Universal and data-adaptive algorithms for model selection in linear contextual bandits

Vidya Muthukumar [1 2] **Akshay Krishnamurthy** [3]

Abstract

Model selection in contextual bandits is an important complementary problem to regret minimization with respect to a fixed model class. We consider the simplest non-trivial instance of model-selection: distinguishing a simple multi-armed bandit problem from a linear contextual bandit problem. Even in this instance, current state-of-the-art methods explore in a suboptimal manner and require strong "feature-diversity" conditions. In this paper, we introduce new algorithms that a) explore in a data-adaptive manner, and b) provide model selection guarantees of the form $\mathcal{O}(d^{\alpha}T^{1-\alpha})$ with *no* feature diversity conditions whatsoever, where d denotes the dimension of the linear model and T denotes the total number of rounds. The first algorithm enjoys a "best-of-both-worlds" property, recovering two prior results that hold under distinct distributional assumptions, simultaneously. The second removes distributional assumptions altogether, expanding the scope for tractable model selection. Our approach extends to model selection among nested linear contextual bandits under some additional assumptions.

1. Introduction

The *contextual bandit (CB) problem* (Woodroofe, 1979; Langford & Zhang, 2008; Li et al., 2010) is a foundational paradigm for online decision-making. In this problem, the decision-maker takes one out of K actions as a function of available contextual information, where this function is chosen from a fixed policy class and is typically learned from past outcomes. Most work on CB has centered around designing algorithms that minimize regret with respect to the best policy in hindsight; a particular non-triviality involves doing this in a computationally efficient manner (Agarwal et al., 2014; Syrgkanis et al., 2016; Foster & Krishnamurthy, 2018). When the rewards are *realizable* under the chosen policy class, this is now an essentially solved problem (Foster & Rakhlin, 2020; Simchi-Levi & Xu, 2020).

A complementary problem to regret minimization with respect to a fixed policy class is *choosing the policy class* that is best for the problem at hand. This constitutes a model selection problem, and its importance is paramount in CB, as selecting a class that either underfits or overfits can lead to highly suboptimal performance. To see why, consider the simplest model selection instance, which involves deciding whether to use the contexts (with, say, a policy class of d-dimensional linear functions) or simply run a multi-armed bandit (MAB) algorithm. Making this choice a priori is suboptimal one way or another: if we choose a MAB algorithm, we obtain the optimal $\mathcal{O}(\sqrt{KT})$ regret to the best fixed action, but the latter may be highly suboptimal if the rewards depend on the context. On the other hand, if we choose a linear CB algorithm, we incur $\mathcal{O}((\sqrt{K}+\sqrt{d})\sqrt{T})$ regret even when the rewards do not depend on the contexts due to overfitting, which is highly suboptimal[1]. To avoid these two failure modes, we seek an algorithm that achieves the best-of-both-worlds, by retaining the linear CB guarantee while adapting to hidden structure if it exists. Focusing on the MAB-vs-linear setting, the strongest variant of the model selection objective asks:

Objective 1. *Can we design a* ***single*** *algorithm that simultaneously achieves the respective minimax-optimal rates of $\mathcal{O}(\sqrt{KT})$ under simple MAB structure (when it exists) and $\mathcal{O}((\sqrt{K}+\sqrt{d})\sqrt{T})$ under d-dimensional linear CB structure?*

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA [2]H. Milton School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, USA [3]Microsoft Research, New York City, USA. Correspondence to: Vidya Muthukumar <vmuthukumar8@gatech.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]Throughout the paper, we consider K to be a small constant and allow the feature dimension d to be quite large. This is in contrast to the typical case of a *linear non-contextual bandit* problem where the number of arms K can be quite large, and the features model structure across arms to ensure a much smaller regret (see, e.g., Chapter 19 of Lattimore & Szepesvári (2019)). There, a rather different model selection problem arises, where the linear bandit class is actually the simpler one. For the CB model selection problem, the $K \geq d$ regime renders the problem uninteresting: it is always better to use a linear CB algorithm as it will better model structure and incur negligible regret overhead.

A related but weaker objective is also proposed in a COLT 2020 open problem (Foster et al., 2020), restated here for the special case of linear CB:

> ***Objective 2.*** *Can we design a* ***single*** *algorithm that simultaneously achieves the rate $\mathcal{O}(K^\beta T^{1-\alpha})$ under simple MAB structure and $\mathcal{O}(K^\beta d^\alpha T^{1-\alpha})$ under linear CB structure, for some $\alpha \in (0, 1/2]$ and some $\beta \in (0, 1)$?*

Foster et al. (2020) highlighted the importance of the $\widetilde{\mathcal{O}}(d^\alpha T^{1-\alpha})$-rate as verifying that model selection is possible whenever the model class is *learnable*. This is the case if and only if $d = o(T)$, assuming, as we do, that K is a small constant (we omit dependence on K in the main paper).

Even in the simplest instance of model selection between a multi-armed bandit and a linear contextual bandit, achieving either Objective 1 or Objective 2 under minimal assumptions remains open. Existing approaches to address either the stronger Objective 1 (Chatterji et al., 2020) or the weaker Objective 2 (Foster et al., 2019) make restrictive assumptions regarding the conditioning (what we will call *diversity*) of the contexts. Other, more data-agnostic approaches (Agarwal et al., 2017; Pacchiano et al., 2020b;a; Lee et al., 2021) achieve neither of the above objectives. This leads us to ask whether we can design a *universal* model selection approach that is data-agnostic (other than requiring a probability model on the contexts) *and* achieves either Objective 1 or 2.

Another important question is the *adaptivity* of approaches to situations in which model selection is especially tractable. At the heart of effective data-driven model selection is a meta-exploration-vs-exploitation tradeoff: while we need to exploit the currently believed simpler model structure, we also need to explore sufficiently to discover potential complex model structure. Most approaches to model selection incorporate forced exploration of an ϵ-greedy type to navigate this tradeoff; however, such exploration may not always be needed. Indeed, (Chatterji et al., 2020) use no forced exploration in their approach and thereby achieve the optimal guarantee of Objective 1, but their approach only works under restrictive diversity assumptions. It is natural to ask whether we can design *data-adaptive* exploration schedules that employ forced exploration only when it is absolutely needed, thus recovering the strongest possible guarantee (Objective 1) under favorable situations and a weaker-but-still-desirable guarantee (Objective 2) otherwise.

1.1. Our contributions

From the above discussion, it is clear that algorithm design for model selection that satisfies the criteria posed in (Foster et al., 2020) involves two non-trivial components: a) designing an efficient statistical test to distinguish between simple and complex model structure that works under minimal assumptions, and b) designing an exploration schedule to ensure that sufficiently expressive data is collected for a) to succeed. In this paper, we advance the state-of-the-art for both of these components in the following ways:

- We design a new test based on eigenvalue thresholding that works for all stochastic sub-Gaussian contexts; in contrast to (Chatterji et al., 2020) and (Foster et al., 2019), it does not require any type of context diversity. We utilize the fact that "low-energy" directions can be thresholded and ignored to estimate the gap in error between model classes. See Theorem 3.2 for our new model selection guarantee, which only requires stochasticity on the contexts to meet Objective 2.
- We also design a data-adaptive exploration schedule that performs forced exploration only when necessary. This approach meets Objective 2 under the *action-averaged* feature diversity assumption that is made in (Foster et al., 2019), but also the stronger Objective 1 under the stronger assumption that the feature of *each action* is diverse (made in (Chatterji et al., 2020)). In fact, our approach not only meets Objective 1 under *action-specific* feature diversity, but even the instance-optimal rate (in line with the results of (Chatterji et al., 2020)). See Theorem 3.3 for a precise statement of our new adaptive guarantee on model selection.

Taken together, our results advance our understanding of model selection for contextual bandits, by demonstrating how statistical approaches can yield universal (i.e., nearly assumption-free) and adaptive guarantees.

1.2. Related work

While model selection is a central and classical topic in machine learning and statistics, most results primarily apply to supervised offline and full-information online learning. Only recently has attention turned to model selection in online partial information settings including contextual bandits. Here, we focus on this growing body of work, which we organize based on overarching algorithmic principles.

Corralling approaches. The first line of work constitutes a hierarchical learning scheme where a meta-algorithm uses bandit techniques to compete with many (contextual) bandit algorithms running as base learners. One of the first such approaches is the CORRAL algorithm of Agarwal et al. (2017), which uses online mirror descent with the log-barrier regularizer as the meta-algorithm. Subsequent work focuses on adapting CORRAL to the stochastic setting (Pacchiano et al., 2020a; Lee et al., 2021) and developing UCB-style meta-algorithms (Cutkosky et al., 2020; Arora et al., 2021). These approaches are quite general and can often be used with abstract non-linear function classes.

However, they do not meet either of Objectives 1 or 2 in general. In our setting, these approaches yield the tuple of rates $(\widetilde{\mathcal{O}}(\sqrt{T}), \widetilde{\mathcal{O}}(d\sqrt{T}))$, which clearly cannot be expressed in the form $(\widetilde{\mathcal{O}}(T^{1-\alpha}), \widetilde{\mathcal{O}}(d^{\alpha}T^{1-\alpha})$ for any value of $\alpha \in (0, 1)$. Consequently, the problem of model selection as described in (Foster et al., 2020) is left open even for linear classes.

Statistical approaches. The second line of algorithmic approaches involves constructing statistical tests for model misspecification. This approach was initially used in the context of model selection concurrently by Foster et al. (2019) and Chatterji et al. (2020), who focus on the linear setting. At a high level, these papers develop efficient misspecification tests under certain covariate assumptions and use these tests to obtain $d^{\alpha}T^{1-\alpha}$-style model selection guarantees. In particular, Foster et al. (2019) use a "sublinear" square loss estimator under somewhat mild covariate assumptions to obtain $d^{1/3}T^{2/3}$ regret, while Chatterji et al. (2020) obtain $\sqrt{dT}$ regret under stronger covariate assumptions. As these two works are the foundation for our results, we discuss these papers in detail in the sequel.

Several recent papers extend statistical testing approaches in several ways. Ghosh et al. (2021a) estimate the support of the parameter vector, which fundamentally incurs a dependence on the magnitude of the smallest non-zero coefficient. Beyond the linear setting, Cutkosky et al. (2021) use the "putative" regret bound for each model class directly to test for misspecification, while Ghosh et al. (2021b); Krishnamurthy & Athey (2021) consider general function classes with realizability. While these latter approaches are more general than ours, they cannot be directly used to obtain our results. Indeed, central to our results (and those of Foster et al. (2019)) is the fact that our statistical test provides a fast rate for detecting misspecification; this guarantee is quantitatively better than what is provided by the putative regret bound, requires carefully adjusting the exploration schedule, and is not available for general function classes.

We also briefly mention two peripherally related lines of work. The first is on representation selection in bandits and reinforcement learning (Papini et al., 2021; Zhang et al., 2021), which involves identifying a feature mapping with favorable properties from a small class of candidate mappings. While this is reminiscent of the model selection problem, the main differences are that in representation selection all mappings are of the same dimensionality and realizable, and the goal is to achieve much faster regret rates by leveraging additional structural properties of the "good" representation.

The second line of work is on Pareto optimality in *non-contextual* bandits and related problems. Beginning with the result of Lattimore (2015), these results show that certain non-uniform regret guarantees are not achievable in various bandit settings. For example, (Lattimore, 2015) shows that, in K-armed bandit problems, one cannot simultaneously achieve $O(\sqrt{T})$ regret to one specific arm, while guaranteeing $O(\sqrt{KT})$ regret to the rest. Such results have been extended to both linear and Lipschitz non-contextual bandits (Zhu & Nowak, 2021; Locatelli & Carpentier, 2018; Krishnamurthy et al., 2020) as well as Lipschitz contextual bandits under margin assumptions (Gur et al., 2021), and they establish that model selection is not possible in these settings. However, these ideas have not been extended to standard contextual bandit settings, which is our focus.

2. Setup

Notation. We use boldface to denote vectors and matrices (e.g. $\boldsymbol{x}$ to denote a vector, and x to denote a scalar). For any value of $M < \infty$, $[M]$ denotes the finite set $\{1, \ldots, M\}$. We use $\|\cdot\|_2$ to denote the ℓ_2-norm of a vector, and $\|\cdot\|_{\mathsf{op}}$ to denote the operator norm of a matrix. We use $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ to denote the Euclidean inner product between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. We use big-Oh notation in the main text; $\widetilde{\mathcal{O}}(\cdot)$ hides dependences on the number of actions K, and $\mathcal{O}_{\delta}(\cdot)$ denotes a bound that hides a $\log\left(\frac{1}{\delta}\right)$ factor and holds with probability at least $1 - \delta$.

2.1. The bandit-vs-contextual bandit problem

The simplest instance of model selection involves a d-dimensional *linear contextual bandit* problem with possibly hidden *multi-armed bandit* structure. This model was proposed as an initial point of study in (Chatterji et al., 2020). Concretely, K actions (which we henceforth call *arms*) are available to the decision-maker at every round, and T denotes the total number of rounds. At round t, the reward of each arm is given by $G_{i,t} = \mu_i + \langle \boldsymbol{x}_{i,t}, \boldsymbol{\theta}^* \rangle + W_{i,t}$, where μ_i denotes the bias of arm i, $\boldsymbol{x}_{i,t} \in \mathbb{R}^d$ denotes the d-dimensional context corresponding to arm i at round t, and $W_{i,t}$ denotes random noise. Finally, $\boldsymbol{\theta}^* \in \mathbb{R}^d$ denotes an unknown parameter. We make the following standard assumptions on the problem parameters.

- The biases $\{\mu_i\}_{i=1}^K$ are assumed to be bounded between -1 and 1.
- The unknown parameter is assumed to be bounded, i.e. $\|\boldsymbol{\theta}^*\|_2 \leq 1$.
- The contexts corresponding to each arm i are assumed to be iid across rounds $t \geq 1$, and 1-sub-Gaussian. We denote by $\boldsymbol{\Sigma}_i$ the covariance matrix of the context $\boldsymbol{x}_{i,t}$, and additionally note that $\boldsymbol{\Sigma}_i \preceq \boldsymbol{I}_d$ as a consequence of the 1-sub-Gaussian assumption. Without loss of generality (since bias can be incorporated into μ_i), we assume that for each arm $i \in [K]$ the mean of the context $\boldsymbol{x}_{i,t}$ is equal to the zero vector.

- The noise $W_{i,t}$ is iid across arms $i \in [K]$ and rounds $t \in [T]$, centered, and 1-sub-Gaussian.

We denote the achieved pseudo-regret with respect to the best fixed arm (the standard metric for a MAB problem) by R_T^S, and the achieved pseudo-regret with respect to the best policy under a d-dimensional model (the standard metric for a linear CB problem) by R_T^C. Notice that in the special case when $\boldsymbol{\theta}^* = \mathbf{0}$, this reduces to a standard multi-armed bandit (MAB) instance. The best possible regret rate is then given by $R_T^S = \mathcal{O}(\sqrt{KT})$ in the worst case, and we also have the instance-dependent rate $R_T^S = \mathcal{O}\left(\sum_{i \neq i^*} \frac{\log T}{\Delta_i}\right)$, where $\Delta_i := \mu^* - \mu_i$. Both of these are known to be information-theoretically optimal (Lai & Robbins, 1985; Audibert et al., 2009). On the other hand, the minimax-optimal rate for the linear contextual bandit (linear CB) problem is given by $R_T^C = \mathcal{O}((\sqrt{d} + \sqrt{K})\sqrt{T})$ (Chu et al., 2011; Abbasi-Yadkori et al., 2011). The following natural dichotomy in algorithm choice presents itself:

1. While the state-of-the-art for the linear CB problem achieves the minimax-optimal rate $R_T^C = \mathcal{O}((\sqrt{d} + \sqrt{K})\sqrt{T})$, it does not adapt automatically to the simpler MAB case. In particular, the regret R_T^S will still scale with the dimension d of the contexts owing both to unnecessary exploration built into linear CB algorithms and overfitting effects. This precludes achieving the minimax-optimal rate of $R_T^S = \mathcal{O}(\sqrt{KT})$ in the MAB setting, let alone the instance-dependent rate.

2. On the other hand, any state-of-the-art algorithm that is tailored to the MAB problem would not achieve any meaningful regret rate for the linear CB problem, simply because it does not incorporate contextual information into its decisions.

The simulations in (Chatterji et al., 2020) empirically illustrate this dichotomy and clearly motivate the model selection problem in its most ambitious form, i.e. Objective 1 as stated in Section 1. Objective 2 constitutes a weaker variant of the model selection problem that was proposed in (Foster et al., 2019; 2020) and justified by the fact that it yields non-trivial model selection guarantees whenever the underlying class is learnable. While Objective 2 is in itself a desirable and non-trivial model selection guarantee, we note that it is strictly weaker than Objective 1. To see this, note that the objectives coincide for $\alpha = 1/2$, and since we require $d < T$ for sublinear regret in the first place, the rate $d^\alpha T^{1-\alpha}$ is a decreasing function in α.

Algorithm 1 Model selection meta-algorithm through one-shot sequential testing. $\widehat{\mathcal{E}}$ denotes an estimator of the square-loss gap between the MAB and linear CB model, $\nu_t \in (0,1)$ denotes a forced-exploration parameter, and $\delta \in (0,1)$ denotes a failure probability.

for $t = 1, \ldots, K$ **do**
  Play arm t and receive reward $g_{t,t}$,
end for
`Current Algorithm` $\leftarrow$ 'MAB'
for $t = K+1, \ldots, n$ **do**
  MAB Algorithm: $i_t =$ arm pulled by UCB
  CB Algorithm: $j_t =$ arm pulled by LINUCB
  if `Current Algorithm` = 'MAB' **then**
    Estimate square loss gap $\widehat{\mathcal{E}}$ and *declare misspecification* if $\widehat{\mathcal{E}} > \alpha_\delta$ (where α_δ is a threshold defined as a function of the filtration $\mathcal{H}_{t-1}$ and the failure probability δ, and is specified for various algorithm choices in Appendices A and B).
    If misspecification detected, then set `Current Algorithm` $\leftarrow$ 'CB'.
  end if
  if `Current Algorithm` = 'MAB' **then**
    Select $U_t = 1$ with probability $1 - \nu_t$, $U_t = 0$ otherwise.
    Play arm $A_t = i_t$ if $U_t = 1$ and $A_t \sim$ Unif$[K]$ if $U_t = 0$.
    Receive reward $g_{A_t,t}$.
  else if `Current Algorithm` = 'CB' **then**
    Play arm j_t and receive $g_{j_t,t}$.
  end if
end for

2.2. Meta-algorithm and prior instantiations

As mentioned in Section 1.2, the vast toolbox of corralling-type approaches does not achieve either Objective 1 or 2 for model selection. (Chatterji et al., 2020) and (Foster et al., 2019), which are concurrent to each other, are among the first approaches to tackle the model selection problem and the only ones that achieve Objectives 1 and 2 respectively—but under additional strong assumptions. Both approaches use the same structure of a statistical test to distinguish between a simple (MAB) and complex (CB) instance. This meta-approach is described in Algorithm 1. Here, A_t denotes the arm that is pulled at round t, and as is standard in bandit literature, $\mathcal{H}_t := \{A_s, G_{A_s,s}\}_{s=1}^t$ is the relevant filtration at round t.

As our results also involve instantiating this meta-algorithm, we now discuss its main elements. The meta-algorithm begins by assuming that the problem is a simple (MAB) instance and primarily uses an optimal MAB algorithm for arm selection: this default choice is denoted by i_t in Algorithm 1. To address model selection, it uses both an

exploration schedule and a *misspecification test*, both of which admit different instantiations. The exploration schedule governs a rate at which the algorithm should choose arms uniformly at random, which can be helpful for detecting misspecification. The misspecification test is simply a surrogate statistical test to check if the instance is, in fact, a complex (CB) instance (i.e. $\boldsymbol{\theta}^* \neq \mathbf{0}$). If the test detects misspecification, we immediately switch to an optimal linear CB algorithm for the remaining time steps.

While (Chatterji et al., 2020) and (Foster et al., 2019) both use the meta-algorithmic structure in Algorithm 1, they instantiate it with difference choices of misspecification test and exploration schedule. The high-level details of where the approaches diverge are summarized in Table 1, and the results that they obtain are summarized in Table 2. We provide a brief description of the salient differences below.

1. Chatterji et al. (2020) do not incorporate any forced exploration in their procedure, as evidenced by the choice of parameter $\nu_t = 0$ above for all values of t. They also use the plug-in estimator of the linear model parameter $\boldsymbol{\theta}^*$ to obtain an estimate of the gap in performance between the two model classes. The error rate of this plug-in estimator scales as $\mathcal{O}(d/n)$ as a function of the number of samples n, and matches the putative regret bound for linear CB. Consequently, they achieve the optimal model selection rate of Objective 1, as well as the stronger instance-optimal rate in the case of MAB, but require a strong assumption of *feature diversity for each arm*; that is, they require $\boldsymbol{\Sigma}_i \succeq \gamma \boldsymbol{I}_d$ for all $i \in [K]$. Intuitively, feature diversity eliminates the need for forced exploration to successfully test for potential complex model structure.

2. Foster et al. (2019) incorporate forced exploration of an ϵ-greedy-style by setting the forced exploration parameter $\nu_t = t^{-1/3}$. This automatically precludes achieving the stronger Objective 1, but leaves the door open to achieving Objective 2 for some smaller choice of α. To do this, they leverage fast estimators (Verzelen et al., 2018; Dicker, 2014; Kong & Valiant, 2018) of the gap between the two model classes, whose error rate can be verified to scale as $\mathcal{O}(\sqrt{d}/n)$ as a function of the number of samples n. This is significantly better in its dependence on d than the standard plug-in estimator. Moreover, forced exploration removes the requirement of restrictive feature diversity assumptions *on each arm*; nevertheless, an *arm-averaged* feature diversity assumption is still required. Specifically, they assume that $\boldsymbol{\Sigma} \succeq \gamma \boldsymbol{I}_d$ where $\boldsymbol{\Sigma} := \frac{1}{K}\sum_{i\in[K]} \boldsymbol{\Sigma}_i$, which is strictly weaker than the arm-specific condition of Chatterji et al. (2020). Above, $\boldsymbol{\Sigma}$ is the covariance matrix of the *mixed* context obtained from uniform exploration, which we denote by $\boldsymbol{x} := \boldsymbol{x}_I$ where $I \sim \mathrm{Unif}[K]$.

This discussion tells us that the initial attempts at model selection (Chatterji et al., 2020; Foster et al., 2019) fall short both in their breadth of applicability and their ability to adapt to structure in the model selection problem. This naturally motivates the question of whether we can design new algorithms with two key properties:

- *Universality*: Can we meet Objective 2 for some value of $\alpha \in (0, 1)$ under stochastic contexts but with no additional diversity assumptions?
- *Adaptivity*: Can we meet Objective 1 under maximally favorable conditions (feature diversity for all arms), and Objective 2 otherwise?

3. Main results

We now introduce and analyze two new algorithms that provide a nearly complete answer to the problems of universality and adaptivity for the MAB-vs-linear CB problem.

3.1. Universal model selection under stochasticity

In this section, we present MODCB.U, a simple variant of MODCB (Foster et al., 2019) that achieves Objective 2 of model selection without requiring *any* feature diversity assumptions, arm-averaged or otherwise. Therefore, this constitutes a *universal model selection* algorithm between an MAB instance and a linear CB instance.

Our starting point is the approach to model selection in (Foster et al., 2019) described above in Section 2.2. Here, we recap the details of the fast estimator $\widehat{\mathcal{E}}(\cdot)$ of the *square-loss-gap*, which is given by $\mathcal{E} := \mathbb{E}\left[(\boldsymbol{x}^\top \boldsymbol{\theta}^*)^2\right] = (\boldsymbol{\theta}^*)^\top \boldsymbol{\Sigma} \boldsymbol{\theta}^*$. The square-loss-gap can be verified to be an upper bound on the expected gap of the best-in-class performance between the CB and MAB models (see (Foster et al., 2019) for details on this upper bound), but is also equal to 0 *iff* $\boldsymbol{\theta}^* = \mathbf{0}$ (and $\boldsymbol{\Sigma}$ is full rank). Therefore, it is a suitable surrogate statistic to test for misspecification, as detailed in the meta-algorithmic structure of Algorithm 1. The estimator is denoted by $\widehat{\mathcal{E}}$, and is described as a black-box procedure in Algorithm 2 with access to an estimator of the covariance matrix $\widehat{\boldsymbol{\Sigma}}_t$ that is constructed from t *unlabeled* samples. The estimator that is used byMODCB (Foster et al., 2019) is simply the sample covariance matrix at round t, defined by

$$\widehat{\boldsymbol{\Sigma}}_t := \frac{1}{Kt} \sum_{s=1}^{t} \sum_{i=1}^{K} \boldsymbol{x}_{i,s} \boldsymbol{x}_{i,s}^\top. \tag{1}$$

Note that such an estimator can be easily constructed as we have access to *all* past contexts $\{\boldsymbol{x}_{i,s}\}_{i\in[K],s\in[t]}$ at any round t. This effective full-information access to contexts, in fact, forms the crux of both of our algorithmic ideas.

This approach is summarized in the sub-routine Algorithm 2,

Algorithm	Estimator $\widehat{\mathcal{E}}(\cdot)$	Forced exploration parameter ν_t
OSOM (Chatterji et al., 2020)	Plug-in estimator	$\nu_t = 0$ (no extra exploration)
MODCB (Foster et al., 2019)	Fast estimator defined in (Foster et al., 2019)	$\nu_t \asymp t^{-1/3}$
MODCB.U	Fast estimator defined in Algorithm 2	$\nu_t \asymp t^{-2/9}$
MODCB.A	Fast estimator defined in (Foster et al., 2019)	Algorithm 3

Table 1. Comparison of model selection algorithms in terms of their estimator for the square loss gap $\widehat{\mathcal{E}}(\cdot)$ and exploration schedule.

which is instantiated in MODCB for any time step t with

$$\{\boldsymbol{x}_i, y_i\}_{i=1}^n := \{\boldsymbol{x}_{A_s,s}, g_{A_s,s} - \widehat{\mu}_{A_s,s}\}_{1\le s\le t: U_s=0}.$$

That is, the set of training examples used is the set of context-reward pairs on all designated *exploration rounds*. Above, $\widehat{\mu}_{i,s}$ constitutes the estimate of the sample means constructed only from past exploration rounds[2]. As a consequence of this choice, we note that for this instantiation of the sub-routine Algorithm 1, we have $n :=$ # of exploration rounds before time step t and $m := t$ at any given time step t.

A key bottleneck lies in the obtainable estimation error rate of $\mathcal{E}$: while the leading dependence is given by $\frac{\sqrt{d}}{\text{\# of exploration rounds}}$ (which is at the heart of the $\widetilde{\mathcal{O}}(d^{1/3}T^{2/3})$ rate that MODCB achieves), there is also an inverse dependence on the minimum eigenvalue of the arm-averaged covariance matrix $\boldsymbol{\Sigma}$, which we denote here by $\gamma_{\min}$. This dependence arises as a consequence of needing to estimate the inverse covariance matrix $\boldsymbol{\Omega} := \boldsymbol{\Sigma}^{-1}$ from unlabeled samples. In essence, this requires $\boldsymbol{\Sigma}$ to be *well-conditioned*, in the sense that we need $\gamma_{\min}$ to be a positive constant to ensure the model selection rate of $\widetilde{\mathcal{O}}(d^{1/3}T^{2/3})$. This precludes non-trivial model selection rates from MODCB for cases where $\gamma_{\min}$ could itself decay with d, the dimension of the contexts, or T, the number of rounds. It also does not allow for cases in which $\boldsymbol{\Sigma}$ may be singular.

Our first main contribution is to adjust MODCB to successfully achieve Objective 2 in model selection with arbitrary stochastic, sub-Gaussian contexts. Because our algorithm achieves a *universal* model selection guarantee over all stochastic context distributions, we name it MODCB.U. The algorithmic procedure is identical to that of MODCB except for the choice of estimator for the inverse covariance matrix, $\widehat{\boldsymbol{\Omega}}$, that is plugged into Algorithm 2. Our key observation is as follows: if certain directions are small in magnitude for the contexts corresponding to *all* arms (as will be the case when $\boldsymbol{\Sigma}$ has vanishingly small eigenvalues), then we may not actually want try to estimate the square loss gap along them: ignoring them might be a better option. Our approach to ignoring low-value directions simply uses eigenvalue thresholding[3] to construct an improved *biased* estimate of the inverse covariance matrix $\widehat{\boldsymbol{\Omega}}$. We formally define the eigenvalue thresholding operator below.

Definition 3.1. Define the clipping operator $[x]_v := \max\{x, v\}$. Then, for any matrix $\boldsymbol{M} \succeq 0$ with diagonalization $\boldsymbol{M} := \boldsymbol{U_M}\boldsymbol{\Lambda_M}\boldsymbol{U_M}^\top$ and any value of $\gamma > 0$, we define the thresholding operator

$$T_\gamma(\boldsymbol{M}) := \boldsymbol{U_M} T_\gamma(\boldsymbol{\Lambda_M})\boldsymbol{U_M}^\top \quad \text{where}$$
$$T_\gamma(\boldsymbol{\Lambda}) := \text{diag}([\lambda_1]_\gamma, \ldots, [\lambda_d]_\gamma).$$

We use Definition 3.1 to specify our (biased) estimators of the covariance and inverse-covariance matrices $\boldsymbol{\Sigma}, \boldsymbol{\Omega} := \boldsymbol{\Sigma}^{-1}$. In particular, we let $\widehat{\boldsymbol{\Sigma}}_t$ denote the sample covariance matrix of $\boldsymbol{\Sigma}$ from t unlabeled samples, given in Eq. (1). Then, our estimators are given by

$$\widehat{\boldsymbol{\Sigma}} := T_\gamma\left(\widehat{\boldsymbol{\Sigma}}_t\right) \quad \text{and} \quad \widehat{\boldsymbol{\Omega}} := \widehat{\boldsymbol{\Sigma}}^{-1}, \tag{2}$$

and we simply plug the estimate $\widehat{\boldsymbol{\Omega}}$ into Algorithm 2. Note that $\widehat{\boldsymbol{\Sigma}}$ is always invertible for any $\gamma > 0$. In essence, this lets us set γ as a tunable parameter to tradeoff the estimation error of a surrogate approximation to the square-loss gap $\mathcal{E}$ (which will *decrease* in γ) and the approximation error that arises from ignoring all directions with value less than γ (which will *increase* in γ). As our first main result shows, we can set a value of γ that scales with d and T and successfully achieve Objective 2 of model selection for *any* stochastic sub-Gaussian context distribution.

Theorem 3.2. MODCB.U *with* $\gamma := (d/T)^{1/3}$ *achieves, with probability at least* $1-\delta$*, model selection rates*

$$R_T^S = \widetilde{\mathcal{O}}_\delta(T^{5/6}) \quad \textit{and} \quad R_T^C = \widetilde{\mathcal{O}}_\delta(d^{1/6}T^{5/6}). \tag{3}$$

Equation (3) clearly demonstrates model selection rates of the form required from Objective 2, and shows that Objective 2 can be met for some value of α with the sole requirement of stochasticity on the contexts. Table 2 allows

[2]This ensures the estimates are unbiased, i.e. $\mathbb{E}[\boldsymbol{x}y] = \boldsymbol{\Sigma}\boldsymbol{\theta}^*$.

[3]Note that eigenvalue thresholding is a type of *hard-thresholding* approach commonly employed in high-dimensional statistics (Bhatia et al., 2015). A plausible alternative approach would be to undertake a *soft-thresholding* approach, by simply adding a non-zero quantity γ to each of the eigenvalues of the sample covariance matrix. This is the approach taken, for example, in ridge regression, and we expect similar results to hold with soft-thresholding.

Algorithm	Obj. 1 (optimal rates)	Obj. 2 ($d^\alpha T^{1-\alpha}$ rates)	context assumption
OSOM (Chatterji et al., 2020)	Yes	Yes ($\alpha = 1/2$)	$\forall i \in [K] : \boldsymbol{\Sigma}_i \succeq \gamma \boldsymbol{I}_d$
MODCB (Foster et al., 2019)	No	Yes ($\alpha = 1/3$)	$\boldsymbol{\Sigma} \succeq \gamma \boldsymbol{I}_d$
MODCB.U	No	Yes ($\alpha = 1/6$)	iid contexts only
CORRAL-STYLE	No	No	iid contexts only

Table 2. Comparison of model selection algorithms in terms of the regret guarantee and assumptions, i.e., "universality." Dependence on the number of arms K is omitted.

Algorithm 2 EstimateResidual

Input: Examples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and second moment matrix estimate $\widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{d\times d}$ (which can be constructed from $m \gg n$ unlabeled sample).
Return estimator

$$\widehat{\mathcal{E}} := \frac{1}{\binom{n}{2}} \sum_{i<j} \left\langle \widehat{\boldsymbol{\Omega}}^{1/2} \boldsymbol{x}_i y_i, \widehat{\boldsymbol{\Omega}}^{1/2} \boldsymbol{x}_j y_j \right\rangle$$

of the square-loss gap $\mathcal{E} := (\boldsymbol{\theta}^*)^\top \boldsymbol{\Sigma} \boldsymbol{\theta}^*$.

Algorithm 3 Data-adaptive exploration schedule

Input: Labeled examples $\{\boldsymbol{x}_{A_s,s}, g_{A_s,s}\}_{s=1}^{t-1}$, current exploration set $\mathcal{W}(t-1)$, hyperparameter $\gamma > 0$.
Construct matrices $\boldsymbol{M}_{t-1} := \sum_{s\in\mathcal{W}(t-1)} \boldsymbol{x}_{A_s,s} \boldsymbol{x}_{A_s,s}^\top$ and $\boldsymbol{M}_t(i_t) := \boldsymbol{M}_{t-1} + \boldsymbol{x}_{i_t,t} \boldsymbol{x}_{i_t,t}^\top$.
Define $Z_t \sim \text{Bernoulli}(1 - t^{-1/3})$ and $Y_t = \mathbb{I}\left[\frac{1}{|\mathcal{W}(t)|+1} \cdot \boldsymbol{M}_t(i_t) \succeq \gamma \boldsymbol{I}_d\right]$.
Add t to "exploration set" $\mathcal{W}(t)$ if ($Y_t = 1$ and $Z_t = 1$) or ($Y_t = 0$ and $Z_t = 0$).
Output: Return $A_t = i_t$ if $Y_t = 1$ *or* $Z_t = 1$; $A_t \sim \text{Unif}[K]$ otherwise.

us to compare the achievable rate to both OSOM (Chatterji et al., 2020) and MODCB (Foster et al., 2019); corralling approaches, which are assumption-free but meet neither Objectives 1 nor 2, are also included as a benchmark. In particular, it is clear from a quick read of the table that as we go from OSOM to MODCB to our approach, the assumptions required on the context distributions weaken, as do the obtainable rates (recall that because the rate $d^\alpha T^{1-\alpha}$ decreases in α, a guarantee with a larger value of α implies one with a smaller value of α).

The proof of Theorem 3.2 is provided in Appendix A. In Appendix C, we describe how this procedure and result extends to the more complex case of linear CB under an additional assumption of block-diagonal structure on the covariance.

3.2. Data-adaptive algorithms for model selection

In this section, we introduce a new *data-adaptive* exploration schedule and show that it provably achieves Objective 1 under the strongest assumption of feature diversity *for each arm* (as in (Chatterji et al., 2020)), but also achieves Objective 2 under the weaker assumption of *arm-averaged* feature diversity (as in (Foster et al., 2019)). Our key insight is that the arm-specific feature diversity condition used by (Chatterji et al., 2020) is itself testable from past contextual information; therefore, it can be tested for *before* we decide on an arm and receive a reward.

To describe this idea formally, we introduce some more notation. At time step t, we denote the exploration set that we have built up thus far by $\mathcal{W}(t-1) \subset [t-1]$. Now, we use an inductive principle. Suppose that the contexts that are present in the exploration set $\mathcal{W}(t-1)$ are already sufficiently "diverse" in a certain quantitative sense (that we will specify shortly). Then, we can easily check whether the arm that we would ideally pull when the true model is simple, i.e. i_t (the "greedy" arm), continues to preserve this property of diversity. Importantly, because we are able to observe the contexts before making a decision, we can check this condition *before* deciding on the value of A_t.

This new sub-routine for data-adaptive exploration, which we call MODCB.A, is described in Algorithm 3. We elaborate on the algorithm description along three critical verticals: a) the decision to forcibly explore, b) the choice of estimator, and c) the designated "exploration rounds" that are used for the estimator.

When to forcibly explore: At time step t, MODCB.A uses the random variables Y_t and Z_t to decide whether to stick with the "greedy" arm $A_t = i_t$, or to forcibly explore, i.e. $A_t \sim \text{Unif}[K]$. For time step t, the random variable Y_t denotes the indicator that the diversity condition continues to be met by context $\boldsymbol{x}_{i_t,t}$. This means that if $Y_t = 1$, we will pick $A_t = i_t$. On the other hand, if the diversity condition is not met (i.e. $Y_t = 0$), we revert to the forced exploration schedule used by MODCB. This schedule sets a variable $Z_t \sim \text{Bernoulli}(1 - t^{-1/3})$, and selects $A_t = i_t$ if $Z_t = 1$ and $A_t \sim \text{Unif}[K]$ otherwise. In summary, we end up picking $A_t = i_t$ if $Z_t = 1$ *or* $Y_t = 1$, while MODCB would have picked $A_t = i_t$ *only if* $Z_t = 1$. As a result, our procedure, called MODCB.A, allows us to adapt on-the-fly to friendly feature diversity structure (and explore much less) while preserving more general guarantees.

The choice of estimator $\widehat{\mathcal{E}}$: First, we specify the choice of estimator of the square loss gap, $\widehat{\mathcal{E}}$ from samples in the designated exploration set $\mathcal{W}(t)$. (We will specify the procedure for construction of this exploration set shortly.) For convenience, we index the elements of the exploration set $\mathcal{W}(t)$ in ascending order by $s_1, \ldots, s_{|\mathcal{W}(t)|}$. We also recall that $\widehat{\boldsymbol{\Sigma}}_t$ denotes the sample covariance matrix as defined in Equation (1). Armed with this exploration set, we define our estimator of $\widehat{\mathcal{E}}(\mathcal{W}(t))$ in accordance with the sub-routine in Algorithm 2 with the examples from the exploration set, i.e. $\{\boldsymbol{x}_{A_{s_j},s_j}, y_{A_{s_j},s_j}\}_{j=1}^{|\mathcal{W}(t)|}$. In particular, we estimate an adjusted square loss gap, given by

$$\overline{\mathcal{E}} := \left\| (\widehat{\boldsymbol{\Sigma}}_t)^{-1/2} \overline{\boldsymbol{\Sigma}}_{|\mathcal{W}(t)|} \boldsymbol{\theta}^* \right\|_2^2, \quad \text{where}$$
$$\overline{\boldsymbol{\Sigma}}_t := \frac{1}{|\mathcal{W}(t)|} \cdot \mathbb{E}\left[\sum_{j=1}^{|\mathcal{W}(t)|} \boldsymbol{x}_{A_{s_j}} \boldsymbol{x}_{A_{s_j}}^\top \right] \tag{4}$$

Note that because $\overline{\boldsymbol{\Sigma}}_t$ is random, the adjusted square loss gap is also random; nevertheless, it turns out that it is *almost surely* a good proxy for the true square loss gap $\mathcal{E}$. We estimate this adjusted squared loss gap with the estimator that is given by

$$\widehat{\mathcal{E}} := \frac{1}{\binom{|\mathcal{W}(t)|}{2}} \sum_{j'<j} \left\langle (\widehat{\boldsymbol{\Sigma}}_t)^{-1/2} \boldsymbol{x}_{s_{j'}} y_{s_{j'}}, (\widehat{\boldsymbol{\Sigma}}_t)^{-1/2} \boldsymbol{x}_{s_j} y_{s_j} \right\rangle. \tag{5}$$

How to build the exploration set $\mathcal{W}(t)$: To complete our description of MODCB.A, we specify the data-adaptive exploration set $\mathcal{W}(t) \subset [t]$ at round t that is used for the estimation subroutine. Notice from Algorithm 3 that we did not include the rounds for which $Y_t \neq Z_t$ in the exploration set. Interestingly, the two cases for which this happens are undesirable for two distinct reasons, as detailed below.

- Rounds on which $Y_t = 0$ and $Z_t = 1$ constitute rounds on which there was no forced exploration and the context corresponding to arm i_t need not be well-conditioned: therefore, we do not want to include these samples for estimation.
- Rounds on which $Y_t = 1$ and $Z_t = 0$ are picked as a sole consequence of well-conditioning on the context $\boldsymbol{x}_{t,i_t}$. When this condition holds, $\boldsymbol{x}_{t,i_t}$ induces good conditioning, however its distribution is affected by the filtering process, inducing bias that complicates estimating the square loss gap. To avoid these complexities, we filter out these rounds. Note that there is no bias when both $Y_t = Z_t = 1$, because the choice $A_t = i_t$ can be attributed to $Z_t = 1$ and not because the context feature induces adequate conditioning. (The proof of Theorem 3.3 highlights that these rounds would make a minimal difference to the ensuing model selection rates.)

Algorithm	Arm-specific diversity	Arm-averaged diversity
OSOM	$\log(T)$/gap and $\sqrt{dT}$	None
MODCB	$T^{2/3}$ and $d^{1/3}T^{2/3}$	$T^{2/3}$ and $d^{1/3}T^{2/3}$
MODCB.A	$\log(T)$/gap and $\sqrt{dT}$	$T^{2/3}$ and $d^{1/3}T^{2/3}$

Table 3. Adaptivity properties of model selection algorithms. We list the regret bounds for simple and complex model under context diversity conditions. By carefully adjusting the exploration schedule, MODCB.A adapts to the favorable "arm-specific diversity" setting. Dependence on the number of arms K is omitted.

This completes our description of our adaptive algorithm, MODCB.A. We show below that MODCB.A achieves the following data-adaptive model selection guarantee.

Theorem 3.3. MODCB.A *with parameter choice $\gamma > 0$ achieves the following model selection rates, each with probability at least $1 - \delta$:*

1. *If feature-diversity holds* ***for every arm*** *with parameter $\gamma' \geq \gamma$, then*

$$R_T^S = \widetilde{\mathcal{O}}_\delta\left(\sum_{i \neq i^*} \frac{\log T}{\Delta_i} \right) \text{ and } R_T^C = \widetilde{\mathcal{O}}_\delta\left(\sqrt{\frac{dT}{\gamma'}} \right). \tag{6}$$

2. *If* ***arm-averaged*** *feature diversity is satisfied with parameter $\gamma' \geq \gamma$, then*

$$R_T^S = \widetilde{\mathcal{O}}_\delta(T^{2/3}) \quad \text{and} \quad R_T^C = \widetilde{\mathcal{O}}_\delta\left(\frac{1}{(\gamma')^2} d^{1/3} T^{2/3} \right). \tag{7}$$

The proof of Theorem 3.3 is provided in Appendix B. Observe that Equation (6) is identical to the OSOM rate, and Equation (7) is identical to the MODCB rate. Consequently, our data-adaptive exploration subroutine results in a *single* algorithm that achieves both rates under the requisite conditions. As summarized in Table 3, OSOM will not work even under arm-averaged feature diversity if arm-specific diversity does not hold. On the other hand, MODCB can be verified not to improve under the stronger condition of arm-specific feature diversity. In conclusion, we can think of MODCB.A as achieving the "best-of-both-worlds" model selection guarantee between the two approaches, by meeting Objective 1 under arm-specific feature diversity and Objective 2 otherwise.

4. Discussion and future work

In this paper, we introduced improved statistical estimation routines and exploration schedules to plug-and-play

with model selection algorithms. The result of these improvements is that we advance the state-of-the-art for model selection along the axes of *universality* and *adaptivity* (as defined at the end of Section 2). Our results are most complete for the model selection problem of MAB-vs-linear CB, but Appendix C presents some extensions to the problem of model selection among linear contextual bandits. Given the recent interest and sharp results for model selection in the linear setting, it is natural to ask whether these ideas extend to any nonlinear setting. Recent work (Marinov & Zimmert, 2021) constructed nonlinear function classes for which Objective 2 cannot be achieved even when the contexts are stochastic. However, identifying nonlinear function classes for which Objective 1 or Objective 2 of model selection is possible remains open and is an important direction for future work. We believe that achieving Objective 2 requires exploiting specific properties of function classes such as the ability to test for misspecification at a fast rate. However, the ideas in OSOM (Chatterji et al., 2020) and our data-adaptive exploration routine MODCB.A are more model-agnostic and instead exploit quantities that depend only on the data distribution. Consequently, we believe they can be generalized to achieve Objective 1 for general function classes when it is possible. To this end, Appendix E presents a generalization of the arm-specific diversity condition that is motivated by statistical concepts in transfer learning/covariate-shift (Quiñonero-Candela et al., 2008), and a consequent possible extension of the principles in OSOM and MODCB.A. Some parts of this extension are relatively straightforward, and we sketch how to do them in Appendix E; however, obtaining strong end-to-end guarantees requires more effort and is an interesting direction for future work.

Acknowledgements

VM acknowledges helpful initial discussions with Weihao Kong, and support from a Simons-Berkeley Research Fellowship for the program "Theory of Reinforcement Learning" in Fall 2020. This work was done in part while the authors were visiting the Simons Institute for the Theory of Computing.

References

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

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, 2014.

Agarwal, A., Luo, H., Neyshabur, B., and Schapire, R. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, 2017.

Arora, R., Marinov, T. V., and Mohri, M. Corralling stochastic bandit algorithms. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Audibert, J.-Y. and Tsybakov, A. B. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 2007.

Audibert, J.-Y., Bubeck, S., et al. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory*, 2009.

Bhatia, K., Jain, P., and Kar, P. Robust regression via hard thresholding. *Advances in Neural Information Processing Systems*, 28, 2015.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Chatterji, N., Muthukumar, V., and Bartlett, P. OSOM: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, 2011.

Cutkosky, A., Das, A., and Purohit, M. Upper confidence bounds for combining stochastic bandits. *arXiv:2012.13115*, 2020.

Cutkosky, A., Dann, C., Das, A., Gentile, C., Pacchiano, A., and Purohit, M. Dynamic balancing for model selection in bandits and RL. In *International Conference on Machine Learning*, 2021.

Dicker, L. H. Variance estimation in high-dimensional linear models. *Biometrika*, 2014.

Dudeja, R. and Hsu, D. Learning single-index models in gaussian space. In *Conference On Learning Theory*, 2018.

Foster, D. and Krishnamurthy, A. Contextual bandits with surrogate losses: Margin bounds and efficient algorithms. In *Advances in Neural Information Processing Systems*, 2018.

Foster, D. and Rakhlin, A. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, 2020.

Foster, D., Krishnamurthy, A., and Luo, H. Model selection for contextual bandits. *Advances in Neural Information Processing Systems*, 2019.

Foster, D., Krishnamurthy, A., and Luo, H. Open problem: Model selection for contextual bandits. In *Conference on Learning Theory*, 2020.

Ghosh, A., Sankararaman, A., and Kannan, R. Problem-complexity adaptive model selection for stochastic linear bandits. In *International Conference on Artificial Intelligence and Statistics*, 2021a.

Ghosh, A., Sankararaman, A., and Ramchandran, K. Model selection for generic contextual bandits. *arXiv:2107.03455*, 2021b.

Gur, Y., Momeni, A., and Wager, S. Smoothness-adaptive contextual bandits. *Available at SSRN*, 2021.

Hanneke, S. and Kpotufe, S. On the value of target data in transfer learning. *arXiv:2002.04747*, 2020.

Kong, W. and Valiant, G. Estimating learnability in the sublinear data regime. *Advances in Neural Information Processing Systems*, 2018.

Kpotufe, S. Lipschitz density-ratios, structured data, and data-driven tuning. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Kpotufe, S. and Martinet, G. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 2021.

Krishnamurthy, A., Langford, J., Slivkins, A., and Zhang, C. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. *Journal of Machine Learning Research*, 2020.

Krishnamurthy, S. K. and Athey, S. Optimal model selection in contextual bandits with many classes via offline oracles. *arXiv:2106.06483*, 2021.

Lai, T. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 1985.

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

Lattimore, T. The pareto regret frontier for bandits. In *Advances in Neural Information Processing Systems*, 2015.

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

Lee, J., Pacchiano, A., Muthukumar, V., Kong, W., and Brunskill, E. Online model selection for reinforcement learning with function approximation. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *International conference on World Wide Web*, 2010.

Lin, Z., Ding, P., and Han, F. Estimation based on nearest neighbor matching: from density ratio to average treatment effect. *arXiv:2112.13506*, 2021.

Locatelli, A. and Carpentier, A. Adaptivity to smoothness in X-armed bandits. In *Conference on Learning Theory*, 2018.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Multiple source adaptation and the rényi divergence. *arXiv:1205.2628*, 2012.

Marinov, T. V. and Zimmert, J. The pareto frontier of model selection for general contextual bandits. *Advances in Neural Information Processing Systems*, 34, 2021.

Mendelson, S. Learning without concentration. In *Conference on Learning Theory*, 2014.

Nemirovski, A. Topics in non-parametric statistics. *Lecture Notes in Mathematics*, 2000.

Pacchiano, A., Dann, C., Gentile, C., and Bartlett, P. Regret bound balancing and elimination for model selection in bandits and RL. *arXiv:2012.13045*, 2020a.

Pacchiano, A., Phan, M., Abbasi-Yadkori, Y., Rao, A., Zimmert, J., Lattimore, T., and Szepesvari, C. Model selection in contextual stochastic bandit problems. *arXiv:2003.01704*, 2020b.

Papini, M., Tirinzoni, A., Restelli, M., Lazaric, A., and Pirotta, M. Leveraging good representations in linear contextual bandits. *arXiv:2104.03781*, 2021.

Peña, V. H., Lai, T. L., and Shao, Q.-M. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. MIT Press, 2008.

Simchi-Levi, D. and Xu, Y. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 2020.

Syrgkanis, V., Krishnamurthy, A., and Schapire, R. Efficient algorithms for adversarial contextual learning. In *International Conference on Machine Learning*, 2016.

Verzelen, N., Gassiat, E., et al. Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, 2018.

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

Woodroofe, M. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 1979.

Yu, L., Jin, Y., and Ermon, S. A unified framework for multi-distribution density ratio estimation. *arXiv:2112.03440*, 2021.

Zhang, W., He, J., Zhou, D., Zhang, A., and Gu, Q. Provably efficient representation learning in low-rank markov decision processes. *arXiv:2106.11935*, 2021.

Zhu, Y. and Nowak, R. Pareto optimal model selection in linear bandits. *arXiv:2102.06593*, 2021.

A. Proof of Theorem 3.2

The following lemma (intended to replace Theorem 2 in Foster et al. (2019)) characterizes how the estimation error of our thresholded estimator $|\widehat{\mathcal{E}} - \mathcal{E}|$ will depend on the choice of γ.

Lemma A.1. *Suppose that we have n labeled samples and m unlabeled samples. Provided that $m \geq C(d + \log(2/\delta))/\gamma$, the estimator provided in Algorithm 2 guarantees that*

$$|\widehat{\mathcal{E}} - \mathcal{E}| \leq \frac{1}{2}\mathcal{E} + \alpha_\delta(n, m) \text{ where} \tag{8}$$
$$\alpha_\delta(n, m) := \mathcal{O}\left(\frac{1}{\gamma} \cdot \frac{d^{1/2}\log^2(2d/\delta)}{n} + \frac{1}{\gamma^4} \cdot \|\mathbb{E}[\boldsymbol{x}y]\|_2^2 \cdot \frac{d + \log(2/\delta)}{m} + \gamma\right)$$

with probability at least $1 - \delta$.

This is similar to the bound in Foster et al. (2019) (with slightly improved inverse dependences on the threshold γ due to the relative simplicity of the MAB-vs-linear CB setting), except that we are not assuming any spectral conditions on $\boldsymbol{\Sigma}$ and we incur an extra additive term of $\mathcal{O}(\gamma)$ in the estimation error arising from the bias induced by the thresholding operator. For comparison, the bound provided in Foster et al. (2019) is for the choice

$$\alpha_\delta(n, m) := \mathcal{O}\left(\frac{1}{\gamma^2} \cdot \frac{d^{1/2}\log^2(2d/\delta)}{n} + \frac{1}{\gamma^4} \cdot \|\mathbb{E}[\boldsymbol{x}y]\|_2^2 \cdot \frac{d + \log(2/\delta)}{m}\right), \tag{9}$$

but only holds if we have $\boldsymbol{\Sigma} \succeq \gamma \boldsymbol{I}_d$.

Before proving Lemma A.1, we sketch how it leads to the statement provided in Theorem 3.2. We follow the outline that is given in Appendix C.2.4 of Foster et al. (2019). An examination of that proof, specialized to the case of 2 model classes (in our case, MAB and linear CB), demonstrates that the dominant terms in the overall regret under the complex model (see, e.g. Eqs. (19), (20), (21) and (22) in Appendix C.2.4 of (Foster et al., 2019)) are given by

$$T\sqrt{K \cdot \alpha_\delta(|\mathcal{W}(T)|, T)} + |\mathcal{W}(T)|,$$

where $\mathcal{W}(T)$ denotes the set of designated exploration rounds. We set $\nu_t = t^{-\kappa}$ to be the forced exploration parameter (as defined in Algorithm 1), and specify a choice of κ subsequently. Just as in (Foster et al., 2019), we then have $|\mathcal{W}(T)| \leq \sqrt{\log(2/\delta)}K^\kappa T^{1-\kappa}$ with probability at least $1 - \delta$. Plugging $n := |\mathcal{W}(T)|$ and $m := T$ into Lemma A.1 then gives us

$$\begin{aligned} R_T &= \mathcal{O}\left(T\sqrt{K \cdot \left(\frac{1}{\gamma} \cdot \frac{d^{1/2}\log^2(2d/\delta)}{K^\kappa T^{1-\kappa}} + \frac{1}{\gamma^4} \cdot \|\mathbb{E}[\boldsymbol{x}y]\|_2^2 \cdot \frac{d + \log(2/\delta)}{T} + \gamma\right)} + \sqrt{\log(2/\delta)}K^\kappa T^{1-\kappa}\right) \\ &= \mathcal{O}_\delta\left(K^\kappa T^{1-\kappa} + \frac{1}{\sqrt{\gamma}} \cdot K^{\frac{1}{2}(1-\kappa)} \cdot d^{1/4} \cdot T^{\frac{1+\kappa}{2}} + \frac{1}{\gamma^2} \cdot \sqrt{KdT} + \gamma T\sqrt{K}\right) \end{aligned}$$

with probability at least $1 - \delta$. Note that the extra $\gamma T\sqrt{K}$ term comes from the estimation error due to misspecification (bias) that we now incur. We now need to select the truncation amount γ and the exploration factor κ to minimize the above expression. One way of doing this is given by equating the third and fourth terms (ignoring universal constants and log factors). This gives us $\gamma^3 = \sqrt{\frac{d}{T}}$. Substituting this into the above gives us

$$R_T = \mathcal{O}_\delta\left(K^\kappa T^{1-\kappa} + \left(\frac{T}{d}\right)^{1/12} K^{\frac{1}{2}(1-\kappa)} T^{\frac{1}{2}(1+\kappa)} \cdot d^{1/4} + \sqrt{K}d^{1/6}T^{5/6}\right),$$

and further substituting $\kappa = 2/9$ gives us

$$R_T = \mathcal{O}_\delta\left(K^{2/9}T^{7/9} + K^{7/18} \cdot d^{1/6} \cdot T^{7/9} + \sqrt{K}d^{1/6}T^{5/6}\right) = \widetilde{\mathcal{O}}(d^{1/6}T^{5/6}),$$

which clearly satisfies the form $R_T = \widetilde{\mathcal{O}}(d^\alpha T^{1-\alpha})$ for the case $\alpha = 1/6$.

Now that we understand how Lemma A.1 leads to Theorem A.1, let us prove it.

Proof. Before beginning the proof, we define a term called the *truncated square-loss gap* as below:

$$\widetilde{\mathcal{E}} := (\boldsymbol{\Sigma}\boldsymbol{\theta}^*)^\top T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}\boldsymbol{\theta}^*. \tag{10}$$

We also recall that we defined $\boldsymbol{\Omega} := \boldsymbol{\Sigma}^{-1}$ and $\widehat{\boldsymbol{\Omega}} := \widehat{\boldsymbol{\Sigma}}^{-1}$, where recall that $\widehat{\boldsymbol{\Sigma}}$ is the truncated second moment estimate. The proof is carried out in three distinct steps:

1. Upper-bounding $\left|\widehat{\mathcal{E}} - \mathbb{E}\left[\widehat{\mathcal{E}}\right]\right|$, the "variance" estimation error arising from n samples.
2. Upper-bounding $\left|\mathbb{E}\left[\widehat{\mathcal{E}}\right] - \widetilde{\mathcal{E}}\right|$, the bias-term with respect to the truncated squared loss gap.
3. Upper-bounding $\left|\widetilde{\mathcal{E}} - \mathcal{E}\right|$, the bias arising from truncation.

1. Upper-bounding $\left|\widehat{\mathcal{E}} - \mathbb{E}\left[\widehat{\mathcal{E}}\right]\right|$. We note that $\mathbb{E}\left[\widehat{\mathcal{E}}\right] = \left\|\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{\Omega}}\mathbb{E}\left[\boldsymbol{x}y\right]\right\|_2^2$. We consider the random vector

$$\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{\Omega}}\boldsymbol{x}y - \widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{\Omega}}\mathbb{E}\left[\boldsymbol{x}y\right],$$

and show that it is sub-exponential with parameter $\mathcal{O}(2/\gamma)$. This follows because $\left\|\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{\Omega}}\right\|_{\mathsf{op}} \leq \left\|\widehat{\boldsymbol{\Omega}}^{1/2}\right\|_{\mathsf{op}} \leq \frac{1}{\sqrt{\gamma}}$, where the second-last inequality follows by the definition of the truncation operator.

Thus, using the sub-exponential tail bound just as in Lemma 17, (Foster et al., 2019), we get

$$\left|\widehat{\mathcal{E}} - \mathbb{E}\left[\widehat{\mathcal{E}}\right]\right| = \mathcal{O}\left(\frac{1}{\gamma}\cdot\frac{d^{1/2}\log^2(2d/\delta)}{n} + \frac{1}{\sqrt{\gamma}}\cdot\frac{\left\|\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{\Omega}}\mathbb{E}\left[\boldsymbol{x}y\right]\right\|_2\log(2/\delta)}{\sqrt{n}}\right).$$

Now, we note that $\left\|\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{\Omega}}\mathbb{E}\left[\boldsymbol{x}y\right]\right\|_2 = \sqrt{\mathbb{E}\left[\widehat{\mathcal{E}}\right]}$. Therefore, we apply the AM-GM inequality to deduce that

$$\left|\widehat{\mathcal{E}} - \mathbb{E}\left[\widehat{\mathcal{E}}\right]\right| \leq \frac{1}{8}\mathbb{E}\left[\widehat{\mathcal{E}}\right] + \mathcal{O}\left(\frac{1}{\gamma}\cdot\frac{d^{1/2}\log^2(2d/\delta)}{n}\right) \tag{11}$$

2. Upper-bounding $\left|\mathbb{E}\left[\widehat{\mathcal{E}}\right] - \widetilde{\mathcal{E}}\right|$. We denote $\boldsymbol{\mu} := \mathbb{E}\left[\boldsymbol{x}y\right]$ as shorthand. It is then easy to verify that $\mathbb{E}\left[\widehat{\mathcal{E}}\right] = \langle\widehat{\boldsymbol{\Omega}}\boldsymbol{\mu}, \boldsymbol{\mu}\rangle$ and $\widetilde{\mathcal{E}} = \langle T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu}, \boldsymbol{\mu}\rangle$. Then, following an identical sequence of steps to (Foster et al., 2019), we get

$$\left|\mathbb{E}\left[\widehat{\mathcal{E}}\right] - \widetilde{\mathcal{E}}\right| \leq \frac{1}{8}\widetilde{\mathcal{E}} + \mathcal{O}\left(\left\|(\widehat{\boldsymbol{\Omega}} - T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1})\boldsymbol{\mu}\right\|_2^2\right)$$

We now state and prove the following lemma on operator norm control.

Lemma A.2. *We have*

$$\left\|\widehat{\boldsymbol{\Omega}} - T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1}\right\|_{\mathsf{op}} \leq \mathcal{O}\left(\frac{\epsilon}{\gamma^2}\right), \tag{12}$$

where we denote $\epsilon := \sqrt{\frac{d+\log(2/\delta)}{m}}$ *as shorthand.*

Note that substituting Lemma A.2 above directly gives us

$$\left|\mathbb{E}\left[\widehat{\mathcal{E}}\right] - \widetilde{\mathcal{E}}\right| \leq \frac{1}{8}\widetilde{\mathcal{E}} + \mathcal{O}\left(\frac{\|\boldsymbol{\mu}\|_2^2\cdot(d+\log(2/\delta))}{\gamma^4 m}\right). \tag{13}$$

We will prove Lemma A.2 at the end of this proof.

3. Upper-bounding $\left|\widetilde{\mathcal{E}} - \mathcal{E}\right|$. Observe that

$$\widetilde{\mathcal{E}} = (\boldsymbol{\theta}^*)^\top \boldsymbol{\Sigma} T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1} \boldsymbol{\Sigma}\boldsymbol{\theta}^* \quad \text{and} \quad \mathcal{E} = (\boldsymbol{\theta}^*)^\top \boldsymbol{\Sigma}\boldsymbol{\theta}^*.$$

This directly implies that

$$|\widetilde{\mathcal{E}} - \mathcal{E}| = |(\boldsymbol{\theta}^*)^\top (\boldsymbol{\Sigma} T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1} \boldsymbol{\Sigma} - \boldsymbol{\Sigma})\boldsymbol{\theta}^*| \leq \left\|(\boldsymbol{\Sigma} T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1} \boldsymbol{\Sigma} - \boldsymbol{\Sigma})\right\|_{\mathsf{op}},$$

where the second inequality follows because we have assumed bounded signal, i.e. $\|\boldsymbol{\theta}^*\|_2 \leq 1$. It remains to control the operator norm terms above. We denote $\boldsymbol{\Sigma} := \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, and note that $T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1} := \boldsymbol{U} T_\gamma\left(\boldsymbol{\Lambda}\right)^{-1} \boldsymbol{U}^\top$. Thus, we get

$$\left\|(\boldsymbol{\Sigma} T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1} \boldsymbol{\Sigma} - \boldsymbol{\Sigma})\right\|_{\mathsf{op}} = \left\|(\boldsymbol{\Lambda} T_\gamma\left(\boldsymbol{\Lambda}\right)^{-1} \boldsymbol{\Lambda} - \boldsymbol{\Lambda})\right\|_{\mathsf{op}} \leq \gamma.$$

Putting these together gives us

$$\left|\widetilde{\mathcal{E}} - \mathcal{E}\right| \leq \gamma. \tag{14}$$

Thus, putting together Equations (11), (13) and (14) completes the proof. It only remains to prove Lemma A.2, which we now do. □

Proof of Lemma A.2. First, recall that $\widehat{\boldsymbol{\Sigma}} := T_\gamma\left(\widehat{\boldsymbol{\Sigma}}_m\right)$, and so we really want to upper bound the quantity $\left\|T_\gamma\left(\widehat{\boldsymbol{\Sigma}}_m\right) - T_\gamma\left(\boldsymbol{\Sigma}\right)\right\|_{\mathsf{op}}$. It is well known (see, e.g. (Boyd et al., 2004)) that for any positive semidefinite matrix $\boldsymbol{M}$, the operator $T_\gamma\left(\boldsymbol{M}\right) - \gamma \boldsymbol{I}_d$ is a proximal operator with respect to the convex nuclear norm functional. The non-expansiveness of proximal operators then gives us

$$\left\|T_\gamma\left(\widehat{\boldsymbol{\Sigma}}_m\right) - T_\gamma\left(\boldsymbol{\Sigma}\right)\right\|_{\mathsf{op}} = \left\|T_\gamma\left(\widehat{\boldsymbol{\Sigma}}_m\right) - \gamma\boldsymbol{I}_d - (T_\gamma\left(\boldsymbol{\Sigma}\right) - \gamma\boldsymbol{I}_d)\right\|_{\mathsf{op}} \leq \left\|\widehat{\boldsymbol{\Sigma}}_m - \boldsymbol{\Sigma}\right\|_{\mathsf{op}} = \mathcal{O}(\epsilon)$$

with probability at least $1 - \delta$. Here, the last step follows by standard arguments on the concentration of the empirical covariance matrix. Recall that we defined $\epsilon := \sqrt{\frac{d+\log(2/\delta)}{m}}$ as shorthand.

We will now use this to show that Equation (12) holds. We have $T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1} - \widehat{\boldsymbol{\Omega}} = T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1} - \widehat{\boldsymbol{\Sigma}}^{-1} = (T_\gamma\left(\boldsymbol{\Sigma}\right))^{-1}(\widehat{\boldsymbol{\Sigma}} - T_\gamma\left(\boldsymbol{\Sigma}\right))\widehat{\boldsymbol{\Sigma}}^{-1}$. By the sub-multiplicative property of the operator norm, we then get

$$\left\|(T_\gamma\left(\boldsymbol{\Sigma}\right))^{-1} - \widehat{\boldsymbol{\Sigma}}^{-1}\right\|_{\mathsf{op}} \leq \left\|(T_\gamma\left(\boldsymbol{\Sigma}\right))^{-1}\right\|_{\mathsf{op}} \left\|\widehat{\boldsymbol{\Sigma}} - T_\gamma\left(\boldsymbol{\Sigma}\right)\right\|_{\mathsf{op}} \left\|\widehat{\boldsymbol{\Sigma}}^{-1}\right\|_{\mathsf{op}} \leq \frac{1}{\gamma^2} \left\|\widehat{\boldsymbol{\Sigma}} - T_\gamma\left(\boldsymbol{\Sigma}\right)\right\|_{\mathsf{op}} = \mathcal{O}\left(\frac{\epsilon}{\gamma^2}\right)$$

where the second-to-last inequality is a consequence of the definition of the truncation operation. This shows Equation (12), and completes the proof. □

B. Proof of Theorem 3.3

Proof. Our proof constitutes a deterministic proof working on various events used in Chatterji et al. (2020) and Foster et al. (2019) as well as additional high-probability events that we will define. Recall that for each value of $t = 1, \ldots, T$, we defined the filtration $\mathcal{H}_t := \{A_s, G_{A_s,s}\}_{s=1}^t$.

Meta-analysis: We begin the analysis by providing a common lemma for both cases that will characterize a high-probability regret bound as a functional of two random quantities: a) $|\mathcal{W}(t)|$, the number of designated exploration rounds *that we use for fast estimation*, and $|\mathcal{T}(t)|$, the total number of *forced-exploration* rounds. Here, we define

$$\mathcal{T}(t) := \{s \in [t] : Z_s = 0 \text{ and } Y_s = 0\}.$$

It is easy to verify that by definition, we have $\mathcal{T}(t) \subset \mathcal{W}(t)$. Indeed, recall from the pseudocode in Algorithm 3 that we defined

$$\mathcal{W}(t) := \{s \in [t] : (Z_s = 0 \text{ and } Y_s = 0) \text{ or } (Z_s = 1 \text{ and } Y_s = 1)\}.$$

We first state our guarantee on estimation error. For any $1 \leq s \leq t$, we define

$$\alpha_\delta(s,t) := \mathcal{O}\left(\frac{1}{\gamma} \cdot \frac{d^{1/2}\log^2(2d/\delta)}{s} + \frac{1}{\gamma^4} \cdot \frac{d + \log(2/\delta)}{t}\right). \tag{15}$$

Lemma B.1. *For every $t \geq 1$, we have*

$$|\widehat{\mathcal{E}}(\mathcal{W}(t)) - \widetilde{\mathcal{E}}| \leq \frac{1}{2}\widetilde{\mathcal{E}} + \alpha_\delta(|\mathcal{W}(t)|, t) \tag{16}$$

with probability at least $1-\delta$, and $\widetilde{\mathcal{E}}$ is the adjusted square loss gap given by

$$\widetilde{\mathcal{E}} := \left\|\boldsymbol{\Sigma}^{-1/2}\overline{\boldsymbol{\Sigma}}_t\boldsymbol{\theta}^*\right\|_2^2,$$

and $\overline{\boldsymbol{\Sigma}}_t$ was defined in Equation (4).

Proof. This proof essentially constitutes a martingale adaptation of the proof of fast estimation in (Foster et al., 2019). Let $\tau(n)$ denote the *random* stopping time at which n exploration samples have been collected. Moreover, let $s_1, \ldots, s_n$ denote the (again random) times at which exploration samples were collected, and $A_{s_1}, \ldots, A_{s_n}$ denote the corresponding actions that were taken. Then, we define a time-averaged covariance matrix as

$$\overline{\boldsymbol{\Sigma}}_n := \frac{1}{n} \cdot \mathbb{E}\left[\sum_{j=1}^n \boldsymbol{x}_{A_{s_j}}\boldsymbol{x}_{A_{s_j}}^\top\right]$$

for every value of $n \geq 1$. We state the following technical lemma, which is proved in Appendix D and critically uses the fact that the rounds on which $A_t = i_t$ is picked as a sole consequence of well-conditioning of the context $\boldsymbol{x}_{i_t,t}$ (i.e. if $Y_t = 1$ and $Z_t = 0$) are filtered out of the considered exploration set $\mathcal{W}(t)$.

Lemma B.2. *Assume that $\boldsymbol{\Sigma}_i \preceq \boldsymbol{I}_d$ for all $i \in [K]$. Then, we have $\gamma\boldsymbol{I}_d \preceq \overline{\boldsymbol{\Sigma}}_n \preceq \boldsymbol{I}_d$ for all values of $n \geq 1$.*

We will use Lemma B.2 to prove Lemma B.1. First, we recall the definition of the adjusted square loss gap,

$$\overline{\mathcal{E}}(\mathcal{W}(t)) := \left\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2}\overline{\boldsymbol{\Sigma}}_t\boldsymbol{\theta}^*\right\|_2^2$$

where we overload notation and define $\overline{\boldsymbol{\Sigma}}_t := \overline{\boldsymbol{\Sigma}}_{|\mathcal{W}(t)|}$ as shorthand. Note that $\overline{\boldsymbol{\Sigma}}_t$ is a random quantity because $|\mathcal{W}(t)|$ is random. By Lemma B.2, we have that $\gamma\boldsymbol{I}_d \succeq \overline{\boldsymbol{\Sigma}}_t \preceq \boldsymbol{I}_d$ *almost surely*.

Similar to the proof of Lemma A.1, the analysis proceeds in two parts:

1. Upper-bounding $\left|\widehat{\mathcal{E}}(\mathcal{W}(t)) - \overline{\mathcal{E}}(\mathcal{W}(t))\right|$*:* For a given index $1 \leq j \leq |\mathcal{W}(t)|$, we consider the random vector

$$\boldsymbol{\xi}_{t,j} := \widehat{\boldsymbol{\Sigma}}_t^{-1/2}\left(\boldsymbol{x}_{s_j,A_{s_j}}y_{A_{s_j}} - \mathbb{E}\left[\boldsymbol{x}_{s_j,A_{s_j}}y_{A_{s_j}}|\mathcal{H}_{s_j-1}\right]\right).$$

Now, by an identical argument to that provided in (Foster et al., 2019), we have that for each $j \in [|\mathcal{W}(t)|]$, the random vector $\boldsymbol{\xi}_{t,j}$ is conditionally sub-exponential with parameter $\mathcal{O}(2/\gamma)$. Consequently, using a martingale version of the sub-exponential tail bound (see, e.g., Chapter 2, (Wainwright, 2019)), we get

$$\left|\widehat{\mathcal{E}}(\mathcal{W}(t)) - \overline{\mathcal{E}}(\mathcal{W}(t))\right| = \mathcal{O}\left(\frac{1}{\gamma} \cdot \frac{d^{1/2}\log^2(2d/\delta)}{|\mathcal{W}(t)|} + \frac{1}{\sqrt{\gamma}} \cdot \frac{\left\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2}\overline{\boldsymbol{\Sigma}}_t\boldsymbol{\theta}^*\right\|_2 \log(2/\delta)}{\sqrt{|\mathcal{W}(t)|}}\right).$$

with probability at least $1-\delta$ for all $t \geq 1$. Therefore, we apply the AM-GM inequality to deduce that

$$\left|\widehat{\mathcal{E}}(\mathcal{W}(t)) - \overline{\mathcal{E}}(\mathcal{W}(t))\right| \leq \frac{1}{8}\overline{\mathcal{E}}(\mathcal{W}(t)) + \mathcal{O}\left(\frac{1}{\gamma} \cdot \frac{d^{1/2}\log^2(2d/\delta)}{|\mathcal{W}(t)|}\right). \tag{17}$$

2. Upper-bounding $\left|\overline{\mathcal{E}}(\mathcal{W}(t)) - \widetilde{\mathcal{E}}\right|$. We note that $\widetilde{\mathcal{E}} = \left\|\boldsymbol{\Sigma}^{-1/2}\overline{\boldsymbol{\Sigma}}_t\boldsymbol{\theta}^*\right\|_2^2$, and $\overline{\mathcal{E}}(\mathcal{W}(t)) := \left\|\widehat{\boldsymbol{\Sigma}}_t^{-1/2}\overline{\boldsymbol{\Sigma}}_t\boldsymbol{\theta}^*\right\|_2^2$, where recall that $\boldsymbol{\Sigma}$ denotes the action-averaged covariance matrix. Further, recall that $\widehat{\boldsymbol{\Sigma}}_t$ is the sample covariance matrix constructed from t *unlabeled* samples. Thus, applying Lemma A.2, following an identical sequence of steps to Foster et al. (2019), and using that $\overline{\boldsymbol{\Sigma}}_t \preceq \boldsymbol{I}_d$ almost surely and $\|\boldsymbol{\theta}^*\|_2 \leq 1$, we get

$$\left|\overline{\mathcal{E}}(\mathcal{W}(t)) - \widetilde{\mathcal{E}}\right| \leq \frac{1}{8}\widetilde{\mathcal{E}} + \mathcal{O}\left(\frac{1}{\gamma^4} \cdot \frac{d+\log(2/\delta)}{t}\right). \tag{18}$$

Now, putting Equations (17) and (18) together, we get

$$|\widehat{\mathcal{E}}(\mathcal{W}(t)) - \widetilde{\mathcal{E}}| \leq \frac{1}{2}\widetilde{\mathcal{E}} + \mathcal{O}\left(\frac{1}{\gamma} \cdot \frac{d^{1/2}\log^2(2d/\delta)}{|\mathcal{W}(t)|} + \frac{1}{\gamma^4} \cdot \frac{d+\log(2/\delta)}{t}\right).$$

This completes the proof. □

Next, we prove the following meta-lemma that characterizes the simple model (SM) and complex model (CM) regret purely in terms of the size of the designated exploration set $\mathcal{W}(T)$ and the forced exploration set $\mathcal{T}(T)$ with high probability.

Lemma B.3. *Let τ_{switch} denote the last round before which the algorithm switches to the complex model, if any (otherwise, we define $\tau_{\mathsf{switch}} := T$). Then, for any $\gamma > 0$ and $\delta > 0$, the following result holds with probability at least $1-\delta$ for model selection between MAB and CB:*

$$R_T^S = \mathcal{O}\left(\sum_{i \neq i^*} \frac{1}{\Delta_i}\log\left(\frac{2KT}{\Delta_i\delta}\right) + |\mathcal{T}(T)|\right) \text{ and}$$

$$R_T^C = \mathcal{O}\left(|\mathcal{T}(T)| + \sqrt{dT}\left(1 + \frac{1}{\gamma^2}\right) + \frac{d^{1/4}}{\gamma^2} \cdot \frac{\tau_{\mathsf{switch}}}{\sqrt{|\mathcal{W}(\tau_{\mathsf{switch}})|}}\right)$$

Proof. For this proof, we work on the event

$$\mathcal{A}_0 := \left\{\left|\widetilde{\mathcal{E}} - \widehat{\mathcal{E}}(\mathcal{W}(t))\right| \leq \frac{1}{2}\widetilde{\mathcal{E}} + \alpha_\delta(|\mathcal{W}(t)|, t) \text{ for all } t = 1, \ldots, T\right\},$$

where $\alpha_\delta(\cdot,\cdot)$ was defined in Equation (15). Lemma B.1 showed that this happens with probability at least $1-\delta$. Further, we denote $g_{\max} = \mathbb{E}\left[\max_{i\in[K]} \boldsymbol{x}_i^\top\boldsymbol{\theta}^* + \mu_i\right]$ where the expectation is over the contexts $\{\boldsymbol{x}_i\}_{i\in[K]}$ drawn identically to the contexts $\{\boldsymbol{x}_{i,t}\}_{i\in[K]}$ for any round $t \geq 1$. We now have two cases to analyze:

1. The case where the true model is SM. In this case, note that $\widetilde{\mathcal{E}} = 0$ by definition and the event $\mathcal{A}_0$ directly gives us, for all $t \geq 1$,

   $$\widehat{\mathcal{E}}(\mathcal{W}(t)) \leq 0 + 0 + \alpha_\delta(|\mathcal{W}(t)|, t)$$

   with probability at least $1-\delta$, and so the condition for elimination is never met. Since under this event, we stay in the simple model, we get pseudo-regret

   $$R_T^S \leq \sum_{t\notin\mathcal{T}(T)} (\mu^* - \mu_{A_t}) + |\mathcal{T}(T)| = \sum_{t\notin\mathcal{T}(T)} (\mu^* - \mu_{i_t}) + |\mathcal{T}(T)|$$
   $$= \mathcal{O}\left(\sum_{i\neq i^*} \frac{1}{\Delta_i}\log\left(\frac{2KT}{\Delta_i\delta}\right) + |\mathcal{T}(T)|\right).$$

Above, the first step uses that rewards are bounded. The second step uses that by the definition of $\mathcal{T}(T)$ and the fact that we never switch, we will have $A_t = i_t$ for all $t \notin \mathcal{T}(T)$. The third step uses the fact that SM updates are only made using the rounds that are not "forced-exploration", i.e. not in $\mathcal{T}(T)$. This completes the proof of the lemma for the case of SM.

2. The case where the true model is CM. Let τ_{switch} denote the last round before which the algorithm switches to CM, if any (otherwise, we define $\tau_{\mathsf{switch}} := T$). It suffices to bound the regret until time τ_{switch} (when we will be playing the SM). First, we note that because because we have not yet switched, we have $\widehat{\mathcal{E}}(\mathcal{W}(\tau_{\mathsf{switch}})) \leq \alpha_\delta(|\mathcal{W}(\tau_{\mathsf{switch}})|, \tau_{\mathsf{switch}})$. Therefore, we get

$$\widetilde{\mathcal{E}} \leq 2\widehat{\mathcal{E}}(\mathcal{W}(\tau_{\mathsf{switch}})) + \alpha_\delta(|\mathcal{W}(\tau_{\mathsf{switch}})|, \tau_{\mathsf{switch}}) \leq 3\alpha_\delta(|\mathcal{W}(\tau_{\mathsf{switch}})|, \tau_{\mathsf{switch}}).$$

Furthermore, we note that $\widetilde{\mathcal{E}} \geq \|\boldsymbol{\theta}^*\|_2^2 \cdot \gamma^2$ (because we have $\overline{\boldsymbol{\Sigma}}_{\tau_{\mathsf{switch}}} \succeq \gamma \boldsymbol{I}_d$ almost surely, and $\boldsymbol{\Sigma}^{-1/2} \succeq I$). This gives us

$$\|\boldsymbol{\theta}^*\|_2^2 \leq \frac{3}{\gamma^2} \cdot \alpha_\delta(|\mathcal{W}(\tau_{\mathsf{switch}})|, \tau_{\mathsf{switch}}). \tag{19}$$

Next, we get

$$\begin{aligned}
R_T^C &= \sum_{t=1}^{\tau_{\mathsf{switch}}} (\boldsymbol{x}_{\kappa_t,t}^\top \boldsymbol{\theta}^* + \mu_{\kappa_t} - \boldsymbol{x}_{A_t,t}^\top \boldsymbol{\theta}^* - \mu_{A_t}) \\
&\leq \sum_{t \notin \mathcal{T}(\tau_{\mathsf{switch}})} (\boldsymbol{x}_{\kappa_t,t}^\top \boldsymbol{\theta}^* + \mu_{\kappa_t} - \boldsymbol{x}_{i_t,t}^\top \boldsymbol{\theta}^* - \mu_{i_t}) + |\mathcal{T}(\tau_{\mathsf{switch}})| \\
&\leq \tau_{\mathsf{switch}} \cdot \|\boldsymbol{\theta}^*\|_2 + \sum_{t \notin \mathcal{T}(\tau_{\mathsf{switch}})} \mu^* - \mu_{i_t} + |\mathcal{T}(\tau_{\mathsf{switch}})| \\
&\leq \frac{T}{\gamma} \cdot \sqrt{\alpha_\delta(\mathcal{W}(\tau_{\mathsf{switch}}), \tau_{\mathsf{switch}})} + \sum_{i \neq i^*} \frac{1}{\Delta_i} \log\left(\frac{2KT}{\Delta_i \delta}\right) + |\mathcal{T}(\tau_{\mathsf{switch}})| \\
&= \mathcal{O}\left(|\mathcal{T}(\tau_{\mathsf{switch}})| + \sqrt{d\tau_{\mathsf{switch}}} \cdot \frac{1}{\gamma^3} + \frac{d^{1/4}}{\gamma^2} \cdot \frac{\tau_{\mathsf{switch}}}{\sqrt{|\mathcal{W}(\tau_{\mathsf{switch}})|}}\right) \\
&= \mathcal{O}\left(|\mathcal{T}(T)| + \sqrt{dT} \cdot \frac{1}{\gamma^3} + \frac{d^{1/4}}{\gamma^2} \cdot \frac{\tau_{\mathsf{switch}}}{\sqrt{|\mathcal{W}(\tau_{\mathsf{switch}})|}}\right)
\end{aligned}$$

This completes the proof for the case of CM. The first step is the definition of regret. The second step uses the definition of $\mathcal{T}(T)$ and the fact that we have not switched yet. The third step uses a sub-Gaussian tail bound (and might incur $\log K$ factors). The fourth step uses Equation (19). The fifth step substitutes the definition of $\alpha_\delta(\cdot, \cdot)$ from Equation (15).

□

Armed with this meta-lemma, we now complete the proof of Theorem 3.3 for the cases under which feature holds for all actions, and action-averaged feature diversity holds respectively.

Case 1: Universal feature diversity holds We need to show that $Y_t = 1$ for all $t \geq 1$ with high probability; if this is the case, the proof is a direct consequence of the techniques that are provided in Chatterji et al. (2020). We consider the following anytime statistical event.

$$\mathcal{A}_1 := \{Y_t = 1 \text{ for all } t \in \{\tau_{\min}(\delta, T), \ldots, T\}\}, \text{ where } \tau_{\min}(\delta, T) := \left(\frac{16}{\gamma^2} + \frac{8}{3\gamma}\right) \log\left(\frac{2dT}{\delta}\right)$$

We now show that the event $\mathcal{A}_1$ occurs with probability at least $1 - \delta$. Since i_t is a deterministic functional of the history $\mathcal{H}_{t-1}$, we can directly apply Lemma 7, (Chatterji et al., 2020) (which is itself an application of the matrix Freedman inequality) to get

$$\gamma_{\min}(\boldsymbol{M}_t(i_t)) \geq 1 + \frac{\gamma t}{2} \text{ for all } t \geq \tau_{\min}(\delta, T)$$

with probability at least $1-\delta$. This clearly ensures that $\frac{1}{t}\cdot\gamma_{\min}(\boldsymbol{M}_t(i_t)) \geq \frac{\gamma}{2}$ for all $t \in \{\tau_{\min}(\delta,T),\ldots,T\}$, which is the required condition.

Connecting this to the meta-analysis above, event $\mathcal{A}_1$ ensures that $|\mathcal{T}(T)| = 0$ by the definition of $\mathcal{T}(t)$. This completes the proof for the case where the true model is SM.

For the case of CM, we also need to *lower bound* $|\mathcal{W}(\tau_{\mathsf{switch}})|$. We denote $\tau_{\min} := \tau_{\min}(\delta,T)$ as shorthand for this portion of the proof. It suffices to consider the case where $\tau_{\mathsf{switch}} \geq 2\tau_{\min}$ (as otherwise, we can simply bound $R_T^C \leq 2\tau_{\min}$). Note that $|\mathcal{W}(\tau_{\mathsf{switch}})|$ is lower bounded by the number of rounds $1 \leq t \leq \tau_{\mathsf{switch}}$ for which $Y_t = 1$ and $Z_t = 1$. Since $Z_t \sim \text{Bernoulli}(1-\nu_t)$ and independent of $\{Y_t\}$, on the event $\mathcal{A}_1$ we have

$$|\mathcal{W}(\tau_{\mathsf{switch}})| \geq \sum_{t=\tau_{\min}}^{\tau_{\mathsf{switch}}} Z_t.$$

Because $\tau_{\min} \geq 8$, it is easy to verify that $1-\nu_t \geq 0.5$ for all $t \geq \tau_{\min}$. Then, we note that $\mathbb{E}\left[\sum_{t=\tau_{\min}}^{\tau_{\mathsf{switch}}} Z_t\right] \geq 0.5(\tau_{\mathsf{switch}} - \tau_{\min})$ since $\nu_t \leq 1/2$ for all $t \geq \tau_{\min}$. Applying Hoeffding's inequality then gives us $\sum_{t=\tau_{\min}}^{\tau_{\mathsf{switch}}} Z_t \geq 0.5(\tau_{\mathsf{switch}} - \tau_{\min}) - \mathcal{O}\left(\sqrt{(\tau_{\mathsf{switch}} - \tau_{\min})\log\left(\frac{1}{\delta}\right)}\right)$ with probability at least $1-\delta$. Putting all of this together gives us

$$|\mathcal{W}(\tau_{\mathsf{switch}})| \geq 0.5(\tau_{\mathsf{switch}} - \tau_{\min}) - \mathcal{O}\left(\sqrt{(\tau_{\mathsf{switch}} - \tau_{\min})\log\left(\frac{1}{\delta}\right)}\right) = \Omega(\tau_{\mathsf{switch}}).$$

Plugging this into Lemma B.3 completes the proof of the theorem.

Case 2: Action-averaged feature diversity holds In the case where action-averaged feature diversity holds, it suffices to provide an upper bound on $|\mathcal{T}(T)|$ and a lower bound on $|\mathcal{W}(\tau_{\mathsf{switch}})|$. We will not define any extra statistical events for this case. First, we note that because $Z_t \in \{0,1\}$, we can apply Hoeffding's inequality to get

$$|\mathcal{T}(T)| \leq \sum_{t=1}^{T}(1-Z_t) \leq 2\sum_{t=1}^{T}\nu_t \leq 4T^{2/3}$$

with probability at least $1-e^{-T^{1/3}}$. This gives us SM regret that scales as $\mathcal{O}(T^{2/3})$. (We completely sacrifice on the instance-dependent guarantees in this case.)

Next, we characterize $|\mathcal{W}(\tau_{\mathsf{switch}})|$. We have

$$|\mathcal{W}(\tau_{\mathsf{switch}})| = \sum_{t=1}^{\tau_{\mathsf{switch}}} Y_t Z_t + (1-Y_t)(1-Z_t),$$

and since $\{Z_t\}_{t\geq 1}$ is an iid sequence completely independent of $\{Y_1,\ldots,Y_{\tau_{\mathsf{switch}}}\}$, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{W}(\tau_{\mathsf{switch}})|\{Y_1,\ldots,Y_{\tau_{\mathsf{switch}}}\}\right] &= \mathbb{E}\left[\sum_{t=1}^{\tau_{\mathsf{switch}}}(1-Z_t) + Y_t(2Z_t-1)|\{Y_1,\ldots,Y_{\tau_{\mathsf{switch}}}\}\right] \\ &= \sum_{t=1}^{\tau_{\mathsf{switch}}}\sum_{t=1}^{\tau_{\mathsf{switch}}} t^{-1/3} + Y_t(1-2t^{-1/3}) \geq \sum_{t=1}^{\tau_{\mathsf{switch}}} t^{-1/3} = \Omega({\tau_{\mathsf{switch}}}^{2/3}).\end{aligned}$$

Applying the tower property of conditional expectations then gives us $\mathbb{E}\left[|\mathcal{W}(\tau_{\mathsf{switch}})|\right] = \Omega((\tau_{\mathsf{switch}})^{2/3})$. The last inequality uses the fact that $Y_t \geq 0$ and $1-2t^{-1/3} > 0$, therefore that term can be lower bounded by 0.

This tells us that it suffices to show that $\mathcal{W}(\tau_{\mathsf{switch}})$ concentrates around its expectation. For this, we note that $Y_tZ_t + (1-Y_t)(1-Z_t)$ is bounded between 0 and 2 and use the Azuma-Hoeffding inequality to get $|\mathcal{W}(\tau_{\mathsf{switch}})| \geq \frac{1}{2}\mathbb{E}\left[|\mathcal{W}(\tau_{\mathsf{switch}})|\right]$ with probability at least $1-e^{-(\tau_{\mathsf{switch}})^{1/3}/2}$. It suffices to consider $\tau_{\mathsf{switch}} \geq (\log T)^3$ (as otherwise, we would just have $R_T^C \leq (\log T)^3$). Then, we get $1-e^{-(\tau_{\mathsf{switch}})^{1/3}/2} \geq 1-\frac{1}{\sqrt{T}}$, and so we have $|\mathcal{W}(\tau_{\mathsf{switch}})| \geq \frac{1}{2}\mathbb{E}\left[|\mathcal{W}(\tau_{\mathsf{switch}})|\right]$ with high probability as desired.

Plugging these bounds into Lemma B.3 gives us

$$R_T^C = \mathcal{O}\left(T^{2/3} + \sqrt{dT} + \frac{1}{\gamma^2}\cdot\sqrt{dT} + \frac{d^{1/4}(\tau_{\mathsf{switch}})^{2/3}}{\gamma^2}\right)$$
$$\leq \mathcal{O}\left(T^{2/3} + \sqrt{dT} + \frac{1}{\gamma^2}\cdot\sqrt{dT} + \frac{d^{1/4}T^{2/3}}{\gamma^2}\right),$$

which completes the proof for this case. □

C. Partial results for model selection in nested linear contextual bandits

In this section, we provide a simple extension of our approach to universal model selection to the case of multiple linear contextual bandits (linear CB). We use the setup that is provided in (Foster et al., 2019), i.e. the reward associated with action i at round t is given by

$$g_{i,t} = \langle \boldsymbol{x}_{i,t}, \boldsymbol{\theta}^* \rangle + W_{i,t}, \tag{20}$$

where $\boldsymbol{\theta}^* \in \mathbb{R}^{d_M}$, and d_M denotes the maximal possible dimension of the model (which is also the dimension of the provided contexts $\boldsymbol{x}_{i,t}$). While this assumes that the rewards are *realizable* under this maximal dimensional model, there may be hidden simpler structure that is unknown a-priori. To model this simpler structure, we consider $0 < d_1 < \ldots < d_M$, and stipulate that the rewards are realizable under *model order* $j^* \in [M]$ if only the first d_{j^*} coordinates of $\boldsymbol{\theta}^*$ are non-zero. This means that we can represent the rewards as

$$g_{i,t} = \langle \boldsymbol{x}_{i,t,d_{j^*}}, \boldsymbol{\theta}^*_{(d_{j^*})} \rangle + W_{i,t} \tag{21}$$

where $\boldsymbol{\theta}^*_{(d_{j^*})}$ denotes the first j^* coordinates of $\boldsymbol{\theta}^*$ and $\boldsymbol{x}_{i,t,d_{j^*}}$ denotes the first d_{j^*} coordinates of the context $\boldsymbol{x}_{i,t}$.

In this setup, Objective 1 of the model selection problem corresponds to achieving a rate of the form $R_T^{j^*} = \widetilde{\mathcal{O}}(\sqrt{d_{j^*}T})$, and Objective 2 corresponds to achieving a rate of the form $R_T^{j^*} = \widetilde{\mathcal{O}}(d_{j^*}^{\alpha}T^{1-\alpha})$. Here, j^* is the true model order and $R_T^{j^*}$ denotes the regret with respect to the best parameter Like in the simpler bandit-vs-CB problem, (Foster et al., 2019) achieves Objective 2 for the case $\alpha = 1/3$, but only under the condition that the action-averaged covariance matrix $\boldsymbol{\Sigma}$ is well-conditioned, i.e. we have $\boldsymbol{\Sigma} \succeq \gamma_{\min}\boldsymbol{I}_d$ for a constant $\gamma_{\min}$ that does not depend on d or T. The hidden factors in the model selection rates $\widetilde{\mathcal{O}}(d_{j^*}^{\alpha}T^{1-\alpha})$ scale inversely with $\gamma_{\min}$.

Algorithm 4 EstimateResidualMultipleModels

Input: Examples $\{(\boldsymbol{x}_s, y_s)\}_{s=1}^n$ and second moment matrix estimates $\widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{d\times d}$ and $\widehat{\boldsymbol{\Sigma}}_1 \in \mathbb{R}^{d_1\times d_1}$.
Return estimator

$$\widehat{\mathcal{E}} := \frac{1}{\binom{n}{2}}\sum_{s<t}\left\langle \widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{R}}\boldsymbol{x}_s y_s, \widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{R}}\boldsymbol{x}_t y_t \right\rangle$$

of the square-loss gap $\mathcal{E} := \mathbb{E}\left[(\boldsymbol{x}^\top\boldsymbol{\theta}^* - \boldsymbol{x}_1^\top\boldsymbol{\theta}_1^*)^2\right]$, where $\widehat{\boldsymbol{R}} := \begin{bmatrix}\widehat{\boldsymbol{\Omega}}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} - \widehat{\boldsymbol{\Omega}}$.

For this extension, we directly use the framework of MODCB (Foster et al., 2019) and do not reproduce the details here, except for the square loss gap estimator provided above in Algorithm 4 for two candidate dimensions $d_1 < d$. Here, $d_1 := d_{i-1}$ and $d := d_i$ for some value of $i \in [M]$, and so this routine will be used at the i^{th} stage of model selection, i.e deciding between model orders i and $i+1$. Note that here, we can write the square loss gap in the form

$$\mathcal{E} = (\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{\theta}^*)^\top\boldsymbol{\Sigma}(\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{\theta}^*), \tag{22}$$

where we define $\boldsymbol{R} := \begin{bmatrix}\boldsymbol{\Omega}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} - \boldsymbol{\Omega}$.

Our extension is simple to describe in this context: we simply use the estimators $\widehat{\boldsymbol{\Sigma}} := T_\gamma\left(\widehat{\boldsymbol{\Sigma}}_t\right)$ and $\widehat{\boldsymbol{\Sigma}}_1 := T_\gamma\left(\widehat{\boldsymbol{\Sigma}}_1^{(t)}\right)$ with the subroutine provided in Algorithm 4. We require an extra assumption of *block-diagonal structure* on the covariance matrix, which we provide below.

Definition C.1. We use the notation $d_1 := d_{i-1}$ and $d := d_i$ as above. Then, for each value of $i \in [M]$ the action-averaged covariance matrix $\boldsymbol{\Sigma}$ is assumed to possess block-diagonal structure of the form

$$\boldsymbol{\Sigma} := \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{-1} \end{bmatrix}. \tag{23}$$

Note that under this block-diagonal assumption, we simply get $\boldsymbol{R} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{-1} \end{bmatrix}$.

The assumption in Definition C.1 is utilized to avoid blow-ups in the approximation error arising from thresholding due to cross-correlation terms. It will not, in general, hold for a linear contextual bandit problem; therefore, it is an important question for future work to provide universal model selection rates in the absence of this assumption. Nevertheless, under this assumption we can extend the universal model selection result of Theorem 3.2 to the case of model selection among nested linear CB models, as stated below.

Theorem C.2. *Algorithm 1 with the residual estimator given in Algorithm 4 and the inverse covariance matrix estimate that uses eigenvalue thresholding with the choice $\gamma_j := (d_j/T)^{1/3}$ corresponding to model order j achieves model selection rate*

$$R_T^{j^*} = \widetilde{\mathcal{O}}_\delta(d_{j^*}^{1/6} T^{5/6}). \tag{24}$$

Here, j^* is the minimal model order under which the rewards are realizable, $\delta \in (0,1)$ denotes a failure probability, and the $\widetilde{\mathcal{O}}_\delta(\cdot)$ hides sublinear dependences in K and polylogarithmic dependences on $1/\delta$.

We conclude this section with a proof of Theorem C.2. The following lemma (intended to replace Theorem 2 in Foster et al. (2019)) characterizes the estimation error $|\widehat{\mathcal{E}} - \mathcal{E}|$ will depend on the choice of γ for $d_1 < d$.

Lemma C.3. *We use the notation d_1 and d from Definition C.1 to denote the currently estimated model dimension and the next model dimension respectively. Suppose that we have n labeled samples and m unlabeled samples, and $m > n$. Provided that $m \geq C(d + \log(2/\delta))/\gamma$, the estimator provided in Algorithm 2 guarantees that*

$$|\widehat{\mathcal{E}} - \mathcal{E}| \leq \frac{1}{2}\mathcal{E} + \mathcal{O}\left(\frac{1}{\gamma^2} \cdot \frac{d^{1/2}\log^2(2d/\delta)}{n} + \frac{1}{\gamma^4} \cdot \|\mathbb{E}[\boldsymbol{x}y]\|_2^2 \cdot \frac{d + \log(2/\delta)}{m} + \gamma\right) \tag{25}$$

with probability at least $1 - \delta$.

Thus, we can utilize Lemma C.3 with different thresholding values γ_i corresponding to each estimated model class $i \in [M]$. In particular, it is clear that in a manner similar to the proof of Theorem 3.2, the choice $\gamma_i := (d_i/T)^{1/3}$ will yield the desired model selection rate. Therefore, we only provide the proof of Lemma C.3 here.

Proof. The proof follows along similar lines to the proof of Lemma A.1, but with several extra terms. Before beginning the proof, we define a term called the *truncated square-loss gap* as below:

$$\widetilde{\mathcal{E}} := ([\boldsymbol{R}]_\gamma \boldsymbol{\Sigma}\boldsymbol{\theta}^*)^\top T_\gamma(\boldsymbol{\Sigma})([\boldsymbol{R}]_\gamma \boldsymbol{\Sigma}\boldsymbol{\theta}^*). \tag{26}$$

where we define $[\boldsymbol{R}]_\gamma := \begin{bmatrix} T_\gamma(\boldsymbol{\Sigma}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - T_\gamma(\boldsymbol{\Sigma})^{-1}$ as shorthand. It is easy to see that $[\boldsymbol{R}]_\gamma = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & T_\gamma(\boldsymbol{\Sigma}_{-1})^{-1} \end{bmatrix}$. We also recall that we defined $\boldsymbol{\Omega} := \boldsymbol{\Sigma}^{-1}$ and $\widehat{\boldsymbol{\Omega}} := \widehat{\boldsymbol{\Sigma}}^{-1}$. Finally, we define $\boldsymbol{\mu} := \mathbb{E}[\boldsymbol{x}y] = \boldsymbol{\Sigma}\boldsymbol{\theta}^*$ as shorthand.

The proof is carried out in three distinct steps:

1. Upper-bounding $\left|\widehat{\mathcal{E}} - \mathbb{E}\left[\widehat{\mathcal{E}}\right]\right|$, the "variance" estimation error arising from n samples.
2. Upper-bounding $\left|\mathbb{E}\left[\widehat{\mathcal{E}}\right] - \widetilde{\mathcal{E}}\right|$, the bias-term with respect to the truncated squared loss gap.

3. Upper-bounding $\left|\widetilde{\mathcal{E}} - \mathcal{E}\right|$, the bias arising from truncation.

1. Upper-bounding $\left|\widehat{\mathcal{E}} - \mathbb{E}\left[\widehat{\mathcal{E}}\right]\right|$. We note that $\mathbb{E}\left[\widehat{\mathcal{E}}\right] = \left\|\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{R}}\boldsymbol{\mu}\right\|_2^2$. We consider the random vector

$$\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{R}}\boldsymbol{x}y - \widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{R}}\mathbb{E}\left[\boldsymbol{x}y\right],$$

and show that it is sub-exponential with parameter $\mathcal{O}(2/\gamma)$. This follows because $\left\|\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{R}}\right\|_{\mathsf{op}} \leq \left\|\widehat{\boldsymbol{\Omega}}^{1/2}\right\|_{\mathsf{op}} + \left\|\widehat{\boldsymbol{\Omega}}_1^{1/2}\right\|_{\mathsf{op}} \leq \frac{2}{\sqrt{\gamma}}$, where the last inequality follows by the definition of the truncation operator.

Thus, using the sub-exponential tail bound just as in Lemma 17, (Foster et al., 2019), we get

$$\left|\widehat{\mathcal{E}} - \mathbb{E}\left[\widehat{\mathcal{E}}\right]\right| = \mathcal{O}\left(\frac{1}{\gamma}\cdot\frac{d^{1/2}\log^2(2d/\delta)}{n} + \frac{1}{\sqrt{\gamma}}\cdot\frac{\left\|\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{R}}\mathbb{E}\left[\boldsymbol{x}y\right])\right\|_2\log(2/\delta)}{\sqrt{n}}\right).$$

Now, we note that $\left\|\widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{R}}\mathbb{E}\left[\boldsymbol{x}y\right]\right\|_2 = \sqrt{\mathbb{E}\left[\widehat{\mathcal{E}}\right]}$. Therefore, we apply the AM-GM inequality to deduce that

$$\left|\widehat{\mathcal{E}} - \mathbb{E}\left[\widehat{\mathcal{E}}\right]\right| \leq \frac{1}{8}\mathbb{E}\left[\widehat{\mathcal{E}}\right] + \mathcal{O}\left(\frac{1}{\gamma}\cdot\frac{d^{1/2}\log^2(2d/\delta)}{n}\right) \tag{27}$$

2. Upper-bounding $\left|\mathbb{E}\left[\widehat{\mathcal{E}}\right] - \widetilde{\mathcal{E}}\right|$. We know that $\mathbb{E}\left[\widehat{\mathcal{E}}\right] = \langle\widehat{\boldsymbol{R}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{R}}\boldsymbol{\mu},\boldsymbol{\mu}\rangle$ and
$\widetilde{\mathcal{E}} = \langle[\boldsymbol{R}]_\gamma T_\gamma\left(\boldsymbol{\Sigma}\right)[\boldsymbol{R}]_\gamma\boldsymbol{\mu},\boldsymbol{\mu}\rangle$. Following an identical sequence of steps to Foster et al. (2019) (in particular, the steps leading up to Eq. (16) in Appendix C.1), we get

$$\begin{aligned}\left|\mathbb{E}\left[\widehat{\mathcal{E}}\right] - \widetilde{\mathcal{E}}\right| \leq \frac{1}{8}\widetilde{\mathcal{E}} + \mathcal{O}\Big(&\left\|T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1/2}\widehat{\boldsymbol{\Sigma}}([\boldsymbol{R}]_\gamma - \widehat{\boldsymbol{R}})\boldsymbol{\mu}\right\|_2^2 \\ &+ \left\|T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1/2}\left(\widehat{\boldsymbol{\Sigma}} - T_\gamma\left(\boldsymbol{\Sigma}\right)\right)\widehat{\boldsymbol{R}}\boldsymbol{\mu}\right\|_2^2 + \left\|\widehat{\boldsymbol{\Sigma}}^{1/2}([\boldsymbol{R}]_\gamma - \widehat{\boldsymbol{R}})\boldsymbol{\mu}\right\|_2^2\Big)\end{aligned}$$

Furthermore, note that

$$\widehat{\boldsymbol{R}} - [\boldsymbol{R}]_\gamma = (\widehat{\boldsymbol{\Omega}} - T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1}) - \begin{bmatrix}\widehat{\boldsymbol{\Omega}}_1 - T_\gamma\left(\boldsymbol{\Sigma}_1\right)^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}.$$

We now state and prove the following lemma, which will help us characterize the sum above.

Lemma C.4. *We have*

$$\left\|T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1/2}\widehat{\boldsymbol{\Sigma}}T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1/2} - \boldsymbol{I}_d\right\|_{\mathsf{op}} = \left\|T_\gamma\left(\boldsymbol{\Sigma}_1\right)^{-1/2}\widehat{\boldsymbol{\Sigma}}_1 T_\gamma\left(\boldsymbol{\Sigma}_1\right)^{-1/2} - \boldsymbol{I}_d\right\|_{\mathsf{op}} = \mathcal{O}\left(\frac{\epsilon}{\gamma}\right) \text{ where} \tag{28}$$

$$\epsilon = \sqrt{\frac{d+\log(2/\delta)}{m}}.$$

with probability at least $1-\delta$. *From the above, we get the following inequalities:*

$$\left\|T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1/2}\widehat{\boldsymbol{\Sigma}}\right\|_{\mathsf{op}} \leq \mathcal{O}(1) \tag{29a}$$

$$\left\|\widehat{\boldsymbol{R}}\boldsymbol{\mu}\right\|_2 \leq \mathcal{O}(1/\gamma) \tag{29b}$$

$$\left\|[\boldsymbol{R}]_\gamma - \widehat{\boldsymbol{R}}\right\|_{\mathsf{op}} \leq \mathcal{O}(\epsilon/\gamma^2) \tag{29c}$$

We will prove Lemma C.4 at the end of this proof. Before that, we use it to complete the proof of Lemma A.1. Using Equation (29) and the sub-multiplicative property of the operator norm, we get

$$\left\|T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1/2}\widehat{\boldsymbol{\Sigma}}([\boldsymbol{R}]_\gamma-\widehat{\boldsymbol{R}})\boldsymbol{\mu}\right\|_2^2 \le \left\|T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1/2}\widehat{\boldsymbol{\Sigma}}\right\|_{\mathsf{op}}^2\|\boldsymbol{\mu}\|_2^2\cdot\left\|[\boldsymbol{R}]_\gamma-\widehat{\boldsymbol{R}}\right\|_{\mathsf{op}}^2 \le \mathcal{O}\left(\frac{\|\boldsymbol{\mu}\|_2^2\cdot\epsilon^2}{\gamma^4}\right)$$

$$\begin{aligned}\left\|T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1/2}\left(\widehat{\boldsymbol{\Sigma}}-T_\gamma\left(\boldsymbol{\Sigma}\right)\right)\widehat{\boldsymbol{R}}\boldsymbol{\mu}\right\|_2^2 &\le \|T_\gamma\left(\boldsymbol{\Sigma}\right)\|_{\mathsf{op}}\left\|\widehat{\boldsymbol{R}}\boldsymbol{\mu}\right\|_2^2\cdot\left\|T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1/2}\widehat{\boldsymbol{\Sigma}}T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1/2}-\boldsymbol{I}_d\right\|_{\mathsf{op}}^2\\ &\le \mathcal{O}\left(\frac{\|\boldsymbol{\mu}\|_2^2\cdot\epsilon^2}{\gamma^4}\right)\end{aligned}$$

$$\left\|\widehat{\boldsymbol{\Sigma}}^{1/2}([\boldsymbol{R}]_\gamma-\widehat{\boldsymbol{R}})\boldsymbol{\mu}\right\|_2^2 \le \left\|\widehat{\boldsymbol{\Sigma}}\right\|_{\mathsf{op}}\|\boldsymbol{\mu}\|_2^2\cdot\left\|[\boldsymbol{R}]_\gamma-\widehat{\boldsymbol{R}}\right\|_{\mathsf{op}}^2 \le \mathcal{O}\left(\frac{\|\boldsymbol{\mu}\|_2^2\cdot\epsilon^2}{\gamma^4}\right)$$

This directly gives us

$$\left|\mathbb{E}\left[\widehat{\mathcal{E}}\right]-\widetilde{\mathcal{E}}\right| \le \frac{1}{8}\widetilde{\mathcal{E}}+\mathcal{O}\left(\frac{\|\boldsymbol{\mu}\|_2^2\cdot\epsilon^2}{\gamma^4}\right). \tag{30}$$

3. Upper-bounding $\left|\widetilde{\mathcal{E}}-\mathcal{E}\right|$. Observe that

$$\begin{aligned}\widetilde{\mathcal{E}} &= (\boldsymbol{\theta}^*)^\top\boldsymbol{\Sigma}[\boldsymbol{R}]_\gamma T_\gamma\left(\boldsymbol{\Sigma}\right)[\boldsymbol{R}]_\gamma\boldsymbol{\Sigma}\boldsymbol{\theta}^* \text{ and}\\ \mathcal{E} &= (\boldsymbol{\theta}^*)^\top\boldsymbol{\Sigma}\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}\boldsymbol{\theta}^*.\end{aligned}$$

This directly implies that

$$\begin{aligned}|\widetilde{\mathcal{E}}-\mathcal{E}| &= |(\boldsymbol{\theta}^*)^\top\boldsymbol{\Sigma}([\boldsymbol{R}]_\gamma T_\gamma\left(\boldsymbol{\Sigma}\right)[\boldsymbol{R}]_\gamma-\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R})\boldsymbol{\Sigma}\boldsymbol{\theta}^*|\\ &\le \|\boldsymbol{\Sigma}([\boldsymbol{R}]_\gamma T_\gamma\left(\boldsymbol{\Sigma}\right)[\boldsymbol{R}]_\gamma-\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R})\boldsymbol{\Sigma}\|_{\mathsf{op}},\end{aligned}$$

where the second inequality follows because we have assumed bounded signal, i.e. $\|\boldsymbol{\theta}^*\|_2 \le 1$.

It remains to control the operator norm terms above. We denote $\boldsymbol{\Sigma} := \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, and write $\boldsymbol{\Lambda} := \begin{bmatrix}\boldsymbol{\Lambda}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Lambda}_{-1}\end{bmatrix}$. First, we note that because of the block-diagonal structure considered in Definition C.1, we can also write $\begin{bmatrix}\boldsymbol{\Sigma}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} := \boldsymbol{U}\begin{bmatrix}\boldsymbol{\Lambda}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{U}^\top$. This directly gives us $\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R} = \boldsymbol{U}\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Lambda}_{-1}^{-1}\end{bmatrix}\boldsymbol{U}^\top$.

Second, we note that $T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1} := \boldsymbol{U}T_\gamma\left(\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{U}^\top$. Recalling the definition of $[\boldsymbol{R}]_\gamma$ and using again the block-diagonal structure in Definition C.1, simple algebra gives us

$$[\boldsymbol{R}]_\gamma = \boldsymbol{U}\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & T_\gamma\left(\boldsymbol{\Lambda}_{-1}\right)^{-1}\end{bmatrix}\boldsymbol{U}^\top$$

and $[\boldsymbol{R}]_\gamma T_\gamma\left(\boldsymbol{\Sigma}\right)[\boldsymbol{R}]_\gamma = \boldsymbol{U}\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & T_\gamma\left(\boldsymbol{\Lambda}_{-1}\right)^{-1}\end{bmatrix}\boldsymbol{U}^\top$. Putting all of this together directly gives us

$$\|\boldsymbol{\Sigma}([\boldsymbol{R}]_\gamma T_\gamma\left(\boldsymbol{\Sigma}\right)[\boldsymbol{R}]_\gamma-\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R})\boldsymbol{\Sigma}\|_{\mathsf{op}} = \left\|\boldsymbol{\Lambda}_{-1}T_\gamma\left(\boldsymbol{\Lambda}_{-1}\right)^{-1}\boldsymbol{\Lambda}_{-1}-\boldsymbol{\Lambda}_{-1}\right\|_{\mathsf{op}} \le \gamma$$

and so, we get

$$\left|\widetilde{\mathcal{E}}-\mathcal{E}\right| \le \gamma. \tag{31}$$

Thus, putting together Equations (27), (30) and (31) completes the proof.

It only remains to prove Lemma C.4, which we now do.

Proof. First, we note that Equation (28) directly implies Equations (29a) and (29b). To see this, we note the following in order:

1. We have $\left\|T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1/2}\widehat{\boldsymbol{\Sigma}}\right\|_{\mathsf{op}} \leq \left\|T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1/2}\widehat{\boldsymbol{\Sigma}}T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1/2}\right\|_{\mathsf{op}}\left\|T_\gamma\left(\boldsymbol{\Sigma}\right)\right\|_{\mathsf{op}} \leq \left(1+\frac{\epsilon}{\gamma}\right) = \mathcal{O}(1)$. This shows that Equation (29a) holds.
2. We have $\left\|\widehat{\boldsymbol{R}}\right\|_{\mathsf{op}} \leq \left\|\widehat{\boldsymbol{\Sigma}}^{-1}\right\|_{\mathsf{op}} = \mathcal{O}(1/\gamma)$. This shows that Equation (29b) holds.

Moreover, Equation (29c) follows directly from Lemma A.2. Therefore, it suffices to prove Equation (28) here. We only provide the argument for $\widehat{\boldsymbol{\Sigma}}$ with respect to $T_\gamma\left(\boldsymbol{\Sigma}\right)$ here: an identical argument holds to bound $\widehat{\boldsymbol{\Sigma}}_1$ with respect to $T_\gamma\left(\boldsymbol{\Sigma}\right)$. First, recall that $\widehat{\boldsymbol{\Sigma}} := T_\gamma\left(\widehat{\boldsymbol{\Sigma}}_m\right)$, and so we really want to upper bound the quantity $\left\|T_\gamma\left(\widehat{\boldsymbol{\Sigma}}_m\right) - T_\gamma\left(\boldsymbol{\Sigma}\right)\right\|_{\mathsf{op}}$. It is well known (see, e.g. (Boyd et al., 2004)) that for any positive semidefinite matrix $\boldsymbol{M}$, the operator $T_\gamma\left(\boldsymbol{M}\right) - \gamma\boldsymbol{I}_d$ is a proximal operator with respect to the convex nuclear norm functional. The non-expansiveness of proximal operators then gives us

$$\begin{aligned}\left\|\widehat{\boldsymbol{\Sigma}} - T_\gamma\left(\boldsymbol{\Sigma}\right)\right\|_{\mathsf{op}} &= \left\|T_\gamma\left(\widehat{\boldsymbol{\Sigma}}_m\right) - T_\gamma\left(\boldsymbol{\Sigma}\right)\right\|_{\mathsf{op}} \\ &= \left\|T_\gamma\left(\widehat{\boldsymbol{\Sigma}}_m\right) - \gamma\boldsymbol{I}_d - \left(T_\gamma\left(\boldsymbol{\Sigma}\right) - \gamma\boldsymbol{I}_d\right)\right\|_{\mathsf{op}} \\ &\leq \left\|\widehat{\boldsymbol{\Sigma}}_m - \boldsymbol{\Sigma}\right\|_{\mathsf{op}} \leq \epsilon,\end{aligned}$$

where the last step follows by standard arguments on the concentration of the empirical covariance matrix. Thus, we have

$$\left\|T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1/2}\widehat{\boldsymbol{\Sigma}}T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1/2} - \boldsymbol{I}_d\right\|_{\mathsf{op}} \leq \left\|T_\gamma\left(\boldsymbol{\Sigma}\right)^{-1}\right\|_{\mathsf{op}} \cdot \left\|\widehat{\boldsymbol{\Sigma}} - T_\gamma\left(\boldsymbol{\Sigma}\right)\right\|_{\mathsf{op}} = \frac{1}{\gamma}\cdot\epsilon = \frac{\epsilon}{\gamma}.$$

This completes the proof. □

With the proof of Lemma C.4 complete, we have completed the proof of Lemma A.1. □

D. Technical lemmas

D.1. Proof of Lemma B.2

It is equivalent to show that $\mathbb{E}\left[\sum_{j=1}^n \boldsymbol{x}_{s_j,A_{s_j}}\boldsymbol{x}_{s_j,A_{s_j}}^\top\right] \preceq n\boldsymbol{I}_d$. We will critically use the fact that on all designated exploration rounds $1 \leq s_1 \leq \ldots \leq s_n := \tau(n)$, the distribution of the action A_s is independent of the context realizations at round s, and in particular independent of $\boldsymbol{x}_{s_j,A_{s_j}}$. Thus, we use the tower property of conditional expectations to get

$$\mathbb{E}\left[\sum_{j=1}^n \boldsymbol{x}_{s_j,A_{s_j}}\boldsymbol{x}_{s_j,A_{s_j}}^\top\right] = \mathbb{E}\left[\sum_{j=1}^n \mathbb{E}\left[\boldsymbol{x}_{s_j,A_{s_j}}\boldsymbol{x}_{s_j,A_{s_j}}^\top\middle|A_{s_j}\right]\right] \preceq \mathbb{E}\left[\sum_{j=1}^n \boldsymbol{I}_d\right] = n\boldsymbol{I}_d.$$

Above, the inequality follows because we have $\boldsymbol{\Sigma}_i \preceq \boldsymbol{I}_d$ for all $i \in [K]$. This completes the proof. On the other hand, we have $\boldsymbol{\Sigma}_{A_{s_j}} \succeq \gamma\boldsymbol{I}_d$ for all $j \in [n]$ by the definition of the data-adaptive exploration set (which maintains feature diversity).

E. Data-adaptive exploration for general contextual bandits: A partial roadmap

In this section, we discuss a partial roadmap to extend the data-adaptive exploration sub-routine in MODCB.A to general contextual bandits by leveraging ideas from function estimation under covariate-shift. Our starting point is the observation that the optimal model selection algorithm (OSOM) (Chatterji et al., 2020) that achieves Objective 1 does not require a fast estimation subroutine; only a favorable diversity condition that is purely *covariate-dependent*. We present a generalization of this arm-specific diversity condition to general function classes that is motivated by considerations in covariate shift and transfer learning. Moreover, since the data-adaptive exploration subroutine in MODCB.A tests for the presence of a feature diversity condition, its principle could be generalized as well. In particular, we could adaptively decide whether to deploy

the "greedy" action selection used by OSOM or a forced-exploration subroutine that is exclusively used to estimate the putative regret bound (Pacchiano et al., 2020b; Lee et al., 2021). Of course, such an adaptive procedure would not achieve even Objective 2 in the worst case — indeed, Objective 2 was shown to not be achievable for general function classes by any algorithm (Marinov & Zimmert, 2021). However, it opens the door to achieving Objective 1 under favorable diversity conditions when they exist — a more realistic, and still desirable goal.

E.1. From benign covariate-shift to optimal model selection

We consider the bandit-vs-contextual bandit model selection formulation, and a general function class $\mathcal{F}$ such that the reward model is given by

$$G_{i,t} = \mu_i + f^*(\boldsymbol{x}_{i,t}) + W_{i,t} \text{ for every } i \in [K] \tag{32}$$

for some (unknown) $f^* \in \mathcal{F}$ and bias terms $\{\mu_i\}_{i\in[K]}$. To preserve the property of nestedness, we assume that the zero function (i.e. $f_0(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$) is contained in the function class $\mathcal{F}$. We also assume, without loss of generality, that all functions in $\mathcal{F}$ are *centered* in the sense that $\mathbb{E}\left[f(\boldsymbol{x}_i)\right] = 0$ for all $f \in \mathcal{F}$ and all $i \in [K]$. We consider *learnable* function classes, in the sense that it is information-theoretically possible to design a contextual bandit algorithm that achieves regret $\mathcal{O}(\sqrt{\text{comp}(\mathcal{F}) \cdot T})$ with respect to the best policy induced by $f^*(\cdot)$, where comp$(\mathcal{F})$ is a standard learning-theoretic measure of function complexity[4]. Thus, Objective 1 of model selection would be to simultaneously achieve regret rates $\mathcal{O}(\sqrt{KT})$ under simple MAB structure and $\mathcal{O}(\sqrt{K \cdot \text{comp}(\mathcal{F}) \cdot T}$ more generally (i.e. under Equation (32)).

We first observe that, in the case of linear function classes, arm-specific feature diversity is a sufficient condition for reliable estimation under covariate-shift across any two arms. This connection is made precise in Example E.2 and allows us to postulate a more general feature-diversity condition. Suppose that for each arm $i \in [K]$, we have $\boldsymbol{x}_{i,t}$ i.i.d. $\sim \mathcal{D}_i$ across $t \geq 1$. Then, a favorable situation for transfer learning would allow an estimator constructed from samples under the distribution $\mathcal{D}_i$ to generalize *only upto a constant factor worse* on samples obtained from the shifted distribution $\mathcal{D}_j$; further, we would like this property to hold for any $j \neq i$. The definition below formalizes such a sufficient condition in terms of the function class $\mathcal{F}$ and data distribution tuple $(\mathcal{D}_1, \ldots, \mathcal{D}_K)$.

Definition E.1. A function class $\mathcal{F}$ is said to be covariate-agnostic with factor $C > 1$ with respect to data distributions $\mathcal{D}_1, \ldots, \mathcal{D}_K$ if, for every $i \neq j \in [K]$ and any two functions $f, f' \in \mathcal{F}$, we have

$$\mathbb{E}\left[(f(\boldsymbol{x}_j) - f'(\boldsymbol{x}_j))^2\right] \leq C \cdot \mathbb{E}\left[(f(\boldsymbol{x}_i) - f'(\boldsymbol{x}_i))^2\right]. \tag{33}$$

Here, C should be a constant that does not directly or indirectly depend on the model complexity measure comp$(\mathcal{F})$, and only depends on the data distribution tuple $(\mathcal{D}_1, \ldots, \mathcal{D}_K)$.

Intuitively, the condition of covariate-agnosticity as defined above should reduce the requirement for forced exploration, raising the possibility of achieving Objective 1 through "greedy" action selection, i.e. $A_t = i_t$ under the MAB hypothesis. This is because samples collected under the action sequence recommended by a MAB algorithm, i.e. $\{A_t = i_t\}_{t=1}^T$, could be used to estimate the putative regret bound that would be obtained through the counterfactual action sequence recommended by a CB algorithm, i.e. $\{A_t = j_t\}_{t=1}^T$, with only a constant factor inflation in estimation error. Indeed, this is the approach at the heart of the OSOM algorithm (Chatterji et al., 2020), and the reason why a feature diversity condition is required for it to work. This suggests a plausible extension of this approach to general function classes under data distributions that satisfy Equation (33). Formally establishing such an extension[5] is an interesting direction for future work.

E.2. Adaptive exploration by testing for benign covariate-shift

The property of covariate-agnosticity turns out to be equivalent to the arm-specific feature-diversity condition in the case of linear models. Moreover, it can be characterized in several examples that go beyond linear function classes. The examples provided below formally demonstrate this.

[4]For example, in the case of linear models we have comp$(\mathcal{F}) = d$ and in the case of unstructured finite function classes, we have comp$(\mathcal{F}) = \log|\mathcal{F}|$.

[5]While feature-diversity in the sense of Equation (33) is the most essential requirement, we note that there are several other non-trivial technical pieces required for such a generalization to work, most notably the ability to obtain self-normalized generalization bounds from adaptively collected samples (well-known for the linear case (Peña et al., 2008; Abbasi-Yadkori et al., 2011)) under general function classes. While this may constrain the set of function classes we can work with, we still expect such bounds to be establishable well beyond the linear case.

Example E.2 (Linear models)**.** Consider the linear function class $\mathcal{F} := \{\langle \boldsymbol{\theta}, \cdot \rangle \text{ for all } \boldsymbol{\theta} \in \Theta\}$, and suppose that $\boldsymbol{I}_d \succeq \boldsymbol{\Sigma}_i \succeq \gamma \boldsymbol{I}_d$ for all $i \in [K]$, which is precisely our definition of arm-specific feature diversity. Then, an elementary calculation shows that Equation (33) holds for the choice $C = \frac{1}{\gamma}$. Specifically, for two linear model parameters $\boldsymbol{\theta}, \boldsymbol{\theta}'$, and any pair $i \neq j \in [K]$, we have

$$\begin{aligned}
\mathbb{E}\left[\langle \boldsymbol{x}_j, \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle^2\right] &= \|\boldsymbol{\Sigma}_j^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}')\|_2^2 \\
&= \|\boldsymbol{\Sigma}_j^{1/2}\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{\Sigma}_i^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}')\|_2^2 \\
&\leq \|\boldsymbol{\Sigma}_j^{1/2}\boldsymbol{\Sigma}_i^{-1/2}\|_{\mathsf{op}}^2 \cdot \mathbb{E}\left[\langle \boldsymbol{x}_i, \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle^2\right] \\
&\leq \frac{1}{\gamma} \cdot \mathbb{E}\left[\langle \boldsymbol{x}_i, \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle^2\right],
\end{aligned}$$

where the last inequality follows because we have $\boldsymbol{\Sigma}_i \succeq \gamma \boldsymbol{I}_d$ and $\boldsymbol{\Sigma}_j \preceq \boldsymbol{I}_d$. Moreover, equality is achieved when $\boldsymbol{\Sigma}_j = \boldsymbol{I}_d$ and $\gamma_{\min}(\boldsymbol{\Sigma}_i) = \gamma$.

Example E.3 (Single-index models and high-dimensional data)**.** The single-index model class can be modeled through functions of the form $f(\boldsymbol{x}) = \sum_{m=1}^{p} a_m \langle \boldsymbol{\theta}^*, \boldsymbol{x} \rangle^p$, where $\{a_m\}_{m=1}^p$ may be known or unknown (Dudeja & Hsu, 2018). We consider the special case under which p is a constant with respect to d. We assume that, for each $i \in [K]$, the distribution $\mathcal{D}_i$ on context $\boldsymbol{x}_i$ is sub-Gaussian with parameter at most $\boldsymbol{\Sigma}_i$, and further satisfies the small-ball property (Mendelson, 2014), i.e. that there exists a universal positive constant $u > 0$ (that does not depend on d or p) such that

$$\Pr\left[|\langle \boldsymbol{\Delta}, \boldsymbol{x}_i \rangle| \geq u \cdot \|\boldsymbol{\Sigma}_i^{1/2}\boldsymbol{\Delta}\|_2\right] \geq \frac{1}{2} \tag{34}$$

for any vector $\boldsymbol{\Delta}$. Finally, we assume that the arm-specific feature diversity condition holds, i.e. $\boldsymbol{\Sigma}_i \succeq \gamma \boldsymbol{I}_d$. Then, we can verify that Equation (33) holds for the choice $C = \left(\frac{1}{u\gamma}\right)^p$. (It suffices to consider the contribution from the highest-order terms (i.e. $a_m = 0$ for all $m < p$), and so we have

$$\begin{aligned}
\mathbb{E}\left[(\langle \boldsymbol{x}_j, \boldsymbol{\theta} \rangle^p - \langle \boldsymbol{x}_j, \boldsymbol{\theta}' \rangle^p)^2\right] &\leq \|\boldsymbol{\Sigma}_j^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}')\|_2^{2p} \\
&= \|\boldsymbol{\Sigma}_j^{1/2}\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{\Sigma}_i^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}')\|_2^{2p} \\
&\leq \frac{1}{\gamma^p} \cdot \|\boldsymbol{\Sigma}_i^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}')\|_2^{2p} \\
&\leq \left(\frac{1}{u\gamma}\right)^p \cdot \mathbb{E}\left[(\langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle^p - \langle \boldsymbol{x}_i, \boldsymbol{\theta}' \rangle^p)^2\right]
\end{aligned}$$

where the first inequality follows from sub-Gaussianity of $\mathcal{D}_j$ and the last inequality follows from the small-ball assumption on $\mathcal{D}_i$.)

Example E.4 (Nonparametric classes and bounded density ratios)**.** Finally, we consider $\mathcal{F}$ to be a general nonparametric function class under which consistent estimation is possible, such as the set of all Holder-smooth functions (Nemirovski, 2000; Audibert & Tsybakov, 2007). For all $i \neq j \in [K]$, we further assume $\mathcal{D}_i$ to be absolutely continuous with respect to $\mathcal{D}_j$, and that $\sup_{\boldsymbol{x} \in \mathbb{R}^d} \frac{p_j(\boldsymbol{x})}{p_i(\boldsymbol{x})} \leq C_d$ (where C_d may depend on the data dimension d). Then, a special case of the results in (Mansour et al., 2012), that express transfer error as a function of the Renyi divergence between distributions $\mathcal{D}_i$ and $\mathcal{D}_j$, shows[6] that Equation (33) holds for $C := \log(C_d)$. For low-dimensional data (i.e $d \ll \text{comp}(\mathcal{F})$), the constant $\log(C_d)$ will be independent of the model complexity $\text{comp}(\mathcal{F})$.

Examples E.2 and E.3 demonstrate that arm-specific feature diversity can constitute a sufficient test statistic for covariate-agnosticity (and the possibility of achieving Objective 1) in both linear and nonlinear models. For these function classes, the data-adaptive exploration subroutine defined in Algorithm 3 could be directly plugged and played with alternative model selection algorithms designed for general function classes (Lee et al., 2021; Pacchiano et al., 2020b). More generally, the multiplicative factor C in Equation (33) can be characterized for very general function classes and data distributions (Kpotufe & Martinet, 2021; Hanneke & Kpotufe, 2020). Moreover, since C is a functional of the data distribution tuple $(\mathcal{D}_1, \ldots, \mathcal{D}_K)$,

[6] It is also possible to obtain transfer exponents for more general distributions, e.g. even when $\mathcal{D}_i$ is not absolutely continuous with respect to $\mathcal{D}_j$ (Kpotufe & Martinet, 2021), but the resulting characterization of transfer error is significantly worse than Equation (33) and may therefore preclude optimal model selection.

it can also be estimated from the unlabeled iid covariates $\{\boldsymbol{x}_{i,t}\}_{i\in[K],t\geq 1}$. In the case of absolutely continuous distributions, Example E.4 shows that C can be characterized through a uniform upper bound on density ratios; indeed, density ratio estimation is an extensively studied topic in its own right (Kpotufe, 2017; Lin et al., 2021; Yu et al., 2021). This raises the possibility of leveraging independent advances in transfer coefficient estimation to create new data-adaptive exploration schedules for model selection among general function classes.