Experiment Planning with Function Approximation

Aldo Pacchiano
Broad Institute & Boston University
apacchia@broadinstitute.org

Jonathan N. Lee
Stanford University
jnl@stanford.edu

Emma Brunskill
Stanford University
ebrun@cs.stanford.edu

Abstract

We study the problem of experiment planning with function approximation in contextual bandit problems. In settings where there is a significant overhead to deploying adaptive algorithms—for example, when the execution of the data collection policies is required to be distributed, or a human in the loop is needed to implement these policies—producing in advance a set of policies for data collection is paramount. We study the setting where a large dataset of contexts but not rewards is available and may be used by the learner to design an effective data collection strategy. Although when rewards are linear this problem has been well studied [53], results are still missing for more complex reward models. In this work we propose two experiment planning strategies compatible with function approximation. The first is an eluder planning and sampling procedure that can recover optimality guarantees depending on the eluder dimension [42] of the reward function class. For the second, we show that a uniform sampler achieves competitive optimality rates in the setting where the number of actions is small. We finalize our results introducing a statistical gap fleshing out the fundamental differences between planning and adaptive learning and provide results for planning with model selection.

1 Introduction

Data-driven decision-making algorithms have achieved impressive empirical success in various domains such as online personalization [4, 48], games [36, 43], dialogue systems [32] and robotics [25, 35]. In many of these decision-making scenarios, it is often advantageous to consider contextual information when making decisions. This recognition has sparked a growing interest in studying adaptive learning algorithms in the setting of contextual bandits [26, 33, 6] and reinforcement learning (RL) [46]. Adaptive learning scenarios involve the deployment of data collection policies, where learners observe rewards or environment information and utilize this knowledge to shape subsequent data collection strategies. Nonetheless, the practical implementation of adaptive policies in real-world experiments currently presents significant challenges. First, there is significant infrastructure requirements and associated overheads which require skills and resources many organizations lack. For example, while there are end-user services that enable organizations to automatically send different text messages to different individuals, such services typically do not offer adaptive bandit algorithms. Second, the resulting reward signal may be significantly delayed. As an example, the effect of personalized health screening reminders on a patient making a doctors appointments may take weeks. Therefore, while it is increasingly recognized that there exist other settings where context-specific policies are likely to be beneficial, including behavioural science[8], many organizations working in these settings may find it infeasible to run experiments with adaptive policies. To bridge this gap, there is a need to explore the deployment of non-adaptive or static strategies that can effectively collect data with minimal or no updates. Surprisingly, limited research has been conducted to investigate this particular setting, highlighting the importance of further exploration in this area.

In this work, we consider how to design a static experimental sampling strategy for contextual multi-armed bandit settings which, when executed, will yield a dataset from which we can compute a near-optimal contextual policy (one with small simple regret). Our framework closely adheres to the experiment planning problem formulation originally introduced by [53]. The problem involves a learner interacting with a contextual bandit problem where the reward function is unknown. Our assumption is that the learner possesses a substantial offline dataset comprising of m sample contexts, none of which include reward information. The learner's objective is to utilize this data to design a static policy sequence of length T (where $m \geqslant T$) that enables the collection of valuable reward information when deployed in the real world. The ultimate goal is to produce an almost optimal policy using the data gathered during deployment. To address this scenario, the authors of [53] propose a two-phase approach involving Planner and Sampler algorithms. In the planner phase, the learner employs the context data to generate a set of sampling policies, which are then utilized in the sampler phase. The primary focus of [53] lies in the analysis of the linear case, assuming the reward function to be a linear function of the context vectors. Their key algorithmic contribution is the linear Planner Algorithm, which encompasses a reward-free instance of LinUCB. The analysis presented in [53] demonstrates that the required number of samples under the static sampler to achieve an ε-optimal policy (also referred to as simple regret) scales as $\mathcal{O}\left(d^2/\varepsilon^2\right)$, where d represents the dimension of the underlying space. This result matches the online minimax sample complexity for the simple regret problem [11, 1].

The algorithm proposed by [53] effectively constructs a static policy sequence of adequate span, utilizing the linear structure of the problem. However, in many problem settings, linearity alone may not be sufficient to capture the true nature of the underlying reward model. This limitation is particularly evident in scenarios such as genetic perturbation experimentation [40] or other similar contexts. In such cases, the availability of algorithms that do not rely on linearity becomes crucial. Unfortunately, extending the results and techniques presented in [53] to the generic function approximation regime is not straightforward.

Here we consider when the reward function is realized by an unknown function $f_\star$ belonging to a known function class $\mathcal{F}$ (which can be more complex than linear).

Adaptive Learning with Function Approximation. In adaptive learning settings where a learner can change the data collection policies as it observes rewards and has much more flexibility than in experiment planning scenarios, various adaptive learning procedures compatible with generic function approximation have been proposed for contextual bandit problems. Among these, two significant methods relevant to our discussion are the Optimistic Least Squares algorithm (OptLS) introduced by [42] and the SquareCB algorithm introduced by [17]. Both of these methods offer guarantees for cumulative regret. Specifically, the cumulative regret of OptLS scales as $\mathcal{O}(\sqrt{d_{\mathrm{eluder}} \log(|\mathcal{F}|)T})$, while the cumulative regret of $\mathrm{SquareCB}$ scales as $\mathcal{O}\left(\sqrt{|\mathcal{A}| \log(|\mathcal{F}|)T}\right)$, where $\mathcal{A}$ corresponds to the set of actions. The eluder dimension[1] (d_{eluder}) is a statistical complexity measure introduced by [42], which enables deriving guarantees for *adaptive* learning algorithms based on the principle of optimism in the face of uncertainty in contextual bandits and reinforcement learning [31, 22, 37, 9]. By employing an online-to-batch conversion strategy, these two methods imply that the number of samples required to achieve ε-simple regret using an adaptive algorithm is at most $\mathcal{O}\left(d_{\mathrm{eluder}} \log(|\mathcal{F}|)/\varepsilon^2\right)$ and $\mathcal{O}\left(|\mathcal{A}| \log(|\mathcal{F}|)/\varepsilon^2\right)$ respectively.

Contributions. In this paper, we address the function approximation setting within the contextual bandit (static) experiment planning problem. Although in experiment planning the data collection policies have to be produced before interacting with the environment, we establish that surprisingly the adaptive ε-simple regret rates achieved by OptLS and $\mathrm{SquareCB}$ are also attainable by a static experiment sampling strategy. To achieve this, we introduce and analyze two experimental planning algorithms. The $\mathrm{PlannerEluder}$ algorithm (see Section 4) utilizes confidence intervals derived from least squares optimization to construct a static set of policies to be executed during the sampling phase. When the algorithm has access to a sufficient number of offline samples to generate $T = \Omega\left(d_{\mathrm{eluder}} \log(|\mathcal{F}|)/\varepsilon^2\right)$ static policies, the learner can produce an ε-optimal policy denoted as $\widehat{\pi}_T$ thus matching the online-to-batch ε-simple regret rates of OptLS. Since the eluder dimension of linear function classes of dimension d is -up to logarithmic factors- of order $\mathcal{O}(d)$, our results

---

[1]We formally introduce this quantity in Section 4. Here we use a simpler notation to avoid confusion.

recover the linear experiment planning rates of [53]. Additionally, in Section 5, we demonstrate that collecting $T = \Omega\left(|\mathcal{A}|\log(|\mathcal{F}|)/\varepsilon^2\right)$ uniform samples is sufficient to obtain an ε-optimal policy $\widehat{\pi}_T$. This matches the online-to-batch ε-simple regret rates of $\mathrm{SquareCB}$. These two results yield conclusions similar to those known in the linear setting [53], but for general function classes.

These results prompt two important questions: (1) are existing adaptive cumulative regret algorithms already optimal for simple regret minimization, and (2) is static experiment planning sufficient to match the simple regret guarantees of adaptive learning for realizable contextual bandits?

In Section 6, we provide a negative answer to both questions. We present a certain structured class of bandit problems where a different adaptive algorithm can require significantly less samples than that implied by the bounds from OptLS and $\mathrm{SquareCB}$. Our result is for the *realizable* setting [2, 18, 17, 44, 9], where the true reward function $f_\star$ belongs to a known function class $\mathcal{F}$, where the true complexity of simple regret problems in adaptive learning scenarios has not be known known in this setting. This result complements recent results [34] that online-to-batch strategies are suboptimal for minimizing simple regret in the agnostic contextual bandits setting.[2]

For the second question, our statistical lower bound shows that for this class of problems, a significantly smaller number of samples are needed to find an ε-optimal policy when using adaptive learning, than when using a static policy. This result complements related work in reinforcement learning with realizable linear state-action value function approximation[52].

Finally, we address the problem of model selection when the learner is presented with a family of reward function classes $\{\mathcal{F}_i\}_{i=1}^M$ and is guaranteed that the true reward function $f_\star$ belongs to $\mathcal{F}_{i_\star}$, where the index $i_\star$ is unknown. We demonstrate that as long as $T = \Omega\left(|\mathcal{A}|\log(\max(M, |\mathcal{F}_{i_\star}|))/\varepsilon^2\right)$ uniform samples are available, it is possible to construct an ε-optimal policy denoted as $\widehat{\pi}_T$. Further details regarding these results can be found in Section 7.

2 Further Related Work

Best Arm Identification and Design of Experiments. Previous work to efficiently produce an almost optimal policy for non-contextual linear and multi-armed bandits settings is based on best-arm identification procedures [16, 20, 13, 45, 47, 51]. These are strategies that react adaptively to the collected rewards and achieve instance dependent sample complexity. Other papers in the design of experiments literature [24, 15, 29, 41] introduce methods for designing a non-adaptive policy that can be used to find an optimal arm with high probability in non-contextual scenarios.

Safe Optimal Design. When safety concerns are imposed in the exploration policy, works such as [54] and [49] have studied the problem of producing a safe and efficient static policy used to collect data that can then be used for policy evaluation in the MAB and linear settings. This is in contrast with the unconstrained contextual function approximation scenario we study.

Reward Free Reinforcement Learning. In Reinforcement Learning settings a related line of research [21, 50, 10] considers the problem of producing enough exploratory data to allow for policy optimization for all functions in a reward class. These works allow for the learner to interact with the world for a number of steps, enough to collect data that will allow for zero-shot estimation of the optimal policy within a reward family. This stands in contrast with the experiment planning setup, where the objective is to produce a sequence of policies for static deployment at test time with the objective of learning the unknown reward function.

3 Problem Definition

We study a two-phase interaction between a learner and its environment consisting of a *Planning* and a *Sampling* phase. During the planning phase the learner has access to $m \geqslant T$ i.i.d. offline

---

[2]In the agnostic case the learner is instead presented with an arbitrary policy class Π that may or may not be related to the mean reward function $f_\star$. In this setting a recent study by [34] introduces an algorithm that achieves sharp rates characterized in terms of a quantity ρ_Π.

context samples[3] from a context distribution $\mathcal{P}$ supported on a context space $\mathcal{X}$ as well as knowledge of a reward function class $\mathcal{F}$ where each element $f \in \mathcal{F}$ has domain $\mathcal{X} \times \mathcal{A}$ where $\mathcal{A}$ is the action set. During the sampling phase the learner interacts with a contextual bandit problem for T steps where at each time-step t a context $x_t \sim \mathcal{P}$ is revealed to the learner, the learner takes an action a_t and receives a reward r_t. We adopt the commonly used realizability assumption, which is widely employed in the contextual bandit literature [2, 18, 17, 44, 9].

Assumption 3.1 (Realizability)**.** *There exists a function $f_\star \in \mathcal{F}$ such that $r_t = f_\star(x_t, a_t) + \xi_t$ where $f_\star \in \mathcal{F}$ and ξ_t is a conditionally zero mean $1-$subgaussian random variable for all $t \in [T]$ during the sampling phase.*

During the planning phase the learner is required to build a static policy sequence $\{\pi_i\}_{i=1}^T$ (a policy is a mapping from $\mathcal{X}$ to $\Delta_\mathcal{A}$, the set of distributions over actions) that, without knowledge of the reward, can be deployed to collect data during the sampling phase. In contrast with adaptive learning procedures, in the experiment planning problem the policies in $\{\pi_i\}_{i=1}^T$ have to be fully specified during the planning phase and cannot depend on the learner's observed reward values during the sampling phase. The learner's objective is to use the reward data collected during the sampling phase to produce an $\varepsilon-$optimal policy $\widehat{\pi}_T$ such that,

$$\mathbb{E}_{x \sim \mathcal{P}, a \sim \widehat{\pi}_T(x)} \left[f_\star(x, a) \right] + \varepsilon \geq \max_\pi \mathbb{E}_{x \sim \mathcal{P}, a \sim \pi(x)} \left[f_\star(x, a) \right].$$

for a suboptimality parameter $\varepsilon > 0$. Because we have assumed realizability of the reward function, the optimal policy among all possible policies $\pi_\star = \arg\max_\pi \mathbb{E}_{x \sim \mathcal{P}, a \sim \pi(x)} \left[f_\star(x, a) \right]$ can be written as $\pi_\star(x) = \arg\max_{a \in \mathcal{A}} f_\star(x, a)$ for all $x \in \mathcal{X}$.

Throughout this work we make the following boundedness assumptions,

Assumption 3.2 (Bounded Function Range)**.** *The range of $\mathcal{F}$ is bounded, i.e. for all $x, a \in \mathcal{X} \times \mathcal{A}$, $|f(x, a)| \leq B$ for some constant $B > 0$.*

Assumption 3.3 (Bounded Noise)**.** *The noise variables are bounded $|\xi_t| \leq \bar{B}$ for all $t \in [T]$ for some constant $\bar{B} > 0$.*

Assumptions 3.2 and 3.3 imply $|r_t| \leq B + \bar{B}$ for all $t \in [T]$.

4 Planning Under the Eluder Dimension

The eluder dimension is a sequential notion of complexity for function classes that was originally introduced by [42]. The eluder dimension can be used to bound the regret of optimistic least squares algorithms in contextual bandits [42, 9] and reinforcement learning [37]. Informally speaking the eluder dimension is the length of the longest sequence of points one must observe to accurately estimate the function value at any other point. We use the eluder dimension definition from [42].

Definition 4.1. *($\varepsilon-$dependence) Let $\mathcal{G}$ be a scalar function class with domain $\mathcal{Z}$ and $\varepsilon > 0$. An element $z \in \mathcal{Z}$ is $\varepsilon-$dependent on $\{z_1, \cdots, z_n\} \subseteq \mathcal{Z}$ w.r.t. $\mathcal{G}$ if any pair of functions $g, g' \in \mathcal{G}$ satisfying $\sqrt{\sum_{i=1}^n (g(z_i) - g'(z_i))^2} \leq \varepsilon$ also satisfies $g(z) - g'(z) \leq \varepsilon$. Furthermore, $z \in \mathcal{Z}$ is $\varepsilon-$independent of $\{z_1, \cdots, z_n\}$ w.r.t. $\mathcal{G}$ if it is not $\varepsilon-$dependent on $\{z_1, \cdots, z_n\}$.*

Definition 4.2. *(ε-eluder) The $\varepsilon-$non monotone eluder dimension $\overline{d_{\text{eluder}}}(\mathcal{G}, \varepsilon)$ of $\mathcal{G}$ is the length of the longest sequence of elements in $\mathcal{Z}$ such that every element is $\varepsilon-$independent of its predecessors. Moreover, we define the $\varepsilon-$eluder dimension $d_{\text{eluder}}(\mathcal{G}, \varepsilon)$ as $d_{\text{eluder}}(\mathcal{G}, \varepsilon) = \max_{\varepsilon' \geq \varepsilon} \overline{d_{\text{eluder}}}(\mathcal{G}, \varepsilon)$.*

By definition the $\varepsilon-$eluder dimension increases as ε is driven down to zero. Let $\mathcal{D}$ be a dataset of $\mathcal{X}$, $\mathcal{A}$ pairs. For any f, f' with support $\mathcal{X} \times \mathcal{A}$ we use the notation $\|f - f'\|_\mathcal{D}$ to denote the data norm of the difference between functions f and f':

$$\|f - f'\|_\mathcal{D} = \sqrt{\sum_{(x,a) \in \mathcal{D}} (f(x, a) - f'(x, a))^2}.$$

---

[3]This assumption is based on the fact that gathering offline context samples can be significantly less costly compared to conducting multiple rounds of experimental deployment.

Algorithm 1 EluderPlanner

1: Input: $m \geqslant T$ samples $\{x_\ell \sim \mathcal{P}\}_{\ell=1}^m$, function class $\mathcal{F}$, confidence radius function $\beta_\mathcal{F} : t, \delta \to \mathbb{R}$.
2: Initialize data buffer $\mathcal{D}_1 = \varnothing$
3: **for** $t = 1 \ldots T$ **do**
4:     For any $x \in \mathcal{X}$ define the uncertainty radius of action $a \in \mathcal{A}$ as,

$$\omega(x, a, \mathcal{D}_t) = \max_{f, f' \in \mathcal{F} \text{ s.t. } \|f - f'\|_{\mathcal{D}_t} \leqslant 4\beta_\mathcal{F}(t,\delta)} f(x, a) - f'(x, a).$$

5:     Define the policy π_t as $\pi_t(x) = \arg\max_{a \in \mathcal{A}} \omega(x, a, \mathcal{D}_t)$ for all $x \in \mathcal{X}$.
6:     Update $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{(x_t, \pi_t(x_t))\}$
7: **end for**

Algorithm 2 Sampler

1: Input: number of time-steps T.
2: Initialize data buffer $\widetilde{\mathcal{D}}_1 = \varnothing$
3: **for** $t = 1 \ldots T$ **do**
4:     Define deployment policy $\tilde{\pi}_t$ as π_t.
5:     Observe context $\tilde{x}_t \sim \mathcal{P}$.
6:     Play action $\tilde{a}_t = \tilde{\pi}_t(\tilde{x}_t)$ and receive reward $\tilde{r}_t$.
7:     Update $\widetilde{\mathcal{D}}_{t+1} \leftarrow \widetilde{\mathcal{D}}_t \cup \{(\tilde{x}_t, \tilde{a}_t, \tilde{r}_t)\}$.
8: **end for**

The EluderPlanner algorithm takes as input $m \geqslant T$ i.i.d. samples from the context distribution $\{x_\ell \sim \mathcal{P}\}_{\ell=1}^m$ and a realizable function class $\mathcal{F}$ satisfying Assumption 3.1. The learner iterates over T out of these m samples in a sequential manner. We use the name x_t to denote the $t-$th input sample and call π_t to the (deterministic) exploration policies produced upon processing samples $1, \cdots, t-1$. We use the notation $\mathcal{D}_t = \{(x_\ell, \pi_\ell(x_\ell))\}_{\ell=1}^{t-1}$ to denote the dataset composed of the first $t-1$ (ordered) samples from $\{x_\ell\}_{\ell=1}^m$ and actions from policies $\{\pi_\ell\}_{\ell=1}^{t-1}$. We adopt the convention that $\mathcal{D}_1 = \varnothing$. The output of the EluderPlanner algorithm is the sequence of policies $\{\pi_t\}_{t=1}^T$. Policy π_t is defined as $\pi_t(x) = \arg\max_{a \in \mathcal{A}} \omega(x, a, \mathcal{D}_t)$ such that $\omega(x, a, \mathcal{D}_t)$ is an uncertainty measure defined as,

$$\omega(x, a, \mathcal{D}_t) = \max_{f, f' \in \mathcal{F} \text{ s.t. } \|f - f'\|_{\mathcal{D}_t} \leqslant 4\beta_\mathcal{F}(t,\delta)} f(x, a) - f'(x, a). \tag{1}$$

Where the confidence radius function $\beta_\mathcal{F}(\delta, t)$ equals

$$\beta_\mathcal{F}(\delta, t) = \bar{C}(B + \bar{B})\sqrt{\log\left(\frac{|\mathcal{F}|t}{\delta}\right)} \tag{2}$$

for some universal constant $\bar{C} > 0$ defined in Proposition A.7. This is a complexity radius implied by the guarantees of least squares regression (Lemma A.8) and constraints imposed by Lemma 4.6. See Appendix A.1 and Proposition A.7 for a detailed discussion about this definition. We call the combination of the EluderPlanner and Sampler as the *eluder planning algorithm*. Its performance guarantees are obtained by relating the regret of a constructed sequence of optimistic policies $\{\widetilde{\pi}_t^{\text{opt}}\}_{t=1}^T$ based on the data collected by $\{\pi_t\}_{t=1}^T$ to the sum of uncertainty radii $\sum_{t=1}^T \omega(x_t, \pi_t(x_t), \mathcal{D}_t)$ of the context-actions collected during the planning phase. Using optimism when constructing this policy sequence is important. In contrast with the sequence of greedy policies $\pi_t^{\text{greedy}}(\cdot|x) = \arg\max_{a \in \mathcal{A}} \hat{f}_t(x, a)$, where $\hat{f}_t = \arg\min_{f \in \mathcal{F}} \sum_{(x,a,r) \in \widetilde{\mathcal{D}}_t} (f(x, a) - r)^2$ resulting from a least squares over $\widetilde{\mathcal{D}}_t$, the sequence of optimistic policies satisfies a regret bound. Thus, playing a uniform policy from the sequence $\{\widetilde{\pi}_t^{\text{opt}}\}_{t=1}^T$ satisfies a simple regret guarantee. The sum of uncertainty radii $\sum_{t=1}^T \omega(x_t, \pi_t(x_t), \mathcal{D}_t)$ is bounded using a Lemma we borrow from [9].

Lemma 4.3. *[Lemma 3 from [9]] Let $\mathcal{F}$ be a function class satisfying Assumption 3.2 with ε-eluder dimension $d_{\text{eluder}}(\mathcal{F}, \varepsilon)$. For all $T \in \mathbb{N}$ and any dataset sequence $\{\bar{D}_t\}_{t=1}^\infty$ with $\bar{D}_1 = \varnothing$ and $\bar{D}_t = \{\{\bar{x}_\ell, \bar{a}_\ell\}\}_{\ell=1}^{t-1}$ of context-action pairs, the following inequality on the sum of the uncertainty radii holds,*

$$\sum_{t=1}^T \omega(\bar{x}_t, \bar{a}_t, \bar{D}_t) \leqslant \mathcal{O}\left(\min\left(BT, Bd_{\text{eluder}}(\mathcal{F}, B/T) + \beta_\mathcal{F}(T, \delta)\sqrt{d_{\text{eluder}}(\mathcal{F}, B/T)T}\right)\right).$$

Extracting Policy $\widehat{\pi}_T$ from Data. At the end of the execution of the Sampler algorithm, the learner will have access to a sequence of datasets $\{\widetilde{D}_t\}_{t=1}^T$ of contexts, actions and reward triplets. The learner will use this sequence of datasets to compute the sequence of regression functions,

$$\widehat{f}_t = \arg\min_{f \in \mathcal{F}} \sum_{\ell=1}^{t-1} (f(\widetilde{x}_\ell, \widetilde{a}_\ell) - \widetilde{r}_\ell)^2. \tag{3}$$

Standard Least Squares (LS) results (see Lemma A.8) imply that $\|\widehat{f}_t - f_\star\|_{\widetilde{\mathcal{D}}_t} \leqslant \beta(t, \delta)$ with high probability for all $t \in \mathbb{N}$. These confidence sets are used to define a sequence of optimistic (deterministic) policies $\{\widetilde{\pi}_t^{\mathrm{opt}}\}_{t=1}^T$:

$$\widetilde{\pi}_t^{\mathrm{opt}}(x) = \arg\max_{a \in \mathcal{A}} \max_{\widetilde{f} \text{ s.t. } \|\widehat{f}_t - f\|_{\widetilde{\mathcal{D}}_t} \leqslant \beta_{\mathcal{F}}(t,\delta)} \widetilde{f}(x, a) \tag{4}$$

The candidate optimal policy $\widehat{\pi}_T$ is then defined as $\widehat{\pi}_T = \mathrm{Uniform}(\widetilde{\pi}_1^{\mathrm{opt}}, \cdots, \widetilde{\pi}_T^{\mathrm{opt}})$. The main result in this section is the following Theorem.

Theorem 4.4. *Let $\varepsilon > 0$. There exists a universal constant $c > 0$ such that if*

$$T \geqslant c \frac{\max^2(B, \bar{B}, 1) d_{\mathrm{eluder}}(\mathcal{F}, B/T) \log\left(|\mathcal{F}|T/\delta\right)}{\varepsilon^2} \tag{5}$$

then with probability at least $1 - \delta$ the eluder planning algorithm's policy $\widehat{\pi}_T$ is $\varepsilon-$optimal .

Remark 4.5. *The results of this section hold for the structured bandits setting [23, 27]; that is, scenarios where $\mathcal{X} = \{\varnothing\}$.*

Both sides of the inequality defining T in Theorem 4.4 depend on T. Provided $d_{\mathrm{eluder}}(\mathcal{F}, B/T)$ is sublinear as a function of T, there exists a finite value of T satisfying Equation 5. When $d_{\mathrm{eluder}}(\mathcal{F}, \bar{\varepsilon}) = \bar{d}_{\mathrm{eluder}} \log\left(1/\bar{\varepsilon}\right)$ for some $\bar{d}_{\mathrm{eluder}}$ and for all $\bar{\varepsilon} > 0$ setting $T \geqslant c \frac{\max^2(B, \bar{B}, 1) d_{\mathrm{eluder}}(\mathcal{F}, B\varepsilon) \log(|\mathcal{F}|/\varepsilon\delta)}{\varepsilon^2}$ is enough to guarantee the conditions of Theorem 4.4 are satisfied. Examples 4 and 5 from [42] show the eluder dimension of linear or generalized linear function classes can be written in such way. In these cases $\bar{d}_{\mathrm{eluder}}$ is a constant multiple of the underlying dimension of the space.

Using standard techniques, the results of Theorem 4.4 can be adapted to the case when the function class $\mathcal{F}$ is infinite but admits a metrization and has a covering number. In this case the sample complexity will scale not with $\log(|\mathcal{F}|)$ but instead with the logarithm of the covering number of $\mathcal{F}$. For example, in the case of linear functions, the logarithm of the covering number of $\mathcal{F}$ under the ℓ_2 norm will scale as d, the ambient dimension of the space up to logarithmic factors of T (see for example Lemma D.1 of [14]). Plugging this into the sample guarantees of Theorem 4.4 recovers the $\widetilde{\mathcal{O}}\left(d^2/\varepsilon^2\right)$ sample complexity for linear experiment planning from [53].

4.1 Proof Sketch of Theorem 4.4

Let $\widetilde{x}_1, \cdots, \widetilde{x}_T$ be the sampler's context sequence. The proof works by showing that with probability at least $1 - \delta$ the regret of the $\{\widetilde{\pi}_t^{\mathrm{opt}}(\cdot)\}_{t=1}^T$ sequence satisfies the regret bound

$$\sum_{t=1}^T \max_{a \in \mathcal{A}} f_\star(\widetilde{x}_t, a) - f_\star(\widetilde{x}_t, \widetilde{\pi}_t^{\mathrm{opt}}(\widetilde{x}_t)) \leqslant \mathcal{O}\left((B + \bar{B})\sqrt{d_{\mathrm{eluder}}(\mathcal{F}, B/T)T \log(|\mathcal{F}|T/\delta)}\right)$$

A key part of the proof involves relating the balls defined by the data norms of the planner datasets $\{\mathcal{D}_t\}_{t=1}^T$ and those induced by the data norms of the sampler datasets $\{\widetilde{\mathcal{D}}_t\}_{t=1}^T$. The precise statement of this relationship is provided in Lemma 4.6, and its proof can be found in Appendix C.

Lemma 4.6. *With probability at least $1 - \frac{\delta}{12}$,*

$$\{(f, f') \text{ s.t. } \|f - f'\|_{\widetilde{\mathcal{D}}_t} \leqslant 2\beta_{\mathcal{F}}(t, \delta)\} \subseteq \{(f, f') \text{ s.t. } \|f - f'\|_{\mathcal{D}_t} \leqslant 4\beta_{\mathcal{F}}(t, \delta)\} \tag{6}$$

Simultaneously for all $t \in [T]$. Where $\{D_t\}_{t=1}^T$ is the dataset sequence resulting of the execution of EluderPlanner *while $\{\widetilde{D}_t\}_{t=1}^T$ is the dataset sequence resulting of the execution of the* Sampler *and $\beta_{\mathcal{F}}(\delta, t)$ is the confidence radius function defined in Equation 2.*

Let $\widehat{f}_t$ as in Equation 3 and define $\mathcal{E}$ as the event where the Standard Least Squares results (see Lemma A.8) hold i.e. $\|\widehat{f}_t - f_\star\|_{\widetilde{\mathcal{D}}_t} \leqslant \beta_{\mathcal{F}}(t,\delta)$ for all $t \in [T]$ and also Equation 6 from Lemma 4.6 holds for all $t \in [T]$. The results of Lemmas A.8 and 4.6 imply $\mathbb{P}(\mathcal{E}) \geqslant 1 - \frac{\delta}{6}$.

For context x and action a let's denote by $\tilde{f}_t^{x,a}$ as a function achieving the inner maximum in the definition of $\tilde{\pi}_t^{\mathrm{opt}}$ (see Equation 4). When $\mathcal{E}$ holds the policies $\{\tilde{\pi}_t^{\mathrm{opt}}\}_{t=1}^T$ are optimistic in the sense that $\tilde{f}_t^{x,\tilde{\pi}_t^{\mathrm{opt}}(x)}(x, \tilde{\pi}_t^{\mathrm{opt}}(x)) \geqslant \max_{a \in \mathcal{A}} f_\star(x,a)$ and therefore,

$$\sum_{t=1}^T \max_{a \in \mathcal{A}} f_\star(\tilde{x}_t, a) - f_\star(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t)) \overset{(i)}{\leqslant} \sum_{t=1}^T \tilde{f}_t^{x,\tilde{\pi}_t^{\mathrm{opt}}(x)}(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t)) - f_\star(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t))$$

$$\overset{(ii)}{\leqslant} \sum_{t=1}^T \max_{f,f' \in \mathbb{B}_t(2, \widetilde{\mathcal{D}}_t)} f(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t)) - f'(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t))$$

$$\overset{(iii)}{\leqslant} \sum_{t=1}^T \max_{f,f' \in \mathbb{B}_t(4, \mathcal{D}_t)} f(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t)) - f'(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t)) = (\star).$$

Where $\mathbb{B}_t(\gamma, \mathcal{D}) = \{f, f' \in \mathcal{F} \text{ s.t. } \|f - f'\|_{\mathcal{D}} \leqslant \gamma \beta_{\mathcal{F}}(\delta, t)\}$. Inequality (i) follows by optimism, (ii) is a consequence of $\tilde{f}_t^{x,\tilde{\pi}_t^{\mathrm{opt}}(x)}, f_\star \in \mathbb{B}_t(2, \widetilde{\mathcal{D}}_t)$ when $\mathcal{E}$ holds since in this case $\|\widehat{f}_t - f_\star\|_{\widetilde{\mathcal{D}}_t} \leqslant \beta_{\mathcal{F}}(t,\delta)$ and $\|\tilde{f}_t^{x,\tilde{\pi}_t^{\mathrm{opt}}(x)} - \widehat{f}_t\|_{\widetilde{\mathcal{D}}_t} \leqslant \beta_{\mathcal{F}}(t,\delta)$, and (iii) follows because when $\mathcal{E}$ holds, Equation 6 of Lemma 4.6 is satisfied and therefore $\mathbb{B}_t(2, \widetilde{\mathcal{D}}_t) \subseteq \mathbb{B}_t(4, \mathcal{D}_t)$. The RHS $(\star)$ of the equation above satisfies,

$$(\star) = \sum_{t=1} \omega(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t), \mathcal{D}_t) \leqslant \sum_{t=1} \max_{a \in \mathcal{A}} \omega(\tilde{x}_t, a, \mathcal{D}_t) = \sum_{t=1} \omega(\tilde{x}_t, \pi_t(\tilde{x}_t), \mathcal{D}_t).$$

We relate the sum of uncertainty radii $\{\omega(\tilde{x}_t, \pi_t(\tilde{x}_t), \mathcal{D}_t)\}_{t=1}^T$ with those of the planner $\{\omega(x_t, \pi_t(x_t), \mathcal{D}_t)\}_{t=1}^T$ via Hoeffding Inequality (see Lemma A.1) and conclude that w.h.p,

$$\sum_{t=1}^T \max_{a \in \mathcal{A}} f_\star(\tilde{x}_t, a) - f_\star(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t)) \leqslant \sum_{t=1}^T \omega(x_t, \pi_t(x_t), \mathcal{D}_t) + \mathcal{O}\left(B\sqrt{T \log(1/\delta)}\right).$$

Lemma 4.3 allows us to bound the sum of these uncertainty radii as

$$\sum_{t=1}^T \omega(x_t, \pi_t(x_t), \mathcal{D}_t) \leqslant \mathcal{O}\left(\min\left(BT, Bd_{\mathrm{eluder}}(\mathcal{F}, B/T) + \beta_{\mathcal{F}}(T, \delta)\sqrt{d_{\mathrm{eluder}}(\mathcal{F}, B/T)T}\right)\right),$$

and therefore w.h.p,

$$\sum_{\ell=1}^T \max_{a \in \mathcal{A}} f_\star(\tilde{x}_\ell, a) - f_\star(\tilde{x}_\ell, \tilde{\pi}_j^{\mathrm{opt}}(\tilde{x}_\ell)) \leqslant \mathcal{O}\left(\min\left(BT, Bd_{\mathrm{eluder}}(\mathcal{F}, B/T) + \beta_{\mathcal{F}}(T, \delta)\sqrt{d_{\mathrm{eluder}}(\mathcal{F}, B/T)T}\right)\right).$$

Converting this cumulative regret bound into a simple regret one (Lemma B.1) finalizes the result. A detailed version of the proof can be found in Appendix C.1.

5 Uniform Sampling Strategies

In this section we show that a uniform sampling strategy can produce an ε optimal policy with probability at least $1 - \delta$ after collecting $\mathcal{O}\left(\frac{\max^2(B,\bar{B})|\mathcal{A}|\log(|\mathcal{F}|/\varepsilon\delta)}{\varepsilon^2}\right)$ samples. This procedure achieves the same simple regret rate as converting SquareCB 's cumulative regret into simple regret[4].

In contrast with the eluder planning algorithm the uniform sampling strategy does not require a planning phase. Instead it consists of running of the Sampler (Algorithm 2) setting $\tilde{\pi}_j = \mathrm{Uniform}(\mathcal{A})$ for all $j \in [T]$. Given a dataset $\widetilde{\mathcal{D}}_T$ of contexts, actions and rewards collected during the sampling phase, we solve the least squares problem:

$$\widehat{f}_T = \arg\min_{f \in \mathcal{F}} \sum_{(x_i, a_i, r_i) \in \widetilde{D}_T} (f(x_i, a_i) - r_i)^2$$

and define the policy $\widehat{\pi}_T$ as $\widehat{\pi}_T(x) = \arg\max_{a \in \mathcal{A}} \widehat{f}_T(x,a)$. The main result of this section is,

---

[4]The regret bound of the SquareCB algorithm scales as $\mathcal{O}\left(\sqrt{|\mathcal{A}|\log(\mathcal{F})T\log(T/\delta)}\right)$.

Theorem 5.1. *There exists a universal constant $\tilde{c} > 0$ such that if $T \geqslant \tilde{c}\frac{\max^2(B,\bar{B})|\mathcal{A}|\log(|\mathcal{F}|/\varepsilon\delta)}{\varepsilon^2}$ then with probability at least $1 - \delta$ the uniform planning algorithm's policy $\widehat{\pi}_T$ is $\varepsilon-$optimal.*

The proof of Theorem 5.1 can be found in Appendix D.

Comparison Between OptLS **and** SquareCB **.** The regret rate after T steps of the OptLS algorithm applied to a contextual bandit problem with discrete function class $\mathcal{F}$ scales[5] (up to logarithmic factors) as $\mathcal{O}\left(\sqrt{d_{\mathrm{eluder}}(\mathcal{F}, B/T)\log(|\mathcal{F}|/\delta)T}\right)$. In contrast, the regret of the SquareCB algorithm satisfies a regret guarantee (up to logarithmic factors) of order $\mathcal{O}\left(\sqrt{|\mathcal{A}|\log(|\mathcal{F}|/\delta)T}\right)$ where the eluder dimension dependence is substituted by a polynomial dependence on the number of actions $|\mathcal{A}|$. When the number of actions is small, or even constant, the regret rate of SquareCB can be much smaller than that of OptLS . The opposite is true when the number of actions is large or even infinite. Converting these cumulative regret bounds to simple regret implies the number of samples required to produce an $\varepsilon-$optimal policy from the *adaptive* OptLS policy sequence scales as $\frac{d_{\mathrm{eluder}}(\mathcal{F},B/T)\log(|\mathcal{F}|/\delta)}{\varepsilon^2}$ whereas for the *adaptive* SquareCB policy sequence it scales as $\frac{|\mathcal{A}|\log(|\mathcal{F}|/\delta)}{\varepsilon^2}$. The results of Theorems 4.4 and 5.1 recover these rates in the experiment planning setting.

6 Gap Between Experiment Planning and Adaptive Learning

The results of Sections 4 and 5 imply planning bounds that are comparable to the corresponding online-to-batch guarantees for OptLS [42] and SquareCB [17]. The main result of this section Theorem 6.1 shows there are problems where the number of samples required for experiment planning can be substantially larger than the number of samples required of an adaptive learning algorithm. This result implies the suboptimality of algorithms such as SquareCB and OptLS for adaptive learning.

In order to state our results we consider an action set $\mathcal{A}_{\mathrm{tree}}$ indexed by the nodes of a height L binary tree defined here as having L levels and $2^L - 1$ nodes.

$a_{1,1}$

$a_{2,1}$ $a_{2,2}$

$a_{3,1}$ $a_{3,2}$ $a_{3,3}$ $a_{3,4}$
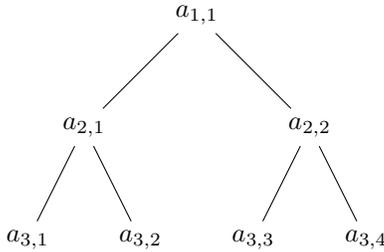
Figure 1: Binary Tree

We call $a_{l,i}$ the i-th action of the l-th level of the tree. For an example see Figure 1. Let $\varepsilon > 0$. We define a function class $\mathcal{F}_{\mathrm{tree}}$ indexed by paths from the root node to a leaf. For any such path $\mathbf{p} = \{a_{1,1}, a_{2,i_2}, \cdots, a_{L,i_L}\}$ the function $f^{(\mathbf{p})}$ equals,

$$f^{(\mathbf{p})}(a) = \begin{cases} 1 & \text{if } a = a_{L,i_L} \\ 1 - 2\varepsilon & \text{if } a \in \mathbf{p}\backslash\{a_{L,i_L}\} \\ 1 - 12\varepsilon & \text{o.w.} \end{cases}$$

The following result fleshes out a separation between planning and adaptive learning with action set $\mathcal{A}_{\mathrm{tree}}$ and function class $\mathcal{F}_{\mathrm{tree}}$ in the setting where at time T the learner will produce a guess for the optimal policy $\widehat{\pi}_T$.

Theorem 6.1. *Let $\varepsilon > 0$, $T \in \mathbb{N}$. Consider the action set $\mathcal{A}_{\mathrm{tree}}$ and function class $\mathcal{F}_{\mathrm{tree}}$ and a reward noise process such that $\xi_t \sim \mathcal{N}(0,1)$ conditionally for all $t \in [T]$. For any planning algorithm* Alg *there is a function $f_\star \in \mathcal{F}_{\mathrm{tree}}$ such that when $T \leqslant \frac{2^{L-5}}{9\varepsilon^2}$ and* Alg *interacts with $f_\star$ then,*

$$\mathbb{E}_{\mathrm{Alg},f_\star}\left[\mathbb{E}_{a\sim\widehat{\pi}_T}\left[f_\star(a)\right]\right] < \max_{a\in\mathcal{A}_{\mathrm{tree}}} f_\star(a) - \varepsilon.$$

Moreover, there is an adaptive algorithm Alg$_{\mathrm{adaptive}}$ *such that if $T \geqslant \frac{2L\log(2L/\varepsilon)}{\varepsilon^2}$,*

$$\mathbb{E}_{\mathrm{Alg}_{\mathrm{adaptive}}}\left[\mathbb{E}_{a\sim\widehat{\pi}_T}\left[f(a)\right]\right] \geqslant \max_{a\in\mathcal{A}_{\mathrm{tree}}} f(a) - \varepsilon.$$

for all $f \in \mathcal{F}_{\mathrm{tree}}$. Where $\mathbb{E}_{\widetilde{\mathrm{Alg}},\widetilde{f}}\left[\cdot\right]$ is the expectation over the randomness of $\widetilde{\mathrm{Alg}}$ and the environment for target function $\widetilde{f}$, and $\widehat{\pi}_T$ is the algorithm's final policy guess after the sampling phase.

---

[5]We obviate the T dependence on d_{eluder} for readability.

The main insight behind Theorem 6.1 is that adaptive strategies to find an optimal action in the $\mathcal{F}_{\text{tree}}$ function class can make use of the conditional structure of the action space by trying to determine a path of actions from the root to the leaf containing the optimal action. An adaptive strategy can determine this path by querying only nodes that are adjacent to it. In contrast, a static experiment planning strategy cannot leverage this structure and has to query all leaf nodes. Theorem 6.1 implies a gap between the adaptive and experiment planning problems. Moreover, since the eluder dimension of $\mathcal{F}_{\text{tree}}$ scales with 2^L (see Lemma D.1), OptLS and SquareCB are suboptimal adaptive algorithms for this model class. In contrast with the $\mathcal{O}\left(\frac{2L \log(2L/\varepsilon)}{\varepsilon^2}\right)$ upper bound in Theorem 6.1, converting the cummulative regret bounds of OptLS and SquareCB yield guarantees scaling as $\mathcal{O}\left(\frac{2^L \log(|\mathcal{F}_{\text{tree}}|)}{\varepsilon^2}\right)$.

7 Model Selection

In this section we consider a setting where the learner has access to $\mathcal{F}_1, \cdots, \mathcal{F}_M$ function (model) classes all with domain $\mathcal{X} \times \mathcal{A}$. The learner is promised there exists an index $i_\star$ such that $f_\star \in \mathcal{F}_{i_\star}$. The value of index $i_\star$ is not known by the learner. We will show the uniform sampling strategy has a sample complexity that scales logarithmically with the number of models and with the complexity of the optimal model class $|\mathcal{F}_{i_\star}|$. In this setting the Sampler will collect two uniform actions datasets $(\widetilde{\mathcal{D}}_T^{(\text{Train})}, \widetilde{\mathcal{D}}_T^{(\text{Test})})$ of size $T/2$ each. Using the "train" dataset $\widetilde{\mathcal{D}}_T^{(\text{Train})}$ the learner computes candidate reward models $\widehat{f}_T^{(i)}$ for all $i \in [M]$ by solving the least squares problems:

$$\widehat{f}_T^{(i)} = \arg\min_{f \in \mathcal{F}_i} \sum_{(x_i, a_i, r_i) \in \widetilde{\mathcal{D}}_T^{(\text{Train})}} (f(x_i, a_i) - r_i)^2.$$

Using the "test" dataset the learner extracts a guess for the optimal model class index $\underline{i}$ by solving,

$$\underline{i} = \arg\min_{i \in [M]} \sum_{(x_\ell, a_\ell, r_\ell) \in \widetilde{\mathcal{D}}_T^{(\text{Test})}} \left(\widehat{f}_T^{(i)}(x_\ell, a_\ell) - r_\ell\right)^2. \tag{7}$$

The candidate policy $\widehat{\pi}_T$ is defined as $\widehat{\pi}_T(x) = \arg\max_{a \in \mathcal{A}} \widehat{f}_T^{(\underline{i})}(x, a)$. Let's start by relating the expected least squares loss of the candidate model $\widehat{f}_T^{(\underline{i})}$ to the size of the optimal model class $|\mathcal{F}_{i_\star}|$,

Proposition 7.1. *There exists a universal constant $\underline{c} > 0$ such that,*

$$\mathbb{E}_{x \sim \mathcal{P}, a \sim \text{Uniform}(\mathcal{A})}\left[\left(f_\star(x, a) - \widehat{f}_T^{(\underline{i})}(x, a)\right)^2\right] \leq \frac{\underline{c}\max^2(B, \bar{B})\log(T\max(M, |\mathcal{F}_{i_\star}|)/\delta)}{T}.$$

With probability at lest $1 - \delta$.

The proof of Proposition 7.1 can be found in Appendix E.1. The uniform sampling model-selection algorithm has the following performance guarantees,

Theorem 7.2. *There exists a universal constant $\widetilde{c}' > 0$ s.t. if $T \geq \widetilde{c}' \frac{\max^2(B, \bar{B})|\mathcal{A}|\log\left(\frac{\max(M, |\mathcal{F}_\star|)}{\varepsilon\delta}\right)}{\varepsilon^2}$ then with probability at least $1 - \delta$, the candidate policy $\widehat{\pi}_T$ of the uniform sampling model-selection algorithm is $\varepsilon-$optimal.*

Proof. Due to the realizability assumption $(r_t = f_\star(x_t, a_t) + \xi_t)$ the instantaneous regret of $\widehat{\pi}_T$ on context $x \in \mathcal{X}$ equals $\max_a f_\star(x, a) - f_\star(x, \widehat{\pi}_T(x))$. Let $\pi_\star(x) = \arg\max_{a \in \mathcal{A}} f_\star(x, a)$. Just like in the proof of Theorem 5.1 (see Equation 25), we can relate the suboptimality of $\widehat{\pi}_T$ with the expected least squares loss of $\widehat{f}_T^{(\underline{i})}$ under the uniform policy,

$$\mathbb{E}_{x \sim \mathcal{P}}\left[\max_{a \in \mathcal{A}} f_\star(x, a) - f_\star(x, \widehat{\pi}_T(x))\right] \leq 2\sqrt{|\mathcal{A}|\mathbb{E}_{x \sim \mathcal{P}, a \sim \text{Uniform}(\mathcal{A})}\left[\left(f_\star(x, a) - \widehat{f}_T^{(\underline{i})}(x, a)\right)^2\right]}. \tag{8}$$

Finally Proposition 7.1 implies thre is a universal constant $\underline{c} > 0$ such that,

$$\mathbb{E}_{x \sim \mathcal{P}, a \sim \text{Uniform}(\mathcal{A})}\left[\left(f_\star(x, a) - \widehat{f}_T^{(\underline{i})}(x, a)\right)^2\right] \leq \frac{\underline{c}(\bar{B}^2 + B^2)\log(T\max(M, |\mathcal{F}_{i_\star}|)/\delta)}{T}$$

9

with probability at least $1 - \delta$. Plugging this result back into Equation 8 we see the suboptimality of $\widehat{\pi}_T$ can be upper bounded as,

$$\mathbb{E}_{x \sim \mathcal{P}} \left[\max_{a \in \mathcal{A}} f_\star(x, a) - f_\star(x, \widehat{\pi}_T(x)) \right] \leqslant 2\sqrt{\frac{\underline{c}|\mathcal{A}|(\bar{B}^2 + B^2) \log(T \max(M, |\mathcal{F}_{i_\star}|)/\delta)}{T}}$$

Setting $g(T) = 4\frac{\underline{c}|\mathcal{A}|(\bar{B}^2 + B^2) \log(T \max(M, |\mathcal{F}_{i_\star}|)/\delta)}{T}$ in Lemma A.6 implies there exists a universal constant $\tilde{c}' > 0$ such that $g(T) \leqslant \varepsilon^2$ for all $T \geqslant \tilde{c}' \frac{\max^2(B, \bar{B})|\mathcal{A}| \log\left(\frac{\max(M, |\mathcal{F}_{i_\star}|)}{\varepsilon\delta}\right)}{\varepsilon^2}$. $\qquad \square$

The results of Theorem 7.2 can be best understood by contrasting them to the uniform sampling algorithm with input model class equal to the union of the function classes $\mathcal{F}_{\text{all}} = \cup_{i \in [M]} \mathcal{F}_i$. Applying the results of Theorem 5.1 to $\mathcal{F}_{\text{all}}$ yields a sample complexity scaling with $\max_{i \in [M]} \log(|\mathcal{F}|_i)$, a quantity that could be much larger than $\log(|\mathcal{F}_{i_\star}|)$. In contrast, the uniform sampling model-selection algorithm achieves a sample complexity scaling with $\log(|\mathcal{F}_{i_\star}|)$ at a price logarithmic in the number of models classes M. This logarithmic dependence on M stands apart from model selection algorithms for cumulative regret scenarios such as Corral [5, 39], ECE [30] and RegretBalancing [38, 12] that instead have a polynomial dependence on M. The uniform sampling model-selection algorithm is agnostic to the value of ε. The results of Theorem 7.2 hold for any ε. If ε is known in advance the learner can compute the model class index $\widehat{i} = \arg\min \left\{ i \in [M] \text{ s.t. } T \geqslant \tilde{c} \frac{\max^2(B, \bar{B})|\mathcal{A}| \log(|\mathcal{F}_i|/\varepsilon\delta)}{\varepsilon^2} \right\}$ and use the uniform sampling strategy for $\mathcal{F}_{\widehat{i}}$. For this choice of ε, Theorem 5.1 guarantees similar bounds to those of Theorem 7.2. Unfortunately in contrast with the uniform sampling model-selection algorithm this method would be valid for a single choice of ε.

8   Conclusion

In this work we have introduced the first set of algorithms for the experiment planning problem for contextual bandits with general function approximation. We have developed the EluderPlanning algorithm that produces a static policy sequence that after deployment can be used to recover an $\varepsilon-$optimal policy. We showed it is enough for the number of static policies and therefore samples during the sampling phase to be as large as the number of samples required from an adaptive procedure based on an online-to-batch conversion of the OptLS algorithm. Similarly we also demonstrated the uniform sampling strategy enjoys the same online-to-batch conversion sample complexity as the SquareCB algorithm. These results seemingly suggest that simple regret rates for adaptive learning may also be achieved in experiment planning scenarios. We show this is not the case. There exist structured bandit problems for which adaptive learning may require a number of samples that is substantially smaller than the number of samples required by a static policy sequence. This is significant because it implies the suboptimality of the rates achieved by existing adaptive learning algorithms such as OptLS and SquareCB and also because it draws a clear distinction between adaptive learning and experiment planning. This implies these algorithms are either suboptimal or their upper bound analysis is not tight. We believe the first to be correct. This is an important open question we hope to see addressed in future research. We have also introduced the first model selection results for the experiment planning problem.

Many important questions remain regarding this setting. Chief among them is to characterize the true statistical complexity of experimental design for contextual bandits with general function approximations. Our results indicate the eluder dimension is not the sharpest statistical complexity measure to characterize learning here. Developing a more new form of complexity, as well as an accompanying algorithm that can achieve the true statistical lower bound for the problem of experiment planning remains an exciting and important open question to tackle in future research.

Acknowledgments and Disclosure of Funding

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

[2] A. Agarwal, M. Dudík, S. Kale, J. Langford, and R. Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pages 19–26. PMLR, 2012.

[3] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

[4] A. Agarwal, S. Bird, M. Cozowicz, L. Hoang, J. Langford, S. Lee, J. Li, D. Melamed, G. Oshri, O. Ribas, et al. Making contextual decisions with low technical debt. *arXiv preprint arXiv:1606.03966*, 2016.

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

[6] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

[7] A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

[8] C. J. Bryan, E. Tipton, and D. S. Yeager. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature human behaviour*, 5(8):980–989, 2021.

[9] J. Chan, A. Pacchiano, N. Tripuraneni, Y. S. Song, P. Bartlett, and M. I. Jordan. Parallelizing contextual bandits. *arXiv preprint arXiv:2105.10590*, 2021.

[10] J. Chen, A. Modi, A. Krishnamurthy, N. Jiang, and A. Agarwal. On the statistical efficiency of reward-free exploration in non-linear rl. *Advances in Neural Information Processing Systems*, 35:20960–20973, 2022.

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

[12] A. Cutkosky, C. Dann, A. Das, C. Gentile, A. Pacchiano, and M. Purohit. Dynamic balancing for model selection in bandits and rl. In *International Conference on Machine Learning*, pages 2276–2285. PMLR, 2021.

[13] R. Degenne, P. Ménard, X. Shang, and M. Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020.

[14] S. S. Du, S. M. Kakade, J. D. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang. Bilinear classes: A structural framework for provable generalization in rl. *arXiv preprint arXiv:2103.10897*, 2021.

[15] H. Esfandiari, A. Karbasi, A. Mehrabian, and V. Mirrokni. Regret bounds for batched bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7340–7348, 2021.

[16] T. Fiez, L. Jain, K. G. Jamieson, and L. Ratliff. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

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

[18] D. Foster, A. Agarwal, M. Dudik, H. Luo, and R. Schapire. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 1539–1548. PMLR, 2018.

[19] D. A. Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[20] Y. Jedra and A. Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020.

[21] C. Jin, A. Krishnamurthy, M. Simchowitz, and T. Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020.

[22] C. Jin, Q. Liu, and S. Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34: 13406–13418, 2021.

[23] K.-S. Jun and C. Zhang. Crush optimism with pessimism: Structured bandits beyond asymptotic optimality. *Advances in Neural Information Processing Systems*, 33:6366–6376, 2020.

[24] J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

[25] J. Kober, J. A. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[26] J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007.

[27] T. Lattimore and R. Munos. Bounded regret for finite-armed structured bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

[28] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[29] T. Lattimore, C. Szepesvari, and G. Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.

[30] J. Lee, A. Pacchiano, V. Muthukumar, W. Kong, and E. Brunskill. Online model selection for reinforcement learning with function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3340–3348. PMLR, 2021.

[31] G. Li, P. Kamath, D. J. Foster, and N. Srebro. Understanding the eluder dimension. *Advances in Neural Information Processing Systems*, 35:23737–23750, 2022.

[32] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

[33] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[34] Z. Li, L. Ratliff, K. G. Jamieson, L. Jain, et al. Instance-optimal pac algorithms for contextual bandits. *Advances in Neural Information Processing Systems*, 35:37590–37603, 2022.

[35] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[36] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[37] I. Osband and B. Van Roy. Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27, 2014.

[38] A. Pacchiano, C. Dann, C. Gentile, and P. Bartlett. Regret bound balancing and elimination for model selection in bandits and rl. *arXiv preprint arXiv:2012.13045*, 2020.

[39] A. Pacchiano, M. Phan, Y. Abbasi Yadkori, A. Rao, J. Zimmert, T. Lattimore, and C. Szepesvari. Model selection in contextual stochastic bandit problems. *Advances in Neural Information Processing Systems*, 33:10328–10337, 2020.

[40] A. Pacchiano, D. Wulsin, R. A. Barton, and L. Voloch. Neural design for genetic perturbation experiments. *arXiv preprint arXiv:2207.12805*, 2022.

[41] Y. Ruan, J. Yang, and Y. Zhou. Linear bandits with limited adaptivity and learning distributional optimal design. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 74–87, 2021.

[42] D. Russo and B. Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

[43] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[44] D. Simchi-Levi and Y. Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 2021.

[45] M. Soare, A. Lazaric, and R. Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

[46] R. S. Sutton. Introduction: The challenge of reinforcement learning. In *Reinforcement Learning*, pages 1–3. Springer, 1992.

[47] C. Tao, S. Blanco, and Y. Zhou. Best arm identification in linear bandits with linear dimension dependency. In *International Conference on Machine Learning*, pages 4877–4886. PMLR, 2018.

[48] A. Tewari and S. A. Murphy. From ads to interventions: Contextual bandits in mobile health. *Mobile Health: Sensors, Analytic Methods, and Applications*, pages 495–517, 2017.

[49] R. Wan, B. Kveton, and R. Song. Safe exploration for efficient policy evaluation and comparison. In *International Conference on Machine Learning*, pages 22491–22511. PMLR, 2022.

[50] R. Wang, S. S. Du, L. Yang, and R. R. Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *Advances in neural information processing systems*, 33: 17816–17826, 2020.

[51] L. Xu, J. Honda, and M. Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 843–851. PMLR, 2018.

[52] A. Zanette. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. In *International Conference on Machine Learning*, pages 12287–12297. PMLR, 2021.

[53] A. Zanette, K. Dong, J. N. Lee, and E. Brunskill. Design of experiments for stochastic contextual linear bandits. *Advances in Neural Information Processing Systems*, 34:22720–22731, 2021.

[54] R. Zhu and B. Kveton. Safe optimal design with applications in off-policy learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2436–2447. PMLR, 2022.

A Supporting Lemmas

Lemma A.1 (Hoeffding Inequality). *Let $\{Y_\ell\}_{\ell=1}^\infty$ be a martingale difference sequence such that Y_ℓ is $Y_\ell \in [a_\ell, b_\ell]$ almost surely for some constants a_ℓ, b_ℓ almost surely for all $\ell = 1, \cdots, t$. then*

$$\sum_{\ell=1}^t Y_\ell \leqslant 4\sqrt{\sum_{\ell=1}^t (b_\ell - a_\ell)^2 \log\left(\frac{1}{\widetilde{\delta}}\right)}$$

with probability at least $1 - \widetilde{\delta}$.

Lemma A.2 (Anytime Hoeffding Inequality [38]). *Let $\{Y_\ell\}_{\ell=1}^\infty$ be a martingale difference sequence such that Y_ℓ is $Y_\ell \in [a_\ell, b_\ell]$ almost surely for some constants a_ℓ, b_ℓ almost surely for all $\ell = 1, \cdots, t$. then*

$$\sum_{\ell=1}^t Y_\ell \leqslant 2\sqrt{\sum_{\ell=1}^t (b_\ell - a_\ell)^2 \log\left(\frac{12t^2}{\widetilde{\delta}}\right)}$$

with probability at least $1 - \widetilde{\delta}$ for all $t \in \mathbb{N}$ simultaneously.

Our results relies on the following variant of Bernstein inequality for martingales, or Freedman's inequality [19], as stated in e.g., [3, 7].

Lemma A.3 (Simplified Freedman's inequality). *Let $X_1, ..., X_T$ be a bounded martingale difference sequence with $|X_\ell| \leqslant R$. For any $\delta' \in (0, 1)$, and $\eta \in (0, 1/R)$, with probability at least $1 - \delta'$,*

$$\sum_{\ell=1}^T X_\ell \leqslant \eta \sum_{\ell=1}^T \mathbb{E}_\ell[X_\ell^2] + \frac{\log(1/\delta')}{\eta}. \tag{9}$$

where $\mathbb{E}_\ell[\cdot]$ is the conditional expectation[6] induced by conditioning on $X_1, \cdots, X_{\ell-1}$.

Lemma A.4 (Anytime Freedman). *Let $\{X_t\}_{t=1}^\infty$ be a bounded martingale difference sequence with $|X_t| \leqslant R$ for all $t \in \mathbb{N}$. For any $\delta' \in (0, 1)$, and $\eta \in (0, 1/R)$, there exists a universal constant $C > 0$ such that for all $t \in \mathbb{N}$ simultaneously with probability at least $1 - \delta'$,*

$$\sum_{\ell=1}^t X_\ell \leqslant \eta \sum_{\ell=1}^t \mathbb{E}_\ell[X_\ell^2] + \frac{C\log(t/\delta')}{\eta}. \tag{10}$$

where $\mathbb{E}_\ell[\cdot]$ is the conditional expectation induced by conditioning on $X_1, \cdots, X_{\ell-1}$.

Proof. This result follows from Lemma A.3. Fix a time-index t and define $\delta_t = \frac{\delta'}{12t^2}$. Lemma A.3 implies that with probability at least $1 - \delta_t$,

$$\sum_{\ell=1}^t X_\ell \leqslant \eta \sum_{\ell=1}^t \mathbb{E}_\ell\left[X_\ell^2\right] + \frac{\log(1/\delta_t)}{\eta}.$$

A union bound implies that with probability at least $1 - \sum_{\ell=1}^t \delta_t \geqslant 1 - \delta'$,

$$\sum_{\ell=1}^t X_\ell \leqslant \eta \sum_{\ell=1}^t \mathbb{E}_\ell\left[X_\ell^2\right] + \frac{\log(12t^2/\delta')}{\eta}$$

$$\stackrel{(i)}{\leqslant} \eta \sum_{\ell=1}^t \mathbb{E}_\ell\left[X_\ell^2\right] + \frac{C\log(t/\delta')}{\eta}.$$

holds for all $t \in \mathbb{N}$. Inequality (i) holds because $\log(12t^2/\delta') = \mathcal{O}\left(\log(t\delta')\right)$.

$\square$

---

[6]We will use this notation to denote conditional expectations throughout this work.

Lemma A.5. *Let* $\{A_\ell\}_{\ell=1}^\infty$ *be an adapted process. We use the notation* $\mathbb{E}_\ell[\cdot]$ *to denote the conditional expectation* $\mathbb{E}[\cdot|A_1, \cdots, A_{\ell-1}]$. *If* $|A_\ell| \leqslant \tilde{B}$ *for all* $\ell \in \mathbb{N}$ *a.s. then*

$$\sum_{\ell=1}^t A_\ell^2 \leqslant \frac{3}{2} \sum_{\ell=1}^t \mathbb{E}_\ell[A_\ell^2] + 2C\tilde{B}^2 \log\left(\frac{t}{\delta}\right).$$

and

$$\frac{\sum_{\ell=1}^t \mathbb{E}_\ell[A_\ell^2]}{2} \leqslant \sum_{\ell=1}^t A_\ell^2 + 2C\tilde{B}^2 \log\left(\frac{t}{\delta}\right)$$

With probability at least $1 - \delta$, *for all* $t \in \mathbb{N}$ *simultaneously where* $C > 0$ *is a universal constant.*

Proof. Consider a martingale difference sequence $\{W_\ell\}_{\ell=1}^\infty$ defined as $W_\ell = A_\ell^2 - \mathbb{E}_\ell[A_\ell^2]$. By definition,

$$|W_\ell| \leqslant \tilde{B}^2, \text{ and } \mathbb{E}_\ell[W_\ell^2] \leqslant \mathbb{E}_\ell[A_\ell^4] \leqslant \tilde{B}^2 \mathbb{E}_\ell[A_\ell^2].$$

for all $\ell \in \mathbb{N}$.

The anytime freedman inequality (Lemma A.4) and a union bound on the martingale sequences $\{W_\ell\}_{\ell=1}^\infty$ and $\{-W_\ell\}_{\ell=1}^\infty$ implies

$$\left|\sum_{\ell=1}^t W_\ell\right| \leqslant \eta \sum_{\ell=1}^t \mathbb{E}_\ell[W_\ell^2] + \frac{C\log(2t/\delta)}{\eta}$$

$$\overset{(i)}{\leqslant} \eta\tilde{B}^2 \sum_{\ell=1}^t \mathbb{E}_\ell[A_\ell^2] + \frac{C\log(2t/\delta)}{\eta}$$

for all $t \in \mathbb{N}$ with probability at least $1 - \delta$. Inequality (i) follows because $\mathbb{E}_\ell[W_\ell^2] \leqslant \mathbb{E}_\ell[A_\ell^4] \leqslant \tilde{B}^2 \mathbb{E}_\ell[A_\ell^2]$. Setting $\eta = \frac{1}{2\tilde{B}^2} < \frac{1}{\tilde{B}^2}$.

$$\left|\sum_{\ell=1}^t W_\ell\right| \leqslant \frac{\sum_{\ell=1}^t \mathbb{E}_\ell[A_\ell^2]}{2} + 2C\tilde{B}^2 \log(2t/\delta) \overset{(i)}{\leqslant} \frac{\sum_{\ell=1}^t \mathbb{E}_\ell[A_\ell^2]}{2} + C'\tilde{B}^2 \log(t/\delta)$$

where inequality (i) follows from $\log(2t/\delta) = \mathcal{O}(\log(t/\delta))$. Substituting $\left|\sum_{\ell=1}^t W_\ell\right|$ with $-\sum_{\ell=1}^t W_\ell$ and $\sum_{\ell=1}^t W_\ell$ and simplifying the resulting expressions yields the result. $\qquad\square$

Lemma A.6. *Let* $c > 0, \alpha \geqslant 1, \xi \in (0, 1]$ *and* $g : \mathbb{R} \to \mathbb{R}$ *be defined as* $g(T) = c\frac{\log^\alpha(T)}{T}$. *Let* $\xi' = \frac{\xi}{c(\alpha+4)^\alpha}$. *For all* $T \geqslant \frac{\log^\alpha(1/\xi')}{\xi'}$

$$g(T) \leqslant \xi.$$

Proof. Let's see that $g(T)$ is an increasing function for all $T \geqslant$. The derivative of $g(T)$ satisfies $\frac{\partial g(T)}{\partial T} = \frac{\log^{\alpha-1}(\alpha - \log(T))}{T^2}$. Hence for all T such that $\log(T) \geqslant \alpha$ (i.e. $T \geqslant e^\alpha$), the derivative is negative and therefore f is decreasing.

Let $T_0 = \frac{\log^\alpha(1/\xi')}{\xi'}$. Substituting $g(T_0)$ we get,

$$g(T_0) = \frac{c\log^\alpha\left(\frac{\log^\alpha(1/\xi')}{\xi'}\right)}{\log^\alpha(1/\xi')}\xi'$$

15

The numerator satisfies,

$$
\begin{aligned}
\log^\alpha\left(\frac{\log^\alpha(1/\xi')}{\xi'}\right) &= \left(\log\left(\log^\alpha(1/\xi')\right) + \log(1/\xi')\right)^\alpha \\
&= \left(\alpha\log(\log(1/\xi')) + \log(1/\xi')\right)^\alpha \\
&\overset{(i)}{\leqslant} \left((\alpha+1)\log(1/\xi')\right)^\alpha \\
&= (\alpha+1)^\alpha\log^\alpha(1/\xi')
\end{aligned}
$$

Inequality (i) holds because $\log(1/\xi') \leqslant 1/\xi'$ since $\xi' \in (0,1]$. Substituting these inequalities in the expression above yields,

$$g(T_0) \leqslant c(\alpha+1)^\alpha\xi'.$$

Setting $\xi' = \frac{\xi}{c(\alpha+4)^\alpha}$ and noting this definition of T_0 satisfies $T_0 \geqslant e^\alpha$ yields the desired result.

$\square$

A.1 Conditions for $\beta_{\mathcal{F}}(t,\delta)$

Proposition A.7. *There exists a universal constant $\bar{C}$ such that $\beta_{\mathcal{F}}(\delta,t) = \bar{C}(B+\bar{B})\sqrt{\log\left(\frac{|\mathcal{F}|t}{\delta}\right)}$ satisfies*

$$
\beta_{\mathcal{F}}^2(\delta,t) \geqslant 4(2C_1^2 + C_2)B^2\log\left(\frac{2t|\mathcal{F}|}{\delta}\right) \quad \text{and} \quad \beta_{\mathcal{F}}^2(\delta,t) \geqslant \Omega\left((B^2+\bar{B}^2)\log\left(\frac{|\mathcal{F}|t}{\delta}\right)\right)
$$

for all $t \in \mathbb{N}$ and $\delta \in (0,1)$. Where C_1, C_2 are the universal constants from equation 23 in Lemma 4.6 (left) and $\beta_{\mathcal{F}}^2(\delta,t) \geqslant C_{t,\delta}$ in Lemma A.8 (right).

Proof. This result follows immediately from the definition. $\square$

Satisfies both conditions.

A.2 Least Squares Guarantees

In this section we provide bounds for the least squares algorithm. We assume a setting where at each step t the learner observes an element $z_t \in \mathcal{Z}$ with a label $y_t = f_\star(z) + \xi_t$ where $f_\star \in \mathcal{F}$, $\max_{f\in\mathcal{F}, z\in\mathcal{Z}}|f(z)| \leqslant B$ and $\xi_i t$ is a conditionally zero mean $1-$subgaussian random variable satisfying $|\xi_t| \leqslant \bar{B}$. The function $f_\star$ is assumed to satisfy $f_\star \in \mathcal{F}$ for a known function class $\mathcal{F}$. We analyze the least squares procedure,

$$\widehat{f}_t = \min_{f\in\mathcal{F}}\sum_{\ell=1}^{t-1}(f(z_\ell) - y_\ell)^2.$$

Lemma A.8 (LS guarantee). *Let $z_1, \cdots, z_T$ be a sequence of queries and values $y_1, \cdots, y_T$. Define $\widehat{f}_t = \arg\min_{f\in\mathcal{F}}\sum_{\ell=1}^{t-1}(f(z_\ell) - y_\ell)^2$ for a function class $\mathcal{F}$ satisfying $\max_{f\in\mathcal{F},z\in\mathcal{Z}}|f(z)| \leqslant B$ and where $y_\ell(z_\ell) = f_\star(z_\ell) + \xi_\ell$ for ξ_ℓ a conditionally zero mean random variable with $|\xi_i| \leqslant \bar{B}$ then,*

$$\sum_{\ell=1}^{t-1}\left(\widehat{f}_t(z_\ell) - f_\star(z_\ell)\right)^2 \leqslant C_{t,\delta} \leqslant C_{T,\delta}.$$

Moreover, if $z_t = (x_t, a_t)$ with $x_t \sim \mathcal{P}_t$ and $a_t \sim \pi_t(\cdot|x_t)$ an adapted probability-policy sequence,

$$\sum_{\ell=1}^{t-1}\mathbb{E}_{x\sim\mathcal{P}_\ell, a\sim\pi_\ell(\cdot|x)}\left[\left(\widehat{f}_t(z_\ell) - f_\star(z_i)\right)^2\right] \leqslant C_{t,\delta} \leqslant C_{T,\delta}.$$

with probability at least $1 - \delta$ for all $t \in \mathbb{N}$ where $C_{t,\delta} = \mathcal{O}\left((B^2 + \bar{B}^2)\log\left(\frac{|\mathcal{F}|t}{\delta}\right)\right)$.

Proof. By definition $\mathbb{E}_\ell[\xi_\ell] = 0$ and $\mathbb{E}_\ell[\xi_\ell^2] \leqslant \bar{B}^2$ where we define these conditional expectations to include all events occuring before ξ_ℓ (including z_ℓ). Since $\widehat{f}_t$ is the empirical minimizer of the least squares loss,

$$\sum_{\ell=1}^{t-1} \left(\widehat{f}_t(z_\ell) - f_\star(z_\ell) - \xi_\ell \right)^2 \leqslant \sum_{\ell=1}^{t} \xi_\ell^2.$$

And therefore,

$$\sum_{\ell=1}^{t-1} \left(\widehat{f}_t(z_\ell) - f_\star(z_\ell) \right)^2 \leqslant 2 \sum_{\ell=1}^{t-1} \xi_\ell \cdot \left(\widehat{f}_t(z_\ell) - f_\star(z_\ell) \right). \tag{11}$$

For any <u>fixed</u> f let's consider the martingale difference sequence $Z_\ell^f = \xi_\ell \left(f(z_\ell) - f_\star(z_\ell) \right)$. Observe that $|Z_\ell^f| \leqslant 2\bar{B}B$ and that,

$$\mathbb{E}_\ell \left[\left(Z_\ell^f \right)^2 \right] \leqslant \bar{B}^2 \left(f(z_\ell) - f_\star(z_\ell) \right)^2. \tag{12}$$

Recall that $\mathbb{E}_\ell[\cdot]$ conditions on all events right before ξ_ℓ including z_ℓ so that $\mathbb{E}_\ell[(f(z_\ell) - f_\star(z_\ell)^2] = (f(z_\ell) - f_\star(z_\ell)^2$. By the Anytime Friedman's inequality (see Lemma A.4) with $\eta = \frac{1}{4B \max(B,\bar{B})} < \min\left(\frac{1}{2\bar{B}B}, \frac{1}{2\bar{B}^2} \right)$ and $\delta' = \frac{\delta}{2|\mathcal{F}|}$, and a union bound over all $f \in \mathcal{F}$

$$\sum_{\ell=1}^{t-1} \xi_\ell \cdot (f(z_\ell) - f_\star(z_\ell)) \leqslant \eta \sum_{\ell=1}^{t-1} \mathbb{E}_\ell \left[\left(Z_\ell^f \right)^2 \right] + \frac{C \log(2|\mathcal{F}|t/\delta)}{\eta}$$

$$\overset{(i)}{\leqslant} \eta \bar{B}^2 \sum_{\ell=1}^{t-1} (f(z_\ell) - f_\star(z_\ell))^2 + \frac{C \log(2|\mathcal{F}|t/\delta)}{\eta}$$

$$\overset{(ii)}{\leqslant} \frac{\sum_{\ell=1}^{t-1} (f(z_\ell) - f_\star(z_\ell))^2}{4} + \mathcal{O}\left((\bar{B}^2 + \bar{B}B) \log\left(\frac{|\mathcal{F}|t}{\delta} \right) \right)$$

for all $f \in \mathcal{F}$ and all $t \in \mathbb{N}$ simultaneously with probability at least $1 - \delta/2$. Where inequality (i) holds because of Equation 12 and inequality (ii) holds because $\eta \bar{B}^2 \leqslant \frac{1}{4}$. In particular,

$$\sum_{\ell=1}^{t-1} \xi_\ell \cdot \left(\widehat{f}_t(z_\ell) - f_\star(z_\ell) \right) \leqslant \frac{\sum_{\ell=1}^{t-1} \left(\widehat{f}_t(z_\ell) - f_\star(z_\ell) \right)^2}{4} + \mathcal{O}\left((\bar{B}^2 + \bar{B}B) \log\left(\frac{|\mathcal{F}|t}{\delta} \right) \right) \tag{13}$$

Combining inequalities 11 and 13 we conclude,

$$\sum_{\ell=1}^{t-1} \left(\widehat{f}_t(z_\ell) - f_\star(z_\ell) \right)^2 \leqslant \mathcal{O}\left(\bar{B} \max(B, \bar{B}) \log\left(\frac{|\mathcal{F}|t}{\delta} \right) \right). \tag{14}$$

for all $f \in \mathcal{F}$ and all $t \in \mathbb{N}$ simultaneously with probability at least $1 - \delta/2$. This finalizes the first part of the result.

When $z_t = (x_t, a_t)$ with $x_t \sim \mathcal{P}_t$ and $a_t \sim \pi_t(\cdot|x_t)$ Lemma A.5 applied to the adapted sequence $\{A_\ell\}_{\ell=1}^\infty$ where $A_\ell = f(z_\ell) - f_\star(z_\ell)$ implies,

$$\frac{1}{2} \sum_{\ell=1}^{t-1} \mathbb{E}_{x \sim \mathcal{P}_\ell, a \sim \pi_\ell(\cdot|x)}[(f(z_\ell) - f_\star(z_\ell))^2] \leqslant \sum_{\ell=1}^{t-1} (f(z_\ell) - f_\star(z_\ell))^2 + \mathcal{O}\left(B^2 \log\left(\frac{|\mathcal{F}|t}{\delta} \right) \right)$$

for all $f \in \mathcal{F}$ and all $t \in \mathbb{N}$ simultaneously with probability at least $1 - \delta/2$. In particular setting $f = \widehat{f}_t$ in the equation above,

$$\frac{1}{2} \sum_{\ell=1}^{t-1} \mathbb{E}_{x \sim \mathcal{P}_\ell, a \sim \pi_\ell(\cdot|x)}[\left(\widehat{f}_t(z_\ell) - f_\star(z_\ell) \right)^2] \leqslant \sum_{\ell=1}^{t-1} \left(\widehat{f}_t(z_\ell) - f_\star(z_\ell) \right)^2 + \mathcal{O}\left(B^2 \log\left(\frac{|\mathcal{F}|t}{\delta} \right) \right) \tag{15}$$

17

Combining inequalities 14 and 15 we conclude,

$$\sum_{\ell=1}^{t-1} \mathbb{E}_{x \sim \mathcal{P}_t, a \sim \pi_t(\cdot|x)}[(f(z_\ell) - f_\star(z_\ell))^2] \leqslant \mathcal{O}\left((B^2 + \bar{B}^2) \log\left(\frac{|\mathcal{F}|t}{\delta}\right)\right).$$

for all $t \in \mathbb{N}$ simultaneously. This finalizes the second part of the result. A union bound finalizes the Lemma's proof. $\qquad\square$

Lemma A.9. *Assume a learner interacts with a function class $\mathcal{F}$ while satisfying realizability (Assumption 3.1), Bounded Function Range (Assumption 3.2) and Bounded Noise (Assumption 3.3). Let $f \in \mathcal{F}$ be an arbitrary (fixed) function and $\widetilde{\mathcal{D}}_T^{(\text{Test})}$ be a size T dataset produced by sampling T i.i.d. contexts from $\mathcal{P}$ and acting according to the uniform policy* Uniform. *There is a universal constant $\underline{C} > 0$ such that,*

$$\mathbb{E}_{x \sim \mathcal{P}, a \sim \text{Uniform}(\mathcal{A})}\left[(f_\star(x,a) - f(x,a))^2\right]$$

$$\leqslant \frac{4}{T}\left(\sum_{(x_\ell, a_\ell, r_\ell) \in \widetilde{\mathcal{D}}_T^{(\text{Test})}} (f(x_\ell, a_\ell) - r_\ell)^2 - \xi_\ell^2\right) + \frac{\underline{C}(\bar{B}^2 + B^2)\log(T/\delta)}{T}$$

$$\leqslant 16\mathbb{E}_{x \sim \mathcal{P}, a \sim \text{Uniform}(\mathcal{A})}\left[(f_\star(x,a) - f(x,a))^2\right] + 4\frac{\underline{C}(\bar{B}^2 + B^2)\log(T/\delta)}{T}.$$

with probability at least $1 - \delta$.

Proof. Throughout this proof we refer to the elements of $\widetilde{\mathcal{D}}_T^{(\text{Test})}$ as the ordered list $x_1, a_1, \cdots, x_T, a_T$. Let's write $r_\ell = f_\star(x_\ell, a_\ell) + \xi_\ell$. Substituting this into the empirical loss we obtain,

$$\sum_{\ell=1}^{T} (f(x_\ell, a_\ell) - r_\ell)^2 = \sum_{\ell=1}^{T} (f(x_\ell, a_\ell) - f_\star(x_\ell, a_\ell))^2 + 2\xi_\ell(f(x_\ell, a_\ell) - f_\star(x_\ell, a_\ell)) + \xi_\ell^2.$$

Therefore,

$$\frac{1}{T}\left(\sum_{\ell=1}^{T}(f(x_\ell, a_\ell) - f_\star(x_\ell, a_\ell))^2 + 2\xi_\ell(f(x_\ell, a_\ell) - f_\star(x_\ell, a_\ell))\right)$$

$$= \frac{1}{T}\sum_{(x_\ell, a_\ell, r_\ell) \in \widetilde{\mathcal{D}}_T^{(\text{Test})}} (f(x_\ell, a_\ell) - r_\ell)^2 - \xi_\ell^2. \qquad (16)$$

Consider the martingale difference sequence $\{Z_\ell\}_{\ell=1}^T$ defined as $Z_\ell = \xi_\ell(f_\star(x_\ell, a_\ell) - f(x_\ell, a_\ell))$. Notice that $|Z_\ell| \leqslant 2\bar{B}B$ and that $\mathbb{E}_\ell[(\xi_\ell(f(x_\ell, a_\ell) - f_\star(x_\ell, a_\ell)))^2] \leqslant \bar{B}^2(f(x_\ell, a_\ell) - f_\star(x_\ell, a_\ell))^2$. The Anytime Freedman inequality (Lemma A.4) with $\eta = \frac{1}{4\bar{B}\max(\bar{B},B)} < \min\left(\frac{1}{2\bar{B}B}, \frac{1}{2\bar{B}^2}\right)$ and $\delta' = \delta/2$ plus a union bound implies that with probability at least $1 - \delta/2$

$$\max\left(\sum_{\ell=1} Z_\ell, -\sum_{\ell=1} Z_\ell\right) \leqslant \eta \sum_{\ell=1}^{T} \mathbb{E}_\ell[(\xi_\ell(f(x_\ell, a_\ell) - f_\star(x_\ell, a_\ell))^2] + \frac{C\log(2T/\delta)}{\eta}$$

$$\leqslant \eta\bar{B}^2 \sum_{\ell=1}^{T}(f(x_\ell, a_\ell) - f_\star(x_\ell, a_\ell))^2 + \frac{C\log(2T/\delta)}{\eta} \qquad (17)$$

$$\overset{(i)}{\leqslant} \frac{1}{4}\sum_{\ell=1}^{T}(f(x_\ell, a_\ell) - f_\star(x_\ell, a_\ell))^2 + C_3(\bar{B}^2 + B^2)\log(T/\delta). \qquad (18)$$

where inequality (i) follows because $\eta\bar{B}^2 2 \leqslant \frac{1}{4}$ and because $C\log(2T/\delta) = \mathcal{O}(\log(T/\delta))$. Substituting $Z_\ell = \xi_\ell(f_\star(x_\ell, a_\ell) - f(x_\ell, a_\ell))$, considering the inequality with $-\sum_{\ell=1} Z_\ell$ on the LHS of the above display and dividing by T the expression above yields,

$$-\frac{1}{2T}\sum_{\ell=1}^{T}(f(x_\ell, a_\ell) - f_\star(x_\ell, a_\ell))^2 - \frac{2C_3}{T}(\bar{B}^2 + B^2)\log(T/\delta) \leqslant \frac{1}{T}\sum_{\ell=1}^{T} 2\xi_\ell(f(x_\ell, a_\ell) - f_\star(x_\ell, a_\ell))$$

18

and therefore, combining this expression above with Equation 16 yields,

$$\frac{1}{2T}\sum_{\ell=1}^{T}(f(x_\ell,a_\ell)-f_\star(x_\ell,a_\ell))^2 - \frac{2C_3}{T}(\bar{B}^2+B^2)\log(T/\delta)$$

$$\leq \frac{1}{T}\sum_{(x_\ell,a_\ell,r_\ell)\in\widetilde{\mathcal{D}}_T^{(\text{Test})}}(f(x_\ell,a_\ell)-r_\ell)^2 - \xi_\ell^2. \qquad (19)$$

Finally, the second item of Lemma A.5 applied to the $\{A_\ell\}_{\ell=1}^T$ sequence with $A_\ell = f_\star(x_\ell,a_\ell) - f_\star(x_\ell,a_\ell)$ implies (after dividing by T) and noting that at all time-steps the context and sampling distribution are the same,

$$\mathbb{E}_{x\sim\mathcal{P},a\sim\text{Uniform}(\mathcal{A})}\left[(f_\star(x,a)-f(x,a))^2\right]$$

$$\leq \frac{2}{T}\sum_{\ell=1}^{T}(f(x_\ell,a_\ell)-f_\star(x_\ell,a_\ell))^2 + \frac{C_4}{T}(\bar{B}^2+B^2)\log(T/\delta) \qquad (20)$$

and

$$\frac{1}{T}\sum_{\ell=1}^{T}(f(x_\ell,a_\ell)-f_\star(x_\ell,a_\ell))^2$$

$$\leq \frac{3}{2}\mathbb{E}_{x\sim\mathcal{P},a\sim\text{Uniform}(\mathcal{A})}\left[(f_\star(x,a)-f(x,a))^2\right] + \frac{C_4}{T}(\bar{B}^2+B^2)\log(T/\delta) \qquad (21)$$

with probability at least $1-\delta/2$. Combining Equations 19 and 20,

$$\mathbb{E}_{x\sim\mathcal{P},a\sim\text{Uniform}(\mathcal{A})}\left[(f_\star(x,a)-f(x,a))^2\right]$$

$$\leq \frac{4}{T}\sum_{(x_\ell,a_\ell,r_\ell)\in\widetilde{\mathcal{D}}_T^{(\text{Test})}}(f(x_\ell,a_\ell)-r_\ell)^2 - \xi_\ell^2 + \frac{\bar{\mathsf{C}}'(\bar{B}^2+B^2)\log(T/\delta)}{T}$$

with probability at least $1-\delta$ for some universal constant $\bar{\mathsf{C}}' > 0$.

To prove the second part of the argument we notice the following sequence of equalities and inequalities hold,

$$\frac{1}{T}\sum_{(x_\ell,a_\ell,r_\ell)\in\widetilde{\mathcal{D}}_T^{(\text{Test})}}(f(x_\ell,a_\ell)-r_\ell)^2 - \xi_\ell^2$$

$$= \frac{1}{T}\left(\sum_{\ell=1}^{T}(f(x_\ell,a_\ell)-f_\star(x_\ell,a_\ell))^2 + 2\xi_\ell(f(x_\ell,a_\ell)-f_\star(x_\ell,a_\ell))\right)$$

$$\overset{(i)}{\leq} \frac{1}{T}\left(\frac{3}{2}\sum_{\ell=1}^{T}(f(x_\ell,a_\ell)-f_\star(x_\ell,a_\ell))^2 + 2C_3(\bar{B}^2+B^2)\log(2T/\delta)\right)$$

$$\overset{(ii)}{\leq} \frac{9}{4}\mathbb{E}_{x\sim\mathcal{P},a\sim\text{Uniform}(\mathcal{A})}\left[(f_\star(x,a)-f(x,a))^2\right] + \frac{\bar{\mathsf{C}}''(\bar{B}^2+B^2)\log(T/\delta)}{T}$$

for some universal constant $\bar{\mathsf{C}}'' > 0$ and where inequality (i) follows from equation 18 and considering the inequality with $-\sum_{\ell=1}^{T}Z_\ell$ on the RHS. Inequality (ii) is a result of Equation 21.

Picking a universal constant $\bar{\mathsf{C}}$ larger than all the universal constants in this discussion plus a union bound finalizes the result.

$\square$

B Online to Batch Conversion

Lemma B.1. *Let $T \in \mathbb{N}$ and $\delta' \in [0,1]$. Assume there is an algorithm* Alg *that interacts sequentially with i.i.d. contexts $x_1,\cdots,x_T \overset{i.i.d.}{\sim} \mathcal{P}$ producing policies $\{\pi_\ell\}_{\ell=1}^T$, playing actions $a_\ell \sim \pi_\ell(x_\ell)$*

19

and observing rewards $r_\ell = f_\star(x_\ell a_\ell) + \xi_\ell$ for all $\ell \leqslant T$ and with ξ_ℓ a conditionally zero mean $1-$subgaussian random variable such that with probability at least $1 - \delta'$,

$$\sum_{\ell=1}^{T} \max_{a \in \mathcal{A}} f_\star(x_\ell, a) - f_\star(x_\ell, a_\ell) \leqslant B_{T,\delta'}. \tag{22}$$

For some $B_{T,\delta'} \in \mathbb{R}$. Then policy $\widehat{\pi}_T = \mathrm{Uniform}(\pi_1, \cdot, \pi_T)$ satisfies,

$$\mathbb{E}_{x \sim \mathcal{P}, a \sim \widehat{\pi}_T(x)}[f_\star(x, a)] + \frac{B_{T,\delta'}}{T} + 8B\sqrt{\frac{\log(1/\delta')}{T}} \geqslant \mathbb{E}_{x \sim \mathcal{P}}\left[\max_{a \in \mathcal{A}} f_\star(x, a)\right].$$

with probability at least $1 - 3\delta'$.

Proof. Consider the martingale sequences $\{Z_\ell^{(1)}\}_{\ell=1}^{T}$ and $\{Z_\ell^{(2)}\}_{\ell=1}^{T}$ defined as $Z_\ell^{(1)} = \max_{a \in \mathcal{A}} f_\star(x_\ell, a) - \mathbb{E}_{x \sim \mathcal{P}}[\max_{a \in \mathcal{A}} f_\star(x, a)]$ and $Z_\ell^{(2)} = f_\star(x_\ell, a_\ell) - \mathbb{E}_{x \sim \mathcal{P}, a \sim \pi_\ell(x)}[f_\star(x, a)]$ for $\ell \in [T]$.

The bounded function range assumption 3.2 implies $\max(|Z_\ell^{(1)}|, |Z_\ell^{(2)}|) \leqslant 2B$ for all $\ell \leqslant T$. Therefore, Hoeffding Inequality applied to these two martingale sequences (see A.1) and a union bound imply

$$\sum_{\ell=1}^{T} \max_{a \in \mathcal{A}} f_\star(x_\ell, a) + 4B\sqrt{T \log(1/\delta')} \geqslant \sum_{\ell=1}^{T} \mathbb{E}_{x \sim \mathcal{P}}[\max_{a \in \mathcal{A}} f_\star(x, a)] = T\mathbb{E}_{x \sim \mathcal{P}}[\max_{a \in \mathcal{A}} f_\star(x, a)]$$

$$\sum_{\ell=1}^{T} \mathbb{E}_{x \sim \mathcal{P}, a \sim \pi_\ell(x)}[f_\star(x, a)] + 4B\sqrt{T \log(1/\delta')} \geqslant \sum_{\ell=1}^{T} f_\star(x_\ell, a_\ell).$$

simultaneously with probability at least $1 - 2\delta'$. Combining this with inequality 22 and a union bound implies,

$$\sum_{\ell=1}^{T} \mathbb{E}_{x \sim \mathcal{P}, a \sim \pi_\ell(x)}[f_\star(x, a)] + B_{T,\delta'} + 8B\sqrt{T \log(1/\delta')} \geqslant T\mathbb{E}_{x \sim \mathcal{P}}[\max_{a \in \mathcal{A}} f_\star(x, a)].$$

with probability at least $1 - 3\delta'$. Therefore,

$$\mathbb{E}_{x \sim \mathcal{P}, a \sim \widehat{\pi}_T(x)}[f_\star(x, a)] + \frac{B_{T,\delta'}}{T} + 8B\sqrt{\frac{\log(1/\delta')}{T}} \geqslant \mathbb{E}_{x \sim \mathcal{P}}\left[\max_{a \in \mathcal{A}} f_\star(x, a)\right]$$

$\square$

C   Proofs of Section 4

Lemma 4.6. *With probability at least $1 - \frac{\delta}{12}$,*

$$\{(f, f') \text{ s.t. } \|f - f'\|_{\widetilde{\mathcal{D}}_t} \leqslant 2\beta_{\mathcal{F}}(t, \delta)\} \subseteq \{(f, f') \text{ s.t. } \|f - f'\|_{\mathcal{D}_t} \leqslant 4\beta_{\mathcal{F}}(t, \delta)\} \tag{6}$$

Simultaneously for all $t \in [T]$. Where $\{D_t\}_{t=1}^{T}$ is the dataset sequence resulting of the execution of EluderPlanner *while $\{\widetilde{D}_t\}_{t=1}^{T}$ is the dataset sequence resulting of the execution of the* Sampler *and $\beta_{\mathcal{F}}(\delta, t)$ is the confidence radius function defined in Equation 2.*

Proof. We will define a process $\{Z_\ell\}_{\ell=1}^{T}$ such that $Z_\ell = (x_\ell, \tilde{x}_\ell, \pi_\ell)$. Here $\tilde{x}_\ell$ is independent and identically distributed to any x_ℓ (and in particular to x_ℓ).

Let $f, f' \in \mathcal{F}$ be two fixed functions. Define two adapted processes $\{f(x_\ell, \pi_\ell(x_\ell)) - f'(x_\ell, \pi_\ell(x_\ell))\}_\ell$ and $\{f(\tilde{x}_\ell, \pi_\ell(\tilde{x}_\ell)) - f'(\tilde{x}_\ell, \pi_\ell(\tilde{x}_\ell))\}_\ell$ (in this definition we have used the fact that π_ℓ is a sequence of deterministic policies).

Lemma A.5 implies,

$$\sum_{\ell=1}^{t} (f(x_\ell, \pi_\ell(x_\ell)) - f'(x_\ell, \pi_\ell(x_\ell)))^2$$

$$\leqslant \frac{3}{2} \sum_{\ell=1}^{t} \mathbb{E}_{x \sim \mathcal{P}}[(f(x, \pi_\ell(x)) - f'(x, \pi_\ell(x)))^2] + (2C_1^2 + C_2)B^2 \log\left(\frac{2t}{\bar{\delta}}\right)$$

and

$$\frac{1}{2} \sum_{\ell=1}^{t} \mathbb{E}_{x \sim \mathcal{P}}[(f(x, \pi_\ell(x)) - f'(x, \pi_\ell(x)))^2]$$

$$\leqslant \sum_{\ell=1}^{t} (f(\tilde{x}_\ell, \pi_\ell(\tilde{x}_\ell)) - f'(\tilde{x}_\ell, \pi_\ell(\tilde{x}_\ell)))^2 + (2C_1^2 + C_2)B^2 \log\left(\frac{2t}{\bar{\delta}}\right)$$

with probability at least $1 - \bar{\delta}$ for all $t \in \mathbb{N}$. Combining these two inequalities we conclude,

$$\sum_{\ell=1}^{t} (f(x_\ell,) - f'(x_\ell))^2 \leqslant 3 \sum_{\ell=1}^{t} (f(\tilde{x}_\ell) - f'(\tilde{x}_\ell))^2 + 4(2C_1^2 + C_2)B^2 \log\left(\frac{2t}{\bar{\delta}}\right)$$

Setting $\bar{\delta} = \delta/|\mathcal{F}|^2$ and a union bound implies that,

$$\|f - f'\|_{\mathcal{D}_t}^2 \leqslant 3\|f - f'\|_{\tilde{\mathcal{D}}_t}^2 + 4(2C_1^2 + C_2)B^2 \log\left(\frac{2|\mathcal{F}|t}{\delta}\right)$$

for all $f, f' \in \mathcal{F}$ simultaneously. Since (see Proposition A.7) we define $\beta_\mathcal{F}(\delta, t)$ such that

$$\beta_\mathcal{F}^2(\delta, t) \geqslant 4(2C_1^2 + C_2)B^2 \log\left(\frac{2|\mathcal{F}|t}{\delta}\right) \tag{23}$$

we conclude that whenever $f, f' \in \mathcal{F}$ satisfy $\|f - f'\|_{\tilde{\mathcal{D}}_t} \leqslant 2\beta_\mathcal{F}(t, \delta)$,

$$\|f - f'\|_{\mathcal{D}_t}^2 \leqslant 16\beta_\mathcal{F}^2(t, \delta)$$

and therefore $\|f - f'\|_{\mathcal{D}_t} \leqslant 4\beta_\mathcal{F}(t, \delta)$. This concludes the result. $\qquad \square$

C.1 Proof of Theorem 4.4

Theorem 4.4. *Let $\varepsilon > 0$. There exists a universal constant $c > 0$ such that if*

$$T \geqslant c \frac{\max^2(B, \bar{B}, 1) d_{\text{eluder}}(\mathcal{F}, B/T) \log\left(|\mathcal{F}|T/\delta\right)}{\varepsilon^2} \tag{5}$$

then with probability at least $1 - \delta$ the eluder planning algorithm's policy $\hat{\pi}_T$ is $\varepsilon-$optimal .

In this section we present a more complete version of the proof from Section 4.1. We will actually prove a 'stronger' version of Theorem 4.4.

Theorem C.1. *Let $\varepsilon > 0$. There exists a universal constant $c_0 > 0$ such that if*

$$T \geqslant c_0 \max\left(\frac{\max^2(B, \bar{B}) d_{\text{eluder}}(\mathcal{F}, B/T) \log\left(|\mathcal{F}|T/\delta\right)}{\varepsilon^2}, \frac{B d_{\text{eluder}}(\mathcal{F}, B/T)}{\varepsilon}\right) \tag{24}$$

then with probability at least $1 - \delta$ the eluder planning algorithm's policy $\hat{\pi}_T$ is $\varepsilon-$optimal .

Proof. For readability we have repeated some of the text of the proof in here.

Let $\hat{f}_t$ as in Equation 3 and define $\mathcal{E}$ as the event where the Standard Least Squares results (see Lemma A.8) hold i.e. $\|\hat{f}_t - f_\star\|_{\tilde{\mathcal{D}}_t} \leqslant \beta(t, \delta)$ for all $t \in [T]$ and also Equation 6 from Lemma 4.6 holds for all $t \in [T]$. The results of Lemmas A.8 and 4.6 imply $\mathbb{P}(\mathcal{E}) \geqslant 1 - \frac{\delta}{6}$.

21

For context x and action a let's denote by $\tilde{f}_t^{x,a}$ as a function achieving the inner maximum in the definition of $\tilde{\pi}_t^{\mathrm{opt}}$ (see Equation 4). When $\mathcal{E}$ holds the policies $\{\tilde{\pi}_t^{\mathrm{opt}}\}_{t=1}^T$ are optimistic in the sense that $\tilde{f}_t^{x,\tilde{\pi}_t^{\mathrm{opt}}(x)}(x, \tilde{\pi}_t^{\mathrm{opt}}(x)) \geqslant \max_{a \in \mathcal{A}} f_\star(x, a)$. To see why this is the case notice that $\mathcal{E}$ holds $f_\star \in \{\|f - \hat{f}_j\|_{\tilde{\mathcal{D}}_t} \leqslant \beta_{\mathcal{F}}(t, \delta)\}$ and therefore

$$\tilde{f}_t^{x,\tilde{\pi}_t^{\mathrm{opt}}(x)}(x, \tilde{\pi}_t^{\mathrm{opt}}(x)) \geqslant \max_{f \in \{\|f - \hat{f}_t\|_{\tilde{\mathcal{D}}_t} \leqslant \beta_{\mathcal{F}}(t, \delta)\}} f(x, \pi_\star(x)) \geqslant f_\star(x, \pi_\star(x)).$$

where $\pi_\star$ corresponds to the optimal policy $\pi_\star(x) = \arg\max_{a \in \mathcal{A}} f_\star(x, a)$. This implies the following sequence of inequalities,

$$\sum_{t=1}^T \max_{a \in \mathcal{A}} f_\star(\tilde{x}_t, a) - f_\star(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t)) \leqslant \sum_{t=1}^T \tilde{f}_t^{x,\tilde{\pi}_t^{\mathrm{opt}}(x)}(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t)) - f_\star(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t))$$

$$\leqslant \sum_{t=1}^T \max_{f,f' \text{ s.t. } \|f-f'\|_{\tilde{\mathcal{D}}_t} \leqslant 2\beta_{\mathcal{F}}(\delta,t)} f(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t)) - f'(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t))$$

$$\overset{(i)}{\leqslant} \sum_{t=1}^T \max_{f,f' \text{ s.t. } \|f-f'\|_{\mathcal{D}_t} \leqslant 4\beta_{\mathcal{F}}(\delta,t)} f(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t)) - f'(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t))$$

$$= \sum_{t=1}^T \omega(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t), \mathcal{D}_t)$$

$$\leqslant \sum_{t=1}^T \max_{a \in \mathcal{A}} \omega(\tilde{x}_t, a, \mathcal{D}_t)$$

$$= \sum_{t=1}^T \omega(\tilde{x}_t, \pi_t(\tilde{x}_t), \mathcal{D}_t).$$

Where inequality (i) is satisfied when $\mathcal{E}$ holds as a result of Lemma 4.6. Using Hoeffding Inequality (see Lemma A.1) yields $\sum_{t=1}^T \omega(\tilde{x}_t, \pi_t(\tilde{x}_t), \mathcal{D}_t) \leqslant \sum_{t=1}^T \omega(x_t, \pi_t(x_t), \mathcal{D}_t) + 16B\sqrt{T \log\left(\frac{3}{\delta}\right)}$ with probability at least $1 - \delta/3$. Finally Lemma 4.3 implies there exists a universal constant $C > 0$ such that $\sum_{t=1}^T \omega(x_t, \pi_t(x_t), \mathcal{D}_t) \leqslant C \cdot \min\left(BT, Bd_{\mathrm{eluder}}(\mathcal{F}, B/T) + \beta_{\mathcal{F}}(T, \delta)\sqrt{d_{\mathrm{eluder}}(\mathcal{F}, B/T)T}\right)$ and therefore,

$$\sum_{t=1}^T \max_{a \in \mathcal{A}} f_\star(\tilde{x}_t, a) - f_\star(\tilde{x}_t, \tilde{\pi}_t^{\mathrm{opt}}(\tilde{x}_t)) \leqslant C \cdot \min\left(BT, Bd_{\mathrm{eluder}}(\mathcal{F}, B/T) + \beta_{\mathcal{F}}(T, \delta)\sqrt{d_{\mathrm{eluder}}(\mathcal{F}, B/T)T}\right).$$

With probability at least $1 - \delta$. The Online to Batch Conversion Lemma B.1 implies there exists a universal constant $C > 0$ such that,

$$\mathbb{E}_{x \sim \mathcal{P}, a \sim \hat{\pi}_T(x)} [f_\star(x, a)] +$$
$$\frac{C \cdot \min\left(BT, Bd_{\mathrm{eluder}}(\mathcal{F}, B/T) + \sqrt{d_{\mathrm{eluder}}(\mathcal{F}, B/T)\beta_{\mathcal{F}}(T, \delta)T}\right)}{T} + 8B\sqrt{\frac{\log(1/\delta)}{T}}$$
$$\geqslant \max_\pi \mathbb{E}_{x \sim \mathcal{P}} [f_\star(x, \pi(x))].$$

with probability at least $1 - 3\delta$ where $\hat{\pi}_T = \mathrm{Uniform}(\pi_1^{\mathrm{opt}}, \cdots, \pi_T^{\mathrm{opt}})$. Finally setting T sufficiently large such that,

$$T \geqslant \frac{4CB}{\varepsilon}$$
$$T \geqslant \frac{4Bd_{\mathrm{eluder}}(\mathcal{F}, B/T)}{\varepsilon}$$
$$T \geqslant \frac{16d_{\mathrm{eluder}}(\mathcal{F}, B/T)\beta_{\mathcal{F}}^2(T, \delta)}{\varepsilon^2}$$
$$T \geqslant \frac{256B \log(1/\delta)}{\varepsilon^2}$$

we conclude there exists a universal constant $c_0 > 0$ such that if

$$T \geqslant c_0 \max \left(\frac{\max^2(B, \bar{B}) d_{\text{eluder}}(\mathcal{F}, B/T) \log (|\mathcal{F}| T/\delta)}{\varepsilon^2}, \frac{B d_{\text{eluder}}(\mathcal{F}, B/T)}{\varepsilon} \right)$$

then with probability at least $1 - \delta$,

$$\mathbb{E}_{x \sim \mathcal{P}} \left[f_\star(x, \hat{\pi}_T(x)) \right] + \varepsilon \geqslant \max_\pi \mathbb{E}_{x \sim \mathcal{P}} \left[f_\star(x, \pi(x)) \right].$$

$\square$

We now return to the proof of Theorem 4.4. We can simplify inequality 24 by instead imposing the condition,

$$T \geqslant c \frac{\max^2(B, \bar{B}, 1) d_{\text{eluder}}(\mathcal{F}, B/T) \log (|\mathcal{F}| T/\delta)}{\varepsilon^2}$$

for some universal constant $c > 0$. This finalizes the proof.

Comparison with Linear Experiment Planning . Using standard techniques the results of Theorem 4.4 can be adapted to the case when the function class $\mathcal{F}$ is infinite but admits a metrization and has a covering number. In this case the sample complexity will scale not with $\log(|\mathcal{F}|)$ but instead with the logarithm of the covering number of $\mathcal{F}$. For example, in the case of linear functions, the logarithm of the covering number of $\mathcal{F}$ under the ℓ_2 norm will scale as d, the ambient dimension of the space up to logarithmic factors of T (see for example Lemma D.1 of [14]). Plugging this into the sample guarantees of Theorem 4.4 recovers the $\tilde{\mathcal{O}} \left(d^2/\varepsilon^2 \right)$ sample complexity for linear experiment planning from [53].

D Proof of Theorem 5.1

Theorem 5.1. *There exists a universal constant $\tilde{c} > 0$ such that if $T \geqslant \tilde{c} \frac{\max^2(B, \bar{B}) |\mathcal{A}| \log(|\mathcal{F}|/\varepsilon \delta)}{\varepsilon^2}$ then with probability at least $1 - \delta$ the uniform planning algorithm's policy $\hat{\pi}_T$ is $\varepsilon-$optimal.*

Proof. Due to the realizability assumption ($r_t = f_\star(x_t, a_t) + \xi_t$) the instantaneous regret of $\hat{\pi}_T$ on context $x \in \mathcal{X}$ equals $\max_a f_\star(x, a) - f_\star(x, \hat{\pi}_T(x))$. Let $\pi_\star(x) = \arg\max_{a \in \mathcal{A}} f_\star(x, a)$. The following inequalities hold,

$$\max_{a \in \mathcal{A}} f_\star(x, a) - f_\star(x, \hat{\pi}_T(x)) = f_\star(x, \pi_\star(x)) - f_\star(x, \hat{\pi}_T(x))$$

$$\overset{(i)}{\leqslant} f_\star(x, \pi_\star(x)) - \hat{f}_T(x, \pi_\star(x)) + \hat{f}_T(x, \hat{\pi}_T(x)) - f_\star(x, \hat{\pi}_T(x))$$

$$\leqslant 2 \max_{a \in \mathcal{A}} |f_\star(x, a) - \hat{f}_T(x, a)|$$

Where (i) holds because $\hat{f}_T(x, \pi_\star(x)) \leqslant \hat{f}_T(x, \hat{\pi}_T(x))$. Therefore,

$$\mathbb{E}_{x \sim \mathcal{P}} \left[\max_{a \in \mathcal{A}} f_\star(x, a) - f_\star(x, \hat{\pi}_T(x)) \right] \leqslant 2 \mathbb{E}_{x \sim \mathcal{P}} \left[\max_{a \in \mathcal{A}} |f_\star(x, a) - \hat{f}_T(x, a)| \right]$$

$$\overset{(i)}{\leqslant} 2 \sqrt{\mathbb{E}_{x \sim \mathcal{P}} \left[\max_{a \in \mathcal{A}} \left(f_\star(x, a) - \hat{f}_T(x, a) \right)^2 \right]}$$

$$\overset{(ii)}{\leqslant} 2 \sqrt{|\mathcal{A}| \mathbb{E}_{x \sim \mathcal{P}, a \sim \text{Uniform}(\mathcal{A})} \left[\left(f_\star(x, a) - \hat{f}_T(x, a) \right)^2 \right]}.$$

(25)

Inequality (i) holds because of Jensen's inequality. Inequality (ii) follows because $\max_{a \in \mathcal{A}} \left(f_\star(x, a) - \hat{f}_T(x, a) \right)^2 \leqslant \sum_{a \in \mathcal{A}} \left(f_\star(x, a) - \hat{f}_T(x, a) \right)^2$.

Finally the context generating distributions equal $\mathcal{P}$ at all timesteps and the policies equal $\text{Uniform}(\mathcal{A})$ standard least squares analysis (Lemma A.8) yields that with probability at least $1 - \delta$

$$\mathbb{E}_{x \sim \mathcal{P}, a \sim \text{Uniform}(\mathcal{A})} \left[\left(f_\star(x, a) - \hat{f}_T(x, a) \right)^2 \right] \leqslant \frac{C_{T, \delta}}{T}.$$

23

Therefore we conclude that,

$$\mathbb{E}_{x \sim \mathcal{P}} \left[\max_{a \in \mathcal{A}} f_\star(x, a) - f_\star(x, \widehat{\pi}_T(x)) \right] \leqslant 2\sqrt{\frac{|\mathcal{A}| C_{T,\delta}}{T}}$$

Since $C_{T,\delta} = c'(B^2 + \bar{B}^2) \log\left(\frac{|\mathcal{F}|t}{\delta}\right)$ for some universal constant $c' > 0$ we conclude that,

$$\mathbb{E}_{x \sim \mathcal{P}} \left[\max_{a \in \mathcal{A}} f_\star(x, a) - f_\star(x, \widehat{\pi}_T(x)) \right] \leqslant 2\sqrt{\frac{c'|\mathcal{A}| \max^2(B, \bar{B}) \log\left(\frac{|\mathcal{F}|t}{\delta}\right)}{T}}$$

Setting $g(T) = 4\frac{c'|\mathcal{A}| \max^2(B,\bar{B}) \log\left(\frac{|\mathcal{F}|t}{\delta}\right)}{T}$ in Lemma A.6 implies there exists a universal constant $\tilde{c} > 0$ such that $g(T) \leqslant \varepsilon^2$ for all $T \geqslant \tilde{c}\frac{\max^2(B,\bar{B})|\mathcal{A}| \log\left(\frac{|\mathcal{F}|}{\varepsilon\delta}\right)}{\varepsilon^2}$. The result follows.

$\square$

D.1 Proof of Theorem 6.1

In order to prove Theorem 6.1 we will assume a gaussian model for the noise variables $\xi_t \sim \mathcal{N}(0, 1)$. This will allow us to use the appropriate information theoretic tools.

Theorem 6.1. *Let $\varepsilon > 0$, $T \in \mathbb{N}$. Consider the action set $\mathcal{A}_{\text{tree}}$ and function class $\mathcal{F}_{\text{tree}}$ and a reward noise process such that $\xi_t \sim \mathcal{N}(0, 1)$ conditionally for all $t \in [T]$. For any planning algorithm* Alg *there is a function $f_\star \in \mathcal{F}_{\text{tree}}$ such that when $T \leqslant \frac{2^{L-5}}{9\varepsilon^2}$ and* Alg *interacts with $f_\star$ then,*

$$\mathbb{E}_{\text{Alg},f_\star} \left[\mathbb{E}_{a \sim \widehat{\pi}_T} \left[f_\star(a) \right] \right] < \max_{a \in \mathcal{A}_{\text{tree}}} f_\star(a) - \varepsilon.$$

Moreover, there is an adaptive algorithm $\text{Alg}_{\text{adaptive}}$ *such that if $T \geqslant \frac{2L \log(2L/\varepsilon)}{\varepsilon^2}$,*

$$\mathbb{E}_{\text{Alg}_{\text{adaptive}}} \left[\mathbb{E}_{a \sim \widehat{\pi}_T} \left[f(a) \right] \right] \geqslant \max_{a \in \mathcal{A}_{\text{tree}}} f(a) - \varepsilon.$$

for all $f \in \mathcal{F}_{\text{tree}}$. Where $\mathbb{E}_{\widetilde{\text{Alg}}, \widetilde{f}} \left[\cdot \right]$ is the expectation over the randomness of $\widetilde{\text{Alg}}$ and the environment for target function $\widetilde{f}$, and $\widehat{\pi}_T$ is the algorithm's final policy guess after the sampling phase.

Proof. We use the notation $a_1, \cdots, a_T$ to denote a (random) sample path of query tuples for algorithm Alg with observed responses $r_1, \cdots, r_T$ where $r_t = f(a_t) + \xi_t$. We'll assume that $\xi_t \sim \mathcal{N}(0, 1)$.

Algorithm Alg interacts for T timesteps during the planning phase with $\mathcal{A}_{\text{tree}}$ and produces a sequence of policies $\{\pi_t\}_{t=1}^T$ to deploy during the sampling phase. The data collected from these deployment policies is then used by the algorithm to define a candidate ε-optimal policy $\widehat{\pi}_T$.

Since we are studying the context-less setting, the underlying context distribution equals a delta mass over the empty context and therefore the distribution over the sampling policy sequence $\{\pi_t\}_{t=1}^T$ is independent of f. Let $n_T(a)$ be the random variable specifying how many times action a was pulled during the sampling phase of algorithm Alg. This random variable is independent of f.

Let $\widetilde{a}$ and $\widetilde{b}$ be two consecutive leaf actions (i.e. inhabiting a size 2 sub-tree of $\mathcal{A}_{\text{tree}}$) such that,

$$\mathbb{E}_{\text{Alg}}[n_T(\widetilde{a}) + n_T(\widetilde{b})] \leqslant \frac{T}{2^{L-2}}.$$

Let $\mathbf{p}_{\widetilde{a}}$ be the path leading to $\widetilde{a}$ and $\mathbf{p}_{\widetilde{b}}$ be the path leading to $\widetilde{b}$. We'll use the notation $\mathbb{P}_a^{\mathbf{P}_{\widetilde{a}}}, \mathbb{P}_a^{\mathbf{P}_{\widetilde{b}}}$ to denote the distribution of arm a in the world specified by $f^{\mathbf{P}_{\widetilde{a}}}$ and $f^{\mathbf{P}_{\widetilde{b}}}$ respectively. The divergence decomposition for markov processes (see for example Lemma 15.1 in [28]) implies

$$\text{KL}(\mathbb{P}_{\text{Alg},f^{(\mathbf{P}_{\widetilde{a}})}} \parallel \mathbb{P}_{\text{Alg},f^{(\mathbf{P}_{\widetilde{b}})}}) = \sum_{a \in \mathcal{A}_{\text{tree}}} \mathbb{E}_{\text{Alg}} \left[n_T(a) \right] \text{KL}(\mathbb{P}_a^{\mathbf{P}_{\widetilde{a}}} \parallel \mathbb{P}_a^{\mathbf{P}_{\widetilde{b}}})$$

We have used the notation $\mathbb{E}_{\text{Alg}} \left[\cdot \right]$ above because the $n_T(a)$ random variables do not depend on f. Notice that for all $a \notin \{\widetilde{a}, \widetilde{b}\}$, $\text{KL}(\mathbb{P}_a^{\mathbf{P}_{\widetilde{a}}} \parallel \mathbb{P}_a^{\mathbf{P}_{\widetilde{b}}}) = 0$. In contrast $\text{KL}(\mathbb{P}_a^{\mathbf{P}_{\widetilde{a}}} \parallel \mathbb{P}_a^{\mathbf{P}_{\widetilde{b}}}) = 72\varepsilon^2$ for $a \in \{\widetilde{a}, \widetilde{b}\}$. Plugging this into the equation above,

$$\text{KL}(\mathbb{P}_{\text{Alg},f^{(\mathbf{P}_{\widetilde{a}})}} \parallel \mathbb{P}_{\text{Alg},f^{(\mathbf{P}_{\widetilde{b}})}}) = 72\mathbb{E}_{\text{Alg}} \left[n_T(\widetilde{a}) + n_T(\widetilde{b}) \right] \varepsilon^2 \leqslant \frac{\varepsilon^2 T}{2^{L-3}}.$$

24

Define the event $\mathcal{E} = \{\widehat{\pi}_T(\widetilde{a}) < \widehat{\pi}_T(\widetilde{b})\}$. The Huber-Bretagnolle inequality implies,

$$\mathbb{P}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{a}})}}\left(\mathcal{E}\right) + \mathbb{P}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{b}})}}\left(\mathcal{E}^c\right) \geqslant \exp\left(-\mathrm{KL}(\mathbb{P}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{a}})}} \parallel \mathbb{P}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{b}})}})\right) \geqslant \exp\left(-\frac{9\varepsilon^2 T}{2^{L-5}}\right). \quad (26)$$

Notice that when $\mathcal{E}$ holds, $2\widehat{\pi}_T(\widetilde{a}) \leqslant \widehat{\pi}_T(\widetilde{a}) + \widehat{\pi}_T(\widetilde{b}) \leqslant 1$ and therefore $\widehat{\pi}_T(\widetilde{a}) \leqslant \frac{1}{2}$. Thus if $\mathcal{E}$ holds,

$$\mathbb{E}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{a}})}}\left[\mathbb{E}_{a\sim\widehat{\pi}_T}\left[f_\star(a)\right]\Big|\mathcal{E}\right] \leqslant \frac{1}{2} + \frac{1}{2}\left(1-12\varepsilon\right) = 1-6\varepsilon. \quad (27)$$

In this case,

$$\mathbb{E}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{a}})}}\left[\mathbb{E}_{a\sim\widehat{\pi}_T}\left[f_\star(a)\right]\right] \leqslant \mathbb{E}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{a}})}}\left[\mathbb{E}_{a\sim\widehat{\pi}_T}\left[f_\star(a)\right]\Big|\mathcal{E}\right]\mathbb{P}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{a}})}}\left(\mathcal{E}\right) +$$

$$\mathbb{E}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{a}})}}\left[\mathbb{E}_{a\sim\widehat{\pi}_T}\left[f_\star(a)\right]\Big|\mathcal{E}^c\right]\mathbb{P}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{a}})}}\left(\mathcal{E}^c\right)$$

$$\leqslant \mathbb{E}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{a}})}}\left[\mathbb{E}_{a\sim\widehat{\pi}_T}\left[f_\star(a)\right]\Big|\mathcal{E}\right]\mathbb{P}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{a}})}}\left(\mathcal{E}\right) + \mathbb{P}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{a}})}}\left(\mathcal{E}^c\right)$$

$$\overset{(i)}{\leqslant} 1-6\varepsilon\mathbb{P}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{a}})}}\left(\mathcal{E}\right) \quad (28)$$

Where inequality (i) holds because of Equation 27. Similarly we can infer that,

$$\mathbb{E}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{b}})}}\left[\mathbb{E}_{a\sim\widehat{\pi}_T}\left[f_\star(a)\right]\right] \leqslant 1-\varepsilon\mathbb{P}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{b}})}}\left(\mathcal{E}\right). \quad (29)$$

Finally, combining Equations 28 and 29 with the result of Huber-Bretagnolle (see Equation 26) implies,

$$\mathbb{E}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{a}})}}\left[\mathbb{E}_{a\sim\widehat{\pi}_T}\left[f_\star(a)\right]\right] + \mathbb{E}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{b}})}}\left[\mathbb{E}_{a\sim\widehat{\pi}_T}\left[f_\star(a)\right]\right] \leqslant 2-6\varepsilon\exp\left(-\frac{9\varepsilon^2 T}{2^{L-5}}\right).$$

Hence, as long as $T \leqslant \frac{2^{L-5}}{9\varepsilon^2}$, then $\exp\left(-\frac{9\varepsilon^2 T}{2^{L-5}}\right) \geqslant \exp(-1)$ and therefore $\mathbb{E}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{a}})}}\left[\mathbb{E}_{a\sim\widehat{\pi}_T}\left[f_\star(a)\right]\right] + \mathbb{E}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{b}})}}\left[\mathbb{E}_{a\sim\widehat{\pi}_T}\left[f_\star(a)\right]\right] \leqslant 2-6\varepsilon\exp(-1)$. And therefore at least one $\widetilde{c} \in \{\widetilde{a}, \widetilde{b}\}$ satisfies

$$\mathbb{E}_{\mathrm{Alg},f^{(\mathbf{P}_{\tilde{c}})}}\left[\mathbb{E}_{a\sim\widehat{\pi}_T}\left[f_\star(a)\right]\right] \leqslant 1-\frac{3\varepsilon}{e} < 1-\varepsilon.$$

This finalizes the first result.

For the second result we analyze the following algorithm,

Algorithm 3 Adaptive Tree Sampling

1: Input: Number of samples per node M.
2: Initialize current node as $a_{\mathrm{curr}} = a_{1,1}$
3: **for** $j = 1 \ldots L-1$ **do**
4:    Expand current node $(b, c) = \mathrm{Children}(a_{\mathrm{curr}})$.
5:    Collect M samples from each node b and c and compute mean rewards $\hat{r}_b$ and $\hat{r}_c$.
6:    Update $a_{\mathrm{curr}} = \arg\max_{e\in(b,c)} \hat{r}_e$.
7: **end for**
8: **Output** $a_{\mathrm{output}} = a_{\mathrm{curr}}$.

Let's assume $f^{(\mathbf{P})}$ is the world's descriptor for some path $\mathbf{p}$. We will analyze the performance of Algorithm 3 where $M = \frac{\log(2L/\delta)}{\varepsilon^2}$. The output of this algorithm will be a policy centered around the output action a.

Let's first observe that Hoeffding inequality (Lemma A.1) and the union bound imply that $|\hat{r}_b - f^{(\mathbf{P})}(b)| \leqslant \varepsilon$, with probability at least $1 - \frac{\delta}{L}$ and $|\hat{r}_e - f^{(\mathbf{P})}(e)| \leqslant \varepsilon$ for $e \in \{b,c\}$ simultaneously. Let's call this event $\mathcal{E}_a$.

Lets assume $a_{\mathrm{curr}} \in \mathbf{p}$ and is not a leaf node, and w.l.o.g. when $b \in \mathbf{p}$ and $c \notin \mathbf{p}$, such that $f^{(\mathbf{P})}(b) \geqslant 1-2\varepsilon$ and $f^{(\mathbf{P})}(c) = 1-12\varepsilon$. If $\mathcal{E}_a$ holds,

$$\hat{r}_b \geqslant f^{(\mathbf{P})}(b) - \varepsilon \geqslant 1-3\varepsilon > 1-11\varepsilon \geqslant f^{(\mathbf{P})}(c) + \varepsilon \geqslant \hat{r}_c.$$

We conclude that in this case $\hat{r}_b > \hat{r}_c$ and therefore the updated node a (Line 6 of Algorithm 3) will also be in $\mathbf{p}$.

Combining these results with union bounds applied conditionally over all a_{curr} in Algorithm 3's sample path, we conclude that with probability at least $1 - \delta$, the output action a_{output} will equal the last action in $\mathbf{p}$'s path, and therefore the optimizer of $f^{(\mathbf{P})}$. Denote this event as $\mathcal{E}_\delta$.

If we denote this algorithm by $\mathrm{Alg}_{\text{adaptive}}$, define $T \geqslant 2(L-1)M$ and $\widehat{\pi}_T = \delta_{a_{\text{output}}}$ where a_{output} is the output action, we obtain

$$\mathbb{E}_{\mathrm{Alg}_{\text{adaptive}}}\left[\mathbb{E}_{a \sim \widehat{\pi}_T}\left[f(a)\right]\right] \geqslant \mathbb{E}_{\mathrm{Alg}_{\text{adaptive}}}\left[\mathbb{E}_{a \sim \widehat{\pi}_T}\left[f(a)\right]\Big|\mathcal{E}_\delta\right]\mathbb{P}_{\mathrm{Alg}_{\text{adaptive}}}(\mathcal{E}_\delta) \geqslant 1 - \delta$$

Thus, setting $\delta = \varepsilon$ and therefore for all $T \geqslant \frac{2L\log(2L/\varepsilon)}{\varepsilon^2}$ we obtain the desired result. $\qquad\square$

D.2 Eluder dimension of $\mathcal{F}_{\text{tree}}$

Lemma D.1. *The eluder dimension of the function class $\mathcal{F}_{\text{tree}}$ satisfies,*

$$d_{\text{eluder}}(\mathcal{F}_{\text{tree}}, \varepsilon) \geqslant 2^{L-1} - 1.$$

Proof. To prove this statement it is sufficient to exhibit an $\varepsilon-$independent sequence of actions that certifies the eluder dimension is large. To do this let's we use the ordered list of leaf actions. For simplicity we will call f_i to the function defined by the path ending at leaf node $a_{L,i}$. The sequence

Consider the sequence of actions $a_{L,1}, \cdots, a_{L,2^{L-1}-2}$. Notice that for all $n \in [2^{L-1} - 2]$ the two functions f_{n+1} and $f_{2^{L-1}}$ agree on $a_{L,1}, \cdots, a_{L,n}$. Thus for all $n \in [2^{L-1} - 2]$,

$$\sum_{i=1}^{n}(f_{n+1}(a_i) - f_{2^{L-1}}(a_i)) = 0,$$

while $f_{n+1}(a_{n+1}) - f_{2^{L-1}}(a_{n+1}) = 12\varepsilon > \varepsilon$. This shows $a_{L,1}, \cdots, a_{L,2^{L-1}-1}$ is a size $2^{L-1} - 1$ independent sequence. Since the eluder dimension equals the size of the largest independence sequence this implies the result. $\qquad\square$

E   Model Selection

E.1   Proof of Proposition 7.1

Proposition 7.1. *There exists a universal constant $\underline{c} > 0$ such that,*

$$\mathbb{E}_{x \sim \mathcal{P}, a \sim \text{Uniform}(\mathcal{A})}\left[\left(f_\star(x,a) - \widehat{f}_T^{(i)}(x,a)\right)^2\right] \leqslant \frac{\underline{c}\max^2(B, \bar{B})\log(T\max(M, |\mathcal{F}_{i_\star}|)/\delta)}{T}.$$

With probability at lest $1 - \delta$.

Proof. Let's start by considering $\widehat{f}_T^{(i_\star)}$. Standard analysis (Lemma A.8) yields that with probability at least $1 - \delta/2$

$$\mathbb{E}_{x \sim \mathcal{P}, a \sim \text{Uniform}(\mathcal{A})}\left[\left(f_\star(x,a) - \widehat{f}_T^{(i_\star)}(x,a)\right)^2\right] \leqslant \frac{2C_{T,\delta/2}^{(i_\star)}}{T}. \tag{30}$$

Let's write $r_\ell = f_\star(x_\ell, a_\ell) + \xi_\ell$ where ξ_ℓ are $1-$subgaussian conditionally zero mean. Lemma A.9 yields that,

$$\mathbb{E}_{x \sim \mathcal{P}, a \sim \text{Uniform}(\mathcal{A})}\left[\left(f_\star(x,a) - \widehat{f}_T^{(i)}(x,a)\right)^2\right]$$

$$\leqslant \frac{8}{T}\sum_{(x_\ell,a_\ell,r_\ell)\in\widetilde{\mathcal{D}}_T^{(\text{Test})}}\left(\widehat{f}_T^{(i)}(x_\ell, a_\ell) - r_\ell\right)^2 - \frac{8}{T}\sum_{\ell=1}^{T/2}\xi_\ell^2 + \frac{\bar{\underline{C}}'(\bar{B}^2 + B^2)\log(TM/\delta)}{T}$$

$$\tag{31}$$

for all $i \in [M]$ simultaneously (union bound) with probability at least $1 - \delta/2$ for some constant $\underline{\bar{C}}' > 0$. Incorporating the definition of $\underline{i}$ as the minimizer of the empirical loss (see Equation 7) into Equation 31 implies

$$
\mathbb{E}_{x \sim \mathcal{P}, a \sim \text{Uniform}(\mathcal{A})} \left[\left(f_\star(x,a) - \widehat{f}_T^{(\underline{i})}(x,a) \right)^2 \right]
$$

$$
\leqslant \frac{8}{T} \sum_{(x_\ell, a_\ell, r_\ell) \in \widetilde{\mathcal{D}}_T^{(\text{Test})}} \left(\widehat{f}_T^{(\underline{i})}(x_\ell, a_\ell) - r_\ell \right)^2 - \frac{8}{T} \sum_{\ell=1}^{T/2} \xi_\ell^2 + \frac{\underline{\bar{C}}'(\bar{B}^2 + B^2) \log(TM/\delta)}{T}
$$

$$
\leqslant \frac{8}{T} \sum_{(x_\ell, a_\ell, r_\ell) \in \widetilde{\mathcal{D}}_T^{(\text{Test})}} \left(\widehat{f}_T^{(i_\star)}(x_\ell, a_\ell) - r_\ell \right)^2 - \frac{8}{T} \sum_{\ell=1}^{T/2} \xi_\ell^2 + \frac{\underline{\bar{C}}'(\bar{B}^2 + B^2) \log(TM/\delta)}{T}
$$

with probability at least $1 - \delta/2$ where $\underline{i}$ is on the LHS and $i_\star$ is on the right hand side. Using the RHS of the inequality from Lemma A.9 yields that,

$$
\mathbb{E}_{x \sim \mathcal{P}, a \sim \text{Uniform}(\mathcal{A})} \left[\left(f_\star(x,a) - \widehat{f}_T^{(\underline{i})}(x,a) \right)^2 \right]
$$

$$
\leqslant 16 \mathbb{E}_{x \sim \mathcal{P}, a \sim \text{Uniform}(\mathcal{A})} \left[\left(f_\star(x,a) - \widehat{f}_T^{(i_\star)}(x,a) \right)^2 \right] + \frac{\underline{\bar{C}}''(\bar{B}^2 + B^2) \log(TM/\delta)}{T}
$$

$$
\tag{32}
$$

for some universal constant $\underline{\bar{C}}'' > 0$ with probability at least $1 - 3\delta/4$. Finally, the Least Squares guarantee of Lemma A.8 applied to $\mathcal{F}_{i_\star}$ and least squares estimator $\widehat{f}_T^{(i_\star)}$ (where all $\mathcal{P}_t = \mathcal{P}$) implies the first term of the RHS of the equation above can be bounded as,

$$
\mathbb{E}_{x \sim \mathcal{P}, a \sim \text{Uniform}(\mathcal{A})} \left[\left(f_\star(x,a) - \widehat{f}_T^{(i_\star)}(x,a) \right)^2 \right] \leqslant \frac{2C_{T/2, \delta/4}}{T}.
$$

with probability at least $1 - \frac{\delta}{4}$. Since in this case $C_{T/2, \delta/4} = \mathcal{O}\left((B^2 + \bar{B}^2) \log\left(\frac{T|\mathcal{F}_{i_\star}|}{\delta} \right) \right)$, combining this with Equation 32 and a union bound implies

$$
\mathbb{E}_{x \sim \mathcal{P}, a \sim \text{Uniform}(\mathcal{A})} \left[\left(f_\star(x,a) - \widehat{f}_T^{(\underline{i})}(x,a) \right)^2 \right] \leqslant \frac{\underline{c}(\bar{B}^2 + B^2) \log(T \max(M, |\mathcal{F}_{i_\star}|)/\delta)}{T}.
$$

for some universal constant $\underline{c} > 0$ and with probability at least $1 - \delta$. The result follows. $\qquad\square$