Tractable contextual bandits beyond realizability

Sanath Kumar Krishnamurthy, Vitor Hadad, and Susan Athey

Stanford University

October 27, 2020

Abstract

Tractable contextual bandit algorithms often rely on the realizability assumption – i.e., that the true expected reward model belongs to a known class, such as linear functions. We investigate issues that arise in the absence of realizability and note that the dynamics of adaptive data collection can lead commonly used bandit algorithms to learn a suboptimal policy. In this work, we present a tractable bandit algorithm that is not sensitive to the realizability assumption and computationally reduces to solving a constrained regression problem in every epoch. When realizability does not hold, our algorithm ensures the same guarantees on regret achieved by realizability-based algorithms under realizability, up to an additive term that accounts for the misspecification error. This extra term is proportional to T times the ⅖-root of the mean squared error between the best model in the class and the true model, where T is the total number of time-steps. Our work sheds light on the bias-variance trade-off for tractable contextual bandits. This trade-off is not captured by algorithms that assume realizability, since under this assumption there exists an estimator in the class that attains zero bias.

1   Introduction

Contextual bandit algorithms serve as a fundamental tool for online decision making and have been used in a wide range of settings from recommendation systems Agarwal et al. (2016) to mobile health Tewari and Murphy (2017), and due to their applicability over the past couple of decades there has been an increasing amount of research in contextual bandits Lattimore and Szepesvári (2020). However, the performance of many common algorithms

relies on an assumption called "realizability", which requires the analyst to possess some knowledge about the underlying data generating process – and often also relies on some luck that the process be relatively simple. When this assumption is satisfied, there exist algorithms that are statistically optimal and computationally tractable (in a sense we'll discuss more below). However, when it is violated, the performance of these algorithms can degrade in unexpected ways. The search for tractable algorithms that do not rely on this assumption an ongoing open problem Foster et al. (2019). In this work, we propose an algorithm that is optimal when the "realizability" assumption is satisfied and whose behavior is accurately characterized in its absence. We will also point to directions of research that may help do away with this assumption entirely.

Our underlying setup is the general stochastic contextual bandit setting. Using potential outcome notation, observations are represented as a sequence of iid random variables (x_t, r_t), where $x_t \in \mathcal{X}$ stands for a context in arbitrary set $\mathcal{X}$ and $r_t \in [0,1]^K$ is a vector of rewards, where $K := |\mathcal{A}|$ is the (finite) number of actions. Upon selecting one of action $a_t \in \mathcal{A}$, the algorithm observes $r_t(a_t)$. Therefore, the sequence of observed data points is $(x_t, a_t, r_t(a_t))$. Here t denotes the time-step which and is also the index for the sequence of observations. This sequence has length T, which may be known or unknown. A "policy" is a deterministic mapping from contexts to arms, representing a particular action selection strategy. Relative to the set of all policies $\mathcal{A}^{\mathcal{X}}$, we define the optimal policy as $\pi^* = \arg\max_{\pi \in \mathcal{A}^{\mathcal{X}}} \mathbb{E}[r_t(\pi(x_t))]$, where the expectation is taken over contexts and rewards.[1] The expressions "reward model" or "outcome model" refer to the conditional expectation of potential outcomes given the action and context, or $\mathbb{E}[r_t(a)|x]$, where the expectation is taken over rewards, and will often be represented as $f(x,a)$. We say that a reward model "induces a policy" π if $\pi(x) \in \arg\max_a \mathbb{E}[r_t(a)|x]$ for every x.

The goal of bandit algorithms is to find a sequence of actions that maximizes the sum of rewards observed during the experiment or, equivalently, to minimize cumulative regret, defined as the difference between the reward that was observed and that that would have been observed under the optimal policy,

$$R_T := \sum_{t=1}^{T} r_t(\pi^*(x_t)) - r_t(a_t). \tag{1}$$

The statistical performance of different algorithms is characterized by the rate at which (1) grows with the length of the experiment T.

Contextual bandit algorithms can often be categorized into three groups, depending on what is assumed about the underlying data-generating process. The first group of algorithms are the "agnostic" algorithms. These algorithms make no assumptions about the reward model, and they learn the best policy in some fixed class $\Pi \subseteq \mathcal{A}^{\mathcal{X}}$ while balancing the exploration-exploitation trade-off. To do this, these algorithms Beygelzimer et al. (2011); Dudik et al. (2011); Agarwal et al. (2014) need to construct a distribution over the policies Π in every epoch. Constructing this distribution is computationally challenging, and hence this

---

[1]Uniqueness of the optimal policy is not important for our results.

approach is colloquially referred to as the "Monster" Langford (2014). We now focus on the results in Agarwal et al. (2014) because computationally and statistically, they provide the state of the art agnostic algorithms. When Π is a finite class, Agarwal et al. (2014) present an algorithm called ILTCB that constructs a distribution with support size of $\mathcal{O}(\log|\Pi|)$ and the regret [2] against the best policy in Π scales at the rate $\tilde{\mathcal{O}}(KT\log|\Pi|)$. Each policy in the support of this distribution can be computed by solving the following cost-sensitive classification problem:

$$\arg\max_{\pi\in\Pi}\sum_{s=1}^{t}\hat{r}_s(\pi(x_s)), \tag{2}$$

where $(x_s,\hat{r}_s)$ is some sequence in $\mathcal{X}\times[0,1]^K$. When Π is large, the support of the distribution needed to be computed in every epoch of ILTCB may be large and hence would still be impractical to implement. To overcome this limitation, Agarwal et al. (2014) propose a heuristic called Online Cover (with parameter l) that computes a distribution over polices using the same approach as ILTCB but stops increasing the support of this distribution after computing some fixed number of policies l for the support. To the best of our knowledge there aren't any theoretical guarantees for Online Cover. Further, finding an exact solutions to (2) is generally intractable, so implementations of Online Cover use heuristics to solve this optimization problem.

The second group of algorithms requires knowledge about some set of functions $\mathcal{F}$ that is assumed to include the true reward model. That is, that there exists a function $f^*\in\mathcal{F}$ such that $f^*(x,a)=\mathbb{E}[r_t(a)|x]$ for all contexts and arms. This assumption is called "realizability", and it often allows for algorithms that computationally tractable and typically easier to implement. Computationally, algorithms in this category rely on being able to solve the regression problem

$$\hat{f}_t=\arg\min_{f\in\mathcal{F}}\sum_{s=1}^{t-1}(f(x_s,a_s)-r_s(a_s))^2, \tag{3}$$

or a weighted version of it, either online or offline. A routine that solves (3) is called "regression oracle". This class includes algorithms built on upper confidence bounds Li et al. (2010); Abbasi-Yadkori et al. (2011); Foster et al. (2018) or Thompson sampling Agrawal and Goyal (2013); Russo et al. (2018), and algorithms built on simple probabilistic selection strategies Abe and Long (1999); Foster and Rakhlin (2020); Simchi-Levi and Xu (2020). Regret rates for this class of algorithms are related to the complexity class of the outcome model $\mathcal{F}$, and under realizability optimal algorithms attain a rate of $\tilde{\mathcal{O}}(\sqrt{TK\log|\mathcal{F}|})$ for finite classes $\mathcal{F}$ (similar results are available for more general classes). In particular, the FALCON algorithm of Simchi-Levi and Xu (2020) will serve as the basis of our method attains this statistically optimal rate (so long as realizability holds) and is computationally tractable, in that the algorithm only needs to solve the problem (3) a small (at most logarithmic) number of times during the experiment.

---

[2]Note that while this notion of regret compares against the best policy in Π, the notion of regret used in this paper compares against the true optimal policy π^*.

A third set of bandit algorithms that does not fall neatly into any of the two categories above are algorithms that allow for a non-parametric model class. For example, in Rigollet and Zeevi (2010), Perchet et al. (2013) the reward model is assumed to be Hölder continuous but non-differentiable, and in Hu et al. (2020), Gur et al. (2019) it satisfies a Hölder smoothness assumption. The main characteristic of this class of algorithms is that they partition the covariate space into hypercubes of appropriate size and run multi-armed bandit algorithms within each cube. Depending on the smoothness of the reward model, there can be some information sharing across cubes that induces correlation across assignments in adjacent hypercubes and decreases regret. Although this is a very interesting direction of research, the structure of these algorithms forces the running time to exponentially depend on the context dimension, making them computationally intractable and hard to implement for most real life problems. Hence, for the rest of the paper, we will focus on the first two classes of algorithms.

1.1 The problem with realizability

As we have mentioned before, realizability is an extremely convenient and pervasive assumption in many tractable contextual bandit algorithms, but it is nevertheless very strong. In this section, we attempt to shed light on some issues that may arise in its absence.
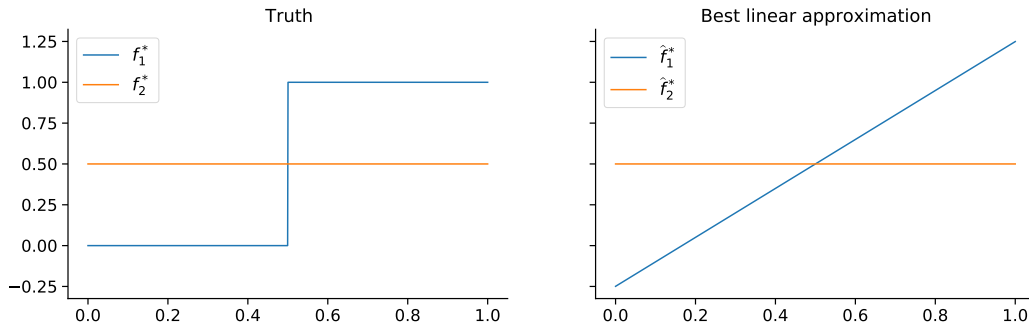

Figure 1: **Example**. Data-generating process (left) and its best approximation in the class of linear functions (right). The induced policies are the same.

To fix ideas, start from the following illustration. There are two arms and a single context is distributed uniformly on the unit interval. However, unbeknownst to the researcher, the conditional average rewards for each arm are a step function $f_1^*(x_t) := \mathbb{I}\{x_t > 0.5\}$ and a constant $f_2^*(x_t) \equiv 0.5$ (Figure 1). Rewards are observed with error $\epsilon_t \sim \mathcal{N}(0, .01)$. The researcher erroneously assumes that both can be realized in the class $\mathcal{F}$ of linear functions. Fortunately, in this example the best linear approximation $(\hat{f}_1^*, \hat{f}_2^*)$ induces a good policy. In fact, it coincides with the one the researcher would obtain if they had knowledge about the true function class – i.e., the policy induced by the best linear approximation $\hat{\pi}^*$ defined by $\hat{\pi}(x) = 1$ if $x > 0.5$ and $\hat{\pi}(x) = 2$ otherwise actually coincides with the policy induces

by the true model $\pi^*(x) := \arg\max_a \mathbb{E}[f^*(x, a)]$. Therefore, if the sequence of fitted models $(\hat{f}_{1,t}, \hat{f}_{2,t})$ converges to the best linear approximation $(\hat{f}_{1,t}^*, \hat{f}_{2,t}^*)$, regret should decay to zero asymptotically. However, as we will see next, this convergence may not happen.
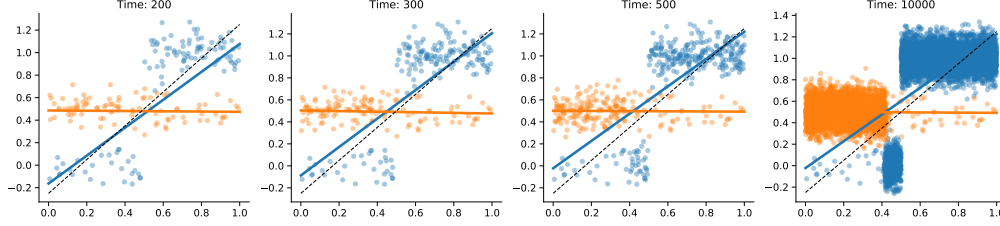

Figure 2: **Evolution of model estimates** for the UCB example. Although the model is initially correct, the distribution shift caused by the assignment mechanism ends up biasing reward estimates over time. Best linear approximation in black, dotted line. In our proposed algorithm, the blue and black lines are kept appropriately close, preventing this phenomenon.

Let us assume that the researcher collects data via LinUCB Li et al. (2010), with model updates in batches of 100 observations. Figure 2 shows the evolution of the estimated models $(\hat{f}_{1,t}, \hat{f}_{2,t})$ over time for a single simulation. After about a few hundred observations, the estimated model approximates the best linear approximation well and regret is small since the induced policy is nearly optimal. However, by continuing to assign treatments following this policy (plus some negligible exploration), the distribution of observations changes, which pushes the model away from the best linear approximation. In turn, this causes per-period regret to increase over time, as we show on Figure 3.
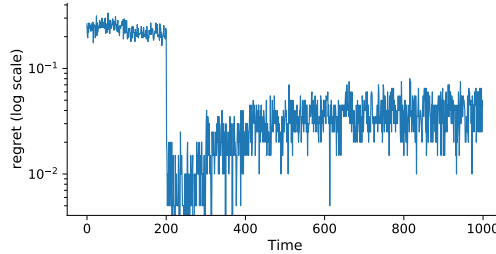

Figure 3: **Evolution of per-period regret** for the UCB example described in the text. Regret initially decreases because the estimated models $(\hat{f}_{1,t}, \hat{f}_{2,t})$ are close to the best linear approximations $(\hat{f}_1^*, \hat{f}_2^*)$, which in this example induces a good policy. However, as we gather more data that was collected with an exploitation objective, the estimated model diverges and the performance of the algorithm degrades. (Average across 100 simulations.)

The previous example demonstrates that in the absence of realizability the dynamics of adaptive data collection can lead the algorithm to learn a policy that is suboptimal relative

to the one that it would have learned under non-adaptive data collection. As an extreme thought example, one may also consider a situation in which actions are assigned via the optimal policy π^*. If we were to fit a linear model using exclusively this data, we would estimate that $\hat{f}_1 \equiv 1$ and $\hat{f}_2 \equiv 0.5$, which would in turn induce the policy $\pi(x) \equiv 1$ – a policy that always assigns arm 1 everywhere and therefore clearly suboptimal. In fact, more can be said. We can construct examples where even when the approximation error b is arbitrarily small, given data from the optimal policy, the confidence intervals used by LinUCB would tightly concentrate around a high regret policy, showing that the confidence intervals used by LinUCB are extremely sensitive to the realizability assumption (See Appendix E).

To prevent this phenomenon, in the next section we consider an algorithm that constraints the estimate of the outcome model $\hat{f}$ to be close to $\hat{f}^*$. This also allows us to derive upper bounds on regret in terms of the deviation of the best in-class model $\hat{f}^*$ from the true model f^*. This characterization is important as it allows us to take into account regret incurred due to model misspecification – a cost that often assumed away under realizability.

2   Main results

We propose an algorithm that we call Epsilon-FALCON, which is a modification of the "FAst Least-squares-regression-oracle CONtextual bandits", or FALCON algorithm described in Simchi-Levi and Xu (2020). The main departure from FALCON is that although we do posit some "tentative" set $\mathcal{F}$ that could contain the true outcome model, our regret guarantees do not depend on this assumption being satisfied. For simplicity of exposition we will initially assume that $\mathcal{F}$ is a convex subset of a d-dimensional linear space [3], but our results can be extended to more complex classes as we show later.

We will need some additional notation. Let f^* represent the true outcome model, i.e., $f^*(x, a) = \mathbb{E}[r_t(a)|x]$ for all x and a. Moreover, let $\hat{f}^*$ denote the best in-class approximation to the true outcome model when data is collected non-adaptively, or

$$\hat{f}^* := \arg\min_{f \in \mathcal{F}} \mathbb{E}_{x \sim D_{\mathcal{X}}} \mathbb{E}_{a \sim \text{Unif}(\mathcal{A})} [(f(x, a) - f^*(x, a))^2], \tag{4}$$

where $D_{\mathcal{X}}$ is the distribution of contexts, and $\text{Unif}(\mathcal{A})$ is a probability distribution that assigns equal probability to every arm. The approximation error between these two functions is denoted as

$$b := \mathbb{E}_{x \sim D_{\mathcal{X}}} \mathbb{E}_{a \sim \text{Unif}(\mathcal{A})} [(\hat{f}^*(x, a) - f^*(x, a))^2]. \tag{5}$$

Naturally, the approximation error (5) will be zero when realizability holds. And when it doesn't hold, we will show that the algorithm will incur some regret whose upper bound

---

[3] Consider the class of estimators $\mathcal{F}$ where linear functions estimate rewards for each arm using a total of d parameters. Note that this is a special case of requiring $\mathcal{F}$ to be a convex subset of a d-dimensional linear space. Hence, the guarantees in Theorem 1 hold for stochastic linear bandits.

increases with the approximation error. This is what allows us to accurately characterize the cost that we pay when we $\mathcal{F}$ is misspecified (i.e., $f^* \notin \mathcal{F}$).

Algorithm: Epsilon-FALCON is implemented in increasing epochs (batches) that are indexed by m. Each epoch m begins at period τ_{m-1}, we set epoch schedule so that $\tau_0 = 0$, $\tau_1 \geq 4$, and $\tau_{m+1} = 2\tau_m$ for any epoch $m \geq 1$. Every epoch m starts out with an estimated reward model $\hat{f}_m$ obtained at the end of the last batch, with $\hat{f}_1 \equiv 0$. For a fraction ϵ of each epoch, called the "passive" phase, Epsilon-FALCON assigns arms uniformly at random. For the remaining $1 - \epsilon$ fraction of the epoch, in what we call the "active" phase, it acts as a modified version of FALCON.[4]

Our action selection mechanism is the same as FALCON's, so let's briefly review it. At each epoch m, given the current reward model estimate $\hat{f}_m$ and a scaling parameter $\gamma_m > 0$, actions are drawn from the probability distribution described by the following "action selection kernel",

$$p_m(a|x) := \begin{cases} \frac{1}{K + \gamma_m \left(\hat{f}_m(x,\hat{a}) - \hat{f}_m(x,a) \right)} & \text{for } a \neq \hat{a} \\ 1 - \sum_{a' \neq \hat{a}} p(a'|x) & \text{for } a = \hat{a}. \end{cases} \tag{6}$$

where $\hat{a} = \max_a \hat{f}_m(a,x)$ is the best predicted action. The assignment rule (6) ensures that actions that are predicted to be good according to the current model estimate $\hat{f}_m$ are given higher probability. The scaling parameter, set to $\gamma_m \simeq \sqrt{K(\tau_{m-1} - \tau_{m-2})/(d \ln(m/\delta))}$ with initial values $\gamma_1 = 1$, control the degree of exploration during the active phase, with higher values of γ_m indicating less exploration. We may sometimes refer to γ_m and ϵ as the active and passive exploration parameters respectively.

The main difference between our method and FALCON is in how we estimate the outcome model $\hat{f}_{m+1}$ from data collected in the previous epoch m. The original algorithm simply uses the estimator that minimizes empirical risk on data collected in the previous time-steps, but as we saw in the example in Section 1.1, when realizability fails the sequence of estimators $\hat{f}_m$ may not converge to $\hat{f}^*$. This is due to the fact that the empirical risk minimizer when data is collected adaptively may be very different from the one attained when data is collected non-adaptively, and its performance may not be well understood (See Figure 4). In order to ensure that our estimates converge to $\hat{f}^*$, our algorithm uses a "*constrained* regression oracle" that ensures that the estimated model is always close to the best approximation $\hat{f}^*$. Let's see how this is done.

Denote the data collected using the passive and active phases of the epoch m by S'_m and S_m respectively. Moreover, let $\mathcal{F}'_m$ denote the subset of functions $f \in \mathcal{F}$ for which the following constraint in satisfied,

$$\sum_{(x,a,r(a)) \in S'_m} (f(x,a) - r(a))^2 \leq \alpha_m + C_1 d \ln(12m^2/\delta) \tag{7}$$

---

[4]More precisely, it acts as a modified version of the FALCON+ algorithm in the same paper, but the distinction is minor enough that we will ignore it for the purposes of naming our method.

where $\alpha_m := \min_{g \in \mathcal{F}} \sum_{(x,a,r(a)) \in S'_m} (g(x,a) - r(a))^2$ is the sum of squared residuals in the model fitted on the data collected in the "passive" phase, and C_1 is a constant chosen appropriately to ensure that $\hat{f}^*$ also lies in $\mathcal{F}'$ with probability at least $1 - \delta/(12m^2)$. [5]

The estimated model $\hat{f}_{m+1}$ will be constrained to lie in this set. More specifically, it is the output of the following constrained regression problem:

$$\min_{f \in \mathcal{F}} \sum_{(x,a,r(a)) \in S_m} (f(x,a) - r(a))^2$$
$$\text{s.t.} \quad f \in \mathcal{F}'_m. \tag{8}$$

The intuition, again, is that since $\hat{f}_{m+1} \in \mathcal{F}'_m$ by construction, and since $\hat{f}^* \in \mathcal{F}'_m$ with high probability, the two will likely remain close. And since $\mathcal{F}'_m$ shrinks over time, $\hat{f}_{m+1}$ must ultimately converge to $\hat{f}^*$. Therefore, the convergence issues we saw in our example in Section 1.1 cannot happen. This is what allows us to derive regret guarantees even when realizability fails (see Figure 4 for an intuitive illustration). The full description and the pseudocode for the general algorithm can be found in the Appendix (Algorithm 1). [6]
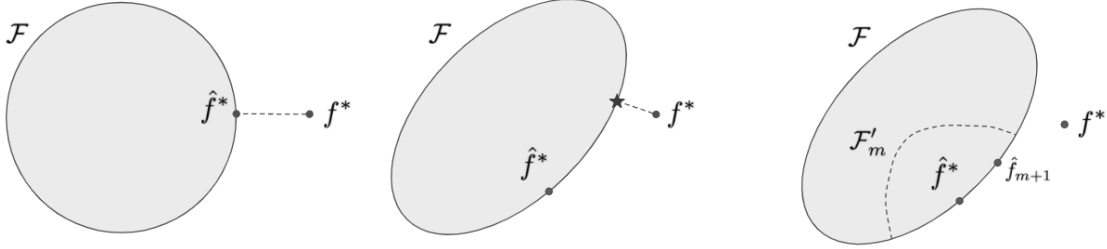

Figure 4: Intuition for our method. Left: the function $\hat{f}^*$ is the best in-class approximation to f^* under non-adaptive data collection. Middle: under a different distribution, the best in-class approximation (starred) may lie very far away from $\hat{f}^*$, and there are no guarantees on its performance. Right: in our method, we construct a shrinking sequence of sets $\mathcal{F}'_m$ that contain $\hat{f}^*$ with high probability, and ensure that our model estimates lie in this set.

Computational tractability of the constrained regression problem: Note that Epsilon-FALCON is very easy to implement given a constrained regression oracle. Hence, for the computational tractability of Epsilon-FALCON, it is sufficient to argue that the constrained regression problem is computationally tractable. When $\mathcal{F}$ is the class of linear reward models, then clearly the constrained regression problem is a convex and can be solved efficiently. In general, when $\mathcal{F}$ is any convex class, we show that the constrained regression problem can be solved efficiently with a weighted regression oracle (see Appendix D). Hence we can use the abundance of existing algorithms for weighted regression as a subroutine to

---

[5]We pin down the value of this constant in the Appendix

[6]Except for the choice of γ_m and the RHS of the constraint Equation (7), the algorithm for general $\mathcal{F}$ is the same as the description in this section.

solve the constrained regression problem. While this is one approach to solve the constrained regression problem, in practice, directly solving the constrained regression problem may be faster.

Theorem 1 provides a high probability regret guarantee for Epsilon-FALCON when $\mathcal{F}$ is a convex subset of some d-dimensional linear space.

Theorem 1 (Linear case). *Suppose $\mathcal{F}$ is a convex subset of a d-dimensional linear space. With probability at least $1 - \delta$, Epsilon-FALCON with passive exploration parameter $\epsilon > 0$ attains the following regret guarantee:*

$$R_T \leq \mathcal{O}\left(\sqrt{KTd\ln\left(\frac{\ln(T)}{\delta}\right)} + KT\sqrt{\frac{b}{\sqrt{\epsilon}}} + \epsilon T \right). \tag{9}$$

The guarantees in (9) consist of three terms. The first term is the regret due to the complexity of the class $\mathcal{F}$, and is the bound guaranteed by realizability based algorithms like FALCON under realizability. The second term can be interpreted as the "cost of misspecification", this term depends on the approximation error b and the passive exploration parameter ϵ. Finally, the third term is the regret incurred in the passive phase and depends only on the passive exploration parameter ϵ.

At first glance, the result in (9) may look rather weak due to the linear dependence in the horizon T. However, we contend that any algorithm that that works with a restricted class of policies or reward models, including agnostic algorithms like ILTCB Agarwal et al. (2014), will incur some linear regret if these restrictions are violated. In Theorem 1 we simply make this issue explicit, as one of our goals is to accurately characterize the bias-variance trade-off in our problem. Our results show that, if the practitioner is willing to spend ϵT regret in the passive phase, then in the active phase excess regret due to misspecification will be $\mathcal{O}\left(KT\sqrt{b/\sqrt{\epsilon}}\right)$. On the other hand, realizability based approaches do not have any guarantees under general misspecification.

As a thought experiment, suppose we knew the approximation error b or could make an educated guess about it, then we could choose ϵ as a function of b so as to optimize (9). In that case, our algorithm would attain the bound in Corollary 1.

Corollary 1 (Linear case with known b). *Under the setting of Theorem 1, if the passive exploration parameter is set to $\epsilon = \Theta(K^{4/5}b^{2/5})$, we have the following corollary.*

$$R_T \leq \mathcal{O}\left(\sqrt{KTd\ln\left(\frac{\ln(T)}{\delta}\right)} + K^{4/5}b^{2/5}T \right). \tag{10}$$

This result is interesting because it tells us that if we were able to tune the passive exploration parameter optimally, we get improved regret rates that only depend on the complexity of $\mathcal{F}$ and the approximation error b. Thus, achieving a bias-variance trade-off over the entire horizon T. This suggests that tuning ϵ by estimating b may be a promising direction for future work to get algorithms with better regret guarantees.

Understanding the constrained regression problem: Having explained the overall algorithm, let's now understand the constrained regression problem in a bit more detail, so the reader will be able to follow the proof steps in the Appendix.

At the end of epoch m, we have the two kinds of data, that is the data from the passive phase of the epoch and the data from the active phase of the epoch. The data from the passive phase is used to construct $\mathcal{F}'_m$, and contains the best in-class approximation of the true outcome model $\hat{f}^*$ with high probability (see Lemma 7). The data from the active phase is used to select a "good" estimate within $\mathcal{F}'_m$ which in turn induces a "good" action selection kernel. A good action selection kernel has low regret, and ensures that the data generated by this kernel can be used to construct "good" estimates in the next epoch. In terms of exploration, there is a trade-off between these two properties as more exploration helps you generate "good" data but incurs higher regret. In terms of estimates, both these properties are related because good estimates come from good data. For simplicity let us focus on arguing that the action selections kernels we estimate generate "good" data and believe that the active exploration parameter γ_m is set optimally. In particular, we say the action selection kernel generates "good" data if the reward of the policy induced by $\hat{f}^*$ can be estimated using the data generated by this kernel. Note that this is trivially ensured when actions are selected uniformly at random, as we did in the first epoch. In later epochs, as the kernel gets less explorative (γ_m increases), to ensure this we need the estimator ($\hat{f}_{m+1}$) that induces this action selection kernel to be close to the best in-class model ($\hat{f}^*$). More mathematically, as shown in Lemma 9, we need the root mean squared difference between $\hat{f}_{m+1}$ and $\hat{f}^*$ to shrink at the same rate as γ_m increases. This property is guaranteed by the fact that both $\hat{f}_{m+1}$ and $\hat{f}^*$ lie in $\mathcal{F}'_m$ with high probability, and by the fact that $\mathcal{F}'_m$ is sufficiently small as we have collected enough data in the passive phase to ensure this (see Lemma 7). Additionally, this property helps us ensure that our action selection kernels p_m are stable over time, in the sense that if the reward of a policy could be estimated from the data generated by p_{m+1} (in expectation) then the reward of this policy could also be estimated by the data generated by p_m (in expectation), see Lemma 10 for a more formal statement. In other words, the set of policies that we implicitly consider do not erratically change over time and only decrease.

General classes of outcome models: Although for concreteness we have explained our results when $\mathcal{F}$ is a convex subset of a d-dimensional linear space, Theorem 1 readily extends to more general classes of functions. In particular we can extend Theorem 1 whenever $\mathcal{F}$ is a convex and satisfies Assumption 1. In therms of the algorithm, except for the choice of γ_m and the RHS of the constraint Equation (7), Epsilon-FALCON for general $\mathcal{F}$ is the same as the description in this section. See Algorithm 1 in the Appendix for more details. Stating Assumption 1 can get cumbersome quickly, here we state an informal version of this assumption, followed by Theorem 2, and applications of this Theorem to various convex classes $\mathcal{F}$. In what follows $\mathbf{comp}(\mathcal{F})$ will denote an appropriate measure of complexity, like VC subgraph dimension or entropy.

Main Assumption: We now state an informal version of Assumption 1. Let n denote the number of data points collected from some distribution. Suppose we have $\rho \in (0, 1]$, $\rho' \in [0, \infty)$, and $C > 0$. Further suppose for any convex subset $\mathcal{F}'$ of $\mathcal{F}$ and $\zeta \in (0, 1/2)$, with probability $1 - \zeta$, for any $\eta \geq C \ln^{\rho'}(n) \ln(1/\zeta) \mathbf{comp}(\mathcal{F})/n^\rho$, the empirical and true risks of any estimators in $\mathcal{F}'$ are "close" in the following sense:

- If the population risk of any estimator in $\mathcal{F}'$ is smaller than η, then its empirical risk is not larger than $3\eta/2$.

- If the empirical risk of any estimator in $\mathcal{F}'$ is smaller than η, then its population risk is not larger than 2η.

Theorem 2 (Main result). *Suppose $\mathcal{F}$ is a convex set and suppose Assumption 1 holds. Then with probability at least $1 - \delta$, Epsilon-FALCON with passive exploration parameter $\epsilon > 0$ attains the following regret guarantee:*

$$R_T \leq \mathcal{O}\left(\sqrt{KT^{2-\rho} \ln^{\rho'}(T) \ln(\frac{\ln(T)}{\delta}) \mathbf{comp}(\mathcal{F})} + KT\sqrt{\frac{b}{\sqrt{\epsilon^\rho}}} + \epsilon T \right). \tag{11}$$

In Appendix C, we provide convenient Lemmas to prove Assumption 1 for various convex classes $\mathcal{F}$. These Lemmas directly follow from results in Koltchinskii (2011). In fact, Theorem 1 is implied by Theorem 2 and results stated in Appendix C. We now go over similar results that follow from Theorem 2 and Appendix C.

Example 1: Suppose $\mathcal{F}$ is convex and has VC-subgraph dimension V. Then with probability $1 - \delta$, Epsilon-FALCON guarantees the following bound on the regret:

$$R_T \leq \mathcal{O}\left(\sqrt{KTV \ln\left(\frac{T}{V}\right) \ln\left(\frac{\ln(T)}{\delta}\right)} + KT\sqrt{\frac{b}{\sqrt{\epsilon}}} + \epsilon T \right).$$

Example 2: Suppose $\mathcal{F}$ is a convex hull of class with VC-subgraph dimension V. Then with probability $1 - \delta$, Epsilon-FALCON guarantees the following bound on the regret:

$$R_T \leq \mathcal{O}\left(\sqrt{KT^{\frac{2+3V}{2+2V}} V^{\frac{2+V}{2+2V}} \ln\left(\frac{\ln(T)}{\delta}\right)} + KT\sqrt{\frac{b}{\sqrt{\epsilon}}} + \epsilon T \right).$$

Example 3: Suppose for some $\rho \in (0, 1)$, the empirical entropy is bounded by $\mathcal{O}(\epsilon^{-2\rho})$ for all empirical distributions. Then with probability $1 - \delta$, Epsilon-FALCON guarantees the following bound on the regret:

$$R_T \leq \mathcal{O}\left(\sqrt{KT^{\frac{1+2\rho}{1+\rho}} \ln\left(\frac{\ln(T)}{\delta}\right)} + KT\sqrt{\frac{b}{\sqrt{\epsilon^{1/(1+\rho)}}}} + \epsilon T \right).$$

3   Discussion

To summarize, our contributions in this work is twofold. First, to illustrate how algorithms that rely on realizability may incur unexpected regret when this assumption is violated. We saw in Section 1.1 that one can construct examples where regret is large even in relatively benign settings.

Our second contribution is to propose a flexible family of computationally tractable algorithm that are less sensitive to realizability. As we see in Section 2, our analysis characterizes the behavior of regret under misspecification, and gives us insight into the bias-variance trade-off in contextual bandits.

In terms of algorithm design, Epsilon-FALCON inherits the computational elegance of realizability based approaches like FALCON. In particular, a single estimator gives you an implicit distribution over policies via the action selection kernel and bypasses the need to explicitly construct a distribution over policies. Our key insight is that by using a constrained regression estimator, we can make this approach robust to misspecification at the expense of some additional regret in the passive phase.

To close, we note that these results hint at the possibility of exploiting the bias-variance trade-off in tractable contextual bandits to perform good model selection. This would be an interesting direction for future work.

References

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

Abe, N. and Long, P. M. (1999). Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pages 3–11. Citeseer.

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

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

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135.

Bertsekas, D. P. and Scientific, A. (2015). *Convex optimization algorithms*. Athena Scientific Belmont.

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

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. (2011). Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*.

Foster, D. J., Agarwal, A., Dudík, M., Luo, H., and Schapire, R. E. (2018). Practical contextual bandits with regression oracles. *arXiv preprint arXiv:1803.01088*.

Foster, D. J., Krishnamurthy, A., and Luo, H. (2019). Model selection for contextual bandits. In *Advances in Neural Information Processing Systems*, pages 14741–14752.

Foster, D. J. and Rakhlin, A. (2020). Beyond ucb: Optimal and efficient contextual bandits with regression oracles. *arXiv preprint arXiv:2002.04926*.

Gur, Y., Momeni, A., and Wager, S. (2019). Smoothness-adaptive stochastic bandits. *arXiv preprint arXiv:1910.09714*.

Hu, Y., Kallus, N., and Mao, X. (2020). Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. In *Conference on Learning Theory*, pages 2007–2010.

Jin, C., Netrapalli, P., and Jordan, M. I. (2019). What is local optimality in nonconvex-nonconcave minimax optimization? *arXiv preprint arXiv:1902.00618*.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media.

Langford, J. (2014). Interactive machine learning.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

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

Perchet, V., Rigollet, P., et al. (2013). The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41(2):693–721.

Rigollet, P. and Zeevi, A. (2010). Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630*.

Russo, D. J., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. (2018). A tutorial on thompson sampling.

Simchi-Levi, D. and Xu, Y. (2020). Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Available at SSRN*.

Tewari, A. and Murphy, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer.

A   Detailed setup

In Appendix B we will prove the claims in the body of the paper. This requires us to establish some additional notation, which we do in Appendix A.1. Most of these symbols and definitions were used in the original FALCON paper Simchi-Levi and Xu (2020). The results in Appendix C use notation and definitions from Koltchinskii (2011) and are stated within Appendix C. Appendix A.2 states the main assumption used in Theorem 2, and Appendix A.3 describes the general version of Epsilon-FALCON.

A.1   Preliminaries

To start, let Γ_t denote the set of observed data points up to and including time t. That is

$$\Gamma_t := \{(x_s, a_s, r_s(a_s))\}_{s=1}^t \tag{12}$$

Recalling the text, an "action selection kernel" p gives us the probability $p(a|x)$ of selecting an arm a given a context x, and a "policy" is a deterministic mapping from contexts to actions. Let $\Psi = \mathcal{A}^{\mathcal{X}}$ denote the universal policy space containing all possible policies. Following Lemma 3 in Simchi-Levi and Xu (2020), given any action selection kernel p we can construct a unique product probability measure on Ψ, given by:

$$Q_p(\pi) := \prod_{x \in \mathcal{X}} p(\pi(x)|x), \tag{13}$$

and it satisfies the following property

$$p(a|x) = \sum_{\pi \in \Psi} \mathbb{I}\{\pi(x) = a\} Q_p(\pi). \tag{14}$$

Property (14) establishes a duality between action selection kernels, which are used in practice in the algorithm implementation, and the probability distribution (13), which is a theoretical object that can be used to simplify the proofs below. For short-hand, we let $Q_m \equiv Q_{p_m}$ denote the product probability measure on Ψ induced by the action selection kernel p_m defined in (6).

Now, for any action selection kernel p and any policy π, we let $V(p, \pi)$ denote the expected inverse probability.

$$V(p, \pi) := \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}} \left[\frac{1}{p(\pi(x)|x)} \right] \tag{15}$$

One can interpret (15) as a measure of average divergence between $p(\cdot|x)$ and $\pi(x)$. Simchi-Levi and Xu (2020) refer to this as the decisional divergence between the randomized policy Q_p and deterministic policy π.

Given an outcome model f and policy π, we can define the expected instantaneous reward of the policy π with respect to the model f as

$$R_f(\pi) := \mathop{\mathbb{E}}_{x \in \mathcal{X}}[f(x, \pi(x))]. \tag{16}$$

When there is no possibility of confusion, we will write $R(\pi)$ to mean $R_{f^*}(\pi)$, the reward with respect to the true model f^*.. The policy π_f induced by the model f is defined by setting $\pi_f(x) := \max_a f(x, a)$ for every x. Note that this policy has the highest instantaneous reward with respect to the model f, that is $\pi_f = \arg\max_{\pi \in \Psi} R_f(\pi)$. We can also define the expected instantaneous regret with respect to the outcome model f as

$$\operatorname{Reg}_f(\pi) := \mathop{\mathbb{E}}_{x \in \mathcal{X}}[f(x, \pi_f(x)) - f(x, \pi(x))]. \tag{17}$$

When there is no possibility of confusion, we will write $\operatorname{Reg}(\pi)$ to mean $\operatorname{Reg}_{f^*}(\pi)$, the regret with respect to the true model f^*.

Recall that we define $\hat{f}^*$ as the best in-class approximation to the true outcome model when arms are sampled uniformly at random. Also recall that we define b as the approximation error or mean squared difference between $\hat{f}^*$ and f^* when arms are sampled uniformly at random. We now define B to be the largest mean squared difference between $\hat{f}^*$ and f^* under any action selection kernel. That is, [7]

$$B := \max_p \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a \sim p(\cdot|x)}[(\hat{f}^*(x, a) - f^*(x, a))^2] = \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}}[\max_a(\hat{f}^*(x, a) - f^*(x, a))^2]. \tag{18}$$

A.2 Main assumption

Assumption 1. *Suppose that our outcome model $\mathcal{F}$ satisfies the following property. There exists constants $C > 0$, $\rho \in (0, 1]$, $\rho' \in [0, \infty)$ such that for any action selection kernel p, any convex subset $\mathcal{F}' \subset \mathcal{F}$, any natural number n, any $\zeta \in (0, 1)$, and any $\eta > C \ln^{\rho'}(n) \ln(1/\zeta) \mathbf{comp}(\mathcal{F})/n^\rho$, the following holds with probability at least $1 - \zeta$:*

$$\mathcal{F}'(\eta, p) \subseteq \widehat{\mathcal{F}'}(3\eta/2, \widetilde{S}) \qquad \text{and} \qquad \widehat{\mathcal{F}'}(\eta, \widetilde{S}) \subseteq \mathcal{F}'(2\eta, p), \tag{19}$$

where the η-minimal set is defined as

$$\mathcal{F}'(\eta, p) := \left\{ f \in \mathcal{F}' \,\middle|\, \mathbb{E}[(f(x_i, a_i) - r_i(a_i))^2] \le \min_{\tilde{f} \in \mathcal{F}'} \mathbb{E}[(\tilde{f}(x_i, a_i) - r_i(a_i))^2] + \eta \right\}, \tag{20}$$

and the empirical η-minimal set *is defined as*

$$\widehat{\mathcal{F}'}(\eta, \widetilde{S}) := \left\{ f \in \mathcal{F} \,\middle|\, \frac{1}{n} \sum_{i=1}^n (f(x_i, a_i) - r_i(a_i))^2 \le \min_{\tilde{f} \in \mathcal{F}'} \frac{1}{n} \sum_{i=1}^n (\tilde{f}(x_i, a_i) - r_i(a_i))^2 + \eta \right\}. \tag{21}$$

and where the data $\widetilde{S} \equiv (x_i, a_i, r_i(a_i))_{i=1}^n$ are drawn independently and identically from $x_i \sim D_{\mathcal{X}}$, $a_i|x_i \sim p(\cdot|x_i)$ and $r_i \sim D_{r_i|x_i, a_i}$, and the expectations are taken with respect to these distributions.

---

[7]Lemma 1 bounds B with Kb.

A.3 Algorithm

Algorithm 1 Epsilon-FALCON

---

input: epoch schedule $\tau_1 \geq 4$, confidence parameter δ, and forced exploration parameter ϵ.

1: Set $\tau_0 = 0$, and $\tau_{m+1} = 2\tau_m$ for all $m \geq 1$.

2: Let $\hat{f}_1 \equiv 0$.

3: **for** epoch $m = 1, 2, \ldots$ **do**

4:      Let $\gamma_m = \sqrt{\dfrac{C_3 K(\tau_{m-1} - \tau_{m-2})^\rho}{\ln^{\rho'}(\tau_{m-1} - \tau_{m-2})\ln((m-1)/\delta)\mathbf{comp}(\mathcal{F})}}$ (for epoch 1, $\gamma_1 = 1$).

5:      **for** round $t = \tau_{m-1} + 1, \ldots, \tau_m - \lceil\epsilon(\tau_m - \tau_{m-1})\rceil$ **do**

6:          Observe context x_t, let $\hat{a}_t = \arg\max_{a \in \mathcal{A}} \hat{f}_m(x_t, a)$, and define:

$$p_t(a) := \begin{cases} \frac{1}{K + \gamma_m(\hat{f}_m(x_t, \hat{a}_t) - \hat{f}_m(x_t, a))}, & \text{for all } a \neq \hat{a}_t \\ 1 - \sum_{a' \neq \hat{a}_t} p(a'|x), & \text{for } a = \hat{a}_t \end{cases}$$

7:          Sample $a_t \sim p_t(\cdot)$ and observe $r_t(a_t)$.

8:      **end for**

9:      **for** round $t = \tau_m - \lceil\epsilon(\tau_m - \tau_{m-1})\rceil + 1, \ldots, \tau_m$ **do**

10:         Observe context x_t, sample a_t uniformly at random from $\mathcal{A}$, and observe $r_t(a_t)$.

11:      **end for**

12:      Let:

$$S_m = \{(x_t, a_t, r_t(a_t))\}_{t=\tau_{m-1}+1}^{\tau_m - \lceil\epsilon(\tau_m - \tau_{m-1})\rceil}$$
$$S'_m = \{(x_t, a_t, r_t(a_t))\}_{t=\tau_m - \lceil\epsilon(\tau_m - \tau_{m-1})\rceil+1}^{\tau_m}.$$

13:      Compute $\hat{f}_{m+1}$ by solving

$$\min_{f \in \mathcal{F}} \sum_{(x,a,r(a)) \in S_m} (f(x,a) - r(a))^2 \tag{22}$$
$$\text{s.t.} \quad f \in \mathcal{F}'_m.$$

     where $\mathcal{F}'_m$ is defined as in (24).

14: **end for**

The general version of our algorithm for general classes of outcome models $\mathcal{F}$ requires three modifications. Note the constants C, ρ, and ρ' mentioned below are rate terms from Assumption 1, $C_3 := 1/(4C_5)$ (see Lemma 8), and $C_5 := 2C \times 4^\rho \times (2 + \ln(12))$ (see Lemma 7).

First, the epoch schedule needs to satisfy $\tau_0 = 0$, $\tau_1 \geq 4$ and for subsequent epochs we set $\tau_{m+1} = 2\tau_m$.

Second, the parameter γ_t is set to $\gamma_1 = 1$ and

$$\gamma_m = \sqrt{\frac{C_3 K(\tau_{m-1} - \tau_{m-2})^\rho}{\ln^{\rho'}(\tau_{m-1} - \tau_{m-2})\ln((m-1)/\delta)\mathbf{comp}(\mathcal{F})}}. \tag{23}$$

Third and finally, the constraint set $\mathcal{F}'_m$ consists of the set of outcome models $f \in \mathcal{F}$ such that

$$\mathcal{F}'_m := \left\{ f \in \mathcal{F} \;\middle|\; \frac{1}{|S'_m|} \sum_{S'_m} (f_{m+1}(x,a) - r(a))^2 \leq \alpha_m + \frac{C_1 \ln^{\rho'}(|S'_m|) \ln(1/\delta') \mathbf{comp}(\mathcal{F})}{|S'_m|^\rho} \right\},$$
(24)

where $\alpha_m := \frac{1}{|S'_m|} \min_{g \in \mathcal{F}} \sum_{S'_m} (g(x,a) - r(a))^2$, $\delta' = \delta/(12m^2)$, and $C_1 = 3C/2$ (see Lemma 7).

B Proofs

The goal of this section is to present our proof of Theorem 2. Section B.1 gives a brief overview of the argument. Section A.2 restates the main assumption. Sections B.2-B.8 prove auxiliary Lemmas, and finally Section B.9 concludes with a proof of the theorem. A small, more technical, portion of the argument is deferred to Section C.

B.1 Overview of the proof for Theorem 2

For convenience, here is an informal, abridged version of the argument used in the proofs. We hope the reader will find it useful to navigate the results that follow.

- First of all, during the passive phase we always incur ϵT regret. For the remainder let's consider the regret incurred during periods occurring in the active phase of each epoch.

- The cumulative regret incurred across the active phases will be close to the sum of its conditional expectations at each period,

$$\sum_{t \in \mathcal{T}_{\text{active}}} r_t(\pi^*(x)) - r_t(a_t) \approx \sum_{t \in \mathcal{T}_{\text{active}}} \mathbb{E}[r_t(\pi^*(x)) - r_t(a_t)|\Gamma_{m(t)-1}] \qquad \text{w.h.p.,}$$

so we only need to bound these conditional expectations.

- By Lemma 3, the conditional expectation of instantaneous regret at period t in the active phase of epoch m can be rewritten in terms of the probability measure Q_m over policies,

$$\mathbb{E}[r_t(\pi^*(x)) - r_t(a_t)|\Gamma_{m(t)-1}] = \sum_{\pi \in \Psi} Q_m(\pi) \text{Reg}_{f^*}(\pi).$$

- By design, our method will produce a sequence of actions such that the *estimated* regret $\text{Reg}_{\hat{f}_m}(\pi)$ is small for the policies that receive high probability under Q_m (see Lemma 4). In order to show that the *expected* regret $\text{Reg}_{f^*}(\pi)$ is also small, we need to show that the two are "close", at least for policies that receive high probability under Q_m.

- Naturally the difference between expected and estimated regret depends on how closely the sequence $\hat{f}_m$ approximates f^*. In Lemma 7, we characterize this approximation as a function of two objects: the expected distance between $\hat{f}_m$ and the best in-class approximation $\hat{f}^*$, and the distance between $\hat{f}^*$ and the true model f^*. The former decreases at a rate characterized by $1/\gamma_m$ due to properties of our constrained regression problem. The latter is upper bounded by B. Therefore,

$$\mathbb{E}_{x\sim D_{\mathcal{X}}} \mathbb{E}_{a\sim \text{Unif}(\mathcal{A})} [(\hat{f}_{m+1}(x,a) - \hat{f}^*(x,a))^2] \lesssim \frac{1}{\epsilon^\rho \gamma_m}$$

$$\mathbb{E}_{x\sim D_{\mathcal{X}}} \mathbb{E}_{a\sim p_m(\cdot|x)} [(\hat{f}_{m+1}(x,a) - f^*(x,a))^2] \lesssim B + \frac{1}{\gamma_m}.$$

- In Lemma 8, we extend these results to bound on the approximation error for any policy π,

$$\left| \mathbb{E}_{x\sim \mathcal{D}_{\mathcal{X}}} [\hat{f}_{m+1}(x,\pi(x)) - f^*(x,\pi(x))] \right| \lesssim \sqrt{V(p_m,\pi)} \left(\sqrt{B} + \frac{\sqrt{K}}{\gamma_m} \right).$$

- Lemmas 9 and 10 characterize the behavior of the object $V(p_m,\pi)$. In Lemma 11 we use these results to show that estimated and expected regret satisfy the following relation, which formalized the notion of "closeness" between the two:

$$\text{Reg}_{f^*}(\pi) \lesssim \text{Reg}_{\hat{f}_m}(\pi) + \frac{K}{\gamma_m} + \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + \sqrt{V(p_m,\pi)B}$$

$$\text{Reg}_{\hat{f}_m}(\pi) \lesssim \text{Reg}(\pi) + \frac{K}{\gamma_m} + \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + \sqrt{V(p_m,\pi)B}.$$

- Lemma 12 concludes that the average expected regret suffered during any point in the active phase is bounded by

$$\sum_{\pi\in\Psi} Q_m(\pi)\text{Reg}(\pi) \lesssim \frac{K}{\gamma_m} + \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}}.$$

- In subsection B.9 we put all of these results together to prove Theorem 2.

B.2 Bounds on best predictor

In this subsection we provide basic bounds on terms involving the best predictor. We start by bounding the empirical mean square error between the best predictor ($\hat{f}^*$) and the true model (f^*) under any action selection kernel, see Lemma 1. We then use this to bound the regret of the policy induced by the best predictor ($\pi_{\hat{f}^*}$), see Lemma 2. Hence indicating that this policy is a reasonable policy to try to converge to.

Lemma 1 (Bounding B). *For any action selection kernel p, we then have that:*

$$\mathbb{E}_{x\sim D_\mathcal{X}} \mathbb{E}_{a\sim p(\cdot|x)} [(\hat{f}^*(x,a) - f^*(x,a))^2] \leq B \leq Kb.$$

Proof. We get the first inequality from the definition of B:

$$\mathbb{E}_{x\sim D_\mathcal{X}} \mathbb{E}_{a\sim p(\cdot|x)} [(\hat{f}^*(x,a) - f^*(x,a))^2] \leq \max_{p'} \mathbb{E}_{x\sim D_\mathcal{X}} \mathbb{E}_{a\sim p'(\cdot|x)} [(\hat{f}^*(x,a) - f^*(x,a))^2] = B.$$

For any context $x \in \mathcal{X}$, note that:

$$\mathbb{E}_{a\sim p'(\cdot|x)} [(\hat{f}^*(x,a) - f^*(x,a))^2] \leq \sum_{a\in\mathcal{A}} (\hat{f}^*(x,a) - f^*(x,a))^2.$$

Now, taking expectations on both sides gives us the second inequality of Lemma 1:

$$B \leq \sum_{a\in\mathcal{A}} \mathbb{E}_{x\sim D_\mathcal{X}} [(\hat{f}^*(x,a) - f^*(x,a))^2] = Kb.$$

$\square$

Lemma 2 (Regret of the policy induced by the best predictor). *We have the following bound on the regret of $\pi_{\hat{f}^*}$:*

$$Reg(\pi_{\hat{f}^*}) := R(\pi_{f^*}) - R(\pi_{\hat{f}^*}) \leq 2\sqrt{B}.$$

Proof. Note that, for any policy π, we have:

$$|R_{\hat{f}^*}(\pi) - R(\pi)|^2 = |\mathbb{E}[\hat{f}^*(x,\pi(x)) - f^*(x,\pi(x))]|^2 \leq \mathbb{E}[(\hat{f}^*(x,\pi(x)) - f^*(x,\pi(x)))^2] \leq B.$$

Where the last inequality follows from Lemma 1. Hence for any policy π, we have that:

$$R(\pi_{\hat{f}^*}) \geq R_{\hat{f}^*}(\pi_{\hat{f}^*}) - \sqrt{B} \geq R_{\hat{f}^*}(\pi) - \sqrt{B} \geq R(\pi) - 2\sqrt{B}.$$

In particular, this implies that $\text{Reg}(\pi_{\hat{f}^*}) := R(\pi_{f^*}) - R(\pi_{\hat{f}^*}) \leq 2\sqrt{B}$. $\square$

B.3 Properties of the action selection kernel

In this subsection, we explore properties of the algorithm that directly follow from the definitions in Appendix A and from the form of the action kernel used in the active phase of Epsilon-FALCON. For this reason, all the properties stated here hold true for the Falcon algorithm as well. Except for Lemma 5 and the lower bound in Lemma 6, all Lemmas in this subsection have been proved for Falcon and can be found in Simchi-Levi and Xu (2020). We state and prove these Lemmas that we use for completeness and to show that they hold for Epsilon-FALCONas well. We start with Lemma 3 which shows that the expected instantanious regret is equal to the regret of the randomized policy Q_m.

Lemma 3 (Conditional expected reward). *For any epoch $m \geq 1$ and time-step $t \geq 1$ in the active phase of epoch m, we have:*

$$\underset{x_t,r_t,a_t}{\mathbb{E}} [r_t(\pi^*(x)) - r_t(a_t)|\Gamma_{t-1}] = \sum_{\pi \in \Psi} Q_m(\pi) Reg(\pi).$$

Proof. Consider any epoch $m \geq 1$ and time-step $t \geq 1$ in the active phase of epoch m, then from Equation (14) we have:

$$\underset{x_t,r_t,a_t}{\mathbb{E}} [r_t(\pi^*(x)) - r_t(a_t)|\Gamma_{t-1}]$$

$$= \underset{x \sim D_\mathcal{X}, a \sim p_m(\cdot|x)}{\mathbb{E}} [f^*(x, \pi^*)) - f^*(x, a)]$$

$$= \underset{x \sim D_\mathcal{X}}{\mathbb{E}} \left[\sum_{a \in \mathcal{A}} p_m(a|x)(f^*(x, \pi^*)) - f^*(x, a)) \right]$$

$$= \underset{x \sim D_\mathcal{X}}{\mathbb{E}} \left[\sum_{a \in \mathcal{A}} \sum_{\pi \in \Psi} \mathbb{I}(\pi(x) = a) Q_m(\pi)(f^*(x, \pi^*)) - f^*(x, a)) \right]$$

$$= \sum_{\pi \in \Psi} Q_m(\pi) \underset{x \sim D_\mathcal{X}}{\mathbb{E}} \left[(f^*(x, \pi^*)) - f^*(x, \pi(x))) \right]$$

$$= \sum_{\pi \in \Psi} Q_m(\pi) Reg(\pi).$$

$\square$

Lemma 4 states a key bound on the estimated regret of the randomized policy Q_m.

Lemma 4 (Action selection kernel has low estimated regret). *For any epoch $m \geq 1$, we have:*

$$\sum_{\pi \in \Psi} Q_m(\pi) Reg_{\hat{f}_m}(\pi) \leq \frac{K}{\gamma_m}.$$

Proof. Note that:

$$\sum_{\pi \in \Psi} Q_m(\pi) Reg_{\hat{f}_m}(\pi) = \sum_{\pi \in \Psi} Q_m(\pi) \underset{x \sim D_\mathcal{X}}{\mathbb{E}} \left[\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, \pi(x)) \right]$$

$$= \underset{x \sim D_\mathcal{X}}{\mathbb{E}} \left[\sum_{\pi \in \Psi} Q_m(\pi) \left(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, \pi(x)) \right) \right]$$

$$= \underset{x \sim D_\mathcal{X}}{\mathbb{E}} \left[\sum_{a \in \mathcal{A}} \sum_{\pi \in \Psi} \mathbb{I}(\pi(x) = a) Q_m(\pi) \left(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, a) \right) \right]$$

$$= \underset{x \sim D_\mathcal{X}}{\mathbb{E}} \left[\sum_{a \in \mathcal{A}} p_m(a|x) \left(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, a) \right) \right]$$

$$= \underset{x \sim D_\mathcal{X}}{\mathbb{E}} \left[\sum_{a \in \mathcal{A}} \frac{\left(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, a) \right)}{K + \gamma_m \left(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, a) \right)} \right] \leq \frac{K}{\gamma_m}.$$

$\square$

Lemma 5 is a direct concequence of Jensen's inequality and helps us in the derivation of Lemma 12, which bounds the true regret of the randomized policy Q_m.

Lemma 5 (An implication of inherent duality between p_m and Q_m). *For any epoch $m \geq 1$, we have:*

$$\sum_{\pi \in \Psi} Q_m(\pi) \sqrt{V(p_m, \pi)} \leq \sqrt{K}.$$

Proof. Note that:

$$\sum_{\pi \in \Psi} Q_m(\pi) \sqrt{V(p_m, \pi)} \leq \sqrt{\sum_{\pi \in \Psi} Q_m(\pi) V(p_m, \pi)} = \sqrt{\sum_{\pi \in \Psi} Q_m(\pi) \mathop{\mathbb{E}}_{x \sim D_\mathcal{X}} \left[\frac{1}{p_m(\pi(x)|x)} \right]}$$

$$= \sqrt{\mathop{\mathbb{E}}_{x \sim D_\mathcal{X}} \left[\sum_{\pi \in \Psi} Q_m(\pi) \sum_{a \in \mathcal{A}} \frac{\mathbb{I}(\pi(x) = a)}{p_m(a|x)} \right]} = \sqrt{\mathop{\mathbb{E}}_{x \sim D_\mathcal{X}} \left[\sum_{a \in \mathcal{A}} \frac{\sum_{\pi \in \Psi} \mathbb{I}(\pi(x) = a) Q_m(\pi)}{p_m(a|x)} \right]}$$

$$= \sqrt{\mathop{\mathbb{E}}_{x \sim D_\mathcal{X}} \left[\sum_{a \in \mathcal{A}} \frac{p_m(a|x)}{p_m(a|x)} \right]} = \sqrt{K}.$$

Where the first inequality is an application of Jensen's inequality, and the other equalities are straight forward. $\square$

For any policy π, Lemma 6 provides key bounds on $V(p_m, \pi)$. These bounds help us understand the average divergence between the action distribution $p_m(\cdot|x)$ and action selected by the policy $\pi(x)$.

Lemma 6 (Bounds on expected inverse probability). *For all policies $\pi \in \Psi$ and epochs $m \geq 1$, we have:*

$$\gamma_m \mathop{\mathbb{E}}_{x \sim D_\mathcal{X}} \left[\left(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, \pi(x)) \right) \right] \leq V(p_m, \pi) \leq K + \gamma_m \mathop{\mathbb{E}}_{x \sim D_\mathcal{X}} \left[\left(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, \pi(x)) \right) \right]$$

Proof. Consider any policy $\pi \in \Psi$ and epoch $m \geq 1$. For any context $x \in \mathcal{X}$ and action $a \in \mathcal{A} \setminus \{\pi_{\hat{f}_m}(x)\}$, from our choice for p_m, we get:

$$\frac{1}{p_m(a|x)} = K + \gamma_m(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, a)).$$

For the action $a = \pi_{\hat{f}_m}(x)$, we have:

$$0 = \gamma_m \left[\left(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, a) \right) \right] \leq \frac{1}{p_m(a|x)} = \frac{1}{1 - \sum_{a' \neq a} \frac{1}{K + \gamma_m \left(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, a') \right)}} \leq K$$

In particular, putting the above inequality together, we get:

$$\gamma_m \left[\left(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, \pi(x)) \right) \right] \leq \frac{1}{p_m(\pi(x)|x)} \leq K + \gamma_m \left[\left(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, \pi(x)) \right) \right].$$

The Lemma now follows by taking expectation over $x \sim D_\mathcal{X}$. $\square$

B.4 Constrained regression oracle guarantees

Lemma 7 (Guarantees on the constrained regression oracle). *Suppose Assumption 1 holds and suppose $\epsilon < 0.5$. Then there exists positive constants C_4 and C_5 such that with probability at least $1 - \delta/2$, the following holds for all epoch $m \geq 1$:*

$$\mathbb{E}_{x \sim D_{\mathcal{X}}} \mathbb{E}_{a \sim Unif(\mathcal{A})} [(\hat{f}_{m+1}(x,a) - \hat{f}^*(x,a))^2] \leq \frac{C_4 \ln^{\rho'}(\tau_m - \tau_{m-1}) \ln(m/\delta) \boldsymbol{comp}(\mathcal{F})}{(\epsilon(\tau_m - \tau_{m-1}))^\rho}.$$

$$\mathbb{E}_{x \sim D_{\mathcal{X}}} \mathbb{E}_{a \sim p_m(\cdot|x)} [(\hat{f}_{m+1}(x,a) - f^*(x,a))^2] \leq B + \frac{C_5 \ln^{\rho'}(\tau_m - \tau_{m-1}) \ln(m/\delta) \boldsymbol{comp}(\mathcal{F})}{(\tau_m - \tau_{m-1})^\rho}.$$

Proof. Let $\mathcal{F}'$ denote the set of estimators in the constraint set at the end of epoch m. Let $\delta' = \delta/(12m^2)$. Since $\hat{f}_{m+1} \in \mathcal{F}'$, we have:

$$\frac{1}{|S'_m|} \sum_{(x,a,r(a)) \in S'_m} (\hat{f}_{m+1}(x,a) - r(a))^2 - \frac{1}{|S'_m|} \min_{g \in \mathcal{F}} \sum_{(x,a,r(a)) \in S'_m} (g(x,a) - r(a))^2$$
$$\leq \frac{C_1 \ln^{\rho'}(|S'_m|) \ln(1/\delta') \boldsymbol{comp}(\mathcal{F})}{|S'_m|^\rho}.$$

The above inequality bounds the empirical excess risk for $\hat{f}_{m+1}$ with respect to the empirical data S'_m and the set of estimators in $\mathcal{F}$. Now note that S'_m is generated by sampling arms uniformly at random, and note that $\mathcal{F}$ is a convex set. Hence from Assumption 1, we get that for some universal constant $L_1 = 2 \max\{C, C_1\}$ [8], with probability at least $1 - \delta'$, we have:

$$\mathbb{E}_{(x,r) \sim D} \mathbb{E}_{a \sim Unif(\mathcal{A})} [(\hat{f}_{m+1}(x,a) - r(a))^2] - \mathbb{E}_{(x,r) \sim D} \mathbb{E}_{a \sim Unif(\mathcal{A})} [(\hat{f}^*(x,a) - r(a))^2]$$
$$\leq \frac{L_1 \ln^{\rho'}(|S'_m|) \ln(1/\delta') \boldsymbol{comp}(\mathcal{F})}{|S'_m|^\rho}. \tag{25}$$

Since $\mathcal{F}$ is a convex class of functions, Lemma 5.1 in [Koltchinskii (2011)](#) gives us that:

$$\mathbb{E}_{x \sim D_{\mathcal{X}}} \mathbb{E}_{a \sim Unif(\mathcal{A})} [(\hat{f}_{m+1}(x,a) - \hat{f}^*(x,a))^2]$$
$$\leq 2 \mathbb{E}_{(x,r) \sim D} \mathbb{E}_{a \sim Unif(\mathcal{A})} [(\hat{f}_{m+1}(x,a) - r(a))^2 - (\hat{f}^*(x,a) - r(a))^2]. \tag{26}$$

Therefore, putting everything together (see eq. (25) and eq. (26)), with probability at least $1 - \delta'$ we have:

$$\mathbb{E}_{x \sim D_{\mathcal{X}}} \mathbb{E}_{a \sim Unif(\mathcal{A})} [(\hat{f}_{m+1}(x,a) - \hat{f}^*(x,a))^2] \leq \frac{2 L_1 \ln^{\rho'}(|S'_m|) \ln(1/\delta') \boldsymbol{comp}(\mathcal{F})}{|S'_m|^\rho}.$$

Hence the first inequality in Lemma 7 follows from noting that $|S'_m| = \lceil \epsilon(\tau_m - \tau_{m-1}) \rceil$, and choosing $C_4 = 2L_1(2 + \ln(12))$.

---

[8] Where C is the constant from Assumption 1.

Note that $\mathcal{F}$ is convex, $\hat{f}^*$ has no population excess risk with respect to the distribution generated from picking arms uniformly at random among estimators in $\mathcal{F}$, and note that S'_m is generated by sampling arms uniformly at random. Hence from Assumption 1, with probability at least $1 - \delta'$, we get that:

$$\frac{1}{|S'_m|} \sum_{(x,a,r(a)) \in S'_m} (\hat{f}^*(x,a) - r(a))^2 - \frac{1}{|S'_m|} \min_{g \in \mathcal{F}} \sum_{(x,a,r(a)) \in S'_m} (g(x,a) - r(a))^2$$
$$\leq \frac{(3C/2) \ln^{\rho'}(|S'_m|) \ln(1/\delta') \mathbf{comp}(\mathcal{F})}{|S'_m|^\rho}.$$

Therefore by choosing $C_1 \geq 3C/2$, with probability at least $1 - \delta'$, we get that $\hat{f}^* \in \mathcal{F}'$. Now recall that:

$$\hat{f}_{m+1} \in \arg \min_{f \in \mathcal{F}'} \frac{1}{|S_m|} \sum_{(x,a,r(a)) \in S_m} (f(x,a) - r(a))^2$$

That is, $\hat{f}_{m+1}$ has no empirical excess risk with respect to the empirical data S'_m among estimators in $\mathcal{F}'$. Also note that $\mathcal{F}'$ is convex subset of $\mathcal{F}$, and S_m is generated by sampling arms according to the action selection kernel p_m. Hence from Assumption 1, with probability at least $1 - \delta'$, we get that:

$$\mathbb{E}_{(x,r) \sim D} \mathbb{E}_{a \sim p_m(\cdot|x)} [(\hat{f}_{m+1}(x,a) - r(a))^2] - \min_{f \in \mathcal{F}'} \mathbb{E}_{(x,r) \sim D} \mathbb{E}_{a \sim p_m(\cdot|x)} [(f(x,a) - r(a))^2]$$
$$\leq \frac{2C \ln^{\rho'}(|S_m|) \ln(1/\delta') \mathbf{comp}(\mathcal{F})}{|S_m|^\rho}. \tag{27}$$

Hence by taking union bound so that eq. (27) holds and $\hat{f}^* \in \mathcal{F}'$, with probability at least $1 - 2\delta'$, we have:

$$\mathbb{E}_{(x,r) \sim D} \mathbb{E}_{a \sim p_m(\cdot|x)} [(\hat{f}_{m+1}(x,a) - r(a))^2] - \mathbb{E}_{(x,r) \sim D} \mathbb{E}_{a \sim p_m(\cdot|x)} [(\hat{f}^*(x,a) - r(a))^2]$$
$$\leq \frac{2C \ln^{\rho'}(|S_m|) \ln(1/\delta') \mathbf{comp}(\mathcal{F})}{|S_m|^\rho}.$$

Recall that B is the worst case excess risk for $\hat{f}^*$ under any kernel. Therefore, with probability at least $1 - 2\delta'$, we have:

$$\mathbb{E}_{x \sim D_\mathcal{X}} \mathbb{E}_{a \sim p_m(\cdot|x)} [(\hat{f}_{m+1}(x,a) - f^*(x,a))^2]$$
$$= \mathbb{E}_{(x,r) \sim D} \mathbb{E}_{a \sim p_m(\cdot|x)} [(\hat{f}_{m+1}(x,a) - r(a))^2 - (f^*(x,a) - r(a))^2]$$
$$\leq \mathbb{E}_{(x,r) \sim D} \mathbb{E}_{a \sim p_m(\cdot|x)} [(\hat{f}^*(x,a) - r(a))^2 - (f^*(x,a) - r(a))^2] + \frac{2C \ln^{\rho'}(|S_m|) \ln(1/\delta') \mathbf{comp}(\mathcal{F})}{|S_m|^\rho}$$
$$= \mathbb{E}_{x \sim D_\mathcal{X}} \mathbb{E}_{a \sim p_m(\cdot|x)} [(\hat{f}^*(x,a) - f^*(x,a))^2] + \frac{2C \ln^{\rho'}(|S_m|) \ln(1/\delta') \mathbf{comp}(\mathcal{F})}{|S_m|^\rho}$$
$$\leq B + \frac{2C \ln^{\rho'}(|S_m|) \ln(1/\delta') \mathbf{comp}(\mathcal{F})}{|S_m|^\rho}.$$

23

For any epoch $m \geq 1$, note that $\tau_m - \tau_{m-1} \geq \tau_1 \geq 4$. Therefore since $\epsilon < 0.5$, we get that:

$$|S_m| = \tau_m - \tau_{m-1} - \lceil \epsilon(\tau_m - \tau_{m-1}) \rceil \geq \frac{1}{4}(\tau_m - \tau_{m-1}).$$

Hence the second inequality in Lemma 7 follows from choosing an appropriate value for $C_5 = 2C \times 4^\rho \times (2 + \ln(12))$. Taking union bound, we finally note that both inequalities in lemma 7 hold for all epochs with probability at least:

$$1 - \sum_{m=1}^{\infty} 3\frac{\delta}{12m^2} \geq 1 - \frac{\delta(\pi^2/6)}{4} \geq 1 - \delta/2.$$

$\square$

Additional notation   For compactness of notation, define the following event:

$$\mathcal{W} := \left\{ \forall m \geq 1, \quad \underset{x \sim D_\mathcal{X}}{\mathbb{E}} \underset{a \sim \mathrm{Unif}(\mathcal{A})}{\mathbb{E}} [(\hat{f}_{m+1}(x,a) - \hat{f}^*(x,a))^2] \leq \frac{C_4 \ln^{\rho'}(\tau_m - \tau_{m-1}) \ln(m/\delta)\mathbf{comp}(\mathcal{F})}{(\epsilon(\tau_m - \tau_{m-1}))^\rho}, \right.$$

$$\left. \underset{x \sim D_\mathcal{X}}{\mathbb{E}} \underset{a \sim p_m(\cdot|x)}{\mathbb{E}} [(\hat{f}_{m+1}(x,a) - f^*(x,a))^2] \leq B + \frac{C_5 \ln^{\rho'}(\tau_m - \tau_{m-1}) \ln(m/\delta)\mathbf{comp}(\mathcal{F})}{(\tau_m - \tau_{m-1})^\rho} \right\},$$

$$(28)$$

for two constants C_4 and C_5 that were defined in Lemma 7.

B.5   Bounding prediction error of implicit rewards

For any policy, Lemma 8 bounds the prediction error of implicit reward estimate of the policy at every epoch. This Lemma and its proof are similar to Lemma 7 in Simchi-Levi and Xu (2020).

Lemma 8 (Accuracy of implicit policy estimate). *Suppose $C_3 \leq 1/(4C_5)$ and suppose the event $\mathcal{W}$ from* (28) *holds. Then, for all policies π and epoch $m \geq 1$, we have:*

$$|R_{\hat{f}_{m+1}}(\pi) - R(\pi)| \leq \sqrt{V(p_m, \pi)}\sqrt{B} + \frac{\sqrt{V(p_m, \pi)}\sqrt{K}}{2\gamma_{m+1}}$$

Proof. For any policy π and epoch $m \geq 1$, note that:

$$|R_{\hat{f}_{m+1}}(\pi) - R(\pi)|^2$$
$$\leq \left(\mathbb{E}_{x \sim D_\mathcal{X}} \left[\left| \hat{f}_{m+1}(x, \pi(x)) - f^*(x, \pi(x)) \right| \right] \right)^2$$
$$= \left(\mathbb{E}_{x \sim D_\mathcal{X}} \left[\sqrt{\frac{1}{p_m(\pi(x)|x)} p_m(\pi(x)|x) \left(\hat{f}_{m+1}(x, \pi(x)) - f^*(x, \pi(x)) \right)^2} \right] \right)^2$$
$$\leq \left(\mathbb{E}_{x \sim D_\mathcal{X}} \left[\sqrt{\frac{1}{p_m(\pi(x)|x)} \mathbb{E}_{a \sim p_m(\cdot|x)} \left[\hat{f}_{m+1}(x, a) - f^*(x, a) \right]^2} \right] \right)^2$$
$$\leq \mathbb{E}_{x \sim D_\mathcal{X}} \left[\frac{1}{p_m(\pi(x)|x)} \right] \mathbb{E}_{x \sim D_\mathcal{X}} \mathbb{E}_{a \sim p_m(\cdot|x)} \left[\left(\hat{f}_{m+1}(x, a) - f^*(x, a) \right)^2 \right]$$
$$\leq V(p_m, \pi) \left(B + \frac{C_5 \ln^{\rho'}(\tau_m - \tau_{m-1}) \ln(1/\delta') \mathbf{comp}(\mathcal{F})}{(\tau_m - \tau_{m-1})^\rho} \right).$$

The first inequality follows from Jensen's inequality, the second inequality is straight forward, the third inequality follows from Cauchy-Schwarz inequality, and the last inequality follows from assuming that $\mathcal{W}$ from (28) holds. Now from the sub-additive property of square-root, we get:

$$|R_{\hat{f}_{m+1}}(\pi) - R(\pi)| \leq \sqrt{V(p_m, \pi)} \left(\sqrt{B} + \sqrt{\frac{C_5 \ln^{\rho'}(\tau_m - \tau_{m-1}) \ln(m/\delta) \mathbf{comp}(\mathcal{F})}{(\tau_m - \tau_{m-1})^\rho}} \right)$$
$$\leq \sqrt{V(p_m, \pi)} \sqrt{B} + \frac{\sqrt{V(p_m, \pi)} \sqrt{K}}{2\gamma_{m+1}}.$$

Where the last inequality follows from the choice of γ_{m+1} and from assuming that $C_3 \leq 1/(4C_5)$. $\qquad\square$

B.6   Bounding decisional divergence

At any epoch m, Lemma 9 bounds the decisional divergence between the active policy at that epoch (Q_m) and the policy induced by the best estimator ($\pi_{\hat{f}^*}$). This implies that even as the active policy is less explorative, Q_m is not very far from $\pi_{\hat{f}^*}$ and hence eventually converges to it.

Lemma 9 (Action selection kernels are always close to target policy). *Suppose the event $\mathcal{W}$ from (28) holds. Then there exists a positive constant C_6 such that, for any epoch $m \geq 1$, we have:*

$$V(p_m, \pi_{\hat{f}^*}) \leq \frac{C_6 K}{\sqrt{\epsilon^\rho}}$$

Proof. Since the action selection kernel $p_1(\cdot|x)$ picks arms uniformly at random for all $x \in \mathcal{X}$, we have that $V(p_1, \pi_{\hat{f}^*}) = K$. Hence, by choosing $C_6 \geq 1$, we get that $V(p_1, \pi_{\hat{f}^*}) \leq C_6 K/\sqrt{\epsilon^\rho}$.

Now consider any epoch $m \geq 2$. Note that from the definition of $\pi_{\hat{f}^*}$, for any context $x \in \mathcal{X}$ we get:
$$\hat{f}^*(x, \pi_{\hat{f}^*}(x)) = \max_{a \in \mathcal{A}} \hat{f}^*(x, a) \geq \hat{f}^*(x, \pi_{\hat{f}_m}(x)).$$
Hence from the above inequality, for any context $x \in \mathcal{X}$ we get:
$$\begin{aligned}
&\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, \pi_{\hat{f}^*}(x)) \\
&= \left(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}^*(x, \pi_{\hat{f}_m}(x))\right) + \left(\hat{f}^*(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, \pi_{\hat{f}^*}(x))\right) \\
&\leq \left(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}^*(x, \pi_{\hat{f}_m}(x))\right) + \left(\hat{f}^*(x, \pi_{\hat{f}^*}(x)) - \hat{f}_m(x, \pi_{\hat{f}^*}(x))\right) \\
&\leq \left|\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}^*(x, \pi_{\hat{f}_m}(x))\right| + \left|\hat{f}^*(x, \pi_{\hat{f}^*}(x)) - \hat{f}_m(x, \pi_{\hat{f}^*}(x))\right| \\
&\leq 2 \max_{a \in \mathcal{A}} \left|\hat{f}^*(x, a) - \hat{f}_m(x, a)\right|.
\end{aligned}$$
Now from Lemma 6, the above inequality, and Jensen's inequality, we get:
$$\begin{aligned}
V(p_m, \pi_{\hat{f}^*}) &\leq \mathbb{E}_{x \sim \mathcal{X}} \left[K + \gamma_m \left(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, \pi_{\hat{f}^*}(x))\right) \right] \\
&\leq K + 2\gamma_m \mathbb{E}_{x \sim \mathcal{X}} \left[\max_{a \in \mathcal{A}} \left|\hat{f}^*(x, a) - \hat{f}_m(x, a)\right| \right] \\
&\leq K + 2\gamma_m \sqrt{\mathbb{E}_{x \sim \mathcal{X}} \left[\max_{a \in \mathcal{A}} \left|\hat{f}^*(x, a) - \hat{f}_m(x, a)\right|^2 \right]} \\
&\leq K + 2\gamma_m \sqrt{\sum_{a \in \mathcal{A}} \mathbb{E}_{x \sim \mathcal{X}} \left[\left(\hat{f}_m(x, a) - \hat{f}^*(x, a)\right)^2 \right]}.
\end{aligned}$$
Now let $C_6 = 1 + 2\sqrt{C_3 C_4}$. From the above inequality, we further get:
$$\begin{aligned}
V(p_m, \pi_{\hat{f}^*}) &\leq K + 2\gamma_m \sqrt{\sum_{a \in \mathcal{A}} \mathbb{E}_{x \sim \mathcal{X}} \left[\left(\hat{f}_m(x, a) - \hat{f}^*(x, a)\right)^2 \right]} \\
&\leq K + 2\gamma_m \sqrt{K \frac{C_4 \ln^{\rho'}(\tau_{m-1} - \tau_{m-2}) \ln((m-1)/\delta) \mathbf{comp}(\mathcal{F})}{(\epsilon(\tau_{m-1} - \tau_{m-2}))^\rho}} \\
&= K + 2K\sqrt{\frac{C_3 C_4}{\epsilon^\rho}} \leq \frac{C_6 K}{\sqrt{\epsilon^\rho}}.
\end{aligned}$$
Where the second inequality follows from the assumption that $\mathcal{W}$ holds. And the last inequality follows from our choice of C_6. □

Lemma 10 shows that for any policy π and epoch m, if the decisional divergence between Q_m and π was large, then the decisional divergence between Q_{m+1} and π must also be large. Hence the Lemma shows that the active phase of Epsilon-FALCON stops exploring in a stable manner.

Lemma 10 (Do not pick up policies that you drop)**.** *Suppose the event $\mathcal{W}$ defined in (28) holds, and $\delta \leq 0.5$. Then there exists a positive constant C_7 such that, for all policies π and epochs m, we have:*
$$V(p_m, \pi) \leq \frac{C_7 K}{\sqrt{\epsilon^\rho}} + V(p_{m+1}, \pi).$$

Proof. Consider any policy π. Since the action selection kernel $p_1(\cdot|x)$ picks arms uniformly at random for all x, we have that $V(p_1, \pi) = K$. Hence, by choosing $C_7 \geq 1$, we get that $V(p_1, , \pi) \leq \frac{C_7 K}{\sqrt{\epsilon^\rho}} + V(p_2, \pi)$. Now consider any epoch $m \geq 2$. For any context $x \in \mathcal{X}$, we get:

$$\hat{f}_{m+1}(x, \pi_{\hat{f}_{m+1}}(x)) = \max_{a \in \mathcal{A}} \hat{f}_{m+1}(x, a) \geq \begin{cases} \hat{f}_{m+1}(x, \pi_{\hat{f}_m}(x)) \\ \hat{f}_{m+1}(x, \pi(x)). \end{cases}$$

From Lemma 6, the fact that $\gamma_{m+1} \geq \gamma_m$, and the above inequality, we get:

$$
\begin{aligned}
& V(p_m, \pi) - K - V(p_{m+1}, \pi) \\
& \leq \gamma_m \mathop{\mathbb{E}}_{x \sim D_\mathcal{X}} [\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, \pi(x))] - \gamma_{m+1} \mathop{\mathbb{E}}_{x \sim D_\mathcal{X}} [\hat{f}_{m+1}(x, \pi_{\hat{f}_{m+1}}(x)) - \hat{f}_{m+1}(x, \pi(x))] \\
& \leq \gamma_m \mathop{\mathbb{E}}_{x \sim D_\mathcal{X}} [(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_{m+1}(x, \pi_{\hat{f}_{m+1}}(x))) + (\hat{f}_{m+1}(x, \pi(x)) - \hat{f}_m(x, \pi(x)))] \\
& \leq \gamma_m \mathop{\mathbb{E}}_{x \sim D_\mathcal{X}} [(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_{m+1}(x, \pi_{\hat{f}_m}(x))) + (\hat{f}_{m+1}(x, \pi(x)) - \hat{f}_m(x, \pi(x)))] \\
& \leq 2\gamma_m \mathop{\mathbb{E}}_{x \sim D_\mathcal{X}} \left[\max_a |\hat{f}_{m+1}(x, a) - \hat{f}_m(x, a)| \right].
\end{aligned}
$$

$$(29)$$

Also note that from Jensen's inequality, we get:

$$
\begin{aligned}
& \mathop{\mathbb{E}}_{x \sim D_\mathcal{X}} \left[\max_a |\hat{f}_{m+1}(x, a) - \hat{f}_m(x, a)| \right] \\
& \leq \sqrt{\mathop{\mathbb{E}}_{x \sim D_\mathcal{X}} \left[\max_a |\hat{f}_{m+1}(x, a) - \hat{f}_m(x, a)|^2 \right]} \\
& \leq \sqrt{\sum_a \mathop{\mathbb{E}}_{x \sim D_\mathcal{X}} \left[(\hat{f}_{m+1}(x, a) - \hat{f}_m(x, a))^2 \right]} \\
& = \sqrt{K \mathop{\mathbb{E}}_{x \sim D_\mathcal{X}} \mathop{\mathbb{E}}_{a \sim \text{Unif}(\mathcal{A})} \left[(\hat{f}_{m+1}(x, a) - \hat{f}_m(x, a))^2 \right]} \\
& \leq \sqrt{2K \mathop{\mathbb{E}}_{x \sim D_\mathcal{X}} \mathop{\mathbb{E}}_{a \sim \text{Unif}(\mathcal{A})} \left[(\hat{f}_{m+1}(x, a) - \hat{f}^*(x, a))^2 + (\hat{f}^*(x, a) - \hat{f}_m(x, a))^2 \right]} \\
& \leq \sqrt{\frac{4KC_4 \ln^{\rho'} (\tau_m - \tau_{m-1}) \ln(m/\delta) \mathbf{comp}(\mathcal{F})}{(\epsilon(\tau_{m-1} - \tau_{m-2}))^\rho}}.
\end{aligned}
$$

$$(30)$$

The second last inequality follows from the identity that for any two real numbers u, v, $(u + v)^2 \leq 2(u^2 + v^2)$. The last inequality follows from the assumption that $\mathcal{W}$ holds, the fact that $m \geq m - 1$, and the fact that epoch lengths are non-decreasing (i.e. $\tau_m - \tau_{m-1} \geq$

$\tau_{m-1} - \tau_{m-1})$. Now, by combining Equation (29) and Equation (30), we get:

$$V(p_m, \pi) \le V(p_{m+1}, \pi) + K + 4\gamma_m \sqrt{\frac{KC_4 \ln^{\rho'}(\tau_m - \tau_{m-1}) \ln(m/\delta)\mathbf{comp}(\mathcal{F})}{(\epsilon(\tau_{m-1} - \tau_{m-2}))^\rho}}$$

$$= V(p_{m+1}, \pi) + K + 4K \sqrt{\frac{C_3 C_4}{\epsilon^\rho} \frac{\ln^{\rho'}(\tau_m - \tau_{m-1})}{\ln^{\rho'}(\tau_{m-1} - \tau_{m-2})} \frac{\ln(m/\delta)}{\ln((m-1)/\delta)}}$$

$$\le V(p_{m+1}, \pi) + \frac{C_7 K}{\sqrt{\epsilon^\rho}}.$$

Where the last inequality follows from choosing $C_7 = 1 + 4\sqrt{2^{1+\rho'}C_3 C_4}$, and from the fact that for $m \ge 2$ and $\delta \le 0.5$ we have: $\frac{\ln(m/\delta)}{\ln((m-1)/\delta)} \le 2$, and $\frac{\ln^{\rho'}(\tau_m - \tau_{m-1})}{\ln^{\rho'}(\tau_{m-1} - \tau_{m-2})} \le \frac{\ln^{\rho'}(\tau_{m-1})}{\ln^{\rho'}(\tau_{m-1}/2)} \le 2^{\rho'}$. $\qquad\square$

B.7   Bounding prediction error of implicit regret

For any policy, Lemma 11 bounds the prediction error of implicit regret estimate of the policy at every epoch. This Lemma and its proof are similar to Lemma 8 in Simchi-Levi and Xu (2020).

Lemma 11 (Bounds on implicit estimates of policy regret). *Suppose the event $\mathcal{W}$ defined in (28) holds, and $\delta \le 0.5$. Then there exists positive constants C_0, C_8, C_9 such that, for all policies π and epochs m, we have:*

$$Reg(\pi) \le 2Reg_{\hat{f}_m}(\pi) + \frac{C_0 K}{\gamma_m} + C_8 \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + C_9 \sqrt{V(p_m, \pi)B}$$

$$Reg_{\hat{f}_m}(\pi) \le 2Reg(\pi) + \frac{C_0 K}{\gamma_m} + C_8 \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + C_9 \sqrt{V(p_m, \pi)B}$$

Proof. We will prove this by induction. Let C_0 be a positive constant such that $C_0 \ge 1 \ge \gamma_1/K$. The base case then follows from the fact that for all policies π, we have:

$$\mathrm{Reg}(\pi) \le 1 \le C_0 K/\gamma_1$$
$$\mathrm{Reg}_{\hat{f}_1}(\pi) \le 1 \le C_0 K/\gamma_1.$$

For the inductive step, fix some $m \ge 1$. Assume for all policies π, we have:

$$\mathrm{Reg}(\pi) \le 2\mathrm{Reg}_{\hat{f}_m}(\pi) + \frac{C_0 K}{\gamma_m} + C_8 \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + C_9 \sqrt{V(p_m, \pi)B}$$

$$\mathrm{Reg}_{\hat{f}_m}(\pi) \le 2\mathrm{Reg}(\pi) + \frac{C_0 K}{\gamma_m} + C_8 \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + C_9 \sqrt{V(p_m, \pi)B} \qquad (31)$$

28

Note that:

$$\text{Reg}(\pi) - \text{Reg}_{\hat{f}_{m+1}}(\pi)$$
$$= \Big(R(\pi_{f^*}) - R(\pi) \Big) - \Big(R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}}) - R_{\hat{f}_{m+1}}(\pi) \Big)$$
$$\leq \Big(R(\pi_{\hat{f}^*}) - R(\pi) \Big) - \Big(R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}}) - R_{\hat{f}_{m+1}}(\pi) \Big) + 2\sqrt{B}$$
$$\leq \Big(R(\pi_{\hat{f}^*}) - R(\pi) \Big) - \Big(R_{\hat{f}_{m+1}}(\pi_{\hat{f}^*}) - R_{\hat{f}_{m+1}}(\pi) \Big) + 2\sqrt{B}$$
$$\leq |R(\pi_{\hat{f}^*}) - R_{\hat{f}_{m+1}}(\pi_{\hat{f}^*})| + |R(\pi) - R_{\hat{f}_{m+1}}(\pi))| + 2\sqrt{B}.$$

Where the first inequality follows from Lemma 2, and the second inequality follows from the definition of $\pi_{\hat{f}_{m+1}}$ which gives us that $R_{\hat{f}_{m+1}}(\pi_{\hat{f}^*}) \leq R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}})$. Now, further simplifying the above inequality we get:

$$\text{Reg}(\pi) - \text{Reg}_{\hat{f}_{m+1}}(\pi)$$
$$\leq |R(\pi_{\hat{f}^*}) - R_{\hat{f}_{m+1}}(\pi_{\hat{f}^*})| + |R(\pi) - R_{\hat{f}_{m+1}}(\pi))| + 2\sqrt{B}$$
$$\leq \sqrt{V(p_m, \pi_{\hat{f}^*})}\sqrt{B} + \frac{\sqrt{V(p_m, \pi_{\hat{f}^*})}\sqrt{K}}{2\gamma_{m+1}} + \sqrt{V(p_m, \pi)}\sqrt{B} + \frac{\sqrt{V(p_m, \pi)}\sqrt{K}}{2\gamma_{m+1}} + 2\sqrt{B}$$
$$\leq \frac{5K}{8\gamma_{m+1}} + \frac{V(p_m, \pi_{\hat{f}^*})}{5\gamma_{m+1}} + \frac{V(p_m, \pi)}{5\gamma_{m+1}} + \sqrt{B}\Big(\sqrt{V(p_m, \pi_{\hat{f}^*})} + \sqrt{V(p_m, \pi)} + 2 \Big)$$
$$\leq \frac{5K}{8\gamma_{m+1}} + \frac{V(p_m, \pi_{\hat{f}^*})}{5\gamma_{m+1}} + \frac{V(p_m, \pi)}{5\gamma_{m+1}} + \Big(\sqrt{C_6} + \sqrt{C_7} \Big)\sqrt{\frac{BK}{\sqrt{\epsilon^\rho}}} + \sqrt{B}\sqrt{V(p_{m+1}, \pi)} + 2\sqrt{B}. \tag{32}$$

Where the second inequality follow from Lemma 8, the third inequality is an application of Cauchy-Schwarz inequality, and the last inequality follows from Lemmas 9 and 10. Now note that:

$$\frac{V(p_m, \pi_{\hat{f}^*})}{5\gamma_{m+1}} \leq \frac{K + \gamma_m \text{Reg}_{\hat{f}_m}(\pi_{\hat{f}^*})}{5\gamma_{m+1}}$$
$$\leq \frac{K + \gamma_m \Big(2\text{Reg}(\pi_{\hat{f}^*}) + \frac{C_0 K}{\gamma_m} + C_8\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + C_9\sqrt{V(p_m, \pi_{\hat{f}^*})B} \Big)}{5\gamma_{m+1}}$$
$$\leq \frac{K(1 + C_0)}{5\gamma_{m+1}} + \frac{2\text{Reg}(\pi_{\hat{f}^*})}{5} + \frac{C_8}{5}\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + \frac{C_9}{5}\sqrt{V(p_m, \pi_{\hat{f}^*})B}$$
$$\leq \frac{K(1 + C_0)}{5\gamma_{m+1}} + \frac{4\sqrt{B}}{5} + \frac{C_8 + C_9\sqrt{C_6}}{5}\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}}. \tag{33}$$

Where the first inequality follows from Lemma 6, the second inequality follows from Equa-

tion (31), and the last inequality follows from Lemmas 2 and 9. Similarly note that:

$$
\begin{aligned}
\frac{V(p_m, \pi)}{5\gamma_{m+1}} &\leq \frac{K + \gamma_m \text{Reg}_{\hat{f}_m}(\pi)}{5\gamma_{m+1}} \\
&\leq \frac{K + \gamma_m \left(2\text{Reg}(\pi) + \frac{C_0 K}{\gamma_m} + C_8 \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + C_9 \sqrt{V(p_m, \pi)B} \right)}{5\gamma_{m+1}} \\
&\leq \frac{K(1 + C_0)}{5\gamma_{m+1}} + \frac{2\text{Reg}(\pi)}{5} + \frac{C_8}{5} \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + \frac{C_9}{5} \sqrt{V(p_m, \pi)B} \\
&\leq \frac{K(1 + C_0)}{5\gamma_{m+1}} + \frac{2\text{Reg}(\pi)}{5} + \frac{C_8 + C_9\sqrt{C_7}}{5} \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + \frac{C_9}{5} \sqrt{V(p_{m+1}, \pi)B}.
\end{aligned}
\tag{34}
$$

Where the first inequality follows from Lemma 6, the second inequality follows from Equation (31), and the last inequality follows from Lemma 10. Now from combining Equation (32), Equation (33), and Equation (34), we get:

$$
\begin{aligned}
\text{Reg}(\pi) - \text{Reg}_{\hat{f}_{m+1}}(\pi) \leq &\frac{5K}{8\gamma_{m+1}} + \frac{2K(1 + C_0)}{5\gamma_{m+1}} + \frac{2\text{Reg}(\pi)}{5} + \frac{14}{5}\sqrt{B} \\
&+ \frac{2C_8 + (C_9 + 5)(\sqrt{C_6} + \sqrt{C_7})}{5} \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + \frac{C_9 + 5}{5} \sqrt{V(p_{m+1}, \pi)B}
\end{aligned}
$$

Which implies:

$$
\begin{aligned}
\text{Reg}(\pi) \leq &\frac{5}{3}\text{Reg}_{\hat{f}_{m+1}}(\pi) + \frac{K(2C_0 + 5.125)}{3\gamma_{m+1}} + \frac{14}{3}\sqrt{B} \\
&+ \frac{2C_8 + (C_9 + 5)(\sqrt{C_6} + \sqrt{C_7})}{3} \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + \frac{C_9 + 5}{3} \sqrt{V(p_{m+1}, \pi)B}
\end{aligned}
$$

Now choosing constants so that $C_0 \geq 5.125$, $C_9 \geq 2.5$, and $C_8 \geq (C_9 + 5)(\sqrt{C_6} + \sqrt{C_7})$. The above inequality then gives us:

$$
\text{Reg}(\pi) \leq 2\text{Reg}_{\hat{f}_{m+1}}(\pi) + \frac{C_0 K}{\gamma_{m+1}} + C_8 \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + C_9 \sqrt{V(p_{m+1}, \pi)B}.
\tag{35}
$$

Hence from our induction hypothesis (Equation (31)), we get Equation (35), which provides the required upper bound on $\text{Reg}(\pi)$ in terms of $\text{Reg}_{\hat{f}_{m+1}}(\pi)$. To complete the inductive argument, we need to show the corresponding upper bound on $\text{Reg}_{\hat{f}_{m+1}}(\pi)$. Similar to

30

Equation (32), we get:

$$
\begin{aligned}
&\operatorname{Reg}_{\hat{f}_{m+1}}(\pi) - \operatorname{Reg}(\pi) \\
=& \left(R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}}) - R_{\hat{f}_{m+1}}(\pi) \right) - \left(R(\pi_{f^*}) - R(\pi) \right) \\
\leq& \left(R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}}) - R_{\hat{f}_{m+1}}(\pi) \right) - \left(R(\pi_{\hat{f}_{m+1}}) - R(\pi) \right) \\
\leq& |R(\pi_{\hat{f}_{m+1}}) - R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}})| + |R(\pi) - R_{\hat{f}_{m+1}}(\pi))| \\
\leq& \sqrt{V(p_m, \pi_{\hat{f}_{m+1}})}\sqrt{B} + \frac{\sqrt{V(p_m, \pi_{\hat{f}_{m+1}})}\sqrt{K}}{2\gamma_{m+1}} + \sqrt{V(p_m, \pi)}\sqrt{B} + \frac{\sqrt{V(p_m, \pi)}\sqrt{K}}{2\gamma_{m+1}} \\
\leq& \frac{5K}{8\gamma_{m+1}} + \frac{V(p_m, \pi_{\hat{f}_{m+1}})}{5\gamma_{m+1}} + \frac{V(p_m, \pi)}{5\gamma_{m+1}} + \sqrt{B}\left(\sqrt{V(p_m, \pi_{\hat{f}_{m+1}})} + \sqrt{V(p_m, \pi)} \right) \\
\leq& \frac{5K}{8\gamma_{m+1}} + \frac{V(p_m, \pi_{\hat{f}_{m+1}})}{5\gamma_{m+1}} + \frac{V(p_m, \pi)}{5\gamma_{m+1}} + 2\sqrt{\frac{C_7 BK}{\sqrt{\epsilon^\rho}}} + \sqrt{B}\left(\sqrt{V(p_{m+1}, \pi_{\hat{f}_{m+1}})} + \sqrt{V(p_{m+1}, \pi)} \right).
\end{aligned}
\tag{36}
$$

Where the first inequality follows from the definition of π_{f^*}, the second inequality is straight forward, the third inequality follows from Lemma 8, the forth inequality is an application of Cauchy-Schwarz inequality, and the last inequality follows from Lemma 10. Similar to Equation (33), we get:

$$
\begin{aligned}
\frac{V(p_m, \pi_{\hat{f}_{m+1}})}{5\gamma_{m+1}} &\leq \frac{K + \gamma_m \operatorname{Reg}_{\hat{f}_m}(\pi_{\hat{f}_{m+1}})}{5\gamma_{m+1}} \\
&\leq \frac{K + \gamma_m \left(2\operatorname{Reg}(\pi_{\hat{f}_{m+1}}) + \frac{C_0 K}{\gamma_m} + C_8 \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + C_9 \sqrt{V(p_m, \pi_{\hat{f}_{m+1}})B} \right)}{5\gamma_{m+1}} \\
&\leq \frac{K(1 + C_0)}{5\gamma_{m+1}} + \frac{2\operatorname{Reg}(\pi_{\hat{f}_{m+1}})}{5} + \frac{C_8}{5}\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + \frac{C_9}{5}\sqrt{V(p_m, \pi_{\hat{f}_{m+1}})B} \\
&\leq \frac{K(1 + C_0)}{5\gamma_{m+1}} + \frac{2\operatorname{Reg}(\pi_{\hat{f}_{m+1}})}{5} + \frac{C_8 + C_9\sqrt{C_7}}{5}\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + \frac{C_9}{5}\sqrt{V(p_{m+1}, \pi_{\hat{f}_{m+1}})B} \\
&\leq \frac{K(1 + 3C_0)}{5\gamma_{m+1}} + \frac{3C_8 + C_9\sqrt{C_7}}{5}\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + \frac{3C_9}{5}\sqrt{V(p_{m+1}, \pi_{\hat{f}_{m+1}})B}.
\end{aligned}
\tag{37}
$$

Where the first inequality follows from Lemma 6, the second inequality follows from Equation (31), the forth inequality follows from Lemma 10, and the last inequality follows from Equation (35). Also note that:

$$
V(p_{m+1}, \pi_{\hat{f}_{m+1}}) \leq K + \gamma_{m+1}\operatorname{Reg}_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}}) = K.
\tag{38}
$$

Combining Equation (34), Equation (36), Equation (37), and Equation (38), we get:

$$\text{Reg}_{\hat{f}_{m+1}}(\pi) \leq \frac{7\text{Reg}(\pi)}{5} + \frac{5K}{8\gamma_{m+1}} + \frac{2K(1+2C_0)}{5\gamma_{m+1}} + \frac{4C_8 + 2\sqrt{C_7}(C_9+5)}{5}\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}}$$
$$+ \frac{C_9+5}{5}\sqrt{V(p_{m+1},\pi)B} + \frac{3C_9+5}{5}\sqrt{KB}.$$

Now choosing constants so that $C_0 \geq 2$, $C_9 \geq 2.5$, and $C_8 \geq 2\sqrt{C_7}(C_9+5)+(3C_9+5)$. The above inequality then gives us:

$$\text{Reg}_{\hat{f}_{m+1}}(\pi) \leq 2\text{Reg}(\pi) + \frac{C_0K}{\gamma_{m+1}} + C_8\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + C_9\sqrt{V(p_{m+1},\pi)B}. \tag{39}$$

This completes the inductive step. $\qquad\square$

B.8 Bounding true regret

For any epoch m, Lemma 12 bounds regret of the randomized policy Q_m.

Lemma 12 (Action selection kernel has low true regret)**.** *Suppose the event $\mathcal{W}$ defined in (28) holds, and $\delta \leq 0.5$. And let $C_{10} := C_8 + C_9$. Then for all epochs m, we have:*

$$\sum_{\pi\in\Psi} Q_m(\pi)Reg(\pi) \leq \frac{(2+C_0)K}{\gamma_m} + C_{10}\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}}.$$

Proof. Note that:

$$\sum_{\pi\in\Psi} Q_m(\pi)\text{Reg}(\pi)$$
$$\leq \sum_{\pi\in\Psi} Q_m(\pi)\left(2\text{Reg}_{\hat{f}_m}(\pi) + \frac{C_0K}{\gamma_m} + C_8\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + C_9\sqrt{V(p_m,\pi)B}\right)$$
$$\leq \frac{2K}{\gamma_m} + \frac{C_0K}{\gamma_m} + C_8\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + C_9\sqrt{KB} \leq \frac{(2+C_0)K}{\gamma_m} + C_{10}\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}}.$$

Where the first inequality follows from Lemma 11, and the second inequality follows from Lemmas 4 and Lemma 5. $\qquad\square$

B.9 Proof of Theorem 2

We can now bound the cumulative regret of Epsilon-FALCON. Fix some (possibly unknown) horizon T. Let $\mathcal{T}_{\text{active}} \subseteq [T]$ be the set of time-steps that are in the active phase of some

epoch. Similarly let $\mathcal{T}_{\text{passive}} \subseteq [T]$ be the set of time-steps that are in the active phase of some epoch. Let $m(t)$ denote the epoch in which the time-step t occurs. For each round $t \in \{1, 2, \ldots, T\}$, define:

$$M_t := r_t(\pi^*(x_t)) - r_t(a_t) - \sum_{\pi \in \Psi} Q_{m(t)}(\pi) \text{Reg}(\pi).$$

Recall that from Lemma 3, for all $t \in \mathcal{T}_{\text{active}}$ we have:

$$\underset{x_t, r_t, a_t}{\mathbb{E}} [r_t(\pi^*(x)) - r_t(a_t)|\Gamma_{t-1}] = \sum_{\pi \in \Psi} Q_{m(t)}(\pi) \text{Reg}(\pi).$$

Hence from Azuma's inequality, with probability at least $1 - \delta/2$, we have:

$$\sum_{t \in \mathcal{T}_{\text{active}}} M_t \leq 2\sqrt{2|\mathcal{T}_{\text{active}}| \log(2/\delta)} \leq 2\sqrt{2T \log(2/\delta)}. \tag{40}$$

Hence when Equation (40) holds, we get:

$$\sum_{t=1}^{T} \left(r_t(\pi^*(x_t)) - r_t(a_t) \right)$$
$$= \sum_{t \in \mathcal{T}_{\text{passive}}} \left(r_t(\pi^*(x_t)) - r_t(a_t) \right) + \sum_{t \in \mathcal{T}_{\text{active}}} \left(r_t(\pi^*(x_t)) - r_t(a_t) \right) \tag{41}$$
$$\leq |\mathcal{T}_{\text{passive}}| + \sum_{t \in \mathcal{T}_{\text{active}}} \sum_{\pi \in \Psi} Q_{m(t)}(\pi) \text{Reg}(\pi) + \sqrt{8T \log(2/\delta)}$$

Since in any epoch $m \geq 1$, there are at most $1 + \epsilon(\tau_m - \tau_{m-1})$ passive time-steps. Therefore:

$$|\mathcal{T}_{\text{passive}}| \leq \epsilon T + m(T) \leq 1 + \log_2(T) + \epsilon T. \tag{42}$$

Further when $\mathcal{W}$ holds, from Lemma 12, we have:

$$\sum_{t \in \mathcal{T}_{\text{active}}} \sum_{\pi \in \Psi} Q_{m(t)}(\pi) \text{Reg}(\pi)$$
$$\leq \tau_1 + \sum_{\{t \in \mathcal{T}_{\text{active}} \mid t \geq \tau_1 + 1\}} \sum_{\pi \in \Psi} Q_{m(t)}(\pi) \text{Reg}(\pi)$$
$$\leq \tau_1 + C_{10} \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} T + \sum_{t=\tau_1+1}^{T} \frac{(2 + C_0)K}{\gamma_{m(t)}} \tag{43}$$
$$\leq \tau_1 + C_{10} \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} T + \sum_{m=2}^{m(T)} \frac{(2 + C_0)K}{\gamma_m} (\tau_m - \tau_{m-1})$$

Since $\tau_1 \geq 4$, $\tau_{m(t)-1} \leq t$ for all $t \geq 1$, and $\tau_m = \tau_1 2^{m-1}$ for all $m \geq 1$. We get that $\tau_{m(T)-1} \leq T$, $\tau_{m(T)} \leq 2T$, and $m-1 \leq \log_2(T)$ for all $m \leq m(T)$. Therefore, we get:

$$
\begin{aligned}
&\sum_{m=2}^{m(T)} \frac{(2+C_0)K}{\gamma_m}(\tau_m - \tau_{m-1}) \\
&= \sum_{m=2}^{m(T)} (2+C_0)K\sqrt{\frac{\ln^{\rho'}(\tau_{m-1} - \tau_{m-2})\ln((m-1)/\delta)\mathbf{comp}(\mathcal{F})}{C_3 K(\tau_{m-1} - \tau_{m-2})^\rho}}(\tau_m - \tau_{m-1}) \qquad (44) \\
&\leq \frac{(2+C_0)}{\sqrt{C_3}}\sqrt{K\ln^{\rho'}(T)\ln(\log_2(T)/\delta)\mathbf{comp}(\mathcal{F})}\sum_{m=2}^{m(T)} \frac{\tau_m - \tau_{m-1}}{\sqrt{(\tau_{m-1} - \tau_{m-2})^\rho}}.
\end{aligned}
$$

Since for all $m \geq 1$, $\tau_{m+1} = 2\tau_m$, we have that:

$$
\begin{aligned}
\sum_{m=2}^{m(T)} \frac{\tau_m - \tau_{m-1}}{\sqrt{(\tau_{m-1} - \tau_{m-2})^\rho}} &= 2^{\rho/2}\sum_{m=2}^{m(T)} \frac{\tau_m - \tau_{m-1}}{\tau_{m-1}^{\rho/2}} \leq 2^{\rho/2}\sum_{m=2}^{m(T)} \int_{\tau_{m-1}}^{\tau_m} \frac{dy}{y^{\rho/2}} \\
&= 2^{\rho/2}\int_{\tau_1}^{\tau_{m(T)}} \frac{dy}{y^{\rho/2}} \leq \frac{2^{\rho/2}}{1-\rho/2}\tau_{m(T)}^{1-\rho/2} \leq \frac{2}{1-\rho/2}T^{1-\rho/2}.
\end{aligned} \qquad (45)
$$

Where the last inequality follows from the fact that $\tau_{m(T)} \leq 2T$. Hence when Equation (40) and $\mathcal{W}$ hold, from Equation (41), Equation (42), Equation (43), Equation (44), and Equation (45), we get:

$$
\begin{aligned}
&\sum_{t=1}^{T}\Big(r_t(\pi^*(x_t)) - r_t(a_t)\Big) \\
&\leq |\mathcal{T}_{\text{passive}}| + \sum_{t\in\mathcal{T}_{\text{active}}}\sum_{\pi\in\Psi} Q_{m(t)}(\pi)\text{Reg}(\pi) + \sqrt{8T\log(2/\delta)} \\
&\leq 1 + \log_2(T) + \epsilon T + \sqrt{8T\log(2/\delta)} + \tau_1 + C_{10}\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}T} \\
&\quad + \frac{(2+C_0)}{\sqrt{C_3}}\frac{4}{2-\rho}\sqrt{KT^{2-\rho}\ln^{\rho'}(T)\ln(\log_2(T)/\delta)\mathbf{comp}(\mathcal{F})} \\
&= \mathcal{O}\left(\left(\epsilon + \sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}}\right)T + \sqrt{KT^{2-\rho}\ln^{\rho'}(T)\ln(\log_2(T)/\delta)\mathbf{comp}(\mathcal{F})}\right).
\end{aligned}
$$

Note that from Lemma 7, we know that $\mathcal{W}$ holds with probability $1-\delta/2$. Also from Azuma's inequality, we showed that Equation (40) holds with probability $1-\delta/2$. Hence from union bound, we get that the above inequality holds with probability $1-\delta$. This concludes the proof of Equation (11).

C Learning rates

In this section, we restate results from Koltchinskii (2011) on bounds for excess risk in a form that is convenient for us to use. We consider the standard machine learning setting. That is, we let (Z, Y) be a random tuple in $\mathcal{Z} \times [0, 1]$ with distribution P. Assume Z is observable and Y is to be predicted based on an observation of Z. Let $l : \mathbb{R} \times \mathbb{R}$ be the squared error loss, that is $l(a, b) = (a - b)^2$. Given a function $g : \mathcal{Z} \to \mathbb{R}$, let $(l \cdot g)(z, y) := l(y, g(z))$ be interpreted as the loss suffered when $g(z)$ is used to predict y. Let $\mathcal{G}$ be a convex class of functions from $\mathcal{Z}$ to $\mathbb{R}$. The problem of optimal prediction can be viewed as finding a solution to the following risk minimization problem:

$$\min_{g \in \mathcal{G}} P(l \cdot g).$$

Where $P(l \cdot g)$ is a short hand for $\mathbb{E}_P[(l \cdot g)(Z, Y)]$. Let $\hat{g}^* \in \mathcal{G}$ be a solution to the above risk minimization problem. Let $g^*(z) := \mathbb{E}_P[Y | Z = z]$. Since the distribution P is unknown, the above risk minimization problem is replaced by the empirical risk minimization problem:

$$\min_{g \in \mathcal{G}} P_n(l \cdot g).$$

Where P_n is an empirical distribution generated from n i.i.d. samples of (Z, Y) from the distribution P. Here $P_n(l \cdot g)$ is a short hand for $\mathbb{E}_{P_n}[(l \cdot g)(Z, Y)]$. In general, we will use $P(\cdot)$ and $P_n(\cdot)$ as a short hand for $\mathbb{E}_P[\cdot]$ and $\mathbb{E}_{P_n}[\cdot]$ respectively. Now, let $\hat{g}_n \in \mathcal{G}$ be a solution to the above empirical risk minimization problem. Also let $\mathcal{G}^l$ denote the loss class, that is $\mathcal{G}^l := \{l \cdot g \mid g \in \mathcal{G}\}$. For any $g \in \mathcal{G}$, we define the excess risk $(\mathcal{E}(l \cdot g))$ and the empirical excess risk $(\hat{\mathcal{E}}(l \cdot g))$, given by:

$$\mathcal{E}(l \cdot g) := P(l \cdot g) - \min_{l \cdot g' \in \mathcal{G}^l} P(l \cdot g') = \mathbb{E}_P[(l \cdot g)(Z, Y)] - \min_{g' \in \mathcal{G}} \mathbb{E}_P[(l \cdot g')(Z, Y)],$$

$$\hat{\mathcal{E}}(l \cdot g) := P_n(l \cdot g) - \min_{l \cdot g' \in \mathcal{G}^l} P_n(l \cdot g') = \mathbb{E}_{P_n}[(l \cdot g)(Z, Y)] - \min_{g' \in \mathcal{G}} \mathbb{E}_{P_n}[(l \cdot g')(Z, Y)].$$

For $\delta \in \mathbb{R}_+$, we define the δ-minimal set $(\mathcal{G}^l(\delta))$ and the empirical δ-minimal set $(\hat{\mathcal{G}}^l(\delta))$, given by:

$$\mathcal{G}^l(\delta) := \left\{ h \in \mathcal{G}^l \mid \mathcal{E}(h) \le \delta \right\}, \quad \hat{\mathcal{G}}^l(\delta) := \left\{ h \in \mathcal{G}^l \mid \hat{\mathcal{E}}(h) \le \delta \right\}.$$

We now define a version of local Rademacher averages (ψ_n). We start by defining the Rademacher process $(R_n(\cdot))$. For any function $h : \mathcal{Z} \to \mathbb{R}$, $R_n(h)$ is given by:

$$R_n(h) := \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h(Z_i).$$

Where $\{Z_i\}_{i=1}^n$ are i.i.d. random samples from the marginal distribution of P on $\mathcal{Z}$. And where $\{\epsilon_i\}_{i=1}^n$ are i.i.d. Rademacher random variables (that is, ϵ_i takes the values $+1$ and -1 with probability $1/2$ each) independent of Z_i. We also define a (pseudo)-metric (ρ_P)

on the set of functions that are square integrable with respect to P, such that: $\rho_P(f,g) := \sqrt{P((f-g)^2)}$. We now define the local Rademacher average (ψ_n) as:

$$\psi_n(\delta) := 16 \, \underset{P,\epsilon}{\mathbb{E}} \, \sup\{|R_n(g - \hat{g}^*)| \mid g \in \mathcal{G}, \rho_P^2(g, \hat{g}^*) \leq 2\delta\}.$$

Finally we define the $\flat$-transform and the $\sharp$-transform. For any $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$, define:

$$\kappa^\flat(\delta) := \sup_{\delta' \geq \delta} \frac{\kappa(\delta')}{\delta'}, \quad \kappa^\sharp(\epsilon) := \inf\{\delta > 0 \mid \kappa^\flat(\delta) \leq \epsilon\}.$$

It is easy to see that $\sharp$-transforms are decreasing functions, and we will use this property in the proof of Lemma 13. For more details and properties of these transformations, see section A.3 in Koltchinskii (2011). We now get to the main Lemma of this section (Lemma 13), which is implicitly evident from results in Koltchinskii (2011). Lemma 13 shows that, with high-probability, the δ-minimal set $(\mathcal{G}^l(\delta))$ and the empirical δ-minimal set $(\hat{\mathcal{G}}^l(\delta))$ approximate each other.

Lemma 13. *Let $\mathcal{G}$ be a convex class of functions from $\mathcal{Z}$ to $[0,1]$. Suppose $\zeta \in (0, 1/2)$. With probability at least $1 - \zeta$, for all $\delta \geq \max\{\psi_n^\sharp(\frac{1}{16}), \frac{16384 \ln(2/\zeta)}{n}\}$ we have:*

$$\mathcal{G}^l(\delta) \subset \hat{\mathcal{G}}^l(3\delta/2), \quad \hat{\mathcal{G}}^l(\delta) \subset \mathcal{G}^l(2\delta).$$

Proof. Lemma 13 is a corollary of a few Lemmas and inequalities in Koltchinskii (2011). In the next few steps, we will define a function U_n and bound $U_n^\sharp(1/2)$. Lemma 13 will follow from Lemma 4.2 in Koltchinskii (2011) and the bounds on $U_n^\sharp(1/2)$. Let $D(\delta)$ denote the ρ_P-diameter of the δ-minimal set $(\mathcal{G}^l(\delta))$. That is:

$$D(\delta) := \sup_{h,h' \in \mathcal{G}^l(\delta)} \rho_P(h, h').$$

Also let ϕ_n be a measure of empirical approximation:

$$\phi_n(\delta) := \mathbb{E}\left[\sup_{h,h' \in \mathcal{G}^l(\delta)} \left|(P_n - P)(h - h')\right|\right].$$

Let $t, \sigma > 0$, and $q > 1$. We will fix the values of t, σ and q later in the proof. Let $\delta_j := q^{-j}$ and $t_j := t\frac{\delta_j}{\sigma}$, for all $j \geq 0$. We will now define a function $U_n : (0, 1] \to \mathbb{R}_+$. For all $j \geq 0$ and $\delta \in (\delta_{j+1}, \delta_j]$, define:

$$U_n(\delta) := \phi_n(\delta_j) + \sqrt{2\frac{t_j}{n}(D^2(\delta_j) + 2\phi_n(\delta_j))} + \frac{t_j}{2n}$$

$$= \phi_n(\delta_j) + \sqrt{2\frac{t}{n}\frac{\delta_j}{\sigma}(D^2(\delta_j) + 2\phi_n(\delta_j))} + \frac{t}{2n}\frac{\delta_j}{\sigma}$$

The reader may have astutely noticed that functions like U_n appear as upper bounds in Talagrand type concentration inequalities, in fact that is where this comes from. We now

bound $U_n^\flat(\eta)$ for all $\eta > 0$:

$$U_n^\flat(\eta) \leq \sup_{\delta_j \geq \eta} \frac{q}{\delta_j} \left\{ \phi_n(\delta_j) + \sqrt{2\frac{t}{n}\frac{\delta_j}{\sigma}(D^2(\delta_j) + 2\phi_n(\delta_j))} + \frac{t}{2n}\frac{\delta_j}{\sigma} \right\}$$

$$\leq q \sup_{\delta_j \geq \eta} \frac{\phi_n(\delta_j)}{\delta_j} + \sup_{\delta_j \geq \eta} q\left\{ \sqrt{\frac{2t}{\sigma n}\frac{D^2(\delta_j)}{\delta_j}} + \sqrt{\frac{4t}{\sigma n}\frac{\phi_n(\delta_j)}{\delta_j}} \right\} + \frac{qt}{2\sigma n} \qquad (46)$$

$$\leq q\phi_n^\flat(\eta) + q\sqrt{\frac{2t}{\sigma n}(D^2)^\flat(\eta)} + q\sqrt{\frac{4t}{\sigma n}\phi_n^\flat(\eta)} + \frac{qt}{2\sigma n}$$

From Equation (46), we get a bound on $U_n^\sharp(\epsilon)$ for all $\epsilon > 0$:

$$U_n^\sharp(\epsilon) := \inf\{\eta > 0 \mid U_n^\flat(\eta) \leq \epsilon\}$$

$$\leq \inf\left\{ \eta > 0 \mid q\phi_n^\flat(\eta) + q\sqrt{\frac{2t}{\sigma n}(D^2)^\flat(\eta)} + q\sqrt{\frac{4t}{\sigma n}\phi_n^\flat(\eta)} + \frac{qt}{2\sigma n} \leq \epsilon \right\}$$

$$\leq \inf\left\{ \eta > 0 \mid \phi_n^\flat(\eta) + \sqrt{\frac{2t}{\sigma n}(D^2)^\flat(\eta)} + \sqrt{\frac{4t}{\sigma n}\phi_n^\flat(\eta)} \leq \frac{1}{q}\left(\epsilon - \frac{qt}{2\sigma n}\right) \right\}$$

$$\leq \max\left\{ \inf\left\{ \eta > 0 \mid \phi_n^\flat(\eta) \leq \frac{1}{3q}\left(\epsilon - \frac{qt}{2\sigma n}\right) \right\}, \inf\left\{ \eta > 0 \mid \sqrt{\frac{2t}{\sigma n}(D^2)^\flat(\eta)} \leq \frac{1}{3q}\left(\epsilon - \frac{qt}{2\sigma n}\right) \right\}, \right.$$

$$\left. \inf\left\{ \eta > 0 \mid \sqrt{\frac{4t}{\sigma n}\phi_n^\flat(\eta)} \leq \frac{1}{3q}\left(\epsilon - \frac{qt}{2\sigma n}\right) \right\} \right\}$$

$$\leq \max\left\{ \phi_n^\sharp\left(\frac{\epsilon}{3q} - \frac{t}{6\sigma n}\right), (D^2)^\sharp\left(\frac{\sigma n}{2t}\left(\frac{\epsilon}{3q} - \frac{t}{6\sigma n}\right)^2\right), \phi_n^\sharp\left(\frac{\sigma n}{4t}\left(\frac{\epsilon}{3q} - \frac{t}{6\sigma n}\right)^2\right) \right\}$$

$$(47)$$

To further bound $U_n^\sharp(\cdot)$, we need to bound the terms in Equation (47). From page 78 in Koltchinskii (2011), we get that the convexity of $\mathcal{G}$ implies a bound on $D(\cdot)$ which further gives us a bound on $(D^2)^\flat(\cdot)$:

$$D(\delta) \leq 4\sqrt{2}\sqrt{\delta}, \quad \text{for all } \delta \geq 0.$$

$$\implies (D^2)^\flat(\eta) = \sup_{\delta' \geq \eta} \frac{D^2(\delta')}{\delta'} \leq 32, \quad \text{for all } \eta \geq 0.$$

Hence we have:

$$(D^2)^\sharp(\epsilon) = 0, \quad \text{for all } \epsilon \geq 32 \qquad (48)$$

To upper-bound $U_n^\sharp(1/2)$, we now bound the $(D^2)^\sharp(\cdot)$ term in Equation (47). To do this we choose $\sigma = 4096tq^2/n$. Hence from the choice of σ and from Equation (48), we get:

$$\sigma \geq \frac{4096tq^2}{n} = 1024\frac{tq^2}{n(1/2)^2} \implies \frac{\sigma n}{2t}\frac{(1/2)^2}{16q^2} \geq 32 \implies \frac{\sigma n}{2t}\left(\frac{1/2}{3q} - \frac{t}{6\sigma n}\right)^2 \geq 32$$

$$\implies (D^2)^\sharp\left(\frac{\sigma n}{2t}\left(\frac{1/2}{3q} - \frac{t}{6\sigma n}\right)^2\right) = 0$$

$$(49)$$

37

We now bound the $\phi_n^\sharp(\cdot)$ terms in Equation (47), in terms of $\psi_n^\sharp(\cdot)$. Again from page 78 in Koltchinskii (2011), we get that the convexity of $\mathcal{G}$ implies a bound on $\phi_n(\cdot)$ which further gives us a bound on $\phi_n^\sharp(\cdot)$:

$$\phi_n(\delta) \leq \psi_n(\delta) \quad \text{for all } \delta \geq 0.$$
$$\implies \phi_n^\sharp(\epsilon) \leq \psi_n^\sharp(\epsilon) \quad \text{for all } \epsilon \geq 0. \tag{50}$$

To upper-bound $U_n^\sharp(1/2)$, we now bound the $\phi_n^\sharp(\cdot)$ terms in Equation (47). From the choice of σ and from Equation (50), we get:

$$\sigma \geq \frac{4qt}{n} = \frac{2qt}{n(1/2)} \implies \frac{1/2}{12q} \geq \frac{t}{6\sigma n} \implies \frac{1/2}{3q} - \frac{t}{6\sigma n} \geq \frac{1/2}{4q}$$
$$\implies \phi_n^\sharp\left(\frac{1/2}{3q} - \frac{t}{6\sigma n}\right) \leq \phi_n^\sharp\left(\frac{1/2}{4q}\right) \leq \psi_n^\sharp\left(\frac{1/2}{4q}\right). \tag{51}$$

Again from the choice of σ and from Equation (50), we get:

$$\sigma \geq \frac{32tq}{n} = \frac{16tq}{n(1/2)} \implies \frac{\sigma n}{4t}\frac{(1/2)^2}{16q^2} \geq \frac{1/2}{4q} \implies \frac{\sigma n}{4t}\left(\frac{1/2}{3q} - \frac{t}{6\sigma n}\right)^2 \geq \frac{1/2}{4q}$$
$$\implies \phi_n^\sharp\left(\frac{\sigma n}{4t}\left(\frac{1/2}{3q} - \frac{t}{6\sigma n}\right)^2\right) \leq \phi_n^\sharp\left(\frac{1/2}{4q}\right) \leq \psi_n^\sharp\left(\frac{1/2}{4q}\right). \tag{52}$$

Combining Equation (47), Equation (49), Equation (51), and Equation (52), we get:

$$U_n^\sharp(1/2) \leq \psi_n^\sharp\left(\frac{1}{8q}\right). \tag{53}$$

Lemma 4.2 in Koltchinskii (2011) states that with probability at least $1 - \sum_{\delta_j \geq \delta_n^\diamond} e^{-t_j}$, for all $\delta \geq \delta_n^\diamond$ we have: $\mathcal{G}^l(\delta) \subset \hat{\mathcal{G}}^l(3\delta/2)$ and $\hat{\mathcal{G}}^l(\delta) \subset \mathcal{G}^l(2\delta)$. Where $\delta_n^\diamond$ is any number such that $\delta_n^\diamond \geq U_n^\sharp(1/2)$. Hence from Equation (53), we can choose:

$$\delta_n^\diamond = \max\left\{\psi_n^\sharp\left(\frac{1}{8q}\right), \frac{4096tq^2}{n}\right\} \geq \max\{U_n^\sharp(1/2), \sigma\}.$$

Now by choosing $q = 2$ and $t = \ln(2/\zeta)$, using the fact that $\zeta \in (0, 1/2)$, we get that $t \geq 1$. Hence, we have that:

$$\sum_{\delta_j \geq \delta_n^\diamond} e^{-t_j} \leq \sum_{\delta_j \geq \sigma} e^{-t_j} = \sum_{\delta_j \geq \sigma} \exp\left\{-t\frac{\delta_j}{\sigma}\right\} \leq \sum_{j \geq 0} e^{-tq^j} =$$
$$e^{-t} + \frac{q}{q-1}\sum_{j=1}^{\infty} q^{-j}(q^j - q^{j-1})e^{-tq^j} \leq e^{-t} + \frac{1}{q-1}\int_1^{\infty} e^{-tx}dx =$$
$$e^{-t} + \frac{1}{q-1}\frac{1}{t}e^{-t} \leq e^{-t} + \frac{1}{q-1}e^{-t} = \frac{q}{q-1}e^{-t} = \zeta.$$

That is, we have shown that with probability at least $1-\zeta$, for all $\delta \geq \max\{\psi_n^\sharp(1/8q), 4096tq^2/n\}$, we have: $\mathcal{G}^l(\delta) \subset \hat{\mathcal{G}}^l(3\delta/2)$ and $\hat{\mathcal{G}}^l(\delta) \subset \mathcal{G}^l(2\delta)$. $\qquad\square$

Corollary 2 uses Lemma 13 and a bound on $\psi_n^\sharp(\cdot)$ when $\mathcal{G}$ is a convex subset of a d-dimensional linear space to show that for all $\delta \geq \frac{Cd\ln(1/\zeta)}{n}$, the δ-minimal set $(\mathcal{G}^l(\delta))$ and the empirical δ-minimal set $(\hat{\mathcal{G}}^l(\delta))$ approximate each other with probability at least $1 - \zeta$.

Corollary 2. *Let $\mathcal{G}$ be a convex class of functions from $\mathcal{Z}$ to $[0,1]$, and a subset d dimensional linear space. Suppose $\zeta \in (0, 1/2)$. With probability at least $1 - \zeta$, for all $\delta \geq \frac{Cd\ln(1/\zeta)}{n}$ we have:*

$$\mathcal{G}^l(\delta) \subset \hat{\mathcal{G}}^l(3\delta/2), \quad \hat{\mathcal{G}}^l(\delta) \subset \mathcal{G}^l(2\delta).$$

Where $C > 0$ is a positive constant.

Proof. Since $\mathcal{G}$ is a convex subset of a d dimensional linear space, we get from proposition 3.2 in Koltchinskii (2011) that:

$$\psi_n(\delta) = 16 \, \mathbb{E}_{P,\epsilon} \sup\{|R_n(g - \hat{g}^*)| \mid g \in \mathcal{G}, \rho_P^2(g, \hat{g}^*) \leq 2\delta\}$$

$$\leq 16\sqrt{2\delta}\sqrt{\frac{d}{n}}.$$

Which implies that:

$$\psi_n^\flat(\delta) = \sup_{\delta' \geq \delta} \frac{\psi_n(\delta')}{\delta'} \leq \sup_{\delta' \geq \delta} 16\sqrt{\frac{2d}{n\delta'}} = 16\sqrt{\frac{2d}{n\delta}}.$$

Hence, we get that:

$$\psi_n^\sharp(\epsilon) = \inf\{\delta > 0 \mid \psi_n^\flat(\delta) \leq \epsilon\}$$

$$\leq \inf\left\{\delta > 0 \mid 16\sqrt{\frac{2d}{n\delta}} \leq \epsilon\right\}$$

$$= \inf\left\{\delta > 0 \mid \frac{512d}{n\epsilon^2} \leq \delta\right\} = \frac{512d}{n\epsilon^2}.$$

Therefore:

$$\psi_n^\sharp(1/16) \leq \frac{512d}{n(1/16)^2} = \frac{131072d}{n}.$$

Hence Corollary 2 follows from Lemma 13 and the above inequality. $\square$

Rates for general classes of functions Lemma 14 provides rates for $\psi_n^\sharp$ for different classes of $\mathcal{G}$. Hence similar to Corollary 2, these bounds imply that for all $\delta \geq \mathcal{O}(\psi_n^\sharp(1/16)\ln(1/\zeta))$, the δ-minimal set $(\mathcal{G}^l(\delta))$ and the empirical δ-minimal set $(\hat{\mathcal{G}}^l(\delta))$ approximate each other with probability at least $1 - \zeta$. The results stated in Lemma 14 are from Koltchinskii (2011) (pages 85 to 87), we state the same results without proof.

Lemma 14. *Let $\mathcal{G}$ be a convex class of functions from $\mathcal{Z}$ to $[0,1]$.*

- Suppose $\mathcal{G}$ is VC-subgraph class of functions with VC-dimension V. Then for all $\epsilon > 0$, we have:

$$\psi_n^\sharp(\epsilon) \leq \mathcal{O}\left(\frac{V}{n\epsilon^2} \log\left(\frac{n\epsilon^2}{V}\right)\right).$$

- Let $N(\mathcal{G}, L_2(P_n), \epsilon)$ denote the number of $L_2(P_n)$ balls of radius ϵ covering $\mathcal{G}$. Suppose the empirical entropy is bounded, that is for some $\rho \in (0,1)$ we have that: $\log(N(\mathcal{G}, L_2(P_n), \epsilon)) \leq \mathcal{O}(\epsilon^{-2\rho})$. Then for all $\epsilon > 0$, we have:

$$\psi_n^\sharp(\epsilon) \leq \mathcal{O}\left((n\epsilon^2)^{\frac{-1}{1+\rho}}\right).$$

- Suppose $\mathcal{G}$ is a convex hull of a VC-subgraph class of functions with VC-dimension V. Then for all $\epsilon > 0$, we have:

$$\psi_n^\sharp(\epsilon) \leq \mathcal{O}\left(\left(\frac{V}{n\epsilon^2}\right)^{\frac{1}{2}\frac{2+V}{1+V}}\right).$$

Proving Assumption 1 We now describe the general outline to prove Assumption 1 using the results in this section for different convex classes $\mathcal{F}$. Note that we need the conditions of Assumption 1 to hold for any convex set $\mathcal{F}' \subseteq \mathcal{F}$, and any action selection kernel p. First let $\mathcal{Z}$ used in this section correspond to $\mathcal{X} \times \mathcal{A}$, and let distribution P correspond to the distribution described by $x_t \sim \mathcal{D}_\mathcal{X}$, $a|x \sim p(a|x)$ and $r_t \sim \mathcal{D}_{r|x,a}$ induced by the action selection kernel p. Also note that the empirical distribution corresponding to $\widetilde{S}$, in fact corresponds to P_n in this section. Hence from lemma 13, to show that the empirical and population η-minimal sets approximate each other with high-probability (as is required in Assumption 1), it is sufficient to bound $\psi_n^\sharp(1/16)$ uniformly for all convex subsets $\mathcal{F}' \subseteq \mathcal{F}$ and all distributions induced by action selection kernels. Such bounds can be proven for many interesting convex classes of estimators because the bounds on $\psi_n^\sharp$ are often distribution-free and we often have that $\mathbf{comp}(\mathcal{F}') \leq \mathbf{comp}(\mathcal{F})$.

For example, say $\mathcal{F}$ is a convex subset of a d dimensional linear space, then any convex subset $\mathcal{F}' \subseteq \mathcal{F}$ is also a convex subset of a d dimensional linear space. Hence, corollary 2 can be used on $\mathcal{F}'$ to show that the empirical and population η-minimal sets approximate each other (as is required in Assumption 1). Note that this along with Theorem 2 gives us Theorem 1. Similarly, say $\mathcal{F}$ is a convex set with VC sub-graph dimension V. Note that, for any convex set $\mathcal{F}' \subseteq \mathcal{F}$, we have that $\mathcal{F}'$ has a VC sub-graph dimension V. Hence, we can then use Lemma 14 to bound $\psi_n^\sharp(1/16)$ in a distribution free manner and then show that the empirical and population η-minimal sets approximate each other (using Lemma 13). Note that this along with Theorem 2 gives us Example 1 in Section 2. We can similarly that Examples 2 and 3 follow from Theorem 2 and the results in this section.

D Solving the constrained regression problem

In this section, we show the constrained regression problem can be solved using a weighted regression oracle. The purpose of this argument is to show that the constrained regression problem is computationally tractable for many class of estimators. Suppose $\mathcal{F}$ is a convex set. Let $S, S' \subseteq \mathcal{X} \times \mathcal{A} \times [0, 1]$, often these sets represent the data collected in the active and passive phases respectively. Consider the following optimization problem:

$$
\begin{aligned}
\min_{f \in \mathcal{F}} \quad & \frac{1}{|S|} \sum_{(x,a,r(a)) \in S} (f(x,a) - r(a))^2 \\
\text{s.t.} \quad & \frac{1}{|S'|} \sum_{(x,a,r(a)) \in S'} (f(x,a) - r(a))^2 \leq \alpha + \beta.
\end{aligned}
\tag{54}
$$

Where $\beta > 0$ is a fixed problem parameter, and $\alpha := \frac{1}{|S'|} \min_{f \in \mathcal{F}} \sum_{(x,a,r(a)) \in S'} (f(x,a) - r(a))^2$. From the definition of α and β, we have that there exists a $g \in \mathcal{F}$ such that:

$$
\frac{1}{|S'|} \sum_{(x,a,r(a)) \in S'} (g(x,a) - r(a))^2 < \alpha + \beta.
\tag{55}
$$

That is there is a $g \in \mathcal{F}$ such that the constraint in the optimization problem (54) is not tight. Hence strong duality holds (See proposition 1.1.3 in Bertsekas and Scientific (2015)). Now consider the lagragian of the constrained regression problem:

$$
L(f, \lambda) := \frac{1}{|S|} \sum_{(x,a,r(a)) \in S} (f(x,a) - r(a))^2 + \lambda \left(\frac{1}{|S'|} \sum_{(x,a,r(a)) \in S'} (f(x,a) - r(a))^2 - \alpha - \beta \right).
$$

Note that probelem 54 can be re-written as, $\min_{f \in \mathcal{F}} \max_{\lambda \geq 0} L(f, \lambda)$. Since strong duality holds, this is equivalent to solving the following dual optimization problem:

$$
\max_{\lambda \geq 0} \min_{f \in \mathcal{F}} L(f, \lambda) \equiv \max_{\lambda \geq 0} g(\lambda).
$$

Where, $g(\lambda) := \min_{f \in \mathcal{F}} L(f, \lambda)$. For any fixed λ, note that evaluating $g(\lambda)$ is equivalent to solving a weighted regression problem:

$$
\arg\min_{f \in \mathcal{F}} L(f, \lambda) = \arg\min_{f \in \mathcal{F}} \frac{1}{|S|} \sum_{(x,a,r(a)) \in S} (f(x,a) - r(a))^2 + \frac{\lambda}{|S'|} \sum_{(x,a,r(a)) \in S^{pass}} (f(x,a) - r(a))^2.
$$

Now, let λ^* be an optimal dual solution. Since the dual problem is a one-dimensional concave maximization problem, we can use a bisection method to find the optimal dual solution. Hence one can solve the dual optimization problem with $\mathcal{O}(\log(\lambda^*))$ calls to evaluate $g(\cdot)$, where each evaluation call corresponds to one call to a weighted regression oracle. Suppose this procedure outputs $\bar{\lambda}$ as the optimal dual solution. We then output the estimator that solves:

$$
\arg\min_{f \in \mathcal{F}} \frac{1}{|S|} \sum_{(x,a,r(a)) \in S} (f(x,a) - r(a))^2 + \frac{\bar{\lambda}}{|S'|} \sum_{(x,a,r(a)) \in S^{pass}} (f(x,a) - r(a))^2.
$$

Note that this estimator must be optimal for the primal problem [9]. Since there are many algorithms and heuristics to solve weighted regression problems, this argument shows that the constrained regression problem is often computationally tractable.

Algorithm 2 Solving constrained regression

---

input: Given a threshold parameter $\kappa > 0$ and a weighted regression oracle to evaluate $g(\cdot)$.

1: Set $\lambda_L = 0$, $\lambda_M = 1$, and $\lambda_R = 2$.
2: **while** $g(\lambda_M) < g(\lambda_R)$ **do**
3:     Set $\lambda_R \leftarrow 2\lambda_R$ and set $\lambda_M \leftarrow 2\lambda_M$.
4: **end while**
5: **while** $|\lambda_R - \lambda_L| \geq \kappa$ **do**
6:     **if** $g(\lambda_M + \kappa) > g(\lambda_M)$ **then**
7:         Set $\lambda_L \leftarrow \lambda_M$.
8:     **else**
9:         Set $\lambda_R \leftarrow \lambda_M$.
10:     **end if**
11:     Set $\lambda_M \leftarrow \frac{1}{2}(\lambda_L + \lambda_R)$.
12: **end while**
13: Return the output of the weighted regression oracle on the following problem:

$$\min_{f \in \mathcal{F}} \frac{1}{|S|} \sum_{(x,a,r(a)) \in S} (f(x,a) - r(a))^2 + \frac{\lambda_M}{|S'|} \sum_{(x,a,r(a)) \in S^{pass}} (f(x,a) - r(a))^2.$$

We note that in practice, rather than solving multiple weighted regression problems, one may prefer to directly find a minimax solution to the lagrangian of the constrained regression problem (see Jin et al. (2019)).

E   Sensitivity of confidence intervals to realizability

In this section, we demonstrate that the confidence intervals used by LinUCB can be extremely sensitive to the realizability assumption. We also point out analogous issues in LinTS and FALCON (with linear estimates). We do this by constructing a family of contextual bandit problems where the approximation error to the class of linear models can be arbitrarily small, but given data from the policy induced by the best linear estimate (which also happens to be optimal), the confidence intervals used by LinUCB tightly concentrate around bad estimators that induce high-regret policies.

Consider a family of two armed contextual bandit problems that are parameterized by $\theta \in (0, 0.05]$. Let $\mathcal{X} = (0,1)$ be the set of contexts, and let $\mathcal{A} = \{1, 2\}$ be the set of arms.

---

[9] Here when we say optimal, we mean optimal up to the accuracy thresholds.

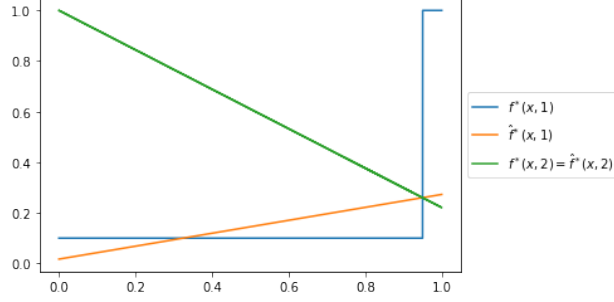Figure 5: This is a plot of the conditional expected reward (f^*) and the best linear estimate $\hat{f}^*$ when arms are sampled uniformly at random. Note that the conditional expected reward for arm 2 is linear. The problem is constructed so that the policy ($\pi_{\hat{f}^*}$) that is induced by the best linear estimates ($\hat{f}$) samples arm 2 for all x such that $f^*(x, 1) = 0.1$, and samples arm 1 for all x such that $f^*(x, 1) = 1$. Note that this policy is also optimal.

At every time-step, the environment draws a context according to the continuous uniform distribution on $\mathcal{X}$. That is, $D_{\mathcal{X}} \equiv \text{Unif}(\mathcal{X})$. To estimate the conditional expected reward (f^*) and select a policy, we pick estimators from a convex class of functions $\mathcal{F}$, where:

$$\mathcal{F} := \{f : \mathcal{X} \times \mathcal{A} \to \mathbb{R} \mid f(\cdot, 1) \text{ and } f(\cdot, 2) \text{ are linear}\}.$$

For $\theta = 0.05$, Figure 5 plots the conditional expected rewards (f^*) and the best linear estimate $\hat{f}^* \in \mathcal{F}$ when arms are sampled uniformly at random. We will now specify these terms more generally, starting with the conditional expected reward for arm 1, which is given by:

$$f^*(x, 1) := \begin{cases} 0.1, & \text{for all } x \leq 1 - \theta \\ 1, & \text{for all } x > 1 - \theta. \end{cases}$$

The conditional expected reward for arm 2 is linear, and is given by $f^*(x, 2) := 1 + m_\theta x$. Where m_θ is such that $\hat{f}^*(x, 1)$ and $f^*(x, 2)$ meet at $x = 1 - \theta$, which is ensured by defining:

$$m_\theta := \frac{\hat{f}^*(1 - \theta, 1) - 1}{1 - \theta}.$$

Since $f^*(\cdot, 2)$ is linear, we get that $\hat{f}^*(\cdot, 2) \equiv f^*(\cdot, 2)$. Further since $m_\theta < 0$, we get that $\hat{f}^*(x, 2)$ is decreasing in x. Similarly, one can show that $\hat{f}^*(x, 1)$ is increasing in x. Therefore, we get that $\pi_{\hat{f}^*}$ is given by:

$$\pi_{\hat{f}^*}(x) := \begin{cases} 2, & \text{for all } x \leq 1 - \theta \\ 1, & \text{for all } x > 1 - \theta. \end{cases}$$

It is interesting to note that $\pi_{\hat{f}^*}$ is optimal for this family of bandit problems, that is $\pi_{f^*} \equiv \pi_{\hat{f}^*}$. Now let $\hat{f}$ be the best predictor of arm rewards under the distribution induced by $\pi_{\hat{f}^*}$. That is:

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \mathbb{E}_{x \sim D_{\mathcal{X}}}[(f(x, \pi_{\hat{f}^*}(x)) - f^*(x, \pi_{\hat{f}^*}(x)))^2].$$

43

Since $f^*(x, 1) = 1$ for all $x > 1 - \theta$, we get that $\hat{f}(\cdot, 1) \equiv 1$. Also since $f^*(\cdot, 2)$ is linear and arm 2 is chosen for all $x \leq 1 - \theta$, we get that $\hat{f}(\cdot, 2) \equiv f^*(\cdot, 2)$. For a more visual understanding, see Figure 6 which plots f^* and $\hat{f}$ for $\theta = 0.05$.
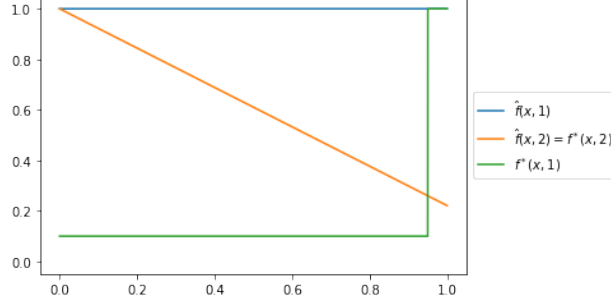

Figure 6: This is a plot of the conditional expected reward (f^*) and the linear estimate $(\hat{f})$ that is learnt from data collected by $\pi_{\hat{f}*}$. Note that $\pi_{\hat{f}*}$ is infact the same as the optimal policy π_{f^*}. Also note that the policy $\pi_{\hat{f}}$ that is induced by the estimate $\hat{f}$ samples arm 1 for all x. Hence, this policy has high regret.

Therefore $\pi_{\hat{f}}(x) = 1$ for all x, and hence incurs high regret:

$$\text{Reg}(\pi_{\hat{f}}) \geq \frac{1}{2}(1 - \theta)(1 - 0.1) \geq 0.4275.$$

For this family of bandit problems, while the regret of $\pi_{\hat{f}}$ is at least 0.4275, the approximation error (b) can be arbitrarily small. In particular, since $f^*(\cdot, 2)$ is linear, we get:

$$b = \min_{f \in \mathcal{F}} \frac{1}{2} \mathbb{E}_{x \sim D_{\mathcal{X}}} [(f(x, 1) - f^*(x, 1))^2] \leq \frac{1}{2} \mathbb{E}_{x \sim D_{\mathcal{X}}} [(0.1 - f^*(x, 1))^2] \leq \frac{\theta}{2}.$$

Further note that for this family of problems, as sufficient data is collected from policy $\pi_{\hat{f}*}$ (which is also optimal), the confidence intervals used by LinUCB tightly concentrate around $\hat{f}$.

Hence even under minor violations of realizability (the approximation error b of the best linear estimator can be arbitrarily small), the confidence intervals that are used by LinUCB are invalid, in the sense that this confidence interval tightly concentrates on a bad linear estimate $(\hat{f})$ that induces a policy $(\pi_{\hat{f}})$ with high regret $(\text{Reg}(\pi_{\hat{f}}) > 0.4275)$. Note that a similar argument can be used to argue that for this family of bandit problems, given data from the optimal policy, the posterior of LinTS concentrates on the same bad linear estimate. Similarly for this family of bandit problems, given data from the optimal policy, the empirical risk minimizer would be the bad linear estimate $\hat{f}$ and the induced randomized policy constructed by FALCON would converge to the high regret policy $(\pi_{\hat{f}})$ induced by this estimate. This example calls into question the validity of any model update step in realizability-based approaches.