Bandits with Costly Reward Observations

Aaron D. Tucker[*]
Dept. of Comp. Sci.
Cornell University
Ithaca, NY

Caleb Biddulph[†]
Cornell University
Ithaca, NY

Claire Wang[†]
Cornell University
Ithaca, NY

Thorsten Joachims
Dept. of Comp. Sci.
Cornell University
Ithaca, NY

Abstract

Many Machine Learning applications are based on clever ways of constructing a large dataset from already existing sources in order to avoid the cost of labeling training examples. However, in settings such as content moderation with rapidly changing distributions without automatic ground-truth feedback, this may not be possible. If an algorithm has to pay for reward information, for example by asking a person for feedback, how does this change the exploration/exploitation tradeoff? We study Bandits with Costly Reward Observations, where a cost needs to be paid in order to observe the reward of the bandit's action. We show the impact of the observation cost on the regret by proving an $\Omega(c^{1/3}T^{2/3})$ lower bound, present a general non-adaptive algorithm which matches the lower bound, and present several competitive adaptive algorithms.

1 Introduction

Machine learning has proven extremely successful on tasks where accurately measurable rewards are readily available and abundant, such as in speech recognition or online advertising. However, there are many crucial settings such as content moderation where obtaining more reward information necessarily involves some costly interaction which would not automatically happen while running the service. In these situations, there is a tradeoff between collecting more labels to achieve better performance and collecting fewer labels to avoid the labeling cost. We study this problem in the classic bandit setting for analyzing exploration/exploitation tradeoffs. We also study a contextual bandit setting where the algorithm can decide that it needs additional oversight in some states but not others. This problem is highly relevant to scalable oversight (3), and more generally to the value alignment problem by explicitly accounting for the cost of collecting additional reward information, as would be the case for learning rewards from a human (6).

We study a setting, studied by prior authors in (7), where a bandit problem is modified by adding a decision at each time step t to pay or not pay a known cost c to observe the otherwise unknown reward r_t. As in standard bandit problems, a_t is the arm chosen at time t and r_t depends on a_t. There are many different types of bandit, and their definitions are deferred to the relevant sections.

We build on prior work in five ways, by 1) proving an information-theoretic lower bound of $\Omega(c^{1/3}T^{2/3})$ on the regret in the Bandits with Costly Reward Observations (BwCRO) (spoken bwick-roh) setting, 2) introducing a method which generalizes BwCRO to many bandit settings by turning suitable $O(T^{1/2})$ regret bandit algorithms into $O(c^{1/3}T^{2/3})$ regret BwCRO algorithms, 3) introducing a novel algorithm for simple multi-armed BwCRO which provably matches these lower bounds up to a logarithmic factor, 4) providing a novel heuristic algorithm for linear contextual BwCRO which can adaptively choose when to query for a label depending on the context, 5) and finally exploring the empirical performance of different algorithms in the BwCRO setting.

[*]corresponding author, contact at `aarondtucker@cs.cornell.edu`
[†]Equal contribution, authors listed alphabetically.

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

2 Lower bound

We prove that the BwCRO setting has a regret lower bound of $\Omega(c^{1/3}T^{2/3})$. This proves a novel rate for the labeling cost c, as well as agreeing with the rate for T proposed in (7) and (4). Our information-theoretic proof is based on the regret lower bound proof in (9). The full proof is in A.2.

Theorem 1. *The Bandits with Costly Observations setting has a regret lower bound of* $\Omega(c^{1/3}T^{2/3})$.

PROOF SKETCH. Define k multi-armed bandits, each with k actions where a coin is flipped with reward 1 for heads and 0 for tails. In each bandit j, coin j is biased with expected reward $(1+\epsilon)/2$, and all other $k-1$ coins are fair. Let Q_j^T denote the number of times coin j is played in T timesteps.

First, consider a base instance 0 where all coins are fair. At least $2/3$ of the k coins are such that $\mathbb{E}[Q_j^T] \leq 3T/k$, with randomness over the algorithm and fair coin flips in instance 0. For each of these coins, the Markov inequality implies that $Q_j^T \leq 6T/k$ with probability at least $1/2$. Therefore, in the base instance 0 for a uniformly randomly chosen j, $\Pr(Q_j^T \leq 6T/k) \geq 1/3$. Second, according to a KL divergence lemma, for any event A with at most n observations of the j^*th coin, the probability of the event A between the base instance and the instance j^* differs by at most $\epsilon\sqrt{n}$. Therefore $\Pr(Q_{j^*}^T < 6T/k) > 1/3 - \epsilon\sqrt{n}$. Finally, setting $\epsilon = \sqrt[3]{c/T}$ and assuming that the algorithm minimizes the regret with respect to n, we have an $\Omega(c^{1/3}T^{2/3})$ regret lower bound. □

3 Algorithms

3.1 Algorithm for all Bandits with Costly Reward Observations

With the lower bound in place we now provide a general method for converting bandit algorithms into algorithms for the corresponding BwCRO settings. Recall that for each of the t time steps, a_t is the arm chosen, r_t is the reward received, and that c is the labeling cost. Define a_t^* as the optimal action at time t and n as the total number of reward labels requested. For the sake of clarity, we define the regret ignoring label costs as $\text{Regret}^\circ = \sum_t \left(\mathbb{E}\left[r_t|a_t^*\right] - \mathbb{E}\left[r_t|a_t\right]\right)$, and we define the regret of the setting to be $\text{cRegret} = \text{Regret}^\circ + cn$.

We now introduce our general BwCRO algorithm for converting suitable bandit algorithms into BwCRO algorithms. It depends on a Uniform Regret assumption that a base bandit algorithm can be stopped after any number of rounds n and suffer an expected regret of $O(n^{1/2})$.

Assumption 3.1 (Uniform Regret Rate)**.** For all $n \leq T$, playing algorithm $\mathcal{A}$ while observing labels at every timestep results in a regret of $\mathbb{E}\left[\text{Regret}_{1:n}^\circ\right] \in O(n^{1/2})$, and the average regret of playing according to $\mathcal{A}$ without further labels is s.t. $\frac{1}{T-n}\mathbb{E}\left[\text{Regret}_{n+1:T}^\circ\right] \leq \frac{1}{n}\mathbb{E}\left[\text{Regret}_{1:n}^\circ\right]$.

Algorithm 1 Fixed-N Algorithm for Multi-armed Bandits

Given: Algorithm $\mathcal{A}$ for setting $\mathcal{B}$ that satisfies the Assumption 3.1 with $\mathbb{E}\left[\text{Regret}_{1:n}^\circ\right] \leq K\sqrt{n}$,
Phase 1: Play according to $\mathcal{A}$ while observing the first $n = \left(\frac{TK}{2c}\right)^{2/3}$ labels.
Phase 2: Play according to $\mathcal{A}$ without observing additional labels.

The following theorem shows that Algorithm 1 matches our regret lower bound in the worst case, though only to a logarithmic factor if the original algorithm has $\tilde{O}(T^{1/2})$ regret.

Theorem 2 (Fixed N Regret Rate)**.** *Assuming that $\mathcal{A}$ satisfies the uniform regret assumption, the Fixed N algorithm based on $\mathcal{A}$ has cRegret $\in O(c^{1/3}T^{2/3})$, matching the lower bound.*

PROOF SKETCH. Assume that $\mathcal{A}$ satisfies the Uniform Regret assumption, and that $\mathbb{E}\left[\text{Regret}_{1:n}^\circ\right] \leq K\sqrt{n}$ for all $n > n_0$ for some n_0. In the BwCRO setting, receiving n labels necessarily incurs a regret of cn, so the total regret of using $\mathcal{A}$ while labeling the first $n > n_0$ can be bounded as follows:

$$\begin{aligned}
\text{cRegret}_{1:T} &= cn + \mathbb{E}\left[\text{Regret}_{1:n}^\circ\right] + (T-n)\,\mathbb{E}\left[\text{Regret}_{n+1:T}^\circ\right]/(T-n) \\
&\leq cn + n\,\mathbb{E}\left[\text{Regret}_{1:n}^\circ\right]/n + (T-n)\,\mathbb{E}\left[\text{Regret}_{1:n}^\circ\right]/n && \text{by Uniform Regret Rate} \\
&\leq cn + TKn^{-1/2} && \text{by def. of } O(\sqrt{n}).
\end{aligned}$$

As shown in Appendix A.3, $cn + TKn^{-1/2}$ is minimized by $n = \left(\frac{TK}{2c}\right)^{2/3}$. Plugging this value of n into the original bound $cn + TKn^{-1/2}$ yields the regret $O(c^{1/3}K^{2/3}T^{2/3}) \subset O(c^{1/3}T^{2/3})$. □

Algorithm 1 generalizes bandit algorithms meeting Assumption 3.1 into corresponding BwCRO algorithms. While this makes it a desirable baseline algorithm for BwCRO settings, it cannot request fewer labels on easier problem instances. We thus present adaptive algorithms next.

3.2 Adaptive BwCRO Algorithm for Multi-armed Bandit

Now we present an adaptive algorithm for a simple multi-armed bandit with costly observations.

Simple Multi-armed Bandit Setting In this setting, there is a fixed set of arms $\mathcal{A}$. At each time step $t \leq T$, the algorithm chooses an arm $a_t \in \mathcal{A}$ and receives a reward r_t which is sampled from a distribution that depends on a_t. We use $\mu^{(a)}$ to denote arm a's expected reward $\mathbb{E}[r_t|a_t = a]$.

Algorithm Explanation The basic idea of our algorithm for simple BwCRO settings is based on the idea that if two arms a and a' have close enough expected values $\mu^{(a)}$ and $\mu^{(a')}$, then it is better to suffer the regret of picking one of them rather than paying the cost to learn the difference between them. If $\Delta = |\mu^{(a)} - \mu^{(a')}|$, then it takes $O(1/\Delta^2)$ labels to realize if $\mu^{(a)} > \mu^{(a')}$ or $\mu^{(a)} < \mu^{(a')}$ (as shown in A.4). However the regret from choosing the wrong arm is ΔT, so if $T\Delta \leq cO(1/\Delta^2)$ then it is not worth it to get enough labels to choose between a and a'.

First, define $\hat{\mu}_t^{(a)}$ as the empirical average reward for arm a at time t, $n_t^{(a)}$ as the number of times arm a was observed at time t, and define $u_t^{(a)}$ and $\ell_t^{(a)}$ as the Hoeffding upper/lower bounds from $|\mu^{(a)} - \hat{\mu}_n^{(a)}| \leq \sqrt{\log(kT/\delta)/n_t^{(a)}}$. Define the holdout reward ν_t which with high probability is a lower bound on the reward we get from committing to the holdout arm a_t^ν with average reward μ_t^ν

$$\nu_t = \max_{a \in A} \left(\hat{\mu}_t^{(a)} - \sqrt{\frac{\log(kT/\delta)}{n_t^{(a)}}} \right) = \max_{a \in \mathcal{A}} \ell_t^{(a)},\ a_t^\nu = \arg\max_{a \in \mathcal{A}} \ell_t^{(a)}, \text{ and } \mu_t^\nu = \mu^{(a_t^\nu)}.$$

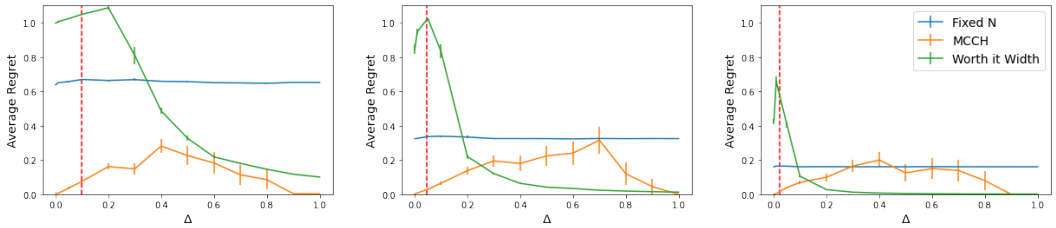
Now, define the gap $g_t^{(a)}$ as $g_t^{(a)} = u_t^{(a)} - \nu_t$, which is an upper bound on the per-step regret for choosing the holdout arm a_t^ν instead of a. Define the maximum gap $\bar{g}_t = \max_{a \in \mathcal{A}} g_t^{(a)}$ and the arm with the maximum gap $a_t^{\bar{g}} = \arg\max_{a \in \mathcal{A}} g_t^{(a)}$. Finally, define the worth it width $w = \sqrt[3]{c\log(kT/\delta)/T}$. Since $\bar{g}_t$ is an upper bound on the per-step regret of choosing the holdout arm a_t^ν, once $\bar{g}_t \leq w$ committing to the holdout arm a_t^ν is better than gathering enough labels to conclude with high probability that some other arm a' has a higher average reward and $\mu^{(a')} \geq \mu_t^\nu$.

Algorithm 2 Worth it Width (WiW) algorithm for simple BwCRO

At each time step t, compute the upper/lower bounds $u_t^{(a)}$ & $\ell_t^{(a)}$, holdout value ν_t, and max gap $\overline{g}_t$.
If $\overline{g}_t \leq w = \sqrt[3]{c\log(kT/\delta)/T}$, then commit to the holdout arm a_t^ν.
Else if $g_t^{(a')} \leq w$ for all $a' \neq a_t^{\bar{g}}$ and the maximum gap $\overline{g}_t^{(a)}$ is such that $a_t^{\bar{g}} = a_t^\nu$, commit to arm a_t^ν.
Otherwise, play and observe the arm $a = \arg\min_{a' \in \{a_t^{\bar{g}}, a_t^\nu\}} n_t^{(a')}$.

Theorem 3 (Regret Rate for Algorithm 2)**.** *Algorithm 2 has a regret rate of $\tilde{O}(kc^{1/3}T^{2/3})$ with high probability.*

Proof. We can play an arm at most $\sqrt[3]{4\log(kT/\delta)}(T/c)^{2/3}$ times before $g_t^{(a)} \leq w$, and since we always play an arm associated with the largest gap we can only gather $k\sqrt[3]{4\log(kT/\delta)}(T/c)^{2/3}$ labels before terminating. Further, with high probability, $g_t^{(a)}$ bounds the regret of committing to the holdout arm instead of arm a since $\mu^{(a)} - \mu^{a_t^\nu} \leq u_t^{(a)} - \ell_t^{(a_t^\nu)} = g_t^{(a)}$. At termination $g_t^{(a)} < \sqrt[3]{c\log(kT/\delta)/T}$, so our regret thereafter is bounded by $\sqrt[3]{c\log(kT/\delta)T^2}$ with high probability. Adding these two terms, the regret is $\tilde{O}(T^{2/3})$. □



(a) $T = 1000$ (b) $T = 10000$ (c) $T = 100000$

Figure 1: Final average per step regret for varying values of gaps Δ, $c = 1$, standard error from 20 trials. Dashed red line is at the predicted worst-case $\Delta = \sqrt[3]{c/T}$.

3.3 Adaptive BwCRO Algorithm for Linear Contextual Bandits

While the previous algorithm is able to adapt to easier problem instances, the basic multi-armed bandit setup does not have a notion of state where the value of an action can change. Now, we present an algorithm which adaptively labels more interesting states, rather than only easier problems.

Linear Contextual Bandit Setting In our linear contextual bandit setting at each time step t the algorithm chooses an action from among k contexts $X_t = \{x_t^{(j)} \in \mathbb{R}^d\}_{j=1}^k$ which are drawn (at each time step) from some distribution $\mathcal{D}$ such that $\|x_t^{(j)}\| \leq B$. The algorithm receives reward $x_t \cdot \mu^* + \eta_t$ for the chosen $x_t \in X_t$, where μ^* is unknown, $\|\mu^*\| \leq W$, and η_t is σ^2 sub-Gaussian noise.

Algorithm Explanation We apply the idea of only requesting labels if the information could be worth it to design a similar algorithm for linear contextual bandits. LinUCB is a well-studied implementation of the "optimism in the face of uncertainty" principle for linear contextual bandits (8; 5; 1). We build on the LinUCB algorithm by using the same uncertainty regions, and deciding whether or not to request labels based on the change in the uncertainty region.

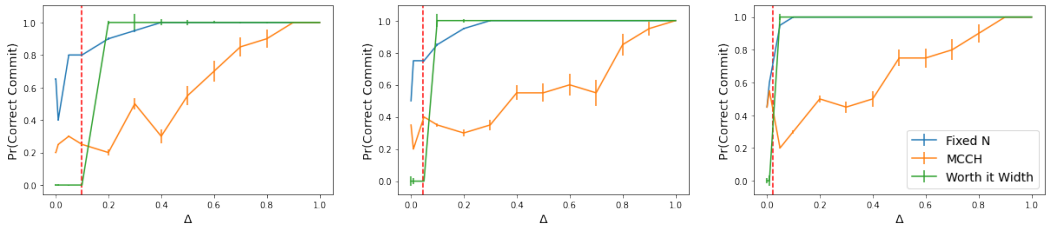
Following (2), define $\Sigma_t = (\sigma^2/W^2)I + \sum_{\tau=1}^{t-1} x_\tau x_\tau^T$, mean $\hat{\mu} = \Sigma_t^{-1} \sum_{\tau=1}^{t-1} r_\tau x_\tau$, and uncertainty region $\text{Ball}_t = \{\mu | \, (\hat{\mu}_t - \mu)^T \Sigma_t^{-1} (\hat{\mu}_t - \mu) \leq \beta_t\}$, where $\beta_t = \sigma^2 \left(2 + 4d \log(1 + tB^2W^2/d) + 8 \log(4/\delta)\right)$. We then upper bound the value of $\mu^* \cdot x$ by adding the width $\text{width}(x) = (\beta_t x^T \Sigma_t^{-1} x)^{1/2}$. The maximum width bounds the per-step regret. Conveniently, the change in the maximum width from observing a reward depends only on the context x. This allows us to compute the reduction in the maximum width that we get from requesting a label based only on the context, which allows us only request a label if the amortized cost of the width reduction is worth it. The max width can be computed as $\max_{\text{eigenvectors}} \text{width}(e_i)$.

Algorithm 3 Delta Max Width Algorithm (DMW)

At each time step t, compute the center $\hat{\mu}_t$, covariance Σ_t, and uncertainty region Ball_t.
Play $x_t = \arg\max_{x \in D} \max_{\mu \in \text{Ball}_t} \mu \cdot x$ as the choice of arm.
If x_t is such that $\left(\text{maxwidth}(\Sigma_t^{-1}) - \text{maxwidth}((\Sigma_t + x_t x_t^T)^{-1})\right)(T - t) > c$, then request label
Otherwise don't request label

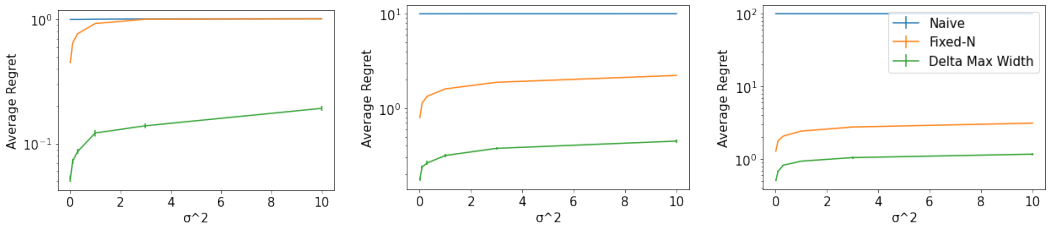
4 Experiments

Comparing simple MAB Algorithms We compare the MAB algorithms in a variety of two armed Bernoulli bandit settings with average reward $0.5 \pm \Delta/2$ and labeling cost $c = 1$. A more detailed study of the cost is in A.1.1. Both the Fixed N and MCCH (7) algorithms make labeling decisions on top of a simple UCB algorithm. As shown in Figure 1, the fixed N algorithm consistently has low variance in its performance and is robust to different environments. Figure 1 also shows that neither the MCCH nor the WiW algorithm dominate the other, with WiW having the advantage for larger Δs and longer episode lengths. This is likely a tradeoff between MCCH committing to an arm earlier and saving on labeling costs, at the expense of being more likely to commit to the wrong arm as shown in Figure 2.



(a) $T = 1000$ (b) $T = 10000$ (c) $T = 100000$

Figure 2: Probability of committing to the higher value arm for varying values of gaps Δ, $c = 1$, standard error from 20 trials. Dashed red line is at the predicted worst-case $\Delta = \sqrt[3]{c/T}$.



(a) $c = 1$ (b) $c = 10$ (c) $c = 100$

Figure 3: Final average per step regret for varying noises σ^2, standard error from 20 trials. Note the logarithmic y scale. The contexts have dimension $d = 5$, drawn from $\mathcal{N}(0, 1)^d$ and rescaled to size 1.

Linear Contextual Bandits We set W and B (the sizes of μ^* and x) to 1, set $T = 10000$, and set the number of arms $k = 5$. Increasing noise forces all non-naive algorithms request more labels, though DMW is able to avoid always requesting labels even with $\eta \sim \sqrt{10} * \mathcal{N}(0, 1)$ with the each r_t constrained within $[-1, 1]$.

While the adaptive algorithms can request fewer labels in easier instances, all of the previous experiments show algorithms which always request labels and then stop. In contextual bandits, it is possible to make context-dependent labeling decisions where the algorithm can choose to get a label for the particular state it is in. As shown in Figure 3, the DMW algorithm is able to substantially save on labeling costs by only requesting labels for informative contexts, demonstrating the value of scalable oversight in linear contextual bandits with costly observations.

5 Related Work

Krueger et. al 2016 (7) studies BwCRO as the active multi-armed bandit problem. Our work differs in that we prove that the BwCRO setting has a regret lower bound of $\Omega(c^{1/3}T^{2/3})$. While this fact was mentioned in (7), they did not provide a proof. This finding is in line with the work in partial monitoring (4) which argues that this problem will have an $\Omega(T^{2/3})$ regret rate because the agent needs to take suboptimal actions in order to confirm that it is following the best strategy. Our finding differs in that it shows that this regret rate holds even with additional known structure that the labeled and unlabeled arm choices differ only by a known cost, and analyzes the impact of the labeling cost to demonstrate the necessity of the $c^{1/3}$ term in the regret rate. Interestingly, this matches the lower bound for non-adaptive exploration in T.

6 Conclusion

In conclusion, we prove $O(c^{1/3}T^{2/3})$ lower bounds on BwCRO, provide an algorithm for generalizing suitable bandit algorithms into BwCRO settings with the Fixed-N algorithm, demonstrate scalable oversight in linear contextual bandits, and compare the performance of discussed algorithms.

Acknowledgements. This research was supported in part by NSF Awards IIS1901168, IIS-2008139, and scholarship funding from Open Philanthropy. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

[1] ABBASI-YADKORI, Y., PÁL, D., AND SZEPESVÁRI, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems* (2011), J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., vol. 24, Curran Associates, Inc.

[2] AGARWAL, A., JIANG, N., AND KAKADE, S. M. Reinforcement learning: Theory and algorithms.

[3] AMODEI, D., OLAH, C., STEINHARDT, J., CHRISTIANO, P., SCHULMAN, J., AND MANÉ, D. Concrete problems in ai safety, 2016.

[4] BARTÓK, G., FOSTER, D. P., PÁL, D., RAKHLIN, A., AND SZEPESVÁRI, C. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research 39*, 4 (2014), 967–997.

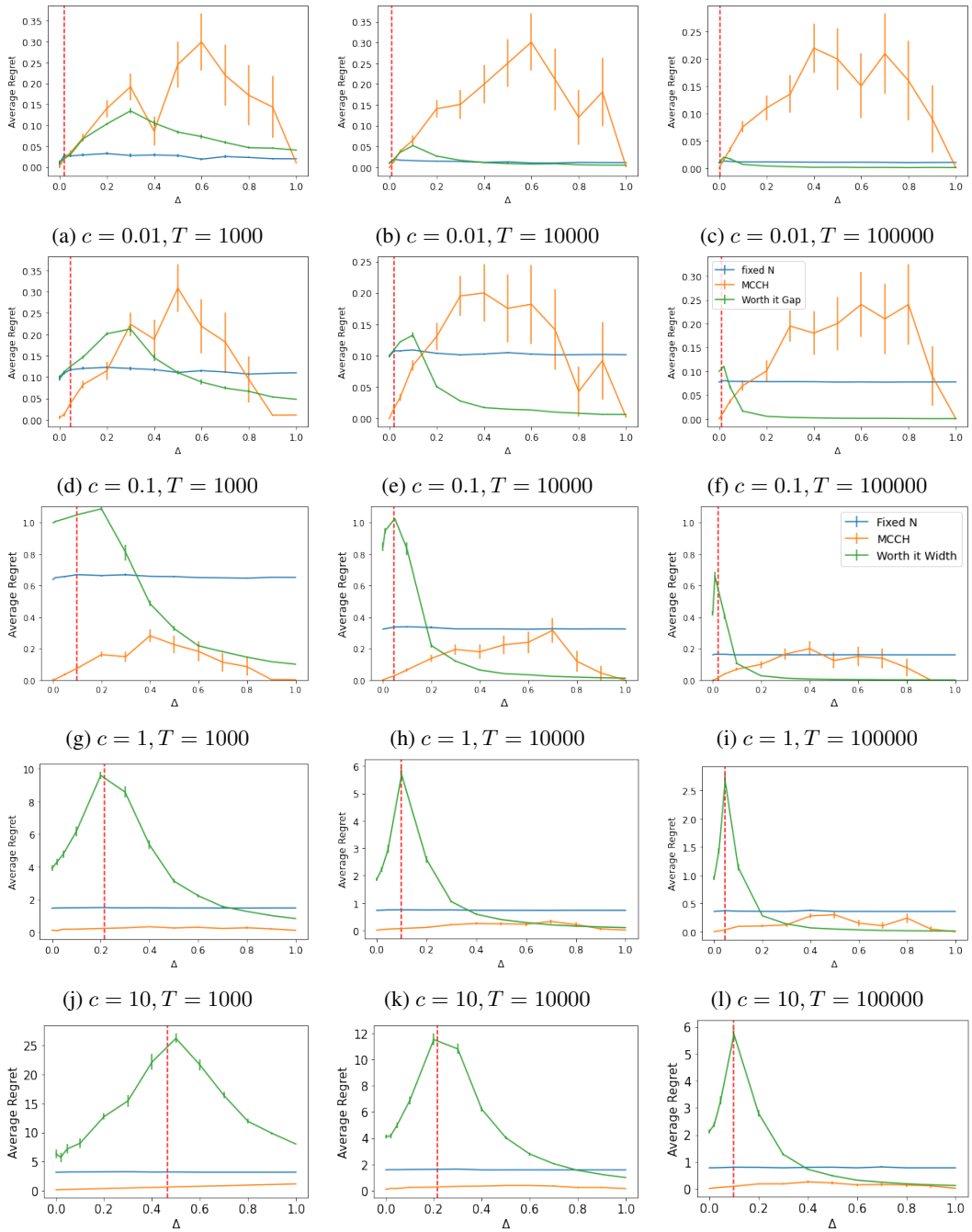
[5] DANI, V., HAYES, T. P., AND KAKADE, S. M. Stochastic linear optimization under bandit feedback. *Conference on Learning Theory* (2008), 355–366.

[6] HENDRYCKS, D., CARLINI, N., SCHULMAN, J., AND STEINHARDT, J. Unsolved problems in ml safety, 2021.

[7] KRUEGER, D., LEIKE, J., EVANS, O., AND SALVATIER, J. Active reinforcement learning: Observing rewards at a cost. *NeurIPS Future of Interactive Learning Machines (FILM) workshop* (2016).

[8] LI, L., CHU, W., LANGFORD, J., AND SCHAPIRE, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web - WWW '10* (2010), ACM Press.

[9] SLIVKINS, A. Introduction to multi-armed bandits, 2019.



(a) $c = 0.01, T = 1000$ (b) $c = 0.01, T = 10000$ (c) $c = 0.01, T = 100000$

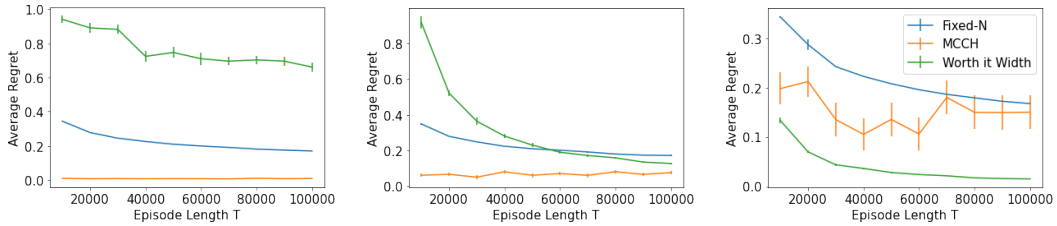
(d) $c = 0.1, T = 1000$ (e) $c = 0.1, T = 10000$ (f) $c = 0.1, T = 100000$

(g) $c = 1, T = 1000$ (h) $c = 1, T = 10000$ (i) $c = 1, T = 100000$

(j) $c = 10, T = 1000$ (k) $c = 10, T = 10000$ (l) $c = 10, T = 100000$

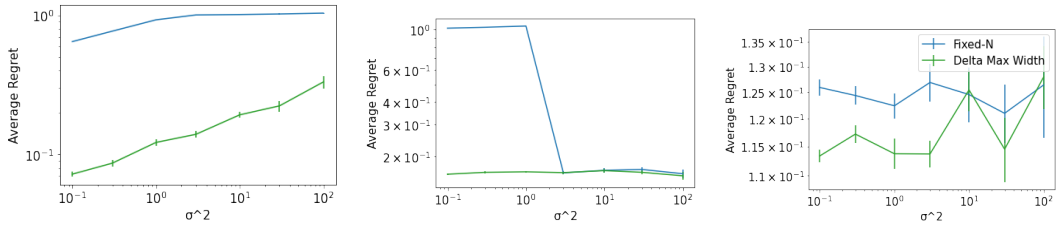
(m) $c = 100, T = 1000$ (n) $c = 100, T = 10000$ (o) $c = 100, T = 100000$

Figure 4: Final average per step regret for varying values of gaps Δ, across many different horizons T and costs c. Standard error from 20 trials. Dashed red line is at the predicted worst-case $\Delta = \sqrt[3]{c/T}$.



(a) $\Delta = 0.01$ (b) $\Delta = 0.1$ (c) $\Delta = 0.3$

Figure 5: Final average per step regret for varying horizons T, $c = 1$, standard error from 20 trials.



(a) $d = 5$ (b) $d = 50$ (c) $d = 100$

Figure 6: Final average per step regret for varying noises σ^2, standard error from 20 trials. Note the logarithmic y scale. $T = 10000$, $k = 5$

A Appendix

A.1 Experimental Appendix

A.1.1 Impact of Labeling Cost

Figure 4 shows that Fixed-N and WiW have an advantage over MCCH in low cost ($c \leq 1$) settings, but that MCCH does better in higher cost settings.

A.1.2 Varying Time

Figure 5 shows that increasing the episode length generally improves the performance of all algorithms, with more dramatic impacts for the Worth it Width algorithm in the regime near the predicted worst-case $\Delta = \sqrt[3]{c/T}$.

A.1.3 Impact of Dimension and Noise in Linear Contextual Setting

Figure 6 shows that DMW maintains its advantage over the Fixed N algorithm over a variety of context dimensions d and noise scales. Note that both have lower regret with higher dimensions, likely because the randomly drawn vectors become more orthogonal with increasing dimension, resulting in smaller differences between the rewards and lower regret.

A.1.4 Hyperparameters

The only hyperparameter for the Fixed N and Worth it Width algorithms is the parameter δ, which is set to 0.5 and not tuned. The MCCH heuristic from (7) also has a single parameter α which is set to 0.1 which appeared to be the best setting in the paper's experiments, though they note that the algorithm appears robust to parameter choice.

A.2 Proof of Theorem 1

Define k bandit problem instances, with each arm being associated with a flip from one of k coins. If the selected coin lands heads then the agent receives reward 1, and otherwise it receives reward 0. Our bandit problem is then drawn with uniform probability from these k settings. We additionally

analyze a base instance 0 in which all coins are unbiased and have reward $1/2$, and in instance j coin j has expected reward $(1+\epsilon)/2$. Denote the probability of an event A in instance j as $\Pr_j(A)$, and the expectation of a random variable X in instance j as $\mathbb{E}_j(X)$.

We will analyze how often an algorithm plays a given arm j^* in the base instance 0, then use the fact that the coins have similar probability distributions to bound the performance in the instance j^* where the coin is preferred. In order to establish the bound, we first need to prove a KL divergence lemma.

Lemma 1 (KL Bound)**.** *For any event A based on n observations of the coin flips, for any $j \in [1..k]$,*

$$|\Pr_0(A) - \Pr_j(A)| \leq \epsilon\sqrt{n}.$$

Proof. First, define p and q to be the probability distributions over n independent $(\epsilon/2)$-biased and fai coin flips respectively, and let p_i be the ith flip from the biased coin and q_i be the ith flip from the fair coin. The KL divergence between a coin flip p_i with bias $\epsilon/2$ and a fair coin flip q_i is as follows:

$$\begin{aligned}
\mathrm{KL}(p_i; q_i) &= \frac{1+\epsilon}{2}\log(1+\epsilon) + \frac{1-\epsilon}{2}\log(1-\epsilon) \\
&= \frac{1}{2}\log\left(1-\epsilon^2\right) + \frac{\epsilon}{2}\log\left(\frac{1+\epsilon}{1-\epsilon}\right) && \pm\frac{1}{2}\log(1+\epsilon) \\
&\leq \frac{\epsilon}{2}\log\left(\frac{1+\epsilon}{1-\epsilon}\right) && \text{since } \frac{1}{2}\log\left(1-\epsilon^2\right) < 0 \\
&= \frac{\epsilon}{2}\log\left(1+\frac{2\epsilon}{1-\epsilon}\right) \\
&\leq \frac{\epsilon}{2}\frac{2\epsilon}{1-\epsilon} && \text{since } \log(1+x) \leq x \text{ for } x > 0 \\
&\leq 2\epsilon^2 && \text{since } 0 \leq \epsilon \leq 1/2
\end{aligned}$$

$$\begin{aligned}
|\Pr_0(A) - \Pr_j(A)| &\leq \sqrt{\frac{1}{2}\mathrm{KL}(p;q)} && \text{by Pinsker's inequality} \\
&\leq \sqrt{\frac{1}{2}\sum_{i=1}\mathrm{KL}(p_i;q_i)} && \text{by KL divergence chain rule for independent draws} \\
&\leq \sqrt{\frac{1}{2}(2n\epsilon^2)} && \text{since } \mathrm{KL}(p_i;q_i) \leq 2\epsilon^2 \\
&\leq \epsilon\sqrt{n}
\end{aligned}$$

□

Theorem 1. *The Bandits with Costly Observations setting has a regret lower bound of $\Omega(c^{1/3}T^{2/3})$.*

Proof. The basic idea of the proof is that for every instance $j^* \neq 0$, we can upper bound how many times we play the optimal arm j^* by looking at how many times we play j^* in instance 0, then using a KL divergence lemma to upper bound the probability of playing coin j^* in instance j^* in terms of the number of observations n. This will establish that we cannot frequently play the coin j^* in the appropriate instance j^* without also playing it in the incorrect instances $j' \neq j^*$, leading to regret.

How many times do we play j^* in instance 0? Let $Q_j^{(t)}$ be the number of times that the algorithm flips coin j by time t. Note that by linearity of expectation

$$\sum_{j=1}^{k}\mathbb{E}_0\left[Q_j^{(t)}\right] = \mathbb{E}_0\left[\sum_{j=1}^{k}Q_j^{(t)}\right] = \mathbb{E}_0[t] = t.$$

Let $J_t = \{j : \mathbb{E}_0[Q_t^{(j)}] \leq 3t/k\}$ be the set of coins that the algorithm has not played more than $3/k$ of the time over the first t timesteps in instance 0. As previously shown $\sum_{j=1}^{k} \mathbb{E}_0\left[Q_t^{(j)}\right] = t$, so J_t must have at least $2k/3$ elements since

$$t = \sum_{j=1}^{k} \mathbb{E}_0\left[Q_t^{(j)}\right] \geq \sum_{j \notin J_t} \mathbb{E}_0\left[Q_t^{(j)}\right] \geq \sum_{j \notin J_t} \frac{3t}{k} \geq |\{j : j \notin J_t\}| \frac{3t}{k} \text{ implies } |\{j : j \notin J_t\}| \leq k/3.$$

By the Markov Inequality $\mathbb{E}_0[Q_t^{(j)}] \leq 3t/k$ implies that for any coin $j \in J_t$ and any a

$$\Pr_0\left(Q_j^{(t)} \geq a\right) \leq \frac{\mathbb{E}_0[Q_t^{(j)}]}{a} \leq \frac{3t/k}{a}, \text{ and therefore } \Pr_0\left(Q_t^{(j)} < a\right) > 1 - \frac{3t}{ka}.$$

Now, we compute the probability that j^* is played less than a times in instance 0. Let $\mathcal{E}_{j^*}$ be the event that a given $j^* \in J_T$ and that $Q_t^{(j^*)} < a$.

$$\begin{aligned}
\Pr_0(\mathcal{E}_j^*) &= \Pr_{\text{inst}}(j^* \in J_T) \Pr_0\left(Q_t^{(j^*)} < a | j \in J_T\right) && \text{(Randomness in } \Pr_{\text{inst}} \text{ is over instances)} \\
&= \frac{2}{3} \Pr_0\left(Q_t^{(j^*)} < a | j \in J_T\right) && \text{since } |J_T| > 2k/3 \\
&> \frac{2}{3}\left(1 - \frac{3T}{ka}\right) && \text{Markov inequality with } \mathbb{E}_0\left[Q_T^{(j^*)}\right] \leq \frac{3T}{k} \\
&= \frac{2}{3} - \frac{2T}{ka}
\end{aligned}$$

As a sanity check, note that increasing the number of arms raises the lower bound and makes $\mathcal{E}_j$ more likely, as does increasing the threshold a. Increasing T on the other hand makes it less likely.

Expected regret in instance j^*? Assume that the algorithm observed n rewards for arm j^* over the entire history. We know from Lemma 1 that for any event A based on n labels $|\Pr_0(A) - \Pr_j(A)| \leq \epsilon\sqrt{n}$, which lower bounds the probability $\Pr_{j^*}(\mathcal{E}_{j^*})$ of playing j^* less than a times as

$$\Pr_{j^*}(\mathcal{E}_{j^*}) > \frac{2}{3} - \frac{2T}{ka} - \epsilon\sqrt{n}.$$

If j^* is the best arm with bias $(1+\epsilon)/2$ and all other coins are fair, then the regret in instance j^* if event $\mathcal{E}_{j^*}$ holds is simply the difference of the two rewards, plus the cost of acquiring n labels.

$$\begin{aligned}
\mathbb{E}_{j^*}[\text{Regret}_T] &= \Pr_{j^*}(\overline{\mathcal{E}_{j^*}}) \mathbb{E}_{j^*}\left[\text{Regret}_T | \overline{\mathcal{E}_{j^*}}\right] + \Pr_{j^*}(\mathcal{E}_{j^*}) \mathbb{E}_{j^*}[\text{Regret}_T | \mathcal{E}_{j^*}] + cn && (1) \\
&\geq \Pr_{j^*}(\mathcal{E}_{j^*}) \mathbb{E}_{j^*}[\text{Regret}_T | \mathcal{E}_{j^*}] + cn && (2) \\
&= \Pr_{j^*}(\mathcal{E}_{j^*}) \left(T\frac{1+\epsilon}{2} - T\frac{1 + Q_{j^*}^{(T)}\epsilon}{2}\right) + cn && (3) \\
&\geq \Pr_{j^*}(\mathcal{E}_{j^*}) \left(T\frac{1+\epsilon}{2} - T\frac{1 + a\epsilon}{2}\right) + cn && (4) \\
&= \Pr_{j^*}(\mathcal{E}_{j^*}) \frac{(T-a)\epsilon}{2} + cn && (5) \\
&> \left(\frac{2}{3} - \frac{2T}{ka} - \epsilon\sqrt{n}\right) \frac{(T-a)\epsilon}{2} + cn && (6)
\end{aligned}$$

Inequality 2 holds because $\Pr_{j^*}(\overline{\mathcal{E}_{j^*}}) \mathbb{E}_{j^*}\left[\text{Regret}_T | \overline{\mathcal{E}_{j^*}}\right]$ is positive, equation 3 holds by the definition of regret, inequality 4 holds since $\mathcal{E}_{j^*}$ is true and so $Q_{j^*}^{(T)} < a$ and $-a < -Q_{j^*}^{(T)}$, and inequality 6 holds from the KL divergence lemmas.

Conclusion Now we can conclude the proof. Recall that a is from the Markov inequality, and so we are free to choose $a = 6T/k$, yielding the bound

$$\begin{aligned}
\mathbb{E}_{j^*}\left[\text{Regret}_T\right] &\geq \left(\frac{2}{3} - \frac{2Tk}{k6T} - \epsilon\sqrt{n}\right)\frac{(T-6T/k)\epsilon}{2} + cn \\
&= \left(\frac{1}{3} - \epsilon\sqrt{n}\right)\frac{(k-6)T\epsilon}{2k} + cn \\
&= \frac{(k-6)T\epsilon}{6k} - \frac{(k-6)T\epsilon^2}{2k}\sqrt{n} + cn.
\end{aligned}$$

Now, choose $\epsilon = \sqrt[3]{c/T}$ for the coin expected rewards, for a regret bound of

$$\mathbb{E}_{j^*}\left[\text{Regret}_T\right] \geq \frac{(k-6)}{6k}\sqrt[3]{cT^2} - \frac{(k-6)}{2k}\sqrt[3]{c^2T}\sqrt{n} + cn.$$

Now, imagine that the algorithm did as well as possible, and minimized this value with respect to n. This yields $\sqrt{n} = \frac{(k-6)}{4k}\sqrt[3]{T/c}$, and a regret of

$$\mathbb{E}_{j^*}\left[\text{Regret}_T\right] \geq \frac{(k-6)}{6k}\sqrt[3]{cT^2} - \frac{(k-6)^2}{16k^2}\sqrt[3]{cT^2},$$

for an $\Omega(c^{1/3}T^{2/3})$ regret lower bound, as desired.

□

A.3 Proof of Theorem 2

With the uniform regret assumption, the $O(c^{1/3}T^{2/3})$ regret rate for the Fixed N algorithm is the result of fairly straightforward algebraic manipulations.

Assumption 3.1 (Uniform Regret Rate)**.** For all $n \leq T$, playing algorithm $\mathcal{A}$ while observing labels at every timestep results in a regret of $\mathbb{E}\left[\text{Regret}^{\circ}_{1:n}\right] \in O(n^{1/2})$, and the average regret of playing according to $\mathcal{A}$ without further labels is s.t. $\frac{1}{T-n}\mathbb{E}\left[\text{Regret}^{\circ}_{n+1:T}\right] \leq \frac{1}{n}\mathbb{E}\left[\text{Regret}^{\circ}_{1:n}\right]$.

Proof. Assume that $\mathcal{A}$ meets the uniform regret assumption, so that

$$\frac{1}{T-n}\mathbb{E}\left[\text{Regret}_{n+1:T}\right] \leq \frac{1}{n}\mathbb{E}\left[\text{Regret}_{1:n}\right].$$

Then, by the definition of $O(n^{1/2})$ regret there is a constant K and n_0 such that for all $n > n_0$

$$\mathbb{E}\left[\text{Regret}_{1:n}\right] \leq K\sqrt{n} \text{ and therefore } \frac{1}{T-n}\mathbb{E}\left[\text{Regret}_{n+1:T}\right] \leq \frac{1}{n}\mathbb{E}\left[\text{Regret}_{1:n}\right] \leq \frac{K}{\sqrt{n}}.$$

In the BwCO setting, receiving n labels necessarily incurs a regret of cn, so the total regret of using $\mathcal{A}$ while labeling the first n observations is simply

$$\begin{aligned}
\text{Regret}_{1:T} &= cn + \mathbb{E}\left[\text{Regret}_{1:n}\right] + (T-n)\frac{1}{T-n}\mathbb{E}\left[\text{Regret}_{n+1:T}\right] \\
&\leq cn + \mathbb{E}\left[\text{Regret}_{1:n}\right] + (T-n)\frac{1}{n}\mathbb{E}\left[\text{Regret}_{1:n}\right] \\
&\leq cn + K\sqrt{n} + (T-n)\frac{K}{\sqrt{n}} \\
&= cn + n\frac{K}{\sqrt{n}} + (T-n)\frac{K}{\sqrt{n}} \\
&= cn + TKn^{-1/2}
\end{aligned}$$

We can now simply minimize this expression with respect to the number of labels n...

$$\frac{d}{dn}\text{Regret}_{1:T} \leq \frac{d}{dn}\left(cn + TKn^{-1/2}\right) = c - \frac{TK}{2}n^{-3/2}$$

Solving for $c - TKn^{-3/2}/2 = 0$, we have

$$n = \left(\frac{TK}{2c}\right)^{2/3}$$

Since the second derivative $3TKn^{-5/2}/4$ is always positive, this is a global minima.

Plugging it back into the original expression, we have the desired regret rate

$$\begin{aligned}
\text{Regret}_{1:T} &\leq cn + TKn^{-1/2} \\
&= c\left(\frac{TK}{2c}\right)^{2/3} + TK\left(\left(\frac{TK}{2c}\right)^{2/3}\right)^{-1/2} \\
&= c\left(\frac{TK}{2c}\right)^{2/3} + TK\left(\frac{TK}{2c}\right)^{-1/3} \\
&= c^{1/3}\left(\frac{TK}{2}\right)^{2/3} + (TK)^{2/3}(2c)^{1/3} \\
&\in O\left(c^{1/3}K^{2/3}T^{2/3}\right).
\end{aligned}$$

Note that as $c \to 0$, $n \to \infty$ which makes sense since if the labels are free and always improve performance then the algorithm should always get the label. In this case, note that n must be less than or equal to T, and therefore we recover the original regret expression.

$$\text{Regret}_{1:T} \leq cn + TKn^{-1/2} = 0n + TKn^{-1/2} = T^{-1/2} = K\sqrt{T}$$

□

A.4 Single-armed Case

Algorithm 2 makes more sense after first considering the simplified problem of how many samples to collect from a costly one-armed bandit, where one arm has a stochastic reward with unknown mean μ^* with rewards bounded in $[0, 1]$, and one arm with a known average reward ν. Define the regret of choosing the wrong arm as $\Delta = \|\mu^* - \nu\|$.

Definition A.1 (Disambiguate)**.** Two arms a and a' with means μ_a^* and $\mu_{a'}^*$ are disambiguated if with probability at least $1 - \delta$ it can be said that either $\mu_a^* > \mu_{a'}^*$ or $\mu_a^* < \mu_{a'}^*$.

Remark. *The stochastic and fixed arms are disambiguated after n stochastic arm reward samples if*

$$n > \frac{\log(T/\delta)}{(\nu - \hat{\mu}_n^*)^2} = \frac{\log(T/\delta)}{\hat{\Delta}_n^2}$$

Proof. Hoeffding's inequality bounds the true average reward μ^* based on the observed rewards $\hat{\mu}_n$.

$$\ell_n = \hat{\mu}_n - \sqrt{\frac{\log(T/\delta)}{n}} \leq \mu^* \leq \hat{\mu}_n + \sqrt{\frac{\log(T/\delta)}{n}} = u_n.$$

If $\mu^* \leq \nu$, then $\nu \leq u_n = \hat{\mu}_n + \sqrt{\log(T/\delta)/n}$ can only hold while $n \leq 4\log(T/\delta)/\hat{\Delta}_n^2$. Similarly, if $\nu \leq \mu^*$, then $\hat{\mu}_n - \sqrt{\log(T/\delta)/n} \leq \ell_n \leq \nu$ can only hold while $n \leq 4\log(T/\delta)/\hat{\Delta}_n^2$. Additionally, since $\mu^* \leq u_n$, $u_n \leq \nu$ implies $\mu^* \leq \nu$, and similarly $\nu \leq \ell_n$ implies $\nu \neq \mu^*$ Therefore, with high probability the two arms will be disambiguated once $n > \log(T/\delta)/\Delta_n^2$. Applying Hoeffding's inequality to bound $\hat{\mu}_n$ demonstrates that this will happen within at most $4\log(T/\delta)/\Delta^2$ steps. □

However, the smaller the gap Δ between the fixed and stochastic arms, the less regret is accumulated by choosing the wrong arm. If Δ is small enough then it is preferable to pay the regret of choosing the wrong arm rather than pay the labeling cost needed to disambiguate between the two arms.

Remark. *It is not worth it to disambiguate between the stochastic and fixed arms if*

$$\hat{\Delta}_n + \sqrt{\log(T/\delta)/T} < \sqrt[3]{\frac{c\log(T/\delta)}{T}}$$

Proof. Requesting n labels yields a regret of at least cn, where c is the labeling cost. Simply choosing an arm at the beginning yields a regret of at most $T\Delta$, which is bounded above by $T(\hat{\Delta}_n + \sqrt{\log(T/\delta)/T})$. With high probability any future $\hat{\Delta}_{n'>n} < \hat{\Delta}_n + \sqrt{\log(T/\delta)/T}$, so as per the previous remark it will cost at least $c\log(T/\delta)/(\Delta_n + \sqrt{\log(T/\delta)/T})^2$ to disambiguate the arms. If the biggest possible regret of committing incorrectly is less than the smallest possible cost of labeling, it is not worth it to disambiguate the arms, so therefore, it is not worth it disambiguate between the two arms if $\hat{\Delta}_n + \sqrt{\log(T/\delta)/T} < \sqrt[3]{c\log(T/\delta)/T}$. □

A.5 Proof of Theorem 4

Theorem 4 (Regret Rate for Algorithm 2)**.** *Algorithm 2 has a regret rate of $\tilde{O}(kc^{1/3}T^{2/3})$ with high probability.*

Proof. The proof has two main claims – that we will hit a termination condition within $\tilde{O}(k(T/c)^{2/3})$ labels, and that upon doing so the regret will be bounded by $\tilde{O}(kT^{2/3})$.

Termination We show that the algorithm terminates after $\tilde{O}(T^{2/3})$ labels by showing that the number of labels necessary for the algorithm to terminate can be bounded by the number of labels necessary for $u_t^{(a)} - \ell_t^{(a)} < w$ to hold for all arms.

First, note that since $g_t^{(a)} = u_t^{(a)} - \nu_t$ and $\nu_t = \max \ell_t^{(a)}$, an arm's gap $g_t^{(a)}$ is bounded above by $u_t^{(a)} - \ell_t^{(a)}$.

$$g_t^{(a)} = u_t^{(a)} - \nu_t = u_t^{(a)} - \max_{a\in\mathcal{A}} \ell_t^{(a)} \leq u_t^{(a)} - \ell_t^{(a)}$$

Therefore, $u_t^{(a)} - \ell_t^{(a)} \leq w$ implies that $g_t^{(a)} \leq w$. Similarly, if $u_t^{(a)} - \ell_t^{(a)} \leq w$ for all arms $a \in \mathcal{A}$ then $g_t^{(a)} \leq w$ for all arms $a \in \mathcal{A}$ and the first termination condition holds.

Now, we solve for how many reward observations for an arm a are necessary for $g_t^{(a)} \leq u_t^{(a)} - \ell_t^{(a)} \leq w$.

$$u_t^{(a)} - \ell_t^{(a)} = \mu_t^{(a)} + \sqrt{\frac{\log(kT/\delta)}{n_t^{(a)}}} - \left(\mu_t^{(a)} - \sqrt{\frac{\log(kT/\delta)}{n_t^{(a)}}}\right) = 2\sqrt{\frac{\log(kT/\delta)}{n_t^{(a)}}} = \sqrt{\frac{4\log(kT/\delta)}{n_t^{(a)}}}$$

$$u_t^{(a)} - \ell_t^{(a)} = \sqrt{\frac{4\log(kT/\delta)}{n_t^{(a)}}} \leq \sqrt[3]{\frac{4c\log(kT/\delta)}{T}} = w$$

$$\sqrt[3]{4\log(kT/\delta)}(T/c)^{2/3} \leq n_t^{(a)}$$

Therefore, an arm a needs to be played at most $\sqrt[3]{4\log(kT/\delta)}(T/c)^{2/3}$ times in order for $g_t^{(a)} \leq w$ to hold.

Second, note that since the arm always plays the least played arm associated with the maximum gap, it takes at most $2\sqrt[3]{4\log(kT/\delta)}(T/c)^{2/3}$ labels for a gap for both of the associated arms to have $u_t^{(a)} - \ell_t^{(a)} \leq w$ hold, and therefore for $g_t^{(a)} \leq w$ to hold. Further, since the algorithm always plays an arm associated with the maximum gap, it will be decreasing all of the k gaps

until it terminates. Therefore, the algorithm will reach the first termination condition after at most $2k\sqrt[3]{4\log(kT/\delta)}(T/c)^{2/3}$ labels. Note that the second termination condition may be reached sooner than this if all but the holdout arm have $g_t^{(a)} \leq w$.

Therefore in conclusion, the algorithm will commit to an arm after at most $2k\sqrt[3]{4\log(kT/\delta)}(T/c)^{2/3}$ labels. We can upper bound the regret incurred during this phase by $(1+c)$ times the length of the labeling phase to represent paying regret for the largest possible reward difference between the arms as well as the labeling cost c, totaling in a regret of at most

$$2(1+c)k\sqrt[3]{4\log(kT/\delta)}(T/c)^{2/3}.$$

Regret There are two regret cases to cover, one for if the first termination is reached, and another for if the second termination condition is reached.

In the first case, we commit to playing the arm a_t^ν associated with ν_t after $g_t^{(a)} \leq w$ for all arms. Since $g_t^{(a)} = u_t^{(a)} - \nu_t = u_t^{(a)} - \nu_t = u_t^{(a)} - \nu_t$ and since with high probability for all arms $a \in \mathcal{A}$, $\ell_t^{(a)} \leq \mu^{*(a)} \leq u_t^{(a)}$, it follows that $g_t^{(a)}$ is an upper bound on the per-turn regret of choosing a_t^ν instead of a.

$$\begin{aligned} g_t^{(a)} &\geq \mu^a - \nu_t && \text{Hoeffding bound} \\ &= \mu^a - \ell_t^{(a_t^\nu)} && \text{Definition of } \nu_t \\ &\geq \mu^a - \mu^{(a_t^\nu)} && \text{Hoeffding bound.} \end{aligned}$$

Since $g_t^{(a)} \leq w$ for all arms, it then follows that the per-turn regret of committing to a^ν is at most $w = \sqrt[3]{\frac{c\log(kT/\delta)}{T}}$. The regret after committing can be bounded by T times the maximum possible per-turn regret, yielding a regret of at most

$$T\sqrt[3]{\frac{c\log(kT/\delta)}{T}} = \sqrt[3]{c\log(kT/\delta)}T^{2/3}.$$

In the second case, the arm a with the maximum gap $g_t^{(a)}$ is the holdout arm, while every other a' is such that $g_t^{(a')} \leq w$. In this case, w still bounds the per-turn regret of choosing a instead of some other a', and has the same regret bound.

Conclusion Adding together the two regret terms, we have $2(1+c)k\sqrt[3]{4\log(kT/\delta)}(T/c)^{2/3} + \sqrt[3]{c\log(kT/\delta)}T^{2/3}$, for a total $\tilde{O}(c^{1/3}T^{2/3})$ regret of

$$k\sqrt[3]{c\log(kT/\delta)}(T/c)^{2/3} + (1+2k)\sqrt[3]{4c\log(kT/\delta)}T^{2/3} \in \tilde{O}(c^{1/3}T^{2/3}).$$

□