Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Online Bandit Learning for a Special Class of Non-convex Losses

Lijun Zhang[1] and **Tianbao Yang**[2] and **Rong Jin**[3] and **Zhi-Hua Zhou**[1]
[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
[2]Department of Computer Science, the University of Iowa, Iowa City, IA 52242, USA
[3]Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA
{zhanglj, zhouzh}@lamda.nju.edu.cn, tianbao-yang@uiowa.edu, rongjin@cse.msu.edu

Abstract

In online bandit learning, the learner aims to minimize a sequence of losses, while only observing the value of each loss at a single point. Although various algorithms and theories have been developed for online bandit learning, most of them are limited to convex losses. In this paper, we investigate the problem of online bandit learning with non-convex losses, and develop an efficient algorithm with formal theoretical guarantees. To be specific, we consider a class of losses which is a composition of a non-increasing scalar function and a linear function. This setting models a wide range of supervised learning applications such as online classification with a non-convex loss. Theoretical analysis shows that our algorithm achieves an $\widetilde{O}(poly(d)T^{2/3})$ regret bound when the variation of the loss function is small. To the best of our knowledge, this is the first work in online bandit learning that does not rely on convexity.

Introduction

Online decision-making has become a popular learning paradigm in many disciplines such as Artificial Intelligence, Economics and Control Theory (Saha and Tewari 2011). At each round of online learning, the learner chooses a decision from a given set, and an adversary responds with a loss function that decides the cost of decisions. The performance of an online learning algorithm is measured by the regret, which is the difference between the total cost of the decisions it chooses, and the cost of the optimal decision chosen in hindsight. According to the amount of information revealed to the learner, online learning can be classified into two categories (Cesa-Bianchi and Lugosi 2006): i) full information setting where the learner observes the entire cost function, and ii) bandit setting where only the cost of the selected action is available.

In the past decades, there have been tremendous progresses made in online bandit learning, ranging from multiarmed bandit (Robbins 1952; Auer et al. 2003), online linear optimization with bandit feedback (Awerbuch and Kleinberg 2004; Dani, Kakade, and Hayes 2008; Abernethy, Hazan, and Rakhlin 2008), to online convex optimization with bandit feedback (Flaxman, Kalai, and McMahan 2005; Saha and Tewari 2011; Agarwal, Dekel, and Xiao 2010). A major limitation of the previous work is that most of them are restricted to convex losses. The drawback of using convex losses has been revealed by several recent studies. In (Calauzènes, Usunier, and Gallinari 2012), the authors show it is impossible to find any convex loss that is calibrated with the standard evaluation metrics for ranking. Similarly, for multi-label learning, no convex loss is consistent with the popular ranking loss (Gao and Zhou 2011). The success of deep learning also indicates the significance of using non-convex losses (Hinton, Osindero, and Teh 2006).

The resurge of non-convex losses in machine learning motives us to investigate online bandit learning with non-convex loss functions. In particular, we consider the following learning protocol.

- At the t-th round, the learner submits a point $\mathbf{x}_t \in \mathbb{R}^d$ with $\|\mathbf{x}_t\|_2 \leq 1$, and simultaneously an oblivious adversary selects a vector $\mathbf{u}_t$ and a non-increasing scalar function $f_t : \mathbb{R} \mapsto \mathbb{R}$ that assesses the consistency between the ground truth $\mathbf{u}_t$ and the answer $\mathbf{x}_t$ by $f_t(\mathbf{x}_t^\top \mathbf{u}_t)$.
- Instead of revealing the loss function $f_t(\langle \mathbf{u}_t, \cdot \rangle)$ directly, the adversary only provides the learner $c_t \in \mathbb{R}$ whose expectation is the cost $f_t(\mathbf{u}_t^\top \mathbf{x}_t)$, i.e.,
$$\mathrm{E}_{t-1}[c_t] = f_t(\mathbf{u}_t^\top \mathbf{x}_t), \tag{1}$$
where $\mathrm{E}_{t-1}[\cdot]$ is the expectation conditioned on the randomness until round $t-1$.

Notice that the above protocol generalizes many machine learning tasks. Taking the online classification as an example, we can set $\mathbf{u}_t = y_t \mathbf{z}_t$, where $\mathbf{z}_t \in \mathbb{R}^d$ is an instance, $y_t \in \{\pm 1\}$ is the assigned class label, and $f_t(\cdot)$ can be any loss function such as the ramp loss (Ertekin, Bottou, and Giles 2011).

We emphasize that in our setting both $\mathbf{u}_t$ and $f_t(\cdot)$ are *unknown* to the learner. More importantly, unlike most online learning that assume the loss function to be convex, in this study, $f_t(\cdot)$ can be non-convex. This relaxation makes our problem significantly more challenging than most online learning problems, including the traditional online bandit learning. Following the convention in online learning, our goal is to generate a sequence of answer vectors $\mathbf{x}_1, \ldots, \mathbf{x}_T$, that leads to a small regret defined below

$$\text{regret} = \sum_{t=1}^{T} f_t(\mathbf{u}_t^\top \mathbf{x}_t) - \min_{\|\mathbf{x}\|_2 \leq 1} \sum_{t=1}^{T} f_t(\mathbf{u}_t^\top \mathbf{x}).$$

We present a simple algorithm for online bandit learning with non-convex losses. It is computationally efficient and achieves a non-trivial regret bound under appropriate conditions. Our approach follows the standard exploration-exploitation framework for bandit learning (Awerbuch and Kleinberg 2004; McMahan and Blum 2004). In an exploration round, the algorithm submits a random vector in order to obtain an unbiased estimate of $\mathbf{u}_t$, and updates the current solution based on the bandit feedback. In an exploitation round, it submits the current solution in order to incur a small loss. Under the assumption that $f_t(\cdot)$ is non-increasing and Lipschitz continuous, we are able to bound the regret by the number of iterations and the variation of the target vectors $\{\mathbf{u}_t\}_{t=1}^T$. To be specific, the regret bound takes the form $\widetilde{O}(poly(d)T^{2/3}+\sqrt{TV_T})$,[1] where V_T is the variation of vectors $\mathbf{u}_1, \ldots, \mathbf{u}_T$. Thus, the proposed algorithm achieves an $\widetilde{O}((poly(d)T^{2/3})$ regret bound if $V_T \leq O(T^{1/3})$.

Related Work

In this section, we briefly review the related work in online convex and non-convex optimizations.

Online Convex Optimization

In the full information setting, online convex optimization has been extensively studied (Kivinen, Smola, and Williamson 2002; Zhang et al. 2013). Zinkevich (2003) shows that a simple online gradient descent algorithm achieves an $O(\sqrt{T})$ regret bound for convex and Lipschitz continuous functions. When the loss function is strongly convex, the regret bound can be improved to $O(\log T)$ (Hazan, Agarwal, and Kale 2007). Both the $O(\sqrt{T})$ and $O(\log T)$ regrets bounds, for convex and strongly convex loss functions respectively, are known to be minimax optimal (Abernethy et al. 2009).

Compared to the full information setting, the regret bound for the bandit setting is usually worse and has an explicit dependence on the dimensionality d. The current best-known regret bounds are $O(dT^{3/4})$, $\widetilde{O}(d^{2/3}T^{2/3})$, $O(d^{2/3}T^{2/3})$, and $O(d\sqrt{T})$ for convex, convex-and-smooth, strongly convex, and strongly-convex-and-smooth functions, respectively (Flaxman, Kalai, and McMahan 2005; Saha and Tewari 2011; Agarwal, Dekel, and Xiao 2010). Notice that when the learner is allowed to query the loss function at multiple points, the regret can be improved to match its counterpart in the full information setting (Agarwal, Dekel, and Xiao 2010).

In the bandit setting, there are two special cases that are well-studied: multiarmed bandit and online linear optimization with bandit feedback. In the first problem, we assume there are K arms, and a gambler pulls one of them to receive a reward in each round. Auer et al. (2003) prove that the gambler's regret can be bounded by $\widetilde{O}(\sqrt{KT})$, which is optimal up to logarithmic factors (Audibert and Bubeck 2009). Furthermore, if the reward function has some structural properties, such as Lipschitz (Magureanu, Combes, and Proutiere 2014), the regret could be further improved. The online linear optimization problem with bandit feedback was first introduced by Awerbuch and Kleinberg (2004) who obtained a $O(d^{3/5}T^{2/3})$ regret bound against an oblivious adversary. Later, McMahan and Blum (2004) achieved an $O(poly(d)T^{3/4})$ regret bound against an adaptive adversary. In (Dani, Kakade, and Hayes 2008; Abernethy, Hazan, and Rakhlin 2008), the regret bound was improved to $O(poly(d)\sqrt{T})$, where the dependence on T is optimal (Bubeck, Cesa-Bianchi, and Kakade 2012).

Online Non-convex Optimization

Several heuristic approaches have been developed for online learning with non-convex loss in the full information setting, such as the online version of the concave-convex procedure (Ertekin, Bottou, and Giles 2011; Gasso et al. 2011). However, none of them are equipped with a formal regret bound. One exception is the online submodular minimization (Hazan and Kale 2012) that achieves $O(\sqrt{dT})$ and $O(dT^{2/3})$ regret bounds in the full information and bandit settings, respectively. But these algorithms rely on the specific property of submodular function (i.e., the Lovász extension is convex), and thus cannot be applied to the problem considered here.

An Efficient Algorithm for Online Bandit Learning

We first describe the proposed algorithm for online bandit learning with non-convex losses, and then state its theoretical guarantees.

The Algorithm

Algorithm 1 summarizes the key steps of the proposed algorithm. We maintain two sequences of vectors during the learning process: the answer vectors $\mathbf{x}_t$ and the auxiliary vector $\mathbf{w}_t$. We initialize the answer vector $\mathbf{x}_1$ to be a random normalized vector, and the auxiliary vector $\mathbf{w}_1$ to be 0. At each iteration t, we generate a Bernoulli random variable Z_t with $\Pr(Z_t = 1) = \eta$ to determine whether to explore or exploit. When $Z_t = 0$, we will simply submit the answer vector $\mathbf{x}_t$ as the solution, and make no update. When $Z_t = 1$, we will first compute a normalized Gaussian random vector $\mathbf{v}_t/\|\mathbf{v}_t\|_2$ and submit it as the answer. Based on the received feedback c_t, we update the auxiliary vector and the answer vector by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{c_t}{\|\mathbf{v}_t\|_2}\mathbf{v}_t \text{ and } \mathbf{x}_{t+1} = \frac{\mathbf{w}_{t+1}}{\|\mathbf{w}_{t+1}\|_2}.$$

We note that for the sake of simplicity, Algorithm 1 follows the early studies of online bandit learning that separate the exploration steps from the exploitation steps (Awerbuch and Kleinberg 2004; McMahan and Blum 2004). This is different from the more recent strategy for exploration-exploitation (Flaxman, Kalai, and McMahan 2005; Abernethy, Hazan, and Rakhlin 2008) that usually combines exploration and exploitation into a single step by adding random perturbation to the submitted solutions. We will exam-

[1] We use the $\widetilde{O}$ notation to hide constant factors as well as polylogarithmic factors in d and T.

Algorithm 1 An Efficient Algorithm for Online Bandit Learning

Input: step size η and number of trials T

Set $\eta = T^{-1/3}$
Initialize $\mathbf{x}_1$ as any random normalized vector and $\mathbf{w}_1 = \mathbf{0}$
for $t = 1, 2, \ldots, T$ **do**
  Sample binary random variable Z_t with $\Pr(Z_t = 1) = \eta$.
  Sample a random vector $\mathbf{v}_t$ from an Gaussian distribution $\mathcal{N}(0, I_d)$
  Submit the solution $\mathbf{x}'_t = Z_t \frac{\mathbf{v}_t}{\|\mathbf{v}_t\|_2} + (1 - Z_t)\mathbf{x}_t$
  Receive c_t from the adversary
  Update the auxiliary vector $\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{Z_t c_t}{\|\mathbf{v}_t\|_2}\mathbf{v}_t$
  Update the answer vector $\mathbf{x}_{t+1} = \mathbf{w}_{t+1}/\|\mathbf{w}_{t+1}\|_2$
end for

ine in the future the second strategy for online bandit learning with non-convex losses.

Finally, we would like to point out that following the idea of discretizing the decision space (Dani, Kakade, and Hayes 2008), our problem can be reduced to the multiarmed bandit and solved by existing methods (Auer et al. 2003). However, this strategy is inefficient because the number of arms is exponential in the dimensionality d, and the regret bound may also have a high dependence on d. In contrast, our algorithm is very efficient and the regret bound only has a polynomial dependence on d.

The Main Results

Besides the basic assumption in (1), we further make the following assumptions in our analysis.

- $f_t(\cdot)$ is non-increasing and L-Lipschitz continuous.
- Both c_t and $f_t(\cdot)$ are upper bounded by a constant B. That is,

$$\sup_{t\in[T]} |c_t| \leq B, \text{ and } \sup_{t\in[T],\|\mathbf{x}\|_2\leq 1} |f_t(\mathbf{u}_t^\top \mathbf{x})| \leq B. \quad (2)$$

- The target vectors $\mathbf{u}_t$'s are of unit length, i.e., $\|\mathbf{u}_t\|_2 = 1$, $t = 1, \ldots T$.
- The adversary is oblivious, meaning that both $\mathbf{u}_t$'s and f_t's are fixed.

As a starting point, we first analyze the regret bound for the simplest case when all the target vectors are the same, and then move to the general case.

Regret Bound for a Single Target Vector We first consider the simplest case when $\mathbf{u}_1 = \mathbf{u}_2 = \cdots = \mathbf{u}_T = \mathbf{u}$ and $f_1 = f_2 = \cdots = f_T = f$.

Define $h(z)$ as the probability density function (PDF) of the inner product of random unit vectors in $\mathbb{R}^d$. When $d = 1$, it is easy to verify

$$h(z) = \frac{1}{2}(\delta(z-1) + \delta(z+1)),$$

where $\delta(\cdot)$ is the Dirac delta function. When $d \geq 2$, we have (Cho 2009)

$$h(z) = \begin{cases} \frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)\sqrt{\pi}}(1-z^2)^{\frac{d-3}{2}}, & \text{for } -1 < z < 1 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $\Gamma(\cdot)$ is the Gamma function. The following proposition, which is inspired by the recent developments in one-bit compressive sensing (Plan and Vershynin 2013; Zhang, Yi, and Jin 2014), provides the key observation for our analysis.

Proposition 1. *We have*

$$\mathrm{E}_{t-1}\left[-\frac{Z_t c_t}{\|\mathbf{v}_t\|_2}\mathbf{v}_t\right] = \gamma\eta\mathbf{u},\ t = 1, \ldots, T \quad (4)$$

where

$$\gamma = -\int_{-1}^{1} f(z)h(z)z\,\mathrm{d}\,z. \quad (5)$$

From our assumption that $f(\cdot)$ is non-increasing, it is easy to verify that $\gamma \geq 0$. We note that γ will have a strong dependence on d. Generally speaking, we have $\gamma = poly(d^{-1})$. For instance, when $f(z) = -z$, we have $\gamma = \mathrm{E}[z^2] = d^{-1}$ (Cho 2009).

Proposition 1 shows that in the exploration step, our algorithm is able to find an unbiased estimate of $\mathbf{u}$, up to a scaling factor. Based on this observation, we obtain the following regret bound.

Theorem 1. *Assume*

$$T \geq \max\left\{e, \left(\sqrt{\log\frac{T(d+1)}{\delta}}\log T\right)^3\right\}. \quad (6)$$

Set $\eta = T^{-1/3}$ in Algorithm 1. Then, with a probability $1 - \delta - \tau$, we have

$$\textit{regret} \leq 4B\left(\log\frac{1}{\tau} + 1\right) + 4B\left(\frac{3L}{\gamma}\sqrt{\log\frac{T(d+1)}{\delta}} + 1\right)T^{\frac{2}{3}}.$$

To simplify the presentation, we assume the horizon T is known so that we can choose $\eta = T^{-1/3}$ in the algorithm. This limitation could be addressed by the well-known "doubling trick" (Cesa-Bianchi and Lugosi 2006). Theorem 1 implies our algorithm achieves an $\widetilde{O}(\gamma^{-1}T^{2/3})$ regret bound, which is even better than the regret bound in the general online convex optimization with bandit feedback (Flaxman, Kalai, and McMahan 2005). From the discussion in (Dani and Hayes 2006; Abernethy, Hazan, and Rakhlin 2008), we also know that $\Omega(T^{2/3})$ regret bound is unavoidable if any algorithm ignores the feedback received during exploitation. We finally note that although the regret bound in Theorem 1 does not have an explicit dependence on d, its dependence on d comes from γ.

We now extend the simple case to a slightly more complicated scenario where a different loss function is used. In this case, we have the following proposition.

Proposition 2. *We have*

$$\mathrm{E}_{t-1}\left[-\frac{Z_t c_t}{\|\mathbf{v}_t\|_2}\mathbf{v}_t\right] = \gamma_t \eta \mathbf{u},\ t = 1, \ldots, T$$

where

$$\gamma_t = -\int_{-1}^{1} f_t(z) h(z) z \,\mathrm{d}\, z.$$

Following almost the same analysis for Theorem 1, we obtain the following theorem to bound the regret.

Theorem 2. *Suppose T and η satisfy the conditions in Theorem 1. With a probability $1-\delta-\tau$, we have*

$$\textit{regret} \leq 4B\left(\log\frac{1}{\tau}+1\right) + 4B\left(\frac{3L}{\bar{\gamma}}\sqrt{\log\frac{T(d+1)}{\delta}}+1\right)T^{\frac{2}{3}}$$

where $\bar{\gamma} = \min_{t\in[T]} \gamma_t$.

The only difference between Theorems 2 and 1 is that γ in Theorem 1 is replaced with $\bar{\gamma}$, the smallest one among $\{\gamma_t\}_{t=1}^T$.

Regret Bound for the General Case We now consider the more general case where each $\mathbf{u}_t$ is a different vector. Let $\bar{\mathbf{u}}_t$ be the average of vectors $\mathbf{u}_1, \ldots, \mathbf{u}_t$, i.e.,

$$\bar{\mathbf{u}}_t = \frac{1}{t}\sum_{i=1}^{t}\mathbf{u}_t.$$

Similar to Theorem 1, we have the following regret bound for the general case when a single loss function is used.

Theorem 3. *Suppose T and η satisfy the conditions in Theorem 1. Then, with a probability $1-\delta-\tau$, we have*

$$\begin{aligned}\textit{regret} \leq & 4B\left(\log\frac{1}{\tau}+1\right) + 2L\sum_{t=2}^{T}\|\mathbf{u}_t - \bar{\mathbf{u}}_{t-1}\|_2 \\ & + 4B\left(\frac{3L}{\gamma\rho_T}\sqrt{\log\frac{T(d+1)}{\delta}}+1\right)T^{\frac{2}{3}},\end{aligned}$$

where γ is defined in (5) and $\rho_T = \min_{t\in[T]}\|\bar{\mathbf{u}}_t\|_2$.

To bound the second term in the regret bound, we need the following inequality (Hazan and Kale 2010)

$$\sum_{t=2}^{T}\|\mathbf{u}_t - \bar{\mathbf{u}}_{t-1}\|_2^2 \leq \sum_{t=1}^{T}\|\mathbf{u}_t - \bar{\mathbf{u}}_T\|_2^2 + 12\sqrt{\sum_{t=1}^{T}\|\mathbf{u}_t - \bar{\mathbf{u}}_T\|_2^2}. \tag{7}$$

From (7) and Theorem 3, we have

$$\begin{aligned}\text{regret} \leq & 4B\left(\log\frac{1}{\tau}+1\right) + 2L\sqrt{T(V_T + 12\sqrt{V_T})} \\ & + 4B\left(\frac{3L}{\gamma\rho_T}\sqrt{\log\frac{T(d+1)}{\delta}}+1\right)T^{\frac{2}{3}}\end{aligned}$$

where

$$V_T = \sum_{t=1}^{T}\|\mathbf{u}_t - \bar{\mathbf{u}}_T\|_2^2.$$

When the total variation $V_T \leq O(T^{1/3})$, the additional term $\sqrt{T(V_T + 12\sqrt{V_T})}$ is on the order of $T^{2/3}$. Furthermore, if we assume a small variation for each iteration, that is, $V_t \leq O(t^{1/3})$ for all $t \in [T]$, each $\|\bar{\mathbf{u}}_t\|_2$ will be lower bounded by some constant,[2] and thus $1/\rho_T$ is upper bounded by some constant. As a result, we still have an $\widetilde{O}(\gamma^{-1}T^{2/3})$ regret bound. On the other hand, the regret bound becomes trivial when $V_T = \Omega(T)$, which is significantly worse than the previous results on variation based regret bound (Hazan and Kale 2010). This is because we are dealing with non-convex optimizations and therefore the gradient does not provide an universal lower bound for the entire function. Thus, we cannot utilize the analysis for convex functions, but only rely on the assumption that $f_t(\cdot)$ is non-increasing and Lipschitz continuous. Finally, it is straightforward to extend the above result to the case when a different loss function is used, by introducing $\bar{\gamma}$ defined in Theorem 2.

Analysis

We here present the proofs of main theorems. The omitted proofs are provided in the supplementary material.

Proof of Theorem 1

Under the assumption $\mathbf{u}_1 = \mathbf{u}_2 = \cdots = \mathbf{u}_T = \mathbf{u}$, we have

$$\begin{aligned}&\text{regret} \\ =&\sum_{t=1}^{T} f(\mathbf{u}^\top \mathbf{x}_t') - T\min_{\|\mathbf{x}\|\leq 1} f(\mathbf{u}^\top \mathbf{x}) \\ =&\sum_{t=1}^{T} Z_t\left(f\left(\frac{\mathbf{u}^\top \mathbf{v}_t}{\|\mathbf{v}_t\|_2}\right) - f(\mathbf{u}^\top \mathbf{x}_t)\right) \\ &+ \sum_{t=1}^{T}\left(f(\mathbf{u}^\top \mathbf{x}_t) - \min_{\|\mathbf{x}\|\leq 1} f(\mathbf{u}^\top \mathbf{x})\right) \\ \overset{(2)}{\leq}& 2B\underbrace{\sum_{t=1}^{T} Z_t}_{\Delta_1} + \underbrace{\sum_{t=1}^{T}\left(f(\mathbf{u}^\top \mathbf{x}_t) - \min_{\|\mathbf{x}\|\leq 1} f(\mathbf{u}^\top \mathbf{x})\right)}_{\Delta_2}.\end{aligned} \tag{8}$$

Then, we discuss how to bound Δ_1 and Δ_2.

According to the Multiplicative Chernoff Bound (Angluin and Valiant 1979) provided in the supplementary, we have with a probability at least $1-\tau$

$$\Delta_1 \leq 2\mathrm{E}[\Delta_1] + 2\log\frac{1}{\tau} = 2\eta T + 2\log\frac{1}{\tau}. \tag{9}$$

To bound Δ_2, we introduce the vector-valued martingale-difference sequence

$$\delta_i = -\frac{Z_i c_i}{\|\mathbf{v}_i\|_2}\mathbf{v}_i - \gamma\eta\mathbf{u},\ i = 1, \ldots, T. \tag{10}$$

The following lemma follows immediately from the Freedman's inequality for matrix martingales (Tropp 2011).

[2] We can prove it by contradiction. Suppose $\|\bar{\mathbf{u}}_t\|_2 \leq c$, then we must have $V_t \geq (1-c)^2 t = \Theta(t)$.

Lemma 1. *With a probability* $1-\delta$*, we have, for any* $t \in [T]$*,*

$$\left\| \sum_{i=1}^{t} \delta_i \right\|_2 \leq \rho_t(\delta)$$

where

$$\rho_t(\delta) = \frac{4B}{3} \log \frac{T(d+1)}{\delta} + B\sqrt{2\eta t \log \frac{T(d+1)}{\delta}}.$$

From the assumption that $f(\cdot)$ is non-increasing and L-Lipschitz continuous, we have the following lemma.

Lemma 2. *We have*

$$f(\mathbf{x}_t^\top \mathbf{u}) - \min_{\|\mathbf{x}\| \leq 1} f(\mathbf{x}^\top \mathbf{u}) \leq \begin{cases} 2B, & \text{if } t = 1; \\ \frac{2L}{\gamma\eta(t-1)} \left\| \sum_{i=1}^{t-1} \delta_i \right\|_2, & \text{otherwise.} \end{cases}$$

Based on Lemmas 1 and 2, we have with a probability at least $1 - \delta$,

$$\begin{aligned} \Delta_2 \leq & 2B + \frac{2L}{\gamma\eta} \sum_{t=2}^{T} \frac{\rho_{t-1}(\delta)}{t-1} \\ = & 2B + \frac{8LB}{3\gamma\eta} \log \frac{T(d+1)}{\delta} \sum_{t=1}^{T-1} \frac{1}{t} \\ & + \frac{2LB}{\gamma} \sqrt{\frac{2}{\eta} \log \frac{T(d+1)}{\delta}} \sum_{t=1}^{T-1} \frac{1}{\sqrt{t}} \qquad (11) \\ \overset{(6)}{\leq} & 2B + \frac{6LB}{\gamma\eta} \log \frac{T(d+1)}{\delta} \log T \\ & + \frac{4LB}{\gamma} \sqrt{\frac{2T}{\eta} \log \frac{T(d+1)}{\delta}}, \end{aligned}$$

where we use the following inequalities

$$\sum_{t=1}^{T} \frac{1}{t} \leq 1 + \int_{t=1}^{T} \frac{1}{t} dt = 1 + \log t|_1^T = \log T + 1, \text{ and}$$

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 1 + \int_{t=1}^{T} \frac{1}{\sqrt{t}} dt = 1 + 2\sqrt{t}|_1^T = 2\sqrt{T} - 1.$$

Combining (8), (9) and (11), we have with a probability at least $1 - \tau - \delta$,

$$\begin{aligned} & \text{regret} \\ \leq & 4B \left(\eta T + \log \frac{1}{\tau} + 1 \right) + \frac{6LB}{\gamma\eta} \log \frac{T(d+1)}{\delta} \log T \\ & + \frac{4LB}{\gamma} \sqrt{\frac{2T}{\eta} \log \frac{T(d+1)}{\delta}} \\ = & 4B \left(T^{\frac{2}{3}} + \log \frac{1}{\tau} + 1 \right) + \frac{6LB}{\gamma} T^{\frac{1}{3}} \log \frac{T(d+1)}{\delta} \log T \\ & + \frac{6LB}{\gamma} \sqrt{\log \frac{T(d+1)}{\delta}} T^{\frac{2}{3}} \\ \overset{(6)}{\leq} & 4B \left(T^{\frac{2}{3}} + \log \frac{1}{\tau} + 1 \right) + \frac{12LB}{\gamma} \sqrt{\log \frac{T(d+1)}{\delta}} T^{\frac{2}{3}}. \end{aligned}$$

Proof of Lemma 1

We first state the Freedman's inequality for matrix martingales below.

Theorem 4. *(Tropp 2011, Corollary 1.3) Let* $\|\cdot\|$ *be the spectral norm of a matrix, which returns its largest singular value. Consider a matrix martingale* $\{Y_i : i = 0, 1, 2, \dots\}$ *whose values are matrices with dimension* $d_1 \times d_2$*. Let* $\{X_i : i = 1, 2, 3, \dots\}$ *be the difference sequence, and assume that the difference sequence is uniformly bounded:*

$$\|X_i\| \leq R \text{ almost surely } i = 1, 2, 3, \dots.$$

Define two predictable quadratic variation processes for this martingale:

$$W_{col,t} := \sum_{i=1}^{t} \mathrm{E}_{i-1}[X_i X_i^\top],$$

$$W_{row,t} := \sum_{i=1}^{t} \mathrm{E}_{i-1}[X_i^\top X_i], \ t = 1, 2, 3, \dots.$$

Then, for all $\gamma \geq 0$ *and* $\sigma^2 > 0$*,*

$$\begin{aligned} & \Pr\left\{ \|Y_t\| \geq \gamma \text{ and } \max\{\|W_{col,t}\|, \|W_{row,t}\|\} \leq \sigma^2 \right\} \\ & \leq (d_1 + d_2) \exp\left(-\frac{\gamma^2/2}{\sigma^2 + R\gamma/3} \right). \end{aligned}$$

By setting $\delta = \exp\left(-\frac{\gamma^2/2}{\sigma^2 + R\gamma/3}\right)$, Theorem 4 implies that with a probability at most δ,

$$\|Y_t\| \geq \frac{2R}{3} \log \frac{d_1 + d_2}{\delta} + \sqrt{2\sigma^2 \log \frac{d_1 + d_2}{\delta}} \text{ and}$$
$$\max\{\|W_{col,t}\|, \|W_{row,t}\|\} \leq \sigma^2.$$

We then introduce several facts that will be used in our analysis. Let $\boldsymbol{\xi}$ be a random vector. Based on Jensen's inequality, we have

$$\|\mathrm{E}[\boldsymbol{\xi}]\|_2 \leq \mathrm{E}[\|\boldsymbol{\xi}\|]_2. \qquad (12)$$

From the property of positive semidefinite (PSD) matrices, we have

$$\begin{aligned} & \left\| E\left[(\boldsymbol{\xi} - E[\boldsymbol{\xi}])(\boldsymbol{\xi} - E[\boldsymbol{\xi}])^\top \right] \right\| \\ = & \left\| \mathrm{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top] - \mathrm{E}[\boldsymbol{\xi}]\mathrm{E}[\boldsymbol{\xi}]^\top \right\| \\ = & \boldsymbol{\alpha}^\top \left(\mathrm{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top] - \mathrm{E}[\boldsymbol{\xi}]\mathrm{E}[\boldsymbol{\xi}]^\top \right) \boldsymbol{\alpha} \\ \leq & \boldsymbol{\alpha}^\top \mathrm{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top] \boldsymbol{\alpha} \leq \|\mathrm{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top]\|, \end{aligned} \qquad (13)$$

where $\boldsymbol{\alpha}$ is the largest eigenvector of the PSD matrix $\mathrm{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top] - \mathrm{E}[\boldsymbol{\xi}]\mathrm{E}[\boldsymbol{\xi}]^\top$. Furthermore, it is easy to verify that

$$\begin{aligned} & \mathrm{E}\left[(\boldsymbol{\xi} - E[\boldsymbol{\xi}])^\top (\boldsymbol{\xi} - E[\boldsymbol{\xi}]) \right] \\ = & \mathrm{E}[\boldsymbol{\xi}^\top \boldsymbol{\xi}] - \mathrm{E}[\boldsymbol{\xi}]^\top \mathrm{E}[\boldsymbol{\xi}] \leq \mathrm{E}[\boldsymbol{\xi}^\top \boldsymbol{\xi}]. \end{aligned} \qquad (14)$$

Notice that the spectral norm of a vector is its ℓ_2-norm.

We bound the ℓ_2-norm of the δ_i as follows

$$\begin{aligned}\|\delta_i\|_2 &= \left\| -\frac{Z_i c_i}{\|\mathbf{v}_i\|_2}\mathbf{v}_i - \gamma\eta\mathbf{u} \right\|_2 \\ &\leq \left\| \frac{Z_i c_i}{\|\mathbf{v}_i\|_2}\mathbf{v}_i \right\|_2 + \|\gamma\eta\mathbf{u}\|_2 \\ &\overset{(4),(12)}{\leq} \left\| \frac{Z_i c_i}{\|\mathbf{v}_i\|_2}\mathbf{v}_i \right\|_2 + \mathrm{E}_{i-1}\left[\left\| \frac{Z_i c_i}{\|\mathbf{v}_i\|_2}\mathbf{v}_i \right\|_2 \right] \\ &\overset{(2)}{\leq} 2B,\ i = 1, \ldots, T.\end{aligned}$$

We then bound the two predictable quadratic variation processes as follows

$$\begin{aligned}&\left\| \sum_{i=1}^{t} \mathrm{E}_{i-1}[\delta_i \delta_i^\top] \right\| \\ &= \left\| \sum_{i=1}^{t} \mathrm{E}_{i-1}\left[\left(\frac{Z_i c_i \mathbf{v}_i}{\|\mathbf{v}_i\|_2} + \gamma\eta\mathbf{u} \right) \left(\frac{Z_i c_i \mathbf{v}_i}{\|\mathbf{v}_i\|_2} + \gamma\eta\mathbf{u} \right)^\top \right] \right\| \\ &\leq \sum_{i=1}^{t} \left\| \mathrm{E}_{i-1}\left[\left(\frac{Z_i c_i \mathbf{v}_i}{\|\mathbf{v}_i\|_2} + \gamma\eta\mathbf{u} \right) \left(\frac{Z_i c_i \mathbf{v}_i}{\|\mathbf{v}_i\|_2} + \gamma\eta\mathbf{u} \right)^\top \right] \right\| \\ &\overset{(4),(13)}{\leq} \sum_{i=1}^{t} \left\| \mathrm{E}_{i-1}\left[\frac{Z_i c_i^2}{\|\mathbf{v}_i\|_2^2}\mathbf{v}_i\mathbf{v}_i^\top \right] \right\| \overset{(2)}{\leq} \eta B^2 t,\end{aligned}$$

$$\begin{aligned}&\left| \sum_{i=1}^{t} \mathrm{E}_{i-1}[\delta_i^\top \delta_i] \right| \\ &= \sum_{i=1}^{t} \mathrm{E}_{i-1}\left[\left(\frac{Z_i c_i}{\|\mathbf{v}_i\|_2}\mathbf{v}_i + \gamma\eta\mathbf{u} \right)^\top \left(\frac{Z_i c_i}{\|\mathbf{v}_i\|_2}\mathbf{v}_i + \gamma\eta\mathbf{u} \right) \right] \\ &\overset{(4),(14)}{\leq} \sum_{i=1}^{t} \mathrm{E}_{i-1}\left[Z_i c_i^2 \right| \overset{(2)}{\leq} \eta B^2 t.\end{aligned}$$

Then, based on Theorem 4, for each $t \in [T]$, we have with a probability at least $1 - \delta$

$$\left\| \sum_{i=1}^{t} \delta_i \right\|_2 \leq \frac{4B}{3} \log \frac{d+1}{\delta} + B\sqrt{2\eta t \log \frac{d+1}{\delta}}.$$

We complete the proof by taking the union bound over $t = 1, \ldots, T$.

Proof of Lemma 2

When $t = 1$, it is clear that

$$f(\mathbf{x}_1^\top \mathbf{u}) - \min_{\|\mathbf{x}\| \leq 1} f(\mathbf{x}^\top \mathbf{u}) \overset{(2)}{\leq} 2B.$$

In the following, we discuss the case when $t \geq 2$. From our assumption that $f(\cdot)$ is a non-increasing function, we have

$$f(\mathbf{x}_t^\top \mathbf{u}) - \min_{\|\mathbf{x}\| \leq 1} f(\mathbf{x}^\top \mathbf{u}) = f(\mathbf{x}_t^\top \mathbf{u}) - f(\mathbf{u}^\top \mathbf{u}).$$

Since $f(\cdot)$ is L-Lipschitz continuous, we further have

$$\begin{aligned}f(\mathbf{x}_t^\top \mathbf{u}) - f(\mathbf{u}^\top \mathbf{u}) &\leq L|\mathbf{x}_t^\top \mathbf{u} - \mathbf{u}^\top \mathbf{u}| \\ &\leq L\|\mathbf{x}_t - \mathbf{u}\|_2 = L\left\| \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|} - \mathbf{u} \right\|_2. \end{aligned} \tag{15}$$

According to the procedure in Algorithm 1, we have

$$\mathbf{w}_t = \sum_{i=1}^{t-1} -\frac{Z_i c_i}{\|\mathbf{v}_i\|_2}\mathbf{v}_i \overset{(10)}{=} \gamma\eta(t-1)\mathbf{u} + \sum_{i=1}^{t-1} \delta_i.$$

Then, we have

$$\begin{aligned}&\|\mathbf{w}_t - \gamma\eta(t-1)\mathbf{u}\|_2 \leq \left\| \sum_{i=1}^{t-1} \delta_i \right\|_2 \\ \Leftrightarrow &\left\| \frac{\mathbf{w}_t}{\gamma\eta(t-1)} - \mathbf{u} \right\|_2 \leq \frac{1}{\gamma\eta(t-1)} \left\| \sum_{i=1}^{t-1} \delta_i \right\|_2.\end{aligned}$$

Following a simple geometric argument, we have

$$\begin{aligned}&\left\| \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|} - \mathbf{u} \right\|_2 \leq \frac{2}{\gamma\eta(t-1)} \left\| \sum_{i=1}^{t-1} \delta_i \right\|_2 \\ \overset{(15)}{\Rightarrow} &f(\mathbf{x}_t^\top \mathbf{u}) - \min_{\|\mathbf{x}\| \leq 1} f(\mathbf{x}^\top \mathbf{u}) \leq \frac{2L}{\gamma\eta(t-1)} \left\| \sum_{i=1}^{t-1} \delta_i \right\|_2.\end{aligned}$$

Proof of Theorem 2

In this case, we define the vector-valued martingale-difference sequence as

$$\delta_i = -\frac{Z_i c_i}{\|\mathbf{v}_i\|_2}\mathbf{v}_i - \gamma_i\eta\mathbf{u},\ i = 1, \ldots, T.$$

It is easy to verify that Lemma 1 still holds and Lemma 2 become the following one.

Lemma 3. *We have*

$$f_t(\mathbf{x}_t^\top \mathbf{u}) - \min_{\|\mathbf{x}\| \leq 1} f_t(\mathbf{x}^\top \mathbf{u}) \leq \begin{cases} 2B, & \text{if } t = 1; \\ \frac{2L}{\gamma\eta(t-1)} \left\| \sum_{i=1}^{t-1} \delta_i \right\|_2, & \text{otherwise.} \end{cases}$$

The rest proof is the same as that for Theorem 1.

Conclusion and Future Work

In this paper, we study the problem of online bandit learning with non-convex losses, and assume the loss function is a composition of a non-increasing scalar function and a linear function. Following the idea of exploration and exploitation, we develop an efficient algorithm which achieves $\widetilde{O}(poly(d)T^{2/3})$ regret bound under appropriate conditions.

One limitation of the current work is that the regret bound only holds against an oblivious adversary. In the future, we will investigate how to extend our results to the adaptive adversary. There are also many open problems for bandit learning with non-convex losses, such as under what condition there exists a Hannan-consistent algorithm and what is the lower bound. We will leave these questions for future investigations.

Acknowledgments

This research was supported by NSFC (61333014, 61321491), the Collaborative Innovation Center of Novel Software Technology and Industrialization, NSF (IIS-1251031), ARO (W911NF-11-1-0383), and Baidu Fund (181315PO5760).

References

Abernethy, J.; Agarwal, A.; Bartlett, P. L.; and Rakhlin, A. 2009. A stochastic view of optimal regret through minimax duality. In *Proceedings of the 22nd Annual Conference on Learning Theory*.

Abernethy, J.; Hazan, E.; and Rakhlin, A. 2008. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning*, 263–274.

Agarwal, A.; Dekel, O.; and Xiao, L. 2010. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the 23rd Annual Conference on Learning*, 28–40.

Angluin, D., and Valiant, L. 1979. Fast probabilistic algorithms for hamiltonian circuits and matchings. *Journal of Computer and System Sciences* 18(2):155–193.

Audibert, J.-Y., and Bubeck, S. 2009. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory*.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2003. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing* 32(1):48–77.

Awerbuch, B., and Kleinberg, R. D. 2004. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, 45–53.

Bubeck, S.; Cesa-Bianchi, N.; and Kakade, S. M. 2012. Towards minimax policies for online linear optimization with bandit feedback. In *Proceedings of the 25th Annual Conference on Learning Theory*.

Calauzènes, C.; Usunier, N.; and Gallinari, P. 2012. On the (non-)existence of convex, calibrated surrogate losses for ranking. In *Advances in Neural Information Processing Systems 25*, 197–205.

Cesa-Bianchi, N., and Lugosi, G. 2006. *Prediction, Learning, and Games*. Cambridge University Press.

Cho, E. 2009. Inner product of random vectors. *International Journal of Pure and Applied Mathematics* 56(2):217–221.

Dani, V., and Hayes, T. P. 2006. Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithm*, 937–943.

Dani, V.; Kakade, S. M.; and Hayes, T. P. 2008. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems 20*, 345–352.

Ertekin, S.; Bottou, L.; and Giles, C. L. 2011. Nonconvex online support vector machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(2):368–381.

Flaxman, A. D.; Kalai, A. T.; and McMahan, H. B. 2005. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, 385–394.

Gao, W., and Zhou, Z.-H. 2011. On the consistency of multi-label learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, 341–358.

Gasso, G.; Pappaioannou, A.; Spivak, M.; and Bottou, L. 2011. Batch and online learning algorithms for nonconvex neyman-pearson classification. *ACM Transactions on Intelligent Systems and Technology* 2(3):28:1–28:19.

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

Hazan, E., and Kale, S. 2010. Extracting certainty from uncertainty: regret bounded by variation in costs. *Machine Learning* 80(2-3):165–188.

Hazan, E., and Kale, S. 2012. Online submodular minimization. In *Journal of Machine Learning Research*, volume 13, 2903–2922.

Hinton, G. E.; Osindero, S.; and Teh, Y.-W. 2006. A fast learning algorithm for deep belief nets. *Neural Computation* 18(7):1527–1554.

Kivinen, J.; Smola, A. J.; and Williamson, R. C. 2002. Online learning with kernels. In *Advances in Neural Information Processing Systems 14*, 785–792.

Magureanu, S.; Combes, R.; and Proutiere, A. 2014. Lipschitz bandits: Regret lower bounds and optimal algorithms. In *Proceedings of the 27th Conference on Learning Theory*, 975–999.

McMahan, H. B., and Blum, A. 2004. Online geometric optimization in the bandit setting against an adaptive adversary. In *Proceedings of the 17th Annual Conference on Learning Theory*, 109–123.

Plan, Y., and Vershynin, R. 2013. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory* 59(1):482–494.

Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 58(5):527–535.

Saha, A., and Tewari, A. 2011. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 636–642.

Tropp, J. A. 2011. Freedman's inequality for matrix martingales. *Electronic Communications in Probability* 16:262–270.

Zhang, L.; Yi, J.; Jin, R.; Lin, M.; and He, X. 2013. Online kernel learning with a near optimal sparsity bound. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*.

Zhang, L.; Yi, J.; and Jin, R. 2014. Efficient algorithms for robust one-bit compressive sensing. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*.

Zinkevich, M. 2003. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, 928–936.