2024 International Conference on Smart Applications, Communications and Networking (SmartNets) | 979-8-3503-8532-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/SMARTNETS61466.2024.10577740

Provably Correct SGD-based Exploration for Generalized Stochastic Bandit Problem

Jialin Dong*, Jiayi Wang†, and Lin F. Yang*
* University of California, Los Angeles, CA, US, 90095
†The University of Utah, Salt Lake City, UT, US, 84112
E-mail: jialind@g.ucla.edu, jiayi.wang@utah.edu, linyang@ee.ucla.edu

***Abstract*—Bandit problems have been widely used in wireless communication systems which involve generalized reward models and may suffer high computational complexity. Despite the success of applying stochastic gradient descent (SGD) in stochastic bandits to reduce computational complexity, several limitations persist in state-of-the-art. Specifically, current papers only consider linear models which is not practical in wireless communication. Their algorithms are only guaranteed by the expected regret bound, which may not be effective when many actions are sub-optimal. Additionally, existing SGD-based approaches raise bias in the estimation due to a greedy action selection strategy, deviating from the conventional SGD approach that uniformly samples. To address these limitations, we propose an online SGD-based algorithm with a high probability regret bound guarantee, which can apply to stochastic bandits with general parametric reward functions. We develop an action-elimination strategy to gradually eliminate sub-optimal actions and uniformly at random select the action from the current action subset. This strategy guarantees an unbiased estimation of model parameters. Theoretically, we prove that our proposed algorithm can achieve the regret of $O(d\sqrt{n\log(n/\delta)})$ with probability at least $1-\delta$, where n is the number of time steps and d is the dimension of model parameters, matching existing near-optimal regret bounds in UCB-type algorithms. We further conduct experiments to demonstrate the advantage of our algorithm.**

I. Introduction

Online stochastic bandits represent a class of sequential decision-making problems where an agent makes actions and receives uncertain rewards. The applications in wireless communication range from client scheduling [1] to channel selection [2], [3], [4]. The goal of the agent is to maximize the cumulative rewards over n time steps by strategically selecting actions based on streaming data. A line of literature has developed effective algorithms for online stochastic bandits. Compared with the common methods, such as the upper confidence bound (UCB) bandit algorithm [5], [6] and online mirror descent (OMD) [7], [8], the SGD-based methods [9], [10], [11] can effectively reduce computational complexity by avoiding the matrix inverse operations when estimating the model parameter. However, several limitations persist in the current SGD-based methods for online stochastic bandits.

Firstly, existing online algorithms predominantly focus on linear models, while the general parametric model is unexplored. Secondly, prior SGD-based approaches only focused on expected regret bounds and did not tackle high probability bounds, leaving uncertainties about their algorithms in achieving desirable regret bounds when involving too many sub-optimal actions. Thirdly, current SGD-based approaches introduce bias in their estimators. This bias arises from a greedy action selection strategy at each time step, deviating from the conventional SGD approach that uniformly samples from all available data points. The presence of bias implies a larger divergence between the estimation and the ground truth, potentially compromising result robustness. In other words, different datasets may yield significantly different estimation results.

To address the above limitations, we consider SGD-based stochastic bandit problems with a general parametric model, emphasizing performance guarantees that hold high probability, an aspect lacking in current literature due to the considerable technical effort and modifications required to establish such guarantees. Specifically, the general parametric models usually involve complex optimization problems. It is vital to make reasonable but not strict assumptions about the model to guarantee feasible solutions. Furthermore, the statistical analysis associated with general parametric models requires precisely establishing corresponding i.i.d. random variables, necessitating the utilization of random matrix theory thorough analysis.

The contributions can be summarized as follows:

979-8-3503-8532-8/24/ $31.00 ©2024 IEEE

1) **General framework.** Our proposed method applies to stochastic bandits with a general parametric reward functions.
2) **High probability bound.** The proposed algorithm endows with a high probability regret-bound guarantee.
3) **Unbiased Action-elimination strategy.** We design a strategy to gradually eliminate sufficiently suboptimal actions. The proposed algorithm uniformly at random selects the action from the current action subset. This strategy guarantees an unbiased estimation of model parameters, yielding a robust and desirable upper regret bound.

A. Related Work

Online algorithms that use streaming data to update the models, rather than waiting for a complete set of data, reduce the need for extensive storage on previously seen data. This idea has been recently applied to the online stochastic bandits problem. Even though papers such as [12], [13], [14] proposed online-mirror-descent-based methods [7], these methods involve matrix inverse operations, which still brings computational complexity of $O(nd^2)$ when the dimension of the feature vector is large.

A notable line of literature ([15]) extensively explores the high-probability expected estimator error of SGD, forming the basis for regret analyses in SGD-based bandit problems ([10], [11]). However, these investigations are confined to the linear setting, lacking generalizability. This paper seeks to fill this gap by undertaking a comprehensive examination of SGD-based algorithms within a broader bandit problem framework, delving into both theoretical and empirical dimensions. The comparison between our result and state-of-the-art is illustrated in Table I. The first four papers lack consideration for high probability bounds. In contrast to [5], our bound demonstrates an improvement by a multiplicative factor of approximately $\sqrt{\log(n)}$. Notably, our method not only attains a near-optimal bound akin to [6] but also enjoys low computational complexity.

II. Preliminaries

Let $\mathcal{D} \subset \mathbb{R}^d$ be a compact set of decisions the environment decides. At each time t, the learning algorithm determines a subset $\mathcal{A}_t \subseteq \mathcal{D}$ and the agent selects an action $x_t \in \mathcal{A}_t$, after which the agent observes a reward y_t.

We denote $\mathcal{H}_t$ as the history $(\mathcal{A}_1, x_1, y_1, \ldots, \mathcal{A}_{t-1}, x_{t-1}, y_{t-1}, \mathcal{A}_t)$ of observations available to the agent when choosing an action x_t. After choosing the action x_t, the agent receives a reward y_t that is a function with respect to a certain parameter $\theta_\star \in \mathbb{R}^d$, i.e., for all $x_t \in \mathcal{D}$, $y_t = r(\theta_\star, x_t) + \epsilon_t$ where ϵ_t denotes the noise. Generally, the vector $\theta_\star$ is unknown, though fixed.

We begin with two standard assumptions for most bandit problems [16]. The first assumption sets the range of the reward function.

Assumption 1 (Reward function)**.** *Define a function $r : \mathbb{R}^d \times \mathcal{D} \to \mathbb{R}$. The reward function for bandit problem is represented as $r(\theta_\star, x)$ for all $x \in \mathcal{D}$ and a certain parameter $\theta_\star \in \mathbb{R}^d$ where $\|\theta_\star\|_2 \le S$ for $S > 0$.*

Our second assumption ensures that observation noise is light-tailed. A wide range of noise, e.g., Gaussian and sub-Gaussian noise, is covered by this assumption.

Assumption 2 (Noise assumption)**.** *For all $t \in [n], \epsilon_t = y_t - r(\theta_\star, x)$ conditioned on $\mathcal{H}_t$ is σ-sub-Gaussian, i.e., $\mathbb{E}[\exp(\lambda \epsilon_t)] \le \exp\left(\lambda^2\sigma^2/2\right)$ almost surely for all λ.*

We let $x^* \in \arg\max_{x \in \mathcal{D}} \; r(\theta_\star, x)$ denote the optimal action. The n period cumulative regret is $\mathrm{Reg}(n) = \sum_{t=1}^{n} [r(\theta_\star, x^*) - r(\theta_\star, x_t)]$ where $\{x_t : t \in [n]\}$ denote the actions.

III. Algorithm

We present our proposed algorithm in this section. To begin with, we need to estimate the model parameter in the bandit problem. The efficient estimation is intractable for a general function $r(\theta, x)$ unless we consider the bandit problem under some reasonable and mild assumptions. These assumptions ensure that stochastic gradient descent can efficiently and effectively apply to model parameter estimation. Before presenting assumptions, we start with the representation of the loss function in estimation.

Suppose that decisions $x_1, \ldots x_n \in \mathcal{D}$ have been made, corresponding rewards are $y_1, \ldots, y_n \in \mathbb{R}$. The loss function at i-th step is defined as $\ell_i(r(\theta, x_i), y_i), \; \forall \; \theta \in \mathbb{R}^d, \; i \in [n]$, which depends on data pair (x_i, y_i) and the reward function $r(\theta, x_i)$. To guarantee efficient and effective parameter estimation, we consider the problem of minimizing a smooth and convex function by stochastic gradient descent. Specifically, we make the following assumptions on the loss function $\ell_i(r(\theta, x_i), y_i)$. We abuse the notation $\ell_i(\theta)$ to represent $\ell_i(r(\theta, x_i), y_i)$ in the following.

Assumption 3 (Loss function)**.** *We assume that the loss function $\ell_i(\theta)$ (III) with $i \in [n]$ satisfies, for some Lipschitz constants $L, L_G, L_H > 0$, convexity constant $\mu > 0$ and any points $\theta, \theta' \in \mathbb{R}^d$*

TABLE I
COMPARISON OF OUR MAIN RESULT AND STATE-OF-THE-ART.

Paper	Model	Algorithm	Regret	Computational complexity
[8]	Linear	OMD-based	$O(d\sqrt{n\log(n)})$	$O(nd^2)$
[10]	Linear	SGD-TS	$O(d\sqrt{n\log(n)})$	$O(nd)$
[11]	Linear	SGD-based	$O(d\sqrt{n\log(n\log(n))})$	$O(nd)$
[9]	Linear	SGD-based	$O(d\log^4 n\cdot\sqrt{n})$	$O(nd)$
[5]	Linear	UCB	$O\Bigg(d\log(n)\sqrt{n}$ $+\sqrt{dn\log\left(\frac{n}{\delta}\right)}\Bigg)$	$O(nd^2)$
[6]	General parametric	UCB	$O(d\sqrt{n\log(n/\delta)})$	$O(n^2d^2)$
Theorem 1	**General parametric**	**SGD-based**	$\mathbf{O(d\sqrt{nlog(n/\delta)})}$	$\mathbf{O(nd)}$

- *Convexity and smoothness:*

$$\ell_i(\theta') - \ell_i(\theta) - \langle\nabla\ell_i(\theta), \theta' - \theta\rangle \geq \frac{\mu}{2}\|\theta' - \theta\|_2^2,$$
$$\|\nabla\ell_i(\theta') - \nabla\ell_i(\theta)\|_2 \leq L_G\|\theta' - \theta\|_2,$$
$$|r(\theta', x_i) - r(\theta, x_i)| \leq L\|\theta' - \theta\|_2.$$

  In other words, $\ell_i(\theta)$ is L_G-smooth with respect to θ, and $r(\theta, x_i)$ is L-smooth with respect to θ.

- *Hessian-Smoothness: $\left\|\nabla^2\ell_i(\theta') - \nabla^2\ell_i(\theta)\right\|_2 \leq L_H\|\theta' - \theta\|_2$, which is equivalent to*

$$\left\|\nabla\ell_i(\theta') - \nabla\ell_i(\theta) - \nabla^2\ell_i(\theta)(\theta' - \theta)\right\|_2 \leq \frac{L_H}{2}\|\theta' - \theta\|_2^2. \quad (1)$$

- *Bounded gradient: For the model parameter $\theta_\star$ in Assumption 1, we have $\mathbb{E}[\|\nabla\ell_i(\theta_\star)\|_2] \leq \mu\|\theta_\star\|_2$.*

Assumption 3 can be easily satisfied by a wide range of loss functions under various scenarios. An example of stochastic linear bandit problems that satisfies Assumption 3 is presented in Example 1 in Section IV.

Based on Assumption 3, we update the estimator of the model parameter via the mini-batch averaged SGD [15]. Additionally, we design an action-elimination strategy to gradually eliminate sufficiently sub-optimal actions in the learning process and maintain near-optimal actions. At each round of mini-batch averaged SGD, we uniformly and randomly select the action from the current action subset to guarantee the corresponding feature vectors are i.i.d. The details on designing proper action subsets are presented in Section IV, which is our main argument to guarantee near-optimal regret bound.

According to the above discussions, the algorithm is presented as follows. An estimate $\hat{\theta}$ to the ground-truth vector $\theta_\star$ can be constructed by

$$\hat{\theta} := \arg\min_{\theta}\ \mathcal{L}(\theta), \text{ where } \mathcal{L}(\theta) := \frac{1}{n}\sum_{i=1}^{n}\ell_i(\theta) \quad (2)$$

where $\ell_i(\theta)$ is the loss function. We apply the mini-batch SGD to estimate the model parameter. Initialized at θ_0, the t-th iteration is computed by the mini-batch SGD with step size η_t. We define B as the mini-batch size and step sizes η_t will be discussed in our proposed algorithm. In each round t, the agent randomly selects an action $x_t \in \mathcal{A}_t$. The proposed SGD for general bandit is illustrated in Algorithm 1.

Our algorithm (Algorithm 1) is an exploration-exploitation modification of mini-batch averaged SGD, where the action-elimination strategy realizes the exploration-exploitation balancing. At a high level, each round consists of one inner loop over all mini-batch sizes B. Before initiating the inner loop, an action subset (referenced in line 4) is established. In line 4, the objective is to filter out actions from $\mathcal{A}_{t-1}$ whose rewards deviate significantly from the maximum action reward within $\mathcal{A}_{t-1}$ by solving the optimization problem (3). The remaining actions are then defined as $\mathcal{A}_t$. In the case of a nonconvex reward function, the problem (3) becomes intractable in general. However, when the reward function adheres to suitable statistical models (e.g., low-rank matrix), straightforward first-order methods are assured to discover a local minimum with a minimal number of iterations. This approach can still achieve low computational and sample complexities.

After getting an action set $\mathcal{A}_t$, the inner loop (line 5-9) uniformly and randomly selects an action from the action set $\mathcal{A}_t$ to guarantee the i.i.d. property of the action x_i and the reward y_i. It ensures the unbiased estimation of stochastic gradient. Subsequently, the inner loop (line

Algorithm 1 SGD-based Algorithm for General Stochastic Bandits

```
Require: Neighborhood radius $\beta_t$, $\forall\ t \in \{1, 2, \cdots, T\}$, number of outer iteration rounds $T$, mini-batch size $B$, step-size $\eta_t = \eta_0 t^{-\alpha}$ where $\alpha \in (0, 1)$.
1: Initialize $\theta_0 \in \mathbb{R}^d$ such that $\|\theta_0\|_2 \leq S$, $\bar{\theta}_0 = 0 \in \mathbb{R}^d$, $\mathcal{A}_0 = \mathcal{D}$, $\beta_0 = S$.
2: for $t = 1$ to $T$ do
3:   Initial $g_t = 0 \in \mathbb{R}^d$.
4:   Update action set as
       $$\mathcal{A}_t := \{x \in \mathcal{A}_{t-1} | \max_{a \in \mathcal{A}_{t-1}} r(\bar{\theta}_{t-1}, a) - r(\bar{\theta}_{t-1}, x) \leq 2L\beta_{t-1}\}, \quad (3)$$
     where $L$ is defined in Assumption 3.
5:   for $i = (t-1)B + 1$ to $tB$ do
6:     Uniformly at random select an action $x_i \in \mathcal{A}_t$.
7:     Observe the reward $y_i$.
8:     Update $g_t = g_t + \frac{1}{B}\nabla \ell_i(\theta_{t-1})$.
9:   end for
10:  Update $\theta_t = \theta_{t-1} - \eta_t g_t$.
11:  Compute $\bar{\theta}_t = t^{-1}(\theta_t + (t-1)\bar{\theta}_{t-1})$
12: end for
13: Output: $\bar{\theta}_T$
```

5-9) updates the stochastic gradient g_t that is used to form the iterate, i.e., θ_t. The analysis of our proposed algorithm is provided in the next section.

IV. Main Theory

In this section, we present our main result, Theorem 1, which provides the sample complexity guarantee for Algorithm 1 in general bandit problem under mild assumptions.

We begin with notations used in the main theorem. Let

$$\Psi_{\chi,\alpha} = \int_1^{\infty} \exp\left(-\chi \int_1^z x^{-\alpha} dx\right) dz \leq C_{\chi,\alpha}, \quad (4)$$

where $C_{\chi,\alpha} > 0$, $\chi > 0$, and $0 < \alpha \leq 1$. We define $\lambda_* = \max_t \lambda_{\max}(\nabla^2 \ell_t(\theta_\star))\ t \in [T]$ in the following. In Theorem 1, let $\sigma > 0$ denote the standard deviation of noise for the gradient caused by the noise $\{\epsilon_t\}$ under Assumption 2.

Theorem 1. *Under Assumptions 1, 2, and 3, we set*

$$\beta_t = C''_{\chi,\alpha}\sqrt{\frac{16 d L_H^2}{t\mu^3}\log\left(\frac{t}{\delta}\right)},\ t \in [T], \quad (5)$$

$\alpha = 1/2$, the mini-batch size $B = \mu\sigma^2 d/L_H^2$, and the initial step-size $\eta_0 \leq 1/\lambda_*$, Algorithm 1 achieves the following regret with probability at least $1 - \delta$,

$$\mathrm{Reg}(n) \leq \frac{2\mu\sigma^2 SLd}{L_H^2} + 32 \cdot \sigma C''_{\chi,\alpha} dL\sqrt{\frac{n}{\mu^2}\log\left(\frac{n}{\delta}\right)}, \quad (6)$$

where $n = TB$ and $C''_{\chi,\alpha} \asymp C_{\chi,\alpha}$ with $C_{\chi,\alpha}$ derived from (4) with $\chi = \mu\eta_0/2$ and $\alpha = 1/2$.

The first term of the regret bound (6) comprises a constant determined by the parameters. Specifically, a more concentrated tail distribution, reflected by a smaller σ^2, results in a diminished upper regret bound. If the slope (first-order derivative) and the curvature (second-order derivative) of the loss function don't change too rapidly across its domain, reflected by a smaller L and a larger L_H, the upper regret bound will be reduced. The dominant factor in the regret bound (6) is the second part. If the loss function's curvature changes more gradually (rapidly) across its domain, reflected by a larger μ, this leads to a reduced upper bound on regret. Thus, there is a trade-off between the first and second terms of the regret bound (6).

Example 1 (Linear Bandits)**.** *Consider linear bandits where the function $r(\theta, x) = \theta^\top x$ is 2-smooth with respect to $\theta \in \mathbb{R}^d$ for a fixed $x \in \mathcal{D}$ since $\nabla_\theta r(\theta, x) = x$. The loss function of estimating θ can be given by*

$$\ell_i(\theta) = \frac{1}{2n}\left(\theta^\top x_i - y_i\right)^2 + \frac{\mu}{2}\|\theta\|_2^2 \quad (7)$$

for $i \in [n]$ and $\mu > 0$. Let $\mu = 1$, Algorithm 1 solves linear bandits and achieves the following regret with probability at least $1 - \delta$,

$$\mathrm{Reg}(n) \leq \frac{4\sigma^2 Sd}{L_H^2} + 64 \cdot \sigma C''_{\chi,\alpha} d\sqrt{n \log\left(\frac{n}{\delta}\right)},$$

where $n = TB$ and $C''_{\chi,\alpha} \asymp C_{\chi,\alpha}$ with $C_{\chi,\alpha}$ defined in (4), which is comparable to the result for linear bandit in [5] $O(d\sqrt{n \log(n/\delta)})$.

V. Proof Sketch

In this section, we provide an overview of the key mechanisms behind the regret bound in Theorem 1. Our analysis is composed of the following three steps.

Step 1: Neighborhood construction. We define the neighborhood of the averaged iterate $\bar{\theta}_t$ at round t as follows.

$$\mathcal{B}_t := \left\{\nu : \left\|\nu - \bar{\theta}_t\right\|_2 \leq \beta_t\right\}. \quad (8)$$

Based on β_t (5) in $\mathcal{B}_t$ (8), $\theta_\star$ stays in $\mathcal{B}_t, \forall t$ with high probability, which ensures that the averaged iterate $\bar{\theta}_t$ converges to the ground truth $\theta_\star$. The value of $\beta_t = \tilde{O}(1/\sqrt{t})$ (5) plays a important role in regret analysis. The primary purpose of our neighborhood $\mathcal{B}$ is to contract and converge towards the ground truth parameter $\theta_\star$, aiming to identify the true parameter rather than quantifying uncertainty about the mean reward of each arm.

Step 2: Action-elimination. We further show that $\forall t \in \mathbb{N}$, the set $\mathcal{A}_t$ (3) contains the optimal action x^* with high probability. It guarantees that our proposed algorithm chooses near-optimal action and eliminates actions that are sufficiently suboptimal as time goes by. At each t round of Algorithm 1, the regret of action $x_i \in \mathcal{A}_t$ is bounded by β_{t-1}, which contributes to bound regret in the following step.

Step 3: Regret bound. Combining the above two steps, we bound the total regret by the sum of the regrets concerning each selected action. At each t round of Algorithm 1, an action is selected from the action set $\mathcal{A}_t$ that is updated at the beginning of the inner loop. Herein, we bound the regret of action $x_i \in \mathcal{A}_t$ with $4L\beta_{t-1}$.

VI. Simulation Results

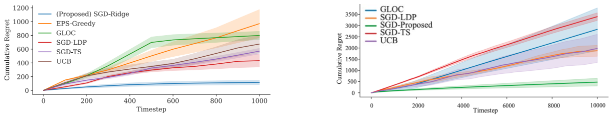
In this section, we provide the experimental results with industry-standard synthetic datasets for both linear and logistic bandit.

- Linear bandit: We set the number of rounds $T = 1000$ and conduct simulations on the parameter: $K = 30$ (K is the number of action) and $d = 2$. We build linear bandit models, where the feature vectors $\{x_i\}$ and the true model parameter $\theta_\star$ are drawn i.i.d. from Gaussian distribution $\mathcal{N}(0, I_d)$ and normalize to $\|x_i\|_2 = 1, \ \|\theta_\star\|_2 = 1$. The loss function for a linear bandit is the form of (7) with the regularization parameter μ.
- Logistic bandit: We set the number of rounds $T = 10000$ and conduct simulations on the parameter: $K = 40$ and $d = 2$. We draw $\{x_i\}$ and the true model parameter $\theta_\star$ iid from uniform distribution in the interval of $[-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}]$. We build a logistic model on the dataset and draw random rewards y_t from a Bernoulli distribution with mean $1/(1+\exp(x_i^\top \theta^*))$. The loss function for logistics is the form of

$$\ell_i(\theta) = \frac{1}{2n}\left(\left(1+\exp(\theta^\top x_i)\right)^{-1} - y_i\right)^2 + \frac{\mu}{2}\|\theta\|_2^2 \tag{9}$$

for $i \in [n]$ and μ, which satisfies Assumption 3.

To ensure a fair comparison, we evaluate SGD-Ridge (for linear bandit) and SGD-Proposed (for logistic bandit) alongside established methods including ϵ-greedy [16], [17], GLOC [18], SGD-LDP [11], SGD-TS [10], and UCB [19], with their codes available publicly. We standardize noise levels, considering both privacy noise [11] and reward noise.



Linear bandit

Logistic bandit

Fig. 1. The cumulative regret vs. time-step of different algorithms.

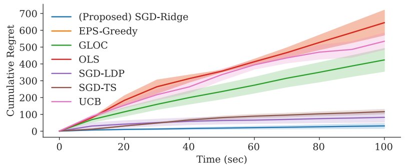
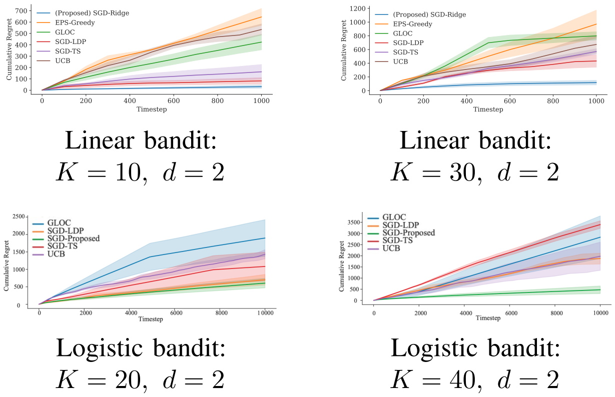


Fig. 2. The cumulative regret vs. computational time of different algorithms.

Parameter tuning is conducted uniformly across all algorithms. For GLOC and UCB-GLM, we explore exploration rates in $0.01, 0.1, 1, 5, 10$. The exploration probability of ϵ-greedy is set as $\frac{c}{\sqrt{t}}$ at the t-th iteration, with c selected from $0.01, 0.1, 1, 5, 10$. For SGD-based algorithms, we set mini-batch size $B = 16d/\lambda$, with λ tuned in $0.01, 0.1, 1, 5, 10$. The parameter β_t (5) is set as $\sqrt{4d\log(\frac{t}{10^{-3}})/t}$. Step size η_t is $\eta_0/\sqrt{t}$, where η_0 is chosen from $0.01, 0.05, 0.1, 0.5, 1, 5, 10$. Regularization parameter μ for each algorithm is searched from $0.01, 0.05, 0.1, 0.15$.

We perform experiments 30 times and plot the mean and standard deviation of their regrets, which are illustrated in Fig 1 and Fig 2. It shows that our proposed algorithms outperform state-of-the-art approaches. The beneficial performance is due to a good balance between exploitation and exploration via an action-elimination strategy and efficient estimation via the mini-batch SGD method. Moreover, random selection during each mini-batch guarantees an unbiased gradient, which outperforms greedy selection used by Han et al. [11].

To further illustrate the computational efficiency of the proposed algorithm, we set $K = 1000$ and $d = 100$ for linear bandits and keep other settings mentioned above. Fig 1 presents timing curves that measure the cumulative regret vs. the computational time each algorithm takes. All the algorithms are required to solve similar optimization problems as (3) which aims to find the proper arm to pull. The advantage of our proposed algorithm on



Linear bandit: $K = 10,\ d = 2$

Linear bandit: $K = 30,\ d = 2$

Logistic bandit: $K = 20,\ d = 2$

Logistic bandit: $K = 40,\ d = 2$

Fig. 3. (a) and (b) illustrate the cumulative regret of different algorithms for linear bandit concerning timestep. (c) and (d) illustrate the cumulative regret of different algorithms for logistic bandit for timestep.

computational time mostly comes from the efficiency of estimating parameters via SGD.

We further provide simulated experiments in industry-standard synthetic datasets for both linear and logistic bandit. We plot the mean and standard deviation of their regrets, which are illustrated in Fig 3. Our proposed algorithms outperform state-of-the-art approaches and maintain advantages with larger K in both linear and logistic bandits. The beneficial performance is due to a good balance between exploitation and exploration via an action-elimination strategy. Moreover, random selection during each mini-batch guarantees an unbiased gradient, which outperforms greedy selection used by Han et al. [11].

VII. Conclusion

In this paper, we present the SGD-based algorithm for generalized stochastic bandits, focusing on regret-bound guarantees that hold with high probability. In addition to theoretical validation, we conducted experiments to showcase the practical effectiveness of our proposed algorithm. The results highlight the improved performance and versatility of our approach in handling stochastic bandits with general parametric reward functions. By addressing the identified limitations of current works, our algorithm presents a promising advancement in the application of SGD to stochastic bandit problems, paving the way for more robust and efficient solutions in real-world scenarios. In addition to the stochastic bandit problems addressed in this paper, it is interesting to investigate more complex and general reward models, such as neural networks in the future.

References

[1] W. Xia, T. Q. Quek, K. Guo, W. Wen, H. H. Yang, and H. Zhu, "Multi-armed bandit-based client scheduling for federated learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7108–7123, 2020.

[2] A. Slivkins *et al.*, "Introduction to multi-armed bandits," *Foundations and Trends® in Machine Learning*, vol. 12, no. 1-2, pp. 1–286, 2019.

[3] S. Maghsudi and S. Stańczak, "Channel selection for network-assisted d2d communication via no-regret bandit learning with calibrated forecasting," *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1309–1322, 2014.

[4] F. Li, D. Yu, H. Yang, J. Yu, H. Karl, and X. Cheng, "Multi-armed-bandit-based spectrum scheduling algorithms in wireless networks: A survey," *IEEE Wireless Communications*, vol. 27, no. 1, pp. 24–30, 2020.

[5] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

[6] D. Russo and B. Van Roy, "Eluder dimension and the sample complexity of optimistic exploration," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[7] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge university press, 2006.

[8] S. Bubeck, N. Cesa-Bianchi, and S. M. Kakade, "Towards minimax policies for online linear optimization with bandit feedback," in *Conference on Learning Theory*, pp. 41–1, JMLR Workshop and Conference Proceedings, 2012.

[9] N. Korda, L. Prashanth, and R. Munos, "Fast gradient descent for drifting least squares regression, with application to bandits," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, 2015.

[10] Q. Ding, C.-J. Hsieh, and J. Sharpnack, "An efficient algorithm for generalized linear bandit: Online stochastic gradient descent and thompson sampling," in *International Conference on Artificial Intelligence and Statistics*, pp. 1585–1593, PMLR, 2021.

[11] Y. Han, Z. Liang, Y. Wang, and J. Zhang, "Generalized linear bandits with local differential privacy," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[12] S. Ito, "An optimal algorithm for bandit convex optimization with strongly-convex and smooth loss," in *International Conference on Artificial Intelligence and Statistics*, pp. 2229–2239, PMLR, 2020.

[13] T. Lattimore, "Improved regret for zeroth-order adversarial bandit convex optimisation," *Mathematical Statistics and Learning*, vol. 2, no. 3, pp. 311–334, 2020.

[14] A. S. Suggala, P. Ravikumar, and P. Netrapalli, "Efficient bandit convex optimization: Beyond linear losses," in *Conference on Learning Theory*, pp. 4008–4067, PMLR, 2021.

[15] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM journal on control and optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[16] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[17] R. S. Sutton, A. G. Barto, *et al.*, *Introduction to reinforcement learning*, vol. 135. MIT press Cambridge, 1998.

[18] K.-S. Jun, A. Bhargava, R. Nowak, and R. Willett, "Scalable generalized linear bandits: Online computation and hashing," in *Advances in Neural Information Processing Systems*, pp. 99–109, 2017.

[19] L. Li, Y. Lu, and D. Zhou, "Provably optimal algorithms for generalized linear contextual bandits," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2071–2080, JMLR. org, 2017.