Qualitative Multi-Armed Bandits: A Quantile-Based Approach

Balázs Szörényi[1,5,6] SZORENYI@INF.U-SZEGED.HU
Róbert Busa-Fekete[2] BUSAROBI@UPB.DE
Paul Weng[3,4] PAWENG@CMU.EDU
Eyke Hüllermeier[2] EYKE@UPB.DE

[1]INRIA Lille - Nord Europe, SequeL project, 40 avenue Halley, 59650 Villeneuve d'Ascq, France
[2]Department of Computer Science, University of Paderborn, Warburger Str. 100, 33098 Paderborn, Germany
[3]SYSU-CMU Joint Institute of Engineering, 132 East Waihuan Road, Guangzhou, 510006, China
[4]SYSU-CMU Shunde International Joint Research Institute, 9 Eastern Nanguo Road, Shunde, 528300, China
[5]MTA-SZTE Research Group on Artificial Intelligence, Tisza Lajos krt. 103., H-6720 Szeged, Hungary
[6]Department of Electrical Engineering, The Technion - Israel Institute of Technology, Haifa, Israel 32000

Abstract

We formalize and study the multi-armed bandit (MAB) problem in a generalized stochastic setting, in which rewards are not assumed to be numerical. Instead, rewards are measured on a qualitative scale that allows for comparison but invalidates arithmetic operations such as averaging. Correspondingly, instead of characterizing an arm in terms of the mean of the underlying distribution, we opt for using a quantile of that distribution as a representative value. We address the problem of quantile-based online learning both for the case of a finite (pure exploration) and infinite time horizon (cumulative regret minimization). For both cases, we propose suitable algorithms and analyze their properties. These properties are also illustrated by means of first experimental studies.

1. Introduction

The multi-armed bandit (MAB) problem (or simply bandit problem) refers to an iterative decision making problem in which an agent repeatedly chooses among K options, metaphorically corresponding to pulling one of K arms of a bandit machine. In each round, the agent receives a random reward that depends on the arm being selected. The agent's goal is to optimize an evaluation metric, e.g., the *error rate* (expected percentage of playing a suboptimal arm) or the *cumulative regret* (difference between the sum of rewards obtained and the (expected) rewards that could have been obtained by selecting the best arm in each round).

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

In the *stochastic* multi-armed bandit setup, the distributions can vary with the arms but do not change with time. To achieve the desired goal, the agent has to tackle the classical exploration/exploitation dilemma: It has to properly balance the pulling of arms that were found to yield high rewards in earlier rounds and the selection of arms that have not yet been tested often enough (Auer et al., 2002; Cesa-Bianchi & Lugosi, 2006; Lai & Robbins, 1985).

MAB algorithms have not only been studied quite thoroughly from a theoretical point of view but have also been used in many real applications, such as medical treatment allocation design (Kuleshov & Precup, 2014), feature selection (Gaudel & Sebag, 2010; Busa-Fekete & Kégl, 2010) and crowdsourced labeling (Zhou et al., 2014). In many practical applications, however, numerical rewards are not provided in a natural way. Consider the example of clinical trials for testing pain medication. Here, the patients are asked to value their pain on a scale such as `no pain—mild—moderate—severe`, which is of qualitative nature. Computing averages on ordinal scales of that kind is clearly invalid and may lead to disputable conclusions.

In this paper, we therefore propose a setting in which the arm distributions are defined over a complete totally ordered set $(L, \preceq)$; the corresponding online learning framework will be introduced formally in Section 3, after reviewing related work in Section 2. The quality of an arm is expressed in terms of a τ-quantile of the arm's distribution over L. Thus, arms are compared in terms of their τ-quantiles, and an arm is considered to be *τ-optimal* if its τ-quantile coincides with the highest τ-quantile of all arms.

We consider two quantile-based learning frameworks that we refer to as the *finite* and the *infinite horizon* cases, respectively. The finite horizon case (Section 4) is formalized in the PAC framework: the goal of the learner is to find a

τ-optimal arm with probability at least $1 - \delta$. As opposed to this, the infinite horizon case (Section 5) is formalized as a regret minimization problem, in which the regret depends on τ and the quantile functions of the arms. The difficulty of both setups stems from the fact that, when for all τ-optimal arms, the probability of getting qualitative rewards lower or equal to the optimal τ-quantile x^* is close or equal to τ, it is hard (or impossible) to guess x^*, which is essential to decide whether an arm is optimal or not.

2. Related Work

Pure exploration algorithms for the stochastic bandit problem sample the arms a certain number of times (not necessarily known in advance) and then output a recommendation, such as the best arm or the m best arms (Bubeck et al., 2009; Even-Dar et al., 2002; Bubeck et al., 2013; Gabillon et al., 2011; Cappé et al., 2013; Kalyanakrishnan et al., 2012). Since our quantile-based learning task in the finite horizon case is formulated in a PAC setting, it can be viewed as a pure exploration strategy, too. Yet, we do not assume that *absolute numerical* feedback can be generated for individual arms; instead, our feedback is of *qualitative* nature. Therefore, since averaging rewards is no longer meaningful, the preference relation over the arms is defined based on τ-quantiles instead of mean values of the underlying distribution on rewards.

Yu & Nikolova (2013) introduce a pure exploration setting where, instead of the means, the goodness value or payoff of the arms is defined based on some notion of risk, such as the *value-at-risk* (Schachter, 1997), a famous risk measure used in finance, which is a particular case of quantiles. Their setup is similar to the best arm identification problem (Bubeck et al., 2009), where the goal of the learner is to control the so-called simple regret, which is the difference between the payoff of the best arm and the expected payoff obtained by its recommendation. The algorithm proposed by Yu & Nikolova (2013) is based on their result concerning the concentration property of the estimators of various risk measures—these properties are preconditioned on the assumption that the density functions of arms are continuously differentiable, and their derivatives are bounded from above. The proposed algorithm is computationally demanding since it solves a non-linear constrained and integer-valued optimization task in each round; moreover, their results regarding the performance of the algorithm assume that the densities are bounded away from zero everywhere. In addition to these limitations on the reward distributions, our learning setup also differs from theirs in that we assume a PAC setting with finite horizon, where the goal is to find a τ-optimal arm with high probability. Thus, since the error of the learner is controlled, the algorithms are evaluated in terms of their sample complexity (the number of samples taken prior to termination).

In the infinite case, the most commonly optimized property is the regret with respect to the maximum mean reward (Bubeck & Cesa-Bianchi, 2012). Nevertheless, alternative targets have already been considered, too, which led to interesting formal tasks. In a recent study, Carpentier & Valko (2014) formulate the regret in terms of the extreme value of the arm distributions. The goal here is to optimize the maximal regret observed. To this end, the learner intends to identify the most "abnormal" arm having the heaviest tail distribution, since the rewards received on a heavy-tailed arm are likely to deviate the most from its mean with highest probability. The learner is evaluated in terms of so-called extreme regret, which is the most extreme value found and compared to the most extreme value possible. The authors devise an algorithm, called EXTREMEHUNTER, based on the optimism in the face of uncertainty principle, which can achieve logarithmic expected regret in this setting. Sani et al. (2012) consider a MAB setting with a regret notion based on the principle of risk-aversion, where the risk is defined based on mean-variance risk. More specifically, there is a trade-off parameter that controls the influence of the mean and variance of the arm distributions. Thus, pulling an arm with high mean might result in a high regret if the variance of the rewards is high. The worst case regret of the proposed algorithm is $\mathcal{O}(KT^{2/3})$, and it is not clear whether it can be improved. As it is known that the worst case regret for the standard mean-based regret is $\mathcal{O}(\sqrt{KT})$, which is achieved by the MOSS algorithm by Audibert & Bubeck (2010), the optimization of regret based on risk aversion is conjectured to be a more complex problem.

In the preference-based bandit setup (Busa-Fekete & Hüllermeier, 2014), also known as duelling bandits (Yue et al., 2012), feedback about arms is not received in terms of absolute numerical rewards either. Instead, the learner is only allowed to compare the arms in a pairwise manner and receives binary feedback informing about the winner of the comparison. From this point of view, the feedback about the arms is even weaker than in our qualitative setting. Moreover, the notion of optimality of an arm can be defined in various ways in the preference-based setup. For example, a commonly used notion is that of a *Condorcet winner*, for which the probability of winning in a pairwise comparison is larger than $1/2$ against each other arm (Yue & Joachims, 2011; Zoghi et al., 2014).

3. Qualitative Multi-Armed Bandits

Formally, a standard *value-based* multi-armed or K-armed bandit problem is specified by real-valued random variables $X_1, \ldots, X_K$ associated, respectively, with K arms (that we simply identify by the numbers $1, \ldots, K$). In each

time step t, the online learner selects one or more arms (depending on the specific problem) and obtains a random sample of the corresponding distributions. These samples, which are called rewards, are assumed to be independent of all previous actions and rewards. The goal of the learner can be defined in different ways, such as maximizing the sum of rewards over time (Lai & Robbins, 1985; Auer et al., 2002) or identifying, with high probability, an arm the expected value of which differs by at most ϵ from the highest one (Even-Dar et al., 2002; Kalyanakrishnan et al., 2012).

In the *qualitative* multi-armed bandit (QMAB) problem, the rewards are not necessarily real-valued. Instead, the arms $X_1, \ldots, X_K$ are random variables over a complete totally ordered set[1] $(L, \preceq)$. Accordingly, when arm k is played in the t-th round, it yields a qualitative payoff $x \in L$. Independence is again assumed to hold across rounds. We will also use the reverse order $\succeq$ over L and the associated asymmetric (hence irreflexive) relations $\prec$ and $\succ$ of $\preceq$ and $\succeq$, respectively. For simplicity, we shall assume that L is a subset of the real numbers and $\preceq$ denotes the ordinary ranking over the reals. However, we shall not make use of the nominal values of the elements in L, only of their ordering.

3.1. Empirical CDF and Quantile Function

Let F^X denote the cumulative distribution function (CDF) of a random variable X. The *quantile function* $Q^X : [0,1] \to L$ of X is defined as

$$Q^X(\tau) = \inf\left\{x \in L : \tau \leq F^X(x)\right\} \quad .$$

We extend the domain of this function to the whole real line by defining $Q^X(\tau) = \inf L$ for $\tau < 0$ and $Q^X(\tau) = \sup L$ for $\tau > 1$.

As already mentioned, our aim is to compare arms in terms of their τ-quantiles, where $\tau \in [0,1]$ is a (user-specified) parameter of the problem—the concrete learning tasks will be detailed in Sections 4 and 5, respectively. As will be seen, the highest τ-quantile of the arms, $x^* = \max_{1 \leq k \leq K} Q^{X_k}(\tau)$, will play a central role in both cases. The difficulty of our quantile-based approach is due to the fact that x^* is unknown and, moreover, hard to guess.[2]

Denote the j-th sample of arm k by $X_{k,j}$. The empirical estimate of the CDF (or empirical CDF) of X_k based on $X_{k,1}, \ldots, X_{k,m}$ is

$$\widehat{F}_m^{X_k}(x) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}\left\{X_{k,j} \preceq x\right\} \quad ,$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. Denoting by $T_t(k)$ the number of times arm k has been pulled up to time t, the empirical CDF of X_k in round t is $\widehat{F}^{X_k}_{T_t(k)}(x)$.

The empirical estimator for the quantile function of arm k is based on the empirical distribution function:

$$\widehat{Q}_m^{X_k}(\tau) = \inf\left\{x \in L : \tau \leq \widehat{F}_m^{X_k}(x)\right\}$$

The accuracy of these empirical estimates can be quantified using a concentration result of Dvoretzky et al. (1956), which upper-bounds the tail distribution of the deviation of the empirical cumulative distribution function in supremum norm. Its improved version (Massart, 1990) (having optimal constants) can be formulated in our case as follows:[3]

$$\mathbf{P}\left(\|F^{X_k} - \widehat{F}_m^{X_k}\|_\infty > c\right) \leq 2\exp\left(-2mc^2\right) \quad , \qquad (1)$$

where $\|\cdot\|_\infty$ denotes the supremum norm.

For the sake of conciseness, we introduce an auxiliary function that determines the size of the confidence intervals:

$$c_m(\delta) = \sqrt{\frac{1}{2m} \log \frac{\pi^2 m^2}{3\delta}} \qquad (2)$$

Proposition 1. *Fix some $1 \leq k \leq K$ and $\delta \in (0,1)$. The following holds with probability at least $1-\delta$: For all $m \geq 1$ and for every $0 \leq \tau \leq 1$,*

$$\widehat{Q}_m^{X_k}(\tau - c_m(\delta)) \preceq Q^{X_k}(\tau) \preceq \widehat{Q}_m^{X_k}(\tau + c_m(\delta)) \quad (3)$$

Proof. To simplify notations, denote F^{X_k} by F and $\widehat{F}_m^{X_k}$ by $\widehat{F}_m$. Combining the bound (1) with the uniform bound and the Basel problem one obtains that, with probability at least $(1-\delta)$, $\|F - \widehat{F}_m\|_\infty \leq c_m(\delta)$ for all $m > 0$. In addition, $\|F - \widehat{F}_m\|_\infty \leq c_m(\delta)$ implies

$$\begin{aligned} Q(\tau) &= \inf\{x \in L : \tau \leq F(x)\} \\ &\preceq \inf\left\{x \in L : \tau \leq \widehat{F}_m(x) - c_m(\delta)\right\} \\ &= \widehat{Q}_m(\tau + c_m(\delta)) \end{aligned}$$

and

$$\begin{aligned} \widehat{Q}_m(\tau - c_m(\delta)) &= \inf\left\{x \in L : \tau \leq \widehat{F}_m(x) + c_m(\delta)\right\} \\ &\preceq \inf\{x \in L : \tau \leq F(x)\} \\ &= Q(\tau) \end{aligned}$$

□

[1] A totally ordered set is *complete* if every subset of it that has an upper bound also has a least upper bound.

[2] If x^* were known, the problem could be simplified to a standard value-based MAB with reward 1 in case the qualitative reward is at least as good as x^* and 0 otherwise.

[3] Each analysis in this paper also goes through using the Chernoff-Hoeffding bounds, essentially without any modification, at the cost of having slightly worse multiplicative constants (see Appendix C).

4. Finite Horizon: A PAC Algorithm

In this section, we consider the task of determining a "best" arm. In accordance with the goal highlighted in the introduction, the optimality of an arm is defined as follows.

Definition 1. *An arm k is said to be* τ-optimal *if* $Q^{X_k}(\tau) = x^*$*, where*

$$x^* = \max_{1 \leq k' \leq K} Q^{X_{k'}}(\tau) \ . \tag{4}$$

Throughout this section, let k^* denote the index of a τ-optimal arm. Requiring the learner to output such an arm might be hard or even impossible to achieve in cases where the probability of getting qualitative rewards lower or equal to the optimal τ-quantile is close or equal to τ for all τ-optimal arms. Therefore, in the spirit of the PAC bandit setting introduced by Even-Dar et al. (2002), we are going to tolerate some approximation error.

Definition 2. *An arm k is said to be* (ϵ, τ)-optimal *iff* $Q^{X_k}(\tau+\epsilon) \succeq x^*$*. Put in words, a slight "negative" perturbation on the distribution of an (ϵ, τ)-optimal arm yields a τ-quantile that is higher or equal to x^*.*

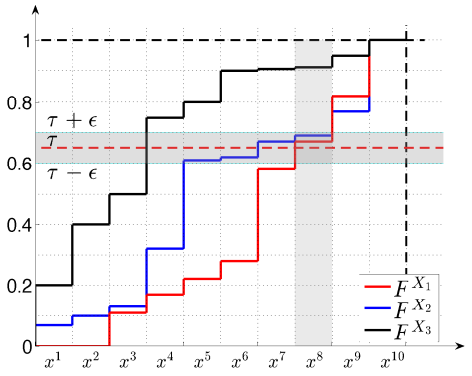


Figure 1. A qualitative MAB setup with three arms. The CDFs of the arms are plotted. The rewards come from $L = \{x^1, \ldots, x^{10}\}$, where $x^1 \prec x^2 \prec \cdots \prec x^{10}$.

Example 1. *To illustrate the notion of (ϵ, τ)-optimality, consider the following simple qualitative MAB problem with three arms and parameters $\tau = 0.65$, $\epsilon = 0.05$. Each arm is a random variable over ten possible qualitative rewards $x^1 \prec x^2 \prec \ldots \prec x^{10}$. Figure 1 depicts their cumulative distributions F^{X_1}, F^{X_2} and F^{X_3}. The τ-quantiles of arms 1, 2 and 3 are x^8, x^7 and x^4, respectively. The first arm (plotted in red) is τ-optimal, whence $k^* = 1$ and $x^* = x^8$. The second arm (plotted in blue) is not τ-optimal, since $F^{X_2}(x^7) > \tau$; yet, it is (ϵ, τ)-optimal since $F^{X_2}(x^7) - \epsilon < \tau$. The third arm (plotted in black) is not (ϵ, τ)-optimal, since the 0.7-quantile of X_3 is still given by $x^4 \prec x^8$.*

In practice, there may be several τ-optimal and several (ϵ, τ)-optimal arms. The goal of the learner is to identify one of them reliably.

Definition 3. *An online learning algorithm is called* (ϵ, τ, δ)-quantile learner *if it outputs an (ϵ, τ)-optimal arm with probability at least $1 - \delta$.*

4.1. The QPAC Algorithm

In Algorithm 1, we present our QPAC (Qualitative Probably Approximately Correct) algorithm, an adaptive elimination strategy inspired by Even-Dar et al. (2002). The algorithm computes lower and upper bounds x_t^- and x_t^+ of the optimal τ-quantile and exploits that (3) holds with high probability, which has several important consequences (as will be shown later in the analysis). One such consequence is that

$$x_t^- \preceq \widehat{Q}_t^{X_k}\left(\tau + \epsilon + c_t\left(\tfrac{\delta}{K}\right)\right) \tag{5}$$

for all (ϵ, τ)-optimal arms k, and thus every arm h with $\widehat{Q}_t^{X_h}\left(\tau + \epsilon + c_t\left(\tfrac{\delta}{K}\right)\right) \ . \prec x_t^-$ can be eliminated. Another important consequence is that an arm k is necessarily (ϵ, τ)-optimal if

$$x_t^+ \preceq \widehat{Q}_t^{X_k}\left(\tau + \epsilon - c_t\left(\tfrac{\delta}{K}\right)\right) \tag{6}$$

The rest will be detailed in the analysis.

Algorithm 1 QPAC(δ, ϵ, τ)

```
1: Set A = {1, ..., K}                                   ▷ Active arms
2: t = 1
3: while A ≠ ∅ do
4:     for k ∈ A do
5:         Pull arm k and observe X_{k,t}
6:     x_t^+ = max_{k∈A} Q̂_t^{X_k}(τ + c_t(δ/K))
7:     x_t^- = max_{k∈A} Q̂_t^{X_k}(τ − c_t(δ/K))
8:     for k ∈ A do
9:         if Q̂_t^{X_k}(τ + ε + c_t(δ/K)) ≺ x_t^- then
10:            A = A \ {k}                    ▷ Discard k based on (5)
11:        if x_t^+ ⪯ Q̂_t^{X_k}(τ + ε − c_t(δ/K)) then
12:            k̂ = k                          ▷ Select k according to (6)
13:            BREAK
14:    t = t + 1
15: return k̂
```

Let us illustrate the algorithm on Example 1.

Example 2. *(Example 1 continued) The non-(ϵ, τ)-optimal arm 3 cannot be eliminated by* QPAC *unless $x_t^- \succ x^4$. This happens when $\widehat{Q}_t^{X_k}(\tau - c_m(\frac{\delta}{K})) \succ x^4$ for arm 1 or arm 2 (see line 7 of Algorithm 1). Therefore, x_t^- needs to be high enough to eliminate a non-(ϵ, τ)-optimal arm. Moreover, for eliminating the third arm (see line 10), we need $\widehat{Q}_t^{X_3}(\tau+\epsilon+c_t(\frac{\delta}{K})) \preceq x^4$ to hold, i.e., $\widehat{F}_t^{X_3}(x^4) - c_t(\frac{\delta}{K}) -$*

$\epsilon > \tau$. Therefore, for eliminating a non-(ϵ, τ)-optimal arm, the estimate of its CDF needs to be tight enough as well. The selection mechanism of QPAC *is based on a very similar argument as the one described above for elimination.*

4.2. Analysis

The sample complexity of the qualitative PAC setting is very similar to the one of the value-based setting (see Even-Dar et al. (2002)). Before discussing the result in more detail, some further notation needs to be introduced.

First of all, denote the set of (ϵ, τ)-optimal arms by $\mathcal{K}_{\epsilon,\tau}$, and define

$$\Delta_k^\epsilon = \sup\left\{\Delta \in [0,1] \,\middle|\, Q^{X_k}(\tau + \epsilon + \Delta) \prec \max_{h \in \mathcal{K}_{\epsilon,\tau}} Q^{X_h}(\tau - \Delta)\right\}$$

for $k = 1, \dots, K$, where $\sup \emptyset = 0$ by definition. Finally, let $\vee$ denote the max operator.

Theorem 1. *Assume that algorithm* QPAC *is run with parameters (ϵ, δ, τ) on a problem with K arms $X_1, \dots, X_K$. 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. Thus, QPAC *is an (ϵ, τ, δ)-quantile learner.*

The proof of Theorem 1 is deferred to Appendix A.

Remark 1. *Note that the sample complexity of* QPAC *depends on the number of arms, K, but not on the number of different rewards (i.e., size of L).*

Remark 2. *Lower bound on sample complexity for value-based PAC bandits had already been investigated by Mannor & Tsitsiklis (2004). A similar lower bound analysis also applies to the qualitative setting resulting in a lower bound of the form $\Omega(\sum_{k=1}^{K} \frac{1}{(\epsilon \vee \Delta_k^\epsilon)^2} \log \frac{1}{\delta})$ (see Appendix A.1). Therefore this lower bound shows that the sample complexity of the* QPAC *algorithm given in Theorem 1 is optimal up to a logarithmic factor.*

5. Infinite Horizon

In this section, we analyze the infinite horizon setting, where the goal of the online learner is normally defined as minimizing the cumulative regret of its actions in the course of time. First of all, this of course presupposes an appropriate definition of the notion of *regret*. Preferably, in order to allow for a simple accumulation, regret should be defined in a quantitative way.

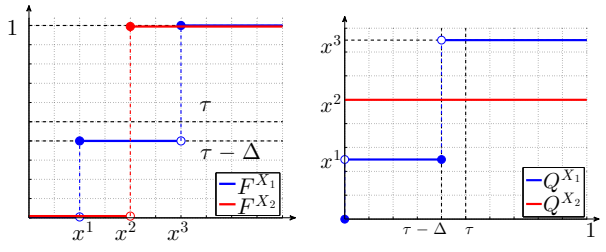
In the standard value-based setting, the regret of choosing an arm is typically defined in terms of the difference $x^* - x$ between the reward observed, x, and the reward that would have been obtained (in expectation) by choosing the best arm, namely the arm with the highest expected reward x^*. In our setting, a quantification of regret in terms of differences is no longer possible, however, since arithmetic operations are not defined on our qualitative scale L. Instead, as explained before, we are only allowed to compare outcomes in terms of "better" or "worse". This leads us quite naturally to a binary regret function $\mathrm{regret}(x, y) = \mathbb{I}\{x \in G\} - \mathbb{I}\{y \in G\}$ for obtaining reward y instead of x, where G is the subset of outcomes in L considered to be "good". Accordingly, the (expected) immediate regret of choosing arm X_k is

$$\begin{aligned}\rho_k &= \max_{k'=1,\dots,K} \mathbf{E}[\mathrm{regret}(X'_k, X_k)] \\ &= \max_{k'=1,\dots,K} \mathbf{P}[X_{k'} \in G] - \mathbf{P}[X_k \in G] \qquad (7)\end{aligned}$$

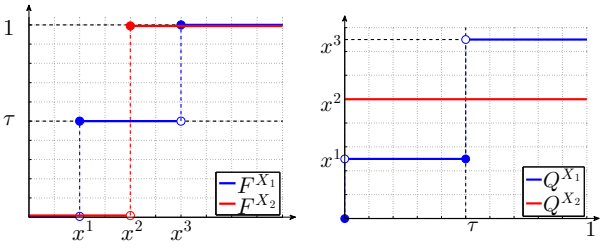
Now, the above definition of regret raises another question: What is the set G of good outcomes? In our setting, a natural answer is to consider an outcome x as good if $x \succeq x^*$, i.e., if x is at least as good as the optimal τ-quantile (4). However, in conjunction with the sharp discontinuity of the regret function (7), this definition renders regret minimization a truly hard problem. In fact, as shown by the following example, no algorithm with sublinear distribution independent regret guarantees exists. A more formal explanation of the linear worst case regret is deferred to the supplementary material (see Appendix B.2). Our worst case analysis is based on the fact that bandit instances given in Example 3 (a) and (b) are hard to distinguish.

Example 3. *Consider the qualitative MAB settings illustrated in Figure 2. In case (a), it is clear that x^3 should be considered as the only "good" reward, and thus $x^* = x^3$ and $G = \{x^3\}$. The second arm thus never returns good rewards, whereas the first arm does with probability $1-\tau+\delta$. Therefore, the regret of arm 2 is $1-\tau+\delta$. On the other hand, in case (c) both x^2 and x^3 should be considered good, so $x^* = x^2$ and $G = \{x^2, x^3\}$. Thus, while arm 2 returns a good reward consistently, the first arm is doing so only with probability $1 - \tau - \delta$. The regret of the first arm is $\tau + \delta$.*

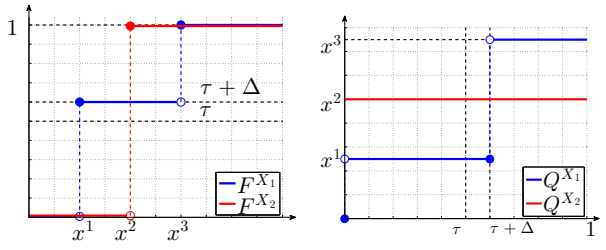
As long as one cannot distinguish between cases (a) and (c) with high enough probability, the choice of which one to optimize for (which is crucial, as arm 2 has at least constant regret in (a), and the same holds for arm 1 in (c)) will remain random. (The problem of the learner, therefore, is to find out whether $\mathbf{P}[X_1 = x^1] \geq \tau$ or $\mathbf{P}[X_1 = x^1] < \tau$ for the first arm.) Thus, until that point is reached, any learner necessarily incurs linear regret in at least one of these examples. Additionally, to distinguish between the examples is getting more and more difficult as δ approaches 0 (for formal results see Appendix B.1 and B.2). This suggests that one cannot hope for sublinear worst case regret bounds.



(a) CDF and Q function for synthetic problem



(b) CDF and Q function for synthetic problem



(c) CDF and Q function for synthetic problem

Figure 2. Synthetic qualitative MAB tasks with two arms.

Example 4 (Examples in Figure 2 continued)**.** *Now, consider cases (a) and (b), and the τ-quantile x^*. In case (b), $x^* = x^2$, thus $\mathbf{P}[X_1 \succeq x^*] = 1 - \tau$ and $\mathbf{P}[X_2 \succeq x^*] = 1$ while in case (a) (see Example 3), $\mathbf{P}[X_1 \succeq x^*] = 1 - \tau + \delta$ and $\mathbf{P}[X_2 \succeq x^*] = 0$. However, in order to distinguish the two cases, the learner needs to pull arm 1, leading to some non-negligible regret. This regret in (b), however, cannot be explained by any natural parameter (like the difference of the means in the quantitative case).*

In order to avoid the problem in Example 4, we propose a slightly different definition of the set G of good outcomes. To this end, let

$$x^*(\tau') = \max_{k=1,\dots,K} Q^{X_k}(\tau')$$

for $\tau' \in [0, 1]$ (thus $x^* = x^*(\tau)$), and define

$$G = L_\tau = \{x \in L : x \succeq x^*(\tau') \text{ for some } \tau' > \tau\} \ .$$

Correspondingly, the best arm with the minimal expected regret is defined by

$$k^* = \operatorname*{argmax}_{1 \le k \le K} \mathbf{P}[X_k \in L_\tau] \ ,$$

Algorithm 2 QUCB(τ)

for rounds $t = 1, \dots, K$ **do**
  set $k_t = t$ and $T(k_t) = 1$
  pull arm k_t and observe sample $X_{k_t,1}$
for rounds $t = K+1, K+2, \dots$ **do**
  $\widehat{x}_t = \sup_{k=1,\dots,K} \widehat{Q}^{X_k}_{T(k)}(\tau + c(t, T(k))$
  $k_t := \operatorname{argmin}_{k=1,\dots,K} \left(\widehat{p}^{X_k}_{T(k)}(\widehat{x}_t) - c(t, T(k)) \right)$
  set $T(k_t) = T(k_t) + 1$
  pull arm k_t, and observe sample $X_{k_t,T(k_t)}$

the (expected) immediate regret of arm k is $\rho_k = \mathbf{P}[X_{k^*} \in L_\tau] - \mathbf{P}[X_k \in L_\tau]$, and $R_t = t\mathbf{P}[X_{k^*} \in L_\tau] - \mathbf{E}[\sum_{t'=1}^{t} \mathbb{I}\{X_{k_{t'}} \in L_\tau\}]$ is the expected cumulative regret, where $k_{t'}$ is the index of the arm chosen in round t'. In our example, this approach renders the first arm optimal in both (a) and (b), since in both cases $L_\tau = x^3$. Note also that in case of a "clear" separation (i.e., when $x^*(\tau) = x^*(\tau+\epsilon)$ for some $\epsilon > 0$) this regret is equivalent to the one based on $\mathbf{P}[X_k \succeq x^*]$.

5.1. Algorithm QUCB

In Algorithm 2, we present the pseudocode of our QUCB (which is short for Qualitative Upper Confidence Bound) algorithm. In each round t, it uses an estimate $\widehat{Q}^{X_k}_{T_t(k)}$ of the τ-quantiles and pulls the arm which maximizes this estimate. The confidence term used is $c(t, s) = \sqrt{2\frac{\ln(t-1)}{s}}$. (The algorithm also needs to break ties, which is carried out in a similar fashion, but using the estimates of the p functions described later.) As a result, the accuracy of the estimate for the most promising arm will be increased. Suboptimal arms will be revealed as such as soon as the accuracy of the corresponding estimate is high enough—mathematically, this can be guaranteed thanks to the right-continuity of the quantile functions.

For selecting the most promising arms between those with $\mathbf{P}[X_k \notin L_\tau] < \tau$, the algorithm additionally keeps track of an estimate of another function, $p^{X_k}(x) = \mathbf{P}[X_k \prec x]$, using $\widehat{p}^{X_k}_m(x) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{I}\{X_{k,j} \prec x\}$. (Thus $\widehat{F}^{X_k}_m(x) = \widehat{p}^{X_k}_m(x)$ for $x \in L \setminus \{X_{k,1}, \dots, X_{k,m}\}$.)

In order to state the regret bound of QUCB, some further notation needs to be introduced. But first we need a technical lemma (see Appendix B for the proof).

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

For arm k with $\mathbf{P}[X_k \notin L_\tau] > \tau$, define $\Delta_k = \mathbf{P}[X_k \notin L_\tau] - \tau$. Now, consider some arm k with $\mathbf{P}[X_k \notin L_\tau] \le \tau$. In case $\mathbf{P}[X_k \notin L_\tau] \le \tau$, X_k is also optimal, it is thus only interesting to upper bound $T_k(t)$ in case $\rho_k = \mathbf{P}[X_k \notin$

$L_\tau] - \mathbf{P}[X_{k^*} \notin L_\tau] > 0$. In that case, $\mathbf{P}[X_{k^*} \notin L_\tau] < \tau$ and Lemma 1 applies. Therefore, $\Delta_0 = \min_{k'} \mathbf{P}[X_{k'} \preceq \inf L_\tau] - \tau > 0$. Based on these, for X_k satisfying $\mathbf{P}[X_k \notin L_\tau] \leq \tau$, define $\Delta_k = \min(\rho_k, \Delta_0)$. Then, we have (see Appendix B for the proof):

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

For regret lower bounds see Appendix B.1 and B.2.

6. Experiments

6.1. Finite Horizon

The goal of the first experiment is to assess the impact of the parameters τ and ϵ on the sample complexity, that is, the number of samples taken by the algorithm prior to termination. We generated random bandit instances for which rewards are taken from a totally ordered discrete set $\{x^1, \ldots, x^{10}\}$. In other words, the arm distributions are multinomial distributions. The parameters of the distributions are drawn uniformly at random from $(0, 1)$ and proportionally scaled so as to sum up to one. The sample complexities for various values of parameters ϵ and τ are shown in Figure 3. As can be seen, the smaller ϵ, the higher the sample complexity—thus, our algorithm scales gracefully with the approximation error allowed. The second observation is that the parameter τ has only a weak influence on the sample complexity. This can be explained by the fact that our confidence intervals are derived for the empirical CDF as a whole, without being tailored for a specific quantile.

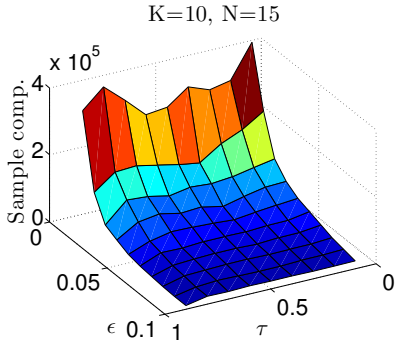


Figure 3. The sample complexity of QPAC for $K = 10$ and different values of the parameter τ and ϵ. The arm distributions are categorical distributions. The results are averaged over 100 repetitions. The confidence parameter δ was set to 0.05 for each run; accordingly, the average accuracy was significantly above $1 - \delta = 0.95$ in each case.

In the second experiment, we compare the performance of the QPAC algorithm with a standard PAC bandit algorithm on bandit problems for which the (ϵ, τ)-optimal arm coincides with the ϵ-best arm in terms of means, thereby assuring that both learners are seeking the same arm. As a baseline, we run the SUCCESSIVEELIMINATION learner (SE) by Even-Dar et al. (2002), which is an (ϵ, δ)-PAC learner (i.e., it returns an ϵ-optimal arm with probability at least $1 - \delta$). To guarantee a fair comparison, the confidence interval defined in (2) is used in our implementation of SE, which differs only in constants from the one of Even-Dar et al. (2002). We tested the algorithm on the "Needle in the Haystack" (NHS) problem, which consists of arms obeying Bernoulli distribution with parameter p—except for one of them, the target, which has a slightly higher parameter $p + p'$. Note that for $\tau = (1 - p) - p'/2$, we have $Q^{X_i}(\tau) = 1$ for the single τ-optimal arm, and $Q^{X_{i'}}(\tau) = 0$ otherwise. We run the experiments with $\tau = 0.1, \ldots, 0.9$ and $p' = 0.1$; correspondingly, p was set to $0.85, 0.75, \ldots, 0.05$, respectively. The approximation error ϵ was set to 0.01, 0.03 and 0.05. As can be seen from the results in Figure 4, QPAC slightly outperforms SE in terms of sample complexity for smaller values of ϵ. This can be explained by the fact that, although both algorithms are using similar elimination strategies, the statistics they use are of different nature.

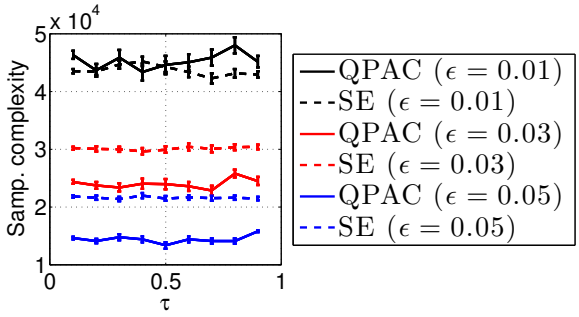
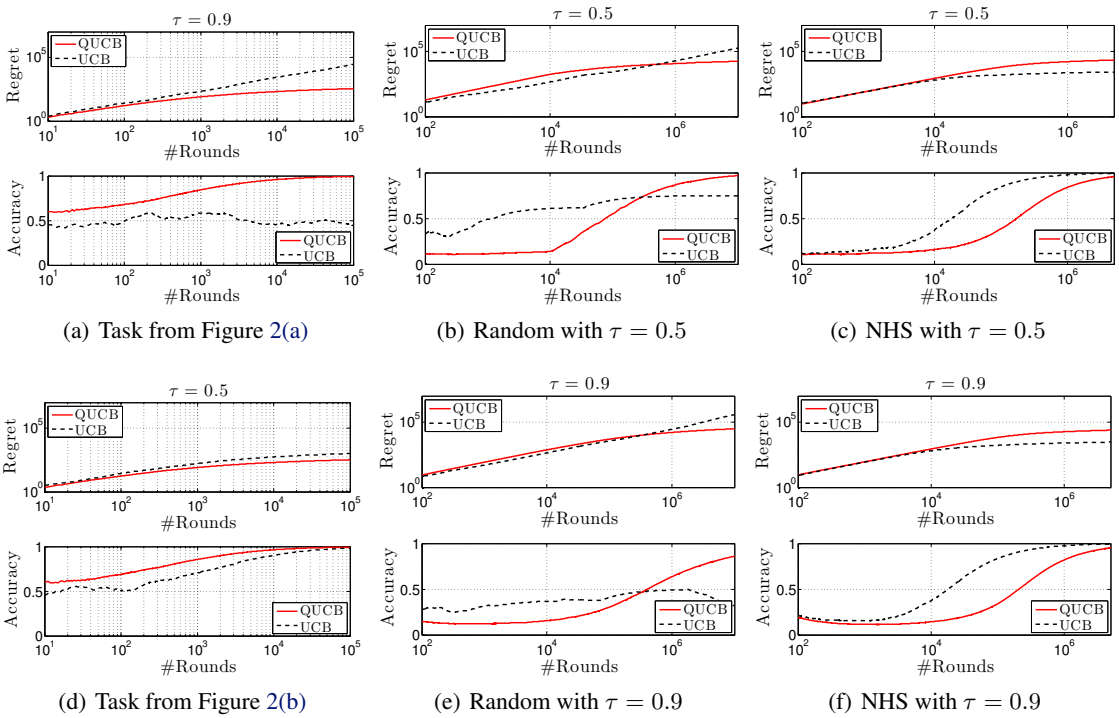


Figure 4. The sample complexity of SE and QPAC for the NHS problem with various parameter setting. The number K of arms was set to 15. The results are averaged over 100 repetitions. The confidence parameter δ was set to 0.05 for each run; accordingly, the average accuracy was significantly above $1 - \delta = 0.95$ in each case.

6.2. Infinite Horizon

In this section, we evaluate the QUCB algorithm on several numerical test cases. As a baseline, we run the standard UCB algorithm (Auer et al., 2002), which maximizes the sum of rewards. In each case, we plot the quantile-based cumulative regret and the average accuracy of the algorithms versus the number of iterations. By definition, the accuracy of an algorithm is 1, if it selects an arm from $\mathcal{K}^*$, and 0 otherwise. We run the algorithms on the following bandit problems:

1. Bandit instances defined in Figure 2(a) and 2(b). The



(a) Task from Figure 2(a) (b) Random with $\tau = 0.5$ (c) NHS with $\tau = 0.5$

(d) Task from Figure 2(b) (e) Random with $\tau = 0.9$ (f) NHS with $\tau = 0.9$

Figure 5. Cumulative regret and accuracy for various test cases.

results are shown in Figure 5(a) and 5(d), respectively. We set $x^1 = 1$, $x^2 = 2$ and $x^3 = 3$ in the case of UCB. The parameter δ was set to 0.1.

2. Multinomial arm distributions as described in the previous section, with parameters drawn uniformly at random. For the quantiles, we used $\tau \in \{0.5, 0.9\}$. The results are shown in Figure 5(b) and 5(e), respectively.

3. NHS problem with parameters $p = 0.45, p' = 0.1, \tau = 0.5$ and $p = 0.85, p' = 0.1, \tau = 0.9$. The results are shown in Figure 5(c) and 5(f).

In the first test case described in Figure 2(a), the mean of both arm distributions is $1/2$. Therefore, since UCB cannot distinguish the arms, its accuracy is fluctuating around $1/2$. As opposed to this, QUCB is able to identify the optimal arm. In the second test case defined in Figure 2(b), the best option is the second arm, both in terms of mean and τ-quantile. Accordingly, QUCB and UCB are both able the identify the optimal arm. On multinomial arm distributions, QUCB significantly outperforms the UCB algorithm. This can be explained by the fact that the median ($\tau = 1/2$) does not necessarily coincide with the mean—the higher τ, the more different the goals of the learners will actually become. As expected, the performance of both algorithms is on par in the case of the NHS problem.

7. Conclusion

We have investigated the setting of quantile-based online bandit learning in the qualitative case, that is, when rewards are coming from a complete totally ordered set but are not necessarily numerical. We introduced and analyzed a PAC algorithm in the finite horizon setting. Moreover, for the infinite horizon setting, we proposed an algorithm the (distribution-dependent) expected regret of which is growing logarithmically with time.

We have showed that sublinear regret in the qualitative setting is not achievable in the worst case (without using properties of the underlying distributions) in general. Since the standard reward expectation maximization problem has a known lower-bound of $\Omega(1/\sqrt{T})$ (Audibert & Bubeck, 2010) and the risk-aversion setup has $\Omega(T^{2/3})$ (Sani et al., 2012), therefore our worst case result implies that minimizing the quantile-based regret in the qualitative setting is intrinsically more difficult than the standard value-based and the risk-aversion bandit problems.

Acknowledgments

This work was supported by the European Community's 7th Framework Programme (FP7/2007-2013) under grant agreement n° 270327 (project CompLACS) and by the German Research Foundation under grant HU 1284/8-1.

References

Audibert, Jean-Yves and Bubeck, Sébastien. Regret bounds and minimax policies under partial monitoring. *J. Mach. Learn. Res.*, 11:2785–2836, 2010.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5 (1):1–122, 2012.

Bubeck, S., Munos, R., and Stoltz, G. Pure exploration in multi-armed bandits problems. In *Proceedings of the 20th ALT*, ALT'09, pp. 23–37, 2009.

Bubeck, S., Wang, T., and Viswanathan, N. Multiple identifications in multi-armed bandits. In *Proceedings of The 30th ICML*, pp. 258–265, 2013.

Busa-Fekete, R. and Hüllermeier, E. A survey of preference-based online learning with bandit algorithms. In *Algorithmic Learning Theory (ALT)*, volume 8776, pp. 18–39, 2014.

Busa-Fekete, R. and Kégl, B. Fast boosting using adversarial bandits. In *International Conference on Machine Learning (ICML)*, volume 27, pp. 143–150, Haifa, Israel, 2010. ACM.

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3): 1516–1541, 2013.

Carpentier, A. and Valko, M. Extreme bandits. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems*, pp. 1089–1097, 2014.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning and Games*. Cambridge university press, 2006.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956.

Even-Dar, E., Mannor, S., and Mansour, Y. PAC bounds for multi-armed bandit and markov decision processes. In *Proceedings of the 15th Conference on Learning Theory (COLT)*, pp. 255–270, 2002.

Gabillon, V., Ghavamzadeh, M., Lazaric, A., and Bubeck, S. Multi-bandit best arm identification. In *Advances in NIPS 24*, pp. 2222–2230, 2011.

Gaudel, Romaric and Sebag, Michèle. Feature selection as a one-player game. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pp. 359–366, 2010.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. Pac subset selection in stochastic multi-armed bandits. In *Proceedings of the Twenty-ninth International Conference on Machine Learning (ICML 2012)*, pp. 655–662, 2012.

Kuleshov, Volodymyr and Precup, Doina. Algorithms for multi-armed bandit problems. *CoRR*, abs/1402.6028, 2014.

Lai, T.L. and Robbins, H. Asymptotically efficient allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Mannor, Shie and Tsitsiklis, John N. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

Massart, P. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, 18(3): 1269–1283, 1990.

Sani, Amir, Lazaric, Alessandro, and Munos, Rémi. Risk-aversion in multi-armed bandits. In *26th Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 3284–3292, 2012.

Schachter, B. An irreverent guide to Value-at-Risk. *Financial Engineering News*, 1, 1997.

Yu, Jia Yuan and Nikolova, Evdokia. Sample complexity of risk-averse bandit-arm selection. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pp. 2576–2582. AAAI Press, 2013.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Yue, Yisong and Joachims, Thorsten. Beat the mean bandit. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 241–248, 2011.

Zhou, Yuan, Chen, Xi, and Li, Jian. Optimal pac multiple arm identification with applications to crowdsourcing. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 217–225, 2014.

Zoghi, M., Whiteson, S., Munos, R., and Rijke, M. Relative upper confidence bound for the k-armed dueling bandit problem. In *Proceedings of the Thirty-First International Conference on Machine Learning (ICML)*, pp. 10–18, 2014.

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.