Supplementary material for "Qualitative Multi-Armed Bandits: A Quantile-Based Approach"

A. Analysis of QPAC

For the reader's convenience, we restate the theorem.

Theorem 3. *Assume that algorithm* QPAC *is run with parameters* (ϵ, δ, τ) *on a problem with* K *arms* $X_1, \ldots, X_K$*. Then, with probability at least* $1-\delta$*,* QPAC *outputs an* (ϵ, τ)*-optimal arm after drawing*

$$\mathcal{O}\left(\sum_{k=1}^{K} \frac{1}{(\epsilon \vee \Delta_k^\epsilon)^2} \log \frac{K}{(\epsilon \vee \Delta_k^\epsilon) \cdot \delta}\right)$$

samples. Consequently, QPAC *is an* (ϵ, τ, δ)*-quantile learner.*

Proof. Throughout the proof, assume that (3) holds for each $\widehat{Q}^{X_1}, \ldots, \widehat{Q}^{X_K}$, every $t = 1, 2, \ldots$, and every $0 \leq \tau \leq 1$, but with $c_t(\delta)$ replaced by $c_t(\delta/K)$. According to Proposition 1, this happens with probability at least $1-\delta$.

Consider some $k \in \mathcal{K}_{\epsilon,\tau}$. As it is (ϵ, τ)-optimal, we have the following:

$$\begin{aligned} \max_{h=1,\ldots,K} \widehat{Q}_t^{X_h}(\tau - c_t(\delta/K)) &\preceq \max_{h=1,\ldots,K} Q^{X_h}(\tau), && (8)\\ &\preceq Q^{X_k}(\tau+\epsilon) \\ &\preceq \widehat{Q}_t^{X_k}(\tau + \epsilon + c_t(\delta/K)) && (9) \end{aligned}$$

It follows that, with high probability, (ϵ, τ)-optimal arms never get discarded. Thus, with $\mathcal{A}_t$ denoting the set of arms in the t-th iteration of the while loop, it holds that

$$(\forall t \geq 1)\ \ \mathcal{K}_{\epsilon,\tau} \subseteq \mathcal{A}_t. \tag{10}$$

Now, let k be some non-(ϵ, τ)-optimal arm. According to our assumption, the following holds for any $t \geq 1$:

$$\widehat{Q}_t^{X_k}(\tau + \epsilon - c_t(\delta/K)) \preceq Q^{X_k}(\tau+\epsilon) \prec x^* \tag{11}$$

On the other hand, because of (10) and our assumption,

$$\begin{aligned} x^* = \max_{h \in \mathcal{K}_{\epsilon,\tau}} Q^{X_h}(\tau) &\preceq \max_{h \in A_t} Q^{X_h}(\tau) \\ &\preceq \max_{h \in A_t} \widehat{Q}_t^{X_h}(\tau + c_t(\delta/K)) && (12) \end{aligned}$$

for any m. It thus follows that, with high probability, a non-(ϵ, τ)-optimal arm is never selected to be $\widehat{k}$.

This proves the correctness of the algorithm.

Now, for a non-(ϵ, τ)-optimal arm k, define $t_k^* = \min\{t \geq 0 : 2c_t(\delta/K) \leq \Delta_k^\epsilon\}$. Then

$$\begin{aligned} \widehat{Q}_{t_k^*}^{X_k}\left(\tau + \epsilon + c_{t_k^*}\left(\tfrac{\delta}{K}\right)\right) &\preceq Q^{X_k}\left(\tau + \epsilon + 2c_{t_k^*}\left(\tfrac{\delta}{K}\right)\right) && (13)\\ &\preceq Q^{X_k}(\tau + \epsilon + \Delta_k^\epsilon) \\ &\prec \max_{h \in \mathcal{K}_{\epsilon,\tau}} Q^{X_h}(\tau - \Delta_k^\epsilon) \\ &\preceq \max_{h \in \mathcal{K}_{\epsilon,\tau}} Q^{X_h}\left(\tau - 2c_{t_k^*}\left(\tfrac{\delta}{K}\right)\right) \\ &\preceq \max_{h \in \mathcal{K}_{\epsilon,\tau}} \widehat{Q}_{t_k^*}^{X_h}\left(\tau - c_{t_k^*}\left(\tfrac{\delta}{K}\right)\right) && (14)\\ &\preceq \max_{h \in A_{t_k^*}} \widehat{Q}_{t_k^*}^{X_h}\left(\tau - c_{t_k^*}\left(\tfrac{\delta}{K}\right)\right) \end{aligned}$$

Thus, unless the algorithm terminates earlier, arm k is discarded at the latest in round t_k^*.

Finally, with $t_0^* = \min\{t \geq 0 : c_t(\delta/K) \leq \epsilon/2\}$ we obviously have

$$\max_{k \in A_{t_0^*}} \widehat{Q}^{X_k}_{t_0^*}\left(\tau + c_{t_0^*}\left(\tfrac{\delta}{K}\right)\right) \preceq \max_{k \in A_{t_0^*}} \widehat{Q}^{X_k}_{t_0^*}\left(\tau + \epsilon - c_{t_0^*}\left(\tfrac{\delta}{K}\right)\right) \quad .$$

This implies that the criterion for choosing $\widehat{k}$ (line 12 in Algorithm 1) is satisfied in round t_0^*, and thus the algorithm terminates at the latest in that round.

The sample complexity bound follows by noting that $\min(t_k^*, t_0^*) \leq \mathcal{O}\left(\frac{1}{(\epsilon \vee \Delta_k^\epsilon)^2} \log \frac{K}{(\epsilon \vee \Delta_k^\epsilon)\cdot\delta})\right)$. □

A.1. Lower bound

We start by invoking a lower bound result by (Mannor & Tsitsiklis, 2004) for the standard, value-based scenario. It considers the simplest setting: when the rewards come from Bernoulli distributions. This is equivalent to having K coins, and where the goal is to find the coin with the highest probability of head as the outcome of a coin flip.

More precisely, fix some $\epsilon' > 0$ and some $m_1, \ldots, m_K \in (0,1)$, denote the bias of the k-th coin by μ_k, and consider the following hypotheses:

$$H_0 : \quad \mu_k = m_k, \text{ for } k = 1, \ldots, K$$

and for $\ell = 1, \ldots, K$,

$$H_\ell : \quad \mu_k = m_k, \text{ for } k = 1, \ldots, \ell-1, \ell+1, \ldots, K, \quad \text{and} \quad \mu_\ell = m^* + \epsilon'$$

where $m^* = \max_{k'=1,\ldots,K} m_{k'}$, (Mannor & Tsitsiklis, 2004) show that it is not possible to distinguish with high certainty between these hypotheses based on only a few coin tosses. In particular, fixing some algorithm and denoting by I the index it recommends at the end of its run and by T the number of coin tosses it used, they show the following result (see Theorem 5 and its proof).

Theorem 4. (Mannor & Tsitsiklis, 2004) *Fix some $m_0 \in (0, 1/2)$. Then there exist $\delta_0 > 0$ and $c_1 > 0$ such that for every $\epsilon' \in (0, 1/2)$, every $\delta \in (0, \delta_0)$, and every $m_1, \ldots, m_K \in [0, 1/2]$, if some algorithm satisfies $\mathbf{P}[\mu_I \geq m^* - \epsilon' | H_0] \geq 1 - \delta$ and $\mathbf{P}[I = \ell | H_\ell] \geq 1 - \delta$ for every $\ell = 1, \ldots, K$, then*

$$\mathbf{E}[T|H_0] \geq c_1 \left(\tfrac{|S_1|}{(\epsilon')^2} + \sum_{k \in S_2} \tfrac{1}{(m^* - m_k)^2} \right) \log \tfrac{1}{8\delta} = c_1 \left(\sum_{k \in S_1 \cup S_2} \frac{1}{\left((m^* - m_k) \vee \epsilon'\right)^2} \right) \log \tfrac{1}{8\delta}$$

where

$$S_1 = \left\{ k : m^* > m_k > m^* - \epsilon', \text{ and } m_k > m_0, \text{ and } m_k \geq \tfrac{\epsilon' + m^*}{1 + \sqrt{1/2}} \right\}$$

and

$$S_2 = \left\{ k : m_k \leq m^* - \epsilon', \text{ and } m_k > m_0, \text{ and } m_k \geq \tfrac{\epsilon' + m^*}{1 + \sqrt{1/2}} \right\} \quad .$$

This can be used to derive the following lower bound result for the QMAB setting.

Proposition 2. *Fix some $m_0 \in (0, 1/2)$, and let $(L, \prec) = ([0,1], <)$. Then there exist $\delta_0 > 0$ and $c_1' > 0$ such that for every $\epsilon \in (0, 1/4)$, every $\delta \in (0, \delta_0)$, and every $m_1, \ldots, m_K \in [0, 1/2 - 2\epsilon]$, then every $(\epsilon, 3/4, \delta)$-quantile learner has expected sample complexity*

$$\mathbf{E}[T|H_0] \geq c_1' \left(\sum_{k \in S} \tfrac{1}{(\Delta_k^\epsilon \vee \epsilon)^2} \right) \log \tfrac{1}{8\delta}$$

where $S = \left\{ k : m_k > m_0, \text{ and } m_k \geq \tfrac{2\epsilon + m^*}{1 + \sqrt{1/2}} \right\}$.

Proof. Pick some $\epsilon' > 0$ and $m_1, \dots, m_K \in (0, 1/2 - \epsilon']$. Denote $m^* = \max_k m_k$ and assume for simplicity that $m^* = 1/2 - \epsilon'$. Consider the following hypotheses:

$$H'_0: \text{ For } k = 1, \dots, K, \quad \mathbf{P}[X_k = 1] = \tfrac{m_k}{2}, \text{ and } \mathbf{P}[X_k \le x] = \tfrac{1-m_k}{2} + \tfrac{x}{2} \text{ for } x \in [0, 1)$$

and for $\ell = 1, \dots, K$,

$$\begin{aligned} H'_\ell: \text{ For } k = 1, \dots, K, k \neq \ell, \quad & \mathbf{P}[X_k = 1] = \tfrac{m_k}{2}, \text{ and } \mathbf{P}[X_k \le x] = \tfrac{1-m_k}{2} + \tfrac{x}{2} \text{ for } x \in [0, 1) \\ & \text{and } \mathbf{P}[X_\ell = 1] = \tfrac{m_\ell + \epsilon'}{2}, \text{ and } \mathbf{P}[X_\ell \le x] = \tfrac{1-m_\ell-\epsilon'}{2} + \tfrac{x}{2} \text{ for } x \in [0, 1) \end{aligned}$$

This can be interpreted as the same coin tosses as in hypotheses $H_0, H_1, \dots, H_K$, with 1 playing the role of having a head, 0 playing the role of having a tail, and with the additional perturbation that with probability $1/2$ there is no return. This last scenario is represented by having the outcome $X_k \in (0, 1)$ as, indeed, this provides no useful information because, under any of the hypotheses, $\mathbf{P}[X_1 \in H] = \cdots = \mathbf{P}[X_K \in H]$ for any measurable $H \subseteq (0, 1)$. Consequently, distinguishing between hypotheses H'_ℓ and $H'_{\ell'}$ implies distinguishing between hypotheses H_ℓ and $H_{\ell'}$ for any $0 \le \ell < \ell' \le K$.

Set $\tau = 1 - (m^* + \epsilon')/2 = 3/4$ and $\epsilon = \epsilon'/2$. Then, for any $\ell = 0, 1, \dots, K$, an arm is (ϵ, τ)-optimal under hypothesis H'_ℓ iff it is ϵ-optimal under hypothesis H_ℓ. Indeed, in the H'_ℓ case for $\ell = 1, \dots, K$, $x^* = 1$ and the only (ϵ, τ)-optimal arm is ℓ. On the other hand, in the H'_0 case, $x^* = 1 - \epsilon'$ and an arm X_k is (ϵ, τ)-optimal iff $1 - \tau - \epsilon \le \mathbf{P}[X_k \succeq x^*] = 1 - (1 - m_k)/2 - x^*/2$. The latter is equivalent to $m^*/2 + \epsilon'/2 - \epsilon \le m_k/2 + \epsilon'/2$, that is, to $m^* \le m_k + \epsilon'$.

To determine Δ_k^ϵ note that, in the H_0 scenario, the definition $\Delta_k^\epsilon = \sup\{\Delta > 0 : Q^{X_k}(\tau + \epsilon + \Delta) < Q^{X_{k^*}}(\tau - \Delta)\}$, where k^* is such that $m_{k^*} = m^*$, implies $\tau + \epsilon + \Delta_k^\epsilon - \frac{1-m_k}{2} = \tau - \Delta_k^\epsilon - \frac{1-m^*}{2}$, and thus $\Delta_k^\epsilon = \frac{m^* - m_k}{4} - \epsilon/2 = \frac{m^* - m_k - \epsilon'}{4}$. It is easy to check that:

$$\Delta_k^\epsilon \vee \epsilon \;\ge\; \tfrac{(m^* - m_k) \vee \epsilon'}{6} \;.$$

The result now follows from Theorem 4. □

Remark 3. *One can derive similar bounds for finite L as well, however the analysis becomes more cumbersome.*

B. Analysis of QUCB

For the reader's convenience, we restate the results.

We start with the proof of Lemma 1.

Lemma 2 (Restatement of Lemma 1)**.** *If $\mathbf{P}[X_k \notin L_\tau] < \tau$ for some $1 \le k \le K$ then $(\inf L_\tau) \in L_\tau$, $\min_{k'} \mathbf{P}[X_{k'} \prec \inf L_\tau] < \tau$ and $\min_{k'} \mathbf{P}[X_{k'} \preceq \inf L_\tau] > \tau$. Additionally, $Q^{X_k}(\tau) = x^*$.*

Proof. By definition, if $x' \preceq x''$ for every $x'' \in L_\tau$, then $x' \preceq \inf L_\tau$. Thus, for every $x' \succ \inf L_\tau$, there must exist some $\tau' > \tau$ such that $x' \succ x^*(\tau')$, and so $F^{X_k}(x') = \mathbf{P}[X_k \preceq x'] \ge \mathbf{P}[X_k \prec x'] \ge \mathbf{P}[X_k \preceq x^*(\tau')] \ge \tau' > \tau$. Therefore, and because a CDF is right-continuous, $\mathbf{P}[X_k \notin (L_\tau \setminus \inf L_\tau)] = F^{X_k}(\inf L_\tau) = \inf_{x \succ \inf L_\tau} F^{X_k}(x) \ge \tau$. Thus $\mathbf{P}[X_k \notin L_\tau] < \tau$ implies $(\inf L_\tau) \in L_\tau$ and $\mathbf{P}[X_k \prec \inf L_\tau] = \mathbf{P}[X_k \notin \inf L_\tau] < \tau$. All this also implies $Q^{X_k}(\tau) = \inf L_\tau = x^*$.

Additionally, $(\inf L_\tau) \in L_\tau$ implies that $(\inf L_\tau) \succeq x^*(\tau_1)$ for some $\tau_1 > \tau$, which further implies that

$$\begin{aligned} \min_{k'} \mathbf{P}[X_{k'} \preceq \inf L_\tau] &\ge \min_{k'} \mathbf{P}[X_{k'} \preceq x^*(\tau_1)] \\ &= \min_{k'} \mathbf{P}[X_{k'} \preceq \max_{k''} Q^{X_{k''}}(\tau_1)] \\ &\ge \min_{k'} \mathbf{P}[X_{k'} \preceq Q^{X_{k'}}(\tau_1)] \\ &= \min_{k'} F^{X'_k}(Q^{X_{k'}}(\tau_1)) \\ &\ge \tau_1 \qquad (15) \\ &> \tau. \end{aligned}$$

where (15) holds because, as a CDF is right-continuous, $F^{X'_k}(Q^{X_{k'}}(\tau_1)) = \inf_{x \succ Q^{X_{k'}}(\tau_1)} F^{X'_k}(x) \ge \tau_1$. □

We continue with the proof of Theorem 2.

Theorem 5 (Restatement of Theorem 2)**.** *The expected cumulative regret of QUCB in round t is* $R_t = \mathcal{O}\left(\sum_{k:\Delta_k>0} \frac{\rho_k}{(\Delta_k)^2} \log t\right)$.

Proof. The structure of the proof follows closely the analysis of UCB1 (Auer et al., 2002).

First of all, similarly as in the proof of Proposition 1, for every $k = 1, \ldots, K$, every $m = 1, 2, \ldots$

$$\mathbf{P}\Big[(\widehat{Q}_m^{X_k}(\tau' - c) \succ Q^{X_k}(\tau')) \text{ or } (Q^{X_k}(\tau') \succ \widehat{Q}_m^{X_k}(\tau' + c)) \text{ for some } \tau' \in (0,1)\Big] \leq 2\exp(-2mc^2) \tag{16}$$

Additionally, (1) also implies that for every $k = 1, \ldots, K$ and every $m = 1, 2, \ldots$

$$\mathbf{P}[\|p^{X_k} - p_m^{X_k}\|_\infty > c] \leq 2\exp(-2mc^2) \tag{17}$$

Define for $k = 1, \ldots, K$

$$E_k(t, s, s_k) = \Big\{(\widehat{Q}_s^{X_{k^*}}(\tau + c(t,s)) \prec \widehat{Q}_{s_k}^{X_k}(\tau + c(t, s_k))) \bigvee \Big((\widehat{Q}_s^{X_{k^*}}(\tau + c(t,s)) = \widehat{Q}_{s_k}^{X_k}(\tau + c(t,s_k)) = \widehat{x}_t) \wedge (\widehat{p}_s^{X_{k^*}}(\widehat{x}_t) - c(t,s) \geq \widehat{p}_{s_k}^{X_k}(\widehat{x}_t) - c(t,s_k))\Big)\Big\}$$

Let ℓ be some positive integer specified later. Then

$$\begin{aligned}
T_t(k) &= 1 + \sum_{t'=K+1}^{t} \mathbb{I}\{k_{t'} = k\} \\
&\leq \ell + \sum_{t'=K+1}^{t} \mathbb{I}\{k_{t'} = k,\ T_{t'-1}(k) \geq \ell\} \\
&\leq \ell + \sum_{t'=K+1}^{t} \mathbb{I}\{E_k(t', T_{t'-1}(k^*), T_{t'-1}(k)),\ T_{t'-1}(k) \geq \ell\} \qquad (18) \\
&\leq \ell + \sum_{t'=K+1}^{t} \sum_{s=1}^{t-1} \sum_{s_k=\ell}^{t-1} \mathbb{I}\{E_k(t', s, s_k)\} \qquad (19)
\end{aligned}$$

where (18) is true because $\widehat{p}_{T_{t'}(k)}^{X_k}(\widehat{x}_{t'}) - c(t, T_{t'}(k)) > \tau \geq \min_{k'=1,\ldots,K}\left(\widehat{p}_{T_{t'}(k')}^{X_{k'}}(\widehat{x}_{t'}) - c(t', T_t(k'))\right)$ whenever $\widehat{Q}_{T_{t'}(k)}^{X_k}(\tau + c(t, T_{t'}(k))) \prec \widehat{x}_{t'}$.

Consider some arm X_k with $\mathbf{P}[X_k \notin L_\tau] > \tau$. Then

$$\begin{aligned}
\mathbb{I}\{E_k(t', s, s_k)\} \leq& \mathbb{I}\left\{\widehat{Q}_s^{X_{k^*}}(\tau + c(t', s)) \preceq \widehat{Q}_{s_k}^{X_k}(\tau + c(t', s_k))\right\} \\
\leq& \mathbb{I}\left\{\widehat{Q}_s^{X_{k^*}}(\tau + c(t', s)) \preceq Q^{X_k}(\tau + \Delta_k - c(t', s_k))\right\} \qquad (20) \\
&+ \mathbb{I}\left\{Q^{X_k}(\tau + \Delta_k - c(t', s_k)) \preceq \widehat{Q}_{s_k}^{X_k}(\tau + \Delta_k - 2c(t', s_k)))\right\} \qquad (21) \\
&+ \mathbb{I}\{\Delta_k \leq 3c(t', s_k)\} \qquad (22)
\end{aligned}$$

Note that (20) is upper bounded by $\mathbb{I}\left\{\widehat{Q}_s^{X_{k^*}}(\tau + c(t', s)) \prec Q^{X_{k^*}}(\tau)\right\}$. Furthermore, (16) entails high probability upper bound on this and (21), whereas (22) is 0 for s_k big enough to satisfy $\Delta_k > 3c(t', s_k)$. Thus, setting $\ell = 9 \cdot \frac{2}{(\Delta_k)^2} \ln(t-1)$ one obtains the following bound

$$\mathbf{E}[T_t(k)] \leq 9 \cdot \tfrac{2}{(\Delta_k)^2} \ln(t-1) + 2 \sum_{t'=K+1}^{t} \sum_{s=1}^{t-1} \sum_{s_k=\ell}^{t-1} (t')^{-4} \leq 9 \cdot \tfrac{2}{(\Delta_k)^2} \ln(t-1) + \pi^2/3.$$

Consider now some arm X_k with $\mathbf{P}[X_k \notin L_\tau] \leq \tau$. In case $\mathbf{P}[X_k \notin L_\tau] = \mathbf{P}[X_{k^*} \notin L_\tau]$, X_k is also optimal, it is thus only interesting to upper bound $T_k(t)$ in case $\rho_k = \mathbf{P}[X_k \notin L_\tau] - \mathbf{P}[X_{k^*} \notin L_\tau] > 0$. However, in that case $\mathbf{P}[X_{k^*} \notin L_\tau] < \tau$, and Lemma 1 applies, and so $\Delta_0 \triangleq \min_{k'} \mathbf{P}[X_{k'} \preceq \inf L_\tau] - \tau > 0$. Then,

$$
\begin{aligned}
&\mathbb{I}\{E_k(t,s,s_k)\} \\
\leq &\mathbb{I}\left\{\widehat{Q}_s^{X_{k^*}}(\tau + c(t,s)) \prec Q^{X_{k^*}}(\tau)\right\} && (23)\\
&+ \mathbb{I}\left\{Q^{X_k}(\tau) \prec \widehat{Q}_{s_k}^{X_k}(\tau + c(t,s_k))\right\} && (24)\\
&\mathbb{I}\left\{\left(\widehat{Q}_{s_k}^{X_k}(\tau + c(t,s_k)) = \widehat{Q}_s^{X_{k^*}}(\tau + c(t,s)) = x^*(\tau)\right) \wedge \left(\widehat{p}_{s_k}^{X_k}(x^*(\tau)) - c(t,s_k) \leq \widehat{p}_s^{X_{k^*}}(x^*(\tau)) - c(t,s)\right)\right\} && (25)\\
\leq &\mathbb{I}\left\{\widehat{Q}_s^{X_{k^*}}(\tau + c(t,s)) \prec Q^{X_{k^*}}(\tau)\right\} && (26)\\
&+ \mathbb{I}\left\{Q^{X_k}(\tau + \Delta_0 - c(t,s_k)) \prec \widehat{Q}_{s_k}^{X_k}(\tau + \Delta_0 - 2c(t,s_k))\right\} + \mathbb{I}\{\Delta_0 \leq 3c(t,s_k)\} && (27)\\
&+ \mathbb{I}\left\{\widehat{p}_{s_k}^{X_k}(x^*(\tau)) - c(t,s_k) \leq p^{X_k}(x^*(\tau)) - 2c(t,s_k)\right\} && (28)\\
&+ \mathbb{I}\left\{p^{X_k}(x^*(\tau)) - 2c(t,s_k) \leq p^{X_{k^*}}(x^*(\tau))\right\} && (29)\\
&+ \mathbb{I}\left\{p^{X_{k^*}}(x^*(\tau)) \leq \widehat{p}_s^{X_{k^*}}(x^*(\tau)) - c(t,s)\right\} && (30)
\end{aligned}
$$

In (23)-(25) we used that $Q^{X_k}(\tau) = Q^{X_{k^*}}(\tau) = x^*$ by Lemma 1. (27) follows because $Q^{X_k}(\tau) \succeq Q^{X_k}(\tau + \Delta_0 - c)$ for every $c > 0$ by the definition of Δ_0. The rest follows similarly as in the previous case: for (26), (28), (30), and the first term in (27) one can give high confidence upper bounds based on (16) and (B), whereas (29) and the second term in (27) is 0 for s_k big enough to satisfy $\rho_k \geq 2c(t,s_k)$ and $\Delta_0 \geq 3c(t,s_k)$ (by Lemma 1 again, $p^{X_{k^*}}(x^*(\tau)) = \mathbf{P}[X_{k^*} \notin L_\tau]$).

□

B.1. Lower bounds

The Δ_k parameters represent the hardness of distinguishing a non-optimal arm X_k from the optimal X_{k^*}. On the other hand, ρ_k represents the actual immediate expected regret. In the classical settings these two parameters coincide, but in the qualitative setting they are more separated. This is represented in the regret bound of QUCB and, as we show, it is also reflected in the lower bounds below.

B.1.1. $\Delta_k = \rho_k$ CASE

First we show lower bounds for some scenario when $\Delta_k = \rho_k$. Let $X_1, \ldots, X_K$ have Bernoulli distributions with parameters $m_1, \ldots, m_K \in (1/2, 3/4)$ respectively and set $\tau = 1/2$. Then $x^* = 1$, $L_\tau = \{x^*\}$, and for each $k = 1, \ldots, K$, $\mathbf{P}[X_k \notin L_\tau] = \mathbf{P}[X_k \neq 1] < \tau$, consequently $\Delta_0 = 1 - \tau = 1/2$ and $\Delta_k = \rho_k = \mathbf{P}[X_{k^*} = 1] - \mathbf{P}[X_k = 1] \leq 1/4$. Consequently, the qualitative setting coincides with the classical one in this case. Therefore, the expected cumulative regret is asymptotically $\Omega\left(\sum_{k:\rho_k>0} \frac{1}{\rho_k} \log t\right) = \Omega\left(\sum_{k:\rho_k>0} \frac{\rho_k}{(\Delta_k)^2} \log t\right)$.

B.1.2. $\Delta_k < \rho_k$ CASE

This is the case when $\mathbf{P}[X_k \notin L_\tau] > \tau$ or when $\mathbf{P}[X_k \notin L_\tau] \leq \tau$ but $\Delta_0 < \rho_k$. For this case we only show some significantly weaker results. Our analysis is based on Theorem 10 of (Mannor & Tsitsiklis, 2004), and considers only two-armed bandits taken from Example 3 (a) and (c).

Fix some $\Delta \in (0, 1/4)$, and consider the following two hypothesis

$$H_0: \quad \mathbf{P}[X_1 = x^1] = \mathbf{P}[X_1 = x^3] = 1/2 \quad \mathbf{P}[X_2 = x^2] = 1$$

and

$$H_1: \quad \mathbf{P}[X_1 = x^1] = 1/2 - \Delta,\ \mathbf{P}[X_1 = x^3] = 1/2 + \Delta, \quad \mathbf{P}[X_2 = x^2] = 1$$

Here we assume that $x^1 \prec x^2$ and $x^2 \prec x^3$. Then, if $\tau = 1/2 - \Delta/2$, then distinguishing between H_0 and H_1 resembles the situation when one had to distinguish between cases (a) and (c) in Example 3. In case of H_0, $x^* = x^2$, $k^* = 2$,

$L_\tau = \{x^2, x^3\}$, $\rho_1 = 1/2$, $\Delta_1 = \mathbf{P}[X_1 \notin L_\tau] = 1/2 - \tau = \Delta/2$, $\Delta_0 = \mathbf{P}[X_1 \preceq 1] - \tau = \Delta/2$. In case of H_1, $x^* = x^3$, $k^* = 1$, $L_\tau = \{x^3\}$, $\Delta_0 = 1 - \tau = 1/2$, $\rho_2 = 1/2$, $\Delta_2 = 1 - \tau = 1/2$.

Now, as in the proof of Theorem 10 in (Mannor & Tsitsiklis, 2004), if $\mathbf{P}[T_t(2) \geq t/2|H_0] < 3/4$, then $\mathbf{E}[R_t|H_0] \geq \rho_1 t/8 = t/16$, which is much more than the desired regret in case of H_0. Otherwise, as they show, by Lemma 4 in (Mannor & Tsitsiklis, 2004), $\mathbf{P}[T_t(2) \geq t/2|H_1] \geq \delta_1$, where δ_1 is the number satisfying $\mathbf{E}[T_t(1)|H_0] = \frac{1}{100\Delta^2} \log \frac{1}{4\delta_1}$. If now $\delta_1 \geq 1/\sqrt{t}$, then $\mathbf{E}[R_t|H_1] \geq t\rho_2/\sqrt{t} = \sqrt{t}/2$ which is, again, larger then desired. If, however, $\delta_1 < 1/\sqrt{t}$, then $\mathbf{E}[T_t(1)|H_0] \geq \frac{1}{200\Delta^2} \log \frac{t}{16}$, and thus

$$\mathbf{E}[R_t|H_0] \geq \frac{\rho_1}{200\Delta^2} \log \frac{t}{16} = \frac{1}{400\Delta^2} \log \frac{t}{16} \ .$$

B.2. Distribution independent analysis

Following the proof of Theorem 10 in (Mannor & Tsitsiklis, 2004) more closer then in Section B.1.2, one can show that $\max\left(\mathbf{E}[R_t|H_0], \mathbf{E}[R_t|H_1]\right) \geq \min\left(c_1 t, c_2 \frac{1}{\Delta^2} \log t\right)$. However, as $\Delta > 0$ can be arbitrarily small, this implies that no sublinear distribution independent upper bound exists. This is the consequence of the phenomenon that was discussed at the beginning of Section B.1.

C. Estimating quantiles using the Chernoff-Hoeffding bound

First, we derive the concentration bounds for the empirical estimate of the quantiles, based on the Chernoff-Hoeffding bounds.

Lemma 3. *For any random variable X over L, any $m \geq 1$ and any $\tau, c \in (0, 1)$,*

$$\mathbf{P}[Q^X(\tau) \prec \widehat{Q}_m^X(\tau - c)] \leq e^{-c^2 m/2} \tag{31}$$

and

$$\mathbf{P}[Q^X(\tau) \succ \widehat{Q}_m^X(\tau + c)] \leq e^{-c^2 m/2} \tag{32}$$

Proof. Let $x_0 = Q^X(\tau)$. Then, by definition, $\tau \leq F^X(x_0)$. Therefore, $F^X(x_0) \leq \widehat{F}_m^X(x_0) + c$ implies $x_0 \in \{x \in L : \tau \leq \widehat{F}_m^X(x) + c\}$, and thus

$$Q^X(\tau) = x_0 \succeq \inf\{x \in L : \tau \leq \widehat{F}_m^X(x) + c\} = \widehat{Q}_m^X(\tau - c)$$

Combining this with the Chernoff-Hoeffding bound $\mathbf{P}[F^X(x_0) > \widehat{F}_m^X(x_0) + c] \leq e^{-c^2 m/2}$ proves (31).

Showing (32) goes similarly, by switching the roles of Q^X and $\widehat{Q}_m^X$, and changing the parameters appropriately. □

Finally, note that this lemma can be directly applied in the proof of Theorem 1 to upper bound the probability that (8), (9), (11), (12), (13), (14) hold. Similarly, it can be directly applied in the proof of Theorem 2 to bound (26) and the first term in (27), whereas for (28), (30) one can directly apply the Chernoff-Hoeffding bound.