Osom: A simultaneously optimal algorithm for multi-armed and linear contextual bandits

Niladri S. Chatterji
UC Berkeley

Vidya Muthukumar
UC Berkeley

Peter L. Bartlett
UC Berkeley

Abstract

We consider the stochastic linear (multi-armed) contextual bandit problem with the possibility of hidden *simple multi-armed bandit* structure in which the rewards are independent of the contextual information. Algorithms that are designed solely for one of the regimes are known to be sub-optimal for their alternate regime. We design a single computationally efficient algorithm that simultaneously obtains problem-dependent optimal regret rates in the simple multi-armed bandit regime and minimax optimal regret rates in the linear contextual bandit regime, without knowing a priori which of the two models generates the rewards. These results are proved under the condition of stochasticity of contextual information over multiple rounds. Our results should be viewed as a step towards principled data-dependent policy class selection for contextual bandits.

1 Introduction

The *contextual bandit* paradigm involves sequential decision-making settings in which we repeatedly pick one out of K actions (or "arms") in the presence of contextual side information. Algorithms for this problem usually involve policies that map the contextual information to a chosen action, and the reward feedback is typically *limited* in the sense that it is only obtained for the action that was chosen. The goal is to maximize the total reward over several (n) rounds of decision-making, and the performance of an online algorithm is typically measured in terms of *regret* with respect to the best policy within some policy class Π that is fixed a priori. Applications of this paradigm include advertisement placement/web article recommendation [Li+10; Aga+16], clinical trials and mobile health-care [Woo79; TM17].

The contextual bandit problem can be thought of as an online supervised learning problem (over policies mapping contexts to actions) with limited information feedback, and so the optimal regret bounds scale like $\mathcal{O}(\sqrt{Kn\log|\Pi|})$, a natural measure of the sample complexity of the policy class [Aue+02; MS09; Bey+11]. These are typically achieved by algorithms that are inefficient (linear in the size of the policy class). Much of the research in contextual bandits has tackled computational efficiency [LZ08; Aga+14; RS16; SKS16; Syr+16; FK18]: do there exist computationally efficient algorithms that achieve the optimal regret guarantee? A question that has received relatively less attention involves the choice of policy class itself. Even for a fixed regret-minimizing algorithm, the choice of policy class is critical to maximize the overall *reward* of the algorithm. As can be seen in applications of contextual bandits models for article recommendation [Li+10], the choice is often made in hindsight, and more complex policy classes are used if the algorithm is run for more rounds. A quantitative understanding of how to do this is still lacking, and intuitively, we should expect the optimal choice of policy class to not be static. Ideally, we could design adaptive contextual bandit algorithms that would initially use simple policies, and switch over to more complex ones as more data is obtained.

Theoretically, what this means is that the regret bounds derived for a contextual bandit algorithm are only meaningful for rewards that are generated by a policy within the policy class to which the algorithm is tailored. If the rewards are derived from a "more complex" policy outside the policy class, even the optimal policy may neglect obvious patterns and obtain a very low reward. If the rewards are derived from a policy that is expressible by a much smaller class, the regret that is accumulated is unnecessary. Let us view this

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

through the lens of the simplest possible example: the standard linear contextual bandits [Chu+11] paradigm, where we can choose one out of K arms and rewards are generated according to the process

$$g_{i,t} = \mu_i + \langle \theta^*, \alpha_{i,t} \rangle + \eta_{i,t}, \text{ for all } i \in [K],$$

where μ_i represents a "bias" of arm i, $\theta^* \in \mathbb{R}^d$ represents the linear parameter of the model (which is shared across all arms[1]), $\alpha_{i,t} \in \mathbb{R}^d$ represents the contextual information and $\{\eta_{i,t}\}_{t=1}^n$ represents noise in the reward observations. It is well-known that variants of linear upper confidence bound algorithms like LinUCB [Chu+11] and OFUL [APS11][2] suffer at most $\widetilde{\mathcal{O}}((\sqrt{d}+\sqrt{K})\sqrt{n})$ regret with respect to the optimal linear policy. However, setting $\theta^* = 0$ yields the important case of the reward distribution being independent from the contextual information. Here, a simple upper confidence bound algorithm like UCB [ACF02] would yield the optimal $\mathcal{O}(\log n)$ regret bound, *which does not depend on the dimension of the contexts d*. Thus, we pay substantial extra regret by using the algorithm meant for linear contextual bandits on such instances with much simpler structure. On the other hand, upper confidence bounds that ignore the contextual information will not guarantee any control on the policy regret: it can even be linear. It is natural to desire a single approach that adapts to the inherent complexity of the reward-generating model and obtains the optimal regret bound as if this complexity was known in hindsight. Specifically, this paper seeks an answer to the following question:

Does there exist a single algorithm that simultaneously achieves the $\mathcal{O}(\log n)$ regret rate on simple multi-armed bandit instances and the $\widetilde{\mathcal{O}}((\sqrt{d}+\sqrt{K})\sqrt{n})$ regret rate on linear contextual bandit instances?

1.1 Our contributions

We answer the question of simultaneously optimal regret rates in the multi-armed ("simple") bandit regime and the linear contextual ("complex") bandit regime affirmatively under the condition that the contexts are generated from a stochastic process that yields covariates that are not ill-conditioned. Our algorithm, OSOM (for Optimistic Selection of Models), essentially exploits the best policy (simply the best arm) that is learned under the assumption of the simple reward model - while conducting a sequential statistical test for the presence of additional complexity in the model, and particularly *whether ignoring this additional complexity would lead to substantial regret*. This is a simple statistical principle that could conceivably be generalized to arbitrary policy classes *that are nested*: we will see that the OSOM algorithm critically exploits the nested structure of the simple bandit model within the linear contextual model.

1.2 Related work

The contextual bandit paradigm was first considered by Woodroofe (1979) to model clinical trials. Since then it has been studied intensely both theoretically and empirically in many different application areas under many different pseudonyms. We point the reader to [TM17] for an extensive survey of the contextual bandits history and literature.

Treating policies as experts (EXP4 [Aue+02]) with careful control on the exploration distribution led to the optimal regret bounds of $\mathcal{O}(\sqrt{Kn\log|\Pi|})$ in a number of settings. From an *efficiency* point of view (where efficiency is defined with respect to an *arg-max-oracle* that is able to compute the best greedy policy in hindsight), the first approach conceived was the epoch-greedy approach [LZ08], that suffers a sub-optimal dependence of $n^{2/3}$ in the regret. More recently, "randomized-UCB" style approaches [Aga+14] have been conceived that retain the optimal regret guarantee with $\widetilde{\mathcal{O}}(\sqrt{n})$ calls to the arg-max-oracle. This question of computational efficiency has generated a lot of research interest [RS16; SKS16; Syr+16; FK18]. The problem of policy class selection itself has received less attention in the research community, and how this is done in practice in a statistically sound manner remains unclear. An application of *linear* contextual bandits was to personalized article recommendation using hand-crafted features of users [Li+10]: two classes of linear contextual bandit models with varying levels of complexity were compared to simple (multi-armed) bandit algorithms in terms of *overall reward* (which in this application represented the click-through rate of ads). A striking observation was that the more complex models won out when the algorithm was run for a longer period of time (eg: 1 day as opposed to half a day). Surveys on contextual bandits as applied to mobile health-care [TM17] have expressed a desire for algorithms that adapt their choice of policy class according to the amount of information they have received (e.g. the number of rounds). At a

[1] This is the model that was described in [Chu+11]. It is worth noting that more complex variants of this model with a separate θ_i^* for every $i \in [K]$ have also been empirically evaluated [Li+10].

[2] Guarantees for OFUL were established under slightly different constraints on θ^* and the context vectors which led to a regret bound of $\widetilde{\mathcal{O}}((d+K)\sqrt{n})$. We show in Lemma 6 that a slight variant of OFUL has its regret bounded by $\widetilde{\mathcal{O}}((\sqrt{d}+\sqrt{K})\sqrt{n})$ in our setting. The regret bound scales with $\sqrt{d}+\sqrt{K}$ in our setting as the linear policy has dimension $d+K$.

high level, we seek a theoretically principled way of doing this.

Perhaps the most relevant work to online policy class selection involves significant attempts to *corral* a band of M base bandit algorithms into a meta-bandit framework [Aga+17]. The idea is to bound the regret of the meta-algorithm in terms of the regret of the best base algorithm in hindsight. (This is clearly useful for policy class selection that we study here – by corralling together an algorithm designed for the linear model and one for the simple multi-armed bandits model.) The Corral framework is very general and can be applied to any set of base algorithms, whether efficient or not. This generality is attractive, but it is not the optimal choice of *computationally efficient* algorithm for the multi-armed-vs-linear-contextual bandit problem for a couple of reasons.

1. It is not clear what (if any) choice of base algorithms would lead to a computationally efficient algorithm that is also statistically optimal in a minimax sense simultaneously for both problems.

2. The meta-algorithm framework uses an experts algorithm (in particular, mirror descent with log-barrier regularizer and importance weighting on the base algorithms) to choose which base algorithm to play in each round. Thus, it is impossible to expect the instance-optimal regret rate of $\mathcal{O}(\log n)$ on the simple bandit instance. More generally, the Corral framework will not yield instance-optimal rates on any policy class[3].

The Corral framework highlights the principal difficulty in contextual bandit model selection that can be thought of as an even finer exploration-exploitation tradeoff: algorithms (designed for particular model classes) that fall out of favor in initial rounds could be picked very rarely and the information required to truly perform model selection may be absent even after many rounds of play. CORRAL tackles this difficulty using the log-barrier regularizer for the meta-algorithm as a natural form of heightened exploration [Fos+16], together with clever learning rate schedules.

Closely related is the concurrent work of [FKL19] which tackles the problem of selecting among a hierarchy of linear classes with growing dimension. They work with stochasticity assumptions on the contexts that are *weaker* than the assumptions that we make in our paper. However, they are only able to establish a sub-optimal bound on the regret of $\widetilde{\mathcal{O}}(d_*^{1/3}n^{2/3})$ (where d_* is dimension of the optimal linear policy) as opposed to the minimax optimal regret rates (that scale with $n^{1/2}$) which we establish in our paper.

Our stylistic approach to the model selection problem is a little different, as we focus on the much more specific case of 2 models: the simple multi-armed bandit model and the linear contextual bandit model. We encounter a similar difficulty and obtain striking clarity on the extent of this difficulty owing to the simplicity of the models. On the other hand, we observe that commonly encountered sequences of contexts can help us carefully navigate the finer exploration-exploitation tradeoff when the model classes are nested.

Our algorithm (OSOM) utilizes a simple "best-of-both-worlds" principle: exploit the possible simple reward structure in the model until (unless) there is significant statistical evidence for the presence of complex reward structure *that would incur substantial complex policy regret if not exploited.* This algorithmic framework is inspired by the initial "best-of-both-worlds" results for stochastic and adversarial multi-armed bandits; in particular, the "Stochastic and Adversarial Optimal" (SAO) algorithm [BS12] (although the details of the phases of the algorithm and the statistical test are very different). In that framework, instances that are not stochastic (and could be thought of as "adversarial") are not always detected as such by the test. The test is designed in an elegant manner such that the regret is optimally bounded on instances that are not detected as adversarial, *even if an algorithm meant for stochastic rewards is used.* Our test to distinguish between simple and complex instances shares this flavor – in fact, all theoretically complex instances ($\theta^* \neq 0$) are not detected as such.

Also related are results on contextual bandits with similarity information *on the contexts*, which automatically encodes a potentially easier learning problem [Sli14]. The main novelty in these results involves adapting to such similarity online.

Technically, our proofs leverage the most recent set of theoretical results on regret bounds for linear bandits [APS11], which can easily be applied to the linear contextual bandit model, and sophisticated *self-normalized* concentration bounds for our estimates of both the bias terms μ_i and the parameter vector θ^*. For the latter, we find that the Matrix Freedman inequality [Oli09; Tro11] is particularly useful.

[3] On our much simpler instance of bandit-vs-linear-bandit, we do obtain instance-optimal rates for at least the simple bandit model.

1.3 Problem statement

At the beginning of each round $t \in [n]$, the learner is required to choose one of K arms and gets a *reward* associated with that arm. To help make this choice the learner is handed a context vector at every round $\alpha_t = [\alpha_{1,t}, \ldots, \alpha_{K,t}] \in \mathbb{R}^{d \times K}$ (this is essentially a concatenation of K vectors, each of dimension equal to d). Let $g_{i,t}$ denote the reward of arm i and let $A_t \in [K]$ denote the choice of the learner in round t. The rewards could be arriving from one of two models that is described below:

Simple Model: Under the simple *multi-armed* bandit model, the mean rewards of K arms are fixed and are *not* a function of the contexts. That is, at each round

$$g_{i,t} = \mu_i + \eta_{i,t}, \qquad \forall i \in [K]$$

where $\mu_i \in [-1, 1]$, $\{\eta_{i,t}\}_{i=1}^K$ are identical, independent, zero mean, σ-sub-Gaussian noise (defined below). Let the arm with the highest reward have mean μ^* and be indexed by i^*. The benchmark that the algorithm hopes to compete against is the *pseudo-regret* (henceforth regret for brevity),

$$R_n^s := n\mu^* - \sum_{s=1}^n \mu_{A_s}.$$

Define the gap as the difference in the mean rewards of the best arm compared to the mean reward of the i^{th} arm, that is, $\Delta_i := \mu^* - \mu_i$. Previous literature on multi-armed bandits [LR85] tells us that the best one can hope to do in this setting in the worst case is $\mathbb{E}[R_n^s] = \Omega(\sum_i \log(n)/\Delta_i)$. Several algorithms like UCB [ACF02] and MOSS [AB10; DP16] achieve this lower bound up to logarithmic (and constant) factors.

Complex Model: In this model the mean reward of each arm is a linear function of the contexts (linear contextual bandits). We work with the following stochastic assumptions on the context vectors. Each of these contexts vectors $\alpha_{i,t} \in \mathbb{B}_2^d(1)$ and are drawn independent of the past from a distribution such that $\alpha_{i,t}$ is independent of $\{\alpha_{j,t}\}_{j \neq i}$ and, $\forall i \in [K]$ and $\forall t \in [n]$,

$$\begin{aligned}
\mathbb{E}_{t-1}[\alpha_{i,t}] &:= \mathbb{E}\left[\alpha_{i,t} \middle| \{\eta_{j,s}, \alpha_{j,s}\}_{j \in [K], s \in [t-1]}\right] = 0, \\
\mathbb{E}_{t-1}\left[\alpha_{i,t}\alpha_{i,t}^\top\right] &:= \mathbb{E}\left[\alpha_{i,t}\alpha_{i,t}^\top \middle| \{\eta_{j,s}, \alpha_{j,s}\}_{j \in [K], s \in [t-1]}\right] \\
&= \Sigma_c \succeq \rho_{\min} \cdot I. \qquad (1)
\end{aligned}$$

The conditional mean of the context vectors are 0 and the co-variance matrix has its minimum eigenvalue *bounded below* by $\rho_{\min}$. Assumptions similar to the one above have also been made in past work on linear contextual bandits [BBK17; Kan+18; Rag+18].

It is important to note here that algorithms designed solely for the linear contextual bandit problem, like OFUL, work for stochastic conditional rewards regardless of the sequence of contexts, which can be chosen *adversarially*. However, our goal here is to optimally adapt to simpler model structure while retaining the contextual bandit regret guarantee. Currently designed algorithms tailored to the linear contextual bandits problem, like OFUL, will fail at this objective even under the stochastic assumption. Our stochastic assumption essentially constitutes a *sufficient* condition for optimal model selection in linear contextual bandits. Whether it is *necessary*, that is, whether model selection is possible for the case of adversarial contexts, is an intriguing question left to future work.

In this complex model, we assume there exists an underlying linear predictor $\theta^* \in \mathbb{R}^d$ and *biases* $[\mu_1, \ldots, \mu_K] \in \mathbb{R}^K$ of the K arms, such that the mean rewards of the arms are affine functions of the contexts, that is,

$$g_{i,t} = \mu_i + \langle \theta^*, \alpha_{i,t} \rangle + \eta_{i,t}.$$

We impose compactness constraints on the parameters: in particular, we have $\mu_i \in [-1, 1]$, $\theta^* \in \mathbb{B}_2^d(1)$. Further, the noise $\{\eta_{i,t}\}_{t=1}^n$ are identical, independent, zero mean, and σ-sub-Gaussian. Clearly, simple model instances (which are parameterized only by the biases $[\mu_1, \ldots, \mu_K] \in \mathbb{R}^K$) can be expressed as complex model instances by setting $\theta^* = 0$.

At each round define $\kappa_t = \text{argmax}_{\kappa \in \{1,\ldots,K\}_{i=1}^K} \{\mu_\kappa + \langle \theta^*, \alpha_{\kappa,t} \rangle\}$ to be the best arm at round t. Here, we define pseudo-regret with respect to the optimal policy under the generative linear model:

$$R_n^c := \sum_{s=1}^n \left[\mu_{\kappa_s} + \langle \theta^*, \alpha_{\kappa_s,s} \rangle - \mu_{A_s} - \langle \theta^*, \alpha_{A_s,s} \rangle\right].$$

As noted above, past literature on this problem yielded algorithms like LinUCB [Chu+11] and OFUL [APS11] that only suffer from the minimax regret of $\widetilde{\mathcal{O}}((\sqrt{d} + \sqrt{K})\sqrt{n})$. As we will see in the simulations, these algorithms actually incur the dependence on the dimension in the regret, even for simple instances.

Notation and definitions. Given a vector v, let v_i denote its i^{th} component. For a vector we let $\|v\|_p$ for $p \in [1, \infty]$ denote the ℓ_p-norm. Given a matrix M we denote it's operator norm by $\|M\|_{op}$, and use $\|M\|_F$ to denote its Frobenius norm. Given a symmetric matrix S

let $\gamma_{\max}(S)$ and $\gamma_{\min}(S)$ denote its largest and smallest eigenvalues. Given a positive definite matrix V we define the norm of a vector w with respect to matrix V as $\|w\|_V^2 = w^\top V w$. Let $\{\mathcal{F}_t\}_{t=1}^{\infty}$ be a filtration. A stochastic process $\{\xi_t\}_{t=1}^{\infty}$ where ξ_t is measurable with respect to $\mathcal{F}_{t-1}$ is defined to be conditionally σ-sub-Gaussian for some $\sigma > 0$ if, for all $\lambda \in \mathbb{R}$, we have, $\mathbb{E}\left[e^{\lambda \xi_t} \middle| \mathcal{F}_{t-1}\right] \le \exp(\lambda^2\sigma^2/2)$.

2 Construction of Confidence Sets

In our algorithm, which is presented subsequently at the *end of round* t, we build an upper confidence estimate for each arm. Let $T_i(t) := \sum_{s=1}^{t} \mathbb{I}\left[A_s = i\right]$ be the number of times arm i was pulled and $\bar{g}_{i,t} := \sum_{s=1}^{t} g_{i,s}\mathbb{I}\left[A_s = i\right]/T_i(t)$ be the average reward of that arm at the end of round t. For each arm we define the upper confidence estimate as follows,

$$\tilde{\mu}_{i,t} := \bar{g}_{i,t} + \sigma\left[\frac{1+T_i(t)}{T_i^2(t)}\left(1 + 2\log\left(\frac{K(1+T_i(t))^{\frac{1}{2}}}{\delta}\right)\right)\right]^{\frac{1}{2}}. \tag{2}$$

Lemma 6 in [APS11] (restated below as Lemma 1 here) uses a refined self-normalized martingale concentration inequality to bound $|\mu_i - \bar{g}_{i,t}|$ across all arms and all rounds.

Lemma 1. *Under the simple model, with probability at least* $1-\delta$ *we have,* $\forall i \in \{1,\dots,K\}, \forall t \ge 0$,

$$|\mu_i - \bar{g}_{i,t}| \le \sigma\left[\frac{1+T_i(t)}{T_i^2(t)}\left(1 + 2\log\left(\frac{K(1+T_i(t))^{\frac{1}{2}}}{\delta}\right)\right)\right]^{\frac{1}{2}}.$$

For any round $t > K$, let $\hat{\theta}_t$ be the ℓ^2-regularized least-squares estimate of θ^* defined below.

$$\hat{\theta}_t = \left(\boldsymbol{\alpha}_{K+1:t}^\top \boldsymbol{\alpha}_{K+1:t} + I\right)^{-1} \boldsymbol{\alpha}_{K+1:t}^\top \mathbf{G}_{K+1:t}, \tag{3}$$

where $\boldsymbol{\alpha}_{K+1:t}$ is the matrix whose rows are the context vectors selected from round $K+1$ up until round t: $\alpha_{A_{K+1},K+1}^\top, \dots, \alpha_{A_t,t}^\top$ and $\mathbf{G}_{K+1:t} = [g_{A_{K+1},K+1} - \tilde{\mu}_{A_{K+1},K}, \dots, g_{A_t,t} - \tilde{\mu}_{A_t,t-1}]^\top$. Here we are regressing on the rewards seen to estimate θ^*, while using the bias estimates $\tilde{\mu}_{i,t-1}$ obtained by our upper confidence estimates defined in Eq. (2).

Lemma 2. *Let* $\hat{\theta}_t$ *be defined as in Eq. (3). Then, with probability at least* $1-3\delta$ *we have that for all* $t > K$, θ^* *lies in the set*

$$\mathcal{C}_t^c := \left\{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|_2 \le \mathcal{K}_\delta(t,n)\right\}, \tag{4}$$

where $\mathcal{K}_\delta(t,n) = \widetilde{\mathcal{O}}(\sigma\sqrt{d\cdot n})$ *is defined in Eq. (8d).*

We prove this lemma in Appendix B.

Algorithm 1: OSOM (Optimistic Selection Of Models)

```
for $t = 1, \dots, K$ do
    Play arm $t$ and receive reward $g_{t,t}$,
    (Play each arm at least once.)
for $t = K+1, \dots, n$ do
    Current Model ← 'Simple'
    Simple Model Estimate:
        $i_t \in \operatorname{argmax}_{i \in \{1,\dots,K\}} \{\tilde{\mu}_{i,t-1}\}$    (5)
    Complex Model Estimate:
        $j_t, \tilde{\theta}_t \in \operatorname{argmax}_{i \in \{1,\dots,K\}, \theta \in \mathcal{C}_{t-1}^c} \{\tilde{\mu}_{i,t-1} + \langle \alpha_{i,t}, \theta \rangle\},$    (6)
    where $\mathcal{C}_{t-1}^c$ defined in Eq. (4).
    if Current Model = 'Simple' and $t > K+1$ then
        Check the condition:
            $\sum_{s=K+1}^{t-1} \left\{\tilde{\mu}_{j_s,s-1} + \langle \alpha_{j_s,s}, \tilde{\theta}_s \rangle - g_{i_s,s}\right\} \le \mathcal{W}_\delta(t,n),$    (7)
        where $\mathcal{W}_\delta(t,n)$ defined in Eq. (8e).
        If violated then:
            Current Model ← 'Complex'.
    If Current Model = 'Simple': Play arm $i_t$ and receive reward $g_{i_t,t}$.
    Else if Current Model = 'Complex': Play arm $j_t$ and receive $g_{j_t,t}$.
    Update $\{\tilde{\mu}_{i,t}\}_{i=1}^{K}$ and $\mathcal{C}_t^c$.
```

3 Algorithm and main result

The intuition behind Algorithm 1 is straightforward. The algorithm starts off by using the simple model estimate of the recommended action, that is, i_t; until it has reason to believe that there is a benefit from switching to the complex model estimates. If the rewards are truly coming from the simple model, *or from a complex model that is well approximated by a simple multi-armed bandit model*, then Condition 7 *will not be violated* and the regret shall continue to be bounded under either model. However, if Condition 7 *is violated* then algorithm switches to the complex estimates – j_t for the remaining rounds. The condition is designed using the function $\mathcal{W}_\delta(t,n)$ which is of the order $\widetilde{\mathcal{O}}(\sigma(\sqrt{d}+\sqrt{K})\sqrt{t})$. This corresponds to the additional regret incurred when we attempt to estimate the extra parameter – $\tilde{\theta}_t \in \mathbb{R}^d$.

At each round Condition 7 compares the algorithm's *estimate* for the cumulative reward that could be obtained by playing according to the complex estimates – $\sum_{s=K+1}^{t-1} \tilde{\mu}_{j_s,s-1} + \langle \alpha_{j_s,s}, \tilde{\theta}_s \rangle$ – with the actual cumulative rewards seen so far $\sum_{s=K+1}^{t-1} g_{i_s,s}$ by sticking to the simple estimates.

Under the simple model, given our construction of the confidence sets the term $\sum_{s=K+1}^{t-1} \langle \alpha_{j_s,s}, \tilde{\theta}_s \rangle$ will be bounded by $\widetilde{\mathcal{O}}((\sqrt{d}+\sqrt{K})\sqrt{t})$ as the true underlying vector $\theta^* = 0$. While the remaining terms $\sum_{s=K+1}^{t-1} \tilde{\mu}_{j_s,s-1} - g_{i_s,s}$ shall be at most $\widetilde{\mathcal{O}}(\sqrt{Kt})$; as the simple estimates (i_s) shall be picking out the best arm quite often under the simple model. In fact under this model we show in Lemma 4 that Condition 7 is *not violated* with high probability and the algorithm shall continue using simple estimates throughout its entire run.

On the other hand, under the complex model, we switch to the complex estimates only if the difference between the algorithm's estimate for the cumulative reward that could be obtained by playing according to the complex estimates – $\sum_{s=K+1}^{t-1} \tilde{\mu}_{j_s,s-1} + \langle \alpha_{j_s,s}, \tilde{\theta}_s \rangle$ – exceeds the rewards seen so far $\sum_{s=K+1}^{t-1} g_{i_s,s}$ by $\widetilde{\mathcal{O}}((\sqrt{d}+\sqrt{K})\sqrt{t})$. That is, only when the algorithm starts to suffer a regret that is equal to the minimax rate of regret. While instead if this condition is not violated under the complex model, that is, our estimated cumulative reward for switching to the complex model is close to the rewards seen far. Then we show that the regret under the complex model is small even by using simple estimates. We do this in Lemma 5.

By combining the arguments outlined above our main theorem optimally bounds the regret of OSOM under either of the two reward-generating models.

Theorem 3. *With probability at least* $1-9\delta$*, we obtain the following upper bounds on regret for the algorithm* OSOM *(Algorithm 1):*

(a) Under the Simple Model*:*

$$R_n^s \le \sigma \cdot \sum_{i:\Delta_i>0} \left[3\Delta_i + \frac{16}{\Delta_i} \log\left(\frac{2K}{\Delta_i \delta}\right) \right].$$

(b) Under the Complex Model*:*

$$R_n^c \le 4(K+1) + 4\mathcal{W}_\delta(n,n) = \widetilde{\mathcal{O}}\left\{\sigma(\sqrt{d}+\sqrt{K})\sqrt{n}\right\},$$

where $\mathcal{W}_\delta(n,n)$ *is defined in Eq.* (8e).

Notice that Theorem 3 establishes regret bounds on the algorithm OSOM which are minimax optimal under both *simple model* and the *complex model* up to logarithmic factors. In fact, under the simple model we are able to obtain *problem-dependent* regret rates.

In the complex model we match the minimax rates obtained by OFUL (which holds for adversarial contexts as well). A natural question is if it is also possible to obtain problem dependent rates in the complex model simultaneously. For example under the complex model by using OFUL it is possible to show that regret grows poly-logarithmically with n: $R_n^c \le \widetilde{\mathcal{O}}\left((d+K)^2/\Delta_\ell\right)$, where Δ_ℓ is an appropriately defined *gap* in the linear model.

Proof and key lemmas. We present the key lemmas below (see a proof of these lemmas in Appendix A) and use them to prove our main theorem. To prove Theorem 3, we need to show that the regret of OSOM is bounded under either underlying model. In Lemma 4 we demonstrate that whenever the rewards are generated under the simple model, Condition 7 is *not violated* with high probability.

Lemma 4. *Assume that rewards are generated under the simple model. Then, with probability at least* $1-5\delta$*, we have for all* $t \in \{K+2,\ldots,n\}$*:*

$$\sum_{s=K+1}^{t-1} \left[\tilde{\mu}_{j_s,s-1} + \langle \alpha_{j_s,s}, \tilde{\theta}_s \rangle\right] - \sum_{s=K+1}^{t-1} g_{i_s,s} < \mathcal{W}_\delta(t-1,n).$$

This ensures that when the data is generated from the simple model, we have that the Boolean variable `Current Model` = 'Simple' throughout the run of the algorithm. Thus, the regret is equal to the regret incurred by the UCB algorithm, which is meant for simple model instances.

On the other hand, when the data is generated according to the complex model, we first demonstrate in Lemma 5 that the regret remains appropriately bounded if Condition 7 is *not violated*.

Lemma 5. *For all* $t \in \{K+1,\ldots,n\}$*. Let Condition 7 not be violated up until round* $t+1$*, that is,*

$$\sum_{s=K+1}^{t} \left\{\tilde{\mu}_{j_s,s-1} + \langle \alpha_{j_s,s}, \tilde{\theta}_s \rangle - g_{i_s,s}\right\} \le \mathcal{W}_\delta(t,n).$$

Then, we have $R_t^c \le 4K + 2\mathcal{W}_\delta(t,n)$*, with probability at least* $1-5\delta$.

While when the data is generated according to the complex model and if the condition does get violated at a certain round, we switch to the estimates of the complex model, that is, j_t. This corresponds to a variant of the algorithm OFUL, which is meant for complex instances. Thus, the regret remains bounded in the subsequent rounds under this event as well (formally proved in Lemma 6 in Appendix A). Combining the results of these three lemmas yields the regret bound.

Proof [of Theorem 3] **Part (a):** We have established in Lemma 4 that Condition 7 is *not violated* with probability at least $1 - 5\delta$ under the simple model. Conditioned on this event, OSOM plays according to the simple model estimate, i_t, for all rounds. Invoking Theorem 7 in [APS11] gives us that with probability at least $1 - \delta$, $R_n^s \leq \sum_{i:\Delta_i > 0} 3\Delta_i + (16/\Delta_i)\log(2K/\Delta_i\delta)$. Applying the union bound over these two events gives this regret bound with probability at least $1 - 6\delta$.

Part (b): One out the two disjoint events are possible under the complex model.

Case 1: In this event Condition 7 is never violated throughout the run of the algorithm. Then by Lemma 5 we have

$$R_n^c \leq 4K + 2\mathcal{W}_\delta(n, n)$$

with probability at least $1 - 5\delta$.

Case 2: The other event is when Condition 7 is violated in round $\tau_* < n$. We know by Lemma 5:

$$R_{\tau_* - 2}^c \leq 4K + 2\mathcal{W}_\delta(n, n)$$

with probability at least $1 - 5\delta$. Also, by Lemma 6:

$$\begin{aligned} R_{\tau_*:n}^c &:= \sum_{s=\tau_*}^{t} \left[\mu_{\kappa_s} + \langle \theta^*, \alpha_{\kappa_s, s}\rangle - \mu_{A_s} - \langle \theta^*, \alpha_{A_s, s}\rangle\right] \\ &\leq 2\mathcal{W}_\delta(n, n) \end{aligned}$$

with probability at least $1 - 4\delta$. We can decompose the cumulative regret up to round n as follows:

$$R_n^c \leq R_{\tau_* - 1}^c + R_{\tau_*:n}^c + 4,$$

where $R_{\tau_*:n}^c$ denotes the regret of the algorithm starting from round τ^* up to round n and the 4 appears as it is the maximum regret that could be incurred in round τ_* by the algorithm under the complex model. By taking a union bound and using the decomposition of the regret above, we get $R_n^c \leq 4(K + 1) + 4\mathcal{W}_\delta(n, n)$, with probability at least $1 - 9\delta$. ■

4 Experiments

To experimentally corroborate our claims, we ran our model-selecting algorithm, OSOM, on both simple and complex instances. We compared its performance to that of UCB (which is optimal up to logarithmic factors under the simple model) and OFUL (which is minimax optimal under the complex model). Complete experimental details are provided in Appendix D.

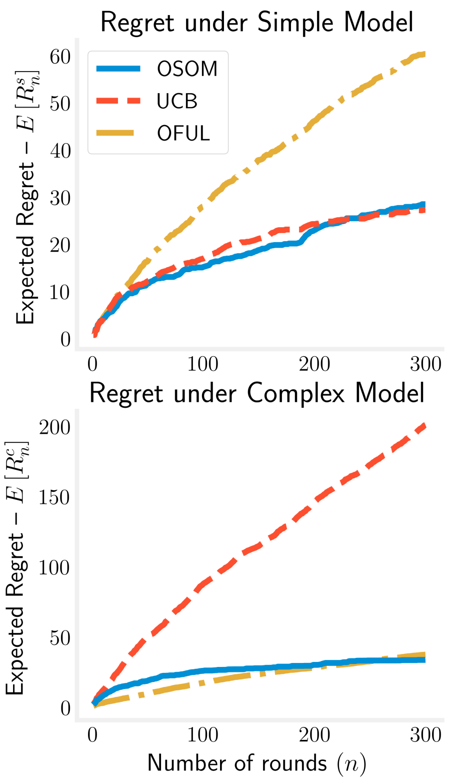

Figure 1. Experiments on synthetic data with $K = 5$, $d = 50$ and $n = 300$. The three algorithms that we ran were OSOM, UCB and OFUL.

When data is generated according to the simple model ($\theta^* = 0$), we see that OSOM and UCB suffer regret that is sub-linear, and is significantly lower than the regret suffered by OFUL whose regret is also sub-linear but pays for the additional variance of estimating a more complex model. While when the data is generated from the complex model ($\|\theta^*\|_2 = 1$) the regret suffered by UCB is *linear* as it does not identify and estimate the linear structure of the mean rewards. Here, the regret suffered by both OFUL and OSOM is sub-linear and almost identical.

5 Discussion

We were able to successfully obtain minimax-optimal rates in both regimes under suitable stochastic conditions on the contextual information. This is a natural step to understanding data-dependent model selection for contextual bandits. A number of exciting directions remain open.

- We relied on the linear structure of the rewards to obtain our regret bounds. It is conceivable that this *linearity* is not essential, and that these algorithmic ideas could be generalized to arbitrary nested models.
- Our guarantees here are under a stochastic assumption on both the rewards and the distribution of the contexts. It would be interesting to understand whether these assumptions can be loosened, or if there exist fundamental limitations to model-selecting under bandit feedback in adversarial settings.

Useful functions

These functions arise by applying the concentration inequalities on terms that appear while controlling the regret. It is straightforward to verify that $\mathcal{W}_\delta(t,n) = \widetilde{\mathcal{O}}\left(\sigma(\sqrt{d}+\sqrt{K})\sqrt{t}\right)$.

$$\tau_{\min}(\delta,n) := \left(\frac{16}{\rho_{\min}^2} + \frac{8}{3\rho_{\min}}\right)\log\left(\frac{2dn}{\delta}\right). \tag{8a}$$

$$\Upsilon_\delta(t,n) := \left(\frac{20}{3} + \frac{10\sigma}{3}\left[1+2\log\left(\frac{2Kn}{\delta}\right)\right]^{\frac{1}{2}}\right)\left[\log\left(\frac{2dn}{\delta}\right) + \sqrt{t\log\left(\frac{2dn}{\delta}\right)+\log^2\left(\frac{2dn}{\delta}\right)}\right]. \tag{8b}$$

$$\mathcal{M}_\delta(t) := \sqrt{2\sigma^2\left(\frac{d}{2}\log\left(1+\frac{t}{d}\right)+\log\left(\frac{1}{\delta}\right)\right)} + 1. \tag{8c}$$

$$\mathcal{K}_\delta(t,n) := \begin{cases} \mathcal{M}_\delta(t) + \Upsilon_\delta(t,n), & \text{if } K < t \leq K + \tau_{\min}(\delta,n), \\ \frac{\mathcal{M}_\delta(t)}{\sqrt{1+\rho_{\min}\cdot(t-K)/2}} + \frac{\Upsilon_\delta(t,n)}{1+\rho_{\min}\cdot(t-K)/2}, & \text{if } K + \tau_{\min}(\delta,n) < t. \end{cases} \tag{8d}$$

$$\mathcal{W}_\delta(t,n) := 2\sum_{s=K+1}^{t}\mathcal{K}_\delta(s-1,n) + \sigma\sqrt{\frac{1+t}{2}\log\left(\frac{1}{\delta}\right)} + \left[2\sigma\sqrt{\left(1+2\log\left(\frac{Kt^{1/2}}{\delta}\right)\right)}\right]\sqrt{Kt}. \tag{8e}$$

Acknowledgements

The authors would like to thank Kush Bhatia, Akshay Krishnamurthy and Anant Sahai for helpful discussions. We gratefully acknowledge the support of the NSF through grants AST-1444078, IIS-1619362 and ECCS-1343398, and to ML4Wireless center member companies. This work was done in part while the authors were visiting the Simons Institute for the Theory of Computing.

References

[AB10] Jean-Yves Audibert and Sébastien Bubeck. “Regret bounds and minimax policies under partial monitoring”. In: *Journal of Machine Learning Research* 11.Oct (2010), pp. 2785–2836 (Cited on page 4).

[ACF02] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. “Finite-time analysis of the multiarmed bandit problem”. In: *Machine Learning* 47.2-3 (2002), pp. 235–256 (Cited on pages 2, 4).

[Aga+14] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. “Taming the monster: A fast and simple algorithm for contextual bandits”. In: *Proceedings of the International Conference on Machine Learning*. 2014 (Cited on pages 1, 2).

[Aga+16] Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, Siddhartha Sen, and Alex Slivkins. “Making contextual decisions with low technical debt”. In: *arXiv preprint arXiv:1606.03966* (2016) (Cited on page 1).

[Aga+17] Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert Schapire. “Corralling a Band of Bandit Algorithms”. In: *Proceedings of the Conference on Learning Theory*. 2017 (Cited on page 3).

[APS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. “Improved algorithms for linear stochastic bandits”. In: *Proceedings of the Advances in Neural Information Processing Systems*. 2011 (Cited on pages 2–5, 7, 19).

[Aue+02] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. “The nonstochastic multi-armed bandit problem”. In: *SIAM Journal on Computing* 32.1 (2002), pp. 48–77 (Cited on pages 1, 2).

[BBK17] Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. "Mostly exploration-free algorithms for contextual bandits". In: *arXiv preprint arXiv:1704.09011* (2017) (Cited on page 4).

[Bey+11] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert E Schapire. "Contextual bandit algorithms with supervised learning guarantees". In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*. 2011 (Cited on page 1).

[BS12] Sébastien Bubeck and Aleksandrs Slivkins. "The best of both worlds: Stochastic and adversarial bandits". In: *Proceedings of the Conference on Learning Theory*. 2012 (Cited on page 3).

[Chu+11] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. "Contextual bandits with linear payoff functions". In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*. 2011 (Cited on pages 2, 4).

[DP16] Rémy Degenne and Vianney Perchet. "Anytime optimal algorithms in stochastic multi-armed bandits". In: *Proceedings of the International Conference on Machine Learning*. 2016 (Cited on page 4).

[FK18] Dylan Foster and Akshay Krishnamurthy. "Contextual bandits with surrogate losses: Margin bounds and efficient algorithms". In: *Proceedings of the Advances in Neural Information Processing Systems*. 2018 (Cited on pages 1, 2).

[FKL19] Dylan Foster, Akshay Krishnamurthy, and Haipeng Luo. "Model selection for contextual bandits". In: *Proceedings of the Advances in Neural Information Processing Systems* (2019) (Cited on page 3).

[Fos+16] Dylan Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. "Learning in games: Robustness of fast convergence". In: *Proceedings of the Advances in Neural Information Processing Systems*. 2016 (Cited on page 3).

[Kan+18] Sampath Kannan, Jamie Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. "A smoothed analysis of the greedy algorithm for the linear contextual bandit problem". In: *Proceedings of the Advances in Neural Information Processing Systems*. 2018 (Cited on page 4).

[Li+10] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. "A contextual-bandit approach to personalized news article recommendation". In: *Proceedings of the International conference on World Wide Web*. ACM. 2010 (Cited on pages 1, 2).

[LR85] T.L Lai and Herbert Robbins. "Asymptotically Efficient Adaptive Allocation Rules". In: *Advances in Applied Mathematics* 6.1 (1985), pp. 4–22 (Cited on page 4).

[LS19] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press (preprint), 2019 (Cited on page 19).

[LZ08] John Langford and Tong Zhang. "The epoch-greedy algorithm for multi-armed bandits with side information". In: *Proceedings of the Advances in Neural Information Processing Systems*. 2008 (Cited on pages 1, 2).

[MS09] H Brendan McMahan and Matthew Streeter. "Tighter bounds for multi-armed bandits with expert advice". In: (2009) (Cited on page 1).

[Oli09] Roberto Imbuzeiro Oliveira. "Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges". In: *arXiv preprint arXiv:0911.0600* (2009) (Cited on page 3).

[Rag+18] Manish Raghavan, Aleksandrs Slivkins, Jennifer Vaughan Wortman, and Zhiwei Steven Wu. "The Externalities of Exploration and How Data Diversity Helps Exploitation". In: *Proceedings of the Conference On Learning Theory*. 2018 (Cited on page 4).

[RS16] Alexander Rakhlin and Karthik Sridharan. "BISTRO: An efficient relaxation-based method for contextual bandits". In: *Proceedings of the International Conference on Machine Learning*. 2016 (Cited on pages 1, 2).

[SKS16] Vasilis Syrgkanis, Akshay Krishnamurthy, and Robert Schapire. "Efficient algorithms for adversarial contextual learning". In: *Proceedings of the International Conference on Machine Learning*. 2016 (Cited on pages 1, 2).

[Sli14] Aleksandrs Slivkins. "Contextual bandits with similarity information". In: *Journal of Machine Learning Research* 15.1 (2014), pp. 2533–2568 (Cited on page 3).

[Syr+16] Vasilis Syrgkanis, Haipeng Luo, Akshay Krishnamurthy, and Robert Schapire. “Improved regret bounds for oracle-based adversarial contextual bandits”. In: *Proceedings of the Advances in Neural Information Processing Systems*. 2016 (Cited on pages 1, 2).

[TM17] Ambuj Tewari and Susan A Murphy. “From ads to interventions: Contextual bandits in mobile health”. In: *Mobile Health*. Springer, 2017, pp. 495–517 (Cited on pages 1, 2).

[Tro11] Joel Tropp. “Freedman’s inequality for matrix martingales”. In: *Electronic Communications in Probability* 16 (2011), pp. 262–270 (Cited on pages 3, 19).

[Woo79] Michael Woodroofe. “A one-armed bandit problem with a concomitant variable”. In: *Journal of the American Statistical Association* 74.368 (1979), pp. 799–806 (Cited on pages 1, 2).