Perturbed-History Exploration in Stochastic Linear Bandits

Branislav Kveton
Google Research

Csaba Szepesvári
DeepMind

Mohammad Ghavamzadeh
Facebook AI Research

Craig Boutilier
Google Research

Abstract

We propose a new online algorithm for cumulative regret minimization in a stochastic linear bandit. The algorithm pulls the arm with the highest estimated reward in a linear model trained on its perturbed history. Therefore, we call it *perturbed-history exploration in a linear bandit (*`LinPHE`*)*. The *perturbed history* is a mixture of observed rewards and randomly generated *i.i.d. pseudo-rewards*. We derive a $\tilde{O}(d\sqrt{n})$ gap-free bound on the n-round regret of `LinPHE`, where d is the number of features. The key steps in our analysis are new concentration and anti-concentration bounds on the weighted sum of Bernoulli random variables. To show the generality of our design, we generalize `LinPHE` to a logistic model. We evaluate our algorithms empirically and show that they are practical.

1 INTRODUCTION

A *multi-armed bandit* [19, 4, 20] is an online learning problem where the *learning agent* acts by pulling *arms*. After the arm is *pulled*, the agent receives its *stochastic reward*. The objective of the agent is to maximize its expected cumulative reward. The agent does not know the mean rewards of the arms in advance and faces the so-called *exploration-exploitation dilemma*: *explore*, and learn about arms; or *exploit*, and pull the arm with the highest estimated reward thus far. This model captures many practical applications. In a clinical trial, for example, the *arm* may be a treatment and its *reward* is the outcome of that treatment on some patient population.

A *stochastic linear bandit* [7, 28, 1] is a generalization of the multi-armed bandit to the setting where each arm is associated with a feature vector. The mean reward of an arm is the dot product of its feature vector and an unknown parameter vector, which is shared by all arms. In our clinical example, the feature vector may be a vector of treatment indicators and the parameter vector may be the effects of individual treatments.

The most popular exploration strategies in stochastic bandits, *optimism in the face of uncertainty (OFU)* [4] and *Thompson sampling (TS)* [32, 2, 29], are relatively well understood in linear bandits [7, 1, 3, 20]. Unfortunately, these designs and their guarantees do not extend easily to complex problems. For instance, in generalized linear bandits [11], all OFU algorithms use *approximate* high-probability confidence sets, which are loose and statistically suboptimal [11, 21, 14]. Also the posterior distribution of model parameters does not have a closed form. Therefore, posterior sampling in TS has to be *approximated*. Posterior approximations are computationally costly in general [12, 16, 24, 27, 22, 23].

In this work, we study a simple exploration strategy that can be easily generalized to complex problems. The key idea is to *explore by perturbing* the training data of a reward generalization model, which is fit by an existing offline oracle. Specifically, the model is fit to a mixture of *history*, features of the pulled arms with their realized rewards; and *pseudo-history*, features of the pulled arms with randomly generated *i.i.d. pseudo-rewards*. In *perturbed-history exploration (PHE)*, the agent pulls the arm with the highest reward in its estimated model and then updates its history with the observed reward.

The key to the generality and optimism in PHE are the pseudo-rewards. They are drawn from the same family of distributions as the actual rewards, and thus we can reuse existing methods for fitting the reward generalization model. They are also *maximum variance randomized data*, which induce suitable exploration. We show that appropriate randomization, not necessarily by posterior sampling, can lead to practical exploration in structured problems.

We make the following contributions in this paper. First, we propose LinPHE, a linear bandit algorithm that estimates the mean rewards of arms using PHE. Second, we prove a $\tilde{O}(d\sqrt{n})$ gap-free bound on the n-round regret of LinPHE, where d is the number of features. Our analysis relies on novel concentration and anti-concentration bounds on the weighted sum of Bernoulli random variables. Third, we propose a generalization of LinPHE to a logistic model and call it LogPHE. Finally, we evaluate both algorithms empirically. They are competitive with Thompson sampling, although they are derived based on different insights.

2 SETTING

We adopt the following notation. The set $\{1, \ldots, n\}$ is denoted by $[n]$. All vectors are column vectors. The minimum and maximum eigenvalues of matrix M are denoted by $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$, respectively. We define $\mathrm{Ber}(x; p) = p^x(1-p)^{1-x}$ and let $\mathrm{Ber}(p)$ be the corresponding Bernoulli distribution. We also define $B(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$ and let $B(n, p)$ be the corresponding binomial distribution. For any event E, $\mathbb{1}\{E\} = 1$ if event E occurs and $\mathbb{1}\{E\} = 0$ otherwise. We denote a $d \times d$ identity matrix by I_d. We use $\tilde{O}$ for the big-O notation up to logarithmic factors.

A *stochastic linear bandit* [7, 28, 1] is an online learning problem where the learning agent sequentially pulls arms, each of which is associated with a feature vector. Let K be the number of arms, $x_i \in \mathbb{R}^d$ be the *feature vector* of arm $i \in [K]$, and $\theta_* \in \mathbb{R}^d$ be an unknown *parameter vector*. The *reward* of arm i in round $t \in [n]$, $Y_{i,t}$, is drawn i.i.d. from a distribution of that arm with mean $\mu_i = x_i^\top \theta_*$. The learning agent acts as follows. In round t, it *pulls* arm $I_t \in [K]$ and receives reward $Y_{I_t,t}$. The agent aims to maximize its *expected cumulative reward* in n rounds. To simplify exposition, we denote by $X_t = x_{I_t}$ and $Y_t = Y_{I_t,t}$ the feature vector of the pulled arm in round t and its reward, respectively.

Without loss of generality, we assume that arm 1 is *optimal*, that is $\mu_1 > \max_{i>1} \mu_i$. Let $\Delta_i = \mu_1 - \mu_i$ denote the *gap* of arm i. Maximization of the expected cumulative reward in n rounds is equivalent to minimizing the *expected n-round regret*,

$$R(n) = \sum_{i=2}^{K} \Delta_i \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}\{I_t = i\}\right].$$

We make several additional assumptions. First, rewards are bounded in $[0, 1]$, that is $Y_{i,t} \in [0, 1]$ for any arm i and round t. This assumption is standard. Second, the last feature is a *bias term*, $x_i(d) = 1$ for all arms i. This is without loss of generality, since such a feature can be always added. Finally, the feature vectors of the last d arms are a basis in $\mathbb{R}^d$. This is without loss of generality, since the arms can be always reordered to satisfy this.

Algorithm 1 Perturbed-history exploration in a linear bandit (LinPHE) with $[0, 1]$ rewards.

1: **Inputs**:
2: Integer perturbation scale $a > 0$
3: Regularization parameter $\lambda > 0$

4: **for** $t = 1, \ldots, n$ **do**
5: **if** $t > d$ **then**
6: Generate $(Z_{j,\ell})_{j \in [a],\, \ell \in [t-1]} \sim \mathrm{Ber}(1/2)$
7: $G_t \leftarrow (a+1) \sum_{\ell=1}^{t-1} X_\ell X_\ell^\top + \lambda(a+1) I_d$
8: $\tilde{\theta}_t \leftarrow G_t^{-1} \sum_{\ell=1}^{t-1} X_\ell \left[Y_\ell + \sum_{j=1}^{a} Z_{j,\ell} \right]$
9: $I_t \leftarrow \arg\max_{i \in [K]} x_i^\top \tilde{\theta}_t$
10: **else**
11: $I_t \leftarrow K - t + 1$
12: Pull arm I_t and get reward $Y_{I_t,t}$
13: $X_t \leftarrow x_{I_t}$, $Y_t \leftarrow Y_{I_t,t}$

3 PERTURBED-HISTORY EXPLORATION

Now we introduce *perturbed-history exploration (PHE)*. Our algorithm, *perturbed-history exploration in a linear bandit* (LinPHE), is presented in Algorithm 1. In round t, LinPHE fits a linear model to its *perturbed history* up to round t (line 8),

$$\tilde{\theta}_t = G_t^{-1} \sum_{\ell=1}^{t-1} X_\ell \left[Y_\ell + \sum_{j=1}^{a} Z_{j,\ell} \right], \tag{1}$$

where

$$G_t = (a+1) \sum_{\ell=1}^{t-1} X_\ell X_\ell^\top + \lambda(a+1) I_d \tag{2}$$

is the *sample covariance matrix* up to round t, $a > 0$ is a tunable integer parameter, $\lambda > 0$ is the regularization parameter, and $(Z_{j,\ell})_{j \in [a],\, \ell \in [t-1]}$ are *i.i.d. pseudo-rewards*, which are freshly sampled in each round. Our model can be viewed as follows. If $Z_{j,\ell}$ were omitted in (1) and $a + 1$ was omitted in (2), we would get a regularized least-squares regression on rewards up to round t. Thus, $\tilde{\theta}_t$ is a regularized least-squares solution on the past $t - 1$ rewards and $a(t-1)$ *i.i.d. pseudo-rewards*.

LinPHE pulls the arm with the highest estimated reward under $\tilde{\theta}_t$ (line 9). Any tie-breaking rule can be used as

needed. LinPHE is initialized by pulling each arm in the basis once (line 11). This guarantees that LinPHE is sufficiently optimistic about any optimal arm (Lemma 7).

LinPHE has two tunable parameters. The *perturbation scale* a dictates the number of pseudo-rewards for each observed reward in the perturbed history. Therefore, it trades off exploration and exploitation, with higher values of a leading to more exploration. We argue informally in Section 3.1 that any $a > 1$ is sufficient for sublinear regret. The formal regret analysis is in Section 4. The *regularization parameter* $\lambda > 0$ ensures that G_t can be inverted and makes LinPHE stable. Regularization is used frequently in linear bandit analyses [1, 3].

3.1 Informal Justification

Before we analyze LinPHE in Section 4, we informally explain how exploration arises in it. To do this, we introduce two least-squares solutions that are closely related to $\tilde{\theta}_t$ in (1). In the first, the pseudo-rewards are replaced by their means,

$$\bar{\theta}_t = G_t^{-1} \sum_{\ell=1}^{t-1} X_\ell \left[Y_\ell + \sum_{j=1}^{a} \bar{Z}_{j,\ell} \right], \tag{3}$$

where $\bar{Z}_{j,\ell} = \mathbb{E}[Z_{j,\ell}] = 1/2$. In the second, both the rewards and pseudo-rewards are the so-replaced,

$$\bar{\bar{\theta}}_t = G_t^{-1} \sum_{\ell=1}^{t-1} X_\ell \left[X_\ell^\top \theta_* + \sum_{j=1}^{a} \bar{Z}_{j,\ell} \right].$$

Let $\mathcal{H} = (I_1, \ldots, I_{t-1})$ be a sequence of pulled arms in the first $t-1$ rounds.

The solution $\tilde{\theta}_t$ has two important properties that allow us to bound the regret of LinPHE. First, it concentrates at $\bar{\bar{\theta}}_t$ given history $\mathcal{H}$, since $\bar{\bar{\theta}}_t$ solves a noiseless variant of the least-squares problem solved by $\tilde{\theta}_t$. Furthermore, $\bar{\bar{\theta}}_t \to \theta'$ as regularization vanishes, where θ' are scaled and shifted parameters of the original problem. That is, $x_i^\top \theta' = (\mu_i + a/2)/(a+1)$ for all arms i.

Second, from the definitions of $\tilde{\theta}_t$, $\bar{\theta}_t$, and $\bar{\bar{\theta}}_t$, we have

$$x_i^\top \bar{\bar{\theta}}_t - x_i^\top \bar{\theta}_t = x_i^\top G_t^{-1} \sum_{\ell=1}^{t-1} X_\ell W_\ell \,,$$
$$x_i^\top \tilde{\theta}_t - x_i^\top \bar{\theta}_t = x_i^\top G_t^{-1} \sum_{\ell=1}^{t-1} X_\ell \sum_{j=1}^{a} (Z_{j,\ell} - \bar{Z}_{j,\ell}) \,,$$

where $W_\ell = X_\ell^\top \theta_* - Y_\ell$ is the "noise" in the reward in round ℓ. The first term is the deviation in the estimated reward of arm i due to reward randomness. The second term represents the deviation in the estimated reward of arm i due to pseudo-reward randomness.

Fix history $\mathcal{H}$ and let $(Y_\ell)_{\ell=1}^{t-1}$ be conditionally independent given $\mathcal{H}$. Then

$$\operatorname{var}\left[x_i^\top \bar{\bar{\theta}}_t - x_i^\top \bar{\theta}_t \,\middle|\, \mathcal{H} \right] < \operatorname{var}\left[x_i^\top \tilde{\theta}_t - x_i^\top \bar{\theta}_t \,\middle|\, \mathcal{H} \right]$$

for $a > 1$, because $x_i^\top \bar{\bar{\theta}}_t - x_i^\top \bar{\theta}_t \mid \mathcal{H}$ is a weighted sum of $t-1$ i.i.d. reward deviations and $x_i^\top \tilde{\theta}_t - x_i^\top \bar{\theta}_t \mid \mathcal{H}$ is a weighted sum, with the same weights, of $a(t-1)$ i.i.d. maximum-variance deviations on $[0, 1]$.

If both $x_i^\top \bar{\bar{\theta}}_t - x_i^\top \bar{\theta}_t$ and $x_i^\top \tilde{\theta}_t - x_i^\top \bar{\theta}_t$ were normally distributed, this would imply that for any $\varepsilon > 0$,

$$\begin{aligned} &\mathbb{P}\left(x_i^\top \bar{\bar{\theta}}_t - x_i^\top \bar{\theta}_t = \varepsilon \,\middle|\, \mathcal{H} \right) \\ &\quad \le \mathbb{P}\left(x_i^\top \bar{\bar{\theta}}_t - x_i^\top \bar{\theta}_t \ge \varepsilon \,\middle|\, \mathcal{H} \right) \\ &\quad < \mathbb{P}\left(x_i^\top \tilde{\theta}_t - x_i^\top \bar{\theta}_t \ge \varepsilon \,\middle|\, \mathcal{H} \right) , \end{aligned}$$

where the first inequality holds trivially. That is, for any potentially harmful deviation $\varepsilon > 0$ in the estimated reward of arm i, $x_i^\top \tilde{\theta}_t$ overestimates the perturbed mean reward with a higher probability than the probability of that deviation. This *optimism* induces exploration and is the key feature of LinPHE.

The idea of offsetting a fixed history of rewards by i.i.d. pseudo-rewards is very general and applies beyond the linear model in this section. In Section 3.3, we apply it to a logistic model. In Section 5, we evaluate our linear and logistic algorithms empirically.

3.2 Efficient Implementation

LinPHE can be implemented such that its expected computational cost in round t is independent of t. In particular, line (8) in LinPHE can be rewritten as

$$\tilde{\theta}_t = G_t^{-1} \sum_{i=1}^{K} x_i [V_{i,t} + U_{i,t}] \,, \tag{4}$$

where $V_{i,t}$ is the cumulative reward of arm i in the first $t-1$ rounds, $U_{i,t} \sim B(a\, T_{i,t-1}, 1/2)$ is the sum of the pseudo-rewards of arm i in round t, and $T_{i,t}$ is the number of pulls of arm i in the first t rounds. The statistics $V_{i,t}$ and G_t can be updated incrementally as

$$\begin{aligned} V_{i,t} &= V_{i,t-1} + \mathbb{1}\{I_{t-1} = i\}\, Y_{t-1} \,, \\ G_t &= G_{t-1} + (a+1) X_{t-1} X_{t-1}^\top \,, \end{aligned}$$

where we assume that $V_{i,0} = 0$ and $G_0 = \lambda(a+1) I_d$. The inverse of G_t can be also updated directly using the Sherman-Morrison formula.

The statistics $V_{i,t}$ and G_t take $O(K + d^2)$ space. If the Sherman-Morrison formula is used, the cost of updating

G_t^{-1} is $O(d^2)$. After that, the cost of computing $\tilde{\theta}_t$ in (4) is $O(Kd^2)$, if $U_{i,t}$ can be sampled in $O(1)$ time. Based on Section 4.4 of Devroye [8], $B(n,p)$ can be sampled from in $O(1)$ time in expectation for any n and p.

3.3 Algorithm LogPHE

While our formal analysis is for linear bandits, the idea of PHE is much more general. To illustrate it, we extend LinPHE to a *logistic bandit*, where the mean reward of arm i is $\mu_i = \sigma(x_i^\top \theta_*)$ and $\sigma(v) = 1/(1+\exp[-v])$ is a *sigmoid* function. The reward of arm i in round t is drawn i.i.d. from $\mathrm{Ber}(\mu_i)$.

To extend LinPHE to this class of problems, we replace $\tilde{\theta}_t$ in LinPHE with the minimizer of

$$\sum_{\ell=1}^{t-1}\left[g(X_\ell^\top\theta, Y_\ell) + \sum_{j=1}^{a} g(X_\ell^\top\theta, Z_{j,\ell})\right] + \lambda\|\theta\|_2^2\,,$$

where $g(s,y) = -y\log(\sigma(s)) - (1-y)\log(1-\sigma(s))$. For $\lambda = 0$, we obtain the maximum likelihood solution.

The above problem is convex. Also the sufficient statistics in this problem, the number of positive and negative observations of arms, can be updated incrementally as in Section 3.2. Therefore, $\tilde{\theta}_t$ in round t can be estimated in a constant time in t. We call this algorithm LogPHE and evaluate it in Section 5.2.

4 ANALYSIS

We now provide a formal analysis of LinPHE. In Section 4.1, we introduce relevant notation. In Section 4.2, we prove a generic regret bound that applies to any randomized algorithm that estimates θ_*. The regret bound of LinPHE in Section 4.3 is an instance of this result.

4.1 Notation

To simplify the analysis of LinPHE, we assume that its sample covariance matrix is not scaled by $a+1$. That is, $G_t = \sum_{\ell=1}^{t-1} X_\ell X_\ell^\top + \lambda I_d$. This does not change the behavior of LinPHE. We also assume that $\theta_* \in \mathbb{R}^d$ is a parameter vector such that $x_i^\top\theta_* = \mu_i + a/2$ holds for any arm i. Note that this transformation does not change the gaps of arms. It only shifts their mean rewards by a factor of $a/2$. Recall that arm 1 is optimal.

Let $\mathcal{F}_t = \sigma(I_1,\ldots,I_t,Y_{I_1,1},\ldots,Y_{I_t,t})$ be the σ-algebra generated by the pulled arms and their rewards by the end of round $t \in [n]\cup\{0\}$. We define $\mathcal{F}_0 = \{\emptyset,\Omega\}$, where Ω is the sample space of the probability space that holds all random variables. We denote by $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot\,|\,\mathcal{F}_{t-1})$ and $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot\,|\,\mathcal{F}_{t-1}]$ the conditional probability and expectation operators, respectively, given the past at the beginning of round t. Let $\|x\|_M = \sqrt{x^\top M x}$. Let

$$E_{1,t} = \left\{\forall i \in [K] : \left|x_i^\top\bar{\theta}_t - x_i^\top\theta_*\right| \le c_1\|x_i\|_{G_t^{-1}}\right\} \quad (5)$$

be the event that $\bar{\theta}_t$ is "close" to θ_* in round t, where $\bar{\theta}_t$ is defined in (3) and $c_1 > 0$ is tuned such that $\bar{E}_{1,t}$, the complement of $E_{1,t}$, is unlikely. Let $E_1 = \bigcap_{t=d+1}^{n} E_{1,t}$ and $\bar{E}_1$ be its complement. Let

$$E_{2,t} = \left\{\forall i \in [K] : \left|x_i^\top\tilde{\theta}_t - x_i^\top\bar{\theta}_t\right| \le c_2\|x_i\|_{G_t^{-1}}\right\} \quad (6)$$

be the event that $\tilde{\theta}_t$ is "close" to $\bar{\theta}_t$ in round t, where $\tilde{\theta}_t$ is defined in (1) and $c_2 > 0$ is tuned such that $\bar{E}_{2,t}$, the complement of $E_{2,t}$, is unlikely given any past.

4.2 General Regret Bound

In this section, we prove a general regret bound for *any* "model-based" linear bandit algorithm. The algorithm is model-based if the pulled arm in round t is chosen as in line 9 of LinPHE, where $\tilde{\theta}_t$ can be computed by any possibly randomized procedure based on past data.

Our regret bound involves three probability constants. The first constant, p_1, is an upper bound on the probability of event $\bar{E}_1$, that is $p_1 \ge \mathbb{P}(\bar{E}_1)$. The second constant, p_2, is an upper bound on the probability of event $\bar{E}_{2,t}$ given any past,

$$\mathbb{P}_t(\bar{E}_{2,t}) \le p_2\,. \quad (7)$$

The last constant, p_3, is a lower bound on the probability that the estimated reward of the optimal arm 1 is optimistic given any past,

$$\mathbb{P}_t\left(x_1^\top\tilde{\theta}_t - x_1^\top\bar{\theta}_t > c_1\|x_1\|_{G_t^{-1}}\right) \ge p_3\,. \quad (8)$$

To simplify exposition, we define $\langle x\rangle = \min\{x,1\}$. The main result of this section is the following regret bound.

Theorem 1. *Let $c_1, c_2 \ge 1$. Let A be any algorithm that pulls arm $I_t = \arg\max_{i\in[K]} x_i^\top\tilde{\theta}_t$ in round t, where $\tilde{\theta}_t$ is estimated from past data. Let the rewards be in $[0,1]$; p_1, p_2, and p_3 be defined as above; and $p_3 > p_2$. Then the expected n-round regret of A is bounded as $R(n) \le$*

$$(c_1+c_2)\left(1+\frac{2}{p_3-p_2}\right)\sqrt{c_3 n} + (p_1+p_2)n + d\,,$$

where c_3 is defined in Table 1.

Theorem 1 is extracted from prior work, where similar randomized algorithms have been analyzed [3, 33]. The proof relies on the following two lemmas.

Lemma 2. *Let $c_1, c_2 \geq 1$. Then for any round $t > d$ and history $\mathcal{F}_{t-1}$, $\mathbb{E}_t\left[\Delta_{I_t} \mathbb{1}\{E_{1,t}\}\right] \leq$*

$$(c_1 + c_2)\left(1 + \frac{2}{p_3 - p_2}\right)\mathbb{E}_t\left[\langle \|x_{I_t}\|_{G_t^{-1}}\rangle\right] + p_2\,.$$

We defer the proof of Lemma 2 to Appendix A. We also use Lemma 11 of Abbasi-Yadkori et al. [1].

Lemma 3. *For any $\lambda > 0$, $\sum_{t=d+1}^{n}\langle \|x_{I_t}\|^2_{G_t^{-1}}\rangle \leq c_3$, where $c_3 = 2d\log(1 + nL^2/(d\lambda))$.*

Proof of Theorem 1. First, we split the regret based on whether event E_1 occurs and obtain

$$\begin{aligned} R(n) &\leq \sum_{t=d+1}^{n}\mathbb{E}\left[\Delta_{I_t}\right] + d \\ &\leq \sum_{t=d+1}^{n}\mathbb{E}\left[\Delta_{I_t}\mathbb{1}\{E_{1,t}\}\right] + n\mathbb{P}\left(\bar{E}_1\right) + d \\ &\leq \sum_{t=d+1}^{n}\mathbb{E}\left[\mathbb{E}_t\left[\Delta_{I_t}\mathbb{1}\{E_{1,t}\}\right]\right] + p_1 n + d\,. \end{aligned}$$

Since $E_{1,t}$ is $\mathcal{F}_{t-1}$ measurable, $\mathbb{E}_t\left[\Delta_{I_t}\mathbb{1}\{E_{1,t}\}\right]$ can be bounded from above by Lemma 2. We apply it and get

$$\begin{aligned} R(n) \leq (c_1 + c_2)\left(1 + \frac{2}{p_3 - p_2}\right) \times \\ \mathbb{E}\left[\sum_{t=d+1}^{n}\langle \|x_{I_t}\|_{G_t^{-1}}\rangle\right] + (p_1 + p_2)n + d\,. \end{aligned}$$

By the Cauchy-Schwarz inequality and Lemma 3,

$$\sum_{t=d+1}^{n}\langle \|x_{I_t}\|_{G_t^{-1}}\rangle \leq \sqrt{n\sum_{t=d+1}^{n}\langle \|x_{I_t}\|^2_{G_t^{-1}}\rangle} \leq \sqrt{c_3 n}\,.$$

The claim follows from chaining the above two inequalities. □

4.3 Expected n-Round Regret of LinPHE

The main result of this section is stated below.

Theorem 4. *Let all parameters be chosen as in Table 1 and $n > \max\left\{34,\, 8\sqrt{2}c_1\right\} = \tilde{O}(d)$. Then the expected n-round regret of* LinPHE *is $R(n) = \tilde{O}(d\sqrt{n})$.*

Our regret bound scales with d and n as that of LinTS [3]. This is unsurprising, since we build on the analysis of LinTS. Our bound also does not improve over those of OFU designs, such as LinUCB [1]. The improvement is in practical performance, as shown in Section 5.

The proof of Theorem 4 follows from Theorem 1 for appropriate choice of c_1, c_2, p_1, p_2, and p_3. These values, together with a number of other constants, are summarized in Table 1. The proof is broken down into lemmas, which are proved in Appendix A.

Constant	Value
L	$\max_{i\in[K]}\|x_i\|_2$
L_*	$\|\theta_*\|_2$
p_1	$1/n$
p_2	$1/n^2$
p_3	$\dfrac{1/2 - 128\,c_1^2 n^{-3}}{16\log n}$
c_1	$\dfrac{1}{2}\sqrt{d\log\left(n + n^2L^2/(d\lambda)\right)} + \lambda^{\frac{1}{2}}L_*$
c_2	$\sqrt{a\log(\sqrt{2K}n)}$
c_3	$2d\log(1 + nL^2/(d\lambda))$
λ	$\lambda_{\min}(G_{d+1})/4$
a	$\lceil 16\,c_1^2\rceil$

Table 1: Summary of the constants in the analysis.

The first lemma guides our choice of c_1. Specifically, we get $p_1 = 1/n$ for c_1 in Table 1.

Lemma 5 (Least-squares concentration)**.** *For any $\lambda > 0$, $\delta > 0$, and*

$$c_1 = \frac{1}{2}\sqrt{d\log\left(\frac{1 + nL^2/(d\lambda)}{\delta}\right)} + \lambda^{\frac{1}{2}}L_*\,,$$

event E_1 occurs with probability at least $1 - \delta$.

The next lemma, together with the union bound over all arms, guarantees that $p_2 = 1/n^2$ for c_2 in Table 1. This lemma is a key part of our analysis.

Lemma 6 (Concentration)**.** *For any $t > d$, $c > 0$, and vector $x \in \mathbb{R}^d$, we have*

$$\mathbb{P}_t\left(\left|x^\top\tilde{\theta}_t - x^\top\bar{\theta}_t\right| \geq c\|x\|_{G_t^{-1}}\right) \leq 2\exp\left[-2c^2/a\right]\,.$$

The next lemma bounds p_3 from below. This lemma is another key part of our analysis.

Lemma 7 (Anti-concentration)**.** *For any round $t > d$, constants a and c such that $2a\log n > c^2 > 0$, and vector $x \in \mathbb{R}^d$ such that $x \neq \mathbf{0}$, we have*

$$\begin{aligned} &\mathbb{P}_t\left(x^\top\tilde{\theta}_t - x^\top\bar{\theta}_t > c\|x\|_{G_t^{-1}}\right) \\ &\quad\geq \frac{1}{16\log n}(1 - \lambda\lambda_{\min}^{-1}(G_{d+1}) - 4a^{-1}c^2 - 8an^{-3})\,. \end{aligned}$$

For $\lambda = \lambda_{\min}(G_{d+1})/4$, $a = \lceil 16\,c_1^2\rceil$, $c = c_1$, and any $x_1 \neq \mathbf{0}$, Lemma 7 implies that

$$p_3 - p_2 \geq \frac{1/2 - 128\,c_1^2 n^{-3} - 16\,n^{-2}\log n}{16\log n}\,.$$

Since $\lambda_{\min}(G_{d+1}) = \lambda + \lambda_{\min}(\sum_{\ell=1}^{d} X_\ell X_\ell^\top)$, we have that $\lambda = \lambda_{\min}(\sum_{\ell=1}^{d} X_\ell X_\ell^\top)/3$. Finally, we set c_3 as in Table 1. Now are ready to prove Theorem 4.

Proof of Theorem 4. If $x_1 = \mathbf{0}$, the proof is trivial. Now suppose that $x_1 \neq \mathbf{0}$. Since $L = O(\sqrt{d})$, $L_* = O(\sqrt{d})$, and $\lambda = O(1)$, we have $c_1 = \tilde{O}(\sqrt{d})$. Moreover, since $a = \lceil 16\, c_1^2 \rceil$, we have $c_2 = \tilde{O}(\sqrt{d})$. Finally, $c_3 = \tilde{O}(d)$.

Now we show that $1 + 1/(p_3 - p_2) = \tilde{O}(1)$. Trivially, $n^{-1} \log n \leq 1$ for $n \geq 1$. In addition, for $n \geq 8\sqrt{2} c_1$, $128\, c_1^2 n^{-2} \leq 1$. So, for any such n,

$$p_3 - p_2 \geq \frac{1/2 - 17\, n^{-1}}{16 \log n}\,.$$

Finally, for any $n > 34$, the above lower bound is positive and $1 + 1/(p_3 - p_2) = \tilde{O}(1)$. This concludes our proof. □

5 EXPERIMENTS

We conduct two experiments to evaluate both LinPHE and LogPHE in terms of their regret. The algorithms are compared to several state-of-the-art baselines.

5.1 Linear Bandit

The first experiment is with linear bandits. We experiment with dimensions d from 5 to 20. The number of arms is $K = 100$. To avoid biases, we randomly generate problem instances. Each instance is generated as follows. The first $d-1$ entries of feature vector x_i are drawn from a unit $(d-2)$-sphere and the last entry is 1. The first $d-1$ entries of parameter vector θ_* are drawn from a $(d-2)$-sphere with radius 0.5 and the last entry is 0.5. This construction guarantees that $x_i^\top \theta_* \in [0, 1]$ for all arms i. The reward of arm i is drawn i.i.d. from $\text{Ber}(x_i^\top \theta_*)$. The horizon is $n = 10000$ rounds and our results are averaged over 100 problem instances.

We compare LinPHE to LinUCB [1], LinTS [3], and the ε-greedy policy [30, 4] with a linear model. LinUCB is an OFU algorithm. We set the regularization parameter in LinUCB as $\lambda = 1$. All other parameters are set as in Abbasi-Yadkori et al. [1]. LinTS is a posterior sampling algorithm. We set its prior to $\mathcal{N}(0, I_d)$. The exploration rate in the ε-greedy policy is $\varepsilon_t = \min\{1, 0.05/(2\sqrt{t})\}$, which results in about 5% exploration. We experiment with three practical perturbation scales a in LinPHE: 2, 1, and 0.5. We implement LinPHE with non-integer perturbation scales a by replacing $B(a\, T_{i,t-1}, 1/2)$ in Section 3.2 with $B(\lceil a\, T_{i,t-1} \rceil, 1/2)$.

Our results are reported in Figure 1. We observe the following trends. First, LinPHE outperforms LinUCB at all perturbation scales a. Second, LinPHE outperforms the ε-greedy policy at all perturbation scales a in the first two problems. In the last problem, this happens only at $a \leq 1$. Finally, LinPHE performs similarly to LinTS at $a = 1$ and outperforms it at $a = 0.5$. However, the run time of LinPHE is less than a half of that of LinTS. For instance, at $d = 5$, the average run times of LinPHE and LinTS are 113 and 273 seconds, respectively. The increased run time of LinTS is due to posterior sampling from the multivariate normal distribution.

5.2 Logistic Bandit

The last experiment is with logistic bandits (Section 3.3). The experimental setup differs from Section 5.1 only in how θ_* is generated. The first $d-1$ entries of parameter vector θ_* are drawn from a $(d-2)$-sphere with radius 1.5 and the last entry is 0. By design, $\left|x_i^\top \theta_*\right| \leq 1.5$.

We compare LogPHE to LogTS [6, 29], GLM-UCB [11], UCB-GLM [21], and the ε-greedy policy [30, 4] with a logistic model. GLM-UCB and UCB-GLM are OFU methods for logistic bandits. We implement them with regularization (Section 3.3) and $\lambda = 1$. The minimum derivative of the mean function, which is tunable in both methods, is set to the most optimistic value of $1/4$. All other parameters are set as suggested by theory. LogTS is a posterior sampling algorithm for logistic regression, which uses the Laplace approximation and has prior $\mathcal{N}(0, I_d)$. The ε-greedy policy is implemented as in Section 5.1.

Our results are reported in Figure 2. We observe similar trends to Section 5.1. In particular, LogPHE usually outperforms OFU algorithms and is competitive with posterior sampling when $a \leq 1$. In summary, our experimental results show that both LinPHE and LogPHE perform well, and are comparable to or better than existing algorithms.

6 RELATED WORK

Our work is motivated by Kveton et al. [17], who proposed a multi-armed bandit algorithm that pulls the arm with the highest average reward in its perturbed history with i.i.d. pseudo-rewards. We generalize this approach to linear, and more generally contextual, bandits. This generalization is important. While the perturbed history is conceptually simple, it is unclear how to extend it to structured problems, and assessing if such a generalization is sound is non-trivial. We propose one generalization, and show it to be both sound and effective.

Our work is related to posterior sampling. In particular, let $\mu \sim \mathcal{N}(\mu_0, \sigma^2)$ and $(Y_\ell)_{\ell=1}^s \sim \mathcal{N}(\mu, \sigma^2)$ be s i.i.d. noisy observations of μ. Then the posterior distribution of μ conditioned on $(Y_\ell)_{\ell=1}^s$ is

$$\mathcal{N}\left(\frac{\mu_0 + \sum_{\ell=1}^s Y_\ell}{s+1},\, \frac{\sigma^2}{s+1}\right)\,. \tag{9}$$

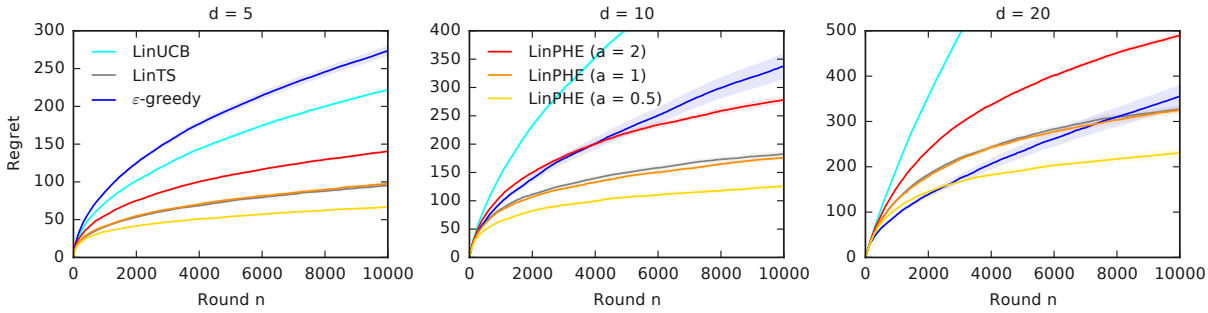

Figure 1: Comparison of LinPHE to several baselines in three linear bandit problems. All results are averaged over 100 random problem instances. The shaded areas are standard errors of the estimates. The legend is split between the first two plots.

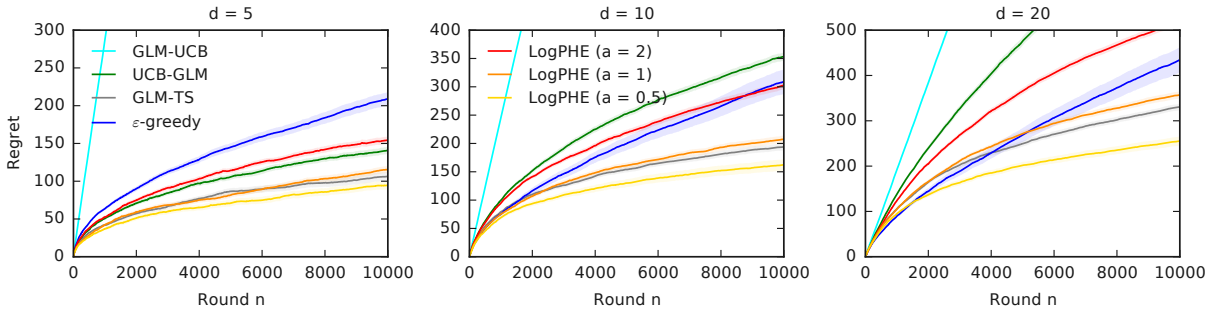

Figure 2: Comparison of LogPHE to several baselines in three logistic bandit problems. All results are averaged over 100 random problem instances. The shaded areas are standard errors of the estimates. The legend is split between the first two plots.

It is easy to see that the above distribution can be indirectly sampled from as follows. First, draw $s+1$ i.i.d. samples $(Z_\ell)_{\ell=0}^s \sim \mathcal{N}(0, \sigma^2)$. Then

$$\frac{\mu_0 + \sum_{\ell=1}^s Y_\ell + \sum_{\ell=0}^s Z_\ell}{s+1}$$

is a sample from (9). This equivalence can be generalized to linear models with Gaussian noise [24]. Unfortunately, it holds only for normal random variables. Therefore, it cannot justify our perturbation scheme as a form of posterior sampling.

The design of LinPHE is similar to *follow the perturbed leader (FPL)* [13, 15]. FPL has been traditionally studied in the *non-stochastic full-information* setting. Neu and Bartok [25] extended it to *semi-bandits* using geometric resampling. Their algorithm cannot solve our problem efficiently because it is for a K-armed bandit with independent arms.

Our work is closely related to bootstrapping exploration [5, 9, 26, 31, 10, 18, 34], where the learning agent perturbs its history of observations by resampling in order to achieve exploration. Contextual bootstrapping algorithms [31, 10, 18, 34] have superior empirical performance but no regret bounds. Our work provides a stepping stone for the analysis of such algorithms, since our perturbation scheme is similar but simpler.

7 CONCLUSIONS

We propose LinPHE, a new online algorithm for cumulative regret minimization in stochastic linear bandits. The key idea in LinPHE is to perturb the history in round t by $O(t)$ i.i.d. pseudo-rewards, which are drawn from the maximum variance distribution. We derive a $\tilde{O}(d\sqrt{n})$ bound on the n-round regret of LinPHE, where d is the number of features. We also propose LogPHE, a natural generalization of LinPHE to a logistic model. We evaluate LinPHE and LogPHE empirically. Both algorithms are competitive with Thompson sampling, although they are derived based on different insights.

LinPHE can be easily extended to any linear model with a bounded support. In particular, if $Y_{i,t} \in [m, M]$, Y_ℓ in LinPHE should be replaced with $(Y_\ell - m)/(M - m)$.

Our work can be extended in several directions. First, although we propose LogPHE for a logistic model, we do

not analyze it. We believe that the regret analysis is possible because generalized linear bandit analyses [11, 21] are similar to linear bandit analyses [7, 1]. Second, the theory-suggested perturbation scale a in Table 1 is too conservative to be practical, for the same reason as the analyzed variant of `LinTS` in Agrawal and Goyal [3]. A tighter analysis should be possible. Third, our key technical lemmas, Lemmas 6 and 7, can be extended to other randomized pseudo-rewards than Bernoulli. This would be necessary for other generalized linear models than logistic. Finally, our design seems conservative since the strategy for adding pseudo-rewards does not adapt over time. More adaptive designs may be possible.

A PROOFS

A.1 Proof of Lemma 2

Let

$$\bar{S}_t = \left\{ i \in [K] : (c_1 + c_2)\|x_i\|_{G_t^{-1}} \geq \Delta_i \right\} \tag{10}$$

be the set of *undersampled arms* in round t. Note that by definition $1 \in \bar{S}_t$. The set of *sufficiently sampled arms* is defined as $S_t = [K] \setminus \bar{S}_t$. Let

$$J_t = \arg\min_{i \in \bar{S}_t} \|x_i\|_{G_t^{-1}} \tag{11}$$

be the *least uncertain undersampled arm* in round t. In all steps below, we assume that event $E_{1,t}$ occurs.

Let $c = c_1 + c_2$. In round t on event $E_{2,t}$,

$$\begin{aligned}
\Delta_{I_t} &= \Delta_{J_t} + \langle x_{J_t}^\top \theta_* - x_{I_t}^\top \theta_* \rangle \\
&\leq \Delta_{J_t} + \langle x_{J_t}^\top \tilde{\theta}_t - x_{I_t}^\top \tilde{\theta}_t \rangle + \\
&\quad c\left(\langle \|x_{I_t}\|_{G_t^{-1}} \rangle + \langle \|x_{J_t}\|_{G_t^{-1}} \rangle\right) \\
&\leq c\left(\langle \|x_{I_t}\|_{G_t^{-1}} \rangle + 2\langle \|x_{J_t}\|_{G_t^{-1}} \rangle\right),
\end{aligned}$$

where the first inequality is by the definitions of events $E_{1,t}$ and $E_{2,t}$, and the second follows from the definitions of I_t and J_t. We also used that $c = c_1 + c_2 \geq 1$. Now we take the expectation of both sides and get

$$\begin{aligned}
&\mathbb{E}_t\left[\Delta_{I_t}\right] \\
&\quad = \mathbb{E}_t\left[\Delta_{I_t} \mathbb{1}\{E_{2,t}\}\right] + \mathbb{E}_t\left[\Delta_{I_t} \mathbb{1}\{\bar{E}_{2,t}\}\right] \\
&\quad \leq c\,\mathbb{E}_t\left[\langle \|x_{I_t}\|_{G_t^{-1}} \rangle + 2\langle \|x_{J_t}\|_{G_t^{-1}} \rangle\right] + \mathbb{P}_t\left(\bar{E}_{2,t}\right).
\end{aligned}$$

The last step is to bound $\mathbb{E}_t\left[\langle \|x_{J_t}\|_{G_t^{-1}} \rangle\right]$ from above. The key observation is that

$$\begin{aligned}
&\mathbb{E}_t\left[\langle \|x_{I_t}\|_{G_t^{-1}} \rangle\right] \\
&\quad \geq \mathbb{E}_t\left[\langle \|x_{I_t}\|_{G_t^{-1}} \rangle \,\middle|\, I_t \in \bar{S}_t\right] \mathbb{P}_t\left(I_t \in \bar{S}_t\right) \\
&\quad \geq \langle \|x_{J_t}\|_{G_t^{-1}} \rangle\, \mathbb{P}_t\left(I_t \in \bar{S}_t\right),
\end{aligned}$$

where the last inequality is from the definition of J_t and that $\bar{S}_t$ is $\mathcal{F}_{t-1}$-measurable. We rearrange the inequality and get

$$\langle \|x_{J_t}\|_{G_t^{-1}} \rangle \leq \mathbb{E}_t\left[\langle \|x_{I_t}\|_{G_t^{-1}} \rangle\right] \Big/ \mathbb{P}_t\left(I_t \in \bar{S}_t\right).$$

Next we bound $\mathbb{P}_t\left(I_t \in \bar{S}_t\right)$ from below. On event $E_{1,t}$,

$$\begin{aligned}
\mathbb{P}_t\left(I_t \in \bar{S}_t\right) &\geq \mathbb{P}_t\left(\exists i \in \bar{S}_t : x_i^\top \tilde{\theta}_t > \max_{j \in \bar{S}_t} x_j^\top \tilde{\theta}_t\right) \\
&\geq \mathbb{P}_t\left(x_1^\top \tilde{\theta}_t > \max_{j \in S_t} x_j^\top \tilde{\theta}_t\right) \\
&\geq \mathbb{P}_t\left(x_1^\top \tilde{\theta}_t > \max_{j \in S_t} x_j^\top \tilde{\theta}_t,\ E_{2,t} \text{ occurs}\right) \\
&\geq \mathbb{P}_t\left(x_1^\top \tilde{\theta}_t > x_1^\top \theta_*,\ E_{2,t} \text{ occurs}\right) \\
&\geq \mathbb{P}_t\left(x_1^\top \tilde{\theta}_t > x_1^\top \theta_*\right) - \mathbb{P}_t\left(\bar{E}_{2,t}\right).
\end{aligned}$$

Note that we require a sharp inequality because $x_i^\top \tilde{\theta}_t \geq \max_{j \in S_t} x_j^\top \tilde{\theta}_t$ does not imply that arm i is pulled. The fourth inequality holds because for any $j \in S_t$,

$$x_j^\top \tilde{\theta}_t \leq x_j^\top \theta_* + c\|x_j\|_{G_t^{-1}} < x_j^\top \theta_* + \Delta_j = x_1^\top \theta_*$$

on event $E_{1,t} \cap E_{2,t}$. Finally,

$$\mathbb{P}_t\left(x_1^\top \tilde{\theta}_t > x_1^\top \theta_*\right) \geq \mathbb{P}_t\left(x_1^\top \tilde{\theta}_t - x_1^\top \bar{\theta}_t > c_1 \|x_1\|_{G_t^{-1}}\right)$$

on event $E_{1,t}$, because $x_1^\top \theta_* \leq x_1^\top \bar{\theta}_t + c_1 \|x_1\|_{G_t^{-1}}$ holds on event $E_{1,t}$. Now we chain all inequalities and use the definitions of p_1, p_2, and p_3 to complete the proof.

A.2 Proof of Lemma 5

By the Cauchy-Schwarz inequality,

$$\begin{aligned}
x_i^\top \bar{\theta}_t - x_i^\top \theta_* &= x_i^\top G_t^{-\frac{1}{2}} G_t^{\frac{1}{2}} (\bar{\theta}_t - \theta_*) \\
&\leq \|\bar{\theta}_t - \theta_*\|_{G_t} \|x_i\|_{G_t^{-1}}.
\end{aligned}$$

Now note that the least-squares estimate $\bar{\theta}_t$ is computed from sub-Gaussian rewards with variance proxy $1/4$. As a result, by Theorem 2 of Abbasi-Yadkori et al. [1] for $R = 1/2$, $\|\bar{\theta}_t - \theta_*\|_{G_t} \leq c_1$ holds jointly in all rounds $t \leq n$ with probability of at least $1 - \delta$. This concludes the proof.

A.3 Proof of Lemma 6

Let

$$\begin{aligned}
U &= \sum_{\ell=1}^{t-1} \sum_{j=1}^{a} x^\top G_t^{-1} X_\ell Z_{j,\ell}, \\
\bar{U} &= \sum_{\ell=1}^{t-1} \sum_{j=1}^{a} x^\top G_t^{-1} X_\ell \bar{Z}_{j,\ell},
\end{aligned}$$

and $D = U - \bar{U}$. Then by Hoeffding's inequality,

$$\begin{aligned}
&\mathbb{P}_t\left(\left|x^\top\tilde{\theta}_t - x^\top\bar{\theta}_t\right| \geq c\|x\|_{G_t^{-1}}\right) \\
&\quad= \mathbb{P}_t\left(|D| \geq c\|x\|_{G_t^{-1}}\right) \\
&\quad\leq 2\exp\left[-\frac{2c^2\|x\|^2_{G_t^{-1}}}{a\sum_{\ell=1}^{t-1} x^\top G_t^{-1}X_\ell X_\ell^\top G_t^{-1}x}\right].
\end{aligned}$$

This step of the proof relies on the fact that new $Z_{j,\ell}$ are generated in each round t. Also note that

$$\begin{aligned}
&\sum_{\ell=1}^{t-1} x^\top G_t^{-1}X_\ell X_\ell^\top G_t^{-1}x \qquad (12)\\
&\quad\leq x^\top G_t^{-1}\left(\sum_{\ell=1}^{t-1} X_\ell X_\ell^\top + \lambda I_d\right)G_t^{-1}x = \|x\|^2_{G_t^{-1}}.
\end{aligned}$$

Our claim follows from chaining all above inequalities.

A.4 Proof of Lemma 7

Let U, $\bar{U}$, and D be defined as in the proof of Lemma 6. Then $x^\top\tilde{\theta}_t - x^\top\bar{\theta}_t = D$. We also define events

$$\begin{aligned}
F_1 &= \left\{|D| \leq c\|x\|_{G_t^{-1}}\right\}, \\
F_2 &= \left\{|D| \leq \sqrt{2a\log n}\|x\|_{G_t^{-1}}\right\}.
\end{aligned}$$

Since $2a\log n > c^2$, $F_1 \subset F_2$. Then

$$\begin{aligned}
\mathrm{var}\left[U \,|\, \mathcal{F}_{t-1}\right] = \mathbb{E}_t\left[D^2\mathbb{1}\{F_1\}\right] + \\
\mathbb{E}_t\left[D^2\mathbb{1}\{\bar{F}_1, F_2\}\right] + \\
\mathbb{E}_t\left[D^2\mathbb{1}\{\bar{F}_2\}\right].
\end{aligned}$$

Now we bound each term on the right-hand side of the above equality from above. From the definition of event F_1, term 1 is bounded as

$$\mathbb{E}_t\left[D^2\mathbb{1}\{F_1\}\right] \leq c^2\|x\|^2_{G_t^{-1}}.$$

By the definition of F_1 and F_2, term 2 is bounded as

$$\begin{aligned}
&\mathbb{E}_t\left[D^2\mathbb{1}\{\bar{F}_1, F_2\}\right] \\
&\quad\leq (2a\|x\|^2_{G_t^{-1}}\log n)\,\mathbb{P}_t\left(\bar{F}_1, F_2 \text{ occur}\right) \\
&\quad\leq (2a\|x\|^2_{G_t^{-1}}\log n)\,\mathbb{P}_t\left(|D| > c\|x\|_{G_t^{-1}}\right).
\end{aligned}$$

Now we bound term 3. First, note that

$$\begin{aligned}
|D| &\leq a\sum_{\ell=1}^{t-1}\left|x^\top G_t^{-1}X_\ell\right| \\
&\leq a\sqrt{n}\sqrt{\sum_{\ell=1}^{t-1} x^\top G_t^{-1}X_\ell X_\ell^\top G_t^{-1}x} \\
&\leq a\sqrt{n}\|x\|_{G_t^{-1}},
\end{aligned}$$

where the last step follows from (12). Then, by the definition of event F_2 and Lemma 6 for $c = \sqrt{2a\log n}$,

$$\begin{aligned}
\mathbb{E}_t\left[D^2\mathbb{1}\{\bar{F}_2\}\right] &\leq a^2 n\|x\|^2_{G_t^{-1}}\mathbb{P}_t\left(\bar{F}_2\right) \\
&\leq \frac{2a^2\|x\|^2_{G_t^{-1}}}{n^3}.
\end{aligned}$$

Finally, by the definition of U,

$$\begin{aligned}
\mathrm{var}\left[U \,|\, \mathcal{F}_{t-1}\right] &= \frac{a}{4}\sum_{\ell=1}^{t-1} x^\top G_t^{-1}X_\ell X_\ell^\top G_t^{-1}x \\
&= \frac{a}{4}\|x\|^2_{G_t^{-1}} - \frac{a}{4}\lambda x^\top G_t^{-2}x.
\end{aligned}$$

We bound the last term from below as follows. For any positive semi-definite matrix $M \in \mathbb{R}^{d\times d}$,

$$\begin{aligned}
x^\top M^2 x &= \lambda^2_{\max}(M)\,x^\top\left(\lambda^{-2}_{\max}(M)M^2\right)x \\
&\leq \lambda^2_{\max}(M)\,x^\top\left(\lambda^{-1}_{\max}(M)M\right)x \\
&= \lambda_{\max}(M)\|x\|^2_M,
\end{aligned}$$

where the inequality follows from the fact that all eigenvalues of $\lambda^{-2}_{\max}(M)M^2$ are in $[0,1]$. We apply this upper bound for $M = G_t^{-1}$ and get that

$$\begin{aligned}
\mathrm{var}\left[U \,|\, \mathcal{F}_{t-1}\right] &\geq \frac{a}{4}\|x\|^2_{G_t^{-1}} - \frac{a\lambda}{4\,\lambda_{\min}(G_t)}\|x\|^2_{G_t^{-1}} \\
&\geq \frac{a}{4}\|x\|^2_{G_t^{-1}} - \frac{a\lambda}{4\,\lambda_{\min}(G_{d+1})}\|x\|^2_{G_t^{-1}},
\end{aligned}$$

where the last inequality is by $\lambda_{\min}(G_t) \geq \lambda_{\min}(G_{d+1})$ and holds for any $t > d$.

Now we combine all above inequalities and get

$$\begin{aligned}
&\left[\frac{a}{4} - \frac{a\lambda}{4\,\lambda_{\min}(G_{d+1})} - c^2 - \frac{2a^2}{n^3}\right]\|x\|^2_{G_t^{-1}} \\
&\quad\leq (2a\|x\|^2_{G_t^{-1}}\log n)\,\mathbb{P}_t\left(|D| > c\|x\|_{G_t^{-1}}\right).
\end{aligned}$$

Since $2a\log n > 0$ and $\|x\|_{G_t^{-1}} > 0$, the above inequality can be simplified as

$$\begin{aligned}
&\mathbb{P}_t\left(|D| > c\|x\|_{G_t^{-1}}\right) \\
&\quad\geq \frac{1}{8\log n}(1 - \lambda\lambda^{-1}_{\min}(G_{d+1}) - 4a^{-1}c^2 - 8an^{-3}).
\end{aligned}$$

Finally, we note that the distribution of D is symmetric. Therefore, for any $\varepsilon > 0$, $\mathbb{P}_t(|D| > \varepsilon) = 2\mathbb{P}_t(D > \varepsilon)$. This completes the proof.

References

[1] Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320, 2011.

[2] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for Thompson sampling. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, pages 99–107, 2013.

[3] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, pages 127–135, 2013.

[4] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

[5] Akram Baransi, Odalric-Ambrym Maillard, and Shie Mannor. Sub-sampling for multi-armed bandits. In *Proceeding of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2014.

[6] Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems 24*, pages 2249–2257, 2011.

[7] Varsha Dani, Thomas Hayes, and Sham Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 355–366, 2008.

[8] Luc Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, NY, 1986.

[9] Dean Eckles and Maurits Kaptein. Thompson sampling with the online bootstrap. *CoRR*, abs/1410.4009, 2014. URL http://arxiv.org/abs/1410.4009.

[10] Adam Elmachtoub, Ryan McNellis, Sechan Oh, and Marek Petrik. A practical method for solving contextual bandit problems using decision trees. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, 2017.

[11] Sarah Filippi, Olivier Cappe, Aurelien Garivier, and Csaba Szepesvari. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23*, pages 586–594, 2010.

[12] Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *Proceedings of the 31st International Conference on Machine Learning*, pages 100–108, 2014.

[13] James Hannan. Approximation to Bayes risk in repeated play. In *Contributions to the Theory of Games*, volume 3, pages 97–140. Princeton University Press, Princeton, NJ, 1957.

[14] Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems 30*, pages 99–109, 2017.

[15] Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

[16] Jaya Kawale, Hung Bui, Branislav Kveton, Long Tran-Thanh, and Sanjay Chawla. Efficient Thompson sampling for online matrix-factorization recommendation. In *Advances in Neural Information Processing Systems 28*, pages 1297–1305, 2015.

[17] Branislav Kveton, Csaba Szepesvari, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic multi-armed bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019.

[18] Branislav Kveton, Csaba Szepesvari, Sharan Vaswani, Zheng Wen, Mohammad Ghavamzadeh, and Tor Lattimore. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3601–3610, 2019.

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

[20] Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2019.

[21] Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2071–2080, 2017.

[22] Zachary Lipton, Xiujun Li, Jianfeng Gao, Lihong Li, Faisal Ahmed, and Li Deng. BBQ-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 5237–5244, 2018.

[23] Bing Liu, Tong Yu, Ian Lane, and Ole Mengshoel. Customized nonlinear bandits for online response selection in neural conversation models. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 5245–5252, 2018.

[24] Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. In *Advances in Neural Information Processing Systems 30*, pages 3258–3266, 2017.

[25] Gergely Neu and Gabor Bartok. An efficient algorithm for learning with semi-bandit feedback. In *Proceedings of the 24th International Conference on Algorithmic Learning Theory*, pages 234–248, 2013.

[26] Ian Osband and Benjamin Van Roy. Bootstrapped Thompson sampling and deep exploration. *CoRR*, abs/1507.00300, 2015. URL http://arxiv.org/abs/1507.00300.

[27] Carlos Riquelme, George Tucker, and Jasper Snoek. Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks for Thompson sampling. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[28] Paat Rusmevichientong and John Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

[29] Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

[30] Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[31] Liang Tang, Yexi Jiang, Lei Li, Chunqiu Zeng, and Tao Li. Personalized recommendation via parameter-free contextual bandits. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 323–332, 2015.

[32] William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[33] Michal Valko, Remi Munos, Branislav Kveton, and Tomas Kocak. Spectral bandits for smooth graph functions. In *Proceedings of the 31st International Conference on Machine Learning*, pages 46–54, 2014.

[34] Sharan Vaswani, Branislav Kveton, Zheng Wen, Anup Rao, Mark Schmidt, and Yasin Abbasi-Yadkori. New insights into bootstrapping for bandits. *CoRR*, abs/1805.09793, 2018. URL http://arxiv.org/abs/1805.09793.