Quick Best Action Identification in Linear Bandit Problems

Jun Geng[1] and Lifeng Lai[2]
[1] School of Electrical and Information Engineering, Harbin Institue of Technology, Harbin, China
[2] Dept. of Electrical and Computer Engineering, University of California, Davis, CA
Emails: jgeng@hit.edu.cn, lflai@ucdavis.edu

***Abstract*—In this paper, we consider a best action identification problem in the stochastic linear bandit setup with a fixed confident constraint. In the considered best action identification problem, instead of minimizing the accumulative regret as done in existing works, the learner aims to obtain an accurate estimate of the underlying parameter based on his action and reward sequences. To improve the estimation efficiency, the learner is allowed to select his action based his historical information; hence the whole procedure is designed in a sequential adaptive manner. We first show that the existing algorithms designed to minimize the accumulative regret is not a consisent estimator and hence is not a good policy for our problem. We then charcaterize a lower bound on the estimation error for any policy. We further design a simple policy and show that the estimation error of the designed policy achieves the same scaling order as that of the derived lower bound.**

I. Introduction

Multi-armed bandit problem is a canonical sequential decision problem that has a wide range of applications [1]–[5]. In the classic multi-armed bandit problem, at each time slot, a decision maker has to choose one of K competing decisions or "arms", and receives a reward related to certain unknown parameters from his selected decision. Based on the knowledge collected from his past decisions and the corresponding rewards, the decision maker can then carefully decide his future actions according to different goals. The most commonly used goal is to minimize the cumulative regret, which is the cumulative difference between the optimal reward that one can achieve when the underlying parameters are known and the reward of the action taken by the decision maker. This setup nicely captures "exploration versus exploitation" phenomena in sequential decision making, as a crucial tradeoff faced by the decision maker at each round is between "exploitation", i.e. to choose the decision with the highest estimated expected rewards, and "exploration", i.e. to choose other decisions so as to obtain better estimates of the expected rewards of these decisions. Recently, another goal named "best arm identification" has received significant attentions [6]–[12]. In the best arm identification problem, instead of minimizing the cumulative regret, the goal is to identify the best arm that provides the highest expected rewards with high probability. This setup is also known as pure exploration since the decision maker now has the freedom to explore all arms without having to worry about regrets incurred in these exploration actions.

A natural generalization of the classic multi-armed bandit problem is so called stochastic linear multi-armed bandit problem [13]. In the stochastic linear multi-armed bandit problem, the decision maker chooses his decision $\mathbf{x}_t$ from an $d-$dimensional compact set D and receives a reward $< \mathbf{x}_t, \boldsymbol{\theta}^* > +\eta_t$, in which $\boldsymbol{\theta}^*$ is a fixed but unknown parameter and η_t is noise. Defining the regret as the difference between the rewards of the best decisions when $\boldsymbol{\theta}^*$ is known and the rewards of the selected decisions, existing works on the stochastic linear multi-armed bandit problem aim to minimize the total regret. For example, [13], [14] have proposed algorithms according to the optimism in the face of uncertainty (OFU) principle, and have shown the proposed algorithms are Hannan consistent.

In this paper, similar to the best arm identification problem studied in the classic multi-armed bandit setup, we consider the best action identification problem in the stochastic linear multi-armed bandit setup. More specifically, instead of aiming to minimize the cumulative regret, we aim to obtain an accurate estimation $\hat{\boldsymbol{\theta}}$ of the unknown parameter $\boldsymbol{\theta}^*$ under a fixed confidence constraint. In particular, the decision maker aims to minimize the total number of actions under the constraint that the estimation error $||\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*||_2$ is under control with a large probability. We call this best action identification problem, as the best action $\mathbf{x}_t$ should have the same direction as $\boldsymbol{\theta}^*$.

In this paper, we first show that existing algorithms based on the OFU principle lead to inconsistent estimators of $\boldsymbol{\theta}^*$ and hence are not suitable for the best action identification. Intuitively, the OFU algorithm keeps selecting the actions that are close to the current estimation $\hat{\boldsymbol{\theta}}_t$ in each round since it aims to minimize the regret. As a result, all selected actions are concentrated in a small cone around the direction of the true underlying parameter $\boldsymbol{\theta}^*$. The decision maker has to use the rewards of selected actions to estimate $\boldsymbol{\theta}^*$, but the actions with similar directions only bring similar rewards. In other words, it is challenging for the decision

The work of J. Geng is supported by the National Natural Science Foundation of China under grant 61601144. The work of L. Lai is supported by the National Science Foundation under grants ECCS-1711468 and CCF-1717943.

maker to tell whether the change of rewards is caused by the different action selection or by the random noise. Hence it is difficult to identify which action is better. Motivated by this intuitive explanation, we propose a scheme that selects actions that are orthogonal to the direction of $\hat{\boldsymbol{\theta}}_t$. We show that the rewards from these different directions are effective in identifying the best action. In particular, we show that the proposed algorithm leads to a consistent estimator of $\boldsymbol{\theta}^*$. Furthermore, we calculate a lower bound of the estimation error of any policy, and further show that the performance of our proposed algorithm achieves this lower bound up to a constant factor.

The remainder of the paper is organized as follows. The mathematical model is given in Section II. In Section III, the limitation of OFU based algorithms is discussed. We further propose a new algorithm and analyze its performance. Section IV provides a numerical example to illustrate the conclusion obtained in this paper. Section V offers some concluding remarks. Due to space limitations, we present main ideas and conclusions without detailed proof.

Notations: $||\mathbf{x}||_p$ denotes the $p-$norm of a vector $\mathbf{x} \in \mathbb{R}^d$. For a positive definite matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, the weighted norm of a vector $\mathbf{x}$ is denoted as $||\mathbf{x}||_{\mathbf{A}} = \sqrt{\mathbf{x}^T\mathbf{A}\mathbf{x}}$, and the weighted inner product of two vectors $\mathbf{x}, \mathbf{y}$ is denoted as $< \mathbf{x}, \mathbf{y} >_{\mathbf{A}} = \mathbf{x}^T\mathbf{A}\mathbf{y}$. $\lambda_{\max}(\mathbf{A})$, $\lambda_{\min}(\mathbf{A})$, det$(\mathbf{A})$ and $\mathrm{trace}(\mathbf{A})$ denote the maximum eigenvalue, the minimum eigenvalue, the determinant and the trace of matrix $\mathbf{A}$, respectively.

II. Problem Formulation

In this paper, we consider the stochastic linear bandit problem which proceeds in rounds $t = 1, 2, \ldots$. In each round t, the decision maker chooses a decision $\mathbf{x}_t$ from a compact decision set $D_t \subset \mathbb{R}^d$, and subsequently obtains a reward

$$y_t =< \mathbf{x}_t, \boldsymbol{\theta}^* > +\eta_t, \tag{1}$$

in which $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is a fixed but unknown parameter with finite l_2-norm $||\boldsymbol{\theta}^*||_2 \leq S$, and η_t is a centered sub-Gaussian random variable with variance proxy σ^2. $\{\eta_t, t = 1, 2, \ldots\}$ is assumed to be a sequence of independently and identically distributed (i.i.d.) random variables.

Let $\mathcal{F}_t = \sigma\{\eta_1, \eta_2, \ldots, \eta_t\}$ be the sigma field at time t. The decision maker is allowed to choose his decision adaptively based on his historical information. Mathematically, $\mathbf{x}_t$ can be expressed as

$$\mathbf{x}_t = f_t(\mathbf{x}_1, y_1, \ldots, \mathbf{x}_{t-1}, y_{t-1}),$$

in which $f_t(\cdot)$ is some $\mathcal{F}_{t-1}$ measurable function. To simplify the derivation, we assume that the decision set $D_t = \{\mathbf{x} \in \mathbb{R}^d : ||\mathbf{x}||_2^2 \leq 1\}$, which is a fixed set over time. Hence in the remainder of this paper, we also denote the decision set as D.

We express the relationship between decisions and corresponding rewards in the matrix form as

$$\mathbf{Y}_t = \mathbf{X}_t\boldsymbol{\theta}^* + \boldsymbol{\eta}_t, \tag{2}$$

in which $\mathbf{Y}_t = [y_1, y_2, \ldots, y_t]^T$, $\boldsymbol{\eta}_t = [\eta_1, \eta_2, \ldots, \eta_t]^T$ and $\mathbf{X}_t = [\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_t^T]^T \in \mathbb{R}^{t\times d}$. Denote $\hat{\boldsymbol{\theta}}_t$ as the estimate of $\boldsymbol{\theta}^*$ at time t. The decision maker aims to design an efficient algorithm to select decisions $\mathbf{X}_t$ and accurately estimate the unknown parameter $\boldsymbol{\theta}^*$ based on his sequential information $\{\mathbf{x}_1, \ldots, \mathbf{x}_t, y_1, \ldots, y_t\}$. The performance metric is specified as

$$P(||\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*||_2^2 \leq \epsilon) \geq 1 - \delta \tag{3}$$

for some given constant $\epsilon > 0$ and $\delta \in (0, 1)$. That is, the decision maker should have strong confidence on the result that the estimation error is less than a small value ϵ when the decision procedure is terminated. Since $\{\eta_t\}$ is a sequence of sub-Gaussian random variable, we expect that ϵ converges to zero and δ decays exponentially with respect to t as $t \to \infty$.

III. Algorithms and Performances

A natural estimator for (2) is the ordinary least squares estimator

$$\hat{\boldsymbol{\theta}}_t = (\mathbf{X}_t^T\mathbf{X}_t)^{-1}\mathbf{X}_t^T\mathbf{Y}_t. \tag{4}$$

One difficulty with the above estimator is that $\mathbf{X}_t^T\mathbf{X}_t$ is not invertible when its rank is deficient (e.g. $t \leq d$). In this paper we focus on the following class of estimators that are slight modification of (4)

$$\hat{\boldsymbol{\theta}}_t = (\mathbf{X}_t^T\mathbf{X}_t + \mathbf{W}_0)^{-1}\mathbf{X}_t^T\mathbf{Y}_t, \tag{5}$$

in which $\mathbf{W}_0$ is a positive definite matrix. This class of estimators are widely used in the regret minimizaiton problems [13], [14]. For notation convenience, we define

$$\mathbf{W}_t := \mathbf{W}_0 + \mathbf{X}_t^T\mathbf{X}_t. \tag{6}$$

It is easy to see that $\mathbf{W}_t$ is always positive definite; hence the inversion in (5) is always valid. We further note that $\mathbf{W}_t$ can be efficiently calculated using the recursive formula $\mathbf{W}_t = \mathbf{W}_{t-1} + \mathbf{x}_t\mathbf{x}_t^T$.

A. Challenges of Existing Algorithms

The most well known algorithm for the stochastic linear bandit problem is designed according to the *optimism in the face of uncertainty principle* [14]. The basic idea of this algorithm is to use observations to construct a confidence set $C_t \subset \mathbb{R}^d$ that contains the unknown parameter $\boldsymbol{\theta}^*$ with a high probability. The confidence set C_t is updated whenever the decision maker obtains a new reward y_t. The algorithm then estimates the unknown parameter by $\hat{\boldsymbol{\theta}}_t = \mathrm{argmax}_{\boldsymbol{\theta}\in C_t}(\max_{\mathbf{x}\in D_t} < \mathbf{x}, \boldsymbol{\theta} >)$ and selects the next decision by solving $\mathbf{x}_t = \mathrm{argmax}_{\mathbf{x}\in D_t} < \mathbf{x}, \hat{\boldsymbol{\theta}}_t >$.

In our context, for $t = 1, 2, \ldots$, the algorithm designed according to the OFU principle can be expressed as:

$$\hat{\boldsymbol{\theta}}_t = (\mathbf{X}_t^T \mathbf{X}_t + \mathbf{W}_0)^{-1} \mathbf{X}_t^T \mathbf{Y}_t, \tag{7}$$

$$C_t = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : ||\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}||_{\mathbf{W}_t} \leq \beta_t \right\}, \tag{8}$$

$$\mathbf{x}_t = \text{argmax}_{\mathbf{x} \in D_t} < \mathbf{x}, \hat{\boldsymbol{\theta}}_t >= \hat{\boldsymbol{\theta}}_t / ||\hat{\boldsymbol{\theta}}_t||_2^2. \tag{9}$$

We note that the confidence region C_t is an ellipsoid with radius β_t. The value of β_t is updated at every time slot according to newly obtained information.

Several existing works [13], [14] have shown that, if β_t is properly designed, the above algorithm has a small cumulative regret. Particularly, let $\mathbf{x}_t^* = \text{argmax}_{\mathbf{x} \in D_t} < \mathbf{x}, \boldsymbol{\theta}^* >$ be the best decision for $\boldsymbol{\theta}^*$, let $r_t =< \mathbf{x}_t^*, \boldsymbol{\theta}^* > - < \mathbf{x}_t, \boldsymbol{\theta}^* >$ be the regret at time t for taking decision $\mathbf{x}_t$ and let $R_n = \sum_{t=1}^n r_t$ be the cumulative regret. [14] proved the following result.

Theorem 1: (Theorem 2 and Theorem 3 in [14]) Let $\mathbf{W}_0 = \kappa \mathbf{I}$, $\kappa > 0$. By setting

$$\beta_t = \sigma^2 \sqrt{2 \log(\det(\mathbf{W}_t)^{1/2} \det(\lambda \mathbf{I})^{-1/2} / \delta)} + \kappa^{1/2} S,$$

then for any $\delta > 0$, with probability at least $1 - \delta$, $\boldsymbol{\theta}^*$ lies in the set C_t. Further more, if for all t and all $\mathbf{x} \in D_t$, $< \mathbf{x}, \boldsymbol{\theta}^* >\in [-1, 1]$, then with probability at least $1 - \delta$, the cumulative regret satisfies

$$\begin{aligned} \forall n \geq 0, \quad R_n \leq 4 \sqrt{nd \log(\kappa + n/d)} \\ (\kappa^{1/2} S + \sigma^2 \sqrt{2 \log(1/\delta) + d \log(1 + n/(\kappa d))}). \end{aligned}$$

Theorem 1 indicates that the OFU algorithm is Hannan consistent, i.e., $\lim_{n \to \infty} R_n / n = 0$. However, in the following, we point out that the OFU algorithm leads to an inconsistent estimate of $\boldsymbol{\theta}^*$. The result is stated in the following theorem.

Theorem 2: If $R_n / n \to 0$ as $n \to \infty$ with probability at least $1 - \delta$, then

$$\lim_{t \to \infty} P \left(||\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*||_2^2 \geq \sigma^2 \right) \geq 1 - \delta. \tag{10}$$

Define event $\mathcal{E} := \left\{ ||\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*||_2^2 \geq \sigma^2 \right\}$. Theorem 2 implies that

$$\mathbb{E}[||\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*||_2^2] \geq \mathbb{E}[||\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*||_2^2 | \mathcal{E}] P(\mathcal{E}) \geq \sigma^2 (1 - \delta).$$

That is, the estimation error of the OFU algorithm does not vanish as the sample size goes to infinity.

In the following, we provide an intuitive explanation of the reason why OFU algorithms lead to inconsistent estimators. The examination of this also provides motivation for the proposed algorithm to be discussed below. Considering the case with $d = 2$, the reward $y_t =< \mathbf{x}_t, \boldsymbol{\theta}^* > + \eta_t = ||\mathbf{x}_t|| ||\boldsymbol{\theta}^*|| \cos \psi + \eta_t$, in which ψ is the angle between $\mathbf{x}_t$ and $\boldsymbol{\theta}^*$. As illustrated in the upper-right subfigure in Fig. 1, the solid cosine curve is $< \mathbf{x}_t, \boldsymbol{\theta}^* >$, and the region bounded by the two dash cosine lines characterizes the possible region of the rewards y_t. In OFU algorithms, the decision maker takes action $\mathbf{x}_t = \hat{\boldsymbol{\theta}}_{t-1}$. As $\hat{\boldsymbol{\theta}}_{t-1}$ and $\boldsymbol{\theta}^*$ are generally close, the regret is small and ψ is close to zero. In this case, an obtained feedback reward y_t leads to a wide possible range for $\boldsymbol{\theta}^*$. That is, any value of ψ in the red region of the upper-right subfigure in Fig. 1 could lead to the same reward y_t. In the regret minimization, this is unavoidable, as we need to select $\mathbf{x}_t$ that has small angle with $\boldsymbol{\theta}^*$. In our problem setup, as the regret is not of primary concern, we can avoid this by selecting $\mathbf{x}_t$ to be orthogonal to $\hat{\boldsymbol{\theta}}_{t-1}$ (and hence has large angle with $\boldsymbol{\theta}^*$). These actions are helpful in improving the estimation accuracy of ψ as their rewards are close to the zero-crossing region of the cosine curve, which infers a much narrower possible region for $\boldsymbol{\theta}^*$. The proposed algorithm to be discussed below is motivated by this observation.

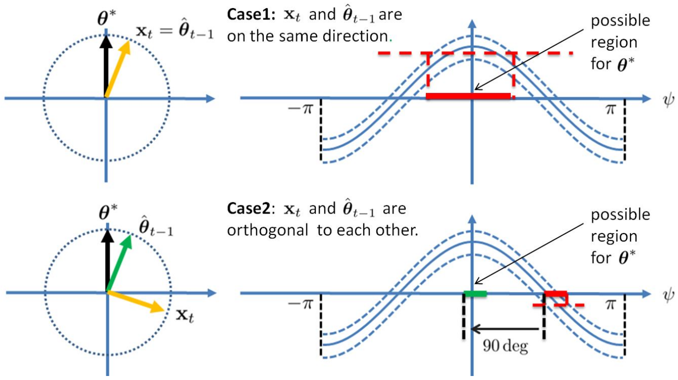
$\boldsymbol{\theta}^*$
$\mathbf{x}_t = \hat{\boldsymbol{\theta}}_{t-1}$
Case1: $\mathbf{x}_t$ and $\hat{\boldsymbol{\theta}}_{t-1}$ are on the same direction.
possible region for $\boldsymbol{\theta}^*$
$-\pi$
π
ψ
$\boldsymbol{\theta}^*$
$\hat{\boldsymbol{\theta}}_{t-1}$
$\mathbf{x}_t$
Case2: $\mathbf{x}_t$ and $\hat{\boldsymbol{\theta}}_{t-1}$ are orthogonal to each other.
possible region for $\boldsymbol{\theta}^*$
$-\pi$
π
ψ
90 deg

Fig. 1: An illustration of the difference between the OFU algorithm and the proposed algorithm. Upper figures: the case for OFU algorithm. Lower figures: the case for proposed algorithm.

B. Proposed Algorithm and Performance Analysis

Motivated by the discussion above, we propose a novel algorithm which leads to a consistent estimator with a fast convergence rate. The proposed algorithm is specified in Algorithm 1. To facilitate the presentation, for $k = 1, 2, \ldots$, we use the following notations in Algorithm 1:

$$\mathbf{X}_{k,d} = \left[\mathbf{x}_{(k-1)d+1}^T, \mathbf{x}_{(k-1)d+2}^T, \cdots, \mathbf{x}_{kd}^T \right]^T,$$

$$\mathbf{Y}_{k,d} = [y_{(k-1)d+1}, y_{(k-1)d+2}, \ldots, y_{kd}]^T,$$

$$\boldsymbol{\eta}_{k,d} = [\eta_{(k-1)d+1}, \eta_{(k-1)d+2}, \ldots, \eta_{kd}]^T.$$

The proposed algorithm adopts batch processing. In particular, the proposed algorithm initializes the first d decisions as a group of standard orthogonal basis. The decision maker updates the estimate $\hat{\boldsymbol{\theta}}_t$ whenever he collects d successive rewards. Furthermore, whenever a new estimate $\hat{\boldsymbol{\theta}}_t$ is calculated, the decision maker chooses next decision $\mathbf{x}_{t+1}$ as the direction of $\hat{\boldsymbol{\theta}}_t$, and selects another $d - 1$ decisions such that these d decisions form another group of orthogonal basis. We emphasize that algorithms according to the OFU principle keep taking decisions that maximize the reward $< \mathbf{x}, \hat{\boldsymbol{\theta}}_t >$. In our context, the OFU algorithm will always select the decision with the same direction of $\hat{\boldsymbol{\theta}}_t$. However, in our proposed algorithm, among every successive d decisions, only one decision is on the direction of $\hat{\boldsymbol{\theta}}_t$; the rest of $d-1$ decisions are orthogonal to $\hat{\boldsymbol{\theta}}_t$. This is the key difference between the OFU algorithm and our algorithm.

```
Data: the adaptively designed decisions $\mathbf{x}_1, \dots, \mathbf{x}_t$ and
      corresponding rewards $y_1, \dots, y_t$
Result: the estimate $\hat{\boldsymbol{\theta}}_t$
Initialization: select $\mathbf{x}_1, \dots, \mathbf{x}_d$ as a set of standard
orthogonal basis ;
for $k = 1, 2, \dots \lceil t/d \rceil$ do
    obtain rewards: $\mathbf{Y}_{k,d} = \mathbf{X}_{k,d}\boldsymbol{\theta}^* + \boldsymbol{\eta}_{k,d}$ ;
    update matrix: $\mathbf{W}_{kd} = \mathbf{W}_{(k-1)d} + \mathbf{X}_{k,d}^T\mathbf{X}_{k,d}$ ;
    estimate parameter: $\hat{\boldsymbol{\theta}}_{kd} = \mathbf{W}_{kd}^{-1}\mathbf{X}_{kd}^T\mathbf{Y}_{kd}$ ;
    choose decision: $\mathbf{x}_{kd+1} = \hat{\boldsymbol{\theta}}_{kd}/||\hat{\boldsymbol{\theta}}_{kd}||_2^2$, select
    $\{\mathbf{x}_{kd+1}, \mathbf{x}_{kd+2}, \dots, \mathbf{x}_{(k+1)d}\}$ to be an orthogonal
    basis;
end
```

Algorithm 1: The Proposed Algorithm

The performance of the proposed algorithm is characterized in the following theorem.

Theorem 3: For the proposed algorithm, we have

$$\mathbb{E}[||\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*||_2^2] \leq \frac{d^2}{t}\sigma^2(1 + o(1)).$$

Furthermore, if $\boldsymbol{\eta}_t$ is a sub-Gaussian vector, then

$$P\left(||\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*||_2^2 \geq \frac{3\sigma^2 d^{3/2}}{t^{1/2}} + O\left(\frac{\sigma^2 d^2}{t}\right)\right) \leq e^{-t}$$

Theorem 3 characterizes our performance metric (3). In particular, δ decays exponentially as $t \to \infty$, and the bound of estimation error ϵ shrinks to zero on the order $O(t^{-1/2})$ for the proposed algorithm.

We now provide a lower bound of the mean square estimation error (MSE) for all possible sequential decision selection strategies and show that MSE reduces at most on order $O(t^{-1})$.

Theorem 4: (Lower Bounds on MSE) Let η_t be a sub-Gaussian random variable with variance proxy σ^2. If estimator (5) is adopted, then for any adaptively selected decision sequence $\{\mathbf{x}_i, i = 1, 2, \dots, t\}$, we have

$$\mathbb{E}[||\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*||_2^2] \geq \frac{1}{t}\sigma^2 + o\left(\frac{1}{t}\right). \tag{11}$$

Theorem 3 indicates that the convergence rate of MSE for the proposed algorithm is on order $O(t^{-1})$, while Theorem 4 shows that the convergence rate of MSE cannot be faster than $O(t^{-1})$. Hence, the proposed algorithm is order optimal.

IV. Numerical Simulation

In this section, we provide a numerical example to illustrate the results obtained in this paper. In this numerical example, we set $d = 5$, and compare the performance of the OFU algorithm and our proposed algorithm. In particular, the MSE of each algorithm is calculated by Monte Carlo method. In the simulation, the estimation procedure proceeds 3000 rounds; hence, for each trial, the decision maker has to adaptively make 3000 decisions. For each algorithm, we conduct 10^5 trials with randomly created underlying parameter $\boldsymbol{\theta}^*$, and we record the estimation error at each round of decision. Then, the logarithm of MSE, which is estimated by the average of estimation error at each trial, at each decision round is illustrated in Figure 2.

In Figure 2, The blue solid line is the performance of the OFU algorithm and the red dash line is the performance of the proposed algorithm. The simulation result shows that the error of the OFU algorithm tends to be a constant as the number of decisions goes large; hence, the corresponding MSE also tends to a constant. However, the error of the proposed algorithm decays when the number of decisions grows, which indicates the estimation error tends to zero as the number of decisions goes to infinity. Hence, the proposed estimator is consistent.

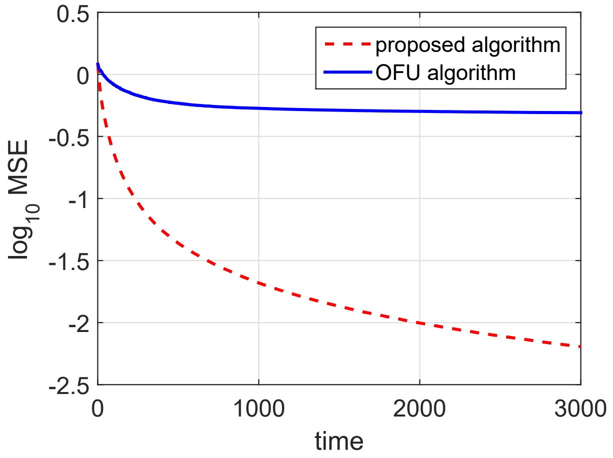

Fig. 2: Estimation error vs. the total number of decisions

V. Conclusion

In this paper, we have studied the problem of identifying the best action in the stochastic linear bandit setup with a fixed confidence constraint. We have shown that the existing OFU algorithm is an inconsistent estimator for the unknown parameter $\boldsymbol{\theta}^*$. We have proposed and analyzed a novel algorithm. We have shown that the proposed algorithm is consistent and that its mean square estimation error reduces on order $O(t^{-1})$. Furthermore, we have shown that the probability that the estimation error is larger than $t^{-1/2}$ decays exponentially with respect to t.

We note that this paper has considered the asymptotic case with $t \to \infty$. In the future, it will be of interest to consider the problem when a finite number of decisions are made. In this case, we expect that tools from optimal stopping [15] and controlled sensing [16] will be useful.

References

[1] W. H. Press, "Bandit solutions provide unified ethical models for randomized clinical trials and comparative effecctiveness research," *Proceedings of the National Academy of Sciences*, vol. 106, p-p. 22387–22392, Dec. 2009.

[2] B. Awerbuch and P. Kleinberg, "Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches," in *Proc. Annual ACM Symp. Theory of Computing*, (Chicago, IL), pp. 45–53, June 2004.

[3] I. Manickam, A. S. Lan, and R. G. Baraniuk, "Contextual multi-armed bandit algorithms for personalized learning action selection," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, (New Orleans, LA), pp. 6344–6348, June 2017.

[4] P. Reverdy and V. Srivastava, "Multi-armed bandits for human-machine decision making," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, (Calagry, CA), pp. 6986–6990, Apr. 2018.

[5] L. Lai, H. E. Gamal, H. Jiang, and H. V. Poor, "Cognitive medium access: Exploration, exploitation, and competition," *IEEE Transactions on Mobile Computing*, vol. 10, pp. 239–253, Feb 2011.

[6] V. Gabillon, M. Ghavamzadeh, and A. Lazaric, "Best arm identification: A unified approach to fixed budget and fixed confidence," in *Advances in Neural Information Processing Systems*, (Lake Tahoe, USA), pp. 1–16, Dec. 2012.

[7] K. Jamieson and R. Nowak, "Best-arm indentification algorithms for multi-armed bandits in the fixed confidence setting," in *Proc. Conf. on Information Science and Systems*, (Princeton, NJ), pp. 1–6, Mar. 2014.

[8] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck, "lil'ucb: An optimal exploration algorithm for multi-armed bandits," *Journal of Machine Learning Research*, vol. 35, pp. 1–17, Dec. 2014.

[9] Z. Karnin, T. Koren, and S. Somekh, "Almost optimal exploration in multi-armed bandits," in *Proc. Intl. Conf. on Machine Learning*, June 2013.

[10] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck, "On finding the largest mean among many," *arXiv preprint arXiv:1306.3917*, 2013.

[11] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research*, vol. 3, pp. 397–422, 2002.

[12] S. Kalyanakrishnan, A. Tewari, and P. Stone, "PAC subset selection in stochastic multi-armed bandits," in *Proc. Intl. Conf. on Machine Learning*, pp. 655–662, June 2012.

[13] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *Proc. Annual Conference on Learning Theory*, pp. 355–366, 2008.

[14] Y. Abbasi-Yadkori, D. Pal, and C. Szepesvari, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems*, (Granada, Spain), pp. 1–19, Dec. 2011.

[15] H. V. Poor and O. Hadjiliadis, *Quickest Detection*. Cambridge, UK: Cambridge University Press, 2008.

[16] S. Nitinawarat, G. K. Atia, and V. V. Veeravalli, "Controlled sensing for multihypothesis testing," *IEEE Trans. Automatic Control*, vol. 58, pp. 2451– 2464, Oct. 2013.