# Best-Arm Identification for Quantile Bandits with Privacy

**Dionysios S. Kalogerias**
Department of ESE
University of Pennsylvania
Philadelphia, PA 19104
dionysis@seas.upenn.edu

**Kontantinos E. Nikolakakis**
Department of ECE
Rutgers University
Piscataway, NJ 08854
k.nikolakakis@rutgers.edu

**Anand D. Sarwate**
Department of ECE
Rutgers University
Piscataway, NJ 08854
anand.sarwate@rutgers.edu

**Or Sheffet**
Faculty of Engineering
Bar-Ilan University
Ramat Gan, Israel 5290002
or.sheffet@biu.ac.il

## Abstract

We study the best-arm identification problem in multi-armed bandits with stochastic, potentially private rewards, when the goal is to identify the arm with the highest quantile at a fixed, prescribed level. First, we propose a (non-private) successive elimination algorithm for strictly optimal best-arm identification, we show that our algorithm is $\delta$-PAC and we characterize its sample complexity. Further, we provide a lower bound on the expected number of pulls, showing that the proposed algorithm is essentially optimal up to logarithmic factors. Both upper and lower complexity bounds depend on a special definition of the associated suboptimality gap, designed in particular for the quantile bandit problem — as we show, when the gap approaches zero, best-arm identification is impossible. Second, motivated by applications where the rewards are private, we provide a differentially private successive elimination algorithm whose sample complexity is finite even for distributions with infinite support-size, and we characterize its sample complexity as well. Our algorithms do not require prior knowledge of either the suboptimality gap, or other statistical information related to the bandit problem at hand.

## 1 Introduction

Multi-armed bandits are an important class of online learning problems with a rich history (see the book by Lattimore and Czepesvári [2020] for a detailed treatment). In a *stochastic* $K$-armed bandit problem, a learner is presented with a set of $K$ different actions (or *arms*) $\{1, 2, \ldots, K\}$ and can sequentially take actions (*pull arms*) to receive random rewards. The reward of arm $i$ at time $t$ is $X_t^i$. The learner may have one of a number of common objectives, such as to find the arm with the maximum $\mu_i = \mathbb{E}[X^i]$ to minimize cumulative regret [Madani et al., 2004, Bubeck et al., 2009].

In this paper, we study a different form of bandit problems in which the figure of merit is the *left-side* $q$*-quantile* of the involved reward distributions, defined, for arm $i$, as $F_i^{-1}(q) = \inf\{x : F_i(x) \geq q\}$, where $F_i(x) = \mathbb{P}[X^i \leq x]$ is the corresponding cumulative distribution function (CDF) [Yu and Nikolova, 2013a, Szörényi et al., 2015]. In particular, we study the problem of *best-arm identification*, i.e., that of identifying the arm with the highest or lowest $q$-quantile, with as few samples as possible.

Authors' names are in alphabetical order | Preprint under review.

The quantile bandit problem arises naturally in the context of *risk-aware* optimization and learning, which has expanded considerably during the last decade [Ruszczyński and Shapiro, 2006, Sani et al., 2012a, Shapiro, 2012, Shapiro et al., 2014, Tamar et al., 2017, Jiang and Powell, 2018, Huang and Haskell, 2018, Kalogerias and Powell, 2018, Vitt et al., 2019, Kalogerias and Powell, 2019, Kagrecha et al., 2019, Cardoso and Xu, 2019, Kim et al., 2019, Zhou and Tokekar, 2020]. There are many application scenarios which fit this quantile-based risk-aware setting.

1. Arms are different feasible asset portfolio allocations [Gaivoronski and Pflug, 2005] and the goal is to find the portfolio with the minimum potential monetary loss, within a target *investment risk* $q$. If such a (random) loss is denoted by $Z$, then this goal may be achieved by choosing $F_Z^{-1}(1-q)$ as the corresponding objective (to be minimized). In this context, the $(1-q)$-quantile is well-known as the *Value-at-Risk at level* $q$, denoted as $\mathrm{V@R}_q(Z) \equiv F_Z^{-1}(1-q)$.
2. Arms are different servers which can be assigned jobs and the rewards are delays. The goal is to identify the server with the highest 95th percentile delay because "waiting for the slowest 5% of the requests to complete is responsible for half of the total 99%-percentile latency" [Dean and Barroso, 2013].
3. Arms are different strains of an illness (e.g. different lung cancer genotypes) and the rewards are effectiveness of a proposed treatment on the strain. We wish to find a treatment which guarantees maximal effectiveness in at least $90\%$ of patients.

From a technical standpoint, the quantile bandit problem differs from the mean (or risk-neutral) bandit problem in a number of important ways. First, for the mean, the *suboptimality gap* between the optimal $i^*$ and a sub-optimal arm $i$ is simply $\mu_{i^*} - \mu_i$, whereas the absolute difference between quantiles is less useful. Instead, the gap depends on the levels of the CDF in the neighborhood of the quantile and prior works have proposed a number of variant definitions [Szörényi et al., 2015, David and Shimkin, 2016, Howard and Ramdas, 2019]. In our work, we define an improved version of this gap and show that it exactly characterizes the problem: when our gap is zero the max-quantile arm is not identifiable.

The advent of wide-scale data analytics has made privacy issues a growing concern. Differential privacy (DP) Dwork et al. [2006] has become the de-facto gold standard for privacy preserving data-analysis. For quantile bandit problems involving the data of individuals (e.g. the financial and medical examples above) it is natural to model the reward information as private or sensitive. We therefore like to both identify the arm with the best quantile and protect the privacy of individuals. The goal is to minimize the "cost of privacy": how many *more* samples does the private algorithm need over the non-private algorithm? To understand this, it is necessary to have a non-private baseline to measure against; our characterization of the sample complexity of the non-private problem establishes such a baseline. We thus dedicate the first part of this work to showing that our notion of gap fully characterizes solvable instances.

In this paper we make the following contributions.

- We provide a *definition of the gap* $\Delta_i$ *at level* $q$ that generalizes those proposed in prior work [Szörényi et al., 2015, David and Shimkin, 2016, Howard and Ramdas, 2019].[1] Our gap precisely captures the difficulty of the problem in the sense that when $\Delta_i = 0$ for all suboptimal arms $i$, no algorithm can hope to identify the arm with the higher $q$-quantile (Lemma 5).
- We introduce a new *pure-exploration successive elimination algorithm for quantile bandits* (Algorithm 1), show that it is $\delta$-PAC (Theorem 6) and provide nearly matching upper (Theorem 7) and lower (Theorem 8) bounds on the sample complexity that depend on our improved gap definition. These results complement prior work on $\varepsilon$-optimal quantile bandits [Szörényi et al., 2015, David and Shimkin, 2016, Howard and Ramdas, 2019], and present a complete picture of the problem.
- Using our modified confidence intervals, we propose the first *differentially private best-arm identification algorithm* (Algorithm 2) for quantile bandits, prove that it is private (Theorem 9), and analyze the trade-off between privacy budget and sample complexity (Theorems 10 and 11). Interestingly, the sample complexity bound for our private algorithm has no dependency on the support size of the distribution, which is *necessary* in the case where one wished to privately

[1] Shortly prior to submission we were informed that a revised (yet unposted) draft of Howard and Ramdas [2019] concurrently proposes the same gap (Ramdas, Personal communication, 2020).

estimate the $q$-quantile [Beimel et al., 2013a, Feldman and Xiao, 2014, Bun et al., 2015] rather than identify which arm has highest quantile.

**Prior work.** The main part of the literature considers the problem of stochastic best-arm identification for the mean. This setting received renewed attention after the work of Even-Dar et al. [2002] on the MAB problem in the PAC learning setting. Later work follows by considering extensions/variations of this problem [Kalyanakrishnan et al., 2012, Gabillon et al., 2012, Karnin et al., 2013, Jamieson et al., 2013, 2014]. Additionally, Mannor and Tsitsiklis [2004b], Anthony and Bartlett [2009] provide lower bounds on the sample complexity in terms of the mean sub-optimality gap. Alternative lower bounds based on the KL-divergence of the arms' distributions are provided by Burnetas and Katehakis [1996], Chen and Li [2015], Kaufmann et al. [2016], Garivier and Kaufmann [2016]. Cappé et al. [2013] present the KL-UCB algorithm that achieves (asymptotically) optimal sample complexity rates by matching known lower bounds. In parallel, prior works encompass non-stochastic approaches [Jamieson and Talwalkar, 2016, Li et al., 2016], as well.

Bandit models models with non-stationary [Allesiardo and Féraud, 2017, Allesiardo et al., 2017], or heavy-tailed [Bubeck et al., 2012, 2013] distributions are most related, since the quantile problem is often of interest in these settings. Further, Kagrecha et al. [2019] consider the unbounded reward best-arm identification problem while variants of regret based approaches include minimization of generalized loss functions [Li et al., 2018, Berthet and Perchet, 2017, Boda et al., 2019, Maillard, 2013]. More recent works also consider risk measures, for instance conditional value-at-risk (CVaR) [Yu and Nikolova, 2013b, Cardoso and Xu, 2019], mean-variance [Even-Dar et al., 2006, Sani et al., 2012b, Vakili and Zhao, 2016] or unified approaches [Cassel et al., 2018]. These are complemented by concentration results on risk measure estimators [Wang and Gao, 2010, Kolla et al., 2019, Bhat and Prashanth, 2019].

Our results complement prior work on quantile bandit problem for best-arm identification [Yu and Nikolova, 2013a, Szörényi et al., 2015, David and Shimkin, 2016, Torossian et al., 2019, Howard and Ramdas, 2019]. Altschuler et al. [2019] specifically study *median* identification for contaminated distributions in the robust statistics sense under distributional assumptions on the arms. Of these, the most highly related works are the beautiful work of Szörényi et al. [2015], the refinement by David and Shimkin [2016], and the preprint of Howard and Ramdas [2019]. These works study $\epsilon$-approximate best-arm identification: the algorithm returns an arm which is within $\epsilon$ of optimal, for some $\epsilon > 0$. Szörényi et al. [2015] and David and Shimkin [2016] use a gap which depends on this $\epsilon$. Howard and Ramdas [2019] have a gap without $\epsilon$ but study $\epsilon$-approximate algorithms. By contrast, our algorithm returns the optimal arm, and we show that when our gap is 0 then a suboptimal distribution (with small $q$-quantile) is actually indistinguishable from a distribution with a larger $q$-quantile (see Section 3.2). Like [Szörényi et al., 2015], our algorithm uses successive elimination (and not the UCB method of Howard and Ramdas [2019]). Both David and Shimkin [2016] and Howard and Ramdas [2019] tighten the upper bounds to a double-logarithmic factor, which we can do as well using a standard epoch modification.

The field of differentially private machine learning is, by now, too large to summarize here, as the following (non-exhaustive) list of works discussing learning quantiles/threshold-functions attests [Nissim et al., 2007, Chaudhuri and Hsu, 2011, Beimel et al., 2013a,b, Feldman and Xiao, 2014, Bun et al., 2015, Alon et al., 2019, Kaplan et al., 2020]. For differentially private multi-armed bandit problems for the mean, Mishra and Thakurta [2015] were the first to analyze a differentially private (DP) algorithm for multi-armed bandit, building a private variant of the UCB-algorithm [Auer et al., 2002] using the tree-based algorithm [Chan et al., 2010, Dwork et al., 2010]. Shariff and Sheffet [2018] have proven that any DP-algorithm for the (mean) multi-armed bandit problem must pull each suboptimal arm $i$ at least $\Omega\left(\log(T)/\epsilon(\mu_{i^*} - \mu_i)\right)$ many times (with $i^*$ denoting the optimal arm, of largest mean-reward $\mu_{i^*} = \max_{i\in\mathcal{A}} \mu_i$) which doesn't quite meet the DP-UCB algorithm's upper bound. Most recently Sajed and Sheffet [2019] gave a DP version of successive elimination whose regret matches the lower bound of Shariff and Sheffet [2018].

## 2 Problem Statement

We consider a $K$-armed unstructured stochastic bandit $\nu = (\nu_i : ci \in \mathcal{A})$, where $\mathcal{A} \triangleq \{1, 2, \ldots, K\}$ is the set of arms and $\nu_i$ are probability measures. For the $i$-th arm, let $X^i$ be a random variable with distribution $\nu_i$. We will describe distributions by their cumulative distribution functions (CDFs).

**Definition 1.** *Let $F_i(\cdot)$ be the CDF of $X^i$ for arm $i$. The* $q$-quantile *$F_i^{-1}(q)$ is defined as*

$$F_i^{-1}(q) \triangleq \inf\{\xi : \mathbb{P}[X^i \leq \xi] \geq q\} \tag{1}$$

*and* best arm *is defined as*

$$i^* \triangleq \arg\max_{i \in \mathcal{A}} F_i^{-1}(q). \tag{2}$$

For simplicity, we assume that the best arm is unique in the set $\mathcal{A}$. We denote the set of sub-optimal arms as $\mathcal{A}^{-i^*} \triangleq \{1, 2, \ldots, K\} \setminus \{i^*\}$. Given $n$ samples the estimated CDF of $X^i$ is $\hat{F}_{n,i}(x) \triangleq \frac{1}{n}\sum_{\ell=1}^{n} \mathbb{I}\{X_\ell^i \leq x\}$. We denote the set of samples from arm $k$ as $\{X_i^k\}$, while for the $j^{\text{th}}$ order statistic of $\{X_i^k\}$ we use the standard notation $X_{(j)}^k$.

An algorithm for our quantile bandit chooses at each time $n$ an arm $i_n \in \mathcal{A}$ and obtains a reward $X_n^{i_n} \sim \nu_{i_n}$. The algorithm terminates by stopping sampling and declaring an arm $\hat{k}$ as the arm with the highest $q$-quantile, and succeeds if actually $\hat{k} = i^*$. We call an algorithm $\delta$-PAC if $\mathbb{P}(\hat{k} = i^*) \geq 1 - \delta$.

To derive our results on differential privacy, we think of the rewards from each of the arms at each time $t$ as coming from different individuals. This means that to protect an individual we are interested in *event-level privacy* [Dwork et al., 2010], defined as follows.

**Definition 2.** *Two sequences of rewards $S$ and $S'$ are called* neighboring *if there exists a single $t$ on which they differ and for any other $t' \neq t$ the rewards are equal. A randomized algorithm $A$ is said to be* $\epsilon$-differentially private ($\epsilon$-DP) under continuous observation *if for any two neighboring input streams $S$ and $S'$ and for any set $\mathcal{C} \subset \mathcal{A}^T$ of sequences of arm-pulls we have that $\mathbb{P}[A(S) \in \mathcal{C}] \leq e^\epsilon \mathbb{P}[A(S') \in \mathcal{C}]$.*

It is important to note that much like the input sequence $S$ which is revealed one reward at a time based on the algorithm's decision, the output $A(S)$ is also revealed one pull at a time, and our requirement holds for all time-steps throughout the execution of the algorithm.

We comment on some differential privacy properties (see also Dwork and Roth [2014]) and we will use them later in the construction of the DP-algorithm: *Basic composition:* if $A_1$ and $A_2$ are $\epsilon_1$- and $\epsilon_2$-DP algorithms resp., then for any $S$ releasing the pair $\langle A_1(S), A_2(S)\rangle$ is $(\epsilon_1 + \epsilon_2)$-DP (provided both algorithms use independent coin tosses). *Parallel composition:* if $A$ is an $\epsilon$-DP algorithm, then for any input $S$ and any partition of $S$ into $S_1, S_2$, outputting $\langle A(S_1), A(S_2)\rangle$ is $\epsilon$-DP (again, provided both instantiations are done using independent coin tosses). *Laplace noise:* if $Q$ is a query where $GS(Q) = \max_{S,S' \text{neighbors}} |Q(S) - Q(S')| \leq 1$, then outputting $Q(S) + Z$ with $Z \sim \mathsf{Lap}(1/\epsilon)$ preserves $\epsilon$-DP, where $\mathsf{Lap}(1/\epsilon)$ is the Laplace distribution with density $\epsilon e^{-\epsilon|x|}/2$.

# 3 Non-Private Best-Arm Identification

## 3.1 A Successive Elimination Algorithm

We choose to study successive elimination (SE) rather than a variant of UCB [Auer et al., 2002] (adopted by Howard and Ramdas [2019] for quantiles) for a number of reasons. Firstly, SE methods are in a sense more natural than UCB since the goal is to identify the best arm. Secondly, we prove matching upper and lower bounds on the sample complexity, showing our algorithm is essentially optimal (up to logarithmic terms). Lastly, because we are interested in developing differentially private algorithms, the SE algorithm is more "privacy friendly" because the sampling strategy is independent of the data and it uses confidence bounds in terms of the order statistics. Finally, there is no private analog to UCB when the distributions have infinite support.

Our Successive Elimination algorithm for Quantiles (SEQ) Algorithm is shown in Algorithm 1. To explain SEQ (Algorithm 1), we define the sequence $\mathrm{D} = \mathrm{D}(n) \triangleq \sqrt{\log(4Kn^2/\delta)/2n}$ and we use the concentration bound on the quantile (Lemma 14, supplementary material, Section A)

$$\mathbb{P}\big(F_i^{-1}(q) \in \big[X^i_{(\lfloor n(q-\mathrm{D})\rfloor)}, X^i_{(\lceil n(q+\mathrm{D})\rceil)}\big]\big) > 1 - \frac{\delta}{2Kn^2}. \tag{3}$$

The latter yields the elimination condition in line 13 of Algorithm 1. Specifically, when the inequality $X^j_{(\lfloor n(q-\mathrm{D})\rfloor)} \geq X^i_{(\lceil n(q+\mathrm{D})\rceil)}$ holds then $F_i^{-1}(q) \leq F_j^{-1}(q)$ with probability at least $1 - \delta$. Thus to identify the arm with the maximum quantile if $X^j_{(\lfloor n(q-\mathrm{D})\rfloor)} \geq X^i_{(\lceil n(q+\mathrm{D})\rceil)}$ we remove $i$ from $\mathcal{A}$.

**Algorithm 1** Successive Elimination for Quantiles (SEQ)

```
Require: $\delta, q$
1: $\mathcal{A} \leftarrow \{1, \dots, K\}$
2: $\mathrm{D} \leftarrow \sqrt{\frac{\log(4Kn^2/\delta)}{2n}}$
3: Find the smallest $n_* \in \mathbb{N} \setminus \{1\}$ such that $\mathrm{D} \leq q$
4: $n \leftarrow n_*$
5: Pull $n$ times each arm $k \in \mathcal{A}$, obtain new samples $X_1^k, \dots, X_n^k$ for all $k \in K$
6: while $|\mathcal{A}| > 1$ do
7:     Increment $n \leftarrow n + 1$
8:     Set $\mathrm{D} \leftarrow \sqrt{\frac{\log(4Kn^2/\delta)}{2n}}$
9:     Pull each arm in $\mathcal{A}$, obtain samples $X_n^k$ for all $k \in \mathcal{A}$
10:    Update the order statistics $X^k_{(\lfloor n(q-\mathrm{D})\rfloor)}$ and $X^k_{(\lceil n(q+\mathrm{D})\rceil)}$ for all $k \in \mathcal{A}$
11:    for each ($j \in \mathcal{A}$) do
12:        for each ($i \in \mathcal{A}$ where $i \neq j$) do
13:            if $X^j_{(\lfloor n(q-\mathrm{D})\rfloor)} \geq X^i_{(\lceil n(q+\mathrm{D})\rceil)}$ then remove $i$ from $\mathcal{A}$
14: return $\mathcal{A}$
```

We can also consider variants of the algorithm. For example, to identify the arm with the minimum quantile, we modify line 13 of the algorithm as follows: if $X^j_{(\lfloor n(q-\mathrm{D})\rfloor)} \geq X^i_{(\lceil n(q+\mathrm{D})\rceil)}$ then remove $j$ from $\mathcal{A}$. A second variant would be to take samples in epochs of increasing size. We take this approach in the development of the differentially private version of Algorithm 1, which reduces to a non-private epoch-based variant of Algorithm 1 (Section 4, Algorithm 2). This epoch-based algorithm improves the bound of Theorem 7 from $\log(\frac{1}{\Delta_i})$ to $\log\log(\frac{1}{\Delta_i})$ and matches asymptotically the bound of Howard and Ramdas [2019] (see the discussion at the end of Section 4).

### 3.2 Sub-optimality gap

We first define the suboptimality gap between the best arm $i^*$ and any sub-optimal arm.

**Definition 3.** *The* suboptimality gap *between the optimal arm* $i^*$ *and any suboptimal arm* $i$ *at level* $q \in (0, 1)$ *is*

$$\Delta_i \triangleq \Delta(F_i, F_{i^*}) = \sup\{\eta \geq 0 : \ F_i^{-1}(q + \eta) \leq F_{i^*}^{-1}(q - \eta)\}. \tag{4}$$

How can we interpret this gap? Roughly speaking, it is the amount of probability mass needed to swap the order of the quantiles. To get further insight into the definition (4), notice that D is decreasing with respect to $n$, and the elimination occurs at the first time (maximum value of D) that gives $X^j_{(\lfloor n(q-\mathrm{D})\rfloor)} \geq X^i_{(\lceil n(q+\mathrm{D})\rceil)}$. As consequence, the value $\Delta_i$ in (4) acts as a threshold on the quantity D. Our definition applies on continuous, discrete, and mixture distributions. For a graphical representation of the suboptimality gap for different values of the level $q$ see Section D in the supplement, Figure 2.

For discrete distributions, we show that the difference between the quantile can become arbitrarily small while the definition of the gap and the sample complexity of Algorithm 1 remain insensitive (Section D, Figure 1, supplement). While the difference between the quantile values is not the correct defintion to use for the gap, the two quantities are related in certain cases (see Section B, supplement).

**Proposition 4.** *Suppose* $F$ *and* $G$ *are two distributions with L-Lipschitz continuous and strictly increasing CDFs. Then the gap* $\Delta(F, G) \leq \frac{L}{2}|F^{-1}(q) - G^{-1}(q)|$.

We provide a discussion to explain differences between the proposed gap in (4) and other definitions in prior work, see Section D. Most importantly, the key point in this definition of the quantile suboptimality-gap is that it *fully characterizes* the pairs of distributions for which we can discern that one has a higher $q$-quantile than another *from i.i.d. samples*. Formally, for a pair of distributions $(F_\mathrm{l}, F_\mathrm{h})$ where the former has a suboptimal $q$-quantile than the latter, namely — $F_\mathrm{l}^{-1}(q) \leq F_\mathrm{h}^{-1}(q)$, we define the notion of *distance to quantile-flip at* $q$ as

$$d_{\mathrm{flip}}(F_\mathrm{l}, F_\mathrm{h}) = \inf_{(G_\mathrm{h}, G_\mathrm{l}): G_\mathrm{h}^{-1}(q) > G_\mathrm{l}^{-1}(q)} \max\{d_{\mathrm{TV}}(F_\mathrm{l}, G_\mathrm{h}), d_{\mathrm{TV}}(F_\mathrm{h}, G_\mathrm{l})\}. \tag{5}$$

In Section B of the supplement, we prove the following lemma.

**Lemma 5.** *For any $0 < q < 1$ and any two distributions $F_i$ and $F_{i^*}$ such that $F_i^{-1}(q) \leq F_{i^*}^{-1}(q)$ it holds that $d_{\text{flip}}(F_i, F_{i^*}) = \Delta(F_i, F_{i^*})$ provided that $\Delta(F_i, F_{i^*}) < \min\{q, 1-q\}$.*

This lemma shows that if $\Delta_i = 0$ then *no algorithm* can distinguish which arm has the higher $q$-quantile, regardless of its sample-size: every batch of samples can be generated by a quantile flip $(G_{\text{h}}, G_{\text{l}})$ with the same probability Conversely, when $\Delta_i > 0$ we devise an algorithm that discerns which arm has the higher $q$-quantile using $\tilde{O}(\Delta_i^{-2})$ many examples from each arm and argue that this bound is optimal in the sense that there exists a collection of $K$ distributions requiring $\tilde{O}(\Delta_i^{-2})$ many examples from each distribution (Section 3.3).

### 3.3 Analysis

Our first result guarantees that the Algorithm 1 eliminates the sub-optimal arms while the unique best arm remains in the set $\mathcal{A}$ with high probability until the algorithm terminates. For the rest of the paper we assume that $\Delta_i > 0$ for all $i \in \{1, 2 \ldots, K\} \setminus i^*$.

**Theorem 6.** *Algorithm 1 is $\delta$-PAC.*

**Theorem 7.** *Fix $\delta \in (0, 1)$. There exists a constant $C > 0$ such that the number of samples $\tau$ (and total number of pulls) of Algorithm 1 satisfies, with probability at least $1 - \delta$,*

$$\tau \leq C \sum_{i \in \mathcal{A}^{-i^*}} \frac{\log \frac{K}{\delta} + \log(\frac{1}{\Delta_i})}{\Delta_i^2}. \tag{6}$$

From the proof of Theorem 7, it also follows that the number of pulls for each suboptimal arm $i \in \mathcal{A}^{-i^*}$ is at most $\mathcal{O}\left(\log(K/\delta\Delta_i)/\Delta_i^2\right)$. The upper bound indicates that the number of pulls (with high probability) is proportional to the quantity $1/\Delta_i^2$ up to a logarithmic factor for each suboptimal arm $i$. In fact, experimental results on SEQ (Algorithm 1) show that the explicit bound in (6) matches the average number of pulls in the experiment. We provide the proofs of Theorem 6 and Theorem 7 in the supplementary material, Sections B.3 and B.4 respectively. For the simulation results we refer the reader to the supplementary material, Section D, Figure 3.

We next provide a lower bound on the expected number of pulls (proof in Section B.5). Szörényi et al. [2015] use the results of Mannor and Tsitsiklis [2004a] to obtain a bound that depends on $\varepsilon \vee \Delta$ (for $\varepsilon > 0$). We use the approach suggested in Lattimore and Czepesvári [2020] on a different class of distributions and get a bound that depends on $\Delta$.

**Theorem 8.** *Fix $\delta \in (0, 1)$ and $\Delta \in (0, 1/4)$. There exists a quantile bandit with $K$-arms and worst-case gap $\min_{i \neq i^*} \Delta_i = \Delta$, such that*

$$\inf_{\delta-\text{PAC policy } \pi} \mathbb{E}_\pi[\tau] \geq C^2 \frac{K-1}{15\Delta^2} \log\left(\frac{1}{4\delta}\right). \tag{7}$$

From Theorem 8, it follows that, up to logarithmic factors depending on $\delta$ (Theorem 8), and also $K$, $\Delta_i, i \in \mathcal{A}^{-i^*}$ (Theorem 7), Algorithm 1 is (almost) optimal relative to the expected number of pulls achieved, and its performance is *necessarily* inversely proportional to the square of our suboptimality gap. Also note that Theorem 8 is consistent with other results in the literature proved for the usual, risk-neutral bandit setting (see, e.g., Cappé et al. [2013]). More interestingly, our lower bound shows that as $\Delta \to 0$ the sample complexity goes to $\infty$ and indeed as Lemma 5 shows, $\Delta = 0$ implies that the best-quantile arm identification problem is impossible.

## 4 A Private Algorithm for Best-Quantile-Arm Identification

### 4.1 Differential Privacy Preliminaries

Releasing a differentially private *estimation* of the $q$-quantile of a given distribution is considered to be a hard task. In particular its accuracy is dependent on the (tight bound for $\epsilon$-differential privacy were given by Beimel et al. [2013a], Feldman and Xiao [2014]). This makes the problem infeasible in the case the distribution's support contains the entire real interval $[0, 1]$. We get around this by never publishing an approximation for $q$-quantile; instead we output an arm $\hat{k}$ that should have a higher $q$-quantile than any other arm $i$. We do this by eliminating an arm based on a query that counts the number of *pairs of draws* attesting for an arm's suboptimal $q$-quantile. This counting query has the key property that its sensitivity is always 1 regardless of the size of the support of the reward distribution of arm $i$. This reformulation is what allows us to obtain a sample complexity bound that is independent of the support size of any arm's distribution and in particular works even for distributions over the reals (of unbounded support). Further details follow.

*On the difficulties with a private UCB quantile algorithm.* Differentially private UCB algorithms for the mean using tree-based algorithms [Chan et al., 2010, Dwork et al., 2010] do not extend straightforwardly to the quantile case, but a carefully designed counting query[2] renders the usage of tree-based algorithms feasible to our problem. However, Algorithm 2 is superior to this approach in two respects, both related to the horizon $T$. First, (as observed by Sajed and Sheffet [2019]) the tree based algorithm's utility bound has a $\mathrm{poly}(\log(T))$ dependence whereas our algorithm's is only $\log\log(T)$.[3] Secondly, the tree-based algorithms require knowing $T$ in advance; this is nontrivial because doubling tricks require either rebudgeting $\epsilon$ (incurring increased sample complexity) or discarding all samples when the next epoch begins, which incurs $\tilde{O}(\Delta_i^{-2})$ pulls per suboptimal arm in *every* epoch because the UCB algorithm never eliminates any arms. Our gap definition and algorithm avoids having any such prior knowledge of $T$ or the value of the gap.

*Notation.* Throughout this section we deal with pure $\epsilon$-DP and use $\delta$ to represent the failure probability of our algorithm. The reader is advised to not be confused with the notion of $(\epsilon, \delta)$-DP.[4] Because of the many indices needed, in this section we index arms by $a$ and $b$ rather than $i$ and $j$.

### 4.2 Differentially Private Successive Elimination for the Highest Quantile Arm

The differentially private algorithm is shown in Algorithm 2. Much like the algorithm in Sajed and Sheffet [2019], our algorithm is also epoch based. In epoch $e$ our goal is to eliminate all arms $i$ with suboptimality quantile-gap (Equation (4)) $\Delta_i \geq \Gamma_e = 2^{-e}$. As we argue, the number of arm pulls in each epoch from each existing arm is $n_e \geq \Gamma_e^{-2}$. The key point is that due to the geometric nature of $\Gamma_e$ it follows that each $n_e$ is proportional to the sum of pulls thus far $\sum_{1 \leq e' < e} n_{e'}$, and so we may as well split the stream into different chunks, starting each epoch anew (discarding all examples drawn in all previous epochs). Because we eliminate arms, this still doesn't cost us a lot in the number of overall pulls, yet allows us to avoid splitting the privacy budget $\epsilon$ due to parallel composition.

**Algorithm 2** Differentially Private Successive Elimination for Quantiles (DP-SEQ)

```
Input: Number of arms $K$, quantile level $q \in (0,1)$, privacy parameter $\epsilon > 0$, failure probability
    $\delta \in (0, 1/2)$.
Initialize $\mathcal{A} \leftarrow \{1, \dots, K\}$, epoch $e \leftarrow 0$.
while $|\mathcal{A}| > 1$ do
    increment $e \leftarrow e + 1$
    Set $\Gamma_e \leftarrow 2^{-e}$ and $\gamma \leftarrow \frac{\Gamma_e}{4}$. Set $n_e \leftarrow \max\left\{\frac{16}{\Gamma_e^2}, \frac{64(|\mathcal{A}|-1)}{\Gamma_e \cdot \epsilon}\right\} \cdot \log(\frac{6|\mathcal{A}|e^2}{\delta})$.
    for each ($i \in \mathcal{A}$) do pull arm $i$ for $n_e$ times to obtain $X_1^i, X_2^i, ..., X_{n_e}^i$.
    Set $i \leftarrow \lfloor n_e(q - 2\gamma) \rfloor$ and $j \leftarrow \lceil n_e(q + 2\gamma) \rceil$.
    for each ($a \in \mathcal{A}$) do
        for each ($b \in \mathcal{A}$ where $b \neq a$) do
            Draw $Z_{a,b} \sim \mathsf{Lap}(\frac{2(|\mathcal{A}|-1)}{\epsilon})$
            $\ell^* \leftarrow \max\{\ell \leq \min\{i, n_e - j\} : \forall \ell' \leq \ell$ we have that $X^b_{(j+\ell')} \leq X^a_{(i-\ell')}\}$
            if $(\max\{\ell^*, 0\} + Z_{a,b} \geq \frac{4(|\mathcal{A}|-1)}{\epsilon} \cdot \log(6|\mathcal{A}|^2 e^2/\delta)$ ) then
                remove $b$ from $\mathcal{A}$
```

We still need a way to privately eliminate arms at the end of each epoch. In the case of the means, Sajed and Sheffet [2019] eliminate arms by computing $\epsilon$-DP approximations of the means and

[2] Count the number of examples required to make the quantile-UCB of this arm the max.

[3] Both utility guarantees also have a $\log(1/\delta)$-factor.

[4] We could have used the notion of $(\epsilon, \delta)$-DP and reduce our bounds by a factor of $\sqrt{K}$ by relaying on the advanced composition theorem. As a matter of style, we opted for pure-DP.

comparing those, leveraging the post-processing invariance of DP. Unfortunately, we cannot find $\epsilon$-DP approximations for $q$-quantiles that do not depend on the cardinality of the support (and in particular, it is infeasible with infinite support). Instead, we resort to the more naive approach of pairwise comparisons between all $K(K-1)/2 = \Theta(K^2)$ pairs of arms. This requires partitioning the $\epsilon$ of our privacy budget into $\epsilon/2(K-1)$ as each arm participates in at most $2(K-1)$ many comparisons. However, using pairwise comparisons we are able to convert the higher-quantile question into a counting query: how many consecutive examples satisfy that $X^a_{(LCB-i)} \geq X^b_{(UCB+i)}$? We prove that under event-level privacy, this query has sensitivity of at most 1, allowing us to eliminate the suboptimal arm $b$ via the standard Laplace-mechanism.

Our first result is a privacy guarantee for Algorithm 2. The proof for this and all results that follow may be found in the Section C of the supplement.

**Theorem 9.** *Algorithm 2 is $\epsilon$-differentially private.*

We continue by providing a high probability guarantee on the first epoch for which the private SEQ (Algorithm 2) terminates.

**Theorem 10.** *For Algorithm 2, the following events occur with probability at least $1-\delta$: (a) it keeps at least one optimal arm in $\mathcal{A}$ and (b) it removes each sub-optimal arm $a$ by epoch $e = \lceil \log_2(1/\Delta_a) \rceil$.*

Lastly, we characterize the sample complexity of DP-SEQ (Algorithm 2), the number of pulls for each suboptimal arm and the total number of pulls at termination.

**Theorem 11.** *With probability at least $1-\delta$, Algorithm 2, pulls each suboptimal arm $a$ at most*

$$\mathcal{O}\left(\left(\frac{1}{\Delta_a^2} + \frac{K}{\epsilon \Delta_a}\right) \log\left(\frac{K}{\delta} \log\left(\frac{1}{\Delta_a}\right)\right)\right) \tag{8}$$

*many times.*

By taking $\epsilon \to \infty$, Theorem 11 provides the utility of the standard (non-private) epoch-based successive elimination variant of Algorithm 1. Indeed, by introducing epochs the concentration bound in (3) becomes

$$\mathbb{P}\big(F_i^{-1}(q) \in \big[X^i_{(\lfloor n_e(q-\mathrm{D}_e)\rfloor)}, X^i_{(\lceil n_e(q+\mathrm{D}_e)\rceil)}\big]\big) > 1 - \frac{\delta}{2Ke^2}, \tag{9}$$

where $e$ denotes the epoch, $\mathrm{D}_e = 2^{-e}$ and $n_e = \mathrm{D}_e^{-2} \log(4Ke^2/\delta)/2$. This yields a bound on the total number of pulls for the epoch-based algorithm of the order of

$$\mathcal{O}\Bigg(\sum_{i \in \mathcal{A}^{-i^*}} \frac{1}{\Delta_i^2} \log\left(\frac{K}{\delta} \log\left(\frac{1}{\Delta_i}\right)\right)\Bigg), \tag{10}$$

matching the ($\varepsilon$-optimal) bounds of Howard and Ramdas [2019]. As a consequence of the epoch-based approach, the dependence $\log(K/(\delta\Delta_i))$ in (6) becomes $\log(K\delta^{-1}\log(\Delta_i^{-1}))$ for $i \in \mathcal{A}\backslash\{i^*\}$. However, this comes at the expense of much larger constants.

## 5 Discussion and Future Directions

In this paper we characterized the sample complexity of the quantile multi-armed bandit problem when the goal is to exactly identify the arm with the highest $q$-quantile in terms of a new measure of suboptimality (gap) between the distributions of each pair of arms. The problem of the lowest $q$-quantile is a simple modification of our method. Motivated by scenarios where the arm rewards are private or carry sensitive information, we also provided the first differentially private algorithm for the quantile bandit problem. These privacy considerations lead to an interesting open problem which we discuss next.

*Open Problem for Privacy.* Algorithm 2 pulls each suboptimal arm $i$ roughly $K/\epsilon\Delta_i$ times more than Algorithm 1. Because we cannot publish approximations of the $q$-quantiles, the factor of $K$ comes because of the need to make private pairwise comparisons. An open question remains: can we avoid this factor of $K$ or is there a converse showing it is necessary? This factor does not appear when looking at the difference between private and non-private best *mean* arm identification. We would like to know if a different elimination procedure would have the same property but for the quantiles.

The bandit literature is vast, with many variations, and for some of these the quantile bandit setting might provide an interesting twist as a form of risk-aware learning. Bandit optimization with risk control is a particularly interesting direction to which this work can apply. For the case of contaminated quantiles Altschuler et al. [2019] our results imply that $\leq \Delta_i/2$ fraction of contaminated examples could be handled for general $q$-quantiles. There are still open fundamental questions one may ask, in particular related to the hardness of best-arm-identification for functionals of the distribution beyond the mean and variance [Cassel et al., 2018] in the private and nonprivate case.

## References

Robin Allesiardo and Raphaël Féraud. Selection of learning experts. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1005–1010. IEEE, 2017.

Robin Allesiardo, Raphaël Féraud, and Odalric-Ambrym Maillard. The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, 3(4):267–283, 2017.

Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private PAC learning implies finite Littlestone dimension. In Moses Charikar and Edith Cohen, editors, *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 852–860. ACM, 2019.

Jason Altschuler, Victor-Emmanuel Brunel, and Alan Malek. Best arm identification for contaminated bandits. *Journal of Machine Learning Research*, 20(91):1–39, 2019.

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *JMLR*, 47(2-3):235–256, 2002.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of private learners. In *ITCS*, pages 97–110. ACM, 2013a.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In Prasad Raghavendra, Sofya Raskhodnikova, Klaus Jansen, and José D. P. Rolim, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 16th International Workshop, APPROX 2013, and 17th International Workshop, RANDOM 2013, Berkeley, CA, USA, August 21-23, 2013. Proceedings*, volume 8096 of *Lecture Notes in Computer Science*, pages 363–378. Springer, 2013b.

Quentin Berthet and Vianney Perchet. Fast rates for bandit optimization with upper-confidence frank-wolfe. In *Advances in Neural Information Processing Systems*, pages 2225–2234, 2017.

Sanjay P Bhat and LA Prashanth. Concentration of risk measures: A wasserstein distance approach. In *Advances in Neural Information Processing Systems*, pages 11739–11748, 2019.

Vinay Praneeth Boda et al. Correlated bandits or: How to minimize mean-squared error online. *arXiv preprint arXiv:1902.02953*, 2019.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

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

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil P. Vadhan. Differentially private release and learning of threshold functions. In *FOCS*, pages 634–649. IEEE Computer Society, 2015.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz, et al. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

Adrian Rivera Cardoso and Huan Xu. Risk-averse stochastic convex bandit. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 39–47. PMLR, 16–18 Apr 2019. URL `http://proceedings.mlr.press/v89/cardoso19a.html`.

Asaf Cassel, Shie Mannor, and Assaf Zeevi. A general approach to multi-armed bandits under risk criteria. *arXiv preprint arXiv:1806.01380*, 2018.

T.-H. Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. In *Automata, Languages and Programming*, Lecture Notes in Computer Science, pages 405–417, 2010.

Kamalika Chaudhuri and Daniel J. Hsu. Sample complexity bounds for differentially private learning. In *COLT*, volume 19 of *JMLR Proceedings*, pages 155–186, 2011.

Lijie Chen and Jian Li. On the optimal sample complexity for best arm identification. *arXiv preprint arXiv:1511.03774*, 2015.

Yahel David and Nahum Shimkin. PAC lower bounds and efficient algorithms for the max $k$-armed bandit problem. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 878–887, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `http://proceedings.mlr.press/v48/david16.html`.

J. Dean and L. A. Barroso. The tail at scale. *Communications of the ACM*, 56(2):74–80, 2013.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, August 2014. doi: 10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, Lecture Notes in Computer Science, pages 265–284. Springer, Berlin, Heidelberg, 2006.

Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. Differential privacy under continual observation. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, STOC '10, pages 715–724, 2010.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and markov decision processes. In Jyrki Kivinen and Robert H. Sloan, editors, *International Conference on Computational Learning Theory*, volume 2375 of *Lecture Notes in Artificial Intelligence*, pages 255–270. Springer, 2002. doi: 10.1007/3-540-45435-7_18. URL `http://dx.doi.org/10.1007/3-540-45435-7_18`.

Eyal Even-Dar, Michael Kearns, and Jennifer Wortman. Risk-sensitive online learning. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer, 2006.

Vitaly Feldman and David Xiao. Sample complexity bounds on differentially private learning via communication complexity. In Maria-Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 1000–1019. JMLR.org, 2014.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, pages 3212–3220, 2012.

Alexei A Gaivoronski and Georg Pflug. Value-at-risk in portfolio optimization: properties and computational approach. *Journal of risk*, 7(2):1–31, 2005.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027, 2016.

Steven Howard and Aaditya Ramdas. Sequential estimation of quantiles with applications to a/b-testing and best-arm identification. *CoRR*, abs/1902.00746, 2019. URL `http://arxiv.org/abs/1906.09712`.

Wenjie Huang and William B. Haskell. Risk-Aware Q-learning for Markov Decision Processes. In *2017 IEEE 56th Annual Conference on Decision and Control, CDC 2017*, volume 2018-Janua, pages 4928–4933. IEEE, dec 2018. ISBN 9781509028733. doi: 10.1109/CDC.2017.8264388. URL `http://ieeexplore.ieee.org/document/8264388/`.

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial Intelligence and Statistics*, pages 240–248, 2016.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sebastien Bubeck. On finding the largest mean among many. *arXiv preprint arXiv:1306.3917*, 2013.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439, 2014.

Daniel R. Jiang and Warren B. Powell. Risk-Averse Approximate Dynamic Programming with Quantile-Based Risk Measures. *Mathematics of Operations Research*, 43(2):554–579, nov 2018. ISSN 15265471. doi: 10.1287/moor.2017.0872. URL `http://pubsonline.informs.org/doi/10.1287/moor.2017.0872`.

Anmol Kagrecha, Jayakrishnan Nair, and Krishna Jagannathan. Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. In *Advances in Neural Information Processing Systems*, pages 11272–11281, 2019.

Dionysios S. Kalogerias and Warren B. Powell. Recursive Optimization of Convex Risk Measures: Mean-Semideviation Models. *arXiv preprint, arXiv:1804.00636*, apr 2018. URL `http://arxiv.org/abs/1804.00636`.

Dionysios S. Kalogerias and Warren B. Powell. Zeroth-order Algorithms for Risk-Aware Learning. *arXiv preprint, arXiv:1912.09484*, 2019.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012.

Haim Kaplan, Katrina Ligett, Yishay Mansour, Moni Naor, and Uri Stemmer. Privately learning thresholds: Closing the exponential gap. *COLT*, 2020.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246, 2013.

Emilie Kaufmann, Olivier Cappe, and Aurelien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Sung-Kyun Kim, Rohan Thakker, and Ali-akbar Agha-mohammadi. Bi-directional Value Learning for Risk-aware Planning Under Uncertainty. *IEEE Robotics and Automation Letters*, 4(3):2493–2500, jul 2019. ISSN 2377-3766. doi: 10.1109/LRA.2019.2903259. URL `https://ieeexplore.ieee.org/document/8661624/http://arxiv.org/abs/1902.05698`.

Ravi Kumar Kolla, LA Prashanth, Sanjay P Bhat, and Krishna Jagannathan. Concentration bounds for empirical conditional value-at-risk: The unbounded case. *Operations Research Letters*, 47(1): 16–20, 2019.

Tor Lattimore and Csaba Czepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK, 2020.

Bingcong Li, Tianyi Chen, and Georgios B Giannakis. Bandit online learning with unknown delays. *arXiv preprint arXiv:1807.03205*, 2018.

Lisha Li, Kevin Jamieson, Giulia De Salvo, Rostamizadeh A Talwalkar, and A Hyperband. A novel bandit-based approach to hyperparameter optimization. *Computer Vision and Pattern Recognition, arXiv: 1603.0656*, 2016.

Omid Madani, Daniel J Lizotte, and Russell Greiner. The budgeted multi-armed bandit problem. In *International Conference on Computational Learning Theory*, pages 643–645. Springer, 2004.

Odalric-Ambrym Maillard. Robust risk-averse stochastic multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 218–233. Springer, 2013.

Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, June 2004a. URL `http://www.jmlr.org/papers/v5/mannor04b.html`.

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

Nikita Mishra and Abhradeep Thakurta. (nearly) optimal differentially private stochastic multi-arm bandits. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 592–601. AUAI Press, 2015.

Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. Smooth sensitivity and sampling in private data analysis. In David S. Johnson and Uriel Feige, editors, *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA, June 11-13, 2007*, pages 75–84. ACM, 2007.

Andrzej Ruszczyński and Alexander Shapiro. Optimization of convex risk functions. *Mathematics of operations research*, 31(3):433–452, 2006.

Touqir Sajed and Or Sheffet. An optimal private stochastic-MAB algorithm based on optimal private stopping rule. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 5579–5588. PMLR, 2019.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3275–3283. Curran Associates, Inc., 2012a. URL `http://papers.nips.cc/paper/4753-risk-aversion-in-multi-armed-bandits.pdf`.

Amir Sani, Alessandro Lazaric, and Remi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012b.

Alexander Shapiro. Minimax and risk averse multistage stochastic programming. *European Journal of Operational Research*, 219(3):719–726, 2012.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. Society for Industrial and Applied Mathematics, 2nd edition, 2014. ISBN 089871687X. doi: http://dx.doi.org/10.1137/1.9780898718751.

Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. In *Advances in Neural Information Processing Systems*, pages 4301–4311, 2018.

Balazs Szörényi, Róbert Busa-Fekete, Paul Weng, and Eyke Hüllermeier. Qualitative multi-armed bandits: A quantile-based approach. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1660–1668, Lille, France, 07–09 Jul 2015. PMLR. URL `http://proceedings.mlr.press/v37/szorenyi15.html`.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Sequential Decision Making with Coherent Risk. *IEEE Transactions on Automatic Control*, 62(7):3323–3338, jul 2017. ISSN 00189286. doi: 10.1109/TAC.2016.2644871.

Léonard Torossian, Aurélien Garivier, and Victor Picheny. $\mathcal{X}$-armed bandits: Optimizing quantiles, cvar and other risks. In Wee Sun Lee and Taiji Suzuki, editors, *Proceedings of The Eleventh Asian Conference on Machine Learning*, volume 101 of *Proceedings of Machine Learning Research*, pages 252–267, Nagoya, Japan, 17–19 Nov 2019. PMLR. URL `http://proceedings.mlr.press/v101/torossian19a.html`.

Sattar Vakili and Qing Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1093–1111, 2016.

Constantine Alexander Vitt, Darinka Dentcheva, and Hui Xiong. Risk-Averse Classification. *Annals of Operations Research*, aug 2019. ISSN 0254-5330. doi: 10.1007/s10479-019-03344-6.

Ying Wang and Fuqing Gao. Deviation inequalities for an estimator of the conditional value-at-risk. *Operations Research Letters*, 38(3):236–239, 2010.

Jia Yuan Yu and Evdokia Nikolova. Sample complexity of risk-averse bandit-arm selection. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013a. URL `https://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/view/6194/7094`.

Jia Yuan Yu and Evdokia Nikolova. Sample complexity of risk-averse bandit-arm selection. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013b.

Lifeng Zhou and Pratap Tokekar. An Approximation Algorithm for Risk-Averse Submodular Optimization. In *Springer Proceedings in Advanced Robotics, vol. 14*, pages 144–159. Springer, Cham, dec 2020. doi: 10.1007/978-3-030-44051-0_9.

# Supplementary material for "Best-Arm Identification for Quantile Bandits with Privacy"

## A Quantile properties

We start by providing two inequalities for the quantile that we will use later.

**Proposition 12.** *Fix $q \in (0,1)$. Let $F_X^{-1}(q)$ be the $q$-quantile of a random variable $X$ defined in* (1) *of Definition 1. Then*

$$\mathbb{P}\left[X < F_X^{-1}(q)\right] \leq q \tag{11}$$

*and*

$$\mathbb{P}[X > F_X^{-1}(q)] \leq 1 - q. \tag{12}$$

*Proof.* Let $x_n$ be a monotonically increasing sequence such that $\lim_{n\to\infty} x_n = F_X^{-1}(q)$, then

$$\mathbb{P}[X < F_X^{-1}(q)] = \lim_{n\to\infty} \mathbb{P}[X \leq x_n] = \lim_{n\to\infty} F(x_n).$$

Consider $x_n = F_X^{-1}(q) - 2^{-n}$ and assume for sake of contradiction that $\mathbb{P}[X < F_X^{-1}(q)] = q + \epsilon > q$ for some $\epsilon > 0$. It follows that for some $n$ it holds that $F(x_n) > q + \epsilon/2$. Assuming $\xi = x_n < F_X^{-1}(q)$ in the set $\{\xi : \mathbb{P}[X \leq \xi] \geq q\}$ contradicts the definition of $F_X^{-1}(q)$.

It is true that $\mathbb{P}[X > F_X^{-1}(q)] = 1 - \mathbb{P}[X \leq F_X^{-1}(q)]$. The second part of the claim follows from $\mathbb{P}[X \leq F_X^{-1}(q)] \geq q$, the definition of the quantile $F_X^{-1}(q)$ and the fact that the CDF $F_X(\cdot)$ is right continuous. □

The next lemma gives a concentration bound for the quantile function. All works require such bounds for their sequential estimation/elimination strategies [Szörényi et al., 2015, Altschuler et al., 2019, Howard and Ramdas, 2019]. Ours differ only in how we set the constants in the inequalities. A detailed proof is given here for the convenience of the reviewer.

**Lemma 13** (Concentration Bound)**.** *Choose a level $q \in (0,1)$. Fix $\delta \in (0,1)$. For any $n \in \mathbb{N}$ if*

$$\sqrt{\frac{\log(2/\delta)}{2n}} \leq \Delta \leq q \tag{13}$$

*then*

$$\mathbb{P}\left(F_X^{-1}(q) \notin \left[X_{(\lfloor n(q-\Delta)\rfloor)}, X_{(\lceil n(q+\Delta)\rceil)}\right]\right) \leq \delta. \tag{14}$$

*Proof.* Hoeffding's inequality gives

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i < F_X^{-1}(q)} < \mathbb{P}[X < F_X^{-1}(q)] + \sqrt{\frac{\log(2/\delta)}{2n}}\right] > 1 - \delta/2 \tag{15}$$

and we have

$$\begin{aligned}
1 - \frac{\delta}{2} &< \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i < F_X^{-1}(q)} < \mathbb{P}[X < F_X^{-1}(q)] + \sqrt{\frac{\log(2/\delta)}{2n}}\right] \\
&\leq \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i < F_X^{-1}(q)} < q + \sqrt{\frac{\log(2/\delta)}{2n}}\right] \qquad (16) \\
&= \mathbb{P}\left[\sum_{i=1}^{n} \mathbf{1}_{X_i < F_X^{-1}(q)} < n\left(q + \sqrt{\frac{\log(2/\delta)}{2n}}\right)\right] \\
&\leq \mathbb{P}\left[\sum_{i=1}^{n} \mathbf{1}_{X_i < F_X^{-1}(q)} < n\left\lceil q + \sqrt{\frac{\log(2/\delta)}{2n}}\right\rceil\right] \\
&\leq \mathbb{P}\left[\sum_{i=1}^{n} \mathbf{1}_{X_i < F_X^{-1}(q)} < n\lceil q + \Delta\rceil\right], \quad \text{for any } \Delta \geq \sqrt{\frac{\log(2/\delta)}{2n}}, \qquad (17)
\end{aligned}$$

(16) comes from Proposition 12. The latter implies that

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{X_i<F_X^{-1}(q)} \geq n\lceil q+\Delta\rceil\right] \leq \delta/2, \quad \forall\Delta \geq \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{18}$$

The following statement holds

$$X_{(\lceil n(q+\Delta)\rceil)} < F_X^{-1}\left(q\right) \iff \sum_{i=1}^{n}\mathbf{1}_{X_i<F_X^{-1}(q)} \geq \lceil n(q+\Delta)\rceil, \tag{19}$$

together with (18) gives

$$\mathbb{P}\left[X_{(\lceil n(q+\Delta)\rceil)} < F_X^{-1}\left(q\right)\right] \leq \delta/2, \quad \text{for any } \Delta \geq \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{20}$$

Similarly,

$$1-\frac{\delta}{2} \leq \mathbb{P}\left[\sum_{i=1}^{n}\mathbf{1}_{X_i>F^{-1}(q)]} < n-n\lfloor q-\Delta\rfloor\right], \quad \forall\Delta \in \left[\sqrt{\frac{\log(2/\delta)}{2n}}, q\right], \tag{21}$$

and

$$X_{(\lfloor n(q-\Delta)\rfloor)} > F_X^{-1}\left(q\right) \iff \sum_{i=1}^{n}\mathbf{1}_{X_i>F^{-1}(q)} \geq n-\lfloor n(q-\Delta)\rfloor, \tag{22}$$

Then (21) gives

$$\mathbb{P}\left[X_{(\lfloor n(q-\Delta)\rfloor)} > F_X^{-1}\left(q\right)\right] \leq \delta/2, \quad \forall\Delta \in \left[\sqrt{\frac{\log(2/\delta)}{2n}}, q\right]. \tag{23}$$

Finally, (20), (23) and the union bound give the statement of the theorem. □

Recall that $n^*$ is the smallest integer that satisfies the inequality $\Delta(n^*) \leq q$. Next we prove that the event $\mathcal{E}$ defined as

$$\mathcal{E} \triangleq \bigcap_{k=1}^{K}\bigcap_{n=n^*}^{\infty}\left\{F_k^{-1}(q) \in \left[X^k_{(\lfloor n(q-\mathrm{D}(n))\rfloor)}, X^k_{(\lceil n(q+\mathrm{D}(n))\rceil)}\right]\right\} \tag{24}$$

occurs with probability at least $1-\delta$.

**Lemma 14.** *Choose a level* $q \in (0,1)$. *Fix* $\delta \in (0,1)$. *Define* $\mathrm{D}(n) \triangleq \sqrt{\frac{\log(4Kn^2/\delta)}{2n}}$ *and let* $n^*$ *be the smallest integer that satisfies the inequality* $\mathrm{D}(n^*) \leq q$. *Then the following holds*

$$\mathbb{P}\left(\mathcal{E}^c\right) = \mathbb{P}\left(\bigcup_{k=1}^{K}\bigcup_{n=n^*}^{\infty}\left\{F_k^{-1}(q) \notin \left[X^k_{(\lfloor n(q-\mathrm{D}(n))\rfloor)}, X^k_{(\lceil n(q+\mathrm{D}(n))\rceil)}\right]\right\}\right) \leq \delta. \tag{25}$$

*Proof.* Hoeffding's inequality gives

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{X_i<F_X^{-1}(q)} < \mathbb{P}[X < F_X^{-1}\left(q\right)] + \epsilon\right] > 1-e^{-2n\epsilon^2} \tag{26}$$

and

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{X_i>F^{-1}(q)} < \mathbb{P}[X > F_X^{-1}\left(q\right)] + \epsilon\right] > 1-e^{-2n\epsilon^2} \tag{27}$$

for any $n \in \mathbb{N}$ and $\epsilon > 0$ independent of $n$. For a fixed $\delta \in (0,1)$ we can choose

$$\epsilon = \sqrt{\frac{\log(4Kn^2/\delta)}{2n}}, \tag{28}$$

then (26) and (27) give

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{X_i<F_X^{-1}(q)} < \mathbb{P}[X < F_X^{-1}(q)] + \sqrt{\frac{\log(2Kn^2/\delta)}{2n}}\right] > 1 - \frac{\delta}{4Kn^2}, \tag{29}$$

and

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{X_i>F_X^{-1}(q)} < \mathbb{P}[X > F_X^{-1}(q)] + \sqrt{\frac{\log(2Kn^2/\delta)}{2n}}\right] > 1 - \frac{\delta}{4Kn^2}. \tag{30}$$

Define

$$\mathrm{D}(n) \triangleq \sqrt{\frac{\log(4Kn^2/\delta)}{2n}} \leq q, \tag{31}$$

then Lemma 13 gives

$$\mathbb{P}\left(\left\{F_k^{-1}(q) \notin \left[X^k_{(\lfloor n(q-\mathrm{D}(n))\rfloor)}, X^k_{(\lceil n(q+\mathrm{D}(n))\rceil)}\right]\right\}\right) \leq \frac{\delta}{2Kn^2} \tag{32}$$

$k \in \{1, 2, \ldots, K\}$. We conclude that

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{k=1}^{K}\bigcup_{n=n^*}^{\infty}\left\{F_k^{-1}(q) \notin \left[X^k_{(\lfloor n(q-\mathrm{D}(n))\rfloor)}, X^k_{(\lceil n(q+\mathrm{D}(n))\rceil)}\right]\right\}\right) \\
&\leq \sum_{k=1}^{K}\sum_{n=n^*}^{\infty}\mathbb{P}\left(\left\{F_k^{-1}(q) \notin \left[X^k_{(\lfloor n(q-\mathrm{D}(n))\rfloor)}, X^k_{(\lceil n(q+\mathrm{D}(n))\rceil)}\right]\right\}\right) \\
&\leq \sum_{k=1}^{K}\sum_{n=n^*}^{\infty}\frac{\delta}{2Kn^2} \\
&= \sum_{n=n^*}^{\infty}\frac{\delta}{2n^2} \\
&\leq \delta\sum_{n=1}^{\infty}\frac{1}{2n^2} \\
&= \delta.
\end{aligned} \tag{33}$$

This proves the result stated in the lemma. □

# B Proofs for the quantile bandit problem

We collect here the proofs of Proposition 4, Lemma 5, Theorems 6, 7, and 8.

## B.1 Proof of Proposition 4

**Proposition 15** (Proposition 4 restated)**.** *Suppose $F$ and $G$ are two distributions with $L$-Lipschitz continuous and strictly increasing CDFs. Then the gap $\Delta(F,G) \leq \frac{L}{2}|F^{-1}(q) - G^{-1}(q)|$.*

*Proof of Proposition 4.* By definition, we have

$$\eta = \left|F\left(F^{-1}(q+\eta)\right) - F\left(F^{-1}(q)\right)\right| \leq L\left|F^{-1}(q+\eta) - F^{-1}(q)\right| = L\left(F^{-1}(q+\eta) - F^{-1}(q)\right)$$

$$\eta = \left|G\left(G^{-1}(q)\right) - G\left(G^{-1}(q-\eta)\right)\right| \leq L\left|G^{-1}(q) - G^{-1}(q-\eta)\right| = L\left(G^{-1}(q) - G^{-1}(q-\eta)\right)$$

So

$$2\eta + L\left(G^{-1}(q-\eta) - F^{-1}(q+\eta)\right) \leq L\left(G^{-1}(q) - F^{-1}(q)\right)$$

From the definition of the gap, taking the supremum over $\eta$ gives

$$\Delta(F,G) \leq \frac{L}{2}\left|F^{-1}(q) - G^{-1}(q)\right|.$$

□

### B.2 Proof of Lemma 5

**Proposition 16.** *Let $F$ and $F'$ be two distributions such that $d_{\mathrm{TV}}(F, F') = \eta$. Then for any $q \in (\eta, 1-\eta)$ it holds that $F^{-1}(q-\eta) \le (F')^{-1}(q) \le F^{-1}(q+\eta)$.*

*Proof.* By definition, for any $x \in \mathbb{R}$ it holds that $|F(x) - F'(x)| \le \eta$. Define the set $S(F, q) = \{\xi :\ F(\xi) \ge q\}$ where $F^{-1}(q) = \inf S(F, q)$. It follows that any $\xi \in S(F', q)$ also satisfies that $F(\xi) \ge F'(\xi) - \eta \ge q - \eta$ which means $\xi \in S(F, q-\eta)$, and so $F^{-1}(q-\eta) = \inf S(F, q-\eta) \le (F')^{-1}(q) = \inf S(F', q)$. Similarly, any $\xi \in S(F, q+\eta)$ also belongs to the set $S(F', q)$ proving that $(F')^{-1}(q) \le F^{-1}(q+\eta)$. □

**Lemma 17** (Lemma 5 restated)**.** *For any $0 < q < 1$ and any two distributions $F_i$ and $F_{i^*}$ such that $F_i^{-1}(q) \le F_{i^*}^{-1}(q)$ it holds that $d_{\text{flip}}(F_i, F_{i^*}) = \Delta(F_i, F_{i^*})$ provided that $\Delta(F_i, F_{i^*}) < \min\{q, 1-q\}$.*

*Proof.* First, recall the definitions of the distance to flip (Equation (5))

$$d_{\text{flip}}(F_{\mathrm{l}}, F_{\mathrm{h}}) = \inf_{(G_{\mathrm{h}}, G_{\mathrm{l}}) : (G_{\mathrm{h}})^{-1}(q) > G_{\mathrm{l}}^{-1}(q)} \max\{d_{\mathrm{TV}}(F_{\mathrm{l}}, G_{\mathrm{h}}), d_{\mathrm{TV}}(F_{\mathrm{h}}, G_{\mathrm{l}})\}$$

and the gap (Equation (4))

$$\Delta(F_{\mathrm{l}}, F_{\mathrm{h}}) = \sup\{\eta \ge 0 :\ F_{\mathrm{l}}^{-1}(q+\eta) \le F_{\mathrm{h}}^{-1}(q-\eta)\}.$$

Now, given $\eta < \min\{q, 1-q\}$ and a distribution $F$ we define two specific shifts. The first is referred to as *$\eta$-push* of $F$ and denoted $F^{\to\eta}$ — we subtract $\eta$ probability mass from the interval $(-\infty, F^{-1}(q))$ and add $\eta$ probability mass to any point or interval in $(F^{-1}(q+2\eta), \infty)$. It is now clear that the $q$-quantile of $F^{\to\eta}$ is in fact $F^{-1}(q+\eta)$ and that $d_{\mathrm{TV}}(F, F^{\to\eta}) = \eta$ The second shift is equivalent and is an *$\eta$-pull*, denoted $F^{\leftarrow\eta}$ — we subtract $\eta$-probability mass from the interval $(F^{-1}(q), \infty)$ and move it to the interval $(-\infty, F^{-1}(q-2\eta))$. One can check that the $q$-quantile of $F^{\leftarrow\eta}$ is in fact $F^{-1}(q-\eta)$ and that $d_{\mathrm{TV}}(F, F^{\leftarrow\eta}) = \eta$.

We now prove the first part of the lemma. Denote that $\Delta(F_{\mathrm{l}}, F_{\mathrm{h}}) = \Delta$. Namely, for every $\eta > 0$ it holds that $F_{\mathrm{l}}^{-1}(q+\Delta+\eta) > F_{\mathrm{h}}^{-1}(q-\Delta-\eta)$. For any $\eta > 0$, consider the $(\Delta+\eta)$-push of $F_{\mathrm{l}}$ so that $(F_{\mathrm{l}}^{\to(\Delta+\eta)})^{-1}(q) = F_{\mathrm{l}}^{-1}(q+\Delta+\eta)$ and the $(\Delta+\eta)$-pull of $F_{\mathrm{h}}$ so that $(F_{\mathrm{h}}^{\leftarrow(\Delta+\eta)})^{-1}(q) = F_{\mathrm{h}}^{-1}(q-\Delta-\eta)$. Putting these inequalities together shows that $(F_{\mathrm{l}}^{\to(\Delta+\eta)})^{-1}(q) > (F_{\mathrm{h}}^{\leftarrow(\Delta+\eta)})^{-1}(q)$. Applying the definition of the distance to quantile flip, this shows that $d_{\text{flip}}(F_{\mathrm{l}}, F_{\mathrm{h}}) < \Delta + \eta$ for any positive $\eta$. Thus, $d_{\text{flip}}(F_{\mathrm{l}}, F_{\mathrm{h}}) \le \Delta$. Specifically, in the case where $\Delta(F_{\mathrm{l}}, F_{\mathrm{h}}) = \Delta = 0$ we have that $d_{\text{flip}}(F_{\mathrm{l}}, F_{\mathrm{h}}) = 0$.

We now show the contrapositive. Assume that $\Delta(F_{\mathrm{l}}, F_{\mathrm{h}}) > 0$. We thus have that for any $0 < \eta < \Delta(F_{\mathrm{l}}, F_{\mathrm{h}})$ it holds that $F_{\mathrm{l}}^{-1}(q+\eta) \le F_{\mathrm{h}}^{-1}(q-\eta)$. Fix any $\tilde{G}$ and $\tilde{H}$ such that $d_{\mathrm{TV}}(F_{\mathrm{l}}, \tilde{G}) \le \eta$ and $d_{\mathrm{TV}}(F_{\mathrm{h}}, \tilde{H}) \le \eta$. It follows from Proposition 16 and the definition of the gap that

$$\tilde{G}^{-1}(q) \le F_{\mathrm{l}}^{-1}(q+\eta) \le F_{\mathrm{h}}^{-1}(q-\eta) \le H^{-1}(q).$$

This shows that any pair of distributions with max TV-distance of $\eta$ to $F_{\mathrm{l}}$ and $F_{\mathrm{h}}$ is such that that the $q$-quantile has not flipped and it still holds that $\tilde{G}^{-1}(q) \le \tilde{H}^{-1}(q)$. Thus the distance to quantile flip of the pair $(F_{\mathrm{l}}, F_{\mathrm{h}})$ has to be at least $\eta$. Since this holds for any $\eta < \Delta(F_{\mathrm{l}}, F_{\mathrm{h}})$ it follows that $d_{\text{flip}}(F_{\mathrm{l}}, F_{\mathrm{h}}) \ge \Delta(F_{\mathrm{l}}, F_{\mathrm{h}})$. □

Note that our proof actually shows that the distance to a flip of a pair of distributions is *proportional* (up to a constant) to the quantile suboptimality-gap.

### B.3 Proof of Theorem 6

*Proof of Theorem 6.* Recall that $\mathrm{D}(n) = \sqrt{\log(4Kn^2/\delta)/2n}$ and $n^*$ is the smallest integer that satisfies the inequality $\mathrm{D}(n^*) \le q$. Consider the event

$$\mathcal{E} \triangleq \bigcap_{k=1}^{K} \bigcap_{n=n^*}^{\infty} \left\{ F_k^{-1}(q) \in \left[ X^k_{(\lfloor n(q-\mathrm{D}(n)) \rfloor)}, X^k_{(\lceil n(q+\mathrm{D}(n)) \rceil)} \right] \right\}. \tag{34}$$

Lemma 14 (see Section A of the supplement) gives $\mathbb{P}(\mathcal{E}) > 1-\delta$. Under the event $\mathcal{E}$ the following inequalities hold

$$F_j^{-1}(q) \geq X^j_{(\lfloor n(q-\mathrm{D}(n))\rfloor)} \text{ and } F_i^{-1}(q) \leq X^i_{(\lceil n(q+\mathrm{D}(n))\rceil)}. \tag{35}$$

Further every time that the stopping condition $X^j_{(\lfloor n(q-\mathrm{D}(n))\rfloor)} \geq X^i_{(\lceil n(q+\mathrm{D}(n))\rceil)}$ occurs we eliminate the arm $i$ and the arm $j$ remains in $\mathcal{A}$. The stopping condition and the inequalities in (35) guarantee that

$$F_j^{-1}(q) \geq F_i^{-1}(q). \tag{36}$$

As a consequence the optimal arm $i^*$ is not eliminated and the Algorithm stops when $\mathcal{A} = \{i^*\}$. □

### B.4 Proof of Theorem 7

*Proof of Theorem 7.* Under the event $\mathcal{E}$ (see definition (34)), we will find a bound on the smallest value of $n$ that satisfies the inequality $X^{i^*}_{(\lfloor n(q-\mathrm{D}(n))\rfloor)} \geq X^i_{(\lceil n(q+\mathrm{D}(n))\rceil)}$. With probability at least $1-\delta$ it is true that

$$X^{i^*}_{(\lfloor n(q-\mathrm{D}(n))\rfloor)} \geq X^{i^*}_{(\lceil n(q-\mathrm{D}(n))-1\rceil)} = X^{i^*}_{(\lceil n(q-\mathrm{D}(n)-1/n)\rceil)} \overset{(\mathrm{A})}{\geq} F_{i^*}^{-1}(q-2\mathrm{D}(n)-1/n), \tag{37}$$

$$F_i^{-1}(q+2\mathrm{D}(n)+1/n) \overset{(\mathrm{B})}{\geq} X^i_{(\lfloor n(q+\mathrm{D}(n)+1/n)\rfloor)} \geq X^i_{(\lceil n(q+\mathrm{D}(n))\rceil)} \tag{38}$$

and (A), (B) come from Lemma 13 (see Section A). From the definition of the sub-optimality gap follows that $F_{i^*}^{-1}(q-\Delta_{i^*}) \geq F_i^{-1}(q+\Delta_i)$. The latter together with (37) and (38) give that it is sufficient to find the smallest value of $n$ that satisfies the inequalities

$$F_{i^*}^{-1}(q-2\mathrm{D}(n)-1/n) \geq F_{i^*}^{-1}(q-\Delta_i) \text{ and } F_i^{-1}(q+\Delta_i) \geq F_i^{-1}(q+2\mathrm{D}(n)+1/n). \tag{39}$$

The monotonicity of $F_{i^*}(\cdot)$, $F_i(\cdot)$ and (39) give

$$\Delta_i \geq 2\mathrm{D}(n) + \frac{1}{n} \implies \Delta_i \geq 2\sqrt{\frac{\log(4Kn^2/\delta)}{2n}} + 1/n, \tag{40}$$

and the values of $n$ that satisfy the inequality above are bounded by

$$\tau_i = \mathcal{O}\left(\frac{\log \frac{K}{\delta \Delta_i}}{\Delta_i^2}\right). \tag{41}$$

To conclude, the total number of samples $\tau$ is $\sum_{i\in\mathcal{A}-i^*} \tau_i$ with probability at least $1-\delta$. □

### B.5 Proofs for the lower bound on the sample complexity

Define the following class of distributions:

$$g^w(x) = w\delta(x) + (1-w) \qquad x \in [0,1] \tag{42}$$

the mixture of a mass (Dirac delta) at 0 and a uniform distribution on $[0,1]$. Let $G^w$ be the cumulative distribution function of $g^w$. Then the mean of $g^w$ is $\frac{1-w}{2}$ and the $q$-quantile is 0 for $q \leq w$ and $\frac{q-w}{1-w}$ for $q > w$. The variance of $g^w$ is

$$\int_{x=0}^{1} (1-w)x^2 dx - \frac{(1-w)^2}{4} = (1-w)\frac{1}{3} - \frac{(1-w)^2}{4} = (1-w)\left(\frac{1}{12} + \frac{1}{4}w\right) \tag{43}$$

The KL-divergence between two such distributions is

$$\boldsymbol{D}_{\mathrm{KL}}(g^w \| g^{w'}) = w \log \frac{w}{w'} + (1-w) \log \frac{1-w}{1-w'}, \tag{44}$$

which is the same as the divergence between two Bernoulli random variables. The gap between $g^w$ and $g^{w+\gamma}$ for $q > w+\gamma$ and small $\gamma < \frac{1}{2}(q-w)$ is $\Theta(\gamma)$. To see this, let $\nu = g^w$ and $\nu' = g^{w+\gamma}$, so $\nu$ has the higher $q$-quantile. We can calculate the $(q-\eta)$-quantile of $\nu$ and the $(q+\eta)$-quantile of $\nu'$:

$$x = \frac{q-\eta-w}{1-w} \tag{45}$$

$$x' = \frac{q+\eta-(w+\gamma)}{1-(w+\gamma)}. \tag{46}$$

We need to find the inf over all $\eta$ such that $x' < x$. Taking the case of equality:

$$\frac{q-\eta-w}{1-w} = \frac{q+\eta-(w+\gamma)}{1-(w+\gamma)} \tag{47}$$

$$(q-\eta-w)(1-w-\gamma) = (q+\eta-w-\gamma)(1-w) \tag{48}$$

$$(q-w)(1-w)-\gamma(q-w)-\eta(1-w-\gamma) = (q-w)(1-w)-\gamma(1-w)+\eta(1-w) \tag{49}$$

$$\gamma(1-q) = \eta(2-2w-\gamma) \tag{50}$$

$$\eta = \frac{1-q}{2-2w-\gamma}\gamma \tag{51}$$

So for small $\gamma$ this is $\Theta(\gamma)$.

We adapt a strategy for the mean-bandit problem appearing in Lattimore and Czepesvári [2020, Section 33.2] to the quantile bandit setting. Let $\mathcal{E}$ denote a class of environments for the bandit problem and $\nu \in \mathcal{E}$ be a particular environment (i.e. setting of the arm distributions). Let $i^*(\nu)$ be the optimal arm[5] which we will denote by $i^*$ when $\nu$ is clear from context.

*Proof of Theorem 8.* Let $\nu^{(1)}$ be defined by the arm CDFs

$$\nu_i^{(1)} = \begin{cases} G^{1/3-\gamma} & i=1 \\ G^{1/3} & i \neq 1 \end{cases} \tag{52}$$

The gap between $\nu_1^{(1)}$ and $\nu_i^{(1)}$ is $\Theta(\gamma)$. For each $j$ define $\nu^{(j)}$

$$\nu_i^{(j)} = \begin{cases} G^{1/3-\gamma} & i=1 \\ G^{1/3-2\gamma} & i=j \\ G^{1/3} & i \neq 1,j \end{cases}. \tag{53}$$

Let $\pi$ be a $\delta$-PAC policy. Then we have $\mathbb{P}_{\nu^{(1)}\pi}(\hat{k} \neq 1) \leq \delta$ and $\mathbb{P}_{\nu^{(j)}\pi}(\hat{k} \neq 1) \leq \delta$. Since $\nu^{(1)}$ and $\nu^{(j)}$ differ in only a single arm distribution, we have [Lattimore and Czepesvári, 2020, Lemma 15.1]

$$\boldsymbol{D}_{\mathrm{KL}}(\mathbb{P}_{\nu^{(1)}} \| \mathbb{P}_{\nu^{(j)}}) = \sum_{i=1}^{K} \mathbb{E}_{\nu^{(1)}}[T_i(n)] \boldsymbol{D}_{\mathrm{KL}}(P_{\nu_i^{(1)}} \| P_{\nu_i^{(j)}}) \tag{54}$$

$$= \mathbb{E}_{\nu^{(1)}}[T_j(\tau)] \boldsymbol{D}_{\mathrm{KL}}(G^{1/3} \| G^{1/3-2\gamma}) \tag{55}$$

and

$$\boldsymbol{D}_{\mathrm{KL}}(G^{1/3} \| G^{1/3-2\gamma}) = \frac{1}{3}\log\frac{1/3}{1/3-2\gamma} + \frac{2}{3}\log\frac{2/3}{1/3+2\gamma} \tag{56}$$

$$= \frac{1}{3}\log\frac{1}{1-6\gamma} + \frac{2}{3}\frac{1}{1+3\gamma} \tag{57}$$

$$\leq \frac{1}{3}\left(6\gamma+18\gamma^2+54\gamma^3\right) + \frac{2}{3}\left(3\gamma+\frac{9}{2}\gamma^2\right) \tag{58}$$

$$= 9\gamma^2+18\gamma^3 \tag{59}$$

where we used the inequalities $\log\frac{1}{1-x} \leq x + \frac{x^2}{2} + \frac{x^3}{3}$ and $-\log(1+x) \leq -x + \frac{x^2}{2} - \frac{x^3}{4} \leq -x + \frac{x^2}{2}$ for $x \in [0, 0.42]$. So for $\gamma < \frac{1}{4}$,

$$\boldsymbol{D}_{\mathrm{KL}}(\mathbb{P}_{\nu^{(1)}} \| \mathbb{P}_{\nu^{(j)}}) \leq 15\gamma^2 \mathbb{E}_{\nu^{(1)}}[T_j(\tau)]. \tag{60}$$

[5] For the example in Theorem 8 there is a unique optimal arm.

Now define the events

$$A = \Big\{\tau < \infty\} \cap \{\hat{k} \neq j\}\Big\} \tag{61}$$

$$A^c = \Big\{\tau = \infty\} \cup \{\hat{k} = j\}\Big\}\,. \tag{62}$$

Then since $\{\hat{k} \neq 1\} \subseteq A^c$ and $\pi$ is $\delta$-PAC policy we have $\mathbb{P}_{\nu^{(1)}}(A^c) + \mathbb{P}_{\nu^{(j)}}(A) \leq 2\delta$.

Now, by the Bretagnolle-Huber Inequality [Lattimore and Czepesvári, 2020, Theorem 14.2],

$$2\delta \geq \frac{1}{2}\exp\left(-\boldsymbol{D}_{\mathrm{KL}}(\mathbb{P}_{\nu^{(1)}} \| \mathbb{P}_{\nu^{(j)}})\right) \tag{63}$$

$$\geq \frac{1}{2}\exp\left(-15\gamma^2 \mathbb{E}_{\nu^{(1)}}[T_j(\tau)]\right). \tag{64}$$

Rearranging,

$$\mathbb{E}_{\nu^{(1)}}[T_j(\tau)] \geq \frac{1}{15\gamma^2}\log\left(\frac{1}{4\delta}\right). \tag{65}$$

Repeating the argument for each $j \in \{2, 3, \dots K\}$ and using the fact that $\Delta = \min_k \Delta_k \geq C\gamma$ we get

$$\mathbb{E}_{\nu^{(1)}}[\tau] = \sum_{j=1}^{K} \mathbb{E}[T_j(\tau)] \geq C^2 \frac{K-1}{15\Delta^2}\log\left(\frac{1}{4\delta}\right). \tag{66}$$

□

# C Proofs for the private algorithm

In the results for the private algorithm, because of the many indices needed, we index arms by $a$ and $b$ rather than $i$ and $j$.

## C.1 Proof of Theorem 9

*Proof of Theorem 9.* Fix two input streams $S$ and $S'$. Since they differ on a single reward-entry, we denote the arm this reward was drawn for as $a$ and $e$ as the epoch on which this different reward is drawn (fix the randomness of previous epochs so that in epoch $e$ the initial set $\mathcal{A}$ of available arms is fixed at the beginning of the epoch).

Now, in epoch $e$, under the stream $S$, for any arm $b \neq a$ we consider the intervals $[i - \ell^*, i]$ and $[j, j + \ell^*]$ which holds samples such that:

$$X^b_{(j)} \leq X^b_{(j+1)} \leq \dots \leq X^b_{(j+\ell^*)} \leq X^a_{(i-\ell^*)} \leq X^a_{(i-\ell^*+1)} \leq \dots \leq X^a_{(i)} \leq X^a_{(i+1)} \tag{67}$$

and moreover, due to the maximality of $\ell^*$ we have that $X^b_{(j+\ell^*+1)} > X^a_{(i-\ell^*-1)}$ implying that

$$X^b_{(j+\ell^*+2)} \geq X^b_{(j+\ell^*+1)} > X^a_{(i-\ell^*-1)} \geq X^a_{(i-\ell^*-2)}$$

The query on the data that we are approximating with differential privacy is $q_{a,b}(S) = \max\{\ell \leq \min\{i, n-j\} : X^b_{(j+\ell)} \leq X^a_{(i-\ell)}\}$. We claim this query has global sensitivity of 1.

Consider a neighboring stream $S'$ where arm $a$ differs from $S$ on one reward (whereas arm $b$ has the exact same reward sequence). That is, under $S'$ we may add or subtract a sample to the sequence of samples that fall in positions $[i - \ell^* - 1, i]$, but the remaining shifted rewards will still satisfy (67). Thus, it must still hold that all samples in positions $[i - \ell^* + 1, i]$ and $[j, j + \ell^* - 1]$ satisfy the chain of inequalities. Moreover, even if under stream $S'$ the reward $X^a_{(i-\ell^*-1)}$ is replaced by a different one (or if it shifts one position up or down) we still have that $X^b_{(j+\ell^*+2)} > X^a_{(i-\ell^*-2)}$ and so the sequence of examples in positions above $j$ of arm $b$ that are smaller than the sequence of examples of arm $a$ that are in positions below $i$ must be contained in the shifted intervals $[i - \ell^* - 1, i]$ and $[j, j + \ell^* + 1]$ respectively. Thus in $S'$ the position of $\ell^*$ may shift by no more than 1. This shows that the query has global sensitivity 1, as desired.

It follows then that the differentially private approximation $q_{a,b}(S) + \mathsf{Lap}(2(|\mathcal{A}| - 1)/\epsilon)$ preserves $\epsilon/2(|\mathcal{A}| - 1)$-DP. Since arm $a$ participates in at most $2(|\mathcal{A}| - 1)$ many such queries in epoch $e$, we have by direct composition that our algorithm is $\epsilon$-DP. □

### C.2 Proofs of Theorem 10 and Theorem 11

We continue by providing the proof of Theorem 10.

#### C.2.1 Proof of Theorem 10

Fix an epoch $e$. We denote the following events:

$E_1$ : There exists an arm $a$ s.t. $X^a_{(i)} > F_a^{-1}\left(i/n_e + \frac{\Gamma_e}{4}\right)$ or $X^a_{(\lfloor i - n_e\frac{\Gamma_e}{4}\rfloor)} < F_a^{-1}\left(i/n_e - \frac{\Gamma_e}{2}\right)$

$E_2$ : There exists an arm $a$ s.t. $X^a_{(j)} < F_a^{-1}\left(j/n_e - \frac{\Gamma_e}{4}\right)$ or $X^a_{(\lceil j + n_e\frac{\Gamma_e}{4}\rceil)} > F_a^{-1}\left(j/n_e + \frac{\Gamma_e}{2}\right)$

$E_3$ : There exists a pair of distinct arms $a, b$ s.t. $|Z_{a,b}| > \frac{2(|\mathcal{A}|-1)}{\epsilon} \cdot \log(6|\mathcal{A}|^2 e^2/\delta)$

By Lemma 13 it holds that for a given arm $a$ and any specific index $k$, it holds that

$$\mathbb{P}\left[X^a_{(k)} < F_a^{-1}\left(k/n - \frac{\Gamma_e}{4}\right)\right] \leq \frac{\delta}{12e^2|\mathcal{A}|}$$

since $n_e \geq 8\log(12e^2|\mathcal{A}|/\delta)/\Gamma_e^2$, and similarly that

$$\mathbb{P}\left[X^a_{(k)} > F_a^{-1}\left(k/n + \frac{\Gamma_e}{4}\right)\right] \leq \frac{\delta}{12e^2|\mathcal{A}|}. \tag{68}$$

Applying the union bound over the $2|\mathcal{A}|$ choices for an arm and the two particular indices $k = i$ and $k = \lfloor i - n_e\frac{\Gamma_e}{4}\rfloor$, we have that $\mathbb{P}[E_1] \leq \delta/6e^2$. Similarly, the same line of reasoning gives that $\mathbb{P}[E_2] \leq \delta/6e^2$. Lastly, due to the properties of the Laplace distribution (or the exponential distribution which dictates the magnitude of $|Z_{a,b}|$ we have that $\mathbb{P}[E_3] \leq |\mathcal{A}|^2\delta/6e^2|\mathcal{A}|^2 = \delta/6e^2$. We apply the union bound again (twice) to infer that $\mathbb{P}[E_1 \cup E_2 \cup E_3] \leq \delta/2e^2$, and thus, the probability that

$$\mathbb{P}\left[\exists e \text{ where either } E_1, E_2 \text{ or } E_3 \text{ hold}\right] \leq \sum_{e\geq 0} \delta/2e^2 \leq \delta. \tag{69}$$

We continue under the assumption that in all epochs all three bad events never occur. Also by our choice of $n_e$ it is true that $8(|\mathcal{A}|-1)\epsilon^{-1}\log(6|\mathcal{A}|^2e^2/\delta) \leq 16(|\mathcal{A}|-1)\epsilon^{-1}\log(6|\mathcal{A}|e^2/\delta) \leq n_e\Gamma_e/4$. It is now fairly straightforward to argue that when comparing a suboptimal arm $a$ and an optimal arm $b$ we never remove $b$: this follows from the fact that in this case we have

$$X^b_{(j)} \geq F_b^{-1}\left(\frac{j}{n_e} - \frac{\Gamma_e}{4}\right) \geq F_b^{-1}(q) > F_a^{-1}(q) \geq F_a^{-1}\left(\frac{i}{n_e} + \frac{\Gamma_e}{4}\right) \geq X^a_{(i)}$$

and so for such a pair $\ell^* = 0$, making $\ell^* + Z_{a,b} \leq 2(|\mathcal{A}|-1)\log(6|\mathcal{A}|^2e^2/\delta)/\epsilon$ under the complement of $E_3$. Thus, we can only eliminate an optimal arm when comparing it to another optimal arm, and so $\mathcal{A}$ must always contain at least one optimal arm. Secondly, when comparing an optimal arm $a$ to a suboptimal arm $b$ where the optimality gap is at least $2^{-e}$ we have that at epoch $e$ it holds that for $\ell = 6(|\mathcal{A}|-1)\log(6|\mathcal{A}|^2e^2/\delta)\epsilon$ we have

$$\begin{aligned} X^b_{(j)} \leq X^b_{(j+1)} \leq ... \leq X^b_{(j+\ell)} \leq & X^b_{(\lceil j + n_e\frac{\Gamma_e}{4})\rceil} \leq F_b^{-1}(q + \Gamma_e) \\ & \leq F_a^{-1}(q - \Gamma_e) \leq X^a_{(\lfloor i - n_e\frac{\Delta}{4}\rfloor)} \leq X^a_{(i-\ell)} \leq ... \leq X^a_{(i)}. \end{aligned} \tag{70}$$

It follows that for such a pair $\ell^* \geq 6(|\mathcal{A}|-1)\log(6|\mathcal{A}|^2e^2/\delta)/\epsilon$, so under $E_3$ we have that $\ell^* + Z_{a,b} \geq 6(|\mathcal{A}|-1)\log(4|\mathcal{A}|^2e^2/\delta)/\epsilon$ so we eliminate arm $b$. The latter and (70) completes the proof. □

#### C.2.2 Proof of Theorem 11

Fix any suboptimal arm $a$. Denote $e^*$ as the first integer $e$ for which $2^{-e} \leq \Delta_a$. Thus $\Delta_{e^*} = 2^{-e^*} \leq \Delta_a < 2\Delta_{e^*}$ making $2^{e^*} \leq 2/\Delta_a$. According to Theorem 10 we have that w.p. at least $1-\delta$ by

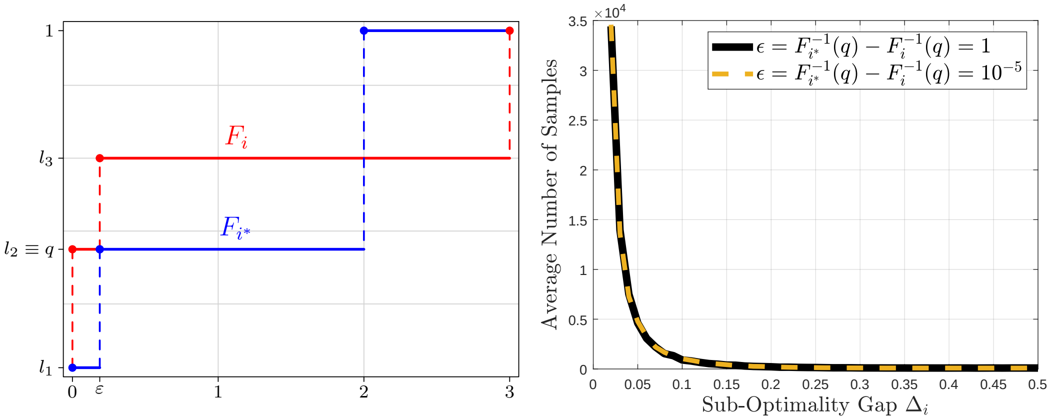

Figure 1: Left: Distributions $F_{i^*}(\cdot), F_i(\cdot)$ of the optimal and suboptimal arm. Note that $0 = F_i^{-1}(q) < F_{i^*}^{-1}(q) = \epsilon$, for some $\epsilon \in (0, 1]$. However, the gap is $\Delta_i = \min\{l_3 - q, q - l_1\}$, independent of $\epsilon$. Right: Experimental evaluation of the average number of samples at termination of Algorithm 1 for $\epsilon = 1$ and $\epsilon = 10^{-5}$.

epoch $e^*$ arm $a$ is eliminated. Since in any epoch we have that $|\mathcal{A}| \le K$, we have that the total number of pulls of arm $a$ is

$$\begin{aligned}
\sum_{0\le e\le e^*} n_e &\le \sum_{0\le e\le e^*} \left(\frac{16}{\Gamma_e^2} + \frac{64(K-1)}{\Gamma_e \cdot \epsilon}\right) \cdot \log\left(\frac{6Ke^2}{\delta}\right) \\
&\le 16 \log\left(\frac{6K(e^*)^2}{\delta}\right) \sum_{0\le e\le e^*} 2^{2e} + \frac{64K \log\left(\frac{6K(e^*)^2}{\delta}\right)}{\epsilon} \sum_{0\le e\le e^*} 2^e \\
&\le 32 \log\left(\frac{6K(e^*)^2}{\delta}\right) 2^{2e^*} + \frac{128K \log\left(\frac{6K(e^*)^2}{\delta}\right)}{\epsilon} 2^{e^*} \\
&\le \frac{128 \log\left(\frac{6K(e^*)^2}{\delta}\right)}{\Delta_a^2} + \frac{256K \log\left(\frac{6K(e^*)^2}{\delta}\right)}{\epsilon \Delta_a} \\
&\le \left(\frac{1}{\Delta_a^2} + \frac{K}{\epsilon \Delta_a}\right) \cdot 512 \log\left(\frac{6K}{\delta} \cdot \log\left(\frac{1}{\Delta_a}\right)\right).
\end{aligned}$$

To conclude, the total number of samples (and pulls) is

$$\mathcal{O}\left(\sum_{a\in\mathcal{A}^{-i^*}} \left(\frac{1}{\Delta_a^2} + \frac{K}{\epsilon \Delta_a}\right) \cdot \log\left(\frac{K}{\delta} \log\left(\frac{1}{\Delta_a}\right)\right)\right) \tag{71}$$

with probability at least $1 - \delta$. □

# D Numerical Illustrations and Further Discussion of the Results

In this section, we provide indicative numerical simulations exploring and confirming various properties related to the proposed elimination algorithms (private and non-private), as well as the proposed definition of the associated suboptimality gap. Because the focus of the paper is primarily theoretical, we do not present empirical comparisons against existing $\varepsilon$-optimal algorithms for quantile best-arm identification [Szörényi et al., 2015, David and Shimkin, 2016, Howard and Ramdas, 2019]. Nevertheless, we would like to emphasize that our work essentially complements existing literature, and provides a comprehensive treatment to the best-arm problem for quantile bandits, by providing strictly optimal and fully data-driven algorithms.

**The difference in quantile values is not informative.** As is well-known, for mean bandits the gap is simply the difference in means. However, for quantile bandits the gap between the quantile values

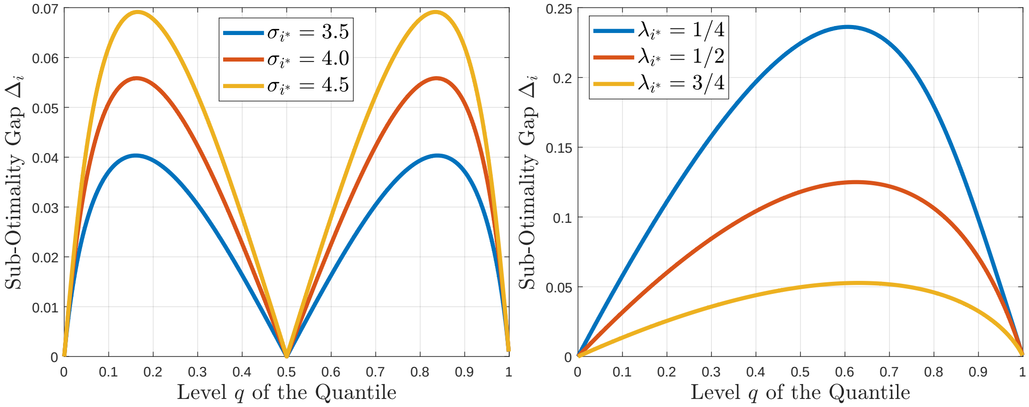

Figure 2: Illustration of the suboptimality gap for two example distributions. Up: Gaussian with $\mu_{i^*} = \mu_i = 0$, $\sigma_i = 2$ and different values of $\sigma_{i^*}$, Down: Exponential with $\lambda_i = 1$.

is not informative, as shown by Szörényi et al. [2015]. In fact, in Figure 1 we provide a case for which the difference between the quantile values can be any value $\epsilon \in (0, 1]$, however the sample complexity of SEQ remains the same. This is reasonable because the suboptimality gap depends on the levels of the CDF (see Figure 1) and not on the difference $\epsilon$ between the quantiles. However, for Lipschitz CDFs the two are related as shown in Proposition 4.

In the following example, we consider two cases (Gaussian and Log-normal distributions), for which the differences between the quantile values are different but the gap is identical for any $q \in (0, 1)$. The latter can be verified by our Definition 4. As a consequence we expect to find the same average number of pulls for Gaussian and Log-normal quantile bandits in our experiment. We can see this by comparing the performance (average termination time averaged over 500 runs) for a normal distribution and a log-normal distribution for large values of $q$, see Figure 4 (left). We take $K = 2$. The suboptimal distribution (normal or log-normal) has mean 0 and parameter $\sigma = 2$. We vary the best arm by changing $\sigma$. In our definition of gap, the gap between two normal distributions with parameters $\sigma_i$ and $\sigma_{i^*}$ is the same as the gap between two log-normal distributions with parameters $\sigma_i$ and $\sigma_{i^*}$. Each curve shows that the sample complexity when comparing normal and log-normal distributions is the same. In the case of the log-normal distributions the difference in the $q$-quantiles may be quite large. However, the sample complexity of the algorithm depends on the gap.

**Difference between the proposed gap and prior work.** Although the suboptimality gap we propose (Definition 3) may look similar to those proposed in prior works [Szörényi et al., 2015, David and Shimkin, 2016, Howard and Ramdas, 2019], there are several important differences. The proposals in Szörényi et al. [2015] and David and Shimkin [2016] explicitly incorporate the approximation parameter, whereas our definition depends only on the arm distributions. The definition in Howard and Ramdas [2019] looks like a "one-sided" version of Definition 3 but it turns out that the values can be arbirtarily different. Indeed, a comparison of our suboptimality gap $\Delta_i$ (Definition 3) with the definition of suboptimality gap $\Delta_i^{\mathrm{HR}}$ by Howard and Ramdas [2019] shows that $\Delta_i$ can be arbitrarily small while $\Delta_i^{\mathrm{HR}}$ is large. For example, in Figure 1 (Left) fix $\ell_3 - q$ and consider the case $q - \ell_1 = \epsilon'$such that $0 < \epsilon' \ll \ell_3 - q$, then $\Delta_i = \min\{l_3 - q, q - l_1\} = \epsilon'$ but $\Delta_i^{\mathrm{HR}} = l_3 - q$. As a consequence $\Delta_i \ll \Delta_i^{\mathrm{HR}}$. Experimental results also support our theory and verify that the Definition 3 of gap $\Delta_i$ is the appropriate see Figure 3 by matching the theoretical and experimental number of pulls. On the other hand, the quantity $\Delta_i^{\mathrm{HR}}$ fails to capture the correct number of pulls in examples like the one that we consider here (and more). To conclude, we also proved a lower bound based on the gap $\Delta_i$ (Theorem 8), the latter shows that our upper bound is optimal up to logarithmic factors.

**The value of the gap for different distribtions.** What does the value of the gap (4) look like? Figure 2 shows the gap as a function of the level $q$ of the quantile for two continuous distributions, the Gaussian and the exponential. We vary the optimal distribution by altering the parameter (variance or rate). For the Gaussian example (left) we look at the gap between a (suboptimal) $\mathcal{N}(0, 2)$ distribution and Gaussians of higher variance. As expected, when looking at the median the gap is 0 since they are both symmetric distributions. More interestingly, the best-arm identification problem becomes easiest

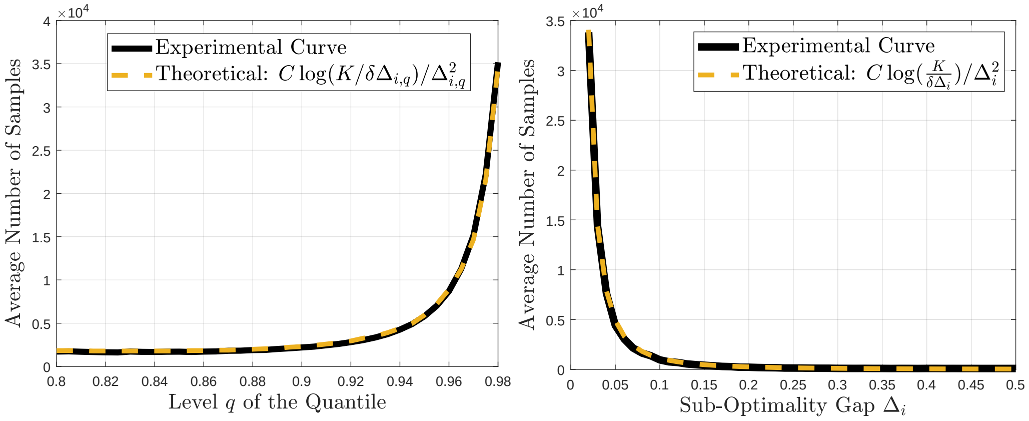

Figure 3: Left: Gaussian data, $\mu_{i^*} = \mu_i = 0$, $\sigma_{i^*} = 1$ and $\sigma_i = 0.5$, illustration of the average number of samples (and pulls) as the level of $q$ increases. Right: Discrete data, $q = 0.4$, illustration of the average number of samples (and pulls) as the suboptimality gap $\Delta_i$ increases. In both cases we use 100 in total independent runs.

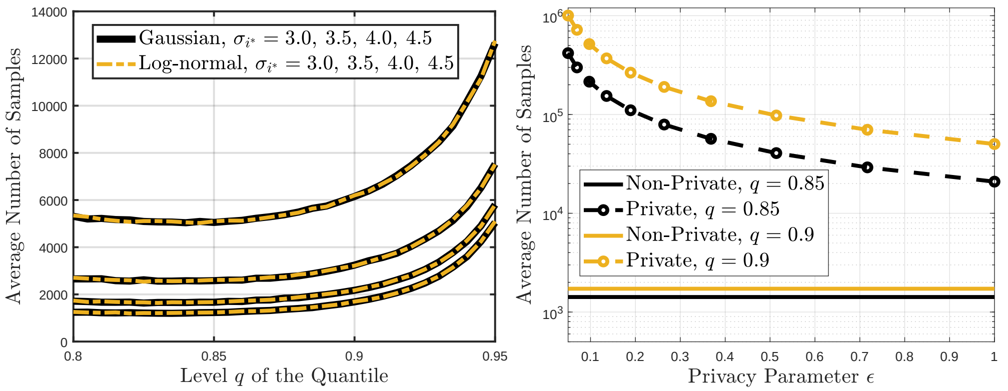

Figure 4: Left: Sample complexity of Algorithm 1 for Gaussian and Log-normal data, $K = 2$. For both cases $\mu = 0$ for all the arms, $\sigma_i = 2$. Right: Comparison of the private (Algorithm 2) and non-private SEQ (Algorithm 1). We consider Log-normal distributions with $K = 10$ arms and parameters $\mu_{i^*} = \mu_i = 0$, $\sigma_i \in [0.1, 0.2, \ldots, 0.9]$ and $\sigma_{i^*} = 2$. We compare the estimated number of samples (averaging over 10 iterations) for different values of the privacy parameter $\epsilon \in [0.05, 1]$ and for quantile levels $q = 0.85$ and $q = 0.9$.

when looking at some quantile $q^*$ (or $1 - q^*$) that lies between $1/2$ and 1. The problem becomes hard again when looking at the tails of the distribution. For the exponential distribution we compare to a rate $\lambda_i = 1$ for smaller values of the rate. As the difference in rates grows, the problem becomes easier, as expected. Here too we see an optimal $q^*$ between $1/2$ and 1 for which the top quantile is easiest to identify. Analytical expressions for these optimal points could possibly be derived through analyzing the corresponding densities. However, we defer this for future work.

**Empirical verification of the tightness of the bounds.** To empirically validate our theoretical results on the sample complexity, we show in Figure 2 the average number of samples to identify the best arm for a Gaussian (left) and discrete (right) problem setting. For both settings the average number of pulls is evaluated through 100 independent runs. These curves show that there exists a constant $C$ such that the sample complexity of the algorithm matches our analysis, specifically $C = 3/2$ for the left and right figure. For the Gaussian distribution, $\mu_{i^*} = \mu_i = 0$, $\sigma_i = 2$ while $\sigma_{i^*}$ varies. The discrete distribution is provided in Figure 1 on the left, $q = 0.4$ while the levels $\ell_1, \ell_3$ vary (see Figure 1, left).

**The cost of privacy for Algorithm 1.** Figure 4 (right) shows the performance of Algorithm 1 as a function of the privacy risk $\epsilon$. As expected, as $\epsilon$ increases the sample complexity decreases. The plots show that as the quantile decreases the gap in expected pulls between the private and non-private algorthms decreases. The high cost of privacy in this example shows that there is potential for improvement in the private algorithm: in order to get the sample complexity scaling we chose to double epoch sizes (a standard technique) but empirically we may choose a less aggressive approach.