A Proofs of Section 3

A.1 Proof of Proposition 1

Proof. We only prove the statement for the optimistic reward, $\widetilde{r}_{\pi,t}$. The proof for the pessimistic cost, $\widetilde{c}_{\pi,t}$, is analogous. From the definition of the confidence set $\mathcal{C}_t^r(\alpha_r)$ in (7), any vector $\theta \in \mathcal{C}_t^r(\alpha_r)$ can be written as $\widehat{\theta}_t + v$, where v satisfying $\|v\|_{\Sigma_t} \leq \alpha_r \beta_t(\delta, d)$. Thus, we may write

$$\begin{aligned}\widetilde{r}_{\pi,t} &= \max_{\theta \in \mathcal{C}_t^r(\alpha_r)} \mathbb{E}_{x\sim\pi}[\langle x, \theta\rangle] = \max_{\theta \in \mathcal{C}_t^r(\alpha_r)} \langle x_\pi, \theta\rangle = \langle x_\pi, \widehat{\theta}_t\rangle + \max_{v:\|v\|_{\Sigma_t} \leq \alpha_r\beta_t(\delta,d)} \langle x_\pi, v\rangle \\ &\overset{(a)}{\leq} \langle x_\pi, \widehat{\theta}_t\rangle + \alpha_r \beta_t(\delta, d)\|x_\pi\|_{\Sigma_t^{-1}}.\end{aligned}$$

(a) By Cauchy-Schwartz, for all v, we have $\langle x_\pi, v\rangle \leq \|x_\pi\|_{\Sigma_t^{-1}}\|v\|_{\Sigma_t}$. The result follows from the condition on v in the maximum, i.e., $\|v\|_{\Sigma_t} \leq \alpha_r\beta_t(\delta, d)$.

Let us define $v^* := \frac{\alpha_r\beta_t(\delta,d)\Sigma_t^{-1}x_\pi}{\|x_\pi\|_{\Sigma_t^{-1}}}$. This value of v^* is feasible because

$$\|v^*\|_{\Sigma_t} = \frac{\alpha_r\beta_t(\delta,d)}{\|x_\pi\|_{\Sigma_t^{-1}}}\sqrt{x_\pi^\top \Sigma_t^{-1}\Sigma_t\Sigma_t^{-1}x_\pi} = \frac{\alpha_r\beta_t(\delta,d)}{\|x_\pi\|_{\Sigma_t^{-1}}}\sqrt{x_\pi^\top \Sigma_t^{-1}x_\pi} = \alpha_r\beta_t(\delta,d).$$

We now show that v^* also achieves the upper-bound in the above inequality resulted from Cauchy-Schwartz

$$\langle x_\pi, v^*\rangle = \frac{\alpha_r\beta_t(\delta,d)x_\pi^\top\Sigma_t^{-1}x_\pi}{\|x_\pi\|_{\Sigma_t^{-1}}} = \alpha_r\beta_t(\delta,d)\|x_\pi\|_{\Sigma_t^{-1}}.$$

Thus, v^* is the maximizer and we can write

$$\widetilde{r}_{\pi,t} = \langle x_\pi, \widehat{\theta}_t\rangle + \langle x_\pi, v^*\rangle = \langle x_\pi, \widehat{\theta}_t\rangle + \alpha_r\beta_t(\delta,d)\|x_\pi\|_{\Sigma_t^{-1}},$$

which concludes the proof. □

A.2 Proof of Proposition 2

Proof. Recall that $\tilde{c}_{\pi,t} = \frac{\langle x_\pi^o, e_0\rangle c_0}{\|x_0\|} + \langle x_\pi^{o,\perp}, \widehat{t}_\pi^{o,\perp}\rangle + \alpha_c\beta_t(\delta, d-1)\|x_\pi^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}} \leq \tau$.

Conditioned on the event $\mathcal{E}$ as defined in equation 16, it follows that:

$$\begin{aligned}|\langle x_\pi^{o,\perp}, \widehat{\mu}_t^{o,\perp} - \mu_*^{o,\perp}\rangle| &\leq \|\mu_*^{o,\perp} - \widehat{\mu}_t^{o,\perp}\|_{\Sigma_t^{o,\perp}}\|x_\pi\|_{(\Sigma_t^{o,\perp})^{-1}} \\ &\leq \langle x_\pi^{o,\perp}, \widehat{\mu}_t^{o,\perp} - \mu_*^{o,\perp}\rangle\beta_t(\delta, d-1)\|x_\pi\|_{(\Sigma_t^{o,\perp})^{-1}}\end{aligned}$$

And therefore:

$$0 \leq \langle x_\pi^{o,\perp}, \widehat{\mu}_t^{o,\perp} - \mu_*^{o,\perp}\rangle + \beta_t(\delta, d-1)\|x_\pi\|_{(\Sigma_t^{o,\perp})^{-1}} \tag{21}$$

Observe that:

$$\begin{aligned}c_\pi &= \frac{\langle x_\pi^o, e_0\rangle c_0}{\|x_0\|} + \langle x_\pi^{o,\perp}, \mu_*^{o,\perp}\rangle \\ &\leq \underbrace{\frac{\langle x_\pi^o, e_0\rangle c_0}{\|x_0\|} + \langle x_\pi^{o,\perp}, \widehat{\mu}_t^{o,\perp}\rangle + \alpha_c\beta_t(\delta, d-1)\|x_\pi^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}}}_{\text{I}}\end{aligned} \tag{22}$$

The last inequality holds by adding Inequality 21 to Inequality 22. Since by assumption for all $\pi \in \Pi_t$ term $I \leq \tau$, we obtain that $c_\pi \leq \tau$. The result follows. □

B Proofs of Section 4

B.1 Proof of Lemma 2

We first state the following proposition that is used in the proof of Lemma 2. This proposition is a direct consequence of Eq. 20.9 and Lemma 19.4 in Lattimore and Szepesvári (2019). Similar result has also been reported in the appendix of Amani et al. (2019).

Proposition 3. *For any sequence of actions $(x_1, \dots, x_t)$, let Σ_t be its corresponding Gram matrix defined by (4) with $\lambda \geq 1$. Then, for all $t \in [T]$, we have*

$$\sum_{s=1}^{T} \|x_s\|_{\Sigma_s^{-1}} \leq \sqrt{2Td \log\left(1 + \frac{TL^2}{\lambda}\right)}.$$

We now state the proof of Lemma 2.

Proof of Lemma 2. We prove this lemma through the following sequence of inequalities:

$$\begin{aligned}\sum_{t=1}^{T} \langle x_t, \widetilde{\theta}_t \rangle - \langle x_t, \theta_* \rangle &\overset{(a)}{\leq} \sum_{t=1}^{T} \|x_t\|_{\Sigma_t^{-1}} \|\widetilde{\theta}_t - \theta_*\|_{\Sigma_t} \overset{(b)}{\leq} \sum_{t=1}^{T} (1+\alpha_r)\beta_t(\delta, d)\|x_t\|_{\Sigma_t^{-1}} \\ &\overset{(c)}{\leq} (1+\alpha_r)\beta_T(\delta, d) \sum_{t=1}^{T} \|x_t\|_{\Sigma_t^{-1}} \overset{(d)}{\leq} (1+\alpha_r)\beta_T(\delta, d)\sqrt{2Td \log\left(1 + \frac{TL^2}{\lambda}\right)}\end{aligned}$$

(a) This is by Cauchy-Schwartz.

(b) This follows from the fact that $\widetilde{\theta}_t \in \mathcal{C}_t^r(\alpha_r)$ and we are on event $\mathcal{E}$.

(c) This is because $\beta_t(\delta, d)$ is an increasing function of t, i.e., $\beta_T(\delta, d) \geq \beta_t(\delta, d)$, $\forall t \in [T]$.

(d) This is a direct result of Proposition 3. □

B.2 Proof of Lemma 3

Proof. In order to prove the desired result it is enough to show that:

$$\left(x_\pi^{o,\perp}\right)^\top \left(\Sigma_t^{o,\perp}\right)^\dagger x_\pi^{o,\perp} \leq x_\pi^\top \Sigma_t^{-1} x_\pi$$

w.l.o.g. we can assume $x_o = e_1$, the first basis vector. Notice that in this case $\Sigma_t^{o,\perp}$ can be thought of as a submatrix of Σ_t such that $\Sigma_t[2:, 2:] = \Sigma_t^{o,\perp}$, where $\Sigma_t[2:, 2:]$ denotes the submatrix with row and column indices from 2 onwards.

Using the following formula for the inverse of a psd symmetric matrix:

$$\begin{bmatrix} Z & \delta \\ \delta^\top & A \end{bmatrix} = \begin{bmatrix} \frac{1}{D} & -\frac{A^{-1}\delta}{D} \\ -\frac{\delta^\top A^{-1}}{D} & A^{-1} + \frac{A^{-1}\delta\delta^\top A^{-1}}{D} \end{bmatrix}$$

Where $D = z - \delta^\top A^{-1}\delta$. In our case $D = \Sigma_t[1,1] - \Sigma_t[2:d]^\top \left(\Sigma_t^{o,\perp}\right)^{-1} \Sigma_t[2:d] \in \mathbb{R}$. Observe that since Σ_t is PSD, $D \geq 0$. Therefore:

$$\Sigma_t^{-1} = \begin{bmatrix} 1/D & -\frac{\left(\Sigma_t^{o,\perp}\right)^{-1}\Sigma_t[2,:d]}{D} \\ -\frac{\Sigma_t^\top[2:d]\left(\Sigma_t^{o,\perp}\right)^{-1}}{D} & \left(\Sigma_t^{o,\perp}\right)^{-1} + \frac{\left(\Sigma_t^{o,\perp}\right)^{-1}\Sigma_t[2:d]\Sigma_t[2:d]\left(\Sigma_t^{o,\perp}\right)^{-1}}{D} \end{bmatrix}$$

Then:

$$
\begin{aligned}
x_\pi^\top \left(\Sigma_t^{-1}\right)^{-1} x_\pi = & \frac{x_\pi(1)^2 - 2x_\pi(1)\Sigma_t[2:d]^\top \left(\Sigma_t^{o,\perp}\right)^{-1} x_\pi[2:d]}{D} + \\
& \frac{x_\pi[2:d]^\top \left(\Sigma_t^{o,\perp}\right)^{-1} \Sigma_t[2:d]\Sigma_t[2:d]^\top \left(\Sigma_t^{o,\perp}\right)^{-1} x_\pi[2:d]}{D} \\
& + x_\pi[2:d]^\top \left(\Sigma_t^{o,\perp}\right)^{-1} x_\pi[2:d] \\
& \geq x_\pi[2:d]^\top \left(\Sigma_t^{o,\perp}\right)^{-1} x_\pi[2:d]
\end{aligned}
$$

The result follows by noting that $x_\pi[2:d] = x_\pi^{o,\perp}$. □

B.3 Proof of Lemma 4

Proof. For any policy π, we have

$$\widetilde{r}_{\pi,t} = \max_{\theta \in \mathcal{C}_t^r(\alpha_r)} \langle x_\pi, \theta \rangle \geq \langle x_\pi, \theta_* \rangle = r_\pi. \tag{23}$$

If $\pi_t^* \in \Pi_t$, then by the definition of π_t (Line 4 of Algorithm 1), we have

$$\widetilde{r}_{\pi_t,t} \geq \widetilde{r}_{\pi_t^*,t}. \tag{24}$$

Combining (23) and (24), we may conclude that $\widetilde{r}_{\pi_t,t} \geq r_{\pi_t^*}$ as desired.

We now focus on the case that $\pi_t^* \notin \Pi_t$, i.e.,

$$\widetilde{c}_{\pi_t^*,t} = \frac{\langle x_{\pi_t^*}^o, e_0 \rangle c_0}{\|x_0\|} + \langle x_{\pi_t^*}^{o,\perp}, \widehat{\mu}_t^{o,\perp} \rangle + \alpha_c \beta_t(\delta, d-1) \|x_{\pi_t^*}^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}} > \tau.$$

We define a mixture policy $\widetilde{\pi}_t = \eta_t \pi_t^* + (1-\eta_t)\pi_0$, where π_0 is the policy that always selects the safe action x_0 and $\eta_t \in [0,1]$ is the maximum value of η such that $\left(\eta \pi_t^* + (1-\eta)\pi_0\right) \in \Pi_t$. Conceptually, η_t shows how close is the optimal policy π_t^* to the set of safe policies Π_t.

By the definition of $\widetilde{\pi}_t$, we have

$$x_{\widetilde{\pi}_t}^o = \eta_t x_{\pi_t^*}^o + (1-\eta_t)x_0, \qquad x_{\widetilde{\pi}_t}^{o,\perp} = \eta_t x_{\pi_t^*}^{o,\perp}, \tag{25}$$

which allows us to write

$$
\begin{aligned}
\widetilde{c}_{\widetilde{\pi}_t,t} &= \frac{\eta_t \langle x_{\pi_t^*}^o, e_0 \rangle + (1-\eta_t)\langle x_0, e_0 \rangle}{\|x_0\|} \cdot c_0 + \eta_t \langle x_{\pi_t^*}^{o,\perp}, \widehat{\mu}_t^{o,\perp} \rangle + \eta_t \alpha_c \beta_t(\delta, d-1) \|x_{\pi_t^*}^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}} \\
&= \frac{(1-\eta_t)\langle x_0, e_0 \rangle c_0}{\|x_0\|} + \eta_t \widetilde{c}_{\pi_t^*,t}.
\end{aligned}
$$

From the definition of η_t, we have $\widetilde{c}_{\widetilde{\pi}_t,t} = \frac{(1-\eta_t)\langle x_0, e_0 \rangle c_0}{\|x_0\|} + \eta_t \widetilde{c}_{\pi_t^*,t} = \tau$, and thus, we may write

$$
\begin{aligned}
\eta_t &= \frac{\tau - \frac{\langle x_0, e_0 \rangle c_0}{\|x_0\|}}{\widetilde{c}_{\pi_t^*,t} - \frac{\langle x_0, e_0 \rangle c_0}{\|x_0\|}} = \frac{\tau - c_0}{\frac{\langle x_{\pi_t^*}^o, e_0 \rangle c_0}{\|x_0\|} + \langle x_{\pi_t^*}^{o,\perp}, \widehat{\mu}_t^{o,\perp} \rangle + \alpha_c \beta_t(\delta, d-1) \|x_{\pi_t^*}^{o,\perp}\|_{(\Sigma_t^{0,\perp})^{-1}} - c_0} \\
&= \frac{\tau - c_0}{\frac{\langle x_{\pi_t^*}^o, e_0 \rangle c_0}{\|x_0\|} + \langle x_{\pi_t^*}^{o,\perp}, \mu_* \rangle + \langle x_{\pi_t^*}^{o,\perp}, \widehat{\mu}_t^{o,\perp} - \mu_* \rangle + \alpha_c \beta_t(\delta, d-1) \|x_{\pi_t^*}^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}} - c_0} \\
&\overset{(a)}{\geq} \frac{\tau - c_0}{\frac{\langle x_{\pi_t^*}^o, e_0 \rangle c_0}{\|x_0\|} + \langle x_{\pi_t^*}^{o,\perp}, \mu_* \rangle + (1+\alpha_c) \beta_t(\delta, d-1) \|x_{\pi_t^*}^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}} - c_0} \\
&\overset{(b)}{\geq} \frac{\tau - c_0}{\tau + (\alpha_c + 1)\beta_t(\delta, d-1) \|x_{\pi_t^*}^{o,\perp}\|_{(\Sigma_t^{o,\perp})^{-1}} - c_0}.
\end{aligned} \tag{26}
$$

(a) This holds because

$$\langle x^{o,\perp}_{\pi^*_t}, \widehat{\mu}^{o,\perp}_t - \mu_* \rangle = \langle x^{o,\perp}_{\pi^*_t}, \widehat{\mu}^{o,\perp}_t - \mu^{o,\perp}_* \rangle \leq \|\widehat{\mu}^{o,\perp}_t - \mu^{o,\perp}_*\|_{\Sigma^{o,\perp}_t} \|x^{o,\perp}_{\pi^*_t}\|_{(\Sigma^{o,\perp}_t)^{-1}} \leq \beta_t(\delta, d-1)\|x^{o,\perp}_{\pi^*_t}\|_{(\Sigma^{o,\perp}_t)^{-1}},$$

where the last inequality is because we are on the event $\mathcal{E}$.

(b) This passage is due to the fact that the optimal policy π^*_t is feasible, and thus, $\mathbb{E}_{x\sim\pi^*_t}[\langle x, \mu_*\rangle] \leq \tau$. Therefore, we may write

$$\begin{aligned}
\mathbb{E}_{x\sim\pi^*_t}[\langle x, \mu_*\rangle] &= \mathbb{E}_{x\sim\pi^*_t}[\langle x^o, \mu_*\rangle] + \langle x^{o,\perp}_{\pi^*_t}, \mu_*\rangle = \mathbb{E}_{x\sim\pi^*_t}[\langle \langle x, e_0\rangle e_0, \mu_*\rangle] + \langle x^{o,\perp}_{\pi^*_t}, \mu_*\rangle \\
&= \mathbb{E}_{x\sim\pi^*_t}[\langle \langle x, e_0\rangle \frac{x_0}{\|x_0\|}, \mu_*\rangle] + \langle x^{o,\perp}_{\pi^*_t}, \mu_*\rangle = \frac{c_0}{\|x_0\|}\mathbb{E}_{x\sim\pi^*_t}[\langle x, e_0\rangle] + \langle x^{o,\perp}_{\pi^*_t}, \mu_*\rangle \\
&= \frac{\langle x^o_{\pi^*_t}, e_0\rangle c_0}{\|x_0\|} + \langle x^{o,\perp}_{\pi^*_t}, \mu_*\rangle \leq \tau.
\end{aligned}$$

Since $\widetilde{\pi}_t \in \Pi_t$, we have

$$\begin{aligned}
\widetilde{r}_{\pi_t,t} \geq \widetilde{r}_{\widetilde{\pi}_t,t} &= \langle x_{\widetilde{\pi}_t}, \widehat{\theta}_t\rangle + \alpha_r\beta_t(\delta,d)\|x_{\widetilde{\pi}_t}\|_{\Sigma_t^{-1}} = \langle x_{\widetilde{\pi}_t}, \theta_*\rangle + \langle x_{\widetilde{\pi}_t}, \widehat{\theta}_t - \theta_*\rangle + \alpha_r\beta_t(\delta,d)\|x_{\widetilde{\pi}_t}\|_{\Sigma_t^{-1}} \\
&\overset{(a)}{\geq} \langle x_{\widetilde{\pi}_t}, \theta_*\rangle + (\alpha_r - 1)\beta_t(\delta,d)\|x_{\widetilde{\pi}_t}\|_{\Sigma_t^{-1}} \overset{(b)}{\geq} \langle x_{\widetilde{\pi}_t}, \theta_*\rangle + (\alpha_r - 1)\beta_t(\delta,d-1)\|x^{o,\perp}_{\widetilde{\pi}_t}\|_{(\Sigma^{o,\perp}_t)^{-1}} \\
&\overset{(c)}{=} \eta_t\langle x_{\pi^*}, \theta_*\rangle + (1-\eta_t)\langle x_0, \theta_*\rangle + \eta_t(\alpha_r - 1)\beta_t(\delta,d-1)\|x^{o,\perp}_{\pi^*_t}\|_{(\Sigma^{o,\perp}_t)^{-1}} \\
&\overset{(d)}{\geq} \eta_t\langle x_{\pi^*_t}, \theta_*\rangle + \eta_t(\alpha_r - 1)\beta_t(\delta,d-1)\|x^{o,\perp}_{\pi^*}\|_{(\Sigma^{o,\perp}_t)^{-1}} \\
&\overset{(e)}{\geq} \underbrace{\Big(\frac{\tau - c_0}{\tau - c_0 + (\alpha_c+1)\beta_t(\delta,d-1)\|x^{o,\perp}_{\pi^*_t}\|_{(\Sigma^{o,\perp}_t)^{-1}}}\Big)\Big(\langle x_{\pi^*_t}, \theta_*\rangle + (\alpha_r-1)\beta_t(\delta,d-1)\|x^{o,\perp}_{\pi^*_t}\|_{(\Sigma^{o,\perp}_t)^{-1}}\Big)}_{C_0}. \qquad (27)
\end{aligned}$$

(a) This is because we may write

$$|\langle x_{\widetilde{\pi}_t}, \widehat{\theta}_t - \theta_*\rangle| \leq \|\widehat{\theta}_t - \theta_*\|_{\Sigma_t}\|x_{\widetilde{\pi}_t}\|_{\Sigma_t^{-1}} \leq \beta_t(\delta,d)\|x_{\widetilde{\pi}_t}\|_{\Sigma_t^{-1}},$$

where the last inequality is due to the fact that we are on the event $\mathcal{E}$. Thus, $\langle x_{\widetilde{\pi}_t}, \widehat{\theta}_t - \theta_*\rangle \geq -\beta_t(\delta,d)\|x_{\widetilde{\pi}_t}\|_{\Sigma_t^{-1}}$.

(b) This is a consequence of Lemma 3 stated in the paper and proved in Appendix B.2.

(c) This is from the definition of $\widetilde{\pi}$ and Eq. 25.

(d) This is because $\eta_t \in [0,1]$ and from Assumption 4 we have that all expected rewards are positive (belong to $[0,1]$), and thus, $\langle x_0, \theta_*\rangle \geq 0$.

(e) This is by lower-bounding η_t from (26).

Let us define the shorthand notation $C_1 := \beta_t(\delta, d-1)\|x^{o,\perp}_{\pi^*_t}\|_{(\Sigma^{o,\perp}_t)^{-1}}$. Thus, we may write C_0 as

$$C_0 = \frac{\tau - c_0}{\tau - c_0 + (1+\alpha_c)C_1} \times \big(\langle x_{\pi^*_t}, \theta_*\rangle + (\alpha_r - 1)C_1\big).$$

Note that $C_0 \geq \langle x_{\pi^*_t}, \theta_*\rangle = r_{\pi^*_t}$ (and as a results $\widetilde{r}_{\pi_t,t} \geq r_{\pi^*_t}$ as desired) iff:

$$(\tau - c_0)r_{\pi^*_t} + (\tau - c_0)(\alpha_r - 1)C_1 \geq (\tau - c_0)r_{\pi^*_t} + (1+\alpha_c)C_1 r_{\pi^*_t},$$

which holds iff: $(\tau - c_0)(\alpha_r - 1)C_1 \geq (1+\alpha_c)C_1 r_{\pi^*_t}$.

Since $r_{\pi^*_t} \leq 1$ from Assumption 4, this holds iff: $1 + \alpha_c \leq (\tau - c_0)(\alpha_r - 1)$. This concludes the proof as for both cases of $\pi^*_t \in \Pi_t$ and $\pi^*_t \notin \Pi_t$, we proved that $\widetilde{r}_{\pi_t,t} \geq r_{\pi^*_t}$. □

B.4 Unknown c_0

In this section we relax Assumption 5, and instead assume we only have the knowledge of a safe arm x_0, but its expected cost c_0 is unknown and needs to be learned. If the cost of the safe arm c_0 is unknown, we start by taking the safe action x_0 for T_0 rounds to produce first an empirical mean estimator $\hat{c}_0$. Notice that for all $\delta \in (0,1)$, $\hat{c}_0$ satisfies:

$$\mathbb{P}\left(\hat{c}_0 \leq c_0 - \sqrt{\frac{2\log(1/\delta)}{T_0}}\right) \leq \delta \tag{28}$$

Let $\tilde{c}_0 = \hat{c}_0 + \sqrt{\frac{2\log(1/\delta)}{T_0}}$. By inequality 28, it follows that with probability at least $1-\delta$:

$$\tilde{c}_0 \geq c_0$$

We select T_0 in an adaptive way. In other words, we do the following:

Let $\delta = \frac{1}{T^2}$. And let $\hat{c}_0(t)$ be the sample mean estimator of c_0, when using only t samples. Similarly define $\tilde{c}_0(t) = \hat{c}_0(t) + \sqrt{\frac{2\log(1/\delta)}{t}}$ Let's condition on the event $\mathcal{E}$ that for all $t \in [T]$:

$$|\hat{c}_0(t) - c_0| \leq \sqrt{\frac{2\log(1/\delta)}{t}}$$

By assumption $\mathbb{P}(\mathcal{E}) \geq 1 - T2\delta = 1 - \frac{2}{T}$. Let T_0 be the first time that $\tilde{c}_0(T_0) + 2\sqrt{\frac{2\log(1/\delta)}{T_0}} \leq \tau$.

Notice that in this case and conditioned on $\mathcal{E}$ and therefore on $\tilde{c}_0(T_0) \geq c_0$:

$$\sqrt{\frac{2\log(1/\delta)}{T_0}} \leq \frac{\tau - c_0}{2} \quad \text{i.e.} \quad T_0 \geq \frac{8\log(1/\delta)}{(\tau - c_0)^2}$$

In other words, this test does not stop until $T_0 \geq \frac{8\log(1/\delta)}{(\tau-c_0)^2}$. Now we see it won't take much longer than that to stop:

Conversely, let $T_0' \geq \frac{32\log(1/\delta)}{(\tau-c_0)^2}$. For any such T_0' we observe that by conditioning on $\mathcal{E}$:

$$\tilde{c}_0(T_0') + 2\sqrt{\frac{2\log(1/\delta)}{T_0'}} \leq c_0 + 4\sqrt{\frac{2\log(1/\delta)}{T_0'}} \leq \tau$$

Thus conditioned on $\mathcal{E}$, we conclude $\frac{8\log(1/\delta)}{(\tau-c_0)^2} \leq T_0 \leq \frac{32\log(1/\delta)}{(\tau-c_0)^2}$. Then,

Therefore $\hat{\delta}_c = \sqrt{\frac{8\log(1/\delta)}{T_0}}$ would serve as a conservative estimator for $\frac{\tau-c_0}{2}$ satisfying:

$$\frac{\tau - c_0}{2} \leq \hat{\delta}_c \leq \tau - c_0$$

We proceed by warm starting our estimators for θ_* and μ_* using the data collected by playing x_0. However, instead of estimating $\mu_*^{o,\perp}$, we build an estimator for μ_* over all its directions, including e_0, similar to what OPLB does for θ_*. We then set $\frac{\alpha_r}{\alpha_c} = 1/\hat{\delta}_c$ and run Algorithm 1 for rounds $t > T_0$. Since the scaling of α_r w.r.t. α_c is optimal up to constants, the same arguments hold.

C Constrained Multi-Armed Bandits

C.1 Optimism Pessimism

Here we reproduce the full pseudo-code for OPB:

Algorithm 2 Optimism-Pessimism

Input: Number of arms K, constants $\alpha_r, \alpha_c \geq 1$.
for $t = 1, \ldots, T$ **do**
1. Compute estimates $\{u_a^r(t)\}_{a\in\mathcal{A}}$, $\{u_a^c(t)\}_{a\in\mathcal{A}}$.
2. Form the approximate LP (20) using these estimates.
3. Find policy π_t by solving (20).
4. Play arm $a \sim \pi_t$

Similar to the case of OPLB, we define $\Pi_t = \{\pi \in \Delta_{\mathcal{A}} : \sum_{a\in\mathcal{A}} \pi_a u_a^c(t) \leq \tau\}$. We also define $\beta_a(0) = 0$ for all $a \in \mathcal{A}$.

C.2 The LP Structure

The main purpose of this section is to prove the optimal solutions of the linear program from (20) are supported on a set of size at most 2. This structural result will prove important to develop simple efficient algorithms to solve for solving it. Let's recall the form of the Linear program in (20), i.e.,

$$\max_{\pi\in\Delta_K} \sum_{a\in\mathcal{A}} \pi_a u_a^r(t), \qquad \text{s.t.} \quad \sum_{a\in\mathcal{A}} \pi_a u_a^c(t) \leq \tau.$$

Let's start by observing that in the case $K = 2$ with $\mathcal{A} = \{a_1, a_2\}$ and $u_{a_1}^c(t) < \tau < u_{a_2}^c(t)$, the optimal policy π^* is a mixture policy satisfying:

$$\pi^*_{a_1} = \frac{u^c_{a_2}(t) - \tau}{u^c_{a_2}(t) - u^c_{a_1}(t)}, \qquad \pi^*_{a_2} = \frac{\tau - u^c_{a_1}(t)}{u^c_{a_2}(t) - u^c_{a_1}(t)}. \tag{29}$$

The main result in this section is the following Lemma:

Lemma 7 (support of π^*)**.** *If (20) is feasible, there exists an optimal solution with at most 2 non-zero entries.*

Proof. We start by inspecting the dual problem of (20):

$$\min_{\lambda\geq 0} \max_a \lambda(\tau - u_a^c(t)) + u_a^r(t) \tag{D}$$

This formulation is easily interpretable. The quantity $\tau - u_a^c(t)$ measures the feasibility gap of arm a, while $u_a^r(t)$ introduces a dependency on the reward signal. Let λ^* be the optimal value of the dual variable λ. Define $\mathcal{A}^* \subseteq \mathcal{A}$ as $\mathcal{A}^* = \arg\max_a \lambda^*(\tau - u_a^c(t)) + u_a^r(t)$. By complementary slackness the set of nonzero entries of π^* must be a subset of $\mathcal{A}^*$.

If $|\mathcal{A}^*| = 1$, complementary slackness immediately implies the desired result. If a_1, a_2 are two elements of $\mathcal{A}^*$, it is easy to see that:

$$u^r_{a_1}(t) - \lambda^* u^c_{a_1}(t) = u^r_{a_2}(t) - \lambda^* u^c_{a_2}(t),$$

and thus,

$$\lambda^* = \frac{u^r_{a_2}(t) - u^r_{a_1}(t)}{u^c_{a_2}(t) - u^c_{a_1}(t)}. \tag{30}$$

If $\lambda^* = 0$, the optimal primal value is achieved by concentrating all mass on any of the arms in $\mathcal{A}^*$. Otherwise, plugging (30) back into the objective of (D) and rearranging the terms, we obtain

$$\text{(D)} = \lambda^*(\tau - u^c_{a_1}(t)) + u^r_{a_1}(t) = u^r_{a_2}(t)\left(\frac{\tau - u^c_{a_1}(t)}{u^c_{a_2}(t) - u^c_{a_1}(t)}\right) + u^r_{a_1}(t)\left(\frac{u^c_{a_2}(t) - \tau}{u^c_{a_2}(t) - u^c_{a_1}(t)}\right).$$

If $u^c_{a_2}(t) \geq \tau \geq u^c_{a_1}(t)$, we obtain a feasible value for the primal variable $\pi^*_{a_1} = \frac{\tau - u^c_{a_1}(t)}{u^c_{a_2}(t) - u^c_{a_1}(t)}$, $\pi^*_{a_2} = \frac{u^c_{a_2}(t) - \tau}{u^c_{a_2}(t) - u^c_{a_1}(t)}$ and zero for all other $a \in \mathcal{A} \backslash \{a_1, a_2\}$. Since we have assumed (20) to be feasible there must be either one arm $a^* \in \mathcal{A}^*$ satisfying $a^* = \arg\max_{a \in \mathcal{A}^*} u^r_a(t)$ and $u^c_{a^*}(t) \leq \tau$ or two such arms a_1 and a_2 in $\mathcal{A}^*$ that satisfy $u^c_{a_2}(t) \geq \tau \geq u^c_{a_1}(t)$, since otherwise it would be impossible to produce a feasible primal solution without having any of its supporting arms a satisfying $u^c_a(t) \leq \tau$, there must exist an arm $a \in \mathcal{A}^*$ with $u^c_a(t) < \tau$. This completes the proof. □

From the proof of Lemma 5 we can conclude the optimal policy is either a delta mass centered at the arm with the largest reward - whenever this arm is feasible - or it is a strict mixture supported on two arms.

A further consequence of Lemma 7 is that it is possible to find the optimal solution π^* to problem 20 by simply enumerating all pairs of arms (a_i, a_j) and all singletons, compute their optimal policies (if feasible) using Equation 29 and their values and selecting the feasible pair (or singleton) achieving the largest value. More sophisticated methods can be developed by taking into account elimination strategies to prune out arms that can be determined in advance not to be optimal nor to belong to an optimal pair. Overall this method is more efficient than running a linear programming solver on (20).

If we had instead m constraints, a similar statement to Lemma 5 holds, namely it is possible to show the optimal policy will have support of size at most $m + 1$. The proof is left as an exercise for the reader.

C.3 Regret Analysis

In order to show a regret bound for Algorithm 2, we start with the following regret decomposition:

$$\mathcal{R}_\Pi(T) = \sum_{t=1}^{T} \mathbb{E}_{a \sim \pi^*}[\bar{r}_a] - \mathbb{E}_{a \sim \pi_t}[\bar{r}_a] = \underbrace{\left(\sum_{t=1}^{T} \mathbb{E}_{a \sim \pi^*}[\bar{r}_a] - \mathbb{E}_{a \sim \pi_t}[u^r_a(t)] \right)}_{\text{(I)}} + \underbrace{\left(\sum_{t=1}^{T} \mathbb{E}_{a \sim \pi_t}[u^r_a(t)] - \mathbb{E}_{a \sim \pi_t}[\bar{r}_a] \right)}_{\text{(II)}}.$$

In order to bound $\mathcal{R}_\Pi(T)$, we independently bound terms (I) and (II). We start by bounding term (I). We proceed by first proving an Lemma 6, the equivalent version of Lemma 4 for the multi armed bandit problem.

C.4 Proof of Lemma 6

Proof. Throughout this proof we denote as π_0 to the delta function over the safe arm 1. We start by noting that under $\mathcal{E}$, and because $\alpha_r, \alpha_c \geq 1$, then:

$$(\alpha_r - 1)\beta_a(t) \leq \xi^r_a(t) \leq (\alpha_r + 1)\beta_a(t) \ \forall a \quad \text{and} \quad (\alpha_c - 1)\beta_a(t) \leq \xi^c_a(t) \leq (\alpha_c + 1)\beta_a(t) \ \forall a \neq 0. \tag{31}$$

If $\pi^* \in \Pi_t$, it immediately follows that:

$$\mathbb{E}_{a \sim \pi^*}[\bar{r}_a] \leq \mathbb{E}_{a \sim \pi^*}[u^r_a(t)] \leq \mathbb{E}_{a \sim \pi_t}[u^r_a(t)]. \tag{32}$$

Let's now assume $\pi^* \notin \Pi_t$, i.e., $\mathbb{E}_{a \sim \pi^*}[u^c_a(t)] > \tau$. Let $\pi^* = \rho^* \bar{\pi}^* + (1 - \rho)\pi_0$ with $\bar{\pi}^* \in \Delta_K[2:K]$[5].

Consider a mixture policy $\widetilde{\pi}_t = \gamma_t \pi^* + (1 - \gamma_t)\pi_0 = \gamma_t \rho^* \bar{\pi}^* + (1 - \gamma_t \rho^*)\pi_0$, where γ_t is the maximum $\gamma_t \in [0, 1]$ such that $\widetilde{\pi}_t \in \Pi_t$. It can be easily established that

$$\gamma_t = \frac{\tau - \bar{c}_1}{\rho^* \mathbb{E}_{a \sim \bar{\pi}^*}[u^c_a(t)] - \rho^* \bar{c}_1} = \frac{\tau - \bar{c}_1}{\mathbb{E}_{a \sim \bar{\pi}^*}[\rho^*(\bar{c}_a + \xi^c_a(t))] - \rho^* \bar{c}_1} \overset{\text{(i)}}{\geq} \frac{\tau - \bar{c}_1}{\tau - \bar{c}_1 + \rho^*(1 + \alpha_c)\mathbb{E}_{a \sim \bar{\pi}^*}[\beta_a(t)]}.$$

(i) is a consequence of (31) and of the observation that since π^* is feasible $\rho^* \mathbb{E}_{a \sim \bar{\pi}^*}[\bar{c}_a] + (1 - \rho^*)\bar{c}_1 \leq \tau$. Since

[5] In other words, the support of $\bar{\pi}^*$ does not contain the safe arm 1.

$\widetilde{\pi}_t \in \Pi_t$, we have

$$\begin{aligned}
\mathbb{E}_{a\sim\pi_t}[u_a^r(t)] &\geq \underbrace{\gamma_t \mathbb{E}_{a\sim\pi^*}[u_a^r(t)] + (1-\gamma_t)u_0^r(t)}_{\mathbb{E}_{a\sim\widetilde{\pi}_t}[u_a^r(t)]} \\
&\overset{\text{(ii)}}{\geq} \frac{\tau-\bar{c}_1}{\tau-\bar{c}_1+\rho^*(1+\alpha_c)\mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)]} \times \mathbb{E}_{a\sim\pi^*}[u_a^r(t)] \\
&= \frac{\tau-\bar{c}_1}{\tau-\bar{c}_1+\rho^*(1+\alpha_c)\mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)]} \times \left(\mathbb{E}_{a\sim\pi^*}[\bar{r}_a] + \mathbb{E}_{a\sim\pi^*}[\xi_a^r(t)]\right) \\
&\overset{\text{(iii)}}{\geq} \frac{\tau-\bar{c}_1}{\tau-\bar{c}_1+\rho^*(1+\alpha_c)\mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)]} \times \left(\mathbb{E}_{a\sim\pi^*}[\bar{r}_a] + (\alpha_r-1)\mathbb{E}_{a\sim\pi^*}[\beta_a(t)]\right) \\
&\overset{\text{(iv)}}{\geq} \underbrace{\frac{\tau-\bar{c}_1}{\tau-\bar{c}_1+(1+\alpha_c)\mathbb{E}_{a\sim\pi^*}[\beta_a(t)]} \times \left(\mathbb{E}_{a\sim\pi^*}[\bar{r}_a] + (\alpha_r-1)\mathbb{E}_{a\sim\pi^*}[\beta_a(t)]\right)}_{C_0}.
\end{aligned}$$

(ii) holds because $u_0^r(t) \geq 0$. **(iii)** is a consequence of (31) and **(iv)** follows because $\mathbb{E}_{a\sim\pi^*}[\beta_a(t)] = \rho^*\mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)] + (1-\rho^*)\beta_0(t) \geq \rho^*\mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)]$ since $\beta_a(t) \geq 0$ for all a and t.

Let $C_1 = \mathbb{E}_{a\sim\pi^*}[\beta_a(t)]$. The following holds:

$$C_0 = \frac{\tau-\bar{c}_1}{\tau-\bar{c}_1+(1+\alpha_c)C_1} \times \left(\mathbb{E}_{a\sim\pi^*}[\bar{r}_a] + (\alpha_r-1)C_1\right).$$

Note that $C_0 \geq \mathbb{E}_{a\sim\pi^*}[\bar{r}_a]$ iff:

$$(\tau-\bar{c}_1)\mathbb{E}_{a\sim\pi^*}[\bar{r}_a] + (\tau-\bar{c}_1)(\alpha_r-1)C_1 \geq (\tau-\bar{c}_1)\mathbb{E}_{a\sim\pi^*}[\bar{r}_a] + (1+\alpha_c)C_1\mathbb{E}_{a\sim\pi^*}[\bar{r}_a],$$

which holds iff:

$$(\tau-\bar{c}_1)(\alpha_r-1)C_1 \geq (1+\alpha_c)C_1\mathbb{E}_{a\sim\pi^*}[\bar{r}_a].$$

Since $\mathbb{E}_{a\sim\pi^*}[\bar{r}_a] \leq 1$, this holds if $1+\alpha_c \leq (\tau-\bar{c}_1)(\alpha_r-1)$. □

Proposition 4. *If $\delta = \frac{\epsilon}{4KT}$ for $\epsilon \in (0,1)$, $\alpha_r, \alpha_c \geq 1$ with $\alpha_c \leq \tau(\alpha_r-1)$, then with probability at least $1-\frac{\epsilon}{2}$, we have*

$$\sum_{t=1}^{T} \mathbb{E}_{a\sim\pi^*}[\bar{r}_a] - \mathbb{E}_{a\sim\pi_t}[u_a^r(t)] \leq 0$$

Proof. A simple union bound implies that $\mathbb{P}(\mathcal{E}) \geq 1-\frac{\epsilon}{2}$. Combining this observation with Lemma 6 yields the result. □

Term (II) can be bounded using the confidence intervals radii:

Proposition 5. *If $\delta = \frac{\epsilon}{4KT}$ for an $\epsilon \in (0,1)$, then with probability at least $1-\frac{\epsilon}{2}$, we have*

$$\sum_{t=1}^{T} \mathbb{E}_{a\sim\pi_t}[u_a^r(t)] - \mathbb{E}_{a\sim\pi_t}[\bar{r}_a] \leq (\alpha_r+1)\left(2\sqrt{2TK\log(1/\delta)} + 4\sqrt{T\log(2/\epsilon)\log(1/\delta)}\right).$$

Proof. Under these conditions $\mathbb{P}(\mathcal{E}) \geq 1-\frac{\epsilon}{2}$. Recall $u_a^r(t) = \widehat{r}_a(t) + \alpha_r\beta_a(t)$ and that conditional on $\mathcal{E}$, $\bar{r}_a \in [\widehat{r}_a(t)-\beta_a(t), \widehat{r}_a(t)+\beta_a(t)]$ for all $t \in [T]$ and $a \in \mathcal{A}$. Thus, for all t, we have

$$\mathbb{E}_{a\sim\pi_t}[u_a^r(t)] - \mathbb{E}_{a\sim\pi_t}[\bar{r}_a] \leq (\alpha_r+1)\mathbb{E}_{a\sim\pi_t}[\beta_a(t)].$$

Let $\mathcal{F}_{t-1}$ be the sigma algebra defined up to the choice of π_t and a_t' be a random variable distributed as $\pi_t \mid \mathcal{F}_{t-1}$ and conditionally independent from a_t, i.e., $a_t' \perp a_t \mid \mathcal{F}_{t-1}$. Note that by definition the following equality holds:

$$\mathbb{E}_{a\sim\pi_t}[\beta_a(t)] = \mathbb{E}_{a_t'\sim\pi_t}[\beta_a(t) \mid \mathcal{F}_{t-1}].$$

Consider the following random variables $A_t = \mathbb{E}_{a'_t \sim \pi_t}[\beta_{a'_t}(t) \mid \mathcal{F}_{t-1}] - \beta_{a_t}(t)$. Note that $M_t = \sum_{i=1}^t A_i$ is a martingale. Since $|A_t| \leq 2\sqrt{2\log(1/\delta)}$, a simple application of Azuma-Hoeffding[6] implies:

$$\mathbb{P}\left(\underbrace{\sum_{t=1}^T \mathbb{E}_{a\sim\pi_t}[\beta_a(t)] \geq \sum_{t=1}^T \beta_{a_t}(t) + 4\sqrt{T\log(2/\epsilon)\log(1/\delta)}}_{\mathcal{E}_A^c}\right) \leq \epsilon/2.$$

We can now upper-bound $\sum_{t=1}^T \beta_{a_t}(t)$. Note that $\sum_{t=1}^T \beta_{a_t}(t) = \sum_{a\in\mathcal{A}} \sum_{t=1}^T \mathbf{1}\{a_t = a\}\beta_a(t)$. We start by bounding for an action $a \in \mathcal{A}$:

$$\sum_{t=1}^T \mathbf{1}\{a_t = a\}\beta_a(t) = \sqrt{2\log(1/\delta)} \sum_{t=1}^{T_a(T)} \frac{1}{\sqrt{t}} \leq 2\sqrt{2T_a(T)\log(1/\delta)}.$$

Since $\sum_{a\in\mathcal{A}} T_a(T) = T$ and by concavity of $\sqrt{\cdot}$, we have

$$\sum_{a\in\mathcal{A}} 2\sqrt{2T_a(T)\log(1/\delta)} \leq 2\sqrt{2TK\log(1/\delta)}.$$

Conditioning on the event $\mathcal{E} \cap \mathcal{E}_A$ whose probability satisfies $\mathbb{P}(\mathcal{E} \cap \mathcal{E}_A) \geq 1 - \epsilon$ yields the result. □

We can combine these two results into our main theorem:

Theorem 4 (Main Theorem)**.** *If $\epsilon \in (0,1)$, $\alpha_c = 1$ and $\alpha_r = \frac{2}{\tau - \bar{c}_1} + 1$, then with probability at least $1-\epsilon$, Algorithm 2 satisfies the following regret guarantee:*

$$\mathcal{R}_\Pi(T) \leq \left(\frac{2}{\tau - \bar{c}_1} + 1\right)\left(2\sqrt{2TK\log(4KT/\epsilon)} + 4\sqrt{T\log(2/\epsilon)\log(4KT/\epsilon)}\right)$$

Proof. This result is a direct consequence of Propositions 4 and 5 by setting $\delta = 4KT\epsilon$. □

C.5 Lower Bound

We start by proving a generalized version of the divergence decomposition lemma for bandits.

Lemma 8. *[Divergence decomposition for constrained multi armed bandits] Let $\nu = ((P_1, Q_1), \cdots, (P_K, Q_K))$ be the reward and constraint distributions associated with one instance of the single constraint multi-armed bandit, and let $\nu' = ((P'_1, Q'_1), \cdots, (P'_K, Q'_K))$ be the reward and constraint distributions associated with another constrained bandit instance. Fix some algorithm $\mathcal{A}$ and let $\mathbb{P}_\nu = \mathbb{P}_{\nu\mathcal{A}}$ and $\mathbb{P}_{\nu'} = \mathbb{P}_{\nu' A}$ be the probability measures on the cannonical bandit model (See section 4.6 of Lattimore and Szepesvári (2019)) induced by the T round interconnection of $\mathcal{A}$ and ν (respectively $\mathcal{A}$ and ν'). Then:*

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{a=1}^K \mathbb{E}_\nu[T_a(T)]\mathrm{KL}((P_a, Q_a), (P'_a, Q'_a))$$

Where $T_a(T)$ denotes the number of times arm a was pulled until by $\mathcal{A}$ and up to time T.

Proof. The same proof as in Lemma 15.1 from Lattimore and Szepesvári (2019) applies in this case. □

The following two lemmas will prove useful as well:

Lemma 9. *[Gaussian Divergence] The divergence between two multivariate normal distributions and means $\mu_1, \mu_2 \in \mathbb{R}^d$ with spherical identity covariance $\mathbb{I}_d$ equals:*

$$\mathrm{KL}(\mathcal{N}(\mu_1, \mathbb{I}_d), \mathcal{N}(\mu_2, \mathbb{I}_d)) = \frac{\|\mu_1 - \mu_2\|^2}{2}$$

[6] We use the following version of Azuma-Hoeffding: if X_n, $n \geq 1$ is a martingale such that $|X_i - X_{i-1}| \leq d_i$, for $1 \leq i \leq n$, then for every $n \geq 1$, we have $\mathbb{P}(X_n > r) \leq \exp\left(-\frac{r^2}{2\sum_{i=1}^n d_i^2}\right)$.

Define the binary relative entropy to be:

$$d(x,y) = x\log\left(\frac{x}{y}\right) + (1-x)\log\left(\frac{1-x}{1-y}\right)$$

and satisfies:

$$d(x,y) \geq (1/2)\log(1/4y) \tag{33}$$

for $x \in [1/2, 1]$ and $y \in (0,1)$. Adapted from Kaufmann et al. (2016), Lemma 1.

Lemma 10. *Let ν, ν' be two constrained bandit models with K arms. Borrow the setup, definitions and notations of Lemma 8, then for any measurable event $\mathcal{B} \in \mathcal{F}_T$:*

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{a=1}^{K} \mathbb{E}_\nu[T_a(T)]\mathrm{KL}((P_a, Q_a), (P'_a, Q'_a)) \geq d(\mathbb{P}_\nu(\mathcal{B}), \mathbb{P}_{\nu'}(\mathcal{B})) \tag{34}$$

We now present a worst-case lower bound for the constrained multi armed bandit problem. We restrict ourselves to Gaussian instances with mean reward and cost vectors $\bar{r}, \bar{c} \in [0,1]^K$. Let $\mathcal{A}$ be an algorithm for policy selection in the constrained MAB problem. For the purpose of this section we denote as $\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}, \bar{c})$ as the constrained regret of algorithm $\mathcal{A}$ in the Gaussian instance $\mathcal{N}(\bar{r}, \mathbb{I})$, $\mathcal{N}(\bar{c}, \mathbb{I})$. The following theorem holds:

Theorem 5. *Let $\tau, \bar{c}_1 \in (0,1)$, $K \geq 4$, and $B := \max\left(\frac{1}{27}\sqrt{(K-1)T}, \frac{1}{6(\tau-\bar{c}_1)^2}\right)$ and assume[7] $T \geq \max(K-1, 24eB)$ and let τ be the maximum allowed cost. Then for any algorithm $\mathcal{A}$ there is a pair of mean vectors $\bar{r}, \bar{c} \in [0,1]^K$ such that:*

$$\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}, \bar{c}) \geq B$$

Proof. If $\max\left(\frac{1}{27}\sqrt{(K-1)T}, \frac{1}{6(\tau-\bar{c}_1)^2}\right) = \sqrt{KT}$, then the argument in Theorem 15.2 of Lattimore and Szepesvári (2019) yields the desired result by noting that the framework of constrained bandits subsumes unconstrained multi armed bandits when all costs other than c_0 equal zero. In this case we conclude there is an instance $\bar{r}, \bar{c}$ with $\bar{c}_a = 0$ for all $a \in \mathcal{A}$ satisfying:

$$\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}, \bar{c}) \geq \frac{1}{27}\sqrt{(K-1)T}$$

Let's instead focus on the case where $B = \max\left(\frac{1}{27}\sqrt{(K-1)T}, \frac{1}{6(\tau-\bar{c}_1)^2}\right) = \frac{1}{6(\tau-\bar{c}_1)^2}$.

Pick any algorithm. We want to show that the algorithm's regret on some environment is as large as B. If there was an instance $\bar{r}, \bar{c}$ such that $\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}, \bar{c}) > B$ there would be nothing to be proven. Hence without loss of generality, we can assume that the algorithm satisfies $\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}, \bar{c}) \leq B$ for all $\bar{r}, \bar{c} \in [0,1]^K$ and having unit variance Gaussian rewards.

Let $c \in (0,1)$ with $c = \tau - \bar{c}_1$. For the reader's convenience we will use the notation $\Delta = 1/2$. By treating the rewards in a symbolic way it is easier to understand the logic of the proof argument. Let's consider the following constrained bandit instance inducing measure ν:

$$\begin{aligned}
\bar{c}^1 &= (\tau - c, && \tau + 2c, && \tau - c, && \tau + 2c, && \cdots, && \tau + 2c) \\
\bar{r}^1 &= (\Delta, && 8\Delta, && 0, && 4\Delta, && \cdots, && 4\Delta)
\end{aligned}$$

Notice that the optimal policy equals a mixture between arm 1 and 2, where arm 1 is chosen with probability 2/3 and arm 2 with probability 1/3. The value of this optimal policy equals $10/3\Delta$.

Recall we use the notation $\bar{T}_j(t)$ denote the total amount of probability mass that $\mathcal{A}$ allocated to arm j up to time t. Notice that the expected reward of all feasible policies that do not have arm 1 in their support have a gap (w.r.t the optimal feasible policy's expected reward) of at least $\frac{2\Delta}{3}$. Since by assumption, $\mathcal{A}$ satisfies $\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}^1, \bar{c}^1) \leq B$, we have

$$B \geq \mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}^1, \bar{c}^1) \geq \frac{2\Delta}{3}\left(\frac{2}{3}T - \frac{1}{2}T\right)\mathbb{P}\left(\bar{T}_1(T) < \frac{T}{2}\right) = \frac{\Delta}{9}T\mathbb{P}\left(\bar{T}_1(T) < \frac{T}{2}\right),$$

[7] This constraint on T translates to $T \geq C$ for some constant C.

and thus, we may write

$$\mathbb{P}\left(\bar{T}_1(T) \geq \frac{T}{2}\right) = 1 - \mathbb{P}\left(\bar{T}_1(T) < \frac{T}{2}\right) \geq 1 - \frac{9B}{\Delta T} \geq 1/2.$$

The last inequality follows from the assumption $T \geq \max(K-1, 24eB)$.

Let's now consider the following constrained bandit instance inducing measure ν':

$$\begin{aligned}
\bar{c}^2 &= (\tau - c, && \tau + 2c, && \tau - c, && \tau - c, && \cdots, && \tau + 2c) \\
\bar{r}^2 &= (\Delta, && 8\Delta, && 0, && 4\Delta, && \cdots, && 4\Delta)
\end{aligned}$$

In this instance the optimal policy is to play arm 4 deterministically, which gets a reward of 4Δ. Notice that the expected reward of any feasible policy that does not contain arm 4 in its support has a gap (w.r.t. the optimal feasible policy's expected reward) of at least $\frac{2\Delta}{3}$. Since by assumption, $\mathcal{A}$ satisfies $\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}^2, \bar{c}^2) \leq B$, we have

$$B \geq \mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}^2, \bar{c}^2) \geq \frac{2\Delta}{3}\left(\frac{1}{2}T\right)\mathbb{P}\left(\bar{T}_1(T) \geq \frac{T}{2}\right) = \frac{\Delta}{3}T\mathbb{P}\left(\bar{T}_1(T) \geq \frac{T}{2}\right),$$

and thus, we may write

$$\mathbb{P}\left(\bar{T}_1(T) \geq \frac{T}{2}\right) \leq \frac{3B}{\Delta T} \leq \frac{1}{4e}.$$

The last inequality follows from the assumption $T \geq \max(K-1, 24eB)$. As a consequence of inequality (33) and Lemmas 9 and 10, we have

$$\mathbb{E}_\nu[T_4(T)]\mathrm{KL}\left(\mathcal{N}\left(\begin{pmatrix}\tau + 2c \\ 4\Delta\end{pmatrix}, \mathbb{I}_d\right), \mathcal{N}\left(\begin{pmatrix}\tau - c \\ 4\Delta\end{pmatrix}, \mathbb{I}_d\right)\right) = \mathbb{E}_\nu[T_4(T)]2c^2 \geq \frac{1}{2},$$

and thus, we can conclude that

$$\mathbb{E}[\bar{T}_4(T)] = \mathbb{E}[T_4(T)] \geq \frac{1}{4c^2}. \tag{35}$$

Since in ν, any feasible policy with support in arm 4 and no support in arm 2 has a sub-optimality gap of $4/3\Delta$, we conclude the regret $\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}^2, \bar{c}^2)$ must satisfy:

$$\mathcal{R}_\Pi(T, \mathcal{A}, \bar{r}^2, \bar{c}^2) \geq \frac{\Delta}{3c^2}.$$

Since $\Delta = \frac{1}{2}$ and noting that in this case $\frac{\Delta}{3c^2} = B$. The result follows. □

C.6 Multiple Constraints

We consider the problem where the learner must satisfy M constraints with threshold values $\tau_1, \ldots, \tau_M$. Borrowing from the notation in the previous sections, we denote by as $\{\bar{r}_a\}_{a\in\mathcal{A}}$ the mean reward signals and $\{\bar{c}_a^{(i)}\}$ the mean cost signals for $i = 1, \ldots, M$. The full information optimal policy can be obtained by solving the following linear program:

$$\max_{\pi\in\Delta_K} \sum_{a\in\mathcal{A}} \pi_a \bar{r}_a, \qquad \text{s.t.} \quad \sum_{a\in\mathcal{A}} \pi_a \bar{c}_a^{(i)} \leq \tau_i, \quad \text{for} \quad i = 1, \ldots, M. \tag{P-M}$$

In order to ensure the learner's ability to produce a feasible policy at all times, we make the following assumption:

Assumption 6. *The learner has knowledge of $\bar{c}_1^{(i)} < \tau_i$ for all $i = 1, \ldots, M$.*

We denote by $\{\widehat{r}_a\}_{a\in\mathcal{A}}$ and $\{\widehat{c}_a^{(i)}\}_{a\in\mathcal{A}}$, for $i = 1, \ldots, M$ the empirical means of the reward and cost signals. We call $\{u_a^r(t)\}_{a\in\mathcal{A}}$ to the upper confidence bounds for our reward signal and $\{u_a^c(t,i)\}_{a\in\mathcal{A}}$, for $i = 1, \ldots, M$ the costs' upper confidence bounds:

$$u_a^r(t) = \widehat{r}_a(t) + \alpha_r\beta_a(t), \qquad u_a^c(t,i) = \widehat{c}_a^{(i)}(t) + \alpha_c\beta_a(t),$$

where $\beta_a(t) = \sqrt{2\log(1/\delta)/T_a(t)}$, $\delta \in (0,1)$ as before. A straightforward extension of Algorithm 2 considers instead the following $M-$constraints LP:

$$\max_{\pi \in \Delta_K} \sum_{a \in \mathcal{A}} \pi_a \; u_a^r(t), \qquad \text{s.t.} \quad \sum_{a \in \mathcal{A}} \pi_a \; u_a^c(t,i) \leq \tau_i, \quad \text{for} \;\; i = 1, \ldots, M. \tag{$\widehat{P-M}$}$$

We now generalize Lemma 6:

Lemma 11. *Let $\alpha_r, \alpha_c \geq 1$ satisfying $\alpha_c \leq \min_i(\tau_i - \bar{c}_1^{(i)})(\alpha_r - 1)$. Conditioning on $\mathcal{E}_a(t)$ ensures that with probability $1-\delta$:*

$$\mathbb{E}_{a \sim \pi_t}\left[u_a^r(t)\right] \geq \mathbb{E}_{a \sim \pi^*}\left[\bar{r}_a\right].$$

Proof. The same argument as in the proof of Lemma 6 follows through, the main ingredient is to realize that γ_t satisfies the sequence of inequalities in the lemma with $\tau - \bar{c}_1$ substituted by $\min \tau_i - \bar{c}_1^{(i)}$. □

The following result follows:

Theorem 6 (Multiple Constraints Main Theorem)**.** *If $\epsilon \in (0,1)$, $\alpha_c = 1$ and $\alpha_r = \frac{2}{\min_i \tau_i - \bar{c}_1^{(i)}} + 1$, then with probability at least $1-\epsilon$, Algorithm 2 satisfies the following regret guarantee:*

$$\mathcal{R}_{\Pi}(T) \leq \left(\frac{2}{\min_i \tau_i - \bar{c}_1^{(i)}} + 1\right)\left(2\sqrt{2TK\log(4KT/\epsilon)} + 4\sqrt{T\log(2/\epsilon)\log(4KT/\epsilon)}\right)$$

Proof. The proof follows the exact same argument we used for the proof of Theorem 3 substituting $\tau - \bar{c}_1$ by $\min_i \tau_i - \bar{c}_1^{(i)}$. □

D Extra Experiments

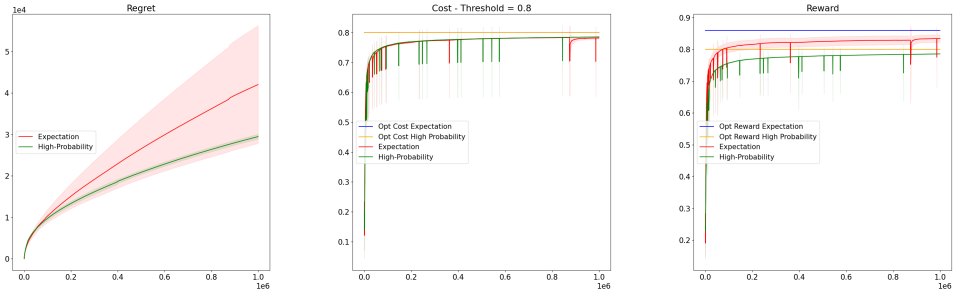

Figure 6: Constraint Threshold $\tau = 0.8$.

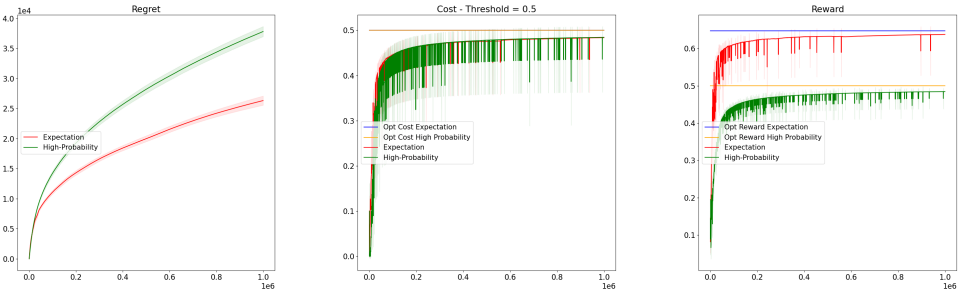

Figure 7: Constraint Threshold $\tau = 0.5$.

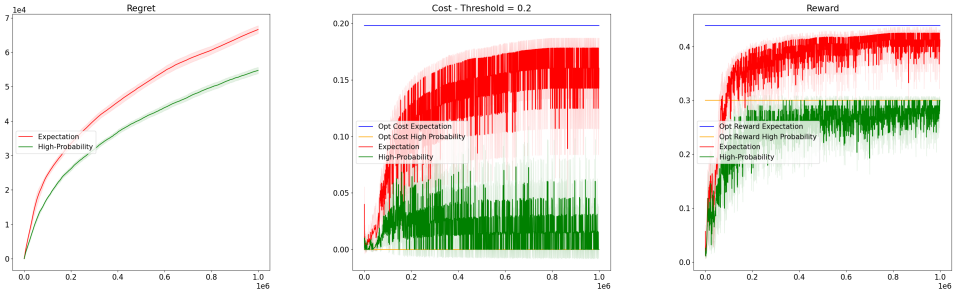

Figure 8: Constraint Threshold $\tau = 0.2$.

Regret *(left)*, cost *(middle)*, and reward *(right)* evolution of OPLB in a Linear Problem. The arms are identified with the rays corresponding to the standard basis vectors $[\mathbf{0}, e_1], [\mathbf{0}, e_2], [\mathbf{0}, e_3]$. The vector $\theta_\star = (1, .2, .3)$ and $\mu_\star = (1, 0, 0)$.

In figures figs. 6 to 8 we show the advantages of relaxing the objective to an expectation constraint. In this problem we let the action set be union of the rays $[\mathbf{0}, e_1], [\mathbf{0}, e_2]$ and $[\mathbf{0}, e_3]$ with reward and cost vectors equal to $\theta_\star = (1, .2, .3)$ and $\mu_\star = (1, 0, 0)$ and the safe action corresponding to the zero vector $\mathbf{0}$ and having 0 cost. In the following table we compare the optimal costs and reward profiles for the optimal policy satisfying the in-expectation constraint, vs the optimal policy satisfying the cost constraints with probability one for the different thresholds values $\tau = .2, .5, .8$. The optimal probability-one cost constrained policy always consists of playing a scaled version of e_i for $i \in \{1, 2, 3\}$. The optimal in-expectation cost constrained policy corresponds to a scaled point of the 3 dimensional simplex.

Threshold τ	Opt Cost Exp	Opt Cost High Prob	Opt Reward Exp	Opt Reward High Prob
0.8	0.8	0.8	0.86	0.8
0.5	0.5	0.5	0.65	0.5
0.2	0.2	0.0	0.44	0.3

We can observe that the optimal reward values can be substantially larger for the optimal in-expectation cost

constrained policy. Nevertheless, the regret curves of OPLB with an expectation cost constraint or a high probability cost constraint are comparable. This points to the fact that learning under a relaxed expectation cost constraint is not substantially harder than under a high probability cost constraint but can allow for much higher levels of accrued reward.

In order to run OPLB w.r.t a high probability cost constraint we modify the Algorithm 1 so that instead of constructing a feasible policy set Π_t as in Equation 13, we compute a safe action set $\tilde{\mathcal{A}}_t$ defined as:

$$\tilde{\mathcal{A}}_t = \{a \in \mathcal{A}_t : \tilde{c}_{a,t} \leq \tau\}.$$

The rest of the algorithm remains the same. In order to make the OPLB algorithm computationally feasible we notice that optimizing a constrained optimistic ellipsoidal reward objective over a ray $[\mathbf{0}, e_i]$ can be done in linear time. This is because the sets $\tilde{E}_i = \{a \in [\mathbf{0}, e_i] : \tilde{c}_{a,t} \leq \tau\}$ are also rays, and therefore the maximization problems $\max_{\theta \in \mathcal{C}_t^r(\alpha_r)} \max_{a \in \tilde{E}_i} \langle x, a \rangle$ are tractable. We approximate the OPLB constrained expectation objective by sampling 1000 uniform random points $\{p_i\}_{i=1}^{1000}$ from the simplex spanned by e_1, e_2, e_3 and adding the 1000 rays (with a start point at $\mathbf{0}$) to the action set yielding an enlarged action set $\{[\mathbf{0}, e_1], [\mathbf{0}, e_1], [\mathbf{0}, e_1]\} \cup \{[\mathbf{0}, p_i]\}_{i=0}^{1000}$. We optimize the high probability OPLB objective over this enlarged action set. In figures figs. 6 to 8 we run each experiment 10 times and report average curves with a shaded region corresponding to the ± 0.5 standard deviation around the regret, cost and reward values.