Bandit Online Linear Optimization with Hints and Queries

Aditya Bhaskara [1] **Ashok Cutkosky** [2] **Ravi Kumar** [3] **Manish Purohit** [3]

Abstract

We study variants of the online linear optimization (OLO) problem with bandit feedback, where the algorithm has access to external information about the unknown cost vector. Our motivation is the recent body of work on using such "hints" towards improving regret bounds for OLO problems in the full-information setting. Unlike in the full-information OLO setting, with bandit feedback, we first show that one cannot improve the standard regret bounds of $\tilde{O}(\sqrt{T})$ by using hints, even if they are always well-correlated with the cost vector. In contrast, if the algorithm is empowered to issue queries and if all the responses are correct, then we show $O(\log T)$ regret is achievable. We then show how to make this result more robust—when some of the query responses can be adversarial—by using a little feedback on the quality of the responses.

1. Introduction

Online linear optimization (OLO) is an elegant abstraction that captures the essence of many online decision making problems (Zinkevich, 2003; Hazan, 2016). Informally speaking, it is a T-round game between an algorithm and an adversary that is played over a convex domain. In each time step the algorithm first plays a vector after which an adversary replies with a cost vector; the loss at this time step is the inner product of the cost and the played vectors. The algorithm's performance, called *regret*, is measured as the difference between the total loss incurred by the algorithm and that of an algorithm that plays the best single vector in hindsight at all time steps. There are two popular OLO settings: in the *full-information* feedback setting, the algorithm gets to see the cost vector and in the more challenging *bandit* feedback setting, only the loss but not the cost vector is visible to the algorithm. OLO is well-studied in both these settings, and the optimal regret is known to be $\Theta(\sqrt{T})$ for both (Abernethy et al., 2008).

A substantial body of research aims to understand variants of OLO where the $\sqrt{T}$ regret bound can be improved, ideally, to $\log T$. A promising such variant is the use of "hints": before the algorithm plays a vector, it receives a hint. There have been recent results achieving logarithmic regret for OLO, even when each hint is only mildly correlated (i.e., "good") with the yet to appear cost vector (Hazan & Megiddo, 2007; Rakhlin & Sridharan, 2013; Dekel et al., 2017; Bhaskara et al., 2020). These algorithms are also robust, namely, their regret gracefully degrades with the number of time steps in which the hints are "bad" and in the limit when all hints are bad, the regret is $O(\sqrt{T})$. An important detail is that these hint-based algorithms operate in the full-information setting and crucially depend on the availability of the cost vector. This poses a natural question: are there hint-based algorithms for bandit OLO and how much hints can help with reducing the regret in this case?

In this paper we study this question. Our first result is somewhat surprising and strong: we show that having access to good hints is *insufficient* to obtain better than $\tilde{O}(\sqrt{T})$ regret for bandit OLO; furthermore, this negative result holds even in two dimensions when the domain is simply the unit ball! This is in stark contrast with the logarithmic regret that is possible in the analogous full-information setting, dashing any hopes of taking advantage of hints for bandit OLO. The proof is based on constructing a pair of distributions on the plane and arguing that no low-regret algorithm can distinguish them (Section 3).

Necessitated by this lower bound, we turn our attention to a different yet natural way of obtaining hints, namely, answers to queries. In this model, the algorithm can actively query the correlation (inner product) of a point of its choice with the cost vector *before* playing. We present such an algorithm that obtains logarithmic regret even if the query points are chosen at random. The main intuition behind the result is that a good response to a random query can be used to provide an unbiased estimate of the cost vector; in addition it can also be used to construct a proxy hint for the algorithm.

[*]Equal contribution [1]University of Utah, Salt Lake City, UT, USA [2]Boston University, Boston, MA, USA [3]Google Research, Mountain View, CA, USA. Correspondence to: Aditya Bhaskara <bhaskaraaditya@gmail.com>, Ashok Cutkosky <ashok@cutkosky.com>, Ravi Kumar <ravi.k53@gmail.com>, Manish Purohit <mpurohit@google.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

We also show that the algorithm is robust, i.e., the regret gracefully degrades if some of the responses to queries are allowed to be bad/incorrect; however, the degradation is linear in the number of bad responses (Section 4).

To improve the robustness so that the regret bounds would still be $O(\sqrt{T})$ even when all the query responses are bad, we aid the algorithm with additional binary feedback about the "goodness" of the response *after* playing. Exploiting this information is challenging because exploring to hedge against a potentially "bad" response is in tension with exploiting a potentially "good" response. This exploration/exploitation tradeoff is more fraught than the standard one encountered in bandits as we need more "exploitation" in order to achieve logarithmic rather than $\sqrt{T}$ regret. Nevertheless, in this enhanced model, we show that we can recover the optimal robustness bounds (Section 5).

1.1. Related work

Our work connects naturally with the recent literature on algorithms that can leverage ML-based predictions (e.g., (Lykouris & Vassilvtiskii, 2018; Kumar et al., 2018)). Much of this line of work assumes that an ML "oracle" makes problem-specific predictions, that are used by an algorithm for obtaining better guarantees, especially for combinatorial optimization (Gollapudi & Panigrahi, 2019; Jiang et al., 2020; Rohatgi, 2020; Bamas et al., 2020; Im et al., 2021; Mitzenmacher, 2020; Kumar et al., 2019; Lavastida et al., 2021). In the online learning community, this setting has been studied under the name of *optimistic* regret bounds (Rakhlin & Sridharan, 2013; Steinhardt & Liang, 2014; Dekel et al., 2017; Wei & Luo, 2018; Bhaskara et al., 2020). Our query model is different in that the algorithm *interacts* with the oracle and is thus able to obtain better regret. Recently, Bhaskara et al. (2023) introduced a query model similar to ours, but could only obtain guarantees in the full-information setting, or in the case of stochastic multi-armed bandits.

Our model is also related to the "observe before play" model introduced in Zuo et al. (2019). However, the key difference is that they consider policy regret, i.e., they compete against policies that also make observations before playing an arm, while our work competes against the more classic benchmark of the best fixed action in hindsight, as in the work on optimistic regret bounds. Another difference is that our focus is on adversarial bandits, while Zuo et al. (2019) primarily study the stochastic case.

One of the challenges we face in Sections 4 and 5 concerns dealing with incorrect query responses. This is a well-known challenge for learning with bandit feedback, as a small number of incorrect responses can throw off the estimates that an algorithm uses to maintain information about the arms. Recent work such as (Lykouris et al., 2018; Gupta et al., 2019; Ito, 2021; Wei et al., 2020) develop different techniques to handle this issue. It is an interesting question to see if such ideas can let us handle incorrect responses without receiving feedback as in Section 5.

2. Formulation

Let $[T] = \{1, \ldots, T\}$. Let $\mathbb{B}^d = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$ denote the unit Euclidean ball in d dimensions. Let $\vec{c} = c_1, \ldots, c_T$ denote the sequence of cost vectors, where each $c_i \in \mathbb{B}^d$.

The online linear optimization (OLO) problem with limited feedback, aka, the *bandit OLO* setting, is a game between an algorithm $\mathcal{A}$ and an adversary over T rounds. In each time step $t \in [T]$, an adversary chooses the cost vector $c_t \in \mathbb{B}^d$; this cost vector is *not* revealed to the algorithm. The algorithm *plays* a point $x_t \in \mathbb{B}^d$ and receives feedback $\langle c_t, x_t \rangle$; it is said to incur a *loss* of $\ell_t = \langle c_t, x_t \rangle$ in this time step. The total loss of the algorithm $\mathcal{A}$ is defined as $\text{loss}_{\mathcal{A}}(\vec{c}) = \sum_{t \in [T]} \ell_t$.

The *regret* of the algorithm $\mathcal{A}$ is the difference between its total loss and that of the best algorithm that is constrained to play the same point in $\mathbb{B}^d$ at all time steps:

$$\mathcal{R}_{\mathcal{A}}(\vec{c}) = \text{loss}_{\mathcal{A}}(\vec{c}) - \min_{x \in \mathbb{B}^d} \sum_{t=1}^{T} \langle c_t, x \rangle.$$

The goal is to design an algorithm with minimum regret.

In this paper we consider the following variants of the bandit OLO setting.

(i) **Hints.** Before playing x_t, the algorithm receives a *hint* $h_t \in \mathbb{B}^d$. A hint is said to be *good* if $\langle h_t, c_t \rangle \geq \alpha$ and *bad* otherwise; here α is a fixed parameter.

(ii) **Queries.** Before playing x_t, the algorithm can query an arbitrary point $s_t \in \mathbb{B}^d$ and receive a response q_t. A response is said to be *good* if $q_t = \langle c_t, s_t \rangle$ and *bad* otherwise.

(iii) **Response Feedback.** The setting is same as (ii)—before playing x_t, the algorithm gets a response q_t to a query s_t. In addition, after it plays x_t, it receives feedback g_t, which is 1 if the response q_t was good and is 0 otherwise.

2.1. Notation

For $a > 0$, we define

$$\text{clip}_a(x) = \max(-a, \min(a, x)).$$

We also let $c_{1:t}$ denote $\sum_{i=1}^{t} c_i$ and let $\|c\|^2_{1:t}$ denote $\sum_{i=1}^{t} \|c_i\|^2$.

For a distribution D, we let $x \sim D$ to denote that the random variable x is drawn from D. For two distributions

P, Q, let $d_{\mathrm{tv}}(P;Q)$ denote the total variation distance and let $d_{\mathrm{KL}}(P;Q)$ denote the KL-divergence. Pinsker's inequality states that $d_{\mathrm{tv}}(P;Q) \le \sqrt{d_{\mathrm{KL}}(P;Q)/2}$. Let $\mathrm{N}(\mu, \sigma^2)$ denote the one-dimensional Gaussian distribution with mean μ and variance σ^2.

3. Limitations of Hints

In this section we show it is not possible to reduce the regret below $\tilde{O}(\sqrt{T})$ for bandit OLO optimization, even if good hints are available to the algorithm at every step.

Theorem 3.1. *For any bandit online learning algorithm $\mathcal{A}$, there is a distribution over a sequence $c_1, \ldots, c_T \in \mathbb{B}^d$ of cost vectors and a sequence $h_1, \ldots, h_T \in \mathbb{B}^d$ of hint vectors such that the following hold:*

1. *$\langle h_t, c_t \rangle \ge 1/4$ for all $t \in [T]$ (with high probability), and*
2. *$1/4 \le \|c_t\| \le 1$ and $\|h_t\| = 1$ for all $t \in [T]$ (with high probability), and*
3. *expected regret of $\mathcal{A}$ is $\Omega(\sqrt{\frac{T}{\log T}})$.*

The proof follows the general template of lower bounds in bandit settings, and proceeds by constructing two distributions and arguing about whether an algorithm can distinguish them or not, and proving a high regret bound in either case. Our lower bound construction will only require two dimensions, and so the theorem holds even for $d = 2$.

Define the following two distributions over cost vectors. Let ϵ, σ be parameters that will be chosen later. We define D_+ to be the distribution over $\mathbb{R}^2$ where a random point is generated as $(\mathrm{N}(\frac{1}{2}, \sigma^2), \mathrm{N}(+\epsilon, \sigma^2))$. Similarly, we define D_- to be the distribution where a random point in $\mathbb{R}^2$ is generated as $(\mathrm{N}(\frac{1}{2}, \sigma^2), \mathrm{N}(-\epsilon, \sigma^2))$.

Our construction will use $\epsilon < 1/4$, and $\sigma^2 \le \frac{1}{100 \log T}$. This will ensure that $c_t \sim D_+$ (or $\sim D_-$) satisfies $\frac{1}{4} \le \|c_t\| \le 1$ for all t, with probability $\ge 1 - T^{-4}$. We also set $h_t = (1, 0)$ for all t. Once again, the choice of σ, ϵ will ensure that $\langle c_t, h_t \rangle \ge 1/4$ for all t, with probability $\ge 1 - T^{-4}$.

We now provide an outline of the proof. Consider any (possibly randomized) algorithm that plays the point $x_t = (a_t, b_t)$ and observes loss ℓ_t at time t. We consider the distribution of the losses $(\ell_1, \ldots, \ell_T)$ under the input distributions D_+ and D_-. We argue that if the loss distributions in the two cases are close, then the algorithm must be playing "similar points" in the two cases, and must therefore incur high regret in one of the two cases. Otherwise, we show that the algorithm must place a significant mass on b_t (in magnitude) on average, which then leads to a high regret for one of the two distributions. We now formalize this argument.

Proof of Theorem 3.1. Let $\mathcal{A}$ be a (possibly randomized) algorithm that is constrained to play a point $x_t = (a_t, b_t) \in \mathbb{B}^2$ at every time step. Let ℓ_t denote the loss that $\mathcal{A}$ incurs at time t. Note that this is a random variable that depends on the costs at time $\le t$, as well as the randomness in $\mathcal{A}$. If the c_t's were drawn from D_+, then we can write

$$\ell_t = \frac{a_t}{2} + \epsilon b_t + \alpha_t a_t + \beta_t b_t, \tag{1}$$

where $\alpha_t, \beta_t \sim \mathrm{N}(0, \sigma^2)$. We let P be the joint distribution of $(\ell_1, \ldots, \ell_T)$ in the case where $c_t \sim D_+$. Similarly, define Q to be the joint distribution in the case where $c_t \sim D_-$. We consider two cases.

Case 1. $d_{\mathrm{tv}}(P;Q) \le 1/3$. Intuitively in this case, $\mathcal{A}$ cannot distinguish between whether $c_t \sim D_+$ or $c_t \sim D_-$. We now argue that in *one* of the cases, $\mathcal{A}$ needs to incur a large regret. Let I_t be the binary variable that indicates if $b_t > 0$. For convenience, let us write $\mathbb{E}_+[Z]$ for a random variable Z to denote the expected value when $c_t \sim D_+$ and write $\mathbb{E}_-[Z]$ when $c_t \sim D_-$.

The first observation is that for every t, $|\mathbb{E}_+[I_t] - \mathbb{E}_-[I_t]| \le 1/3$. This is true by the assumption $d_{\mathrm{tv}}(P;Q) \le 1/3$, and by using the fact that the points played by $\mathcal{A}$ (and hence I_t) only depend on the losses observed by the algorithm so far. Thus, if we define $N = \sum_t I_t$, we have $|\mathbb{E}_+[N] - \mathbb{E}_-[N]| \le T/3$. Thus, we must either have $\mathbb{E}_+[N] \ge T/3$ or $\mathbb{E}_-[N] \le 2T/3$ (because of neither holds, the difference will be $> T/3$).

Assume first that $\mathbb{E}_+[N] \ge T/3$. In this case, we have $\sum_t \mathbb{E}[\ell_t] = \sum_t \mathbb{E}[\frac{a_t}{2} + \epsilon b_t]$ (this is because a_t, b_t only depend on the losses and costs at time steps $< t$). Since $(a_t, b_t) \in \mathbb{B}^2$, we always (i.e., regardless of I_t) have

$$\frac{a_t}{2} + \epsilon b_t \ge -\sqrt{\frac{1}{4} + \epsilon^2},$$

and furthermore, if $I_t = 1$, we get a stronger bound of

$$\frac{a_t}{2} + \epsilon b_t \ge -\frac{1}{2} \ge -\sqrt{\frac{1}{4} + \epsilon^2} + \frac{\epsilon^2}{2},$$

where the second inequality is from Taylor expansion. Next, note that if $\mathbb{E}_+[N] \ge T/3$, we must have $I_t = 1$ for at least $T/3$ steps, and thus

$$\sum_t \mathbb{E}[\ell_t] \ge -T\sqrt{\frac{1}{4} + \epsilon^2} + \frac{T\epsilon^2}{6}.$$

However, the optimal comparator in hindsight incurs an expected loss of $-\mathbb{E}[|\sum_{t=1}^T c_t|] \le -|\mathbb{E}\sum_{t=1}^T c_t| = -T\sqrt{\frac{1}{4} + \epsilon^2}$ by Jensen's inequality. This shows that the expected regret is $\Omega(T\epsilon^2)$. The proof for the case $\mathbb{E}_-[N] \le 2T/3$ is similar (and here, the argument implies that the regret is high for $c_t \sim D_-$). Together, this completes the proof for Case 1.

Case 2. $d_{\mathrm{tv}}(P;Q) > 1/3$. By Pinsker's inequality, note that we must also have $d_{\mathrm{KL}}(P;Q) > 2/9$.

Let us now obtain a bound on $d_{\mathrm{KL}}(P;Q)$ using the chain rule for KL-divergence. For convenience, denote $\ell_1^t = (\ell_1, \ldots, \ell_t)$. Then we have

$$d_{\mathrm{KL}}(P;Q) \le \sum_t \mathbb{E}_{\ell_1^{t-1}} d_{\mathrm{KL}}(\ell_t \mid D_+, \ell_1^{t-1}; \ell_t \mid D_-, \ell_1^{t-1}).$$

Here, $\ell_t \mid D_+, \ell_1^{t-1}$ refers to the distribution of ℓ_t when c_t is drawn from D_+ and we condition over the losses in the first $t-1$ steps. Also note that the expectation over ℓ_1^{t-1} corresponds to $c_t \sim D_+$; this is due to the asymmetry in the definition of d_{KL}. Since ℓ_1^{t-1} determine the values of a_t, b_t, we can use the expression in (1) (and the analogous expression for the case of D_-) to obtain:

$$d_{\mathrm{KL}}(\ell_t \mid D_+, \ell_1^{t-1}; \ell_t \mid D_-, \ell_1^{t-1}) = \frac{4\epsilon^2 b_t^2}{\sigma^2(a_t^2 + b_t^2)}.$$

(We are using the standard formula for the KL-divergence between univariate Gaussians (Tsybakov, 2009).)

Whenever $a_t^2 + b_t^2 \ge 1/2$, this quantity is $\le \frac{8\epsilon^2 b_t^2}{\sigma^2}$, and further, it is *always* $\le \frac{4\epsilon^2}{\sigma^2}$. So, if we denote by J_t the binary variable that indicates if $a_t^2 + b_t^2 < 1/2$, we have that

$$\frac{4\epsilon^2 b_t^2}{\sigma^2(a_t^2 + b_t^2)} \le \frac{4\epsilon^2}{\sigma^2} J_t + \frac{8\epsilon^2 b_t^2}{\sigma^2}.$$

Let us write $M = \sum_t J_t$. Then, the bound $d_{\mathrm{KL}}(P;Q) > 2/9$ implies that

$$\mathbb{E}_+[M] + 2\sum_t \mathbb{E}_+[b_t^2] \ge \frac{\sigma^2}{18\epsilon^2}.$$

Thus, one of the terms on the LHS must be $\ge \frac{\sigma^2}{36\epsilon^2}$, and we consider two sub-cases. For both the cases, we choose the parameters σ, ϵ such that $\frac{\sigma^2}{36\epsilon^2} \ge 40T\epsilon^2$.

Case 2a. $\mathbb{E}_+[M] \ge \frac{\sigma^2}{36\epsilon^2}$. In this case, it is easy to see that whenever $J_t = 1$,

$$\frac{a_t}{2} + \epsilon b_t \ge -\frac{1}{\sqrt{2}}\sqrt{\frac{1}{4} + \epsilon^2} > -\sqrt{\frac{1}{4} + \epsilon^2} + \frac{1}{3}.$$

This implies that the total regret in this case is $\ge 40T\epsilon^2/3$.

Case 2b. $\mathbb{E}_+[\sum_t b_t^2] \ge \frac{\sigma^2}{72\epsilon^2} > 20T\epsilon^2$.

In this case, the idea is to argue that b_t^2 is "too large" on average, and use this to conclude that we have high regret. Let (u_1, u_2) be the unit vector along $(-1/2, -\epsilon)$. Thus, to minimize $\frac{a_t}{2} + \epsilon b_t$, we must have $b_t = u_2$. We start with the following easy claim that quantifies the regret when $b_t \ne u_2$.

Claim. Suppose $v = (x, y)$, and $x^2 + y^2 \le 1$. Then,

$$\frac{x}{2} + \epsilon y \ge -\sqrt{\frac{1}{4} + \epsilon^2} + \frac{(y - u_2)^2}{4}.$$

To prove the claim, note that for any given y, in order to minimize the LHS, x must be made as small as possible, i.e., we can set $x = -\sqrt{1 - y^2}$, and thus we may assume that v is a unit vector. Then, if we denote $u = (u_1, u_2)$ to be the unit vector along $(-1/2, -\epsilon)$, we have (since both are unit vectors),

$$1 - \langle u, v \rangle = \frac{\|u - v\|^2}{2} \ge \frac{(y - u_2)^2}{2}.$$

Thus, we have

$$\sqrt{\frac{1}{4} + \epsilon^2} + \left(\frac{x}{2} + \epsilon y\right) \ge \sqrt{\frac{1}{4} + \epsilon^2} \cdot \frac{(y - u_2)^2}{2},$$

completing the proof of the claim. □

Using this for all t, we have that the expected regret is at least $\mathbb{E}_+\left[\sum_t \frac{(b_t - u_2)^2}{4}\right]$. To bound this, we note that

$$(b_t - u_2)^2 \ge \begin{cases} 0 & \text{if } b_t^2 \le 4u_2^2 \\ b_t^2/4 & \text{otherwise.} \end{cases}$$

Thus, the regret can be lower bounded by $\sum_{t: b_t^2 \ge 4u_2^2} \frac{b_t^2}{16}$. Finally, since

$$\sum_{t: b_t^2 < 4u_2^2} b_t^2 + \sum_{t: b_t^2 \ge 4u_2^2} b_t^2 \ge 20\epsilon^2 T,$$

and the first sum is at most $4u_2^2 T = \frac{4\epsilon^2 T}{\frac{1}{4} + \epsilon^2} \le 16\epsilon^2 T$, the second term must be $\ge 4\epsilon^2 T$, and thus the regret is $\ge \frac{\epsilon^2 T}{4}$.

This completes the proof of Case 2, and hence also the proof of the theorem.

Choice of parameters. Note that in order for all the inequalities needed in the proof to hold, we can set $\sigma^2 = \frac{1}{100 \log T}$ and $\epsilon^2 = \frac{\sigma}{50\sqrt{T}}$. With this setting, we get a regret lower bound of $T\epsilon^2/4$, i.e., $\Omega\left(\sqrt{\frac{T}{\log T}}\right)$. □

4. Sublinear Regret with Queries

Theorem 3.1 shows that, unlike the full-information setting (Bhaskara et al., 2020; 2021), *passive* hints regarding upcoming cost vectors are not sufficient to obtain logarithmic regret guarantees in the bandit setting. Motivated by recent work (Bhaskara et al., 2023), we now consider a setup where the algorithm can actively query the value of the cost function at a point of its choice before playing. Our main result is the following:

Algorithm 1 Bandit OLO with Queries.

for $t = 1, \dots, T$ **do**
  $s_t \leftarrow$ uniform random vector on unit sphere in $\mathbb{R}^d$
  $q_t \leftarrow \mathcal{Q}(s_t)$ ▷ Query and response
  $\hat{c}_t \leftarrow d \cdot q_t \cdot s_t$ ▷ Estimate cost
  $\alpha_t \leftarrow \mathrm{clip}_{\frac{4}{\sqrt{d}}}(q_t)$
  $h_t \leftarrow \frac{\sqrt{d}}{4} \cdot \alpha_t \cdot s_t$ ▷ Construct hint
  $\bar{x}_t \leftarrow \mathrm{argmin}_{x \in \mathbb{B}^d} \tilde{\ell}_{1:t-1}(x) + \|x\|^2$ ▷ FTRL step
  Play $x_t = \bar{x}_t + \dfrac{\|\bar{x}_t\|^2 - 1}{2} \cdot h_t$
  Incur loss $\langle c_t, x_t \rangle$
  Define $\tilde{\ell}_t(\cdot) := \langle \hat{c}_t, \cdot \rangle + \frac{\sqrt{d}}{4} \cdot \dfrac{q_t \alpha_t}{2}(\|\cdot\|^2 - 1)$
end for

Theorem 4.1. *For the bandit OLO problem with queries, Algorithm 1 obtains expected regret $O(d^{3/2} \log T + d^2 \log(B+1) + dB)$, where B is the number of bad responses.*

In particular, when the responses to all queries are good (i.e., $B = 0$), the regret is $O(d^{3/2} \log T)$. We also remark that while our algorithm assumes that the query responses are perfectly accurate (in all but B steps), our methods can also be used in weaker settings. For example, if we know that $\|c_t\| = 1$ for all t, then simply receiving the *sign* of $\langle c_t, s_t \rangle$ suffices. This is because the sign suffices to obtain the guarantees of Lemma 4.2, and thereby the desired regret bounds. We omit these details for brevity.

4.1. Algorithm

The algorithm exploits the simple fact that a random query can be used to get a low variance estimate of the cost vector as well as to construct a good hint. The details are presented in Algorithm 1. The random query s_t and the response q_t to this query are used to obtain an unbiased estimate $\hat{c}_t$ of the cost vector. They are also used to construct a good hint h_t. Note that we also need to "clip" q_t for constructing the hint; this is important for achieving the claimed regret bound but makes the analysis tricky. The hint h_t is then leveraged to construct a strongly-convex surrogate loss function, as in prior works (Bhaskara et al., 2020).

4.2. Analysis

We first establish some simple and useful properties of the estimated cost vector and the constructed hint. Let $\mathbb{E}_{t-1}[\cdot]$ denote the expected value conditioned on the history until time $t-1$. The following lemma relies on properties of a point sampled uniformly from the unit sphere and we defer the proof to Appendix A.

Lemma 4.2. *In Algorithm 1, the following hold: (i) $\|h_t\| \le 1$ and $\|x_t\| \le 1$. If the response is good at time t, then*

(ii) $\mathbb{E}[\hat{c}_t] = c_t$,
(iii) $\mathbb{E}_{t-1}[d\alpha_t^2] \ge (1/4)\|c_t\|^2$,
(iv) $\mathbb{E}[\|\hat{c}_t\|^2] = d\|c_t\|^2$.

We next bound the regret incurred by Algorithm 1 by the regret incurred by the FTRL procedure against the constructed surrogate loss functions, during time steps when the query response is good.

Lemma 4.3. *If the response q_t at time t is good, then*

$$\mathbb{E}[\langle c_t, x_t - u \rangle] \le \mathbb{E}[\tilde{\ell}_t(\bar{x}_t) - \tilde{\ell}_t(u)],\ \forall u \in \mathbb{B}^d.$$

Proof.

$$\mathbb{E}[\tilde{\ell}_t(\bar{x}_t)] = \mathbb{E}\left[\langle \hat{c}_t, \bar{x}_t \rangle + \frac{\sqrt{d}}{4} \cdot \frac{\langle c_t, s_t \rangle \alpha_t}{2}(\|\bar{x}_t\|^2 - 1)\right].$$

When the response q_t is good, from Lemma 4.2(ii), we have $\mathbb{E}[\hat{c}_t] = c_t$. Since $\hat{c}_t$ and $\bar{x}_t$ are independent, the expectation of the first term is exactly $\langle c_t, \bar{x}_t \rangle$. This implies:

$$\mathbb{E}[\tilde{\ell}_t(\bar{x}_t)] = \mathbb{E}[\langle c_t, x_t \rangle].$$

Next, note that for any $\|u\| \le 1$,

$$\begin{aligned}\mathbb{E}[\tilde{\ell}_t(u)] &= \mathbb{E}\left[\langle \hat{c}_t, u \rangle + \frac{\langle c_t, s_t \rangle \alpha_t}{2}(\|u\|^2 - 1)\right] \\ &= \langle c_t, u \rangle + \mathbb{E}\left[\frac{\langle c_t, s_t \rangle \alpha_t}{2}(\|u\|^2 - 1)\right] < \langle c_t, u \rangle,\end{aligned}$$

once again because $\langle c_t, s_t \rangle \alpha_t \ge 0$ if $q_t = \langle c_t, s_t \rangle$. Putting these together completes the proof of the claim. □

So, now all we need to show is that the FTRL procedure obtains low regret on the surrogate losses $\tilde{\ell}_t(\cdot)$.

Lemma 4.4. $\mathbb{E}[\sum_{t=1}^T \tilde{\ell}_t(\bar{x}_t) - \tilde{\ell}_t(u)] \le O(d^{3/2} \log T + d^2 \log(B+1))$ *where B denotes the number of time steps when the response is bad.*

Proof. For convenience, let $\sigma_t := \frac{\sqrt{d}}{4} q_t \alpha_t$. Then by definition, $\tilde{\ell}_t$ is σ_t-strongly convex wrt norm $\|\cdot\|$. Consider the regularizer $r(x) = \|x\|^2$.

We have $\tilde{\ell}_{1:t+1} + r$ is $(\sigma_{1:t+1} + 2)$-strongly convex wrt norm $\|\cdot\|$. Equivalently $\tilde{\ell}_{1:t+1} + r$ is 1-strongly convex wrt norm $\|\cdot\|_{(t)} := \sqrt{\sigma_{1:t+1} + 2}\,\|\cdot\|$. Let $\|\cdot\|_{(t),\star} = (\frac{1}{\sqrt{\sigma_{1:t+1}+2}})\cdot\|\cdot\|$ be the corresponding dual norm.

Let $\tilde{g}_t$ denote the gradient of $\tilde{\ell}_t$. Then applying (McMahan, 2017, Theorem 1), we get

$$\mathbb{E}\left[\sum_{t=1}^T \tilde{\ell}_t(\bar{x}_t) - \tilde{\ell}_t(u)\right] \le \mathbb{E}\left[r(u) + \frac{1}{2}\sum_{t=1}^T \|\tilde{g}_t\|^2_{(t-1),\star}\right]$$

$$= \mathbb{E}\left[r(u) + \frac{1}{2}\sum_{t=1}^{T}\frac{\|\tilde{g}_t\|^2}{2+\sigma_{1:t}}\right]$$
$$\le \mathbb{E}\left[r(u) + \frac{1}{2}\sum_{t=1}^{T}\frac{\|\tilde{g}_t\|^2}{2+\sigma_{1:t-1}}\right]$$
$$= \mathbb{E}\left[r(u) + \frac{1}{2}\sum_{t\in I}\frac{\|\tilde{g}_t\|^2}{2+\sigma_{1:t-1}} + \frac{1}{2}\sum_{t\notin I}\frac{\|\tilde{g}_t\|^2}{2+\sigma_{1:t-1}}\right],$$

where $I \subseteq T$ denotes the set of times at which the responses are bad.

The first term is simply $\|u\|^2$. For the second term, we note that even for $t \in I$:

$$\|\tilde{g}_t\|^2 \le 2\|\hat{c}_t\|^2 + 2\sigma_t^2 \le 2d^2q_t^2 + 2\sigma_t \le 2(d^2+1)\sigma_t,$$

where the last inequality uses $|q_t| \le \frac{\sqrt{d}}{4}|\alpha_t|$. Substituting, we have the following

$$\mathbb{E}\left[\frac{1}{2}\sum_{t\in I}\frac{\|\tilde{g}_t\|^2}{2+\sigma_{1:t-1}}\right] \le \mathbb{E}\left[\frac{1}{2}\sum_{t\in I}\frac{\|\tilde{g}_t\|^2}{1+\sum_{s\le t|s\in I}\sigma_s}\right]$$
$$\le \mathbb{E}\left[\frac{1}{2}\sum_{t\in I}\frac{2(d^2+1)\sigma_t}{1+\sum_{s\le t|s\in I}\sigma_s}\right]$$
$$\le (d^2+1)\mathbb{E}[\log(1+\sum_{t\in I}\sigma_t)] \le (d^2+1)\log(B+1).$$

Finally for the third term, we have:

$$\mathbb{E}\left[\sum_{t\notin I}\frac{\|\tilde{g}_t\|^2}{2+\sum_{s<t|s\notin I}\sigma_s}\right] = \mathbb{E}^{t-1}\left[\sum_{t\notin I}\frac{\mathbb{E}_{t-1}[\|\tilde{g}_t\|^2]}{2+\sum_{s<t|s\notin I}\sigma_s}\right]$$

where $\mathbb{E}_{t-1}$ indicates the expectation *given* the history $x_1,\dots,x_{t-1}, q_1,\dots,q_{t-1}$ and $\mathbb{E}^{t-1}$ indicates expectation over *only* the history $x_1,\dots,x_{t-1}, q_1,\dots,q_{t-1}$ so that by the tower rule, $\mathbb{E} = \mathbb{E}^{t-1}\mathbb{E}_{t-1}$. The equality follows from the tower rule of expectation and the fact that $\sum_{s<t|s\notin I}\sigma_s$ is deterministic given the history $x_1,\dots,x_{t-1}$, so that $\frac{\mathbb{E}_{t-1}[\|\tilde{g}_t\|^2]}{2+\sum_{s<t|s\notin I}\sigma_s} = \mathbb{E}_{t-1}\left[\frac{\|\tilde{g}_t\|^2}{2+\sum_{s<t|s\notin I}\sigma_s}\right]$. Now, to bound $\mathbb{E}_{t-1}[\|\tilde{g}_t\|^2]$, we utilize the fact that all $t \notin I$ we have the following:

$$\|\tilde{g}_t\| \le \|\hat{c}_t\| + \frac{\sqrt{d}}{4}q_t\alpha_t, \text{ yielding}$$
$$\mathbb{E}_{t-1}[\|\tilde{g}_t\|^2] \le 2\mathbb{E}_{t-1}[\|\hat{c}_t\|^2] + \frac{d}{8}\mathbb{E}_{t-1}\left[q_t^2\alpha_t^2\right].$$

By Lemma 4.2 (parts (iii) and (iv)), we have $\mathbb{E}_{t-1}[\|\hat{c}_t\|^2] = d\|c_t\|^2 \le 4d^2\mathbb{E}_{t-1}[\alpha_t^2]$. Also, since $q_t^2 \le 1$, the second term is at most $\frac{d}{8}\mathbb{E}_{t-1}[\alpha_t^2]$. Thus,

$$\mathbb{E}_{t-1}[\|\tilde{g}_t\|^2] \le 8d^2\mathbb{E}_{t-1}[\alpha_t^2] + \frac{d}{8}\mathbb{E}_{t-1}[\alpha_t^2] \le 5d^2\mathbb{E}_{t-1}[\alpha_t^2]$$
$$= 9d^2\mathbb{E}_{t-1}\left[\alpha_t \cdot \frac{4\sigma_t}{\sqrt{d}q_t}\right] \le 36d^{3/2}\mathbb{E}_{t-1}[\sigma_t].$$

Hence we get,

$$\mathbb{E}\left[\sum_{t=1}^{T}\tilde{\ell}_t(\bar{x}_t) - \tilde{\ell}_t(u)\right] \le 1 + (d^2+1)\log(B+1) + \frac{1}{2}\mathbb{E}^{t-1}\left[\sum_{t\notin I}\frac{36d^{3/2}\mathbb{E}_{t-1}[\sigma_t]}{2+\sum_{s<t|s\notin I}\sigma_s}\right]$$

since $\sum_{s<t|s\notin I}\sigma_s$ is constant given the history up to time $t-1$, and $\mathbb{E}^{t-1}\mathbb{E}_{t-1} = \mathbb{E}$ by the tower rule,

$$\le 1 + (d^2+1)\log(B+1) + \mathbb{E}\left[\frac{1}{2}\sum_{t\notin I}\frac{36d^{3/2}\sigma_t}{2+\sum_{s<t|s\notin I}\sigma_s}\right]$$

since $\sigma_t \le 1$,

$$\le 1 + (d^2+1)\log(B+1) + \mathbb{E}\left[\frac{1}{2}\sum_{t\notin I}\frac{36d^{3/2}\sigma_t}{1+\sum_{s\le t|s\notin I}\sigma_s}\right]$$
$$\le 1 + (d^2+1)\log(B+1) + 18d^{3/2}\mathbb{E}\left[\log(1+\sum_{s\notin I}\sigma_s)\right],$$

where the last step follows from the inequality $\sum_{t=1}^{T}\frac{a_t}{1+a_{1:t}} \le \log(1+a_{1:T})$ for any non-negative real numbers $a_1,\dots,a_T$. □

Proof of Theorem 4.1. Lemma 4.2, Lemma 4.3, and Lemma 4.4 complete the proof when there are no bad responses (i.e., $B = 0$).

Now we focus on the case when some of the responses can be bad (i.e., $B \neq 0$). The intuitive difficulty in this case is that if an estimate $\hat{c}_t$ is incorrect for a certain time t, the value $\bar{x}_t$ will continue to be "incorrect" even for time steps after t. Our core observation is that this can be managed.

Define $I \subseteq [T]$ to be the set of times at which the responses are bad. First, note that even with bad responses, the algorithm always plays a feasible point x_t. We have the following.

$$\mathbb{E}[\sum_t\langle c_t, x_t - u\rangle] = \sum_{t\in I}\mathbb{E}[\langle c_t, x_t - u\rangle] + \sum_{t\notin I}\mathbb{E}[\langle c_t, x_t - u\rangle]$$
$$\le 2B + \sum_{t\notin I}\mathbb{E}[\tilde{\ell}_t(\bar{x}_t) - \tilde{\ell}_t(u)]$$
$$\le 2B + \mathbb{E}\left[\left|\sum_{t\in I}\tilde{\ell}_t(\bar{x}_t) - \tilde{\ell}_t(u)\right| + \sum_{t=1}^{T}(\tilde{\ell}_t(\bar{x}_t) - \tilde{\ell}_t(u))\right].$$

The last term can, once again, be bounded by Lemma 4.4. To bound the middle term, note that we always have $\tilde{\ell}_t(x) \le \langle\hat{c}_t, x\rangle \le \|\hat{c}_t\| \le d$. Thus, we have $\tilde{\ell}_t(\bar{x}_t) - \tilde{\ell}_t(u) \le 2d$ for any $t \in I$, and this completes the proof. □

5. Robustness with Response Feedback

In this section we enhance our model in order to improve our robustness to bad responses. Previously, we achieved logarithmic regret when all responses are good, with a linear decay in the number of bad responses B. If B is small this is still a non-trivial robustness guarantee, but if $B = T$ then we could simply ignore all responses and run a standard bandit algorithm to achieve regret $O(\sqrt{B})$. It is interesting to ask if there is a single algorithm that can achieve this "best of both worlds" guarantee. Unfortunately, we are not aware of such an algorithm, and designing one is an interesting open problem (note that this is known to be possible in the full-information setting (Bhaskara et al., 2023)). The challenge arises because in a bandit setting, the algorithm may never know if a query response was good or bad.

We thus study a new model in which we allow our algorithm some extra knowledge about the bad responses. Specifically, the algorithm is told whether a response q_t is bad, but only *after* it has played x_t. Formally, in each time step, the algorithm first makes a query s_t, receives a response q_t, and then must play a point x_t. After this, the algorithm the receives the loss $\langle c_t, x_t\rangle$ as well as a *feedback* $g_t \in \{0, 1\}$ such that if $g_t = 1$, then $q_t = \langle c_t, s_t\rangle$.

Even with this extra knowledge, it is unclear how to handle bad responses. Recall that in Algorithm 1, we use the response to generate an unbiased estimate of the unknown cost. If the response is revealed to be bad, then we are at liberty to ignore this corrupted estimate. However, if we ignore an estimate, then we cannot make an update in that time step, which again leads to a linear dependence on B.

Ideally, we might hope to divide the time steps into two groups: those where $g_t = 1$ and those where $g_t = 0$. Then, we could run a standard bandit algorithm on the time steps where $g_t = 0$ and use Algorithm 1 on the rest. The problem, of course, is that we do not know the value of g_t before playing x_t, and so we do not know which algorithm to use.

A more nuanced approach would be to instead incorporate some kind of extra "exploration" into Algorithm 1. That is, we play $\hat{x}_t = x_t + e_t$ for some mean-zero random vector e_t on each time step. This is a common tactic in bandit analysis, and would allow us to form an unbiased estimate of c_t via the one-point regression $\hat{c}_t = \langle \hat{x}_t, c_t\rangle \mathbb{E}[e_t e_t^\top]^{-1} e_t$. The key difficulty is choosing an appropriate distribution for e_t such that the variance of the estimate is small.

Intuitively, in standard bandit algorithms, the variance is kept small by forcing x_t not to get too close to the boundary of the domain, e.g., x_t is usually implicitly constrained to be inside the ball of radius $1 - \frac{1}{\sqrt{T}}$. However, such a constraint would not allow us to obtain logarithmic regret when all the query responses are good. Moreover, typically there is a very intricate relationship between the update step to generate x_{t+1} from x_t and $\hat{c}_t$ and the exploration distribution used to sample e_t. Thus, our challenge is to incorporate this careful exploration alongside exploitation of the queries.

Our approach is again inspired by previous literature on using hints in the full-information setting, but via the very different algorithmic construction of Bhaskara et al. (2021). This more recent construction is designed to make use of only $O(\sqrt{T})$ hints, and yet still obtain logarithmic regret. This is plausibly useful for our purposes because it suggests that the algorithm's actions depend only very mildly on the hints and so are less likely to disturb the delicate balance required for bandit exploration/exploitation tradeoffs. The formal specification is provided in Algorithms 2 and 3.

Algorithm 2 Bandit OLO with Response Feedback.

Require: Parameters $\eta, \gamma \geq 0$
  Define $\phi(x) := -\log(1 - \|x\|^2)$ and $r_0(x) := \frac{\sqrt{d^2}}{2\eta}\|x\|^2$
  $\bar{x}_1 \leftarrow 0$
  for $t = 1, \ldots, T$ **do**
    $s_t \leftarrow$ uniform random vector on unit sphere in $\mathbb{R}^d$
    $q_t \leftarrow \mathcal{Q}(s_t)$ ▷ Query and response
    $\alpha_t \leftarrow \mathrm{clip}_{\frac{4}{\sqrt{d}}}(q_t)$
    $h_t \leftarrow \frac{\sqrt{d}}{4}\alpha_t s_t$
    $w_t \leftarrow$ uniform random vector on unit sphere in $\mathbb{R}^d$
    $z_t \leftarrow \bar{x}_t + \nabla^2\phi(\bar{x}_t)^{-1/2} w_t$
    Get $p_t \in [0, 1/2]$ from Algorithm 3
    Play $x_t = -p_t h_t + (1 - p_t) z_t$
    Incur loss $\ell_t = \langle c_t, x_t\rangle$
    Receive $g_t \in \{0, 1\}$, ($g_t = 1$ if the response is good)
    ▷ Feedback
    $\hat{c}_t \leftarrow \begin{cases} \frac{d\ell_t}{1-p_t} \cdot \nabla^2\phi(\bar{x}_t)^{1/2} w_t & \text{if } g_t = 0 \\ d \cdot q_t \cdot s_t & \text{if } g_t = 1 \end{cases}$
    Define $\sigma_t := g_t\|\hat{c}_t\|^2$ and $r_t(x) := \frac{\delta_t}{2}\|x\|^2$ with $\delta_t = \frac{\sqrt{d^2+\sigma_{1:t}} - \sqrt{d^2+\sigma_{1:t-1}}}{\eta}$
    $\bar{x}_{t+1} \leftarrow \mathrm{argmin}_{\|x\|\leq 1} \langle \hat{c}_{1:t}, x\rangle + r_{0:t}(x) + \gamma\phi(x)$
    Send $\langle c_t, -p_t h_t + (1 - p_t) z_t\rangle = \ell_t$ to Algorithm 3
    If $g_t = 1$, also send $\langle c_t, h_t\rangle = \frac{\sqrt{d}}{4}\alpha_t q_t$ to Algorithm 3
  end for

We informally discuss the main ideas here and defer the formal proof and technical details to Appendix C. At each time step, we will play a *linear combination*:

$$x_t = -p_t h_t + (1 - p_t) z_t$$
$$z_t = \bar{x}_t + e_t,$$

where $p_t \approx 1/\sqrt{T}$ is a weighting factor, $\bar{x}_t$ is the output of a more "standard" base bandit algorithm, and e_t is a random exploration term. Note that since p_t is rather small, the algorithm actually does not deviate much from the predictions of the base bandit algorithm. This property will allow us to blend bandit analysis with queries.

Algorithm 3 Hint Weight Learner

Require: $B =$ number of bad queries , parameter λ
$p_1 \leftarrow 0$
$D_1 \leftarrow \frac{\lambda}{\sqrt{B+4\lambda}} \leq 1/2.$
for $t = 1, \dots, T$ **do**
  Play p_t
  Receive feedback $\langle c_t, -p_t h_t + (1-p_t) z_t \rangle$
  If $g_t = 1$, also receive feedback $\langle c_t, h_t \rangle$
  Define $v_t = -\langle c_t, h_t \rangle - \langle c_t, z_t \rangle$ ▷ for analysis
  $\hat{v}_t \leftarrow g_t v_t$ ▷ v_t can be computed if $g_t = 1$
  Define $\eta_t = \frac{\lambda}{4\lambda + B + \hat{v}_{1:t}^2}$
  Define $D_{t+1} = \min\left(1, \frac{\lambda}{\sqrt{4\lambda+B+\hat{v}_{1:t}^2}}\right)$
  Output $p_{t+1} = \max(0, \operatorname{clip}_{D_{t+1}}(p_t - \eta_t \hat{v}_t))$
end for

In time steps in which $g_t = 1$, we can form a low-variance estimate of the cost c_t via $dq_t s_t$, exactly as in Algorithm 1. However, in time steps in which $g_t = 0$, we can still form an unbiased estimate of c_t via $\hat{c}_t = \frac{\langle c_t, x_t \rangle \mathbb{E}[e_t e_t^\top]^{-1} e_t}{1-p_t}$. As is typical in bandit OLO, we will rely on tools from self-concordant analysis to ensure that $\langle c_t, x_t \rangle \mathbb{E}[e_t e_t^\top]^{-1} e_t$ does not have prohibitively high variance, and then rely on the fact that $p_t \approx 1/\sqrt{T}$, and in particular $p_t \leq 1/2$, to ensure that $\hat{c}_t$ will continue to not have high variance.

In more detail, $\bar{x}_t$ will be set via the FTRL update:

$$\bar{x}_{t+1} = \operatorname*{argmin}_{\|x\| \leq 1} \langle \hat{c}_{1:t}, x \rangle + \frac{\sqrt{d^2 + \sum_{i=1}^t g_i \|\hat{c}_i\|^2}}{4} \|x\|^2 - d^{3/4}\sqrt{B}\log(1 - \|x\|^2).$$

This is an FTRL update with a regularizer that is a mixture of the standard quadratic regularizer popular in full-information settings and the self-concordant barrier regularizer $-\log(1-\|x\|^2)$ popular in bandit settings. The exploration e_t is generated by $\nabla^2 \phi(\bar{x}_t)^{-1/2} w_t$ where $\phi(x) = -\log(1-\|x\|^2)$ and w_t is uniform on the unit sphere. This is the classical Dikin ellipsoid exploration that is ubiquitous in bandit analysis (Abernethy et al., 2008; Bubeck et al., 2012; Lattimore & Szepesvári, 2020).

To analyze this procedure, we write the regret as follows:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \langle c_t, -p_t h_t + (1-p_t) z_t - u \rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T p_t \langle c_t, -h_t - z_t \rangle + \sum_{t=1}^T \langle c_t, z_t - u \rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T (p_t - p_t^\star) \langle c_t, -h_t - z_t \rangle + \sum_{t=1}^T p_t^\star \langle c_t, -h_t - z_t \rangle \right. \\
&\qquad \left. + \sum_{t=1}^T \langle c_t, x_t - u \rangle\right],
\end{aligned}$$

where $p_t^\star$ is an arbitrary sequence of scalars and we move from z_t to x_t in the final equation because $\mathbb{E}[z_t] = x_t$. Our analysis then proceeds in several steps.

First, we show that for any sequence $p_t^\star$ with $p_t^\star$ set to an unknown constant that is $O(1/\sqrt{S})$ for the first S iterations and then 0 afterwards, there is a strategy for choosing p_t such that $\sum_{t=1}^T (p_t - p_t^\star)\langle c_t, -h_t - z_t \rangle = \tilde{O}(\sqrt{dB})$. This is accomplished by choosing p_t itself via an OLO algorithm operating on the linear losses $p \mapsto p\langle c_t, -h_t - z_t \rangle$ (described formally by Algorithm 3). Critically, when $g_t = 1$ we can exactly compute $\langle c_t, -h_t - z_t \rangle$. However, for time steps with $g_t = 0$, we cannot compute this loss, but we bound the influence of these B time steps by restricting p_t to the range $[0, 1/\sqrt{B}]$. This result is formalized in Lemma C.1 Thus, after this step is finished, we need only show that there exists an appropriate $p_t^\star$ that causes the remaining terms to be small.

Next, we consider the regret of the "base" FTRL algorithm: $\mathbb{E}\left[\sum_{t=1}^T \langle c_t, z_t - u \rangle\right] = \mathbb{E}\left[\sum_{t=1}^T \langle c_t, \bar{x}_t - u \rangle\right]$. In standard bandit analysis, we would use exclusively the estimate $\hat{c}_t = d\langle c_t, z_t \rangle \nabla^2 \phi(\bar{x}_t)^{1/2} w_t$. In this case, we instead have the estimate $\hat{c}_t = \frac{d}{1-p_t}\langle c_t, z_t \rangle \nabla^2 \phi(\bar{x}_t)^{1/2} w_t$ when $g_t = 0$ and $dq_t s_t$ when $g_t = 1$. Our analysis will also partition the iterates by the value of g_t. When $g_t = 0$, notice that since $p_t \leq 1/2$, the variance of our $\hat{c}_t$ is only a factor of 4 worse than the variance of the standard bandit estimator. Therefore, classical bandit analysis based on *self-concordance* allows us to control the total regret over these time steps at a rate of $\tilde{O}(d^{5/4}\sqrt{B})$ using the $-d^{3/4}\sqrt{B}\log(1-\|x\|^2)$ term of the regularizer.

For time steps with good responses, we observe that the variance of $\hat{c}_t$ is bounded by d (and in particular does *not* depend on any careful exploration/exploitation tradeoff) so we can deploy techniques from full-information analysis of FTRL (e.g., (McMahan, 2017)) based on *strong convexity* to bound the regret of the x_t using the $\|x\|^2$ term in the regularizer by $O(\sqrt{dT})$. Overall then, we see that the regret of the FTRL iterates $\bar{x}_t$ can be bounded by $O(\sqrt{dT} + d^{5/4}\sqrt{B})$. The $d^{5/4}$ arises from the $d^{3/4}$ coefficient on the self-concordant barrier, and is required to balance a stability term that appears in the next step of the analysis.

Now, for the final most technical challenge: we need to show that by an appropriate choice of $p_t^\star$, the term $\sum_{t=1}^T p_t^\star \langle c_t, -h_t - z_t \rangle$ will be a negative value that "cancels out" the $O(\sqrt{dT})$ term in FTRL regret. The first step of this is to actually improve the bound of FTRL. Specifically, we show that if there is a time point S for which $\|\sum_{t=1}^{t'} \hat{c}_t\| \geq \Omega(d^{3/2}\sqrt{d^2 + \sum_{t=1}^{t'} \|\hat{c}_t\|^2})$, $\forall t' \geq S$, then

the $\sqrt{dT}$ term in the FTRL bound may be improved to $\sqrt{dS}$. The intuition for this result is the following: when the sum of the costs is $\Omega(d^{3/2}\sqrt{d^2 + \sum_{t=1}^{t'} \|\hat{c}_t\|^2})$, then the FTRL update "looks like" a projection onto a ball of radius roughly $1 - d^{5/4}\sqrt{B}/t$. This $d^{5/4}$ in the numerator arises from the $d^{3/4}$ coefficient in the self-concordant barrier and is in tension with the $d^{5/4}\sqrt{B}$ term in the regret of FTRL, justifying our use of this non-standard coefficient (rather than the d one might expect). Now, past analysis of this projection algorithm (Huang et al., 2017; Bhaskara et al., 2021) show that the regret accumulated over indices $t' > S$ will be only $\tilde{O}(d^{3/2})$. These results are captured in Lemma C.4.

Finally, we choose the correct values for $p_t^\star$. The idea is to set $p_t^\star \approx d^{3/2}/\sqrt{S}$ for $t \le S$, and $p_t^\star = 0$ for $t \ge S$. With this setting, we can show that for the first S steps, $\mathbb{E}[\sum_{t=1}^{S}\langle c_t, h_t + z_t\rangle] \ge \Omega(S/d)$, so that $-p_t^\star \sum_{t=1}^{S}\langle c_t, h_t + z_t\rangle \le -\sqrt{dS}$, which is enough to completely cancel the $\sqrt{dS}$ term in the regret accumulated by FTRL. The final Theorem is formally presented in Theorem 5.1:

Theorem 5.1. *Suppose we run Algorithm 2 with $\eta = 4$, $\gamma = d^{3/4}\sqrt{B}$ and $\lambda = 3\cdot 4\cdot(32\cdot 52)^2\cdot d + (32\cdot 52)\sqrt{3dB} + \sqrt{3}\cdot 32\cdot 52\cdot d^{3/2}$. Suppose also that the number of times $g_t = 0$ is at most B. Then:*

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle c_t, x_t - u\rangle\right] \le \tilde{O}(d^{3/2} + d^{5/4}\sqrt{B}).$$

Notice that this result requires an upper bound on B as input. However, this can be easily removed via a doubling trick (e.g., see (Shalev-Shwartz et al., 2012)): we maintain a "guess" for the final value of B, and every time our guess is violated we restart the algorithm and double the guess. This will worsen the constants, but not the asymptotics.

6. Conclusions

In this paper we study OLO with bandit feedback when the algorithm has access to additional information via hints and queries to the upcoming cost vector. Surprisingly, unlike the full-information setting, we show that even receiving good hints at all time steps is not sufficient to obtain regret better than $\tilde{O}(\sqrt{T})$. We then introduce the query model and show that it is possible to obtain the desired logarithmic regret bounds in this setting. Extending our robustness results from Section 5 when the algorithm does not receive feedback on the response is an interesting research question.

Acknowledgements

Aditya Bhaskara is partially supported by NSF awards CCF-2008688 and CCF-2047288.

References

Abernethy, J. D., Hazan, E., and Rakhlin, A. Competing in the dark: An efficient algorithm for bandit linear optimization. In *COLT*, 2008.

Bamas, É., Maggiori, A., Rohwedder, L., and Svensson, O. Learning augmented energy minimization via speed scaling. In *NeurIPS*, 2020.

Bhaskara, A., Cutkosky, A., Kumar, R., and Purohit, M. Online learning with imperfect hints. In *ICML*, 2020.

Bhaskara, A., Cutkosky, A., Kumar, R., and Purohit, M. Logarithmic regret from sublinear hints. In *NeurIPS*, pp. 28222–28232, 2021.

Bhaskara, A., Gollapudi, S., Im, S., Kollias, K., and Munagala, K. Online learning and bandits with queried hints. In *ITCS*, 2023.

Bubeck, S., Cesa-Bianchi, N., and Kakade, S. M. Towards minimax policies for online linear optimization with bandit feedback. In *COLT*, pp. 41–1, 2012.

Dekel, O., Flajolet, A., Haghtalab, N., and Jaillet, P. Online learning with a hint. In *NIPS*, pp. 5299–5308, 2017.

Gollapudi, S. and Panigrahi, D. Online algorithms for rent-or-buy with expert advice. In *ICML*, pp. 2319–2327, 2019.

Gupta, A., Koren, T., and Talwar, K. Better algorithms for stochastic bandits with adversarial corruptions. In *COLT*, pp. 1562–1578, 2019.

Hazan, E. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Hazan, E. and Megiddo, N. Online learning with prior knowledge. In *COLT*, pp. 499–513, 2007.

Huang, R., Lattimore, T., György, A., and Szepesvári, C. Following the leader and fast rates in online linear prediction: Curved constraint sets and other regularities. *JMLR*, 18(145):1–31, 2017.

Im, S., Kumar, R., Montazer Qaem, M., and Purohit, M. Non-clairvoyant scheduling with predictions. In *SPAA*, pp. 285–294, 2021.

Ito, S. On optimal robustness to adversarial corruption in online decision problems. In *NeurIPS*, 2021.

Jiang, Z., Panigrahi, D., and Sun, K. Online algorithms for weighted paging with predictions. *ICALP*, pp. 69:1–69:18, 2020.

Kumar, R., Purohit, M., and Svitkina, Z. Improving online algorithms using ML predictions. In *NeurIPS*, pp. 9661–9670, 2018.

Kumar, R., Purohit, M., Schild, A., Svitkina, Z., and Vee, E. Semi-online bipartite matching. In *ITCS*, pp. 50:1–50:20, 2019.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020.

Lavastida, T., Moseley, B., Ravi, R., and Xu, C. Learnable and instance-robust predictions for online matching, flows and load balancing. In *ESA*, pp. 59:1–59:17, 2021.

Lykouris, T. and Vassilvtiskii, S. Competitive caching with machine learned advice. In *ICML*, pp. 3296–3305, 2018.

Lykouris, T., Mirrokni, V., and Paes Leme, R. Stochastic bandits robust to adversarial corruptions. In *STOC*, pp. 114–122, 2018.

McMahan, H. B. A survey of algorithms and analysis for adaptive online learning. *JMLR*, 18(1):3117–3166, 2017.

Mitzenmacher, M. Scheduling with predictions and the price of misprediction. In *ITCS*, pp. 14:1–14:18, 2020.

Rakhlin, A. and Sridharan, K. Online learning with predictable sequences. In *COLT*, pp. 993–1019, 2013.

Ranosova, H. Spherically Symmetric Measures. Bachelor's thesis, Department of Probability and Mathematical Statistics, Charles University, 2021.

Rohatgi, D. Near-optimal bounds for online caching with machine learned advice. In *SODA*, pp. 1834–1845, 2020.

Shalev-Shwartz, S. et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Steinhardt, J. and Liang, P. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *ICML*, 2014.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer, 2009.

Wei, C. and Luo, H. More adaptive algorithms for adversarial bandits. In *COLT*, pp. 1263–1291, 2018.

Wei, C.-Y., Luo, H., and Agarwal, A. Taking a hint: How to leverage loss predictors in contextual bandits? In *COLT*, pp. 3583–3634, 2020.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pp. 928–936, 2003.

Zuo, J., Zhang, X., and Joe-Wong, C. Observe before play: Multi-armed bandit with pre-observations. *ACM SIGMETRICS Performance Evaluation Review*, 46(2): 89–90, 2019.

A. Missing Proofs

Lemma 4.2. *In Algorithm 1, the following hold: (i) $\|h_t\| \le 1$ and $\|x_t\| \le 1$. If the response is good at time t, then*

(ii) $\mathbb{E}[\hat{c}_t] = c_t$,
(iii) $\mathbb{E}_{t-1}[d\alpha_t^2] \ge (1/4)\|c_t\|^2$,
(iv) $\mathbb{E}[\|\hat{c}_t\|^2] = d\|c_t\|^2$.

Proof. For part (i), we note that $\|h_t\| = \frac{\sqrt{d}}{4} \cdot |\alpha_t| \le 1$ by the definition of α_t. To bound $\|x_t\|$, we use the following simple argument (e.g., Bhaskara et al., 2020, Lemma 3.2),

$$\|x_t\| \le \|\bar{x}_t\| + \frac{(\|\bar{x}_t\|^2 - 1)}{2}\|h_t\| \le \|\bar{x}_t\| + \frac{(\|\bar{x}_t\|^2 - 1)}{2} \le 1,$$

where the last inequality uses $\|\bar{x}_t\| \le 1$.

We next focus on parts (ii)–(iv). Due to rotational symmetry, we may assume that $c_t = (\gamma, 0, \dots, 0) \in \mathbb{R}^d$ for some fixed $\gamma \in [0, 1]$. Also, let us write $s_t = (g_1, \dots, g_d)$, for convenience. So we have $q_t = \langle c_t, s_t \rangle = \gamma g_1$. Since s_t is a unit vector, g_i, g_j are not independent for $i \ne j$. However, we still have the property that $\mathbb{E}[g_j \mid g_1 = z] = 0$ for all $z \in [-1, 1]$ and $j \ne 1$. To see part (ii), observe that:

$$\mathbb{E}[\hat{c}_t] = \mathbb{E}[dq_t s_t] = d \cdot (\gamma\mathbb{E}[g_1^2], \mathbb{E}[g_1 g_2], \dots, \mathbb{E}[g_1 g_d]) = (\gamma, 0, \dots, 0) = c_t,$$

where because of symmetry, $\mathbb{E}[g_1^2] = \mathbb{E}[(g_1^2 + \cdots + g_d^2)/d] = 1/d$ and $\mathbb{E}[g_1 g_j] = \mathbb{E}[g_1 \mathbb{E}[g_j | g_1]] = 0,\ \forall j \ne 1$.

Now, for part (iii), observe that $\mathbb{E}[dq_t^2] = \|c_t\|^2$, so that it suffices to show $\mathbb{E}[q_t^2] - \mathbb{E}[\alpha_t^2] \le (3\|c_t\|^2)/(4d)$. To this end,

$$\begin{aligned}\mathbb{E}[q_t^2 - \alpha_t^2] &\le \Pr\left[q_t^2 \in \left(\frac{16 \cdot 2^0}{d}, \frac{16 \cdot 2^1}{d}\right]\right] \frac{16 \cdot 2^1}{d} + \Pr\left[q_t^2 \in \left(\frac{16 \cdot 2^1}{d}, \frac{16 \cdot 2^2}{d}\right]\right] \frac{16 \cdot 2^2}{d} + \cdots \\ &\le \sum_{k=1}^{\infty} \Pr\left[q_t^2 > \frac{16 \cdot 2^{k-1}}{d}\right] \left(\frac{16 \cdot 2^k}{d}\right).\end{aligned}$$

But, by Markov's inequality, we have

$$\Pr\left[q_t^2 > \frac{16 \cdot 2^{k-1}}{d}\right] = \Pr\left[q_t^4 > \frac{16^2 \cdot 2^{2k-2}}{d^2}\right] \le \frac{\mathbb{E}[q_t^4] \cdot d^2}{16^2 \cdot 2^{2k-2}} \le \frac{\|c_t\|^2 \cdot 3}{16^2 \cdot 2^{2k-2}},$$

where the last inequality follows from $\mathbb{E}[q_t^4] = \gamma^4 \mathbb{E}[g_1^4] \le \gamma^2 \cdot \frac{3}{d(d+2)}$ (see Lemma B.1). Substituting, we get

$$\mathbb{E}[q_t^2 - \alpha_t^2] \le \left(\frac{3\|c_t\|^2}{d}\right) \sum_{k=1}^{\infty} \frac{16 \cdot 2^k}{16^2 \cdot 2^{2k-2}} = \frac{3\|c_t\|^2}{4d}.$$

Finally, for part (iv), we have

$$\mathbb{E}[\|\hat{c}_t\|^2] = \mathbb{E}[d^2 q_t^2] = d^2 \mathbb{E}[\gamma^2 g_1^2] = d\gamma^2.$$ □

B. Properties of Uniform Distribution on the Sphere

Lemma B.1. *If $(x_1, \dots, x_d) \in \mathbb{R}^d$ is a uniform random point on the unit sphere, then (i) $\mathbb{E}[x_1^4] = \frac{3}{d(d+2)}$ and (ii) $\mathbb{E}[x_1^2 x_2^2] = \frac{1}{d(d+2)}$.*

Proof. It is known that for any $i \in [d]$,

$$x_i^2 \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{d-1}{2}\right),$$

see, e.g., (Ranosova, 2021, Theorem 13 and Remark 1) for a proof. Also, if $Z \sim \mathrm{Beta}(\alpha, \beta)$, then its second moment is

$$\mathbb{E}[Z^2] = \frac{\alpha(\alpha+1)}{(\alpha+\beta+1)(\alpha+\beta)}. \tag{2}$$

Hence, $\mathbb{E}[x_1^4]$ is the second moment of x_1^2, which can be computed from (2) setting $\alpha = 1/2$ and $\beta = (d-1)/2$. This, after simplification, yields (i). For (ii), notice that since $(x_1, \ldots, x_d)$ is on the unit sphere, we get

$$1 = \left(\sum_{i\in[d]} x_i^2\right)^2 = \sum_{i\in[d]} x_i^4 + 2\sum_{i\neq j\in[d]} x_i^2 x_j^2.$$

Taking expectation, using (i), and by rotational symmetry, we obtain (ii). □

C. Improved Robustness via Response Feedback

Theorem 5.1. *Suppose we run Algorithm 2 with $\eta = 4$, $\gamma = d^{3/4}\sqrt{B}$ and $\lambda = 3\cdot 4\cdot (32\cdot 52)^2\cdot d + (32\cdot 52)\sqrt{3dB} + \sqrt{3}\cdot 32\cdot 52\cdot d^{3/2}$. Suppose also that the number of times $g_t = 0$ is at most B. Then:*

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle c_t, x_t - u\rangle\right] \le \tilde{O}(d^{3/2} + d^{5/4}\sqrt{B}).$$

Proof. First, observe that since $\mathbb{E}[z_t] = \bar{x}_t$, we have for any sequence $p_1^\star, \ldots, p_T^\star$:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle c_t, x_t - u\rangle\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\langle c_t, -p_t h_t - p_t z_t + \bar{x}_t - u\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T} p_t(\langle c_t, -h_t\rangle - \langle c_t, z_t\rangle) + \sum_{t=1}^{T}\langle c_t, \bar{x}_t - u\rangle\right]\\ &= \mathbb{E}\left[\sum_{t=1}^{T} p_t v_t + \sum_{t=1}^{T}\langle c_t, \bar{x}_t - u\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T} v_t(p_t - p_t^\star) + \sum_{t=1}^{T} v_t p_t^\star + \sum_{t=1}^{T}\langle c_t, \bar{x}_t - u\rangle\right].\end{aligned}$$

Let $\alpha = 1/(16d^{3/2})$ and let S be the smallest index such that $\|\hat{c}_{1:t}\| \ge \alpha(d^2 + \sigma_{1:t})/\eta$ for all $t > S$. Applying Lemma C.4:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t - u\rangle\right] \le \mathbb{E}\left[1 + \frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \gamma\log(T) + \frac{16d^2B}{\gamma} + \frac{16(\eta+1+\gamma/d)}{\alpha} + \left(\frac{8\eta}{\alpha}+4\right)\log\left(1+\frac{\sigma_{1:t}}{d^2}\right)\right]\\ + \mathbb{E}\left[\|\hat{c}_{1:S}\| + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle\right].\end{aligned}$$

Next, by Lemma C.1, if we set $p_t^\star$ such that $p_t^\star = \delta$ for some fixed $\delta \le D_S$ for all $t \le S$ and $p_t^\star = 0$ for $t > S$, we have:

$$\sum_{t=1}^{T}\langle v_t, p_t - p_t^\star\rangle \le \lambda + 2\lambda\log(4\lambda + B + T) + \sqrt{B},$$

and for $t \le S$ and $g_t = 1$:

$$\begin{aligned}\mathbb{E}[\langle v_t, p_t^\star\rangle] &= \mathbb{E}[\delta\langle c_t, -h_t\rangle - \delta\langle c_t, z_t\rangle] = \mathbb{E}[\delta\langle c_t, -h_t\rangle - \delta\langle c_t, \bar{x}_t\rangle] = \mathbb{E}\left[-\frac{\delta\sqrt{d}}{4}q_t\alpha_t - \delta\langle c_t, \bar{x}_t\rangle\right]\\ &\le \mathbb{E}\left[-\frac{\delta\sqrt{d}}{4}\alpha_t^2 - \delta\langle c_t, \bar{x}_t\rangle\right] \le \mathbb{E}\left[-\frac{\delta}{16\sqrt{d}}\|c_t\|^2 - \delta\langle c_t, \bar{x}_t\rangle\right]\\ &= \mathbb{E}\left[-\frac{\delta}{16d^{3/2}}\|\hat{c}_t\|^2 - \delta\langle c_t, \bar{x}_t\rangle\right] = \mathbb{E}\left[-\frac{\delta}{16d^{3/2}}\sigma_t - \delta\langle c_t, \bar{x}_t\rangle\right],\end{aligned}$$

where the second inequality and the penultimate equality follow using Lemma 4.2. Alternatively, when $g_t = 0$,

$$\mathbb{E}[\langle v_t, p_t^\star\rangle] \le \mathbb{E}[2\delta] \le \mathbb{E}\left[3\delta - \delta\langle c_t, \bar{x}_t\rangle\right] \le \mathbb{E}\left[3\delta - \delta\langle \hat{c}_t, \bar{x}_t\rangle\right] \le \mathbb{E}\left[\frac{3}{\sqrt{B}} - \delta\langle \hat{c}_t, \bar{x}_t\rangle\right].$$

Therefore:

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle v_t, p_t^\star\rangle\right] \le \mathbb{E}\left[-\delta\frac{\sigma_{1:S}}{16d^{3/2}} - \delta\sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t\rangle + 3\sqrt{B}\right].$$

Thus, overall we have:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\langle c_t, x_t - u\rangle\right] \\
&\le \mathbb{E}\left[1 + \gamma\log(T) + \frac{16d^2B}{\gamma} + \frac{16(\eta+1+\gamma/d)}{\alpha} + \left(\frac{8\eta}{\alpha}+4\right)\log\left(1+\frac{\sigma_{1:t}}{d^2}\right)\right] \\
&\quad + \mathbb{E}\left[\inf_{0\le\delta\le D_S}\frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \|\hat{c}_{1:S}\| + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle + \lambda + 2\lambda\log(4\lambda+B+T) + 4\sqrt{B} - \delta\frac{\sigma_{1:S}}{16d^{3/2}} - \delta\sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t\rangle\right] \\
&\le \mathbb{E}\left[1 + \gamma\log(T) + \frac{16d^2B}{\gamma} + \frac{80+16\gamma/d}{\alpha} + \left(\frac{32}{\alpha}+4\right)\log\left(1+\frac{\sigma_{1:t}}{d^2}\right)\right] \\
&\quad + \mathbb{E}\left[\inf_{0\le\delta\le D_S}\frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \|\hat{c}_{1:S}\| + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle + \lambda + 2\lambda\log(4\lambda+B+T) + 4\sqrt{B} - \delta\frac{\sigma_{1:S}}{16d^{3/2}} - \delta\sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t\rangle\right].
\end{aligned} \tag{3}$$

Further, observe that by Lemma C.3:

$$\|\hat{c}_{1:S}\| + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle \le \gamma\log(S) + 16\left(\eta+\frac{1}{\eta}\right)\sqrt{d^2+\sigma_{1:S}} + \frac{4d^2B}{\gamma}.$$

With $\lambda = 3\cdot 4\cdot(24\cdot 52)^2\cdot d + (24\cdot 52)\sqrt{3dB} + \sqrt{3}\cdot 24\cdot 52\cdot d^{3/2}$, suppose $\sqrt{4\lambda+B+d^2+\sigma_{1:S}} \le \lambda$. Then we have:

$$12\left(\eta+\frac{1}{\eta}\right)\sqrt{d^2+\sigma_{1:S}} + \frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} \le \left(12\eta+\frac{13}{\eta}\right)\lambda \le 52\lambda,$$

where the last inequality follows since $\eta = 4$. With $\delta = 0$:

$$\begin{aligned}
&\frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \|\hat{c}_{1:S}\| + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle + \lambda + 2\lambda\log(4\lambda+B+T) + 4\sqrt{B} - \delta\frac{\sigma_{1:S}}{d^2} - \delta\sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t\rangle \\
&\quad\le 53\lambda + 2\lambda\log(4\lambda+B+T) + 4\sqrt{B}.
\end{aligned} \tag{4}$$

In the remaining argument, we assume $\sqrt{4\lambda+B+d^2+\sigma_{1:S}} > \lambda$. Now, observe that even in this scenario we still have:

$$\begin{aligned}
\frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \|\hat{c}_{1:S}\| + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle &\le \gamma\log(S) + \left(12\eta+\frac{13}{\eta}\right)\sqrt{d^2+\sigma_{1:S}} + \frac{4d^2B}{\gamma} \\
&\le \gamma\log(S) + 52\sqrt{d^2+\sigma_{1:S}} + \frac{4d^2B}{\gamma}.
\end{aligned}$$

Our goal will be to show that either the term $\sqrt{d^2+\sigma_{1:S}}$ is small, or that it can be canceled out by negative terms multiplied by δ. To do this, we consider a few cases. First, suppose that $\sqrt{d^2+\sigma_{1:S}} \le \frac{10}{\alpha}$. Then, observe that by setting $\delta = 0$:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\langle c_t, x_t - u\rangle\right] &\le \mathbb{E}\left[1 + \gamma\log(T) + \frac{16d^2B}{\gamma} + \frac{8400+16\gamma/d}{\alpha} + \left(\frac{32}{\alpha}+4\right)\log\left(1+\frac{\sigma_{1:t}}{d^2}\right)\right] \\
&\quad + \mathbb{E}\left[\lambda + 2\lambda\log(4\lambda+B+T) + 4\sqrt{B}\right].
\end{aligned} \tag{5}$$

Moreover, if alternatively we have $\frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \|\hat{c}_{1:S}\| + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle \le \frac{\sqrt{d}}{16}$, then we also can set $\delta = 0$ to obtain:

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle c_t, x_t - u\rangle\right] \le \mathbb{E}\left[1 + \frac{\sqrt{d}}{16} + \gamma\log(T) + \frac{16d^2B}{\gamma} + \frac{80+16\gamma/d)}{\alpha} + \left(\frac{8}{\alpha}+4\right)\log\left(1+\frac{\sigma_{1:t}}{d^2}\right)\right] + \mathbb{E}\left[\lambda + 2\lambda\log(4\lambda+B+T) + 4\sqrt{B}\right]. \tag{6}$$

Let us now consider the situation in which $\sqrt{d^2+\sigma_{1:S}} > \frac{10}{\alpha}$ and also $\frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \|\hat{c}_{1:S}\| + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle > \frac{\sqrt{d}}{16}$. In this case, we want to choose δ such that (recalling $\eta = 4$):

$$\delta\frac{\sigma_{1:S}}{16d^{3/2}} + \delta\sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t\rangle \ge 12(\eta + 1/\eta)\sqrt{d^2+\sigma_{1:S}} = 52\sqrt{d^2+\sigma_{1:S}}.$$

We claim that in this case (i.e., $\sqrt{d^2+\sigma_{1:S}} > \frac{10}{\alpha}$ and also $\frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \|\hat{c}_{1:S}\| + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle > \frac{\sqrt{d}}{16}$), we have:

$$\mathbb{E}\left[\frac{\sigma_{1:S}}{16d^{3/2}} + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle\right] \ge \mathbb{E}\left[\frac{1}{32d^{3/2}}(d^2+\sigma_{1:S})\right].$$

To see this claim, suppose otherwise. Then $\mathbb{E}\left[\sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle\right] \le \mathbb{E}\left[\frac{\sqrt{d}}{32} - \frac{1}{32d^{3/2}}\sigma_{1:S}\right]$. This in turn implies (recalling $\alpha = 1/16d^{3/2}$ and $\eta = 4$ and that $\|\hat{c}_{1:S}\| \le \frac{\alpha(d^2+\sigma_{1:S})}{\eta}$ by definition of S):

$$\mathbb{E}\left[\frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \|\hat{c}_{1:S}\| + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle\right] \le \mathbb{E}\left[\frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \frac{\alpha(d^2+\sigma_{1:S})}{\eta} + \frac{\sqrt{d}}{32} - \frac{1}{32d^{3/2}}\sigma_{1:S}\right]$$
$$= \mathbb{E}\left[\frac{\sqrt{d^2+\sigma_{1:S}}}{4} + \frac{(d^2+\sigma_{1:S})}{64d^{3/2}} + \frac{\sqrt{d}}{32} - \frac{1}{32d^{3/2}}\sigma_{1:S}\right] = \mathbb{E}\left[\frac{\sqrt{d^2+\sigma_{1:S}}}{4} - \frac{(d^2+\sigma_{1:S})}{64d^{3/2}} + \frac{\sqrt{d}}{16}\right] \le \frac{\sqrt{d}}{16},$$

where the last inequality follows from the assumption $\sqrt{d^2+\sigma_{1:S}} > \frac{10}{\alpha} = 160d^{3/2}$. But, this contradicts our assumption $\mathbb{E}\left[\frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \|\hat{c}_{1:S}\| + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle\right] > \frac{\sqrt{d}}{16}$.

Now, we set $\lambda = 3\cdot 4\cdot(32\cdot 52)^2\cdot d + (32\cdot 52)\sqrt{3dB} + \sqrt{3}\cdot 32\cdot 52\cdot d^{3/2}$ and $\delta = \frac{\lambda}{\sqrt{4\lambda+B+d^2+\sigma_{1:S}}} = D_S$. Notice that $\delta < 1$ since we previously dispensed with the case $\lambda \ge \sqrt{4\lambda+B+d^2+\sigma_{1:S}}$. Then we have:

$$\lambda^2(d^2+\sigma_{1:S}) \ge 3\cdot 4\cdot(32\cdot 52)^2 d\lambda(d^2+\sigma_{1:S}) \ge 3\cdot 4\cdot(32\cdot 52)^2 d^3\lambda,$$
$$\lambda^2(d^2+\sigma_{1:S}) \ge 3\cdot(32\cdot 52)^2\cdot dB(d^2+\sigma_{1:S}) \ge 3\cdot(32\cdot 52)^2 Bd^3,$$
$$\lambda^2(d^2+\sigma_{1:S}) \ge 3\cdot(32\cdot 52)^2 d^3\cdot d^2(d^2+\sigma_{1:S}) \ge 3\cdot(32\cdot 52)^2 d^3\sigma_{1:T}.$$

Putting these together yields:

$$\lambda^2(d^2+\sigma_{1:S}) \ge 4\cdot(32\cdot 52)^2\lambda d^3 + (32\cdot 52)^2 Bd^3 + (32\cdot 52)^2 d^3\sigma_{1:S},$$
$$\implies \lambda\sqrt{d^2+\sigma_{1:S}} \ge (32\cdot 52)d^{3/2}\sqrt{4\lambda+B+\sigma_{1:S}}.$$

Thus, we have:

$$\delta\frac{\sigma_{1:S}}{16d^{3/2}} + \delta\sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle \ge \frac{\delta}{32d^{3/2}}(d^2+\sigma_{1:S}) = \frac{\lambda(d^2+\sigma_{1:S})}{32d^{3/2}\sqrt{4\lambda+B+\sigma_{1:T}}} \ge 52\sqrt{d^2+\sigma_{1:S}}.$$

So that in this last case we obtain:

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle c_t, x_t - u\rangle\right] \le \mathbb{E}\left[1 + \gamma\log(T) + \frac{16d^2B}{\gamma} + \frac{80+16\gamma/d}{\alpha} + \left(\frac{32}{\alpha}+4\right)\log\left(1+\frac{\sigma_{1:t}}{d^2}\right)\right]$$

$$
\begin{aligned}
&+ \mathbb{E}\left[\frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \|\hat{c}_{1:S}\| + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle + \lambda + 2\lambda\log(4\lambda+B+T) + 4\sqrt{B} - \delta\frac{\sigma_{1:S}}{d^2} - \delta\sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t\rangle\right] \\
&\le \mathbb{E}\left[1 + \gamma\log(T) + \frac{16d^2B}{\gamma} + \frac{80+16\gamma/d}{\alpha} + \left(\frac{32}{\alpha}+4\right)\log\left(1+\frac{\sigma_{1:t}}{d^2}\right)\right] \\
&\quad + \mathbb{E}\left[52\sqrt{d^2+\sigma_{1:S}} + \lambda + 2\lambda\log(4\lambda+B+T) + 4\sqrt{B} - \delta\frac{\sigma_{1:S}}{d^2} - \delta\sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t\rangle\right] \\
&\le \mathbb{E}\left[1 + \gamma\log(T) + \frac{16d^2B}{\gamma} + \frac{80+16\gamma/d}{\alpha} + \left(\frac{32}{\alpha}+4\right)\log\left(1+\frac{\sigma_{1:t}}{d^2}\right)\right] \\
&\quad + \mathbb{E}\left[\lambda + 2\lambda\log(4\lambda+B+T) + 4\sqrt{B}\right]. \qquad (7)
\end{aligned}
$$

Putting (4), (5), (6), and (7) together, we have:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\langle c_t, x_t - u\rangle\right] &\le \mathbb{E}\left[1 + \gamma\log(T) + \frac{16d^2B}{\gamma} + \frac{80+16\gamma/d}{\alpha} + \left(\frac{32}{\alpha}+4\right)\log\left(1+\frac{\sigma_{1:t}}{d^2}\right)\right] \\
&\quad + \mathbb{E}\left[\inf_{\delta}\left\{\frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \|\hat{c}_{1:S}\| + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle + \lambda + 2\lambda\log(4\lambda+B+T) + 4\sqrt{B} - \delta\frac{\sigma_{1:S}}{d} - \delta\sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t\rangle\right\}\right] \\
&\le \mathbb{E}\left[1 + \frac{\sqrt{d}}{12} + \gamma\log(T) + \frac{16d^2B}{\gamma} + \frac{8400+16\gamma/d}{\alpha} + \left(\frac{8\eta}{\alpha}+4\right)\log\left(1+\frac{\sigma_{1:t}}{d^2}\right)\right] \\
&\quad + \mathbb{E}\left[53\lambda + 2\lambda\log(4\lambda+B+T) + 4\sqrt{B}\right].
\end{aligned}
$$

Recalling $\lambda = 3\cdot 4\cdot(24\cdot 52)^2\cdot d + (24\cdot 52)\sqrt{3dB} + \sqrt{3}\cdot 24\cdot 52\cdot d^{3/2}$, $\eta = 4$, $\alpha = 1/16d^{3/2}$ and $\gamma = d^{3/4}\sqrt{B}$, we obtain the desired result. □

Lemma C.1. *Let $(p_1^*, \ldots, p_T^*)$ be any sequence such that for some time index S, we have $p_t^* = p_* \le D_S, \forall t \le S$, and $p_t^* = 0$ for all $t > S$. Then:*

$$
\sum_{t=1}^{T} v_t(p_t - p_t^\star) \le 2\lambda + 2\lambda\log(4\lambda+B+T) + \sqrt{B}.
$$

Proof. Observe that $|v_t| \le 2$. Then $|\hat{v}_t - v_t| \le 2$, and further $|\hat{v}_t - v_t| = 0$ for all but B indices t. Following the standard analysis of online gradient descent yields:

$$
\begin{aligned}
(p_{t+1} - p_t^\star)^2 &\le (p_t - \eta_t\hat{v}_t - p_t^\star)^2 = (p_t - p_t^\star)^2 - 2\eta_t\hat{v}_t(p_t - p_t^\star) + \eta_t^2\hat{v}_t^2, \\
\hat{v}_t(p_t - p_t^\star) &\le \frac{(p_t - p_t^\star)^2}{2\eta_t} - \frac{(p_{t+1} - p_t^\star)^2}{2\eta_t} + \frac{\eta_t\hat{v}_t^2}{2}, \\
v_t(p_t - p_t^\star) &\le \frac{(p_t - p_t^\star)^2}{2\eta_t} - \frac{(p_{t+1} - p_t^\star)^2}{2\eta_t} + \frac{\eta_t\hat{v}_t^2}{2} + D_t|\hat{v}_t - v_t|.
\end{aligned}
$$

Using these,

$$
\begin{aligned}
\sum_{t=1}^{T} v_t(p_t - p_t^\star) &\le \frac{(p_1 - p_1^\star)^2}{2\eta_1} + \sum_{t=2}^{T}(p_t - p_t^\star)^2\left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}}\right) + \sum_{t=1}^{T-1}\frac{(p_{t+1} - p_{t+1}^\star)^2 - (p_{t+1} - p_t^\star)^2}{2\eta_t} \\
&\quad + \sum_{t=1}^{T}\frac{\eta_t\hat{v}_t^2}{2} + D_1\sum_{t=1}^{T}|\hat{v}_t - v_t| \\
&\le \frac{D_1^2}{2\eta_1} + \sum_{t=2}^{T}D_t^2\left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}}\right) + \frac{p_{S+1}^2 - (p_{S+1} - p_S^\star)2}{2\eta_S} + \sum_{t=1}^{T}\frac{\eta_t\hat{v}_t^2}{2} + BD_1
\end{aligned}
$$

$$\leq \frac{D_1^2}{2\eta_1} + \sum_{t=2}^{T} D_t^2 \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) + \frac{D_{S+1}^2}{2\eta_S} + \sum_{t=1}^{T} \frac{\eta_t \hat{v}_t^2}{2} + BD_1.$$

We bound each term in the RHS separately as follows:

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t \hat{v}_t^2 &= \lambda \sum_{t=1}^{T} \frac{\hat{v}_t^2}{4\lambda + B + \sum_{i=1}^{t} \hat{v}_i^2} \leq \lambda \log\left(4\lambda + B + \hat{v}_{1:T}^2\right), \\
D_1 B &\leq \sqrt{B}, \\
\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} &= \frac{\hat{v}_t^2}{\lambda}, \\
\frac{D_1^2}{2\eta_1} + \frac{D_{S+1}^2}{2\eta_S} + \sum_{t=2}^{T} D_t^2 \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) &\leq \frac{\lambda}{2} \left(\frac{4\lambda + B + \hat{v}_1^2}{4\lambda + B} + 1 + \sum_{t=2}^{T} \frac{\hat{v}_t^2}{4\lambda + B + \hat{v}_{1:t}^2} \right) \\
&\leq \frac{\lambda}{2} \left(\frac{4\lambda + B + \hat{v}_1^2}{4\lambda + B} + 1 + \log\left(4\lambda + B + \hat{v}_{1:T}^2\right) \right) \\
&\leq 2\lambda + \lambda \log(4\lambda + B + T).
\end{aligned}$$

Putting all these together shows the claim. □

Proposition C.2. *In Algorithm 2, the following hold for all $t \in [T]$:*

(i) $\|x_t\| \leq 1$.
(ii) $\sigma_t \leq d^2$.
(iii) $\mathbb{E}[\sigma_t] \leq d$.
(iv) $\mathbb{E}[\hat{c}_t | x_1, \ldots, x_t] = c_t$.
(v) If $g_t = 1$, $\mathbb{E}[\langle h_t, c_t \rangle] \leq \|c_t\|^2 / 4\sqrt{d}$.
(vi) When $g_t = 0$, $\hat{c}_t^\top \nabla^2 \phi(\bar{x}_t)^{-1} \hat{c}_t \leq 4d^2$.
(vii) For all $\|x\| \leq 1$, $\sqrt{x^\top \nabla^2 \phi(x)^{-1} x} \leq \frac{1}{2}$.
(viii) $\delta_t \leq \frac{\sigma_t}{\eta \sqrt{d^2 + \sigma_{1:t}}}$.

Proof. (i) Since $|\alpha_t| \leq \frac{4}{\sqrt{d}}$, it is clear that $\|h_t\| \leq 1$. Therefore, it suffices to show $\|z_t\| \leq 1$. Now, observe ϕ is a self-concordant barrier for the unit ball, and that by definition z_t is on the Dikin ellipsoid centered at $\bar{x}_t$, so that $\|z_t\| \leq 1$.

(ii) By definition, $\sigma_t = g_t \|\hat{c}_t\|^2$. If $g_t = 0$, we have $\sigma_t = 0$. Otherwise, we have $\sigma_t = \|\hat{c}_t\|^2 = d^2 q_t^2 \leq d^2$.

(iii) We have $\mathbb{E}[\sigma_t] \leq \mathbb{E}[d^2 c_t^\top s_t s_t^\top c_t] = \mathbb{E}\left[d^2 \frac{c_t^\top I c_t}{d} \right] \leq d$.

(iv) When $g_t = 1$, we have $q_t = \langle c_t, s_t \rangle$ so that $\mathbb{E}[\hat{c}_t] = \mathbb{E}\left[d \cdot q_t \cdot s_t \right] = c_t$ by Lemma 4.2. When $g_t = 0$, we instead have:

$$\begin{aligned}
\mathbb{E}[\hat{c}_t] &= \mathbb{E}\left[d \frac{\langle -p_t h_t + (1 - p_t)\bar{x}_t), c_t \rangle \nabla^2 \phi(\bar{x}_t)^{1/2} w_t}{1 - p_t} \right] + \mathbb{E}[d \nabla^2 \phi(\bar{x}_t)^{1/2} w_t w_t^\top \nabla^2 \phi(\bar{x}_t)^{-1/2} c_t] \\
&= \mathbb{E}\left[d \nabla^2 \phi(\bar{x}_t)^{1/2} \frac{I}{d} \nabla^2 \phi(\bar{x}_t)^{-1/2} c_t \right] = c_t.
\end{aligned}$$

(v) When $g_t = 1$, we have

$$\mathbb{E}[\langle c_t, h_t \rangle] = \frac{\sqrt{d}}{4} \mathbb{E}\left[\langle c_t, s_t \rangle \cdot \mathrm{clip}_{\frac{4}{\sqrt{d}}} \left(\langle c_t, s_t \rangle \right) \right] \leq \frac{\sqrt{d}}{4} \mathbb{E}[\langle c_t, s_t \rangle^2] = \frac{\sqrt{d}}{4} \mathbb{E}[c_t \top s_t s_t^\top c_t] = \frac{\sqrt{d}}{4} \mathbb{E}\left[c_t^\top \frac{I}{d} c_t \right] = \frac{\|c_t\|^2}{4\sqrt{d}}.$$

(vi) We have

$$\hat{c}_t^\top \nabla^2 \phi(\bar{x}_t)^{-1} \hat{c}_t = d^2 \frac{\ell_t^2}{(1 - p_t)^2} w_t^\top \nabla^2 \phi(\bar{x}_t)^{1/2} \nabla^2 \phi(\bar{x}_t)^{-1} \nabla^2 \phi(\bar{x}_t)^{1/2} w_t \leq 4d^2 w_t^\top w_t = 4d^2,$$

where the inequality follows from $|\ell_t| \leq 1$ and $p_t \in [0, 1/2]$.

(vii) A simple calculation shows:

$$\nabla^2\phi(x) = \frac{2I}{1-\|x\|^2} + \frac{4xx^\top}{(1-\|x\|^2)^2}.$$

Clearly, x is an eigenvector of this matrix with eigenvalue $\frac{2+2\|x\|^2}{(1-\|x\|^2)^2}$. Thus,

$$\sqrt{x^\top \nabla^2\phi(x)^{-1}x} = \sqrt{\frac{\|x\|^2(1-\|x\|^2)^2}{2+2\|x\|^2}}.$$

Numerical evaluation of this expression for $\|x\| \in [0,1]$ shows that its maximum is less than 0.5.

(viii) By concavity of the square root function:

$$\delta_t = \frac{\sqrt{d^2+\sigma_{1:t}} - \sqrt{d^2+\sigma_{1:t-1}}}{\eta} \le \frac{\sigma_t}{2\eta\sqrt{d^2+\sigma_{1:t-1}}} \le \frac{\sigma_t}{\eta\sqrt{4d^2+\sigma_{1:t-1}}} \le \frac{\sigma_t}{\eta\sqrt{d^2+\sigma_{1:t}}},$$

where the last step follows from $\sigma_t \le d^2$. □

Next, we obtain a bound on the regret incurred by Algorithm 2. Since Algorithm 2 is an instance of the classic FTRL, we can utilize tools from (Bhaskara et al., 2021) for a tight analysis of FTRL.

Lemma C.3. *For any S and $\|u\| \le 1$:*

$$\sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t - u\rangle \le \gamma\log(S) + 13\left(\eta + \frac{1}{\eta}\right)\sqrt{d^2+\sigma_{1:S}} + \frac{4d^2B}{.}\gamma$$

Proof. First, define $\hat{u}$ to be the projection of u to the ball of radius $1-1/S$. Then clearly we have:

$$\mathbb{E}\left[\sum_{t=1}^{S}\langle c_t, \bar{x}_t - u\rangle\right] \le \mathbb{E}\left[\sum_{t=1}^{S}\langle c_t, \bar{x}_t - \hat{u}\rangle + \frac{\|c_{1:S}\|}{T}\right] \le \mathbb{E}\left[1 + \sum_{t=1}^{S}\langle c_t, \bar{x}_t - \hat{u}\rangle\right] = \mathbb{E}\left[1 + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t - \hat{u}\rangle\right]. \quad (8)$$

We now focus on bounding $\sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t - \hat{u}\rangle$. From (Bhaskara et al., 2021, Lemma B.2), i.e., the FTRL Lemma, we have:

$$r_{0:S}(\bar{x}_{S+1}) + \gamma\phi(\bar{x}_{S+1}) + \langle \hat{c}_{1:S}, \bar{x}_{S+1}\rangle + \sum_{t=1}^{S} r_t(\bar{x}_{t+1}) + \langle \hat{c}_t, \bar{x}_{t+1}\rangle \le r_{0:t}(u) + \gamma\phi(\hat{u}) + \langle \hat{c}_{1:S}, u\rangle,$$

so that we have:

$$\sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t - \hat{u}\rangle \le \gamma\phi(\hat{u}) + r_{0:S}(\hat{u}) + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1}\rangle \le \gamma\log(S) + \frac{\sqrt{d^2+\sigma_{1:S}}}{2\eta} + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1}\rangle.$$

Now by the standard FTRL Lemma, e.g., (Bhaskara et al., 2021, Lemma B.1), we have:

$$\sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t - \hat{u}\rangle \le \gamma\phi(\hat{u}) + r_{0:T}(u) + \sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1}\rangle \le \gamma\log(T) + \frac{\sqrt{d^2+\sigma_{1:T}}}{2\eta} + \sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1}\rangle.$$

By Lemma C.8:

$$\langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1}\rangle \le \|\hat{c}_t\|_{\nabla^2\psi(\bar{x}_t)^{-1}}\|\bar{x}_t - \bar{x}_{t+1}\|_{\nabla^2\psi(\bar{x}_t)} \le 4\|\hat{c}_t\|^2_{\nabla^2\psi(\bar{x}_t)^{-1}} + \frac{4\sigma_t\|\hat{c}_t\|_{\nabla^2\psi(\bar{x}_t)^{-1}}}{\sqrt{\eta}(d^2+\sigma_{1:t})^{3/4}}.$$

When $g_t = 0$, we have $\sigma_t = 0$ and hence

$$\langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1}\rangle \le 4\|\hat{c}_t\|^2_{\nabla^2\psi(\bar{x}_t)^{-1}}.$$

We also have by Proposition C.2, $\|\hat{c}_t\|^2_{\nabla^2\psi(\bar{x}_t)^{-1}} \le \frac{4d^2}{\gamma}$, and therefore

$$\sum_{g_t=0}\langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1}\rangle \le \frac{4d^2B}{\gamma}.$$

When $g_t = 1$, since $\nabla^2\psi(\bar{x}_t) \succeq \frac{d^2+\sigma_{1:t}}{\eta} I$, we have:

$$\|\hat{c}_t\|^2_{\nabla^2\psi(\bar{x}_t)^{-1}} \le \frac{\eta\|\hat{c}_t\|^2}{\sqrt{d^2+\sigma_{1:t}}}.$$

Therefore:

$$\langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1}\rangle \le 4\|\hat{c}_t\|^2_{\nabla^2\psi(\bar{x}_t)^{-1}} + \frac{4\sigma_t\|\hat{c}_t\|_{\nabla^2\psi(\bar{x}_t)^{-1}}}{\sqrt{\eta}(d^2+\sigma_{1:t})^{3/4}} \le \frac{4\eta\sigma_t}{\sqrt{d^2+\sigma_{1:t}}} + \frac{4\sigma_t}{\sqrt{d^2+\sigma_{1:t}}}\frac{\sqrt{\sigma_t}}{\sqrt{d^2+\sigma_{1:t}}} \le \frac{(4\eta+4)\sigma_t}{\sqrt{d^2+\sigma_{1:t}}}.$$

Thus:

$$\sum_{g_t=1}\langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1}\rangle \le \sum_{t=1}^{S}\frac{(4\eta+4)\sigma_t}{\sqrt{d^2+\sigma_{1:t}}} \le (8\eta+8)\sqrt{d^2+\sigma_{1:T}} \le \left(12\eta + \frac{12}{\eta}\right)\sqrt{d^2+\sigma_{1:T}},$$

where the last step follows since $1 \le \frac{\eta}{2} + \frac{1}{2\eta}$. □

Lemma C.4. *Suppose $\gamma \ge 1$, let S be the smallest index such that $\|\hat{c}_{1:t}\| \ge \alpha\frac{d^2+\sigma_{1:t}}{\eta}$ for all $t > S$ and let $\|u\| \le 1$ (note that S is a random variable). Then Algorithm 2 ensures:*

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t - u\rangle\right] \le \mathbb{E}\left[1 + \frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \gamma\log(T) + \frac{16d^2B}{\gamma} + \frac{16(\eta+1+\gamma/d)}{\alpha} + \left(\frac{8\eta}{\alpha}+4\right)\log\left(1+\frac{\sigma_{1:t}}{d^2}\right)\right] + \mathbb{E}\left[\|\hat{c}_{1:S}\| + \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_t\rangle\right].$$

Proof. First, define $\hat{u}$ to be the projection of u to the ball of radius $1 - 1/T$. Then clearly we have:

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle c_t, \bar{x}_t - u\rangle\right] \le \mathbb{E}\left[\sum_{t=1}^{T}\langle c_t, \bar{x}_t - \hat{u}\rangle + \frac{\|c_{1:T}\|}{T}\right] \le \mathbb{E}[1 + \sum_{t=1}^{T}\langle c_t, \bar{x}_t - \hat{u}\rangle] = \mathbb{E}[1 + \sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t - \hat{u}\rangle]. \tag{9}$$

We now focus on bounding $\sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t - \hat{u}\rangle$. Since $\|\hat{c}_{1:S}\| \ge \sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_{S+1}\rangle$, it suffices to show that

$$\sum_{t=1}^{S}\langle \hat{c}_t, \bar{x}_{S+1} - \hat{u}\rangle + \sum_{t>S}\langle \hat{c}_t, \bar{x}_t - u\rangle \le \frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \gamma\log(T) + \frac{16d^2B}{\gamma} + \frac{16(\eta+\gamma)}{\alpha} + \left(\frac{8\eta}{\alpha}+4\right)\log\left(1+\frac{\sigma_{1:t}}{d^2}\right).$$

From (Bhaskara et al., 2021, Lemma B.2), i.e., the FTRL Lemma, we have:

$$r_{0:S}(\bar{x}_{S+1}) + \gamma\phi(\bar{x}_{S+1}) + \langle \hat{c}_{1:S}, \bar{x}_{S+1}\rangle + \sum_{t=1}^{T} r_t(\bar{x}_{t+1}) + \langle \hat{c}_t, \bar{x}_{t+1}\rangle \le r_{0:t}(u) + \gamma\phi(\hat{u}) + \langle \hat{c}_{1:T}, u\rangle.$$

Dropping negative terms and observing that $\phi(\hat{u}) \le -\log(1-(1-1/T)^2) = \log(T^2/(2T-1))) \le \log(T)$, we have:

$$\begin{aligned}\langle \hat{c}_{1:S}, \bar{x}_{S+1} - \hat{u}\rangle + \sum_{t>S}\langle \hat{c}_t, \bar{x}_t - \hat{u}\rangle &\le r_{0:S}(u) + \sum_{t>S} r_t(\hat{u}) - r_t(\bar{x}_{t+1}) + \gamma\phi(\hat{u}) + \sum_{t>S}\langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1}\rangle \\ &= \frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \gamma\log(T) + \sum_{t>S}\frac{\delta_t}{2}(\|\hat{u}\|^2 - \|\bar{x}_{t+1}\|^2) + \sum_{t>S}\langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1}\rangle\end{aligned}$$

$$\leq \frac{\sqrt{d^2+\sigma_{1:S}}}{\eta} + \gamma\log(T) + \sum_{t>S}\frac{\delta_t}{2}(1-\|\bar{x}_{t+1}\|^2) + \sum_{t>S}\langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1}\rangle. \quad (10)$$

Let us focus on bounding first $\sum_{t>S}\frac{\delta_t}{2}(1-\|\bar{x}_{t+1}\|^2)$. The high-level intuition is that for $t > S$, $\|\bar{x}_{t+1}\|$ is close to 1, so that the $\|\hat{u}\|^2 - \|\bar{x}_{t+1}\|^2$ is very small.

To get started on this, we need to understand $\|\bar{x}_{t+1}\|$. To this end, observe that since $r_{0:t}(x) + \gamma\phi(x)$ is a radially-symmetric function that achieves its minimum at the origin, we must have that $\bar{x}_{t+1} = -k\frac{\hat{c}_{1:t}}{\|\hat{c}_{1:t}\|}$ for $k = \|\bar{x}_{t+1}\|$. Further, since ϕ is a barrier function, by first-order optimality conditions we have:

$$\begin{aligned}
\hat{c}_{1:t} + \nabla r_{0:t}(\bar{x}_{t+1}) + \gamma\nabla\phi(\bar{x}_{t+1}) &= 0,\\
\hat{c}_{1:t} + \frac{\bar{x}_{t+1}}{\eta}\sqrt{d^2+\sigma_{1:t}} + 2\gamma\frac{\bar{x}_{t+1}}{1-k^2} &= 0,\\
-\|\hat{c}_{1:t}\| + \frac{k}{\eta}\sqrt{d^2+\sigma_{1:t}} + 2\gamma\frac{k}{1-k^2} &= 0.
\end{aligned}$$

Let M be the smallest index greater than S such that $\sqrt{d^2+\sigma_{1:M}} \geq \frac{4}{\alpha}$. Then for $t \geq M$, we have

$$d + \sum_{i=1}^{t}\frac{\sigma_i}{\sqrt{d^2+\sigma_{1:i}}} \leq 2\sqrt{d^2+\sigma_{1:t}} \leq \frac{\alpha(d^2+\sigma_{1:t})}{2}.$$

Thus, for $t \geq M$, since $\|\hat{c}_{1:t}\| > \frac{\alpha(d^2+\sigma_{1:t})}{\eta}$, we have

$$2\gamma\frac{k}{1-k^2} = \|\hat{c}_{1:t}\| - \frac{k}{\eta}\sqrt{d^2+\sigma_{1:t}} \geq \|\hat{c}_{1:t}\| - \frac{1}{\eta}\sqrt{d^2+\sigma_{1:t}} \geq \frac{\alpha(d^2+\sigma_{1:t})}{2\eta},$$

where the first inequality follows since $k \leq 1$. Using $k \leq 1$ again, we obtain

$$2\gamma\frac{1}{1-k^2} \geq \frac{\alpha(d^2+\sigma_{1:t})}{2\eta} \implies 1-k^2 \leq \frac{4\eta\gamma}{\alpha(d^2+\sigma_{1:t})}.$$

Therefore, using Proposition C.2 which tells us that $\frac{\delta_t}{2} \leq \frac{\sigma_t}{\eta\sqrt{d^2+\sigma_{1:t}}}$, we have:

$$\begin{aligned}
\sum_{t>S}\frac{\delta_t}{2}(1-\|\bar{x}_{t+1}\|^2) &\leq \sum_{t=S+1}^{M-1}\frac{\sqrt{d^2+\sigma_{1:t}} - \sqrt{d^2+\sigma_{1:t-1}}}{2\eta}(1-\|\bar{x}_{t+1}\|^2) + \sum_{t\geq M}\frac{\sigma_t}{2\eta\sqrt{d^2+\sigma_{1:t}}}(1-\|\bar{x}_{t+1}\|^2)\\
&\leq \sum_{t=S+1}^{M-1}\frac{\sqrt{d^2+\sigma_{1:t}} - \sqrt{d^2+\sigma_{1:t-1}}}{2\eta} + \sum_{t\geq M}\frac{2\gamma\sigma_t}{\alpha(d^2+\sigma_{1:t})^{3/2}}\\
&\leq \sqrt{d^2+\sigma_{1:M-1}} + \frac{2\gamma}{\alpha}\int_{d^2}^{\infty}\frac{dx}{x^{3/2}} \;\leq\; \frac{4}{\alpha} + \frac{4\gamma}{\alpha d}, \quad (11)
\end{aligned}$$

where the last step follows from the definition of M. We focus next on bounding $\sum_{t>S}\langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1}\rangle$. To this end, observe that by Lemma C.8, we have:

$$\|\bar{x}_t - \bar{x}_{t+1}\|_{\nabla^2\psi_t(\bar{x}_t)} \leq 4\|\hat{c}_t + \nabla r_t(\bar{x}_t))\|_{\nabla^2\psi_t(\bar{x}_t)^{-1}} \leq 4\|\hat{c}_t\|_{\nabla^2\psi_t(\bar{x}_t)^{-1}} + \frac{4\sigma_t}{\sqrt{\eta}(d^2+\sigma_{1:t})^{3/4}},$$

where $\|x\|_A^2 = x^\top Ax$ for any matrix A. Thus,

$$\langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1}\rangle \leq \|\hat{c}_t\|_{\nabla^2\psi_t(\bar{x}_t)^{-1}}\|\bar{x}_t - \bar{x}_{t+1}\|_{\nabla^2\psi_t(\bar{x}_t)} \leq 4\|\hat{c}_t\|^2_{\nabla^2\psi_t(\bar{x}_t)^{-1}} + \frac{4\sigma_t\|\hat{c}_t\|_{\nabla^2\psi_t(\bar{x}_t)^{-1}}}{\sqrt{\eta}(d^2+\sigma_{1:t})^{3/4}}. \quad (12)$$

Now since $\nabla^2\psi_t(\bar{x}_t) = \frac{\sqrt{d^2+\sigma_{1:t}}}{\eta}I + \gamma\nabla^2\phi(\bar{x}_t)$, we can apply Proposition C.2 to obtain:

$$\|\hat{c}_t\|^2_{\nabla^2\psi_t(\bar{x}_t)^{-1}} \leq \frac{1}{\gamma}\hat{c}_t\nabla^2\phi(\bar{x}_t)\hat{c}_t \leq \frac{4d^2}{\gamma}.$$

Whenever $g_t = 0$, we have $\sigma_t = 0$ and hence (12) yields:

$$\sum_{t>S, g_t=0} \langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1} \rangle \leq \sum_{g_t=0} \frac{16d^2}{\gamma} \leq \frac{16d^2 B}{\gamma}. \tag{13}$$

For the time steps when $g_t = 1$, we have:

$$\|\hat{c}_t\|^2_{\nabla^2 \psi_t(\bar{x}_t)^{-1}} \leq \frac{\eta \|\hat{c}_t\|^2}{\sqrt{d^2 + \sigma_{1:t}}} \leq \frac{\eta}{\sqrt{d^2 + \sigma_{1:t}}}.$$

Using this in (12), we obtain:

$$\langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1} \rangle \leq 4\|\hat{c}_t\|^2_{\nabla^2 \psi_t(\bar{x}_t)^{-1}} + \frac{4\sigma_t}{d^2 + \sigma_{1:t}}.$$

Since

$$\sum_{t=1}^{T} \frac{4\sigma_t}{d^2 + \sigma_{1:t}} \leq 4 \log \left(1 + \frac{\sigma_{1:t}}{d^2}\right), \tag{14}$$

it remains to bound:

$$\sum_{t>S,\ g_t=0} 4\|\hat{c}_t\|^2_{\nabla^2 \psi_t(\bar{x}_t)^{-1}}.$$

Recall that by definition of M, we have $\sqrt{d^2 + \sigma_{1:M-1}} \leq \frac{4}{\alpha}$. Thus:

$$\begin{aligned} \sum_{S<t<M,\ g_t=1} \langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1} \rangle &\leq \sum_{t=1}^{M-1} 4\|\hat{c}_t\|^2_{\nabla^2 \psi_t(\bar{x}_t)^{-1}} \leq \sum_{t=1}^{M-1} \frac{4\eta \|\hat{c}_t\|^2}{\sqrt{d^2 + \sigma_{1:t}}} = \sum_{t=1}^{M-1} \frac{4\eta \sigma_t}{\sqrt{d^2 + \sigma_{1:t}}} \\ &\leq 4\eta \sqrt{2d^2 + 2\sigma_{1:M-1}} \leq \frac{16\eta}{\alpha}. \end{aligned} \tag{15}$$

Thus, we need to bound the remaining sum:

$$\sum_{t\geq M,\ g_t=1} \langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1} \rangle \leq \sum_{S<t<M,\ g_t=1} 4\|\hat{c}_t\|^2_{\nabla^2 \psi_t(\bar{x}_t)^{-1}}.$$

For these, we recall that $1 - \|\bar{x}_t\|^2 \leq \frac{4\eta\gamma}{\alpha(d^2+\sigma_{1:t})}$ for all $t \geq M$. Then, we have:

$$\nabla^2 \phi(\bar{x}_t) = 2\frac{I}{1 - \|\bar{x}_t\|^2} + 4\frac{\bar{x}_t \bar{x}_t^\top}{(1 - \|\bar{x}_t\|^2)^2} \succeq \frac{\alpha(d^2 + \sigma_{1:t})}{2\eta} I.$$

Therefore for all $t \geq M$ and $g_t = 1$:

$$\begin{aligned} 4\|\hat{c}_t\|^2_{\nabla^2 \psi_t(\bar{x}_t)^{-1}} &\leq 8\eta \frac{\|\hat{c}_t\|^2}{\alpha(d^2 + \sigma_{1:t})} = 8\eta \frac{\sigma_t}{\alpha(d^2 + \sigma_{1:t})} \\ \implies \sum_{t\geq M, g_t=1} 4\|\hat{c}_t\|^2_{\nabla^2 \psi_t(\bar{x}_t)^{-1}} &\leq \frac{8\eta}{\alpha} \log \left(1 + \frac{\sigma_{1:t}}{d^2}\right). \end{aligned} \tag{16}$$

Combining (10, 11, 13, 14,15, 16), we obtain:

$$\begin{aligned} &\langle \hat{c}_{1:S}, \bar{x}_{S+1} - \hat{u} \rangle + \sum_{t>S} \langle \hat{c}_t, \bar{x}_t - \hat{u} \rangle \\ &\leq \frac{\sqrt{1 + \sigma_{1:S}}}{\eta} + \gamma \log(T) + \sum_{t>S} \frac{\delta_t}{2}(1 - \|\bar{x}_{t+1}\|^2) + \sum_{t>S} \langle \hat{c}_t, \bar{x}_t - \bar{x}_{t+1} \rangle \end{aligned}$$

$$\leq \frac{\sqrt{1+\sigma_{1:S}}}{\eta} + \gamma\log(T) + \frac{4+\gamma/d}{\alpha} + \frac{16d^2B}{\gamma} + 4\log\left(1+\frac{\sigma_{1:t}}{d^2}\right) + \frac{16\eta}{\alpha} + \frac{8\eta}{\alpha}\log\left(1+\frac{\sigma_{1:t}}{d^2}\right)$$
$$\leq \frac{\sqrt{1+\sigma_{1:S}}}{\eta} + \gamma\log(T) + \frac{16d^2B}{\gamma} + \frac{16(\eta+1+\gamma/d)}{\alpha} + \left(\frac{8\eta}{\alpha}+4\right)\log\left(1+\frac{\sigma_{1:t}}{d^2}\right).$$

So, finally adding back the 1 from (9), we have:

$$\sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t - u\rangle \leq 1 + \frac{\sqrt{1+\sigma_{1:S}}}{\eta} + \gamma\log(T) + \frac{16d^2B}{\gamma} + \frac{16(\eta+1+\gamma/d)}{\alpha} + \left(\frac{8\eta}{\alpha}+4\right)\log\left(1+\frac{\sigma_{1:t}}{d^2}\right)$$
$$+ \|\hat{c}_{1:S}\| + \sum_{t=1}^{T}\langle \hat{c}_t, \bar{x}_t\rangle. \qquad \square$$

For the remainder of this section, let ϕ be a self-concordant barrier on a space Ω. For a symmetric positive-definite matrix M, define the norm $\|h\|_M = h^\top M h$. The following is a standard fact about self-concordant barriers:

Proposition C.5. *For all $x, x' \in \Omega$ with $\|x - x'\|_{\nabla^2\phi(x)} < 1$ and all vectors h:*

$$\|h\|_{\nabla^2\phi(x)}(1 - \|x-x'\|_{\nabla^2\phi(x)}) \leq \|h\|_{\nabla^2\phi(x')} \leq \frac{\|h\|_{\nabla^2\phi(x)}}{1-\|x-x'\|_{\nabla^2\phi(x)}}.$$

This proposition has the following immediate corollary:

Corollary C.6. *For all $x, x' \in \Omega$ with $\|x - x'\|_{\nabla^2\phi(x)} < 1$, we have:*

$$\|h\|_{\nabla^2\phi^{-1}(x)}(1 - \|x-x'\|_{\nabla^2\phi(x)}) \leq \|h\|_{\nabla^2\phi(x')^{-1}} \leq \frac{\|h\|_{\nabla^2\phi(x)^{-1}}}{1-\|x-x'\|_{\nabla^2\phi(x)}}.$$

Proof. Set $z' = \frac{\nabla^2\phi(x')^{-1}h}{\|h\|_{\nabla^2\phi(x')^{-1}}}$. Observe that $\langle z', h\rangle = \|h\|_{\nabla^2\phi(x')^{-1}}$ and also $\|z'\|_{\nabla^2\phi(x')} = 1$. Now, by the Cauchy–Schwarz inequality,

$$\|h\|_{\nabla^2\phi(x')^{-1}} = \langle z', h\rangle \leq \|h\|_{\nabla^2\phi(x)^{-1}}\|z'\|_{\nabla^2\phi(x)} \leq \|h\|_{\nabla^2\phi(x)^{-1}}\frac{\|z'\|_{\nabla^2\phi(x')}}{1-\|x-x\|_{\nabla^2\phi(x)}} = \frac{\|h\|_{\nabla^2\phi(x)^{-1}}}{1-\|x-x\|_{\nabla^2\phi(x)}}.$$

Similarly, let $z = \frac{\nabla^2\phi(x)^{-1}h}{\|h\|_{\nabla^2\phi(x)^{-1}}}$, so that $\langle z, h\rangle = \|h\|_{\nabla^2\phi(x)^{-1}}$ and also $\|z\|_{\nabla^2\phi(x)} = 1$. Then:

$$\|h\|_{\nabla^2\phi(x)^{-1}} = \langle z, h\rangle \leq \|h\|_{\nabla^2\phi(x')^{-1}}\|z'\|_{\nabla^2\phi(x')} \leq \|h\|_{\nabla^2\phi(x')^{-1}}\frac{\|z'\|_{\nabla^2\phi(x)}}{1-\|x-x\|_{\nabla^2\phi(x)}} = \frac{\|h\|_{\nabla^2\phi(x')^{-1}}}{1-\|x-x\|_{\nabla^2\phi(x)}}$$
$$\implies \|h\|_{\nabla^2\phi(x)^{-1}}(1-\|x-x\|_{\nabla^2\phi(x)}) \leq \|h\|_{\nabla^2\phi(x')^{-1}}. \qquad \square$$

Now, we can generalize this:

Proposition C.7. *Let $\psi(x) = \frac{\lambda}{2}\|x\|^2 + \gamma\phi(x)$. Then for all $x, x' \in \Omega$ with $\|x - x'\|_{\nabla^2\phi(x)} < 1$ and all vectors h:*

$$\|h\|_{\nabla^2\psi(x)}(1 - \|x-x'\|_{\nabla^2\phi(x)}) \leq \|h\|_{\nabla^2\psi(x')} \leq \frac{\|h\|_{\nabla^2\psi(x)}}{1-\|x-x'\|_{\nabla^2\phi(x)}}.$$

Proof. By Proposition C.5:

$$\|h\|^2_{\nabla^2\phi(x)}(1 - \|x-x'\|_{\nabla^2\phi(x)})^2 \leq \|h\|^2_{\nabla^2\phi(x')} \leq \frac{\|h\|^2_{\nabla^2\phi(x)}}{(1-\|x-x'\|_{\nabla^2\phi(x)})^2}.$$

Further,

$$\|h\|^2_{\nabla^2\psi(x')} = \lambda\|h\|_2^2 + \gamma\|h\|^2_{\nabla^2\phi(x')}.$$

Combining these observations:

$$\lambda\|h\|^2 + \gamma\|h\|^2_{\nabla^2\phi(x)}(1 - \|x - x'\|_{\nabla^2\phi(x)})^2 \le \|h\|^2_{\nabla^2\psi(x')} \le \lambda\|h\|^2 + \gamma\frac{\|h\|^2_{\nabla^2\phi(x)}}{(1 - \|x - x'\|_{\nabla^2\phi(x)})^2}$$

observing that $1 - \|x - x'\|_{\nabla^2\phi(x)} \in [0, 1)$:

$$(\lambda\|h\|^2 + \gamma\|h\|^2_{\nabla^2\phi(x)})(1 - \|x - x'\|_{\nabla^2\phi(x)})^2 \le \|h\|^2_{\nabla^2\psi(x')} \le \frac{\lambda\|h\|^2 + \gamma\|h\|^2_{\nabla^2\phi(x)}}{(1 - \|x - x'\|_{\nabla^2\phi(x)})^2},$$

which implies the desired result. □

Now, we prove a some key bounds on $\bar{x}_t - \bar{x}_{t+1}$:

Lemma C.8. *Define $\psi_t(x) = r_{0:t}(x) + \gamma\phi(x)$ for $\phi(x) = -\log(1 - \|x\|^2)$. Then:*

$$\|\bar{x}_t - \bar{x}_{t+1}\|_{\nabla^2\psi_t(\bar{x}_t)} \le 4\|\hat{c}_t + \nabla r_t(\bar{x}_t))\|_{\nabla^2\psi_t(\bar{x}_t)^{-1}} \le 4\|\hat{c}_t\|_{\nabla^2\psi_t(\bar{x}_t)^{-1}} + \frac{4\sigma_t}{\sqrt{\eta}(d^2 + \sigma_{1:t})^{3/4}}.$$

Proof. By definition, we have:

$$\bar{x}_t = \operatorname{argmin}\langle\hat{c}_{1:t-1}, x\rangle + \psi_{t-1}(x) \quad \text{and} \quad \bar{x}_{t+1} = \operatorname{argmin}\langle\hat{c}_{1:t}, x\rangle + \psi_t(x).$$

Therefore, by since $\lim_{\|x\|\to 1}\psi_t(x) = \infty$, by first-order optimality conditions we have:

$$\nabla\psi_{t-1}(\bar{x}_t) = -\hat{c}_{1:t-1} \quad \text{and} \quad \nabla\psi_t(\bar{x}_{t+1}) = -\hat{c}_{1:t}.$$

By mean-value theorem, there are two points y and y' on the line segment connecting $\bar{x}_t$ and $\bar{x}_{t+1}$ such that:

$$\begin{aligned}
\psi_t(\bar{x}_t) &= \psi_t(\bar{x}_{t+1}) + \langle\nabla\psi_t(\bar{x}_{t+1}), \bar{x}_t - \bar{x}_{t+1}\rangle + \frac{\|\bar{x}_t - \bar{x}_{t+1}\|^2_{\nabla^2\psi_t(y)}}{2}.\\
\psi_t(\bar{x}_{t+1}) &= \psi_t(\bar{x}_t) + \langle\nabla\psi_t(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle + \frac{\|\bar{x}_t - \bar{x}_{t+1}\|^2_{\nabla^2\psi_t(y')}}{2}\\
&= \psi_t(\bar{x}_t) + \langle\nabla\psi_{t-1}(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle + \langle\nabla r_t(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle + \frac{\|\bar{x}_t - \bar{x}_{t+1}\|^2_{\nabla^2\psi_t(y')}}{2}.
\end{aligned}$$

Adding these equations and simplifying, we have:

$$\begin{aligned}
0 &= \langle\nabla\psi_{t-1}(\bar{x}_t) - \psi_t(\bar{x}_{t+1}) + \nabla r_t(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle + \frac{\|\bar{x}_t - \bar{x}_{t+1}\|^2_{\nabla^2\psi_t(y') + \nabla^2\psi_t(y)}}{2}\\
&\implies \langle\hat{c}_t + \nabla r_t(\bar{x}_t), \bar{x}_t - \bar{x}_{t+1}\rangle = \frac{\|\bar{x}_t - \bar{x}_{t+1}\|^2_{\nabla^2\psi_t(y')}}{2} + \frac{\|\bar{x}_t - \bar{x}_{t+1}\|^2_{\nabla^2\psi_t(y)}}{2}.
\end{aligned}$$

By Lemma C.9, $\|\bar{x}_t - \bar{x}_{t+1}\|^2_{\nabla^2\psi_t(\bar{x}_t)} \le \frac{1}{2}$, and so $\|\bar{x}_t - y\|^2_{\nabla^2\psi_t(\bar{x}_t)} \le \frac{1}{2}$ and $\|\bar{x}_t - y'\|^2_{\nabla^2\psi_t(\bar{x}_t)} \le \frac{1}{2}$. Thus by Proposition C.5:

$$\begin{aligned}
&\frac{\|\bar{x}_t - \bar{x}_{t+1}\|^2_{\nabla^2\psi_t(y')}}{2} + \frac{\|\bar{x}_t - \bar{x}_{t+1}\|^2_{\nabla^2\psi_t(y)}}{2} \ge \frac{\|\bar{x}_t - \bar{x}_{t+1}\|^2_{\nabla^2\psi_t(\bar{x}_t)}}{8} + \frac{\|\bar{x}_t - \bar{x}_{t+1}\|^2_{\nabla^2\psi_t(\bar{x}_t)}}{8} = \frac{1}{4}\|\bar{x}_t - \bar{x}_{t+1}\|^2_{\nabla^2\psi_t(\bar{x}_t)}\\
&\implies 4\langle\hat{c}_t + \nabla r_t(\bar{x}_t), \bar{x}_t - \bar{x}_{t+1}\rangle \ge \|\bar{x}_t - \bar{x}_{t+1}\|^2_{\nabla^2\psi_t(\bar{x}_t)}.
\end{aligned}$$

Now, applying the Cauchy–Schwarz inequality, we have:

$$\begin{aligned}
&\|\bar{x}_t - \bar{x}_{t+1}\|^2_{\nabla^2\psi_t(\bar{x}_t)} \le 4\|\hat{c}_t + \nabla r_t(\bar{x}_t)\|_{\nabla^2\psi_t(\bar{x}_t)^{-1}}\|\bar{x}_t - \bar{x}_{t+1}\|_{\nabla^2\psi_t(\bar{x}_t)}\\
&\implies \|\bar{x}_t - \bar{x}_{t+1}\|_{\nabla^2\psi_t(\bar{x}_t)} \le 4\|\hat{c}_t + \nabla r_t(\bar{x}_t)\|_{\nabla^2\psi_t(\bar{x}_t)^{-1}}.
\end{aligned}$$

Finally, we consider two cases depending on the value of g_t. First, if $g_t = 0$, then $r_t = 0$ and so $\|\hat{c}_t + \nabla r_t(\bar{x}_t))\|_{\nabla^2 \psi_t(\bar{x}_t)^{-1}} = \|\hat{c}_t\|_{\nabla^2 \psi_t(\bar{x}_t)^{-1}}$. Alternatively, if $g_t = 1$, then if we define $\delta_t = \frac{\sqrt{1+\sigma_{1:t}} - \sqrt{1+\sigma_{1:t-1}}}{\eta}$, we have:

$$\nabla r_t(\bar{x}_t) = \delta_t \bar{x}_t \quad \implies \|\nabla r_t(\bar{x}_t)\|_{\nabla^2 \psi_t(\bar{x}_t)^{-1}} \quad = \delta_t \|\bar{x}_t\|_{\nabla^2 \psi_t(\bar{x}_t)^{-1}} \le \delta_t \sqrt{\frac{\eta}{\sqrt{d^2 + \sigma_{1:t}}}} \le \frac{\sigma_t}{\sqrt{\eta}(d^2 + \sigma_{1:t})^{3/4}},$$

where the first inequality follows since $\nabla^2 \psi_t(\bar{x}_t) \ge \frac{\sqrt{d^2+\sigma_{1:t}}}{\eta} I$. □

The following technical statement is helpful in the proof of Lemma C.8.

Lemma C.9. *Define $\psi_t(x) = r_{0:t}(x) + \gamma\phi(x)$ for $\phi(x) = -\log(1 - \|x\|^2)$. Then $\|\bar{x}_t - \bar{x}_{t+1}\|_{\nabla^2 \phi(\bar{x}_t)} \le \frac{1}{2}$.*

Proof. To prove this, we claim for all v with $\|v\|_{\nabla^2 \phi(\bar{x}_t)} = \frac{1}{2}$, $\psi_t(\bar{x}_t + v) + \langle \hat{c}_{1:t}, \bar{x}_t + v\rangle \ge \psi_t(\bar{x}_t) + \langle \hat{c}_{1:t}, \bar{x}_t\rangle$. This will establish $\|\bar{x}_t - \bar{x}_{t+1}\|_{\nabla^2 \phi(\bar{x}_t)} \le \frac{1}{2}$ since $\bar{x}_{t+1} = \operatorname{argmin}\langle \hat{c}_{1:t}, x\rangle + \psi_t(x)$. To establish the claim, define $\delta_t = \frac{\sqrt{d^2+\sigma_{1:t}} - \sqrt{d^2+\sigma_{1:t-1}}}{\eta}$ so that $r_t(x) = \frac{\delta_t}{2}\|x\|^2$. Further, notice that $\|v\| \le 1$ since the Dikin ellipsoid centered at $\bar{x}_t$ must be contained in the unit ball. Then, we have:

$$\psi_t(\bar{x}_t + v) + \langle \hat{c}_{1:t}, \bar{x}_t + v\rangle = \psi_{t-1}(\bar{x}_t + v) + \frac{\delta_t}{2}\|\bar{x}_t + v\|^2 + \langle \hat{c}_{1:t}, \bar{x}_t + h\rangle$$

by mean value theorem, there is some $y \in [\bar{x}_t, \bar{x}_t + v]$ such that:

$$\begin{aligned}
&= \psi_{t-1}(\bar{x}_t) + \langle \nabla\psi_{t-1}(\bar{x}_t), v\rangle + \frac{\|v\|^2_{\nabla^2\psi_{t-1}(y)}}{2} + \frac{\delta_t}{2}\|\bar{x}_t + v\|^2 + \langle \hat{c}_{1:t}, \bar{x}_t + v\rangle \\
&= \psi_{t-1}(\bar{x}_t) + \langle \hat{c}_{1:t}, \bar{x}_t\rangle + \frac{\|v\|^2_{\nabla^2\psi_{t-1}(y)}}{2} + \frac{\delta_t}{2}\|\bar{x}_t + v\|^2 + \langle \hat{c}_t, v\rangle \\
&= \psi_t(\bar{x}_t) + \langle \hat{c}_{1:t}, \bar{x}_t\rangle + \delta_t\langle \bar{x}_t, v\rangle + \frac{\delta_t\|v\|^2}{2} + \langle \hat{c}_t, v\rangle + \frac{\|v\|^2_{\nabla^2\psi_{t-1}(y)}}{2} \\
&= \psi_t(\bar{x}_t) + \langle \hat{c}_{1:t}, \bar{x}_t\rangle + \delta_t\langle \bar{x}_t, v\rangle + \frac{\delta_t\|v\|^2}{2} + \langle \hat{c}_t, v\rangle + \frac{\delta_{0:t-1}\|v\|^2 + \gamma\|v\|^2_{\nabla^2\phi(y)}}{2} \\
&\ge \psi_t(\bar{x}_t) + \langle \hat{c}_{1:t}, \bar{x}_t\rangle + \delta_t\langle \bar{x}_t, v\rangle + \frac{\delta_t\|v\|^2}{2} + \langle \hat{c}_t, v\rangle + \frac{\delta_{0:t-1}\|v\|^2}{2} + \frac{\gamma\|v\|^2_{\nabla^2\phi(\bar{x}_t)}}{2}(1 - \|y - \bar{x}_t\|_{\nabla^2\phi(\bar{x}_t)})^2 \\
&= \psi_t(\bar{x}_t) + \langle \hat{c}_{1:t}, \bar{x}_t\rangle + \delta_t\langle \bar{x}_t, v\rangle + \langle \hat{c}_t, v\rangle + \frac{\delta_{0:t}\|v\|^2}{2} + \frac{\gamma\|v\|^2_{\nabla^2\phi(\bar{x}_t)}}{8},
\end{aligned}$$

where the inequality follows from Proposition C.5. Applying the Cauchy–Schwarz inequality:

$$\begin{aligned}
&\psi_t(\bar{x}_t + v) + \langle \hat{c}_{1:t}, \bar{x}_t + v\rangle \le \psi_t(\bar{x}_t) + \langle \hat{c}_{1:t}, \bar{x}_t\rangle + \delta_t\langle \bar{x}_t, v\rangle + \langle \hat{c}_t, v\rangle + \frac{\delta_{0:t}\|v\|^2}{2} + \frac{\gamma\|v\|^2_{\nabla^2\phi(\bar{x}_t)}}{8} \\
&\le \psi_t(\bar{x}_t) + \langle \hat{c}_{1:t}, \bar{x}_t\rangle - \delta_t\|\bar{x}_t\|_{\nabla^2\phi(\bar{x}_t)^{-1}}\|v\|_{\nabla^2\phi(\bar{x}_t)} - \|\hat{c}_t\|_{\nabla^2\phi(\bar{x}_t)^{-1}}\|v\|_{\nabla^2\phi(\bar{x}_t)} + \frac{\delta_{0:t-1}\|v\|^2}{8} + \frac{\gamma\|v\|^2_{\nabla^2\phi(\bar{x}_t)}}{8} \\
&= \psi_t(\bar{x}_t) + \langle \hat{c}_{1:t}, \bar{x}_t\rangle - \frac{\delta_t}{2}\|\bar{x}_t\|_{\nabla^2\phi(\bar{x}_t)^{-1}} - \frac{\|\hat{c}_t\|_{\nabla^2\phi(\bar{x}_t)^{-1}}}{2} + \frac{\gamma}{32}.
\end{aligned}$$

From Proposition C.2, we have $\|\hat{c}_t\|_{\nabla^2\phi(\bar{x}_t)^{-1}} \le 2\sqrt{d}$ and $\|\bar{x}_t\|_{\nabla^2\phi(\bar{x}_t)^{-1}} \le 1/2$. Now, notice that $\delta_t \le \frac{\sigma_t}{\eta\sqrt{1+\sigma_{1:t}}} \le \sqrt{\sigma_t}/\eta \le d/\eta$ since $\sigma_t \le d^2$. Therefore, we have:

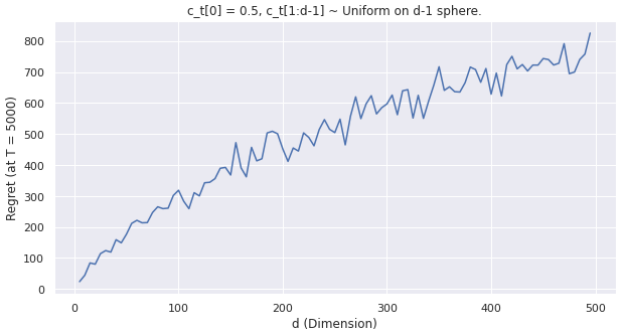
$$\psi_t(\bar{x}_t + v) + \langle \hat{c}_{1:t}, \bar{x}_t + v\rangle \le \psi_t(\bar{x}_t) + \langle \hat{c}_{1:t}, \bar{x}_t\rangle - \frac{d}{4\eta} - \sqrt{d} + \frac{\gamma}{32}.$$

Since $\gamma \ge 8\frac{d}{\eta} + 32\sqrt{d}$, the claim follows. □

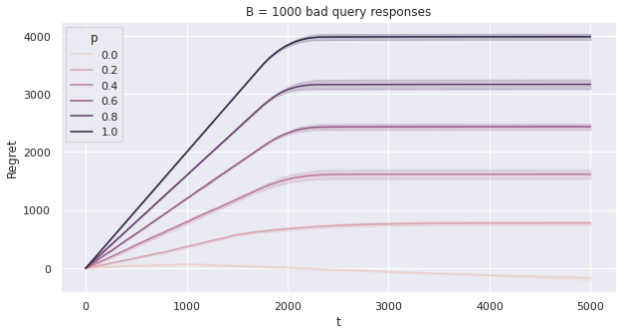
D. Experimental Results

In this section, we include an experimental evaluation of Algorithm 1 on synthetic data.

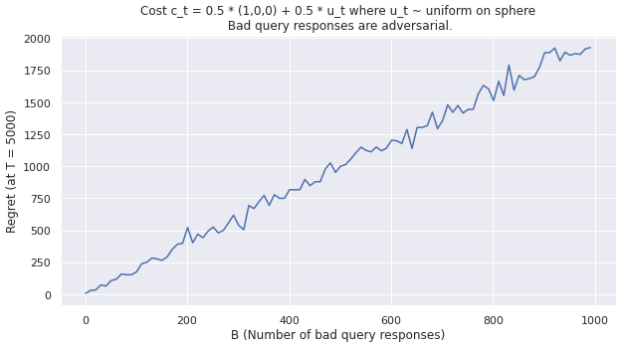
Dependence on the dimension. The regret bound provided by Theorem 4.1 degrades with the dimension. Note that this is in contrast with the dimension-independent regret guarantees available with hints in the full-information setting (Bhaskara et al., 2020; 2021). We consider the following experimental setup. For each time step t independently, the cost vector c_t is generated as follows: the first coordinate of c_t is fixed to be 0.5, and the remaining $d-1$ coordinates are drawn uniformly at random from a $(d-1)$-dimensional sphere of radius $\sqrt{1-0.5^2}$ so that each cost vector has unit length. We set $B = 0$, i.e., there are no bad query responses and set the time horizon $T = 5000$. Figure 1a shows a plot of the regret incurred after $T = 5000$ time steps for varying dimensions. Intriguingly, the regret degrades sublinearly with the dimension even though Theorem 4.1 suggests a superlinear dependence.



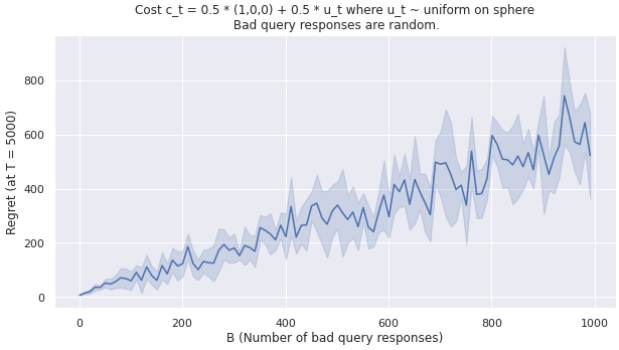
(a) Regret degrades with the dimension.



(b) Regret scales linearly with the number of bad responses.



(c) Linear regret with adversarially chosen responses.



(d) Linear regret with randomly chosen bad responses.

Figure 1. Experimental results.

Dependence on number of bad responses. We now demonstrate that the regret incurred by Algorithm 1 indeed does depend linearly on the number of bad query responses. Intuitively, since the algorithm only uses the query responses to construct estimates of the cost vector (and does not use any additional exploration), it is unable to be robust to bad query responses. We consider the following setup. We fix $d = 3$ and for each time step t independently, the cost vector c_t is generated as follows: $c_t = p \cdot (1, 0, 0) + (1 - p) \cdot u_t$ where u_t is a uniformly random unit vector on the sphere in $\mathbb{R}^3$. We set the $T = 5000$ and let the first $B = 1000$ query responses to be adversarially bad. More precisely, we let $\mathcal{Q}(s_t) = \langle s_t, -c_t \rangle$ for the first $B = 1000$ time steps. Figure 1b shows the regret curve obtained for different values of p. As p increases, the adversarial responses hurt the algorithm more, but in either case the regret increases linearly for the first B time steps.

Non-adversarial bad responses. In this set of experiments, we evaluate the effect of bad but non-adversarial query responses on the regret obtained by Algorithm 1. The experimental setup is same as the one above but we fix $p = 0.5$. For different values of $B \in \{0, 1, \ldots, 1000\}$, we repeat the experiment and record the regret incurred after $T = 5000$ time steps. We consider two scenarios as follows: (i) Figure 1c shows the regret incurred when the bad responses are chosen adversarially (i.e., $\mathcal{Q}(s_t) = \langle s_t, -c_t \rangle$); (ii) Figure 1d shows the regret incurred when the bad responses are chosen randomly, i.e., $\mathcal{Q}(s_t) = \langle s_t, y_t \rangle$ where y_t is a chosen uniformly at random from the unit sphere. In either case, we observe that the regret increases linearly with the number of bad hints.