Problem-Complexity Adaptive Model Selection for Stochastic Linear Bandits

Avishek Ghosh
Dept. of EECS, UC Berkeley

Abishek Sankararaman
AWS AI[1]

Kannan Ramchandran
Dept. of EECS, UC Berkeley

Abstract

We consider the problem of *model selection* for two popular stochastic linear bandit settings, and propose algorithms that adapts to the *unknown* problem complexity. In the first setting, we consider the K armed mixture bandits, where the mean reward of arm $i \in [K]$[2], is $\mu_i + \langle \alpha_{i,t}, \theta^* \rangle$, with $\alpha_{i,t} \in \mathbb{R}^d$ being the known context vector and $\mu_i \in [-1, 1]$ and θ^* are unknown parameters. We define[3] $\|\theta^*\|$ as the problem complexity and consider a sequence of nested hypothesis classes, each positing a different upper bound on $\|\theta^*\|$. Exploiting this, we propose Adaptive Linear Bandit (`ALB`), a novel phase based algorithm that adapts to the true problem complexity, $\|\theta^*\|$. We show that `ALB` achieves regret scaling of[4] $\widetilde{O}(\|\theta^*\|\sqrt{T})$, where $\|\theta^*\|$ is apriori unknown. As a corollary, when $\theta^* = 0$, `ALB` recovers the minimax regret for the simple bandit algorithm without such knowledge of θ^*. `ALB` is the first algorithm that uses parameter norm as model section criteria for linear bandits. Prior state of art algorithms (Chatterji et al. (2019)) achieve a regret of $\widetilde{O}(L\sqrt{T})$, where L is the upper bound on $\|\theta^*\|$, fed as an input to the problem. In the second setting, we consider the standard linear bandit problem (with possibly an infinite number of arms) where the sparsity of θ^*, denoted by $d^* \leq d$, is unknown to the algorithm. Defining d^* as the problem complexity (similar to Foster et al. (2019)), we show that `ALB` achieves $\widetilde{O}(d^*\sqrt{T})$ regret, matching that of an oracle who knew the true sparsity level. This is the first algorithm that achieves such model selection guarantees. This methodology is then extended to the case of finitely many arms and similar results are proven. We further verify through synthetic and real-data experiments that the performance gains are fundamental and not artifacts of mathematical bounds. In particular, we show $1.5 - 3$x drop in cumulative regret over non-adaptive algorithms.

1 INTRODUCTION

We study model selection for MAB, which refers to choosing the appropriate hypothesis class, to model the mapping from arms to expected rewards. Model selection for MAB plays an important role in applications such as personalized recommendations, as we explain in the sequel. Formally, a family of nested hypothesis classes $\mathcal{H}_f$, $f \in \mathcal{F}$ needs to be specified, where each class posits a plausible model for mapping arms to expected rewards. The true model is assumed to be contained in the family $\mathcal{F}$ which is totally ordered, where if $f_1 \leq f_2$, then $\mathcal{H}_{f_1} \subseteq \mathcal{H}_{f_2}$. Model selection guarantees then refers to algorithms whose regret scales in the complexity of the *smallest hypothesis class containing the true model*, even though the algorithm was not aware apriori.

We consider two canonical settings for the stochastic MAB problem. The first is the K *armed mixture MAB* setting, in which the mean reward from any arm $i \in [K]$ is given by $\mu_i + \langle \theta^*, \alpha_{i,t} \rangle$, where $\alpha_{i,t} \in \mathbb{R}^d$ is the known context vector of arm i at time t, and the arm bias $\mu_i \in \mathbb{R}$, $\theta^* \in \mathbb{R}^d$ are unknown and needs to be estimated. This setting also contains the standard MAB (Lai and Robbins (1985); Auer et al. (2002)) when $\theta^* = 0$. Popular linear bandit algorithms, like LinUCB, OFUL (see Chu et al. (2011); Dani et al. (2008); Abbasi-Yadkori et al. (2011)) handle the case with no bias ($\mu_i = 0$), while OSOM Chatterji et al. (2019), the recent improvement can handle arm-bias. Implicitly, all the above algorithms assume an upper bound on

[1]Work done while affiliated with UC Berkeley
[2]We denote $[r] = \{1, 2, \ldots, r\}$; for $r > 0$.
[3]We use $\|.\|$ for the ℓ_2 norm unless otherwise specified.
[4]The notation $\widetilde{\mathcal{O}}$ hides the logarithmic dependence.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

the norm of $\|\theta^*\| \leq L$, which is supplied as an input. Crucially however, the regret guarantees scale linearly in the upper bound L. In contrast, we choose $\|\theta^*\|$ as the problem complexity, and provide a novel phase based algorithm, that, without any upper bound on the norm $\|\theta^*\|$, *adapts to the true complexity* of the problem instance, and achieves a regret linearly in the true norm $\|\theta^*\|$. As a corollary, our algorithm's performance matches the minimax regret of simple MAB when $\theta^* = 0$, even though the algorithm did not apriori know that $\theta^* = 0$. Formally, we consider a continuum of hypothesis classes, with each class positing a different upper bound on the norm $\|\theta^*\|$, where the complexity of a class is the upper bound posited. As our regret bound scales linearly in $\|\theta^*\|$ (the complexity of the smallest hypothesis class containing the instance) as opposed to an upper bound on $\|\theta^*\|$, our algorithm achieves model selection guarantees.

The second setting we consider is the standard linear stochastic bandit (Abbasi-Yadkori et al. (2011)) with possibly an infinite number of arms, and the mean reward of any arm $x \in \mathbb{R}^d$ (arms are vectors in this case) given by $\langle x, \theta^* \rangle$, where $\theta^* \in \mathbb{R}^d$ is unknown. We consider model selection from a total of d hypothesis classes, with each class positing a different cardinality for the support of θ^*. We exhibit a novel algorithm, where the regret scales linearly in the unknown cardinality of the support of θ^*. The regret of our algorithm matches that of an oracle that knows the support of θ^* (Carpentier and Munos (2012),Bastani and Bayati (2020)), thereby achieving model selection guarantees.

This setting with dimension as a measure of complexity was also studied by Carpentier and Munos (2012). However, our regret bounds are stronger (by a logarithm in d factor). Furthermore, our algorithmic paradigm is more broadly applicable – for eg. we can also handle the finite arm case (see Section 4.4), and obtain similar model selection regret guarantees that match the regret of an oracle that knows the true dimension. Model selection with dimension as a measure of complexity was also recently studied by Foster et al. (2019), in which the classical contextual bandit (Chu et al. (2011)) with a finite number of arms was considered. We clarify here that although our results for the finite arm setting yields a better (optimal) regret scaling with respect to the time horizon T and the support of θ^* (denoted by d^*), our guarantee depends on a problem dependent parameter and thus not uniform over all instances. In contrast, the results of Foster et al. (2019), although sub-optimal in d^* and T, is uniform over all problem instances. Closing this gap is an interesting future direction.

1.1 Our Contributions

1. Successive Refinement Algorithms - We present two novel epoch based algorithms, `ALB (Adaptive Linear Bandit) - Norm` and `ALB - Dim`, that achieve model selection guarantees for both families of hypothesis respectively. For the K armed mixture MAB, `ALB-Norm`, at the beginning of each phase, estimates an *upper bound* on $\|\theta^*\|$. Subsequently, the algorithm assumes this bound to be true during the phase, and the upper bound is re-estimated at the end of a phase. Similarly for the linear bandit setting, `ALB-Dim` estimates the support of θ^* at the beginning of each phase and subsequently only plays from this estimated support during the phase. In both settings, we show the estimates converge to the true underlying value —in the first case, the estimate of norm $||\theta^*||$ converges to the true norm, and in the second case, for all time after a random time with finite expectation, the estimated support equals the true support. Our algorithms are reminiscent of successive rejects algorithm of Audibert and Bubeck (2010) for standard MAB, with the crucial difference being that our algorithm is *non-monotone*. Once rejected, an arm is never pulled in the classical successive rejects. In contrast, our algorithm is successive *refinement* and is not necessarily monotone —a hypothesis class discarded earlier can be considered at a later point of time.

2. Regret depending on the Complexity of the smallest Hypothesis Class - In the K armed mixture MAB setting, `ALB-Norm`'s regret scale as $\widetilde{O}(\|\theta^*\|\sqrt{T})$, which is superior compared to state of art algorithms such as `OSOM` Chatterji et al. (2019), whose regret scales as $\widetilde{O}(L\sqrt{T})$, where L is an upper bound on $\|\theta^*\|$ that is supplied as an input[5] . As a corollary, we get the 'best of both worlds' guarantee of Chatterji et al. (2019), where if $\theta^* = 0$, our regret bound recovers known minimax regret guarantee of simple MAB. Similarly, for the linear bandit setting with unknown support, `ALB-Dim` achieves a regret of $\widetilde{O}(d^*\sqrt{T})$, where $d^* \leq d$ is the true sparsity of θ^*. This matches the regret obtained by oracle algorithms that know of the true sparsity d^* (Carpentier and Munos (2012); Bastani and Bayati (2020)). We also apply our methodology to the case when there is a finite number of arms and obtain similar regret scaling as the oracle. `ALB-Dim` is the first algorithm to obtain such model selection guarantees. Prior state of art algorithm `ModCB` for model selection with dimension as a measure of complexity was proposed in Foster et al. (2019), with a

[5]We clarify that, without the arm bias $\{\mu_i\}_{i=1}^K$, the regret dependence on L can be improved with a careful choice of tuning parameters (Abbasi-Yadkori et al. (2011)). However, in the K armed mixture model with arm bias, a linear dependence on L is proved in Chatterji et al. (2019).

finite set of arms, where the regret guarantee was sub-optimal compared to the oracle. However, our regret bounds for dimension, though matches the oracle, depends on the minimum non-zero coordinate value and is thus not uniform over θ^*. Obtaining regret rates in this case that matches the oracle and is uniform over all θ^* is an interesting future work.

Motivating Example: Our model selection framework is applicable to personalized news recommendation platforms, that recommend one of K news outlets, to each of its users. The recommendation decisions made to any fixed user, can be modeled as an instance of a MAB; the arms are the K different news outlets, and the platforms recommendation decision (made to this user) on day t is the arm played at time t. On each day t, each news outlet i reports a story, that can be modeled by the vectors $\alpha_{i,t}$, which can be obtained by embedding the stories into a fixed-dimensional vector space based on some common embedding schemes. The reward obtained by the platform in recommending news outlet i to this user on day t can be modeled as $\mu_i + \langle \alpha_{i,t}, \theta^* \rangle$, where μ_i captures the preference of this user to news outlet i and the vector θ^* captures the 'interest' of the user. If a channel i on day t, publishes a news article $\alpha_{i,t}$, that this user 'likes', then most likely the content $\alpha_{i,t}$ is 'aligned' to θ^* and will have a large inner product $\langle \alpha_{i,t}, \theta^* \rangle$. Different users on the platform however may have different biases and θ^*. Some users have strong preference towards certain topics and will read content written by any outlet on this topic (these users will have a large $\|\theta^*\|$). Other users may be agnostic to topics, but may prefer a particular news outlet a lot (e.g., some users like Fox News exclusively or CNN exclusively, regardless of the topic). These users will have low $\|\theta^*\|$.

In such a multi-user recommendation application, we show that `ALB-Norm` which tailors to the model class for each user separately is more effective than employing a (non-adaptive) linear bandit algorithm for each user. We further show that our algorithms are also more effective than state of art model selection algorithms such as `OSOM` (Chatterji et al. (2019)), which posits a 'binary' model - users either assign a 0 weight to topic or assign a potentially large weight to topic. Furthermore the heterogeneous complexity in this application can also be captured by the cardinality of the support of θ^*; different people are interested in different sub-vectors of θ^* which the recommendation platform is not aware of apriori. In this context, our algorithm `ALB-Dim` that tailors to the interest of the individual user achieves $1.5 - 3$x reduction in cumulative regret compared to its non-adaptive counterparts.

2 RELATED WORK

Model selection for MAB are only recently being studied (Agarwal et al. (2016); Ghosh et al. (2017)), with Chatterji et al. (2019), Foster et al. (2019) being the closest to our work. `OSOM` was proposed in Chatterji et al. (2019) for model selection in the K armed mixture MAB from two hypothesis classes —a "simple model" where $\|\theta^*\| = 0$, or a "complex model", where $0 < \|\theta^*\| \leq L$. `OSOM` was shown to obtain a regret guarantee of $O(\log(T))$ when the instance is simple and $\widetilde{O}(L\sqrt{T})$ otherwise. We refine this to consider a *continuum* of hypothesis classes and propose `ALB-Norm`, which achieves regret $\widetilde{O}(\|\theta^*\|\sqrt{T})$, a superior guarantee (which we also empirically verify) compared to `OSOM`. Model selection with dimension as a measure of complexity was recently initiated in Foster et al. (2019), where an algorithm `ModCB` was proposed. The setup considered in Foster et al. (2019) was that of contextual bandits (Chu et al. (2011)) with a fixed and finite number of arms. `ModCB` in this setting was shown to achieve a regret that is sub-optimal compared to the oracle. In contrast, we consider the linear bandit setting with a continuum of arms (Abbasi-Yadkori et al. (2011)), and `ALB-Dim` achieves a regret scaling matching that of an oracle. The continuum of arms allows `ALB-Dim` a finer exploration of arms, that enables it to learn the support of θ^* reliably and thus obtain regret matching that of the oracle. However, our regret bounds depend on the magnitude of the minimum non-zero value of θ^* and is thus not uniform over all θ^*. Obtaining regret matching the oracle that holds uniformly over all θ^* is an interesting future work.

`Corral` was proposed in Agarwal et al. (2016), by casting the optimal algorithm for each hypothesis class as an expert, with the forecaster's performance having low regret with respect to the best expert (best model class). However, `Corral` can only handle finitely many hypothesis classes and is not suited to our setting with continuum hypothesis classes.

Adaptive algorithms for linear bandits have also been studied in different contexts from ours. The papers of Locatelli and Carpentier (2018); Krishnamurthy et al. (2018) consider problems where the arms have an unknown structure, and propose algorithms adapting to this structure to yield low regret. The paper Lykouris et al. (2017) proposes an algorithm in the adversarial bandit setup that adapt to an unknown structure in the adversary's loss sequence, to obtain low regret. The paper of Auer et al. (2018) consider adaptive algorithms, when the distribution changes over time. In the context of online learning with full feedback, there have been several works addressing model selection (Luo and Schapire (2015); McMahan and Abernethy (2013); Orabona (2014); Cutkosky and

Boahen (2017)). In the context of statistical learning, model selection has a long line of work (for eg. Vapnik (2006), Birgé et al. (1998), Lugosi et al. (1999), Arlot et al. (2011), Cherkassky (2002) Devroye et al. (2013)). However, the bandit feedback in our setups is much more challenging and a straightforward adaptation of algorithms developed for either statistical learning or full information to the setting with bandit feedback is not feasible.

3 NORM AS A MEASURE OF COMPLEXITY

3.1 Problem Formulation

In this section, we formally define the problem. At each round $t \in [T]$, the player chooses one of the K available arms. Each arm has a context $\{\alpha_{i,t} \in \mathbb{R}^d\}_{i=1}^K$ that changes over time t. Similar to the standard stochastic contextual bandit framework, the context vectors for each arm is chosen independently of all other arms and of the past time instances.

We assume that there exists an underlying parameter $\theta^* \in \mathbb{R}^d$ and biases $\{\mu_1, \ldots, \mu_K\}$ each taking value in $[-1, 1]$ such that the mean reward of an arm is a linear function of the context of the arm. The reward for playing arm i at time t is given by, $g_{i,t} = \mu_i + \langle \alpha_{i,t}, \theta^* \rangle + \eta_{i,t},$, where $\{\eta_{i,t}\}_{t=1}^T$ are i.i.d zero mean and σ sub-Gaussian noise. The context vector satisfies $\mathbb{E}[\alpha_{i,t} | \{\alpha_{j,s}, \eta_{j,s}\}_{j\in[K], s\in[t-1]}\}] = 0$, and $\mathbb{E}[\alpha_{i,t}\alpha_{i,t}^\top | \{\alpha_{j,s}, \eta_{j,s}\}_{j\in[K], s\in[t-1]}\}] \succcurlyeq \rho_{\min} I$. The above setting is popularly known as stochastic linear bandit (Chatterji et al. (2019)). In the special case of $\theta^* = 0$, the above model reduces to $g_{i,t} = \mu_i + \eta_{i,t}$. Note that in this setting, the mean reward of arms are fixed, and not dependent on the context. Hence, this corresponds to a simple *multi-armed bandit* setup and standard algorithms (like UCB, Auer et al. (2002)) can be used as a learning rule. At round t, we define $i_t^* = \text{argmax}_{i\in[K]} [\mu_i + \langle \theta^*, \alpha_{i,t} \rangle]$ as the best arm. Also let an algorithm play arm A_t at round t. The regret of the algorithm upto time T is given by[6],

$$R(T) = \sum_{s=1}^{T} \left[\mu_{i_s^*} + \langle \theta^*, \alpha_{i_s^*,s} \rangle - \mu_{A_s} - \langle \theta^*, \alpha_{A_s,s} \rangle \right].$$

We define a new notion of complexity for stochastic linear bandits; and propose an algorithm that adapts to it. We define $\|\theta^*\|$ as the problem complexity for the linear bandit instance. Note that if $\|\theta^*\| = 0$, the linear bandit model reduces to the simple multi-armed bandit setting. Furthermore, the cumulative regret $R(T)$ of linear bandit algorithms (like OFUL Abbasi-Yadkori et al. (2011) and OSOM Chatterji et al. (2019)) scales linearly with $\|\theta^*\|$ (Chatterji et al. (2019)). Hence, $\|\theta^*\|$ constitutes a natural notion of model complexity. In Algorithm 1, we propose a scheme which adapts to the true complexity of the problem, $\|\theta^*\|$. Instead of assuming an upper-bound on $\|\theta^*\|$, we use an initial exploration phase to obtain a rough estimate of $\|\theta^*\|$ and then successively refine it over multiple epochs. The cumulative regret of our proposed algorithm actually scales linearly with $\|\theta^*\|$.

3.2 `ALB-Norm` algorithm

We present the adaptive scheme in Algorithm 1. Note that Algorithm 1 depends on the subroutine OFUL$^+$. Observe that at each iteration, we estimate the bias $\{\mu_1, \ldots, \mu_K\}$ and θ^* separately. The estimation of the bias involves a simple sample mean estimate with upper confidence level, and the estimation of θ^* involves building a confidence set that shrinks over time.

In order to estimate θ^*, we use a variant of the popular OFUL (Abbasi-Yadkori et al. (2011)) algorithm with arm bias. We refer to the algorithm as OFUL$^+$. Algorithm 1 is epoch based, and over multiple epochs, we successively refine the estimate of $\|\theta^*\|$. We start with a rough over-estimate of $\|\theta^*\|$ (obtained from a pure exploration phase), and based on the confidence set constructed at the end of the epoch, we update the estimate of $\|\theta^*\|$. We argue that this approach indeed correctly estimates $\|\theta^*\|$ with high probability over a sufficiently large time horizon T.

We now discuss the algorithm OFUL$^+$. A variation of this is proposed in Chatterji et al. (2019) in the context of model selection between linear and standard multi-armed bandits. We use $\tilde{\mu}_{i,t}$ to address the bias term, which we define shortly. The parameters b and δ are used in the construction of the confidence set $\mathcal{C}_t$. Suppose OFUL$^+$ is run for a total of $\tilde{T}$ rounds and plays arm A_s at time s. Let $T_i(t)$ be the number of times OFUL$^+$ plays arm i until time t. Also, let b be the current estimate of $\|\theta^*\|$. We define, $\bar{g}_{i,t} = \frac{1}{T_i(t)} \sum_{s=1}^t g_{i,s} \mathbf{1}\{A_s = t\}$. With this, we have [7] $\tilde{\mu}_{i,t} = \bar{g}_{i,t} + c(\sigma + b)\sqrt{\frac{d}{T_i(t)} \log\left(\frac{1}{\delta}\right)}$. The confidence interval $\mathcal{C}_t$, is defined as $\mathcal{C}_t = \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\| \le \mathcal{K}_\delta(b, t, \tilde{T})\}$, where $\hat{\theta}_t$ is the least squares estimate defined as, $\hat{\theta}_t = \left(\alpha_{K+1:t}^\top \alpha_{K+1:t} + I\right)^{-1} \alpha_{K+1:t}^\top G_{K+1:t}$ with $\alpha_{K+1:t}$ as a matrix having rows $\alpha_{A_{K+1},K+1}^\top, \ldots, \alpha_{A_t,t}^\top$ and $G_{K+1:t} = [g_{A_{K+1},K+1} - \tilde{\mu}_{A_{K+1},K+1}, \ldots, g_{A_t,t} - \tilde{\mu}_{A_t,t}]^\top$. The radius of $\mathcal{C}_t$ is given by, $\mathcal{K}_\delta(b, t, \tilde{T}) = c\frac{(\sigma\sqrt{d}+b)}{\rho_{\min}\sqrt{t}}\sqrt{\log(K\tilde{T}/\delta)}$ (see Appendix A for complete

[6] Throughout the paper, we use $C, C_1, .., c, c_1, ..$ to denote positive universal constants, the value of which may differ in different instances.

[7] For complete expression, see Appendix A

Algorithm 1: Adaptive Linear Bandit (norm)–`ALB-Norm`

Input: Initial exploration period τ, the phase length T_1, $\delta_1 > 0$, $\delta_s > 0$.
Select an arm at random, sample 2τ rewards
Obtain initial estimate (b_1) of $\|\theta^*\|$ according to Section 3.3
for $t = 1, 2, \ldots, K$ **do**
  Play arm t, receive reward $g_{t,t}$
end for
Define $\mathcal{S} = \{g_{i,i}\}_{i=1}^K$
for epochs $i = 1, 2 \ldots, N$ **do**
  Use $\mathcal{S}$ as pure-exploration reward
  Play $\text{OFUL}^+_{\delta_i}(b_i)$ until the end of epoch i (denoted by $\mathcal{E}_i$)
  At $t = \mathcal{E}_i$, refine estimate of $\|\theta^*\|$ as, $b_{i+1} = \max_{\theta \in \mathcal{C}_{\mathcal{E}_i}} \|\theta\|$
  Set $T_{i+1} = 2T_i$, $\delta_{i+1} = \frac{\delta_i}{2}$.
end for
$\underline{\texttt{OFUL}^+_\delta(b)\texttt{:}}$
Input: Parameters b, $\delta > 0$, number of rounds $\tilde{T}$
for $t = 1, 2, \ldots, \tilde{T}$ **do**
  Select the best arm estimate as $j_t = \text{argmax}_{i \in [K]} \left[\max_{\theta \in \mathcal{C}_{t-1}} \{\tilde{\mu}_{i,t-1} + \langle \alpha_{i,t}, \theta \rangle\}\right]$, where $\tilde{\mu}_{i,t}$ and $\mathcal{C}_t$ are given in Section 3.2.
  Play arm j_t, and update $\{\tilde{\mu}_{i,t}\}_{i=1}^K$ and $\mathcal{C}_t$
end for

expressions). Lemma 2 of Chatterji et al. (2019) shows that $\theta^* \in \mathcal{C}_t$ with probability [8] at least $1 - 4\delta$.

3.3 Construction of initial estimate b_1

We select an arm at random (WLOG, assume that this is arm 1), and sample rewards (in an i.i.d fashion) for 2τ times, where $\tau > 0$ is a parameter to be fed to the Algorithm 1. In order to kill the bias of arm 1, we take pairwise differences and form: $y(1) = g_{1,1} - g_{1,2}$, $y(2) = g_{1,3} - g_{1,4}$ and so on. Augmenting $y(.)$, we obtain: $Y = \tilde{X}\theta^* + \tilde{\eta}$, where the i-th row of $\tilde{X}$ is $(\alpha_{1,2i+1} - \alpha_{1,2i+2})^\top$, the i-th element of $\tilde{\eta}$ is $\eta_{1,2i+1} - \eta_{1,2i+2}$. Hence, the least squares estimate, $\widehat{\theta}^{(\ell s)}$ satisfies $\|\widehat{\theta}^{(\ell s)} - \theta^*\| \leq \sqrt{2}\sigma\sqrt{\frac{d}{\tau}\log(1/\delta_s)}$, with probability exceeding $1 - \delta_s$ (Wainwright (2019)). We set the initial estimate $b_1 = \max\{\|\widehat{\theta}^{(\ell s)}\| + \sqrt{2}\sigma\sqrt{\frac{d}{\tau}\log(1/\delta_s)},\ 1\}$ satisfying $b_1 \geq \|\theta^*\|$ and $b_1 \geq 1$ with high probability.

[8] There is a typo in the proof of regret in Chatterji et al. (2019). We correct the typo, and modify the definition of $\tilde{\mu}_{i,t}$ and $\mathcal{K}_\delta(b, t, \tilde{T})$. As a consequence, the high probability bounds change a little.

3.4 Regret Guarantee of Algorithm 1

We now obtain an upper bound on the cumulative $R(T)$ with Algorithm 1. For theoretical tractability, we assume that OFUL^+ restarts at the start of each epoch. We have the following lemma regarding the sequence $\{b_i\}_{i=1}^\infty$ of estimates of $\|\theta^*\|$:

Lemma 1. *With probability exceeding $1 - 8\delta_1 - \delta_s$, the sequence $\{b_i\}_{i=1}^\infty$ converges to $\|\theta^*\|$ at a rate $\mathcal{O}(\frac{i}{2^i})$, and we obtain $b_i \leq (c_1\|\theta^*\| + c_2)$ for all i, provided $T_1 \geq C_1 \left(\max\{p, q\}\, b_1\right)^2 d$, where $C_1 > 9$, and $p = [\frac{14\log(\frac{2KT_1}{\delta_1})}{\sqrt{\rho_{\min}}}]$, $q = [\frac{2C\sigma\log(\frac{2KT_1}{\delta_1})}{\sqrt{\rho_{\min}}}]$.*

Hence, the sequence converges to $\|\theta^*\|$ at an exponential rate. We have the following guarantee on the cumulative regret $R(T)$:

Theorem 1. *Let $C_1 > 9$ and $T_{\min}(\delta, T) = (\frac{16}{\rho^2_{\min}} + \frac{8}{3\rho_{\min}})\log(\frac{2dT}{\delta})$, and suppose $T_1 > \max\{T_{\min}(\delta, T), C_1 \left(\max\{p, q\}\, b_1\right)^2 d\}$. Then, with probability at least $1 - 18\delta_1 - \delta_s$, we have*

$$R(T) \leq C(2\tau + K)\|\theta^*\| + C_2(\|\theta^*\| + 1) \times (\sqrt{K} + \sqrt{d})\sqrt{T}\,\log(KT_1/\delta_1)\log(T/T_1).$$

Remark 1. *Note that the regret bound depends on the problem complexity $\|\theta^*\|$, and we prove that Algorithm 1 adapts to this complexity. Ignoring the log factors, Algorithm 1 has a regret of $\tilde{\mathcal{O}}((1+\|\theta^*\|)(\sqrt{K}+\sqrt{d})\sqrt{T})$ with high probability.*

Remark 2. *(Matches Linear Bandits) The above bound matches the regret guarantee of the linear bandit algorithm with bias as presented in Chatterji et al. (2019).*

Remark 3. *(Matches UCB when $\theta^* = 0$) When $\theta^* = 0$ (the simplest model, without any contextual information), Algorithm 1 recovers the minimax regret of UCB algorithm. Indeed, substituting $\|\theta^*\| = 0$ in the above regret bound yields $R(T) = \mathcal{O}(\sqrt{KT})$, with high probability, provided $K > d$. Hence, we obtain the "best of both worlds" results with simple model ($\theta^* = 0$) and contextual bandit model ($\theta^* \neq 0$).*

4 DIMENSION AS A MEASURE OF COMPLEXITY

4.1 Continuum Armed Setting

In this section, we consider the standard stochastic linear bandit model in d dimensions (Abbasi-Yadkori et al. (2011)), with dimension as a measure of complexity. The setup in this section is almost identical to that in Section 3.1, with 0 arm biases and a continuum collection of arms denoted by the set

$\mathcal{A} := \{x \in \mathbb{R}^d : \|x\| \leq 1\}$[9] Thus, the mean reward from any arm $x \in \mathcal{A}$ is $\langle x, \theta^* \rangle$, where $\|\theta^*\| \leq 1$. We assume that θ^* is $d^* \leq d$ sparse, where d^* is apriori unknown to the algorithm. Thus, unlike in Section 3.1, there is no i.i.d. context sampling in this section. We consider a sequence of d nested hypothesis classes, where each hypothesis class $i \leq d$, models θ^* as a i sparse vector. The goal of the forecaster is to minimize the regret, $R(T) := \sum_{t=1}^{T} [\langle x_t^* - x_t, \theta^* \rangle]$, where at any time t, x_t is the action recommended by an algorithm and $x_t^* = \text{argmax}_{x \in \mathcal{A}} \langle x, \theta^* \rangle$. The regret $R(T)$ measures the loss in reward of the forecaster with that of an oracle that knows θ^* and thus can compute x_t^* at each time.

4.2 `ALB-Dim` Algorithm

The algorithm is parametrized by $T_0 \in \mathbb{N}$, given in Equation (1) in the sequel and slack $\delta \in (0, 1)$. As in the previous case, `ALB-Dim` proceeds in phases numbered $0, 1, \cdots$ which are non-decreasing with time. At the beginning of each phase, `ALB-Dim` makes an estimate of the set of non-zero coordinates of θ^*, which is kept fixed throughout the phase. Concretely, each phase i is divided into two blocks - *(i)* a regret minimization block lasting $25^i T_0$ time slots, *(ii)* followed by a random exploration phase lasting $5^i \lceil \sqrt{T_0} \rceil$ time slots. At the beginning of each phase $i \geq 0$, $\mathcal{D}_i \subseteq [d]$ denotes the set of 'active coordinates', namely the estimate of the non-zero coordinates of θ^*. Subsequently, in the regret minimization block of phase i, a fresh instance of OFUL (Abbasi-Yadkori et al. (2011)) is spawned, with the dimensions restricted only to the set $\mathcal{D}_i$ and probability parameter $\delta_i := \frac{\delta}{2^i}$. In the random exploration phase, at each time, one of the possible arms from the set $\mathcal{A}$ is played chosen uniformly and independently at random. At the end of each phase $i \geq 0$, `ALB-Dim` forms an estimate $\widehat{\theta}_{i+1}$ of θ^*, by solving a least squares problem using all the random exploration samples collected till the end of phase i. The active coordinate set $\mathcal{D}_{i+1}$, is then the coordinates of $\widehat{\theta}_{i+1}$ with magnitude exceeding $2^{-(i+1)}$. The pseudo-code is provided in Algorithm 2, where, $\forall i \geq 0$, S_i in lines 15 and 16 is the total number of random-exploration samples in all phases upto and including i. By this careful choice of exploration periods and thresholds, we show that the estimated support of θ^* is equal to the true support, for all but finitely many phases. Thus, after a finite amount of time, the support set 'locks in' to the correct support. Once the lock-in occurs, the agent incurs the optimal regret that an oracle knowing the true support would incur. Thus, the difference in the regret between our algorithm and that of the oracle is the additive regret incurred until the lock-in occurs, which is a constant independent of time (depending on the smallest non-zero value of θ^*).

[9]Our algorithm can be applied to any compact set $\mathcal{A} \subset \mathbb{R}^d$, including the finite set as shown in Appendix C.

Algorithm 2: Adaptive Linear Bandit (Dimension)–`ALB-Dim`

1: **Input:** Initial Phase length T_0, slack $\delta > 0$.
2: $\widehat{\theta}_0 = \mathbf{1}$, $T_{-1} = 0$
3: **for** Each epoch $i \in \{0, 1, 2, \cdots\}$ **do**
4: $\quad T_i = 25^i T_0, \quad \varepsilon_i \leftarrow \frac{1}{2^i}, \quad \delta_i \leftarrow \frac{\delta}{2^i}$
5: $\quad \mathcal{D}_i := \{i : |\widehat{\theta}_i| \geq \frac{\varepsilon_i}{2}\}$
6: $\quad$ **for** Times $t \in \{T_{i-1} + 1, \cdots, T_i\}$ **do**
7: $\quad\quad$ Play OFUL$(1, \delta_i)$ restricted to coordinates in $\mathcal{D}_i$. δ_i is the probability slack parameter and 1 represents $\|\theta^*\| \leq 1$.
8: $\quad$ **end for**
9: $\quad$ **for** Times $t \in \{T_i + 1, \cdots, T_i + 5^i \sqrt{T_0}\}$ **do**
10: $\quad\quad$ Play an arm from the action set $\mathcal{A}$ chosen uniformly and independently at random.
11: $\quad$ **end for**
12: $\quad \boldsymbol{\alpha}_i \in \mathbb{R}^{S_i \times d}$; each row being the arm played during all random explorations in the past.
13: $\quad \boldsymbol{y}_i \in \mathbb{R}^{S_i}$; i-th entry being the reward at the i-th random exploration in the past
14: $\quad \widehat{\theta}_{i+1} \leftarrow (\boldsymbol{\alpha}_i^T \boldsymbol{\alpha}_i)^{-1} \boldsymbol{\alpha}_i \mathbf{y}_i$, $(\in \mathbb{R}^d)$
15: **end for**

4.3 Main Result

We first specify, how to set the input parameter T_0, as function of δ. For any $N \geq d$, denote by A_N to be the $N \times d$ random matrix with each row being a vector sampled uniformly and independently from the unit sphere in d dimensions. Denote by $M_N := \frac{1}{N} \mathbb{E}[A_N^T A_N]$, and by $\lambda_{max}^{(N)}, \lambda_{min}^{(N)}$, to be the largest and smallest eigenvalues of M_N. Observe that as M_N is positive semi-definite ($0 \leq \lambda_{min}^{(N)} \leq \lambda_{max}^{(N)}$) and almost-surely full rank, i.e., $\mathbb{P}[\lambda_{min}^{(N)} > 0] = 1$. The constant T_0 is the smallest integer such that

$$\sqrt{T_0} \geq \max\Big(\frac{32\sigma^2}{(\lambda_{min}^{(\lceil\sqrt{T_0}\rceil)})^2} \ln(2d/\delta), \frac{4}{3} \frac{(6\lambda_{max}^{(\lceil\sqrt{T_0}\rceil)} + \lambda_{min}^{(\lceil\sqrt{T_0}\rceil)})(d + \lambda_{max}^{(\lceil\sqrt{T_0}\rceil)})}{(\lambda_{min}^{(\lceil\sqrt{T_0}\rceil)})^2} \ln(2d/\delta)\Big). \quad (1)$$

Remark 4. *T_0 in Equation (1) is chosen such that, at the end of phase 0, $\mathbb{P}[||\widehat{\theta}_0 - \theta^*||_\infty \geq 1/2] \leq \delta$.*

A formal statement of the Remark is provided in Lemma 2 in Appendix B.

Theorem 2. *Suppose Algorithm 2 is run with input parameters $\delta \in (0, 1)$, and T_0 as given in Equation (1), then with probability at-least $1 - \delta$, the regret after a*

total of T arm-pulls satisfies

$$R_T \leq \frac{50}{\gamma^{4.65}} T_0 + 25\sqrt{T}[1 + 4\sqrt{d^* \ln(1 + \frac{25T}{d^*})} \times (1 + \sigma\sqrt{2\ln(\frac{T}{T_0\delta}) + d^* \ln(1 + \frac{25T}{d^*})})].$$

The parameter $\gamma > 0$ is the minimum magnitude of the non-zero coordinate of θ^*, i.e., $\gamma = \min\{|\theta_i^*| : \theta_i^* \neq 0\}$ and d^* the sparsity of θ^*, i.e., $d^* = |\{i : \theta_i^* \neq 0\}|$.

To parse this result, we give the following corollary.

Corollary 1. *Suppose Algorithm 2 is run with input parameters $\delta \in (0,1)$, and $T_0 = \widetilde{O}\left(d^2 \ln^2\left(\frac{1}{\delta}\right)\right)$ given in Equation (1), then with probability at-least $1-\delta$, the regret after T times satisfies*

$$R_T \leq O(\frac{d^2}{\gamma^{4.65}} \ln^2(d/\delta)) + \widetilde{O}(d^*\sqrt{T}).$$

Remark 5. *The constants in the Theorem are not optimized. The exponent of γ can be made arbitrarily close to 4, by setting $\varepsilon_i = C^{-i}$ in Line 4 of Algorithm 2, for some appropriately large constant $C > 1$, and increasing $T_i = (C')^i T_0$, for large C' ($C' \approx C^4$).*

Discussion - The regret of an oracle algorithm that knows the true complexity d^* scales as $\widetilde{O}(d^*\sqrt{T})$ (Carpentier and Munos (2012); Bastani and Bayati (2020)), matching `ALB-Dim`'s regret, upto an additive constant (the price of model selection) independent of T. On the other hand, standard linear bandit algorithms such as `OFUL` achieve a regret scaling $\widetilde{O}(d\sqrt{T})$, which is much larger compared to that of `ALB-Dim`, especially when $d^* << d$, and γ is a constant. Numerical simulations further confirms this deduction, thereby indicating that our improvements are fundamental and not from mathematical bounds. Corollary 1 also indicates that `ALB-Dim` has higher regret if γ is lower. A small value of γ makes it harder to distinguish a non-zero coordinate from a zero coordinate, which is reflected in the regret scaling. Nevertheless, this only affects the *second order term as a constant*, and the dominant scaling term only depends on the true complexity d^*, and not on the underlying dimension d. However, the regret guarantee is not uniform over all θ^* as it depends on γ. Obtaining regret rates matching the oracles and that hold uniformly over all θ^* is an interesting avenue of future work.

4.4 Finite Armed Setting

Setup: In this section, we consider the model selection problem for the setting with finitely many arms in the framework studied in Foster et al. (2019). At each time $t \in [T]$, the forecaster is shown a context $X_t \in \mathcal{X}$, where $\mathcal{X}$ is some arbitrary 'feature space'. The set of contexts $(X_t)_{t=1}^T$ are i.i.d. with $X_t \sim \mathcal{D}$, a probability distribution over $\mathcal{X}$ that is known to the forecaster. Subsequently, the forecaster chooses an action $A_t \in \mathcal{A}$, where the set $\mathcal{A} \coloneqq \{1, \cdots, K\}$ are the K possible actions chosen by the forecaster. The forecaster then receives a reward $Y_t := \langle \theta^*, \phi^M(X_t, A_t)\rangle + \eta_t$. Here $(\eta_t)_{t=1}^T$ is an i.i.d. sequence of 0 mean sub-gaussian random variables with sub-gaussian parameter σ^2 that is known to the forecaster. The function[10] $\phi^M : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ is a known feature map, and $\theta^* \in \mathbb{R}^d$ is an unknown vector. The goal of the forecaster is to minimize its regret, namely $R(T) \coloneqq \sum_{t=1}^T \mathbb{E}\left[\langle A_t^* - A_t, \theta^*\rangle\right]$, where at any time t, conditional on the context X_t, $A_t^* \in \text{argmax}_{a\in\mathcal{A}} \langle \theta^*, \phi^M(X_t, a)\rangle$. Thus, A_t^* is a random variable as X_t is random.

To describe the model selection, we consider a sequence of M dimensions $1 \leq d_1 < d_2, \cdots < d_M \coloneqq d$ and an associated set of feature maps $(\phi^m)_{m=1}^M$, where for any $m \in [M]$, $\phi^m(\cdot,\cdot) : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^{d_m}$, is a feature map embedding into d_m dimensions. Moreover, these feature maps are nested, namely, for all $m \in [M-1]$, for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, the first d_m coordinates of $\phi^{m+1}(x,a)$ equals $\phi^m(x,a)$. The forecaster is assumed to have knowledge of these feature maps. The unknown vector θ^* is such that its first d_{m^*} coordinates are non-zero, while the rest are 0. The forecaster does not know the true dimension d_{m^*}. If this were known, then standard contextual bandit algorithms such as LinUCB (Chu et al. (2011)) can yield a regret of $\widetilde{O}(\sqrt{d_{m^*}T})$. In this section, we provide an algorithm in which, even when the forecaster is unaware of d_{m^*}, the regret scales as $\widetilde{O}(\sqrt{d_{m^*}T})$. However, this result is not uniform over all θ^* as, we will show, depends on the minimum non-zero coordinate value in θ^*.

Model Assumptions We will require some assumptions identical to the ones stated in Foster et al. (2019). Let $\|\theta^*\|_2 \leq 1$, which is known to the forecaster. The distribution $\mathcal{D}$ is assumed to be known to the forecaster. Associated with the distribution $\mathcal{D}$ is a matrix $\Sigma_M \coloneqq \frac{1}{K}\sum_{a\in\mathcal{A}} \mathbb{E}\left[\phi^M(x,a)\phi^M(x,a)^T\right]$ (where $x \sim \mathcal{D}$), where we assume its minimum eigen value $\lambda_{min}(\Sigma_M) > 0$ is strictly positive. Further, we assume that, for all $a \in \mathcal{A}$, the random variable $\phi^M(x,a)$ (where $x \sim \mathcal{D}$ is random) is a sub-gaussian random variable with (known) parameter τ^2.

4.5 `ALB-Dim` Algorithm

The algorithm here is identical to that of Algorithm 2, except that in place of OFUL, we use `SupLinRel` of Chu et al. (2011) as the black-box. The details of the Algorithm are provided in Appendix C.

[10] Superscript M will become clear shortly

4.6 Main Result

For brevity, we only state the Corollary of our main Theorem (Theorem 3) which is stated in Appendix C.

Corollary 2. *Suppose Algorithm 3 is run with input parameters $\delta \in (0,1)$, and $T_0 = \widetilde{O}\left(d^2 \ln^2\left(\frac{1}{\delta}\right)\right)$ given in Equation (15) , then with probability at-least $1-\delta$, the regret after T times satisfies*

$$R_T \leq O\left(\frac{d^2}{\gamma^{4.65}} \ln^2(d/\delta)\tau^2 \ln\left(\frac{TK}{\delta}\right)\right) + \widetilde{O}(\sqrt{Td^*_m}),$$

where $\gamma = \min\{|\theta^*_i| : \theta^*_i \neq 0\}$ and θ^* is d^* sparse.

Discussion - Our regret scaling matches that of an oracle that knows the true problem complexity and thus obtains a regret of $\widetilde{O}(\sqrt{d_{m^*}T})$. This, thus improves on the rate compared to that obtained in Foster et al. (2019), whose regret scaling is sub-optimal compared to the oracle. On the other hand however, our regret bound depends on γ and is thus not uniform over all θ^*, unlike Foster et al. (2019) that is uniform over θ^*. Thus, in general, our results are not directly comparable to that of Foster et al. (2019). It is an interesting future work to close the gap and in particular, obtain the regret matching that of an oracle to hold uniformly over all θ^*.

5 SIMULATIONS

5.1 Synthetic Experiments

We compare `ALB-Norm` with the (non-adaptive) OFUL$^+$ and an *oracle* that knows the problem complexity apriori. The oracle just runs OFUL$^+$ with the known problem complexity. We choose the bias $\sim \mathcal{U}[-1,1]$, and the additive noise to be zero-mean Gaussian random variable with variance 0.5. At each round of the learning algorithm, we sample the context vectors from a d-dimensional standard Gaussian, $\mathcal{N}(0, I_d)$. We select $d = 50$, the number of arms, $K = 75$, and the initial epoch length as 100. In particular, we generate the true θ^* in 2 different ways: (i) $\|\theta^*\| = 0.1$, but the initial estimate $b_1 = 10$, and (ii) $\|\theta^*\| = 1$, with the initial estimate $b_1 = 10$.

In panel (a) and (b) of Figure 1, we observe that, in setting (i), OFUL$^+$ performs poorly owing to the gap between $\|\theta^*\|$ and b_1. On the other hand, `ALB-Norm` is sandwiched between the OFUL$^+$ and the oracle. Similar things happen in setting (ii). In panel (c), we show that the norm estimates of `ALB-Norm` improves over epochs, and converges to the true norm very quickly.

In panel (d)-(f), we compare the performance of `ALB-Dim` with OFUL (Abbasi-Yadkori et al. (2011)) and an *oracle* who knows the true support of θ^* apriori. For computational ease, we set $\varepsilon_i = 2^{-i}$ in simulations. We select θ^* to be $d^* = 20$-sparse, with the smallest non-zero component, $\gamma = 0.12$. We have 2 settings: (i) $d = 500$ and (ii) $d = 200$. In panel (d) and (e), we observe a huge gap in cumulative regret between `ALB-Dim` and OFUL, thus showing the effectiveness of dimension adaptation. In panel (f), we plot the successive dimension refinement over epochs. We observe that within $4-5$ epochs, `ALB-Dim` finds the sparsity of θ^*.

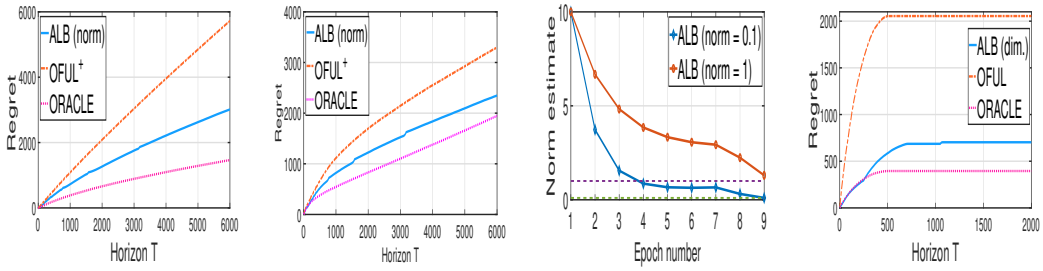
5.2 Real-data experiment

Here, we evaluate the performance of `ALB-Norm` on Yahoo! 'Learning to Rank Challenge' dataset (Chapelle and Chang (2010)). In particular, we use the file `set2.test.txt`, which consists of 103174 rows and 702 columns. The first column denotes the rating, $\{0, 1, ., 4\}$ given by the user (which is taken as reward); the second column denotes the user id, and the rest 700 columns denote the context of the user. After selecting $20,000$ rows and 50 columns at random (several other random selections yield similar results), we cluster the data by running k means algorithm with $k = 500$. We treat each cluster as a bandit arm with mean reward as the empirical mean of the individual rating in the cluster, and the context as the centroid of the cluster. This way, we obtain a bandit setting with $K = 500$ and $d = 50$.

Assuming (reward, context) coming from a linear model (with bias, see Section 3.1), we use `ALB-Norm` to estimate the bias and θ^* simultaneously. In panel (g), we plot the cumulative reward accumulated over time. We observe that the reward is accumulated over time in an almost linear fashion. We also plot the norm estimate, $\|\theta^*\|$ over epochs in panel (h), starting with an initial estimate of 25. We observe that within 6 epochs the estimate stabilizes to a value of 11.1. This shows that `ALB-Norm` adapts to the actual $\|\theta^*\|$.

6 Conclusion

We considered refined model selection for linear bandits, by defining new notions of complexity. We gave two novel algorithms `ALB-Norm` and `ALB-Dim` that successively refines the hypothesis class and achieves model selection guarantees; regret scaling in the complexity of the smallest class containing the true model. This is the first such algorithm to achieve regret scaling similar to an oracle that knew the problem complexity. An interesting direction of future work is to derive regret bounds for the case when the dimension is a measure of complexity, that hold uniformly over all θ^*, i.e., have no explicit dependence on γ.



(a) $||\theta^*|| = 0.1,\ b_1 = 10$ (b) $||\theta^*|| = 1,\ b_1 = 10$ (c) Estimates of $||\theta^*||$ (d) $d^* = 20,\ d = 500$

(e) $d^* = 20,\ d = 200$ (f) Dimension refinement (g) Yahoo; $b_1 = 25$ (h) Yahoo; $b_1 = 25$

Figure 1: Synthetic and real-data experiments, validating the effectiveness of Algorithm 1 and 2. All the results are averaged over 25 trials.

Acknowledgements: The authors would like to acknowledge Akshay Krishnamurthy, Dylan Foster and Haipeng Luo for insightful comments and suggestions.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320.

Agarwal, A., Luo, H., Neyshabur, B., and Schapire, R. E. (2016). Corralling a band of bandit algorithms. *arXiv preprint arXiv:1612.06246*.

Arlot, S., Bartlett, P. L., et al. (2011). Margin-adaptive model selection in statistical learning. *Bernoulli*, 17(2):687–713.

Audibert, J.-Y. and Bubeck, S. (2010). Best arm identification in multi-armed bandits.

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

Auer, P., Gajane, P., and Ortner, R. (2018). Adaptively tracking the best arm with an unknown number of distribution changes. In *European Workshop on Reinforcement Learning*, volume 14, page 375.

Bastani, H. and Bayati, M. (2020). Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294.

Birgé, L., Massart, P., et al. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375.

Carpentier, A. and Munos, R. (2012). Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *Artificial Intelligence and Statistics*, pages 190–198.

Chapelle, O. and Chang, Y. (2010). Yahoo! learning to rank challenge overview. In *Proceedings of the 2010 International Conference on Yahoo! Learning to Rank Challenge - Volume 14*, YLRC'10, page 1–24. JMLR.org.

Chatterji, N. S., Muthukumar, V., and Bartlett, P. L. (2019). Osom: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. *arXiv preprint arXiv:1905.10040*.

Cherkassky, V. (2002). Model complexity control and statistical learning theory. *Natural computing*, 1(1):109–133.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214.

Cutkosky, A. and Boahen, K. (2017). Online learning without prior information. *arXiv preprint arXiv:1703.02629*.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Foster, D. J., Krishnamurthy, A., and Luo, H. (2019). Model selection for contextual bandits. In *Advances in Neural Information Processing Systems*, pages 14714–14725.

Ghosh, A., Chowdhury, S. R., and Gopalan, A. (2017). Misspecified linear bandits. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Krikheli, M. and Leshem, A. (2018). Finite sample performance of linear least squares estimators under sub-gaussian martingale difference noise. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4444–4448. IEEE.

Krishnamurthy, A., Wu, Z. S., and Syrgkanis, V. (2018). Semiparametric contextual bandits. *arXiv preprint arXiv:1803.04204*.

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

Locatelli, A. and Carpentier, A. (2018). Adaptivity to smoothness in x-armed bandits. In *Conference on Learning Theory*, pages 1463–1492.

Lugosi, G., Nobel, A. B., et al. (1999). Adaptive model selection using empirical complexities. *The Annals of Statistics*, 27(6):1830–1864.

Luo, H. and Schapire, R. E. (2015). Achieving all with no parameters: Adanormalhedge. In *Conference on Learning Theory*, pages 1286–1304.

Lykouris, T., Sridharan, K., and Tardos, É. (2017). Small-loss bounds for online learning with partial information. *arXiv preprint arXiv:1711.03639*.

McMahan, B. and Abernethy, J. (2013). Minimax optimal algorithms for unconstrained linear optimization. In *Advances in Neural Information Processing Systems*, pages 2724–2732.

Orabona, F. (2014). Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems*, pages 1116–1124.

Vapnik, V. (2006). *Estimation of dependences based on empirical data*. Springer Science & Business Media.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.