

# Model Selection for Generic Contextual Bandits

**Avishek Ghosh**

*Department of Electrical Eng. and Computer Science, UC Berkeley*

AVISHEK\_GHOSH@BERKELEY.EDU

**Abishek Sankararaman**

*AWS AI Labs, Palo Alto, USA*

ABISANKA@AMAZON.COM

**Kannan Ramchandran**

*Department of Electrical Eng. and Computer Science, UC Berkeley*

KANNANR@BERKELEY.EDU

**Editor:**

## Abstract

We consider the problem of model selection for the general stochastic contextual bandits under the realizability assumption. We propose a successive refinement based algorithm called Adaptive Contextual Bandit (ACB), that works in phases and successively eliminates model classes that are too simple to fit the given instance. We prove that this algorithm is adaptive, i.e., the regret rate order-wise matches that of FALCON, the state-of-art contextual bandit algorithm of Simchi-Levi and Xu (2020), that needs knowledge of the true model class. The price of not knowing the correct model class is only an additive term contributing to the second order term in the regret bound. This cost possess the intuitive property that it becomes smaller as the model class becomes easier to identify, and vice-versa. We then show that a much simpler explore-then-commit (ETC) style algorithm also obtains a regret rate of matching that of FALCON, despite not knowing the true model class. However, the cost of model selection is higher in ETC as opposed to in ACB, as expected. Furthermore, ACB applied to the linear bandit setting with unknown sparsity, order-wise recovers the model selection guarantees previously established by algorithms tailored to the linear setting.

## 1. Introduction

The Contextual Multi Armed Bandit (MAB) problem is a fundamental online learning setting that aims to capture the exploration-exploitation trade-offs associated with sequential decision making (c.f. Cesa-Bianchi and Lugosi (2006); Chu et al. (2011)). It consists of an agent, who at each time is shown a context by nature, and subsequently makes an irrevocable decision from a set of available decisions (arms) and collects a noisy reward depending on the arm chosen and the observed context. The agent initially has no knowledge of the rewards of the various actions, and has to learn by repeated interaction over time, the mapping from the set of context and arms to rewards. The agent's goal is to minimize regret—the expected difference between the reward collected by an oracle that knows the expected rewards of all actions under all possible observed contexts and that of the agent. The recent books of Lattimore and Szepesvári (2020), Slivkins (2019) and the references therein provide comprehensive state-of-art on the general bandit problem.

We study the model selection question in general stochastic contextual bandits (c.f. Agarwal et al. (2014a), Agarwal et al. (2012), Simchi-Levi and Xu (2020), Foster and Rakhlin

(2020a)) . In this setting, at the beginning of each round  $t \in [T]$ , nature reveals a context  $x_t \in \mathcal{X}$  to an agent, who then subsequently takes an action  $a_t \in \mathcal{A}$  from a finite set, and obtains a reward  $r_t$ . In the stochastic setting (the focus of the present paper), the set of contexts  $(x_t)_{t=1}^T$  are assumed to be i.i.d. random variables, with an arbitrary and apriori unknown probability distribution over  $\mathcal{X}$ . At each time  $t$ , conditional on the context  $x_t$  and the action taken  $a_t$ , the observed reward  $r_t$  is independent of everything else, with the mean  $\mathbb{E}[r_t|x_t, a_t] = f^*(x_t, a_t)$ , where  $f^* : \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$ , is an apriori unknown function. The agent is given a finite, nested family  $(\mathcal{F}_m)_{m=1}^M$  of hypothesis classes<sup>1</sup>, where  $1 \leq m_1 < m_2 \leq M$  implies  $\mathcal{F}_{m_1} \subseteq \mathcal{F}_{m_2}$ . Further, there exists an optimal class  $d^* := \inf\{1 \leq m \leq M : f^* \in \mathcal{F}_m\}$ , i.e.,  $\mathcal{F}_{d^*}$  is the smallest hypothesis class containing the unknown reward function  $f^*$ . The agent is not aware of  $d^*$  apriori and needs to estimate it. Model selection guarantees then refers to algorithms for the agent whose regret scales in the complexity of the *smallest hypothesis class* ( $\mathcal{F}_{d^*}$  in the above notation) containing the true model, even though the algorithm was not aware apriori.

In the case when the agent has knowledge that  $f^* \in \mathcal{F}_{d^*}$  but does not know  $f^*$ , Simchi-Levi and Xu (2020) recently gave the first computationally efficient algorithm FALCON, that achieves regret-rate scaling as  $\sqrt{T}$ . It was shown in Simchi-Levi and Xu (2020), that since  $g \in \mathcal{F}_{d^*}$  (i.e., realizable model), the stochastic contextual bandit can be reduced to an offline regression problem, which can be efficiently solved for many well known function classes (for ex. the set of all convex functions (Ghosh et al., 2019)). The regret of FALCON was shown to scale proportional to the square root of the complexity of the function class  $\mathcal{F}_{d^*}$  times  $T$ , the time horizon. In the case when  $\mathcal{F}_{d^*}$  is a finite set, the complexity equals the logarithm of the cardinality, while if the class is infinite (either countable or uncountable), complexity is analogously defined (c.f. Section 5).

The study in this paper is reliant on two assumptions: (i) *Realizability* (Assumption 1), —the true model belongs to at-least one of the many nested hypothesis classes, and (ii) *Separation* (Assumption 2) —the excess risk under any of the plausible model classes not containing the true model is strictly positive. Realizability, has been a standard assumption in stochastic contextual bandits (Foster et al. (2019), Foster and Rakhlin (2020a), Simchi-Levi and Xu (2020)), and is used in our setup to define the optimal model class that needs to be selected. The separation assumption is needed to ensure that not selecting a reliable model class leads to regret scaling linear in time. The separation assumption is analogous to that used in standard multi-armed bandits (Lattimore and Szepesvári, 2020), where the mean reward of the best arm is strictly larger than that of the second best arm.

In parallel independent work, Krishnamurthy and Athey (2021) also study model selection problem, under the same assumptions of realizability and separation that we make. They propose ModIGW algorithm that is built on FALCON and shares similarity to our algorithm ACB; both algorithms run in epochs of doubling length, where at the beginning of each epoch, an appropriate model class is selected, and the rest of the epoch consists of playing FALCON on the selected model class. In order to select the appropriate class, the nested structure of model classes along with the fact that the largest class  $M$  is realizable by definition is used. The regret guarantees are similar for both ACB and ModIGW, with ModIGW having a better second order term, as they have a stronger assumption on the regression

---

1. We use the term hypothesis class and model class interchangeably

oracle. Remark 5 highlights that under the same assumption on the regression oracle, the second order term of ACB will match (order-wise) that of ModIGW. The ACB Algorithm and its analysis can be viewed as a meta-algorithm, that uses FALCON, the state-of-art contextual bandit algorithm as a black-box to yield model selection guarantees. Thus any improvement to the contextual bandit problem, automatically yields a model selection result through ACB.

### 1.1 Our Contributions

We classify our contributions into three.

**1. A Successive Refinement Algorithm for General Contextual Bandit** - We present **Adaptive Contextual Bandit (ACB)** that matches (order-wise), the regret rate of FALCON (Simchi-Levi and Xu (2020)), the state of art algorithm in contextual bandits which assumes knowledge of the true model class. ACB proceeds in epochs, with the first step in every epoch being a statistical test on the samples from the previous epoch to identify the smallest model class, followed by FALCON on this identified class in the epoch. We show that, with probability 1, eventually, ACB identifies the true model class (Lemma 1), and thus its regret rate matches that of FALCON. ACB can be viewed as a meta-algorithm, that uses FALCON, the state-of-art contextual bandit algorithm as a black-box. Thus any improvement to the contextual bandit problem, automatically yields a model selection result through ACB.

**Cost of model selection:** The second order regret term in ACB scales as  $O\left(\frac{\log(T)}{\Delta^2}\right)$ , where  $\Delta > 0$ , is the gap (formally defined in Assumption 2) between the smallest class containing the true model and the highest model class not containing the true model. This term can be interpreted as the *cost of model selection*. Furthermore, as this term is inversely proportional to the gap  $\Delta$ , we see that an ‘easier’ instance ( $\Delta$  being high), incurs lower cost of model selection than an instance with smaller  $\Delta$ . Furthermore, the model selection cost can be reduced to  $\mathcal{O}\left(\frac{\log \log T}{\Delta^2}\right)$  if  $T$  is known in advance.

**2. An Explore-then-commit (ETC) algorithm, also achieves model selection, but has a larger second order regret compared to ACB .** We show that a ETC algorithm also performs model selection, i.e., has a regret rate scaling as that of FALCON on the optimal model class. However, the cost of model selection in ETC is  $O\left(\frac{\log(T)}{\Delta^4}\right)$ , which is larger than that of ACB. Nevertheless, asymptotically, a simple ETC algorithm suffices to obtain model selection.

**3. Improved Regret Guarantee with Linear Structure**—In the special setup of stochastic linear bandits, where the reward is a linear map of the context, we propose and analyze an adaptive algorithm, namely Adaptive Linear Bandits-Dimension (ALB-Dim). We show that the regret of ALB-Dim is independent of the number of actions (arms), which is an improvement over the regret of ACB. Furthermore, we show that, when ACB is applied in the linear bandit setting where each hypothesis classes specifies the sparsity of the linear bandit parameter, the regret guarantee matches order-wise, upto a  $\sqrt{|\mathcal{A}|}$  factor to that of ALB-Dim. On the other hand, ALB-Dim provides model selection guarantees even when the number of actions (arms) is infinite.

**Motivating Example:** Model selection in contextual bandits plays a key role in applications such as personalized recommendation systems, which we sketch. Consider a system

(such as news recommendation) that on each day, recommends one out of  $K$  possible outlets to a user. On each day, an event is realized in nature, which can be modeled as the context vector on that day. The true model function  $f^*$  encodes the user’s preference; for example the user prefers one outlet for sports oriented articles, while another for international events. This apriori unknown to the system and needs to learn this through repeated interactions. The multiple nested hypothesis classes corresponds to a variety of possible neural network architectures to learn the mapping from contexts (event of the day) to rewards (which can be engagement with the recommended item). In practice, these nested hypothesis classes range from simple logistic regression to multi-layer perceptrons (Cheng et al., 2016). Complex network architectures although has the potential for increased accuracy, incurs undesirable overheads such as requiring larger offline training to deliver accuracy gains (Cheng et al., 2016), computational complexity in hyper-parameter tuning (Caselles-Dupré et al., 2018) and challenges of explainability in predictions (McInerney et al., 2018; Balog et al., 2019). Model selection provides a framework to trade-off between accuracy and the overheads.

## 2. Related Work

Model selection for MAB have received increased attention in recent times owing to its applicability in a variety of large-scale settings such as recommendation systems and personalization. The special case of linear contextual bandits was studied in Chatterji et al. (2019), Ghosh et al. (2021) and Foster et al. (2019), where both instance dependent and instance independent algorithms achieving model selection were given. In this linear bandit framework, similar to the present paper, Foster et al. (2019) and Ghosh et al. (2021) considered the family of nested hypothesis classes, with each class positing the sparsity of the unknown linear bandit parameter. In this setup, Foster et al. (2019) proposed `ModCB` which achieves regret rate uniformly for all instances, a rate that is sub-optimal compared to the oracle that knows the true sparsity. In contrast, both our paper and Ghosh et al. (2021) propose an algorithm that achieves regret rate matching that of the oracle that knows the true sparsity. The cost of model selection contributes only a constant that depends on the instance but independent of the time horizon. However, unlike `ModCB`, our regret guarantees are problem dependent and do not hold uniformly for all instances. A parallel line of work on linear bandits has focused on simple LASSO type algorithms under strong stochastic assumptions on the distribution of the contexts that achieve model selection guarantees (Bastani and Bayati, 2020a; Bastani et al., 2021; Oh et al., 2020; Ariu et al., 2020; Li et al., 2021).

A black-box model selection framework for MABs called `Corral` was proposed in Agarwal et al. (2017), where the optimal algorithm for each hypothesis class is treated as an expert and the task of the forecaster is to have low regret with respect to the best expert (best model class). The generality of this framework has rendered it fruitful in a variety of different settings; for example Agarwal et al. (2017); Arora et al. (2021) considered unstructured MABs, which was then extended to both linear and contextual bandits and linear reinforcement learning in a series of works (Pacchiano et al., 2020a,b) and lately to even reinforcement learning Lee et al. (2021). However, the price for this versatility is that the regret rates the cost of model selection is multiplicative rather than additive. In particular, for the special case of linear bandits and linear reinforcement learning, the regret scales as  $\sqrt{T}$  in time with an

additional multiplicative factor of  $\sqrt{M}$ , while the regret scaling with time is strictly larger than  $\sqrt{T}$  in the general contextual bandit.

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.

**Notation:** Throughout the paper, we use  $C, C_1, C_2, \dots, c, c_1, c_2, \dots$  to denote universal positive constants, the value of which may change from instance to instance. Also, for a positive integer  $r$ , we denote the set  $\{1, 2, \dots, r\}$  by the shorthand  $[r]$ . Also,  $a \lesssim b$  means  $a \leq Cb$  for a universal constant  $C$ . Similarly  $a \gtrsim b$  implies  $a \geq Cb$  for a positive constant  $C$ . Also,  $\|\cdot\|$  denotes  $\ell_2$  norm of a vector unless otherwise specified. For a symmetric matrix  $A$ , we denote the maximum and minimum eigenvalues as  $\lambda_{\max}(A)$  and  $\lambda_{\min}(A)$  respectively.

### 3. Problem Setup

#### 3.1 Preliminaries

**Setup:** Let  $\mathcal{A}$  be the set of  $K$  actions, and let  $\mathcal{X} \subseteq \mathbb{R}^d$  be the set of  $d$  dimensional contexts. At time  $t$ , nature picks  $(x_t, r_t)$  in an i.i.d fashion, where  $x_t \in \mathcal{X}$  and a context dependent  $r_t \in [0, 1]^K$ . All expectation operators in this section are with respect to this i.i.d. sequence  $(x, r)$ . Upon observing the context, the agent takes action  $a_t \in \mathcal{A}$ , and obtains the reward of  $r_t(a_t)$ . Note that, the reward  $r_t(a_t, x_t)$  depends on the context  $x_t$  and the action  $a_t$ . Furthermore, it is standard (Foster et al. (2019); Simchi-Levi and Xu (2020)) to have a realizability assumption on the conditional expectation of the reward, i.e., there exists a predictor  $f^* \in \mathcal{F}$ , such that  $\mathbb{E}[r_t(a, x)|x, a] = f^*(x, a)$ , for all  $x$  and  $a$ . We suppress the dependence of the reward on the context  $x_t$  and denote the reward at time  $t$  from action  $a \in \mathcal{A}$  as  $r_t(a)$ .

In the contextual bandit literature (Agarwal et al. (2012); Simchi-Levi and Xu (2020)) it is generally assumed that the true regression function  $f^*$  is unknown, but the function class  $\mathcal{F}$  where it belongs, is known to the learner. The price of not knowing  $f^*$  is characterized by regret, which we define now. To set up notation, for any  $f \in \mathcal{F}$ , we define a policy induced by the function  $f$ ,  $\pi_f : \mathcal{X} \rightarrow \mathcal{A}$  as  $\pi_f(x) = \operatorname{argmin}_{a \in \mathcal{A}} f(x, a)$ <sup>2</sup>, for all  $x \in \mathcal{X}$ . We

---

2. Ties are broken arbitrarily, for example the lexicographic ordering of  $\mathcal{A}$

define the regret over  $T$  rounds as the following:

$$R(T) = \sum_{t=1}^T [r_t(\pi_{f^*}(x_t)) - r_t(a_t)].$$

#### 4. Model Selection for General Contextual Bandits

In this section, we focus on the main contribution of the paper—a provable model selection guarantee for the (generic) stochastic contextual bandit problem. In contrast to the standard setting, in the model selection framework, we do not know  $\mathcal{F}$ . Instead, we are given a nested class of  $M$  function classes,  $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots \subset \mathcal{F}_M$ . Let the smallest function class where the true regressor,  $f^*$  lies be denoted by  $\mathcal{F}_{d^*}$ , where  $d^* \in [M]$ .

From the above discussion, since  $f^* \in \mathcal{F}_{d^*}$ , the regret of an *adaptive* contextual bandit algorithm should depend on the function class  $\mathcal{F}_{d^*}$ . However, we do not know  $d^*$ , and our goal is to propose adaptive algorithms such that the regret depends on the *actual* problem complexity  $\mathcal{F}_{d^*}$ . First, let us write the realizability assumption with the nested function classes.

**Assumption 1** (Realizability). *There exists  $1 \leq d^* \leq M$ , and a predictor  $f^* \in \mathcal{F}_{d^*}$ , such that  $\mathbb{E}[r_t(a)|x] = f^*(x, a)$ , for all  $x$  and  $a$ .*

Furthermore, in order to identify the correct model class within the given  $M$  hypothesis classes, we also require the following separability condition. Note that similar separability condition is also witnessed in Krishnamurthy and Athey (2021).

**Assumption 2.** *There exists a  $\Delta > 0$ , such that,*

$$\inf_{f \in \mathcal{F}_{d^*-1}} \inf_{a \in \mathcal{A}} \mathbb{E}[f(x, a) - f^*(x, a)]^2 \geq \Delta.$$

*The parameter  $\Delta > 0$  is the minimum separation across the function classes. The expectation above is with respect to the distribution by which the contexts are selected.*

The above condition implies that there is a (non-zero) gap, between the regressor functions belonging to the realizable classes and non-realizable classes. Since, we have nested structure,  $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots \subset \mathcal{F}_M$ , condition on the biggest non-realizable class,  $\mathcal{F}_{d^*-1}$  is sufficient. Note that separability condition is quite standard in statistics, specially in the area of clustering (Lu and Zhou (2016)), analysis of Expectation Maximization (EM) algorithm (Kwon and Caramanis (2020); Balakrishnan et al. (2017), understanding the behavior of Alternating Minimization (AM) algorithms (Yi et al. (2016); Ghosh et al. (2019)). Model selection without separability condition is kept as an interesting future work. Note that although we require the gap assumption for theoretical analysis, our algorithm (described next) does not require any knowledge of  $\Delta$ .

##### 4.1 Algorithm—Adaptive Contextual Bandits (ACB)

In this section, we provide a novel model selection algorithm that use successive refinements over epochs. We use a provable contextual bandit algorithm, namely FALCON (stands for

FAst Least-squares-regression-oracle CONtextual bandits) of Simchi-Levi and Xu (2020), as a baseline, and add a model selection phase at the beginning of each epoch. In other words, over multiple epochs, we successively refine our estimates of the *proper* model class where the true regressor function  $f^*$  lies. The details are provided in Algorithm 1. Note that ACB does not require any knowledge of the separation  $\Delta$ . We first briefly discuss the FALCON algorithm.

**The Base Algorithm:** Recently, Simchi-Levi and Xu (2020) proposed and analyzed a contextual bandit algorithm, FALCON, which gives provable guarantees for contextual bandits beyond linear structure. FALCON is an epoch based algorithm, and depends only on an *offline regression oracle*, which outputs an estimate  $\hat{f}$  of the regression function  $f^*$  at the beginning of each epoch. FALCON then uses a randomization scheme, that depends on the inverse gap with respect to the estimate of the best action. Suppose that the true regressor  $f^* \in \mathcal{F}$ , and the realizability condition (Assumption 1) holds. With a proper choice of learning rate, with probability  $1 - \delta$ , FALCON yields a regret of

$$R(T) \leq \mathcal{O}(\sqrt{KT \log(|\mathcal{F}|T/\delta)}).$$

Although the above result makes sense only for the finite  $\mathcal{F}$ , an extension to the infinite  $\mathcal{F}$  is possible and was addressed in the same paper (see Simchi-Levi and Xu (2020)).

**Our Approach:** We use successive refinement based model selection strategy along with the base algorithm FALCON. The details of our algorithm, namely Adaptive Contextual Bandits (ACB) are given in Algorithm 1. We break the time horizon into several epochs with doubling epoch length. Let  $\tau_0, \tau_1, \dots$  be epoch instances, with  $\tau_0 = 0$ , and  $\tau_m = 2^m$ . Before the beginning of the  $m$ -th epoch, using all the data of the  $m - 1$ -th epoch, we add a model selection module, as shown in Algorithm 1 (lines 4-8).

Note that, in ACB, we feed the samples of the  $m - 1$ -th epoch to the offline regression oracle. Moreover, we split the samples in 2 equal halves. We use the first half to compute the regression estimate

$$\hat{f}_j^m = \operatorname{argmin}_{f \in \mathcal{F}_j} \sum_{t=\tau_{m-2}+1}^{\tau_{m-1}/2} (f(x_t, a_t) - r_t(a_t))^2$$

via offline regression oracle, for all  $m \in [M]$ . ACB then use the rest of the samples to construct the test statistics given by,

$$S_j^m = \frac{1}{2^{m-2}} \sum_{t=\tau_{m-1}/2+1}^{\tau_{m-1}} (\hat{f}_j^m(x_t, a_t) - r_t(a_t))^2$$

for all  $m \in [M]$ . We do not use the same set of samples to remove any dependence issues with  $\hat{f}_j^m$  and the samples  $\{x_t, a_t, r_t(a_t)\}_{t=\tau_{m-1}/2+1}^{\tau_{m-1}}$ .

ACB then compares the test statistics  $\{S_j^m\}_{m=1}^M$  in Line 8 of Algorithm 1 to pick the model class. Intuitively, we expect  $S_j^m$  to be small for all hypothesis classes that contain  $f_{d^*}^*$ . Otherwise, thanks to the separation condition in Assumption 2, we expect  $S_j^m$  to be large. Realizability, i.e., Assumption 1 ensures that  $\mathcal{F}_M$ , the largest hypothesis class by

**Algorithm 1:** Adaptive Cotextual Bandits (ACB)

- 
- 1: **Input:** epochs  $0 = \tau_0 < \tau_1 < \tau_2 < \dots$ , confidence parameter  $\delta$ , Function classes  $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots \subset \mathcal{F}_M$
  - 2: **for** epoch  $m = 1, 2, \dots$ , **do**
  - 3:    $\delta_m = \delta/2^m$
  - 4:   **for** function classes  $j = 1, 2, \dots, M$  **do**
  - 5:     Compute  $\hat{f}_j^m = \operatorname{argmin}_{f \in \mathcal{F}_j} \sum_{t=\tau_{m-2}+1}^{\tau_{m-1}/2} (f(x_t, a_t) - r_t(a_t))^2$  via offline regression oracle
  - 6:     Construct  $S_j^m = \frac{1}{2^{m-2}} \sum_{t=\tau_{m-1}/2+1}^{\tau_{m-1}} (\hat{f}_j^m(x_t, a_t) - r_t(a_t))^2$
  - 7:   **end for**
  - 8:   **Model Selection:** Find the minimum index  $\ell \in [M]$  such that  $S_j^m \leq S_M^m + \frac{\sqrt{m}}{2^{m/2}}$ . Let this class be denoted by  $\mathcal{F}_\ell^m$
  - 9:   Set learning rate  $\rho_m = (1/30) \sqrt{K(\tau_{m-1} - \tau_{m-2}) / \log(|\mathcal{F}_\ell^m|(\tau_{m-1} - \tau_{m-2})m/\delta_m)}$
  - 10:   **for** round  $t = \tau_{m-1} + 1, \dots, \tau_m$  **do**
  - 11:     Observe context  $x_t \in \mathcal{X}$
  - 12:     Compute  $\hat{f}_\ell^m(a)$  for all action  $a \in \mathcal{A}$ , set  $\hat{a}_t = \operatorname{argmax}_{a \in \mathcal{A}} \hat{f}_\ell^m(a)$
  - 13:     Define  $p_t(a) = \frac{1}{K + \rho_m(\hat{f}_\ell^m(x_t, \hat{a}_t) - \hat{f}_\ell^m(x_t, a))} \forall a \neq \hat{a}_t$ ,  $p_t(\hat{a}_t) = 1 - \sum_{a \neq \hat{a}_t} p_t(a)$ .
  - 14:     Sample  $a_t \sim p_t(\cdot)$  and observe reward  $r_t(a_t)$ .
  - 15:   **end for**
  - 16: **end for**
- 

definition contains the true model  $f^*$ . Thus  $S_M^m$  serves as an estimate of how small the excess risk of any realizable class must be. We set the threshold to be a small addition to  $S_M^m$ . The additional term of  $\frac{\sqrt{m}}{2^{m/2}}$  in Line 8 of Algorithm 1 is chosen so that it is not too small, but nevertheless goes to 0, as  $m \rightarrow \infty$ . In particular, we choose the threshold in ACB such that it is large enough to ensure all realizable classes have excess risk smaller than this threshold, but also not so large that it exceeds the excess risk of the non-realizable classes.

Let  $\mathcal{F}_\ell^m$  be function class selected by this procedure in epoch  $m$ . ACB now uses *inverse gap* randomization with properly chosen learning rate (see Simchi-Levi and Xu (2020); Foster and Rakhlin (2020b); Sen et al. (2021)) to select the action  $a_t$ . In particular, with  $\hat{f}_\ell^m$  as the regressor function, let  $\hat{a}_t = \operatorname{argmax}_{a \in \mathcal{A}} \hat{f}_\ell^m(a)$  be the greedy action. The *inverse gap* randomization  $p_t(\cdot)$  is defined in the following way:

$$p_t(a) = \frac{1}{K + \rho_m(\hat{f}_\ell^m(x_t, \hat{a}_t) - \hat{f}_\ell^m(x_t, a))} \forall a \neq \hat{a}_t, \quad p_t(\hat{a}_t) = 1 - \sum_{a \neq \hat{a}_t} p_t(a),$$

where  $K$  is the number of arms (actions) and  $\rho_m$  is the learning rate. Finally, we sample action  $a_t \sim p_t(\cdot)$  and henceforth observe reward  $r_t(a_t)$ .

## 4.2 Analysis of ACB

We now analyze the performance of the model selection procedure of Algorithm 1. We have the doubling epochs, i.e.,  $\tau_m = 2^m$ . Without loss of generality, we simply assume  $\tau_1 = 2$ . Also, assume that we are at the beginning of epoch  $m$ , and hence we have the samples from epoch  $m - 1$ . So, we have total of  $2^{m-1}$  samples, out of which, we use  $2^{m-2}$  to construct



the regression functions and the rest  $2^{m-2}$  to obtain the testing function  $S_j^m$ . Furthermore, we want the model selection procedure to succeed with probability at least  $1 - \delta/2^m$ , since we want a guarantee that holds for all  $m$ , and a simple application of the union bound yields that.

We first show that ACB identifies the correct function class with high probability after a few epochs. We have the following Lemma.

**Lemma 1** (Model Selection of ACB). *Suppose Assumptions 1 and 2 holds and we run Algorithm 1. Then, in all phases  $m$  such that*

$$2^m \gtrsim \max\left\{\frac{\log T}{\Delta^2}, \log(|\mathcal{F}_M|), \log(1/\delta)\right\}$$

*Algorithm 1 identifies the correct model class  $\mathcal{F}_{d^*}$  in Line 8, with probability exceeding  $1 - 2M\delta$ .*

*Proof sketch.* In order select the correct function class, we first obtain upper bounds on the test statistics  $S_j^{(m)}$  for model classes that includes the true regressor  $f_{d^*}^*$ . We accomplish this by first carefully bounding the expectation of  $S_j^{(m)}$  and then using concentration. We then obtain a lower bound on  $S_j^{(m)}$  for model classes not containing  $f_{d^*}^*$  via leveraging Assumption 2 (separability) along with Assumption 1. Combining the above two bounds yields the desired result.  $\square$

**Regret Guarantee:** With the above lemma, we obtain the following regret bound for Algorithm 1.

**Theorem 1.** *Suppose the conditions of Lemma 1 hold. Then with probability at least  $1 - 2M\delta - \delta$ , running Algorithm 1 for  $T$  iterations yield*

$$R(T) \leq C \max\left\{\frac{\log T}{\Delta^2}, \log(|\mathcal{F}_M|), \log(1/\delta)\right\} + \mathcal{O}\left(\sqrt{KT \log(|\mathcal{F}_{d^*}|T/\delta)}\right).$$

Several remarks are in order:

**Remark 1.** *The first term of the regret scales weakly with  $T$  (as  $\mathcal{O}(\log T)$ ). Hence, the regret scaling (with respect to  $T$ ) is  $\tilde{\mathcal{O}}(\sqrt{KT \log(|\mathcal{F}_{d^*}|T/\delta)})$ , with high probability, which matches (upto a log factor) the regret of an oracle knowing the true function class  $\mathcal{F}_{d^*}$ .*

**Remark 2.** *The first term can be interpreted as the cost of model selection. Hence, the model selection procedure only adds a  $\mathcal{O}\left(\frac{\log T}{\Delta^2}\right)$  term (minor term compared to the  $\sqrt{T}$  scaling) to the regret.*

**Remark 3.** *Algorithm 1 is parameter-free, i.e., does not need knowledge of  $\Delta$ . Nevertheless, the regret guarantee adapts to the problem hardness, i.e., if  $\Delta$  is small, the regret is larger and vice-versa.*

**Remark 4** (Improvement from  $\mathcal{O}(\log T)$  to  $\mathcal{O}(\log \log T)$  in the model selection cost). *We emphasize that the  $\mathcal{O}(\log T)$  factor in the cost of model selection term can be improved, if we have the knowledge of  $T$  apriori. In that setting, instead of substituting  $\delta_m = \delta/2^m$ , we substitute  $\delta_m = \delta/\log T$  for all phases  $m$ . Since the doubling epoch ensures a total of  $\mathcal{O}(\log T)$  epochs, this choice of  $\delta_m$  indeed works with a regret of*

$$R(T) \leq C \max\left\{\frac{1}{\Delta^2}, \log(|\mathcal{F}_M|), \log(\log T/\delta)\right\} + \mathcal{O}(\sqrt{KT \log(|\mathcal{F}_{d^*}|T \log T/\delta)}),$$

with probability at least  $1 - 2M\delta - \delta$ .

**Remark 5.** *The cost of model selection in Theorem 1, depends on the complexity of the largest model class  $\mathcal{F}_M$ . Under a stronger assumption on the regression oracle used in Line 5 of Algorithm 1 (for example Assumption 5 of Krishnamurthy and Athey (2021)), then the cost of model selection only depends on  $\mathcal{F}_{d^*}$  as opposed to  $\mathcal{F}_M$ .*

### 4.3 A Simple Explore-Then-Commit (ETC) Algorithm for Model Selection

In the previous section, we analyze ACB, which successively estimates the function class over epochs and use FALCON as a base algorithm. In this section, instead, we use a simple Explore-Then-Commit (ETC) algorithm for selecting the correct function class, and then commit to it during the exploitation phase. After a round of exploration, we do a (one-time) threshold based testing to estimate the function class, and after that, exploit the estimated function class for the rest of the iterations. We show that this simple strategy finds the optimal function class  $\mathcal{F}_{d^*}$  with high probability. The details are given in Algorithm 2. We now explain the exploration and exploitation phases of this algorithm.

For the first  $2\sqrt{T}$  time epochs, we do the exploration (i.e., sample randomly). Precisely, the context-reward pair  $(x_t, r_t)$  is being sampled by nature in an i.i.d fashion, and the action the agent takes is chosen uniformly at random from the action set  $\mathcal{A}$ . In particular, the action is chosen independent of the context  $x_t$ . Hence, this is a pure exploration strategy.

Based on the samples of the first  $\sqrt{T}$  rounds, we estimate the regression function  $\{\hat{f}_j\}_{j=1}^M$  for all the (hypothesis) function classes  $\mathcal{F}_1, \dots, \mathcal{F}_M$  via offline regression oracle (see Simchi-Levi and Xu (2020)) and obtain  $\hat{f}_j = \operatorname{argmin}_{f \in \mathcal{F}_j} (\sum_{t=1}^{\sqrt{T}} f(x_t, a_t) - r_t(a_t))^2$  for all  $j \in [M]$ .

To remove dependence issues, we use the remaining  $\sqrt{T}$  samples obtained from the sampling phase. Here we actually compute the following test statistic for all hypothesis classes, namely  $S_j = \frac{1}{\sqrt{T}} \sum_{t=1}^{\sqrt{T}} (\hat{f}_j(x_t, a_t) - r_t(a_t))^2$  for all  $j \in [M]$ . We then perform a thresholding on  $\{S_j\}_{j=1}^M$ . We pick the smallest index  $j$  such that  $S_j \leq S_M + \sqrt{\frac{\log(T)}{\sqrt{T}}}$ . We then commit to this function class for the rest  $T - 2\sqrt{T}$  time steps. Hence, in Algorithm 2, we perform one step thresholding and commit to it. We show that simple scheme obtains the correct model with high probability.

### 4.4 Regret Guarantee of ETC

Here, we analyze Algorithm 2 with large gap assumption. We have the following lemma on model selection with ETC:

---

**Algorithm 2:** ETC for model selection for contextual bandits
 

---

- 1: **Input:** Function classes  $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots \subset \mathcal{F}_M$ , time horizon  $T$ , confidence parameter  $\delta$
  - 2: **Explore:**
  - 3: **for**  $t = 1, 2, \dots, \lceil \sqrt{T} \rceil$  **do**
  - 4:   Observe context reward pair  $(x_t, r_t)$
  - 5:   Select action  $a_t$  uniformly at random from  $\mathcal{A}$ , independent of  $x_t$
  - 6:   Observe reward  $r_t(a_t)$
  - 7: **end for**
  - 8: Compute regression estimator  $\hat{f}_j = \operatorname{argmin}_{f \in \mathcal{F}_j} \frac{1}{\sqrt{T}} \sum_{t=1}^{\lceil \sqrt{T} \rceil} [f(x_t, a_t) - r_t(a_t)]^2$  (via offline regression oracle) for all  $j \in [M]$
  - 9: **Model Selection test:**
  - 10: Obtain another set of  $\lceil \sqrt{T} \rceil$  fresh samples of  $(x_t, r_t, a_t)$  via pure exploration (similar to line 4-6 )
  - 11: Construct the test statistic  $S_j = \frac{1}{\lceil \sqrt{T} \rceil} \sum_{t=1}^{\lceil \sqrt{T} \rceil} (\hat{f}_j(x_t, a_t) - r_t(a_t))^2$  for all  $j \in [M]$
  - 12: **Thresholding:** Find the minimum index  $\ell \in [M]$  such that  $S_j \leq S_M + \sqrt{\frac{\log T}{\sqrt{T}}}$ . We obtain the regressor  $\hat{f}_\ell \in \mathcal{F}_\ell$
  - 13: **Commit:**
  - 14: **for**  $t = 2\lceil \sqrt{T} \rceil + 1, \dots, T$  **do**
  - 15:   Observe context reward pair  $(x_t, r_t)$
  - 16:   Select action  $a_t$  according to *inverse gap* randomization of Algorithm 1 (lines 9-15) with the function class  $\mathcal{F}_\ell$  and observe reward  $r_t(a_t)$
  - 17: **end for**
- 

**Lemma 2** (Model Selection for ETC). *Suppose the time horizon satisfies*

$$T \gtrsim (\log T) \max \left( \log \left( \sqrt{T} |\mathcal{F}_M| \right), \Delta^{-4}, \log(1/\delta) \right).$$

*Then with probability at least  $1 - 4M\delta$ , line 11 in Algorithm 2 identifies the correct model class  $\mathcal{F}_{d^*}$ .*

We now analyze the regret performance of Algorithm 2. The regret  $R(T)$  is comprised of 2 stages; (a) exploration and (b) commit (exploitation). We have the following result.

**Theorem 2.** *Suppose Assumptions 2 and 3 hold. Then with probability at least  $1 - 4M\delta - \delta$ , running Algorithm 2 for  $T$  iterations yield*

$$R(T) \leq C (\log T) \max \left( \log \left( \sqrt{T} |\mathcal{F}_M| \right), \Delta^{-4}, \log(1/\delta) \right) + \mathcal{O} \left( \sqrt{KT \log(|\mathcal{F}_{d^*}|T/\delta)} \right).$$

**Remark 6.** *Asymptotically in time, ETC matches the guarantee of an oracle which knows the true model class a priori.*

**Remark 7.** *The cost of running the simpler ETC algorithm is a worse dependence on the additive constant which scales as  $\frac{\log(T)}{\Delta^4}$  as opposed to  $\frac{\log(T)}{\Delta^2}$  in Algorithm 1. This shows that although asymptotically Algorithms 1 and 2 have identical scaling, the regret guarantee of Algorithm 1 is better, as expected.*

## 5. General Contextual Bandits with Infinite Function Classes

The results in Section 4 hold for finite function classes, since the regret bound depends on the cardinality of the function class. However, as shown in Simchi-Levi and Xu (2020), it can be easily extended to the infinite function class setting. Exploiting the notion of the complexity of infinite function classes, this reduction is done.

Like before, we consider a nested sequence of  $M$  function classes  $\mathcal{F}_1 \subset \dots \subset \mathcal{F}_M$ . The reward is sampled from an unknown function  $f_{d^*}^*$  lying in the (smallest) function class indexed by  $d^* \in [M]$ , which is unknown. Given the function classes, our job is to find the function class  $\mathcal{F}_{d^*}$ , and subsequently exploit the class to obtain sub-linear regret. Let us first rewrite the separability assumption.

We assume that the function classes  $\mathcal{F}_1 \subset \dots \subset \mathcal{F}_M$  are compact. This, in conjunction with the extreme value theorem, it is ensured that the following minimizers exist: for  $j < d^*$ , we define

$$\bar{f}_j = \operatorname{arginf}_{f \in \mathcal{F}_j} \mathbb{E}_{x,a}[f(x,a) - f_{d^*}^*(x,a)]^2$$

for all pairs  $(x,a)$ . For  $j \geq d^*$ , we know that this minimizer is indeed  $f_{d^*}^*$ . This comes directly from the realizability assumption. Note that we require the existence of the minimizer (regression function) in order to use it for selecting actions in the contextual bandit framework (see Simchi-Levi and Xu (2020))

Having defined the minimizers, we rewrite the separability assumption as following:

**Assumption 3.** *For any  $\bar{f}_j$ , where  $j < d^*$ , we have*

$$\mathbb{E}_{x,a}[\bar{f}_j(x,a) - f_{d^*}^*(x,a)]^2 \geq \Delta,$$

for all pairs  $(x,a) \in \mathcal{X} \times \mathcal{A}$ .

Similar to Simchi-Levi and Xu (2020), here, we are not worried about the explicit form of the regression functions  $\bar{f}_j$ . Rather, we assume the following performance guarantee of the offline regressor. For  $j \geq d^*$  (meaning, the class containing the true regressor  $f_{d^*}^*$ ), we have the following assumption.

**Assumption 4.** *Given  $n$  i.i.d data samples  $(x_1, a_1, r_1(a_1)), (x_2, a_2, r_2(a_2)), \dots, (x_n, a_n, r_n(a_n))$ , the offline regression oracle returns a function  $\hat{f}_j$ , such that for  $\delta > 0$ , with probability at least  $1 - \delta$ ,*

$$\mathbb{E}_{x,a}[\hat{f}_j(x,a) - f_{d^*}^*(x,a)]^2 \leq \xi_{\mathcal{F}_j, \delta}(n)$$

This assumption is taken from (Simchi-Levi and Xu, 2020, Assumption 2). As discussed in the above-mentioned paper, the quantity  $\xi_{(\cdot, \cdot)}(n)$  is a decreasing function of  $n$ , e.g.,  $\xi_{(\cdot, \cdot)}(n) = \tilde{O}(1/n)$ . As an instance, consider the class of all linear regressors in  $\mathbb{R}^d$ . In that case,  $\xi_{(\cdot, \cdot)}(n) \sim \tilde{O}(d/n)$ . For function classes with finite VC dimension (or related quantities like VC-sub graph or fat-shattering dimension; pseudo dimension in general, denoted by  $\tilde{d}$ ), we have  $\xi_{(\cdot, \cdot)}(n) \sim \tilde{O}(\tilde{d}/n)$ .

Here, to avoid repetition, we do not present all our previous results in the infinite function class setting. We consider two instances:

1. The adaptive contextual bandit (ACB) algorithm ( Algorithm 1).
2. The ETC algorithm ( Algorithm 2).

The model-selection algorithm remains more-or-less the same overall.

For Option I, we collect all the samples from the previous epoch of the FALCON algorithm, split the samples, to obtain the regression estimate  $\hat{f}_j^m$  and similarly construct test statistic  $S_j^m$  for all  $j \in [M]$ . In this setting, for the  $m$ -th epoch, with model chosen as  $\mathcal{F}_\ell$ , we set the learning rate (similar to the FALCON+ algorithm of Simchi-Levi and Xu (2020)) as

$$\rho_m = (1/30) \sqrt{K/\xi_{\mathcal{F}_\ell^m, \delta/2m^2}(\tau_{m-1} - \tau_{m-2})}.$$

For Option II, we explore for the first  $2\sqrt{T}$  rounds. The first  $\sqrt{T}$  rounds are used to collect samples  $(x_t, r_t, a_t)$  via pure exploration. Feeding this samples to the offline regression oracle, and focusing on the individual function classes  $\{\mathcal{F}_j\}_{j=1}^M$  separately, we obtain  $(\hat{f}_j, \xi_{\mathcal{F}_j, \delta}(\sqrt{T}))$  for all  $j \in [M]$ . Thereafter, we perform another round of pure exploration, and obtain  $\sqrt{T}$  fresh samples. Like in the finite case, we construct statistic  $S_j$  for all  $j \in [M]$ .

Similar to Algorithms 1 and 2, we choose the correct model based on a threshold on the test statistic  $S_j^m$  (for Option II, it is  $S_j$ ) and the threshold in phase  $m$  is  $\gamma^m := S_M^m \sqrt{\frac{m}{2^m}}$  ( $\gamma := S_M + \sqrt{\frac{\log T}{\sqrt{T}}}$  for Option II). We show that for all sufficiently large phase numbers, for all  $j \geq d^*$ ,  $S_j^m \leq \gamma^m$ , and for all  $j < d^*$ ,  $S_j^m > \gamma^m$  with high probability. Once this is shown, the model selection procedure follows exactly as Algorithm 2, i.e., we find the smallest index  $\ell \in [M]$ , for which  $S_\ell \leq \gamma^m$ . With high probability, we show that  $\ell = d^*$ .

**Regret Guarantee** We first show the guarantees for Option I, and Option II.

**Theorem 3.** (ACB with infinite function classes) *Suppose Assumptions 1, 3 and 4 hold. Then, with probability at least  $1 - 2M\delta - \delta$ , running Algorithm 1 for  $T$  iterations yield*

$$R(T) \leq C(\log T) \max\{\max_m 2^{m/2} \xi_{\mathcal{F}_M, 1/2^{m/2}}(2^{m-2}), \log(1/\delta), \Delta^{-2}\} + \mathcal{O}\left(\sqrt{K\xi_{\mathcal{F}_{d^*}, \delta/2T}(T) T}\right).$$

**Theorem 4.** (ETC with infinite function classes) *Suppose Assumptions 1, 3 and 4 hold. Then, provided,*

$$T \gtrsim (\log T) \max\left(T^{1/4} \xi_{\mathcal{F}_M, (1/T^{1/4})}, \Delta^{-4}, \log(1/\delta)\right),$$

*with probability at least  $1 - 4M\delta$ , line 11 in Algorithm 2 identifies the correct model class  $\mathcal{F}_{d^*}$ . Furthermore, running Algorithm 2 for  $T$  iterations yields, with probability at least  $1 - 2M\delta - \delta$ , the regret*

$$R(T) \leq (\log T) \max\left(T^{1/4} \xi_{\mathcal{F}_M, (1/T^{1/4})}, \Delta^{-4}, \log(1/\delta)\right) + \mathcal{O}\left(\sqrt{K\xi_{\mathcal{F}_{d^*}, \delta/2T}(T) T}\right).$$

**Remark 8.** *In both the settings, we match the regret of an oracle knowing the correct function class. We pay a small price for model selection.*

**Remark 9.** *The proof of these theorems parallels exactly similar to the finite function class setting. The only difference is that instead of upper-bounding the prediction error using Agarwal et al. (2012), we use the the definition of  $\xi(\cdot)$  to accomplish this.*

## 6. Model Selection in Stochastic Linear Bandits

In the previous sections, we consider the problem of model selection for general contextual bandits. Moreover, we assumed that the function classes are separable, and leveraging that we have several provable model selection algorithms. In this section, we consider a special case of model selection for stochastic linear bandits. We observe that with this linear structure, assumption like separability across function classes is not required.

In the linear bandit settings, we consider 2 different setup—(a) continuum (infinite) arm setting and (b) finite arm setting. We first start with the continuum arm setup.

### 6.1 Model Selection for Continuum (infinite) Arm Stochastic Linear bandits

#### 6.1.1 SETUP

We consider the standard stochastic linear bandit model in  $d$  dimensions (see Abbasi-Yadkori et al. (2011)), with the dimension as a measure of complexity. The setup comprises of a continuum collection of arms denoted by the set  $\mathcal{A} := \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ <sup>3</sup>. Thus, the mean reward from any arm  $x \in \mathcal{A}$  is  $\langle x, \theta^* \rangle$ , where  $\|\theta^*\| \leq 1$ . We assume that  $\theta^*$  is  $d^* \leq d$  sparse, where  $d^*$  is a priori unknown to the algorithm. For each time  $t \in [T]$ , if an algorithm chooses an arm  $x_t \in \mathcal{A}$ , the observed reward is denoted by  $y_t := \langle x_t, \theta^* \rangle + \eta_t$ , where  $\{\eta_t\}_{t \geq 1}$  is an i.i.d. sequence of 0 mean sub-gaussian random variables with known parameter  $\sigma^2$ .

We consider a sequence of  $d$  nested hypothesis classes, where each hypothesis class  $i \leq d$ , models  $\theta^*$  as a  $i$  sparse vector. The goal of the forecaster is to minimize the regret, namely

$$R(T) = \sum_{t=1}^T [\langle x_t^* - x_t, \theta^* \rangle],$$

where at any time  $t$ ,  $x_t$  is the action recommended by an algorithm and  $x_t^* = \operatorname{argmax}_{x \in \mathcal{A}} \langle x, \theta^* \rangle$ . The regret  $R(T)$  measures the loss in reward of the forecaster with that of an oracle that knows  $\theta^*$  and thus can compute  $x_t^*$  at each time.

Note that, we assume that the *true complexity* (dimension)  $d^* \leq d$  is initially unknown, and we seek algorithms that adapts to this unknown true dimension, rather than assume that the problem is  $d$  dimensional. This is in contrast to both the standard linear bandit setup (Chu et al., 2011; Abbasi-Yadkori et al., 2011), where there is no notion of complexity, as well as the line of work on sparse linear bandits (Bastani and Bayati, 2020b), where the *true sparsity (dimension)* is known, but only the set of which of the  $d^*$  out of the  $d$  coordinates is non-zero is unknown.

#### 6.1.2 ALGORITHM: ADAPTIVE LINEAR BANDITS (DIMENSION) [ALB-DIM]

We present our adaptive scheme in Algorithm 3. The algorithm is parametrized by  $T_0 \in \mathbb{N}$ , which is given in Equation (1) in the sequel and slack  $\delta \in (0, 1)$ . ALB-Dim proceeds in phases numbered  $0, 1, \dots$  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  $\theta^*$ , which is kept fixed throughout the phase. Concretely, each phase  $i$  is divided into two blocks:

---

3. 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 3:** Adaptive Linear Bandit (Dimension)
 

---

- 1: **Input:** Initial Phase length  $T_0$  and slack  $\delta > 0$ .
  - 2:  $\hat{\theta}_0 = \mathbf{1}$ ,  $T_{-1} = 0$
  - 3: **for** Each epoch  $i \in \{0, 1, 2, \dots\}$  **do**
  - 4:    $T_i = 36^i T_0$ ,    $\varepsilon_i \leftarrow \frac{1}{2^i}$ ,    $\delta_i \leftarrow \frac{\delta}{2^i}$
  - 5:    $\mathcal{D}_i := \{i : |\hat{\theta}_i| \geq \frac{\varepsilon_i}{2}\}$
  - 6:   **for** Times  $t \in \{T_{i-1} + 1, \dots, T_i\}$  **do**
  - 7:     Play OFUL(1,  $\delta_i$ ) only restricted to coordinates in  $\mathcal{D}_i$ . Here  $\delta_i$  is the probability slack parameter and 1 represents  $\|\theta^*\| \leq 1$ .
  - 8:   **end for**
  - 9:   **for** Times  $t \in \{T_i + 1, \dots, T_i + 6^i \lceil \sqrt{T_0} \rceil\}$  **do**
  - 10:     Play an arm from the action set  $\mathcal{A}$  chosen uniformly and independently at random.
  - 11:   **end for**
  - 12:    $\alpha_i \in \mathbb{R}^{S_i \times d}$  with each row being the arm played during all random explorations in the past.
  - 13:    $\mathbf{y}_i \in \mathbb{R}^{S_i}$  with  $i$ -th entry being the observed reward at the  $i$ -th random exploration in the past
  - 14:    $\hat{\theta}_{i+1} \leftarrow (\alpha_i^T \alpha_i)^{-1} \alpha_i \mathbf{y}_i$ , is a  $d$  dimensional vector
  - 15: **end for**
- 

1. a regret minimization block lasting  $36^i T_0$  time slots<sup>4</sup>,
2. followed by a random exploration phase lasting  $6^i \lceil \sqrt{T_0} \rceil$  time slots.

Thus, each phase  $i$  lasts for a total of  $36^i T_0 + 6^i \lceil \sqrt{T_0} \rceil$  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  $\theta^*$ . By notation,  $\mathcal{D}_0 = [d]$  and at the start of phase 0, the algorithm assumes that  $\theta^*$  is  $d$  sparse. 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  $\hat{\theta}_{i+1}$  of  $\theta^*$ , 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  $\hat{\theta}_{i+1}$  with magnitude exceeding  $2^{-(i+1)}$ . The pseudo-code is provided in Algorithm 3, 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  $i$ .

### 6.1.3 REGRET GUARANTEE

We first specify, how to set the input parameter  $T_0$ , as function of  $\delta$ . 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

---

4. We have not optimized over the constants like 36 and 6. Please refer to Remark 11 on this.

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

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

**Remark 10.**  $T_0$  in Equation (1) is chosen such that, at the end of phase 0,  $\mathbb{P}[\|\hat{\theta}_0 - \theta^*\|_\infty \geq 1/2] \leq \delta$  (Krikheli and Leshem, 2018). A formal statement of the Remark is provided in Lemma 3 in Appendix A.

**Theorem 5.** Suppose Algorithm 3 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

$$R_T \leq C \frac{T_0}{\gamma^{5.18}} T_0 + C_1 \sqrt{T} [1 + \sqrt{d^* \ln(1 + \frac{T}{d^*})} (1 + \sigma \sqrt{\ln(\frac{T}{T_0 \delta}) + d^* \ln(1 + \frac{T}{d^*})})].$$

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

In order to parse this result, we give the following corollary.

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

$$R_T \leq O\left(\frac{d^2}{\gamma^{5.18}} \ln^2(d/\delta)\right) + \tilde{O}(d^* \sqrt{T}).$$

**Remark 11.** The constants in the above Theorem are not optimized. The epoch length and the threshold parameter  $\varepsilon_i$  can be chosen more carefully. For example, if we set the epoch length as  $4^i T_0 + 2^i \sqrt{T_0}$  and the threshold  $\varepsilon_i$  as  $(0.9)^i$ , we obtain a worse dependence on  $\gamma$ . Furthermore, the exponent of  $\gamma$  can be made arbitrarily close to 4, by setting  $\varepsilon_i = C^{-i}$  in Line 4 of Algorithm 3, for some appropriately large constant  $C > 1$ , and increasing  $T_i = (C')^i T_0$ , for appropriately large  $C'$  ( $C' \approx C^4$ ).

**Discussion -** The regret of an oracle algorithm that knows the true complexity  $d^*$  scales as  $\tilde{O}(d^* \sqrt{T})$  (Carpentier and Munos, 2012; Bastani and Bayati, 2020b), matching ALB-Dim's regret, upto an additive constant independent of time. ALB-Dim is the first algorithm to achieve such model selection guarantees. On the other hand, standard linear bandit algorithms such as OFUL achieve a regret scaling  $\tilde{O}(d\sqrt{T})$ , which is much larger compared to that of ALB-Dim, especially when  $d^* \ll d$ , and  $\gamma$  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  $\gamma$  is lower. A small value of  $\gamma$  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  $\theta^*$  as it depends on  $\gamma$ . Obtaining regret rates matching the oracles and that hold uniformly over all  $\theta^*$  is an interesting avenue of future work.

## 6.2 Dimension as a Measure of Complexity - Finite Armed Setting

### 6.2.1 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  $\mathcal{A} := \{1, \dots, K\}$  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  $\sigma^2$  that is known to the forecaster. The function<sup>5</sup>  $\phi^M : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^d$  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  $R(T) := \sum_{t=1}^T \mathbb{E} [\langle A_t^* - A_t, \theta^* \rangle]$ , where at any time  $t$ , conditional on the context  $X_t$ ,  $A_t^* \in \operatorname{argmax}_{a \in \mathcal{A}} \langle \theta^*, \phi^M(X_t, a) \rangle$ . Thus,  $A_t^*$  is a random variable as  $X_t$  is random.

To describe the model selection, we consider a sequence of  $M$  dimensions  $1 \leq d_1 < d_2, \dots < d_M := d$  and an associated set of feature maps  $(\phi^m)_{m=1}^M$ , where for any  $m \in [M]$ ,  $\phi^m(\cdot, \cdot) : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^{d_m}$ , 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  $\theta^*$  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, than standard contextual bandit algorithms such as LinUCB Chu et al. (2011) can guarantee a regret scaling as  $\tilde{O}(\sqrt{d_{m^*} T})$ . In this section, we provide an algorithm in which, even when the forecaster is unaware of  $d_{m^*}$ , the regret scales as  $\tilde{O}(\sqrt{d_{m^*} T})$ . However, this result is non uniform over all  $\theta^*$  as, we will show, depends on the minimum non-zero coordinate value in  $\theta^*$ .

**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  $\Sigma_M := \frac{1}{K} \sum_{a \in \mathcal{A}} \mathbb{E} [\phi^M(x, a) \phi^M(x, a)^T]$  (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  $\tau^2$ .

---

5. Superscript  $M$  will become clear shortly

### 6.2.2 ALB-DIM ALGORITHM

The algorithm here is identical to that of Algorithm 3, 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.

### 6.2.3 REGRET GUARANTEE

For brevity, we only state the Corollary of our main Theorem (Theorem 6) which is stated in Appendix C.

**Corollary 2.** *Suppose Algorithm 4 is run with input parameters  $\delta \in (0, 1)$ , and  $T_0 = \tilde{O}(d^2 \ln^2(\frac{1}{\delta}))$  given in Equation (16), then with probability at-least  $1 - \delta$ , the regret after  $T$  times satisfies*

$$R_T \leq O\left(\frac{d^2}{\gamma^{5.18}} \ln^2(d/\delta) \tau^2 \ln\left(\frac{TK}{\delta}\right)\right) + \tilde{O}(\sqrt{Td_m^*}),$$

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

**Discussion -** Our regret scaling matches that of an oracle that knows the true problem complexity and thus obtains a regret of  $\tilde{O}(\sqrt{d_m^* T})$ . 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  $\gamma$  and is thus not uniform over all  $\theta^*$ , unlike Foster et al. (2019) that is uniform over  $\theta^*$ . 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  $\theta^*$ .

## 7. Comparison Study: ACB vs. ALB-DIM for Finite Armed Stochastic Linear Bandits

In this section, we study the model selection algorithm for generic contextual bandits, ACB (see Algorithm 1) in the special setting of stochastic linear bandits with finite number of arms. We see that order-wise, the generic Algorithm, ACB recovers the regret guarantees of the linear bandit setup (with finite number of arms).

Recall the problem setup of Section 6.2. Additionally, for simplicity, we also assume that the context embeddings  $\phi^j(x, a)$  is a  $d_j$  dimensional standard Gaussian random variable. Note that this is stronger than the sub-Gaussian assumption of Section 6.2.

Rewriting the nested hypothesis class, this corresponds to setting the function classes  $\mathcal{F}_j$  to be the linear class as the following:

$$\mathcal{F}_j = \{(x, a) \mapsto \langle \theta_j, \phi^j(x, a) \rangle \mid \theta_j \in \mathbb{R}^{d_j}, \|\theta\| \leq 1\},$$

where  $d_1 \leq d_2 \leq \dots \leq d_M = d$ . Also, let  $m^*$  is the smallest index such that the optimal regressor is realized, i.e.,

$$f_{m^*}^*(x, a) = \langle \theta^*, \phi^{m^*}(x, a) \rangle,$$

and hence this can be cast-ed as a model selection problem in our framework. Let us look at the separability condition (Assumption 2). In order to do that, we take  $j = m^* - 1$ , we first compute

$$\bar{f}_j = \operatorname{arginf}_{f \in \mathcal{F}_j} \mathbb{E}_{x,a} [f(x, a) - f_{m^*}^*(x, a)]^2,$$

and then compute the quantity

$$\mathbb{E}_{x,a} [\bar{f}_j(x, a) - f_{m^*}^*(x, a)]^2.$$

Substituting  $f_{m^*}^*(x, a) = \langle \theta^*, \phi^{m^*}(x, a) \rangle$  and optimizing over  $\theta_j$ , we obtain

$$\mathbb{E}_{x,a} [\bar{f}_j(x, a) - f_{m^*}^*(x, a)]^2 \geq \gamma^2,$$

where  $\gamma$  is the minimum magnitude of the non-zero coordinate of  $\theta^*$ , i.e.,  $\gamma = \min\{|\theta_i^*| : \theta_i^* \neq 0\}$ .

Note that in the above calculation (i) the minimizer  $\theta_j$  corresponding to  $\bar{f}_j$ , is precisely a  $m^* - 1$  dimensional vector with entries equal to the first  $m^* - 1$  coordinates of  $\theta^*$ ; (ii) using the fact that  $\phi(\cdot, \cdot)$  is distributed as a standard Gaussian and has nested structure, we obtain the lower bound.

Hence for stochastic linear bandits, one may take  $\Delta = \gamma^2$ . Suppose we run ACB (Algorithm 1) in this setup for  $T$  iterations. Observe that the linear function class has infinite cardinality, but the class of linear functions of dimension  $d_i$  has a VC-dimension is  $d_i + 1$  for all  $i$ . Hence, substituting in the regret expression of Theorem 3, with  $\xi_{\mathcal{F}_{d,1/2^{m/2}}}(T) = \tilde{\mathcal{O}}(d/T)$  (linear class of  $d$  dimensional functions, VC-dimension is  $d + 1$ ), we obtain

$$R(T) \leq \tilde{\mathcal{O}} \left[ \left( \frac{d}{\gamma^4} \right) + \sqrt{KTd_{m^*}} \right].$$

with high probability. Here, we ignore the log factors.

On the other hand, if we use Theorem 6, we obtain a regret of

$$R(T) \leq \tilde{\mathcal{O}} \left[ \left( \frac{d^2}{\gamma^{5.18}} \right) + \sqrt{Td_{m^*}} \right],$$

with high probability. Several remarks are in order:

**Remark 12.** *If the number of actions  $K = \mathcal{O}(1)$ , both the regret scaling are order-wise same. Hence, in this setting, ACB recovers the performance of ALB-DIM, as we claim in the introduction.*

**Remark 13.** *Note that, the performance of ACB Algorithm 1 specialized to the linear setting is worse than ALB-DIM Theorem 6 when  $K$  is large. In particular, there is no  $K$  dependence in Theorem 6, but we have a  $\sqrt{K}$  term in the leading factor here. Note that ACB (Algorithm 1) is applicable for any generic contextual bandit problem, whereas Algorithm 4 is specialized to the linear case only. The price of  $\sqrt{K}$  in regret can be viewed as the cost of generalization.*

**Remark 14.** Let us now focus on the additive (minor) term with no  $T$  dependence. It scales as  $1/\gamma^4$  for *ACB*, whereas for *ALB-DIM* it scales as  $1/\gamma^{5.18}$ . Note that, we remarked (after Theorem 4) that via carefully choosing the problem constants, the dependence in Theorem 6 can be made arbitrarily close to  $1/\gamma^4$ . In that setting, we have the same dependence on  $\gamma$  in both the cases.

**Remark 15.** Finally, note that the additive term is linearly dependent ( $d$ ) in *ACB*, whereas it has a quadratic dependence ( $d^2$ ) in *ALB-DIM*. We believe this stems from the analysis of *ALB-DIM*. In Algorithm 4, we successively estimate the support of the underlying true parameter  $\theta^*$ , and it is not clear whether support recovery is indeed required to ensure low regret.

## 8. Numerical Experiments

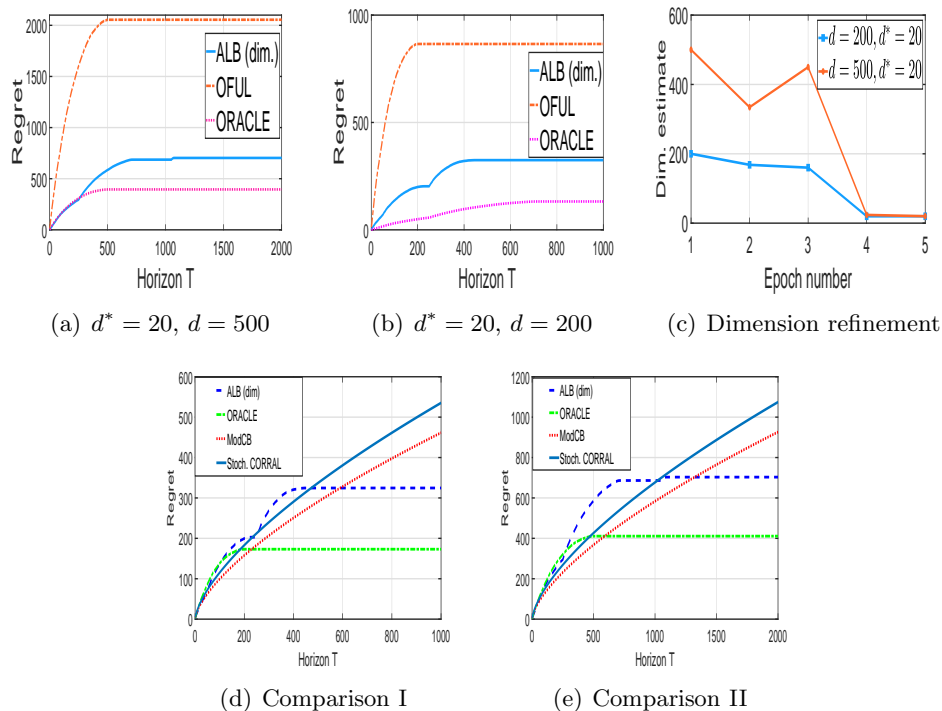


Figure 1: Synthetic experiments, validating the effectiveness of Algorithm 3 and comparisons with several baselines. All the results are averaged over 25 trials.

In this section we will verify the theoretical findings. We concentrate on the linear contextual bandit setup. We compare *ALB-Dim* with the (non-adaptive) *OFUL* algorithm of Abbasi-Yadkori et al. (2011) and an *oracle* that knows the problem complexity a priori. The oracle just runs *OFUL* with the known problem complexity. At each round of the learning algorithm, we sample the context vectors from a  $d$ -dimensional standard Gaussian,  $\mathcal{N}(0, I_d)$ . The additive noise to be zero-mean Gaussian random variable with variance 0.5.

In panel (a)-(c), we compare the performance of **ALB-Dim** with OFUL (Abbasi-Yadkori et al. (2011)) and an *oracle* who knows the true support of  $\theta^*$  apriori. For computational ease, we set  $\varepsilon_i = 2^{-i}$  in simulations. We select  $\theta^*$  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 (c), we plot the successive dimension refinement over epochs. We observe that within 4 – 5 epochs, **ALB-Dim** finds the sparsity of  $\theta^*$ .

**Comparison of ALB (dim):** When  $\theta^*$  is sparse, we compare ALB-Dim with 3 baselines: (i) the ModCB algorithm of Foster et al. (2019) (ii) the Stochastic Corral algorithm of Pacchiano et al. (2020b) and (iii) an oracle which knows the support of  $\theta^*$ . We select  $\theta^*$  to be  $d^* = 20$  sparse, with dimension  $d = 200$  and  $d = 500$ . The smallest non-zero component of  $\theta^*$  is 0.12. For ModCB, we use ILOVETOCORBANDITS algorithm, similar to Agarwal et al. (2014b). We select the cardinality of action set as 2 and select the sub-Gaussian parameter of the embedding as unity. In Figures 1(d) and 1(e), we observe that, the regret of ALB (dim) is better than ModCB and Stochastic Corral. The theoretical regret bound for ModCB scales as  $\mathcal{O}(T^{2/3})$  (which is much larger than the ALB-Dim algorithm we propose), and Figure 1(c), validates this. The Stochastic Corral algorithm treats the base algorithms as bandit arms (with bandit feedback), as opposed to ALB-Dim which, at each arm-pull, updates the information about all the base algorithms. Thus, (Figs 1(d), 1(e)), ALB-Dim has a superior performance compared to Stochastic Corral.

## Appendix

### Appendix A. Model Selection for Contextual Bandits

#### A.1 Proof of Lemma 1

Let us first show that  $S_j^m$  concentrates around its expectation. We show it via a simple application of the Hoeffdings inequality.

Fix a particular  $m$  and  $j \in [M]$ . Note that  $\hat{f}_j^m$  is computed based on  $2^{m-2}$  samples. Also, in the testing phase, we use a fresh set of  $2^{m-2}$  samples, and so  $\hat{f}_j^m$  is independent of the second set of samples, used in constructing  $S_j^m$ . Note that since we have  $r(\cdot) \in [0, 1]$ , we may restrict the offline regression oracle to search over functions having range  $[0, 1]$ . This implies that, we have  $\hat{f}_j^m(\cdot) \in [0, 1]$ . Note that this restricted search assumption is justified since our goal is obtain an estimate of the reward function via regression function, and this assumption also features in Simchi-Levi and Xu (2020). So the random variable  $(\hat{f}_j^m(x_t, a_t) - r_t(a_t))^2$  is upper-bounded by 4, and hence sub-Gaussian with a constant parameter.

Note that we are using only the samples from the previous epoch. Note that in **ACB**, the regression estimate actually remains fixed over an entire epoch. Hence, conditioning on the filtration consisting of (context, action, reward) triplet upto the end of the  $m - 2$ -th epoch, the random variables  $\{(\hat{f}_j^m(x_t, a_t) - r_t(a_t))^2\}_{t=\tau_{m-1}/2+1}^{\tau_m-1}$  (a total of  $2^{m-2}$  samples) are independent. Note that similar argument is given in (Simchi-Levi and Xu, 2020, Section 3.1) (the **FALCON+** algorithm) to argue the independence of the (context, action, reward) triplet, accumulated over just the previous epoch.

Hence using Hoeffdings inequality for sub-Gaussian random variables, we have

$$\mathbb{P}(|S_j - \mathbb{E}S_j| \geq \ell) \leq 2 \exp(-n\ell^2/32).$$

Recall Assumption 1. Note that, the conditional variance of  $r_t(\cdot)$  is finite, i.e., given  $x \in \mathcal{X}$ ,  $\mathbb{E}[r_t(a) - f_{d^*}^*(x, a)]^2 \leq 1$ , for all  $a \in \mathcal{A}$ . Let us define<sup>6</sup>  $\mathbb{E}[r_t(a) - f_{d^*}^*(x, a)]^2 = \sigma^2$ . With this new notation, let us first look at the expression  $\mathbb{E}S_j$ .

**Realizable classes:** Fix  $m$  and consider  $j \in [M]$  such that  $j \geq d^*$ . So, for this realizable setting, we obtain the excess risk as:

$$\begin{aligned} & \mathbb{E}_{x,r,a}[\hat{f}_j^m(x, a) - r(a)]^2 - \inf_{f \in \mathcal{F}_j} \mathbb{E}_{x,r,a}[f(x, a) - r(a)]^2 \\ &= \mathbb{E}_{x,r,a}[\hat{f}_j^m(x, a) - r(a)]^2 - \mathbb{E}_{x,r,a}[f_{d^*}^*(x, a) - r(a)]^2 \\ &= \mathbb{E}_{x,a}[\hat{f}_j^m(x, a) - f_{d^*}^*(x, a)]^2. \end{aligned}$$

So, we have, for the realizable function class,

$$\begin{aligned} \mathbb{E}S_j^m &= \frac{1}{2^{m-2}} \mathbb{E}_{x_t, r_t, a_t} \sum_{t=1}^{2^{m-2}} [\hat{f}_j^m(x_t, a_t) - r_t(a_t)]^2 \\ &= \frac{1}{2^{m-2}} \sum_{t=1}^{2^{m-2}} \mathbb{E}_{x_t, r_t, a_t} [f_{d^*}^*(x_t, a_t) - r_t(a_t)]^2 + \frac{1}{2^{m-2}} \sum_{t=1}^{2^{m-2}} \mathbb{E}_{x_t, a_t} [\hat{f}_j^m(x, a) - f_{d^*}^*(x, a)]^2 \\ &\leq \sigma^2 + C_1 \log(2^{m/2} |\mathcal{F}_j|) / (2^{m-2}), \end{aligned}$$

Here, the first term comes from the second moment bound of  $\sigma^2$ , and the second term comes by setting the high probability slack as  $2^{-m/2}$  into (Agarwal et al., 2012, Lemma 4.1). So, by applying Hoeffding's inequality, we finally have (using the bound  $\mathbb{E}S_j^m \geq \sigma^2$ ):

$$\begin{aligned} \sigma^2 - C_3 \frac{\sqrt{\log(1/\delta)}}{2^{m/2}} - C_4 \frac{\sqrt{m}}{2^{m/2}} \leq S_j^m \leq \sigma^2 + \\ C_1 \frac{\log(|\mathcal{F}_j|)}{2^m} + C_2 \frac{m}{2^m} + C_3 \frac{\sqrt{\log(1/\delta)}}{2^{m/2}} + C_4 \frac{\sqrt{m}}{2^{m/2}} \end{aligned}$$

with probability at least  $1 - \delta/2^m$ . Since we have doubling epoch, we have

$$\sum_{m=1}^N 2^m \leq T,$$

where  $N$  is the number of epochs and  $T$  is the time horizon. From above, we obtain  $N = \mathcal{O}(\log_2 T)$ . Using the bound,  $m \leq N$ , note that, provided

$$2^m \gtrsim \max\{\log T, \log(|\mathcal{F}_M|), \log(1/\delta)\}, \quad (2)$$

we have for some absolute global constant  $c_0$ , for any  $j \geq d^*$ ,

$$\sigma^2 - \frac{c_0}{2^{m/2}} \leq S_j^m \leq \sigma^2 + \frac{c_0}{2^{m/2}} \quad (3)$$

with probability at least  $1 - \delta/2^m$ .

---

6. We use the notation  $\sigma^2$  throughout the rest of the paper.

**Non-Realizable classes:** For the non realizable classes, we have the following calculation. For any  $f \in \mathcal{F}_j$ , where  $j < d^*$ , we have

$$\begin{aligned}
 & \mathbb{E}_{x,r,a}[f(x,a) - r(a)]^2 - \mathbb{E}_{x,r,a}[r(a) - f_{d^*}^*(x,a)]^2 \\
 &= \mathbb{E}_{x,a,r}[(f(x,a) - f_{d^*}^*(x,a))(f(x,a) + f_{d^*}^*(x,a) - 2r(a))] \\
 &= \mathbb{E}_{x,a}\mathbb{E}_{r|x}[(f(x,a) - f_{d^*}^*(x,a))(f(x,a) + f_{d^*}^*(x,a) - 2r(a))] \\
 &= \mathbb{E}_{x,a}[(f(x,a) - f_{d^*}^*(x,a))(f(x,a) + f_{d^*}^*(x,a) - 2\mathbb{E}_{r|x}r(a))] \\
 &= \mathbb{E}_{x,a}[f(x,a) - f_{d^*}^*(x,a)]^2,
 \end{aligned}$$

where the third inequality follows from the fact that given context  $x$ , the distribution of  $r$  is independent of  $a$  (see (Agarwal et al., 2012, Lemma 4.1)).

So, we have

$$\begin{aligned}
 \mathbb{E}_{x,r,a}[f(x,a) - r(a)]^2 &\geq \mathbb{E}_{x,r,a}[r(a) - f_{d^*}^*(x,a)]^2 + \mathbb{E}_{x,a}[f(x,a) - f_{d^*}^*(x,a)]^2 \\
 &\geq \Delta + \sigma^2,
 \end{aligned}$$

where the last inequality comes from the separability assumption along with the assumption on the second moment. Since the regressor  $\hat{f}_j^m \in \mathcal{F}_j$ , we have

$$\begin{aligned}
 \mathbb{E}_{x,r,a}[\hat{f}_j^m(x,a) - r(a)]^2 &\geq \mathbb{E}_{x,r,a}[r(a) - f_{d^*}^*(x,a)]^2 + \mathbb{E}_{x,a}[f(x,a) - f_{d^*}^*(x,a)]^2 \\
 &\geq \Delta + \sigma^2.
 \end{aligned}$$

Now, using  $2^{m-2}$  samples, we obtain from Hoeffding's inequality that

$$S_j^m \geq \Delta + \sigma^2 - C_5 \frac{\sqrt{\log(1/\delta)}}{2^{m/2}} - C_6 \frac{\sqrt{m}}{2^{m/2}}$$

with probability at least  $1 - \delta/2^m$ . In particular, since  $2^m \gtrsim \max\{\log T, \log(|\mathcal{F}_M|), \log(1/\delta)\}$ , there is a global constant  $c_1$  such that, for any  $j < d^*$ ,

$$S_j^m \geq \Delta + \sigma^2 - \frac{c_1}{2^{m/2}}, \tag{4}$$

holds with probability at least  $1 - \delta/2^m$ .

In every phase  $m$ , denote by the threshold  $\gamma_m := S_M^m + \frac{\sqrt{m}}{2^{m/2}}$ , i.e., the Model Selection parameter in Line 8 of Algorithm 1. Now, let  $m_0$  be the smallest value of  $m$  satisfying  $2^m \gtrsim \max\{\frac{\log T}{\Delta^2}, \log(|\mathcal{F}_M|), \log(1/\delta)\}$ . We have from Equations (3) and (4) and a union bound over the  $M$  classes that, with probability at-least  $1 - \sum_{m \geq 1} 2M\delta 2^{-m}$ , for all phases  $m \geq m_0$ ,

$$\begin{aligned}
 S_j^m &\geq \sigma^2 + \Delta - \frac{c_1}{2^{m/2}}, \text{ for all } 1 \leq j < d^*, \\
 \sigma^2 - \frac{c_0}{2^{m/2}} &\leq S_j^m \leq \sigma^2 + \frac{c_0}{2^{m/2}}, \text{ for all } j \geq d^*.
 \end{aligned}$$

The preceding display, along with the fact that the threshold  $\gamma_m = S_m^M + \frac{\sqrt{m}}{2^{m/2}}$ , gives that, with probability at-least  $1 - 2M\delta$  and all phases  $m \geq m_0$ ,

$$\begin{aligned}
 S_{d^*}^m &\leq \sigma^2 + \frac{c_0}{2^{m/2}} \leq \sigma^2 - \frac{c_0}{2^{m/2}} + \frac{\sqrt{m}}{2^{m/2}} \leq \gamma_m \leq \sigma^2 + \frac{c_0}{2^{m/2}} + \frac{\sqrt{m}}{2^{m/2}}, \\
 &\leq \sigma^2 + \Delta - \frac{c_1}{2^{m/2}}.
 \end{aligned}$$

The second inequality follows since  $2^m \gtrsim \frac{\log T}{\Delta^2}$ , by definition of  $m_0$ . The above equations guarantee that, with probability at-least  $1 - 2M\delta$ , in all phases  $m \geq m_0$ , the model selection procedure in Line 8 of Algorithm 1, identifies the correct class  $d^*$ .

## A.2 Proof of Theorem 1

The above calculation shows that as soon as

$$2^m \gtrsim \log T \max\{\log(|\mathcal{F}_M|), \log(1/\delta), \Delta^{-2}\},$$

the model selection procedure will succeed with high probability. Until the above condition is satisfied, we do not have any handle on the regret and hence the regret in that phase will be linear. This corresponds the first term in the regret expression. Suppose  $m^*$  be the epoch index where the conditions of Lemma 1 hold. Then, for  $m > m^*$ , the regret is given by (see Simchi-Levi and Xu (2020)):

$$\sum_{m=m^*}^N \mathcal{O}\left(\sqrt{K(\tau_m - \tau_{m-1}) \log(|\mathcal{F}_{d^*}|(\tau_m - \tau_{m-1})/\delta)}\right) \leq \mathcal{O}\left(\sqrt{KT \log(|\mathcal{F}_{d^*}|T/\delta)}\right),$$

with probability exceeding  $1 - \delta$ , where  $N$  is the number of epochs. Lemma 1 gives that the total number of rounds till the beginning of phase  $m^*$  is upper bounded by  $\mathcal{O}(\log T \max\{\log(|\mathcal{F}_M|), \log(1/\delta), \Delta^{-2}\})$ , where  $\mathcal{O}$  hides global absolute constants. So, the total regret is given by

$$R(T) \leq \mathcal{O}(\log T \max\{\log(|\mathcal{F}_M|), \log(1/\delta), \Delta^{-2}\}) + \sum_{m=m^*}^N \mathcal{O}\left(\sqrt{K(\tau_m - \tau_{m-1}) \log(|\mathcal{F}_{d^*}|(\tau_m - \tau_{m-1})/\delta)}\right),$$

with probability at-least  $1 - \delta - 2M\delta$ . Simplifying the summation, we get

$$\begin{aligned} & \sum_{m=m^*}^N \mathcal{O}\left(\sqrt{K(\tau_m - \tau_{m-1}) \log(|\mathcal{F}_{d^*}|(\tau_m - \tau_{m-1})/\delta)}\right) \\ & \leq \sum_{m=1}^N \mathcal{O}\left(\sqrt{K(\tau_m - \tau_{m-1}) \log(|\mathcal{F}_{d^*}|(\tau_m - \tau_{m-1})/\delta)}\right) \\ & \leq \mathcal{O}(\sqrt{K \log(|\mathcal{F}_{d^*}|(T)/\delta)}) \sum_{m=1}^N \sqrt{\tau_m - \tau_{m-1}}. \end{aligned}$$

Note that, with  $\tau_m = 2^m$ , the epoch length  $\tau_m - \tau_{m-1}$  doubles with  $m$ . Let the length of the  $N$ -th epoch is  $T_N$ . We have

$$\begin{aligned} \sum_{i=1}^N \sqrt{\tau_m - \tau_{m-1}} &= \sqrt{T_N} \left(1 + \frac{1}{\sqrt{2}} + \frac{1}{2} + \dots N\text{-th term}\right) \\ &\leq \sqrt{T_N} \left(1 + \frac{1}{\sqrt{2}} + \frac{1}{2} + \dots\right) = \frac{\sqrt{2}}{\sqrt{2}-1} \sqrt{T_N} \leq \frac{\sqrt{2}}{\sqrt{2}-1} \sqrt{T}, \end{aligned}$$

and this completes the proof of the theorem.



### A.3 Proof of Lemma 2

Since, we have samples from pure exploration, let us first show that  $S_j$  concentrates around its expectation. We show it via a simple application of the Hoeffdings inequality.

Fix a particular  $j \in [M]$ . Note that  $\hat{f}_j$  is computed based on the first set of  $\lceil \sqrt{T} \rceil$  samples. Also, in the testing phase, we again sample  $\lceil \sqrt{T} \rceil$  samples, and so  $\hat{f}$  is independent of the second set of  $\lceil \sqrt{T} \rceil$  samples, used in constructing  $S_j$ . Note that we have  $r(\cdot) \in [0, 1]$ . Furthermore, as explained in the proof of Lemma 1, it is sufficient to have  $\hat{f}(\cdot) \in [0, 1]$ . So the random variable  $(\hat{f}_j(x_t, a_t) - r_t(a_t))^2$  is upper-bounded by 4, and hence sub-Gaussian with a constant parameter. Also, note that since we are choosing an action independent of the context, the random variables  $\{(\hat{f}_j(x_t, a_t) - r_t(a_t))^2\}_{t=1}^{\lceil \sqrt{T} \rceil}$  are independent. Hence using Hoeffdings inequality for sub-Gaussian random variables, we have

$$\mathbb{P}(|S_j - \mathbb{E}S_j| \geq \ell) \leq 2 \exp(-n\ell^2/32).$$

Re-writing the above, we obtain

$$|S_j - \mathbb{E}S_j| \leq C \sqrt{\frac{\log(1/\delta)}{\sqrt{T}}} \quad (5)$$

with probability at least  $1 - 2\delta$ . Let us look at the expression  $\mathbb{E}S_j$ .

$$\mathbb{E}S_j = \mathbb{E} \left( \frac{1}{\lceil \sqrt{T} \rceil} \sum_{t=1}^{\lceil \sqrt{T} \rceil} (\hat{f}_j(x_t, a_t) - r_t(a_t))^2 \right).$$

**Case I: Realizable Class** First consider the case that  $j \geq d^*$ , meaning that  $f_{d^*}^* \in \mathcal{F}_j$ . So, for this realizable setting, we obtain the excess risk as (using Agarwal et al. (2012))

$$\begin{aligned} & \mathbb{E}_{x,r,a}[\hat{f}_j(x, a) - r(a)]^2 - \inf_{f \in \mathcal{F}_j} \mathbb{E}_{x,r,a}[f(x, a) - r(a)]^2 \\ &= \mathbb{E}_{x,r,a}[\hat{f}_j(x, a) - r(a)]^2 - \mathbb{E}_{x,r,a}[f_{d^*}^*(x, a) - r(a)]^2 \\ &= \mathbb{E}_{x,a}[\hat{f}_j(x, a) - f_{d^*}^*(x, a)]^2. \end{aligned}$$

So, we have, for the realizable function class,

$$\begin{aligned} \mathbb{E}S_j &= \frac{1}{\lceil \sqrt{T} \rceil} \mathbb{E}_{x_t, r_t, a_t} \sum_{t=1}^{\lceil \sqrt{T} \rceil} [\hat{f}_j(x_t, a_t) - r_t(a_t)]^2 \\ &= \frac{1}{\lceil \sqrt{T} \rceil} \sum_{t=1}^{\lceil \sqrt{T} \rceil} \mathbb{E}_{x_t, r_t, a_t} [f_{d^*}^*(x_t, a_t) - r_t(a_t)]^2 + \frac{1}{\lceil \sqrt{T} \rceil} \sum_{t=1}^{\lceil \sqrt{T} \rceil} \mathbb{E}_{x_t, a_t} [\hat{f}_j(x, a) - f_{d^*}^*(x, a)]^2 \\ &\leq \sigma^2 + C_1 \frac{\log(\sqrt{T} |\mathcal{F}_j|)}{\sqrt{T}} \end{aligned}$$

where  $C_1$  is an absolute constant. The second term is obtained by setting the high probability slack, as  $2^{-m/2}$  into (Agarwal et al., 2012, Lemma 4.1). So, we finally have from the

preceding display and Equation (5) that

$$\sigma^2 - C_2 \sqrt{\frac{\log(1/\delta)}{\sqrt{T}}} \leq S_j \leq \sigma^2 + C_2 \frac{\log(\sqrt{T}|\mathcal{F}_j|)}{\sqrt{T}} + C_3 \sqrt{\frac{\log(1/\delta)}{\sqrt{T}}} \quad (6)$$

with probability at least  $1 - 2\delta$ .

**Case II: Non-realizable class** We now consider the case when  $j < d^*$ , meaning that  $f_{d^*}^*$  does not lie in  $\mathcal{F}_j$ . Similar to the proof of Lemma 1, we have

$$\begin{aligned} \mathbb{E}_{x,r,a}[f(x,a) - r(a)]^2 &\geq \mathbb{E}_{x,r,a}[r(a) - f_{d^*}^*(x,a)]^2 + \mathbb{E}_{x,a}[f(x,a) - f_{d^*}^*(x,a)]^2 \\ &\geq \Delta + \sigma^2, \end{aligned}$$

where the last inequality comes from the separability assumption along with the assumption on the second moment. Since the regressor  $\hat{f}_j \in \mathcal{F}_j$ , we have

$$\begin{aligned} \mathbb{E}_{x,r,a}[\hat{f}_j(x,a) - r(a)]^2 &\geq \mathbb{E}_{x,r,a}[r(a) - f_{d^*}^*(x,a)]^2 + \mathbb{E}_{x,a}[f(x,a) - f_{d^*}^*(x,a)]^2 \\ &\geq \Delta + \sigma^2, \end{aligned}$$

and hence

$$\mathbb{E}S_j \geq \Delta + \sigma^2$$

So, in this setting, with probability  $1 - 2\delta$ ,

$$S_j \geq \mathbb{E}S_j - C_4 \sqrt{\frac{\log(1/\delta)}{\sqrt{T}}} \quad (7)$$

$$\geq \Delta + \sigma^2 - C_4 \sqrt{\frac{\log(1/\delta)}{\sqrt{T}}}. \quad (8)$$

where  $C$  is an absolute global constant. Thus, from Equations (6) and (8) and an union bound over the  $M$  classes, we have with probability at-least  $1 - 4M\delta$ ,

$$\begin{aligned} S_j &\geq \sigma^2 - C_2 \frac{\log(\sqrt{T}|\mathcal{F}_j|)}{\sqrt{T}} - C_3 \sqrt{\frac{\log(1/\delta)}{\sqrt{T}}}, \text{ for all } j \geq d^*, \\ S_j &\leq \sigma^2 + C_2 \frac{\log(\sqrt{T}|\mathcal{F}_j|)}{\sqrt{T}} + C_3 \sqrt{\frac{\log(1/\delta)}{\sqrt{T}}}, \text{ for all } j \geq d^*, \\ S_j &\geq \Delta + \sigma^2 - C_4 \sqrt{\frac{\log(1/\delta)}{\sqrt{T}}}, \text{ for all } j < d^*. \end{aligned} \quad (9)$$

**Choice of Threshold** Notice from Line 11 of Algorithm 2, that the threshold for model selection is  $\gamma := S_M + \sqrt{\frac{\log(T)}{\sqrt{T}}}$ . Thus, if the event in Equations (9) holds, then the model selection stage will succeed in identifying the correct model class if the threshold  $\gamma$  satisfies

$$\begin{aligned} \gamma &< \Delta + \sigma^2 - C_4 \sqrt{\frac{\log(1/\delta)}{\sqrt{T}}}, \\ \gamma &> \sigma^2 + C_2 \frac{\log(\sqrt{T}|\mathcal{F}_j|)}{\sqrt{T}} + C_3 \sqrt{\frac{\log(1/\delta)}{\sqrt{T}}} \end{aligned} \quad (10)$$

The first item ensures that no-non realizable class will be selected as the true model, and the second item ensures that the smallest realizable class will be selected as the true model. Thus, if the time horizon  $T$  satisfies

$$\begin{aligned} \sqrt{\frac{\log(T)}{\sqrt{T}}} &\geq 2 \left( C_2 \frac{\log(\sqrt{T}|\mathcal{F}_j|)}{\sqrt{T}} + C_3 \sqrt{\frac{\log(1/\delta)}{\sqrt{T}}} \right), \\ \sqrt{\frac{\log(T)}{\sqrt{T}}} + C_2 \frac{\log(\sqrt{T}|\mathcal{F}_j|)}{\sqrt{T}} + C_3 \sqrt{\frac{\log(1/\delta)}{\sqrt{T}}} &\leq \Delta - C_4 \sqrt{\frac{\log(1/\delta)}{\sqrt{T}}}, \end{aligned} \quad (11)$$

then the threshold  $\gamma$  satisfies the conditions in Equations (10). It is easy to verify that for  $T \gtrsim (\log T) \max \left( \log \left( \sqrt{T}|\mathcal{F}_M| \right), \Delta^{-4}, \log(1/\delta) \right)$ , then the conditions in Equations (11) holds. Thus, Equations (9), (10) and (11) yield that, if

$$T \gtrsim (\log T) \max \left( \log \left( \sqrt{T}|\mathcal{F}_M| \right), \Delta^{-4}, \log(1/\delta) \right),$$

with probability at-least  $1 - 4M\delta$ , the model selection test in Line 11 of Algorithm 2 correctly identifies the smallest model class containing the true model.

#### A.4 Proof of Theorem 2

The regret  $R(T)$  can be decomposed in 2 stages, namely exploration and exploitation.

$$R(T) = R_{\text{explore}} + R_{\text{exploit}}$$

Since we spend  $2\lceil\sqrt{T}\rceil$  time steps in exploration, and  $r_t(\cdot) \in [0, 1]$ , the regret incurred in this stage

$$R_{\text{explore}} \leq C_1 \sqrt{T}.$$

Now, at the end of the explore stage, provided Assumptions 2 and 3, we know, with probability at least  $1 - 4M\delta$ , we obtain the true function class  $\mathcal{F}_{d^*}$ . The threshold is set in such a way that we obtain the above result. Now, we would just commit to the function class and use the contextual bandit algorithm, namely FALCON. From Simchi-Levi and Xu (2020), the regret guarantee of FALCON is

$$\begin{aligned} R_{\text{exploit}} &\leq \mathcal{O} \left( \sqrt{K(T - 2\lceil\sqrt{T}\rceil) \log(|\mathcal{F}_{d^*}|(T - 2\lceil\sqrt{T}\rceil)/\delta)} \right) \\ &\leq \mathcal{O} \left( \sqrt{KT \log(|\mathcal{F}_{d^*}|T/\delta)} \right), \end{aligned}$$

with probability exceeding  $1 - \delta$ . Combining the above expressions yield the result.

### A.5 Proof of Theorem 3

The proof follows by combining the proof of Theorem 1 and 4.

For the realizable classes, we have (from Assumption 4 and converting the conditional expectation to unconditional one with probability slack as  $1/2^{m/2}$ , similar to the proof of Lemma 1),

$$\mathbb{E}S_j^m \leq \sigma^2 + \xi_{\mathcal{F}_j, 1/2^{m/2}}(2^{m-2}) + 2\left(\frac{1}{2^{m/2}}\right),$$

and as a result

$$S_j^m \leq \sigma^2 + \xi_{\mathcal{F}_j, 1/2^{m/2}}(2^{m-2}) + C_1 \frac{\sqrt{\log(1/\delta)}}{2^{m/2}} + C_2 \frac{\sqrt{m}}{2^{m/2}}$$

with probability at least  $1 - 2\delta/2^m$ .

Similarly, for non-realizable classes we obtain

$$S_j^m \geq \Delta + \sigma^2 - C_3 \frac{\sqrt{\log(1/\delta)}}{2^{m/2}} - C_4 \frac{\sqrt{m}}{2^{m/2}}$$

with probability at least  $1 - \delta/2^m$ .

Now, suppose we choose the threshold  $\gamma = S_M^m + \frac{\sqrt{m}}{2^{m/2}}$ . Finally, we say that provided

$$2^m \gtrsim (\log T) \max\left\{\max_m 2^{m/2} \xi_{\mathcal{F}_M, 1/2^{m/2}}(2^{m-2}), \log(1/\delta), \Delta^{-2}\right\},$$

the model selection procedure succeeds with probability exceeding

$$1 - \sum_{m=1}^{\infty} 2M\delta/2^m \geq 1 - 2M\delta.$$

The rest of the proof follows similarly to Theorem 1, and we omit the details here.

### A.6 Proof of Theorem 4

**Case I: Realizable Class** Consider  $j \geq d^*$ . Using calculations similar to the finite cardinality setting, we obtain

$$\mathbb{E}S_j \leq \sigma^2 + \xi_{\mathcal{F}_j, (1/T^{1/4})}(\sqrt{T}) + 2(1/T^{1/4}),$$

where we use the definition of  $\xi(\cdot)$ , as given in Assumption 4. Hence, invoking Hoeffding's inequality, we obtain

$$\begin{aligned} S_j &\leq \sigma^2 + \xi_{\mathcal{F}_j, (1/T^{1/4})}(\sqrt{T}) + 2(1/T^{1/4}) + C_1 T^{-1/4} \sqrt{\log(1/\delta)} \\ &\leq \sigma^2 + \xi_{\mathcal{F}_j, (1/T^{1/4})}(\sqrt{T}) + C_1 T^{-1/4} \sqrt{\log(1/\delta)} \end{aligned}$$

with probability at least  $1 - 2\delta$ . We also have (from 2-sided Hoeffding's)

$$S_j \geq \sigma^2 - C_2 T^{-1/4} \sqrt{\log(1/\delta)}$$

**Case II: Non-realizable Class** We now consider the setting where  $j < d^*$ , meaning that  $f_{d^*}^*$  does not lie in  $\mathcal{F}_j$ . In this case, similar to above, we have

$$\mathbb{E}S_j \geq \Delta + \sigma^2,$$

and hence

$$\begin{aligned} S_j &\geq \mathbb{E}S_j - \sqrt{\frac{32 \log(1/\delta)}{\sqrt{T}}} \\ &\geq \Delta + \sigma^2 - \sqrt{\frac{32 \log(1/\delta)}{\sqrt{T}}}. \end{aligned}$$

Now, with the threshold,  $\gamma = S_M + \sqrt{\frac{\log T}{\sqrt{T}}}$ , provided

$$T \gtrsim (\log T) \max \left( \log \left( T^{1/4} \xi_{\mathcal{F}_M, (1/T^{1/4})} \right), \Delta^{-4}, \log(1/\delta) \right),$$

the model selection procedure succeeds with probability at least  $1 - 2M\delta$ , where we do a calculation similar to the proof of Lemma 2.

After obtaining the correct model class, the regret expression comes directly from Simchi-Levi and Xu (2020) in the infinite function class setting.

## Appendix B. Model Selection for Linear Stochastic bandits

### B.1 Proof of Theorem 5

We shall need the following lemma from Krikheli and Leshem (2018), on the behaviour of linear regression estimates.

**Lemma 3.** *If  $M \geq d$  and satisfies  $M = O\left(\left(\frac{1}{\varepsilon^2} + d\right) \ln\left(\frac{1}{\delta}\right)\right)$ , and  $\hat{\theta}^{(M)}$  is the least-squares estimate of  $\theta^*$ , using the  $M$  random samples for feature, where each feature is chosen uniformly and independently on the unit sphere in  $d$  dimensions, then with probability 1,  $\hat{\theta}^{(M)}$  is well defined (the least squares regression has a unique solution). Furthermore,*

$$\mathbb{P}[\|\hat{\theta}^{(M)} - \theta^*\|_\infty \geq \varepsilon] \leq \delta.$$

We shall now apply the theorem as follows. Denote by  $\hat{\theta}_i$  to be the estimate of  $\theta^*$  at the beginning of any phase  $i$ , using all the samples from random explorations in all phases less than or equal to  $i - 1$ .

**Remark 16.** *The choice  $T_0 := O\left(d^2 \ln^2\left(\frac{1}{\delta}\right)\right)$  in Equation (1) is chosen such that from Lemma 4, we have that*

$$\mathbb{P} \left[ \|\hat{\theta}^{(\lceil \sqrt{T_0} \rceil)} - \theta^*\|_\infty \geq \frac{1}{2} \right] \leq \delta$$

**Lemma 4.** *Suppose  $T_0 = O(d^2 \ln^2(\frac{1}{\delta}))$  is set according to Equation (1). Then, for all phases  $i \geq 4$ ,*

$$\mathbb{P} \left[ \|\hat{\theta}_i - \theta^*\|_\infty \geq 2^{-i} \right] \leq \frac{\delta}{2^i}, \quad (12)$$

where  $\hat{\theta}_i$  is the estimate of  $\theta^*$  obtained by solving the least squares estimate using all random exploration samples until the beginning of phase  $i$ .

*Proof.* The above lemma follows directly from Lemma 3. Lemma 3 gives that if  $\hat{\theta}_i$  is formed by solving the least squares estimate with at-least  $M_i := O\left((4^i + d) \ln\left(\frac{2^i}{\delta}\right)\right)$  samples, then the guarantee in Equation (12) holds. However, as  $T_0 = O\left((d+1) \ln\left(\frac{2}{\delta}\right)\right)$ , we have naturally that  $M_i \leq 4^i i \sqrt{T_0}$ . The proof is concluded if we show that at the beginning of phase  $i \geq 4$ , the total number of random explorations performed by the algorithm exceeds  $i4^i \lceil \sqrt{T_0} \rceil$ . Notice that at the beginning of any phase  $i \geq 4$ , the total number of random explorations that have been performed is

$$\begin{aligned} \sum_{j=0}^{i-1} 6^j \lceil \sqrt{T_0} \rceil &= \lceil \sqrt{T_0} \rceil \frac{6^i - 1}{4}, \\ &\geq i4^i \lceil \sqrt{T_0} \rceil, \end{aligned}$$

where the last inequality holds for all  $i \geq 10$ . □

The following corollary follows from a straightforward union bound.

**Corollary 3.**

$$\mathbb{P} \left[ \bigcap_{i \geq 4} \|\hat{\theta}_i - \theta^*\|_\infty \leq 2^{-i} \right] \geq 1 - \delta.$$

*Proof.* This follows from a simple union bound as follows.

$$\begin{aligned} \mathbb{P} \left[ \bigcap_{i \geq 4} \|\hat{\theta}_i - \theta^*\|_\infty \leq 2^{-i} \right] &= 1 - \mathbb{P} \left[ \bigcup_{i \geq 4} \|\hat{\theta}_i - \theta^*\|_\infty \geq 2^{-i} \right], \\ &\geq 1 - \sum_{i \geq 4} \mathbb{P} \left[ \|\hat{\theta}_i - \theta^*\|_\infty \geq 2^{-i} \right], \\ &\geq 1 - \sum_{i \geq 4} \frac{\delta}{2^i}, \\ &\geq 1 - \sum_{i \geq 2} \frac{\delta}{2^i}, \\ &= 1 - \frac{\delta}{2}. \end{aligned}$$

□

We are now ready to conclude the proof of Theorem 5.

*Proof of Theorem 5.* We know from Corollary 3, that with probability at-least  $1 - \delta$ , for all phases  $i \geq 10$ , we have  $\|\hat{\theta}_i - \theta^*\|_\infty \leq 2^{-i}$ . Call this event  $\mathcal{E}$ . Now, consider the phase  $i(\gamma) := \max\left(10, \log_2\left(\frac{1}{\gamma}\right)\right)$ . Now, when event  $\mathcal{E}$  holds, then for all phases  $i \geq i(\gamma)$ ,  $\mathcal{D}_i$  is the correct set of  $d^*$  non-zero coordinates of  $\theta^*$ . Thus, with probability at-least  $1 - \delta$ , the total regret upto time  $T$  can be upper bounded as follows

$$\begin{aligned}
 R_T \leq & \sum_{j=0}^{i(\gamma)-1} \left(36^j T_0 + 6^j \lceil \sqrt{T_0} \rceil\right) + \sum_{j \geq i(\gamma)}^{\left\lceil \log_{36}\left(\frac{T}{T_0}\right) \right\rceil} \text{Regret}(\text{OFUL}(1, \delta_i; 36^j T_0)) \\
 & + \sum_{j=i(\gamma)}^{\left\lceil \log_{36}\left(\frac{T}{T_0}\right) \right\rceil} 6^j \lceil \sqrt{T_0} \rceil. \tag{13}
 \end{aligned}$$

The term  $\text{Regret}(\text{OFUL}(L, \delta, T))$  denotes the regret of the OFUL algorithm Abbasi-Yadkori et al. (2011), when run with parameters  $L \in \mathbb{R}_+$ , such that  $\|\theta^*\| \leq L$ , and  $\delta \in (0, 1)$  denotes the probability slack and  $T$  is the time horizon. Equation (13) follows, since the total number of phases is at-most  $\left\lceil \log_{36}\left(\frac{T}{T_0}\right) \right\rceil$ . Standard result from Abbasi-Yadkori et al. (2011) give us that, with probability at-least  $1 - \delta$ , we have

$$\text{Regret}(\text{OFUL}(1, \delta; T)) \leq 4\sqrt{Td^* \ln\left(1 + \frac{T}{d^*}\right)} \left(1 + \sigma\sqrt{2\ln\left(\frac{1}{\delta}\right) + d^* \ln\left(1 + \frac{T}{d^*}\right)}\right).$$

Thus, we know that with probability at-least  $1 - \sum_{i \geq 4} \delta_i \geq 1 - \frac{\delta}{2}$ , for all phases  $i \geq i(\gamma)$ , the regret in the exploration phase satisfies

$$\begin{aligned}
 \text{Regret}(\text{OFUL}(1, \delta_i; 36^i T_0)) & \leq 4\sqrt{d^* 36^i T_0 \ln\left(1 + \frac{36^i T_0}{d^*}\right)} \\
 & \times \left(1 + \sigma\sqrt{2\ln\left(\frac{2^i}{\delta}\right) + d^* \ln\left(1 + \frac{36^i T_0}{d^*}\right)}\right). \tag{14}
 \end{aligned}$$

In particular, for all phases  $i \in [i(\gamma), \lceil \log_{36}\left(\frac{T}{T_0}\right) \rceil]$ , with probability at-least  $1 - \frac{\delta}{2}$ , we have

$$\begin{aligned}
 \text{Regret}(\text{OFUL}(1, \delta_i; 36^i T_0)) & \leq 4\sqrt{d^* 36^i T_0 \ln\left(1 + \frac{T}{d^*}\right)} \\
 & \times \left(1 + \sigma\sqrt{2\ln\left(\frac{T}{T_0 \delta}\right) + d^* \ln\left(1 + \frac{T}{d^*}\right)}\right), \\
 & = \mathcal{C}(T, \delta, d^*) \sqrt{36^i T_0}, \tag{15}
 \end{aligned}$$

where the constant captures all the terms that only depend on  $T$ ,  $\delta$  and  $d^*$ . We can write that constant as

$$\mathcal{C}(T, \delta, d^*) = 4\sqrt{d^* \ln \left(1 + \frac{T}{d^*}\right)} \left(1 + \sigma \sqrt{2 \ln \left(\frac{T}{T_0 \delta}\right) + d^* \ln \left(1 + \frac{T}{d^*}\right)}\right).$$

Equation (15) follows, by substituting  $i \leq \log_{36} \left(\frac{T}{T_0}\right)$  in all terms except the first  $36^i$  term in Equation (14). As Equations (15) and (13) each hold with probability at-least  $1 - \frac{\delta}{2}$ , we can combine them to get that with probability at-least  $1 - \delta$ ,

$$\begin{aligned} R_T &\leq 2T_0 36^{i(\gamma)} + \sum_{j=0}^{\log_{36} \left(\frac{T}{T_0}\right) + 1} \mathcal{C}(T, \delta, d^*) \sqrt{36^j T_0} + \lceil \sqrt{T_0} \rceil 6^{\log_{36} \left(\frac{T}{T_0}\right)}, \\ &\leq \mathcal{O} \left( T_0 36^{i(\gamma)} + \sqrt{T} + \mathcal{C}(T, \delta, d^*) \sum_{j=0}^{\log_{36} \left(\frac{T}{T_0}\right) + 1} \sqrt{36^j T_0} \right), \\ &\stackrel{(a)}{\leq} \mathcal{O} \left( T_0 \frac{2}{\gamma^{5.18}} + \sqrt{T} + \sqrt{T} \mathcal{C}(T, \delta, d^*) \right), \\ &= \mathcal{O} \left( \frac{d^2}{\gamma^{5.18}} \ln^2 \left( \frac{1}{\delta} \right) \right) + \tilde{O} \left( d^* \sqrt{T \ln \left( \frac{1}{\delta} \right)} \right). \end{aligned}$$

Step (a) follows from  $36 \leq 2^{5.18}$ . □

## Appendix C. ALB-Dim for Stochastic Contextual Bandits with Finite Arms

### C.1 ALB-Dim Algorithm for the Finite Armed Case

The algorithm given in Algorithm 4 is identical to the earlier Algorithm 3, except in Line 8, this algorithm uses `SupLinRel` of Chu et al. (2011) as opposed to `OFUL` used in the previous algorithm. In practice, one could also use `LinUCB` of Chu et al. (2011) in place of `SupLinRel`. However, we choose to present the theoretical argument using `SupLinRel`, as unlike `LinUCB`, has an explicit closed form regret bound (see Chu et al. (2011)). The pseudocode is provided in Algorithm 4.

In phase  $i \in \mathbb{N}$ , the `SupLinRel` algorithm is instantiated with input parameter  $36^i T_0$  denoting the time horizon, slack parameter  $\delta_i \in (0, 1)$ , dimension  $d_{\mathcal{M}_i}$  and feature scaling  $b(\delta)$ . We explain the role of these input parameters. The dimension ensures that `SupLinRel` plays from the restricted dimension  $d_{\mathcal{M}_i}$ . The feature scaling implies that when a context  $x \in \mathcal{X}$  is presented to the algorithm, the set of  $K$  feature vectors, each of which is  $d_{\mathcal{M}_i}$  dimensional are  $\frac{\phi^{d_{\mathcal{M}_i}(x,1)}}{b(\delta)}, \dots, \frac{\phi^{d_{\mathcal{M}_i}(x,K)}}{b(\delta)}$ . The constant  $b(\delta) := O \left( \tau \sqrt{\log \left( \frac{TK}{\delta} \right)} \right)$  is chosen such that

$$\mathbb{P} \left[ \sup_{t \in [0, T], a \in \mathcal{A}} \|\phi^M(x_t, a)\|_2 \geq b(\delta) \right] \leq \frac{\delta}{4}.$$



---

**Algorithm 4:** Adaptive Linear Bandit (Dimension) with Finitely Many arms
 

---

- 1: **Input:** Initial Phase length  $T_0$  and slack  $\delta > 0$ .
  - 2:  $\widehat{\beta}_0 = \mathbf{1}$ ,  $T_{-1} = 0$
  - 3: **for** Each epoch  $i \in \{0, 1, 2, \dots\}$  **do**
  - 4:    $T_i = 36^i T_0$ ,    $\varepsilon_i \leftarrow \frac{1}{2^i}$ ,    $\delta_i \leftarrow \frac{\delta}{2^i}$
  - 5:    $\mathcal{D}_i := \{i : |\widehat{\beta}_i| \geq \frac{\varepsilon_i}{2}\}$
  - 6:    $\mathcal{M}_i := \inf\{m : d_m \geq \max \mathcal{D}_i\}$ .
  - 7:   **for** Times  $t \in \{T_{i-1} + 1, \dots, T_i\}$  **do**
  - 8:     Play according to SupLinRel of Auer (2002) with time horizon of  $36^i T_0$  with parameters  $\delta_i \in (0, 1)$ , dimension  $d_{\mathcal{M}_i}$  and feature scaling  $b(\delta) := O\left(\tau \sqrt{\log\left(\frac{TK}{\delta}\right)}\right)$ .
  - 9:   **end for**
  - 10: **for** Times  $t \in \{T_i + 1, \dots, T_i + 6^i \sqrt{T_0}\}$  **do**
  - 11:   Play an arm from the action set  $\mathcal{A}$  chosen uniformly and independently at random.
  - 12: **end for**
  - 13:    $\alpha_i \in \mathbb{R}^{S_i \times d}$  with each row being the arm played during all random explorations in the past.
  - 14:    $\mathbf{y}_i \in \mathbb{R}^{S_i}$  with  $i$ -th entry being the observed reward at the  $i$ -th random exploration in the past
  - 15:    $\widehat{\beta}_{i+1} \leftarrow (\alpha_i^T \alpha_i)^{-1} \alpha_i \mathbf{y}_i$ , is a  $d$  dimensional vector
  - 16: **end for**
- 

Such a constant exists since  $(x_t)_{t \in [0, T]}$  are i.i.d. and  $\phi^M(x, a)$  is a sub-gaussian random variable with parameter  $4\tau^2$ , for all  $a \in \mathcal{A}$ . Similar idea was used in Foster et al. (2019).

### C.2 Regret Guarantee for Algorithm 4

In order to specify a regret guarantee, we will need to specify the value of  $T_0$ . We do so as before. For any  $N$ , denote by  $\lambda_{max}^{(N)}$  and  $\lambda_{min}^{(N)}$  to be the maximum and minimum eigen values of the following matrix:  $\Sigma^N := \mathbb{E} \left[ \frac{1}{K} \sum_{j=1}^K \sum_{t=1}^N \phi^M(x_t, j) \phi^M(x_t, j)^T \right]$ , where the expectation is with respect to  $(x_t)_{t \in [T]}$  which is an i.i.d. sequence with distribution  $\mathcal{D}$ . First, given the distribution of  $x \sim \mathcal{D}$ , one can (in principle) compute  $\lambda_{max}^{(N)}$  and  $\lambda_{min}^{(N)}$  for any  $N \geq 1$ . Furthermore, from the assumption on  $\mathcal{D}$ ,  $\lambda_{min}^{(N)} = \widetilde{O}\left(\frac{1}{\sqrt{d}}\right) > 0$  for all  $N \geq 1$ . Choose  $T_0 \in \mathbb{N}$  to be the smallest integer such that

$$\sqrt{T_0} \geq b(\delta) \max \left( \frac{32\sigma^2}{(\lambda_{min}^{(\lceil \sqrt{T_0} \rceil)})^2} \ln(2d/\delta), \frac{4}{3} \frac{(6\lambda_{max}^{(\lceil \sqrt{T_0} \rceil)} + \lambda_{min}^{(\lceil \sqrt{T_0} \rceil)})(d + \lambda_{max}^{(\lceil \sqrt{T_0} \rceil)})}{(\lambda_{min}^{(\lceil \sqrt{T_0} \rceil)})^2} \ln(2d/\delta) \right). \quad (16)$$

As before, it is easy to see that

$$T_0 = O \left( d^2 \ln^2 \left( \frac{1}{\delta} \right) \tau^2 \ln \left( \frac{TK}{\delta} \right) \right).$$

Furthermore, following the same reasoning as in Lemmas 4 and 3, one can verify that for all  $i \geq 4$ ,  $\mathbb{P} \left[ \|\widehat{\beta}_{i-1} - \beta^*\|_\infty \geq 2^{-i} \right] \leq \frac{\delta}{2^i}$ .

**Theorem 6.** *Suppose Algorithm 4 is run with input parameters  $\delta \in (0, 1)$ , and  $T_0$  as given in Equation (16), then with probability at-least  $1 - \delta$ , the regret after a total of  $T$  arm-pulls satisfies*

$$R_T \leq CT_0 \frac{1}{\gamma^{5.18}} + (1 + \ln(2KT \ln T))^{3/2} \sqrt{Td_{m^*}} + \sqrt{T}.$$

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

In order to parse the above theorem, the following corollary is presented.

**Corollary 4.** *Suppose Algorithm 4 is run with input parameters  $\delta \in (0, 1)$ , and  $T_0 = \widetilde{O}(d^2 \ln^2(\frac{1}{\delta}))$  given in Equation (16), then with probability at-least  $1 - \delta$ , the regret after  $T$  times satisfies*

$$R_T \leq O\left(\frac{d^2}{\gamma^{5.18}} \ln^2(d/\delta) \tau^2 \ln\left(\frac{TK}{\delta}\right)\right) + \widetilde{O}(\sqrt{Td_m^*}).$$

*Proof of Theorem 6.* The proof proceeds identical to that of Theorem 5. Observe from Lemmas 3 and 4, that the choice of  $T_0$  is such that for all phases  $i \geq 1$ , the estimate  $\mathbb{P} \left[ \|\widehat{\beta}_{i-1} - \beta^*\|_\infty \geq 2^{-i} \right] \leq \frac{\delta}{2^i}$ . Thus, from an union bound, we can conclude that

$$\mathbb{P} \left[ \cup_{i \geq 4} \|\widehat{\beta}_{i-1} - \beta^*\|_\infty \geq 2^{-i} \right] \leq \frac{\delta}{4}.$$

Thus at this stage, with probability at-least  $1 - \frac{\delta}{2}$ , the following events holds.

- $\sup_{t \in [0, T], a \in \mathcal{A}} \|\phi^M(x_t, a)\|_2 \leq b(\delta)$
- $\|\widehat{\beta}_{i-1} - \beta^*\|_\infty \leq 2^{-i}$ , for all  $i \geq 10$ .

Call these events as  $\mathcal{E}$ . As before, let  $\gamma > 0$  be the smallest value of the non-zero coordinate of  $\beta^*$ . Denote by the phase  $i(\gamma) := \max\left(10, \log_2\left(\frac{2}{\gamma}\right)\right)$ . Thus, under the event  $\mathcal{E}$ , for all phases  $i \geq i(\gamma)$ , the dimension  $d_{\mathcal{M}_i} = d_m^*$ , i.e., the SupLinRel is run with the correct set of dimensions.

It thus remains to bound the error by summing over the phases, which is done identical to that in Theorem 5. With probability, at-least  $1 - \frac{\delta}{2} - \sum_{i \geq 4} \delta_i \geq 1 - \delta$ ,

$$\begin{aligned} R_T &\leq \sum_{j=0}^{i(\gamma)-1} \left(36^j T_0 + 6^j \sqrt{T_0}\right) + \sum_{j=i(\gamma)}^{\left\lceil \log_{36}\left(\frac{T}{T_0}\right) \right\rceil} \text{Regret}(\text{SupLinRel})(36^j T_0, \delta_i, d_{\mathcal{M}_i, b(\delta)}) \\ &\quad + \sum_{j=i(\gamma)}^{\left\lceil \log_{36}\left(\frac{T}{T_0}\right) \right\rceil} 6^j \sqrt{T_0}, \end{aligned}$$

where  $\text{Regret}(\text{SupLinRel})(36^i T_0, \delta_i, d_{\mathcal{M}_i, b(\delta)}) \leq C(1 + \ln(2K36^i T_0 \ln 36^i T_0))^{3/2} \sqrt{36^i T_0 d_{\mathcal{M}_i}} + 2\sqrt{36^i T_0}$ . This expression follows from Theorem 6 in Auer (2002). We now use this to bound each of the three terms in the display above. Notice from straightforward calculations that the first term is bounded by  $2T_0 36^{i(\gamma)}$  and the last term is bounded above by  $36 \lceil \sqrt{T_0} \rceil 6^{\log_{36}(\frac{T}{T_0})}$  respectively. We now bound the middle term as

$$\begin{aligned} & \left[ \log_{36} \left( \frac{T}{T_0} \right) \right] \\ & \sum_{j=i(\gamma)} \text{Reg}(\text{SupLinRel})(36^j T_0, \delta_i, d_m^*, b(\delta)) \\ & \leq b(\delta) \left( \left[ \log_{36} \left( \frac{T}{T_0} \right) \right] \sum_{j=i(\gamma)} C(1 + \ln(2K36^i T_0 \ln 36^i T_0))^{3/2} \sqrt{36^i T_0 d_{\mathcal{M}_i}} + 2\sqrt{36^i T_0} \right). \end{aligned}$$

The first summation can be bounded as

$$\begin{aligned} & \left[ \log_{36} \left( \frac{T}{T_0} \right) \right] \\ & \sum_{j=i(\gamma)} C(1 + \ln(2K36^i T_0 \ln 36^i T_0))^{3/2} \sqrt{36^i T_0 d_{\mathcal{M}_i}} \\ & \leq \left[ \log_{36} \left( \frac{T}{T_0} \right) \right] \sum_{j=i(\gamma)} C(1 + \ln(2KT \ln T))^{3/2} \sqrt{36^i T_0 d_m^*}, \\ & = C_1(1 + \ln(2KT \ln T))^{3/2} \sqrt{T d_m^*}, \end{aligned}$$

and the second by

$$\left[ \log_{36} \left( \frac{T}{T_0} \right) \right] \sum_{j=i(\gamma)} 2\sqrt{36^i T_0} \leq C_1 \sqrt{T}.$$

Thus, with probability at-least  $1 - \delta$ , the regret of Algorithm 4 satisfies

$$R_T \leq 2T_0 36^{i(\gamma)} + C(1 + \ln(2KT \ln T))^{3/2} \sqrt{T d_m^*} + C_2 \sqrt{T},$$

where  $i(\gamma) := \max \left( 10, \log_2 \left( \frac{2}{\gamma} \right) \right)$ . Thus,

$$R_T \leq CT_0 \frac{2}{\gamma^{5.18}} + C(1 + \ln(2KT \ln T))^{3/2} \sqrt{T d_m^*} + C_1 \sqrt{T},$$

as  $36 \leq 2^{5.18}$

□

## References

- Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.
- Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pages 19–26. PMLR, 2012.
- 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 *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014a.
- 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 Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1638–1646, Beijing, China, 22–24 Jun 2014b. PMLR. URL <http://proceedings.mlr.press/v32/agarwalb14.html>.
- Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corraling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.
- Kaito Ariu, Kenshi Abe, and Alexandre Proutière. Thresholded lasso bandit. *arXiv preprint arXiv:2010.11994*, 2020.
- Sylvain Arlot, Peter L Bartlett, et al. Margin-adaptive model selection in statistical learning. *Bernoulli*, 17(2):687–713, 2011.
- Raman Arora, Teodor Vanislavov Marinov, and Mehryar Mohri. Corraling stochastic bandit algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 2116–2124. PMLR, 2021.
- Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.
- Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best arm with an unknown number of distribution changes. In *European Workshop on Reinforcement Learning*, volume 14, page 375, 2018.
- Sivaraman Balakrishnan, Martin J Wainwright, Bin Yu, et al. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Annals of Statistics*, 45(1):77–120, 2017.
- Krisztian Balog, Filip Radlinski, and Shushan Arakelyan. Transparent, scrutable and explainable user models for personalized recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 265–274, 2019.

- Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020a.
- Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020b.
- Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.
- Lucien Birgé, Pascal Massart, et al. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.
- Alexandra Carpentier and Rémi Munos. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *Artificial Intelligence and Statistics*, pages 190–198, 2012.
- Hugo Caselles-Dupré, Florian Lesaint, and Jimena Royo-Letelier. Word2vec applied to recommendation: Hyperparameters matter. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 352–356, 2018.
- Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.
- Niladri S Chatterji, Vidya Muthukumar, and Peter L Bartlett. Osom: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. *arXiv preprint arXiv:1905.10040*, 2019.
- Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10, 2016.
- Vladimir Cherkassky. Model complexity control and statistical learning theory. *Natural computing*, 1(1):109–133, 2002.
- Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.
- Ashok Cutkosky and Kwabena Boahen. Online learning without prior information. *arXiv preprint arXiv:1703.02629*, 2017.
- Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.
- Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020a.

- Dylan Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3199–3210. PMLR, 13–18 Jul 2020b. URL <http://proceedings.mlr.press/v119/foster20a.html>.
- Dylan J Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. In *Advances in Neural Information Processing Systems*, pages 14714–14725, 2019.
- Avishek Ghosh, Ashwin Pananjady, Adityanand Guntuboyina, and Kannan Ramchandran. Max-affine regression: Provable, tractable, and near-optimal statistical estimation. *arXiv preprint arXiv:1906.09255*, 2019.
- Avishek Ghosh, Abishek Sankararaman, and Ramchandran Kannan. Problem-complexity adaptive model selection for stochastic linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 1396–1404. PMLR, 2021.
- Michael Krikheli and Amir Leshem. 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, 2018.
- Akshay Krishnamurthy, Zhiwei Steven Wu, and Vasilis Syrgkanis. Semiparametric contextual bandits. *arXiv preprint arXiv:1803.04204*, 2018.
- Sanath Kumar Krishnamurthy and Susan Athey. Optimal model selection in contextual bandits with many classes via offline oracles. *arXiv preprint arXiv:2106.06483*, 2021.
- Jeongyeol Kwon and Constantine Caramanis. The em algorithm gives sample-optimality for learning mixtures of well-separated gaussians. In *Conference on Learning Theory*, pages 2425–2487. PMLR, 2020.
- Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
- Jonathan Lee, Aldo Pacchiano, Vidya Muthukumar, Weihao Kong, and Emma Brunskill. Online model selection for reinforcement learning with function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3340–3348. PMLR, 2021.
- Wenjie Li, Adarsh Barik, and Jean Honorio. A simple unified framework for high dimensional bandit problems. *arXiv preprint arXiv:2102.09626*, 2021.
- Andrea Locatelli and Alexandra Carpentier. Adaptivity to smoothness in x-armed bandits. In *Conference on Learning Theory*, pages 1463–1492, 2018.
- Yu Lu and Harrison H Zhou. Statistical and computational guarantees of lloyd’s algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

- Gábor Lugosi, Andrew B Nobel, et al. Adaptive model selection using empirical complexities. *The Annals of Statistics*, 27(6):1830–1864, 1999.
- Haipeng Luo and Robert E Schapire. Achieving all with no parameters: Adanormalhedge. In *Conference on Learning Theory*, pages 1286–1304, 2015.
- Thodoris Lykouris, Karthik Sridharan, and Éva Tardos. Small-loss bounds for online learning with partial information. *arXiv preprint arXiv:1711.03639*, 2017.
- James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM conference on recommender systems*, pages 31–39, 2018.
- Brendan McMahan and Jacob Abernethy. Minimax optimal algorithms for unconstrained linear optimization. In *Advances in Neural Information Processing Systems*, pages 2724–2732, 2013.
- Min-hwan Oh, Garud Iyengar, and Assaf Zeevi. Sparsity-agnostic lasso bandit. *arXiv preprint arXiv:2007.08477*, 2020.
- Francesco Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems*, pages 1116–1124, 2014.
- Aldo Pacchiano, Christoph Dann, Claudio Gentile, and Peter Bartlett. Regret bound balancing and elimination for model selection in bandits and rl. *arXiv preprint arXiv:2012.13045*, 2020a.
- Aldo Pacchiano, My Phan, Yasin Abbasi Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari. Model selection in contextual stochastic bandit problems. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10328–10337. Curran Associates, Inc., 2020b. URL <https://proceedings.neurips.cc/paper/2020/file/751d51528afe5e6f7fe95dece4ed32ba-Paper.pdf>
- Rajat Sen, Alexander Rakhlin, Lexing Ying, Rahul Kidambi, Dean Foster, Daniel Hill, and Inderjit Dhillon. Top- $k$  extreme contextual bandits with arm hierarchy. *arXiv preprint arXiv:2102.07800*, 2021.
- David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability, 2020.
- Aleksandrs Slivkins. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*, 2019.
- Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *CoRR*, abs/1608.05749, 2016. URL <http://arxiv.org/abs/1608.05749>.