High-Probability Regret Bounds for Bandit Online Linear Optimization

Peter L. Bartlett *
UC Berkeley
bartlett@cs.berkeley.edu

Varsha Dani
University of Chicago
varsha@cs.uchicago.edu

Thomas P. Hayes
TTI Chicago
hayest@tti-c.org

Sham M. Kakade
TTI Chicago
sham@tti-c.org

Alexander Rakhlin *
UC Berkeley
rakhlin@cs.berkeley.edu

Ambuj Tewari *
TTI Chicago
tewari@tti-c.org

Abstract

We present a modification of the algorithm of Dani et al. [8] for the online linear optimization problem in the bandit setting, which *with high probability* has regret at most $O^*(\sqrt{T})$ against an *adaptive* adversary. This improves on the previous algorithm [8] whose regret is bounded *in expectation* against an *oblivious* adversary. 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 [8] and the remarkable technique of Auer et al. [2] 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, there is a fixed decision set $D \in \mathbb{R}^n$ and the player (or decision maker) makes a decision x_t at time $t \in \{1, \ldots, T\}$. Simultaneously, an adversary chooses a loss vector L_t and the player suffers loss $L_t^\dagger x_t$. The goal is to minimize *regret* which measures how much worse the player did as compared to any fixed decision, even one chosen with complete knowledge of the sequence $L_1, \ldots, L_T$,

$$R = \sum_{t=1}^{T} L_t^\dagger x_t - \min_{x \in D} \sum_{t=1}^{T} L_t^\dagger x \, .$$

The adversary can be *oblivious* to the player's moves in which case it chooses the entire sequence $L_1, \ldots, L_t$ in advance of the player's moves. An *adaptive* adversary can, however, choose L_t based on the player's moves $x_1, \ldots, x_{t-1}$ up to that point.

In the full information version of the problem, the loss vector L_t is revealed to the player at the end of round t. For this case, Kalai and Vempala [12] gave an efficient algorithm assuming that the *offline* problem (given L minimize $L^\dagger x$ over $x \in D$) can be solved efficiently. Note that the standard "experts" problem is a special case of this problem because we can choose the set D to be $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, the unit vectors forming the standard basis of $\mathbb{R}^n$. Kalai and Vempala separated the issue of the number of available decisions from the dimensionality of the problem and gave an algorithm with expected regret $O(\mathrm{poly}(n)\sqrt{T})$. In many important cases, for example the *online shortest path problem* [15], the size of the decision set can be exponential in the dimensionality. So, it is important to design algorithms that have polynomial dependence on the dimension.

In the partial information or "bandit" version of the problem, the only feedback that the player receives at the end of round t is its own loss $L_t^\dagger x_t$. The bandit version of the experts problem was considered by Auer et al. [2] who gave a number of algorithms for the problem. Their **Exp3** algorithm achieves $O(\sqrt{T})$ expected regret against oblivious adversaries. However, due to the large variance of the estimates kept by **Exp3** it fails to enjoy a similar regret bound with high probability. To address this issue, the authors used the idea of *high confidence upper bounds* to derive the **Exp3.P** algorithm which achieves $O(\sqrt{T})$ regret with high probability. The regret of these algorithms also has a $\sqrt{|D|}$ dependence on the number $|D|$ of available actions. Hence, these cannot be used directly if $|D|$ is large.

Awerbuch and Kleinberg [4] were the first to consider the general online linear optimization problem in the bandit setting. For oblivious adversaries, they proved a regret bound of $O^*(\mathrm{poly}(n)T^{2/3})$. The case of a general adaptive adversary was handled by McMahan and Blum [14] but they could only prove a regret bound of $O^*(\mathrm{poly}(n)T^{3/4})$. Dani and Hayes [7] later showed that McMahan and Blum's algorithm actually enjoys a regret bound of $O^*(\mathrm{poly}(n)T^{2/3})$. However, the known lower bound for the bandit problem was the same as that in the full information case, namely $\Omega(\sqrt{T})$. Therefore, it was an important open question if there is an algorithm with a regret bound of $O(\mathrm{poly}(n)\sqrt{T})$ for the bandit online linear optimization problem. An affirmative answer was recently given by Dani et al. [8]. Their algorithm has expected regret at most $O^*(\mathrm{poly}(n)\sqrt{T})$ against an oblivious adversary. It was still not known if the same bounds could be achieved with high probability and against adaptive adversaries as well. In this paper, we show how to do this by combining Dani et al.'s techniques with those of Auer et al. [2]. Like **Exp3.P**, our GEOMETRICHEDGE.P algorithm

*PB, AR and AT gratefully acknowledge the support of DARPA under grant FA8750-05-2-0249.

keeps biased estimates of the losses of different actions such that, with high probability, the sums of these estimates are *lower bounds* (because we use losses not gains) on the actual unknown cumulative losses (Lemma 5).

The bandit version of the online shortest path problem has recently received a lot of attention. It can be used to model, for example, routing in ad hoc wireless networks. If we want to make our routing algorithm secure against adversarial attacks, it is necessary to design algorithms that work against adaptive adversaries [3, 13]. Therefore, obtaining low regret against adaptive adversaries is not only an important theoretical problem but it also has practical implications. The algorithm with the best regret guarantee so far is by György et al. [11]. There the authors consider a number of feedback models. Our feedback model in this paper corresponds to what they call the "path-bandit" model. For this model, they give an efficient algorithm specially designed for the bandit online shortest path problem that achieves $O^*(\mathrm{poly}(n)T^{2/3})$ regret with high probability against an adaptive adversary where n is the number of edges in the graph. Our results imply that it is actually possible to achieve $O^*(n^{3/2}\sqrt{T})$ regret with high probability. However, since our algorithm is not efficient, the quest for an efficient algorithm with the same regret, even for this special problem, is still on.

The key tools from probability theory that we use in our proofs are Bernstein-type inequalities, such as Freedman's. These provide sharper concentration bounds for martingales in the presence of variance information. There is a simple corollary of Freedman's inequality that we think is useful not just in our setting but more generally. We state it as Lemma 2 in Section 4.

The present work closes the gap between full information and bandit online optimization against the adaptive adversary in terms of the growth of regret with T. As we said above, our algorithm is not necessarily efficient, because the decision space might need to be discretized to a fine level. We mention that a parallel work by Abernethy, Hazan, and Rakhlin [1] provides an efficient algorithm for the setting; however, their result holds in expectation only (against an oblivious adversary). The present paper and [1] are addressing disparate aspects of the problem and neither result can be concluded from the other. It remains an open question whether there exists an efficient algorithm which enjoys high probability bounds on the regret.

2 Preliminaries

Let $D \subset [-1,1]^n$ denote the decision space. At each t of T time steps, the environment selects a cost vector L_t, and simultaneously, the player (decision maker) selects $x_t \in D$. The loss incurred by the decision maker for this prediction is $L_t^\dagger x_t$. Let

$$L_{\min} := \min_{x \in D} \sum_{t=1}^{T} L_t^\dagger x$$

be the loss of the best single decision in hindsight. The goal of the decision maker is to minimize the *regret*,

$$R = \sum_{t=1}^{T} L_t^\dagger x_t - L_{\min} .$$

We assume that $L_t^\dagger x \in [0,1]$ for all $x \in D$. We also assume that the environment is *adaptive*, *i.e.*, the cost vector L_t selected by the environment at time t may depend arbitrarily on the history $(L_1, x_1, \ldots, L_{t-1}, x_{t-1})$ (note that without loss of generality this dependence may be assumed to be deterministic.) We show that even against such a powerful environment, it is possible to ensure that R is small with high probability.

As in [8], we will require a barycentric spanner for D. Recall that a *barycentric spanner* for D is a set

$$\{y_1, \ldots, y_n\} \subseteq D$$

such that every $x \in D$ can be written as a linear combination of y_i's with coefficients in $[-1,1]$. A c-barycentric spanner is defined similarly where we allow coefficients to be in $[-c,c]$. For $c > 1$, c-barycentric spanners for D may be found efficiently (see [4].) However, for ease of exposition we'll assume that we have an actual barycentric spanner. (Using a c-barycentric spanner instead will only affect the constants.) Finally, if the set D is too large (for example if it is infinite) we can replace it by a cover of size at most $(4nT)^{n/2}$, as the loss of the optimal decision in this cover is within an additive $\sqrt{nT}$ of the optimal loss in D; see [8][Lemma 3.1] for details. Accordingly, after doing this transformation if necessary, we may assume that D is finite and $\ln |D| = O(n \ln T)$. Only the logarithm of the cardinality of the set will enter in our bounds.

3 Algorithm and Main Result

The algorithm presented below is a modification of the algorithm in [8]. Note that the difference is in the way we update weights w_t, using lower confidence intervals. This idea of using confidence intervals is motivated by the **Exp3.P** algorithm of Auer et al. [2]. Feeding in confidence bounds, as opposed to unbiased estimates of the losses, to the exponential updates is the crucial change we make to the algorithm of Dani et al [8]. Lemma 5 below shows that, with high probability, for any $x \in D$, $\sum_t \widetilde{L}_t(x)$ lower bounds $\sum_t L_t^\dagger x$ (up to an additive $O(\sqrt{T})$ term). Our algorithm reduces to **Exp3.P** in the special case of the n-armed bandit problem (when $D = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$. As we point out in the next section, Auer et al.'s proof can be simplified by using the simple corollary of Freedman's inequality [10] that we state as Lemma 2 below.

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

Theorem 1 *Let $T \geq 4$, $n \geq 2$ and $\delta \leq \frac{1}{e}$. If we set $\gamma = \frac{n^{3/2}}{\sqrt{T}}$, $\delta' = \frac{\delta}{|D| \log_2 T}$, and $\eta = \frac{1}{\sqrt{nT}+2\sqrt{\ln(1/\delta')}}$, then against any* adaptive *adversary with probability at least $1 - 4\delta$,*

$$R = O(n^{3/2}\sqrt{T}\ln(nT/\delta)).$$

The dependence on T is optimal (up to logarithmic factors). We get the same dependence on n as Dani et al. [8]. The lower bound known for this problem is $\Omega(n\sqrt{T})$ [8]. Recently, $O(n\sqrt{T})$ regret bounds have been obtained for the stochastic version of the problem [9]. This leads us to conjecture that the lower bound is tight and it remains an open

Algorithm 3.1: GEOMETRICHEDGE.P(D, γ, η, δ')

$\forall x \in D, w_1(x) := 1$
$W_1 := |D|$
for $t = 1$ **to** T
  $\forall x \in D$,
    $p_t(x) = (1-\gamma)\frac{w_t(x)}{W_t} + \frac{\gamma}{n}\mathbf{I}\{x \in \text{spanner}\}$
  Sample x_t according to distribution p_t
  Incur and observe loss $\ell_t := L_t^\dagger x_t$
  $\mathbf{C}_t := \mathbb{E}_{p_t}\left[xx^\dagger\right]$
  $\widehat{L}_t := \ell_t \mathbf{C}_t^{-1} x_t$
  $\forall x \in D, \widetilde{L}_t(x) := \widehat{L}_t^\dagger x - 2x^\dagger \mathbf{C}_t^{-1} x \sqrt{\frac{\ln(1/\delta')}{nT}}$
  $\forall x \in D, w_{t+1}(x) := w_t(x)\exp\{-\eta \widetilde{L}_t(x)\}$
  $W_{t+1} = \sum_{x \in D} w_{t+1}(x)$

question to close the gap (for the dependence on n) between upper and lower bounds. We also note here that although the analysis we provide is for losses, essentially the same algorithm, with a similar analysis, works for gains. We just have to make a few obvious changes to the algorithm: instead of subtracting, we *add* the correction term to the gain estimates and replace $-\eta$ with η in the exponential update.

4 Concentration for Martingales

In this section we derive a concentration inequality for martingale difference sequences. It is a direct application of Freedman's inequality.

Lemma 2 *Suppose $X_1, \ldots, X_T$ is a martingale difference sequence with $|X_t| \le b$. Let*

$$\mathrm{Var}_t X_t = \mathbf{Var}\left(X_t \,|\, X_1, \ldots, X_{t-1}\right).$$

Let $V = \sum_{t=1}^T \mathrm{Var}_t X_t$ be the sum of conditional variances of X_t's. Further, let $\sigma = \sqrt{V}$. Then we have, for any $\delta < 1/e$ and $T \ge 4$,

$$\mathrm{Prob}\left(\sum_{t=1}^T X_t > 2\max\left\{2\sigma, b\sqrt{\ln(1/\delta)}\right\}\sqrt{\ln(1/\delta)}\right) \le \log_2(T)\delta\,.$$

Proof: Note that a crude upper bound on $\mathrm{Var}_t X_t$ is b^2. Thus, $\sigma \le b\sqrt{T}$. We choose a discretization $0 = \alpha_{-1} < \alpha_0 < \ldots < \alpha_l$ such that $\alpha_{i+1} = 2\alpha_i$ for $i \ge 0$ and $\alpha_l \ge b\sqrt{T}$. We will specify the choice of α_0 shortly. We then have,

$$\begin{aligned}
&\mathrm{Prob}\left(\sum_t X_t > 2\max\{2\sigma, \alpha_0\}\sqrt{\ln(1/\delta)}\right) \\
&= \sum_{j=0}^{l} \mathrm{Prob}\left(\begin{array}{c}\sum_t X_t > 2\max\{2\sigma,\alpha_0\}\sqrt{\ln(1/\delta)} \\ \&\ \alpha_{j-1} < \sigma \le \alpha_j\end{array}\right) \\
&\le \sum_{j=0}^{l} \mathrm{Prob}\left(\begin{array}{c}\sum_t X_t > 2\alpha_j\sqrt{\ln(1/\delta)} \\ \&\ \alpha_{j-1}^2 < V \le \alpha_j^2\end{array}\right) \\
&\le \sum_{j=0}^{l} \mathrm{Prob}\left(\sum_t X_t > 2\alpha_j\sqrt{\ln(1/\delta)}\ \&\ V \le \alpha_j^2\right) \\
&\overset{(\star)}{\le} \sum_{j=0}^{l} \exp\left(\frac{-4\alpha_j^2\ln(1/\delta)}{2\alpha_j^2 + \frac{2}{3}\left(2\alpha_j\sqrt{\ln(1/\delta)}\right)b}\right) \\
&= \sum_{j=0}^{l} \exp\left(\frac{-2\alpha_j\ln(1/\delta)}{\alpha_j + \frac{2}{3}\left(\sqrt{\ln(1/\delta)}\right)b}\right)
\end{aligned}$$

where the inequality $(\star)$ follows from Freedman's inequality (Theorem 9). If we now choose $\alpha_0 = b\sqrt{\ln(1/\delta)}$ then $\alpha_j \ge b\sqrt{\ln(1/\delta)}$ for all j and hence every term in the above summation is bounded by $\exp\left(\frac{-2\ln(1/\delta)}{1+2/3}\right) < \delta$. Choosing $l = \log_2(\sqrt{T})$ ensures that $\alpha_l \ge b\sqrt{T}$. Thus we have

$$\begin{aligned}
&\mathrm{Prob}\left(\sum_{t=1}^T X_t > 2\max\{2\sigma, b\sqrt{\ln(1/\delta)}\}\sqrt{\ln(1/\delta)}\right) \\
&= \mathrm{Prob}\left(\sum_t X_t > 2\max\{2\sigma,\alpha_0\}\sqrt{\ln(1/\delta)}\right) \\
&\le (l+1)\delta = (\log_2(\sqrt{T})+1)\delta \le \log_2(T)\delta\,.
\end{aligned}$$

■

This inequality says that, roughly speaking, $\sum_t X_t$ is of the order of $\sigma\sqrt{\ln(1/\delta)}$ which is a central limit theorem-like behavior except that σ here is not fixed but is the actual sum of conditional variances, a random quantity. The overall constant in front of σ is 4. This can be improved to 2 by a slightly more careful analysis. We already know of two instances in the literature where Lemma 2 can be used to give shorter proofs of certain probabilistic upper bounds.

1. The first is in the proof of **Exp3.P**'s regret bound itself. To show that the estimates are upper bounds on the actual losses of an action, the authors explicitly use the exponential moment method in the proof of their Lemma 6.1. Essentially the same lemma can be proved by a direction application of the above lemma.

2. The other instance is in Cesa-Bianchi and Gentile's paper [5] on online to batch conversions. When an online algorithm is run on i.i.d. data with a non-negative and bounded loss function, the conditional variance of the loss at time t can immediately be bounded by the risk of the hypothesis at time $t-1$. The authors use this fact along with an application of Freedman's inequality to prove a sharp upper bound (Proposition 2 in their paper) on the average risk of the hypotheses generated by the online algorithm in terms of its actual cumulative loss. The same result can be quickly derived by an application of the above lemma.

5 Analysis

The remainder of the paper is devoted to the proof of Theorem 1. We first state several results obtained in Dani et al [8] which will be important in our proofs.

Lemma 3 *For any $x \in D$ and $t \in \{1, \ldots, T\}$, it holds that*

1. $|\widehat{L}_t^\dagger x| \le n^2/\gamma$
2. $x^\dagger \mathbf{C}_t^{-1} x \le n^2/\gamma$.
3. $\sum_{x \in D} p_t(x) x^\dagger \mathbf{C}_t^{-1} x = n$.
4. $\mathbb{E}_t \left(\widehat{L}_t^\dagger x\right)^2 \le x^\dagger \mathbf{C}_t^{-1} x$.

We now prove a bound on the perturbed estimated costs, $\widetilde{L}_t$, which are used to update the distribution.

Lemma 4 *For all $x \in D$, $|\widetilde{L}_t(x)| \le \sqrt{nT} + 2\sqrt{\ln(1/\delta')}$.*

Proof: For each $x \in D$,

$$\begin{aligned}|\widetilde{L}_t(x)| &\le |\widehat{L}_t \cdot x| + \left|2x^\dagger C_t^{-1} x \sqrt{\frac{\ln(1/\delta')}{nT}}\right| \\ &\le \frac{n^2}{\gamma} + 2\frac{n^2}{\gamma}\sqrt{\frac{\ln(1/\delta')}{nT}} \\ &\le \sqrt{nT} + 2\sqrt{\ln(1/\delta')}\end{aligned}$$

using Lemma 3 and the choice of $\gamma = \frac{n^{3/2}}{\sqrt{T}}$. ∎

5.1 High Confidence Bounds

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

Lemma 5 *Assume $T \ge 4$. Let $\delta' = \frac{\delta}{|D| \log_2 T}$. Then with probability at least $1 - \delta$, simultaneously for all $x \in D$,*

$$\sum_t \widetilde{L}_t(x) \le \sum_t L_t^\dagger x + 2\left(1 + \sqrt{nT}\right)\ln(1/\delta')$$

Proof: Fix $x \in D$. Let $M_t = M_t(x) = \widehat{L}_t^\dagger x - L_t^\dagger x$. Then (M_t) is a martingale difference sequence. Using Lemma 3, $|M_t| \le \frac{n^2}{\gamma} + 1 = \sqrt{nT} + 1$. Let $V = \sum_t \mathrm{Var}_t(M_t)$ and let $\sigma = \sqrt{V}$. Using Lemma 2, we have that with probability at least $1 - \delta' \log_2 T$,

$$\begin{aligned}\sum_t \widehat{L}_t^\dagger x \le \sum_t L_t^\dagger x + 2\max\{2\sigma, \\ (1 + \sqrt{nT})\sqrt{\ln(1/\delta')}\}\sqrt{\ln(1/\delta')}\end{aligned} \tag{1}$$

Now note that

$$\sigma \le \sqrt{\sum_t x^\dagger C_t^{-1} x} \le \frac{1}{2}\left(\frac{\sum_t x^\dagger C_t^{-1} x}{\sqrt{nT}} + \sqrt{nT}\right),$$

by the arithmetic mean-geometric mean inequality.

Substituting this into (1), we have

$$\begin{aligned}\sum_t \widehat{L}_t^\dagger x \le \sum_t L_t^\dagger x + 2\max\left\{\left(1 + \sqrt{nT}\right)\sqrt{\ln(1/\delta')}, \right. \\ \left.\left(\frac{\sum_t x^\dagger C_t^{-1} x}{\sqrt{nT}} + \sqrt{nT}\right)\right\}\sqrt{\ln(1/\delta')}\end{aligned}$$

with probability at least $1 - \delta' \log_2 T$.

Finally, taking a union bound over all $x \in D$ and rearranging (using the fact that $\max\{a + b, c\} \le a + \max\{b, c\}$ if $a \ge 0$) gives the required result. ∎

5.2 Potential Function Analysis

By Lemma 4 and our choice of $\eta = \frac{1}{\sqrt{nT} + 2\sqrt{\ln(1/\delta')}}$, we have

$$|\eta \widetilde{L}_t(x)| \le 1\,.$$

In the following computation, we will use the facts that $e^{-a} \le 1 - a + a^2$ whenever $|a| \le 1$.

$$\begin{aligned}\frac{W_{t+1}}{W_t} &= \sum_{x \in D} \frac{w_t(x)\exp(-\eta \widetilde{L}_t(x))}{W_t} \\ &\le \sum_{x \in D} \frac{w_t(x)}{W_t}(1 - \eta \widetilde{L}_t(x) + \eta^2 (\widetilde{L}_t(x))^2) \\ &\le 1 + \frac{\eta}{1 - \gamma}\left(-\sum_{x \in D} p_t(x)\widetilde{L}_t(x)\right. \\ &\qquad \left. + \sum_{x \in \text{spanner}} \frac{\gamma}{n}\widetilde{L}_t(x) + \sum_{x \in D} p_t(x)\eta(\widetilde{L}_t(x))^2\right)\end{aligned}$$

since by definition of p_t,

$$\frac{w_t(x)}{W_t} = \frac{p_t(x) - \frac{\gamma}{n}\mathbf{I}\{x \in \text{spanner}\}}{1 - \gamma}\,.$$

Note that we have,

$$\begin{aligned}&-\sum_{x \in D} p_t(x)\widetilde{L}_t(x) \\ &= -\sum_{x \in D} p_t(x)\widehat{L}_t^\dagger x + 2\sum_{x \in D} p_t(x) x^\dagger \mathbf{C}_t^{-1} x \sqrt{\frac{\ln(1/\delta')}{nT}} \\ &= -\sum_{x \in D} p_t(x)\widehat{L}_t^\dagger x + 2n\sqrt{\frac{\ln(1/\delta')}{nT}}\end{aligned}$$

where the last step is by Lemma 3.

Further, since $(b + c)^2 \le 2(b^2 + c^2)$ for every b, c, apply-

ing the definition of $\widetilde{L}_t(x)$, we also have

$$\begin{aligned}
&\sum_{x\in D} p_t(x)\eta(\widetilde{L}_t(x))^2 \\
&\le 2\eta \sum_{x\in D} p_t(x)\left((\widehat{L}_t^\dagger x)^2 + (2x^\dagger \mathbf{C}_t^{-1}x)^2\frac{\ln(1/\delta')}{nT}\right) \\
&\le 2\eta \sum_{x\in D} p_t(x)\left((\widehat{L}_t^\dagger x)^2 + 4x^\dagger \mathbf{C}_t^{-1}x\frac{n^2\ln(1/\delta')}{\gamma nT}\right) \\
&= 2\eta\left[\sum_{x\in D} p_t(x)(\widehat{L}_t^\dagger x)^2 + \frac{4\ln(1/\delta')}{\sqrt{nT}}\sum_{x\in D} p_t(x)x^\dagger \mathbf{C}_t^{-1}x\right] \\
&= 2\eta\left[\sum_{x\in D} p_t(x)(\widehat{L}_t^\dagger x)^2 + \frac{4\sqrt{n}\ln(1/\delta')}{\sqrt{T}}\right]
\end{aligned}$$

by successive applications of Lemma 3.

Putting these together, we have

$$\begin{aligned}
\frac{W_{t+1}}{W_t} \le 1 + \frac{\eta}{1-\gamma}\Bigg(&- \sum_{x\in D} p_t(x)\widehat{L}_t^\dagger x \\
&+ 2\sqrt{\frac{n\ln(1/\delta')}{T}} \\
&+ \sum_{x\in\text{spanner}} \frac{\gamma}{n}\widetilde{L}_t(x) \\
&+ 2\eta\sum_{x\in D} p_t(x)(\widehat{L}_t^\dagger x)^2 \\
&+ 8\eta\frac{\sqrt{n}\ln(1/\delta')}{\sqrt{T}}\Bigg)
\end{aligned}$$

Taking logs, using the fact that $\ln(1+x)\le x$, and summing over t, we have

$$\begin{aligned}
\ln\left(\frac{W_{T+1}}{W_1}\right) \le \frac{\eta}{1-\gamma}\Bigg[&- \sum_{t=1}^{T}\sum_{x\in D} p_t(x)\widehat{L}_t^\dagger x \\
&+ 2\sqrt{nT\ln(1/\delta')} \\
&+ \sum_{t=1}^{T}\sum_{x\in\text{spanner}} \frac{\gamma}{n}\widetilde{L}_t(x) \\
&+ 2\eta\sum_{t=1}^{T}\sum_{x\in D} p_t(x)(\widehat{L}_t^\dagger x)^2 \\
&+ 8\eta\ln(1/\delta')\sqrt{nT}\Bigg] \qquad (2)
\end{aligned}$$

The next three lemmas will bound the three summations that appear on the right hand side above.

Lemma 6 *With probability at least $1-\delta$,*

$$\begin{aligned}
&\sum_{t=1}^{T} L_t^\dagger x_t - \sum_{t=1}^{T}\sum_x p_t(x)\widehat{L}_t^\dagger x \\
&\quad\le (\sqrt{n}+1)\sqrt{2T\ln(1/\delta)} + \frac{4}{3}\ln(1/\delta)\left(\frac{n^2}{\gamma}+1\right).
\end{aligned}$$

Proof: Let us define $\overline{x} := \mathbb{E}_{x\sim p_t} x = \sum_{x\in D} p_t(x)x$ and $Y_t := \ell_t - \widehat{L}_t^\dagger\overline{x}$. Note that $\mathbb{E}_t\,\widehat{L}_t^\dagger\overline{x} = \mathbb{E}_t\,\ell_t$ and therefore Y_t is a martingale difference sequence.

We bound the conditional variance of Y_t as follows.

$$\begin{aligned}
&\sqrt{\mathrm{Var}_t\, Y_t} = \sqrt{\mathbb{E}_t(Y_t^2)} \\
&= \sqrt{\mathbb{E}_t\left(\left(\widehat{L}_t^\dagger\overline{x} - \ell_t\right)^2\right)} \\
&\le \sqrt{\mathbb{E}_t\left(\widehat{L}_t^\dagger\overline{x}\right)^2} + \sqrt{\mathbb{E}_t\left(\ell_t^2\right)} && \text{by Cauchy-Schwarz} \\
&\le \sqrt{\mathbb{E}_t\left(\widehat{L}_t^\dagger\overline{x}\right)^2} + 1 && \text{since } |\ell_t|\le 1 \\
&\le \sqrt{\overline{x}^\dagger C_t^{-1}\overline{x}} + 1 && \text{by Lemma 3} \\
&\le \sqrt{\mathop{\mathbb{E}}_{x\sim p_t} x^\dagger C_t^{-1}x} + 1 && \text{by Jensen's inequality} \\
&= \sqrt{n} + 1 && \text{by Lemma 3.}
\end{aligned}$$

Moreover, $|Y_t| \le n^2/\gamma + 1$ by Lemma 3. Applying Bernstein's inequality for martingale differences (see Appendix) to the sequence Y_t, we obtain that with probability at least $1-\delta$,

$$\sum_{t=1}^{T} Y_t \le (\sqrt{n}+1)\sqrt{2T\ln(1/\delta)} + \frac{4}{3}\ln(1/\delta)\left(\frac{n^2}{\gamma}+1\right),$$

which is the desired bound. ∎

Lemma 7 *With probability at least $1-\delta$,*

$$\sum_{t=1}^{T}\sum_{x\in\text{spanner}} \frac{\gamma}{n}\widetilde{L}_t(x) \le \gamma T + 2\gamma\left(1+\sqrt{nT}\right)\ln(1/\delta').$$

Proof: Using Lemma 5, with probability at least $1-\delta$, we have, for all $x\in$ spanner,

$$\begin{aligned}
\frac{\gamma}{n}\sum_t \widetilde{L}_t(x) &\le \frac{\gamma}{n}\sum_t L_t^\dagger x + \frac{2\gamma}{n}\left(1+\sqrt{nT}\right)\ln(1/\delta') \\
&\le \frac{\gamma T}{n} + \frac{2\gamma}{n}\left(1+\sqrt{nT}\right)\ln(1/\delta'),
\end{aligned}$$

because $L_t^\dagger x$, being the loss of an element of the spanner, is bounded by 1. Summing over the n elements of the spanner, we get the desired bound. ∎

Lemma 8 *With probability at least $1-\delta$,*

$$\sum_{t=1}^{T}\sum_x p_t(x)(\widehat{L}_t^\dagger x)^2 \le nT + T\sqrt{2n\ln(1/\delta)}.$$

Proof: First we observe that for $1\le t\le T$,

$$\begin{aligned}
\sum_x p_t(x)(\widehat{L}_t^\dagger x)^2 &= \sum_x p_t(x)\widehat{L}_t^\dagger x x^\dagger \widehat{L}_t \\
&= \widehat{L}_t^\dagger\left(\sum_x p_t(x)xx^\dagger\right)\widehat{L}_t \\
&= \ell_t^2 x_t^\dagger \mathbf{C}_t^{-1}\mathbf{C}_t\mathbf{C}_t^{-1}x_t \\
&\le x_t^\dagger \mathbf{C}_t^{-1}x_t
\end{aligned}$$

Summing over t,

$$\sum_{t=1}^{T}\sum_{x} p_t(x)(\widehat{L}_t^\dagger x)^2 \le \sum_{t=1}^{T} x_t^\dagger \mathbf{C}_t^{-1} x_t \ .$$

Lemma 3 tells us that, on the one hand, the summands $x_t^\dagger \mathbf{C}_t^{-1} x_t$ are uniformly bounded by $n^2/\gamma = \sqrt{nT}$, and on the other hand, that each one has expectation n, even conditioned on the previous ones.

Applying the Hoeffding-Azuma inequality to the martingale difference sequence

$$x_t^\dagger \mathbf{C}_t^{-1} x_t - \mathbb{E}_{x\sim p_t} x^\dagger \mathbf{C}_t^{-1} x$$

it follows that, with probability at least $1-\delta$,

$$\sum_{t=1}^{T} x_t^\dagger \mathbf{C}_t^{-1} x_t \le nT + T\sqrt{2n\ln(1/\delta)},$$

completing the proof. ■

Substituting the bounds of Lemmas 6, 7 and 8 into (2), we obtain that with probability at least $1-3\delta$,

$$\begin{aligned}\ln\left(\frac{W_{T+1}}{W_1}\right) \le \frac{\eta}{1-\gamma}\Bigg[&-\sum_{t=1}^{T} L_t^\dagger x_t \\ &+ (\sqrt{n}+1)\sqrt{2T\ln(1/\delta)} \\ &+ \frac{4}{3}\ln(1/\delta)\left(\frac{n^2}{\gamma}+1\right) \\ &+ 2\sqrt{nT\ln(1/\delta')} + \gamma T \\ &+ 2\gamma\left(1+\sqrt{nT}\right)\ln(1/\delta') \\ &+ 2\eta nT + 2\eta T\sqrt{2n\ln(1/\delta)} \\ &+ 8\eta\ln(1/\delta')\sqrt{nT}\Bigg] \end{aligned} \quad (3)$$

On the other hand, using Lemma 5, we have with probability at least $1-\delta$, for all $x \in D$,

$$\begin{aligned}\ln\frac{W_{T+1}}{W_1} &\ge -\eta\left(\sum_{t=1}^{T}\widetilde{L}_t(x)\right) - \ln|D| \\ &\ge -\eta\sum_{t=1}^{T} L_t^\dagger x - 2\eta(1+\sqrt{nT})\ln(1/\delta') - \ln|D|. \end{aligned} \quad (4)$$

Combining (3) with (4), we have that with probability at least $1-4\delta$, for every $x \in D$,

$$\begin{aligned}\sum_{t=1}^{T} L_t^\dagger x_t \le &\sum_{t=1}^{T} L_t^\dagger x \\ &+ 2(1+\sqrt{nT})\ln(1/\delta') \\ &+ \frac{1}{\eta}\ln|D| \\ &+ (\sqrt{n}+1)\sqrt{2T\ln(1/\delta)} \\ &+ \frac{4}{3}\ln(1/\delta)\left(\frac{n^2}{\gamma}+1\right) \\ &+ 2\sqrt{nT\ln(1/\delta')} + \gamma T \\ &+ 2\gamma\left(1+\sqrt{nT}\right)\ln(1/\delta') \\ &+ 2\eta nT + 2\eta T\sqrt{2n\ln(1/\delta)} \\ &+ 8\eta\ln(1/\delta')\sqrt{nT}\end{aligned}$$

Recall that $\eta = \frac{1}{\sqrt{nT}+2\sqrt{\ln(1/\delta')}}$, $\gamma = \frac{n^{3/2}}{\sqrt{T}}$, $\delta' = \delta/(|D|\log_2 T)$, and $\ln|D| = O(n\ln T)$. Plugging in these values yields

$$\sum_{t=1}^{T} L_t^\dagger x_t \le L_{\min} + O(n^{3/2}\sqrt{T}\ln(nT/\delta)),$$

completing the proof of Theorem 1.

6 Conclusions and Open Problems

We presented an algorithm that achieves the desired regret bound of $O^*(\sqrt{T})$ with high probability. However, the quest for an *efficient* algorithm with the same high-probability guarantee, even for the special case of bandit online shortest paths, is still open. Achieving similar results for general convex functions is also an intriguing open question.

A Concentration Inequalities

The following inequalities are well known. Theorem 9 is from [10]. Lemmas 10 and 11 can be found, for instance, in [6], Appendix A.

Theorem 9 (Freedman) *Suppose $X_1, \ldots, X_T$ is a martingale difference sequence, and b is an uniform upper bound on the steps X_i. Let V denote the sum of conditional variances,*

$$V = \sum_{i=1}^{n} \mathbf{Var}\,(X_i \mid X_1, \ldots, X_{i-1}).$$

Then, for every $a, v > 0$,

$$\mathrm{Prob}\left(\sum X_i \ge a \text{ and } V \le v\right) \le \exp\left(\frac{-a^2}{2v + 2ab/3}\right).$$

Lemma 10 (Bernstein's inequality for martingales) *Let Y_1, ..., Y_T be a martingale difference sequence. Suppose that $Y_t \in [a,b]$ and*

$$\mathbb{E}[Y_t^2 | X_{t-1}, \ldots, X_1] \le v \text{ a.s.}$$

for all $t \in \{1, \ldots, T\}$. Then for all $\delta > 0$,

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

Lemma 11 (Hoeffding-Azuma inequality) *Let $Y_1, \ldots, Y_T$ be a martingale difference sequence. Suppose that $|Y_t| \leq c$ almost surely for all $t \in \{1, \ldots, T\}$. Then for all $\delta > 0$,*

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

References

[1] Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, 2008. to appear.

[2] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2003.

[3] Baruch Awerbuch, David Holmer, Herb Rubens, and Robert Kleinberg. Provably competitive adaptive routing. In *Proceedings of the 31st IEEE INFOCOM*, volume 1, pages 631–641, 2005.

[4] Baruch Awerbuch and Robert Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the 36th ACM Symposium on Theory of Computing (STOC)*, 2004.

[5] Nicolò Cesa-Bianchi and Claudio Gentile. Improved risk tail bounds for on-line algorithms. *IEEE Transactions on Information Theory*, 54(1):386–39, Jan 2008.

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

[7] Varsha Dani and Thomas P. Hayes. Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary. In *Proceedings of the 17th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2006.

[8] Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. The price of bandit information for online optimization. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20 (NIPS 2007)*. MIT Press, 2008.

[9] Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, 2008. to appear.

[10] David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, Feb 1975.

[11] András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8:2369–2403, 2007.

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

[13] Robert Kleinberg. *Online decision problems with large strategy sets*. PhD thesis, MIT, 2005.

[14] H. Brendan McMahan and Avrim Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *Proceedings of the 17th Annual Conference on Learning Theory (COLT)*, 2004.

[15] Eiji Takimoto and Manfred K. Warmuth. Path kernels and multiplicative updates. *Journal of Machine Learning Research*, 4:773–818, 2003.