Online Stochastic Linear Optimization under One-bit Feedback

Lijun Zhang ZHANGLJ@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

Tianbao Yang TIANBAO-YANG@UIOWA.EDU
Department of Computer Science, The University of Iowa, Iowa City, IA 52242, USA

Rong Jin JINRONG.JR@ALIBABA-INC.COM
Alibaba Group, Seattle, USA

Yichi Xiao XIAOYC@LAMDA.NJU.EDU.CN
Zhi-Hua Zhou ZHOUZH@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

Abstract

In this paper, we study a special bandit setting of online stochastic linear optimization, where only one-bit of information is revealed to the learner at each round. This problem has found many applications including online advertisement and online recommendation. We assume the binary feedback is a random variable generated from the *logit* model, and aim to minimize the regret defined by the unknown linear function. Although the existing method for generalized linear bandit can be applied to our problem, the high computational cost makes it impractical for real-world applications. To address this challenge, we develop an efficient online learning algorithm by exploiting particular structures of the observation model. Specifically, we adopt online Newton step to estimate the unknown parameter and derive a tight confidence region based on the exponential concavity of the logistic loss. Our analysis shows that the proposed algorithm achieves a regret bound of $\widetilde{O}(d\sqrt{T})$, which matches the optimal result of stochastic linear bandits.

1. Introduction

Online learning with bandit feedback plays an important role in several industrial domains, such as ad placement, website optimization, and packet routing (Bubeck & Cesa-Bianchi, 2012). A canonical framework for studying this problem is the multi-armed bandits (MAB), which models the situation that a gambler must choose which of K slot machines to play (Robbins, 1952). In the basic stochastic MAB, each arm is assumed to deliver rewards that are drawn from a fixed but unknown distribution. The goal of the gambler is to minimize the regret, namely the difference between his expected cumulative reward and that of the best single arm in hindsight (Auer et al., 2002).

Although MAB is a powerful framework for modeling online decision problems, it becomes intractable when the number of arms is very large or even infinite. To address this challenge, various algorithms have been designed to exploit different structure properties of the reward function, such as Lipschitz (Kleinberg et al., 2008) and convex (Flaxman et al., 2005; Agarwal et al., 2013). Among them, stochastic linear bandits (SLB) has received considerable attentions during the past decade (Auer, 2002; Dani et al., 2008; Abbasi-yadkori et al., 2011). In each round of SLB, the learner is asked to choose an action $\mathbf{x}_t$ from a decision set $\mathcal{D} \subseteq \mathbb{R}^d$, then he observes y_t such that

$$\mathrm{E}[y_t|\mathbf{x}_t] = \mathbf{x}_t^\top \mathbf{w}_*, \tag{1}$$

where $\mathbf{w}_* \in \mathbb{R}^d$ is a vector of unknown parameters. The goal of the learner is to minimize the (pseudo) regret

$$T \max_{\mathbf{x}\in\mathcal{D}} \mathbf{x}^\top \mathbf{w}_* - \sum_{t=1}^{T} \mathbf{x}_t^\top \mathbf{w}_*. \tag{2}$$

In this paper, we consider a special bandit setting of online linear optimization where the feedback y_t only contains one-bit of information. In particular, $y_t \in \{\pm 1\}$. Our setting is motivated from the fact that in many real-world applications, such as online advertising and recommender

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

systems, user feedback (e.g., click or not, like or dislike) is usually binary. Since the feedback is binary-valued, we assume it is generated according to the logit model (Hastie et al., 2009), i.e.,

$$\Pr[y_t = \pm 1|\mathbf{x}_t] = \frac{1}{1+\exp(-y_t \mathbf{x}_t^\top \mathbf{w}_*)}. \tag{3}$$

Without loss of generality, suppose 1 is the preferred outcome. Then, it is natural to define the regret in terms of the expected times that 1 is observed, i.e.,

$$T \max_{\mathbf{x}\in\mathcal{D}} \frac{\exp(\mathbf{x}^\top \mathbf{w}_*)}{1+\exp(\mathbf{x}^\top \mathbf{w}_*)} - \sum_{t=1}^{T} \frac{\exp(\mathbf{x}_t^\top \mathbf{w}_*)}{1+\exp(\mathbf{x}_t^\top \mathbf{w}_*)}. \tag{4}$$

The observation model in (3) and the nonlinear regret in (4) can be treated as a special case of the Generalized Linear Bandit (GLB) (Filippi et al., 2010). However, the existing algorithm for GLB is inefficient in the sense that: i) it is not a truly online algorithm since the whole learning history is stored in memory and used to estimate $\mathbf{w}_*$; and ii) it is limited to the case that the number of arms is finite because an upper bound for each arm needs to be calculated explicitly in each round.

The main contribution of this paper is an efficient online learning algorithm that effectively exploits particular structures of the logit model. Based on the analytical properties of the logistic function, we first show that the linear regret defined in (2) and the nonlinear regret in (4) only differs by a constant factor, and then focus on minimizing the former one due to its simplicity. Similar to previous studies (Bubeck & Cesa-Bianchi, 2012), we follow the principle of "optimism in face of uncertainty" to deal with the exploration-exploitation dilemma. The basic idea is to maintain a confidence region for $\mathbf{w}_*$, and choose an estimate from the confidence region and an action so that the linear reward is maximized. Thus, the problem reduces to the construction of the confidence region from one-bit feedback that satisfies (3). Based on the exponential concavity of the logistic loss, we propose to use a variant of the online Newton step (Hazan et al., 2007) to find the center of the confidence region and derive its width by a rather technical analysis of the updating rule. Theoretical analysis shows that our algorithm achieves a regret bound of $\widetilde{O}(d\sqrt{T})$,[1] which matches the result for SLB (Dani et al., 2008). Furthermore, we provide several strategies to reduce the computational cost of the proposed algorithm.

2. Related Work

The stochastic multi-armed bandits (MAB) (Robbins, 1952), has become the canonical formalism for studying the problem of decision-making under uncertainty. A long line of successive problems have been extensively studied in statistics (Berry & Fristedt, 1985) and computer science (Bubeck & Cesa-Bianchi, 2012).

2.1. Stochastic Multi-armed Bandits (MAB)

In their seminal paper, Lai & Robbins (1985) establish an asymptotic lower bound of $O(K \log T)$ for the expected cumulative regret over T periods, under the assumption that the expected rewards of the best and second best arms are well-separated. By making use of *upper confidence bounds* (UCB), they further construct policies which achieve the lower bound asymptotically. However, this initial algorithm is quite involved, because the computation of UCB relies on the entire sequence of rewards obtained so far. To address this limitation, Agrawal (1995) introduces a family of simpler policies that only needs to calculate the sample mean of rewards, and the regret retains the optimal logarithmic behavior. A finite time analysis of stochastic MAB is conducted by Auer et al. (2002). In particular, they propose an UCB-type algorithm based on the Chernoff-Hoeffding bound, and demonstrate it achieves the optimal logarithmic regret uniformly over time.

2.2. Stochastic Linear Bandits (SLB)

SLB is first studied by Auer (2002), who considers the case $\mathcal{D}$ is finite. Although an elegant UCB-type algorithm named LinRel is developed, he fails to bound its regret due to independence issues. Instead, he designs a complicated master algorithm which uses LinRel as a subroutine, and achieves a regret bound of $\widetilde{O}((\log|\mathcal{D}|)^{3/2}\sqrt{Td})$, where $|\mathcal{D}|$ is the number of feasible decisions. In a subsequent work, Dani et al. (2008) generalize LinRel slightly so that it can be applied in settings where $\mathcal{D}$ may be infinite. They refer to the new algorithm as ConfidenceBall$_2$, and show it enjoys a bound of $\widetilde{O}(d\sqrt{T})$, which does not depend on the cardinality of $\mathcal{D}$. Later, Abbasi-yadkori et al. (2011) improve the theoretical analysis of ConfidenceBall$_2$ by employing tools from the self-normalized processes. Specifically, the worst case bound is improved by a logarithmic factor and the constant is also improved.

2.3. Generalized Linear Bandit (GLB)

Filippi et al. (2010) extend SLB to the nonlinear case based on the Generalized Linear Model framework of statistics. In the so-called GLB model, y_t is assumed to satisfy $\mathrm{E}[y_t|\mathbf{x}_t] = \mu(\mathbf{x}_t^\top \mathbf{w}_*)$ where $\mu : \mathbb{R} \mapsto \mathbb{R}$ is certain link function. The regret is also defined in terms of $\mu(\cdot)$ and given by

$$T \max_{\mathbf{x}\in\mathcal{D}} \mu(\mathbf{x}^\top \mathbf{w}_*) - \sum_{t=1}^{T} \mu(\mathbf{x}_t^\top \mathbf{w}_*). \tag{5}$$

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

Note that by setting $\mu(x) = \exp(x)/[1+\exp(x)]$, the problem considered in this paper becomes a special case of GLB. An UCB-type algorithm has been proposed for GLB and achieves a regret bound of $\widetilde{O}(d\sqrt{T})$. Different from ConfidenceBall$_2$ which constructs a confidence region in the parameter space, the algorithm of Filippi et al. (2010) operates only in the reward space. However, the space and time complexities of that algorithm in the t-th iteration are $O(t)$ and $O(t+|\mathcal{D}|)$, respectively. The $O(t)$ factor comes from the fact it needs to store the past action-feedback pairs $(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_{t-1}, y_{t-1})$ and use all of them to estimate $\mathbf{w}_*$. The $O(|\mathcal{D}|)$ factor is due to the fact it needs to calculate an upper bound for each arm in order to decide the next action $\mathbf{x}_t$.

Although we can reduce the computational cost of GLB by replacing the batch optimization with online algorithms, such as online gradient descent (Zinkevich, 2003), the theoretical guarantee of the new algorithm needs to be developed. Furthermore, it is also unclear how to extend GLB to the case of infinite number of arms.

2.4. Bandit Learning with One-bit Feedback

There are several new variants of bandit learning that also rely on one one-bit feedback, such as multi-class bandits (Kakade et al., 2008; Chen et al., 2014) and K-armed dueling bandits (Yue et al., 2009; Ailon et al., 2014). For example, in multi-class bandits, the feedback is whether the predicted label is correct or not, and in K-armed dueling bandits, the feedback is the comparison between the rewards from two arms. However, none of them are designed for online linear optimization.

2.5. One-bit Compressive Sensing (CS)

Finally, we would like to discuss one closely related work in signal processing—one-bit Compressive Sensing (CS) (Boufounos & Baraniuk, 2008; Plan & Vershynin, 2013; Zhang et al., 2014). One-bit CS aims to recover a sparse vectors $\mathbf{w}_*$ from a set of one-bit measurements $\{y_i\}$ where y_i is generated from $\mathbf{x}_i^\top \mathbf{w}_*$ according to certain observation model such as (3). The main difference is that one-bit CS is studied in batch setting with the goal to minimize the recovery error, while our problem is studied in online setting with the goal to minimize the regret. Another difference is that $\mathbf{w}_*$ needs to be sparse in one-bit CS, but could be dense in our case.

3. Online Learning for Logit Model (OL^{2}M)

We first describe the proposed algorithm for online stochastic linear optimization given one-bit feedback, next compare it with existing methods, then state its theoretical guarantees, and finally discuss implementation issues.

3.1. The Algorithm

For a positive definite matrix $A \in \mathbb{R}^{d\times d}$, the weighted ℓ_2-norm is defined by $\|\mathbf{x}\|_A^2 = \mathbf{x}^\top A \mathbf{x}$. Without loss of generality, we assume the decision space $\mathcal{D}$ is contained in the unit ball, that is,

$$\|\mathbf{x}\|_2 \leq 1, \ \forall \mathbf{x} \in \mathcal{D}. \tag{6}$$

We further assume the ℓ_2-norm of $\mathbf{w}_*$ is upper bounded by some constant R, which is known to the learner. Our first observation is that the linear regret in (2) and the nonlinear regret in (4) only differs by a constant factor as indicated below.

Lemma 1. *Let R_L and R_N be the linear and nonlinear regrets in (2) and (4), respectively. We have*

$$\frac{1}{2(1+\exp(R))} R_L \leq R_N \leq \frac{1}{4} R_L \tag{7}$$

In the following, we will develop an efficient algorithm that minimizes the linear regret, which in turn minimizes the nonlinear regret as well.

The algorithm is motivated as follows. Suppose actions $\mathbf{x}_1, \ldots, \mathbf{x}_t$ have been submitted to the oracle, and let $y_1, \ldots, y_t$ be the one-bit feedback from the oracle. To approximate $\mathbf{w}_*$, the most straightforward way is to find the maximum likelihood estimator by solving the following logistic regression problem

$$\min_{\|\mathbf{w}\|_2 \leq R} \frac{1}{t} \sum_{i=1}^{t} \log\left(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})\right).$$

However, this approach does not scale well since it requires the leaner to store the entire learning history. Instead, we propose an online algorithm to find an approximate solution. The key observation is that the logistic loss

$$f_t(\mathbf{w}) = \log\left(1 + \exp(-y_t \mathbf{x}_t^\top \mathbf{w})\right)$$

is exponentially concave over bounded domain (Hazan et al., 2014), which motivates us to apply a variant of the online Newton step (Hazan et al., 2007). Specifically, we propose to find an approximate solution $\mathbf{w}_{t+1}$ by solving the following problem

$$\min_{\|\mathbf{w}\|_2 \leq R} \frac{\|\mathbf{w} - \mathbf{w}_t\|_{Z_{t+1}}^2}{2} + (\mathbf{w} - \mathbf{w}_t)^\top \nabla f_t(\mathbf{w}_t) \tag{8}$$

where

$$Z_{t+1} = Z_t + \frac{\beta}{2} \mathbf{x}_t \mathbf{x}_t^\top, \tag{9}$$

and β is defined in (14). Although our updating rule is similar to the method in (Hazan et al., 2007), there also exist some differences. As indicated by (9), in our case

Algorithm 1 Online Learning for Logit Model (OL^{2}M)

```
Input: Regularization Parameter $\lambda$
$Z_1 = \lambda I$, $\mathbf{w}_1 = 0$
for $t = 1, 2, \ldots$ do
    $$(\mathbf{x}_t, \widehat{\mathbf{w}}_t) = \operatorname*{argmax}_{\mathbf{x}\in\mathcal{D}, \mathbf{w}\in\mathcal{C}_t} \mathbf{x}^\top \mathbf{w}$$
    Submit $\mathbf{x}_t$ and observe $y_t \in \{\pm 1\}$
    Solve the optimization problem in (8) to find $\mathbf{w}_{t+1}$
end for
```

$\mathbf{x}_t\mathbf{x}_t^\top$ is used to approximate the Hessian matrix, while in Hazan et al. (2007) $\nabla f_t(\mathbf{w}_t)[\nabla f_t(\mathbf{w}_t)^\top]$ is used.

After a theoretical analysis, we are able to show that with a high probability

$$\mathbf{w}_* \in \mathcal{C}_{t+1} = \left\{\mathbf{w} : \|\mathbf{w} - \mathbf{w}_{t+1}\|_{Z_{t+1}} \leq \sqrt{\gamma_{t+1}}\right\} \quad (10)$$

where the value of γ_{t+1} is given in (12). Given the confidence region, we adopt the principle of "optimism in face of uncertainty", and the next action $\mathbf{x}_{t+1}$ is given by

$$(\mathbf{x}_{t+1}, \widehat{\mathbf{w}}_{t+1}) = \operatorname*{argmax}_{\mathbf{x}\in\mathcal{D}, \mathbf{w}\in\mathcal{C}_{t+1}} \mathbf{x}^\top \mathbf{w}. \quad (11)$$

At the beginning, we set

$$Z_1 = \lambda I, \text{ and } \mathbf{w}_1 = 0.$$

The above procedure is summarized in Algorithm 1, and is refer to as Online Learning for Logit Model (OL^{2}M).

Since both ConfidenceBall$_2$ (Dani et al., 2008) and our OL^{2}M are UCB-type algorithms, their overall frameworks are similar. The main difference lies in the construction of the confidence region and the related analysis. While ConfidenceBall$_2$ uses online least square to update the center of the confidence region, OL^{2}M resorts to online Newton step. Due to the difference in the updating rule and the observation model, the self-normalized bound for vector-valued martingales (Abbasi-yadkori et al., 2011) can not be applied here.

Although our observation model in (3) can be handled by the Generalized Linear Bandit (GLB) (Filippi et al., 2010), this paper differs from GLB in the following aspects.

- To estimate $\mathbf{w}_*$, GLB needs to store the learning history and perform batch updating in each round. In contrast, the proposed OL^{2}M performs online updating.
- While GLB only considers a finite number of arms, we allow the number of arms to be infinite.
- Our algorithm follows the learning framework of SLB. Thus, existing techniques for speeding up SLB can also be used to accelerate our algorithm, which is discussed in Section 3.3.

3.2. Theoretical Guarantees

The main theoretical contribution of this paper is the following theorem regarding the confidence region of $\mathbf{w}_*$ at each round.

Theorem 1. *With a probability at least $1-\delta$, we have*

$$\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_{Z_{t+1}} \leq \sqrt{\gamma_{t+1}}, \ \forall t > 0$$

where

$$\gamma_{t+1} = \left[8R + \left(\frac{8}{\beta} + \frac{16}{3}R\right)\tau_t + \frac{2}{\beta}\log\frac{\det(Z_{t+1})}{\det(Z_1)}\right] + \lambda R^2, \quad (12)$$

$$\tau_t = \log\left(\frac{2\lceil 2\log_2 t\rceil t^2}{\delta}\right), \quad (13)$$

$$\beta = \frac{1}{2(1+\exp(R))}. \quad (14)$$

The main idea is to analyze the growth of $\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_{Z_{t+1}}^2$ by exploring the properties of the logistic loss (Lemmas 2 and 4) and concentration inequalities for martingales (Lemma 5). By a simple upper bound of $\log\det(Z_{t+1})/\det(Z_1)$, we can show that the width of the confidence region is $O(\sqrt{d\log t})$.

Corollary 2. *We have*

$$\log\frac{\det(Z_{t+1})}{\det(Z_1)} \leq d\log\left(1 + \frac{\beta t}{2\lambda d}\right)$$

and thus

$$\gamma_{t+1} \leq O(d\log t), \ \forall t > 0.$$

Based on Theorem 1, we have the following regret bound for OL^{2}M.

Theorem 3. *With a probability at least $1-\delta$, we have*

$$T\max_{\mathbf{x}\in\mathcal{D}}\mathbf{x}^\top\mathbf{w}_* - \sum_{t=1}^{T}\mathbf{x}_t^\top\mathbf{w}_* \leq 4\max\left(1, \sqrt{\frac{\beta}{2}}R\right)\sqrt{\frac{\gamma_T T}{\beta}\log\frac{\det(Z_{T+1})}{\det(Z_1)}}$$

holds for all $T > 0$.

Combining with the upper bound in Corollary 2, the above theorem implies our algorithm achieves a regret bound of $\widetilde{O}(d\sqrt{T})$ which matches the bound for Stochastic Linear Bandits (Dani et al., 2008). One limitation of Theorem 3 is that the upper bound has an exponential dependence on R, which is an upper bound of $\|\mathbf{w}_*\|_2$. That is because our algorithm is built upon online Newton step (Hazan et al., 2007), the regret of which has such a undesirable dependence on R. From the recent studies on logistic regression (Bach & Moulines, 2013; Bach, 2014; Hazan et al., 2014), we conjecture that it is possible to obtain a polynomial dependence on R, but with a higher dependence on T. We will investigate this issue in the future.

3.3. Implementation Issues

The main computational cost of OL^{2}M comes from (11) which is NP-hard in general (Dani et al., 2008). In the following, we discuss two strategies for reducing the computational cost. More results can be found in the supplementary material.

Finite Decision Set If the decision set $\mathcal{D}$ is finite, (11) can be solved by computing an upper bound for each decision in $\mathcal{D}$. Specifically, we have

$$\begin{aligned}\mathbf{x}_{t+1} &= \operatorname*{argmax}_{\mathbf{x}\in\mathcal{D}} \max_{\|\mathbf{w}-\mathbf{w}_{t+1}\|_{Z_{t+1}}\leq\sqrt{\gamma_{t+1}}} \mathbf{x}^\top\mathbf{w}\\ &= \operatorname*{argmax}_{\mathbf{x}\in\mathcal{D}} \max_{\|\mathbf{z}\|_{Z_{t+1}}\leq\sqrt{\gamma_{t+1}}} \mathbf{x}^\top(\mathbf{w}_{t+1}+\mathbf{z})\\ &= \operatorname*{argmax}_{\mathbf{x}\in\mathcal{D}} \left(\mathbf{x}^\top\mathbf{w}_{t+1}+\sqrt{\gamma_{t+1}}\|\mathbf{x}\|_{Z_{t+1}^{-1}}\right).\end{aligned}$$

Optimization Over Ball As mentioned by Dani et al. (2008), in the special case that $\mathcal{D}$ is the unit ball, (11) could be solved in time $O(poly(d))$. Here, we provide an explanation using techniques from convex optimization. To this end, we rewrite the optimization problem in (11) as follows

$$\begin{aligned}&\max_{\|\mathbf{x}\|_2\leq 1, \|\mathbf{w}-\mathbf{w}_{t+1}\|_{Z_{t+1}}\leq\sqrt{\gamma_{t+1}}} \mathbf{x}^\top\mathbf{w}\\ =&\max_{\|\mathbf{w}-\mathbf{w}_{t+1}\|_{Z_{t+1}}\leq\sqrt{\gamma_{t+1}}} \|\mathbf{w}\|_2\end{aligned}$$

which is equivalent to

$$\min_{\|\mathbf{w}-\mathbf{w}_{t+1}\|^2_{Z_{t+1}}\leq\gamma_{t+1}} -\|\mathbf{w}\|_2^2.$$

The above problem is an optimization problem with a quadratic objective and one quadratic inequality constraint, it is well-known that strong duality holds provided there exists a strictly feasible point (Boyd & Vandenberghe, 2004). Thus, we can solve its dual problem which is convex and given by

$$\begin{aligned}\max\quad & \gamma\\ \text{s. t.}\quad & \lambda\geq 0\\ & \begin{bmatrix} -I+\lambda Z_{t+1} & -\lambda Z_{t+1}\mathbf{w}_{t+1}\\ -\lambda\mathbf{w}_{t+1}^\top Z_{t+1} & \lambda(\|\mathbf{w}_{t+1}\|^2_{Z_{t+1}}-\gamma_{t+1})-\gamma\end{bmatrix}\succeq 0\end{aligned}$$

After obtaining the dual solution, we can get the primal solution based on KKT conditions.

4. Analysis

Due to the limitation of space, we only prove Theorem 1. The omitted proofs are provided in the supplementary material.

4.1. Proof of Theorem 1

We begin with several lemmas that are central to our analysis.

Although the application of online Newton step (Hazan et al., 2007) in Algorithm 1 is motivated from the fact that $f_t(\mathbf{w})$ is exponentially concave over bounded domain, our analysis is built upon a related but different property that the logistic loss $\log(1+\exp(x))$ is strongly convex over bounded domain, from which we obtain the following lemma.

Lemma 2. *Denote the ball of radius R by $\mathcal{B}_R$, i.e., $\mathcal{B}_R=\{\mathbf{w}:\|\mathbf{w}\|_2\leq R\}$. The following holds for $\beta\leq\frac{1}{2(1+\exp(R))}$:*

$$\begin{aligned}f_t(\mathbf{w}_2)\geq & f_t(\mathbf{w}_1)+[\nabla f_t(\mathbf{w}_1)]^\top(\mathbf{w}_2-\mathbf{w}_1)\\ &+\frac{\beta}{2}\left((\mathbf{w}_2-\mathbf{w}_1)^\top\mathbf{x}_t\right)^2,\ \forall\mathbf{w}_1,\mathbf{w}_2\in\mathcal{B}_R.\end{aligned}$$

Comparing Lemma 2 with Lemma 3 in (Hazan et al., 2007), we can see that the quadratic term in our inequality does not depends on y_t. This independence allows us to simplify the subsequent analysis involving martingales.

Our second lemma is devoted to analyzing the property of the updating rule in (8).

Lemma 3.

$$\begin{aligned}&\langle\mathbf{w}_t-\mathbf{w}_*,\nabla f_t(\mathbf{w}_t)\rangle-\frac{1}{2}\|\nabla f_t(\mathbf{w}_t)\|^2_{Z_{t+1}^{-1}}\\ \leq&\frac{\|\mathbf{w}_t-\mathbf{w}_*\|^2_{Z_{t+1}}}{2}-\frac{\|\mathbf{w}_{t+1}-\mathbf{w}_*\|^2_{Z_{t+1}}}{2}.\end{aligned}\tag{15}$$

For each function $f_t(\cdot)$, we denote its conditional expectation over y_t by $\bar{f}_t(\mathbf{w})$, i.e.,

$$\bar{f}_t(\mathbf{w})=\mathrm{E}_{y_t}\left[\log\left(1+\exp\left(-y_t\mathbf{x}_t^\top\mathbf{w}\right)\right)\right].\tag{16}$$

According to the Leibniz integral rule, we have

$$\nabla\bar{f}_t(\mathbf{w})=\mathrm{E}_{y_t}\left[\nabla f_t(\mathbf{w})\right].\tag{17}$$

Based the property of Kullback–Leibler divergence (Cover & Thomas, 2006), we obtain the following lemma.

Lemma 4. *We have*

$$\bar{f}_t(\mathbf{w})\geq\bar{f}_t(\mathbf{w}_*),\ \forall\mathbf{w}\in\mathbb{R}^d.$$

Next, we introduce one inequality for bounding the weighted ℓ_2-norm of the gradient

$$\begin{aligned}\|\nabla f_t(\mathbf{w})\|_A^2 =& \left(\frac{\exp(-y_t\mathbf{x}_t^\top\mathbf{w})}{1+\exp(-y_t\mathbf{x}_t^\top\mathbf{w})}\right)^2 \mathbf{x}_t^\top A\mathbf{x}_t \\ \leq& \|\mathbf{x}_t\|_A^2,\ \forall A\succeq 0,\ \mathbf{w}\in\mathbb{R}^d.\end{aligned} \tag{18}$$

We continue the proof of Theorem 1 in the following. Our updating rule in (8) ensures $\|\mathbf{w}_t\|_2 \leq R$, $\forall t>0$. Combining with the assumption $\|\mathbf{w}_*\|_2 \leq R$, Lemma 2 implies

$$\begin{aligned}f_t(\mathbf{w}_t) \leq& f_t(\mathbf{w}_*) + [\nabla f_t(\mathbf{w}_t)]^\top(\mathbf{w}_t-\mathbf{w}_*) \\ &- \frac{\beta}{2}\left((\mathbf{w}_*-\mathbf{w}_t)^\top\mathbf{x}_t\right)^2.\end{aligned} \tag{19}$$

By taking expectation over y_t, (19) becomes

$$\begin{aligned}\bar{f}_t(\mathbf{w}_t) \overset{(16),(17)}{\leq}& \bar{f}_t(\mathbf{w}_*) + [\nabla \bar{f}_t(\mathbf{w}_t)]^\top(\mathbf{w}_t-\mathbf{w}_*) \\ &- \frac{\beta}{2}\left[\left((\mathbf{w}_*-\mathbf{w}_t)^\top\mathbf{x}_t\right)^2\right].\end{aligned}$$

Combining with Lemma 4, we have

$$\begin{aligned}0 \leq& [\nabla \bar{f}_t(\mathbf{w}_t)]^\top(\mathbf{w}_t-\mathbf{w}_*) - \frac{\beta}{2}\underbrace{\left((\mathbf{w}_*-\mathbf{w}_t)^\top\mathbf{x}_t\right)^2}_{:=a_t} \\ =& [\nabla f_t(\mathbf{w}_t)]^\top(\mathbf{w}_t-\mathbf{w}_*) - \frac{\beta}{2}a_t \\ &+ \underbrace{[\nabla \bar{f}_t(\mathbf{w}_t) - \nabla f_t(\mathbf{w}_t)]^\top(\mathbf{w}_t-\mathbf{w}_*)}_{:=b_t} \\ =& [\nabla f_t(\mathbf{w}_t)]^\top(\mathbf{w}_t-\mathbf{w}_*) - \frac{\|\mathbf{w}_t-\mathbf{w}_*\|_{Z_{t+1}}^2}{2} \\ &+ \frac{\|\mathbf{w}_t-\mathbf{w}_*\|_{Z_{t+1}}^2}{2} - \frac{\beta}{2}a_t + b_t \\ \overset{(15)}{\leq}& - \frac{\|\mathbf{w}_{t+1}-\mathbf{w}_*\|_{Z_{t+1}}^2}{2} + \frac{1}{2}\|\nabla f_t(\mathbf{w}_t)\|_{Z_{t+1}^{-1}}^2 \\ &+ \frac{\|\mathbf{w}_t-\mathbf{w}_*\|_{Z_{t+1}}^2}{2} - \frac{\beta}{2}a_t + b_t \\ \overset{(18)}{\leq}& - \frac{\|\mathbf{w}_{t+1}-\mathbf{w}_*\|_{Z_{t+1}}^2}{2} + \frac{1}{2}\underbrace{\|\mathbf{x}_t\|_{Z_{t+1}^{-1}}^2}_{:=c_t} \\ &+ \frac{\|\mathbf{w}_t-\mathbf{w}_*\|_{Z_{t+1}}^2}{2} - \frac{\beta}{2}a_t + b_t \\ \overset{(9)}{=}& - \frac{\|\mathbf{w}_{t+1}-\mathbf{w}_*\|_{Z_{t+1}}^2}{2} - \frac{\beta}{2}a_t + b_t + \frac{1}{2}c_t \\ &+ \frac{\|\mathbf{w}_t-\mathbf{w}_*\|_{Z_t}^2}{2} + \frac{\beta}{4}\left(\mathbf{x}_t^\top(\mathbf{w}_t-\mathbf{w}_*)\right)^2 \\ =& - \frac{\|\mathbf{w}_{t+1}-\mathbf{w}_*\|_{Z_{t+1}}^2}{2} - \frac{\beta}{4}a_t + b_t + \frac{1}{2}c_t \\ &+ \frac{\|\mathbf{w}_t-\mathbf{w}_*\|_{Z_t}^2}{2}.\end{aligned}$$

We thus have

$$\|\mathbf{w}_{t+1}-\mathbf{w}_*\|_{Z_{t+1}}^2 \leq \|\mathbf{w}_t-\mathbf{w}_*\|_{Z_t}^2 - \frac{\beta}{2}a_t + 2b_t + c_t$$

Summing the above inequality over iterations 1 to t, we obtain

$$\begin{aligned}&\|\mathbf{w}_{t+1}-\mathbf{w}_*\|_{Z_{t+1}}^2 + \frac{\beta}{2}\sum_{i=1}^t a_i \\ \leq& \lambda R^2 + 2\sum_{i=1}^t b_i + \sum_{i=1}^t c_i.\end{aligned} \tag{20}$$

Next, we discuss how to bound the summation of martingale difference sequence $\sum_{i=1}^t b_i$. To this end, we prove the following lemma, which is built up the Bernstein's inequality for martingales (Cesa-Bianchi & Lugosi, 2006) and the peeling technique (Bartlett et al., 2005).

Lemma 5. *With a probability at least $1-\delta$, we have*

$$\sum_{i=1}^t b_i \leq 4R + 2\sqrt{\tau_t\sum_{i=1}^t a_i} + \frac{8}{3}R\tau_t,\ \forall t>0$$

where τ_t is defined in (13).

From Lemma 5 and the basic inequality

$$2\sqrt{\tau_t\sum_{i=1}^t a_i} \leq \frac{\beta}{4}\sum_{i=1}^t a_i + \frac{4}{\beta}\tau_t,$$

with a probability at least $1-\delta$, we have

$$\sum_{i=1}^t b_i \leq 4R + \frac{\beta}{4}\sum_{i=1}^t a_i + \left(\frac{4}{\beta}+\frac{8}{3}R\right)\tau_t \tag{21}$$

holds for all $t>0$. Substituting (21) into (20), we obtain

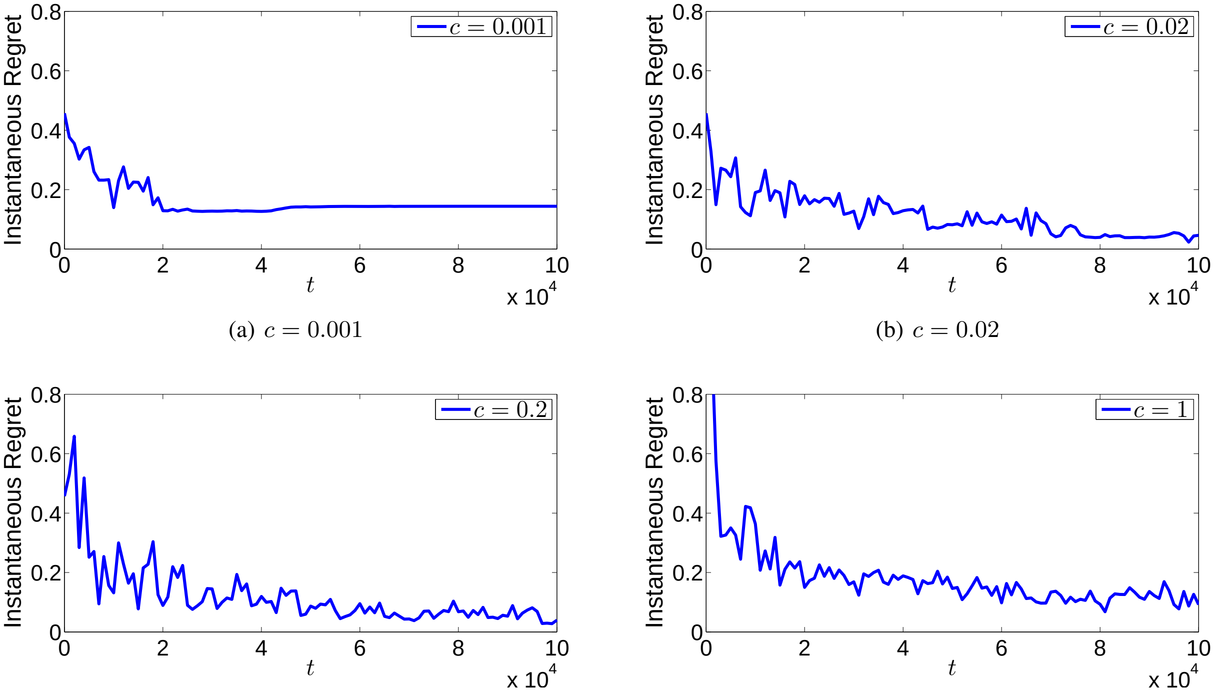
$$\begin{aligned}&\|\mathbf{w}_{t+1}-\mathbf{w}_*\|_{Z_{t+1}}^2 \\ \leq& \lambda R^2 + 2\left[4R + \left(\frac{4}{\beta}+\frac{8}{3}R\right)\tau_t\right] + \sum_{i=1}^t c_i.\end{aligned} \tag{22}$$

Finally, we show an upper bound for $\sum_{i=1}^t c_i$, which is a direct consequence of Lemma 12 in Hazan et al. (2007).

Lemma 6. *We have*

$$\sum_{i=1}^t \|\mathbf{x}_i\|_{Z_{i+1}^{-1}}^2 \leq \frac{2}{\beta}\log\frac{\det(Z_{t+1})}{\det(Z_1)}.$$

We complete the proof by combining (22) with the above lemma.



(a) $c = 0.001$

(b) $c = 0.02$

(c) $c = 0.2$

(d) $c = 1$

Figure 1. Instantaneous regret of OL^{2}M when $\mathcal{D}$ is the unit ball in $\mathbb{R}^{10}$.

5. Experiments

In this section, we present experimental results to demonstrate the effectiveness of the proposed algorithm.

5.1. Experimental Setting

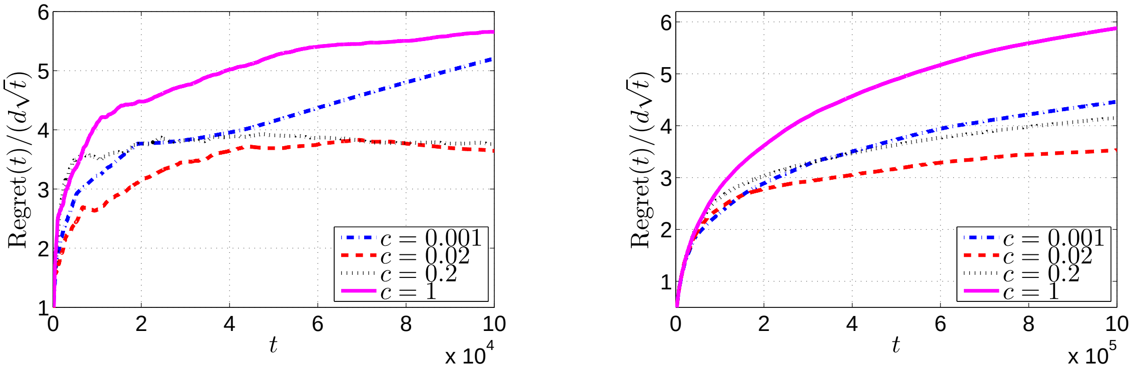
We sample a point uniformly at random from the $(d-1)$-sphere as $\mathbf{w}_*$, and each time the learner submits an action $\mathbf{x}_t$, a one-bit feedback $y_t \in \{\pm 1\}$ is generated according to the logit model in (3). To apply our algorithm, we need to determine the values of two parameters: λ and γ_t. λ is introduced to make Z_t invertible, and the performance of our algorithm is insensitive to its value. Thus, we simply choose $\lambda = 1$ in the following. γ_t is an essential parameter which is the width of the confidence region, and its value is tuned as $c \log \frac{\det(Z_t)}{\det(Z_1)}$ according to (12), where c is searched in the range of $[1e{-}3, 1]$.

5.2. Experimental Results

In the first experiment, we choose the unit ball as the decision set, i.e., $\mathcal{D} = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1\} \subseteq \mathbb{R}^d$, which contains infinite number of actions. As discussed in Section 3.3, in this case, (11) can be cast as a convex optimization problem, which is then solved by the CVX package (Grant & Boyd, 2008; 2014). We first investigate how the instantaneous regret $\mathbf{x}_*^\top \mathbf{w}_* - \mathbf{x}_t^\top \mathbf{w}_*$ varies with t during the learning process. The results for $d = 10$ with different settings of c are shown in Fig. 1. As can be seen, the instantaneous regret decreases overall, although exhibits some local fluctuations. These fluctuations actually reflect the switches between exploitation and exploration. Generally speaking, valley and peak of the curve correspond to exploitation and exploration, respectively.

The value of c determines the width of the confidence region, which in turn controls the exploitation-exploration trade-off. A small value of c prefers exploitation, which may select an action which is not optimal because of too little exploration. For example, in Fig. 1(a) where $c = 0.001$, after 2×10^4 rounds, the learner always submits a suboptimal action and suffers a constant instantaneous regret. On the other hand, a larger value of c favors exploration, which might results in a large regret because too much exploration prevents the algorithm from playing the optimal action. This phenomenon can also be observed in Fig. 1(d) where $c = 1$. From Fig. 1(b) and Fig. 1(c), we see that a good trade-off between exploitation-exploration is achieved when $c = 0.02$ or 0.2, for which the instantaneous regret approaches 0 gradually. The behavior of the instantaneous regret for $d = 100$ is similar and can be found in the supplementary.

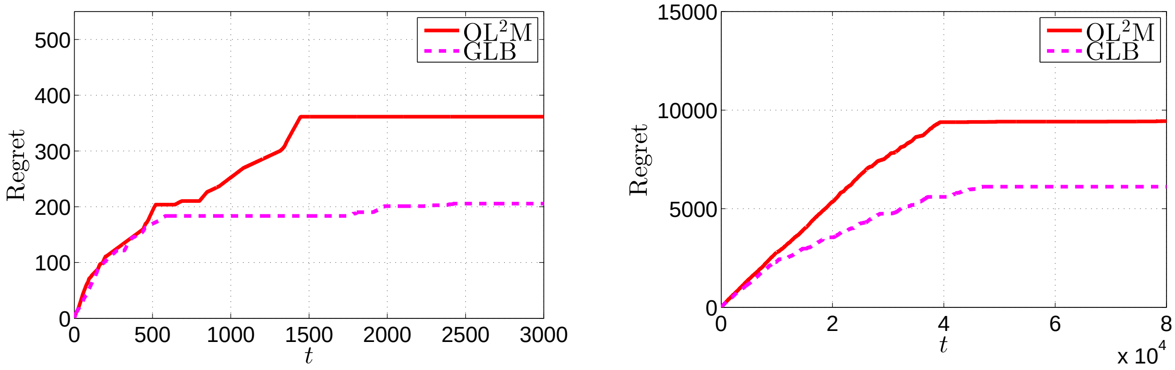
Next, we examine the $\widetilde{O}(d\sqrt{T})$ regret bound indicated by Theorem 3. Let Regret(t) be the regret till round t, i.e.,



(a) $d = 10$

(b) $d = 100$

Figure 2. Regret$(t)/(d\sqrt{t})$ of OL^{2}M when $\mathcal{D}$ is the unit ball in $\mathbb{R}^d$



(a) $d = 10$ and $|\mathcal{D}| = 100$

(b) $d = 100$ and $|\mathcal{D}| = 1000$

Figure 3. Regret of OL^{2}M and GLB when $\mathcal{D}$ contains finite number of actions.

Regret$(t) = \sum_{i=1}^{t} \mathbf{x}_*^\top \mathbf{w}_* - \mathbf{x}_i^\top \mathbf{w}_*$. If the learner achieves an $\widetilde{O}(d\sqrt{T})$ regret bound, the curve of Regret$(t)/(d\sqrt{t})$ should increase *at most* polylogarithmically. Fig. 2 plots the curve of Regret$(t)/(d\sqrt{t})$ with respect to t for $d = 10$ and 100. As can be seen, with a suitable choice of c, the curve indeed increases very slowly (e.g., $d = 100$ and $c = 0.02$), or even decreases slightly after certain rounds (e.g., $d = 10$ and $c = 0.02$).

In the last experiment, we study the case that $\mathcal{D}$ is finite, so that the GLB algorithm (Filippi et al., 2010) can also be applied. In the experiments, the parameter of GLB is also manually tuned. The decision set $\mathcal{D} \subseteq \mathbb{R}^d$ is constructed by sampling $10d$ points uniformly at random from the $(d-1)$-sphere. In Fig. 3, we plot the regret of OL^{2}M and GLB with respect to t. Note that in each round, GLB solves a logistic regression problem that utilizes the whole learning history to estimate $\mathbf{w}_*$. Thus, it is not surprising that the regret of GLB is smaller than OL^{2}M by a constant factor. On the other hand, OL^{2}M performs online updating, which is more efficient when t is large.

6. Conclusions

In this paper, we consider the problem of online linear optimization under one-bit feedback. Under the assumption that the binary feedback is generated from the logit model, we develop a variant of the online Newton step to approximate the unknown vector, and discuss how to construct the confidence region theoretically. Given the confidence region, we choose the action that produces maximal reward in each round. Theoretical analysis reveals that our algorithm achieves a regret bound of $\widetilde{O}(d\sqrt{T})$.

The current algorithm assumes that the one-bit feedback is generated from a logit model. In contrast, a much broader class of observation models are allowed in one-bit compressive sensing (Plan & Vershynin, 2013). In the future, we will investigate how to extend our algorithm to other observation models. Recently studies in online learning have shown that Thompson sampling is both competitive and efficient for addressing the exploration-exploitation dilemma (Chapelle & Li, 2011; Li, 2013). We leave the application of Thompson sampling to our problem for future work.

Acknowledgements

This work was partially supported by the NSFC (61333014), NSF (1463988, 1545995), and the Collaborative Innovation Center of Novel Software Technology and Industrialization of Nanjing University.

References

Abbasi-yadkori, Yasin, Pál, Dávid, and Szepesvári, Csaba. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pp. 2312–2320, 2011.

Agarwal, Alekh, Foster, Dean P., Hsu, Daniel, Kakade, Sham M., and Rakhlin, Alexander. Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, 23(1):213–240, 2013.

Agrawal, Rajeev. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

Ailon, Nir, Karnin, Zohar, and Joachims, Thorsten. Reducing dueling bandits to cardinal bandits. In *Proceedings of The 31st International Conference on Machine Learning*, 2014.

Auer, Peter. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

Auer, Peter, Cesa-Bianchi, Nicolò, and Fischer, Paul. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Bach, Francis. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15:595–627, 2014.

Bach, Francis and Moulines, Eric. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). In *Advances in Neural Information Processing Systems 26*, pp. 773–781, 2013.

Bartlett, Peter L., Bousquet, Olivier, and Mendelson, Shahar. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Berry, Donald A. and Fristedt, Bert. *Bandit problems: Sequential Allocation of Experiments*. Monographs on Statistics and Applied Probability. Springer Netherlands, 1985.

Boufounos, Petros T. and Baraniuk, Richard G. 1-bit compressive sensing. In *Proceedings of the 42nd Annual Conference on Information Sciences and Systems*, pp. 16–21, 2008.

Boyd, Stephen and Vandenberghe, Lieven. *Convex Optimization*. Cambridge University Press, 2004.

Bubeck, Sébastien and Cesa-Bianchi, Nicolò. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

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

Chapelle, Olivier and Li, Lihong. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems 24*, pp. 2249–2257, 2011.

Chen, Shang-Tse, Lin, Hsuan-Tien, and Lu, Chi-Jen. Boosting with online binary learners for the multiclass bandit problem. In *Proceedings of The 31st International Conference on Machine Learning*, 2014.

Cover, Thomas M. and Thomas, Joy A. *Elements of Information Theory*. Wiley-Interscience, second edition edition, 2006.

Dani, Varsha, Hayes, Thomas P., and Kakade, Sham M. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning*, pp. 355–366, 2008.

Filippi, Sarah, Cappe, Olivier, Garivier, Aurélien, and Szepesvári, Csaba. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23*, pp. 586–594, 2010.

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

Grant, Michael and Boyd, Stephen. Graph implementations for nonsmooth convex programs. In Blondel, V., Boyd, S., and Kimura, H. (eds.), *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pp. 95–110. Springer-Verlag Limited, 2008.

Grant, Michael and Boyd, Stephen. CVX: Matlab software for disciplined convex programming, version 2.1, March 2014.

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, 2009.

Hazan, Elad, Agarwal, Amit, and Kale, Satyen. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Hazan, Elad, Koren, Tomer, and Levy, Kfir Y. Logistic regression: Tight bounds for stochastic and online optimization. In *Proceedings of The 27th Conference on Learning Theory*, pp. 197–209, 2014.

Kakade, Sham M., Shalev-Shwartz, Shai, and Tewari, Ambuj. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 440–447, 2008.

Kleinberg, Robert, Slivkins, Aleksandrs, and Upfal, Eli. Multi-armed bandits in metric spaces. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pp. 681–690, 2008.

Lai, T. L. and Robbins, Herbert. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Li, Lihong. Generalized thompson sampling for contextual bandits. *ArXiv e-prints*, arXiv:1310.7163, 2013.

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

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

Yue, Yisong, Broder, Josef, Kleinberg, Robert, and Joachims, Thorsten. The K-armed dueling bandits problem. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Zhang, Lijun, Yi, Jinfeng, and Jin, Rong. Efficient algorithms for robust one-bit compressive sensing. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

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