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

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

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 $\pi_f(x) := \arg\max_a f(x, a)$ 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

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

When there is no possibility of confusion, we will write $\mathrm{Reg}(\pi)$ to mean $\mathrm{Reg}_{f^*}(\pi)$, 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 actions 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 actions 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 $\eta > C \ln^{\rho'}(n) \ln(1/\zeta)\textbf{comp}(\mathcal{F})/n^{\rho}$, 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

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

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 $x_i \sim \mathcal{D}_{\mathcal{X}}$, $a_i|x_i \sim p(\cdot|x_i)$ and $r_i \sim \mathcal{D}_{r_i|x_i,a_i}$, and the expectations are taken with respect to these distributions.

A.3 Algorithm

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 $\tau_0 = 0$, $\tau_1 \ge 4$ and for subsequent epochs we set $\tau_{m+1} = 2\tau_m$.

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

$$\gamma_m = \sqrt{\frac{C_3 K(\tau_{m-1} - \tau_{m-2})^{\rho}}{\ln^{\rho'}(\tau_{m-1} - \tau_{m-2}) \ln((m-1)/\delta)\textbf{comp}(\mathcal{F})}}. \tag{22}$$

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

$$\mathcal{F}'_m := \left\{ f \in \mathcal{F} \;\middle|\; \frac{1}{|S'_m|}\sum_{S'_m} (f_{m+1}(x,a) - r(a))^2 \le \alpha_m + \frac{C_1 \ln^{\rho'}(|S'_m|) \ln(1/\delta')\textbf{comp}(\mathcal{F})}{|S'_m|^{\rho}} \right\}, \tag{23}$$

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).

[7] Lemma 1 bounds B with Kb.

Algorithm 1 Epsilon-FALCON

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

Set $\tau_0 = 0$, and $\tau_{m+1} = 2\tau_m$ for all $m \geq 1$.
Let $\hat{f}_1 \equiv 0$.
for epoch $m = 1, 2, \ldots$ **do**
  Let $\gamma_m = \sqrt{\frac{C_3 K (\tau_{m-1} - \tau_{m-2})^{\rho}}{\ln^{\rho'}(\tau_{m-1} - \tau_{m-2}) \ln((m-1)/\delta) \mathbf{comp}(\mathcal{F})}}$ (for epoch 1, $\gamma_1 = 1$).
  for round $t = \tau_{m-1} + 1, \ldots, \tau_m - \lceil \epsilon(\tau_m - \tau_{m-1}) \rceil$ **do**
    Observe context x_t, let $\hat{a}_t = \arg\max_{a \in \mathcal{A}} \hat{f}_m(x_t, a)$, and define:

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

    Sample $a_t \sim p_t(\cdot)$ and observe $r_t(a_t)$.
  end for
  for round $t = \tau_m - \lceil \epsilon(\tau_m - \tau_{m-1}) \rceil + 1, \ldots, \tau_m$ **do**
    Observe context x_t, sample a_t uniformly at random from $\mathcal{A}$, and observe $r_t(a_t)$.
  end for
  Let:

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

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

$$\begin{aligned} \min_{f \in \mathcal{F}} \quad & \sum_{(x,a,r(a)) \in S_m} (f(x,a) - r(a))^2 \\ \text{s.t.} \quad & f \in \mathcal{F}'_m. \end{aligned} \tag{24}$$

  where $\mathcal{F}'_m$ is defined as in (23).
end for

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,

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

  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,

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

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

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

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

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

- 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:

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

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

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

- 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:*

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

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

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

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

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

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

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

□

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:

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

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

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:*

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

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:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_t,r_t,a_t}[r_t(\pi^*(x)) - r_t(a_t)|\Gamma_{t-1}] \\
&= \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}},a\sim p_m(\cdot|x)}[f^*(x,\pi^*)) - f^*(x,a)] \\
&= \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[\sum_{a\in\mathcal{A}} p_m(a|x)(f^*(x,\pi^*)) - f^*(x,a))\right] \\
&= \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[\sum_{a\in\mathcal{A}}\sum_{\pi\in\Psi} \mathbb{I}(\pi(x)=a)Q_m(\pi)(f^*(x,\pi^*)) - f^*(x,a))\right] \\
&= \sum_{\pi\in\Psi} Q_m(\pi) \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[(f^*(x,\pi^*)) - f^*(x,\pi(x)))\right] \\
&= \sum_{\pi\in\Psi} Q_m(\pi)\mathrm{Reg}(\pi).
\end{aligned}$$

□

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:

$$\begin{aligned}
&\sum_{\pi\in\Psi} Q_m(\pi)\mathrm{Reg}_{\hat{f}_m}(\pi) = \sum_{\pi\in\Psi} Q_m(\pi) \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[\hat{f}_m(x,\pi_{\hat{f}_m}(x)) - \hat{f}_m(x,\pi(x))\right] \\
&= \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[\sum_{\pi\in\Psi} Q_m(\pi)\Big(\hat{f}_m(x,\pi_{\hat{f}_m}(x)) - \hat{f}_m(x,\pi(x))\Big)\right] \\
&= \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[\sum_{a\in\mathcal{A}}\sum_{\pi\in\Psi} \mathbb{I}(\pi(x)=a)Q_m(\pi)\Big(\hat{f}_m(x,\pi_{\hat{f}_m}(x)) - \hat{f}_m(x,a)\Big)\right] \\
&= \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[\sum_{a\in\mathcal{A}} p_m(a|x)\Big(\hat{f}_m(x,\pi_{\hat{f}_m}(x)) - \hat{f}_m(x,a)\Big)\right] \\
&= \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[\sum_{a\in\mathcal{A}} \frac{\Big(\hat{f}_m(x,\pi_{\hat{f}_m}(x)) - \hat{f}_m(x,a)\Big)}{K + \gamma_m\Big(\hat{f}_m(x,\pi_{\hat{f}_m}(x)) - \hat{f}_m(x,a)\Big)}\right] \leq \frac{K}{\gamma_m}.
\end{aligned}$$

□

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:

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

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

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 $m\ge 1$, we have:*

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

Proof. Consider any policy $\pi\in\Psi$ and epoch $m\ge 1$. 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:

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

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

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

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

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 $m\ge 1$:*

$$\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\mathop{\mathbb{E}}_{a\sim Unif(\mathcal{A})}[(\hat{f}_{m+1}(x,a)-\hat{f}^*(x,a))^2] \le \frac{C_4\ln^{\rho'}(\tau_m-\tau_{m-1})\ln(m/\delta)\boldsymbol{comp}(\mathcal{F})}{(\epsilon(\tau_m-\tau_{m-1}))^{\rho}}.$$
$$\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\mathop{\mathbb{E}}_{a\sim p_m(\cdot|x)}[(\hat{f}_{m+1}(x,a)-f^*(x,a))^2] \le B + \frac{C_5\ln^{\rho'}(\tau_m-\tau_{m-1})\ln(m/\delta)\boldsymbol{comp}(\mathcal{F})}{(\tau_m-\tau_{m-1})^{\rho}}.$$

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:

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

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 actions 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:

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

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

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

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

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

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

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

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

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 actions according to the action selection kernel p_m. Hence from Assumption 1, with probability at least $1-\delta'$, we get that:

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

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:

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

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

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:

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

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:

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

□

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

$$\begin{aligned}
\mathcal{W}:=\Bigg\{&\forall m\geq 1,\ \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\mathop{\mathbb{E}}_{a\sim\mathrm{Unif}(\mathcal{A})}[(\hat{f}_{m+1}(x,a)-\hat{f}^*(x,a))^2]\leq\frac{C_4\ln^{\rho'}(\tau_m-\tau_{m-1})\ln(m/\delta)\mathbf{comp}(\mathcal{F})}{(\epsilon(\tau_m-\tau_{m-1}))^{\rho}},\\
&\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\mathop{\mathbb{E}}_{a\sim p_m(\cdot|x)}[(\hat{f}_{m+1}(x,a)-f^*(x,a))^2]\leq B+\frac{C_5\ln^{\rho'}(\tau_m-\tau_{m-1})\ln(m/\delta)\mathbf{comp}(\mathcal{F})}{(\tau_m-\tau_{m-1})^{\rho}}\Bigg\},
\end{aligned}\tag{28}$$

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:

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

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:

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

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

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 $p_1(\cdot|x)$ draws actions 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:

$$\begin{aligned}
&\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, \pi_{\hat{f}^*}(x)) \\
&= \big(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}^*(x, \pi_{\hat{f}_m}(x)) \big) + \big(\hat{f}^*(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, \pi_{\hat{f}^*}(x)) \big) \\
&\leq \big(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}^*(x, \pi_{\hat{f}_m}(x)) \big) + \big(\hat{f}^*(x, \pi_{\hat{f}^*}(x)) - \hat{f}_m(x, \pi_{\hat{f}^*}(x)) \big) \\
&\leq \big| \hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}^*(x, \pi_{\hat{f}_m}(x)) \big| + \big| \hat{f}^*(x, \pi_{\hat{f}^*}(x)) - \hat{f}_m(x, \pi_{\hat{f}^*}(x)) \big| \\
&\leq 2 \max_{a\in\mathcal{A}} \big| \hat{f}^*(x, a) - \hat{f}_m(x, a) \big|.
\end{aligned}$$

Now from Lemma 6, the above inequality, and Jensen's inequality, we get:

$$\begin{aligned}
V(p_m, \pi_{\hat{f}^*}) &\leq \mathop{\mathbb{E}}_{x\sim\mathcal{X}}\Big[K + \gamma_m\big(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, \pi_{\hat{f}^*}(x))\big)\Big] \\
&\leq K + 2\gamma_m \mathop{\mathbb{E}}_{x\sim\mathcal{X}}\Big[\max_{a\in\mathcal{A}} |\hat{f}^*(x,a) - \hat{f}_m(x,a)|\Big] \\
&\leq K + 2\gamma_m \sqrt{\mathop{\mathbb{E}}_{x\sim\mathcal{X}}\Big[\max_{a\in\mathcal{A}} |\hat{f}^*(x,a) - \hat{f}_m(x,a)|^2\Big]} \\
&\leq K + 2\gamma_m \sqrt{\sum_{a\in\mathcal{A}} \mathop{\mathbb{E}}_{x\sim\mathcal{X}}\Big[\big(\hat{f}_m(x,a) - \hat{f}^*(x,a)\big)^2\Big]}.
\end{aligned}$$

Now let $C_6 = 1 + 2\sqrt{C_3 C_4}$. From the above inequality, we further get:

$$\begin{aligned}
V(p_m, \pi_{\hat{f}^*}) &\leq K + 2\gamma_m \sqrt{\sum_{a\in\mathcal{A}} \mathop{\mathbb{E}}_{x\sim\mathcal{X}}\Big[\big(\hat{f}_m(x,a) - \hat{f}^*(x,a)\big)^2\Big]} \\
&\leq K + 2\gamma_m \sqrt{K \frac{C_4 \ln^{\rho'}(\tau_{m-1} - \tau_{m-2}) \ln((m-1)/\delta)\mathbf{comp}(\mathcal{F})}{(\epsilon(\tau_{m-1} - \tau_{m-2}))^\rho}} \\
&= K + 2K\sqrt{\frac{C_3 C_4}{\epsilon^\rho}} \leq \frac{C_6 K}{\sqrt{\epsilon^\rho}}.
\end{aligned}$$

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 $p_1(\cdot|x)$ draws actions 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:

$$\begin{aligned}
&V(p_m, \pi) - K - V(p_{m+1}, \pi) \\
&\leq \gamma_m \mathop{\mathbb{E}}_{x\sim D_\mathcal{X}}[\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_m(x, \pi(x))] - \gamma_{m+1} \mathop{\mathbb{E}}_{x\sim D_\mathcal{X}}[\hat{f}_{m+1}(x, \pi_{\hat{f}_{m+1}}(x)) - \hat{f}_{m+1}(x, \pi(x))] \\
&\leq \gamma_m \mathop{\mathbb{E}}_{x\sim D_\mathcal{X}}[(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_{m+1}(x, \pi_{\hat{f}_{m+1}}(x))) + (\hat{f}_{m+1}(x, \pi(x)) - \hat{f}_m(x, \pi(x)))] \\
&\leq \gamma_m \mathop{\mathbb{E}}_{x\sim D_\mathcal{X}}[(\hat{f}_m(x, \pi_{\hat{f}_m}(x)) - \hat{f}_{m+1}(x, \pi_{\hat{f}_m}(x))) + (\hat{f}_{m+1}(x, \pi(x)) - \hat{f}_m(x, \pi(x)))] \\
&\leq 2\gamma_m \mathop{\mathbb{E}}_{x\sim D_\mathcal{X}}\Big[\max_a |\hat{f}_{m+1}(x,a) - \hat{f}_m(x,a)|\Big].
\end{aligned} \tag{29}$$

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

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[\max_a |\hat{f}_{m+1}(x,a)-\hat{f}_m(x,a)|\right] \\
&\leq \sqrt{\mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[\max_a |\hat{f}_{m+1}(x,a)-\hat{f}_m(x,a)|^2\right]} \\
&\leq \sqrt{\sum_a \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[(\hat{f}_{m+1}(x,a)-\hat{f}_m(x,a))^2\right]} \\
&= \sqrt{K \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a\sim \mathrm{Unif}(\mathcal{A})}\left[(\hat{f}_{m+1}(x,a)-\hat{f}_m(x,a))^2\right]} \\
&\leq \sqrt{2K \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a\sim \mathrm{Unif}(\mathcal{A})}\left[(\hat{f}_{m+1}(x,a)-\hat{f}^*(x,a))^2+(\hat{f}^*(x,a)-\hat{f}_m(x,a))^2\right]} \\
&\leq \sqrt{\frac{4KC_4 \ln^{\rho'}(\tau_m-\tau_{m-1})\ln(m/\delta)\mathbf{comp}(\mathcal{F})}{(\epsilon(\tau_{m-1}-\tau_{m-2}))^{\rho}}}.
\end{aligned} \tag{30}$$

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:

$$\begin{aligned}
V(p_m,\pi) &\leq V(p_{m+1},\pi) + K + 4\gamma_m \sqrt{\frac{KC_4 \ln^{\rho'}(\tau_m-\tau_{m-1})\ln(m/\delta)\mathbf{comp}(\mathcal{F})}{(\epsilon(\tau_{m-1}-\tau_{m-2}))^{\rho}}} \\
&= V(p_{m+1},\pi) + K + 4K\sqrt{\frac{C_3C_4}{\epsilon^{\rho}} \frac{\ln^{\rho'}(\tau_m-\tau_{m-1})}{\ln^{\rho'}(\tau_{m-1}-\tau_{m-2})} \frac{\ln(m/\delta)}{\ln((m-1)/\delta)}} \\
&\leq V(p_{m+1},\pi) + \frac{C_7K}{\sqrt{\epsilon^{\rho}}}.
\end{aligned}$$

Where the last inequality follows from choosing $C_7 = 1 + 4\sqrt{2^{1+\rho'}C_3C_4}$, and from the fact that for $m \geq 2$ and $\delta \leq 0.5$ we have: $\frac{\ln(m/\delta)}{\ln((m-1)/\delta)} \leq 2$, and $\frac{\ln^{\rho'}(\tau_m-\tau_{m-1})}{\ln^{\rho'}(\tau_{m-1}-\tau_{m-2})} \leq \frac{\ln^{\rho'}(\tau_{m-1})}{\ln^{\rho'}(\tau_{m-1}/2)} \leq 2^{\rho'}$. □

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 $\delta \leq 0.5$. Then there exists positive constants C_0, C_8, C_9 such that, for all policies π and epochs m, we have:*

$$\begin{aligned}
Reg(\pi) &\leq 2Reg_{\hat{f}_m}(\pi) + \frac{C_0K}{\gamma_m} + C_8\sqrt{\frac{KB}{\sqrt{\epsilon^{\rho}}}} + C_9\sqrt{V(p_m,\pi)B} \\
Reg_{\hat{f}_m}(\pi) &\leq 2Reg(\pi) + \frac{C_0K}{\gamma_m} + C_8\sqrt{\frac{KB}{\sqrt{\epsilon^{\rho}}}} + C_9\sqrt{V(p_m,\pi)B}
\end{aligned}$$

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

$$\begin{aligned}
\mathrm{Reg}(\pi) &\leq 1 \leq C_0K/\gamma_1 \\
\mathrm{Reg}_{\hat{f}_1}(\pi) &\leq 1 \leq C_0K/\gamma_1.
\end{aligned}$$

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

$$\begin{aligned}
\text{Reg}(\pi) &\leq 2\text{Reg}_{\hat{f}_m}(\pi) + \frac{C_0 K}{\gamma_m} + C_8\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + C_9\sqrt{V(p_m,\pi)B} \\
\text{Reg}_{\hat{f}_m}(\pi) &\leq 2\text{Reg}(\pi) + \frac{C_0 K}{\gamma_m} + C_8\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + C_9\sqrt{V(p_m,\pi)B}
\end{aligned} \tag{31}$$

Note that:

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

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:

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

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:

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

Where the first inequality follows from Lemma 6, the second inequality follows from Equation (31), and the

last inequality follows from Lemmas 2 and 9. Similarly note that:

$$\begin{aligned}
\frac{V(p_m,\pi)}{5\gamma_{m+1}} &\leq \frac{K+\gamma_m \mathrm{Reg}_{\hat{f}_m}(\pi)}{5\gamma_{m+1}} \\
&\leq \frac{K+\gamma_m\Big(2\mathrm{Reg}(\pi)+\frac{C_0K}{\gamma_m}+C_8\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}}+C_9\sqrt{V(p_m,\pi)B}\Big)}{5\gamma_{m+1}} \\
&\leq \frac{K(1+C_0)}{5\gamma_{m+1}}+\frac{2\mathrm{Reg}(\pi)}{5}+\frac{C_8}{5}\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}}+\frac{C_9}{5}\sqrt{V(p_m,\pi)B} \\
&\leq \frac{K(1+C_0)}{5\gamma_{m+1}}+\frac{2\mathrm{Reg}(\pi)}{5}+\frac{C_8+C_9\sqrt{C_7}}{5}\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}}+\frac{C_9}{5}\sqrt{V(p_{m+1},\pi)B}. \qquad (34)
\end{aligned}$$

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:

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

Which implies:

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

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:

$$\mathrm{Reg}(\pi) \leq 2\mathrm{Reg}_{\hat{f}_{m+1}}(\pi)+\frac{C_0K}{\gamma_{m+1}}+C_8\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}}+C_9\sqrt{V(p_{m+1},\pi)B}. \qquad (35)$$

Hence from our induction hypothesis (Equation (31)), we get Equation (35), which provides the required upper bound on $\mathrm{Reg}(\pi)$ in terms of $\mathrm{Reg}_{\hat{f}_{m+1}}(\pi)$. To complete the inductive argument, we need to show the corresponding upper bound on $\mathrm{Reg}_{\hat{f}_{m+1}}(\pi)$. Similar to Equation (32), we get:

$$\begin{aligned}
&\mathrm{Reg}_{\hat{f}_{m+1}}(\pi)-\mathrm{Reg}(\pi) \\
=&\Big(R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}})-R_{\hat{f}_{m+1}}(\pi)\Big)-\Big(R(\pi_{f^*})-R(\pi)\Big) \\
\leq&\Big(R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}})-R_{\hat{f}_{m+1}}(\pi)\Big)-\Big(R(\pi_{\hat{f}_{m+1}})-R(\pi)\Big) \\
\leq&|R(\pi_{\hat{f}_{m+1}})-R_{\hat{f}_{m+1}}(\pi_{\hat{f}_{m+1}})|+|R(\pi)-R_{\hat{f}_{m+1}}(\pi))| \\
\leq&\sqrt{V(p_m,\pi_{\hat{f}_{m+1}})}\sqrt{B}+\frac{\sqrt{V(p_m,\pi_{\hat{f}_{m+1}})}\sqrt{K}}{2\gamma_{m+1}}+\sqrt{V(p_m,\pi)}\sqrt{B}+\frac{\sqrt{V(p_m,\pi)}\sqrt{K}}{2\gamma_{m+1}} \\
\leq&\frac{5K}{8\gamma_{m+1}}+\frac{V(p_m,\pi_{\hat{f}_{m+1}})}{5\gamma_{m+1}}+\frac{V(p_m,\pi)}{5\gamma_{m+1}}+\sqrt{B}\Big(\sqrt{V(p_m,\pi_{\hat{f}_{m+1}})}+\sqrt{V(p_m,\pi)}\Big) \\
\leq&\frac{5K}{8\gamma_{m+1}}+\frac{V(p_m,\pi_{\hat{f}_{m+1}})}{5\gamma_{m+1}}+\frac{V(p_m,\pi)}{5\gamma_{m+1}}+2\sqrt{\frac{C_7BK}{\sqrt{\epsilon^\rho}}}+\sqrt{B}\Big(\sqrt{V(p_{m+1},\pi_{\hat{f}_{m+1}})}+\sqrt{V(p_{m+1},\pi)}\Big). \qquad (36)
\end{aligned}$$

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:

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

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:

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

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

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

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:

$$\text{Reg}_{\hat{f}_{m+1}}(\pi) \leq 2\text{Reg}(\pi) + \frac{C_0 K}{\gamma_{m+1}} + C_8\sqrt{\frac{KB}{\sqrt{\epsilon^\rho}}} + C_9\sqrt{V(p_{m+1}, \pi)B}. \qquad (39)$$

This completes the inductive step. □

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:

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

Where the first inequality follows from Lemma 11, and the second inequality follows from Lemmas 4 and Lemma 5. □

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:

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

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:

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

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:

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

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:

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

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

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

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:

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

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:

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

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:

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

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:

$$\begin{aligned} U_n(\delta) &:= \phi_n(\delta_j) + \sqrt{2\frac{t_j}{n}(D^2(\delta_j) + 2\phi_n(\delta_j))} + \frac{t_j}{2n} \\ &= \phi_n(\delta_j) + \sqrt{2\frac{t}{n}\frac{\delta_j}{\sigma}(D^2(\delta_j) + 2\phi_n(\delta_j))} + \frac{t}{2n}\frac{\delta_j}{\sigma} \end{aligned}$$

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 $\eta > 0$:

$$\begin{aligned} U_n^\flat(\eta) &\leq \sup_{\delta_j \geq \eta} \frac{q}{\delta_j}\left\{\phi_n(\delta_j) + \sqrt{2\frac{t}{n}\frac{\delta_j}{\sigma}(D^2(\delta_j) + 2\phi_n(\delta_j))} + \frac{t}{2n}\frac{\delta_j}{\sigma}\right\} \\ &\leq q \sup_{\delta_j \geq \eta} \frac{\phi_n(\delta_j)}{\delta_j} + \sup_{\delta_j \geq \eta} q\left\{\sqrt{\frac{2t}{\sigma n}\frac{D^2(\delta_j)}{\delta_j}} + \sqrt{\frac{4t}{\sigma n}\frac{\phi_n(\delta_j)}{\delta_j}}\right\} + \frac{qt}{2\sigma n} \\ &\leq q\phi_n^\flat(\eta) + q\sqrt{\frac{2t}{\sigma n}(D^2)^\flat(\eta)} + q\sqrt{\frac{4t}{\sigma n}\phi_n^\flat(\eta)} + \frac{qt}{2\sigma n} \end{aligned} \tag{46}$$

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

$$\begin{aligned} &U_n^\sharp(\epsilon) := \inf\{\eta > 0 \mid U_n^\flat(\eta) \leq \epsilon\} \\ &\leq \inf\left\{\eta > 0 \mid q\phi_n^\flat(\eta) + q\sqrt{\frac{2t}{\sigma n}(D^2)^\flat(\eta)} + q\sqrt{\frac{4t}{\sigma n}\phi_n^\flat(\eta)} + \frac{qt}{2\sigma n} \leq \epsilon\right\} \\ &\leq \inf\left\{\eta > 0 \mid \phi_n^\flat(\eta) + \sqrt{\frac{2t}{\sigma n}(D^2)^\flat(\eta)} + \sqrt{\frac{4t}{\sigma n}\phi_n^\flat(\eta)} \leq \frac{1}{q}\left(\epsilon - \frac{qt}{2\sigma n}\right)\right\} \\ &\leq \max\left\{\inf\left\{\eta > 0 \mid \phi_n^\flat(\eta) \leq \frac{1}{3q}\left(\epsilon - \frac{qt}{2\sigma n}\right)\right\}, \inf\left\{\eta > 0 \mid \sqrt{\frac{2t}{\sigma n}(D^2)^\flat(\eta)} \leq \frac{1}{3q}\left(\epsilon - \frac{qt}{2\sigma n}\right)\right\},\right. \\ &\qquad \left.\inf\left\{\eta > 0 \mid \sqrt{\frac{4t}{\sigma n}\phi_n^\flat(\eta)} \leq \frac{1}{3q}\left(\epsilon - \frac{qt}{2\sigma n}\right)\right\}\right\} \\ &\leq \max\left\{\phi_n^\sharp\left(\frac{\epsilon}{3q} - \frac{t}{6\sigma n}\right), (D^2)^\sharp\left(\frac{\sigma n}{2t}\left(\frac{\epsilon}{3q} - \frac{t}{6\sigma n}\right)^2\right), \phi_n^\sharp\left(\frac{\sigma n}{4t}\left(\frac{\epsilon}{3q} - \frac{t}{6\sigma n}\right)^2\right)\right\} \end{aligned} \tag{47}$$

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 $(D^2)^\flat(\cdot)$:

$$\begin{aligned} &D(\delta) \leq 4\sqrt{2}\sqrt{\delta}, \quad \text{for all } \delta \geq 0. \\ \implies &(D^2)^\flat(\eta) = \sup_{\delta' \geq \eta} \frac{D^2(\delta')}{\delta'} \leq 32, \quad \text{for all } \eta \geq 0. \end{aligned}$$

Hence we have:

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

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:

$$\begin{aligned}\sigma \geq \frac{4096tq^2}{n} = 1024\frac{tq^2}{n(1/2)^2} \implies \frac{\sigma n}{2t}\frac{(1/2)^2}{16q^2} \geq 32 \implies \frac{\sigma n}{2t}\left(\frac{1/2}{3q} - \frac{t}{6\sigma n}\right)^2 \geq 32 \\ \implies (D^2)^\sharp\left(\frac{\sigma n}{2t}\left(\frac{1/2}{3q} - \frac{t}{6\sigma n}\right)^2\right) = 0\end{aligned} \tag{49}$$

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 $\phi_n^\sharp(\cdot)$:

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

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:

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

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

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

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:

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

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 $\hat{\mathcal{G}}^l(\delta) \subset \mathcal{G}^l(2\delta)$. □

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:

$$\begin{aligned}\psi_n(\delta) &= 16 \mathop{\mathbb{E}}_{P,\epsilon} \sup\{|R_n(g-\hat{g}^*)| \mid g \in \mathcal{G}, \rho_P^2(g,\hat{g}^*) \leq 2\delta\} \\ &\leq 16\sqrt{2\delta}\sqrt{\frac{d}{n}}.\end{aligned}$$

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:

$$\begin{aligned}\psi_n^\sharp(\epsilon) &= \inf\{\delta > 0 \mid \psi_n^\flat(\delta) \leq \epsilon\} \\ &\leq \inf\left\{\delta > 0 \mid 16\sqrt{\frac{2d}{n\delta}} \leq \epsilon\right\} \\ &= \inf\left\{\delta > 0 \mid \frac{512d}{n\epsilon^2} \leq \delta\right\} = \frac{512d}{n\epsilon^2}.\end{aligned}$$

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. □

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 [9]. 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:

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

[9] See proposition 1.1.3 in [Bertsekas and Scientific, 2015].

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 [10]. 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 $g(\cdot)$.

Set $\lambda_L = 0$, $\lambda_M = 1$, and $\lambda_R = 2$.
while $g(\lambda_M) < g(\lambda_R)$ **do**
    Set $\lambda_R \leftarrow 2\lambda_R$ and set $\lambda_M \leftarrow 2\lambda_M$.
end while
while $|\lambda_R - \lambda_L| \geq \kappa$ **do**
    if $g(\lambda_M + \kappa) > g(\lambda_M)$ **then**
        Set $\lambda_L \leftarrow \lambda_M$.
    else
        Set $\lambda_R \leftarrow \lambda_M$.
    end if
    Set $\lambda_M \leftarrow \frac{1}{2}(\lambda_L + \lambda_R)$.
end while
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.

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

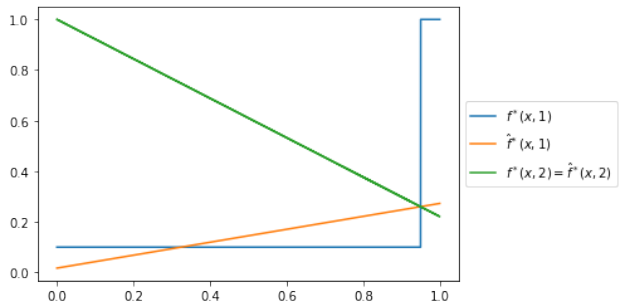

Figure 5: This is a plot of the conditional expected reward (f^*) and the best linear estimate $\hat{f}^*$ when actions are sampled uniformly at random. Note that the conditional expected reward for arm 2 is linear. The problem is constructed so that the policy ($\pi_{\hat{f}^*}$) that is induced by 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.

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 actions. 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 actions 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:

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

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 $\theta = 0.05$.

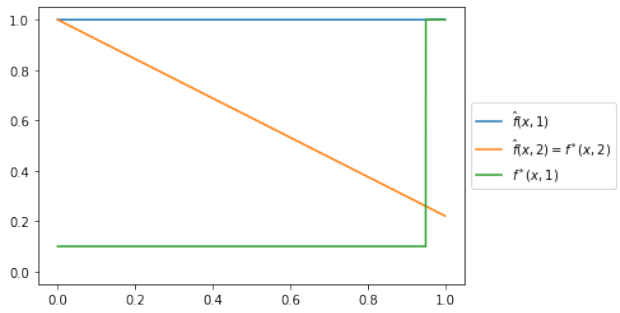

Figure 6: This is a plot of the conditional expected reward (f^*) and the linear estimate ($\hat{f}$) that is learnt from data collected by $\pi_{\hat{f}^*}$. Note that $\pi_{\hat{f}^*}$ is infact the same as the optimal policy π_{f^*}. Also note that the policy $\pi_{\hat{f}}$ that is induced by the 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:

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

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:

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

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 ($\mathrm{Reg}(\pi_{\hat{f}}) > 0.4275$). 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.