Multi-Armed Bandits

Karl Stratos

1 Problem

MDP. The action space is $\mathcal{A} = \{1 \dots K\}$ for some fixed K ("arms"). The state space $\mathcal{S} = (\{1 \dots K\} \times \{0, 1\})^+$ is used to keep track of past actions and rewards, with a deterministic Markov transition $s_t = s_{t-1} \cup (a_t, r_t)$ starting from $s_0 = \varnothing$. The reward function $r \sim D_a$ is stochastic, non-adversarial, and only a function of the choice of an arm $a \in \{1 \dots K\}$, disregarding the state. Here, D_a is some unknown distribution over $[0, 1]$ with mean μ_a. We will denote the best arm by $a^\star = \arg\max_{k=1}^K \mu_k$ associated with the reward distribution $D^\star = D_{a^\star}$ and the mean reward $\mu^\star = \mu_{a^\star}$.

Policy. The policy is an algorithm, not a parameteric model. For $t = 1 \dots T$, the algorithm specifies a conditional distribution $\pi(\cdot|s_{t-1})$ over the K arms given $s_{t-1} = (a_1, r_1) \dots (a_{t-1}, r_{t-1})$, samples $a_t \sim \pi(\cdot|s_{t-1})$, and receives $r_t \sim D_{a_t}$. The badness of this session is measured by the "regret" $R = \sum_{t=1}^T r_t^\star - r_t$ where $r_t^\star \sim D^\star$. The goal of the algorithm is to minimize the expected regret:[1]

$$\mathbf{E}[R] = \sum_{t=1}^T \mathbf{E}[\mu^\star - \mu_{a_t}] \geq 0 \tag{1}$$

We assume $T \geq K$, but T is generally unknown to the algorithm. We are mostly interested in bounding (1) as an asymptotic function of T (and K, less importantly). A meaningful bound must be sublinear in T, since linear is trivial (e.g., always choose $a_t = 1$). On the other hand, the bound cannot be smaller than $\Omega(\sqrt{KT})$ (Appendix A). If we know the best arm $a^\star$, we can achieve $\mathbf{E}[R] = 0$ by always choosing $a_t = a^\star$. But since we do not know $a^\star$, we need to explore unknown arms. This is the exploration-exploitation tradeoff.

1.1 Reward Estimation

The algorithm maintains a running estimate of μ_a at step t by averaging the $N_{a,t}$ samples $r_{a,1} \dots r_{a,N_{a,t}}$ from D_a collected so far,

$$\hat{\mu}_{a,t} = \frac{1}{N_{a,t}} \sum_{i=1}^{N_{a,t}} r_{a,i} \tag{2}$$

Note that $N_{a,t}$ is a random variable. Without conditioning on the choice of arms, $r_{a,i}$ may not even be independent (e.g., the algorithm may sample from arm a only if its previous rewards never fell below 0.1). This makes a naive application of Hoeffding's inequality difficult. A straightforward solution is to consider all possible cases ahead in a "grid" (Slivkins *et al.*, 2019).

Lemma 1.1. Pick any algorithm for sampling $a_1 \dots a_T$. The mean estimates (2) satisfy

$$\Pr\left(|\mu_a - \hat{\mu}_{a,t}| < \sqrt{\frac{2 \log T}{N_{a,t}}} \text{ for all } a \in \{1 \dots K\} \text{ and } t \in \{1 \dots T\}\right) > 1 - \frac{2}{T^2} \tag{3}$$

Proof. Let $G \in \mathbb{R}^{K \times T}$ be a random grid where $G(a, i) \sim D_a$ independently. Let $\hat{\nu}_{a,i}$ denote the average of $G(a, 1) \dots G(a, i)$. By Hoeffding's inequality,

$$\Pr\left(|\mu_a - \hat{\nu}_{a,i}| \geq \sqrt{\frac{2 \log T}{i}}\right) \leq \frac{2}{T^4}$$

[1] This is equivalent to maximizing the (finite-horizon) value function with no discount factor.

Then by the union bound,

$$\Pr\left(|\mu_a - \hat{\nu}_{a,i}| < \sqrt{\frac{2\log T}{i}} \text{ for all } a \in \{1\dots K\} \text{ and } i \in \{1\dots T\}\right) > 1 - \frac{2}{T^2} \tag{4}$$

Now we view the grid G as having "precomputed" all samples and the algorithm as retrieving $r_{a,i} = G_{a,i}$ upon selecting the arm a for the i-th time. At any step t, the estimate (2) for a corresponds to $\hat{\nu}_{a,i}$ for $i = N_{a,t}$. Thus (4) implies (3). □

1.2 Regret Decomposition

Lemma 1.2. The expected regret (1) has the same asymptotic bound when conditioned on the "clean event"

$$|\mu_a - \hat{\mu}_{a,t}| < \sqrt{\frac{2\log T}{N_{a,t}}} = \rho_{a,t} \qquad \forall a \in \{1\dots K\},\ t \in \{1\dots T\} \tag{5}$$

(i.e., every $\hat{\mu}_{a,t}$ captures μ_a in a confidence interval of radius $\rho_{a,t}$).

Proof. Let $E \in \{0,1\}$ be 1 if the clean event happens and 0 otherwise. Lemma 1.1 gives us $\Pr(E=0) \le \frac{2}{T^2}$. Thus

$$\begin{aligned}\mathbf{E}[R] &= \Pr(E=1)\mathbf{E}[R|E=1] + \Pr(E=0)\mathbf{E}[R|E=0] \\ &\le \mathbf{E}[R|E=1] + \frac{2}{T^2}O(T) \\ &\equiv \mathbf{E}[R|E=1]\end{aligned}$$

where $\equiv$ denotes equivalence as asymptotic functions of T. □

2 Algorithms

2.1 Greedy

1. Pull each of the K arms N times. Let $\hat{a}$ denote the arm that has the highest reward estimate.
2. For $t = KN+1\dots T$, pull $a_t = \hat{a}$.

If N is not a function of T, the greedy algorithm clearly incurs a linear regret since the probability of $\hat{a} \neq a^\star$ is nonzero no matter how large N is. Intuitively, the algorithm is "bad" because it stops exploring after the first KN steps. If we use the knowledge of T, we can achieve the following sublinear bound.

Lemma 2.1. The greedy algorithm has a regret bound of $O(T^{2/3}(K\log T)^{1/3})$ for $N = \lfloor (T/K)^{2/3}(\log T)^{1/3} \rfloor$.

Proof. We condition on the clean event (5) without loss of generality. In this case, every arm has the same confidence radius $\rho = \sqrt{\frac{2\log T}{N}}$ after the first KN steps. Suppose $\hat{a} \neq a^\star$. Then

$$\begin{aligned}\hat{\mu}_{a^\star} \le \hat{\mu}_{\hat{a}} \quad &\Longrightarrow \quad \mu^\star - \rho < \mu_{\hat{a}} + \rho \\ &\Longrightarrow \quad \mu^\star - \mu_{\hat{a}} < 2\rho\end{aligned}$$

Thus

$$\mathbf{E}[R|E=1] \le KN + (T-KN)\mathbf{E}[\mu^\star - \mu_{\hat{a}}] \le KN + 2T\rho = KN + 2\alpha N^{-1/2}$$

where $\alpha = T\sqrt{2\log T}$. The function $f(N) = KN + 2\alpha N^{-1/2}$ is convex and minimized by $N^\star = K^{-2/3}\alpha^{2/3}$ at $f(N^\star) = 3K^{1/3}\alpha^{2/3}$. This implies the statement. □

2.2 UCB

For $t = 1 \dots T$, pull $a_t = \arg\max_{a=1}^{K} \hat{\mu}_{a,t} + \underbrace{\sqrt{\frac{2 \log T}{N_{a,t}}}}_{\rho_{a,t}}$.

Unlike the greedy algorithm, the UCB algorithm adaptively controls the exploration-exploitation tradeoff by using the upper confidence bound $\hat{\mu}_{a,t} + \rho_{a,t}$ as the selection criterion. Eventually all confidence intervals will shrink to point estimates and the UCB algorithm will only exploit. The UCB algorithm is "optimistic" because it assumes the best possible reward for any estimate. Intuitively, this optimism encourages exploring less explored arms (i.e., when $N_{a,t}$ is small, the algorithm recommends selecting a).

Lemma 2.2. The UCB algorithm has a regret bound of $O(\sqrt{KT \log T})$.

Proof. Again, we condition on the clean event (5) without loss of generality. Let a be any arm such that $\mu_a < \mu^\star$ and suppose we picked a at some step t. It must be the case that

$$\begin{aligned}
& \hat{\mu}_{a^\star} + \rho_{a^\star,t} \le \hat{\mu}_a + \rho_{a,t} \\
\implies \quad & \mu^\star \le \mu_a + 2\rho_{a,t} && \text{(since } \mu^\star \le \hat{\mu}_{a^\star} + \rho_{a^\star,t} \text{ and } \mu_a \ge \hat{\mu}_a - \rho_{a,t}\text{)} \\
\implies \quad & \mu^\star - \mu_a \le 2\rho_{a,t} = \sqrt{\frac{8 \log T}{N_{a,t}}} && (6)
\end{aligned}$$

Thus the regret caused by a up to step t is at most

$$R_{a,t} = N_{a,t}(\mu^\star - \mu_a) \le \sqrt{8 N_{a,t} \log T} \tag{7}$$

The regret up to step t is then at most

$$\begin{aligned}
R_t = \sum_{a=1}^{K} R_{a,t} &\le \sqrt{8 \log T} \left(\sum_{a=1}^{K} \sqrt{N_{a,t}} \right) \\
&\le \sqrt{8K \log T} \sqrt{\sum_{a=1}^{K} N_{a,t}} && \left(\text{Jensen's inequality: } \frac{1}{K} \sum_k \sqrt{x_k} \le \sqrt{\frac{1}{K} \sum_k x_k} \right) \\
&= \sqrt{8Kt \log T} && \left(\text{using } \sum_{a=1}^{K} N_{a,t} = t \right)
\end{aligned}$$

In particular, $E[R] = O(\sqrt{KT \log T})$. □

Interestingly, the optimism in UCB seems necessary. Consider the "LCB algorithm" which assumes the worst possible reward and selects $a_t = \arg\max_{a=1}^{K} \hat{\mu}_{a,t} - \rho_{a,t}$. Its pessimism discourages exploring less explored arms (and encourages exploiting more explored arms). Taking similar analytic steps, we can bound $\mu^\star - \mu_a \le 2\rho_{a^\star,t}$ only as a function of the best arm, which prevents us from deriving an arm-wise regret bound like (7).

While the exact formula in the UCB algorithm is not essential in practice, the general idea of exploring action a which has been chosen $N_{s,a}$ times at state s by considering a term $\propto \frac{1}{N_{s,a}}$ is useful. For instance, in Monte Carlo Tree Search (Appendix C), we aim to improve upon a raw policy $\pi_\theta(\cdot|s)$ by performing a few rounds of UCB-style exploration (31), then picking the most frequently selected action.

2.2.1 Instance-specific regret upper bound

Lemma 2.3. The UCB algorithm has a regret bound of $O(C_{\mathcal{I}} \log T)$, where $C_{\mathcal{I}}$ is an instance-specific constant.

Proof. Following the same steps to (6), conditioning on the clean event, if the UCB algorithm picks a suboptimal arm a at step t, we must have

$$N_{a,t} \le \frac{8 \log T}{(\mu^\star - \mu_a)^2}$$

(i.e., the suboptimal arm must have been chosen "not that many times"). Thus

$$R_{a,t} = N_{a,t}(\mu^\star - \mu_a) \le \frac{8 \log T}{\mu^\star - \mu_a}$$

and

$$R_t = \sum_{a=1}^{K} R_{a,t} \le \underbrace{\left(\sum_{a=1}^{K} \frac{8}{\mu^\star - \mu_a} \right)}_{C_{\mathcal{I}}} \log T$$

□

Theorem A.1 implies that the UCB algorithm is asymptotically optimal (for problem instances with $\mu^\star \in (0,1)$).

2.3 Thompson Sampling

Thompson sampling derives the policy $\pi(\cdot|s_{t-1})$ by Bayesian principles. First, it assumes a known reward distributional form $D_a = \text{Known}(\mu_a)$, where the only unknown is the mean parameter $\mu_a \in [0,1]$ (e.g., $D_a = \text{Ber}(\mu_a)$ with binary rewards), for each arm a. Thus a bandit environment is completely specified by the mean vector $\mu \in [0,1]^K$. Next, at each step t with history s_{t-1}, it assumes a posterior distribution $q_a(\cdot|s_{t-1})$ over the arm a's mean parameter $\mu_a^{(t)} \in [0,1]$.[2] We define the policy as the associated distribution over the best arm, specifically

$$\pi(a|s_{t-1}) = \Pr_{\mu_a^{(t)} \sim q_a(\cdot|s_{t-1}) \; \forall a} \left(a = \arg\max_{a=1}^{K} \mu_a^{(t)} \right)$$

After sampling $a_t \sim \pi(\cdot|s_{t-1})$ and receiving $r_t \sim D_{a_t}$, we update the posterior distribution for a_t by the Bayes rule:

$$q_{a_t}(\mu'|s_t) \propto q_{a_t}(\mu'|s_{t-1}) \times \text{Known}(\mu'_{a_t})(r_t) \tag{8}$$

By definition, Thompson sampling samples from the best arm distribution, which leads to favorable regret analysis (e.g., it admits a tight regret upper bound in certain settings). We omit regret analysis, but give an instantiation of Thompson sampling with (1) a Bernoulli reward distribution $D_a = \text{Ber}(\mu_a)$, and (2) a beta poterior $q_a(\cdot|s_{t-1}) = \text{Beta}(\alpha_a, \beta_a)$ where $\alpha_a, \beta_a > 0$. In this setting, (8) is given in closed form by the conjugacy of the beta-binomial distribution:

$$\text{Beta}(\alpha + r_t, \beta + (1 - r_t))(\mu') \propto \text{Beta}(\alpha, \beta)(\mu') \times \text{Ber}(\mu'_{a_t})(r_t)$$

(i.e., just keep track of the number of heads vs tails in each arm). Starting with $\alpha_a = \beta_a = 1$ yields the uniform prior $q_a(\cdot|s_0) = \text{Unif}_{[0,1]}$. The resulting algorithm is given below:

1. Start from the uniform prior: $\alpha \leftarrow 1_K$, $\beta \leftarrow 1_K$
2. For $t = 1 \dots T$,
   (a) Posterior sampling: $\mu_a^{(t)} \sim \text{Beta}(\alpha_a, \beta_a)$ for $a = 1 \dots K$
   (b) Pull $a_t \leftarrow \arg\max_{a=1}^{K} \mu_a^{(t)}$ and receive $r_t \in \{0,1\}$ from $\text{Ber}(\mu_{a_t})$ (μ_{a_t} unknown).
   (c) Update the posterior: $\alpha_{a_t} \leftarrow \alpha_{a_t} + r_t$, $\beta_{a_t} \leftarrow \beta_{a_t} + (1 - r_t)$.

We can consider other conjugate distributions (e.g., Gaussian). If conjugacy is not an option, there are many methods for approximate Bayesian inference (e.g., Gibbs sampling, gradient-based Langevin Monte Carlo).

References

Shaw, P., Joshi, M., Cohan, J., Berant, J., Pasupat, P., Hu, H., Khandelwal, U., Lee, K., and Toutanova, K. (2023). From pixels to ui actions: Learning to follow instructions via graphical user interfaces. *arXiv preprint arXiv:2306.00245*.

Slivkins, A. *et al.* (2019). Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, **12**(1-2), 1–286.

[2] For simplicity, we use independent arms for the posterior ("mean field approximation"), instead of a joint posterior $q(\cdot|s_{t-1})$ over $(\mu_1^{(t)} \dots \mu_K^{(t)}) \in [0,1]^K$. In this formulation, $q_a(\cdot|s_{t-1})$ only needs to use samples of arm a in the history s_{t-1}.

A Lower Bounds

We assume binary rewards (i.e., D_a is a Bernoulli distribution), which is sufficient for the purpose of lower bounds.

A.1 $K = 2$

For any $\epsilon \in (0, \frac{1}{2})$ we may construct an adversarial environment by

$$\mu_{a^\star \sim \text{Unif}\{1,2\}} = \frac{1+\epsilon}{2} \qquad \mu_{a \neq a^\star} = \frac{1}{2} \tag{9}$$

In this case, if $T < \frac{1}{16\epsilon^2}$, then $\Pr(a_t \neq a^\star) > \frac{1}{8}$ for all $t = 1 \dots T$ (Section A.1.1). Thus

$$\mathbf{E}[R] > \sum_{t=1}^{T} \left(\frac{1}{8}\right)\left(\frac{\epsilon}{2}\right) = \frac{\epsilon T}{16}$$

Given any T, we can choose $\epsilon = \frac{1}{5\sqrt{T}}$ and construct (9). This choice satisfies $T < \frac{1}{16\epsilon^2}$, thus $\mathbf{E}[R] > \frac{\sqrt{T}}{80}$.

A.1.1 Risk analysis

Let $G \in \{0,1\}^{2\times T}$ denote the "reward situation". Any bandit algorithm can be seen as first sampling the entire G from the environment, then looking up $G_{a,i} \in \{0,1\}$ as the reward of pulling arm a for the i-th time. The sampling process for G is: for each $a \in \{1,2\}$ independently, draw $G_{a,1} \dots G_{a,T} \sim D_a$ independently. For the ϵ-adversarial environment in (9), there are only two possible distributions over G corresponding to $a^\star = 1$ vs $a^\star = 2$. For a *deterministic* algorithm (i.e., conditioned on the past reward information, the current arm selection is a point-mass distribution), a particular reward situation $G = g$ fully determines the choice of the arm $a_t \in \{1,2\}$ at any step t.[3] Thus a_t is random only with respect to the choice of G, in particular

$$\Pr(a_t = 1) = \sum_{g \in \{0,1\}^{2\times T}:\ a_t = 1} \Pr(G = g)$$

(i.e., $a_t = 1$ is a measurable event under a distribution over G). This allows us to use the following flurry of logic:

$$\begin{aligned}
&2\left(\Pr(a_t = 1|a^\star = 1) - \Pr(a_t = 1|a^\star = 2)\right)^2 \\
&\leq \text{KL}\left(\Pr(G|a^\star = 1), \Pr(G|a^\star = 2)\right) && (10)\\
&= \sum_{a\in\{1,2\}} \sum_{t=1}^{T} \text{KL}\left(\Pr(G_{a,t}|a^\star = 1), \Pr(G_{a,t}|a^\star = 2)\right) && (11)\\
&= T \times \text{KL}\left(\text{Ber}\left(\frac{1+\epsilon}{2}\right), \text{Ber}\left(\frac{1}{2}\right)\right) + T \times \text{KL}\left(\text{Ber}\left(\frac{1}{2}\right), \text{Ber}\left(\frac{1+\epsilon}{2}\right)\right) && (12)\\
&\leq 3\epsilon^2 T && (13)
\end{aligned}$$

where (10) is Pinsker's inequality, (11) is by the iid assumption on rewards, (12) is by the definition of the environment, and (13) is by the property of KL divergence between Bernoulli distributions (Lemma B.1). All this work is driven to obtain the bound

$$|\Pr(a_t = 1|a^\star = 1) - \Pr(a_t = 1|a^\star = 2)| \leq 2\epsilon\sqrt{T} < \frac{1}{2} \tag{14}$$

where the last inequality follows with the (retrospective) premise that $T < \frac{1}{16\epsilon^2}$. Now suppose there exists an algorithm with the property that

$$\Pr(a_t = a|a^\star = a) > \frac{3}{4} \qquad \forall a \in \{1,2\} \tag{15}$$

[3] The proof is by strong induction. When $t = 1$, there is no past reward and the claim holds by definition. When $t > 1$, assume that the claim holds for all $t' < t$. This means $a_1 \dots a_{t-1}$ are fixed. Thus, conditioned on a particular $G = g$, the corresponding rewards $r_1 \dots r_{t-1}$ are fixed where $r_{t'} = g_{a_{t'},t'}$. By definition, the deterministic algorithm predicts some a_t with probability 1.

But it implies $\Pr(a_t = 1|a^\star = 1) - \Pr(a_t = 1|a^\star = 2) > \frac{1}{2}$, contradicting (14). Therefore, for any algorithm, there is some arm $a_{\text{tricky}} \in \{1, 2\}$ such that

$$\Pr(a_t = a_{\text{tricky}}|a^\star = a_{\text{tricky}}) \leq \frac{3}{4}$$

This allows us to make the main conclusion: for any step $t \leq T < \frac{1}{16\epsilon^2}$,

$$\Pr(a_t \neq a^\star) \geq \underbrace{\Pr(a^\star = a_{\text{tricky}})}_{=\frac{1}{2}} \times \underbrace{\Pr(a_t \neq a_{\text{tricky}}|a^\star = a_{\text{tricky}})}_{>\frac{1}{4}} > \frac{1}{8} \tag{16}$$

Finally, the whole argument assumes only deterministic bandit algorithms (e.g., greedy and UCB, with ties broken by some deterministic procedure). However, the main conclusion also applies to stochastic algorithms because a stochastic algorithm can be expressed as an expectation over determinstic algorithms Dt. Formally, suppose there is some stochastic algorithm that achieves $\Pr(a_t \neq a^\star) \leq \frac{1}{8}$. Since $\Pr(a_t \neq a^\star) = \mathbf{E}_{\text{Dt}}[\Pr(a_t \neq a^\star|\text{Dt})]$, there must be some deterministic algorithm $\text{Dt} = \text{dt}$ such that $\Pr(a_t \neq a^\star|\text{Dt} = \text{dt}) \leq \frac{1}{8}$, contradicting (16).

A.2 General K

The case with $K = 2$ arms (Section A.1) already proves that no bandit algorithm can obtain sub-$\Omega(\sqrt{T})$ expected regret for all environments. We can improve the bound by incorporating K. The proof has a similar spirit but is substantially more subtle. Fix any bandit algorithm. For $\epsilon \in (0, \frac{1}{2})$ we construct an adversarial environment by

$$\mu_{a^\star \sim \text{Unif}\{1,\ldots,K\}} = \frac{1+\epsilon}{2} \qquad \mu_a = \frac{1}{2} \quad \forall a \neq a^\star \tag{17}$$

We will also need to consider the following "no-winner" environment

$$\mu_a = \frac{1}{2} \qquad \forall a \in \{1 \ldots K\} \tag{18}$$

For clarity, we will write

- P_0 to denote the distribution under the no-winner environment (18)
- P_a to denote the distribution under the adversarial environment (17) with $a^\star = a$

The high-level argument goes: at any step t, some arms are "neglected" under P_0 by inherent classification risks.[4] But P_a is difficult to distinguish from P_0.[5] Therefore, if $a^\star = a$ is one of the neglected arms in P_0, it will be neglected in P_a as well. But this means the algorithm is suboptimal.

A.2.1 Risk analysis: naive version

Pick any t. Assuming $K \geq 12$, by Lemma B.3 at least 66% of the arms $\mathcal{A}_0 \subset \{1 \ldots K\}$ satisfy

$$P_0(a_t = a_0) < \frac{1}{4} \qquad \forall a_0 \in \mathcal{A}_0 \tag{19}$$

Fix any $a_0 \in \mathcal{A}_0$. Let $G \in \{0, 1\}^{K \times T}$ denote the reward situation with $G_{a,t} \sim D_a$ independently. Since the event $a_t = a_0$ is measurable under any distribution over G with a (WLOG) deterministic algorithm (see A.1.1 for a recap), we can use Pinsker's inequality to argue

$$\begin{aligned}
2\left(P_{a_0}(a_t = a_0) - P_0(a_t = a_0)\right)^2 &\leq \text{KL}\left(P_{a_0}(G), P_0(G)\right) \\
&= \sum_{a=1}^{K}\sum_{t=1}^{T} \text{KL}\left(P_{a_0}(G_{a,t}), P_0(G_{a,t})\right) \\
&= \sum_{t=1}^{T} \text{KL}\left(P_{a_0}(G_{a_0,t}), P_0(G_{a_0,t})\right) \qquad (20) \\
&\leq 2\epsilon^2 T \qquad (21)
\end{aligned}$$

[4] Here, we are not using any special properties of P_0 to impose the classification risks. Any distribution is subject to the risks.

[5] Note that we have to go through P_0 to apply classification risks, since P_a is conditioned on some particular arm a which may not be at risk.

where (20) and (21) crucially exploit the definition of P_{a_0} and P_0 (and Lemma B.1). Thus $|P_{a_0}(a_t = a_0) - P_0(a_t = a_0)| \leq \epsilon\sqrt{T}$, which implies if $T \leq \frac{1}{16\epsilon^2}$:

$$\begin{aligned} P_{a_0}(a_t = a_0) &\leq P_0(a_t = a_0) + \frac{1}{4} \\ &\leq \frac{1}{2} \qquad \text{(by the choice of } a_0 \in \mathcal{A}_0\text{, see (19))} \end{aligned}$$

By construction, the optimal arm $a^\star$ in the adversarial environment is chosen uniformly at random, thus $\Pr(a^\star \in \mathcal{A}_0) \geq \frac{2}{3}$. Thus if $K \geq 12$ and $T \leq \frac{1}{16\epsilon^2}$, for any $t \leq T$, the algorithm satisfies

$$\Pr(a_t \neq a^\star) \geq \underbrace{\Pr(a^\star = a_0)}_{\geq \frac{2}{3}} \times \underbrace{P_{a_0}(a_t \neq a_0)}_{> \frac{1}{2}} > \frac{1}{3}$$

where a_0 is any neglected arm in $\mathcal{A}_0$. Then

$$\mathbf{E}\left[R\right] > \sum_{t=1}^{T} \left(\frac{1}{3}\right)\left(\frac{\epsilon}{2}\right) = \frac{\epsilon T}{6}$$

Given T, choose $\epsilon = \frac{1}{5\sqrt{T}}$. This choice satisfies $T < \frac{1}{16\epsilon^2}$, thus $\mathbf{E}\left[R\right] > \frac{\sqrt{T}}{30}$ under the environment (17).

A.2.2 Risk analysis: improved version

The naive argument only yields a $\Omega(\sqrt{T})$ regret lower bound. To introduce a dependence on K, we must avoid considering all T samples for the neglected arm a_0 in (20). Achieving this requires an artificial surgery of the sample space and strengthening the notion of neglection. Define a restricted sample space

$$\Omega^\star = \{0,1\}^T \times \cdots \times \underbrace{\{0,1\}^m}_{(j\text{-th space})} \times \cdots \times \{0,1\}^T$$

where $j \in \{1 \dots K\}$ is the **budget arm** and $m \in [1, T]$ is the **budget**.[6] For any full reward distribution P over $G \in \{0,1\}^{K \times T}$ (e.g., P_a or P_0), we define an associated restricted distribution $P^\star$ over $G^\star \in \Omega^\star$ by the independent sampling process

$$G^\star_{j,1} \dots G^\star_{j,m} \overset{\text{iid}}{\sim} D_j \qquad\qquad G^\star_{a,1} \dots G^\star_{a,T} \overset{\text{iid}}{\sim} D_a \quad \forall a \neq j$$

where $D_1 \dots D_K$ are given by P. Here is the central trick: if O is any measurable event under P whose outcome is completely determined by m samples of arm j and T samples of each $a \neq j$ (henceforth **budget event**), then

$$P(O = o) = \sum_{g \in \{0,1\}^{K \times T} \colon O = o} \Pr\left(G = g\right)$$

and

$$P^\star(O = o) = \sum_{g \in \Omega^\star \colon O = o} \Pr\left(G^\star = g\right)$$

are the same. In the latter version, we simply change the *perspective* on the event and "precompute" the iid rewards more economically for the budget arm since we do not need more than m samples from it (i.e., it does not affect viewing the algorithm as retrieving $G^\star_{a,t}$ upon selecting the arm a for the t-th time). Consequently, for any budget event O we can follow similar steps as before and provide the budget bound (for any choice of $a^\star = a$)

$$\begin{aligned} 2\left(P^\star_a(O = o) - P^\star_0(O = o)\right)^2 &\leq \mathrm{KL}\left(P^\star_a(G^\star), P^\star_0(G^\star)\right) \\ &= \sum_{t=1}^{m} \mathrm{KL}\left(P^\star_a(G^\star_{j,t}), P^\star_0(G^\star_{j,t})\right) + \sum_{a \neq j} \sum_{t=1}^{T} \mathrm{KL}\left(P^\star_a(G^\star_{a,t}), P^\star_0(G^\star_{a,t})\right) \\ &\leq 2\epsilon^2 m \end{aligned}$$

[6] We suppress the notation since otherwise it becomes unwieldy. We will make the choice of j and m extremely clear in the context.

which implies

$$m \le \frac{1}{64\epsilon^2} \qquad \Rightarrow \qquad |P^\star_a(O=o) - P^\star_0(O=o)| \le \frac{1}{8} \tag{22}$$

Now pick any $t \le T$. Assuming $K \ge 24$, by Corollary B.6 at least 33% of the arms $\mathcal{A}_0 \subset \{1 \dots K\}$ satisfy

$$P_0(a_t = a_0) \le \frac{1}{8} \quad \bigwedge \quad P_0\left(N_{a_0,t} \ge \frac{24t}{K}\right) < \frac{1}{8} \qquad \forall a_0 \in \mathcal{A}_0 \tag{23}$$

Pick any $a_0 \in \mathcal{A}_0$. In the following, we assume the budget arm $j = a_0$ and the budget $m = \frac{24t}{K} \le t$ with the superscript $\star$. If $T \le \frac{K}{1536\epsilon^2}$, then

$$\begin{aligned}
P_{a_0}(a_t = a_0) &\le P_{a_0}\left(a_t = a_0 \wedge N_{a_0,t} < m\right) + P_{a_0}\left(N_{a_0,t} \ge m\right) && (24)\\
&= P^\star_{a_0}\left(a_t = a_0 \wedge N_{a_0,t} < m\right) + P^\star_{a_0}\left(N_{a_0,t} \ge m\right) && (25)\\
&\le P^\star_0\left(a_t = a_0 \wedge N_{a_0,t} < m\right) + P^\star_0\left(N_{a_0,t} \ge m\right) + \frac{1}{4} && (26)\\
&= P_0\left(a_t = a_0 \wedge N_{a_0,t} < m\right) + P_0\left(N_{a_0,t} \ge m\right) + \frac{1}{4}\\
&\le P_0\left(a_t = a_0\right) + P_0\left(N_{a_0,t} \ge m\right) + \frac{1}{4}\\
&\le \frac{1}{2} && (27)
\end{aligned}$$

where (25) follows because $N_{a_0,t} < m$ is a budget event;[7] (26) uses (22) with the assumption on T; and (27) exploits the choice of $a_0 \in \mathcal{A}_0$. The first step (24) is necessary to apply $\star$ because $a_t = a_0$ is not a budget event. To see why, suppose $K \gg 24$ so that $m \ll t$. Then the choice of arm at step t will generally require more than m samples from a_0. The rest of the argument is similar to the naive version. By construction, the optimal arm $a^\star$ in the adversarial environment is chosen uniformly at random, thus $\Pr(a^\star \in \mathcal{A}_0) \ge \frac{1}{3}$. Thus if $K \ge 24$ and $T \le \frac{K}{1536\epsilon^2}$, for any $t \le T$, the algorithm satisfies

$$\Pr(a_t \ne a^\star) \ge \underbrace{\Pr(a^\star = a_0)}_{\ge \frac{1}{3}} \times \underbrace{P_{a_0}(a_t \ne a_0)}_{> \frac{1}{2}} > \frac{1}{6}$$

where a_0 is any neglected arm in $\mathcal{A}_0$. Then

$$\mathbf{E}\left[R\right] > \sum_{t=1}^{T} \left(\frac{1}{6}\right)\left(\frac{\epsilon}{2}\right) = \frac{\epsilon T}{12}$$

Given T, choose $\epsilon = \frac{1}{40}\sqrt{\frac{K}{T}}$. This choice satisfies $T < \frac{K}{1536\epsilon^2}$, thus $\mathbf{E}\left[R\right] > \frac{\sqrt{KT}}{480}$ under the environment (17).

A.3 Instance-Depedent Lower Bounds

Previous sections show that no algorithm can achieve sub-$\Omega(\sqrt{t})$ expected regret on *all* problem instances—because we can always produce an adversarial instance where it must have that much expected regret. More formally,

$$\mathbf{E}\left[R_t\right] \ge \Omega(C\sqrt{t})$$

where R_t is the regret at step t and $C > 0$ is constant for all problem instances. However, we may do better if we consider instance-dependent lower bounds. For instance, the UCB algorithm achieves $\mathbf{E}\left[R_t\right] \le O(C_{\mathcal{I}} \log t)$ where $C_{\mathcal{I}}$ is an constant specific for the problem instance $\mathcal{I}$ (Lemma 2.3). Is it possible to do even better than the logarithmic dependence in this setting? The answer is no, as dictated by the following theorem (without proof).

Theorem A.1. Pick a problem instance $\mathcal{I}$ and any bandit algorithm that achieves $\mathbf{E}[R_t] \le O(C_{\mathcal{I},\alpha} t^\alpha)$ for some $\alpha > 0$. Then $\mathbf{E}[R_t] \ge \Omega(C_{\mathcal{I}} \log t)$, which holds for any of the two constants:

$$C_{\mathcal{I}} = \sum_{a=1:\ \mu^\star - \mu_a > 0}^{K} \frac{\mu^\star(1-\mu^\star)}{\mu^\star - \mu_a} \qquad \text{(weaker version)} \tag{28}$$

$$C_{\mathcal{I}} = \sum_{a=1:\ \mu^\star - \mu_a > 0}^{K} \frac{\mu^\star - \mu_a}{\mathrm{KL}\left(D_a, D^\star\right)} - 2\epsilon \qquad \forall \epsilon > 0 \qquad \text{(stronger version)} \tag{29}$$

[7] It is a bit strange to think about it explicitly, but $N_{a_0,t} < m$ is indeed a budget event because the algorithm does not require more than m samples of arm a_0 for the task of selecting the arm fewer than m times.

The condition in the theorem is necessary to rule out trivial cases. Without such a condition, we may select an instance with $a^\star = 1$ and an algorithm that selects $a_t = 1$, which achieves $\mathbf{E}[R_t] = 0$.

B Lemmas

Lemma B.1. For $\epsilon \in [0, \frac{1}{2}]$, $\mathrm{KL}(\mathrm{Ber}(\frac{1+\epsilon}{2}), \mathrm{Ber}(\frac{1}{2})) \leq 2\epsilon^2$ and $\mathrm{KL}(\mathrm{Ber}(\frac{1}{2}), \mathrm{Ber}(\frac{1+\epsilon}{2})) \leq \epsilon^2$.

Proof.

$$\begin{aligned}
\mathrm{KL}\left(\mathrm{Ber}\left(\frac{1+\epsilon}{2}\right), \mathrm{Ber}\left(\frac{1}{2}\right)\right) &= \left(\frac{1+\epsilon}{2}\right)\log(1+\epsilon) + \left(\frac{1-\epsilon}{2}\right)\log(1-\epsilon) \\
&= \frac{1}{2}\underbrace{\log(1-\epsilon^2)}_{\leq 0} + \frac{\epsilon}{2}\underbrace{\log\left(\frac{1+\epsilon}{1-\epsilon}\right)}_{=\log\left(1+\frac{2\epsilon}{1-\epsilon}\right)\leq\frac{2\epsilon}{1-\epsilon}\leq 4\epsilon} \leq 2\epsilon^2 \\
\mathrm{KL}\left(\mathrm{Ber}\left(\frac{1}{2}\right), \mathrm{Ber}\left(\frac{1+\epsilon}{2}\right)\right) &= \frac{1}{2}\underbrace{\log\left(\frac{1}{1-\epsilon^2}\right)}_{=\log\left(1+\frac{\epsilon^2}{1-\epsilon^2}\right)\leq\frac{\epsilon^2}{1-\epsilon^2}\leq 2\epsilon^2} \leq \epsilon^2
\end{aligned}$$

□

Lemma B.2 (Warmup)**.** Fix $\epsilon \in (0, \frac{1}{2})$. Let $p = \mathrm{Ber}(\frac{1+\epsilon}{2})$ and $q = \mathrm{Ber}(\frac{1}{2})$. Let $Z \sim \mathbf{Unk}$ and $X_1 \dots X_T \overset{\text{iid}}{\sim} Z$ where $\mathbf{Unk}$ is some unknown distribution over $\{p, q\}$. Then for any predictor $f : \{0,1\}^T \to \{p, q\}$,

$$\Pr\left(f(X_1 \dots X_T) = Z\right) > \frac{3}{4} \qquad \Longrightarrow \qquad T \geq \frac{1}{4\epsilon^2}$$

Proof. The sample space of possible configurations is $\Omega = \{p, q\} \times \{0,1\}^T$, with the joint distribution

$$l_\Omega(z, x) = \mathbf{Unk}(z) \times \prod_{t=1}^{T} z(x_t)$$

and the conditional distribution $l_\Omega(x|z) = \prod_{t=1}^{T} z(x_t)$. Assuming a deterministic f, the event $f(X_1 \dots X_T) = Z$ is measurable under l_Ω because a sample $(z, x) \sim l_\Omega$ completely determines if the event happens. Formally,

$$\Pr\left(f(X_1 \dots X_T) = Z\right) = \sum_{(z,x)\in\Omega:\ f(x)=z} l_\Omega(z, x)$$

If f is nondeterministic, the event is still measurable with

$$\Pr\left(f(X_1 \dots X_T) = Z\right) = \sum_{(z,x)\in\Omega} l_\Omega(z, x) \times \Pr\left(f(x) = z\right) = \mathop{\mathbf{E}}_{(z,x)\sim l_\Omega}\left[\Pr\left(f(x) = z\right)\right]$$

(which clearly subsumes the deterministic f as a special case). This enables the use of Pinsker's inequality:

$$\begin{aligned}
2\left(\Pr\left(f(X_1 \dots X_T) = Z | Z = p\right) - \Pr\left(f(X_1 \dots X_T) = Z | Z = q\right)\right)^2 &\leq \mathrm{KL}\left(l_\Omega(\cdot|p), l_\Omega(\cdot|q)\right) \\
&= \mathrm{KL}\left(p^T, q^T\right) \\
&= T \times \mathrm{KL}\left(p, q\right) \\
&\leq 2T\epsilon^2 \qquad (30)
\end{aligned}$$

Now, if $\Pr\left(f(X_1 \dots X_T) = Z\right) > \frac{3}{4}$, we must have $\Pr\left(f(X_1 \dots X_T) = p | Z = p\right) - \Pr\left(f(X_1 \dots X_T) = p | Z = q\right) > \frac{1}{2}$. To avoid contradicting the upper bound (30), we must have $\epsilon\sqrt{T} > \frac{1}{2}$, or $T > \frac{1}{4\epsilon^2}$. □

B.1 Classification Risk

Under any distribution over K values, at least 66% of the values have the probability at most $\frac{3}{K}$. Suppose otherwise. Then more than 33% have probability greater than $\frac{3}{K}$, and the total probability mass will be greater than 1. This risk is better understood at a high level like this, but a formal statement is given below for completeness.

Lemma B.3. Let $X \in \{1 \dots K\}$ be any random variable. There is some $S \subset \{1 \dots K\}$ with $|S| > \frac{2K}{3}$ such that $\Pr(X = x) < \frac{3}{K}$ for all $x \in S$.

Proof. Suppose that for all $S \subset \{1 \dots K\}$ with $|S| > \frac{2K}{3}$, we have $\Pr(X = x) \geq \frac{3}{K}$ for some $x \in S$. Take the largest subset $S^\star$ such that $\Pr(X = x) < \frac{3}{K}$ for all $x \in S^\star$, which implies $|S^\star| \leq \frac{2K}{3}$. Take $S' = \{1 \dots K\} \setminus S^\star$, whereupon $|S'| > \frac{K}{3}$. It must also be the case that $\Pr(X = x) \geq \frac{3}{K}$ for all $x \in S'$, since otherwise $S^\star$ could have been larger. Thus $\sum_{x=1}^{K} \Pr(X = x) \geq \sum_{x \in S'} \Pr(X = x) > \left(\frac{K}{3}\right)\left(\frac{3}{K}\right) = 1$. □

We have a similar risk in terms of counts. For the statements below, let $(X_1 \dots X_T) \in \{1 \dots K\}^T$ be any random sequence and define $N_x = \sum_{t=1}^{T} [[X_t = x]]$.

Lemma B.4. There is some $S \subset \{1 \dots K\}$ with $|S| > \frac{2K}{3}$ such that $\mathbf{E}[N_x] < \frac{3T}{K}$ for all $x \in S$.

Proof. Suppose that for all $S \subset \{1 \dots K\}$ with $|S| > \frac{2K}{3}$, we have $\mathbf{E}[N_x] \geq \frac{3T}{K}$ for some $x \in S$. Taking the same steps in the proof of Leamm B.3, we can construct a subset $S' \subset \{1 \dots K\}$ such that $|S'| > \frac{K}{3}$ and $\mathbf{E}[N_x] \geq \frac{3}{K}$ for all $x \in S'$. Thus $\sum_{x=1}^{K} \mathbf{E}[N_x] \geq \sum_{x \in S'} \mathbf{E}[N_x] > \left(\frac{K}{3}\right)\left(\frac{3T}{K}\right) = T$. □

Corollary B.5. There is some $S \subset \{1 \dots K\}$ with $|S| > \frac{2K}{3}$ such that $\Pr(N_x \geq \frac{24T}{K}) < \frac{1}{8}$ for all $x \in S$.

Proof. Since $N_x \geq 0$, we can apply Markov's inequality to have $\Pr(N_x \geq \epsilon) \leq \frac{\mathbf{E}[N_x]}{\epsilon}$ for any $\epsilon > 0$. Combining with Lemma B.4, there is some $S \subset \{1 \dots K\}$ with $|S| > \frac{2K}{3}$ such that $\Pr(N_x \geq \epsilon) < \frac{3T}{K\epsilon}$ for all $x \in S$ and $\epsilon > 0$. Solving for ϵ in $\frac{3T}{K\epsilon} = \frac{1}{8}$, we have $\Pr(N_x \geq \frac{24T}{K}) < \frac{1}{8}$. □

Corollary B.6. There is some $S \subset \{1 \dots K\}$ with $|S| > \frac{K}{3}$ such that $\Pr(X_T = x) < \frac{3}{K}$ and $\Pr(N_x \geq \frac{24T}{K}) < \frac{1}{8}$ for all $x \in S$.

Proof. At least $\frac{K}{3}$ of the values in Lemma B.3 and Corollary B.5 must overlap. □

C Monte Carlo Tree Search

MCTS (based on the variation in Shaw *et al.* (2023))
Input: policy $\pi_\theta(a|s)$, value network $v_\phi(s) \in \mathbb{R}$ (e.g., trained on human demonstrations labeled with short-path-encouraging surrogate rewards $r(s) = -\frac{1}{30} + [[s = \text{terminal}]] r_s$), state transition $\tau(s, a) \in \mathcal{S}$, beginning state $s_{\text{begin}} \in \mathcal{S}$, number of tree expansions K_{expand}, exploration weight $c = 0.1$, value network weight $\lambda = 0.1$
Output: next action $a_{\text{next}} \in \mathcal{A}$
State-action value estimates: $Q(s, a) = \text{Average}(r_0 = v_\phi(s), r_1, \dots, r_M)$ where r_i is the value associated with visiting (s, a) the i-th time

1. For K_{expand} times:
   (a) $s \leftarrow s_{\text{begin}}$
   (b) While $s = s_{\text{begin}}$ or s has been visited before, repeat:

$$a' \leftarrow \arg\max_{a \in \mathcal{A}} \; Q(s, a) + c \times \pi_\theta(a|s) \times \frac{\sqrt{N_s}}{1 + N_{s,a}} \tag{31}$$
$$N_s \leftarrow N_s + 1$$
$$N_{s,a'} \leftarrow N_{s,a'} + 1$$
$$s \leftarrow \tau(s, a')$$

   Let $(s_1, a_1), \dots, (s_T, a_T), s_{\text{leaf}}$ denote the traveled path.
   (c) $v(s_{\text{leaf}}) \leftarrow \lambda \times v_\phi(s_{\text{leaf}}) + (1 - \lambda) \times \text{Rollout}(\pi_\theta, s_{\text{leaf}})$
   (d) For $t = 1 \dots T$, update $Q(s_t, a_t) \leftarrow \text{NewAverage}(Q(s_t, a_t), v(s_{\text{leaf}}))$.
2. Return $a_{\text{next}} \leftarrow \arg\max_{a \in \mathcal{A}} \; N_{s_{\text{begin}}, a}$.