)}(w+1)$ can be seen as a length $(w+1)$ fixed horizon control starting at position $k$. Given $(w+1)$ versions of FHC, AFHC is defined as the following:

ALGORITHM 2 (AVERAGING FIXED HORIZON CONTROL). *At timeslot $t \in 1, ..., T$, $AFHC(w+1)$ sets*

$$x_{AFHC,t} = \frac{1}{w+1} \sum_{k=0}^{w} x^{(k)}_{FHC,t}. \tag{18}$$

## 5. AVERAGE-CASE ANALYSIS

We first consider the average-case performance of AFHC (in this section), and then consider distributional analysis (in Section 6). We focus on the tracking problem in (2) for concreteness and conciseness, though our proof techniques generalize. Note that, unless otherwise specified, we use $||\cdot|| = ||\cdot||_2$.

Our main result shows that AFHC can simultaneously achieve sublinear regret and a constant competitive ratio using only a *constant-sized* prediction window in nearly all cases that it is feasible for an online algorithm to do so. This is in stark contrast with Theorem 1 for the worst-case prediction model.

THEOREM 2. *Let $w$ be a constant. AFHC$(w+1)$ is constant-competitive whenever $\inf_{\hat{y}} \mathbb{E}_e[OPT] = \Omega(T)$ and has sublinear regret whenever $\inf_{\hat{y}} \mathbb{E}_e[STA] \geq \alpha_1 T - o(T)$, for $\alpha_1 = 4V + 8B^2$, where*

$$V = \frac{\beta||K^\dagger||_1||f_w|| + 3\beta^2||(K^TK)^{-1}\mathbb{1}|| + F(w)/2}{w+1} \tag{19}$$

$$B = \beta||(K^T)^\dagger \mathbb{1}||. \tag{20}$$

[3] Note that this result assumes that the action set is bounded, i.e., for all feasible action $x, y$, there exists $D > 0$, such that $||x - y|| < D$, and that there exists $e_0 > 0$, s.t. $c_t(0) \geq e_0, \forall t$. The results we prove in this paper make neither of these assumptions.

and $||M||_1$ denotes the induced 1-norm of a matrix M

Theorem 2 imposes bounds on the expected costs of the dynamic and static optimal in order to guarantee a constant competitive ratio and sublinear regret. These bounds come about as a result of the noise in predictions. In particular, prediction noise makes it impossible for an online algorithm to achieve sublinear expected cost, and thus makes it infeasible for an online algorithm to compete with dynamic and static optimal solutions that perform too well. This is made formal in Theorems 3 and 4, which are proven in Appendix B. Recall that $R_e$ is the covariance of an estimation error vector, $e(t)$.

THEOREM 3. *Any online algorithm ALG that chooses $x_t$ using only (i) internal randomness independent of $e(\cdot)$ and (ii) predictions made up until time $t$, has expected cost $\mathbb{E}_e[\text{cost}(ALG)] \geq \alpha_2 T + o(T)$, where $\alpha_2 = \frac{1}{2}||R_e^{1/2}||^2_{KK^\dagger}$.*

THEOREM 4. *Consider an online algorithm ALG such that $\mathbb{E}_e[\text{cost}(ALG)] \in o(T)$. The actions, $x_t$, of ALG can be used to produce one-step predictions $y'_{t|t-1}$, such that mean square of the one-step prediction error is smaller than that for $y_{t|t-1}$, i.e., $\mathbb{E}_e||y_t - y'_{t|t-1}||^2 \leq \mathbb{E}_e||y_t - y_{t|t-1}||^2$, for all but sublinearly many $t$.*

Theorem 3 implies that it is impossible for any online algorithm that uses extra information (e.g., randomness) independent of the prediction noise to be constant competitive if $\mathbb{E}_e[\text{cost}(OPT)] = o(T)$ or to have sublinear regret if $\mathbb{E}_e[\text{cost}(STA)] \leq (\alpha_2 - \varepsilon)T + o(T)$, for $\varepsilon > 0$.

Further, Theorem 4 states that if an online algorithm does somehow obtain asymptotically smaller cost than possible using only randomness independent of the prediction error, then it must be using more information about future $y_t$ than is available from the predictions. This means that the algorithm can be used to build a better predictor.

Thus, the consequence of Theorems 3 and 4 is the observation that the condition in Theorem 2 for the competitive ratio is tight and the condition in Theorem 2 for regret is tight up to a constant factor, i.e., $\alpha_1$ versus $\alpha_2$. (Attempting to prove matching bounds here is an interesting, but very challenging, open question.)

In the remainder of the section, we outline the analysis needed to obtain Theorem 2, which is proven by combining Theorem 5 bounding the competitive difference of AFHC and Theorem 9 bounding the regret of AFHC. The analysis exposes the importance of the correlation in prediction errors for tasks such as determining the optimal prediction window size for AFHC. Specifically, the window size that minimizes the performance bounds we derive is determined not by the quality of predictions, but rather by how quickly error correlates, i.e., by $||f_\omega||^2$.

## Proof of Theorem 2

The first step in our proof of Theorem 2 is to bound the competitive difference of AFHC. This immediately yields a bound on the competitive ratio and, since it is additive, it can easily be adapted to bound regret as well.

The main result in our analysis of competitive difference is the following. This is the key both to bounding the competitive ratio and regret.

THEOREM 5. *The competitive difference of $AFHC(w+1)$ is $O(T)$ and bounded by:*

$$\sup_{\hat{y}} \mathbb{E}_e[\text{cost}(AFHC) - \text{cost}(OPT)] \leq VT \qquad (21)$$

*where $V$ is given by* (19)

Theorem 5 implies that the competitive ratio of $AFHC$ is bounded by a constant when $\text{cost}(OPT) \in \Omega(T)$.

The following corollary of Theorem 5 is obtained by minimizing $V$ with respect to $w$.

COROLLARY 6. *For $AFHC$, the prediction window size that minimizes the bound in Theorem 5 on competitive difference is a finite constant (independent of $T$) if $F(T) \in \omega(T)$ and is $T$ if there is i.i.d noise*[4].

The intuition behind this result is that if the prediction model causes noise to correlate rapidly, then a prediction for a time step too far into the future will be so noisy that it would be best to ignore it when choosing an action under AFHC. However, if the prediction model is nearly independent, then it is optimal for AFHC to look over the entire time horizon, $T$, since there is little risk from aggregating predictions. Importantly, notice that the quality (variance) of the predictions is not determinant, only the correlation.

Theorem 5 is proven using the following lemma (proven in the appendix) by taking expectation over noise.

LEMMA 7. *The cost of $AFHC(w+1)$ for any realization of $y_t$ satisfies*

$$\text{cost}(AFHC) - \text{cost}(OPT) \leq$$
$$\frac{1}{w+1}\sum_{k=0}^{w}\sum_{\tau\in\Omega_k}\left(\beta||x^*_{\tau-1} - x^{(k)}_{\tau-1}||_1 + \sum_{t=\tau}^{\tau+w}\frac{1}{2}||y_t - y_{t|\tau-1}||^2_{KK^\dagger}\right).$$

Next, we use the analysis of the competitive difference in order to characterize the regret of AFHC. In particular, to bound the regret we simply need a bound on the gap between the dynamic and static optimal solutions.

LEMMA 8. *The suboptimality of the offline static optimal solution $STA$ can be bounded below on each sample path by*

$$\begin{aligned}&\text{cost}(STA) - \text{cost}(OPT)\\ \geq&\frac{1}{2}\left(\sqrt{\sum_{t=1}^{T}||y_t - \bar{y}||^2_{KK^\dagger}} - 2B\sqrt{T})\right)^2 - 2B^2T - C\end{aligned}$$

*where $\bar{y} = \frac{\sum_{t=1}^{T} y_t}{T}$, $B = \beta||(K^T)^\dagger \mathbb{1}||_2$ and $C = \frac{\beta^2 \mathbb{1}^T (K^T K)^{-1}\mathbb{1}}{2T}$.*

Note that the bound above is in terms of $||(y_t - \bar{y})||^2_{KK^\dagger}$, which can be interpreted as a measure of the variability $y_t$. Specifically, it is the projection of the variation onto the range space of $K$.

Combining Theorem 5 with Lemma 8 gives a bound on the regret of AFHC, proven in Appendix B.

THEOREM 9. *$AHFC$ has sublinear expected regret if*

$$\inf_{\hat{y}} \mathbb{E}_e \sum_{t=1}^{T} ||KK^\dagger(y_t - \bar{y})||^2 \geq (8V + 16B^2)T,$$

*where $V$ and $B$ are defined in* (19) *and* (20).

Finally, we make the observation that, for all instances of $y$:

$$\begin{aligned}\text{cost}(STA) &= \frac{1}{2}\sum_{t=1}^{T}||y_t - K_x||^2 + \beta||x||_1\\ &\geq \frac{1}{2}\sum_{t=1}^{T}||(I - KK^\dagger)y_t + KK^\dagger y_t - Kx||^2\\ &= \frac{1}{2}\sum_{t=1}^{T}||(I - KK^\dagger)y_t||^2 + \frac{1}{2}||KK^\dagger y_t - Kx||^2\\ &\geq \frac{1}{2}||KK^\dagger(y_t - \bar{y})||^2.\end{aligned}$$

Hence by Theorem 9, we have the condition of the Theorem.

[4]Specifically $f(0) = I, f(t) = 0 \forall t > 0$

# 6. CONCENTRATION BOUNDS

The previous section shows that AFHC performs well in expectation, but it is also important to understand the distribution of the cost under AFHC. In this section, we show that, with a mild additional assumption on the prediction error $e(t)$, the event when there is a large deviation from the expected performance bound proven in Theorem 5 decays exponentially fast.

The intuitive idea behind the result is the observation that the competitive difference of AFHC is a function of the uncorrelated prediction error $e(1), \dots, e(T)$ that does not put too much emphasis on any one of the random variables $e(t)$. This type of function normally has sharp concentration around its mean because the effect of each $e(t)$ tends to cancel out.

For simplicity of presentation, we state and prove the concentration result for AFHC for the one dimensional tracking cost function

$$\frac{1}{2}\sum_{t=1}^{T}(y_t - x_t)^2 + \beta|x_t - x_{t-1}|.$$

In this case, $R_e = \sigma^2$, and the correlation function $f : \mathbb{N} \to R$ is a scalar valued function. The results can all be generalized to the multidimensional setting.

Additionally, for simplicity of presentation, we assume (for this section only) that $\{e(t)\}_{t=1}^{T}$ are uniformly bounded, i.e., $\exists \epsilon > 0$, s.t. $\forall t$, $|e(t)| < \epsilon$. Note that, with additional effort, the boundedness assumption can be relaxed to the case of $e(t)$ being subgaussian, i.e., $\mathbb{E}[\exp(e(t)^2/\epsilon^2)] \leq 2$, for some $\epsilon > 0$.[5]

To state the theorem formally, let $VT$ be the upper bound of the expected competitive difference of $AFHC$ in (21). Given $\{\hat{y}_t\}_{t=1}^{T}$, the competitive difference of $AFHC$ is a random variable that is a function of the prediction error $e(t)$. The following theorem shows that the probability that the cost of $AFHC$ exceeds that of $OPT$ by much more than the expected value $VT$ decays rapidly.

THEOREM 10. *The probability that the competitive difference of AFHC exceeds VT is exponentially small, i.e., for any $u > 0$:*

$$\begin{aligned}
&\mathbb{P}(\text{cost}(AFHC) - \text{cost}(OPT) > VT + u)\\
\leq &\exp\left(\frac{-u^2}{8\epsilon^2\frac{\beta^2 T}{(w+1)\sigma^2}||f_w||^2}\right) + \exp\left(\frac{-u^2}{16\epsilon^2\lambda(2\frac{T}{w+1}F(w)+u)}\right)\\
\leq &2\exp\left(\frac{-u^2}{a+bu}\right),
\end{aligned}$$

*where $||f_w||^2 = (\sum_{t=0}^{w}|f(t)|^2)$, the parameter $\lambda$ of concentration*

$$\lambda \leq \sum_{t=0}^{w}(w-t)f(t)^2 = \frac{1}{\sigma^2}F(w),$$

*and $a = 8\epsilon^2[T/(w+1)]\max(\frac{\beta^2}{\sigma^2}||f_w||^2, 4\lambda F(w))$, $b = 16\epsilon^2\lambda$.*

The theorem implies that the tail of the competitive difference of $AFHC$ has a Bernstein type bound. The bound decays much faster than the normal large deviation bounds obtained by bounding moments, i.e., Markov Inequality or Chebyshev Inequality. This is done by more detailed analysis of the structure of the competitive difference of $AFHC$ as a function of $e = (e(1), \dots, e(T))^T$.

Note that smaller values of $a$ and $b$ in Theorem 10 imply a sharper tail bound. We can see that smaller $||f_w||$ and smaller $F(w)$ implies the tail bound decays faster. Since higher prediction error correlation implies higher $||f_w||$ and $F(w)$, Theorem 10 quantifies the intuitive idea that, the performance of AFHC concentrates more tightly around its mean when the prediction error is less correlated.

[5]This involves more computation and worse constants in the concentration bounds. Interested readers are referred to Theorem 12 and the following remark of [9] for a way to generalize the concentration bound for the switching cost (Lemma 11), and Theorem 1.1 of [37] for a way to generalize the concentration bound for prediction error (Lemma 15).

## Proof of Theorem 10

To prove Theorem 10, we start by decomposing the bound in Lemma 7. In particular, Lemma 7 gives

$$\text{cost}(AFHC) - \text{cost}(OPT) \leq g_1 + g_2 \tag{22}$$

where

$$g_1 = \frac{1}{w+1}\sum_{k=0}^{w}\sum_{\tau\in\Omega_k}\beta|x^*_{\tau-1} - x^{(k)}_{\tau-1}|,$$

represents loss due to the switching cost, and

$$g_2 = \frac{1}{w+1}\sum_{k=0}^{w}\sum_{\tau\in\Omega_k}\sum_{t=\tau}^{\tau+w}\frac{1}{2}(y_t - y_{t|\tau-1})^2,$$

represents the loss due to the prediction error.

Let $V_1 = \frac{3\beta^2 T}{w+1} + \frac{\beta T}{w+1}||f_w||_2$, and $V_2 = \frac{T}{2(w+1)}F(w)$. Note that $VT = V_1 + V_2$. Then, by (22),

$$\begin{aligned}
&\mathbb{P}(\text{cost}(AFHC) - \text{cost}(OPT) > u + VT)\\
\leq &\mathbb{P}(g_1 > u/2 + V_1 \text{ or } g_2 > u/2 + V_2)\\
\leq &\mathbb{P}(g_1 > u/2 + V_1) + \mathbb{P}(g_2 > u/2 + V_2).
\end{aligned} \tag{23}$$

Thus, it suffices to prove concentration bounds for the loss due to switching cost, $g_1$, and the loss due to prediction error, $g_2$, deviating from $V_1$ and $V_2$ respectively. This is done in the following. The idea is to first prove that $g_1$ and $g_2$ are functions of $e = (e(1), \dots, e(T))^T$ that are not "too sensitive" to any of the elements of $e$, and then apply the method of bounded difference [34] and Log-Sobolev inequality [27]. Combining (23) with Lemmas 11 and 15 below will complete the proof of Theorem 10.

***Bounding the loss due to switching cost.*** This section establishes the following bound on the loss due to switching:

LEMMA 11. *The loss due to switching cost has a sub-Gaussian tail: for any $u > 0$,*

$$\mathbb{P}(g_1 > u + V_1) \leq \exp\left(\frac{-u^2}{2\epsilon^2\beta^2\frac{T}{w+1}(||f_w||)^2}\right). \tag{24}$$

To prove Lemma 11, we introduce two lemmas. Firstly, we use the first order optimality condition to bound $g_1$ above by a linear function of $e = (e(1), \dots, e(T))^T$ using the following lemma proved in the Appendix.

LEMMA 12. *The loss due to switching cost can be bounded above by*

$$g_1 \leq \frac{3\beta^2 T}{w+1} + \frac{\beta}{w+1}\sum_{k=0}^{w}\sum_{\tau\in\Omega_k}\left|\sum_{s=1\vee(\tau-w-2)}^{\tau-1} f(\tau-1-s)e(s)\right| \tag{25}$$

Let $g_1'(e)$ be the second term of $g_1$. Note that the only randomness in the upper bound (25) comes from $g_1'$.

LEMMA 13. *The expectation of $g_1'(e)$ is bounded above by*

$$\mathbb{E}_e g_1'(e) \leq \frac{\beta T}{w+1}||f_w||_2.$$

With Lemma 13, we can reduce (24) to proving a concentration bound on $g_1'(e)$, since

$$\mathbb{P}(g_1 > u + V_1) \leq \mathbb{P}(g_1' - \mathbb{E}g_1'(e) \leq u). \tag{26}$$

To prove concentration of $g_1'(e)$, which is a function of a collection of independent random variables, we use the method of bounded difference, i.e., we bound the difference of $g_1'(e)$ where one component of $e$ is replaced by an identically-distributed copy. Specifically, we use the following lemma, the one-sided version of one due to McDiarmid:

LEMMA 14 ([34], LEMMA 1.2). *Let $X = (X_1, \ldots, X_n)$ be independent random variables and $Y$ be the random variable $f(X_1, \ldots, X_n)$, where function $f$ satisfies*

$$|f(x) - f(x_k')| \leq c_k$$

*whenever $x$ and $x_k'$ differ in the $k$th coordinate. Then for any $t > 0$,*

$$\mathbb{P}(Y - \mathbb{E}Y > t) \leq \exp\left(\frac{-2t^2}{\sum_{k=1}^n c_k^2}\right).$$

PROOF OF LEMMA 11. Let $e = (e(1), \ldots, e(T))^T$, and $e_k' = (e(1), \ldots, e'(k), \ldots, e(T))^T$ be formed by replacing $e(k)$ with an independent and identically distributed copy $e'(k)$. Then

$$\begin{aligned}|g_1(e) - g_1(e_k')| &\leq \frac{1}{w+1}\beta \sum_{m=0}^{w} |f(m)(e(k) - e'(k))| \\ &\leq \frac{2}{w+1}\epsilon\beta \sum_{m=0}^{w} |f(m)| =: c_k.\end{aligned}$$

Hence

$$\sum_{k=1}^{T} c_k^2 = \frac{4\epsilon^2\beta^2 T}{(w+1)^2}\left(\sum_{m=0}^{w}|f(m)|\right)^2 \leq 4\epsilon^2\beta^2\frac{T}{(w+1)\sigma^2}||f_w||^2.$$

By Lemma 14,

$$\mathbb{P}(g_1'(e) - \mathbb{E}g_1'(e) > u) \leq \exp\left(\frac{-u^2}{2\epsilon^2\beta^2\frac{T}{(w+1)\sigma^2}(||f_w||)^2}\right).$$

Substituting this into (26) and (24) finishes the proof. □

***Bounding the loss due to prediction error.*** In this section we prove the following concentration result for the loss due to correlated prediction error.

LEMMA 15. *The loss due to prediction error has Berstein type tail: for any $u > 0$,*

$$\mathbb{P}(g_2 > u + V_2) \leq \exp\left(\frac{-u^2}{8\epsilon^2\lambda(\frac{T}{w+1}F(w) + u)}\right). \tag{27}$$

To prove Lemma 15, we characterize $g_2$ as a convex function of $e$ in Lemma 16. We then show that this is a *self-bounding* function. Combining convexity and self-bounding property of $g_2$, Lemma 17 makes use of the convex Log-Sobolev inequality to prove concentration of $g_2$.

LEMMA 16. *The expectation of $g_2$ is $\mathbb{E}g_2 = V_2$, and $g_2$ is a convex quadratic form of $e$. Specifically, there exists a matrix $A \in \mathbb{R}^{T\times T}$, such that $g_2 = \frac{1}{2}||Ae||^2$. Furthermore, the spectral radius of $\lambda$ of $AA^T$ satisfies $\lambda \leq F(w)$.*

Hence, (27) is equivalent to a concentration result of $g_2$:

$$\mathbb{P}(g_2 > V_2 + u) = \mathbb{P}(g_2 - \mathbb{E}g_2 > u).$$

The method of bounded difference used in the previous section is not good for a quadratic function of $e$ because the uniform bound of $|g_2(e) - g_2(e_k')|$ is too large since

$$|g_2(e) - g_2(e_k')| = \frac{1}{2}|(e - e'(k))^T A^T A(e + e'(k))|,$$

where the $(e + e'(k))$ term has $T$ non-zero entries and a uniform upper bound of this will be in $\Omega(T)$. Instead, we use the fact that the quadratic form is self-bounding. Let $h(e) = g_2(e) - \mathbb{E}g_2(e)$. Then

$$\begin{aligned}||\nabla h(e)||^2 &= ||A^T Ae||^2 = (Ae)^T(AA^T)(Ae) \\ &\leq \lambda(Ae)^T(Ae) = 2\lambda[h(e) + \mathbb{E}V_2].\end{aligned}$$

We now introduce the concentration bound for a self-bounding function of a collection of random variables. The proof uses the convex Log-Sobolev inequality [27].

LEMMA 17. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and random variable $X$ be supported on $[-d/2, d/2]^n$. If $\mathbb{E}[f(X)] = 0$ and $f$ satisfies the self-bounding property*

$$||\nabla f||^2 \leq af + b, \tag{28}$$

*for $a, b > 0$, then the tail of $f(X)$ can be bounded as*

$$\mathbb{P}\{f(X) > t\} \leq \exp\left(\frac{-t^2}{d^2(2b + at)}\right). \tag{29}$$

Now to complete the proof of Lemma 15, apply Lemma 17 to the random variable $Z = h(e)$ to obtain

$$\mathbb{P}\{g_2 - \mathbb{E}g_2 > u\} \leq \exp\left(-\frac{u^2}{8\lambda_{\max}\epsilon^2(2V_2 + u)}\right)$$

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

$$\begin{aligned}\mathbb{P}\{g_2 > u + v_2\} &\leq \exp\left(-\frac{u^2}{8\lambda_{\max}\epsilon^2(2V_2 + t)}\right) \\ &= \exp\left(\frac{-u^2}{8\epsilon^2\lambda(\frac{T}{w+1}F(w) + u)}\right).\end{aligned}$$

# 7. CONCLUDING REMARKS

Making use of predictions about the future is a crucial, but under-explored, area of online algorithms. In this paper, we have introduced a general colored noise model for studying predictions. This model captures a range of important phenomena for prediction errors including, general correlation structures, prediction noise that increases with the prediction horizon, and refinement of predictions as time passes. Further it allows for worst-case analysis of online algorithms in the context of stochastic prediction errors.

To illustrate the insights that can be gained from incorporating a general model of prediction noise into online algorithms, we have focused on online optimization problems with switching costs, specifically, an online LASSO formulation. Our results highlight that a simple online algorithm, AFHC, can simultaneously achieve a constant competitive ratio and a sublinear regret in expectation in nearly any situation where it is feasible for an online algorithm to do so. Further, we show that the cost of AFHC is tightly concentrated around its mean.

We view this paper as a first step toward understanding the role of predictions in the design of online optimization

algorithms and, more generally, the design of online algorithms. In particular, while we have focused on a particular, promising algorithm, AFHC, it is quite interesting to ask if it is possible to design online algorithms that outperform AFHC. We have proven that AFHC uses the asymptotically minimal amount of predictions to achieve constant competitive ratio and sublinear regret; however, the cost of other algorithms may be lower if they can use the predictions more efficiently.

In addition to studying the performance of algorithms other than AFHC, it would also be interesting to generalize the prediction model further, e.g., by considering non-stationary processes or heterogeneous $e(t)$.

# 8. REFERENCES


[1] M. A. Adnan, R. Sugihara, and R. K. Gupta. Energy efficient geographical load balancing via dynamic deferral of workload. In *IEEE Int. Conf. Cloud Computing (CLOUD)*, pages 188–195, 2012.

[2] H. Amur, J. Cipar, V. Gupta, G. R. Ganger, M. A. Kozuch, and K. Schwan. Robust and flexible power-proportional storage. In *Proc. ACM Symp. Cloud computing*, pages 217–228, 2010.

[3] G. Ananthanarayanan, A. Ghodsi, S. Shenker, and I. Stoica. Effective straggler mitigation: Attack of the clones. In *Proc. NSDI*, volume 13, pages 185–198, 2013.

[4] L. Andrew, S. Barman, K. Ligett, M. Lin, A. Meyerson, A. Roytman, and A. Wierman. A tale of two metrics: Simultaneous bounds on competitiveness and regret. In *Conf. on Learning Theory (COLT)*, pages 741–763, 2013.

[5] M. F. Arlitt and C. L. Williamson. Web server workload characterization: The search for invariants. In *Proc. ACM SIGMETRICS*, pages 126–137, 1996.

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

[7] A. Blum, H. Karloff, Y. Rabani, and M. Saks. A decomposition theorem and bounds for randomized server problems. In *Proc. Symp. Found. Comp. Sci. (FOCS)*, pages 197–207, Oct 1992.

[8] A. Borodin, N. Linial, and M. E. Saks. An optimal on-line algorithm for metrical task system. *J. ACM*, 39(4):745–763, 1992.

[9] S. Boucheron, G. Lugosi, and O. Bousquet. Concentration inequalities. In *Advanced Lectures on Machine Learning*, pages 208–240. Springer, 2004.

[10] S. Boyd, M. Mueller, B. OÕDonoghue, and Y. Wang. Performance bounds and suboptimal policies for multi-period investment. *Foundations and Trends in Optimization*, 1(1):1–69, 2012.

[11] E. F. Camacho and C. B. Alba. *Model predictive control*. Springer, 2013.

[12] J. Camacho, Y. Zhang, M. Chen, and D. Chiu. Balance your bids before your bits: The economics of geographic load-balancing. *Proc. of ACM e-Energy*, 2014.

[13] E. J. Candès, Y. Plan, et al. Near-ideal model selection by $\ell 1$ minimization. *Ann. Stat.*, 37(5A):2145–2177, 2009.

[14] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[15] M. A. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE J. Sel. Topics Signal Processing*, 1(4):586–597, 2007.

[16] L. Gan, U. Topcu, and S. Low. Optimal decentralized protocol for electric vehicle charging. *IEEE Transactions on Power Systems*, 28(2):940–951, 2013.

[17] C. E. Garcia, D. M. Prett, and M. Morari. Model predictive control: theory and practice—a survey. *Automatica*, 25(3):335–348, 1989.

[18] D. Gmach, J. Rolia, C. Bash, Y. Chen, T. Christian, A. Shah, R. Sharma, and Z. Wang. Capacity planning and power management to exploit sustainable energy. In *Proc. IEEE Int. Conf. Network and Service Management (CNSM)*, pages 96–103, 2010.

[19] D. Gmach, J. Rolia, L. Cherkasova, and A. Kemper. Workload analysis and demand prediction of enterprise data center applications. In *Proc. IEEE Int. Symp. Workload Characterization*, IISWC '07, pages 171–180. IEEE Computer Society, 2007.

[20] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[21] V. Joseph and G. de Veciana. Variability aware network utility maximization. *arXiv preprint arXiv:1111.3728*, 2011.

[22] V. Joseph and G. de Veciana. Jointly optimizing multi-user rate adaptation for video transport over wireless systems: Mean-fairness-variability tradeoffs. In *Proc. IEEE INFOCOM*, pages 567–575, 2012.

[23] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear estimation*. Prentice-Hall, Inc., 2000.

[24] R. E. Kalman. A new approach to linear filtering and prediction problems. *J. Fluids Engineering*, 82(1):35–45, 1960.

[25] S.-J. Kim and G. B. Giannakis. Real-time electricity pricing for demand response using online convex optimization. In *IEEE Innovative Smart Grid Tech. Conf. (ISGT)*, pages 1–5, 2014.

[26] D. Kusic, J. O. Kephart, J. E. Hanson, N. Kandasamy, and G. Jiang. Power and performance management of virtualized computing environments via lookahead control. *Cluster computing*, 12(1):1–15, 2009.

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

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

[29] M. Lin, A. Wierman, L. Andrew, and E. Thereska. Dynamic right-sizing for power-proportional data centers. *IEEE/ACM Trans. Networking*, 21(5):1378–1391, Oct 2013.

[30] Z. Liu, I. Liu, S. Low, and A. Wierman. Pricing data center demand response. In *Proc. ACM Sigmetrics*, 2014.

[31] Z. Liu, A. Wierman, Y. Chen, B. Razon, and N. Chen. Data center demand response: Avoiding the coincident peak via workload shifting and local generation. *Performance Evaluation*, 70(10):770–791, 2013.

[32] L. Lu, J. Tu, C.-K. Chau, M. Chen, and X. Lin. Online energy generation scheduling for microgrids with intermittent energy sources and co-generation. In *Proc. ACM SIGMETRICS*, pages 53–66, 2013.

[33] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. Scokaert. Constrained model predictive control: Stability and optimality. *Automatica*, 36(6):789–814, 2000.

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

[35] B. Narayanaswamy, V. K. Garg, and T. Jayram. Online optimization for the smart (micro) grid. In *Proc. ACM Int. Conf. on Future Energy Systems*, page 19, 2012.

[36] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs. Cutting the electric bill for internet-scale systems. *ACM SIGCOMM Computer Communication Review*, 39(4):123–134, 2009.

[37] M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.*, 18(82):1–9, 2013.

[38] A. P. Sage and J. L. Melsa. Estimation theory with applications to communications and control. Technical report, DTIC Document, 1971.

[39] S. Sastry and M. Bodson. *Adaptive control: stability, convergence and robustness*. Courier Dover Publications, 2011.

[40] E. Thereska, A. Donnelly, and D. Narayanan. Sierra: a power-proportional, distributed storage system. *Microsoft Research, Cambridge, UK, Tech. Rep. MSR-TR-2009-153*, 2009.

[41] R. Tibshirani. Regression shrinkage and selection via the LASSO. *J. Royal Stat. Soc. Ser. B*, pages 267–288, 1996.

[42] H. Wang, J. Huang, X. Lin, and H. Mohsenian-Rad. Exploring smart grid and data center interactions for electric power load balancing. *ACM SIGMETRICS Performance Evaluation Review*, 41(3):89–94, 2014.

[43] X. Wang and M. Chen. Cluster-level feedback power control for performance optimization. In *High Performance Computer Architecture*, pages 101–110. IEEE, 2008.

[44] N. Wiener. *Extrapolation, interpolation, and smoothing of stationary time series*, volume 2. MIT press, Cambridge, MA, 1949.

[45] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Machine Learning Research*, 11:2543–2596, 2010.

[46] F. Zanini, D. Atienza, L. Benini, and G. De Micheli. Multicore thermal management with model predictive control. In *Proc. IEEE. European Conf. Circuit Theory and Design (ECCTD)*, pages 711–714, 2009.

[47] F. Zanini, D. Atienza, G. De Micheli, and S. P. Boyd. Online convex optimization-based algorithm for thermal management of MPSoCs. In *Proc. ACM Great Lakes Symp. VLSI*, pages 203–208, 2010.

[48] X. Y. Zhou and D. Li. Continuous-time mean-variance portfolio selection: A stochastic LQ framework. *Applied Mathematics & Optimization*, 42(1):19–33, 2000.
[49] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proc. Int. Conf. Machine Learning (ICML)*, pages 928–936. AAAI Press, 2003.

# APPENDIX

# A. PROOFS FOR SECTION 3

## A.1 Proof of Theorem 1

For a contradiction, assume that there exists an algorithm $\mathcal{A}'$ that achieves constant competitive ratio *and* sublinear regret with constant lookahead. We can use algorithm $\mathcal{A}'$ to obtain another online algorithm $\mathcal{A}$ that achieves constant competitive ratio *and* sublinear regret *without* lookahead. This contradicts Theorem 4 of [4], and we get the claim.

Consider an instance $\{c_1, c_2, \ldots, c_T\}$ without lookahead. We simply "pad" the input with $\ell$ copies of the zero function $\mathbf{0}$ if $\mathcal{A}'$ has a lookahead of $\ell$. That is the input to $\mathcal{A}'$ is: $c_1, \mathbf{0}, \ldots, \mathbf{0}, c_2, \mathbf{0}, \ldots, \mathbf{0}, c_3, \mathbf{0}, \ldots$

We simulate $\mathcal{A}'$ and set the $t$th action of $\mathcal{A}$ equal to the $((t-1)(\ell+1)+1)$th action of $\mathcal{A}'$. Note that the optimal values of the padded instance are equal to the optimal values of the given instance. Also, by construction, $\text{cost}(\mathcal{A}) \leq \text{cost}(\mathcal{A}')$. Therefore, if $\mathcal{A}'$ achieves constant competitive ratio *and* sublinear regret then so does $\mathcal{A}$, and the claim follows.

# B. PROOFS FOR SECTION 5

## B.1 Proof of Theorem 3

PROOF. Let $(x_{ALG,t})_{t=1}^T$ be the solution produced by online algorithm $ALG$. Then

$$\begin{aligned}\text{cost}(ALG) \geq& \frac{1}{2}\sum_{t=1}^{T} ||y_t - Kx_{ALG,t}||^2 \\ =& \frac{1}{2}\sum_{t=1}^{T} ||(I-KK^\dagger)y_t||^2 + ||KK^\dagger y_t - Kx_{ALG,t}||^2,\end{aligned}$$

by the identity $(I-KK^\dagger)K = 0$. Let $\epsilon_t = x_{ALG,t} - K^\dagger y_{t|t-1}$, i.e., $\epsilon_t = x_{ALG,t} - K^\dagger(y_{t|0} - \sum_{s=1}^{t-1} f(t-s)e(s))$. Since all predictions made up until $t$ can be expressed in terms of $y_{\cdot|0}$ and $e(\tau)$ for $\tau < t$, which are independent of $e(t)$, and all other information (internal randomness) available to ALG is independent of $e(t)$ by assumption, $\epsilon_t$ is independent of $e(t)$. It follows that

$$\begin{aligned}&\mathbb{E}_e[\text{cost}(ALG)] \geq \mathbb{E}_e[||KK^\dagger y_t - K(K^\dagger y_{t|t-1} + \epsilon_t)||^2] \\ =& \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}_{e\setminus e(t)}\mathbb{E}_{e(t)|e\setminus e(t)}||KK^\dagger e(t)^T - K\epsilon_t||^2 \qquad (30)\\ =& \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}_{e\setminus e(t)}(||R_e^{1/2}||^2_{KK^\dagger} + ||(\mathbb{E}_{e(t)}\epsilon_t\epsilon_t^T)^{1/2}||^2_{K^TK}) \\ \geq& \frac{T}{2}||R_e^{1/2}||^2_{KK^\dagger},\end{aligned}$$

where the first equality uses the identity $(I-KK^\dagger)K = 0$, and the second uses the independence of $\epsilon_t$ and $e(t)$. □

## B.2 Proof of Theorem 4

By Theorem 3, if $\mathbb{E}[\text{cost}(ALG)] \in o(T)$, there must be some $t$ such that $\epsilon_t$ is not independent of $e(t)$. By expanding the square term in (30) and noting it is nonnegative:

$$\mathbb{E}[e(t)^T K\epsilon_t] \leq \frac{1}{2}||R_e^{1/2}||^2_{KK^\dagger,F} + \frac{1}{2}||(\mathbb{E}\epsilon_t\epsilon_t^T)^{1/2}||^2_{K^TK,F}.$$

Each nonzero $\mathbb{E}[e(t)^T K\epsilon_t]$ can at most make one term in (30) zero, since there are $T$ terms in (30) that are each lower bounded by $\frac{1}{2}||R_e^{1/2}||^2$, and by assumption $\mathbb{E}[\text{cost}(ALG)]$ is sublinear. There can be at most a sublinear number of $t$ such that $\mathbb{E}[e(t)^T K\epsilon_t] = 0$.

For every other $t$, we must have $\mathbb{E}[e(t)^T K\epsilon_t] > 0$. Let $l_t = \mathbb{E}[e(t)^T K\epsilon_t] > 0$, and $a_t = ||(\mathbb{E}\epsilon_t\epsilon_t^T)^{1/2}||^2_{KK^T,F} > 0$.

Hence, at time $t$, the algorithm can produce prediction $y'_{t|t-1} = y_{t|t-1} + \frac{1}{w_t}K\epsilon_t$, where the coefficient $w_t$ is chosen later. Then the one step prediction error variance:

$$\begin{aligned}\mathbb{E}||y_t - y'_{t|t-1}||^2 &= \mathbb{E}||e(t) - \frac{1}{w_t}K\epsilon_t||^2 \\ &= ||R_e^{1/2}||^2_F - \frac{2}{w_t}l_t + \frac{1}{w_t^2}a_t.\end{aligned}$$

Pick any $w_t > a_t/2l_t$. Then $\mathbb{E}||y_t - y'_{t|t-1}||^2 < ||R_e^{1/2}||^2_F = \mathbb{E}||y_t - y_{t|t-1}||^2$. Hence $ALG$ can produce better one-step prediction for all but sublinearly many $t$.

## B.3 Proof of Lemma 7

To prove Lemma 7, we use the following Lemma.

LEMMA 18. *The competitive difference of $FHC$ with fixed $(w+1)$-lookahead for any realization is given by*

$$\begin{aligned}\text{cost}(FHC^{(k)}) \leq& \text{cost}(OPT) + \sum_{\tau\in\Omega_k} \beta||(x^*_{\tau-1} - x^{(k)}_{\tau-1})||_1 \\ &+ \frac{1}{2}\sum_{\tau\in\Omega_k}\sum_{t=\tau}^{\tau+w} ||KK^\dagger(y_t - y_{t|\tau-1})||^2.\end{aligned}$$

*where $x^*_t$ is the action chosen by the dynamic offline optimal.*

PROOF OF LEMMA 7. Note that $\text{cost}(FHC^{(k)})$ is convex. The result then follows with a straightforward application of Jensen's inequality to Lemma 18. By the definition of $AFHC$, we have the following inequality:

$$\text{cost}(AFHC) \leq \frac{1}{w+1}\sum_{k=0}^{w}\text{cost}(FHC^{(k)})$$

By substituting the expression for $\text{cost}(FHC^{(k)})$ into the equation above and simplifying, we get the desired result. □

Before we prove Lemma 18, we first introduce a new algorithm we term $OPEN$. This algorithm runs an open loop control over the entire time horizon, $T$. Specifically, it chooses actions $x_t$, for $t \in 1, ..., T$, that solves the following optimization problem:

$$\min \frac{1}{2}\sum_{t=1}^{T}(y_{t|0} - Kx_t)^2 + \beta||(x_t - x_{t-1})||_1$$

$FHC^{(k)}$ can be seen as starting at $x^k_{FHC,\tau-1}$, using prediction $y_{\cdot|\tau-1}$, and running $OPEN$ from $\tau$ to $\tau+w$. Then repeating with updated prediction $y_{\cdot|\tau+\omega}$. We first prove the following Lemma characterizing the performance of $OPEN$.

LEMMA 19. *Competitive difference of $OPEN$ over a time horizon, $T$, is given by*

$$\text{cost}(OPEN) - \text{cost}(OPT) \leq \sum_{t=1}^{T}\frac{1}{2}||\hat{y}_t - y_t||^2_{KK^\dagger}$$

PROOF. Recall that the specific OCO we are studying is

$$\min_x \sum_{t=1}^{T}\frac{1}{2}||y_t - Kx_t||^2 + \beta||(x_t - x_{t-1})||_1 \qquad (31)$$

where $x_t \in \mathbb{R}^n$, $y_t \in \mathbb{R}^m$, $K \in \mathbb{R}^{m\times n}$ and the switching cost, $\beta \in \mathbb{R}_+$.

We first derive the dual of (31) by linearizing the $l_1$ norm which leads to the following equivalent expression of the objective above:

$$\min_{x,z} \frac{1}{2}\sum_{t=1}^{T} ||y_t - Kx_t||^2 + \beta \mathbb{1}^T z_t$$
$$\text{s.t.} \quad z_t \geq x_t - x_{t-1}, z_t \geq x_{t-1} - x_t, \quad \forall t.$$

Hence the Lagrangian is

$$L(x,z;\bar{\lambda},\underline{\lambda}) = \frac{1}{2}\sum_{t=1}^{T} ||y_t - Kx_t||^2 + \langle \bar{\lambda}_t - \underline{\lambda}_t, x_t - x_{t-1}\rangle + \langle \beta\mathbb{1} - (\bar{\lambda}_t + \underline{\lambda}_t), z_t\rangle.$$

where we take $\lambda_{T+1} = 0$ and $x_0 = 0$.

Let $\lambda_t = \bar{\lambda}_t - \underline{\lambda}_t$ and $w_t = \bar{\lambda}_t + \underline{\lambda}_t$. Dual feasibility requires $w_t \leq \beta\mathbb{1}, \forall t$, which implies $-\beta\mathbb{1} \leq \lambda_t \leq \beta\mathbb{1}, \forall t$. Dual feasibility also requires $\langle \beta\mathbb{1} - w_t, z_t\rangle = 0, \forall t$.

Now by defining $s_t = \lambda_t - \lambda_{t+1}$ and equating the derivative with respect to $x_t$ to zero, the primal and dual optimal $x_t^*, s_t^*$ must satisfy $K^T K x_t^* = K^T y_t - s_t^*$.

Note by premultiplying the equation above by $x_t^{*T}$, we have $\langle x_t^*, s_t^*\rangle = \langle Kx_t^*, y_t\rangle - ||Kx_t^*||^2$. If instead we premultiply the same equation by $(K^T)^\dagger$, we have after some simplification that $Kx_t^* = (KK^\dagger)y_t - (K^T)^\dagger s_t^*$. We can now simplify the expression for the optimal value of the objective by using the above two equations:

$$\begin{aligned}\text{cost}(OPT) &= \sum_{t=1}^{T} \frac{1}{2}||y_t - Kx_t^*||^2 + \langle x_t^*, s_t^*\rangle \\ &= \sum_{t=1}^{T} \frac{1}{2}||y_t||^2 - \frac{1}{2}||KK^\dagger y_t - (K^T)^\dagger s_t^*||^2 \end{aligned} \tag{32}$$

Observe that (32) implies $s_t^*$ minimizes the following expression $\sum_{t=1}^{T} ||KK^\dagger y_t - (K^T)^\dagger s_t^*||^2$ over the constraint set $\mathcal{S} = \{s_t | s_t = \lambda_t - \lambda_{t+1}, -\beta\mathbb{1} \leq \lambda_t \leq \beta\mathbb{1} \text{ for } 1 \leq t \leq T, \lambda_{T+1} = 0\}$.

$$\begin{aligned}&\text{cost}(OPEN) - \text{cost}(OPT) = p(\hat{x};y) - p(x;y) \\ =&p(\hat{x};\hat{y}) - p(x;y) + p(\hat{x};y) - p(\hat{x};\hat{y}_t) \\ =&\sum_{t=1}^{T} \frac{1}{2}||\hat{y}_t||^2 - \frac{1}{2}||K\hat{x}_t||^2 - \frac{1}{2}||y_t||^2 + \frac{1}{2}||Kx_t^*||^2 \\ &+ \frac{1}{2}||y_t - K\hat{x}_t||^2 - \frac{1}{2}||\hat{y}_t - K\hat{x}_t||^2\end{aligned}$$

Expanding the quadratic terms, using the property of the pseudo-inverse that $K^\dagger K K^\dagger = K^\dagger$, and using the fact that $Kx_t^* = KK^\dagger y_t - (K^T)^\dagger s_t^*$, we have

$$\begin{aligned}&\text{cost}(OPEN) - \text{cost}(OPT) \\ =&\sum_{t=1}^{T} \frac{1}{2}\left(||KK^\dagger y_t - (K^T)^\dagger s_t^*||^2 - ||(KK^\dagger y_t - (K^T)^\dagger \hat{s}_t)||^2\right) \\ &+ \frac{1}{2}\left(||\hat{y}_t - y_t||^2 - ||(I - KK^\dagger)(\hat{y}_t - y_t)||^2\right) \\ \leq&\sum_{t=1}^{T} \frac{1}{2}||\hat{y}_t - y_t||^2 - \frac{1}{2}||(I - KK^\dagger)(\hat{y}_t - y_t)||^2 \\ =&\sum_{t=1}^{T} \frac{1}{2}||KK^\dagger(\hat{y}_t - y_t)||^2.\end{aligned}$$

where the first inequality is because of the characterization of $s_t^*$ following (32). □

Proof of Lemma 18. The proof is a straightforward application of Lemma 19. Summing the cost of $OPEN$ for all $\tau \in \Omega_k$ and noting that the switching cost term satisfying the triangle inequality gives us the desired result. □

## B.4 Proof of Theorem 5

We first define the sub-optimality of the open loop algorithm over expectation of the noise. $\mathbb{E}[||(y_t - \hat{y}_t)||^2_{KK^\dagger}]$ is the expectation of the projection of the prediction error $t+1$ time steps away onto the range space of $K$, given by:

$$\begin{aligned}&\mathbb{E}[||(y_t - \hat{y}_t)||^2_{KK^\dagger}] = \mathbb{E}||\sum_{s=1}^{t} KK^\dagger(f(t-s)e(s))||^2 \\ =&\mathbb{E}[\sum_{s_1=1}^{t}\sum_{s_2=1}^{t} e(s_1)^T f(t-s_1)^T (KK^\dagger)^T (KK^\dagger) f(t-s_2)e(s_2)] \\ =&\text{tr}(\sum_{s_1=1}^{t}\sum_{s_2=1}^{t} f(t-s_1)^T (KK^\dagger)(KK^\dagger) f(t-s_2)\mathbb{E}[e(s_2)e(s_1)^T]) \\ =&\sum_{s=0}^{t-1} \text{tr}(f(s)^T (KK^\dagger) f(s) R_e),\end{aligned}$$

where the last line is because $\mathbb{E}[e(s_1)e(s_2)^T] = 0$ for all $s_1 \neq s_2$, and $KK^\dagger K = K$. Note that this implies $||f_{t-1}||^2 = \sum_{s=0}^{t-1} \text{tr}(f(s)^T f(s) R_e)$. We now write the expected suboptimality of the open loop algorithm as

$$\begin{aligned}&\mathbb{E}[\text{cost}(OPEN) - \text{cost}(OPT)] \leq \sum_{t=1}^{T} \frac{1}{2}\mathbb{E}[||y_t - \hat{y}_t||^2_{KK^\dagger}] \\ &= \frac{1}{2}\sum_{s=0}^{T-1}\sum_{t=s}^{T-1} \text{tr}(f(s)^T KK^\dagger f(s) R_e) \\ &= \frac{1}{2}\sum_{s=0}^{T-1} (T-s)\text{tr}(f(s)^T KK^\dagger f(s) R_e) = F(T-1)\end{aligned}$$

where the first equality is by rearranging the summation.

Now we take expectation of the expression we have in Lemma 7. Taking expectation of the second penalty term (prediction error term), we have:

$$\begin{aligned}&\frac{1}{w+1}\sum_{k=0}^{w}\sum_{\tau\in\Omega_k}\sum_{t=\tau}^{\tau+w} \mathbb{E}\frac{1}{2}||(KK^\dagger)(y_t - y_{t|\tau-1})||^2 \\ =&\frac{1}{2(w+1)}\sum_{k=0}^{w}\sum_{\tau\in\Omega_k} F(w) = \frac{T}{2(w+1)}F(w)\end{aligned}$$

We now need to bound the first penalty term (switching cost term). By taking the subgradient with respect to $x_t$ and by optimality we have $\forall t = 1, \ldots, T$

$$\begin{aligned}&0 \in K^T(Kx_t^* - y_t) + \beta\partial||(x_t^* - x_{t-1}^*)||_1 + \beta\partial||(x_{t+1}^* - x_t^*)||_1 \\ \Rightarrow &x_t^* \in [(K^TK)^{-1}(K^Ty_t - 2\beta\mathbb{1}), (K^TK)^{-1}(K^Ty_t + 2\beta\mathbb{1})]\end{aligned}$$

where the implication is because the sub-gradient of a 1-norm function $||\cdot||_1$ is between $-\mathbb{1}$ to $\mathbb{1}$.

Similarly, since $x_{\tau-1}^{(k)}$ is the last action taken over a $FHC$ horizon, we have that for all $\tau \in \Omega_k$,

$$\begin{aligned}x_{\tau-1}^{(k)} \in [&(K^TK)^{-1}(K^Ty_{\tau-1|\tau-w-2} - \beta\mathbb{1}), \\ &(K^TK)^{-1}(K^Ty_{\tau-1|\tau-w-2} + \beta\mathbb{1})]\end{aligned}$$

Taking expectation of one of the switching cost term and upper bounding with triangle inequality:

$$\begin{aligned}&\mathbb{E}\left|\left|(x^*_{\tau-1}-x^{(k)}_{\tau-1})\right|\right|_1\\ \le&||K^\dagger||_1\mathbb{E}||y_{\tau-1}-y_{\tau-1|\tau-2-w}||_1+3\beta||(K^TK)^{-1}\mathbb{1}||_1\\ \le&||K^\dagger||_1||f_w||+3\beta||(K^TK)^{-1}\mathbb{1}||_1 \qquad (33)\end{aligned}$$

where the first inequality is by the definition of induced norm, the second inequality is due to concavity of the square-root function and Jensen's inequality. Summing (33) over $k$ and $\tau$, we have the expectation of the switching cost term. Adding the expectation of both penalty terms (loss due to prediction error and loss due to switching cost) together, we get the desired result.

## B.5 Proof of Lemma 8

We first characterize $\text{cost}(STA)$:

$$\text{cost}(STA)=\min_x \frac{1}{2}\sum_{t=1}^T||y_t-Kx||_2^2+\beta\mathbb{1}^Tx$$

By first order conditions, we have the optimal static solution $x=K^\dagger\bar{y}-\frac{\beta}{T}(K^TK)^{-1}\mathbb{1}$. Substituting this to $\text{cost}(STA)$ and simplifying, we have:

$$\begin{aligned}cost(STA)=&\frac{1}{2}\sum_{t=1}^T\left(||(I-KK^\dagger)y_t||_2^2+||KK^\dagger(y_t-\bar{y})||_2^2\right)\\ &-\frac{\beta^2\mathbb{1}^T(K^TK)^{-1}\mathbb{1}}{2T}+\beta\mathbb{1}^TK^\dagger\bar{y}\end{aligned}$$

Let $C=\frac{\beta^2\mathbb{1}^T(K^TK)^{-1}\mathbb{1}}{2T}$. Subtracting $\text{cost}(OPT)$ in (32) from the above, we have $\text{cost}(STA)-\text{cost}(OPT)$ equals:

$$\begin{aligned}&\frac{1}{2}\sum_{t=1}^T(||KK^\dagger(y_t-\bar{y})||_2^2-||KK^\dagger y_t||_2^2+||KK^\dagger y_t-(K^T)^\dagger s_t^*||_2^2)\\ &+\beta\mathbb{1}^TK^\dagger\bar{y}-C\\ =&\frac{1}{2}\sum_{t=1}^T\left(||KK^\dagger(y_t-\bar{y})-(K^T)^\dagger s_t^*||_2^2\right)+\langle K^\dagger\bar{y},\beta\mathbb{1}-\lambda_1\rangle-C\\ \ge&\frac{1}{2}\sum_{t=1}^T\left(||KK^\dagger(y_t-\bar{y})-(K^T)^\dagger s_t^*||_2^2\right)-C\end{aligned}$$

The first equality is by expanding the square terms and noting $s_t=\lambda_t-\lambda_{t+1}$. The last inequality is because $-\beta\mathbb{1}\le\lambda_t\le\beta\mathbb{1}$ and $\beta\mathbb{1}^TK^\dagger\bar{y}$ being positive by assumption that the optimal static solution is positive. Now we bound the first term of the inequality above:

$$\begin{aligned}&\frac{1}{2}\sum_{t=1}^T\left(||KK^\dagger(y_t-\bar{y})-(K^T)^\dagger s_t^*||_2^2\right)\\ \ge&\frac{1}{2}\sum_{t=1}^T\left(||KK^\dagger(y_t-\bar{y})||^2\right)-\sum_{t=1}^T\langle KK^\dagger(y_t-\bar{y}),(K^T)^\dagger s_t^*\rangle\\ \ge&\frac{1}{2}\sum_{t=1}^T\left(||KK^\dagger(y_t-\bar{y})||^2\right)-2\beta\sum_{t=1}^T||KK^\dagger(y_t-\bar{y})||\,||(K^T)^\dagger\mathbb{1}||\\ \ge&\frac{1}{2}\sum_{t=1}^T\left(||KK^\dagger(y_t-\bar{y})||^2\right)-2B\sqrt{T\sum_{t=1}^T||(KK^\dagger)(y_t-\bar{y})||^2}\end{aligned}$$

where $B=\beta||(K^T)^\dagger\mathbb{1}||_2$.

By subtracting $C$ from the expression above and completing the sqaure, we have the desired result.

## B.6 Proof of Theorem 9

Using the results of Lemma 8, taking expectation and applying Jensen's inequality, we have:

$$\begin{aligned}&\mathbb{E}_e\left[\text{cost}(STA)-\text{cost}(OPT)\right]\\ \ge&\mathbb{E}_e\big[\frac{1}{2}\sum_{t=1}^T||KK^\dagger(y_t-\bar{y})||^2-2B\sqrt{T\sum_{t=1}^T||(KK^\dagger)(y_t-\bar{y})||^2}-C\big]\\ \ge&\frac{1}{2}\left(\sqrt{\mathbb{E}_e\sum_{t=1}^T\left(||KK^\dagger(y_t-\bar{y})||^2\right)}-2B\sqrt{T}\right)^2-2B^2T-C.\end{aligned}$$

Hence by Theorem 5, the regret of $AFHC$ is

$$\begin{aligned}&\sup_{\hat{y}}\left(\mathbb{E}_e[\text{cost}(AFHC)\text{–cost}(OPT)+\text{cost}(OPT)\text{–cost}(STA)]\right)\\ &\le VT+2B^2T+C-\frac{1}{2}\inf_{\hat{y}}\left(\sqrt{\mathbb{E}_e\sum_{t=1}^T||(y_t-\bar{y})||^2_{KK^\dagger}}-2B\sqrt{T}\right)^2.\end{aligned}$$

Let $S(T)=\mathbb{E}_e\sum_{t=1}^T||y_t-\bar{y}||^2_{KK^\dagger}$. By the above, to prove $AFHC$ has sublinear regret, it is sufficient that

$$VT+2B^2T-\frac{1}{2}\inf_{\hat{y}}(\sqrt{S(T)}-2B\sqrt{T})^2<g(T) \qquad (34)$$

for some sublinear $g(T)$. By the hypothesis of Theorem 9, we have $\inf_{\hat{y}}S(T)\ge(8A+16B^2)T$.

Then, $S(T)\ge(\sqrt{2VT+4B^2T}+2B\sqrt{T})^2$, and (34) holds since $VT+2B^2T-\frac{1}{2}\inf_{\hat{y}}(\sqrt{S(T)}-2B\sqrt{T})^2\le VT+2B^2T-\frac{1}{2}\left(\sqrt{2VT+4B^2T}\right)^2=0$.

# C. PROOFS FOR SECTION 6

## C.1 Proof of Lemma 12

By the triangle inequality, we have

$$\begin{aligned}g_1=&\frac{1}{w+1}\sum_{k=0}^w\sum_{\tau\in\Omega_k}\beta|x^*_{\tau-1}-x^{(k)}_{\tau-1}|\\ \le&\frac{1}{w+1}\sum_{k=0}^w\sum_{\tau\in\Omega_k}\beta\Big(|x^*_{\tau-1}-y_{\tau-1}|+|y_{\tau-1}-y_{\tau-1|\tau-w-2}|\\ &+|y_{\tau-1|\tau-w-2}-x^{(k)}_{\tau-1}|\Big).\end{aligned}$$

By first order optimality condition, we have $x^*_{\tau-1}\in\{y_{\tau-1}-2\beta,y_{\tau-1}+2\beta\}$, and $x^{(k)}_{\tau-1}\in\{y_{\tau-1|\tau-w-2}-\beta,y_{\tau-1|\tau-w-2}+\beta\}$. Hence, by the prediction model,

$$g_1\le\frac{3\beta^2T}{w+1}+\frac{\beta}{w+1}\sum_{k=0}^w\sum_{\tau\in\Omega_k}\left|\sum_{s=1\vee(\tau-w-2)}^{\tau-1}f(\tau-1-s)e(s)\right|$$

## C.2 Proof of Lemma 13

Note that by Lemma 12, we have

$$\begin{aligned}\mathbb{E}g_1'(e)\le&\frac{\beta}{w+1}\sum_{k=0}^w\sum_{\tau\in\Omega_k}\mathbb{E}\left|\sum_{s=1\vee(\tau-w-2)}^{\tau-1}f(\tau-1-s)e(s)\right|\\ \le&\frac{\beta}{w+1}\sum_{k=0}^w\sum_{\tau\in\Omega_k}\sqrt{\sigma^2\sum_{s=0}^w f^2(s)}=\frac{\beta T}{w+1}||f_w||^2.\end{aligned}$$

where the second inequality is by Jensen's inequality and taking expectation.

## C.3 Proof of Lemma 16

By definition of $g_2$ and unraveling the prediction model, we have

$$g_2 = \frac{1}{w+1}\sum_{k=0}^{w}\sum_{\tau\in\Omega_k}\sum_{t=\tau}^{\tau+w}\frac{1}{2}(y_t - y_{t|\tau-1})^2$$
$$= \frac{1}{w+1}\sum_{k=0}^{w}\sum_{\tau\in\Omega_k}\sum_{t=\tau}^{\tau+w}\frac{1}{2}(\sum_{s=\tau}^{t} f(t-s)e(s))^2.$$

Writing it in matrix form, it is not hard to see that

$$g_2 = \frac{1}{w+1}\sum_{k=0}^{w}\frac{1}{2}||A_k e||^2,$$

where $A_k$ has the block diagonal structure given by

$$A_k = \mathrm{diag}(A_k^1, A_k^2, \ldots, A_k^2, A_k^3),^6 \tag{35}$$

and there are the types of submatrices in $A_k$ given by, for $i = 1, 2, 3$:

$$A_k^i = \begin{pmatrix} f(0) & 0 & \ldots & 0 \\ f(1) & f(0) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ f(v_i) & f(v_{i-1}) & \ldots & f(0) \end{pmatrix},$$

where $v_1 = k-2$ if $k \geq 2$ and $v_1 = k+w-1$ otherwise. $v_2 = w$, $v_3 = (T-k+1) \mod (w+1)$. Note that in fact, the matrix $A_k^2$ is the same for all $k$. Hence, we have

$$g_2 = \frac{1}{2}e^T(\frac{1}{w+1}\sum_{k=0}^{w}A_k^T A_k)e = \frac{1}{2}||Ae||^2,$$

where we define $A$ to be such that $A^TA = \frac{1}{w+1}\sum_{k=0}^{w}A_k^T A_k$, this can be done because the right-hand side is positive semidefinite, since $A_k$ is lower triangular. The last equality is because all $A_l^2$ has the same structure. Let $\lambda$ be the maximum eigenvalue of $AA^T$, which can be expressed by

$$\lambda = \max_{||x||=1} x^T AA^T x$$
$$= \frac{1}{w+1}\max_{||x||=1}\sum_{k=0}^{w}x^T A_k A_k^T x \leq \frac{1}{w+1}\sum_{k=0}^{w}\lambda_k,$$

where $\lambda_k$ is the maximum eigenvalue of $A_k A_k^T$. Note that $A_k$ has a block diagonal structure, hence $A_k A_k^T$ also has block diagonal structure, and if we divide the vector $x = (x_1, x_2, \ldots, x_m)$ into sub-vectors of appropriate dimension, then by the block diagonal nature of $A_k A_k^T$, we have

$$x^T A_k A_k^T x = x_1^T A_k^1 {A_k^1}^T x_1 + x_2^T A_k^2 {A_k^2}^T x_2 + \ldots + x_{m-1}^T A_k^2 {A_k^2}^T x_{m-1} + x_m^T A_k^3 {A_k^3}^T x_m.$$

Hence, if we denote the maximum eigenvalues of $\lambda_k^i$ as the maximum eigenvalue of the matrix $A_k^i {A_k^i}^T$, then we have

$$\lambda_k = \max_x \frac{x^T A_k A_k^T x}{x^T x}$$
$$= \max_{x_1,\ldots,x_m} \frac{x_1^T A_k^1 {A_k^1}^T x_1 + x_2^T A_k^2 {A_k^2}^T x_2 + \ldots + x_m^T A_k^3 {A_k^3}^T x_m}{x_1^T x_1 + \ldots + x_m^T x_m}$$
$$\leq \max_{x_1,\ldots,x_m} \frac{\max(\lambda_k^1, \lambda_k^2, \lambda_k^3)\cdot(x_1^T x_1 + \ldots + x_m^T x_m)}{x_1^T x_1 + \ldots + x_m^T x_m}$$
$$\leq \max(\lambda_k^1, \lambda_k^2, \lambda_k^3),$$

[6] The submatrix $A_k^2$ is repeated $\left\lfloor \frac{T-k+1}{\omega+1} \right\rfloor$ times in $A_k$ for $k \geq 2$, and $\left\lfloor \frac{T-k-\omega}{\omega+1} \right\rfloor$ times for otherwise.

where $\lambda_k^i$ is the maximum eigenvalue of $A_k^i$ for $i \in \{1, 2, 3\}$. As $A_k^i {A_k^i}^T$ are all positive semidefinite, we can bound the maximum eigenvalue by trace, and note that $A_k^1$ and $A_k^3$ are submatrix of $A_k^2$, we have

$$\lambda_k \leq \max(\lambda_k^1, \lambda_k^2, \lambda_k^3) \leq \mathrm{tr}(A_k^2 {A_k^2}^T) = \frac{1}{\sigma^2}F(w).$$

## C.4 Proof of Lemma 17

To prove the lemma, we use the following variant of Log-Sobolev inequality

LEMMA 20 (THEOREM 3.2, [27]). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex, and random variable $X$ be supported on $[-d/2, d/2]^n$, then*

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

We will use Lemma 20 to prove Lemma 17. Denote the moment generating function of $f(X)$ by

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

The function $\theta f : \mathbb{R}^n \to \mathbb{R}$ is convex, and therefore it follows from Lemma 20 that

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

By to the self-bounding property (28),

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

Since $m(\theta) > 0$, dividing by $\theta^2 m(\theta)$ gives

$$\frac{d}{d\theta}\left[\left(\frac{1}{\theta} - \frac{ad^2}{2}\right)\ln m(\theta)\right] \leq \frac{bd^2}{2}. \tag{36}$$

Since $m(0) = 1$ and $m'(0) = \mathbb{E}f(X) = 0$, we have

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

and therefore integrating both sides of (36) from 0 to $s$ gives

$$\left(\frac{1}{s} - \frac{ad^2}{2}\right)\ln m(s) \leq \frac{1}{2}bd^2 s, \tag{37}$$

for $s \geq 0$. We can bound the tail probability $\mathbb{P}\{f > t\}$ with the control (37) over the moment generating function $m(s)$.

In particular,

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

for $s \in [0, 2/(ad^2)]$. Choose $s = t/(bd^2 + ad^2 t/2)$ to get

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