Towards an Optimization Perspective for Bandits Problem

Junyan Liu [1] and Sijie Wang [1]

[1]University of California San Diego
{jul037,siw019}@ucsd.edu

Abstract

In this paper, we study the multi-armed bandit problem from an optimization perspective. In this problem, a decision-maker is faced with a fixed arm set and needs to design a strategy to pull an arm to minimize the cumulative loss, termed regret. At each round, the decision-maker adapts the pulling strategy by solving an optimization problem (OP), and the solution of OP is a probability distribution over all arms. Specifically, we consider two online optimization frameworks for OP, including follow-the-regularized-leader (FTRL) and online mirror descent (OMD). Our results show that both FTRL and OMD with a time-invariant learning rate and a proper entropy regularizer, will be reduced to a conventional bandit algorithm, called Exp3. Apart from equivalence, we also show that both methods are not the same when adopting a time-variant learning rate. We conduct numerical experiments to show that OP-based approach outperforms upper confidence bound (UCB) which is designed from a frequentist-statistical perspective. Our experiments also show that FTRL-based method, in general, outperforms OMD-based method, which is tuned with existing theoretical analysis.

1 Introduction

The multi-armed bandit (MAB) problem is a powerful framework to capture the exploration-exploitation tradeoff in the decision-making problems, including channel allocation Shi et al. (2020); Darak and Hanawal (2019), path routing Zhou et al. (2019), and crowdsourcing Abraham et al. (2013). In the standard MAB model, the decision-maker sequentially pulls arms from an arm set, and then observes the corresponding loss. The objective of the decision-maker is to design a strategy to pull arms, in order to maximize the cumulative rewards, or in turn minimize the cumulative learning loss. Most existing works typically study the MAB problem from a frequentist-statistical perspective, i.e., the pulling strategy is established upon a high-probability confidence interval. In particular, the confidence interval serves as the exploration purpose, whereas the available observations are used to exploit. A famous example is Upper Confidence Bound (UCB) algorithm Auer et al. (2002a) that explores unknown arms within a confidence interval that will gradually shrink with the increase of available information of unknown arms.

Recently, some works Zimmert and Seldin (2019); Amir et al. (2020); Zimmert and Seldin (2021); Ito (2021); Masoudian and Seldin (2021); Erez and Koren (2021) show surprising results that the pulling strategy can also be designed from the optimization-based perspective, and it achieves superior results, compared with conventional bandit algorithms, e.g., UCB. Motivated by this, this paper aims to survey recent progress which regards the exploration-exploitation trade-off in MAB problem as an optimization problem. Then, we show some equivalence between OP-based bandit algorithm and classical algorithms. Finally, we use extensive experiments to corroborate the superiority of optimization-based bandit algorithm.

Organization. The rest of this paper is organized as follows. In Section 2, we review related works. In Section 3, we formulate the setup and the primal problem. Section 4 presents a general OP-based framework of bandit learning and shows the equivalence between OP-based algorithm and two classical bandit algorithms. In Section 5, we present the regret bound and computational complexity. Finally, in Section 6, we conduct numerical experiments to corroborate our analysis.

2 Related Work

Multi-armed bandit problem is typically studied from a frequentist-statistical perspective Auer et al. (2002a); Bubeck et al. (2013); Agrawal and Goyal (2012); Garivier et al. (2016) until recent progress in optimization-based framework which achieves surprising results in not only basic K-armed bandit setup Zimmert et al. (2019); Zimmert and Seldin (2021); Masoudian and Seldin (2021); Ito (2021), but also many variants Erez and Koren (2021); Zimmert and Seldin (2020). Abernethy et al. (2015) makes an early attempt in this direction starts from follow-the-regularized-leader (FTRL) scheme and shows $O(\sqrt{T})$ regret bound, where T is a finite time horizon. This result matches the regret bound of the most classical adversarial bandit algorithm, Exp3 Auer et al. (2002b). However, when the environment generates the stochastic loss, i.e., the losses of arms are generated from fixed and unknown distributions, it cannot be improved to the $O(\log T)$ regret bound (ignoring the arm gap for simplicity). To address this issue, Zimmert and Seldin (2020) shows that when choosing a hybrid regularizer and adopting a time-variant learning rate, FTRL framework can achieve the best of both worlds, i.e., achieving $O(\log T)$ bound if the loss is stochastic, and achieving $O(\sqrt{T})$ bound if the loss is adversarial. Subsequently, Zimmert and Seldin (2021) shows that incorporating the Tsalis entropy with online mirror descent (OMD) framework can also achieve the best of both worlds. However, both optimization-based scheme requires an assumption that the optimal arm is unique, which in general, is not the case. The experiments in Zimmert and Seldin (2021) show that OMD works well even if there exist multiple optimal arms. Ito (2021) resolves this open problem by adopting a key technique, skewed Bregman divergence.

Apart from the most relevant works in the basic K-armed bandit setting, FTRL/OMD technique is also used in many variants. Zimmert and Seldin (2021); Ito (2021) shows that FTRL/OMD-based technique can be naturally used to achieve the optimal result in the corrupted bandit problem Lykouris et al. (2018), thanks to the self-bounding property (refer to Zimmert and Seldin (2021) for more details). Erez and Koren (2021) studies the bandit problem with graph feedback, and it achieves a bound with the dependence of clique covering number. The FTRL-based technique is extended to combinatorial bandit problem Zimmert et al. (2019) where the decision-maker needs to select an assortment from base arms, and also extended to delayed feedback setup Zimmert and Seldin (2020). Besides, these techniques are also well applicable for Markov decision process (MDP) with transition prior Jin and Luo (2020) and unknown transition Jin et al. (2021).

3 Problem Setting

The decision-maker is faced with a fixed arm set $[K] = \{1, 2, \cdots, K\}$. The interaction between the decision-maker and environment proceeds with $t = 1, ..., T$ as

1 Environment picks a loss[1] vector $\ell_t \in [0, 1]^K$ with i-th entry $\ell_{t,i} \in [0, 1]$.

2 Decision-maker designs a probability distribution $p_t = (p_{t,1}, \cdots, p_{t,K})$ by solving an optimization problem (OP).

[1]Some literature use reward $r_{t,i}$, and we note that analyzing both are equivalent, in the sense that $\ell_{t,i} = 1 - r_{t,i}$.

Algorithm 1 OP-based bandit framework

Input: Time horizon T, learning rate η_t, confidence $\delta \in (0,1)$.
Initialization: Set $p_{t,i} = \frac{1}{K}$ for $\forall i \in [K]$.

1: **for** $t = 1, ..., T$ **do**
2: Sample $I_t \sim p_t$ and observe loss ℓ_{I_t}.
3: Construct estimator $\hat{\ell}_{t,i}$ for each $i \in [K]$ by Eq. (6) or Eq. (7).
4: Update probability distribution $p_{t,i}$ for each $i \in [K]$ by Eq. (4) or Eq. (5).
5: **end for**

3 Decision-maker pulls an arm $I_t \sim p_t$ and observes loss ℓ_{I_t}.

At each round t, the decision-maker solves OP to get p_{t+1} by solving OP. Let Φ be a regularizer that will be chosen later and $D_\Phi(x, y)$ is a Φ-induced measure between x and y. This paper considers two optimization schemes, including FTRL and OMD as:

$$\text{FTRL:} \quad p_1 = \arg\min_{p \in \Delta} \Phi(p), \quad p_{t+1} = \arg\min_{p \in \Delta} \eta_t \langle p, \widehat{X}_t \rangle + \Phi(p),$$
$$\text{OMD:} \quad p_1 = \arg\min_{p \in \Delta} \Phi(p), \quad p_{t+1} = \arg\min_{p \in \Delta} \eta_t \langle p, \widehat{X}_t \rangle + D_\Phi(p, p_t),$$

where $\widehat{X}_t$ is the estimation up to round t that will be specified later and $\eta_t > 0$ is the learning rate and Δ is a probability simplex given as:

$$\Delta = \left\{ (p_{t,1}, \cdots, p_{t,K}) : \sum_{i=1}^{K} p_{t,i} \text{ and } p_{t,i} \geq 0 \text{ for } \forall i \in [K] \right\}.$$

Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration such that $\mathcal{F}_t = \sigma(\ell_1, \cdots, \ell_t, I_1, \cdots, I_t)$. We assume that the generation of loss is adaptive, in the sense that the loss vector ℓ_t selected by the environment at round t is $\mathcal{F}_{t-1}$-measurable. The performance of the learning process is evaluated by pseudo-regret $\overline{R}_T$ as:

$$\overline{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{I_t} \right] - \min_{i \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i} \right].$$

4 Algorithm

In this section, we present OP-based algorithm for bandit learning in Algorithm 1. In what follows, we first show the regularizer used for OP, and then provide two ways to constructing the estimator, both of which enjoy different properties. Finally, we theoretically show the equivalence of OMD, FTRL, and Exp3.

4.1 Entropy-based regularizer

FTRL optimization. We choose Φ as the negative Shannon entropy. The primal problem is as:

$$\min_{w \in \Delta} F(w), \quad \text{where} \quad F(w) = \eta_t \langle w, \widehat{X}_t \rangle + \sum_{i=1}^{K} p_i \log p_i.$$

Define Lagrangian $L(w, \lambda, \beta)$ as:

$$L(w, \lambda, \beta) = \eta_t \langle w, \widehat{X}_t \rangle + \sum_{i=1}^{K} p_i \log p_i + \lambda \left(w^\top \mathbb{1} - 1 \right) - \beta^\top w.$$

Define $g(\lambda, \beta)$ as the Lagrangian dual function. The dual problem is as:

$$\max_{\lambda, \beta \geq 0} g(\lambda, \beta) \quad \text{, where} \quad g(\lambda, \beta) = \min_{w \in \Delta} L(w, \lambda, \beta).$$

From KKT conditions, we know that

$$\begin{aligned} \eta_t \widehat{X}_{t,i} + 1 + \log(p_i^\star) + \lambda - \beta_i = 0, & \quad (1) \\ \beta_i^\star \geq 0, & \\ \sum_{i=1}^{K} p_i^\star = 1, & \quad (2) \\ \beta_i^\star p_i^\star = 0, & \quad (3) \\ p_i^\star \geq 0. & \end{aligned}$$

Note that Eq. (1) yields

$$p_i^* = \exp\left(-\eta_t \widehat{X}_{t,i} - 1 - \lambda^\star + \beta_i^\star \right).$$

Using Eq. (3), we have that $\beta_i^\star = 0$, and thus, the above equally gives

$$p_i^* = \frac{\exp\left(-\eta_t \widehat{X}_{t,i} \right)}{\exp\left(\lambda^\star + 1 \right)}.$$

Then, Eq. (2) gives

$$\sum_{i=1}^{K} \exp\left(-\eta_t \widehat{X}_{t,i} - 1 - \lambda^\star \right) = 1 \longrightarrow \sum_{i=1}^{K} \exp\left(-\eta_t \widehat{X}_{t,i} \right) = \exp\left(\lambda^\star + 1 \right).$$

Combining all the above, we have that

$$p_i^* = \frac{\exp\left(-\eta_t \widehat{X}_{t,i} \right)}{\sum_{j=1}^{K} \exp\left(-\eta_t \widehat{X}_{t,i} \right)}. \quad (4)$$

OMD optimization. We choose $D_\Phi(x, y) = \Phi(x) - \Phi(y) - \langle x - y, \nabla \Phi(y) \rangle$ as Bregman divergence between x and y induced by function Φ, and pick negative Shannon entropy as Φ. From these choices, one can get

$$D_\Phi(p, p_t) = \Phi(p) - \Phi(p_t) - \langle p - p_t, \nabla \Phi(p_t) \rangle = \sum_{i=1}^{K} p_i \log \frac{p_i}{p_{t,i}}.$$

The *primal problem* is written as:

$$\min_{w\in\Delta} F(w), \text{where } F(w) = \eta_t\langle w, \widehat{X}_t\rangle + \sum_{i=1}^{K} p_i \log\frac{p_i}{p_{t,i}}.$$

Define Lagrangian $L(w,\lambda,\beta)$ as:

$$L(w,\lambda,\beta) = \eta_t\langle w, \widehat{X}_t\rangle + \sum_{i=1}^{K} p_i \log\frac{p_i}{p_{t,i}} + \lambda\left(w^\top \mathbb{1} - 1\right) - \beta^\top w.$$

Define $g(\lambda,\beta)$ as the Lagrangian dual function. The dual problem is as:

$$\max_{\lambda,\beta\geq 0} g(\lambda,\beta), \text{where } g(\lambda,\beta) = \min_{w\in\Delta} L(w,\lambda,\beta).$$

Here, one can repeat the arguments in FTRL part and get the result as:

$$p_i^* = \frac{p_{t,i}\exp\left(-\eta_t \widehat{X}_{t,i}\right)}{\sum_{j=1}^{K} p_{t,j}\exp\left(-\eta_t \widehat{X}_{t,j}\right)}. \tag{5}$$

4.2 Estimator

In this subsection, we consider the way to constructing the estimator. Let $\hat{\ell}_t$ be the estimated loss of the observed loss ℓ_t. We adopt a widely-used way, called importance-weighted sampling, to construct unbiased estimation Auer et al. (2002b) as:

$$\hat{\ell}_{t,i} = \frac{\mathbb{1}\{I_t = i\}\ell_{t,i}}{p_{t,i}}. \tag{6}$$

One can see that Eq. (6) enjoys the conditionally unbiased property because

$$\mathbb{E}[\hat{\ell}_{t,i}|\mathcal{F}_{t-1}] = \mathbb{E}\left[\frac{\mathbb{1}\{I_t = i\}\ell_{t,i}}{p_{t,i}}\middle|\mathcal{F}_{t-1}\right] = \ell_{t,i}, \quad \mathbb{E}[\hat{\ell}_{t,i}^2|\mathcal{F}_{t-1}] = \mathbb{E}\left[\frac{\mathbb{1}\{I_t = i\}\ell_{t,i}^2}{w_{t,i}^2}\middle|\mathcal{F}_{t-1}\right] = \frac{\ell_{t,i}^2}{p_{t,i}},$$

where $\mathbb{E}[\ell_{t,i}|\mathcal{F}_{t-1}] = \ell_{t,i}$ holds since we assume $\ell_{t,i}$ is $\mathcal{F}_{t-1}$-measurable. One can see that the second moment scales with $1/p_{t,i}$, and thus the variance will become significantly large when $p_{t,i}$ is sufficiently small. This implies that the algorithm may become unstable, even though it would hold a sublinear regret in expectation. To overcome this issue, we, in light of Neu (2015), introduce an estimator with Implicit-eXploration (IX),

$$\hat{\ell}_{t,i} = \frac{\mathbb{1}\{I_t = i\}\ell_{t,i}}{p_{t,i} + \gamma_t}, \tag{7}$$

where $\gamma_t > 0$ is a time-varying exploration parameter. Again, one can check the conditional variance:

$$\mathbb{E}[\hat{\ell}_{t,i}^2|\mathcal{F}_{t-1}] = \mathbb{E}\left[\frac{\mathbb{1}\{I_t = i\}\ell_{t,i}^2}{(p_{t,i}+\gamma_t)^2}\middle|\mathcal{F}_{t-1}\right] = \frac{\ell_{t,i}^2 p_{t,i}^2}{(p_{t,i}+\gamma_t)^2} \geq \frac{\ell_{t,i}^2 p_{t,i}^2}{2w_{t,i}^2 + 2\gamma_t^2},$$

where the last inequality uses AM-GM inequality. Though this construct is not unbiased, one can see

that such an estimator guarantees a lower-bound of the second moment due to a newly-introduced parameter γ_t in the denominator. By choosing a proper γ_t, the estimation will not deviate far away from the unbiased one, while it can guarantee a reduced-variance performance.

4.3 On Equivalence between Exp3, FTRL, and OMD

Now, we show the way to choosing $\widehat{X}_{t,i}$ in Eq. (4), and also show the equivalence between Exp3, FTRL, and OMD when choosing different $\widehat{X}_{t,i}$. Let us first consider

$$\widehat{X}_{t,i} = \sum_{s=1}^{t} \hat{\ell}_{s,i}, \tag{8}$$

which immediately indicates that the OP can be rewritten as:

$$\min_{w \in \Delta} \eta_t \sum_{s=1}^{t} \langle w, \hat{\ell}_s \rangle + D_\Phi(w, p_t).$$

One can see that the above OP falls into FTRL scheme. Note also that if we apply Eq. (8) into Eq. (4), the update rule in Eq. (4) coincides with Exp3 algorithm (see (Bubeck and Cesa-Bianchi, 2012, Chapter 3.1)).
Suppose that we choose

$$\widehat{X}_{t,i} = \hat{\ell}_{t,i}. \tag{9}$$

Plugging the above results in the OMD update rule Eq. (5) to obtain

$$p_{t+1,i} = \frac{p_{t,i} \exp\left(-\eta_t \hat{\ell}_{t,i}\right)}{\sum_{j=1}^{K} p_{t,i} \exp\left(-\eta_t \hat{\ell}_{t,i}\right)}.$$

Now, we choose a time time-invariant learning rate $\eta_t = \eta > 0$, and recursively expand $p_{s,i}$ for $s \leq t$.

$$p_{t+1,i} = \frac{\exp\left(-\eta \hat{\ell}_{1,i}\right) \cdots \exp\left(-\eta \hat{\ell}_{t,i}\right)}{\sum_{j=1}^{K} \exp\left(-\eta \hat{\ell}_{1,i}\right) \cdots \exp\left(-\eta \hat{\ell}_{t,i}\right)} = \frac{\exp\left(-\eta \sum_{s=1}^{t} \hat{\ell}_{s,i}\right)}{\sum_{j=1}^{K} \exp\left(-\eta \sum_{s=1}^{t} \hat{\ell}_{s,i}\right)}.$$

Remark 1. *An interesting observation is that if we choose time-invariant learning rate $\eta_t = \eta > 0$, then, both update rules will coincide with each other. On the contrary, when adopting time-variant learning rate $\eta_t > 0$, both update rules do not match. Although both update rules have already been shown to enjoy the same regret $\widetilde{O}(\sqrt{KT})$[2] ignoring constant, they yield different empirical performance. Moreover, when considering some variants of basic MAB, e.g., full-information MAB, FTRL has been proved Amir et al. (2020) to always outperform OMD.*

5 Regret and Computational Complexity Analysis

In this section, we provide the regret analysis with different estimators, including unbiased one Eq. (6), and biased one with IX Eq. (6). From Bubeck and Cesa-Bianchi (2012), we know that

[2]We use $\widetilde{O}(\cdot)$ to hide all (poly)-logarithmic factors.

Theorem 1. *Running Algorithm 1 with learning rate* $\eta_t = \eta = \sqrt{\frac{2\log K}{TK}}$ *and adopting the unbiased estimator Eq. (7), incurs pseudo-regret* $\overline{R}_T$ *upper-bounded by*

$$\overline{R}_T = \sqrt{2TK\log K}.$$

One can see that the regret bound in Theorem 1 holds for regret notion $\overline{R}_T$, i.e., holds in expectation. A natural idea is that if one does not use pseudo-regret and instead measures the real loss as:

$$R_T = \sum_{t=1}^{T} \ell_{I_t} - \min_{i\in[K]} \sum_{t=1}^{T} \ell_{t,i}.$$

Neu (2015) gives a high-probability regret bound of R_T as follows.

Theorem 2. *Running Algorithm 1 with learning rate* $\eta_t = \eta = \sqrt{\frac{2\log K}{TK}}$ *and adopting the biased estimator with IX Eq. (7), with probability at least* $1-\delta$*, incurs regret* R_T *upper-bounded by*

$$R_T = 2\sqrt{2TK\log K} + \left(2\sqrt{\frac{2KT}{\log K}} + 1\right)\log(2/\delta).$$

Remark 2. *We note that a high probability regret bound can be trivially converted to the expected regret bound because one can choose* $\delta = 1/T$*, and thus the expected regret incurred by failure event can be upper-bounded by* $O(1)$*. As a consequence, the expected regret enjoys the same order. However, one cannot get a high-probability regret bound from the pseudo-regret bound. This is because* R_T *is a random variable, and we have no knowledge of the distribution of* R_T*.*

Remark 3. *Note also that expected regret* $\mathbb{E}[R_T]$ *does not coincide with pseudo-regret* $\overline{R}_T$ *in general cases. However, the upper bound of* $\mathbb{E}[R_T]$ *directly leads to the upper bound of* $\overline{R}_T$*. One can use Jensen's inequality to verify* $\overline{R}_T \leq \mathbb{E}[R_T]$ *always holds, and the equality holds iff the loss vectors are picked obliviously by the environment/adversary (refer to Bubeck and Cesa-Bianchi (2012); Lattimore and Szepesvári (2020) for more details).*

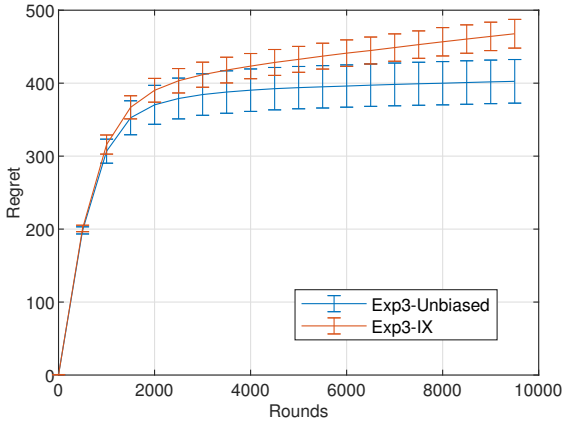
Computational complexity. As the algorithm needs to solve the optimization problem at each round, and it needs to update probability distribution over all K arms, and thus, the computational complexity is $O(KT)$. Note that existing works that use OP-based methods all need this complexity, and the way to reducing the complexity while maintaining a low regret remains open.

6 Numerical Experiments

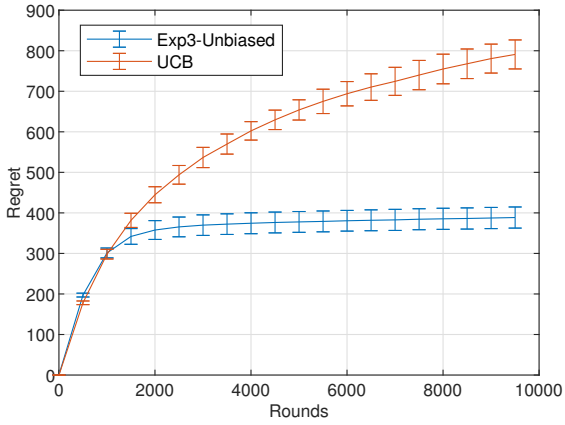
6.1 Experimental Setup

We here conduct numerical experiments to compare the performance of OP-based algorithm with other classical bandit algorithms, UCB Auer et al. (2002a). In the experiments, we consider $K = 10$ cases, and we consider stochastic loss and adversarial loss, respectively. For all the following experiments, we fix $T = 10^4$, $\delta = 1/T$. We assume that the loss across all arms are sampled from Bernoulli distributions. Let μ_i be the mean of Bernoulli distribution of arm i. Our experiments use the setup the same as the following table.

Arm i	1	2	3	4	5	6	7	8	9	10
Mean μ_i	0.8	0.4	0.4	0.4	0.4	0.4	0.1	0.1	0.1	0.1



(a) Evaluation of different estimators

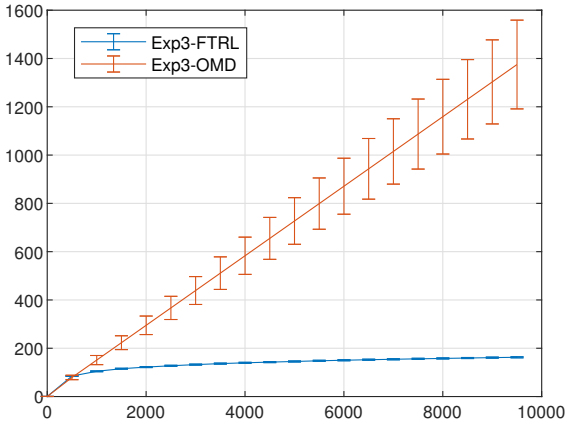


(b) Exp3 VS. UCB

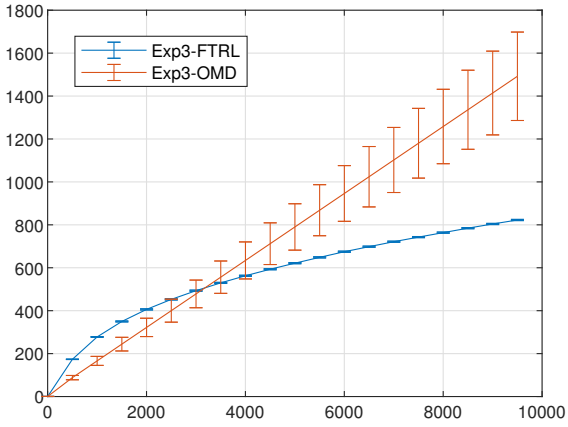
Figure 1: Comparison of regret performance with $K = 10$ and $T = 10^4$, where the error bar is the standard deviation.

6.2 Experimental Results

Fig. 1a compares the regret performance by using the time-invariant learning rate $\eta_t = \eta = \sqrt{2 \log K/(KT)}$. In such a choice of η, FTRL-based Exp3 and OMD-based Exp3 are identical. For brevity, we call this as Exp3. Fig. 1a shows the regret for Exp3 with unbiased and IX estimators. One can see that Exp3-Unbiased outperforms Exp3-IX, but Exp3-IX enjoys a smaller standard deviation than the one of Exp3-unbiased. This numeral results are tuned with our theoretical analysis. In particular, Theorem 1 shows that Exp3-Unbiased has the constant multiplication of $\sqrt{2}$ in upper bound, whereas Exp3-IX is with $2\sqrt{2}$ constant multiplication in upper bound. Moreover, Exp3-IX suffers an additional $\log(1/\delta)$, and when choosing sufficiently small δ, the regret bound will logarithmic times larger than the upper bound of Exp3-Unbiased. Both aspects explain why the unbiased estimator can incur a better regret result than the one with IX estimator.
Fig. 1b shows the regret performance between Exp3-Unbiased and UCB algorithm. We again use the time-invariant learning rate $\eta_t = \eta = \sqrt{2 \log K/(KT)}$, and we apply the same confidence interval $3\sqrt{\log(1/\delta)/(2T_i(t))}$ where $T_i(t)$ is the number of pulls of arm i up to time t. We can observe that Exp3-Unbiased significantly outperform UCB. This result shows that even in the stochastic setting, the OP-based bandit algorithm is superior to frequentist-statistical based bandit algorithm, UCB.
Fig. 2a presents the regret performance by adopting time-variant learning rate $\eta_t = \sqrt{2 \log K/(Kt)}$. As we now use time-variant η_t, the update rules of FTRL-based Exp3 and OMD-based Exp3 are different. One can observe that FTRL-based Exp3 not only achieves smaller regret, but also incurs less standard deviation. Fig. 2b evaluates the regret for two versions of Exp3 under IX estimation. The regret performance is similar to the result in Fig. 2a, i.e., FTRL-based Exp3 again outperforms OMD-based Exp3. These results match existing theoretical analysis Amir et al. (2020) that FTRL in general outperforms OMD, when both methods are not equivalent.
Another interesting observation from Fig. 1 and Fig. 2 is that time-variant learning rate, in most cases, is significantly better than time-invariant learning rate. This experimental results are tuned with the results of some recent progress Zimmert and Seldin (2021); Ito (2021); Masoudian and Seldin (2021). Specifically, those works show a surprising power of a dynamic η_t which can achieve the minimax optimal rate, and the constant factor in upper bound still remains small. More specifically, these works show that by using a dynamic learning rate scaling with $1/\sqrt{t}$, there is a hope to achieve $O(\sum_i \log T/\Delta_i)$ ($\Delta_i \in [0, 1]$ is the arm gap, see Bubeck and Cesa-Bianchi (2012))



(a) Unbiased estimator



(b) IX estimator

Figure 2: Comparison of regret performance with $K = 10$ and $T = 10^4$, where the error bar is the standard deviation (std). Note that std is divided by 10 for exposition simplicity.

rather than $\widetilde{O}(\sqrt{T})$ when the loss is stochastic. One can see that when the arm gap is not too small, $O(\sum_i \log T/\Delta_i)$ is with logarithmic order, significantly better than square-root bound $\widetilde{O}(\sqrt{T})$, if T is sufficiently large.

7 Conclusion and Future Work

In this paper, we study the basic K-armed bandit problem from an optimization perspective. Particularly, we consider two online optimization frameworks, including FTRL and OMD. We show that FTRL-based, OMD-based methods and classical Exp3 are equivalent when choosing a proper regularizer.

There are a number of possible extensions. Though FTRL/OMD-based methods have shown surprising results for both regret and empirical performance, it only has theoretical guarantees in basic K-armed bandit Zimmert and Seldin (2021) and linear bandit Lee et al. (2021). It remains open whether OMD or FTRL-based approaches can be extended to the contextual linear bandit setup, while achieving the best of both worlds, i.e., stochastic and adversarial setting. Besides, Erez and Koren (2021) show that FTRL-based algorithm can achieve the best of both worlds in bandit problem with graph feedback, but it requires the graph to be strongly observable. It is interesting to generalize FTRL-based method to the case that the graph is weakly observable. Finally, as mentioned in Section 5, designing a lazy OP-based method (e.g., based on doubling trick with epoch-based design) is also compelling.

8 Task assignment

Junyan is responsible for the problem formulation and literature review. Sijie is responsible for conducting experiments, and coding. They do together for the paper writing and discuss for the theoretical analysis.

References

J. D. Abernethy, C. Lee, and A. Tewari. Fighting bandits with a new kind of smoothness. *NeurIPS*, 28, 2015.

I. Abraham, O. Alonso, V. Kandylas, and A. Slivkins. Adaptive crowdsourcing algorithms for the bandit survey problem. In *COLT*, pages 882–910. PMLR, 2013.

S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *COLT*, pages 39–1, 2012.

I. Amir, I. Attias, T. Koren, R. Livni, and Y. Mansour. Prediction with corrupted expert advice. In *NeurIPS*, 2020.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002a.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b.

S. Bubeck and N. Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*, volume 5, chapter 3.1, pages 1–122. 2012.

S. Bubeck, T. Wang, and N. Viswanathan. Multiple identifications in multi-armed bandits. In *ICML*, volume 28, pages 258–265, 2013.

S. J. Darak and M. K. Hanawal. Multi-player multi-armed bandits for stable allocation in heterogeneous ad-hoc networks. *IEEE Journal on Selected Areas in Communications*, 37(10):2350–2363, 2019.

L. Erez and T. Koren. Towards best-of-all-worlds online learning with feedback graphs. *NeurIPS*, 34, 2021.

A. Garivier, T. Lattimore, and E. Kaufmann. On explore-then-commit strategies. *NeurIPS*, 29: 784–792, 2016.

S. Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *COLT*, volume 134, pages 2552–2583, 2021.

T. Jin and H. Luo. Simultaneously learning stochastic and adversarial episodic mdps with known transition. In *NeurIPS*, volume 33, 2020.

T. Jin, L. Huang, and H. Luo. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. *NeruIPS*, 34, 2021.

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

C.-W. Lee, H. Luo, C.-Y. Wei, M. Zhang, and X. Zhang. Achieving Near Instance-Optimality and Minimax-Optimality in Stochastic and Adversarial Linear Bandits Simultaneously. In *arXiv:2102.05858*, 2021.

T. Lykouris, V. S. Mirrokni, and R. P. Leme. Stochastic bandits robust to adversarial corruptions. In *STOC*, 2018.

S. Masoudian and Y. Seldin. Improved analysis of robustness of the tsallis-inf algorithm to adversarial corruptions in stochastic multiarmed bandits. In *COLT*, 2021.

G. Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *NeurIPS*, volume 28, 2015.

C. Shi, W. Xiong, C. Shen, and J. Yang. Decentralized multi-player multi-armed bandits with no collision information. In *AISTATS*, pages 1519–1528. PMLR, 2020.

P. Zhou, J. Xu, W. Wang, Y. Hu, D. O. Wu, and S. Ji. Toward optimal adaptive online shortest path routing with acceleration under jamming attack. *IEEE/ACM Trans. Netw.*, 27(5):1815–1829, 2019.

J. Zimmert and Y. Seldin. An optimal algorithm for stochastic and adversarial bandits. In *AISTATS*, volume 89, pages 467–475, 2019.

J. Zimmert and Y. Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *AISTATS*, pages 3285–3294. PMLR, 2020.

J. Zimmert and Y. Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *J. Mach. Learn. Res.*, 22:28:1–28:49, 2021.

J. Zimmert, H. Luo, and C. Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *ICML*, volume 97, pages 7683–7692, 2019.