Bandits with Costly Reward Observations Supplementary Material

Aaron D. Tucker **Caleb Biddulph*** **Claire Wang*** **Thorsten Joachims**

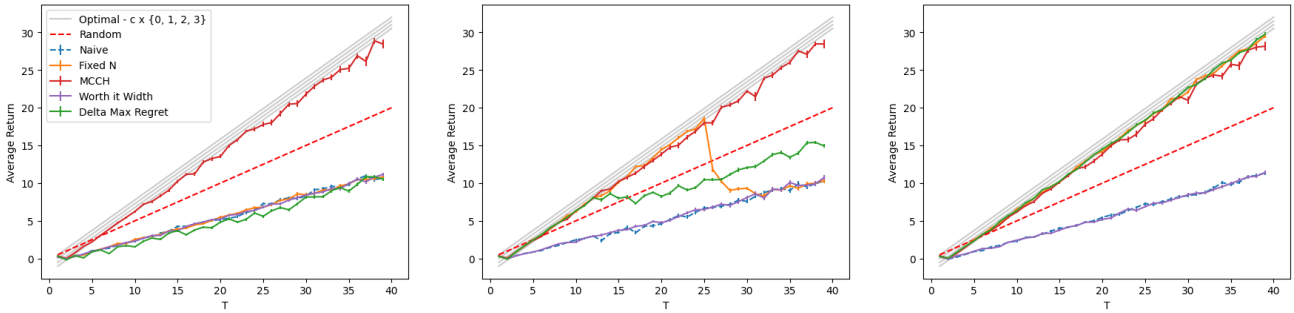
Department of Computer Science, Cornell University, Ithaca NY USA
*Equal contribution, authors listed alphabetically

A APPENDIX

A.1 EXPERIMENTAL APPENDIX

A.1.1 Comparison to BAMCP++

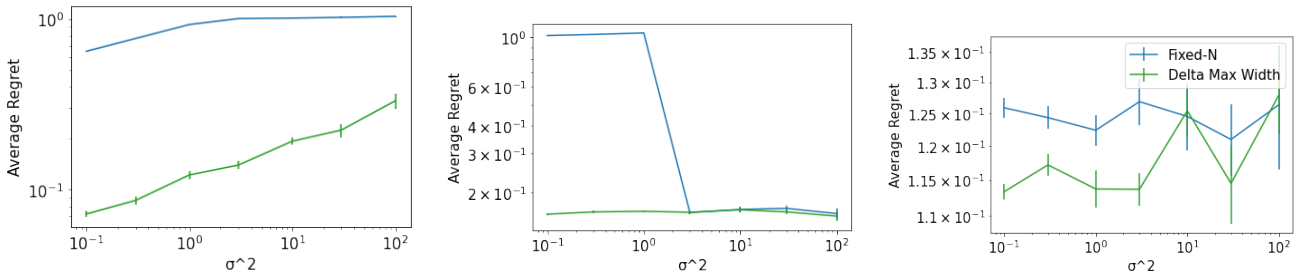
Schulze and Evans [2018] presents the BAMCP++ algorithm for the Active RL setting, which is built on top of Bayesian Monte-Carlo Tree Search, and is applicable to MDP settings as well as bandits. However, it is much more computationally expensive than the algorithms discussed throughout this paper, and the original publication only evaluated its performance on bandits up to 40 timesteps. In Figure 6 we show experiments which are directly comparable to the experiment presented in Figure 3 of Schulze and Evans [2018]. We find that the MCCH heuristic Krueger et al. [2016] is able to achieve higher performance than BAMCP++, since it also stays close to the line corresponding to requesting 3 labels then performing optimally, however it also does so in earlier horizons rather than performing at chance until roughly $T = 15$. All other algorithms presented perform below chance with their typical hyperparameter settings. While the δ hyperparameter for the UCB algorithms represents a bound on the probability of the Azuma-Hoeffding bounds failing [Agarwal et al., 2023], treating it as a freely-chosen hyperparameter and setting δ to higher values causes the DMR and Fixed-N algorithms to perform comparably to or better than MCCH and BAMCP++.



(a) $\delta = 0.5$ (b) $\delta = 50$ (c) $\delta = 250$

Figure 6: Replication of Figure 3 of Schulze and Evans [2018] with varying settings for the hyperparameter δ. As in Schulze and Evans [2018], mean and standard error are presented over 100 trials.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).



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

Figure 7: 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.1.2 Impact of Dimension on Linear Contextual Bandit Results

Figure 7 shows that DMR 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.1.3 Hyperparameters

The only hyperparameter for the Fixed N and Worth it Width algorithms is the parameter δ, which is set to 0.5. The MCCH heuristic from Krueger et al. [2016] 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.1.4 Impact of Labeling Cost

Figure 8 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. Increasing the episode length generally improves the performance of all algorithms, with more dramatic impacts for the WiW algorithm in the regime near the predicted worst-case $\Delta = \sqrt[3]{c/T}$.

A.1.5 Worth-it-Width Ablations

Figure 9 repeats the Worth-it-Width ablation experiments across a variety of parameter settings, demonstrating that all steps of the Worth-it-Width algorithm are necessary for best performance.

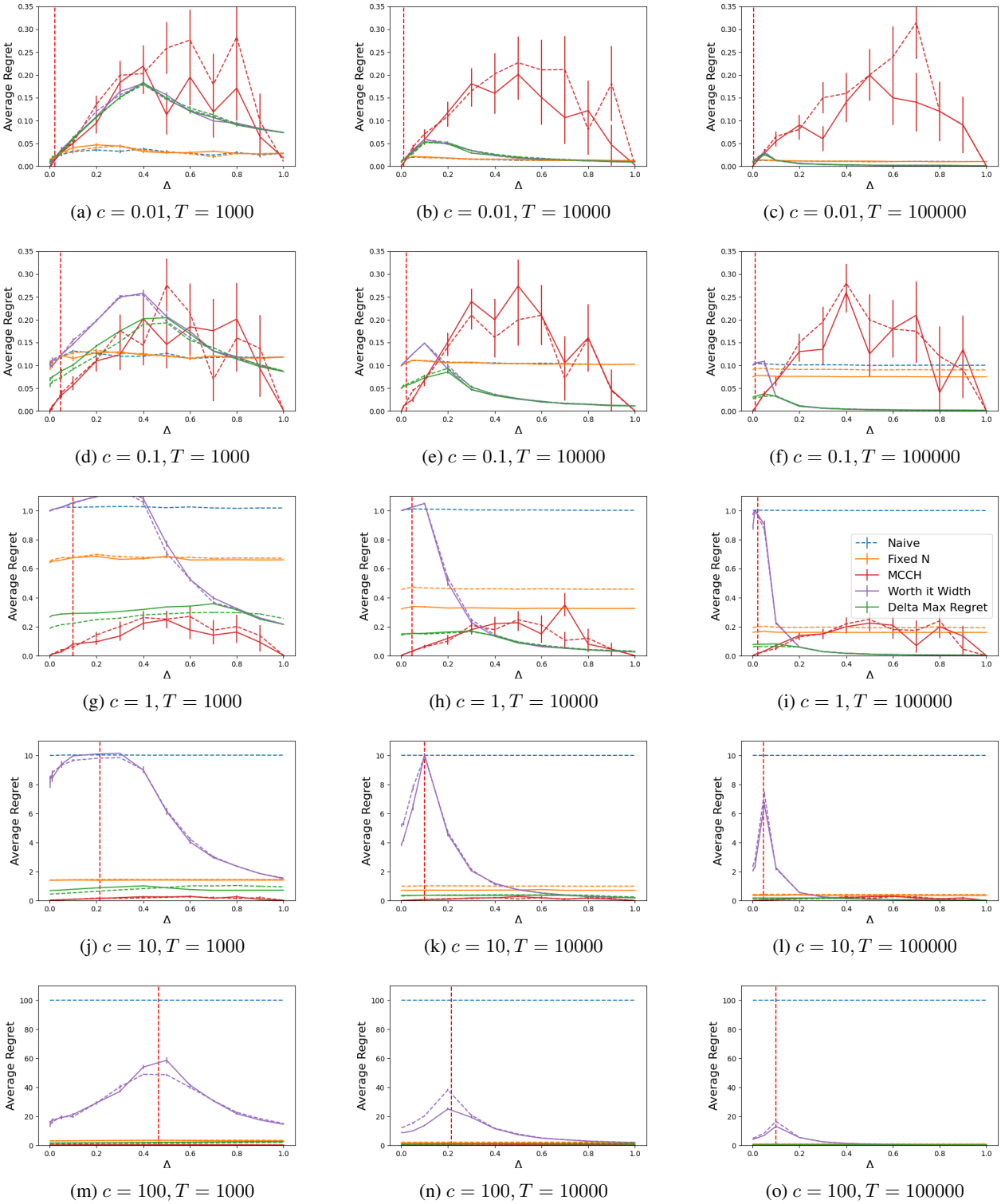
A.2 PROOF OF THEOREM 3

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. This proof and lemma are again based on Slivkins [2019], and adapted to the BwCRO setting.

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

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



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

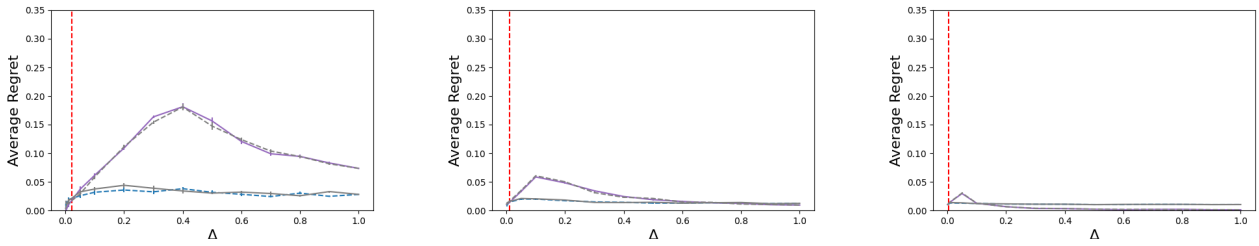
(d) $c = 0.1, T = 1000$ (e) $c = 0.1, T = 10000$ (f) $c = 0.1, T = 100000$

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

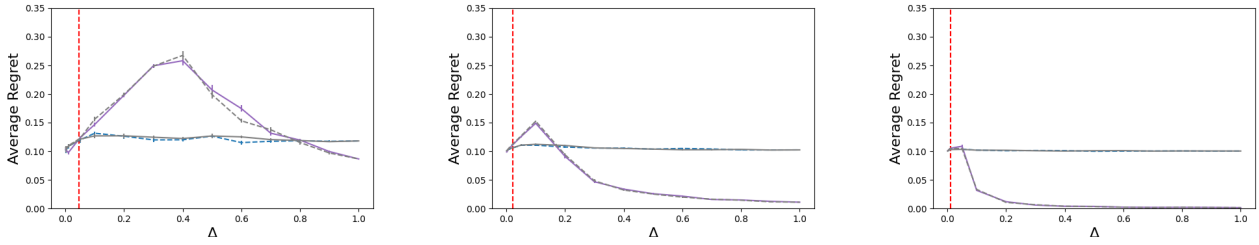
(j) $c = 10, T = 1000$ (k) $c = 10, T = 10000$ (l) $c = 10, T = 100000$

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

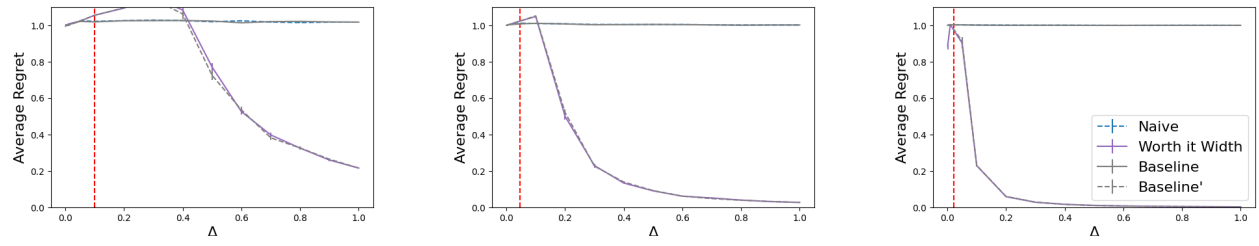
Figure 8: Final average per step regret for varying values of gaps Δ, across many different horizons T and costs c. Standard error from 20 trials. Dashed vertical red line is at the predicted worst-case $\Delta = \sqrt[3]{c/T}$. Other dashed lines correspond to using the doubling trick.



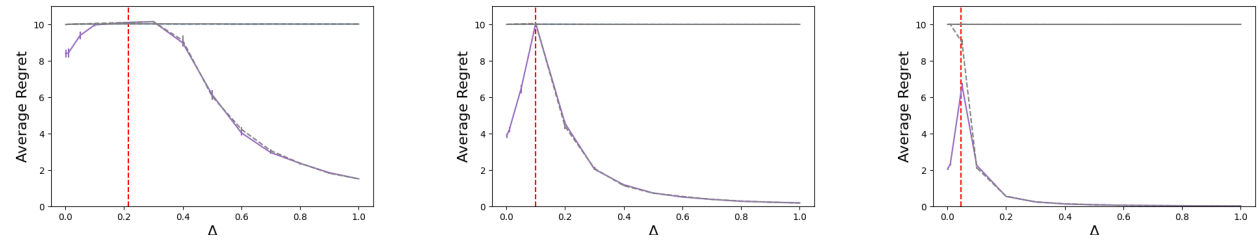
(a) $c = 0.01, T = 1000$ (b) $c = 0.01, T = 10000$ (c) $c = 0.01, T = 100000$



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



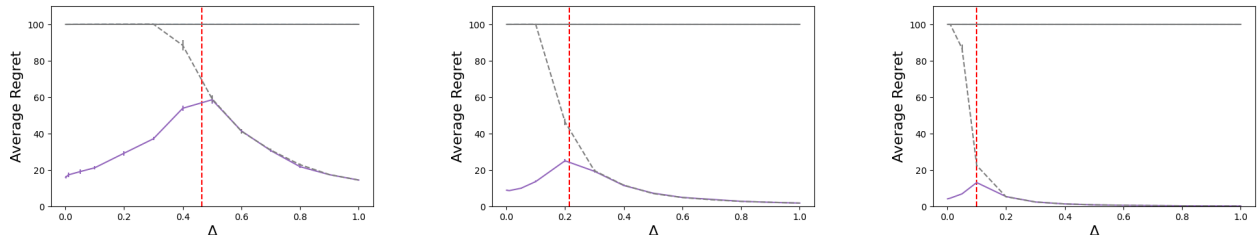
(g) $c = 1, T = 1000$ (h) $c = 1, T = 10000$ (i) $c = 1, T = 100000$



(j) $c = 10, T = 1000$. Note that Baseline' tracks performance of Baseline (rather than WiW) for $\Delta \leq 0.3$.

(k) $c = 10, T = 10000$. Note that Baseline' tracks performance of Baseline (rather than WiW) for $\Delta \leq 0.1$.

(l) $c = 10, T = 100000$



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

Figure 9: Simpler baseline comparisons. 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 $\Delta = \sqrt[3]{c/T}$. Note the similar performance of the Baseline algorithm to Worth-it-Width for $c < 10$, and worse performance in the small arm difference Δ regime whenever $c \geq 10$.

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 3 *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.

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

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

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

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). *An algorithm $\mathcal{A}$ meets the uniform regret assumption if, for all $n \leq T$ and with randomness taken over the algorithm's choices and environment, a) playing according to $\mathcal{A}$ while observing labels for the first n timesteps results in $\mathbb{E}\left[\textit{Regret}^{\circ}_{1:n}\right] \in O(n^{1/2})$ and b) with randomness taken over the algorithm's choices and environment, and if requesting no further labels after the first n timesteps results in*

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

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...

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

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.

Since the regret $\text{Regret}_{1:T}$ is bounded by $cn + TKn^{-1/2}$, and since $cn + TKn^{-1/2}$ is minimized by setting $n = \left(\frac{TK}{2c}\right)^{2/3}$, we can minimize the upper bound on regret by requesting $n = \left(\frac{TK}{2c}\right)^{2/3}$ labels.

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.

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

□

A.4 PROOF OF THEOREM 1

Theorem 1 (Regret Rate for WiW Algorithm). *Algorithm 1 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

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

□

References

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. *Reinforcement learning: Theory and algorithms*. 2023.

David Krueger, Jan Leike, Owain Evans, and John Salvatier. Active reinforcement learning: Observing rewards at a cost. *NeurIPS Future of Interactive Learning Machines (FILM) workshop*, 2016.

Sebastian Schulze and Owain Evans. Active reinforcement learning with monte-carlo tree search. *CoRR*, abs/1803.04926, 2018. URL http://arxiv.org/abs/1803.04926.

Aleksandrs Slivkins. Introduction to multi-armed bandits, 2019. URL https://arxiv.org/abs/1904.07272.