A. Outline

In Appendix A.1, we establish additional notation that will useful for our proofs. We detail the proofs for the upper and lower bounds in Appendix B and Appendix C respectively. Finally in Appendix D we explain the introductory example in more detail and provide some more intuition.

A.1. Preliminaries

Most of the notation and definitions described below will be the same as in (Simchi-Levi & Xu, 2020).

A "policy" is a deterministic mapping from contexts to actions. Let $\Psi = \mathcal{A}^{\mathcal{X}}$ be the universal policy space containing all possible policies. The expected instantaneous reward of the policy π with respect to the model f is defined as

$$R_f(\pi) := \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}[f(x, \pi(x))]. \tag{25}$$

Recalling that f^* is the true conditional means of rewards, we write $R(\pi)$ to mean $R_{f^*}(\pi)$, the true expected instantaneous reward for policy π. The policy π_f that is induced by model f is given by $\pi_f(x) := \arg\max_a f(x, a)$ for every x. This policy has the highest instantaneous reward with respect to the model f, that is $\pi_f = \arg\max_{\pi\in\Psi} R_f(\pi)$.

The expected instantaneous regret of a policy π with respect to the outcome model f is defined as

$$\mathrm{Reg}_f(\pi) := \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}[f(x, \pi_f(x)) - f(x, \pi(x))]. \tag{26}$$

We write $\mathrm{Reg}(\pi)$ to mean $\mathrm{Reg}_{f^*}(\pi)$, the true expected instantaneous regret for policy π. We also let Γ_t denote the set of observations up to and including time t. That is

$$\Gamma_t := \{(x_s, a_s, r_s(a_s))\}_{s=1}^t \tag{27}$$

Given any probability kernel p from $\mathcal{A}\times\mathcal{X}$ to $[0, 1]$, from Lemma 3 in (Simchi-Levi & Xu, 2020), there exists a unique product probability measure on Ψ, given by:

$$Q_p(\pi) := \prod_{x\in\mathcal{X}} p(\pi(x)|x). \tag{28}$$

This measure satisfies the following property

$$p(a|x) = \sum_{\pi\in\Psi} \mathbb{I}\{\pi(x) = a\} Q_p(\pi). \tag{29}$$

Since any probability kernel p from $\mathcal{A}\times\mathcal{X}$ to $[0, 1]$ induces the distribution Q_p over the set of deterministic policies Ψ, we can think of Q_p as a randomized policy induced by p. Equations (29) and (28) establish a correspondence between the probability kernel p and the induced randomized policy Q_p. For any probability kernel p and any policy π, we let $V(p, \pi)$ denote the expected inverse probability. [13]

$$V(p, \pi) := \mathop{\mathbb{E}}_{x\sim D_{\mathcal{X}}}\left[\frac{1}{p(\pi(x)|x)}\right] \tag{30}$$

B. Proof for upper bound

In this section we prove Theorem 1, we start by establishing some more additional notation. We say an epoch m is safe when the status at the end of the epoch is safe, that is the variable **safe** is still set to **True** at the end of this epoch. Let $\hat{m}$ be the last safe epoch, that is:

$$\hat{m} := \max\left\{m \mid \text{ the epoch } m \text{ is safe}\right\}. \tag{31}$$

[13] In (Simchi-Levi & Xu, 2020), this term is called the decisional divergence between the randomized policy Q_p and deterministic policy π.

Note that, for all $m \leq \hat{m}$, the epoch m is safe. Now let m^* be such that:

$$m^* := \max \left\{ m \mid B \leq \xi\Big(\tau_m - \tau_{m-1}, \frac{\delta'}{m^2}\Big) \right\}. \tag{32}$$

Where $\delta' = \delta/13$. As discussed in Section 2.3, the epoch m^* is critical to our theoretical analysis and we will show that with high-probability $m^* + 1$ is safe. We also let $\mathcal{T}$ be the set of time-steps where algorithm 2 checks the safety condition (see Check-is-safe).

$$\mathcal{T} := \left\{ t \in [T] | \text{ status is safe at the start of the time step, } t = \tau_{m(t)} \text{ or } \log_2(t - \tau_{m(t)-1}) \text{ is integral} \right\}. \tag{33}$$

For short hand, we let $Q_m \equiv Q_{p_m}$. With some abuse of notation, we let p_t denote the action selection kernel used at time-step t. Again with some abuse of notation, we let $Q_t \equiv Q_{p_t}$.

B.1. High probability events

Theorem 1 provides certain high probability bounds on cumulative regret. As a preliminary step in these proofs, it will be helpful to show that the events $\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3$ defined below hold with high-probability. Event $\mathcal{W}_1$ ensures that our regression estimates are "good" models for the first few epochs. Event $\mathcal{W}_2$ ensures that the lower bounds constructed by Choose-safe are valid, this event also includes symmetric upper bounds. Event $\mathcal{W}_3$ helps us show that the misspecification tests that we use in Check-is-safe are valid.

We start with the event $\mathcal{W}_1$. This describes the event where, for any epoch $m \in [m^*] \cap [\hat{m}]$, the expected squared error difference between the true model (f^*) and the estimated model ($\hat{f}_{m+1}$) can be bounded purely in terms of the known estimation rate of the regression algorithm.

$$\mathcal{W}_1 := \left\{ \forall m \in [m^*] \cap [\hat{m}], \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a \sim p_m(\cdot|x)} [(\hat{f}_{m+1}(x,a) - f^*(x,a))^2] \leq 2\xi\Big(\tau_m - \tau_{m-1}, \frac{\delta'}{m^2}\Big) \right\}. \tag{34}$$

Lemma 1. *Suppose the regression algorithm used in Safe-FALCON satisfies Assumption 1. Then the event $\mathcal{W}_1$ holds with probability at least $1 - 2\delta'$.*

Proof. Consider any epoch m such that epoch $m-1$ was safe. Since epoch $m-1$ was safe, for the first time-step in epoch m the algorithm samples an action from the probability kernel p_m. It may so happen that the status of the algorithm switches at the end of some time-step in epoch m and the algorithm no longer samples actions according to the kernel p_m. As long as the status of the algorithm does not switch in epoch m, for the purposes of estimating $\hat{f}_{m+1}$, we want to argue that the data collected in epoch m can be treated as iid samples from the distribution $D(p_m)$.

One way to see this is by considering the following thought experiment. At the start of epoch m, the environment generates $\tau_m - \tau_{m-1}$ data points that are iid sampled from the distribution $D(p_m)$. The environment then runs the regression algorithm on this data, and generates the model $\hat{f}_{m+1}$. The environment sequentially shows us these data points throughout epoch m, as long as the status of the algorithm is safe. Additionally, we also observe $\hat{f}_{m+1}$ at the end of epoch m if the status of the algorithm was safe throughout the epoch. Regardless of whether we observe the model $\hat{f}_{m+1}$, the environment constructs $\hat{f}_{m+1}$ by running the regression algorithm on iid samples from $D(p_m)$.

Further note that the kernel p_m lies in $\mathcal{K}(\mathcal{F})$. Hence from Assumption 1, with probability $1 - \delta'/m^2$, we have:

$$\mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a \sim p_m(\cdot|x)} [(\hat{f}_{m+1}(x,a) - f^*(x,a))^2] \leq B + \xi\Big(\tau_m - \tau_{m-1}, \frac{\delta'}{m^2}\Big). \tag{35}$$

Therefore, the probability that (35) does not hold for some epoch $m \in [\hat{m}]$ can be bounded by:

$$\sum_{m=1}^{\infty} \frac{\delta'}{m^2} \leq 2\delta'.$$

Hence from the definition of m^*, we get that $\mathcal{W}_1$ holds with probability at least $1 - 2\delta'$. □

Lemma 2. *For any time-step $t \geq 1$, we have:*

$$\mathop{\mathbb{E}}_{x_t, r_t, a_t}[r_t(a_t)|\Gamma_{t-1}] = \sum_{\pi \in \Psi} Q_t(\pi) R(\pi).$$

Proof. Consider any time-step $t \geq 1$, then from eq. (29) relating $p(\cdot|\cdot)$ and Q_p we have the following equalities:

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

□

The event $\mathcal{W}_2$ provides upper and lower bounds on the expected reward of the randomized policy Q_m, for all $m \in [\hat{m}]$.

$$\begin{aligned}
\mathcal{W}_2 := \Bigg\{ \forall m \in [\hat{m}],\ &\sum_{\pi \in \Psi} Q_m(\pi) R(\pi) \geq \frac{1}{|S_m|} \sum_{(x,a,r) \in S_m} r\ - \sqrt{\frac{1}{2|S_m|} \ln\left(\frac{m^2}{\delta'}\right)}, \\
&\sum_{\pi \in \Psi} Q_m(\pi) R(\pi) \leq \frac{1}{|S_m|} \sum_{(x,a,r) \in S_m} r\ + \sqrt{\frac{1}{2|S_m|} \ln\left(\frac{m^2}{\delta'}\right)} \Bigg\}.
\end{aligned} \tag{36}$$

Lemma 3. *The event $\mathcal{W}_2$ holds with probability at least $1 - 4\delta'$.*

Proof. Consider any safe epoch m. Similar to the argument in Lemma 1, for the purpose of estimating the expected reward of the randomized policy Q_m, we can treat the data generated in epoch m as iid samples from the distribution $D(p_m)$. Since rewards lie in the range $[0,1]$, from Hoeffding's inequality, with probability at least $1 - 2\delta'/m^2$, we get:

$$\begin{aligned}
&\sum_{\pi \in \Psi} Q_m(\pi) R(\pi) \geq \frac{1}{|S_m|} \sum_{(x,a,r) \in S_m} r\ - \sqrt{\frac{1}{2|S_m|} \ln\left(\frac{m^2}{\delta'}\right)}, \\
&\sum_{\pi \in \Psi} Q_m(\pi) R(\pi) \leq \frac{1}{|S_m|} \sum_{(x,a,r) \in S_m} r\ + \sqrt{\frac{1}{2|S_m|} \ln\left(\frac{m^2}{\delta'}\right)}.
\end{aligned}$$

Therefore, we get that these bounds hold for all $m \in [\hat{m}]$ with probability at least:

$$1 - \sum_{m=1}^{\infty} \frac{2\delta'}{m^2} \geq 1 - 4\delta'.$$

□

For all time-steps $t' \in \mathcal{T}$, the event $\mathcal{W}_3$ provides a lower bound on the cumulative reward up to time t' in terms of the expected cumulative reward.

$$\mathcal{W}_3 := \left\{ \forall t' \in [\mathcal{T}],\ \sum_{t=1}^{t'} \left(\sum_{\pi \in \Psi} Q_t(\pi) R(\pi) - r_t(a_t) \right) \leq \sqrt{2t' \ln\left(\frac{\lceil m(t') + \log_2(\tau_1) \rceil^3}{\delta'} \right)} \right\}. \tag{37}$$

Lemma 4. *The event $\mathcal{W}_3$ holds with probability at least $1-\delta'$.*

Proof. For any time-step $t \in [T]$, define:

$$M_t := \sum_{\pi \in \Psi} Q_t(\pi) R(\pi) - r_t(a_t).$$

From Lemma 2, we have that $\mathbb{E}[M_t | \Gamma_{t-1}] = 0$. Also note that for any time-step t, we have $|M_t| \leq 1$. Now consider any time-step $t' \in \mathcal{T}$. From Azuma's inequality, with probability $1 - \delta'/\lceil m(t') + \log_2(\tau_1)\rceil^3$, we get:

$$\sum_{t=1}^{t'} M_t \leq \sqrt{2t' \ln\left(\frac{(m(t') + \log_2(\tau_1))^3}{\delta'}\right)}.$$

Any epoch m has at most $\lceil m + \log_2(\tau_1)\rceil$ time-steps in $\mathcal{T}$. Therefore $\mathcal{W}_3$ holds with probability at least:

$$1 - \sum_{t' \in \mathcal{T}} \frac{\delta'}{\lceil m(t') + \log_2(\tau_1)\rceil^3} \geq 1 - \sum_{m=1}^{\infty} \frac{\delta'}{(m + \log_2(\tau_1))^2} \geq 1 - \delta'.$$

□

B.2. Adapting FALCON+

As we have stated before, both our algorithm and analysis builds on the work of (Simchi-Levi & Xu, 2020). In this section, without assuming realizability, we show that the analysis of FALCON+ continues to hold for the first few epochs of Safe-FALCON. The simple observation that allows us to make this extension is that Lemma 6 can be proved without assuming realizability.

Lemma 5 is more or less a restatement of Lemma 5 in (Simchi-Levi & Xu, 2020), and the proof stays the same. We only include the proof for completeness, as it states a key bound on the estimated regret of the randomized policy Q_m.

Lemma 5. *For any safe epoch m, we have:*

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

Proof. This follows essentially from unpacking the definitions of regret (26) and the representation of the action selection kernel p_m as in (9) and (29).

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

□

For any policy, Lemma 6 bounds the prediction error of the implicit reward estimate for the first few epochs. This Lemma and its proof are similar to those of Lemma 12 in (Simchi-Levi & Xu, 2020). Our definition of m^* and our choice of γ_{m+1} allows us to prove this Lemma without assuming realizability.

Lemma 6. *Suppose the event $\mathcal{W}_1$ defined in (34) holds. Then, for all policies π and epochs $m \leq \min\{m^*, \hat{m}\}$, we have:*

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

Proof. For any policy π and epoch $m \in [m^*]$, note that:

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

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 $\mathcal{W}_1$ from (34) holds. □

The next Lemma implies that before misspecification becomes a problem we are able to bound regret in the same manner as (Simchi-Levi & Xu, 2020). Note that for any epoch $m \leq \hat{m}$, the action selection kernel used in epoch m is given by (9). Further note that since the regression rates are valid (Assumption 1), from (6) and (10), we have that γ_m is increasing in m. Finally, since Lemma 6 holds for all $m \leq \min\{m^*, \hat{m}\}$, following the proof of Lemma 13 in (Simchi-Levi & Xu, 2020), we get:

Lemma 7. *Suppose the event $\mathcal{W}_1$ defined in (34) holds. Let $C_0 = 5.15$. For all policies π and epochs $m \leq \min\{m^*, \hat{m}\}+1$, we have:*

$$\begin{aligned}
\textit{Reg}(\pi) &\leq 2\textit{Reg}_{\hat{f}_m}(\pi) + \frac{C_0 K}{\gamma_m} \\
\textit{Reg}_{\hat{f}_m}(\pi) &\leq 2\textit{Reg}(\pi) + \frac{C_0 K}{\gamma_m}
\end{aligned}$$

That is, for any policy, Lemma 7 bounds the prediction error of the implicit regret estimate for the first few epochs. Lemma 8 bounds the expected regret of the randomized policy Q_m for the first few epochs. Lemma 8 and its proof is more or less the same as the statement and the proof of Lemma 9 in (Simchi-Levi & Xu, 2020).

Lemma 8. *Suppose the event $\mathcal{W}_1$ defined in (34) holds. Then for all epochs $m \leq \min\{m^*, \hat{m}\} + 1$, we have:*

$$\sum_{\pi\in\Psi} Q_m(\pi)\textit{Reg}(\pi) \leq \frac{(2 + C_0)K}{\gamma_m}.$$

Proof. For any $m \leq \min\{m^*, \hat{m}\}$:

$$\sum_{\pi\in\Psi} Q_m(\pi)\text{Reg}(\pi) \leq \sum_{\pi\in\Psi} Q_m(\pi)\left(2\text{Reg}_{\hat{f}_m}(\pi) + \frac{C_0 K}{\gamma_m}\right) \leq \frac{2K}{\gamma_m} + \frac{C_0 K}{\gamma_m}.$$

Where the first inequality follows from Lemma 7, and the second inequality follows from Lemma 5. □

B.3. Bounding $\hat{m}$ and l_m

Lemma 9 shows that when the events defined in Appendix B.1 hold, then $\hat{m}$ is at least as large as $m^* + 1$. In particular, this means that $m^* + 1$ is deemed a safe epoch with high probability.

Lemma 9. *Suppose the events $\mathcal{W}_1$, $\mathcal{W}_2$, and $\mathcal{W}_3$ hold. When $T \leq \tau_{m^*+1}$, the status of the algorithm at the end of round T is safe. When $T > \tau_{m^*+1}$, we have that $m^* + 1 \leq \hat{m}$.*

Proof. We first prove that under the assumptions of the theorem, $m^* + 1 \leq \hat{m}$ when $T > \tau_{m^*+1}$. Suppose for contradiction that $T > \tau_{m^*+1}$ and $\hat{m} \leq m^*$. Now consider the epoch $m' = \hat{m} + 1$. By assumption and choice of m', we have $\hat{m} < m'$ and $m' \leq \min\{m^*, \hat{m}\} + 1$. Therefore m' is not a safe epoch, hence there exists a time-step t' in epoch m' such that $t' \in \mathcal{T}$ and we have that:

$$\sum_{t=1}^{t'} r_t(a_t) < L_{t'}. \tag{38}$$

Since $\mathcal{W}_3$ holds and $t' \in \mathcal{T}$, we have that:

$$\sum_{t=1}^{t'} r_t(a_t) \geq \sum_{t=1}^{t'} \sum_{\pi \in \Psi} Q_t(\pi) R(\pi) - \sqrt{2t' \ln\left(\frac{(m' + \log_2(\tau_1))^3}{\delta'}\right)}. \tag{39}$$

For all $t \leq t'$, note that $m(t) \leq m'$. Therefore from Lemma 8, we have:

$$\begin{aligned}
&\sum_{t=1}^{t'} \sum_{\pi \in \Psi} Q_t(\pi) R(\pi) \\
&= t' R(\pi^*) - \sum_{t=1}^{t'} \sum_{\pi \in \Psi} Q_t(\pi) \mathrm{Reg}(\pi) \\
&\geq t' \cdot l_{m'-1} - \tau_1 - \sum_{t=\tau_1+1}^{t'} \frac{(2 + C_0)K}{\gamma_{m(t)}} \\
&\geq t' \cdot l_{m'-1} - \tau_1 - 20.3\sqrt{K} \sum_{t=\tau_1+1}^{t'} \sqrt{\xi\Big(\tau_{m(t)-1} - \tau_{m(t)-2}, \frac{\delta'}{(m(t))^2}\Big)}.
\end{aligned} \tag{40}$$

Here, the first equality follows from the definition of $\mathrm{Reg}(\cdot)$. The first inequality follows from $\mathcal{W}_2$ and Lemma 8. The result from Lemma 8 can be used here since $\mathcal{W}_1$ holds and since $m(t) \leq m' \leq \min\{m^*, \hat{m}\} + 1$ for all time-steps $t \leq t'$. The last inequality follows from substituting the value for C_0 and $\gamma_{m(t)}$. Combining (39) and (40) contradicts (38). Therefore when $T > \tau_{m^*+1}$, we have that $m^* + 1 \leq \hat{m}$.

The proof of the fact that the status of the algorithm at the end of round T is safe when $T \leq \tau_{m^*+1}$ is similar. Suppose for contradiction, $T \leq \tau_{m^*+1}$ and the status at the end of round T is not safe. We define $t' \in \mathcal{T}$ to be the first round where the status of the algorithm switches to "not safe" and we let $m' = m(t')$. Since t' is the first such time-step we have that $m' = \hat{m} + 1$. Further, since $m' \leq m(T) \leq m^* + 1$, we again have $\hat{m} \leq m^*$ and $m' \leq \min\{m^*, \hat{m}\} + 1$. Hence (38), (39), and (40) still hold. Giving us the same contradiction, because combining (39) and (40) contradicts (38). This completes the proof of Lemma 9. □

For all $m \geq m^* + 1$, Lemma 10 lower bounds l_m in terms of the optimal expected reward and the average misspecification error. Hence lower bounding the expected instantaneous reward of the algorithm when the status is "not safe".

Lemma 10. *Suppose the events $\mathcal{W}_1$ and $\mathcal{W}_2$ hold. For any epoch $m \geq m^* + 1$, we then have that:*

$$l_m \geq R(\pi^*) - 20.3\sqrt{KB} - \sqrt{2B}.$$

Proof. For compactness, let S denote the set S_{m^*+1}. From Lemma 9, we have that m^*+1 is a safe epoch. Hence at any epoch $m \geq m^*+1$, from the update rule in Choose-safe we have:

$$\begin{aligned} l_m &\geq l_{m^*+1} \\ &\geq \frac{1}{|S|}\sum_{(x,a,r)\in S} r \; - \sqrt{\frac{1}{2|S|}\ln\left(\frac{(m^*+1)^2}{\delta'}\right)} \\ &\geq \sum_{\pi\in\Psi} Q_{m^*+1}(\pi)R(\pi) - \sqrt{\frac{2}{|S|}\ln\left(\frac{(m^*+1)^2}{\delta'}\right)} \\ &\geq R(\pi^*) - \frac{(2+C_0)K}{\gamma_{m^*+1}} - \sqrt{\frac{2}{\tau_{m^*+1}-\tau_{m^*}}\ln\left(\frac{(m^*+1)^2}{\delta'}\right)}. \end{aligned} \tag{41}$$

Where the second last inequality follows from the fact that $\mathcal{W}_2$ holds and the last inequality follows from Lemma 8. Now from the definition of m^*, we have:

$$\begin{aligned} &\xi\Big(\tau_{m^*+1}-\tau_{m^*}, \frac{\delta'}{(m^*+1)^2}\Big) < B \\ \implies &\frac{K}{\gamma_{m^*+1}} \leq \sqrt{8K\xi\Big(\tau_{m^*+1}-\tau_{m^*}, \frac{\delta'}{(m^*+1)^2}\Big)} < \sqrt{8KB}. \end{aligned} \tag{42}$$

From Assumption 1, we have that ξ is a valid rate, hence (7) holds. Therefore, from (7) and the definition of m^*, we have:

$$\sqrt{\frac{2}{\tau_{m^*+1}-\tau_{m^*}}\ln\left(\frac{(m^*+1)^2}{\delta'}\right)} \leq \sqrt{2\xi\Big(\tau_{m^*+1}-\tau_{m^*}, \frac{\delta'}{(m^*+1)^2}\Big)} < \sqrt{2B}. \tag{43}$$

The result follows from combining (41), (42), and (43). □

B.4. Additional high probability events

In this section, we show that events $\mathcal{W}_4$ and $\mathcal{W}_5$ hold with high-probability. The event $\mathcal{W}_4$ provide upper and lower bound the difference between the expected regret and average regret at epochs that begin with a "safe" status.

$$\begin{aligned} \mathcal{W}_4 := \Bigg\{ \forall t' \in [\mathcal{T}], \; &\sqrt{\frac{2}{t'-\tau_{m(t')-1}}\ln\left(\frac{\lceil m(t')+\log_2(\tau_1)\rceil^3}{\delta'}\right)} \\ &\geq \sum_{\pi\in\Psi} Q_{t'}(\pi)\mathrm{Reg}(\pi) - \frac{1}{t'-\tau_{m(t')-1}}\sum_{t=\tau_{m(t')-1}+1}^{t'} (r_t(\pi^*(x_t)) - r_t(a_t)) \\ &\geq -\sqrt{\frac{2}{t'-\tau_{m(t')-1}}\ln\left(\frac{\lceil m(t')+\log_2(\tau_1)\rceil^3}{\delta'}\right)} \Bigg\}. \end{aligned} \tag{44}$$

Lemma 11. *The event $\mathcal{W}_4$ holds with probability at least $1-2\delta'$.*

Proof. Consider any time-step $t' \in \mathcal{T}$. Similar to the argument in Lemma 1, for the purpose of estimating the expected reward of the randomized policy $Q_{t'}$, we can treat the data generated in the first $t'-\tau_{m(t')-1}$ time-steps of epoch $m(t')$ as iid samples from the distribution $D(p_{m(t')})$.

Since $t' \in \mathcal{T}$, for all time-steps $t \leq t'$ from epoch $m(t')$, status is "safe" and actions are chosen according to the action

selection kernel $p_{m(t')}$. Hence from Hoeffding's inequality, with probability at least $1 - 2\delta'/\lceil m(t') + \log_2(\tau_1)\rceil^3$, we get:

$$\begin{aligned}
&\sqrt{\frac{2}{t' - \tau_{m(t')-1}} \ln\left(\frac{\lceil m(t') + \log_2(\tau_1)\rceil^3}{\delta'}\right)} \\
&\geq \sum_{\pi \in \Psi} Q_{t'}(\pi)\text{Reg}(\pi) - \frac{1}{t' - \tau_{m(t')-1}} \sum_{t=\tau_{m(t')-1}+1}^{t'} (r_t(\pi^*(x_t)) - r_t(a_t)) \\
&\geq -\sqrt{\frac{2}{t' - \tau_{m(t')-1}} \ln\left(\frac{\lceil m(t') + \log_2(\tau_1)\rceil^3}{\delta'}\right)}.
\end{aligned}$$

Any epoch m has at most $\lceil m + \log_2(\tau_1)\rceil$ time-steps in $\mathcal{T}$. Therefore $\mathcal{W}_4$ holds with probability at least:

$$1 - \sum_{t' \in \mathcal{T}} \frac{2\delta'}{\lceil m(t') + \log_2(\tau_1)\rceil^3} \geq 1 - \sum_{m=1}^{\infty} \frac{2\delta'}{(m + \log_2(\tau_1))^2} \geq 1 - 2\delta'.$$

□

The event $\mathcal{W}_5$ provides lower and upper bounds on the cumulative reward of the algorithm and optimal policy for various ranges of time-steps.

$$\begin{aligned}
\mathcal{W}_5 := \Bigg\{ \forall t' \in [\mathcal{T}] \cup \{T\},\ & \sum_{t=1}^{t'} (r_t(\pi^*(x_t)) - R(\pi^*)) \leq \sqrt{2t' \ln\left(\frac{\lceil m(t') + \log_2(\tau_1)\rceil^3}{\delta'}\right)} \\
& \sum_{t=1}^{t'} (\sum_{\pi \in \Psi} Q_t(\pi) R(\pi) - r_t(a_t)) \leq \sqrt{2t' \ln\left(\frac{\lceil m(t') + \log_2(\tau_1)\rceil^3}{\delta'}\right)} \\
& \sum_{t=t'+1}^{T} (r_t(\pi^*(x_t)) - R(\pi^*)) \leq \sqrt{2(T-t') \ln\left(\frac{\lceil m(t') + \log_2(\tau_1)\rceil^3}{\delta'}\right)} \\
& \sum_{t=t'+1}^{T} (\sum_{\pi \in \Psi} Q_t(\pi) R(\pi) - r_t(a_t)) \leq \sqrt{2(T-t') \ln\left(\frac{\lceil m(t') + \log_2(\tau_1)\rceil^3}{\delta'}\right)} \Bigg\}.
\end{aligned} \tag{45}$$

Lemma 12. *The event $\mathcal{W}_5$ holds with probability at least $1 - 4\delta'$.*

Proof. For each round t, we define:

$$\begin{aligned}
M_t &:= r_t(\pi^*(x_t)) - R(\pi^*), \\
M'_t &:= \sum_{\pi \in \Psi} Q_t(\pi) R(\pi) - r_t(a_t).
\end{aligned}$$

It is straightforward to see that $\mathbb{E}[M_t | \Gamma_{t-1}] = 0$. Further from Lemma 2, we have that $\mathbb{E}[M'_t | \Gamma_{t-1}] = 0$. Now consider any time-step $t' \in \mathcal{T} \cup \{T\}$. Since $|M_t|, |M'_t| \leq 1$ for all t, from Azuma's inequality, with probability at least $1 - 4\delta'/\lceil m(t') +$

$\log_2(\tau_1)\rceil^3$, we have: [14]

$$\begin{aligned}
\sum_{t=1}^{t'} M_t &\le \sqrt{2t' \ln\left(\frac{\lceil m(t') + \log_2(\tau_1)\rceil^3}{\delta'}\right)}, \\
\sum_{t=1}^{t'} M'_t &\le \sqrt{2t' \ln\left(\frac{\lceil m(t') + \log_2(\tau_1)\rceil^3}{\delta'}\right)}, \\
\sum_{t=t'+1}^{T} M_t &\le \sqrt{2(T-t') \ln\left(\frac{\lceil m(t') + \log_2(\tau_1)\rceil^3}{\delta'}\right)}, \\
\sum_{t=t'+1}^{T} M'_t &\le \sqrt{2(T-t') \ln\left(\frac{\lceil m(t') + \log_2(\tau_1)\rceil^3}{\delta'}\right)}.
\end{aligned} \tag{46}$$

Any epoch m has at most $\lceil m + \log_2(\tau_1)\rceil$ time-steps in $\mathcal{T} \cup \{T\}$. Therefore $\mathcal{W}_5$ holds with probability at least:

$$1 - \sum_{t' \in \mathcal{T}} \frac{4\delta'}{\lceil m(t') + \log_2(\tau_1)\rceil^3} \ge 1 - \sum_{m=1}^{\infty} \frac{4\delta'}{(m(t') + \log_2(\tau_1))^2} \ge 1 - 4\delta'.$$

□

B.5. Proof of Theorem 1

Note that $\mathcal{W}_1$, $\mathcal{W}_2$, $\mathcal{W}_3$, $\mathcal{W}_4$, and $\mathcal{W}_5$ all hold with probability at least $1-\delta$. We now split our analysis into two cases and bound the cumulative regret for each case, while assuming all these high-probability events hold.

Case 1 ($T \le \tau_{m^*+1}$):
From $\mathcal{W}_5$, we have that:

$$\sum_{t=1}^{T} \Big(r_t(\pi^*(x_t)) - r_t(a_t)\Big) \le \sum_{t=1}^{T} \sum_{\pi \in \Psi} Q_t(\pi)\text{Reg}(\pi) + \sqrt{8T \ln\left(\frac{\lceil m(T) + \log_2(\tau_1)\rceil^3}{\delta'}\right)}. \tag{47}$$

Since $\mathcal{W}_1$, $\mathcal{W}_2$, $\mathcal{W}_3$ hold, from Lemma 9 we have that the status at the end of round T is safe. Since $\mathcal{W}_1$ holds, from Lemma 8, we have:

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{\pi \in \Psi} Q_t(\pi)\text{Reg}(\pi) \\
&\le \tau_1 + \sum_{t=\tau_1+1}^{T} \frac{(2+C_0)K}{\gamma_{m(t)}} \\
&\le \tau_1 + 20.3\sqrt{K} \sum_{t=\tau_1+1}^{T} \sqrt{\xi\Big(\tau_{m(t)-1} - \tau_{m(t)-2}, \frac{\delta'}{(m(t))^2}\Big)}.
\end{aligned} \tag{48}$$

Combining eq. (47) and eq. (48) completes the analysis for the first case.

Case 2 ($T > \tau_{m^*+1}$):
Let t' be the last time-step where the conditions in Check-is-safe were checked and verified to be true. That is, [15]

$$t' = \max\{t \in \mathcal{T} \cap [T] | \text{ algorithm status is safe at the end of round } t\}. \tag{49}$$

Since $\mathcal{W}_1$, $\mathcal{W}_2$, $\mathcal{W}_3$ hold, from Lemma 9, we have that epoch m^*+1 is safe. Therefore, the last round of this epoch is safe and hence $t' \ge \tau_{m^*+1}$. We now prove a bound on the cumulative regret up to time t' (see (53)). We start by deriving a bound for the cumulative regret up to time τ_{m^*+1} and then derive a bound for the cumulative regret up to time t' when $t' > \tau_{m^*+1}$.

[14] For $t' = T$, the last two inequalities in (46) are trivial.

[15] In the definition of t' (see (49)), it may seem redundant to consider the set $\mathcal{T} \cap [T]$ when $\mathcal{T} \subseteq [T]$. We do this because, later in the proof, we use a thought experiment where we make the bandit run for more that T time-steps.

Since τ_{m^*+1} lies in $\mathcal{T}$, the event $\mathcal{W}_5$ bounds the cumulative regret upto time τ_{m^*+1} in terms of the expected cumulative regret. Hence following the arguments in case 1, we get:

$$\sum_{t=1}^{\tau_{m^*+1}} \Big(r_t(\pi^*(x_t)) - r_t(a_t)\Big) \leq \tau_1 + 20.3\sqrt{K} \sum_{t=\tau_1+1}^{\tau_{m^*+1}} \sqrt{\xi\Big(\tau_{m(t)-1} - \tau_{m(t)-2}, \frac{\delta'}{(m(t))^2}\Big)}. \tag{50}$$

Hence, we now only need to bound the cumulative regret up to time t' when $t' > \tau_{m^*+1}$. For compactness, let m' denote the epoch $m(t')$. Since the status of the algorithm is safe at the end of round t' and $t' \in \mathcal{T}$, we have that:

$$\sum_{t=1}^{t'} r_t(a_t) \geq t' \cdot l_{m'-1} - \tau_1 - \sqrt{2t' \ln\Big(\frac{(m' + \log_2(\tau_1))^3}{\delta'}\Big)} - 20.3\sqrt{K} \sum_{i=\tau_2}^{t'} \sqrt{\xi\Big(\tau_{m(i)-1} - \tau_{m(i)-2}, \frac{\delta'}{(m(i))^2}\Big)}. \tag{51}$$

Now note that when $t' > \tau_{m^*+1}$, we have:

$$\begin{aligned}
&\sum_{t=1}^{t'} (r_t(\pi^*(x_t)) - r_t(a_t)) \\
&\leq \sum_{t=1}^{t'} (R(\pi^*) - r_t(a_t)) + \sqrt{2t' \ln\left(\frac{\lceil m(t') + \log_2(\tau_1)\rceil^3}{\delta'}\right)} \\
&\leq t' \cdot (20.3\sqrt{KB} + \sqrt{2B}) + \tau_1 + \sqrt{8t' \ln\left(\frac{\lceil m(t') + \log_2(\tau_1)\rceil^3}{\delta'}\right)} \\
&\quad + 20.3\sqrt{K} \sum_{i=\tau_2}^{t'} \sqrt{\xi\Big(\tau_{m(i)-1} - \tau_{m(i)-2}, \frac{\delta'}{(m(i))^2}\Big)}.
\end{aligned} \tag{52}$$

Where the first inequality follows from the fact that $\mathcal{W}_5$ holds. When $t' > \tau_{m^*+1}$, note that $m' > m^* + 1$, hence Lemma 10 bounds $l_{m'-1}$. The last inequality follows from (51), and Lemma 10.

To recap, we already argued that $t' \geq \tau_{m^*+1}$. In (50), we also bounded cumulative regret up to time τ_{m^*+1}. Finally, (52) bounds cumulative regret up to time t' when $t' > \tau_{m^*+1}$. Combining everything together, we get an unconditional bound on the cumulative regret up to time t':

$$\begin{aligned}
&\sum_{t=1}^{t'} (r_t(\pi^*(x_t)) - r_t(a_t)) \\
&\leq t' \cdot (20.3\sqrt{KB} + \sqrt{2B}) + \tau_1 + \sqrt{8t' \ln\left(\frac{\lceil m(t') + \log_2(\tau_1)\rceil^3}{\delta'}\right)} \\
&\quad + 20.3\sqrt{K} \sum_{i=\tau_2}^{t'} \sqrt{\xi\Big(\tau_{m(i)-1} - \tau_{m(i)-2}, \frac{\delta'}{(m(i))^2}\Big)}.
\end{aligned} \tag{53}$$

Note that (53) bounds the cumulative regret up to the last time-step where the conditions in Check-is-safe were verified to be true. Also if $t' = T$, then (53) gives us the required bound on the cumulative regret up to time T. Hence, moving forward, we only need to bound cumulative regret when $T > t'$.

Recall that we defined t' to be the last time-step where the conditions in Check-is-safe were verified to be true. Let t'' be the next time-step where the Check-is-safe conditions would be checked. We now bound the cumulative regret up to time t'' in terms of the cumulative regret up to time t'.

Note that if $T \geq t''$, then $t'' \in \mathcal{T}$. On the other hand if $T < t''$, it is easy to see that the cumulative regret up to time T would be roughly smaller than the cumulative regret up to time t'' had the bandit run up to round t''. Since we are going to bound the cumulative regret up to time t'' in terms of the cumulative regret up to time t' and $T > t'$, for the purposes of bounding cumulative regret up to time T, we can assume that the bandit runs up to time t'' and $t'' \in \mathcal{T}$.

Note that if $t'' = t' + 1$, we have:

$$\sum_{t=1}^{t''}(r_t(\pi^*(x_t)) - r_t(a_t)) \leq 2 + \sum_{t=1}^{t'}(r_t(\pi^*(x_t)) - r_t(a_t)) \tag{54}$$

We now want to bound the cumulative regret up to time t'' in terms of the cumulative regret up to time t' when $t'' > t' + 1$. Since both t' and t'' are consecutive rounds in $\mathcal{T}$, when $t'' > t' + 1$, we have that both rounds lie in the same epoch. That is, $m(t'') = m(t') = m'$. When both rounds lie in the same epoch, since the status of the algorithm is safe at the end of round t', we have that the action selection kernel $p_{m'}$ is used to pick actions at every time-step $t \in [\tau_{m'-1} + 1, t'']$. Therefore, when $t'' > t' + 1$, we have:

$$\begin{aligned}
&\sum_{t=\tau_{m'-1}+1}^{t''}(r_t(\pi^*(x_t)) - r_t(a_t)) \\
&\leq (t'' - \tau_{m'-1})\sum_{\pi\in\Psi} Q_{m'}(\pi)\text{Reg}(\pi) + \sqrt{2(t'' - \tau_{m'-1})\ln\left(\frac{\lceil m' + \log_2(\tau_1)\rceil^3}{\delta'}\right)} \\
&\leq 2(t' - \tau_{m'-1})\sum_{\pi\in\Psi} Q_{m'}(\pi)\text{Reg}(\pi) + \sqrt{4(t' - \tau_{m'-1})\ln\left(\frac{\lceil m' + \log_2(\tau_1)\rceil^3}{\delta'}\right)} \\
&\leq 2\sum_{t=\tau_{m'-1}+1}^{t'}(r_t(\pi^*(x_t)) - r_t(a_t)) + 5\sqrt{(t' - \tau_{m'-1})\ln\left(\frac{\lceil m' + \log_2(\tau_1)\rceil^3}{\delta'}\right)}.
\end{aligned} \tag{55}$$

The following arguments assume $t'' > t' + 1$. The first inequality follows from the fact that $t'' \in \mathcal{T}$, the fact that $\mathcal{W}_4$ holds, and the fact that the algorithm uses the action selection kernel $p_{m'}$ to pick actions at every time-step $t \in [\tau_{m'-1} + 1, t'']$. The second inequality follows from the fact that t', t'' are consecutive rounds in $\mathcal{T}$, and are in the same epoch. The last inequality follows from the fact that $t' \in \mathcal{T}$, $\mathcal{W}_4$ holds, and the fact that the algorithm uses the action selection kernel $p_{m'}$ to pick actions at every time-step $t \in [\tau_{m'-1} + 1, t']$.

Therefore, when $t'' > t' + 1$, from (55) we have:

$$\sum_{t=1}^{t''}(r_t(\pi^*(x_t)) - r_t(a_t)) \leq 3\sum_{t=1}^{t'}(r_t(\pi^*(x_t)) - r_t(a_t)) + 5\sqrt{T\ln\left(\frac{(m' + \log_2(\tau_1))^3}{\delta'}\right)} \tag{56}$$

To recap, we know that $t'' > t'$. When $t'' = t' + 1$, (54) bounds the cumulative regret up to time t'' in terms of the cumulative regret up to time t'. When $t'' > t' + 1$, (56) bounds the cumulative regret up to time t'' in terms of the cumulative regret up to time t'. Combining everything together, we get the following unconditional bound on the cumulative regret up to time t'' in terms of the cumulative regret up to time t':

$$\sum_{t=1}^{t''}(r_t(\pi^*(x_t)) - r_t(a_t)) \leq 2 + 3\sum_{t=1}^{t'}(r_t(\pi^*(x_t)) - r_t(a_t)) + 5\sqrt{T\ln\left(\frac{(m' + \log_2(\tau_1))^3}{\delta'}\right)} \tag{57}$$

Case 2.1 ($T > \tau_{m^*+1}$ and $T < t''$):

Note that:

$$\begin{aligned}
&\sum_{t=1}^{T}\Big(r_t(\pi^*(x_t)) - r_t(a_t)\Big)\\
&\leq \sum_{t=1}^{T}\sum_{\pi\in\Psi} Q_t(\pi)\text{Reg}(\pi) + \sqrt{8T\ln\left(\frac{\lceil m(T)+\log_2(\tau_1)\rceil^3}{\delta'}\right)}\\
&\leq \sum_{t=1}^{t''}\sum_{\pi\in\Psi} Q_t(\pi)\text{Reg}(\pi) + \sqrt{8T\ln\left(\frac{\lceil m(T)+\log_2(\tau_1)\rceil^3}{\delta'}\right)}\\
&\leq \sum_{t=1}^{t''}\Big(r_t(\pi^*(x_t)) - r_t(a_t)\Big) + 7\sqrt{T\ln\left(\frac{\lceil 1+m(T)+\log_2(\tau_1)\rceil^3}{\delta'}\right)}.
\end{aligned} \tag{58}$$

Where the first inequality follows from the fact that $\mathcal{W}_5$ holds. The second inequality follows from the fact that $t'' > T$ and the fact that $\text{Reg}(\pi) \geq 0$. The last inequality follows from $\mathcal{W}_5$ and the fact that $t'' \leq 2t' \leq 2T$.

Combining (53), (57), and (58) gives us the required bound on the cumulative regret for this case.

Case 2.2 ($T > \tau_{m^*+1}$ and $T \geq t''$):
From the definition of t' and t'', the status of the algorithm switches to "not safe" at the end of round t''. Thereafter, all actions will be selected according to the action selection kernel $p_{\hat{m}}$. Now note that:

$$\begin{aligned}
&\sum_{t=t''+1}^{T}\Big(r_t(\pi^*(x_t)) - r_t(a_t)\Big)\\
&\leq \sum_{t=t''+1}^{T}\sum_{\pi\in\Psi} Q_{\hat{m}}(\pi)\text{Reg}(\pi) + \sqrt{8(T-t'')\ln\left(\frac{\lceil m(t'')+\log_2(\tau_1)\rceil^3}{\delta'}\right)}\\
&\leq (R(\pi^*) - l_{\hat{m}})(T-t'') + \sqrt{8(T-t'')\ln\left(\frac{\lceil m(t'')+\log_2(\tau_1)\rceil^3}{\delta'}\right)}\\
&\leq (20.3\sqrt{KB} + \sqrt{2B})(T-t'') + \sqrt{8(T-t'')\ln\left(\frac{\lceil m(t'')+\log_2(\tau_1)\rceil^3}{\delta'}\right)}.
\end{aligned} \tag{59}$$

Where the first inequality follows from $\mathcal{W}_5$ and the fact that the kernel $p_{\hat{m}}$ is used for all rounds $t \in [t''+1, T]$. The second inequality follows from $\mathcal{W}_2$, which gives us that the expected reward of the randomized policy $Q_{\hat{m}}$ is lower bounded by $l_{\hat{m}}$. From Lemma 9 we have that $\hat{m} \geq m^* + 1$. Therefore the final inequality follows from Lemma 10, which gives us a lower bound on $l_{\hat{m}}$ since $\hat{m} \geq m^* + 1$.

Combining (53), (57), and (59) gives us the required bound on the cumulative regret for this case.

C. Proof for lower bound

In this section, we prove Theorem 2, which we restate bellow for convenience.

Theorem 2 (Lower bound)**.** *Consider any $K \geq 2$ and $B \in [0, 1/(2K)]$. One can construct a model class $\mathcal{F}$ and a stochastic contextual bandit instance with K arms. Such that the average misspecification error is $\sqrt{B}$. And for any probability kernel p in the convex hull of the kernel set $\mathcal{K}(\mathcal{F})$, the expected instantaneous regret of the induced randomized policy can be lower bounded by:*

$$\mathop{\mathbb{E}}_{(x,r)\sim D}\mathop{\mathbb{E}}_{a\sim p(\cdot|x)}\big[r(\pi^*(x)) - r(a)\big] \geq \Omega(\sqrt{KB}) \tag{23}$$

Proof. Consider any $K \geq 2$ and $B \in [0, 1/(2K)]$. We start by constructing a K arm stochastic contextual bandit instance. Let $\mathcal{A} = [K]$ be the set of arms, $\mathcal{X} = (0, K)$ be the set of contexts, and D denote the joint distribution of rewards and contexts. Such that $D_{\mathcal{X}}$, the marginal distribution of D on the set of contexts, is uniformly distributed over over the context

set $\mathcal{X}$. That is, $D_{\mathcal{X}} \equiv \text{Unif}(0, K)$. For all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, we let the conditional expected reward be given by,

$$f^*(x, a) := \begin{cases} \alpha & \text{for } x \in (a-1, a], \\ 0 & \text{otherwise.} \end{cases} \tag{60}$$

Where α is given by,

$$\alpha := \sqrt{\frac{K^2}{K-1} B}. \tag{61}$$

Since $K \geq 2$ and $B \in [0, 1/(2K)]$, we have that $\alpha \in [0, 1]$. Hence the conditional expected reward for every context and action lies in $[0, 1]^K$. We also let the rewards be noiseless. That is,

$$\Pr_{(x,r)\sim D} \big(r(a) = f^*(x, a)\big) = 1. \tag{62}$$

This completes our description of the stochastic contextual bandit setup that we consider. We now let the model class $\mathcal{F}$ be given by,

$$\mathcal{F} = \{f : \mathcal{X} \times \mathcal{A} \to [0, 1] \mid \forall a \in \mathcal{A},\ \exists f_a \in [0, 1] \text{ such that, } \forall x \in \mathcal{X},\ \text{ we have } f(x, a) = f_a\}. \tag{63}$$

That is, $\mathcal{F}$ is the class of all models that do not depend on contexts. Therefore, the set $\mathcal{K}(\mathcal{F})$ contains all probability kernels that do not depend on contexts.

$$\begin{aligned} \mathcal{K}(\mathcal{F}) =& \left\{ p \text{ is a probability kernel} \middle| \ \exists f_p \in \mathcal{F},\ g_p : \mathcal{A} \times \mathbb{R}^{\mathcal{A}} \to [0, 1],\ \forall\ (x, a) \in \mathcal{X} \times \mathcal{A},\ p(a|x) = g_p(a|f_p(x)) \right\} \\ =& \left\{ p \text{ is a probability kernel} \middle| \ \exists\ g_p : \mathcal{A} \to [0, 1],\ \forall\ (x, a) \in \mathcal{X} \times \mathcal{A},\ p(a|x) = g_p(a) \right\}. \end{aligned} \tag{64}$$

That is, the set $\mathcal{K}(\mathcal{F})$ simply reduces to the set of distributions over $\mathcal{A}$ Which also implies that $\mathcal{K}(\mathcal{F})$ is convex. For notational convenience, we let $\mathcal{G}$ denote the set of all distributions over the action set $\mathcal{A}$. That is,

$$\mathcal{G} = \left\{ g : \mathcal{A} \to [0, 1] \middle| \ \sum_{a \in \mathcal{A}} g(a) = 1 \right\}. \tag{65}$$

Now note that for any arm $a \in \mathcal{A}$, from (60), we have that:

$$\mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}} [f^*(x, a)] = \frac{\alpha}{K} \in [0, 1]. \tag{66}$$

Now note that, the average misspecification is given by:

$$\begin{aligned} & \sqrt{\max_{p \in \mathcal{K}(\mathcal{F})} \min_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}} \mathop{\mathbb{E}}_{a \sim p(\cdot|x)} [(f(x, a) - f^*(x, a))^2]} \\ =& \sqrt{\max_{g \in \mathcal{G}} \sum_{a \in \mathcal{A}} g(a) \min_{f_a \in [0,1]} \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}} [(f_a - f^*(x, a))^2]} \\ =& \sqrt{\max_{g \in \mathcal{G}} \sum_{a \in \mathcal{A}} g(a) \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}} [(f^*(x, a) - \alpha/K)^2]} \\ =& \sqrt{\max_{g \in \mathcal{G}} \sum_{a \in \mathcal{A}} g(a) \frac{K-1}{K^2} \alpha} = \sqrt{B}. \end{aligned} \tag{67}$$

Here the first equality follows from (64) and (65). The second equality follows from (66) and the fact that the mean minimizes the mean squared error. The third equality follows from substituting the value for $f^*(x, a)$ from (60). Finally, the last equality follows from (61) and the fact that $g \in \mathcal{G}$.

It is easy to see that the optimal policy is given by π^* (defined in (68)). Further, the expected reward of π^* is α. That is, $R(\pi^*) = \alpha$.

$$\text{For all } a \in \mathcal{A} \text{ and } x \in (a-1, a],\ \pi^*(x) := a. \tag{68}$$

Now consider any arm $a \in \mathcal{A}$. With some abuse of notation, let a also denote the policy that selects arm a for all contexts. Note that the expected reward for this policy is α/K, that is $R(a) = \alpha/K$. For any probability kernel p in $\mathcal{K}(\mathcal{F})$, since it does not depend on contexts, the randomized policy Q_p is only supported by policies in $\mathcal{A}$. Therefore, we have that the expected regret of any randomized policy Q_p that is induced by some probability kernel in $\mathcal{K}(\mathcal{F})$ is given by:

$$\begin{aligned}&\sum_{\pi\in\Psi} Q_p(\pi)\text{Reg}(\pi) = \sum_{a\in\mathcal{A}} Q_p(a)\text{Reg}(a) = \sum_{a\in\mathcal{A}} Q_p(a)\big(R(\pi^*) - R(a)\big) = \sum_{a\in\mathcal{A}} Q_p(a)\Big(1-\frac{1}{K}\Big)\alpha \\ &= \Big(1-\frac{1}{K}\Big)\alpha = \sqrt{(K-1)B} \geq \sqrt{KB/2}.\end{aligned} \tag{69}$$

This completes the proof of Theorem 2. □

D. Detailed introduction example

To generate both Figure 1 and Figure 2, we implement FALCON+ and Safe-FALCON respectively. Both implementations require knowledge of the estimation rate function (ξ). In this section, we detail our choice of estimation rates and give some more intuition for the oscillatory regret behavior we see in that Figure 1.

FALCON+ Setup As explained in the introduction, we use the FALCON+ algorithm in (Simchi-Levi & Xu, 2020). This algorithm requires knowledge of a function $\xi_{\mathcal{F},\delta}(n)$ representing the estimation rate. This is defined in the following assumption.

Assumption 2 in (Simchi-Levi & Xu, 2020) Given n data samples $\{(x_1, a_1, r_1(a_1)), \cdots, (x_n, a_n, r_n(a_n))\}$ generated *iid* from an arbitrary distribution $\mathcal{D}_{\text{data}}$, the offline regression oracle return a function $\hat{f}$. For all $\delta > 0$, with probability at least $1-\delta$, we have

$$\mathbb{E}[(\hat{f}(x,a) - f^*(x,a))^2] \leq \xi_{\mathcal{F},\delta}(n) \tag{70}$$

We set the class of functions $\mathcal{F}$ to be the set of linear functions (i.e., linear regressions of outcomes on contexts and an intercept). For this model, it is straightforward to show analytically that the random variable on the left-hand side of (70) is distributed as a random variable $\frac{1}{n}\chi^2_2$, where χ^2_2 is a random variable distributed as chi-squared with two degrees of freedom. Therefore, we set $\xi_{\mathcal{F},\delta}(n)$ to be the $1-\delta$-quantile of the distribution of $\frac{1}{n}\chi^2_2$. Note that this quantity is decreasing with n.

Explaining results In Figure 1, we saw that average per-epoch regret oscillates between very low and very high levels. Let's explain this phenomenon.

First, note that the optimal treatment assignment policy for this setting is

$$\pi^*(x) = \begin{cases} 1 & \text{if } x \leq .5, \\ 2 & \text{otherwise.} \end{cases} \tag{71}$$

Moreover, if arms were assigned uniformly at random, the the best linear approximation to $f^*(x,a)$ would be given by

$$\hat{f}^*(x,a) := \begin{cases} -.25 + 1.5x & \text{if } a = 1, \\ .5 & \text{if } a = 2. \end{cases} \tag{72}$$

Finally, note that the data-generating process was selected so that the policy $\pi_{\hat{f}^*}$ induced by the best linear approximation coincides with $\pi^*(x)$.

With this in mind, we are ready to understand the oscillatory behavior in Figure 3. During the first seven or so epochs, arms are assigned roughly at random, and the linear outcome models fit on this data are not too different from $\hat{f}^*(x,a)$, but the exploitation parameter $\gamma_m \propto \xi_{\mathcal{F},\delta}(n)^{-1}$ is small, so the induced action selection kernel $p_m(a|x)$ does not concentrate very much.

However, since epochs have increasing size, for later epochs $\gamma_m \propto \xi_{\mathcal{F},\delta}(n)^{-1}$ can be large, and $p_m(a|x)$ concentrates and becomes approximately $\pi_{\hat{f}^*}$. This is what happens in Epoch 8 in Figure 1, for example. However, this skews the data

distribution, so that in Epoch 9 when we fit an outcome model using data from Epoch 8 we get something that is very different from $\hat{f}^*(x, a)$. In turn, this accidentally increases the amount of exploration happening in this epoch. So in Epoch 10 we are able to return to an outcome model that is similar to $\hat{f}^*(x, a)$. But by then the exploitation parameter γ_m is even larger, so $p_m(a|x)$ concentrates again, causing the cycle to repeat.

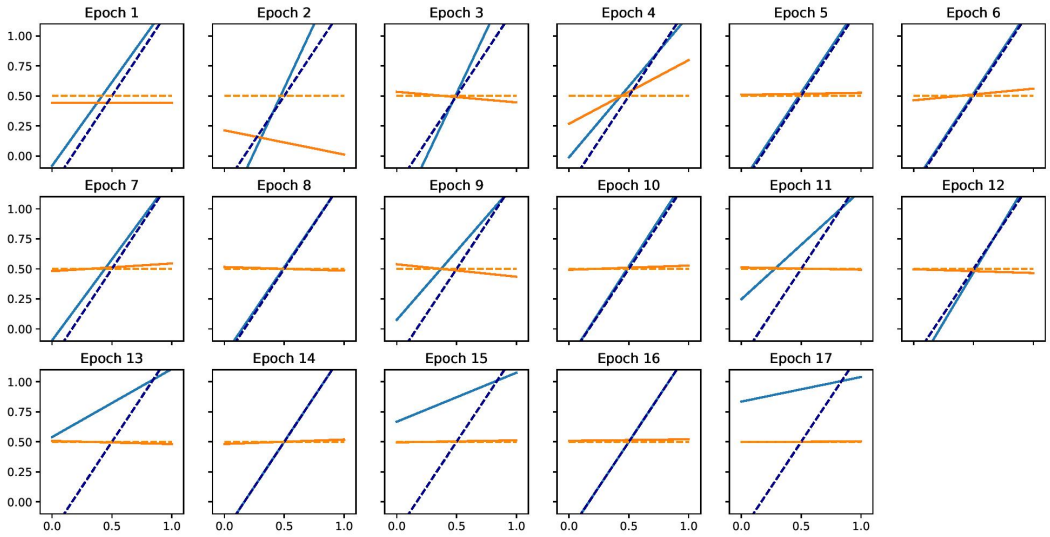

Figure 3. Linear models at the end of each epoch m and action $a = 1$ (blue) and $a = 2$ (orange). Solid lines are fitted models $\hat{f}(\cdot, a)$. Dashed lines are oracle best linear approximation under uniform action sampling.

The takeaway from this example is that in the presence of misspecification we must curb the amount of exploitation. This insight underpinned the construction of the algorithm presented here.