A Proof of Theorem 1: Analysis of the diversity-preserving UCB policy

The $\ln T$ bound of Theorem 1 is proved in Appendix A.1 by showing, in view of (2), that suboptimal distributions $\underline{p} \in \mathrm{Ext}(\mathcal{P})$ are unlikely to be played more than $\ln T$ times. The analysis mimics and adapts the proof scheme corresponding to UCB run on $\mathrm{Ext}(\mathcal{P})$, with three new ingredients specifically underlined.

The proof of the constant regret bound of Theorem 1 may be found in Appendix A.2 and follows a completely different logic. We first show that optimal distributions are typically played at least half of the time. This entails, because $p^\star_{\min}(\underline{\nu}) > 0$, that each pure action $a \in [K]$ is played linearly many times. Therefore, all estimates are sharp, and little regret is suffered.

A.1 Proof of the $\ln T$ bound in Theorem 1

We want to control the $\mathbb{E}\big[N_{\underline{p}}(t)\big]$ by $\ln T$, however, the favorable events at round $t \geqslant 1$ hold rather for quantities based on how often the pure actions were pulled:

$$\mathcal{E}(t) = \left\{ \forall a \in [K], \quad \big|\mu_a - \widehat{\mu}_a(t)\big| \leqslant \sqrt{\frac{8\sigma^2 \ln t}{\max\big\{N_a(t), 1\big\}}} \right\} \quad \text{and} \quad \mathcal{E}'(t) = \left\{ \forall a \in [K], \quad \sqrt{\frac{8\sigma^2 \ln t}{\max\big\{N_a(t), 1\big\}}} < \frac{\Delta}{2} \right\}$$

We also introduce the following events, for any $\underline{p} \in \mathcal{P}$, though we will use them only for $\underline{p} \in \mathrm{Ext}(\mathcal{P}) \setminus \mathrm{Opt}(\underline{\nu}, \mathcal{P})$ in the sequel:

$$\mathcal{E}''(\underline{p}, t) = \left\{ \sum_{a=1}^{K} p_a \sqrt{\frac{8\sigma^2 \ln t}{\max\big\{N_a(t), 1\big\}}} < \frac{\Delta}{2} \right\}.$$

A first new ingredient consists of the following inequalities, obtained by distinguishing whether $p_a = 0$ or $p_a > 0$ and by Jensen's equality for the square root: for all $\underline{p} \in \mathcal{P}$, all $t \leqslant T$, and all $n \geqslant (65K\sigma^2/\Delta^2) \ln T$,

$$2\sqrt{8\sigma^2 \ln t} \sum_{a \in [K]} p_a \frac{1}{\sqrt{\max\big\{np_a/2,\ 1\big\}}} \leqslant \frac{8\sqrt{\sigma^2 \ln t}}{\sqrt{n}} \sum_{a \in [K]} \sqrt{p_a} \leqslant \frac{8\sqrt{\sigma^2 \ln t}}{\sqrt{n}} \sqrt{K} < \Delta\,,$$

thus the following inclusion:

$$\bigcap_{a \in [K]} \big\{N_a(t) \geqslant np_a/2\big\} = \bigcap_{a : p_a > 0} \big\{N_a(t) \geqslant np_a/2\big\} \subseteq \mathcal{E}''(\underline{p}, t)\,. \tag{6}$$

Note also the inclusion $\mathcal{E}'(t) \subseteq \mathcal{E}''(\underline{p}, t)$, valid for all $\underline{p} \in \mathcal{P}$.

Now, the second new ingredient, consisting of the lemma below, is the key to relate the numbers of times $N_{\underline{p}}(t)$ a suboptimal distribution $\underline{p} \in \mathrm{Ext}(\mathcal{P})$ is picked to the numbers of draws $N_a(t)$ of pure actions $a \in [K]$.

Lemma 1. *Fix $\underline{p} \in \mathrm{Ext}(\mathcal{P})$, and denote by $p_{\min>0} = \min\big\{p_a : a \in [K] \text{ s.t. } p_a > 0\big\} > 0$ its minimal positive component. Then, for all $t \geqslant 1$, all $n \geqslant (10/p_{\min>0}) \ln T$, and all $a \in [K]$,*

$$\mathbb{P}\Big(\big\{N_{\underline{p}}(t) \geqslant n\big\} \cap \big\{N_a(t) < np_a/2\big\}\Big) \leqslant \frac{1}{T}\,.$$

Proof. We only need to show the inequality for $a \in [K]$ such that $p_a > 0$. We note that

$$N_a(t) \geqslant \sum_{s=1}^{t} \mathbb{1}_{\{\underline{p}_s = \underline{p}\}}\, \mathbb{1}_{\{A_s = a\}}\,; \tag{7}$$

thus, by optional skipping[2] (see Theorem 5.2 of Doob, 1953, Chapter III, p. 145, see also Chow & Teicher, 1988, Section 5.3), the distribution of $N_a(t)$ on the event $\big\{N_{\underline{p}}(t) \geqslant n\big\}$ is larger than the distribution of a

[2]Sometimes called optional sampling.

random variable $B_{n,a}$ with binomial distribution of parameters n and p_a. In particular,

$$\mathbb{P}\Big(\big\{N_{\underline{p}}(t) \geqslant n\big\} \cap \big\{N_a(t) < np_a/2\big\}\Big) \leqslant \mathbb{P}\big(B_{n,a} < np_a/2\big) = \mathbb{P}\big(B_{n,a} - np_a < -np_a/2\big) \leqslant \exp\left(-\frac{\varepsilon^2}{2(v + b\varepsilon/3)}\right) \leqslant \exp\left(-\frac{np_a}{8(1+1/6)}\right),$$

where, for the final inequality, we applied Bernstein's inequality (see, e.g., Boucheron et al., 2013, end of Section 2.7, Equation 2.10) with variance $v = n\,p_a(1-p_a)$, upper bound $b = 1$ on the range, and deviation $\varepsilon = np_a/2$. Substituting the bound on n concludes the proof. □

The rest of the analysis is essentially standard. The aim is to control each of the following expectations, for $\underline{p} \in \mathrm{Ext}(\mathcal{P}) \setminus \mathrm{Opt}(\underline{\nu}, \mathcal{P})$ and where $n_{\underline{p}} \geqslant 1$ is defined later:

$$\mathbb{E}\big[N_{\underline{p}}(T)\big] \leqslant n_{\underline{p}} + \sum_{t=n_{\underline{p}}}^{T-1} \mathbb{P}\big\{\underline{p}_{t+1} = \underline{p} \;\text{ and }\; N_{\underline{p}}(t) \geqslant n_{\underline{p}}\big\}. \tag{8}$$

We first note that for $t \geqslant 1$, for all $\underline{p} \in \mathrm{Ext}(\mathcal{P}) \setminus \mathrm{Opt}(\underline{\nu}, \mathcal{P})$,

$$\big\{\underline{p}_{t+1} = \underline{p}\big\} \subseteq \overline{\mathcal{E}(t)} \cup \overline{\mathcal{E}''(\underline{p}, t)} \subseteq \overline{\mathcal{E}(t)} \cup \overline{\mathcal{E}'(t)}\,; \tag{9}$$

indeed, on $\mathcal{E}(t) \cap \mathcal{E}''(\underline{p}, t)$, for $\underline{p}^\star \in \mathrm{Opt}(\underline{\nu}, \mathcal{P})$, by definitions of these sets and of $\underline{U}(t)$,

$$\begin{aligned}\langle \underline{p}, \underline{U}(t)\rangle &= \langle \underline{p}, \widehat{\underline{\mu}}(t)\rangle + \sum_{a=1}^{K} p_a \sqrt{\frac{8\sigma^2 \ln t}{\max\big\{N_a(t), 1\big\}}} \leqslant \langle \underline{p}, \underline{\mu}\rangle + 2\sum_{a=1}^{K} p_a \sqrt{\frac{8\sigma^2 \ln t}{\max\big\{N_a(t), 1\big\}}} \\ &< \langle \underline{p}, \underline{\mu}\rangle + \Delta \leqslant \langle \underline{p}, \underline{\mu}\rangle + \Delta(\underline{p}) = \langle \underline{p}^\star, \underline{\mu}\rangle \leqslant \langle \underline{p}^\star, \widehat{\underline{\mu}}(t)\rangle + \sum_{a=1}^{K} p^\star_a \sqrt{\frac{8\sigma^2 \ln t}{\max\big\{N_a(t), 1\big\}}} = \langle \underline{p}^\star, \underline{U}(t)\rangle\,,\end{aligned}$$

while $\big\{\underline{p}_{t+1} = \underline{p}\big\}$ requires $\langle \underline{p}, \underline{U}(t)\rangle \geqslant \langle \underline{p}^\star, \underline{U}(t)\rangle$. Let

$$n_{\underline{p}} = \max\left\{\frac{65K}{\Delta^2}\ln T,\; \frac{10}{p_{\min>0}}\ln T,\; 1 + \frac{1}{8\sigma^2}\max_{a\in[K]}(\mu_a - u_0)^2\right\}; \tag{10}$$

the third element in the maximum will turn useful in the application of Lemma 2 below. For each distribution $\underline{p} \in \mathrm{Ext}(\mathcal{P}) \setminus \mathrm{Opt}(\underline{\nu}, \mathcal{P})$, the inclusions (9) and then (6) entail

$$\big\{\underline{p}_{t+1} = \underline{p}\big\} \cap \big\{N_{\underline{p}}(t) \geqslant n_{\underline{p}}\big\} \subseteq \overline{\mathcal{E}(t)} \cup \Big(\overline{\mathcal{E}''(\underline{p}, t)} \cap \big\{N_{\underline{p}}(t) \geqslant n_{\underline{p}}\big\}\Big) \subseteq \overline{\mathcal{E}(t)} \cup \bigcup_{a\in[K]} \big\{N_{\underline{p}}(t) \geqslant n_{\underline{p}}\big\} \cap \big\{N_a(t) < n_{\underline{p}} p_a/2\big\}.$$

Substituting this bound into (8), resorting to unions bounds and to Lemma 1, yields

$$\mathbb{E}\big[N_{\underline{p}}(T)\big] \leqslant n_{\underline{p}} + K + \sum_{t=n_{\underline{p}}}^{T-1} \mathbb{P}\Big(\overline{\mathcal{E}(t)}\Big) \leqslant n_{\underline{p}} + K + K\sum_{t=n_{\underline{p}}}^{T-1} (2t\,t^{-4}) \leqslant n_{\underline{p}} + 2K\,, \tag{11}$$

where we applied Lemma 2 below for each $a \in [K]$ and with $\delta = t^{-4}$, which satisfies the condition required therein given that $t \geqslant n_{\underline{p}}$. The proof is concluded by resorting to the decomposition (2), to obtain

$$R_T \leqslant \sum_{\underline{p}\in\mathrm{Ext}(\mathcal{P})\setminus\mathrm{Opt}(\underline{\nu},\mathcal{P})} \Delta_{\underline{p}}(n_{\underline{p}} + 2K)\,, \tag{12}$$

which is of the claimed form $C_{\underline{\nu}} \ln T + c_{\underline{\nu}}$. In the derivation of this regret bound, we targeted simplicity and did not try to improve the constants $C_{\underline{\nu}}$ and $c_{\underline{\nu}}$.

Lemma 2 is an essentially standard concentration result for stochastic bandits; the only adaptation therein (the third new ingredient) is handling the case where $N_a(t) = 0$.

Lemma 2. *Consider a model $\mathcal{D}_{\sigma^2}$ with σ^2–sub-Gaussian distributions, and fix a bandit problem $\underline{\nu}$ in $\mathcal{D}_{\sigma^2}$. For $t \geqslant 1$, if the actions $A_1, \dots, A_t$ and rewards $Y_1, \dots, Y_t$ were generated according to the protocol of Box A, then, for all $a \in [K]$, for all $\delta > 0$ with $2\ln(1/\delta) > (\mu_a - u_0)^2/\sigma^2$,*

$$\mathbb{P}\left\{ \left|\mu_a - \widehat{\mu}_a(t)\right| \geqslant \sqrt{\frac{2\sigma^2 \ln(1/\delta)}{\max\left\{N_a(t), 1\right\}}} \right\} \leqslant 2t\delta \,.$$

Proof. Again by optional skipping (see the proof of Lemma 1), by denoting by $\widehat{\mu}_{a,n}$ an empirical average of $n \geqslant 1$ i.i.d. random variables with distribution ν_a, and by using the convention $\widehat{\mu}_{a,0} = u_0$, we have

$$\begin{aligned}
\mathbb{P}\left\{ \left|\mu_a - \widehat{\mu}_a(t)\right| \geqslant \sqrt{\frac{2\sigma^2 \ln(1/\delta)}{\max\{N_a(t), 1\}}} \right\} &\leqslant \mathbb{P}\left\{ \exists n \in \{0, 1, \dots, t\} \ : \ \left|\mu_a - \widehat{\mu}_{a,n}\right| \geqslant \sqrt{\frac{2\sigma^2 \ln(1/\delta)}{\max\{n, 1\}}} \right\} \\
&\leqslant 0 + \sum_{n=1}^{t} \mathbb{P}\left\{ \left|\mu_a - \widehat{\mu}_{a,n}\right| \geqslant \sqrt{\frac{2\sigma^2 \ln(1/\delta)}{n}} \right\} \leqslant \sum_{n=1}^{t} 2\delta = 2t\delta \,,
\end{aligned}$$

where the case $n = 0$ was dropped in the union bound because $|\mu_a - u_0| < \sqrt{2\sigma^2 \ln(1/\delta)}$ by assumption, and where the final inequalities follow from the Cramér–Chernoff inequality (see, e.g., Lattimore & Szepesvári, 2020, Corollary 5.1). □

A.2 Proof of the constant regret bound in Theorem 1

As indicated at the beginning of Appendix A, the proof of the constant regret bound of Theorem 1 follows a completely different logic. For instance, in Appendix A.1, the sets $\mathcal{E}'(t)$ were instrumental in the proof but we had not controlled their probabilities—which constitutes the core of the analysis here. To do so, we show that optimal distributions are typically played at least half of the time; this is the main contribution of this proof. Then, because $p^\star_{\min}(\underline{\nu}) > 0$, we know that each pure action $a \in [K]$ is played linearly many times, which cannot happen on the events $\overline{\mathcal{E}'(t)}$, where at least one action is only played logarithmically many times. This proof strategy for bounded regret was used in Lattimore & Munos (2014). We face the additional technical challenge here that we do not control with certainty the number of pulls of every pure arm because of the randomness in generating the A_t from the p_t; we handle this by carefully applying Bernstein's inequality.

Step 1: Preparation. We fix a threshold $t_0 \geqslant 8 + \max_{a \in [K]} (\mu_a - u_0)^2/(8\sigma^2)$ such that

$$\forall t \geqslant t_0, \qquad \frac{t}{2}\, p^\star_{\min}(\underline{\nu}) - \frac{32\sigma^2 \ln t}{\Delta^2} \geqslant \sqrt{t \ln t} \qquad \text{and} \qquad \frac{\Delta t}{4} - \sqrt{8\sigma^2 \ln t}\big(1 + 2\sqrt{t-1}\big) > \sqrt{8\sigma^2 t \ln^2 t}\,. \tag{13}$$

For example, with the convention that the $\ln\ln x = -\infty$ if $x \leqslant 1$, the constraints above are satisfied with the threshold t_0 such that

$$\ln t_0 = \max\left\{ 2 + \frac{1}{8\sigma^2},\ \ln \frac{\sigma^2}{\Delta^2\, p^\star_{\min}(\underline{\nu})^2} + 3\ln\ln \frac{18\,432\,\sigma^2}{\Delta^2\, p^\star_{\min}(\underline{\nu})^2} + 10 \right\}. \tag{14}$$

(*Note to reviewers: for the sake of concision, we decided to omit the half-page of calculations that lead to this bound. We could of course add it if deemed necessary.*)

By (9), we first note that

$$R_T \leqslant R_{t_0} + \max_{a \in [K]} \mu_a \sum_{t=t_0}^{T-1} \left(\mathbb{P}\Big(\overline{\mathcal{E}(t)}\Big) + \mathbb{P}\Big(\overline{\mathcal{E}'(t)}\Big) \right) \leqslant R_{t_0} + \max_{a \in [K]} \mu_a \left(K + \sum_{t=t_0}^{T-1} \mathbb{P}\Big(\overline{\mathcal{E}'(t)}\Big) \right),$$

where the final inequality follows from a bound proved in (11), given the first condition on t_0. The key step is the decomposition

$$\overline{\mathcal{E}'(t)} \subseteq \big\{ N_\star(t) < t/2 \big\} \cup \Big(\overline{\mathcal{E}'(t)} \cap \big\{ N_\star(t) \geqslant t/2 \big\} \Big), \qquad \text{where} \qquad N_\star(t) = \sum_{s=1}^{t} \sum_{\underline{p} \in \mathrm{Opt}(\underline{\nu}, \mathcal{P})} \mathbb{1}_{\{\underline{p}_s = \underline{p}\}}$$