Model Selection in Contextual Stochastic Bandit Problems

Aldo Pacchiano∗
pacchiano@berkeley.edu
UC Berkeley

My Phan∗
myphan@cs.umass.edu
University of Massachusetts

Yasin Abbasi-Yadkori
yasin.abbasi@gmail.com
VinAI

Anup Rao
anuprao@adobe.com
Adobe

Julian Zimmert
zimmert@google.com
Google Research

Tor Lattimore
lattimore@google.com
DeepMind

Csaba Szepesvari
szepi@google.com
DeepMind

March 4, 2020

Abstract

We study model selection in stochastic bandit problems. Our approach relies on a master algorithm that selects its actions among candidate base algorithms. While this problem is studied for specific classes of stochastic base algorithms, our objective is to provide a method that can work with more general classes of stochastic base algorithms. We propose a master algorithm inspired by CORRAL [Agarwal et al., 2017] and introduce a novel and generic smoothing transformation for stochastic bandit algorithms that permits us to obtain $O(\sqrt{T})$ regret guarantees for a wide class of base algorithms when working along with our master. We exhibit a lower bound showing that even when one of the base algorithms has $O(\log T)$ regret, in general it is impossible to get better than $\Omega(\sqrt{T})$ regret in model selection, even asymptotically. We apply our algorithm to choose among different values of $\epsilon$ for the $\epsilon$-greedy algorithm, and to choose between the $k$-armed UCB and linear UCB algorithms. Our empirical studies further confirm the effectiveness of our model-selection method.

1 Introduction

Bandit algorithms have been applied in a variety of decision making and personalization problems in industry. There are many specialized algorithms each designed to perform well in specific environments. For example, algorithms are designed to exploit low variance [Audibert et al., 2009], extra context information and linear reward structure [Dani et al., 2008; Li et al., 2010; Abbasi-Yadkori et al., 2011], sparsity [Abbasi-Yadkori et al., 2012; Carpentier and Munos, 2012], etc. The exact properties of the current environment however might not be known in advance, and we might not know which algorithm is going to perform best. Given the online nature of the problem, batch model selection is not possible in many practical situations. Therefore, it is desired to develop a method to perform model-selection with bandit information in an online fashion.

As an example, consider the application of bandit algorithms in online personalization problems where the task is to assign one of the available offers to each visiting user. Often a context vector is available that provides extra information about the user (such as location, browser type, etc). Contextual bandit algorithms such as LinUCB [Li et al., 2010] are designed for such problems. When the context vectors are high-dimensional and arrive in an i.i.d fashion, and the time horizon is small, then by the bias-variance

∗Equal contribution.
trade-off we might be better off using a simpler non-contextual bandit algorithm instead of a contextual algorithm. Here, we might want to choose between UCB and LinUCB in an adaptive fashion.

As another application, consider the problem of tuning the exploration rate of bandit algorithms such as $\epsilon$-greedy, UCB, etc. The exploration rate recommended by the theoretical analysis can be overly conservative. It might be tempting to decrease the exploration rate manually when deploying the algorithm in practice. The danger is that if the exploration rate is too small, the algorithm might perform poorly. We would like to design a mechanism to tune the exploration rate in an adaptive data-dependant fashion.

Maillard and Munos (2011) are perhaps the first to address the bandit model-selection problem. These results are improved by Agarwal et al. (2017). The main idea of Agarwal et al. (2017) is to combine the base algorithms using an online mirror descent master algorithm that sends importance weighted rewards to the base algorithms. Given the application of importance weighting, the approach is better suited for combining adversarial base algorithms.

Chatterji et al. (2019) and Foster et al. (2019) study bandit model-selection problem when the reward is stochastic and has a linear structure of unknown order. Chatterji et al. (2019) propose an algorithm for model-selection and show strong guarantees but under strong conditions. More specifically, Chatterji et al. (2019) assume that the contexts are sampled in an i.i.d. fashion from a distribution and the smallest eigenvalue of the covariance matrix of the distribution is sufficiently large. Under such assumptions, Bastani et al. (2017) and Kannan et al. (2018) suggest that advanced exploration might not be necessary. Foster et al. (2019) consider the linear contextual bandit problem with multiple policy classes of different dimensions. Foster et al. (2019) show $O(T^{2/3}d^{1/3})$ and $O(T^{3/4} + \sqrt{Td})$ regret guarantees where $T$ is the time horizon and $d^*$ is the true dimension of the reward function. These bounds are sub-optimal when $d^*$ is not too large. Further, Foster et al. (2019) require a lower bound on the average eigenvalues of the co-variance matrices of all actions. They pose the question of whether model selection is possible without eigenvalue conditions. Apart from strong assumptions, the above results are limited to model-selection among linear classes. A general and efficient method to combine multiple stochastic base algorithms is missing.

In this work, we focus on bandit model-selection in general stochastic environments. Notice that for the approach of Agarwal et al. (2017) to be applicable, a base algorithm needs to be properly modified to satisfy the stability condition of Agarwal et al. (2017). For example, for the UCB algorithm we would need to use a Bernstein type concentration inequality instead of the usual Hoeffding bound. This approach is tedious as each algorithm needs to be individually modified. We would like to provide a generic procedure applicable to most base algorithms in a stochastic environment. We provide such result by introducing a smoothing technique that transforms almost any algorithm in a stochastic environment to one that satisfies a stability condition so that it can be used along with the model selection master algorithm. In particular, we show how the approach can be used to combine UCB and LinUCB in contextual problems. We can also use our model-selection procedure to obtain a near optimal exploration rate for $\epsilon$-greedy algorithms without a prior knowledge of the smallest gap. Our empirical studies confirm the effectiveness of the proposed approach in these two applications.

In the stochastic domain, an important question is whether a model selection procedure can inherit the $O(\log T)$ regret of a fast stochastic base algorithm (such as UCB when the “gap” is large). We show via a lower bound construction that such a result is impossible in general.

Let us also mention the literature on the “best of the both worlds” problems. These papers aim to design a single algorithm that can handle both stochastic and adversarial environments (Audibert and Bubeck, 2009; Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Auer and Chiang, 2016; Abbasi-Yadkori et al., 2018; Zimmert and Seldin, 2019).

2 Problem statement

We consider a contextual multi-armed bandit problem with $K$ actions. In round $t$, the learner observes a $d$-dimensional context vector $x_t \in \mathcal{X}$, that arrives in an i.i.d fashion. Let $h_t \in \mathcal{H}_t$ denote the history at time
We assume $M$ base algorithms are available. Let $R_i(T)$ be the regret of the $i$th base algorithm. We want to design a bandit method that plays one of these base algorithms in each round and its overall regret satisfies $R(T) \leq O(\min_i R_i(T))$. For any algorithm $B$ we define its instantaneous regret at time $t$ as $r_t = f(x_t, \pi^*) - f(x_t, \pi_t)$ where $\pi^*$ is the optimal policy in Equation 1 and $\pi_t$ is the possibly path dependent policy $B$ uses at time $t$.

3 Stochastic Corral

In this Section we introduce our algorithm and provide its regret analysis. In Section 3.1 we describe our algorithm. In Section 4.3 we show the regret analysis provided the base algorithm satisfies a condition. In Section 4.4 we show a "smoothing" procedure that will transform a wide class of algorithms to satisfy the condition.

3.1 Algorithm

Our algorithm is a variant of the CORRAL algorithm [Agarwal et al., 2017] modified for stochastic environments. First, we explain the CORRAL algorithm and then introduce the new variant.

The basic structure of the CORRAL algorithm is the following: The master receives $M$ base algorithms $\{B_i\}_{i=1}^M$. During any time $t$ of the algorithm’s execution, CORRAL maintains a distribution $p_t$ over $\Delta_M$ used to select the index $i_t \sim p_t$ of the algorithm to use during that round. After an algorithm $B_{i_t}$ is selected, its policy $\pi_{t,i_t}$ is used by the master to select an action $a_t \sim \pi_{t,i_t}(x_t)$ where $x_t \sim D$ is the context sampled at time $t$. The resulting reward signal $g_t = f(x_t, \delta_{a_t}) + \xi_t$ where $\xi_t$ denotes a zero-mean random noise and $\delta_i$ denotes the Dirac distribution at action $i$ is fed back to all the $M$ base algorithms in the form of an importance weighted estimate $\hat{g}_i^t = \frac{\prod_{t=1}^T (i_t = i)}{p_t} g_t$ for all $i$. Subsequently each of the base algorithms $i \in [M]$ update their internal state based on the feedback $\hat{g}_i^t$ received.

CORRAL requires its base algorithms to satisfy a stability condition to work along with the importance weighting feedback. Because importance weighting can change the loss range and distribution throughout the run of the algorithm in an unpredictable fashion, it is not directly compatible with a stochastic reward environment. [Agarwal et al., 2017] change the details of many algorithms in a case by case basis to make them stable. To avoid having to know the specific workings inside each base algorithm, we introduce a variant of the CORRAL algorithm and a generic algorithmic smoothing transformation that allows us to prove model selection regret guarantees for a wide class of algorithms in a stochastic reward environment.

Stochastic CORRAL (see Algorithm 1) preserves most of the structure from CORRAL with 2 main differences. First, the base algorithms receive an unweighted feedback and updates their internal policy only when they are chosen and repeat their recommended policy to the master until they are chosen again. This ensures our algorithm is compatible with the internal workings of many algorithms without requiring major modifications. Intuitively, if the base is chosen every $c$ time steps where $c > 1$ is a constant, its regret can

\[ R(T) = \max_{\pi \in \Pi} \mathbb{E} \left[ \sum_{t=1}^T f(x_t, \pi_t) - \sum_{t=1}^T f(x_t, \pi^*) \right]. \]
Algorithm 1 Stochastic Corral

**Input:** Base Algorithms $\{B_i\}_{i=1}^M$, learning rate $\eta$.
Let $\pi_{t,i}$ be the policy of $B_i$ indexed by round $t$.
Let $\pi_{t,i}^{(1)}, \pi_{t,i}^{(2)}$ be the policies used by $B_i$ in round $t$.
Initialize $p_{t,i}$ to $p_{t,i}^1 = \frac{1}{M}$ for all $i \in [M]$.

For $t = 1, \ldots, T$ do

**Step 1**
- Receive context $x_{t,i}^{(1)} \sim D$.
- Receive policy $\pi_{t,i}^{(1)} = \pi_{t,i}^{(1)}$ from $B_i$.
- Play action $a_{t,i}^{(1)} \sim \pi_{t,i}^{(1)}(x_{t,i}^{(1)})$.
- Receive feedback $g_t^{(1)} = f(x_t^{(1)}, \delta_{a_t^{(1)}}) + \xi_t^{(1)}$.
- Update $B_i$ using $g_t^{(1)}$.

**Step 2**
- Receive context $x_{t}^{(2)} \sim D$.
- Sample $s \sim \text{Uniform}(0, \cdots, t)$.
- Receive policy $\pi_{t,s,i}^{(2)} = \pi_{s,s,i}$ from $B_i$.
- Play action $a_{t,i}^{(2)} \sim \pi_{t,s,i}^{(2)}(x_t^{(2)})$.
- Receive feedback $g_t^{(2)} = f(x_t^{(2)}, \delta_{a_t^{(2)}}) + \xi_t^{(2)}$.

Update $p_t$ using $g_t^{(1)} + g_t^{(2)}$ via the Corral Update. See Appendix or Algorithm 1 in Agarwal et al. (2017).

be upper bounded by $cR(T/c)$ because it updates its policy $T/c$ times and repeats a policy for $c$ time steps between two updates.

Second, we introduce a "smoothing" procedure which converts any algorithm to one with non-increasing instantaneous regret with high probability (Definition 4.2). The reason is as follows. Let $\{p_{t,i}^1, \ldots, p_{t,i}^T\}$ be the (random) probabilities that $M$ chooses the $i$-th base algorithm and let $\rho_{t,i}^* = \frac{1}{\min_i p_{t,i}^*}$. We need the instantaneous regret to decrease with high probability so that using $\min_i p_{t,i}^*$ at every time step is the worst case because the base will be updated the least often. Therefore the regret can be upper bounded by $E[\rho_{t,i}^* R(T/\rho_{t,i}^*)]$.

We use a two time step structure, Step 1 to update the policy of the base algorithm, and Step 2 to play a smoothed decision, ensuring its conditional instantaneous regret of Step 2 is upper bounded by a decreasing function with high probability. Details of the smoothing trick is given in Section 4.4.

Henceforth we refer to Algorithm 1 as Stochastic CORRAL and to CORRAL Agarwal et al. (2017) as Vanilla CORRAL. We use $\eta$ to denote the input learning rate of Stochastic CORRAL. The distribution $p_t$ is updated using a log barrier that follows the same update rules as in Vanilla CORRAL. We reproduce the full Vanilla CORRAL algorithm in Appendix A. We use $\mathcal{M}$ to denote the master algorithm. In the remainder we call each time indexed by $t$ a round. Each round is split in two steps of type 1 and 2. The master treats each round’s two rewards $g_t^{(1)}, g_t^{(2)}$ as one $g_t^{(1)} + g_t^{(2)}$.

Let $\{p_{t,1}^1, \ldots, p_{t,2}^T\}$ be the (random) probabilities that $M$ chooses the $i$-th base algorithm and let $\rho_{t,i}^* = \frac{1}{\min_i p_{t,i}^*}$.

We drop the superscript $i$ when it is clear. We use $n_{t,i}$ to denote the number of rounds base $i$ is chosen up to time $t$. Let $t_{i,j}$ be the round index of the $j$-th round the master chooses algorithm $B_i$ and let $b_{i,j} = t_{i,j} - t_{i,j-1}$ with $t_{i,0} = 0$ and $t_{i,j+1} = T + 1$. If a base algorithm is ran for $T$ rounds, we use $\mathcal{T}^{(j)}$ to denote its $T$ steps of type $j$ for $j \in \{1, 2\}$. We use $r_t^{(j)}$ to denote the master algorithm’s instantaneous regret in step $j$ of round $t$. Similarly, we denote by $x_{t}^{(1)}$ and $x_{t}^{(2)}$, $\pi_{t,i}^{(1)}$ and $\pi_{t,i}^{(2)}$ the contexts and policies used by the master during round $t$ step 1 and 2. Analogously we call $\pi_{t,i}^{(1)}$ and $\pi_{t,i}^{(2)}$ the policies proposed by base algorithm $B_i$ at time $t$, even when it is not selected ($i \neq i$) by $M$.

**Base repeated policies.** During the round when $B_i$ is not selected, we assume it repeats its future Step 2’s policy. More precisely for $j \in \{1, 2\}$ and $t = t_{l-1} + 1, \cdots, t_{l-1}$, $\pi_{t,i}^{(j)} = \pi_{t,i}^{(2)}$ for all $l \leq n_{t,i}^T + 1$. For all rounds $t$ and steps $j$, regardless of whether the master selected $B_i$ or not we denote base $i$ instantaneous regret by $r_{t,i}^{(j)}$.

Our main result implies the following:

**Theorem 3.1** (Informal). Let $\alpha \in [1/2, 1)$, if base algorithm $B_i$ satisfies a high probability regret bound $R_i(t) = O(t^\alpha)$ for all $t \in [T]$, the regret of $M$ when running with the smoothed version of $B_i$ satisfies,
\(R(T) \leq \tilde{O}\left(\frac{M}{\eta} + T\eta + T\eta^{1-\alpha}\right)\). Choosing \(\eta \approx \frac{M}{\alpha T}\) yields \(R(T) = \tilde{O}\left(M^{1-\alpha}T^\alpha\right)\).

## 4 Regret Analysis

In this section we analyze the regret of our Stochastic CORRAL algorithm. Our regret analysis follows a similar structure as in [Agarwal et al., 2017]. We split the regret in two terms (Section 4.2): the regret of the master algorithm with respect to a fixed base (I) and the regret of this base algorithm with respect to the optimal policy (II). Controlling term I makes use of the repeated policy structure of modification 1) and Lemma 13 of [Agarwal et al., 2017]. Bounding term II (Section 4.3) is the main focus of the regret analysis in this paper. In Section 4.1 we define the condition necessary for a base algorithm to have low regret while running with our Stochastic CORRAL master.

### 4.1 Non-increasing instantaneous regret

As explained above, we require the base algorithms to satisfy a smoothness condition ensuring an upper bound on the conditional instantaneous regret to be non-increasing. Since this condition need not be true for general bandit algorithms, we produce a generic procedure (Step 2 of the proposed algorithm) to modify an input base algorithm \(B\) into what we term a "smoothed" version \(\tilde{B}\) that satisfies it. Given an algorithm with concave (in \(t\)) cumulative regret bound \(U(t, \delta)\) that holds with high probability, we construct a new algorithm with instantaneous regret bound \(u(t, \delta) = U(t, \delta)/t\). With high probability, since \(U(t, \delta)\) is concave, \(u(t, \delta)\) will be non-increasing in \(t\).

The smoothed version \(\tilde{B}_i\) of a base algorithm \(B_i\) works as follows. We have two steps at each round \(t\). In step 1, we play \(B_i\). In step 2, at time \(t\), we pick a time step \(s \in [1, 2, ..., t]\) uniformly at random, and re-play the policy made by \(B_i\) at time \(s\). Since the policy of \(B_i\) at each round \([1, 2, ..., t]\) is chosen with probability \(1/t\) to be played at step 2, the instantaneous regret of step 2 at round \(t\) is \(1/t\) times the cumulative regret of \(B\) up to time \(t\).

The following three properties will ensure low regret for the overall algorithm: 1) The regret of Step 1 is bounded by \(U(t, \delta)\) with high probability (Definition 4.1). 2) Since the instantaneous regret of Step 2 is \(1/t\) times the cumulative regret of Step 1, the cumulative regret of Step 2 is bounded roughly by \(\sum_{t=1}^{T} 1/t \approx \log(T)\) times that of step 1. 3) The instantaneous regret of step 2 is \(U(t, \delta)/t\), which is non-increasing (Definition 4.2) if \(U(t, \delta)\) is concave. The master receives this feedback from Step 2.

Properties (1) and (2) ensure that regret of the smoothed version is low. Property (3) ensures that using the smallest \(\rho_j^t\) at time \(t\) results in the largest regret. Therefore, regret when running with a master can be upper bounded by \(E[\rho_i U(T/\rho_i^t)]\). We define these properties more precisely:

**Definition 4.1** \((U, \delta, S)\)-Boundedness. Let \(U : \mathbb{R} \times [0, 1] \rightarrow \mathbb{R}^+\) and \(S \subset [T]\). We say an algorithm \(B\) is \((U, \delta, S)\)-bounded if it is updated only on rounds \(S\) and with probability at least \(1 - \delta\) and for all rounds \(t \in S\), the cumulative pseudo-regret of rounds in \(S\) is bounded above by \(U(t, \delta)\):

\[
\sum_{j \in S, j \leq t} f(x_j^{(1)}, \pi^*) - f(x_j^{(1)}, \pi_j) \leq U(t, \delta), \forall t \in S. \tag{2}
\]

**Definition 4.2** \((U, \delta, T)\)-Smoothness. Let \(U : \mathbb{R} \times [0, 1] \rightarrow \mathbb{R}^+\). We say an algorithm \(\tilde{B}\) is \((U, \delta, T)\)-smooth if with probability \(\delta\) and for all rounds \(t \in [T]\), the conditional expected instantaneous regret of type 2 steps is bounded above by \(U(t, \delta)/t\). In other words, with probability \(1 - \delta\):

\[
E_{x_t \sim \mathbb{D}}[r_t^{(2)}|F_{t-1}] \leq \frac{U(t, \delta)}{t}, \forall t \in [T]. \tag{3}
\]

Here \(F_{t-1}\) denotes the sigma algebra of all randomness up to the beginning of round \(t\).

Throughout proofs we assume that base algorithms satisfy \((U, \delta'', T)\)-boundedness (on all type 1 steps) and the smoothed versions satisfy \((U, \delta, T)\)-smoothness for an appropriate function \(U\) and constants \(\delta, \delta'' \in [0, 1]\):
Assumption A1 (Base Boundedness and Smoothness) All input algorithms \( \{B_i\}_{i=1}^M \) are \((U_i, \frac{1}{\sqrt{T}}, T^{(1)})\)–bounded and their smooth versions \( \{\tilde{B}_i\}_{i=1}^M \) are \((U_i, \frac{1}{\sqrt{T}}, T^{(2)})\)–smooth where \( U_i(t, \delta) \) is a concave function in term of \( t \).

In Proposition 4.7 we show that the above assumption is satisfied if all base algorithms are \((U, \delta, [T])\)–bounded for an appropriate function \( U \).

Let \( E_i \) be the event that Equations 3 and 2 hold for all \( t \in [T] \). Throughout the paper we condition on the event \( E = \cap_{i=1}^M E_i \). A simple application of the union bound yields \( P(E) \geq 1 - \delta \). In the remainder of this section we prove our main result (Theorem 3.1).

4.2 Regret Decomposition

In order to analyze the regret of Algorithm 1 we use the same decomposition as in (Agarwal et al., 2017): we decompose the regret into the regret of the master algorithm with respect to base algorithm \( i \), and the regret of this base algorithm with respect to the optimal choice. Algorithm \( B_i \)'s internal state is updated only during steps of type 1 in rounds \( t \in T_i \). Recall that by our construction, a base algorithm repeats its recommended policy in rounds that it is not being selected, i.e. for all \( l \leq n_i \) and during both steps of rounds \( t = t_{i-1} + 1, \cdots, t_{i} - 1, \pi_{t_{i-1}}^{(1)} = \pi_{t_{i-1}}^{(2)} = \pi_{t_{i}}^{(2)} \). In what follows we drop the \( i \) subscript from \( \{t_{l,i}\} \) and \( \{b_{l,i}\} \) whenever clear. The following holds:

\[
R(T) = \max_{\pi \in \Pi} \mathbb{E} \left[ \sum_{t=1}^T \sum_{j=1}^2 f(x_t^{(j)}, \pi) - f(x_t^{(j)}, \pi_t^{(j)}) \right] = \mathbb{E} \left[ \sum_{t=1}^T \sum_{j=1}^2 f(x_t^{(j)}, \pi_t^{(1)}) - f(x_t^{(j)}, \pi_t^{(2)}) \right] + \mathbb{E} \left[ \sum_{t=1}^T \sum_{j=1}^2 f(x_t^{(j)}, \pi^*) - f(x_t^{(j)}, \pi_{t_{i}}^{(1)}) \right]
\]

(4)

We identify the maximizing policy, a deterministic object, with \( \pi^* \). We bound first the expectation of term I in Equation 4 (the regret of the master algorithm with respect to the base).

Term I can be upper bounded following Lemma 13 of Agarwal et al. (2017):

Lemma 4.1 (Lemma 13 of Agarwal et al. (2017)). We have

\[
\mathbb{E}[I] \leq O \left( \frac{M \ln T}{\eta} + T \eta \right) - \mathbb{E} \left[ \rho_\ast \right]
\]

Crucially, this result holds since the importance weighted update of \( p_t \), along with the base repeated policy structure ensures that the master’s loss estimates are indeed unbiased estimators of the base algorithm’s rewards. We discuss in more detail in the Appendix. The rest of the paper is devoted to bounding \( \mathbb{E}[II] \).

4.3 Main results

We split the proof of the main result of this section (Theorem 4.3) in two parts. First we show in Lemma 4.2 an upper bound on the base algorithm’s regret provided the \( p_t \) sequence is lower bounded with probability 1 by a constant \( p \). We leverage this result to prove Theorem 4.3 that shows a bound on the expected regret.
of an algorithm satisfying Assumption A1 whose invocations are controlled by a \( p'_l \) sequence resulting from running Stochastic CORRAL on top.

Let \( T_i \subset [T] \) be the set of rounds where base \( i \) is chosen and \( T^c_i = [T]\setminus T_i \). For \( S \subset [T] \) and \( j \in \{1, 2\} \), we define the regret of the base algorithm during Step \( j \) of rounds \( S \) as \( R_i^{(j)}(S) = \sum_{t \in S} f(x^{(j)}_t, \pi^*) - f(x^{(j)}_t, \pi_{i,t}) \).

The following decomposition of \( E [II] \) holds:

\[
E [II] = E \left[ R_i^{(1)}(T_i) + R_i^{(2)}(T_i) + R_i^{(1)}(T^c_i) + R_i^{(2)}(T^c_i) \right] .
\]

\( R_i^{(1)}(T_i) \) consists of the regret when the base was updated in Step 1 while the remaining 3 terms consists of the regret when the policies are reused. Since \( E [II] \leq E [III(E)] + \delta T \), we focus on bounding \( E [III(E)] \).

Under Assumption A1, \( E [R_i^{(1)}(T_i)1(E)] \leq E [U_i(\delta, n^{(1)}_T)1(E)] \). We proceed to bound the regret corresponding to the remaining terms in \( II_0 \):

\[
E [II_01(E)] = E \left[ \sum_{i=1}^{n^+_1+1} 1\{E\}(2b_i - 1)E \left[ r_i^{(2)(E)}_{i,t-1} \right] \right] \\
\leq E \left[ \sum_{i=1}^{n^+_1+1} 1\{E\}(2b_i - 1)\frac{U_i(l, \delta/2M)}{l} \right] .
\]

The multiplier \( 2b_i - 1 \) arises because the policies proposed by the base algorithm during the rounds it is not selected by \( M \) satisfy \( \pi_{t,\cdot} = \pi_{t,\cdot} = \pi_{t,\cdot}^{(2)} \) for all \( l \leq n^+_T + 1 \) and \( t = t_{l-1} + 1, \ldots, t_l - 1 \). The factorization is a result of conditional independence between \( E \left[ r_i^{(2)(E)}_{i,t-1} \right] \) and \( E [b_{i,l}|F_{t-1}] \) where \( F_{t-1} \) already includes algorithm \( \tilde{B}_i \) update right after round \( t_{l-1} \). The inequality holds because \( \tilde{B}_i \) is \( (U_i, \delta/2M, T^{(2)}) \)-smooth and therefore satisfies Equation 8 on event \( E \).

**Lemma 4.2 (Fixed \( p \)).** If \( \frac{1}{p} = p \leq p'_1, \ldots, p'_T \) with probability one, then, \( E [II] \leq 4\rho U_i(T/\rho, \delta) \log T + \delta T \).

Since the conditional instantaneous regret (Definition 4.2) has a non-increasing upper bound, using \( p \) at every time step will result in the largest upper bound on its regret because the base is updated the least often (see length of \( b_i \) intervals in Eq. 5). In this case the base will be updated every \( \rho \) time-steps and the regret upper bound will be roughly \( \rho U_i(T/\rho, \delta) \). The proof is in Appendix D.

Notice the bound in Lemma 4.2 in addition to Lemma 4.1 would yield a regret guarantee for Stochastic Corral vs the base algorithm in terms of a deterministic lower bound \( p \) for the probabilities \( p'_1, \ldots, p'_T \). This is of course unsatisfactory because these probability values are random. We use a restarting trick to address this concern. **Restarting trick:** Initialize \( p = \frac{T}{TM} \). If \( p'_l < p \), set \( p = \frac{p'_l}{2} \) and restart the base.

Therefore the time horizon is divided into phases, and in each phase the lower bound \( p \) is deterministic. We provide the analysis below:

**Theorem 4.3.** [Path dependent \( p \)] When we run the base algorithm with the CORRAL master algorithm, and restart the base every time Line 10 of the vanilla CORRAL algorithm [Agarwal et al. 2017] is executed (as described above), \( E [II] \leq 4E [\rho U_i(T/\rho, \delta) \log T] + \delta T(\log T + 1) \).

Here, the expectation is over the random variable \( \rho^* = \max_l \frac{1}{p'_l} \). If \( U(t, \delta) = t^\alpha g(\delta) \) for some function \( g : \mathbb{R} \to \mathbb{R}^+ \), and \( \alpha \in [1/2, 1) \) then, \( I \leq 4\frac{2^{1-n}}{\alpha(1-\alpha)}T^{\alpha}g(\delta)E \left[ \rho^{1-n}_l \right] + \delta T(\log T + 1) \).

The proof in Appendix E follows that of Theorem 15 in [Agarwal et al. 2017]. Putting it all together we conclude our main theorem:
Theorem 4.4. Let \( U(t, \delta) = t^\alpha g(\delta) \) for some \( 0 \leq \alpha \leq 1 \) and some function \( g : \mathbb{R} \to \mathbb{R}^+ \). If Algorithm \( B_i \) satisfies \((U, \frac{1}{\Delta}, T^{(1)})\)-boundedness and its smooth version is \((U, \frac{1}{\Delta}, T^{(2)})\)-smooth, stochastic CORRAL with the restarting trick satisfies:

\[
R(T) \leq O \left( \frac{M \ln T}{\eta} + T \eta \right) - E \left[ \frac{\rho_\ast}{40 \eta \ln T} - 2 \rho_\ast U(T/\rho_\ast, \delta) \log T \right] + \delta T.
\]

Proof. The result follows from Equation \[4.1\] and the bounds of Lemma \[4.1\] and Theorem \[4.3\].

Maximizing over \( \rho_\ast \) gives us the following worst-case bound:

Corollary 4.5. If a base algorithm is \((U, \delta, T^{(1)})\)-bounded and its smooth version is \((U, \delta, T^{(2)})\)-smooth for \( U(T, \delta) = T^\alpha g(\delta) \) for some \( \alpha \in [1/2, 1) \), then the regret of the master algorithm is bounded as

\[
R(T) \leq \widetilde{O} \left( \frac{M}{\eta} + T \eta + T g(\delta) \frac{1}{\eta} \frac{1}{\rho_\ast} \right) + \delta T.
\]

When \( \eta = \frac{M^\alpha}{g(\delta) T^\alpha} \) then \( R(T) \leq \widetilde{O} \left( M^{1-\alpha} g(\delta) T^\alpha \right) + \delta T \).

The proof is in Appendix \[3\].

In Section \[6\] we show explicit bounds for some applications.

4.4 Algorithm smoothing

Recall from Section \[4.1\] that the smoothed version \( \mathcal{B}_i \) of a base algorithm \( B_i \) works as follows. In step 1, we play \( B_i \) and use the feedback to update its internal structure. In step 2, at time \( \ell \), we uniformly pick at random a time step \( s \) in \([1, 2, \ldots, \ell]\), and re-play the policy that was made by \( B_i \) in Step 1 at time \( s \). Since the policy of \( B_i \) at each time step \([1, 2, \ldots, \ell]\) is chosen with probability \( 1/\ell \) to be played at step 2, the instantaneous regret of step 2 at time \( \ell \) is \( 1/\ell \) of the cumulative regret of \( B_i \) up to time \( \ell \).

Note that we are re-playing the decision of \( B_i \) at time \( s \) learned from a sequence of contexts \( x^{(1)}_1, \ldots, x^{(1)}_s \) to another context \( x^{(2)}_\ell \). Since the contexts are sampled i.i.d from the same distribution, in Lemma \[4.6\] we will show that when we reuse the policy learned from a series of contexts \( x_1, \ldots, x_i \) to another series of context \( x'_1, \ldots, x'_i \), the regret is multiplied only by a constant factor. We call the regret when using a policy learned from a series of context to another series of contexts "replay regret".

Definition 4.3 (Expected Replay Regret). Let \( h \) be a generic history of algorithm \( B \) and \( h(t) \) the history \( h \) up to time \( t \). If \( x_1, \ldots, x_t \) are i.i.d. contexts from \( D \) and \( \pi_1, \ldots, \pi_t \) is the sequence of policies used by \( B \) on these contexts, the "expected replay regret" \( R(t, h) \) is:

\[
R(t, h) = \mathbb{E}_{x'_1, \ldots, x'_t} \left[ \sum_{l=1}^{t} f(x'_l, \pi_\ast) - f(x'_l, \pi_l) \right]
\]

Lemma 4.6. If \( B \) is \((U, \delta, [T])\)-bounded, \( \max_{x, \pi} |f(x, \pi)| \leq 1 \), \( U(t, \delta) > 8 \sqrt{\log \left( \frac{\ell \delta}{\rho_\ast} \right)} \), and \( \delta \leq \frac{1}{\sqrt{\rho_\ast}} \), then \( B \)'s expected replay regret satisfies:

\[
R(t, h) \leq 4U(t, \delta) + 2\delta t \leq 5U(t, \delta).
\]

Lemma \[4.6\] is a consequence of a simple martingale concentration bound (The proof is in Appendix \[5\]). We now present the main result of this section:

Proposition 4.7. If \( U(t, \delta) > 8 \sqrt{t \log \left( \frac{\ell \delta}{\rho_\ast} \right)} \), \( \delta \leq \frac{1}{\sqrt{\rho_\ast}} \) and \( B \) satisfies \((U, \delta, [T])\)-boundedness, then \( \mathcal{B} \) is \((5U, \delta, T^{(2)})\)-smooth.
Proof. Since the conditional instantaneous regret on Step 2 of round $t$ equals the average replay regret of the type 1 steps up to $t$, Lemma 4.6 implies $E[r_t^{(2)} | F_{t-1}] \leq \frac{5U(t, \delta)}{t}$.

Consequently:

Corollary 4.8 (Informal). All $(U, \delta, [T])$-bounded algorithms can be smoothed and used with Stochastic Corral.

In Appendix [H] and [I] we show that several algorithms are $(U, \delta, [T])$-bounded for appropriate functions $U$.

Lemma 4.9. Assuming that the noise $\xi_t$ is conditionally 1-sub-Gaussian, UCB is $(U, \delta, [T])$-bounded with $U(t, \delta) = O(\sqrt{tK \log \frac{t}{\delta}})$.

Lemma 4.10 (Theorem 3 in [Abbasi-Yadkori et al., 2011]). LinUCB is $(U, \delta, [T])$-bounded with $U(t, \delta) = O(d\sqrt{t \log(1/\delta)})$.

Lemma 4.11. If $c = \frac{10K \log(\frac{1}{\delta})}{\Delta^2}$ where $\Delta_j$ is the gap between the optimal arm and arm $j$ and $\Delta_* = \min_j \Delta_j$, then $\epsilon$-greedy with $\epsilon_t = \frac{c}{t}$ satisfies a $(U, \delta, [T])$-bounded for $\delta \leq \frac{\Delta^2}{T}$ and:

1. $U(t, \delta) = 16\sqrt{\log(\frac{1}{\delta})}t$ when $K = 2$.
2. $U(t, \delta) = 20 \left( K \log(\frac{1}{\delta}) \left( \sum_{j=2}^{K} \Delta_j \right) \right)^{1/3} t^{2/3}$ when $K > 2$.

Lemma 4.12 (Theorem 1 in [Seldin et al., 2013]). Exp3 is $(U, \delta, [T])$-bounded where $U(t, \delta) = O(\sqrt{K t \log \frac{1}{\delta}})$.

5 Lower bound

In stochastic environments with sufficiently large “gap”, algorithms such as UCB achieve logarithmic regret bounds. Our model selection procedure has a $O(\sqrt{T})$ overall regret even in stochastic problems. In this section, we show that in general it is impossible to obtain a regret better than $\Omega(\sqrt{T})$. More specifically, we construct an example in which there are 2 base algorithms, one of which has 0 regret, and show that when running these 2 base algorithms with any master, it is impossible to have better than $\tilde{\Omega}(\sqrt{T})$ regret.

Theorem 5.1. There exists an algorithm selection problem, such that the regret for any time $T$ is lower bounded by $R(T) = \Omega(\sqrt{T \log(T)})$.

Proof sketch, full proof in Appendix [K]. The two base algorithms are constructed such that the gap between the algorithms closes at a rate of $\Theta(1/\sqrt{t \log(t)})$. We show that at this rate, any master will have a constant probability of misidentifying the optimal algorithm even after observing infinite pulls. Hence the regret of the master is of order $\Omega\left( \sum_{t=1}^{T} 1/(\sqrt{t \log(t)}) \right) = \tilde{\Omega}(\sqrt{T})$.

6 Applications

In this section, we show two applications of our results. First, we show how the results can be used to combine contextual and non-contextual stochastic algorithms and match the regret lower bound. Second, we design a method to find a near optimal exploration rate for $\epsilon$-greedy in an adaptive fashion.
6.1 Contextual vs non-contextual UCB

The regret of UCB \cite{lattimore2020bandit} for \( k \)-armed bandit problem is \( \tilde{O}(\sqrt{kT}) \) where \( k \) is the number of arms. The regret of LinUCB \cite{lattimore2020bandit} for linear bandit problem is \( \tilde{O}(d\sqrt{T}) \) where \( d \) is the dimension of the context vectors. In this section we show how to run our Stochastic CORRAL with UCB and LinUCB as base algorithms and achieve the regret matching the lower bound.

Lemma 6.1 (Implied by Theorem 24.4 in \cite{lattimore2020bandit}). Let \( R_\nu(T) \) denote the cumulative regret at time \( T \) on bandit environment \( \nu \). For any algorithm there exist a 1-dimensional linear bandit environment \( \nu_1 \) and a \( k \)-armed bandit environment \( \nu_2 \) such that: \( R_{\nu_1}(T) \cdot R_{\nu_2}(T) \geq T(k-1)e^{-2} \). Without knowing the environment, the regret is at least \( \max\{R_{\nu_1}(T),R_{\nu_2}(T)\} \).

We show that the regret of Stochastic CORRAL with base algorithms LinUCB and UCB matches the lower bound. The proof is in Section J.2.

Theorem 6.2. The regret of stochastic CORRAL with base algorithms LinUCB and UCB and rate \( \eta = \sqrt{\frac{2}{Td\sqrt{k}}} \) is upper bounded by

\[ \tilde{O}\left(\sqrt{2T} \left( d^{0.5}k^{0.25} + d^{0.75}k^{0.25} \right) \right). \]

In terms of dependence in \( k \) and \( T \), the product of the two terms in the bound matches the lower bound in Lemma 6.1 (the product of the two terms being of order \( kT \)).

As we mentioned earlier, Chatterji et al. (2019) and Foster et al. (2019) study related problems. However, the assumptions in Chatterji et al. (2019) appear to be too strong, while the regret bounds in Foster et al. (2019) are not optimal in scaling with \( T \). Our approach has the additional advantage that it can also handle model mis-specification.

6.2 Tuning the exploration rate of \( \epsilon \)-greedy

For a given positive constant \( c \), the \( \epsilon \)-greedy algorithm pulls a random arm in round \( t \) with probability \( \epsilon_t = c/t \), and otherwise pulls the arm with the largest empirical average reward. It can be proven that the optimal value for \( \epsilon_t \) is \( \min(1, \frac{5K}{\Delta_*^2}) \) where \( \Delta_* \) is the smallest gap between the optimal arm and the sub-optimal arms. We would like to find the optimal value of \( c \) without knowing \( \Delta_* \). In this section we will use stochastic CORRAL to find the best \( c \).

Given the time horizon \( T \) as an input, we divide the interval \([1, KT]\) into an exponential grid \([1, 2, 2^2, ..., 2^{\log(KT)}]\). We use \( \epsilon \)-greedy with each value of \( c \) in the grid as a base algorithm for CORRAL. The following theorem shows the regret of CORRAL using these base algorithms. The proof is in Appendix J.1.

Theorem 6.3. The regret of stochastic CORRAL using \( \epsilon \)-greedy base algorithms defined on the grid with \( \eta \) chosen as in Corollary 4.5 is bounded by \( \tilde{O}(T^{2/3}) \) when \( K > 2 \), and by \( \tilde{O}(T^{1/2}) \) when \( K = 2 \).

7 Experiments

Now we show results of running the proposed algorithm (referred to as Corral in the plots) in a couple of scenarios. We do this in two settings. The first problem we consider is the choice of \( \epsilon \) parameter in \( \epsilon \)-greedy. The second problem we consider is model selection between UCB and LinUCB. In all the experiments, we take the initial learning rate of CORRAL to be \( \eta = 20/\sqrt{T} \). We repeat each experiment several times and the shading denotes the mean squared error. Note that we have not implemented smoothing of the base algorithms in these experiments.
For $\epsilon$-greedy, we consider the case of 2 Bernoulli arms with means $p_1 = 0.5$ and $p_2 = 0.5 - \Delta$. We consider two cases, $\Delta = 0.1$ and $\Delta = 0.05$. We consider eighteen base algorithms differing in their choice of $\epsilon$ in the exploration rate $\epsilon_t = \epsilon/t$. We take $T = 100,000$ and $\epsilon$'s to lie on a geometric grid in $[1, 2T]$. The results are shown in Figure 1. While performance of $\epsilon$-greedy with a fixed $\epsilon$ can be sensitive to the environment, CORRAL shows stability and a relatively good performance in different environments.

![Figure 1: $\epsilon$-Greedy. Plot of cumulative regret as a function of time (t).](image)

In the other experiment, we take UCB and LinUCB as the base algorithms, and consider a contextual bandit environment. The contexts are independent and identically distributed. We again have two arms, and each arm $i$ has an associated vector $\theta_i \in \mathbb{R}^{d-1}$. We also let $x_i \sim N(0, 1)$ and independent of each other. We let the reward of arm $i$ be $\mu_i + x_i^\top \theta_i + \eta_i$, where $\eta_i \sim N(0, c)$, for some $c$. Regret is defined as in the contextual case. We consider two different choices of $\theta_i$ and $c$: in the first case, $\theta_i$ is small, which suggests a non-contextual algorithm might perform better. The two plots are given in Figure 2.

![Figure 2: UCB vs. LinUCB. Plot of cumulative regret as a function of time (t). Parameters: $d = 10$, $K = 2$.](image)

The plots indicate the master has sub-linear regret. More importantly, the regret of the master lies in
between the best and worst base algorithms.

8 Conclusions

We study the bandit model-selection problem in stochastic environments. Our approach is general and applicable to a diverse set of stochastic base algorithms. We introduce a smoothing trick that is applicable under mild conditions and can transform a bandit algorithm so that the algorithm is stable with respect to frequency of updates. The smoothing trick allows us to perform model selection using an online mirror descent as the master algorithm. Finally, we perform empirical studies in synthetic environments and we show the effectiveness of the approach in two cases: tuning the exploration rate of $\epsilon$-greedy, and model selection between contextual and non-contextual UCB algorithms.

References

Y. Abbasi-Yadkori, P. Bartlett, V. Gabillon, A. Malek, and M. Valko. Best of both worlds: Stochastic and adversarial best-arm identification. In In Proceedings of the International Conference on Computational Learning Theory (COLT), 2018.

Yasin Abbasi-Yadkori, David Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In NIPS, 2011.

Yasin Abbasi-Yadkori, David Pál, and Csaba Szepesvári. Online-to-confidence-set conversions and application to sparse stochastic bandits. In AISTATS, 2012.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E. Schapire. Corralling a band of bandit algorithms. In COLT, volume 65 of Proceedings of Machine Learning Research, pages 12–38. PMLR, 2017.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In In Proceedings of the International Conference on Computational Learning Theory (COLT), 2009.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. Theoretical Computer Science, 410(19):1876–1902, 2009.

P. Auer and C.-K. Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In In Proceedings of the International Conference on Computational Learning Theory (COLT), 2016.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. Mach. Learn., 47(2-3):235–256, 2002.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits, 2017.

S. Bubeck and A. Slivkins. The best of both worlds: stochastic and adversarial bandits. In In Proceedings of the International Conference on Computational Learning Theory (COLT), 2012.

Alexandra Carpentier and Remi Munos. Bandit theory meets compressed sensing for high-dimensional stochastic linear bandit. In AISTATS, 2012.

Niladri Chatterji, Vidya Muthukumar, and Peter L. Bartlett. Osom: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In arXiv, 2019.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In COLT, 2008.
Dylan Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. In *Advances in Neural Information Processing Systems*, 2019.

Sampath Kannan, Jamie Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS’18, pages 2231–2241, USA, 2018. Curran Associates Inc. URL [http://dl.acm.org/citation.cfm?id=3327144.3327150](http://dl.acm.org/citation.cfm?id=3327144.3327150).

Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW*, 2010.

Odalric-Ambrym Maillard and Rémi Munos. Adaptive bandits: Towards the best history-dependent strategy. In AISTATS, 2011.

Y. Seldin and A. Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *In Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

Yevgeny Seldin, Csaba Szepesvari, Peter Auer, and Yasin Abbasi-Yadkori. Evaluation and analysis of the performance of the exp3 algorithm in stochastic environments. In Marc Peter Deisenroth, Csaba Szepesvári, and Jan Peters, editors, *Proceedings of the Tenth European Workshop on Reinforcement Learning*, volume 24 of *Proceedings of Machine Learning Research*, pages 103–116, Edinburgh, Scotland, 30 Jun–01 Jul 2013. PMLR. URL [http://proceedings.mlr.press/v24/seldin12a.html](http://proceedings.mlr.press/v24/seldin12a.html).

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In AISTATS, 2019.
A Corral Algorithm

A.1 Original Corral

The original Corral algorithm (Agarwal et al., 2017) is reproduced below.

**Algorithm 2 Original Corral**

**Input:** Base Algorithms \(\{B_i\}^M_{i=1}\), learning rate \(\eta\).

Initialize: \(\gamma = 1/T, \beta = e^{\ln T}, \eta_i, \rho^i_1 = 2M, p^i_1 = \frac{1}{M}, p^i_1 = 1/M\) for all \(i \in [M]\).

Initialize all base algorithms.

for \(t = 1, \ldots, T\) do

Receive context \(x_t \sim \mathcal{D}\).

Receive policy \(\pi_t,i\) from \(B_i\) for all \(i \in [M]\).

Sample \(i_t \sim p_t\).

Play action \(a_t \sim \pi_t,i_t(x_t)\).

Receive feedback \(g_t = f(x_t, \delta_{a_t}) + \xi_t\).

Send feedback \(g_{p_t,i_t}\) to \(B_i\) for all \(i \in [M]\).

Update \(p_t, \eta_t\) and \(p_{t+1}\) using Corral-Update.

for \(i = 1, \ldots, M\) do

Set \(\rho^i_{t+1} = \frac{1}{p^i_{t+1}}\).

Return \(p_{t+1}, \eta_{t+1}\) and \(p_{t+1}\).

**Algorithm 3 Corral-Update**

**Input:** learning rate vector \(\eta_t\), distribution \(p_t\), lower bound \(p_{t+1}\) and current loss \(g_t\)

**Output:** updated distribution \(\pi_{t+1}\), learning rate \(\eta_{t+1}\) and loss range \(\rho_{t+1}\)

Update \(p_{t+1} = \text{Log-Barrier-OMD}(p_t, \eta_t, \ell_t, \eta_t)\).

Set \(p_{t+1} = (1 - \gamma)p_{t+1} + \gamma \frac{1}{M}\).

for \(i = 1, \ldots, M\) do

if \(p^i_t > p^i_{t+1}\) then

Set \(p^i_{t+1} = \frac{p^i_{t+1}}{2}, \eta_{t+1,i} = \beta \eta_{t,i}\),

else

Set \(p^i_{t+1} = p^i_{t}, \eta_{t+1,i} = \eta_{t,i}\).

Return \(p_{t+1}, \eta_{t+1}\) and \(p_{t+1}\).

**Algorithm 4 Log-Barrier-OMD\(\left(p_t, \ell_t, \eta_t\right)\)**

**Input:** learning rate vector \(\eta_t\), previous distribution \(p_t\) and current loss \(\ell_t\)

**Output:** updated distribution \(p_{t+1}\)

Find \(\lambda \in [\min_i \ell_t,i, \max_i \ell_t,i]\) such that \(\sum_{i=1}^{M} \frac{1}{p^i_t + \eta_{t,i}(\ell_t,i - \lambda)} = 1\)

Return \(p_{t+1}\) such that \(\frac{1}{p^i_{t+1}} = \frac{1}{p^i_t} + \eta_{t,i}(\ell_t,i - \lambda)\)

A.2 Stochastic Corral

We reproduce our stochastic Corral algorithm below for reference.
Algorithm 5 Stochastic Corral

**Input:** Base Algorithms \(\{\mathcal{B}_i\}_{i=1}^M\), learning rate \(\eta\).

Let \(\pi_{t,i}\) be the policy of \(\mathcal{B}_i\) indexed by round \(t\).
Initialize: \(\gamma = 1/T, \beta = e^{\frac{1}{2\eta^2}}, \eta_{t,i} = \eta, p_i^t = 1/M, p_i^t = \frac{1}{2\eta^2}\) for all \(i \in [M]\).

for \(t = 1, \ldots, T\) do

**Step 1**
- Sample \(i_t \sim p_t\).
- Receive context \(x_t^{(1)} \sim \mathcal{D}\).
- Receive policy \(\pi_{t,i_t}^{(1)} = \pi_{t,i_t}\) from \(\mathcal{B}_{i_t}\).
- Play action \(a_t^{(1)} = \pi_{t,i_t}^{(1)}(x_t^{(1)})\).
- Receive feedback \(g_t^{(1)} = f(x_t^{(1)}, \delta_{a_t^{(1)}}) + \xi_t^{(1)}\).
- Update \(\mathcal{B}_{i_t}\) using \(g_t^{(1)}\).

**Step 2**
- Receive context \(x_t^{(2)} \sim \mathcal{D}\).
- Sample \(s \sim \text{Uniform}(0, \ldots, t)\).
- Receive policy \(\pi_{t,i_t}^{(2)} = \pi_{s,i_t}\) from \(\mathcal{B}_{i_t}\).
- Play action \(a_t^{(2)} = \pi_{t,i_t}^{(2)}(x_t^{(2)})\).
- Receive feedback \(g_t^{(2)} = f(x_t^{(2)}, \delta_{a_t^{(2)}}) + \xi_t^{(2)}\).
- Update \(p_t, \eta_t\) and \(p_i^t\) to \(p_{t+1}, \eta_{t+1}\) and \(p_i^{t+1}\) using \(g_t^{(1)}, g_t^{(2)}\) via Corral-Update.

### B Some useful lemmas

**Lemma B.1.** If \(U(t, \delta) = t^\beta g(\delta)\), for \(0 \leq \beta \leq 1\) then:

\[
U(l, \delta) \leq \sum_{t=1}^{l} \frac{U(t, \delta)}{t} \leq \frac{1}{\beta^2} U(l, \delta)
\]

**Proof.** The LHS follows immediately from observing \(\frac{U(t, \delta)}{t}\) is decreasing as a function of \(t\) and therefore \(\sum_{t=1}^{l} \frac{U(t, \delta)}{t} \geq l \frac{U(l, \delta)}{l} = U(l, \delta)\). The RHS is a consequence of bounding the sum by the integral \(\int_{0}^{l} \frac{U(t, \delta)}{t} dt\), substituting the definition \(U(t, \delta) = t^\beta g(\delta)\) and solving it.

**Lemma B.2.** If \(f(x)\) is a concave and doubly differentiable function on \(x > 0\) and \(f(0) \geq 0\) then \(f(x)/x\) is decreasing on \(x > 0\).

**Proof.** In order to show that \(f(x)/x\) is decreasing when \(x > 0\), we want to show that \(\left(\frac{f(x)}{x}\right)' = \frac{x f'(x) - f(x)}{x^2} < 0\) when \(x > 0\). Since \(0 f'(0) - f(0) \leq 0\), we will show that \(g(x) = x f'(x) - f(x)\) is a non-increasing function on \(x > 0\). We have \(g'(x) = x f''(x) \leq 0\) when \(x \geq 0\) because \(f(x)\) is concave. Therefore \(x f'(x) - f(x) \leq 0 f'(0) - f(0) \leq 0\) for all \(x \geq 0\), which completes the proof.

**Lemma B.3.** For any \(\Delta \leq \frac{1}{4}\), \(kl\left(\frac{1}{2}, \frac{1}{2} - \Delta\right) \leq 3\Delta^2\).

**Proof.** By definition \(kl(p, q) = p \log(p/q) + (1 - p) \log(\frac{1 - p}{1 - q})\), so

\[
kl\left(\frac{1}{2}, \frac{1}{2} - \Delta\right) = \frac{1}{2} \left(\log\left(\frac{1}{1 - 2\Delta}\right) + \log\left(\frac{1}{1 + 2\Delta}\right)\right)
\]

\[
= \frac{1}{2} \log\left(\frac{1}{1 - 4\Delta^2}\right) = \frac{1}{2} \log\left(1 + \frac{4\Delta^2}{1 - 4\Delta^2}\right) \leq \frac{2\Delta^2}{1 - 4\Delta^2} \leq \frac{2\Delta^2}{\frac{3}{4}} \leq 3\Delta^2
\]

\[\square\]
C Bounding $E[I]$

Recall $I = \sum_{t=1}^{T} \sum_{j=1}^{2} f(x_t^{(j)}, \pi_t^{(j)}) - f(x_t^{(j)}, \pi_t^{(j)})$ and that $\mathbb{T}_t$ equals the subset of random rounds where $\mathcal{M}$ listened to $\mathcal{B}_i$ ($i_t = i$). We split this term in two as follows:

\[
I = \sum_{t=1}^{T} \sum_{j=1}^{2} f(x_t^{(j)}, \pi_t^{(j)}) - f(x_t^{(j)}, \pi_t^{(j)})
= \sum_{t\in\mathbb{T}_1} \sum_{j=1}^{2} f(x_t^{(j)}, \pi_t^{(j)}) - f(x_t^{(j)}, \pi_t^{(j)}) + \sum_{t\in\mathbb{T}_2} \sum_{j=1}^{2} f(x_t^{(j)}, \pi_t^{(j)}) - f(x_t^{(j)}, \pi_t^{(j)})
\]

Equality $(i)$ holds because term $I_0$ equals zero and therefore $I_0 = I_0'$ and in all steps $t \in \mathbb{T}_1$, base $i$ repeated a policy of Type 2 so that $I_1 = I_1'$. Equality $(ii)$ follows from adding and subtracting term $I_B$. We now focus on bounding $E[I_A]$ and $E[I_B]$.

C.1 Bounding $E[I_A]$

Notice that:

\[
E[I_A] = E \left[ \sum_{t=1}^{T} 2f(x_t^{(2)}, \pi_t^{(2)}) - 2f(x_t^{(2)}, \pi_t^{(2)}) \right]
\]

We can easily bound this term using Lemma 13 from [Agarwal et al. 2017]. Indeed, in term $I_A$, the policy choice for all base algorithms $\{\mathcal{B}_m\}_{m=1}^{M}$ during any round $t$ is chosen before the value of $i_t$ is revealed. This ensures the estimates $\frac{2g}{p_t}$ and $0$ for all $i \neq i_t$ are indeed unbiased estimators of the base algorithm’s rewards.

We conclude:

\[
E[I_A] \leq O \left( \frac{M \ln T}{\eta} + T\eta \right) - \frac{E[\rho]}{40\eta \ln T}
\]

C.2 Bounding $E[I_B]$

Notice that:

\[
E[I_B] = E \left[ \sum_{t\in\mathbb{T}_1} f(x_t^{(1)}, \pi_t^{(1)}) - f(x_t^{(1)}, \pi_t^{(1)}) \right]
= E \left[ \sum_{t\in\mathbb{T}_1} f(x_t^{(1)}, \pi_t^{(1)}) - f(x_t^{(1)}, \pi_t^{(1)}) \right]
\]

In order to bound this term we will make an extra assumption. We avoided this discussion in the main to avoid deterring the reader from the main discussion: our bound on term $E[I]$.
Assumption A2 (Bounded rewards) We assume the norm of the rewards is bounded by 1. \footnote{This can be relaxed to subgaussianity}

Assumption A3 (Sub-optimal policy) The learner has access to a sub-optimal policy $\pi$ with a known lower bound in its sub-optimality gap of $\frac{1}{2}$.

We will show that under the right definition for Steps 2, term $E[I_B] \leq 0$. We will need to modify slightly the definition of the policy played during steps 2. Instead of playing the "original" $(U_m, \frac{\delta}{M}, T^{(2)})$-smooth policy given by any base algorithm $m$, we play a mixture policy. If we denote by $\pi_{t,m}^{(2)}(0)$ the original smooth policy of $\tilde{B}_m$ at round $t$, step 2, we now declare instead $\pi_{t,m}^{(2)}$ to be $\alpha_t \pi + (1 - \alpha_t) \pi_{t,m}^{(2)}(0)$ for $\alpha_t = \min(1, \frac{8U_m(t, \delta)}{M \delta})$. We also declare that $\pi_{m,t}^{(2)}$ be now this mixture policy instead. Crucially this is possible because the base update is performed based on the feedback obtained from steps of type (1) so that this redefinition does not affect the previous bound on $E[I_A]$.

Notice that under this new definition for $\pi_{t,m}^{(2)}$, if an algorithm $\tilde{B}$ satisfies $(U_m, \frac{\delta}{M}, T^{(2)})$-smoothness then with probability at least $1 - \frac{1}{M}$ its conditional instantaneous regret satisfies both the following upper and lower bounds during all rounds $t \in T_m$:

$$
\frac{1}{2} \min \left( 1, \frac{8U_m(t, \frac{\delta}{M})}{M} \right) \leq E_{x_t \sim D} \left[ r_t^{(2)}(1) | F_{t-1} \right] \leq \frac{6U_m(t, \frac{\delta}{M})}{M}, \quad \forall t \in T_m
$$

Notice that by definition of this modified Step 2 policy and under Assumption A3, the lower bound on the conditional expectation of the instantaneous conditional regret holds in Equation \footnote{This can be relaxed to subgaussianity} holds. For convenience, in the remainder of this section we will use this definition of $(U_m, \frac{\delta}{M}, T^{(2)})$-smoothness. We revert back to its usual definition in the later sections.

We now show Equation \footnote{This can be relaxed to subgaussianity} implies that for any base algorithm $m \in [M]$ the cumulative regret of terms of Steps 2 upper bounds the cumulative regret of Steps 1 provided $\tilde{B}_m$ satisfies $(U_m, \frac{\delta}{M}, T^{(1)})$-boundedness. We use these observations to bound term $E[I_B]$.

Let’s introduce a convenient notational definition, for all $m \in [M]$ let $T_m(l)$ be the first set of $l$ rounds where $i_l = m$. It is easy to see that:

**Proposition C.1.** If $\tilde{B}_m$ is $(U_m, \frac{\delta}{M}, T^{(1)})$-bounded and $(U_m, \frac{\delta}{M}, T^{(2)})$-smooth with $U_m(t, \frac{\delta}{M}) = t^\beta g(\frac{\delta}{M})$:

$$
E \left[ \sum_{l \in T_m} I_B^l \right] \leq \frac{\delta}{M} T - 3E \left[ U_m([T_m], \delta) | E_m \right] + \frac{D_m}{2} - \frac{C_m}{2}.
$$

Where $C_m$ is the index of the first round where $\frac{8U_m(t, \frac{\delta}{M})}{M \delta} \leq 1$ and $D_m = \sum_{l=1}^{T_m} \frac{8U_m(l, \frac{\delta}{M})}{l \delta}$.

**Proof.** We start by conditioning on the event given to us by the boundedness and smoothness assumptions, let’s call it $E_m$. On the complementary event $E_m^c$ (an event with probability $\frac{\delta}{M}$) we pay a linear regret of $T$. In order to upper bound the expected regret on $E_m$ we make use of the observations above.

As a consequence of boundedness, for all $l \leq T_m$ conditioned on $E_m$, $\sum_{l \in T_m(l)} f(x_t^{(1)}, \pi^*) - f(x_t^{(1)}, \pi_t^{(1)}) \leq U_m(l, \frac{\delta}{M})$. Therefore,

$$
E \left[ \sum_{l \in T_m(l)} f(x_t^{(1)}, \pi^*) - f(x_t^{(1)}, \pi_t^{(1)}) \right] \leq U_m \left( l, \frac{\delta}{M} \right) | E_m \right] \leq E \left[ U_m \left( [T_m], \frac{\delta}{M} \right) | E_m \right].
$$

On the other hand as a consequence of smoothness, conditioned on $E_m$, the tower property implies

$$
E \left[ \sum_{l \in T_m} f(x_t^{(2)}, \pi^*) - f(x_t^{(2)}, \pi_t^{(2)}) | E_m \right] \geq E \left[ \sum_{l=1}^{[T_m]} \frac{4U_m(l, \frac{\delta}{M})}{l} \right] - \frac{D_m}{2} + \frac{C_m}{2}.
$$

\footnote{We can extend this to a sub-optimality gap $\Delta$, at the cost of propagating this value throughout the remaining theorems.}
Since \( \mathbb{E} \left[ \sum_{t \in T_m} f(x_t^{(1)}, \pi^*) \right] = \mathbb{E} \left[ \sum_{t \in T_m} f(x_t^{(2)}, \pi^*) \right] \), these inequalities\( \textcircled{8} \) and \( \textcircled{9} \) we conclude:

\[
\mathbb{E} \left[ \sum_{t \in T_m} I'_{B_t} \right] \leq \mathbb{E} \left[ U_m(\lfloor T_m \rfloor, \delta) | \mathcal{E}_m \right] - \mathbb{E} \left[ \sum_{t=1}^{\lfloor T_m \rfloor} 4U_m(t, \delta) \right] + \frac{\delta}{M} T + \frac{D_m}{2} - \frac{C_m}{2}.
\]

By Lemma\( \textcircled{B.1} \)

\[
\mathbb{E} \left[ U_m(\lfloor T_m \rfloor, \delta) | \mathcal{E}_m \right] - \mathbb{E} \left[ \sum_{t=1}^{\lfloor T_m \rfloor} 4U_m(t, \delta) \right] \leq - \mathbb{E} \left[ 3U_m(\lfloor T_m \rfloor, \delta) | \mathcal{E}_m \right]
\]

The result follows.

These guarantees have been derived under the assumption that base algorithm \( \tilde{B}_m \) satisfies both the boundedness and smoothness properties. Although we conditioned on these properties in the main, it should be noted that these may not hold for base algorithms that are not "adapted" to the environment at hand. Nevertheless, in case a base algorithm violates boundedness of smoothness, there is no reason to keep said base algorithm as an option for the master. We formalize how to use this intuition in what follows.

The following lemma can be used to show a high probability lower bound on the regret of steps 2 (under mixture policy incorporating a known suboptimal arm):

**Lemma C.2.** Let \( U_m(t, \frac{\delta}{M}) = t^\beta g(\frac{\delta}{M}) \) with \( \beta \in [\frac{1}{2}, 1] \). For any \( l \in [T] \), with probability at least \( 1 - \tau \):

\[
\sum_{t \in T_m(l)} f(x_t^{(2)}, \pi^*) - f(x_t^{(2)}, \pi_t^{(2)}) \geq \sum_{t=1}^{l} \frac{4U_m(t, \frac{\delta}{M})}{t} - 8g(\frac{\delta}{M}) \Lambda - D_m + C_m
\]

Where \( \Lambda = \left\{ \begin{array}{ll}
\sqrt{\log(\frac{1}{\tau}) \frac{1}{2\beta-1}} t^{\beta - \frac{1}{2}} & \text{if } \beta > \frac{1}{2} \\
\sqrt{\log(\frac{1}{\tau}) \log(t)} & \text{if } \beta = \frac{1}{2}.
\end{array} \right. \)

Where \( C_m \) is the index of the first round where \( \frac{8U_m(t, \frac{\delta}{M})}{t} \leq 1 \) and \( D_m = \sum_{t=1}^{C_m} \frac{8U_m(t, \frac{\delta}{M})}{t} \).

**Proof:**

\[
\sum_{t \in T_m(l)} f(x_t^{(2)}, \pi^*) - f(x_t^{(2)}, \pi_t^{(2)}) \geq \alpha_t \sum_{t \in T_m(l)} f(x_t^{(1)}, \pi^*) - f(x_t^{(2)}, \pi_t)
\]

Notice that \( \sum_{t=1}^{l} \alpha_t^2 \leq \left( g(\frac{\delta}{M}) \right)^2 \sum_{t=1}^{l} t^{2\beta - 2} \leq \left( g(\frac{\delta}{M}) \right)^2 \int_1^l t^{2\beta - 2} dt \leq \left( g(\frac{\delta}{M}) \right)^2 \log(l \beta) \leq \frac{1}{2}. \)

By the Azuma-Hoeffding inequality\( \textcircled{10} \) (we use the fact \( \alpha_t |f(x, \pi^*) - f(x, \pi)| \leq 2\alpha_t \) with probability \( 1 - \tau \):

\[
\alpha_t \sum_{t \in T_m(l)} f(x_t^{(2)}, \pi^*) - f(x_t^{(2)}, \pi_t) \geq \frac{1}{2} \sum_{t=1}^{l} \min(1, \frac{8U_m(t, \delta)}{t}) - 8g(\frac{\delta}{M}) \Lambda
\]

\[
= \frac{1}{2} \sum_{t=1}^{l} 4U_m(t, \delta) - D_m + C_m - 8g(\frac{\delta}{M}) \Lambda
\]

\( ^5 \) We use the following version of Azuma-Hoeffding: If \( X_n, n \geq 1 \) is a martingale such that \( |X_i - X_{i-1}| \leq d_i, 1 \leq i \leq n \) then for every \( t > 0, P(X_n > \tau) \leq \exp \left( \frac{-\tau^2}{\sum_{i=1}^{n} d_i^2} \right) \)

18
As a consequence of Equations 10 and 11 we conclude that with probability at least 

\[ \Pr \geq 1 - 2T\tau \]

The result follows.

Proof. By Lemma B.1, 

\[ \sum_{t=1}^{T} U_{m}(t, \frac{\delta}{M}) \geq 4U_{m}(l, \frac{\delta}{M}). \]

And it is easy to see that for all \( l \) sufficiently large:

\[ U_{m}(l, \frac{\delta}{M}) \geq 8g \left( \frac{\delta}{M} \right) \left\{ \sqrt{\log \left( \frac{1}{\tau} \right)} \right\} \]

The result follows.

Again making use of Martingale Concentration bounds we can show:

Remark C.4. For any \( l \in [T] \) with probability at least 1 - \( \tau \) we have that \( \sum_{t \in T_{m}(l)} f(x_{t}^{(2)}, \pi^{*}) - f(x_{t}^{(1)}, \pi^{*}) < \sqrt{2 \log \left( \frac{1}{\tau} \right)} l \).

We assume there is an \( l'_{m} \) such that \( U_{m}(l, \frac{\delta}{M}) \geq \sqrt{2 \log \left( \frac{1}{\tau} \right)} l \) for all \( l \geq l'_{m} \). This is possible since we consider \( U_{m}(l, \frac{\delta}{M}) = g(\delta)t^{\beta} \) for \( \beta \geq \frac{1}{2} \) and can be achieved by modulating the leading constants included in \( g \). We will subsequently show that \( g(\delta) \) will be required to be of the form \( \beta \) for some \( \gamma \geq 0 \). Let \( s \) define \( G_{m} = \sum_{r=1}^{l'_{m}} \sqrt{2 \log \left( \frac{1}{\tau} \right)} l - U_{m}(r, \frac{\delta}{M}) \).

As a consequence of remark C.4 and proposition C.3, with probability at least 1 - \( \frac{\delta}{M} \):

\[ \sum_{t \in T_{m}(l)} f(x_{t}^{(1)}, \pi^{*}) - f(x_{t}^{(2)}, \pi_{t}^{(2)}) \geq 3U_{m} \left( l, \frac{\delta}{M} \right) - \sqrt{2 \log \left( \frac{1}{\tau} \right)} l - D_{m} + C_{m} - L_{m} \]

\[ \geq 2U(l, \frac{\delta}{M}) - D_{m} + C_{m} - L_{m} - G_{m}. \]

If \( \tilde{E}_{m} \) satisfies \( (U_{m}, \frac{\delta}{M}, T^{(1)}) \)-boundedness, it must hold that, conditioned on \( \mathcal{E}_{m} \):

\[ \sum_{t \in T_{m}(l)} f(x_{t}^{(1)}, \pi^{*}) - f(x_{t}^{(1)}, \pi_{t}^{(1)}) \leq U \left( l, \frac{\delta}{M} \right) \]

As a consequence of Equations 10 and 11 we conclude that with probability at least 1 - \( \frac{\delta}{M} - 2T\tau \), for all \( l \in [T] \):
\[ \sum_{t \in \mathcal{V}_m(t)} f(x_t^{(1)}, \pi^*) - f(x_t^{(1)}, \pi_t^{(1)}) \leq U(l, \frac{\delta}{M}) \]

\[ 2U \left( l, \frac{\delta}{M} \right) - D_m + C_m - L_m - G_m \leq \sum_{t \in \mathcal{V}_m(t)} f(x_t^{(1)}, \pi^*) - f(x_t^{(2)}, \pi_t^{(2)}) \]

And therefore for all \( l \in [T] \) with probability \( 1 - \frac{\delta}{M} - 2T\tau \):

\[ \sum_{t \in \mathcal{V}_m(t)} f(x_t^{(1)}, \pi_t^{(1)}) - f(x_t^{(2)}, \pi_t^{(2)}) \geq U \left( t, \frac{\delta}{M} \right) - D_m + C_m - L_m - G_m \]

Equation 12 holds with probability \( 1 - \frac{\delta}{M} - 2T\tau \) as long as \((U, \frac{\delta}{M}, \mathcal{T}^{(1)})\)-boundedness holds. Furthermore, notice that whenever this event holds, \( \mathbb{E}[I_B] \leq T\delta + 2T^2\tau + \sum_{t=1}^\infty \left( \delta t \right) \), thus preserving our regret guarantees. Crucially notice that all terms \( D_m, C_m, L_m, G_m \) are constant (possibly dependent on \( U(t, \frac{\delta}{M}) \), but independent on \( T \)). Since boundedness is only true for algorithms \( \tilde{B}_m \) that are "adapted" to the current environment, we introduce a small change to make sure the "domination" property between rewards of steps 1 and steps 2 holds for all times so that the bound on the expectation of term \( I_B \) remains valid.

1. For all base algorithms \( \{\tilde{B}_m\}_{m=1}^M \), if \( i_t = m \), suppose algorithm \( m \) has been selected \( l \)-times by the master, and whenever Equation 12 starts failing for base algorithm \( \tilde{B}_m \), we "drop" this base algorithm by declaring the rewards of subsequent steps 2 in subsequent rounds to be 0 before sending this feedback into the master and repeating step 1 twice, alternatively we can send a zero feedback into the master and execute the policy supported on the remaining algorithms whenever a "dropped" base is selected.

If this extra step 1 is executed for algorithm \( \tilde{B}_m \), we know that (with high probability) \((U, \delta, \mathcal{T}^{(1)})\)-boundedness was violated, and therefore that algorithm \( \tilde{B}_m \) couldn't have possibly been optimal for the environment at hand.

After all these arguments we can conclude that:

\[ \mathbb{E}[I] \leq O \left( \frac{M \ln T}{\eta} + T\eta \right) - \frac{\mathbb{E}[\rho_n]}{40\eta \ln T} \]

Up to a \( \delta T + 2T\tau + E_m \), (where \( E_m \) is a constant) additive factor.

**D Proof of Lemma 4.2**

**Proof of Lemma 4.2** Recall \( \mathbb{E}[I] \leq \mathbb{E} \left[ R^{(1)}_i (T_i) 1(\mathcal{E}) + I_0 1(\mathcal{E}) \right] + \delta T \). The first term is bounded by \( \mathbb{E} \left[ U_i(\delta, n^{\mathcal{E}}_f) 1(\mathcal{E}) \right] \) while the second term satisfies the bound in (5). Let \( u_t = \frac{U_t(\delta, \delta/2M)}{t} \). By Lemma B.1, \( \sum_{t=1}^\infty u_t \geq U_i(t, \delta/M) \) for all \( t \), and so,

\[ \mathbb{E} \left[ 1(\mathcal{E}) U_i(\delta, n^{\mathcal{E}}_f) \right] \leq \mathbb{E} \left[ \sum_{t=1}^{n^{\mathcal{E}}_f+1} 1(\mathcal{E}) u_t \right]. \]

By (5) and (13),

\[ \mathbb{E} \left[ R^{(1)}_i (T_i) 1(\mathcal{E}) + I_0 1(\mathcal{E}) \right] \leq \mathbb{E} \left[ \sum_{t=1}^{n^{\mathcal{E}}_l+1} 1(\mathcal{E}) 2b_t u_t \right]. \]

Let \( a_l = \mathbb{E}[b_l] \) for all \( l \). Consider a master algorithm that uses \( \rho \) instead of \( p^l_i \). In this new process let \( t'_l \) be the corresponding rounds when the base is selected, \( n^{l} \) be the total number of rounds the base is selected,
and \( c_t = \mathbb{E} [ t'_t - t'_{t-1} ] \). Since \( p \leq p'_t \) for all \( t \) it holds that \( \sum_{i=1}^{j} a_i \leq \sum_{i=1}^{j} c_i \) for all \( j \). If we use the same coin flips used to generate \( t_i \) to generate \( t'_i \), we observe that \( t'_i \subset t_i \) and \( \bar{n}_i \geq n'_i \). Let \( f : \mathbb{R} \rightarrow [0,1] \) be a decreasing function such that for integer \( i \), \( f(i) = u_i \). Then \( \sum_{i=1}^{n_i} a_i u_i \) and \( \sum_{i=1}^{\bar{n}_i} c_i u_i \) are two estimates of integral \( \int_0^T f(x) \, dx \). Given that \( t'_i \subset t_i \) and \( u_i \) is a decreasing sequence in \( t \),

\[
\sum_{i=1}^{n_i+1} E [ t_i - t_{i-1} ] u_i \leq \sum_{i=1}^{\bar{n}_i+1} E [ t'_i - t'_{i-1} ] u_i,
\]

and thus

\[
E \left[ R^{(1)}_t (T_0 I) + I_0 I \right] \leq E \sum_{i=1}^{n_i+1} 2E [ t'_i - t'_{i-1} ] u_i.
\]

We proceed to upper bound the right hand side of this inequality:

\[
E \left[ \sum_{i=1}^{n_i+1} u_i E [ t'_i - t'_{i-1} ] \right] \leq E \left[ \sum_{i=1}^{\bar{n}_i+1} \frac{u_i}{p} \right] \leq 2p U_t(T/p, \delta) \log(T).
\]

The first inequality holds because \( E [ t'_i - t'_{i-1} ] \leq \frac{1}{p} \) and the second inequality follows by concavity of \( U_t(t, \delta) \) as a function of \( t \). The proof follows. \( \square \)

## E Proof of Lemma 4.6

We restate Lemma 4.6 for readability.

**Lemma E.1.** If \( B \) is \((U, \delta, [T])\)–bounded, \( \max_{x, \pi} |f(x, \pi)| \leq 1, U(t, \delta) > 8\sqrt{t \log \left( \frac{t}{\delta} \right)} \), and \( \delta \leq \frac{1}{\sqrt{T}} \), then \( B \)'s expected replay regret \( R(t, h) \) satisfies:

\[
R(t, h) \leq 4U(t, \delta) + 2\delta t \leq 5U(t, \delta)
\]

**Proof.** Consider the following two martingale sequences:

\[
\{ M^1_i : = f(x_i, \pi^*) - f(x'_i, \pi^*) \}_{i=1}^{t} \quad \{ M^2_i : = f(x'_i, \pi_t) - f(x_i, \pi_t) \}_{i=1}^{t}
\]

Since \( \max (|M^1_i|, |M^2_i|) \leq 2 \) for all \( t \), a simple use of Azuma-Hoeffding yields:

\[
P \left( \left| \sum_i M^1_i \right| \geq U(t, \delta) \right) \leq P \left( \left| \sum_i M^1_i \right| \geq \sqrt{8t \log \left( \frac{8t^2}{\delta} \right)} \right) \leq 2 \exp \left( - \frac{8t \log \left( \frac{8t^2}{\delta} \right)}{8t} \right) = \frac{\delta}{4t^2}
\]

Summing over all \( t \), and all \( i \in \{1, 2\} \) and applying the union bound, using the fact that \( \sum_{i=1}^{T} \frac{1}{T} < 2 \) implies that for all \( t \), with probability \( 1 - \delta \),

\[
\left| \sum_{i=1}^{t} f(x_i, \pi^*) - \sum_{i=1}^{t} f(x_i, \pi_t) \right| - \left| \sum_{i=1}^{t} f(x'_i, \pi^*) - \sum_{i=1}^{t} f(x'_i, \pi_t) \right| \leq 2U(t, 2\delta)
\]
F Proof of Theorem 4.3

Proof of Theorem 4.3. The proof follows that of Theorem 15 in (Agarwal et al., 2017). Let \( \ell_1, \ldots, \ell_{d_i} < T \) be the rounds where Line 10 of the CORRAL is executed. Let \( \ell_0 = 0 \) and \( \ell_{d_i+1} = T \) for notational convenience. Let \( e_i = [\ell_{i-1} + 1, \ldots, \ell_i] \). Denote by \( p_{\ell_i} \) the probability lower bound maintained by Corral during timesteps \( t \in [\ell_{i-1}, \ldots, \ell_i] \) and \( \rho_{\ell_i} = 1/p_{\ell_i} \). In the proof of Lemma 13 in (Agarwal et al., 2017), the authors prove \( d_i \leq \log(T) \) with probability one. Therefore,

\[
E[I] = \sum_{l=1}^{[\log(T)]} \mathbb{P}(d_i + 1 \geq l) E\left[R^{(1)}(e_i) + R^{(2)}(e_i)|d_i + 1 \geq l\right]
\]

\[
\leq \log T \sum_{l=1}^{[\log(T)]} \mathbb{P}(I(l)) E\left[4\rho_{\ell_i} U_i(T/\rho_{\ell_i}, \delta)|I(l)\right] + \delta T(\log T + 1)
\]

\[
= \log T E\left[\sum_{l=1}^{b_i+1} 4\rho_{\ell_i} U_i(T/\rho_{\ell_i}, \delta)\right] + \delta T(\log T + 1).
\]

The inequality is a consequence of Lemma 4.2 applied to the restarted segment \([\ell_{i-1}, \ldots, \ell_i]\). This step is valid because assumption \( \frac{1}{\rho_{\ell_i}} \leq \min_{t \in [\ell_{i-1}, \ldots, \ell_i]} p_t \).

If \( U_i(t, \delta) = t^\alpha g(\delta) \) for some function \( g : \mathbb{R} \to \mathbb{R}^+ \), then \( \rho U(T/\rho, \delta) = \rho^{1-\alpha} T^\alpha g(\delta) \). And therefore:

\[
E\left[\sum_{l=1}^{b_i+1} \rho_{\ell_i} U_i(T/\rho_{\ell_i}, \delta)\right] \leq T^\alpha g(\delta) E\left[\sum_{l=1}^{b_i+1} \rho_{\ell_i}^{1-\alpha}\right]
\]

\[
\leq \frac{2^\alpha}{2^\alpha - 1} T^\alpha g(\delta) E\left[\rho_{\ell_i}^{1-\alpha}\right]
\]

Where \( \bar{\alpha} = 1 - \alpha \). The last inequality follows from the same argument as in Theorem 15 in (Agarwal et al., 2017).

G Proof of Corollary 4.5

Proof of Corollary 4.5. By Theorem 4.4,

\[
R(T) \leq O\left(\frac{M \ln T}{\eta} + T\eta\right)
\]

\[
- E\left[\frac{\rho}{40 \eta \ln T} - 2\rho U(T/\rho, \delta) \log T\right] + \delta T
\]

\[
\leq O\left(\frac{M \ln T}{\eta} + T\eta\right)
\]

\[
- E\left[\frac{\rho}{40 \eta \ln T} - 2\rho^{1-\alpha} T^\alpha g(\delta) \log T\right] + \delta T
\]

\[
\leq \tilde{O}\left(\frac{M}{\eta} + T\eta + T g(\delta) \frac{1}{\eta} \frac{1-\alpha}{\bar{\alpha}}\right) + \delta T,
\]

where the last step is by maximizing the function over \( \rho \). Substituting \( \eta = \frac{M^{\alpha}}{g(\delta)\eta^{\bar{\alpha}}} \) finishes the proof. \( \square \)
H High probability regret bounds for $\epsilon$-greedy

In this section we show that epsilon greedy satisfies a high probability regret bound. We adapt the notation to this setup. Let $\mu_1, \ldots, \mu_K$ be the unknown means of the $K$ arms. Recall that at time $t$ the epsilon Greedy algorithm selects with probability $\epsilon_t = \min(c/t, 1)$ an arm uniformly at random, and with probability $1 - \epsilon_t$ it selects the arm whose empirical estimate of the mean is largest so far. Let $\hat{\mu}_j(t)$ denote the empirical estimate of the mean of arm $j$ after using $t$ samples.

Without loss of generality let $\mu_1$ be the optimal arm. We denote the gaps as $\Delta_j = \mu_1 - \mu_j$ for all $j$. Let $\Delta_*$ be the smallest nonzero gap. We follow the discussion in [Auer et al., 2002] and start by showing that under the right assumptions, and for a horizon of size $T$, the algorithm satisfies a high probability regret bound for all $t \leq T$. The objective of this section is to prove the following Lemma:

**Lemma H.1.** If $c = \frac{10K \log(\frac{1}{\delta})}{\Delta_*^2}$ then $\epsilon$-greedy with $\epsilon_t = \frac{c}{t}$ is $(\delta, U, T)$-stable for $\delta \leq \frac{\Delta_*^2}{T}$ and $U(t, \delta) = \frac{30K \log(\frac{1}{\delta})}{\Delta_*^2} \left( \sum_{j=2}^{K} \frac{\Delta_j}{\Delta_*^2} + \Delta_j \right) \log(t+1)$.

**Proof.** Let $E(t) = \frac{1}{K} \sum_{l=1}^{t} \epsilon_l$ and denote by $T_j(t)$ the random variable denoting the number of times arm $j$ was selected up to time $t$. We start by analyzing the probability that a suboptimal arm $j > 1$ is selected at time $t$:

$$\mathbb{P}(j \text{ is selected at time } t) \leq \frac{\epsilon_t}{K} + \left(1 - \frac{\epsilon_t}{K}\right) \mathbb{P} \left( \hat{\mu}_j^{(T_j(t))} \geq \hat{\mu}_1^{(T_1(t))} \right)$$

(14)

Let’s bound the second term.

$$\mathbb{P} \left( \hat{\mu}_j^{(T_j(t))} \geq \hat{\mu}_1^{(T_1(t))} \right) \leq \mathbb{P} \left( \hat{\mu}_j^{(T_j(t))} \geq \mu_j + \frac{\Delta_j}{2} \right) + \mathbb{P} \left( \hat{\mu}_1^{(T_1(t))} \leq \mu_1 - \frac{\Delta_j}{2} \right)$$

The analysis of these two terms is the same. Denote by $T_j^R(t)$ the number of times arm $j$ was played as a result of a random epsilon greedy move. We have:

$$\mathbb{P} \left( \hat{\mu}_j^{(T_j(t))} \geq \mu_j + \frac{\Delta_j}{2} \right) = \sum_{l=1}^{t} \mathbb{P} \left( T_j(t) = l \text{ and } \hat{\mu}_j^{(l)} \geq \mu_j + \frac{\Delta_j}{2} \right)$$

$$= \sum_{l=1}^{t} \mathbb{P} \left( T_j(t) = l | \hat{\mu}_j^{(l)} \geq \mu_j + \frac{\Delta_j}{2} \right) \mathbb{P} \left( \hat{\mu}_j^{(l)} \geq \mu_j + \frac{\Delta_j}{2} \right)$$

$$\leq \sum_{l=1}^{t} \mathbb{P} \left( T_j(t) = l | \hat{\mu}_j^{(l)} \geq \mu_j + \frac{\Delta_j}{2} \right) \exp(-\Delta_j^2 t/2)$$

$$\leq \sum_{l=1}^{t} \mathbb{P} \left( T_j(t) = l | \hat{\mu}_j^{(l)} \geq \mu_j + \frac{\Delta_j}{2} \right) + \frac{2}{\Delta_j^2} \exp(-\Delta_j^2 E(t)/2)$$

$$\leq \sum_{l=1}^{t} \mathbb{P} \left( T_j^{R}(t) = l | \hat{\mu}_j^{(l)} \geq \mu_j + \frac{\Delta_j}{2} \right) + \frac{2}{\Delta_j^2} \exp(-\Delta_j^2 E(t)/2)$$

$$\leq \frac{\mathbb{E}(t)}{\mathbb{P} \left( T_j^{R}(t) \leq \mathbb{E}(t) \right)} + \frac{2}{\Delta_j^2} \exp(-\Delta_j^2 E(t)/2)$$

Inequality 1 is a consequence of a Chernoff bound. Inequality 2 follows because $\sum_{l=1}^{\infty} \mathbb{P}(s + a l) - \frac{1}{a} \exp(-a E(t))$. Term (1) corresponds to the probability that within the interval $[1, \cdots, t]$, the number of greedy pulls to arm $j$ is at most half its expectation. Term (2) is already "small".

---

6This choice of $c$ is robust to multiplication by a constant.
Recall $\epsilon_t = \min(c/t, 1)$. Let $c = \frac{10K \log(T^3/\gamma)}{\Delta_j^2}$ for some $\gamma \in (0, 1)$ satisfying $\gamma \leq \Delta_j^2$. Under these assumptions we can lower bound $E(t)$: Indeed if $t \geq \frac{10K \log(T^3/\gamma)}{\Delta_j^2}$:

$$
\frac{1}{2K} \sum_{l=1}^{t} \epsilon_l = \frac{5 \log(T^3/\gamma)}{\Delta_j^2} + \frac{5 \log(T^3/\delta)}{\Delta_j^2} \sum_{l=\log(T^3/\gamma)}^{t} \frac{1}{T} \\
\geq \frac{5 \log(T^3/\gamma)}{\Delta_j^2} + \frac{5 \log(T^3/\gamma) \log(t)}{2\Delta_j^2} \\
\geq \frac{5 \log(T^3/\gamma)}{\Delta_j^2}
$$

By Bernstein’s inequality (see derivation of equation (13) in [Auer et al., 2002]) it is possible to show that:

$$
P(T_j^R(t) \leq E(t)) \leq \exp(-E(t)/5) \quad (15)
$$

Hence for $t \geq \frac{10K \log(T^3/\gamma)}{\Delta_j^2}$:

$$
P(T_j^R(t) \leq E(t)) \leq \left(\frac{\gamma}{T^3}\right)^{\frac{1}{\Delta_j^2}}
$$

And therefore since $E(t) \leq T$ and $\frac{1}{\Delta_j^2} \geq 1$ we can upper bound (1) as:

$$
|E(t)|P(T_j^R(t) \leq |E(t)|) \leq \left(\frac{\gamma}{T^2}\right)^{\frac{1}{\Delta_j^2}} \leq \frac{\gamma}{T^2}
$$

Now we proceed with term (2):

$$
\frac{2}{\Delta_j} \exp \left(-\frac{\Delta_j^2}{2} |E(t)| / 2 \right) \leq \frac{2}{\Delta_j} \exp \left(-5K \log\left(\frac{T^3}{\gamma}\right) \frac{\Delta_j^2}{2} \right) \\
\leq \frac{2}{\Delta_j} \exp \left(-5K \log\left(\frac{T^3}{\gamma}\right) \right) \\
= \frac{2}{\Delta_j} \left(\frac{\gamma}{T^3}\right)^{5K}
$$

By the assumption $\gamma \leq \Delta_j^2$ the last term is upper bounded by $\frac{\gamma}{T^2}$.

The previous discussion implies that for $c = \frac{10K \log(T^3/\gamma)}{\Delta_j^2}$, the probability of choosing a suboptimal arm $j \geq 2$ at time $t$ for $t \geq \frac{10K \log(T^3/\gamma)}{\Delta_j^2}$ as a greedy choice is upper bounded by $2\frac{\gamma}{T^2}$. In other words after $t \geq \frac{10K \log(T^3/\gamma)}{\Delta_j^2}$, suboptimal arms with probability $1 - \frac{1}{T}$ over all $t$ are only chosen as a result of a exploration uniformly random epsilon greedy action.

A similar argument as the one that gave us Equation (15) can be used to upper bound the probability that at a round $t$, $T_j^R(t)$ be much larger than its mean:

$$
P(T_j^R(t) \geq 3E(j)) \leq \exp(-E(t)/5)
$$

We can conclude that with probability more than $1 - \frac{K\gamma}{T}$ and for all $t$ and arms $j$, $T_j^R(t) \leq 3E(t)$. Combining this with the observation that after $t \geq \frac{10K \log(T^3/\gamma)}{\Delta_j^2}$ and with probability $1 - \frac{K\gamma}{T}$ over all $t$ simultaneously (by union bound) regret is only incurred by random exploration pulls (and not greedy actions), we can conclude that with probability $1 - \frac{2K\gamma}{T}$ simultaneously for all $t \geq \frac{10K \log(T^3/\gamma)}{\Delta_j^2}$ the regret incurred is upper bounded by:
When \( K \) is satisfied for \( \gamma \) \( \geq 20K \log(T^3/\gamma) \) \( \Delta^2 \) rounds and \( II \) is an upper bound for the regret incurred in the subsequent rounds.

Since \( E(t) \leq \frac{10K \log(T^3/\gamma)}{\Delta^2} \log(t) \) we can conclude that with probability \( 1 - \frac{2K \gamma}{t} \) for all \( t \leq T \) the cumulative regret of epsilon greedy is upper bounded by

\[
R(t) = \sum_{j=2}^{K} \Delta_j + 3E(t) \sum_{j=2}^{K} \Delta_j
\]

We now show the proof of Lemma \ref{lemma:inst_opt_regret_2} the instance-independent regret bound for \( \epsilon \)-greedy:

**Lemma H.2** (Lemma \ref{lemma:inst_opt_regret_2}). If \( c = \frac{10K \log(T^3/\gamma)}{\Delta^2} \), then \( \epsilon \)-greedy with \( \epsilon_t = \frac{\gamma}{t} \) is \((\delta, U, T)\)-stable for \( \delta \leq \frac{\Delta^2}{T^2} \) and:

1. \( U(t, \delta) = 16 \sqrt{\log(\frac{1}{\delta})} t \) when \( K = 2 \).
2. \( U(t, \delta) = 20 \left( K \log(\frac{1}{\delta}) \left( \sum_{j=2}^{K} \Delta_j \right) \right)^{1/3} t^{2/3} \) when \( K > 2 \).

**Proof.** Let \( \Delta \) be some arbitrary gap value. Let \( R(t) \) denote the expected regret up to round \( t \). We recycle the notation from the proof of Lemma \ref{lemma:inst_opt_regret_2} recall \( \delta = \gamma/T^3 \).

\[
R(t) = \sum_{\Delta_j \leq \Delta} \Delta_j E[T_j(t)] + \sum_{\Delta_j > \Delta} \Delta_j E[T_j(t)]
\]

\[
\leq \Delta t + \sum_{\Delta_j > \Delta} \Delta_j E[T_j(t)]
\]

\[
\leq \Delta t + 30K \log(T^3/\gamma) \left( \sum_{\Delta_j > \Delta} \frac{\Delta_j}{\Delta^2} + \Delta_j \right) \log(t)
\]

\[
\leq \Delta t + 30K \log(T^3/\gamma) \left( \sum_{\Delta_j > \Delta} \frac{\Delta_j}{\Delta^2} \right) + 30K \log(T^3/\gamma) \log(t) \left( \sum_{\Delta_j > \Delta} \Delta_j \right)
\]

When \( K = 2 \), \( \Delta_2 = \Delta_\ast \) and therefore (assuming \( \Delta < \Delta_2 \)):

\[
R(t) \leq \Delta t + \frac{30K \log(T^3/\gamma)}{\Delta_2} + 30K \log(T^3/\gamma) \log(t) \Delta_2
\]

\[
\leq \Delta t + \frac{30K \log(T^3/\gamma)}{\Delta} + 30K \log(T^3/\gamma) \log(t) \Delta_2
\]

\[
\leq \frac{\bar{I}}{\bar{I}} 8 \sqrt{30K \log(T^3/\gamma) t} + 30K \log(T^3/\gamma) \log(t) \Delta_2
\]

\[
\leq 16 \sqrt{\log(T^3/\gamma) t}
\]

Inequality \( I \) follows from setting \( \Delta \) to the optimizer, which equals \( \Delta = \sqrt{30K \log(T^3/\gamma)} t \). The second inequality \( II \) is satisfied for \( T \) large enough. We choose this expression for simplicity of exposition.

When \( K > 2 \) notice that we can arrive to a bound similar to \( 16 \).
\begin{align*}
R(t) &\leq \Delta t + 30K \log\left(\frac{T^3}{\gamma}\right) \left(\sum_{\Delta_j \geq \Delta^*} \frac{\Delta_j}{\Delta^*}\right) + 30K \log\left(\frac{T^3}{\gamma}\right) \log(t) \left(\sum_{\Delta_j \geq \Delta} \Delta_j\right)
\end{align*}

Where \(\Delta^*\) is substituted by \(\Delta\). This can be obtained from Lemma H.1 by simply substituting \(\Delta^*\) with \(\Delta\) in the argument for arms \(j: \Delta_j \geq \Delta\).

We upper bound \(\sum_{\Delta_j \geq \Delta} \Delta_j\) by \(\sum_{j=2}^{K \Delta} \Delta_j\). Setting \(\Delta\) to the optimizer of the expression yields \(\Delta = \left(\frac{30K \log\left(\frac{T^3}{\gamma}\right) \left(\sum_{j=2}^{K \Delta} \Delta_j\right)}{t} \right)^{1/3}\), and plugging this back into the equation we obtain:

\begin{align*}
R(t) &\leq 2 \left(30K \log\left(\frac{T^3}{\gamma}\right) \left(\sum_{j=2}^{K \Delta} \Delta_j\right)\right)^{1/3} t^{2/3} + 30K \log\left(\frac{T^3}{\gamma}\right) \log(t) \left(\sum_{j=2}^{K \Delta} \Delta_j\right) \\
\xi &\leq 20 \left(K \log\left(\frac{T^3}{\gamma}\right) \left(\sum_{j=2}^{K \Delta} \Delta_j\right)\right)^{1/3} t^{2/3}
\end{align*}

The inequality \(\xi\) is true for \(T\) large enough. We choose this expression for simplicity of exposition.

### I High Probability Regret Bound for UCB

**Lemma I.1** (Lemma 4.9). Assuming that the noise \(\xi_t\) is conditionally 1-sub-Gaussian, UCB is \((U, \delta, [T])\)-bounded with \(U(t, \delta) = O(\sqrt{tk \log \frac{tk}{\delta}})\).

**Proof.** The regret of UCB is bounded as \(\sum_{i: \Delta_i > 0} \left(3\Delta_i + \frac{16}{\Delta_i} \log \frac{2k}{\Delta_i \delta}\right)\) (Theorem 7 of Abbasi-Yadkori et al. (2011)) where \(\Delta_i\) is the gap between arm \(i\) and the best arm. By substituting the worst-case \(\Delta_i\) in the regret bound, \(U(T, \delta) = O(\sqrt{Tk \log \frac{Tk}{\delta}})\).

### J Section 6 Proofs.

#### J.1 Proof of Theorem 6.3

From Corollary 4.11, we lower bound the smallest gap by \(1/T\) (because the gaps smaller than \(1/T\) will cause constant regret in \(T\) time steps) and choose \(\delta = 1/T^5\). Using an appropriate \(\eta\) as discussed in Corollary 4.5 will result in the regret of the same order \(O(T^{2/3})\) when \(K > 2\) and \(O(T^{1/2})\) when \(K = 2\) with the base running alone.

Next we show that the best value of \(c\) in the exponential grid gives a regret that is within a constant factor of the regret above where we known the smallest non-zero gap \(\Delta^*\). An exploration rates can be at most \(KT\). Since \(\frac{5K}{\Delta^*} > 1\), we need to search only in the interval \([1, KT]\). Let \(c_1\) be the element in the exponential grid such that \(c_1 \leq c^* \leq 2c_1\). Then \(2c_1 = \gamma c^*\) where \(\gamma < 2\) is a constant, and therefore using \(2c_1 = \gamma c^*\) will give a regret up to a constant factor of the optimal regret.

#### J.2 Proof of Theorem 6.2

**Proof.** Using Corollary 4.5, we obtain the regret of stochastic CORRAL with the smooth versions of UCB and LinUCB. From Lemma 4.9 for UCB, \(U(T, \delta) = O(\sqrt{Tk \log \frac{Tk}{\delta}})\). Therefore from Corollary 4.5 running
stochastic CORRAL with smooth UCB results in the following regret bound:

$$O\left(\frac{2\ln T}{\eta} + T\eta + T\left(\sqrt{k\log \frac{Tk}{\delta}}\right)^2\right) + \delta T.$$ 

If we choose $\delta = 1/T$ and hide some log factors, we get $\tilde{O}\left(\frac{T}{\eta} + Tk\eta\right)$.

Similarly, for LinUCB, $U(t, \delta) = O(d\sqrt{t\log(1/\delta)})$ and running stochastic CORRAL with smooth LinUCB results in $\tilde{O}\left(\frac{T}{\eta} + Td^2\eta\right)$ regret bound.

By the choice of $\eta = \sqrt{\frac{2}{Td\sqrt{k}}}$, the regret of stochastic CORRAL with the smooth UCB and LinUCB will be bounded as

$$\tilde{O}\left(\max\left\{\sqrt{2T} \left(\frac{k^{0.75}}{d^{0.5}} + \frac{k^{0.25}}{d^{1.5}}\right), \sqrt{2T} \left(\frac{k^{0.75}}{d^{0.5}} + \frac{k^{0.25}}{d^{1.5}}\right)\right\}\right).$$

\[\square\]

K Proof of Lower Bound

Proof of Theorem 5.1. Consider a stochastic 2-arm bandit problem where the best arm has expected reward $1/2$ and the second best arm has expected reward $1/4$. We construct base algorithms $B_1, B_2$ as follows. $B_1$ always chooses the optimal arm and its expected instantaneous reward is $1/2$. $B_2$ chooses the second best arm at time step $t$ with probability $\frac{4c}{\sqrt{t+2\log(t+2)}}$ (will be specified later), and chooses the best arm otherwise. The expected reward at time step $t$ of $B_2$ is $\frac{1}{2} - \frac{c}{\sqrt{t+2\log(t+2)}}$.

Let $A^*$ be uniformly sampled from $\{1, 2\}$. Consider two environments $\nu_1$ and $\nu_2$ for the master, each made up of two base algorithms $\tilde{B}_1, \tilde{B}_2$. Under $\nu_1$, $\tilde{B}_1$ and $\tilde{B}_2$ are both instantiations of $B_1$. Under $\nu_2$, $\tilde{B}_{A_*}$, where $A^*$ is a uniformly sampled index in $\{1, 2\}$, is a copy of $B_1$ and $\tilde{B}_{1-A_*}$ is a copy of $B_2$.

Let $P_1, P_2$ denote the probability measures induced by interaction of the master with $\nu_1$ and $\nu_2$ respectively. Let $B_{A_t}$ denote the base algorithm chosen by the master at time $t$. We have $P_1(A_t \neq A^*) = \frac{1}{2}$ for all $t$, since the learner has no information available to identify which algorithm is considered optimal. By Pinskers’ inequality we have

$$P_2(A_t \neq A^*) \geq P_1(A_t \neq A^*) - \sqrt{\frac{1}{2} KL(P_1||P_2)}$$

By the divergence decomposition (see [Lattimore and Szepesvari] proof of Lemma 15.1 for the decomposition technique) and using that for $\Delta < \frac{1}{4}$: $kl(\frac{1}{2}, \frac{1}{2} - \Delta) \leq 3\Delta^2$ (Lemma B.3), we have

$$KL(P_1||P_2) = \sum_{t=2}^{\infty} \frac{1}{2} kl\left(\frac{1}{2}, \frac{1}{2} - \frac{c}{\sqrt{t+1\log(t+1)}}\right) \leq \sum_{t=2}^{\infty} \frac{3c^2}{2t\log(t)^2} \leq 3c^2.$$
Picking \( c = \sqrt{\frac{1}{24}} \) leads to \( \mathbb{P}_2(A_t \neq A^*) \geq \frac{1}{4} \), and the regret in environment \( \nu_2 \) is lower bounded by

\[
R(T) \geq \sum_{t=1}^{T} \mathbb{P}_2(A_t \neq A^*) \frac{c}{\sqrt{t + 1} \log(t + 1)} \geq \frac{c}{4 \log(T + 1)} \sum_{t=1}^{T} \frac{1}{\sqrt{t + 1}} = \Omega\left( \frac{\sqrt{T}}{\log(T)} \right).
\]