Stochastic Linear Contextual Bandits with Diverse Contexts

Weiqiang Wu
London Stock Exchange

Jing Yang
The Pennsylvania State University

Cong Shen
University of Virginia

Abstract

In this paper, we investigate the impact of context diversity on stochastic linear contextual bandits. As opposed to the previous view that contexts lead to more difficult bandit learning, we show that when the contexts are sufficiently diverse, the learner is able to utilize the information obtained during exploitation to shorten the exploration process, thus achieving reduced regret. We design the LinUCB-d algorithm, and propose a novel approach to analyze its regret performance. The main theoretical result is that under the diverse context assumption, the cumulative expected regret of LinUCB-d is bounded by a constant. As a by-product, our results improve the previous understanding of LinUCB and strengthen its performance guarantee.

1 Introduction

In many applications, such as resource allocation in cloud computing platforms, or treatment selection for patients in clinical trials, the diverse user preferences and characteristics impose urgent need of personalized decision-making. In order to make optimal decisions for different individuals, the decision maker must learn a model to predict the reward when a decision is taken under different contexts. This problem is often formulated as a contextual bandit problem (Auer, 2003; Langford and Zhang, 2008), which generalizes the classical multi-armed bandit (MAB) framework (Lai and Robbins, 1985; Auer et al., 2002; Bubeck and Cesa-Bianchi, 2012; Agrawal and Goyal, 2012, 2013).

The inclusion of contextual information in the decision-making process introduces more uncertainty into the MAB framework and creates significant challenges for the learning problem. Part of the difficulty in contextual MAB comes from the increased problem dimension, as context is added as part of the unknown environment. Existing literature mostly focuses on developing bandit algorithms and providing theoretical analysis by treating the context as *structureless* side information. The resulting models and algorithms are generic, but represent a "worst case" scenario since very little, if any, structure of the context is exploited.

In many real-world applications, however, context often exhibits sufficient level of *diversity*, which has been largely overlooked in the existing studies. For example, user profile is considered as the context in recommendation systems (Li et al., 2010). When the system serves a large number of users, the group of user profiles is likely to be very diverse. As another example, contextual MAB has been adopted in service placement of mobile edge computing, which utilizes the time of day and mobile user types as the context (Chen and Xu, 2019).

It is not difficult to see that in these applications, the context arrival exhibits sufficient diversity that may be beneficial to the bandit algorithm design. Intuitively, the diverse contexts create opportunities for the learner to reduce the learning regret: when an arm is pulled frequently as the optimal arm for certain contexts, its parameters can be estimated accurately with the rewards obtained during exploitation. Therefore, the learner does not have to spend much time exploring it when the instantaneous context is not in favor for it, thus shortening the exploration stage and speeding up the convergence.

In this paper, we demonstrate this optimistic view of context diversity by investigating the impact of diverse contextual information on the learning regret under the stochastic linear contextual bandits framework (Bastani et al., 2017). We show that, instead of considering the context as part of the "uncontrollable" environment and passively "reacting" to the incoming context, proactively interacting between context and arm exploration allows learning to transfer between different contexts and leads to much better overall performance. Specifically, we consider a set of K arms, where the parameter

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

of each arm is represented by a d-dimensional vector unknown to the learner. When an arm a is pulled under a context c, the obtained reward is the inner product of the parameter of the arm and a feature vector determined by arm a and context c, corrupted by noise. The objective of the learner is to use the information contained in the observed rewards to decide an arm to pull in response to the instantaneous context. Assuming independent and identically distributed (i.i.d.) contexts, we aim to show that when the contexts are sufficiently diverse, the cumulative learning regret in expectation can be bounded by a constant.

The main contributions of this paper are three-fold. First, we formally introduce the concept of *context diversity* into the stochastic linear contextual bandit framework and present a novel geometric interpretation. Such geometric interpretation provides an intuitive viewpoint to understand and analyze the impact of context diversity on the learning performance of stochastic linear contextual bandits.

Second, we propose an Upper Confidence Bound (UCB) type algorithm, termed as LinUCB-d, for the contextual bandit model. The new formulation of LinUCB-d enables a unique approach to characterize the impact of context diversity and achieve finite cumulative regret. The results also extend the existing understanding of LinUCB and strengthen its performance guarantee.

Third, we design a novel approach to analyze the performance of LinUCB-d. There are two distinct features of our approach: First, we relate the uncertainty in the estimated rewards with the solution to a constrained optimization problem, and leverage the optimality of the estimator to bound the corresponding frequency of bad decisions during the learning process. Second, we propose a frame-based approach to isolate the error events on a frame basis and make the regret tractable. These techniques are novel and may find useful in other related settings.

2 Problem Formulation

We consider a set of K items (arms) denoted as $[K] = \{1, 2, \dots, K\}$. Assume for each $a \in [K]$ there is a fixed but unknown parameter vector $\boldsymbol{\theta}(a) \in \mathbb{R}^d$. At each time t, the learner observes a random context c_t, which is generated according to an unknown distribution. Next, the learner decides to pull an arm $a_t \in [K]$ based on the information available. The incurred reward y_t is given by $y_t = r(a_t, c_t) + \eta_t$, where η_t is a random noise, and $r(a_t, c_t)$ is a linear function of $\boldsymbol{\theta}(a_t)$ and the feature vector $\mathbf{x}(a_t, c_t) \in \mathbb{R}^d$, i.e.,

$$r(a_t, c_t) := \boldsymbol{\theta}^\intercal(a_t)\mathbf{x}(a_t, c_t). \tag{1}$$

Here we use $\mathbf{x}^\intercal$ to denote the transpose of vector $\mathbf{x}$. Let $\mathcal{C}$ be the set of all contexts. For any $a \in [K]$, define

$$\mathcal{C}_a := \{c \in \mathcal{C} \mid r(a, c) > r(b, c), \forall b \in [K]\backslash\{a\}\}, \tag{2}$$

$$\mathcal{X}_a := \left\{\mathbf{x}(a, c) \in \mathbb{R}^d \mid c \in \mathcal{C}_a\right\}, \tag{3}$$

i.e., $\mathcal{C}_a$ is the subset of contexts under which arm a is the best arm rendering the maximum expected reward, and $\mathcal{X}_a$ is the set of feature vectors when arm a is pulled under contexts in $\mathcal{C}_a$. Let $\mathcal{F}_t = \sigma(c_1, a_1, y_1, \dots, c_{t-1}, a_{t-1}, y_{t-1}, c_t, a_t)$ be the σ-field summarizing the information available just before y_t is observed. We make the following assumptions throughout the paper.

Assumption 1 *1)* ***Bounded parameters:*** *For any $a \in [K]$, $c \in \mathcal{C}$, we have $\|\boldsymbol{\theta}(a)\|_2 \leq s$, $\|\mathbf{x}(a, c)\|_2 \leq l$.*

2) ***Minimum reward gap:*** *For any $a, b \in [K]$, $b \neq a$, $c \in \mathcal{C}_a$, $r(a, c) - r(b, c) \geq \Delta > 0$.*

3) ***Conditionally 1-subgaussian noise:*** *Given $\mathcal{F}_t$, η_t is conditionally 1-subgaussian with $\mathbb{E}[\eta_t|\mathcal{F}_t] = 0$, $\mathbb{E}[\exp(\lambda\eta_t)|\mathcal{F}_t] \leq \exp(\frac{\lambda^2}{2})$ for any $\lambda > 0$.*

4) ***Stochastic context arrivals:*** *In each time t, c_t is drawn from the context set $\mathcal{C}$ in an i.i.d. fashion according to a distribution ν.*

5) ***Diverse contexts:*** *For any arm $a \in [K]$, $\lambda_{\min}(\mathbb{E}_{c\sim\nu}[\mathbf{x}(a, c)\mathbf{x}(a, c)^\intercal | c \in \mathcal{C}_a]) > 0$, where $\lambda_{\min}(A)$ is the smallest eigenvalue of A.*

Assumption 1.1 ensures that the maximum regret at any step is bounded. Assumption 1.2 indicates that under a given context c, the optimal arm is strictly better than any other sub-optimal arms. Such a reward gap affects the convergence rate of the proposed algorithm (similar to the stochastic MAB setting). Assumption 1.3 allows us to utilize the induced super-martingale to derive exponentially decaying tail bound of the estimation error. We note that all three assumptions on the bandit model are standard in the bandit literature or in the study of linear bandits.

Assumption 1.4 is a non-critical assumption made for ease of exposition. Essentially, what is required for the main results to hold is the ergodicity of the context arrival process, i.e., contexts lying in certain favorable subsets recur frequently in time.

Assumption 1.5 is however critical for our main results to hold. It is equivalent to the condition that

$$\mathbb{P}[\lambda_{\min}(\Phi_a^\intercal\Phi_a) > 0] > 0, \quad \forall a \in [K], \tag{4}$$

where Φ_a is a random matrix whose columns are d feature vectors associated with arm a and d i.i.d. contexts drawn according to the conditional distribution of c given $c \in \mathcal{C}_a$. It implies two conditions: First, all arms in $[K]$ could be optimal under certain contexts, i.e., $\mathcal{C}_a \neq \emptyset$. Second, for all contexts in favor of the same arm (i.e., contexts in $\mathcal{C}_a$), they are sufficiently diverse so that the corresponding feature vectors span $\mathbb{R}^d$. If the first condition does not hold, the arms that are strictly sub-optimal have to be explored sufficiently frequently in order to be distinguished from the optimal arms, thus an $O(\log T)$ regret is unavoidable in this situation. For the second condition, although it seems strict at first sight, it is actually quite reasonable in practice. This is because if a feature vector falls in $\mathcal{X}_a$, we would expect that feature vectors drawn from a small neighborhood of it fall in $\mathcal{X}_a$ as well. Since small perturbations of a vector can form a full-rank matrix, it is thus reasonable to assume $\text{span}(\mathcal{X}_a) = \mathbb{R}^d$.

Assume $\{\mathbf{x}(a,c)\}_{a,c}$ is given and $\{\boldsymbol{\theta}(a)\}_a$ is unknown a priori. The cumulative regret of an online learning algorithm is defined as

$$R_T := \sum_{t=1}^{T} r(a_t^*, c_t) - \sum_{t=1}^{T} y_t, \tag{5}$$

where $a_t^* := \arg\max_{a\in[K]} r(a, c_t)$. While sublinear learning regret has been established for such linear contextual bandits (Abbasi-Yadkori et al., 2011; Chu et al., 2011), our objective is to investigate the fundamental impact of context diversity on the expected regret $\mathbb{E}[R_T]$.

3 Algorithm

The existing linear contextual bandit algorithms such as the celebrated LinUCB (Li et al., 2010) can be directly applied to the considered bandit problem. However, such approaches ignore the diversity in context arrivals and offer little insight to the understanding of diversity. In this section, we propose a Linear Upper Confidence Bound algorithm to manifest the impact of the diversity of context on the scaling of the learning regret. To distinguish it from LinUCB, we term it LinUCB-d.

We label all contexts that have appeared in the order of their first appearances. We assume there are n_t different contexts that have appeared before time t. With a slight abuse of notation, we denote the subset of those contexts as $\mathcal{C}_t$. Besides, we add d dummy contexts, and denote the subset as $\mathcal{C}_0$. In the following, we use $c \in \{1, 2, \ldots, n_t + d\}$ to index the contexts, while the first n_t contexts are in $\mathcal{C}_t$ and the last d are the added dummy ones. For the added dummy contexts, we assume the corresponding feature vector $\mathbf{x}(a, n_t + j) = l\mathbf{e}_j$, $j = 1, \ldots, d$, where $\mathbf{e}_j \in \mathbb{R}^d$ is the unit vector whose jth entry is 1, and l is the upper bound on $\|\mathbf{x}(a,c)\|_2$.

Let $\mathbf{1}\{\mathcal{E}\}$ be an indicator function that takes value one when $\mathcal{E}$ is true and zero otherwise. Define $N_t(a,c) := \sum_{\tau=1}^{t-1} \mathbf{1}\{a_\tau = a, c_\tau = c\}$, i.e., the number of times that arm a is pulled under context c up to time t. Denote $S_t(a,c)$ as the cumulative reward of pulling arm a under context c right before time t, i.e. $S_t(a,c) = \sum_{\tau=1}^{t-1} y_\tau \cdot \mathbf{1}\{a_\tau = a, c_\tau = c\}$ for any $c \in \mathcal{C}$. We point out that for the dummy contexts, i.e., $c = n_t+1, \ldots, n_t+d$, $S_t(i,c) = 0$ at any time t since the dummy contexts never appear.

Algorithm 1 LinUCB-d

```
Initialization: Set $\mathbf{N}_1(a) = \text{diag}[\mathbf{1}_d]$, $\mathbf{s}_1(a) = \mathbf{0}_d^\intercal$ for all $a \in [K]$.
for $t = 1 \ldots, T$ do
    Observe the incoming context $c_t$ and set
        $\alpha_t = ls + \sqrt{(2+d)\log f(t)}$.
    for $a = 1, 2, \ldots, K$ do
        Compute $\boldsymbol{\beta}_t(a) = \mathbf{N}_t(a)\mathbf{X}_t^\intercal(a)\mathbf{V}_t^{-1}(a)\mathbf{x}(a, c_t)$.
        Set
            $\hat{r}_t(a) = \mathbf{s}_t^\intercal(a)\mathbf{N}_t^{-1}(a)\boldsymbol{\beta}_t(a)$,
            $\hat{\sigma}_t(a) = \sqrt{\boldsymbol{\beta}_t^\intercal(a)\mathbf{N}_t^{-1}(a)\boldsymbol{\beta}_t(a)}$.
    end for
    Pull arm $a_t = \arg\max_{a\in[K]} \hat{r}_t(a) + \alpha_t\hat{\sigma}_t(a)$, and observe the reward $y_t$.
    Update $\mathbf{X}_{t+1}(a_t)$, $\mathbf{N}_{t+1}(a_t)$, $\mathbf{s}_{t+1}(a_t)$.
end for
```

To simplify the notation, we let $\mathbf{1}_d$ be the row vector with d 1's, and $\mathbf{0}_d$ be the row vector with d 0's. We also introduce the following matrix(vector)-form notations:

$$\begin{aligned}
\mathbf{X}_t(a) &:= [\mathbf{x}(a,1), \ldots, \mathbf{x}(a,n_t), l\mathbf{e}_1, \ldots, l\mathbf{e}_d], \\
\mathbf{N}_t(a) &:= \text{diag}[N_t(a,1), \ldots, N_t(a,n_t), \mathbf{1}_d], \\
\mathbf{s}_t(a) &:= [S_t(a,1), \ldots, S_t(a,n_t), \mathbf{0}_d]^\intercal, \\
\mathbf{V}_t(a) &:= \mathbf{X}_t(a)\mathbf{N}_t(a)\mathbf{X}_t^\intercal(a).
\end{aligned}$$

Besides, we use $\mathbf{N}_t^{-1}(a)$ to denote the pseudo-inverse of $\mathbf{N}_t(a)$ obtained by flipping its *non-zero* entries, i.e.,

$$\begin{aligned}
&\mathbf{N}_t^{-1}(a) \\
&= \text{diag}\left[\frac{\mathbf{1}\{N_t(a,1)>0\}}{N_t(a,1)}, \ldots, \frac{\mathbf{1}\{N_t(a,n_t)>0\}}{N_t(a,n_t)}, \mathbf{1}_d\right].
\end{aligned}$$

The proposed LinUCB-d algorithm is presented in Algorithm 1, where we set $f(t) := 1 + t\log^2 t$ in the expression of α_t. It adopts the Optimism in Face of Uncertainty (OFU) principle where the learner always chooses the arm with the highest potential reward after padding a UCB term.

We have a critical observation about LinUCB-d, as summarized in Proposition 1, whose proof can be found in Appendix A.

Proposition 1 *$\boldsymbol{\beta}_t(a)$ in Algorithm 1 is the solution to the following optimization problem:*

$$\begin{aligned} \underset{\boldsymbol{\beta}\in\mathbb{R}^{n_t+d}}{\text{minimize}} \quad & \boldsymbol{\beta}^\intercal \mathbf{N}_t^{-1}(a)\boldsymbol{\beta} \\ \text{subject to} \quad & \mathbf{x}(a,c_t) = \mathbf{X}_t(a)\boldsymbol{\beta}. \end{aligned} \tag{6}$$

Remark: The rationale behind Algorithm 1 can be intuitively explained as follows: For each incoming c_t, the learner needs to estimate the expected reward for each of the arms before it decides which one to pull. Due to the linear reward structure in (1), if we are able to express $\mathbf{x}(a,c_t)$ as a linear combination of the feature vectors in $\{\mathbf{x}(a,c)\}_{c\in\mathcal{C}_t\cup\mathcal{C}_0}$ in the form of $\mathbf{X}_t(a)\boldsymbol{\beta}$, then, the expected reward $r(a,c_t)$ can be expressed as $\mathbf{r}(a)\boldsymbol{\beta}$, where $\mathbf{r}(a) := \boldsymbol{\theta}(a)^\intercal\mathbf{X}(a)$. Since $\mathbf{r}(a)$ can be estimated based on observed rewards generated by pulling arm a in the past, we can then estimate $r(a,c_t)$ directly without trying to estimate $\boldsymbol{\theta}(a)$ first.

Thus, the problem boils down to obtaining a valid representation of $\mathbf{x}(a,c_t)$ in the form of $\mathbf{X}_t(a)\boldsymbol{\beta}$. The existence of such a representation can be guaranteed by including the d unit vectors associated with the dummy contexts in $\mathbf{X}_t(a)$. On the other hand, such a representation may not be unique when arm a is pulled and more feature vectors are added to $\mathbf{X}_t(a)$. That is when Proposition 1 comes into play: by minimizing the objective function in (6) subject to the linear constraint, we pick the representation that minimizes the uncertainty in the estimated $r(a,c_t)$.

We point out that inclusion of the dummy contexts introduces bias to the estimation. However, as t increases and $\mathbf{X}_t(a)$ gets expanded by including more feature vectors, the bias caused by the dummy contexts will vanish gradually. This is because under Assumptions 1.4 and 1.5, the optimal solution to (6) will put more and more weights on feature vectors associated with the observed contexts instead of the dummy ones.

Proposition 1 provides a brand new angle to view the linear contextual bandit problem. Leveraging this new viewpoint and the additional diversity assumption on the contexts, we will show that a constant regret can be achieved under LinUCB-d.

We note that LinUCB-d turns out to have deep connections with LinUCB. In order to avoid diversion from the main focus of this work, which is to elucidate the fundamental impact of context diversity on learning regret, we leave the comparison with LinUCB to Appendix B.

4 Analysis: Finite Contexts

In order to obtain some insights on how the diversity of context could help reducing the learning regret, in this section, we focus on a scenario where the context c_t is drawn in an i.i.d. fashion from a finite set $\mathcal{C}$ according to a uniform distribution. With insight obtained for this scenario, we will extend the result and analysis to a general context distribution setting in Section 5.

According to Assumption 1.5, there must exist at least one subset of d distinct contexts in $\mathcal{C}_a$, such that the corresponding feature vectors span $\mathcal{X}_a$. Denote

$$\begin{aligned} \bar{\Phi}_a &:= \arg\max_{\Phi_a} \lambda_{\min}(\Phi_a^\intercal\Phi_a), \\ \lambda_0 &:= \min_{a\in[K]} \lambda_{\min}(\bar{\Phi}_a^\intercal\bar{\Phi}_a), \end{aligned}$$

and $\bar{\mathcal{C}}_a$ as the d contexts associated with the feature vectors in $\bar{\Phi}_a$. Then, under Assumption 1.5, $\lambda_0 > 0$. Intuitively, λ_0 can be used as a metric for the diversity of context under this setting. We present our main theoretical result for the finite contexts setting as follows.

Theorem 1 *Under Assumption 1, if the context arrival c_t is uniformly distributed over a finite set $\mathcal{C}$ with $|\mathcal{C}| = n$, the expected regret under Algorithm 1 can be bounded by $O\left(Kdn^2 + \frac{dn(K+\delta^2)}{\Delta^2}\log\left(\frac{dn(K+\delta^2)}{\Delta^2}\right)\right)$, where $\delta = l\sqrt{d/\lambda_0}$.*

Theorem 1 indicates that the expected regret is bounded by a constant, which is in stark contrast to the state-of-the-art results on linear contextual bandits. It indicates that diverse contexts can indeed help to accelerate the learning process and make it converge to the optimal solution within finite steps on average. Besides, the constant bound monotonically decreases as λ_0 increases, which is consistent with our intuition that larger diversity of context is more advantageous in learning.

We point out that the dependence on the number of contexts n in the upper bound can be further reduced to a constant that does not scale in the total number of contexts, as we will show in the general context distribution setting in Section 5.

4.1 Sketch of the Proof of Theorem 1

The complete proof of Theorem 1 can be found in Appendix C. In this section, we provide a sketch of the proof to highlight the key ideas and shed light on the profound impact of context diversity to the learning performance.

The bounded regret in Theorem 1 can be intuitively explained in this way: thanks to context diversity under Assumption 1.5, arms that are suboptimal for a given context are optimal for some other contexts. Since contexts are drawn in an i.i.d. fashion, then, with high probability, each arm will be played as an optimal arm for a linear fraction of time. Context diversity then ensures that for any arm a, the feature vector $\mathbf{x}(a, c_t)$ for any incoming context c_t can be expressed as a linear combination (denote the coefficient vector as $\bar{\boldsymbol{\beta}}(a, c_t)$) of the columns of $\bar{\Phi}_a$. We note that $\{r(a, c)\}_{c \in \bar{\mathcal{C}}_a}$ can be estimated accurately based on the rewards collected when a is pulled as an optimal arm. Hence, if $\bar{\mathcal{C}}_a$ were given a priori, the error of using the linear combination of $\{r(a, c)\}_{c \in \bar{\mathcal{C}}_a}$ to predict $r(a, c_t)$ would decrease in the order of $O(1/\sqrt{t})$. To overcome the difficulty that $\bar{\mathcal{C}}_a$ is unknown beforehand, LinUCB-d greedily selects the linear combination (with coefficient vector $\boldsymbol{\beta}_t(a)$) to minimize the estimation uncertainty. Then, according to Proposition 1, the corresponding estimation uncertainty must be lower than that if $\bar{\boldsymbol{\beta}}(a, c_t)$ were used, leading to a faster decay of the prediction error.

As explained above, the key to the result in Theorem 1 is to show that each arm will be played as an optimal arm for a linear fraction of time. In order to show this, we propose a novel frame-based approach.

Specifically, we divide the time axis into frames with lengths 2^k, $k = 1, 2, \ldots$, starting at $t = 1$. Denote F_k as the time slots lying in the k-th frame, i.e., $F_k := \left\{t \mid 2^{k-1} \le t \le \min(2^k - 1, T)\right\}$. Denote $N_t(c)$ as the number of times that context c appears up to time t, and $N_{F_k}(c)$ as the number of times context c appears in F_k, i.e., $N_{F_k}(c) := N_{2^k}(c) - N_{2^{k-1}}(c)$. Similarly, we define $N_{F_k}(a, c)$ as the number of times arm a is pulled under context c in F_k. We consider the following error events:

Irregular context arrivals. For each arm $a \in [K]$, we focus on the d contexts in $\bar{\mathcal{C}}_a$. Within a frame, if the total number of arrivals of any context $c \in \bar{\mathcal{C}}_a$ is smaller than half of its *expected* number of arrivals in that frame, we term it irregular context arrivals. If irregular context arrivals happen in frame k, we will put all time indices in the $(k+1)$th frame in $\mathcal{A}_T$, i.e., $\mathcal{A}_T := \cup_k \left\{F_{k+1} \middle| \exists a, c \in \bar{\mathcal{C}}_a, \text{s.t. } N_{F_k}(c) \le \frac{1}{2n} \cdot 2^{k-1}\right\}$.

Intuitively, due to the i.i.d. context arrival assumption, the probability of having irregular context arrivals in the kth frame decays exponentially in the length of frame k. Thus, the corresponding regret over $\mathcal{A}_T$ can be bounded by a constant. The detailed analysis can be found in Appendix C.1.

Bad estimates. At time t, if the estimated reward $\hat{r}_t(a)$ deviates from its expected value $r(a, c_t)$ by more than $\alpha_t \hat{\sigma}_t(a)$, we term it a bad estimate. We group the time slots with bad estimates over $(0, T]$ in $\mathcal{B}_T$, i.e., $\mathcal{B}_T := \{t \mid \exists a \in [K], s.t. |\hat{r}_t(a) - r(a, c_t)| > \alpha_t \hat{\sigma}_t(a)\}$. The regret over $\mathcal{B}_T$ can be bounded by a constant by adapting the Laplace method (Lattimore and Szepesvári, 2019) to our setting. The detailed analysis is deferred to Appendix C.2.

Bad presence of good estimates. Within a frame, if the total number of time slots with bad estimates exceeds $\frac{1}{4n}$ of the frame length, we term the event bad presence of good estimates. If such an event happens in frame k, we put all time indices in the $(k+1)$th frame in $\mathcal{C}_T$, i.e., $\mathcal{C}_T := \cup_k \left\{F_{k+1} \middle| |\mathcal{B}_T \cap F_k| \ge \frac{1}{4n} \cdot 2^{k-1}\right\}$, where $|\mathcal{B}_T \cap F_k| := B_k$ is the number of bad estimates in frame F_k. As shown in Appendix C.3, $|\mathcal{C}_T|$ can be upper bounded by a linear function of $|\mathcal{B}_T|$. The regret over $\mathcal{C}_T$ can thus be bounded as a linear function of the regret over $\mathcal{B}_T$.

Pulling sub-optimal arms in good time slots. For any time slot t not included in $\mathcal{A}_T$, $\mathcal{B}_T$ or $\mathcal{C}_T$, we call it a good time slot. The learner may still pull a sub-optimal arm in a good time slot, due to the overlap of the confidence intervals of $r(a, c_t)$. We group the time slots when such event happens in $\mathcal{D}_T$. Specifically, $\mathcal{D}_T := \{t \mid t \notin \mathcal{A}_T \cup \mathcal{B}_T \cup \mathcal{C}_T, a_t \ne a_t^*\}$.

While the regrets over $\mathcal{A}_T$, $\mathcal{B}_T$ or $\mathcal{C}_T$ can be bounded in a relatively straightforward way, characterizing the regret over $\mathcal{D}_T$ relies on the context diversity, and is the most critical step towards the constant regret in Theorem 1. The detailed analysis is provided in Appendix C.4. It involves the following major steps:

1) Up to time t, the number of times that an arm $a \in [K]$ is chosen as a sub-optimal arm scales as $O(\log t)$ (Lemma 3).
2) Based on the definition of $\mathcal{D}_T$, for any $t \in \mathcal{D}_T$, the number of times a is pulled as an optimal arm before t scales linearly in t (Lemma 4).
3) Leveraging Proposition 1, the prediction error thus decreases in $O(1/\sqrt{t})$ (Lemma 6), which implies that $\mathcal{D}_T$ can only happen before a fixed time (Theorem 4).

After assembling the regrets over $\mathcal{A}_T$, $\mathcal{B}_T$, $\mathcal{C}_T$ and $\mathcal{D}_T$ together, the result in Theorem 1 can be obtained.

Remark: We point out that the operation of LinUCB-d itself does not depend on frames. We introduce them for the purpose of analysis only. Besides, LinUCB-d does not require the knowledge of $\bar{\mathcal{C}}_a$, $\bar{\Phi}_a$ or the distribution of c_t. It can operate under general context arrival processes, even if Assumption 1 does not hold.

5 Analysis: General Context Arrivals

In this section, we extend the analysis for the finite uniform context distribution setting to the general context distribution setting. Compared with the finite contexts case, the major difference for the general setting is that the context set $\mathcal{C}$ could be infinite and even uncountable. Although LinUCB-d still works in the same way, the corresponding performance analysis becomes much more challenging. For the finite contexts case, we choose a set of feature vectors (columns in $\bar{\Phi}_a$) as the basis for $\mathcal{X}_a$, and show that a linear combination of the corresponding empirical average rewards leads to a fast decaying estimation error, as the number of times a is pulled under contexts in $\bar{\mathcal{C}}_a$ scales linearly in time. However, for general context arrivals, the recurrence of any finite subset of contexts may have probability zero. Thus, the previous analysis cannot be extended straightfowardly to handle such case.

To overcome such challenges, we make the following modifications: First, we extend the definition of $\bar{\mathcal{C}}_a$ from d distinct contexts to d non-overlapping *meta-contexts*, where each meta-context consists of a subset of contexts with a non-zero probability mass. Thus, the meta-contexts recur frequently, similar to the finite contexts setting. One subsequent challenge associated with the meta-contexts is that feature vectors associated with the contexts in a meta-context are different and occur randomly. Thus, we cannot fix a basis (such as the columns in $\bar{\Phi}_a$ as in the finite contexts case) beforehand for $\mathcal{X}_a$, as the corresponding contexts may not appear frequently in time. Rather, it needs to be adaptively selected based on up-to-date observations. How to ensure the existence of such a valid basis at each time is thus challenging.

We construct the meta-contexts and a basis for each arm a as follows. First, we select a matrix Φ_a with $\lambda_{\min}(\Phi_a^\intercal \Phi_a) > 0$, and denote its columns as $\{\phi_a^{(i)}\}_{i=1}^d$. Assumption 1.5 ensures the existence of such Φ_a for each $a \in [K]$ according to (4). Let

$$\lambda_0(\{\Phi_a\}) := \min_{a \in [K]} \lambda_{\min}(\Phi_a^\intercal \Phi_a). \tag{7}$$

Then, we have $\lambda_0(\{\Phi_a\}) > 0$ with the selected Φ_as.

We then divide $\mathcal{X}_a$ into d disjoint groups $\{\mathcal{X}_a^{(i)}\}_{i=1}^d$ based on their closeness to $\{\phi_a^{(i)}\}_{i=1}^d$, and break the tie arbitrarily, e.g.,

$$\mathcal{X}_a^{(i)} = \left\{ \mathbf{x} \in \mathcal{X}_a \middle| \frac{\mathbf{x}^\intercal \phi_a^{(i)}}{\|\phi_a^{(i)}\|_2} < \frac{\mathbf{x}^\intercal \phi_a^{(j)}}{\|\phi_a^{(j)}\|_2} \text{ for } j < i, \right. \\ \left. \frac{\mathbf{x}^\intercal \phi_a^{(i)}}{\|\phi_a^{(i)}\|_2} \le \frac{\mathbf{x}^\intercal \phi_a^{(j)}}{\|\phi_a^{(j)}\|_2} \text{ for } j > i \right\}. \tag{8}$$

Let $r = \frac{1}{2}\sqrt{\frac{\lambda_0(\{\Phi_a\})}{d}}$, and $B(\phi_a^{(i)}, r)$ be an ℓ_2 ball centered at $\phi_a^{(i)}$ with radius r. Let $\bar{\mathcal{X}}_a^{(i)} := \mathcal{X}_a^{(i)} \cap B(\phi_a^{(i)}, r)$. Then, as shown in Lemma 7 in Appendix D, a valid basis for $\mathcal{X}_a$ can be formed if an arbitrary vector is picked from each of the subsets $\{\bar{\mathcal{X}}_a^{(i)}\}_{i=1}^d$. We then take the sample average of the previously observed feature vectors in $\bar{\mathcal{X}}_a^{(i)}$ (denoted as $\hat{\phi}_a^{(i)}$) as the corresponding basis vector.

The definition of $\bar{\mathcal{X}}_a^{(i)}$ induces the definition of meta-contexts $\bar{\mathcal{C}}_a^{(i)}$ as follows:

$$\bar{\mathcal{C}}_a^{(i)} := \{c \in \mathcal{C}_a \mid \mathbf{x}(a, c) \in \bar{\mathcal{X}}_a^{(i)}\}.$$

Let

$$p(\{\Phi_a\}) := \min_{a,i} \mathbb{P}[c_t \in \bar{\mathcal{C}}_a^{(i)}]. \tag{9}$$

Then, Assumption 1.5 ensures that there exists $\{\Phi_a\}$ such that $p(\{\Phi_a\})$ is bounded away from zero. For ease of exposition, in the following, we simply use p to denote $p(\{\Phi_a\})$ without causing ambiguity.

Denote $N_{F_k}(\bar{\mathcal{C}}_a^{(i)})$ as the total number of times that the contexts in meta-context $\bar{\mathcal{C}}_a^{(i)}$ appear up to time t. We then keep the definitions of $\mathcal{B}_T$ and $\mathcal{D}_T$ the same as in the finite context set setting and modify the definition of $\mathcal{A}_T$ and $\mathcal{C}_T$ as follows:

$$\mathcal{A}_T := \cup_k \left\{ F_{k+1} \mid \exists i, a, \text{ s.t. } N_{F_k}(\bar{\mathcal{C}}_a^{(i)}) \le \left(\frac{p}{2}\right) 2^{k-1} \right\},$$
$$\mathcal{C}_T := \cup_k \left\{ |\mathcal{B}_T \cap F_k| \ge \left(\frac{p}{4}\right) 2^{k-1} \right\}.$$

Intuitively, the regret over $\mathcal{B}_T$ remains unchanged, while the regrets over $\mathcal{A}_T$ and $\mathcal{C}_T$ can be obtained through a straightforward extension of the previous results in the finite contexts case. The challenge of the analysis thus lies in the analysis of the regret over $\mathcal{D}_T$, whose major steps are listed as follows.

1) We first show that over $\mathcal{D}_T$, the number of times that arm a is pulled as a sub-optimal arm under contexts in $\bar{\mathcal{C}}_a^{(i)}$, $b \neq a$, grows sublinearly in t (Lemma 9). Compared with Lemma 3, the random occurrences of the multiple contexts included in each meta-context incur an extra factor of $2d \log \frac{d+t}{d}$ in the upper bound.
2) We then show that the total number of times that arm a is pulled as the optimal arm under contexts in $\bar{\mathcal{C}}_a^{(i)}$ scales linearly in t (Lemma 10).
3) Since $\{\hat{\phi}_a^{(i)}\}$ is a valid basis, by leveraging Proposition 1, we show that the estimation uncertainty under LinUCB-d decays in $O(1/\sqrt{t})$ (Lemma 11).

Putting everything together, we have the following bounded regret for the general case. The detailed proof is provided in Appendix D.

Theorem 2 *Under Assumption 1, the regret under Algorithm 1 is upper bounded by* $O\left(\frac{Kd}{p^2} + \frac{d(2\delta^2+Kd)}{\Delta^2 p} \log^2\left(\frac{d(2\delta^2+Kd)}{\Delta^2 p}\right)\right)$ *for any valid choice of* $\{\Phi_a\}$*, where* $\delta := l\sqrt{d/\lambda_0(\{\Phi_a\})}$*, and* λ_0 *and* p *are defined in Eqn. (7) and (9), respectively.*

Theorem 2 indicates that even for the general context distribution setting where the contexts are drawn from a continuous set, we are still able to obtain a constant regret bound. Compared with the result in Theorem 1, the scaling in terms of d and δ is larger, due to the inclusion of multiple contexts in the meta-contexts.

Remark: Similar to the finite contexts case, Φ_a, $\hat{\bar{\phi}}_a^{(i)}$, $\bar{\mathcal{X}}_a^{(i)}$, $\bar{\mathcal{C}}_a^{(i)}$, and p are introduced for the purpose of analysis only, and are not required for LinUCB-d.

6 Experimental Evaluation

6.1 Uniform Context Arrivals

First, we consider a simplified scenario with 2 arms and 4 contexts for a proof of concept. We assume the arm parameters are $\boldsymbol{\theta}(1) = (0.8, 0.4)$, $\boldsymbol{\theta}(2) = (0.5, 0.7)$. The arm-context feature vectors are as follows: $\mathbf{x}(1,1) = (0.9, 0.1)^\intercal$, $\mathbf{x}(1,2) = (0.75, 0.25)^\intercal$, $\mathbf{x}(1,3) = (0.25, 0.75)^\intercal$, $\mathbf{x}(1,4) = (0.1, 0.9)^\intercal$, $\mathbf{x}(2,1) = (0.8, 0.2)^\intercal$, $\mathbf{x}(2,2) = (0.7, 0.3)^\intercal$, $\mathbf{x}(2,3) = (0.3, 0.7)^\intercal$, $\mathbf{x}(2,4) = (0.2, 0.8)^\intercal$. The expected rewards for pulling the arms under the four different contexts can be calculated accordingly. Therefore, arm 1 is the optimal arm under contexts 1 and 2 and arm 2 is the optimal arm under contexts 3 and 4. We can verify that $\{\mathbf{x}(1,1), \mathbf{x}(1,2)\}$ and $\{\mathbf{x}(2,3), \mathbf{x}(2,4)\}$ both span $\mathbb{R}^2$, thus they are valid basis for $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively.

With the selected parameters, we compare LinUCB-d with the following baseline algorithms through simulation: 1) UCB with $\alpha_t = \sqrt{2 \log f(t)}$ for individual contexts. We treat the arms under each context as a standard MAB and perform UCB for each context. 2) LinUCB with the same choice of α_t as in LinUCB-d. 3) A greedy LinUCB with $\alpha_t = 0$. This is the pure exploitation algorithm considered in Bastani et al. (2017) essentially.

For each algorithm, we randomly pick one out of those four contexts with probability 1/4 each time, and add i.i.d. noise according to a standard Gaussian distribution $\mathcal{N}(0, 1)$ to generate the reward. We run the simulation 100 times for each algorithm over 500,000 time slots. The sample average pseudo regrets are plotted in Fig. 1(a), where the pseudo regret is obtained by replacing y_t in the definition of regret by $r(a_t, c_t)$, and the shaded area corresponds to twice of the standard deviation. As we expect, LinUCB-d with the same choice of α_t behaves exactly the same as LinUCB, and shows bounded regret. However, the greedy algorithm and UCB do not achieve constant regret. This indicates the following: First, the pure exploitation strategy does not work well in this case. This is because the selected parameters do not satisfy the *covariate diversity* defined in Bastani et al. (2017). The covariate diversity in Bastani et al. (2017) requires that the correlation matrix of the feature vectors lying in any half space is positive definite. It requires that there are feature vectors at least in any half space. Since the feature vectors in our example only lie in the first orthant, the covariate diversity condition is not satisfied and hence the greedy approach does not work well. Second, treating each context individually does not utilize the information obtained under other contexts about the same arm, thus cannot leverage the diversity of context to reduce the regret.

Next, we evaluate how λ_0 affects the regrets. We modify the feature vectors associated with contexts 2 and 3 while keeping the rest parameters the same. Specifically, we let $\mathbf{x}(1,2) = (0.45, 0.65)^\intercal$, $\mathbf{x}(1,3) = (0.55, 0.35)^\intercal$, $\mathbf{x}(2,2) = (0.3, 0.5)^\intercal$, $\mathbf{x}(2,3) = (0.7, 0.5)^\intercal$. Compared with the previous setting, λ_0 increases from 0.00799 to 0.16917, while the reward gap Δ stays approximately the same. Intuitively, the basis vectors for each arm now point to more perpendicular directions and are more diverse in this sense. As indicated in Fig. 1(b), the increased diversity leads to much faster convergence and lower regret.

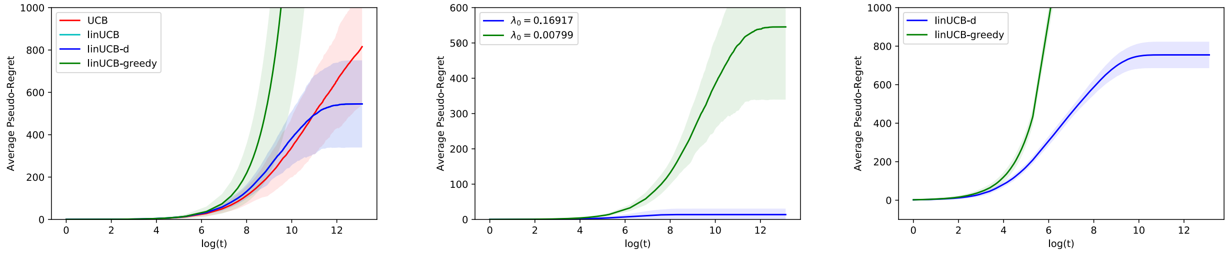
6.2 General Context Arrivals

In this part, we investigate the performance of LinUCB-d with a more general context distribution. We first randomly generate parameter vectors in $\mathbb{R}^4$ for 5 arms under the constraint that $\|\boldsymbol{\theta}(a)\|_2 = 10$. Thus, the arms are randomly located on a sphere in $\mathbb{R}^4$ with radius 10, which ensures that each of them can be optimal under certain contexts. For the feature vectors, we randomly draw $\mathbf{x}(a, c_t) \in [0,1]^4$ for $a \in \{1,2,3,4,5\}$ at each time t and make sure the reward gap condition in Assumption 1.2 is satisfied. We set $\Delta = 0.5$ throughout the simulation. The contexts are drawn from a continuous set which includes infinite many contexts.

We only compare LinUCB-d with greedy LinUCB under this setup. This is because UCB for individual contexts cannot be run without recurring contexts, and LinUCB with the same α_t behaves the same as linUCB-d. The sample average pseudo regrets are plotted in Fig. 1(c). As we observe, LinUCB-d still achieves constant regret, while the greedy algorithm does not converge.

7 Related Work

The model considered in this paper falls in the contextual bandits framework. In the contextual MAB setting, the learner repeatedly takes one of K actions in response to the observed context (Auer, 2003). Efficient exploration based on instantaneous context is of critical importance for contextual bandit algorithms to achieve small learning regret. The strongest known results (Auer, 2003; Langford and Zhang, 2008; McMa-



(a) Uniform context arrivals.

(b) Uniform context arrivals with different context diversity.

(c) General context arrivals.

Figure 1: Pseudo-regret over $\log T$. Shaded area indicates twice the standard deviation.

han and Streeter, 2009; Beygelzimer et al., 2011; Dudík et al., 2011; Agarwal et al., 2014) achieve an optimal regret after T rounds of $O(\sqrt{KT})$ with high probability.

More specifically, our reward model is similar to that of linear contextual bandits in the literature. This setting is first introduced in Auer (2003) through the LinRel algorithm and is subsequently improved through the OFUL algorithm in Dani et al. (2008) and the LinUCB algorithm in Li et al. (2010). Rusmevichientong and Tsitsiklis (2010) extend the work of Dani et al. (2008) by considering both optimistic and explore-then-commit strategies. It is shown in Abbasi-Yadkori et al. (2011) that the regret can be upper bounded by $O(d\sqrt{T})$, where d is the dimension of the context. A modified version of LinUCB, named SupLinUCB, is considered in Chu et al. (2011), and shown to achieve $O(\sqrt{dT})$ regret. Later, Valko et al. (2013) mix LinUCB and SupLinUCB with kernel functions and propose an algorithm to further reduce the regret to $O(\sqrt{\tilde{d}T})$, where $\tilde{d}$ is the effective dimension of the kernel feature space. This line of literature typically allows for arbitrary (adversarial) context sequences, and the $O(\sqrt{T})$ regret persists.

Recently, a few works start to take the diversity in contexts into consideration. Goldenshluger and Zeevi (2013) introduce a notion of diversity similar to Assumption 1.5 to a two-armed linear bandits setting. They show that the regret scales in $O(\log T)$ when a margin condition is satisfied, where the contribution from the "large-margin" covariates scales in $O(\log T)$. Bastani and Bayati (2015) generalize the notation to a so called "compatibility condition" in a contextual linear bandits model with high-dimensional covariates, and investigate a LASSO based approach. They show that the regret can be bounded by a polynomial of $\log T$ under the margin condition. The $O(\log T)$ regret persists for error events associated with large-margin covariate vectors. In contrast, we show that a *bounded* regret can be achieved, by leveraging the geometric interpretation of the diversity condition and the reward gap condition.

Bastani et al. (2017) propose a concept called *covariate diversity*, which requires that the correlation matrix of the covariate vectors lying in any half space is positive definite. Under this condition, it shows that the exploration-free greedy algorithm is near-optimal for a two-armed bandit under the stochastic setting and achieves regret in $O(\log T)$. A perturbed adversarial setting with a similar notion of diversity is studied in Kannan et al. (2018). It shows that greedy algorithms can achieve regrets in $O(\sqrt{dT})$. We note that such condition is stronger than Assumption 1.5. As illustrated through simulations in Section 6, a greedy strategy may not work well under our setting, due to the difference between the diversity definitions.

8 Conclusions

The main purpose of this paper was to study the impact of *context diversity* on the learning performance in stochastic linear contextual bandits. We have shown that, by adding an assumption that the context arrivals satisfy some diversity conditions, it is possible to significantly reduce the learning regret of contextual bandits. We proposed an algorithm called LinUCB-d and showed that when the diversity assumption is satisfied, the expected regret can in fact be upper bounded by a constant. This study illustrates the power of incorporating structure in the contexts to the bandit problem. It is of interest to evaluate whether other structures of the context can be similarly considered, and what their impacts would be. Another interesting problem is to study the impact of context diversity in other settings, such as the perturbed adversarial setting (Kannan et al., 2018).

Acknowledgements

JY acknowledges the support from U.S. National Science Foundation under Grant ECCS-1650299.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 2312–2320.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. E. (2014). Taming the monster: A fast and simple algorithm for contextual bandits. In *In Proceedings of the 31st International Conference on Machine Learning*, pages 1638–1646.

Agrawal, S. and Goyal, N. (2012). Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1.

Agrawal, S. and Goyal, N. (2013). Further optimal regret bounds for Thompson sampling. In *Artificial Intelligence and Statistics*, pages 99–107.

Auer, P. (2003). Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3:397–422.

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

Bastani, H. and Bayati, M. (2015). Online decision-making with high-dimensional covariates. *SSRN Electronic Journal*.

Bastani, H., Bayati, M., and Khosravi, K. (2017). Mostly exploration-free algorithms for contextual bandits. *CoRR*, abs/1704.09011.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 19–26, Fort Lauderdale, FL, USA.

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

Chen, L. and Xu, J. (2019). Budget-constrained edge service provisioning with demand estimation via bandit learning. *IEEE Journal on Selected Areas in Communications*, 37(10):2364–2376.

Chu, W., Li, L., Reyzin, L., and Schapire, R. E. (2011). Contextual bandits with linear payoff functions. In *AISTATS*, volume 15, pages 208–214.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory - COLT*, pages 355–366.

Dudík, M., Hsu, D. J., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. (2011). Efficient optimal learning for contextual bandits. *CoRR*, abs/1106.2369.

Goldenshluger, A. and Zeevi, A. (2013). A linear response bandit problem. *Stoch. Syst.*, 3(1):230–261.

Kannan, S., Morgenstern, J. H., Roth, A., Waggoner, B., and Wu, Z. S. (2018). A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems 31*, pages 2227–2236.

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

Langford, J. and Zhang, T. (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, pages 817–824.

Lattimore, T. and Szepesvári, C. (2019). *Bandit Algorithms*. Cambridge University Press (preprint).

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670.

McMahan, H. B. and Streeter, M. J. (2009). Tighter bounds for multi-armed bandits with expert advice. In *COLT*.

Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Math. Oper. Res.*, 35:395–411.

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. (2013). Finite-time analysis of kernelised contextual bandits. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pages 654–663.