Model Selection in Batch Policy Optimization

Jonathan N. Lee [1 2] **George Tucker** [2] **Ofir Nachum** [2] **Bo Dai** [2]

Abstract

We study the problem of model selection in batch policy optimization: given a fixed, partial-feedback dataset and M model classes, learn a policy with performance that is competitive with the policy derived from the best model class. We formalize the problem in the contextual bandit setting with linear model classes by identifying three sources of error that any model selection algorithm should optimally trade off in order to be competitive: (1) approximation error, (2) statistical complexity, and (3) coverage. The first two sources are common in model selection for supervised learning, where optimally trading off these two is well-studied. In contrast, the third source is unique to batch policy optimization and is due to dataset shift inherent to the setting. We first show that no batch policy optimization algorithm can achieve a guarantee addressing all three simultaneously, revealing a stark contrast between difficulties in batch policy optimization and the positive results available in supervised learning. Despite this negative result, we show that relaxing any one of the three error sources enables the design of algorithms achieving near-oracle inequalities for the remaining two. We conclude with experiments demonstrating the efficacy of these algorithms.

1. Introduction

Model selection and hyperparameter tuning are fundamental tasks in supervised learning and statistical learning theory. In these settings, given a model class, the standard goal is to minimize risk, which can always be decomposed as a sum of approximation error (i.e., bias) and estimation error (i.e., variance) of the model class. A trade-off between these two often exists: a large model class may require a lot of data for generalization while a small one may suffer from large approximation error. At its core, model selection describes the problem of choosing a model class so as to automatically balance these quantities. A vast literature exists on algorithms and selection rules – such as hold-out methods, cross-validation, and structural risk minimization – that nearly optimally achieve this trade-off in supervised learning (Massart, 2007; Lugosi & Nobel, 1999; Bartlett et al., 2002; Bartlett, 2008). That is, one can select a model class from a collection that nearly matches the performance of the best model class. The implications in practice have been equally, if not more, impactful as evidenced by the widespread use of model validation and selection in machine learning applications, where methods like cross-validation on held-out data are standard and essential steps for practitioners.

In recent years, interest has turned to model selection in online bandits and reinforcement learning (Agarwal et al., 2017; Foster et al., 2019; Pacchiano et al., 2020; Lee et al., 2021; Modi et al., 2020; Chatterji et al., 2020; Muthukumar & Krishnamurthy, 2021; Ghosh et al., 2021). However, in contrast to the extensive understanding of model selection in supervised learning and the growing literature in online learning, relatively little is known about model selection in the context of *batch* (or offline) bandits and reinforcement learning. Batch policy optimization, or offline policy optimization, is a promising paradigm for learning decision-making policies by leveraging large datasets of interactions with the environment (Lange et al., 2012; Levine et al., 2020). The goal is for the learner to find a good policy without interacting with the environment in an online fashion so as to avoid dangerous or costly deployments of sub-optimal policies. While a number of works provide theoretically sample-efficient algorithms for batch policy optimization (Munos & Szepesvári, 2008; Jin et al., 2019; Nachum et al., 2019b; Liu et al., 2020; Xie et al., 2021), their effectiveness in practice has been limited due to the lack of tools for validation and selection when faced with multiple options for model classes or hyperparameter settings.

It is clear that a need exists in batch policy optimization for an analogue to methods like cross-validation in supervised learning. Traditionally, this need has motivated a large body of literature dedicated to the problem of batch policy *evaluation* (e.g., (Precup, 2000; Jiang & Li, 2016; Nachum et al.,

[1] Department of Computer Science, Stanford University, USA [2] Google Research, Mountain View, USA. Correspondence to: Jonathan N. Lee <jnl@stanford.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

2019a)). These works aim to estimate the values (i.e., online performance) of a set of arbitrary candidate policies, typically via value function regression on the fixed batch dataset or some form of importance sampling. Unfortunately, in practice these methods warrant their own hyperparameter selection (e.g. choice of model class used to estimate value functions), an observation that has been noted by several authors (Tang & Wiens, 2021; Kumar et al., 2021; Paine et al., 2020). The policy evaluator could thus suffer from the same issues as batch policy optimization algorithms. The question of how to achieve effective model selection methods *in the batch setting* thus remains open.

Motivated by this challenge, in this paper we formalize and study theoretically the problem of model selection for batch policy optimization in the setting of contextual bandits with linear models and make progress towards understanding what is and is not possible in the batch setting compared to supervised learning. The problem is as follows. The learner is given access to a collection of model classes $\mathcal{F}_1, \dots \mathcal{F}_M$ in order to estimate value functions. Equipped with a single model class $\mathcal{F}_k$, a base algorithm produces a policy $\hat{\pi}_k$. The learner's goal is thus to leverage all M classes to produce a policy $\hat{\pi}$ that nearly matches the performance of the best $\hat{\pi}_k$. That is, the learner should perform as if the "optimal" model class $\mathcal{F}_{k_*}$ that produces the best $\hat{\pi}_{k_*}$ were known in advance. To ground our results, we focus on the contextual bandit setting and model classes $\mathcal{F}_k$ that are linear with respect to a collection of known feature maps.

1.1. Contributions

Three Model Selection Criteria In Section 3, we identify three sources of error that a model selection algorithm should trade off in order to be competitive with $\hat{\pi}_{k_*}$ for linear model classes. Two natural sources, borrowed from supervised learning bounds, are **approximation error** and statistical **complexity** of the model class. However, unlike supervised learning, a third source that contributes to error is the **coverage** of the fixed dataset in conjunction with the properties of the model class. This can be interpreted as the effect of dataset shift. The fixed dataset may not sufficiently cover the relevant states and actions[1] and some model classes may be better equipped to handle this. Finally, we aim to ensure that model selection preserves the property that the learned policy competes well against any well-covered comparator policy (Jin et al., 2021; Zanette et al., 2021; Xie et al., 2021).

Hardness of Model Selection Our first technical contribution (Theorem 2) shows that it is provably impossible to perform model selection so as to optimally trade off all three error sources. This is perhaps surprising for two reasons. Firstly, in supervised learning and general risk minimization, such oracle inequalities that nearly optimally balance approximation error and statistical complexity *are* achievable through myriad procedures, and a vast literature exists on this topic. Secondly, recent work by Su et al. (2020) shows that the analogous problem of estimator selection for batch policy *evaluation* is possible up to an inexact oracle inequality. These observations suggest that the difficulty of model selection is a unique characteristic of the batch policy *selection/optimization* problem. Furthermore, since the negative result applies to the more restrictive setting of linear contextual bandits, it automatically implies the same for broader classes of problems such as general function approximators and multi-step reinforcement learning.

Positive Results Despite this negative result, we show in Section 5 that positive results for model selection are possible in some instances. Namely, as long as just one of the three error sources is ignored, there exist algorithms to achieve an oracle inequality that optimally trades off the remaining two. We provide experimental results demonstrating the effectiveness of these algorithms.

1.2. Related Work

Pessimistic Policy Optimization The principle of pessimism in batch policy optimization has recently attracted great interest as both a heuristic and principled method that performs (provably) well on the given data distribution in order combat the dataset shift problem. Work by Jin et al. (2021); Xiao et al. (2021); Xie et al. (2021); Zanette et al. (2021); Uehara & Sun (2021) has sought to theoretically quantify the benefit of pessimistic methods for batch learning with different coverage conditions. In particular, Jin et al. (2021) show it is possible to recover regret bounds that are stronger when compared to policies that are well-covered by the data. However, these theoretical studies typically assume a single model class and either assume realizability (Jin et al., 2021; Uehara & Sun, 2021) or require that the approximation error is known in advance (Xie et al., 2021), which is often not the case. The goal of our paper is to investigate and make progress on these unaddressed issues when multiple model classes are available with unknown approximation error.

Policy Evaluation and Selection Another related line of work has considered the problem of *batch policy evaluation* where one seeks to estimate the value of a target policy using a fixed dataset (Duan et al., 2020). Often, the end goal is to evaluate a number of policies and select the one with maximal value (Yang et al., 2020). In principle, one could use such an evaluation method to select the estimated best policy from several candidates, which are generated

[1] In contextual bandits, one need only worry about action distribution mismatch.

from some batch policy optimization method using different model classes. However, one of the main drawbacks of these policy evaluation methods is that they are subject to their own hyperparameters and modeling errors which may compromise the final model selection regret bound. For example, to ensure accurate policy *evaluation*, one might be tempted to use a large model class, but the resulting estimation error can easily leak into the regret bound if only a small model class is sufficient, rendering any model selection guarantee infeasible. Tang & Wiens (2021); Kumar et al. (2021); Paine et al. (2020); Zhang & Jiang (2021) have previously pointed out this problem and attempted to address it with practical solutions, but theoretical guarantees are less understood. This is precisely what we wish to understand in the current paper.

Farahmand & Szepesvári (2011) studied the problem of selecting action-value functions for reinforcement learning, but also assumed access to an estimator, which may suffer from the aforementioned problems. Xie & Jiang (2021) also considered selecting action-value functions for RL as an application of their algorithm, but did not consider the end-to-end model selection problem with coverage and the resulting guarantee is not competitive with an oracle. Representation learning (Agarwal et al., 2020; Papini et al., 2021; Zhang et al., 2021) is also related, but most current work is in the online setting with assumed realizability or fixed statistical complexity. Most similar in setting to our work is that of Su et al. (2020) who show that it is possible to select from a collection of batch policy *evaluators* in a way that optimally trades off approximation error and estimation error up to constants, as long as the estimators are properly ordered. However, they did not consider the *selection* problem, where the best policy will be selected among the candidates that are evaluated upon given data. We show in Section 4 that this additional task raises complications since different target policies may yield different coverage properties, but these complications can be overcome if the problem is slightly relaxed.

2. Preliminaries

Notation For $n \in \mathbb{N}$, we define the set $[n] := \{1, \dots, n\}$. Let $S \in \mathbb{R}^{d \times d}$ be a positive semi-definite matrix and $x \in \mathbb{R}^d$ be a vector. By default, $\|x\|$ denotes the ℓ_2-norm of x and $\|S\|$ denotes the spectral norm of S. $\|x\|_S = \sqrt{x^\top S x}$ is the Mahalanobis norm. $\| \cdot \|_{\psi_2}$ and $\| \cdot \|_{\psi_1}$ denote the sub-Gaussian and sub-exponential norms, respectively (see Appendix A for precise definitions). We use $C, C_1, C_2, \dots$ to denote absolute constants that do not depend on problem-relevant parameters. We use $\delta > 0$ to denote a desired failure probability and assume $\delta \leq 1/e$. We write $a \lesssim b$ to mean there exists a constant $C > 0$ such that $a \leq Cb$.

2.1. Contextual Bandits

We consider the contextual bandit setting (Lattimore & Szepesvári, 2018) with state space $\mathcal{X}$, action space $\mathcal{A}$, and a fixed distribution $\mathcal{D}$ over $\mathcal{X} \times \mathbb{R}^{\mathcal{A}}$. A learner interacts with the environment through the following protocol: the environment samples a pair $(X, Y) \sim \mathcal{D}$ with $X \in \mathcal{X}$ and $Y \in \mathbb{R}^{\mathcal{A}}$. The learner observes X and then commits to an action $a \in \mathcal{A}$. A reward $Y(a)$ is incurred and subsequently only the value of $Y(a)$ is revealed to the learner. The learner's objective is to determine a policy $\hat{\pi} : \mathcal{X} \to \Delta_{\mathcal{A}}$, which maps states to distributions over actions, such that the expected reward $\mathbb{E}_{\hat{\pi}} Y(A)$ is large. Here $\mathbb{E}_{\hat{\pi}}$ denotes the expectation over state-action-rewards induced by $\hat{\pi}$ with $A \sim \hat{\pi}(\cdot|X)$. Similarly, we use $P_{\hat{\pi}}$ to denote the probability measure under $\hat{\pi}$. Given a state x and action a, we write the expected reward function as $f(x, a) = \mathbb{E}\left[Y(a) \mid x\right]$. We assume $f(x, a) \in [-1, 1]$ and the noise terms $\eta(a) = Y(a) - f(X, a)$ are independent across actions and sub-Gaussian with $\|\eta(a)\|_{\psi_2} \leq 1$. We use $\mathbb{E}_X$ and P_X to denote the expectation and measure over just the marginal distribution over states X.

2.2. Batch Learning

As mentioned before, we consider the batch setting for policy optimization (Lange et al., 2012; Levine et al., 2020; Xiao et al., 2021). Rather than learning via direct interaction with the environment, the learner is given access to a fixed dataset $D = \{x_i, a_i, y_i\}_{i \in [n]}$ of $n \in \mathbb{N}$ prior interactions where $\bar{y}_i$ is drawn from $\mathcal{D}$ conditioned on x_i and $y_i = \bar{y}_i(a_i)$ as in the aforementioned interface. Similar to the potential outcomes framework under unconfoundedness (Lin, 2013; Imbens & Rubin, 2015), we assume that $a_i \perp (y_j(a))_{j \in [n], a \in \mathcal{A}} \mid x_i$. Intuitively, this ensures the process that selects actions does not peek at the outcomes directly or through a confounding variable. For example, this could be collected from a behavior policy. From this data, the learner produces a policy $\hat{\pi}$ to minimize regret with respect to a comparator policy π[2]:

$$\mathrm{Reg}(\pi, \hat{\pi}) = \mathbb{E}_X \left[f(X, \pi(X)) - f(X, \hat{\pi}(X)) \right].$$

The comparator π can be any deterministic policy, including the optimal policy. The typical regret is obtained by maximizing over π: $\mathrm{Reg}(\hat{\pi}) := \max_\pi \mathrm{Reg}(\pi, \hat{\pi})$. Like recent work (Jin et al., 2021; Zanette et al., 2021; Uehara & Sun, 2021; Xie et al., 2021), the motivation for this flexibility is that the globally optimal policy may not be well-covered in the dataset, and in these cases, we are interested in proving regret bounds that compete well against comparators that are nearly optimal but that are also well-covered by D.

[2] For deterministic policies, we write $\pi(x) \in \mathcal{A}$ to denote the action on which all the probability mass lies given the state $x \in \mathcal{X}$. We assume all comparator and learned policies are deterministic but all our results can be easily extended.

Algorithm 1 Pessimistic Linear Learner

```
Input: Dataset $D$, linear model class $\mathcal{F}$, confidence parameter $0 < \delta \leq 1/e$, regularization parameter $\lambda > 0$.
Set $V \leftarrow \frac{\lambda}{n}\mathbb{I}_d + \frac{1}{n}\sum_{i\in[n]}\phi(x_i,a_i)\phi(x_i,a_i)^\top$
Set $\hat{\theta} \leftarrow V^{-1}\left(\frac{1}{n}\sum_{i\in[n]}\phi(x_i,a_i)y_i\right)$
for $x \in \mathcal{X}$ do
    $\hat{f}(x,a) \leftarrow \left\langle \phi(x,a), \hat{\theta} \right\rangle - \beta_{\lambda,\delta}(n,d)\cdot\|\phi(x,a)\|_{V^{-1}}$
    Set $\hat{\pi}(x) \leftarrow \arg\max_{a\in\mathcal{A}} \hat{f}(x,a)$ with ties broken arbitrarily
end for
Return: $\hat{\pi}$
```

In the batch setting, it is often assumed that, for the data in D, states are generated as $x_i \sim \mathcal{D}$ marginally and a_i comes from a (potentially unknown) fixed stochastic behavior policy μ conditioned on x_i (Levine et al., 2020). However, our aforementioned assumption about D throughout most of the paper is more general as it does not require that x_i or a_i come from a random distribution at all such as in a fixed design setting. Nevertheless, we will eventually address this behavior policy setting in Section 5, so we will briefly introduce context and a definition. It is common to assume that μ sufficiently covers the state-action space, often by means of a concentrability coefficient which captures the worst-case ratio of the density under an arbitrary policy π and μ (Munos & Szepesvári, 2008).

Definition 1. *The concentrability coefficient with respect to data collection policy μ is defined as $\mathcal{C}(\mu) := \sup_{\pi,x,a}\pi(a|x)/\mu(a|x)$.*

It should be noted that the assumption that such concentrability coefficients are small can be very strong. When using function approximation, it may not be necessary to require that μ have non-zero density everywhere as long as one can still sample-efficiently find a good fit on the given data distribution. Exploiting the properties of function approximation can thus overcome many coverage issues.

3. Model Selection for Batch Linear Bandit

In this section, we introduce *linear* model classes for the contextual bandit problem (Chu et al., 2011; Abbasi-Yadkori et al., 2011; Jin et al., 2019; 2021) and present a corresponding batch regret bound for a single model class. We identify three sources of sub-optimality in a resulting regret bound and then formally state the goal of model selection to balance these three sources.

3.1. Batch Regret for a Single Model Class

Let $\mathcal{F} \subseteq (\mathcal{X}\times\mathcal{A}\to\mathbb{R})$ be a model class defined by a *known* d-dimensional feature mapping $\phi : \mathcal{X}\times\mathcal{A}\to\mathbb{R}^d$ such that

$$\mathcal{F} := \left\{(x,a)\mapsto\langle\phi(x,a),\theta\rangle \ : \ \theta\in\mathbb{R}^d\right\}. \qquad (1)$$

Leveraging the batch dataset D, Algorithm 1 (which is standard and essentially a restatement of that of Jin et al. (2021)) performs ridge regression with regularizer λ/n on the rewards observed in D and returns a policy $\hat{\pi}$ that conservatively chooses actions based on both the best-fit estimator and a penalty term determined by the coverage. The penalty term is modulated by a coefficient that we set to be equal to

$$\beta_{\lambda,\delta}(n,d) = \sqrt{\tfrac{\lambda d}{n}} + \sqrt{\tfrac{5d+10d^{1/2}\log^{1/2}(1/\delta)+10\log(1/\delta)}{n}}.$$

In general $\mathcal{F}$ may not actually contain the optimal regressor f that defines the true reward function. In such cases, we say that $\mathcal{F}$ may suffer from misspecification or approximation error. It is thus important to ensure that the sub-optimality of the extracted policy scales gracefully with the approximation error of the model, even when the approximation error is not known. The below result, which to the best of our knowledge has not been explicitly stated in the literature, follows as a simple application of a standard regression analysis (e.g. Hsu et al. (2012b, Section 3.1)) to handle concentration and the penalized action-selection method akin to that of Jin et al. (2021).

Theorem 1. *Let $\hat{\pi}$ be the output policy of Algorithm 1 with $\lambda > 0$. Define,*

$$\begin{aligned}\epsilon(\pi,\hat{\pi}) = \ &\mathbb{E}_X\left[|f(X,\pi(X)) - \langle\phi(X,\pi(X)),\theta_*\rangle\ |\right] \\ &+ \mathbb{E}_X\left[|\ \langle\phi(X,\hat{\pi}(X)),\theta_*\rangle - f(X,\hat{\pi}(X))|\right]\end{aligned}$$

where $\theta_* \in \arg\min_{\theta\in\mathbb{R}^d}\sum_{i\in[n]}\left(\phi(x_i,a_i)^\top\theta - f(x_i,a_i)\right)^2$. If $\|\theta_*\| \leq \sqrt{d}$, then with probability at least $1-\delta$, for any policy π (including the optimal policy), Reg$(\pi,\hat{\pi})$ is bounded above by

$$\begin{aligned}&\underbrace{\epsilon(\pi,\hat{\pi})}_{\text{approx. error}} + \underbrace{2\beta_{\lambda,\delta}(n,d)}_{\text{complexity}}\cdot\underbrace{\mathbb{E}_X\|\phi(X,\pi(X))\|_{V^{-1}}}_{\text{coverage}} \\ &= \widetilde{\mathcal{O}}\left(\epsilon(\pi,\hat{\pi}) + \sqrt{\frac{d}{n}}\cdot\mathbb{E}_X\|\phi(X,\pi(X))\|_{V^{-1}}\right)\end{aligned}$$

where V is the regularized empirical covariance matrix of the data, defined in Algorithm 1[3].

The theorem reveals that sub-optimality in the regret bound is due primarily to three sources:

[3] Throughout the paper, we use $\widetilde{\mathcal{O}}$ to omit polylog factors of problem-dependent parameters such as δ^{-1}, dimension d, and number of model classes M (defined in Section 3.2).

1. **Approximation error**: this represents how far the closest function in $\mathcal{F}$ is from representing the true reward function f. Here, "closest" means the solution, θ_*, to the fixed design regression problem. When $f(x,a) = \phi(x,a)^\top \theta_*$[4], we say the data is realizable and clearly $\epsilon(\pi, \hat{\pi}) = 0$. We discuss approximation error and its various forms in more detail in Appendix B.
2. **Statistical complexity**: this represents the learnability size of $\mathcal{F}$. In the linear case, this is concisely encapsulated by the dimension d of the feature map ϕ. From the definition of β, we have $\beta_{\lambda,\delta}(n,d) = \widetilde{\mathcal{O}}(\sqrt{d/n})$
3. **Coverage properties**: this source plays an important role in the batch learning setting where sub-optimality can be due to the dataset shift between D and the learned policy $\hat{\pi}$. In the linear setting, this is represented as the mismatch between the covariance matrix V and the feature distribution of the comparator policy π. That is, if directions of features that π visits do not coincide with directions covered in V, then one should expect the error due to mismatch to be large. One critical factor due to pessimism is that we need only consider the mismatch with the comparator π (e.g. the optimal policy) as opposed to a worst-case policy or the maximum eigenvalue of V^{-1} or a concentrability coefficient, all of which could lead to a substantially larger regret bound. This ensures that $\hat{\pi}$ is competitive against all well-covered comparator policies π under the dataset D, simultaneously (Jin et al., 2021).

We note that θ_* depends on the states and actions in the dataset D. Several additional remarks are in order.

Remark 1. *The bound in Theorem 1 can be viewed as a data-dependent bound (dependent on the state-action pairs in the training dataset D), which is desirable for two main reasons. The first is that it is consistent and easily comparable with the prior work of Jin et al. (2021); Zanette et al. (2021) who proved similar bounds (albeit under the assumption of realizability). The approximation error $\epsilon(\pi, \hat{\pi})$ can be naturally viewed as all of the sub-optimality not accounted for in the usual estimation error, i.e. it recovers the prior work in the realizable case with $\epsilon(\pi, \hat{\pi}) = 0$ since θ_* can be set to the realizing value. The second reason is that, as we will see, the data-dependent bound is convenient for model selection since we can evaluate it directly and it avoids any distributional assumptions and quantities until absolutely necessary (Duan et al., 2020; Xie et al., 2021). In statistical learning settings, data-dependent generalization bounds are desirable precisely for the purpose of model selection. They also tend to be tighter than worst-case counterparts (Antos et al., 2002; Bartlett et al., 2002).*

Remark 2. *In general, one does not know the approximation error $\epsilon(\pi, \hat{\pi})$ or even non-trivial upper bounds on it, which is the initial motivation for model selection both here and in supervised learning. We work with this particular quantity since it is naturally one of the tightest. Alternatives can be easily derived with some work but there is little reason to artificially inflate the bound. See Appendix B for further discussion.*

3.2. Model Selection Objectives

With these sources of error in mind, we now introduce the general model selection problem for batch policy optimization. We assume a collection of M linear model classes $\mathcal{F}_1, \dots, \mathcal{F}_M$, such that, for $k \in [M]$, $\mathcal{F}_k = \left\{(x,a) \mapsto \langle \phi_k(x,a), \theta \rangle \ : \ \theta \in \mathbb{R}^{d_k}\right\}$ where ϕ_k is a known d_k-dimensional feature map for model class $\mathcal{F}_k$. We desire an algorithm with the following guarantee: given an input dataset D of n interactions and model classes $\mathcal{F}_1, \dots, \mathcal{F}_M$, the algorithm outputs a policy $\hat{\pi}$ such that, with probability at least $1-\delta$, $\mathrm{Reg}(\pi, \hat{\pi})$ is bounded above by

$$\widetilde{\mathcal{O}}\left(\min_{k\in[M]}\left\{\epsilon_k(\pi,\hat{\pi}) + \sqrt{\tfrac{d_k}{n}} \cdot \mathbb{E}_X \|\phi_k(X,\pi(X))\|_{V_k^{-1}}\right\}\right) \tag{2}$$

for all deterministic policies π. Here, ϵ_k and V_k are the corresponding approximation error and regularized empirical covariance matrix for class k, as defined in the previous section for the single model class.

Interpretation In words, the main goal of model selection is to achieve performance that is nearly as good as the performance that could be achieved had the optimal model class been known in advance. Observe that this desired bound is essentially the best single model class guarantee from Theorem 1 applied to each of the M model classes. Such an inequality is often referred to as an *oracle inequality* because an oracle with knowledge of the best class could simply choose it. We emphasize that achieving the desired bound in (2) requires careful balancing of all three error sources (approximation, complexity, coverage). Importantly, note that we aim to maintain the property that $\hat{\pi}$ is competitive against any well-covered comparator π. This stands in stark contrast to oracle inequalities in supervised learning, which typically require only balancing approximation error and statistical complexity.

4. A Negative Result for Model Selection

With the main goal of model selection in the batch problem having been introduced, we present our first major contribution, which establishes a fundamental hardness of the model selection problem in the batch policy optimization setting. We show that, unlike standard learning problems, it is actually *impossible* to optimally trade off all three error sources that comprise the oracle inequality in (2).

[4] There could be multiple elements in the arg min, but this is purely analytical so we can set θ_* correctly in the realizable case.

Before arriving at the theorem, we identify a condition to impose additional structure on the model classes $\mathcal{F}_1, \dots, \mathcal{F}_M$ so as to actually make model selection *easier*. Otherwise, without structure, the selection problem is trivially impossible. In particular, we consider the setting where the model classes are *nested.*

Definition 2. *The collection of linear model classes $\mathcal{F}_1, \dots, \mathcal{F}_M$ with respective feature maps $\phi_1, \dots, \phi_M$ is said to be nested if, for each map ϕ_{k+1}, the first d_k coordinates of ϕ_{k+1} are the same as ϕ_k for all $k \in [M-1]$.*

The nestedness condition imposes structure sufficient for adaptive policy value *estimation* via a variant of the algorithm by Su et al. (2020). Nestedness effectively requires that the model classes are ordered by complexity. Nonetheless, for policy *optimization*, we will see the additional structure is still insufficient.

We denote A as a model selection algorithm which takes as input the nested model classes $\mathcal{F}_1, \dots, \mathcal{F}_M$ and a dataset D of n interactions and outputs a learned policy $\hat{\pi} = A(\mathcal{F}_{1:M}, D)$. The following theorem states that even for such nested model classes, near-optimal model selection is impossible, and performance can be arbitrarily worse in fact.

Theorem 2. *Let $\mathcal{F}_1, \dots, \mathcal{F}_M$ be a particular nested collection of linear model classes (Definition 2). For any $\alpha > 0$, there is $n = \Theta(\alpha^2)$ such that for any algorithm A there is a contextual bandit instance with comparator π and dataset D with n interactions consistent with $\mathcal{D}$ that satisfies*

$$\frac{\mathbb{E}_D\left[\textit{Reg}(\pi, A(\mathcal{F}_{1:M}, D))\right]}{\min_k \left\{ \epsilon_k(\pi, \hat{\pi}) + \sqrt{\frac{d_k}{n}} \cdot \mathbb{E}_X \|\phi_k(X, \pi(X))\|_{V_k^{-1}} \right\}} \geq \alpha.$$

Here, $\mathbb{E}_D$ denotes the expectation with respect to the randomness of the observed rewards in the dataset that are distributed according to the bandit instance $\mathcal{D}$ (conditioned on the states and actions). The result follows by reducing the problem to a batch multi-armed bandit problem and designing the appropriate nested model classes that ensure the oracle inequality is much better than what is achievable by any algorithm on the bandit problem. The dataset is constructed in order to be sufficiently imbalanced while still being true to the underlying conditional distribution induced by $\mathcal{D}$. The construction is illustrative of the core problems that lead to the oracle achieving very small regret. We remark that the result does not necessarily apply to all collections of model classes; it relies on existence of such a nested collection. See Appendix C for a detailed proof.

The theorem shows that in general, the oracle (encapsulated by the denominator), which picks the best regret bound induced by the model classes from Theorem 1, can be made to achieve regret that is arbitrarily better than the regret of any model selection algorithm for a sufficiently imbalanced dataset. The result suggests that the desired bound of (2) is too ambitious and it highlights a separation of difficulty between model selection in the batch policy optimization setting (where there is an additional error – coverage – involved in the oracle inequality) and standard statistical learning.

Note that because this is a hardness result in the restricted class of linear contextual bandits, the implication applies much more broadly to policy optimization, for example with general function classes and multi-step reinforcement learning. Interestingly, the nestedness condition should make positive results for model selection easier (by restricting the set of problem instances that an algorithm must deal with), yet the negative result still holds under this condition. This suggests that policy *optimization* is deceptively hard compared to policy *evaluation* where nestedness is indeed sufficient (Su et al., 2020).

We remark briefly that the lower bound by α cannot be trivially due to e.g. polylogarithmic factors in n that may be omitted from the denominator. This is because n can be chosen to be quadratic in α, not exponential. The lower bound, as shown explicitly in the proof of Theorem 2, is indeed due to an imbalanced dataset that causes worse coverage error in the numerator.

5. Positive Results for Special Cases

In the last section, we showed that attempting to optimally trade off all three error sources – (1) approximation error, (2) complexity, and (3) the coverage property – is not possible in general for any model selection algorithm and thus we cannot hope to make progress without further assumptions. The question remains of whether there are other settings sufficient for model selection. In this section, we explore relaxations of the model selection objective. Rather than requiring all three error sources be addressed, we aim to trade off two at a time. We will show that non-trivial model selection guarantees are indeed possible in certain settings.

5.1. Balancing complexity and coverage

We consider the problem of minimizing regret when the approximation error is zero or we are willing to ignore its contribution to the regret. For example, when we have multiple feature representations $\{\phi_k\}$ that all satisfy realizability, but some may handle coverage better or induce more favorable distributions under the behavior policy μ.

Algorithm 2 shows a simple selection rule for this case. The main idea is that we will use each model class $\mathcal{F}_k$ to generate an estimate $\hat{\theta}_k$ and covariance matrix V_k. Then, when extracting the policy, actions are chosen pessimistically, but the action that achieves the highest pessimistic value among all classes is selected. The following theorem shows that this procedure optimally trades off complexity and coverage.

Algorithm 2 Complexity-Coverage Selection

```
Input: Dataset $D$, linear model classes $\mathcal{F}_1, \dots, \mathcal{F}_M$, confidence parameter $\delta > 0$, regularization parameter $\lambda > 0$.
for $k \in [M]$ do
    $V_k \leftarrow \frac{\lambda}{n}\mathbb{I}_d + \frac{1}{n}\sum_{i\in[n]} \phi_k(x_i, a_i)\phi_k(x_i, a_i)^\top$.
    $\hat{\theta}_k \leftarrow V_k^{-1}\left(\frac{1}{n}\sum_{i\in[n]} \phi_k(x_i, a_i) y_i\right)$
end for
for $x \in \mathcal{X}$ do
    $\hat{f}_k(x, a) \leftarrow \langle \phi_k(x, a), \hat{\theta}_k \rangle - \beta_{\lambda,\delta}(n, d_k)\|\phi_k(x, a)\|_{V_k^{-1}}$
    Set $\hat{\pi}(x), \hat{k}(x) \leftarrow \arg\max_{a\in\mathcal{A}, k\in[M]} \hat{f}_k(x, a)$ with ties broken arbitrarily
end for
Return: $\hat{\pi}$
```

Theorem 3. *Given arbitrary linear model classes $\mathcal{F}_1, \dots, \mathcal{F}_M$, Algorithm 2 outputs a policy $\hat{\pi}$ such that, with probability at least $1 - \delta$, for any comparator policy π, the regret Reg$(\pi, \hat{\pi})$ is bounded above by*

$$\min_{k\in[M]} \left\{2\beta_{\lambda,\delta/M}(n, d_k) \cdot \mathbb{E}_X \|\phi_k(X, \pi(X))\|_{V_k^{-1}}\right\} \\ + 2\sum_{k\in[M]} \epsilon_k(\pi, \hat{\pi}).$$

The detailed proof is shown in Appendix D. In the case where realizability holds for all of the model classes ($\epsilon_k = 0$ for all k), an exact oracle inequality is achieved and we have, with high probability, $\mathrm{Reg}(\pi, \hat{\pi})$ is bounded above by

$$\mathcal{O}\left(\min_{k\in[M]} \sqrt{\frac{d_k \log(M/\delta)}{n}} \cdot \mathbb{E}_X \|\phi_k(X, \pi(X)\|_{V_k^{-1}}\right).$$

However, when positive approximation error is involved, it can be cumulative in the regret bound. We note that Algorithm 2 is similar in concept to the online representation selection algorithm of Papini et al. (2021).

5.2. Balancing complexity and approximation error

We now consider the setting where the worst-case coverage properties are tolerable, but we would like to optimally trade off approximation error and statistical complexity. We will examine two methods: **i)** an adaptive method inspired by the SLOPE estimator (Su et al., 2020) and **ii)** the classical hold-out method. While the hold-out method is desirable for its simplicity, the SLOPE method is potentially capable of achieving stronger theoretical guarantees under a slightly stronger nestedness condition on the model classes.

Hitherto, for the dataset D, we required only that the actions a_i are chosen independent of all potential outcomes conditional on x_i without regard to any behavior policy μ. We will now explicitly assume that the each (x_i, a_i, y_i) in the dataset D is sampled jointly from $\mathcal{D}$ and a fixed behavior policy μ as discussed in Section 2.2. This is stronger than before, but it is a standard setting for batch learning (Xie et al., 2021). Henceforth, we simply use $\mathbb{E}_\mu[\cdot]$ to denote the expectation over the joint distribution $\mathbb{E}_{X,A\sim\mathcal{D}\times\mu}[\cdot]$. We also consider a relaxed version of the approximation error, which can be written in terms of the statistical approximation error between $\mathcal{F}_k$ and the true reward function f:

$$\tilde{\epsilon}_k = \min_{\theta\in\mathbb{R}^{d_k}} 2\sqrt{\mathcal{C}(\mu)\mathbb{E}_\mu\left(\langle\phi_k(X, A), \theta\rangle - f(X, A)\right)^2}.$$

Define $\bar{\theta}_k = \arg\min_{\theta\in\mathbb{R}^{d_k}} \mathbb{E}_\mu\left(\langle\phi_k(X, A), \theta\rangle - f(X, A)\right)^2$ when $\mathbb{E}_\mu\left[\phi(X, A)\phi(X, A)^\top\right] \succ 0$. If $\mathcal{F}_k$ satisfies realizability, then $\tilde{\epsilon}_k = 0$ as before. However, this version has dependence on the concentrability coefficient $\mathcal{C}(\mu)$ for the worst-case dataset shift (Definition 1), which we are willing to tolerate in this section.

Algorithm 3 SLOPE Method

```
Input: Dataset $D$, linear model classes $\mathcal{F}_1, \dots, \mathcal{F}_M$, confidence parameter $\delta > 0$
for $k \in [M]$ do
    Estimate covariance matrix $V_k$ and parameters $\hat{\theta}_k$ as in Algorithm 2.
    Set $\hat{f}_k(x, a) \leftarrow \left\langle \phi_k(x, a), \hat{\theta}_k \right\rangle$
    Set $\hat{\pi}_k(x) \leftarrow \arg\max_{a\in\mathcal{A}} \hat{f}(x, a)$ with ties broken arbitrarily
end for
for $\ell \in [M]$ do
    for $k \in [M]$ do
        Set $\hat{v}_k(\hat{\pi}_\ell) \leftarrow \mathbb{E}_X\left[\hat{f}_k(X, \hat{\pi}_\ell(X))\right]$
        Set $\xi_k \leftarrow \zeta_k(\delta/M) \cdot \mathbb{E}_X \max_a \|\phi_k(X, a)\|_{V_k^{-1}}$
        Define intervals
            $$\mathcal{I}_{k,\ell} = [\hat{v}_k(\hat{\pi}_\ell) - 2\xi_k,\ \hat{v}_k(\hat{\pi}_\ell) + 2\xi_k]$$
    end for
    Select model class for for evaluating $\hat{\pi}_\ell$:
        $$\hat{k}(\ell) = \min\left\{k\ :\ \bigcap_{j=k}^{M} \mathcal{I}_{j,\ell} \text{ is non-empty}\right\}$$
    Set $\hat{v}(\hat{\pi}_\ell) \leftarrow \hat{v}_{\hat{k}(\ell)}(\hat{\pi}_\ell)$
end for
Set $\hat{k} = \arg\max_{k\in[M]} \hat{v}(\hat{\pi}_k)$
Return: $\hat{\pi} = \hat{\pi}_{\hat{k}}$
```

5.2.1. SLOPE Method

The first method, which we state in Algorithm 3, is inspired by the SLOPE estimator, originally designed for evaluation (Su et al., 2020). The algorithm begins by generating estimates $\hat{\theta}_k$ using the dataset D. The main idea is to then estimate the values of the $\hat{\pi}_k$ policies using an improved variant of the SLOPE estimator (which may be of independent interest) to achieve the optimal trade-off between approximation error and complexity. We describe this sub-procedure and its differences from the original in Appendix E. However, we require additional structure on the model classes in order to meet the pre-conditions of the improved SLOPE estimator. In particular, we assume the model classes are nested in the sense of Definition 2 (see precedents Bartlett et al. (2002); Foster et al. (2019)). We also assume the following distributional conditions.

Assumption 1. *For all* $k \in [M]$, $\mathbb{E}_\mu\left[\phi(X, A)\right] = 0$ *and* $\Sigma_k := \mathbb{E}_\mu\left[\phi(X, A)\phi(X, A)^\top\right] \succ 0$. *Furthermore, under* $\mathcal{D} \times \mu$, *the features* $\phi_k(X, A)$ *are sub-Gaussian with* $\|\Sigma_k^{-1/2}\phi_k(X, A)\|_{\psi_2} \leq 1$.

Assumption 2. *For all* $k \in [M]$, $\|\bar{\theta}_k\| \leq 1$.

Centering is done for ease of exposition. The eigenvalue lower bound is useful for random design linear regression analysis. The sub-Gaussian condition is standard. Assumption 2 is like the precondition of Theorem 1 but it is distribution-dependent rather than data-dependent. Define confidence coefficient (used in Algorithm 3) as

$$\zeta_k(\delta) = \sqrt{\frac{\lambda d_k}{n}} + C_1\sqrt{\frac{d_k}{n}}\|V_k^{-1/2}\| \log(4d_k/\delta)$$
$$+\sqrt{\frac{C_2 d_k + C_3 d_k^{1/2}\log^{1/2}(4d_k/\delta) + C_4\log(4d_k/\delta)}{n}}$$

for sufficiently large constants $C_1, C_2, C_3, C_4 > 0$ defined in Appendix E. Under these assumptions, we have the following theorem.

Theorem 4. *Let* $\mathcal{F}_1, \ldots, \mathcal{F}_M$ *be a nested collection of linear model classes. For* $\lambda_k = 1$ *for all* $k \in [M]$, *Algorithm 3 outputs a policy* $\hat{\pi}$ *such that, with probability at least* $1 - 4\delta$, *for any comparator policy* π, *Reg*$(\pi, \hat{\pi})$ *is bouned above by*

$$12 \min_{k \in [M]}\left\{\tilde{\epsilon}_k + \zeta_k(\delta/M) \cdot \mathbb{E}_X \max_a \|\phi_k(X, a)\|_{V_k^{-1}}\right\}.$$

We show the detailed proof in Appendix E. Theorem 4 shows that it is possible to obtain a near-oracle inequality when we are willing to forgo the coverage property and focus solely on trading off approximation error and complexity. Concisely, this is bounded above by

$$\mathcal{O}\left(\min_k \tilde{\epsilon}_k + \sqrt{\frac{d_k\|V_k^{-1}\|\log(d_k M/\delta)}{n}}\mathbb{E}\max_a\|\phi_k(X, a)\|_{V_k^{-1}}\right)$$

The coefficient $\zeta_k(\delta/M)$ on the second term is slightly larger than $\beta_{\lambda_k, \delta/M}(n, d_k)$ due to the dependence on $\|V_k^{-1/2}\|$, but they are of approximately the same order in d and n^{-1}. We emphasize the factor $\mathbb{E}_X \max_a \|\phi_k(X, a)\|_{V_k^{-1}}$, is different from the coverage error as defined in (2), which is $\mathbb{E}_X\|\phi_k(X, \pi(X))\|_{V_k^{-1}}$. This version demonstrates the distribution shift effect on features, but depends on the worst-case policy rather than the comparator π. Thus, we are unable to maintain competitiveness against well-covered policies. Note that only the approximation error $\tilde{\epsilon}_k$ depends on $\mathcal{C}(\mu)$.

Note that $\tilde{\epsilon}_k$ is a weaker form of the approximation than the previously used $\epsilon_k(\pi, \hat{\pi})$. A natural question is whether a bound of the form $\min_k\{\tilde{\epsilon}_k + \zeta_k(\delta/M) \cdot \mathbb{E}\|\phi(X, \pi(X))\|_{V_k^{-1}}\}$, which satisfies all three criteria, is possible with this slightly weaker approximation error. We show in Appendix C that the argument in the proof of Theorem 2 still applies in this case and thus a bound of this form is still not possible.

5.2.2. Hold-out Method

We now analyze the performance of the hold-out method, a classical model selection tool in supervised learning and risk minimization. The dataset D is partitioned into D_{in} and D_{out} where D_{out} is used to estimate the regression error as a proxy and selecting among candidates $\hat{\theta}_k$ for each model class trained on D_{in}. As before, we define $\hat{\pi}_k(x) \in \arg\max_{a \in \mathcal{A}}\left\langle\phi_k(x, a), \hat{\theta}_k\right\rangle$. We denote the empirical regression loss on the independent hold-out set as:

$$\hat{L}_k(\theta) = \frac{1}{|D_{out}|}\sum_{(x_i, a_i, y_i) \in D_{out}}\left(\langle\phi_k(x_i, a_i), \theta\rangle - y_i\right)^2$$

The hold-out method simply chooses the model class with smallest empirical loss: $\hat{k} \in \arg\min_{k \in [M]} \hat{L}_k(\hat{\theta}_k)$.

Theorem 5. *Given arbitrary linear model classes* $\mathcal{F}_1, \ldots, \mathcal{F}_M$, *let* $\hat{\pi} = \hat{\pi}_{\hat{k}}$ *where* $\hat{k} \in \arg\min_{k \in [M]} \hat{L}_k(\hat{\theta}_k)$. *Then, there is a constant* $C > 0$ *such that, with probability at least* $1 - 2\delta$, *Reg*$(\pi, \hat{\pi})$ *is bounded above by*

$$\min_k\left\{\tilde{\epsilon}_k + C\sqrt{\mathcal{C}(\mu)}\|\hat{\theta}_k - \bar{\theta}_k\|_{\Sigma_k}\right\}$$
$$+\mathcal{O}\left(\sqrt{\mathcal{C}(\mu)} \cdot \frac{(1 \vee \max_\ell\|\hat{\theta}_\ell\|)\log^{1/2}(M/\delta)}{n_{out}^{1/4}}\right).$$

The detailed proof is provided in Appendix F. Note that, for simplicity, we have stated the bound abstractly in terms of its estimation error $\|\hat{\theta}_k - \bar{\theta}_k\|_{\Sigma_k}$ and the norm of $\max_\ell\|\hat{\theta}_\ell\|$, where Σ_k is defined in Assumption 1. Standard analyses of random design linear regression (Hsu et al., 2012b) can provide further bounds on these. While we are able to select to achieve error on the order of the best model class, this is only achieved when the estimation error depends on the concentrability coefficient $\mathcal{C}(\mu)$. There is some residual estimation error on the order of $O(1/n_{out}^{1/4})$, which is slower than the typical $O(1/\sqrt{n})$; however, this term does not have any dependence on d, assuming $\|\bar{\theta}_\ell\|$ is of constant size.

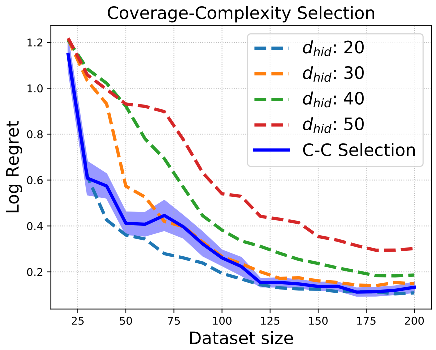

Figure 1: In the fully realizable case, the performance of the Algorithm 2 for model selection is compared to base algorithms that only use a single model class. Each model class is defined by a different feature representation generated with underlying dimension d_{hid}. The error band represents standard error.

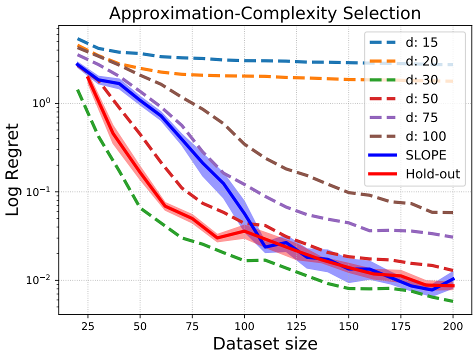

Figure 2: The performance of the SLOPE and hold-out methods are compared against base algorithms that only use a single model class of varying dimension. Both are eventually able to nearly match the performance of the best model class, but the hold-out method is consistently better with less data. Error bands represent standard error.

6. Experiments

To complement our primarily theoretical results, we study the utility of the above model selection algorithms in synthetic experiments and empirically compare them. In individual experiments on balancing complexity-coverage and approximation-complexity, we generated a collection of feature maps, each defining a linear model class. We first evaluated the performance of the base algorithms using Algorithm 1 with each model class. We then compare this performance to the proposed selection algorithms. All results were averaged over 20 trials. For clarity, error bands represent standard error only for the model selection algorithms. Further details can be found in the appendix.

Complexity-Coverage Trade-off In this setting, we studied the case where all feature maps are capable of realizing the true reward function. That is, the learner need not deal with any approximation error, but it can benefit from selecting a good representation. We let $|\mathcal{X}| = 20$ and $|\mathcal{A}| = 10$ and generated $d = |\mathcal{X}||\mathcal{A}|$-dimensional feature maps through the following procedure: for model class k a random collection of $d_{hid,k}$ vectors of size $|\mathcal{X}||\mathcal{A}|$ are generated ensuring that a linear combination exactly equals f. We then randomly project them to d-dimensions to produce ϕ_k. For model selection, we implemented Algorithm 2. Figure 1 compares the performance. As expected, the model class with smallest d_{hid} performs best. The model selection is nearly able to match this performance, even without knowing which feature map uses the smallest d_{hid}.

Approximation Error-Complexity Trade-off Next, we consider trading off approximation error and complexity with nested function classes. We again let $|\mathcal{A}| = 10$, but allowed $\mathcal{X}$ to be infinite with feature vectors generated from zero-mean normal distributions with different covariance matrices. The feature vectors were given ambient dimension $d = 100$, but the reward function was designed using only the first $d_* = 30$ coordinates. Model classes were generated by truncating full feature vectors to the following dimensions $\{15, 20, 30, 50, 75, 100\}$. Thus, the first two suffer from approximation error while the last three are excessively large. For model selection, we considered both the SLOPE method and the hold-out method with an 80/20 data split. Figure 2 shows that both model selection algorithms are eventually able to match performance of the best model class. Interestingly, the hold-out performs consistently better than the SLOPE method with small data.

7. Discussion

We introduced the theoretical study of model selection for batch policy optimization, identifying three sources of error to consider when selecting model classes. We showed that balancing all three is not possible in general while remaining competitive with an oracle, but relaxing any one allows the design of effective model selection algorithms. Several open questions remain. First, while the negative result is general, the positive results thus far have applied only to the linear contextual bandit setting. We expect that the challenges become more complex for reinforcement learning and general model classes, but similar trade-offs may be observed for general function classes that handle coverage in terms of comparator-specific concentrability coefficients (Xie et al., 2021; Uehara & Sun, 2021). Another interesting direction is to understand more formally the strong empirical performance of the hold-out method.

Acknowledgements

We thank Christoph Dann and anonymous reviewers for their detailed feedback in improving the paper. Most of the work was completed while JNL was an intern at Google Research. JNL was partially supported by NSF GRFP.

References

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

Agarwal, A., Luo, H., Neyshabur, B., and Schapire, R. E. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pp. 12–38. PMLR, 2017.

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *arXiv preprint arXiv:2006.10814*, 2020.

Antos, A., Kégl, B., Linder, T., and Lugosi, G. Data-dependent margin-based generalization bounds for classification. *Journal of Machine Learning Research*, 3(Jul): 73–98, 2002.

Bartlett, P. L. Fast rates for estimation error and oracle inequalities for model selection. *Econometric Theory*, pp. 545–552, 2008.

Bartlett, P. L., Boucheron, S., and Lugosi, G. Model selection and error estimation. *Machine Learning*, 48(1): 85–113, 2002.

Bubeck, S., Perchet, V., and Rigollet, P. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pp. 122–134. PMLR, 2013.

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*, pp. 1844–1854, 2020.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

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

Duan, Y., Jia, Z., and Wang, M. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pp. 2701–2709. PMLR, 2020.

Duchi, J. Lecture notes for statistics 311/electrical engineering 377. *Stanford University*, 2019.

Farahmand, A.-m. and Szepesvári, C. Model selection in reinforcement learning. *Machine learning*, 85(3):299–332, 2011.

Foster, D. and Rakhlin, A. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 3199–3210. PMLR, 2020.

Foster, D., Krishnamurthy, A., and Luo, H. Model selection for contextual bandits. *arXiv preprint arXiv:1906.00531*, 2019.

Ghosh, A., Sankararaman, A., and Kannan, R. Problem-complexity adaptive model selection for stochastic linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 1396–1404. PMLR, 2021.

Hsu, D., Kakade, S., Zhang, T., et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012a.

Hsu, D., Kakade, S. M., and Zhang, T. Random design analysis of ridge regression. In *Conference on learning theory*, pp. 9–1. JMLR Workshop and Conference Proceedings, 2012b.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*, 2019.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

Kumar, A., Singh, A., Tian, S., Finn, C., and Levine, S. A workflow for offline model-free robotic reinforcement learning. *arXiv preprint arXiv:2109.10813*, 2021.

Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement learning*, pp. 45–73. Springer, 2012.

Lattimore, T. and Szepesvári, C. Bandit algorithms. *preprint*, 2018.

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*, pp. 3340–3348. PMLR, 2021.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Lin, W. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Provably good batch off-policy reinforcement learning without great exploration. *Advances in Neural Information Processing Systems*, 33:1264–1274, 2020.

Lugosi, G. and Nobel, A. B. Adaptive model selection using empirical complexities. *Annals of Statistics*, pp. 1830–1864, 1999.

Massart, P. Concentration inequalities and model selection. 2007.

Modi, A., Jiang, N., Tewari, A., and Singh, S. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 2010–2020. PMLR, 2020.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (5), 2008.

Muthukumar, V. and Krishnamurthy, A. Universal and data-adaptive algorithms for model selection in linear contextual bandits. *arXiv preprint arXiv:2111.04688*, 2021.

Nachum, O., Chow, Y., Dai, B., and Li, L. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *arXiv preprint arXiv:1906.04733*, 2019a.

Nachum, O., Dai, B., Kostrikov, I., Chow, Y., Li, L., and Schuurmans, D. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019b.

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

Paine, T. L., Paduraru, C., Michi, A., Gulcehre, C., Zolna, K., Novikov, A., Wang, Z., and de Freitas, N. Hyperparameter selection for offline reinforcement learning. *arXiv preprint arXiv:2007.09055*, 2020.

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

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Rudelson, M. and Vershynin, R. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

Su, Y., Srinath, P., and Krishnamurthy, A. Adaptive estimator selection for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9196–9205. PMLR, 2020.

Tang, S. and Wiens, J. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. *arXiv preprint arXiv:2107.11003*, 2021.

Uehara, M. and Sun, W. Pessimistic model-based offline rl: Pac bounds and posterior sampling under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Xiao, C., Wu, Y., Mei, J., Dai, B., Lattimore, T., Li, L., Szepesvari, C., and Schuurmans, D. On the optimality of batch policy optimization algorithms. In *International Conference on Machine Learning*, pp. 11362–11371. PMLR, 2021.

Xie, T. and Jiang, N. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pp. 11404–11413. PMLR, 2021.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021.

Yang, M., Dai, B., Nachum, O., Tucker, G., and Schuurmans, D. Offline policy selection under uncertainty. *arXiv preprint arXiv:2012.06919*, 2020.

Zanette, A., Wainwright, M. J., and Brunskill, E. Provable benefits of actor-critic methods for offline reinforcement learning. *arXiv preprint arXiv:2108.08812*, 2021.

Zhang, S. and Jiang, N. Towards hyperparameter-free policy selection for offline reinforcement learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

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

A. Sub-Gaussian and Sub-Exponential Random Variables

In this section, we review basic definitions and properties of sub-Gaussian and sub-exponential random variables. See (Vershynin, 2018) for a comprehensive introduction.

Let $X \in \mathbb{R}$ be a random variable. We define the norms:

$$\|X\|_{\psi_2} := \sup_{p\in\mathbb{N}} p^{-1/2} \left(\mathbb{E}|X|^p\right)^{1/p} \tag{3}$$

$$\|X\|_{\psi_1} := \sup_{p\in\mathbb{N}} p^{-1} \left(\mathbb{E}|X|^p\right)^{1/p} \tag{4}$$

Definition 3. *The random variable X is sub-Gaussian with parameter $\|X\|_{\psi_2}$ if $\|X\|_{\psi_2} < \infty$. It is sub-exponential with parameter $\|X\|_{\psi_1}$ if $\|X\|_{\psi_1} < \infty$.*

For a non-negative real value $\tau \geq 0$ we write $X \sim \mathrm{subG}(\tau^2)$ to indicate that $\|X\|_{\psi_2} \leq \tau$. Similarly, we write $X \sim \mathrm{subE}(\tau)$ to suggest $\|X\|_{\psi_1} \leq \tau$. We note that this definition of sub-Gaussian random variables is equivalent up to constant factors with an alternative popular definition when $\mathbb{E}X = 0$. This definition requires that $\mathbb{E}e^{\lambda X} \leq \exp\left(\frac{\lambda^2\tau^2}{2}\right)$ for all $\lambda \in \mathbb{R}$. Let $X \in \mathbb{R}^d$ be a random vector. Then, we write $\|X\|_{\psi_2} = \sup_{v\in\mathbb{R}^d \,:\, \|v\|\leq 1} \|v^\top X\|_{\psi_2}$. The same notational conventions above apply to the vector X.

We now state several basic results concerning sub-Gaussian and sub-exponential random variables that will be used throughout the remaining proofs.

Lemma 1 ((Vershynin, 2010), Lemma 2.7.7)**.** *Let X and Y be (potentially dependent) real-valued random variables. Then, the following holds: $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$.*

Lemma 2. *Let $X \in \mathbb{R}^d$ be a random vector with second moment matrix $\Sigma = \mathbb{E}XX^\top$. Let $v \in \mathbb{R}^d$ be such that $\|v\| \leq 1$. Then, $v^\top \Sigma v \leq 2\|X\|_{\psi_2}^2$.*

Proof. We have $v^\top \Sigma v = \mathbb{E}v^\top XX^\top v = \mathbb{E}(v^\top X)^2$. From the definition of $\|\cdot\|_{\psi_2}$, we have that $\mathbb{E}|v^\top X|^2 \leq 2\|v^\top X\|_{\psi_2}^2$. Finally, we note that $\|v^\top X\|_{\psi_2} \leq \|X\|_{\psi_2}$ since $\|v\| \leq 1$. □

Lemma 3. *Let X satisfy $\mathbb{E}X = 0$ and $\|X\|_{\psi_2} \leq \tau$. Then, $\mathbb{E}\exp(\lambda X) \leq \exp\left(\frac{5\lambda^2\tau^2}{2}\right)$ for all $\lambda \in \mathbb{R}$.*

Proof. Note that $\|X\|_{\psi_2} \leq \tau$ implies that $\left(\mathbb{E}|X|^p\right)^{1/p} \leq \tau\sqrt{p}$. for all p. Theorem 3.10 of (Duchi, 2019) shows that this implies the stated condition. □

B. Proof of Theorem 1

Theorem 1. *Let $\hat{\pi}$ be the output policy of Algorithm 1 with $\lambda > 0$. Define,*

$$\begin{aligned}\epsilon(\pi, \hat{\pi}) = \mathbb{E}_X \left[|f(X, \pi(X)) - \langle \phi(X, \pi(X)), \theta_* \rangle |\right] \\ + \mathbb{E}_X \left[| \langle \phi(X, \hat{\pi}(X)), \theta_* \rangle - f(X, \hat{\pi}(X))|\right]\end{aligned}$$

where $\theta_* \in \arg\min_{\theta\in\mathbb{R}^d} \sum_{i\in[n]} \left(\phi(x_i, a_i)^\top \theta - f(x_i, a_i)\right)^2$. If $\|\theta_*\| \leq \sqrt{d}$, then with probability at least $1-\delta$, for any policy π (including the optimal policy), Reg$(\pi, \hat{\pi})$ is bounded above by

$$\begin{aligned}&\underbrace{\epsilon(\pi, \hat{\pi})}_{\textit{approx. error}} + \underbrace{2\beta_{\lambda,\delta}(n, d)}_{\textit{complexity}} \cdot \underbrace{\mathbb{E}_X\|\phi(X, \pi(X))\|_{V^{-1}}}_{\textit{coverage}} \\ &= \widetilde{\mathcal{O}}\left(\epsilon(\pi, \hat{\pi}) + \sqrt{\frac{d}{n}} \cdot \mathbb{E}_X\|\phi(X, \pi(X))\|_{V^{-1}}\right)\end{aligned}$$

where V is the regularized empirical covariance matrix of the data, defined in Algorithm 1[5].

[5] Throughout the paper, we use $\widetilde{\mathcal{O}}$ to omit polylog factors of problem-dependent parameters such as δ^{-1}, dimension d, and number of model classes M (defined in Section 3.2).

We first leverage a basic result in the analysis of fixed design linear regression problems. For convenience, we will write $\phi_i = \phi(x_i, a_i)$, $f_i = f(x_i, a_i)$ and noise $\eta_i = \eta_i(a_i)$. That is, $y_i = f_i + \eta_i$. Recall the following definitions:

$$V = \frac{\lambda}{n}\mathbb{I}_d + \frac{1}{n}\sum_{i\in[n]} \phi_i \phi_i^\top \tag{5}$$

$$\hat{\theta} = V^{-1}\left(\frac{1}{n}\sum_{i\in[n]} \phi_i y_i\right) \tag{6}$$

$$\theta_* \in \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i\in[n]} \left(\phi_i^\top \theta - f_i\right)^2 \tag{7}$$

B.1. Concentration

Lemma 4. *Conditioned on $(x_i, a_i)_{i\in[n]}$, with probability at least $1-\delta$,*

$$\|\hat{\theta} - \theta_*\|_V \le \sqrt{\frac{\lambda\|\theta_*\|^2}{n}} + \sqrt{\frac{C_1 d + C_2 d^{1/2}\log^{1/2}(1/\delta) + C_3\log(1/\delta)}{n}} \tag{8}$$

where $C_1 = 5$, $C_2 = 10$, and $C_3 = 10$.

Proof. From the definition of $\hat{\theta}$, we have

$$\|\hat{\theta} - \theta_*\|_V = \|\frac{1}{n}V^{-1}\sum_i \phi_i\left(f_i + \eta_i\right) - \theta_*\|_V \tag{9}$$

$$= \|\frac{1}{n}V^{-1}\sum_{i\in[n]} \phi_i\phi_i^\top\theta_* + \frac{1}{n}V^{-1}\sum_i \phi_i\eta_i - \theta_*\|_V \tag{10}$$

where in the last equality we have used the fact that θ_* is the solution to $\min_\theta \sum_i (\phi_i^\top \theta_* - f_i)^2$ and therefore satisfies the normal equations:

$$\sum_i \phi_i\phi_i^\top\theta_* = \sum_i \phi_i f_i \tag{11}$$

Then,

$$\|\hat{\theta} - \theta_*\|_V = \| -\frac{\lambda}{n}V^{-1}\theta_* + \frac{1}{n}V^{-1}\sum_i \phi_i\eta_i\|_V \tag{12}$$

$$\le \|\frac{\lambda}{n}V^{-1}\theta_*\|_V + \|\frac{1}{n}V^{-1}\sum_i \phi_i\eta_i\|_V \tag{13}$$

$$\le \sqrt{\frac{\lambda}{n}}\|\theta_*\| + \|\frac{1}{n}V^{-1/2}\sum_i \phi_i\eta_i\| \tag{14}$$

where the last inequality follows since $\sigma_{\max}^{1/2}(V^{-1}) \le (\lambda/n)^{-1/2}$.

To bound the second term, we apply the Lemma 5, stated below, which is an application of standard fixed-design linear regression results of (Hsu et al., 2012b). This shows that

$$\|\frac{1}{n}V^{-1/2}\sum_i \phi_i\eta_i\|^2 \le \frac{5d + 10\sqrt{d\log(1/\delta)} + 10\log(1/\delta)}{n}$$

with probability at least $1-\delta$. Applying this to the previous bound on $\|\theta - \theta_*\|_V$ gives the result. □

Lemma 5.

$$P\left(\|\frac{1}{n}V^{-1/2}\sum_i \phi_i\eta_i\|^2 > \frac{5d+10\sqrt{d\log(1/\delta)}+10\log(1/\delta)}{n} \mid x_{1:n}, a_{1:n}\right) \leq \delta$$

Proof of Lemma 5. Let $\eta = (\eta_1, \dots, \eta_n)^\top$ and $\Phi = \begin{bmatrix}\phi_1 & \cdots & \phi_n\end{bmatrix}^\top$. Note then that $\|\frac{1}{n}V^{-1/2}\sum_i \phi_i\eta_i\| = \frac{1}{\sqrt{n}}\|\left(\lambda\mathbb{I}_d + \Phi^\top\Phi\right)^{-1/2}\Phi^\top\eta\|$, where η is a random 1-sub-Gaussian random vector consisting of independent entries. Note that we have that $\mathbb{E}\eta_i = 0$ and $\|\eta_i\|_{\psi_2} \leq 1$, which, by Lemma 3 implies $\mathbb{E}\exp(\lambda X) \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$ where $\sigma^2 \leq 5$. The concentration result then follows from a version of the Hanson-Wright inequality due to (Hsu et al., 2012a), a restatement of which can be found in Lemma 12:

$$P_\eta\left(\|A\eta\|^2 > C_1\operatorname{tr}(AA^\top) + C_2\sqrt{\operatorname{tr}((AA^\top)^2)\log(1/\delta)} + C_3\sigma^2\|AA^\top\|\log(1/\delta)\right) \leq \delta \tag{15}$$

where $A = \left(\lambda\mathbb{I}_d + \Phi^\top\Phi\right)^{-1/2}\Phi^\top$ and $C_1 = 5$, $C_2 = 10$, and $C_3 = 10$. We may then bound the relevant quantities. Let $\Phi^\top\Phi = U\Lambda U^\top$ be the spectral decomposition of $\Phi^\top\Phi$ where U is unitary and $\Lambda \succeq 0$ is diagonal.

$$\begin{aligned}
\operatorname{tr}(AA^\top) &= \operatorname{tr}\left(\left(\lambda\mathbb{I}_d + \Phi^\top\Phi\right)^{-1/2}\Phi^\top\Phi\left(\lambda\mathbb{I}_d + \Phi^\top\Phi\right)^{-1/2}\right) && (16)\\
&= \operatorname{tr}\left(\left(\lambda\mathbb{I}_d + \Phi^\top\Phi\right)^{-1}\Phi^\top\Phi\right) && (17)\\
&\leq \operatorname{tr}\left((\lambda\mathbb{I}_d + \Lambda)^{-1}\Lambda\right) && (18)\\
&\leq d && (19)\\
\operatorname{tr}((AA^\top)^2) &= \operatorname{tr}\left(\left(\lambda\mathbb{I}_d + \Phi^\top\Phi\right)^{-1/2}\Phi^\top\Phi\left(\lambda\mathbb{I}_d + \Phi^\top\Phi\right)^{-1}\Phi^\top\Phi\left(\lambda\mathbb{I}_d + \Phi^\top\Phi\right)^{-1/2}\right) && (20)\\
&\leq \operatorname{tr}\left((\lambda\mathbb{I}_d + \Lambda)^{-1}\Lambda(\lambda\mathbb{I}_d + \Lambda)^{-1}\Lambda\right) && (21)\\
&\leq d && (22)\\
\|AA^\top\| &= \|\left(\lambda\mathbb{I}_d + \Phi^\top\Phi\right)^{-1/2}\Phi^\top\Phi\left(\lambda\mathbb{I}_d + \Phi^\top\Phi\right)^{-1/2}\| && (23)\\
&= \|U(\lambda\mathbb{I}_d + \Lambda)^{-1/2}\Lambda(\lambda\mathbb{I}_d + \Lambda)^{-1/2}U^\top\| && (24)\\
&\leq 1 && (25)
\end{aligned}$$

where the very last inequality uses the that the unitary matrix preserves the norm an the maximal eigenvalue of $(\lambda\mathbb{I}_d + \Lambda)^{-1/2}\Lambda(\lambda\mathbb{I}_d + \Lambda)^{-1/2}$ is at most 1.

We therefore conclude that

$$\|\frac{1}{n}V^{-1/2}\sum_i \phi_i\eta_i\|^2 \leq \frac{C_1 d + C_2\sqrt{d\log(1/\delta)} + C_3\log(1/\delta)}{n} \tag{26}$$

with probability at least $1-\delta$.

□

B.2. Full Proof

Proof of Theorem 1. for this proof, all expectations $\mathbb{E}$ denote $\mathbb{E}_X$, the expectation over the state random variable X from $\mathcal{D}$. By adding and subtracting, the regret may be decomposed simply as

$$\begin{aligned}
\operatorname{Reg}(\pi, \hat{\pi}) &= \mathbb{E}\left[f(X, \pi(X)) - \langle\phi(X, \pi(X)), \theta_*\rangle\right] && (27)\\
&\quad + \mathbb{E}\left[\langle\phi(X, \hat{\pi}(X)), \theta_*\rangle - f(X, \hat{\pi}(X))\right] && (28)\\
&\quad + \mathbb{E}\left[\langle\phi(X, \pi(X)), \theta_*\rangle - \langle\phi(X, \hat{\pi}(X)), \theta_*\rangle\right] && (29)\\
&\leq \epsilon(\pi, \hat{\pi}) + \mathbb{E}\left[\langle\phi(X, \pi(X)), \theta_*\rangle - \langle\phi(X, \hat{\pi}(X)), \theta_*\rangle\right] && (30)
\end{aligned}$$

For the remainder of the proof, we focus on bounding the second term. Adding and subtracting again, we have

$$
\begin{aligned}
&\mathbb{E}\left[\langle \phi(X,\pi(X)) - \phi(X,\hat{\pi}(X)), \theta_* \rangle\right] \qquad (31)\\
&\leq \mathbb{E}\left[\phi(X,\pi(X))^\top \theta_* - \phi(X,\hat{\pi}(X))^\top \hat{\theta}\right] + \mathbb{E}\left[\phi(X,\hat{\pi}(X))^\top \hat{\theta} - \phi(X,\hat{\pi}(X))^\top \theta_*\right] \qquad (32)\\
&\leq \mathbb{E}\left[\phi(X,\pi(X))^\top \theta_* - \phi(X,\hat{\pi}(X))^\top \hat{\theta}\right] + \|\hat{\theta} - \theta_*\|_V \cdot \mathbb{E}\|\phi(X,\hat{\pi}(X))\|_{V^{-1}} \qquad (33)
\end{aligned}
$$

where the last inequality is due to Cauchy-Schwarz.

Now, we apply the result of Lemma 4 to get that the event

$$
\begin{aligned}
\|\hat{\theta} - \theta_*\|_V &\leq \sqrt{\frac{\lambda \|\theta_*\|^2}{n}} + \sqrt{\frac{C_1 d + C_2 d^{1/2} \log^{1/2}(1/\delta) + C_3 \log(1/\delta)}{n}} \qquad (34)\\
&\leq \beta_{\lambda,\delta}(n,d) \qquad (35)
\end{aligned}
$$

occurs with probability at least $1-\delta$ for absolute constants $C_1, C_2, C_3 > 0$ defined there. Conditioning on this event, we have

$$
\begin{aligned}
\mathbb{E}\left[\langle \phi(X,\pi(X)) - \phi(X,\hat{\pi}(X)), \theta_* \rangle\right] &\leq \mathbb{E}\left[\phi(X,\pi(X))^\top \theta_* - \phi(X,\hat{\pi}(X))^\top \hat{\theta}\right] \\
&\quad + \beta_{\lambda,\delta}(n,d) \cdot \mathbb{E}\|\phi(X,\hat{\pi}(X))\|_{V^{-1}} \\
&\leq \mathbb{E}\left[\phi(X,\pi(X))^\top \theta_* - \phi(X,\pi(X))^\top \hat{\theta}\right] \\
&\quad + \beta_{\lambda,\delta}(n,d) \cdot \mathbb{E}\|\phi(X,\pi(X))\|_{V^{-1}} \\
&\leq \left(\|\hat{\theta} - \theta_*\|_V + \beta_{\lambda,\delta}(n,d)\right) \cdot \mathbb{E}\|\phi(X,\pi(X))\|_{V^{-1}} \\
&\leq 2\beta_{\lambda,\delta}(n,d) \cdot \mathbb{E}\|\phi(X,\pi(X))\|_{V^{-1}}
\end{aligned}
$$

where the second inequality applies the penalized action-selection for policy $\hat{\pi}$, the third inequality applies Cauchy-Schwarz, and the last inequality once again applies the condition on the concentration of $\|\hat{\theta} - \theta_*\|_V$. □

B.3. Discussion of Approximation Error

In this paper, we work with a fairly general notion of approximation error $\epsilon_k(\pi, \hat{\pi})$. Note that this depends both on the comparator policy π and the learned policy $\hat{\pi}$ and it tends to be small when θ_* outputs similar rewards to f on both of these policies. It is exactly zero when realizability is satisfied, $f \in \mathcal{F}_k$, as is assumed in most related work. The reason for this choice is that it allows a large degree of flexibility as many natural alternatives may upper bound it, for example those given below.

Here we point out a couple alternatives that appear frequently in bandit and RL theory.

1. Perhaps the most common assumption is a worst-case difference between f and the model class $\mathcal{F}_k$ (Jin et al., 2019; Foster & Rakhlin, 2020):

   $$\epsilon_{k,\text{worst-case}} = \min_{\hat{f} \in \mathcal{F}_k} \sup_{x \in \mathcal{X}, a \in \mathcal{A}} |f(x,a) - \hat{f}(x,a)|$$

   The obvious disadvantage of this version is that certain states or contexts might be irrelevant but still lead to large prediction errors. Furthermore, the minimizing $\hat{f}$ does not generally have any convenient statistical properties (e.g. satisfying first-order optimality conditions in the linear case with respect to some relevant distribution).

2. Versions of the minimum squared error are also commonly used (Chen & Jiang, 2019) when it is assumed that the data is generated from a behavior policy μ, as assumption we do not consider until Section 5:

   $$\epsilon_{k,\text{sq}} = \min_{\hat{f} \in \mathcal{F}_k} \mathbb{E}_\mu \left(\hat{f}(X,A) - f(X,A)\right)^2$$

This is a natural formulation from a statistical perspective as well and it partially remedies some of the problems of the worst-case approximation error since we care only about those states and actions induced under μ. Unfortunately, this typically brings a concentrability coefficient into the mix. We leverage this as an upper bound to achieve some positive results in Section 5.2.

We remark that numerous prior works make the assumption that an upper bound on the approximation error is *known*. However, it is unrealistic in practice to expect that such information is available, and it furthermore trivializes the model selection problem. While we consider upper bounds to $\epsilon_k(\pi, \hat{\pi})$ in Section 5.2, we make no such assumption about knowing the value of this upper bound.

C. Proof of Theorem 2

Theorem 2. *Let $\mathcal{F}_1, \ldots, \mathcal{F}_M$ be a particular nested collection of linear model classes (Definition 2). For any $\alpha > 0$, there is $n = \Theta(\alpha^2)$ such that for any algorithm A there is a contextual bandit instance with comparator π and dataset D with n interactions consistent with $\mathcal{D}$ that satisfies*

$$\frac{\mathbb{E}_D\left[Reg(\pi, A(\mathcal{F}_{1:M}, D))\right]}{\min_k \left\{\epsilon_k(\pi, \hat{\pi}) + \sqrt{\frac{d_k}{n}} \cdot \mathbb{E}_X \|\phi_k(X, \pi(X))\|_{V_k^{-1}}\right\}} \geq \alpha.$$

Let us begin by first describing the dataset and observations for a given contextual bandit instance. The set of contextual bandit instances is determined by possible distributions $\mathcal{D}$ over state-reward pairs. For our construction, we fix the state-action pairs in the dataset $\{(x_i, a_i)\}_{i=1}^n$ across all instances that we consider. To be clear, the distribution of the rewards in the dataset under different instances will be different, but the covariates are held fixed. Note that this is not necessary for the lower bound, but it will suffice in our example.

Proof of Theorem 2. The main idea of the proof is to construct a difficult contextual bandit problem with $\mathcal{A} = \{a_1, a_2\}$ and show that the oracle can leverage a pair of model classes satisfying Definition 2 to achieve small regret. The hardness of the contextual bandit problem will be shown via a reduction to a multi-armed bandit (MAB). The model classes will be chosen such that one is well-specified while the other has no approximation error in some instances but large approximation error in others.

To start, consider a class of two two-armed bandit instances $\mathcal{E} = (\nu_1, \nu_2)$ (i.e. no states) which are identified by their product distributions over rewards of both arms. We let $\nu_1 = \mathcal{N}(-\Delta, 1) \times \mathcal{N}(-2\Delta, 1)$ and $\nu_2 = \mathcal{N}(-\Delta, 1) \times \mathcal{N}(0, 1)$ for $\Delta > 0$ to be determined later. That is, across both instances, arm a_1 has the same reward distribution, but arm a_2 can have either mean 0 or -2Δ. We let $\mathbb{E}_{\nu_i}$ denote the expectation associated with instance ν_i. We let D be a dataset consisting of $n_1 > 0$ samples from a_1 and $n_2 > 0$ samples from a_2 with n_1 and n_2 to be determined precisely later. Such a construction is similar to that of standard lower bounds in bandits (Bubeck et al., 2013; Lattimore & Szepesvári, 2018); however, since we are in the offline setting, the dataset is given.

We now establish a regret lower bound for any arbitrary algorithm A that outputs an arm $A(D) \in \{a_1, a_2\}$ as a function of the data D. The next lemma follows from a standard application of Le Cam's two-point method and similar results for the offline multi-armed bandit problem (Xiao et al., 2021).

Lemma 6. *Let $\Delta = \frac{1}{2\sqrt{n_2}}$. Then, for any algorithm A, $\max_{i,j} \mathbb{E}_{\nu_i}\left[Y(a_j) - Y(A(D))\right] \geq \frac{1}{8\sqrt{n_2}}$*

Henceforth, we will define $\Delta := \frac{1}{2\sqrt{n_2}}$. We now construct a linear contextual bandit instance and apply a reduction to the MAB setting so that we may leverage the stated lower bound. Let $\mathcal{X}$ be a singleton (that is, states have no effect) and again $\mathcal{A} = \{a_1, a_2\}$. Since there is only a single state, we omit notational dependence of policies[6] and functions on the state.

We again consider two instances $\mathcal{E} = \{\nu_1, \nu_2\}$ which each govern the data distribution denote by $\mathcal{D}_{\nu_i}$. For $\mathcal{D}_{\nu_1}$, we set $Y \sim \mathcal{N}(-\Delta, 1) \times \mathcal{N}(-2\Delta, 1)$ and, for $\mathcal{D}_{\nu_2}$, we set $Y \sim \mathcal{N}(-\Delta, 1) \times \mathcal{N}(0, 1)$ where Δ is defined above. Note that this ensures that the noise for either instance is given by the centered standard normal distribution $(Y(a) - f(a)) \sim \mathcal{N}(0, 1)$ for all $a \in \mathcal{A}$. We use π_* to denote the optimal policy (action), which depends on the instance. In ν_1, we have $\pi_* = a_1$ and

[6] For a deterministic policy π, we simply use $\pi \in \mathcal{A}$ to denote the selected action and n_π denotes the number of samples to arm π.

in ν_2, we have $\pi_* = a_2$. Finally, we assume that the batch dataset D again consists of $n_1 > 0$ samples of a_1 and $n_2 > 0$ samples of a_2 with $n_1 \geq n_2$ giving a total of $n = n_1 + n_2$ samples. Exact quantities will be determined at the end.

Next, we construct two linear model classes $\mathcal{F}_1$ and $\mathcal{F}_2$. For $\mathcal{F}_1$, we use the following 1-dimensional feature map $\phi_1 : \mathcal{A} \to \mathbb{R}$:

$$\phi_1(a) = \begin{cases} 1 & a = a_1 \\ 0 & a = a_2 \end{cases}.$$

$\mathcal{F}_1$ thus has some opportunity to make predictions about the mean of a_1 but the features are trivial for a_2, potentially leading to approximation error. For $\mathcal{F}_2$, we set $\phi_2 : \mathcal{A} \to \mathbb{R}^2$ as

$$\phi_2(a) = \begin{cases} (1,0)^\top & a = a_1 \\ (0,1)^\top & a = a_2 \end{cases}$$

Note that this model is well-specified as $f(a) = \phi_2(a)^\top \theta$ by setting $\theta = (f(a_1), f(a_2))^\top$. It is also evident that $\mathcal{F}_1$ and $\mathcal{F}_2$ are nested according to Definition 2.

Let

$$\theta_{k,*} = \arg\min_{\theta \in \mathbb{R}^{d_k}} \sum_{i \in [n]} \left(\phi_k(a^{(i)})^\top \theta - f(a^{(i)}) \right)^2$$

for $k \in \{1, 2\}$ where $a^{(i)}$ denotes the action of the ith datapoint in D. It is easy to verify that the following conditions are true as long as $n_1 > 0$ and $n_2 > 0$:

1. In ν_1: $\theta_{1,*} = -\Delta$ and $\theta_{2,*} = (-\Delta, -2\Delta)$.
2. In ν_2: $\theta_{1,*} = -\Delta$ and $\theta_{2,*} = (-\Delta, 0)$.

We now summarize the estimation error and approximation error for both model classes. The contribution of estimation error follows directly from the definitions. We summarize the results in the following fact. Note that we need not include expectations over the state since there is only one state.

Fact 1. *Let $V_k = \frac{\lambda}{n} + \frac{1}{n} \sum_{i \in [n]} \phi_k(a^{(i)}) \phi_k(a^{(i)})^\top$ for $k \in \{1, 2\}$ and $\lambda > 0$. For any instance in $\mathcal{E}$ and comparator $\pi \in \mathcal{A}$, the following inequalities hold:*

$$\|\phi_1(\pi)\|_{V_1^{-1}} \leq \sqrt{\frac{n}{n_1}} \qquad\qquad \|\phi_2(\pi)\|_{V_2^{-1}} \leq \sqrt{\frac{n}{n_\pi}}$$

Proof. The results follow by direct calculation and using the fact that $n_1, n_2 > 0$. □

For the approximation error, we can clearly see that, for model class $\mathcal{F}_2$, $\epsilon_2(\pi, \hat{\pi}) = 0$ for all instances in $\mathcal{E}$ and all $\hat{\pi}$ and π since the model is well-specified. For $\mathcal{F}_1$, the approximation error will depend on the instance. Observe that in ν_2, we have

$$\epsilon_1(\pi, \hat{\pi}) = |\langle \phi_1(\hat{\pi}), \theta_{1,*} \rangle - f(\hat{\pi})| + |\langle \phi_1(\pi), \theta_{1,*} \rangle - f(\pi)| = 0$$

regardless of what π and $\hat{\pi}$ are. Furthermore, for ν_1, we have $\epsilon_1(\pi, \hat{\pi}) \leq 2\Delta$ in the worst case.

Combining the results for both estimation error and approximation error, we have

$$\epsilon_1(\pi_*, \hat{\pi}) + \sqrt{\frac{d_1}{n}} \cdot \mathbb{E}_X \|\phi_1(\pi_*)\|_{V_1^{-1}} \leq 2\Delta + \frac{1}{\sqrt{n_1}} \leq \frac{2}{\sqrt{n_2}}$$
$$\epsilon_2(\pi_*, \hat{\pi}) + \sqrt{\frac{d_2}{n}} \cdot \mathbb{E}_X \|\phi_2(\pi_*)\|_{V_2^{-1}} \leq \sqrt{\frac{2}{n_1}}$$

in instance ν_1 and

$$\epsilon_1(\pi_*, \hat{\pi}) + \sqrt{\frac{d_1}{n}} \cdot \mathbb{E}_X \|\phi_1(\pi_*)\|_{V_1^{-1}} \leq \frac{1}{\sqrt{n_1}}$$
$$\epsilon_2(\pi_*, \hat{\pi}) + \sqrt{\frac{d_2}{n}} \cdot \mathbb{E}_X \|\phi_2(\pi_*)\|_{V_2^{-1}} \leq \sqrt{\frac{2}{n_2}}$$

in instance ν_2. Therefore, we conclude that

$$\min_{k \in [2]} \left\{ \epsilon_k(\pi_*, \hat{\pi}) + \sqrt{\frac{d_k}{n}} \cdot \mathbb{E}_X \|\phi_k(X, \pi_*)\|_{V_k^{-1}} \right\} \leq \frac{2}{\sqrt{n_1}}$$

for all $\nu \in \mathcal{E}$. Furthermore, note that this contextual bandit setting exactly reduces to the MAB problem and thus Lemma 6 requires that the regret of any algorithm A be lower bounded as $\mathbb{E}_\nu \left[\text{Reg}(\pi, A(D)) \right] \geq \frac{1}{8\sqrt{n_2}}$ for some $\nu \in \mathcal{E}$ with our given choice of Δ. Therefore, there is a constant $C_1 > 0$ such that for any algorithm A, there exists $\nu \in \mathcal{E}$ satisfying

$$\frac{\mathbb{E}_\nu \left[\text{Reg}(\pi_*, A(D)) \right]}{\min_{k \in [2]} \left\{ \epsilon_k(\pi_*, \hat{\pi}_k) + \sqrt{\frac{d_k}{n}} \cdot \mathbb{E}_X \|\phi_k(X, \pi_*)\|_{V_k^{-1}} \right\}} \geq \sqrt{\frac{n_1}{64 n_2}}$$

Finally, we are left with choosing n_1 and n_2 as the number of samples in the dataset (which is the same across all of the instances). Choosing $n_1 = \Omega\left(\alpha^2 n_2\right)$ ensures the claim, which is possible since it was assumed that $n = \Omega(\alpha^2)$.

□

C.1. Proof of Lemma 6

Proof. Note that

$$\max_{i,j} \mathbb{E}_{\nu_i} \left[Y(j) - Y(A(D)) \right] = \max_i \mathbb{E}_{\nu_i} \left[Y(i) - Y(A(D)) \right] \tag{36}$$

by definition of the instances ν_1 and ν_2. For convenience, we just write A instead of $A(D)$. Then,

$$\max_i \mathbb{E}_{\nu_i} \left[Y(i) - Y(A) \right] \geq \frac{1}{2} \left(\mathbb{E}_{\nu_1} \left[Y(1) - Y(A) \right] + \mathbb{E}_{\nu_2} \left[Y(2) - Y(A) \right] \right) \tag{37}$$
$$= \frac{\Delta}{2} \left(P_{\nu_1}(A \neq 1) + P_{\nu_2}(A \neq 2) \right) \tag{38}$$
$$\geq \frac{\Delta}{2} \left(1 - \|P_{\nu_1} - P_{\nu_2}\|_{TV} \right) \tag{39}$$
$$\geq \frac{\Delta}{2} \left(1 - \sqrt{\frac{1}{2} D_{KL}(P_{\nu_1} \| P_{\nu_2})} \right) \tag{40}$$

where the last two inequalities follow from the definition of the total-variation distance and Pinsker's inequality, respectively. Then, we apply the tensorization of the KL-divergence over the product distribution induced by the dataset to get:

$$D_{KL}\left(P_{\nu_1} \| P_{\nu_2}\right) = n_1 D_{KL}\left(\mathcal{N}(\Delta, 1) \| \mathcal{N}(\Delta, 1)\right) \tag{41}$$
$$+ n_2 D_{KL}\left(\mathcal{N}(0, 1) \| \mathcal{N}(2\Delta, 1)\right) \tag{42}$$
$$= n_2 D_{KL}\left(\mathcal{N}(0, 1) \| \mathcal{N}(2\Delta, 1)\right) \tag{43}$$

For normal distribution, $D_{KL}\left(\mathcal{N}(0, 1) \| \mathcal{N}(2\Delta, 1)\right) = 2\Delta^2$. Therefore, choosing $\Delta = \frac{1}{2\sqrt{n_2}}$, we have

$$\max_i \mathbb{E}_{\nu_i} \left[Y(i) - Y(A) \right] \geq \frac{\Delta}{2} (1 - \Delta) \tag{44}$$
$$= \frac{1}{8\sqrt{n_2}} \tag{45}$$

□

D. Proof of Theorem 3

Theorem 3. *Given arbitrary linear model classes $\mathcal{F}_1, \dots, \mathcal{F}_M$, Algorithm 2 outputs a policy $\hat\pi$ such that, with probability at least $1-\delta$, for any comparator policy π, the regret Reg$(\pi, \hat\pi)$ is bounded above by*

$$\begin{aligned}&\min\nolimits_{k\in[M]}\left\{2\beta_{\lambda,\delta/M}(n,d_k)\cdot\mathbb{E}_X\|\phi_k(X,\pi(X))\|_{V_k^{-1}}\right\}\\ &\quad+2\sum\nolimits_{k\in[M]}\epsilon_k(\pi,\hat\pi).\end{aligned}$$

Proof of Theorem 3. For all $k\in[M]$, let $\theta_{k,*}$ denote the solution to

$$\min_{\theta\in\mathbb{R}^{d_k}}\sum_{i\in[n]}\left(\phi_k(x_i,a_i)^\top\theta-f(x_i,a_i)\right)^2$$

Throughout the proof, all expectations $\mathbb{E}$ denote $\mathbb{E}_X$ over $\mathcal{D}$ and, within expectations over X, we will write $\hat k := \hat k(X)$ for shorthand. From the definition of regret, we have

$$\begin{aligned}\text{Reg}(\pi,\hat\pi)&=\mathbb{E}\left[f(X,\pi(X))-f(X,\hat\pi(X))\right]\\ &\le\mathbb{E}\left[f(X,\pi(X))-\left\langle\phi_{\hat k}(X,\pi(X)),\theta_{\hat k,*}\right\rangle\right]+\mathbb{E}\left[\left\langle\phi_{\hat k}(X,\hat\pi(X)),\theta_{\hat k,*}\right\rangle-f(X,\hat\pi(X))\right]\\ &\quad+\mathbb{E}\left[\left\langle\phi_{\hat k}(X,\pi(X)),\theta_{\hat k,*}\right\rangle-\left\langle\phi_{\hat k}(X,\hat\pi(X)),\theta_{\hat k,*}\right\rangle\right]\end{aligned}$$

We now focus on bounding the contribution of the estimation error to the regret. Using the condition in Lemma 4 to bound each $\|\hat\theta_k-\theta_{*,k}\|_{V_k}$, we have

$$\mathbb{E}\left[\left\langle\phi_{\hat k}(X,\pi(X)),\theta_{\hat k,*}\right\rangle-\left\langle\phi_{\hat k}(X,\hat\pi(X)),\theta_{\hat k,*}\right\rangle\right]\tag{46}$$

$$\le\mathbb{E}\left[\left\langle\phi_{\hat k}(X,\pi(X)),\theta_{\hat k,*}\right\rangle-\left\langle\phi_{\hat k}(X,\hat\pi(X)),\hat\theta_{\hat k}\right\rangle\right]+\|\hat\theta_{\hat k}-\theta_{\hat k,*}\|_{V_{\hat k}}\cdot\mathbb{E}\|\phi_{\hat k}(X,\hat\pi(X))\|_{V_{\hat k}^{-1}}\tag{47}$$

$$\le\mathbb{E}\left[\left\langle\phi_{\hat k}(X,\pi(X)),\theta_{\hat k,*}\right\rangle-\left\langle\phi_{\hat k}(X,\hat\pi(X)),\hat\theta_{\hat k}\right\rangle\right]+\beta_{\lambda,\delta}(n,d_{\hat k})\cdot\mathbb{E}\|\phi_{\hat k}(X,\hat\pi(X))\|_{V_{\hat k}^{-1}}\tag{48}$$

Next, we apply the selection rule that determines $\hat k$ and $\hat\pi$ simultaneously, both of which are designed to maximize the penalized value estimate across actions and model classes. For any fixed $k\in[M]$, the previous display is bounded by

$$\mathbb{E}_X\left[\left\langle\phi_{\hat k}(X,\pi(X)),\theta_{\hat k,*}\right\rangle-\left\langle\phi_k(X,\pi(X)),\hat\theta_k\right\rangle\right]+\beta_{\lambda,\delta}(n,d_k)\cdot\mathbb{E}\|\phi_k(X,\pi(X))\|_{V_k^{-1}}\tag{49}$$

There is now a potential mismatch between the predictions under $\theta_{*,\hat k}$ and the predictions under $\hat\theta_k$. To handle this, we will turn to the approximation error. For $k\in[M]$, define $\epsilon_k(X):=|f(X,\pi(X))-\phi_k(X,\pi(X))^\top\theta_{k,*}|$.

The first term in the previous display can be bounded using predictions under model class k up to additive factors in $\epsilon_{\hat k}(X)$ and $\epsilon_k(X)$.

$$\begin{aligned}&\left\langle\phi_{\hat k}(X,\pi(X)),\theta_{\hat k,*}\right\rangle\\ &\le\left|\left\langle\phi_{\hat k}(X,\pi(X)),\theta_{\hat k,*}\right\rangle-f(X,\pi(X))\right|+\left|\langle\phi_k(X,\pi(X)),\theta_{k,*}\rangle-f(X,\pi(X))\right|+\langle\phi_k(X,\pi(X)),\theta_{k,*}\rangle\\ &\le\epsilon_{\hat k}(X)+\epsilon_k(X)+\langle\phi_k(X,\pi(X)),\theta_{k,*}\rangle\end{aligned}$$

Then, conditioned on the same event from Lemma 4 and using the approximation error above, we may further bound (49) with

$$\begin{aligned}&\mathbb{E}\left[\left\langle\phi_{\hat k}(X,\pi(X)),\theta_{\hat k,*}\right\rangle-\left\langle\phi_k(X,\pi(X)),\hat\theta_k\right\rangle\right]+\beta_{\lambda,\delta}(n,d_k)\cdot\mathbb{E}\|\phi_k(X,\pi(X))\|_{V_k^{-1}}\\ &\le\mathbb{E}\left[\langle\phi_k(X,\pi(X)),\theta_{k,*}\rangle-\left\langle\phi_k(X,\pi(X)),\hat\theta_k\right\rangle\right]+\mathbb{E}\left[\epsilon_{\hat k}(X)+\epsilon_k(X)\right]+\beta_{\lambda,\delta}(n,d_k)\cdot\mathbb{E}\|\phi_k(X,\pi(X))\|_{V_k^{-1}}\\ &\le\mathbb{E}\left[\epsilon_{\hat k}(X)+\epsilon_k(X)\right]+2\beta_{\lambda,\delta}(n,d_k)\cdot\mathbb{E}\|\phi_k(X,\pi(X))\|_{V_k^{-1}}\end{aligned}$$

Applying this upper bound to the regret and using the fact that this holds for any fixed $k \in [M]$, we get

$$
\begin{aligned}
&\mathrm{Reg}(\pi, \hat{\pi}) && (50)\\
&\le \epsilon_k(\pi, \hat{\pi}) + \mathbb{E}_X \left[\epsilon_{\hat{k}}(X)\right] + \min_{k \in [M]} \mathbb{E}_X \left\{ \epsilon_k(X) + 2\beta_{\lambda,\delta}(n, d_k) \cdot \|\phi_k(X, \pi(X))\|_{V_k^{-1}} \right\} && (51)\\
&\le 2 \sum_{k \in [M]} \epsilon_k(\hat{\pi}, \pi) + \min_{k \in [M]} \left\{ 2\beta_{\lambda,\delta}(n, d_k) \cdot \mathbb{E}_X \|\phi_k(X, \pi(X))\|_{V_k^{-1}} \right\} && (52)
\end{aligned}
$$

Note here that $\hat{k}$ depends on X and thus we cannot readily replace $\mathbb{E}_X \left[\epsilon_{\hat{k}}(X)\right]$ with $\max_{k'} \epsilon_{k'}(\pi, \hat{\pi})$ in the first inequality. The sum over approximation errors is thus done to simplify the bound in terms of just $\epsilon_{k'}(\pi, \hat{\pi})$ terms. Finally, note that Lemma 4 establishes concentration of $\|\hat{\theta}_k - \theta_{*,k}\|_{V_k}$ for all $k \in [M]$ with probability at least $1 - M\delta$. Changing variables to $\delta' = M\delta$ proves the result. □

E. Proof of Theorem 4

The SLOPE-inspired algorithm for balancing complexity and approximation error is stated completely in Algorithm 3.

E.1. Single Model Guarantee

We first begin with an independent result for a single d-dimensional $\mathcal{F}$ that shows that one can bound the regret of a learned policy $\hat{\pi}$ by the approximation error $\tilde{\epsilon}$ of the optimal parameter $\bar{\theta}$ plus an estimation error term that depends on the complexity of the model class d. For clarity of notation, we drop dependence on the model class index k in the subscript. Recall the definitions $\phi_i := \phi(x_i, a_i)$ and

$$
\begin{aligned}
V &= \frac{\lambda}{n}\mathbb{I}_d + \frac{1}{n}\sum_i \phi_i \phi_i^\top \\
\hat{\theta} &= V^{-1}\left(\frac{1}{n}\sum_i \phi_i y_i\right) \\
\hat{\pi}(x) &\in \arg\max_{a \in \mathcal{A}} \left\langle \hat{\theta}, \phi(x, a) \right\rangle \\
\bar{\theta} &\in \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}_\mu \left(\langle \phi(X, A), \theta \rangle - f(X, A)\right)^2 \\
\tilde{\epsilon} &= \min_{\theta \in \mathbb{R}^d} 2\sqrt{\mathcal{C}(\mu)\mathbb{E}_\mu \left(\langle \phi(X, A), \theta \rangle - f(X, A)\right)^2}
\end{aligned}
$$

Proposition 1. *For the above definitions, the following inequality holds with probability at least $1 - \delta$:*

$$
\|\hat{\theta} - \bar{\theta}\|_V \le \sqrt{\frac{\lambda\|\bar{\theta}\|^2}{n}} + C_4\sqrt{\frac{d}{n}}\|V^{-1/2}\| \cdot \log(4d/\delta) + \sqrt{\frac{C_1 d + C_2\sqrt{d\log(4d/\delta)} + C_3\log(4d/\delta)}{n}}
$$

Furthermore, under the same event, the regret Reg$(\pi, \hat{\pi})$ is bounded above by:

$$
\tilde{\epsilon} + \left(\sqrt{\tfrac{\lambda\|\bar{\theta}\|^2}{n}} + C_4\sqrt{\tfrac{d}{n}}\|V^{-1/2}\| \cdot \log(4d/\delta) + \sqrt{\tfrac{C_1 d + C_2\sqrt{d\log(4d/\delta)} + C_3\log(4d/\delta)}{n}}\right) \cdot \mathbb{E}_X \max_a \|\phi(X, a)\|_{V^{-1}}
$$

where $C_4 = 192$ and $C_{1:3}$ are defined in Lemma 4.

Proof. The regret decomposes as

$$
\begin{aligned}
\mathrm{Reg}(\pi, \hat{\pi}) &= \mathbb{E}_X \left[f(X, \pi(X)) - f(X, \hat{\pi}(X))\right] \\
&= \mathbb{E}_X \left[\left(f(X, \pi(X)) - \left\langle \phi(X, \pi(X)), \bar{\theta} \right\rangle\right) + \left(\left\langle \phi(X, \hat{\pi}(X)), \bar{\theta} \right\rangle - f(X, \hat{\pi}(X))\right)\right] \\
&\quad + \mathbb{E}_X \left[\left\langle \phi(X, \pi(X)) - \phi(X, \hat{\pi}(X)), \bar{\theta} \right\rangle\right]
\end{aligned}
$$

By Jensen's inequality and Definition 1, the first expectation can be bounded as

$$\begin{aligned}&\mathbb{E}_X\left[\left(f(X,\pi(X))-\langle\phi(X,\pi(X)),\bar\theta\rangle\right)+\left(\langle\phi(X,\hat\pi(X)),\bar\theta\rangle-f(X,\hat\pi(X))\right)\right]\\&\le 2\sqrt{\mathcal{C}(\mu)\mathbb{E}_\mu\left(\langle\phi(X,A),\bar\theta\rangle-f(X,A)\right)^2}\\&=\tilde\epsilon\end{aligned}$$

For the second expectation, it remains to show that the policy $\hat\pi$ selects actions nearly as well as π with respect to $\bar\theta$. For convenience, define $\phi(\pi):=\mathbb{E}_X\phi(X,\pi(X))$ for features ϕ and policy π. Then,

$$\begin{aligned}\left\langle\phi(\pi),\bar\theta\right\rangle-\left\langle\phi(\hat\pi),\bar\theta\right\rangle&=\left\langle\phi(\pi),\bar\theta\right\rangle-\left\langle\phi(\hat\pi),\hat\theta\right\rangle+\left\langle\phi(\hat\pi),\hat\theta\right\rangle-\left\langle\phi(\hat\pi),\bar\theta\right\rangle\\&\le\left\langle\phi(\pi),\bar\theta\right\rangle-\left\langle\phi(\pi),\hat\theta\right\rangle+\left\langle\phi(\hat\pi),\hat\theta\right\rangle-\left\langle\phi(\hat\pi),\bar\theta\right\rangle\\&\le\|\bar\theta-\hat\theta\|_V\cdot\mathbb{E}_X\|\phi(X,\pi(X))\|_{V^{-1}}+\|\bar\theta-\hat\theta\|_V\cdot\mathbb{E}_X\|\phi(X,\hat\pi(X))\|_{V^{-1}}\\&\le 2\|\bar\theta-\hat\theta\|_V\cdot\mathbb{E}_X\max_a\|\phi(X,a)\|_{V^{-1}}\end{aligned}$$

where the first inequality uses the fact that $\hat\pi$ selects actions to maximize the reward predicted with $\hat\theta$, the second inequality applies Cauchy-Schwarz and the last inequality takes the worst-case action. Thus, we focus on the concentration of $\|\hat\theta-\bar\theta\|_V$. We use the previous definitions of $\phi_i=\phi(x_i,a_i)$, $f_i:=f(x_i,a_i)$, and $\eta_i:=\eta_i(a_i)$. Furthermore, we define the error term $e_i=f_i-\phi_i^\top\bar\theta$.

$$\begin{aligned}\|\hat\theta-\bar\theta\|_V&=\left\|\frac{1}{n}V^{-1}\sum_i\phi_iy_i-\bar\theta\right\|_V\\&=\left\|\frac{1}{n}V^{-1}\sum_i\phi_i\left(\phi_i^\top\bar\theta+e_i+\eta_i\right)-\bar\theta\right\|_V\\&=\left\|\frac{1}{n}V^{-1}\sum_i\phi_i\eta_i+\frac{1}{n}V^{-1}\sum_i\phi_ie_i-\lambda V^{-1}\bar\theta\right\|_V\\&\le\frac{1}{n}\left\|\lambda V^{-1}\bar\theta\right\|_V+\left\|\frac{1}{n}V^{-1}\sum_i\phi_i\eta_i\right\|_V+\left\|\frac{1}{n}V^{-1}\sum_i\phi_ie_i\right\|_V\end{aligned}$$

The first term is bounded above as $\frac{1}{n}\|\lambda V^{-1}\bar\theta\|_V\le\sqrt{\frac{\lambda\|\bar\theta\|^2}{n}}$. For the second term, we appeal to Lemma 5 to show that

$$\begin{aligned}\left\|\frac{1}{n}V^{-1}\sum_i\phi_i\eta_i\right\|_V^2&=\left\|\frac{1}{n}V^{-1/2}\sum_i\phi_i\eta_i\right\|^2\\&\le\frac{C_1d+C_2\sqrt{d\log(1/\delta)}+C_3\log(1/\delta)}{n}\end{aligned}$$

conditional on $x_{1:n}$ and $a_{1:n}$ for constants $C_1,C_2,C_3>0$ defined in Lemma 5 with probability at least $1-\delta$. For the third term, we note that the expectation inside the norm is zero and use Lemma 9 to show concentration, yielding:

$$\begin{aligned}\left\|\frac{1}{n}V^{-1}\sum_i\phi_ie_i\right\|_V&\le 64(1+2\|\bar\theta\|)\sqrt{d/n}\|V^{-1/2}\|\cdot\log(2d/\delta)\\&\le C_4\sqrt{d/n}\|V^{-1/2}\|\cdot\log(2d/\delta)\end{aligned}$$

with probability at least $1-\delta$ for a constant $C_4=192$ since $\|\bar\theta\|\le 1$ by assumption.

Therefore, by the union bound, we are able to conclude that

$$\|\hat\theta-\bar\theta\|_V\le\sqrt{\frac{\lambda\|\bar\theta\|^2}{n}}+C_4\sqrt{\frac{d}{n}}\|V^{-1/2}\|\cdot\log(4d/\delta)+\sqrt{\frac{C_1d+C_2\sqrt{d\log(4d/\delta)}+C_3\log(4d/\delta)}{n}}$$

with probability at least $1-\delta$ for $C_4=192$ and $C_{1:3}$ are defined in Lemma 4. □

Proposition 1 ensures the validity of choosing

$$\zeta_k(\delta) = \sqrt{\frac{\lambda}{n}} + 192\sqrt{\frac{d_k}{n}}\|V_k^{-1/2}\| \cdot \log(4d_k/\delta) + \sqrt{\frac{5d_k + 10\sqrt{d_k \log(4d_k/\delta)} + 10\log(4d_k/\delta)}{n}}$$

Note that we have used the assumption that $\|\bar{\theta}_k\| \leq 1$.

E.2. An Improved Analysis of the SLOPE Estimator

In order to simplify notation, we will denote $f(\pi) = \mathbb{E}_X f(X, \pi(X))$ and $\phi(\pi) = \mathbb{E}_X \phi(X, \pi(X))$ for deterministic policies π and features ϕ.

Algorithm 3 relies on the validity of a version of the SLOPE estimator introduced by (Su et al., 2020). Recall that we construct the following value estimators:

$$\hat{v}_k(\pi) = \mathbb{E}_X \left\langle \phi(X, \pi(X)), \hat{\theta}_k \right\rangle$$

Proposition 1 ensures that they satisfy the following guarantee.

Lemma 7. *Let the event of Proposition 1 hold for all model classes $k \in [M]$. Then, for any k and policy π,*

$$|\hat{v}_k(\pi) - f(\pi)| \leq \tilde{\epsilon}_k + \zeta_k(\delta) \cdot \mathbb{E}_X \max_a \|\phi(X, a)\|_{V_k^{-1}}$$

Proof. The event ensures that $\|\hat{\theta}_k - \bar{\theta}_k\| \leq \zeta_k(\delta)$. This holds for all model classes with probability at least $1 - M\delta$. Therefore, we have

$$\begin{aligned}
\hat{v}_k(\pi) - f(\pi) &= \left\langle \phi_k(\pi), \hat{\theta}_k \right\rangle - f(\pi) \\
&= \left\langle \phi_k(\pi), \hat{\theta}_k \right\rangle - \langle \phi_k(\pi), \bar{\theta}_k \rangle + \langle \phi_k(\pi), \bar{\theta}_k \rangle - f(\pi) \\
&\leq \left\langle \phi_k(\pi), \hat{\theta}_k \right\rangle - \langle \phi_k(\pi), \bar{\theta}_k \rangle + \sqrt{\mathcal{C}(\mu)\mathbb{E}_\mu \left(\langle \phi_k(X, A), \bar{\theta}_k \rangle - f(X, A) \right)^2} \\
&\leq \tilde{\epsilon}_k + \|\hat{\theta}_k - \bar{\theta}_k\|_{V_k} \cdot \mathbb{E}_X \max_a \|\phi_k(X, a)\|_{V_k^{-1}} \\
&\leq \tilde{\epsilon}_k + \zeta_k(\delta) \cdot \mathbb{E}_X \max_a \|\phi_k(X, a)\|_{V_k^{-1}}
\end{aligned}$$

where we have again used Jensen's inequality, concentrability in Definition 1, and Cauchy-Schwarz. □

Next, we verify that $\tilde{\epsilon}_k$ is decreasing in k while $\zeta_k(\delta) \cdot \mathbb{E}_X \max_a \|\phi(X, a)\|_{V_k^{-1}}$.

Lemma 8. *The following conditions hold for all $k \in [M - 1]$:*

1. $\tilde{\epsilon}_k \geq \tilde{\epsilon}_{k+1}$ *and*
2. $\zeta_k(\delta) \cdot \mathbb{E}_X \max_a \|\phi_k(X, a)\|_{V_k^{-1}} \leq \zeta_{k+1}(\delta) \cdot \mathbb{E}_X \max_a \|\phi_{k+1}(X, a)\|_{V_{k+1}^{-1}}$.

Proof. The first condition is trivially true since for any $\theta \in \mathbb{R}^{d_k}$ we have $\theta' \in \mathbb{R}^{d_{k+1}}$, which equals θ in the top d_k coordinates and is zero in the bottom $d_{k+1} - d_k$ coordinates since the model classes are nested. This at least achieves the excess risk of θ and therefore $\tilde{\epsilon}_k \geq \tilde{\epsilon}_{k+1}$.

For the second condition, observe that one immediately has $\xi_k(\delta) \leq \xi_{k+1}(\delta)$. It suffices to show that the second factor is also increasing. Lemma 11 shows that in general for nested vectors and positive definite matrices:

$$\|\phi_k(X, a)\|_{V_k^{-1}} \leq \|\phi_{k+1}(X, a)\|_{V_{k+1}^{-1}}, \tag{53}$$

proving the claim. □

We are now ready to prove that the SLOPE estimator from Algorithm 3 is adaptive. Note that this proof is done in general and may be of independent interest as it requires fewer assumptions than that of (Su et al., 2020). Consider estimators $\hat{v}_1, \dots, \hat{v}_M$ of a quantity $v \in \mathbb{R}$ and define

$$\hat{k} = \min\{k : |\hat{v}_k - \hat{v}_\ell| \leq 2\xi_k, \quad \forall \ell > k\}$$

Theorem 6. *Let $\hat{v}_1, \dots, \hat{v}_M$ be estimators of a quantity $v \in \mathbb{R}$ with parameters $(\psi_k)_{k\in[M]}$ and $(\xi_k)_{k\in[M]}$ satisfying*

1. $|\hat{v}_k - v| \leq \psi_k + \xi_k$ *for all* $k \in [M]$
2. $\psi_k \geq \psi_{k+1}$ *for all* $k \in [M-1]$
3. $\xi_k \leq \xi_{k+1}$ *for all* $k \in [M-1]$

Then, the estimator $\hat{v}$ defined above satisfies

$$|\hat{v} - v| \leq C \min_k \{\psi_k + \xi_k\}$$

where $C = 5$.

Proof. Let $k_* = \arg\min_{k\in[M]} \{\psi_k + \xi_k\}$. To prove the claim, we handle to cases: (1) $\hat{k} < k_*$ and (2) $\hat{k} > k_*$. Otherwise, the selection is already correct. In the first case, we have that $\hat{k}$ intersects all intervals above it including k_*. Therefore

$$\begin{aligned}|v - \hat{v}_{\hat{k}}| &\leq |\hat{v}_{\hat{k}} - \hat{v}_{k_*}| + |v - \hat{v}_{k_*}| \\ &\leq 2\xi_{\hat{k}} + 2\xi_{k_*} + \psi_{k_*} + \xi_{k_*} \\ &\leq 5\left(\psi_{k_*} + \xi_{k_*}\right)\end{aligned}$$

For the second case, we have that $i = \hat{k} - 1$, which satisfies $i_* \geq i$ does intersect with some $j \in [\hat{k}, M]$. Therefore

$$2\xi_i + 2\xi_j \leq |\hat{v}_i - \hat{v}_j| \leq \psi_i + \xi_i + \psi_j + \xi_j$$

by definition. It follows then that

$$\xi_i + \xi_j \leq \psi_i + \psi_j \leq 2\psi_{k_*}$$

since $k_* \leq i, j$. Therefore,

$$\begin{aligned}|v - \hat{v}_{\hat{k}}| &\leq \psi_{\hat{k}} + \xi_{\hat{k}} \\ &\leq \psi_{k_*} + \xi_j \\ &\leq \psi_{k_*} + \xi_j + \xi_i \\ &\leq 3\psi_{k_*} \\ &\leq 3\left(\psi_{k_*} + \xi_{k_*}\right)\end{aligned}$$

Since all cases have been handled, we see that the claim is satisfied with $C = 5$. □

E.3. Proof of Model Selection Guarantee

Equipped with the single model class guarantees and the SLOPE estimator guarantees, we are ready to prove the final result, which is restated below.

Theorem 4. *Let $\mathcal{F}_1, \dots, \mathcal{F}_M$ be a nested collection of linear model classes. For $\lambda_k = 1$ for all $k \in [M]$, Algorithm 3 outputs a policy $\hat{\pi}$ such that, with probability at least $1 - 4\delta$, for any comparator policy π, Reg$(\pi, \hat{\pi})$ is bouned above by*

$$12 \min_{k\in[M]} \left\{\tilde{\epsilon}_k + \zeta_k(\delta/M) \cdot \mathbb{E}_X \max_a \|\phi_k(X, a)\|_{V_k^{-1}}\right\}.$$

Proof of Theorem 4. Recall that Proposition 1 guarantees that $\|\hat{\theta}_k - \bar{\theta}\|_V \leq \zeta_k(\delta)$ for all k with probability at least $1 - M\delta$. From here on, assume this event holds. In order to derive the results, we first note that Lemma 8 ensures the ordering properties of ϵ_k and $\zeta_k(\delta) \cdot \mathbb{E} \max \|\phi_k(X,a)\|_{V_k^{-1}}$. Therefore, it is valid to apply Theorem 6 with $\psi_k = \tilde{\epsilon}_k$ and $\xi_k = \zeta_k \cdot \mathbb{E} \max \|\phi_k(X,a)\|_{V_k^{-1}}$.

Note that the selection rule ensures that $\hat{\pi} \equiv \hat{\pi}_{\hat{\ell}}$ where $\hat{\ell} = \arg\max_\ell \hat{v}(\hat{\pi}_\ell)$. From the regret decomposition, we have, for any fixed $\ell \in [M]$,

$$
\begin{aligned}
\mathrm{Reg}(\pi, \hat{\pi}) &= f(\pi) - f(\hat{\pi}) \\
&= f(\pi) - \hat{v}(\hat{\pi}) + \hat{v}(\hat{\pi}) - f(\hat{\pi}) \\
&\leq f(\pi) - \hat{v}(\hat{\pi}_{\hat{\ell}}) + C' \min_k \left\{ \tilde{\epsilon}_k + \zeta_k(\delta) \cdot \mathbb{E} \max \|\phi_k(X,a)\|_{V_k^{-1}} \right\} \\
&\leq f(\pi) - \hat{v}(\hat{\pi}_\ell) + C' \min_k \left\{ \tilde{\epsilon}_k + \zeta_k(\delta) \cdot \mathbb{E} \max \|\phi_k(X,a)\|_{V_k^{-1}} \right\} \\
&= f(\pi) - f(\hat{\pi}_\ell) + f(\hat{\pi}_\ell) - \hat{v}(\hat{\pi}_\ell) + C' \min_k \left\{ \tilde{\epsilon}_k + \zeta_k(\delta) \cdot \mathbb{E} \max \|\phi_k(X,a)\|_{V_k^{-1}} \right\} \\
&\leq f(\pi) - f(\hat{\pi}_\ell) + 2C' \min_k \left\{ \tilde{\epsilon}_k + \zeta_k(\delta) \cdot \mathbb{E} \max \|\phi_k(X,a)\|_{V_k^{-1}} \right\} \\
&= f(\pi) - f(\hat{\pi}_\ell) + 2C' \min_k \left\{ \tilde{\epsilon}_k + \zeta_k(\delta) \cdot \mathbb{E} \max \|\phi_k(X,a)\|_{V_k^{-1}} \right\} \\
&\leq \tilde{\epsilon}_\ell + 2\zeta_\ell(\delta) \cdot \mathbb{E} \max \|\phi_\ell(X,a)\|_{V_\ell^{-1}} + 2C' \min_k \left\{ \tilde{\epsilon}_k + \zeta_k(\delta) \cdot \mathbb{E} \max \|\phi_k(X,a)\|_{V_k^{-1}} \right\}
\end{aligned}
$$

where the first and third inequalities follow from Theorem 6 with the constant $C' = 5$ and the second uses the selection rule of $\hat{\ell}$. The last inequality uses Proposition 1. Therefore,

$$
\mathrm{Reg}(\pi, \hat{\pi}) \leq (2C+1)\tilde{\epsilon}_\ell + 2(C+1)\zeta_\ell(\delta) \cdot \mathbb{E} \max \|\phi_\ell(X,a)\|_{V_\ell^{-1}}
$$

Recall that $\ell \in [M]$ was arbitrary and the assumed event occurs with probability at least $1 - M\delta$. The proof of the claim is completed by a change of variables $\delta' = M\delta$. Therefore, the claim is satisfied by choosing $C = 12$.

□

E.4. Technical Lemmas

Lemma 9. *With probability at least* $1 - \delta$*,*

$$
\left\| V^{-1/2} \sum_i \phi_i e_i \right\| \leq C(1 + 2\|\bar{\theta}\|)\sqrt{nd}\|V^{-1/2}\| \cdot \log(2d/\delta) \tag{54}
$$

where $C = 64$

Proof. Note that we have $\mathbb{E}_\mu[\phi_i e_i] = \mathbb{E}_\mu\left[\phi_i(f_i - \phi_i^\top \bar{\theta})\right] = \mathbb{E}_\mu \phi(X,A) f(X,A) - \Sigma\bar{\theta}$ which equals zero by first order optimality conditions applied to the minimizer $\bar{\theta}$. Therefore it suffices to show concentration of $\sum_i \Sigma^{-1/2}\phi_i e_i$ around its mean.

Define $\tilde{\phi}_i = \Sigma^{-1/2}\phi_i$ and define $\tilde{\phi}_i^j$ as the jth coordinate of the sample $\tilde{\phi}_i$.

Note that $Z_i^j := \tilde{\phi}_i^j \left(f_i - \tilde{\phi}_i^\top \Sigma^{1/2}\bar{\theta} \right)$ is sub-exponential with parameter $\|Z_i^j\|_{\psi_1} \leq C_1 + C_2\|\Sigma^{1/2}\bar{\theta}\| \leq C_1 + 2C_2\|\bar{\theta}\|$ where $C_1 = 1$ and $C_2 = 1$ $C_1, C_2 > 0$ by Lemma 1 and Lemma 2. This follows because $\|\tilde{\phi}_i\|_{\psi_2} \leq 1$ and $f_i \in [-1, 1]$ by assumption. Therefore by Bernstein's inequality and multiplying by $\Sigma^{-1/2}$,

$$
\begin{aligned}
|\sum_i \phi_i^j e_i| &\leq 32\sqrt{(1 + 2\|\bar{\theta}\|)^2 n \cdot \log(2/\delta)} + 32(1 + 2\|\bar{\theta}\|)\log(2/\delta) \\
&\leq 64(1 + 2\|\bar{\theta}\|)\sqrt{n} \cdot \log(2/\delta)
\end{aligned}
$$

with probability at least $1 - \delta$. We have used the fact that $\delta \leq 1/e$ so that $\sqrt{\log(2/\delta)} \leq \log(2/\delta)$. Taking the union bound over all coordinates $j \in [d]$ and applying standard norm inequalities, we have

$$\| \sum_i \phi_i e_i \| \leq 64(1 + 2\|\bar{\theta}\|)\sqrt{nd} \cdot \log(2/\delta)$$

with probability at least $1 - d\delta$ by the union bound. The result then follows by change of variables with $\delta' = d\delta$ and applying the ℓ_2 matrix norm inequality. □

F. Proof of Theorem 5

We define the expected regression loss as $L_k(\theta) = \mathbb{E}\hat{L}_k(\theta)$ for $\theta \in \mathbb{R}^{d_k}$. We first require the following concentration result which is immediate from the selection rule via Bernstein's inequality.

Lemma 10. *There is a constant $C > 0$ such that with probability at least $1 - \delta$,*

$$|\hat{L}_k(\hat{\theta}_k) - L_k(\hat{\theta}_k)| \leq C\sqrt{\frac{(1 + \|\hat{\theta}_k\|)^4}{n_{out}}} \cdot \log(2M/\delta)$$

for all $k \in [M]$.

Proof. Define $Z_{k,i} = \left\langle \phi_{k,i}, \hat{\theta}_k \right\rangle - y_i$. Note that $\|Z_{k,i}\|_{\psi_2} \leq 2\|\hat{\theta}_k\| + 2$ since $\|\Sigma^{-1/2}\phi_{k,i}\|_{\psi_2} \leq 1$ and $\|y_i\|_{\psi_2} \leq 2$. Therefore $\|Z_{k,i}^2\|_{\psi_1} \leq (2 + 2\|\hat{\theta}_k\|)^2$. By Bernstein's inequality, we have

$$\begin{aligned}
|\hat{L}_k(\hat{\theta}_k) - L_k(\hat{\theta}_k)| &= |\frac{1}{n_{out}} \sum_i Z_{k,i}^2 - \mathbb{E}[Z_{k,i}^2]| \\
&\leq 32\sqrt{\frac{4(2 + 2\|\hat{\theta}_k\|)^4 \log(2M/\delta)}{n_{out}}} + \frac{64(2 + 2\|\hat{\theta}_k\|)^2 \log(2M/\delta)}{n_{out}}
\end{aligned}$$

with probability at least $1 - \delta$ for all k. It is assumed that $\delta \leq 1/e$. Therefore, $\frac{1}{n_{out}} \leq \frac{1}{\sqrt{n_{out}}}$ and $\sqrt{\log(M/\delta)} \leq \log(M/\delta)$. Applying these two upper bounds to the terms above and then summing them gives the result.

□

Armed with this result, we turn to the proof of Theorem 5, restated below.

Theorem 5. *Given arbitrary linear model classes $\mathcal{F}_1, \ldots, \mathcal{F}_M$, let $\hat{\pi} = \hat{\pi}_{\hat{k}}$ where $\hat{k} \in \arg\min_{k \in [M]} \hat{L}_k(\hat{\theta}_k)$. Then, there is a constant $C > 0$ such that, with probability at least $1 - 2\delta$, Reg$(\pi, \hat{\pi})$ is bounded above by*

$$\begin{aligned}
&\min_k \left\{ \tilde{\epsilon}_k + C\sqrt{\mathcal{C}(\mu)}\|\hat{\theta}_k - \bar{\theta}_k\|_{\Sigma_k} \right\} \\
&\quad + \mathcal{O}\left(\sqrt{\mathcal{C}(\mu)} \cdot \frac{(1 \vee \max_\ell \|\hat{\theta}_\ell\|) \log^{1/2}(M/\delta)}{n_{out}^{1/4}} \right).
\end{aligned}$$

Proof of Theorem 5. Recall that $Y(A) = f(X, A) + \eta(A)$ is the observed random reward taking action A where η is the noise vector. By linearity of expectation, for any $\hat{k} \in [M]$,

$$\begin{aligned}
L_{\hat{k}}(\theta) &= \mathbb{E}_\mu \left(\langle \phi_{\hat{k}}(X, A), \theta \rangle - Y(A) \right)^2 \\
&= \mathbb{E}_\mu \left(\langle \phi_{\hat{k}}(X, A), \theta \rangle - f(X, A) + f(X, A) - Y(A) \right)^2 \\
&= \mathbb{E}_\mu \left(\langle \phi_{\hat{k}}(X, A), \theta \rangle - f(X, A) \right)^2 + \mathbb{E}_\mu \left(f(X, A) - Y(A) \right)^2 \\
&\quad + 2\mathbb{E}_\mu \left(\langle \phi_{\hat{k}}(X, A), \theta \rangle - f(X, A) \right) \left(f(X, A) - Y(A) \right) \\
&= \mathbb{E}_\mu \left(\langle \phi_{\hat{k}}(X, A), \theta \rangle - f(X, A) \right)^2 + \mathbb{E}_\mu \left(f(X, A) - Y(A) \right)^2
\end{aligned}$$

Applying the tower rule of the expectation to the last term conditioned on (X, A) yields the above results since η is zero-mean independent noise. Then,

$$
\begin{aligned}
\mathbb{E}_\mu \left(\left\langle \phi_{\hat{k}}(X, A), \hat{\theta}_{\hat{k}} \right\rangle - f(X, A) \right)^2 &= L_{\hat{k}}(\hat{\theta}_{\hat{k}}) - \mathbb{E}_\mu \left(f(X, A) - Y(A) \right)^2 && (55) \\
&\le \hat{L}_{\hat{k}}(\hat{\theta}_{\hat{k}}) - \mathbb{E}_\mu \left(f(X, A) - Y(A) \right)^2 + C \sqrt{\frac{(1 + \|\hat{\theta}_{\hat{k}}\|)^4}{n_{out}}} \cdot \log(M/\delta) && (56) \\
&\le \hat{L}_k(\hat{\theta}_k) - \mathbb{E}_\mu \left(f(X, A) - Y(A) \right)^2 + C \sqrt{\frac{(1 + \max_\ell \|\hat{\theta}_\ell\|)^4}{n_{out}}} \cdot \log(M/\delta) && (57) \\
&\le L_k(\hat{\theta}_k) - \mathbb{E}_\mu \left(f(X, A) - Y(A) \right)^2 + 2C \sqrt{\frac{(1 + \max_\ell \|\hat{\theta}_\ell\|)^4}{n_{out}}} \cdot \log(M/\delta) && (58) \\
&\le \mathbb{E}_\mu \left(\left\langle \phi_k(X, A), \hat{\theta}_k \right\rangle - f(X, A) \right)^2 + 2C \sqrt{\frac{(1 + \max_\ell \|\hat{\theta}_\ell\|)^4}{n_{out}}} \cdot \log(M/\delta) && (59)
\end{aligned}
$$

with probability at least $1 - \delta$. The first inequality follows from Lemma 10. The second inequality follows from the choice of $\hat{k}$ to minimize $\hat{L}_k(\hat{\theta}_k)$.

Therefore, we can apply the following regret bound for any $k \in [M]$:

$$
\begin{aligned}
\mathrm{Reg}(\pi, \hat{\pi}) &= f(\pi) - f(\hat{\pi}) \\
&\le f(\pi) - \left\langle \phi(\hat{\pi}), \hat{\theta}_{\hat{k}} \right\rangle + \left\langle \phi(\hat{\pi}), \hat{\theta}_{\hat{k}} \right\rangle - f(\hat{\pi}) \\
&\le f(\pi) - \left\langle \phi(\pi), \hat{\theta}_{\hat{k}} \right\rangle + \left\langle \phi(\hat{\pi}), \hat{\theta}_{\hat{k}} \right\rangle - f(\hat{\pi}) \\
&\le 2 \sqrt{\mathcal{C}(\mu) \mathbb{E}_\mu \left(\left\langle \phi_{\hat{k}}(X, A), \hat{\theta}_{\hat{k}} \right\rangle - f(X, A) \right)^2} \\
&\le 2 \sqrt{\mathcal{C}(\mu) \mathbb{E}_\mu \left(\left\langle \phi_k(X, A), \hat{\theta}_k \right\rangle - f(X, A) \right)^2 + 2C\mathcal{C}(\mu) \sqrt{\frac{(1 + \max_\ell \|\hat{\theta}_\ell\|)^4}{n_{out}}} \cdot \log(M/\delta)} \\
&\le 2 \sqrt{\mathcal{C}(\mu) \mathbb{E}_\mu \left(\langle \phi_k(X, A), \bar{\theta}_k \rangle - f(X, A) \right)^2} + 2 \sqrt{\mathcal{C}(\mu)} \|\hat{\theta}_k - \bar{\theta}_k\|_{\Sigma_k} \\
&\quad + 2 \sqrt{\mathcal{C}(\mu)} \left(\frac{C_1 \left(1 + \max_\ell \|\hat{\theta}_\ell\| \right)^4 \log^2(M/\delta)}{n_{out}} \right)^{1/4} \\
&= \tilde{\epsilon}_k + 2 \sqrt{\mathcal{C}(\mu)} \|\hat{\theta}_k - \bar{\theta}_k\|_{\Sigma_k} + 2 \sqrt{\mathcal{C}(\mu)} \left(\frac{C_1 \left(1 + \max_\ell \|\hat{\theta}_\ell\| \right)^4 \log^2(M/\delta)}{n_{out}} \right)^{1/4}
\end{aligned}
$$

Note that the third inequality follows from applying Jensen's inequality and Definition 1. The fourth inequality applies the previous display.

□

Balancing approximation error and coverage. We conclude by remarking that a final case may be considered when we ignore the model selection criterion of statistical complexity and aim to balance only approximation error and coverage. In this case, the problem becomes trivial since we are "permitted" to take arbitrarily large model classes until realizability is achieved.

G. Supporting Lemmas

In this section, we state several independent results that support the proofs of the main results.

G.1. Nestedness Properties

Let $M \in \mathbb{R}^{d \times d}$ be a positive definite matrix of the form

$$M = \begin{bmatrix} A & B \\ B^\top & D \end{bmatrix} \tag{60}$$

where $A \in \mathbb{R}^{d_1 \times d_1}$ is also a positive definite matrix and $D \in \mathbb{R}^{d_2 \times d_2}$ and $B \in \mathbb{R}^{d_1 \times d_2}$.

Lemma 11. *Let $a \in \mathbb{R}^{d_1}$ and $b \in \mathbb{R}^{d_2}$ be arbitrary. The following inequality holds:*

$$\begin{bmatrix} a \\ b \end{bmatrix}^\top M^{-1} \begin{bmatrix} a \\ b \end{bmatrix} \geq a^\top A^{-1} a \tag{61}$$

Proof. By Schur complement inverse rules:

$$\begin{aligned}
\begin{bmatrix} a \\ b \end{bmatrix}^\top M^{-1} \begin{bmatrix} a \\ b \end{bmatrix} &= a^\top A^{-1} a + a^\top A^{-1} B (M/A)^{-1} B^\top A^{-1} a - 2b^\top (M/A)^{-1} B^\top A^{-1} a + b^\top (M/A)^{-1} b && (62) \\
&\geq a^\top A^{-1} a + a^\top A^{-1} B (M/A)^{-1} B^\top A^{-1} a - a^\top A^{-1} B (M/A)^{-1} B^\top A^{-1} a && (63) \\
&= a^\top A^{-1} a && (64)
\end{aligned}$$

where the inequality follows from optimizing over b. □

G.2. Concentration of Quadratic Forms

The following is a restatement of Lemma 14 of (Hsu et al., 2012b) for convenience, which can be interpreted as a version of the Hanson-Wright inequality (Rudelson & Vershynin, 2013).

Lemma 12. *Let $A \in \mathbb{R}^{m \times n}$ be a matrix and $\Sigma = AA^\top$. Let $X \in \mathbb{R}^n$ be a random vector with independent coordinates $X_1, \ldots, X_n$ such that $\mathbb{E} X_i = 0$ and $\mathbb{E} \exp(\lambda X_i) \leq \mathbb{E} \exp\left(\frac{\lambda^2 \sigma^2}{2}\right)$ for all λ^2. Then,*

$$P\left(\|AX\|^2 \leq \sigma^2 \operatorname{tr} \Sigma + 2\sigma^2 \sqrt{\operatorname{tr}(\Sigma^2) \log(1/\delta)} + 2\sigma^2 \|\Sigma\| \log(1/\delta) \right) \leq \delta \tag{65}$$

H. Additional Experiment Details

In this section, we provide some additional details regarding the experimental results presented in Section 6. We start with details that are common to both of the settings considered. In order to evaluate the performance of algorithms, within each trial, we generated a test set of $n_{\text{test}} = 500$ samples. All algorithms were thus compared on the same data within a trial. For both the batch dataset and the test set, noise was artificially generated on rewards by sampling from a standard normal distribution $\mathcal{N}(0, 1)$ such that $\eta(a) \sim \operatorname{subG}(1)$ for all $a \in \mathcal{A}$. Regret was computed by taking the difference between the optimal policy π_* and the learned policy evaluated on the same test set. Thus, the points approximately (up to noise) represent $\operatorname{Reg}(\pi_*, \hat{\pi}_n)$ where $\hat{\pi}_n$ is the learned policy after n batch samples.

The data collection policy was generated as a policy independent of the observed state. Thus $\mu(a|x) = \mu(a'|x)$ for all a, a', x. We generated the probabilities of sampling arms by sampling from a standard Dirichlet distribution of $|\mathcal{A}|$ values. For the algorithms, penalization terms (i.e. the estimation error) typically depends on constants being chosen sufficiently large to ensure a confidence interval is valid. However, choosing large values in practice can lead to unnecessarily poor convergence. We found that multiplying by $C = 0.1$ yielded good performance in most settings.

H.1. Complexity-Coverage Setting

In this section, all random quantities were generated by sampling multivariate normal distributions. We first generated a random vector $(f(x, a))_{x \in \mathcal{X}, a \in \mathcal{A}}$, which specifies the average reward for each state-action pair. In order to generate a set of linear models (feature maps) that all satisfy realizability, we began with an input d_{hid} and randomly generated $d_{hid} - 1$ vectors $v_1, \ldots, v_{d_{hid}-1}$ of length $|\mathcal{X}||\mathcal{A}|$ and solved for the last $v_{d_{hid}}$ by subtracting these off the reward. This ensures a particular linear combination of the vs equals the vector $(f(x, a))_{x \in \mathcal{X}, a \in \mathcal{A}}$. This procedure was repeated for various values of d_{hid} and the resulting feature vectors were scaled up to $d = |\mathcal{X}||\mathcal{A}|$ by multiplying by a random matrix A with elements generated from $\mathcal{N}(0, 1)$. This ensures that the feature maps are not simply equivalent linear transformations of each other.

H.2. Approximation-Complexity Setting

In contrast the previous setting, we considered an infinite state space where, for each action, a $d = 100$ dimensional covariate vector is sampled from a multivariate normal distribution with mean 0 and covariance matrix Σ_a, where Σ_a was also randomly generated. As mentioned in the main text, we constructed model classes by truncating the original covariate vector to small dimensions, thus inducing a nested structure. Since $d_* = 30$, some of these choices result in misspecified models. For the SLOPE method, in order to estimated the predicted values to generate each $\hat{v}_k$, we used a validation set of unlabeled samples (i.e. no revealed reward).