Closing the Gap between Bandit and Full-Information Online Optimization: High-Probability Regret Bound

Alexander Rakhlin
Ambuj Tewari
Peter Bartlett

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-109
http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-109.html

August 26, 2007

Closing the Gap between Bandit and Full-Information Online Optimization: High-Probability Regret Bound

Alexander Rakhlin
Computer Science Division
UC Berkeley

Ambuj Tewari
Computer Science Division
UC Berkeley

Peter L. Bartlett
Computer Science Division
Department of Statistics
UC Berkeley

August 17, 2007

Abstract

We demonstrate a modification of the algorithm of Dani et al [5] for the online linear optimization problem in the bandit setting, which allows us to achieve an $O(\sqrt{T \ln T})$ regret bound *in high probability* against an *adaptive* adversary, as opposed to the *in expectation* result against an *oblivious* adversary of [5]. We obtain the same dependence on the dimension ($n^{3/2}$) as that exhibited by Dani et al. The results of this paper rest firmly on those of [5] and the remarkable technique of Auer et al [1] for obtaining high-probability bounds via optimistic estimates. This paper answers an open question: it eliminates the gap between the high-probability bounds obtained in the full-information vs bandit settings.

1 Introduction

In the *online linear optimization* problem as studied, for instance, by [6, 8], the decision maker (player) at each time $t \in \{1, \ldots, T\}$ chooses a vector x_t from a decision space $D \subset \mathbb{R}^n$. The reward or gain obtained at time t is a linear function $G_t^\intercal x_t$ of x_t, chosen by the adversary at the same time that x_t is chosen. An *oblivious* adversary is a fixed sequence $\{G_t\}$ of linear functions, while an *adaptive* adversary can choose G_t depending on $x_1, \ldots, x_{t-1}$. The goal of the decision maker is to minimize the regret

$$\max_{x \in D} \sum_{t=1}^{T} G_t^\intercal x - \sum_{t=1}^{T} G_t^\intercal x_t \ ,$$

which is the difference between the total gain of the best decision in hindsight and that of the decision maker. In the *full-information* setting, the linear function G_t is revealed to the decision maker at the end of each round and gradient methods, for instance, can be employed to achieve $O(\sqrt{T})$ bounds on the regret if the decision space is convex. The efficient randomized method of Kalai and Vempala [6] also enjoys this guarantee on the expected regret.

The partial information or *bandit* version of the problem has been receiving increasing attention in the recent literature. In the bandit version, the decision maker only gets to observe $G_t^\intercal x_t$. In

other words, only the value of the linear function at the decision made, not the full function, is revealed. For the special case when $D = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is the standard basis in $\mathbb{R}^n$, the problem is referred to as the *adversarial n-armed bandit problem*. Auer et al. [1] gave several algorithms for the n-armed bandit problem with near-optimal regret bounds. Their **Exp3** algorithm achieves $O^*(\sqrt{nT})$ regret in expectation.[1] However, the gains accumulated by **Exp3** turn out to have high variance and hence the authors proposed another algorithm **Exp3.P** that achieves $O^*(\sqrt{nT})$ regret with high probability. The key idea used by **Exp3.P** is that of maintaining biased estimates of the actual gains of various bandit arms. The bias is made large enough such these estimates are upper bounds on the actual gains with high probability.

One can use Auer et al.'s algorithms in the online linear optimization setting by associating an action with element of D (or an appropriate discretization of D). However, since the size of D is usually exponential in n for most problems of interest (e.g. online shortest paths), the regret bounds thus obtained are loose. Several authors [2, 7, 4, 5] have obtained stronger bounds. Some of these results even hold for the case of *adaptive adversaries*, that is adversaries that do not a priori fix a sequence of functions to be played. Nevertheless, even for an oblivious adversary (one whose moves do not depend on player's actions) the best regret bounds were $O^*(\mathrm{poly}(n)T^{2/3})$. Recently, Dani et al. [5] proposed a breakthrough algorithm that achieves $O^*(n^{3/2}\sqrt{T})$ bound on the expected regret. The dependence on T is optimal ignoring logarithmic factors. Their algorithm is similar in flavor to the **Exp3** algorithm of Auer et al. Moreover, like the original algorithm, the estimates kept by their algorithm can also have high variance and therefore the question of achieving $O^*(\mathrm{poly}(n)\sqrt{T})$ regret bounds that hold with high probability remained open.

We propose an algorithm called **GH.P** that does achieve the desired regret bounds *with high probability*. It is based on the idea of using high confidence upper bounds that was also utilized by **Exp3.P**. Our results are obtained by combining this idea with the insights obtained in Dani et al.'s paper [5]. A key step in our analysis involves the use of strong concentration inequalities like Bernstein's that take variance information into account. We also mention that our bounds scale as $n^{3/2}$ with the dimension, which matches the *in expectation* results of [5]. Curiously, the bounds in the full-information setting of prediction with expert advice (see e.g. [3]) have a $\log n$ dependence on the dimension.

It is known that, without loss of generality, we can assume that the adversary's choice of G_t is a deterministic function of $x_1, \ldots, x_{t-1}$ (see e.g. [4, Remark 3.1], [7]). This makes G_t a constant when conditioned on $x_1, \ldots, x_{t-1}$, a property essential for our proofs.

While the result of the paper closes the gap between the lower and upper bounds in high probability on the regret in the bandit setting with linear functions, the quest for an efficient algorithm is still open.

The plan of the rest of the paper is as follows. We present the algorithm **GH.P** and the high-probability performance guarantee in Section 2. This is followed by the proof of the bound in Section 3. The proof consists of two steps: (i) proving that certain upper bounds do hold with high probability, and (ii) the usual analysis of exponential updates using the exponential potential function. We discuss further open questions in Section 4.

[1] In the rest of the paper, we will use $O^*(\cdot)$ notation to hide constant and logarithmic factors.

2 Algorithm GH. P and a $\sqrt{T}$-Bound

In this paper we mostly follow the notation of Dani et al [5]. We will, however, set the game in terms of gains instead of losses. To this end, denote the decision space by D and suppose that $D \subset [-1,1]^n$. Denote the linear function played by the environment at time $t \in \{1,\dots,T\}$ by $G_t^\intercal x$ and the prediction of the player (decision maker) by $x_t \in D$. Define

$$G_{\max} := \max_{x \in D} \sum_{t=1}^{T} G_t^\intercal x \qquad \text{and} \qquad G_{\text{alg}} := \sum_{t=1}^{T} G_t^\intercal x_t \ .$$

Hence, the goal of the decision maker is to minimize $G_{\max} - G_{\text{alg}}$.

We assume that $G_t^\intercal x \in [0,1]$ for all $x \in D$ and that the standard basis (spanner) is in D (or that there exists a 1-barycentric spanner). According to Lemma 3.1 of [5], we can further assume that D is discretized and of size at most $(4nT)^{n/2}$, as the optimal gain of a decision in this discrete set is within an additive $\sqrt{nT}$ of the best gain in the corresponding continuous decision space. Only the logarithm of the cardinality of the set will enter in our bounds.

The algorithm presented below is a modification of the algorithm in [5]. Note that the difference is in the way we update weights w_t, using upper confidence intervals. This modification was crucial in the results of Auer et al [1] for obtaining high-probability bounds in the n-armed bandit setting. This update is also a crucial change to the algorithm of Dani et al [5].

Algorithm 1 GeometricHedge.P algorithm (**GH. P**)

Input: parameters η, γ, α, T

$\forall x \in D$,

$$w_1(x) = \exp(\eta\alpha\sqrt{T})$$

for $t = 1$ to T **do**

$\forall x \in D$,

$$p_t(x) = (1-\gamma)\frac{w_t(x)}{\sum_{x \in D} w_t(x)} + \frac{\gamma}{n}\mathbf{I}_{\{x \in \text{ spanner}\}}$$

Sample $x_t \sim p_t$

Observe gain $g_t = G_t^\intercal x_t$

Let $\mathbf{C}_t = \sum_x p_t(x) x x^\intercal = \mathbb{E}_{x \sim p_t}[x x^\intercal]$

Let $\hat{G}_t = g_t \mathbf{C}_t^{-1} x_t$

$\forall x \in D$,

$$w_{t+1}(x) = w_t(x) \exp\left[\eta\left(\hat{G}_t^\intercal x + \frac{\alpha}{\sqrt{T}} x^\intercal \mathbf{C}_t^{-1} x\right)\right]$$

end for

The main result of this paper is the following guarantee on the algorithm **GH. P**.

Theorem 2.1. *If we set* $\alpha = 2\sqrt{\ln\frac{T}{\delta} + (n/2)\ln(4nT)}$, $\gamma = \frac{3n^{3/2}\sqrt{\ln 4nT}}{2\sqrt{T}}$, *and* $\eta = \gamma/3n^2$, *then against any* adaptive *adversary with probability at least* $1 - 4\delta$,

$$G_{\max} \le G_{\text{alg}} + 9n\sqrt{\ln \delta^{-1}}\sqrt{T} + 12n^{3/2}\sqrt{\ln 4nT}\sqrt{T} + Q,$$

where $Q = 4n^{3/2}\ln\delta^{-1}\left(\frac{T}{\ln 4nT}\right)^{1/4} + 3n\ln T + 2n^2 \ln 4nT$, *a remainder term.*

Next section is devoted to the proof of Theorem 2.1.

3 Analysis

We first state several results obtained in Dani et al [5] which will be important in our proofs.

Lemma 3.1. *For any $x \in D$ and $t \in \{1, \ldots, T\}$, it holds that*

- $|\hat{G}_t^\intercal x| \leq n^2/\gamma$
- $x^\intercal \mathbf{C}_t^{-1} x \leq n^2/\gamma$.
- $\sum_{x \in D} p_t(x) x^\intercal \mathbf{C}_t^{-1} x = n$.

3.1 High Confidence Upper Bounds

Let $\mathbb{E}_t[\cdot]$ denote $\mathbb{E}[\cdot|x_1, \ldots, x_{t-1}]$. Since we are considering adaptive (but deterministic) adversaries, G_t is not random given $x_1, \ldots, x_{t-1}$. Observe that $\mathbb{E}_t[x_t x_t^\intercal] = \mathbb{E}_{x \sim p_t}[xx^\intercal]$ and, thus, $\mathbb{E}_t[\hat{G}_t] = G_t$. However, the fluctuations of the random variable $\hat{G}_t$ are very large. The following lemma provides a bound on these fluctuations. Its proof closely follows that of Auer et al [1].

Lemma 3.2. *If $\alpha \leq 2\sqrt{T}$ then for any $x \in D$,*

$$\Pr\left(\sum_{t=1}^{T} G_t^\intercal x \geq \sum_{t=1}^{T} \hat{G}_t^\intercal x + \alpha\left(\sqrt{T} + \frac{1}{\sqrt{T}} \sum_{t=1}^{T} x^\intercal \mathbf{C}_t^{-1} x\right)\right) \leq T e^{-\alpha^2/4}$$

Proof. Fix an $x \in D$ and let

$$\sigma_x(t+1) := \sqrt{T} + \frac{1}{\sqrt{T}} \sum_{\tau=1}^{t} x^\intercal \mathbf{C}_\tau^{-1} x \ ,$$

$$s_t := \frac{\alpha}{2\sigma_x(t+1)} \leq \frac{\alpha}{2\sqrt{T}} \ .$$

Since $\alpha \leq 2\sqrt{T}$ and $\sigma_x(t+1) \geq \sqrt{T}$, we have $s_t \leq 1$. Now,

$$\begin{aligned}
&\Pr\left(\sum_{t=1}^{T} (G_t - \hat{G}_t)^\intercal x - \alpha\sigma_x(T+1)/2 \geq \alpha\sigma_x(T+1)/2\right) \\
&\leq \Pr\left(s_T \sum_{t=1}^{T} \left((G_t - \hat{G}_t)^\intercal x - \frac{\alpha}{\sqrt{T}} x^\intercal \mathbf{C}_t^{-1} x\right) \geq \alpha^2/4\right) \\
&= \Pr\left(\exp\left(s_T \sum_{t=1}^{T} \left((G_t - \hat{G}_t)^\intercal x - \frac{\alpha}{\sqrt{T}} x^\intercal \mathbf{C}_t^{-1} x\right)\right) \geq \exp(\alpha^2/4)\right) \\
&\leq e^{-\alpha^2/4} \mathbb{E} \underbrace{\exp\left(s_T \sum_{t=1}^{T} \left((G_t - \hat{G}_t)^\intercal x - \frac{\alpha}{\sqrt{T}} x^\intercal \mathbf{C}_t^{-1} x\right)\right)}_{Z_T} \qquad (1)
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{E}_t Z_t &= Z_{t-1}^{s_t/s_{t-1}} \mathbb{E}_t \exp\left[s_t(G_t - \hat{G}_t)^\intercal x\right] \exp\left[-s_t \frac{\alpha}{\sqrt{T}} x^\intercal \mathbf{C}_t^{-1} x\right] \\
&\leq Z_{t-1}^{s_t/s_{t-1}} \mathbb{E}_t \left[1 + s_t(G_t^\intercal x - \hat{G}_t^\intercal x) + s_t^2 (G_t^\intercal x - \hat{G}_t^\intercal x)^2\right] \exp\left[-s_t \frac{\alpha}{\sqrt{T}} x^\intercal \mathbf{C}_t^{-1} x\right] \qquad (2) \\
&\leq Z_{t-1}^{s_t/s_{t-1}} \left[1 + 0 + s_t^2 x^\intercal \mathbf{C}_t^{-1} x\right] \exp\left[-s_t \frac{\alpha}{\sqrt{T}} x^\intercal \mathbf{C}_t^{-1} x\right] \qquad (3) \\
&\leq Z_{t-1}^{s_t/s_{t-1}} \left[1 + s_t^2 x^\intercal \mathbf{C}_t^{-1} x\right] \exp\left[-s_t^2 x^\intercal \mathbf{C}_t^{-1} x\right] \qquad (4) \\
&\leq Z_{t-1}^{s_t/s_{t-1}} \qquad (5) \\
&\leq 1 + Z_{t-1} \qquad (6)
\end{aligned}$$

Equation (2) uses $s_t(G_t - \hat{G}_t)^\intercal x \leq 1$ and $e^a \leq 1 + a + a^2$ for $a \leq 1$. Equation (3) uses $\mathbb{E}_t[\hat{G}_t] = G_t$ and

$$\begin{aligned}
\mathbb{E}_t(G_t^\intercal x - \hat{G}_t^\intercal x)^2 &= \mathrm{Var}_t(\hat{G}_t^\intercal x) \\
&\leq \mathbb{E}_t(\hat{G}_t^\intercal x)^2 \\
&= \mathbb{E}_t\left[g_t^2 x^\intercal \mathbf{C}_t^{-1} x_t x_t^\intercal \mathbf{C}_t^{-1} x\right] \\
&\leq x^\intercal \mathbf{C}_t^{-1} \mathbb{E}_t\left[x_t x_t^\intercal\right] \mathbf{C}_t^{-1} x \\
&= x^\intercal \mathbf{C}_t^{-1} \mathbf{C}_t \mathbf{C}_t^{-1} x = x^\intercal \mathbf{C}_t^{-1} x \ .
\end{aligned}$$

Equation (4) uses $s_t \leq \alpha/(2\sqrt{T})$. Equation (5) uses $1 + a \leq e^a$ for any a. Finally, equation (6) uses $z^r \leq 1 + z$ for any $z > 0$ and $r \in [0, 1]$. Note that s_t is a decreasing sequence and hence $s_t/s_{t-1} \leq 1$. Noting that $Z_1 \leq 1$ we get $\mathbb{E}Z_T \leq T$. Thus the lemma follows from (1). □

3.2 Potential Function Analysis

Our choice $\eta = \gamma/3n^2$ implies that $\eta \hat{G}_t^\intercal x \leq 1/3$. Moreover, if $\alpha \leq 2\sqrt{T}$ then $\eta\alpha x^\intercal \mathbf{C}_t^{-1} x/\sqrt{T} \leq 2/3$ (by Lemma 3.1) and hence

$$\eta\left(\hat{G}_t^\intercal x + \frac{\alpha}{\sqrt{T}} x^\intercal \mathbf{C}_t^{-1} x\right) \leq 1 \ .$$

Now using the fact that $e^a \leq 1 + a + a^2$ for any $a \leq 1$ and that $(b + c)^2 \leq 2(b^2 + c^2)$ for any b, c, we get

$$\begin{aligned}
\frac{W_{t+1}}{W_t} &= \sum_{x \in D} \frac{w_t(x) \exp\left[\eta\left(\hat{G}_t^\intercal x + \frac{\alpha}{\sqrt{T}} x^\intercal \mathbf{C}_t^{-1} x\right)\right]}{\sum_{x \in D} w_t(x)} \\
&= \sum_{x \in D} \left(\frac{p_t(x) - \frac{\gamma}{n}\mathbf{I}_{\{x \in \text{ spanner}\}}}{1 - \gamma}\right)\left(1 + \eta \hat{G}_t^\intercal x + \frac{\alpha\eta}{\sqrt{T}} x^\intercal \mathbf{C}_t^{-1} x + 2\eta^2 (\hat{G}_t^\intercal x)^2 + \frac{2\alpha^2\eta^2}{T}(x^\intercal \mathbf{C}_t^{-1} x)^2\right) \\
&\leq 1 + \frac{\eta}{1 - \gamma}\left[\sum_x p_t(x) \hat{G}_t^\intercal x + \frac{\alpha}{\sqrt{T}} p_t(x) x^\intercal \mathbf{C}_t^{-1} x + 2\eta p_t(x)(\hat{G}_t^\intercal x)^2 + \frac{2\alpha^2\eta}{T} p_t(x)(x^\intercal \mathbf{C}_t^{-1} x)^2\right]
\end{aligned}$$

Summing over t and using the fact that $\log(1+x) \le x$,

$$\ln \frac{W_{T+1}}{W_1} \le \frac{\eta}{1-\gamma} \left[\sum_{t=1}^{T} \sum_x p_t(x) \hat{G}_t^\intercal x + \frac{\alpha}{\sqrt{T}} p_t(x) x^\intercal \mathbf{C}_t^{-1} x + 2\eta p_t(x) (\hat{G}_t^\intercal x)^2 + \frac{2\alpha^2 \eta}{T} p_t(x) (x^\intercal \mathbf{C}_t^{-1} x)^2 \right] \tag{7}$$

In the remainder of the section, we bound each of the four terms in the above upper bound, summed over t and x. The first term is the most difficult to deal with. We would like to relate it to the total loss of the algorithm, $\sum_{t=1}^{T} G_t^\intercal x_t$. However, the magnitude of the random quantity $\hat{G}_t^\intercal x$ grows with T and a direct approach yields rates worse than the desired $\sqrt{T}$ dependence. It turns out that a more careful analysis can be applied to show that the variance of $\sum_x p_t(x) \hat{G}_t^\intercal x$ is in fact small. The key is to utilize the Bernstein inequality for martingale differences.

3.2.1 First term

Lemma 3.3. *With probability at least* $1 - 2\delta$

$$\sum_{t=1}^{T} \sum_x p_t(x) \hat{G}_t^\intercal x - \sum_{t=1}^{T} G_t^\intercal x_t \le (n+1)\sqrt{2T \ln \delta^{-1}} + \frac{8}{3} \ln \delta^{-1} n^2 \gamma^{-1/2}$$

Proof. We can write $\mathbf{C}_t^{-1} = \mathbf{V}_t \mathbf{\Lambda}_t^{-1} \mathbf{V}_t^\intercal$ where $\mathbf{\Lambda}_t^{-1}$ is a diagonal matrix of $\lambda_{t,i}^{-1}$'s and columns of $\mathbf{V}_t$ are orthonormal bases. Then

$$x^\intercal \mathbf{C}_t^{-1} x_t = \sum_{i=1}^{n} \lambda_{t,i}^{-1} (x^\intercal v_{t,i})(x_t^\intercal v_{t,i}).$$

Let us adopt the following notation. For any y, let $y = \sum_{i=1}^{n} \beta_{t,i}(y) v_{t,i}$. In other words, $\beta_{t,i}(y)$ is the projection of y onto the ith basis vector. Then

$$x^\intercal \mathbf{C}_t^{-1} x_t = \sum_{i=1}^{n} \lambda_{t,i}^{-1} \beta_{t,i}(x) \beta_{t,i}(x_t).$$

We will need the following inequalities:

$$\mathbb{E}_{x \sim p_t} \beta_{t,i}(x)^2 = \sum_x p_t(x) \beta_{t,i}(x)^2 = \sum_x p_t(x) (x^\intercal v_{t,i})^2 = \lambda_{t,i} \tag{8}$$

as shown in equation (1) of [5]. It then follows that

$$\mathbb{E}_{x \sim p_t} \beta_{t,i}(x) \le \sqrt{\lambda_{t,i}} \ . \tag{9}$$

Furthermore, for any $x \in D$,

$$\beta_{t,i}(x)^2 \le \sum_{j=1}^{n} \beta_{t,j}(x)^2 = \|x\|^2 \le n \ .$$

Let us define $Y_t := \sum_{x \in D} p_t(x) \hat{G}_t^\intercal x$. With a bit of calculation, we get the following equality:

$$\begin{aligned}
Y_t &= \sum_x p_t(x) \hat{G}_t^\intercal x \\
&= \sum_x p_t(x) g_t x^\intercal \mathbf{C}_t^{-1} x_t \\
&= \sum_x p_t(x) g_t \sum_{i=1}^n \lambda_{t,i}^{-1} \beta_{t,i}(x) \beta_{t,i}(x_t) \\
&= g_t \sum_{i=1}^n \lambda_{t,i}^{-1} \beta_{t,i}(x_t) \sum_x p_t(x) \beta_{t,i}(x) \\
&= g_t \sum_{i=1}^n \lambda_{t,i}^{-1} \beta_{t,i}(x_t) \mathbb{E}_{x \sim p_t} \beta_{t,i}(x)
\end{aligned}$$

The goal is to show that $\sum_t Y_t$ is close to $\sum_t \mathbb{E}_t Y_t$. Let us bound the conditional variance of Y_t. Note that $0 \leq g_t \leq 1$ by assumption.

$$\begin{aligned}
\mathrm{Var}_t(Y_t) \leq \mathbb{E}_t Y_t^2 &= \mathbb{E}_t \left(g_t \sum_{i=1}^n \lambda_{t,i}^{-1} \beta_{t,i}(x_t) \mathbb{E}_{x \sim p_t} \beta_{t,i}(x) \right)^2 \\
&\leq \mathbb{E}_t \left(\sum_{i=1}^n \lambda_{t,i}^{-1} \beta_{t,i}(x_t) \mathbb{E}_{x \sim p_t} \beta_{t,i}(x) \right)^2 && (10) \\
&\leq n \mathbb{E}_t \sum_{i=1}^n \left(\lambda_{t,i}^{-1} \beta_{t,i}(x_t) \mathbb{E}_{x \sim p_t} \beta_{t,i}(x) \right)^2 && (11) \\
&= n \sum_{i=1}^n \lambda_{t,i}^{-2} \mathbb{E}_t \beta_{t,i}(x_t)^2 (\mathbb{E}_{x \sim p_t} \beta_{t,i}(x))^2 \\
&= n \sum_{i=1}^n \lambda_{t,i}^{-2} \mathbb{E}_{x \sim p_t} \beta_{t,i}(x)^2 (\mathbb{E}_{x \sim p_t} \beta_{t,i}(x))^2 && (12) \\
&\leq n \sum_{i=1}^n \lambda_{t,i}^{-2} \lambda_{t,i}^2 \leq n^2 && (13)
\end{aligned}$$

Equation (10) uses $g_t^2 \leq 1$. Equation (11) uses $(\sum_{i=1}^n a_i)^2 \leq n(\sum_{i=1}^n a_i^2)$. Equation (12) uses

$$\mathbb{E}_t \beta_{t,i}(x_t)^2 = \mathbb{E}_{x \sim p_t} \beta_{t,i}(x)^2 \ .$$

Equation (13) uses (8) and (9). Moreover,

$$\begin{aligned}
|Y_t| &= \left| G_t^\intercal x_t \sum_{i=1}^n \lambda_{t,i}^{-1} \beta_{t,i}(x_t) \mathbb{E}_{x\sim p_t} \beta_{t,i}(x) \right| \\
&\le |G_t^\intercal x_t| \cdot \left| \sum_{i=1}^n \lambda_{t,i}^{-1} \beta_{t,i}(x_t) \mathbb{E}_{x\sim p_t} \beta_{t,i}(x) \right| \\
&\le \sum_{i=1}^n \lambda_{t,i}^{-1} \sqrt{n} \sqrt{\lambda_{t,i}} \\
&\le n\sqrt{n} \max_i \lambda_{t,i}^{-1/2} \le \frac{n^2}{\sqrt{\gamma}}.
\end{aligned}$$

Hence, we also have $|Y_t - \mathbb{E}_t Y_t| \le 2n^2/\sqrt{\gamma}$. Applying Bernstein's inequality (see Appendix) for martingale differences to the sequence $Y_t - \mathbb{E}_t Y_t$, we obtain that with probability at least $1-\delta$,

$$\sum_{t=1}^T Y_t - \sum_{t=1}^T \mathbb{E}_t Y_t \le \sqrt{2T \ln \delta^{-1}} \sqrt{n^2} + 2 \ln \delta^{-1} \frac{4n^2}{3\sqrt{\gamma}}$$

Jumping ahead, we will be setting $\gamma = T^{-1/2}$ so the above rate is correct. We have

$$\sum_{t=1}^T Y_t - \sum_{t=1}^T \mathbb{E}_t Y_t \le \sqrt{2T \ln \delta^{-1}} n + \frac{8}{3} \ln \delta^{-1} n^2 \gamma^{-1/2}$$

Observe that

$$\sum_{t=1}^T \mathbb{E}_t Y_t = \sum_{t=1}^T \mathbb{E}_t \left[\sum_x p_t(x) \hat{G}_t^\intercal x \right] = \sum_{t=1}^T \sum_x p_t(x) G_t^\intercal x = \sum_{t=1}^T G_t^\intercal \mathbb{E}_{x\sim p_t} x$$

and that $\sum_{t=1}^T G_t^\intercal x_t$ is concentrated around this value, as we now show. Indeed, define $Z_t := G_t^\intercal (x_t - \mathbb{E}_{x\sim p_t} x)$. Since

$$\mathbb{E}_t G_t^\intercal x_t = G_t^\intercal \mathbb{E}_t x_t = G_t^\intercal \mathbb{E}_{x\sim p_t} x \ ,$$

Z_t is a martingale difference sequence. Note that $|Z_t| \le 1$ and therefore, by Azuma-Hoeffding inequality (see Appendix), with probability at least $1-\delta$,

$$\sum_{t=1}^T G_t^\intercal \mathbb{E}_{x\sim p_t} x - \sum_{t=1}^T G_t^\intercal x_t \le \sqrt{2T \ln \delta^{-1}}.$$

Combining these two results, with probability at least $1-2\delta$

$$\sum_{t=1}^T \sum_x p_t(x) \hat{G}_t^\intercal x - \sum_{t=1}^T G_t^\intercal x_t \le \sqrt{2T \ln \delta^{-1}} + \sqrt{2T \ln \delta^{-1}} n + \frac{8}{3} \ln \delta^{-1} n^2 \gamma^{-1/2}$$

□

3.2.2 Second term

The second term in (7) is simply bounded as

$$\frac{\alpha}{\sqrt{T}}\sum_{t=1}^{T}\sum_{x}p_t(x)x^\intercal\mathbf{C}_t^{-1}x \le \frac{\alpha}{\sqrt{T}}nT = \alpha n\sqrt{T},$$

using Lemma 3.1.

3.2.3 Third term

Considering the third term in (7), we get

$$\sum_{x}p_t(x)(\hat{G}_t^\intercal x)^2 = \sum_{x}p_t(x)\hat{G}_t^\intercal xx^\intercal\hat{G}_t = \hat{G}_t^\intercal\left(\sum_{x}p_t(x)xx^\intercal\right)\hat{G}_t = g_t^2x_t^\intercal\mathbf{C}_t^{-1}\mathbf{C}_t\mathbf{C}_t^{-1}x_t \le x_t^\intercal\mathbf{C}_t^{-1}x_t. \tag{14}$$

From Lemma 3.1,

$$\mathbb{E}_{x\sim p_t}x^\intercal\mathbf{C}_t^{-1}x = n\ .$$

Define a martingale difference sequence with respect to x_t:

$$Y_t = x_t^\intercal\mathbf{C}_t^{-1}x_t - \mathbb{E}_{x\sim p_t}x^\intercal\mathbf{C}_t^{-1}x.$$

Observe that $|Y_t| \le \frac{n^2}{\gamma}$ and therefore, with probability at least $1-\delta$,

$$\sum_{t=1}^{T}x_t^\intercal\mathbf{C}_t^{-1}x_t \le nT + \sqrt{2\ln\delta^{-1}}\sqrt{T}\frac{n^2}{\gamma}$$

Combining this result with (14), with probability at least $1-\delta$,

$$2\eta\sum_{t=1}^{T}\sum_{x}p_t(x)(\hat{G}_t^\intercal x)^2 \le 2\eta nT + 2\eta\sqrt{2\ln\delta^{-1}}\sqrt{T}\frac{n^2}{\gamma}$$

3.2.4 Fourth term

Finally, the last term in (7) is bounded as

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{x\in D}\frac{2\alpha^2\eta p_t(x)}{T}(x^\intercal\mathbf{C}_t^{-1}x)^2 &\le \frac{2\alpha^2\eta}{T}\sum_{t=1}^{T}\sum_{x\in D}p_t(x)\frac{n^2}{\gamma}x^\intercal\mathbf{C}_t^{-1}x \\
&\le \frac{2\alpha^2\eta n^2}{\gamma T}\sum_{t=1}^{T}\sum_{x\in D}p_t(x)x^\intercal\mathbf{C}_t^{-1}x \\
&\le \frac{2\alpha^2\eta n^2}{\gamma T}\sum_{t=1}^{T}n \\
&\le \frac{2\alpha^2\eta n^3}{\gamma},
\end{aligned}$$

where in the first inequality we used the fact that $0 \le x^\intercal\mathbf{C}_t^{-1}x \le n^2/\gamma$.

3.3 Plugging in

Putting all the results together, we rewrite (7) as follows. With probability at least $1 - 3\delta$,

$$\ln \frac{W_T}{W_1} \leq \frac{\eta}{1-\gamma} \left[\left(G_{\text{alg}} + \sqrt{2T \ln \delta^{-1}} + \sqrt{2T \ln \delta^{-1}} n + \frac{8}{3} \ln \delta^{-1} n^2 \gamma^{-1/2} \right) + \left(\alpha n \sqrt{T} \right) \right.$$
$$\left. + \left(2\eta n T + 2\eta \sqrt{2 \ln \delta^{-1}} \sqrt{T} \frac{n^2}{\gamma} \right) + \left(\frac{2\alpha^2 \eta n^3}{\gamma} \right) \right]. \tag{15}$$

On the other hand, if we choose α such that $T \exp(-\alpha^2/4) \leq \delta/|D|$ then with probability at least $1 - \delta$, we have for all $x \in D$,

$$\ln \frac{W_T}{W_1} \geq \eta \left(\sum_{t=1}^{T} \hat{G}_t^{\intercal} x + \alpha \sigma_x (T+1) \right) - \eta \alpha \sqrt{T} - \ln |D|$$
$$\geq \eta \sum_{t=1}^{T} G_t^{\intercal} x - \eta \alpha \sqrt{T} - \ln |D|. \tag{16}$$

Such a choice of α corresponds to $\alpha \geq 2\sqrt{\ln T + \ln \delta^{-1} + \ln |D|}$. Recall that $\ln |D| \leq (n/2) \ln(4nT)$. Combining (15) with (16), we have that with probability at least $1 - 4\delta$,

$$(1-\gamma) G_{\max} \leq G_{\text{alg}} + \alpha \sqrt{T} + \frac{\ln |D|}{\eta}$$
$$+ \left((n+1) \sqrt{2T \ln \delta^{-1}} + \frac{8}{3} \ln \delta^{-1} n^2 \gamma^{-1/2} \right) + \left(\alpha n \sqrt{T} \right)$$
$$+ \left(2\eta n T + 2\eta \sqrt{2 \ln \delta^{-1}} \sqrt{T} \frac{n^2}{\gamma} \right) + \left(\frac{2\alpha^2 \eta n^3}{\gamma} \right)$$

Plugging in our choice of $\eta = \gamma/3n^2$ and $\alpha = 2\sqrt{\ln T + \ln \delta^{-1} + (n/2) \ln(4nT)}$, and noting that $G_{\max} \leq T$, with probability at least $1 - 4\delta$,

$$G_{\max} \leq G_{\text{alg}} + 2(n+1) \sqrt{\ln T + \ln \delta^{-1} + (n/2) \ln(4nT)} \sqrt{T} + 3n^2 \frac{(n/2) \ln(4nT)}{\gamma}$$
$$+ \left((n+1) \sqrt{2T \ln \delta^{-1}} + \frac{8}{3} \ln \delta^{-1} n^2 \gamma^{-1/2} \right)$$
$$+ \left(\frac{2}{3n} \gamma T + \frac{2}{3} \sqrt{2 \ln \delta^{-1}} \sqrt{T} \right) + \frac{8}{3} n \left(\ln T + \ln \delta^{-1} + (n/2) \ln(4nT) \right) + T\gamma$$

With a choice of $\gamma = \frac{3n^{3/2} \sqrt{\ln 4nT}}{2\sqrt{T}}$. Substituting,

$$G_{\max} \leq G_{\text{alg}} + 2(n+1) \sqrt{\ln T + \ln \delta^{-1} + (n/2) \ln(4nT)} \sqrt{T} + n^{3/2} \sqrt{T} \sqrt{\ln 4nT}$$
$$+ (n+1) \sqrt{2T \ln \delta^{-1}} + \frac{8\sqrt{2}}{3\sqrt{3}} \ln \delta^{-1} n^{3/2} \left(\frac{T}{\ln 4nT} \right)^{1/4}$$
$$+ \sqrt{n} \sqrt{T} \sqrt{\ln(4nT)} + \frac{2}{3} \sqrt{2 \ln \delta^{-1}} \sqrt{T} + \frac{8}{3} n \left(\ln T + \ln \delta^{-1} + (n/2) \ln(4nT) \right) + \frac{3}{2} n^{3/2} \sqrt{\ln 4nT} \sqrt{T}.$$

By making simple over-approximations, we obtain that with probability at least $1 - 4\delta$

$$G_{\max} \leq G_{\text{alg}} + 9n\sqrt{\ln \delta^{-1}}\sqrt{T} + 12n^{3/2}\sqrt{\ln 4nT}\sqrt{T} + Q,$$

where a non-significant remainder term $Q = 4n^{3/2} \ln \delta^{-1} \left(\frac{T}{\ln 4nT}\right)^{1/4} + 3n \ln T + 2n^2 \ln 4nT$. We observe that the dependence on time horizon is $\sqrt{T \ln T}$ and the dependence on the dimension is $n^{3/2}$, matching the corresponding results of [5], obtained in expectation.

4 Conclusions and Open Problems

While the algorithm we presented achieves the desired $O^*(\sqrt{T})$ bound with high probability, the quest for an efficient algorithm is still open. Achieving similar results for general convex functions is also an intriguing open question.

A Concentration Inequalities

The following well-known inequalities can be found, for instance, in [3], Appendix A.

Lemma A.1 (Bernstein's inequality for martingale differences)**.** *Let $Y_1, \ldots, Y_T$ be a martingale difference sequence with respect to $X_1, \ldots, X_T$. Suppose that $Y_t \in [a, b]$ and $\mathbb{E}[Y_t^2 | X_{t-1}, \ldots, X_1] \leq \sigma^2$ almost surely for all $t \in \{1, \ldots, T\}$. Then for all $\epsilon > 0$,*

$$\Pr\left(\sum_{t=1}^{T} Y_t > \sqrt{2T\sigma^2 \ln \delta^{-1}} + 2 \ln \delta^{-1}(b-a)/3\right) \leq \delta$$

Lemma A.2 (Azuma-Hoeffding inequality for martingale differences)**.** *Let $Y_1, \ldots, Y_T$ be a martingale difference sequence with respect to $X_1, \ldots, X_T$. Suppose that $|Y_t| \leq c$ almost surely for all $t \in \{1, \ldots, T\}$. Then for all $\epsilon > 0$,*

$$\Pr\left(\sum_{t=1}^{T} Y_t > \sqrt{2Tc^2 \ln \delta^{-1}}\right) \leq \delta$$

References

[1] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[2] Baruch Awerbuch and Robert D. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of computing*, pages 45–53. ACM Press, 2004.

[3] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[4] Varsha Dani and Thomas P. Hayes. Robbing the bandit: less regret in online geometric optimization against an adaptive adversary. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 937–943. ACM Press, 2006.

[5] Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. The price of bandit information for online optimization. 2007. Submitted. Available on Sham Kakade's website.

[6] Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

[7] H. Brendan McMahan and Avrim Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *Proceedings of the Seventeenth Annual Conference on Learning Theory*, volume 3120 of *Lecture Notes in Computer Science*, pages 109–123. Springer, 2004.

[8] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.