Bernoulli **28**(3), 2022, 1704–1728
https://doi.org/10.3150/21-BEJ1388

Sequential estimation of quantiles with applications to A/B testing and best-arm identification

STEVEN R. HOWARD[1,a] and AADITYA RAMDAS[2,b]

[1]*The Voleon Group, 2150 Dwight Way, Berkley, CA 94704.* [a]*steve@stevehoward.org*
[2]*Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213.*
[b]*aramdas@stat.cmu.edu*

We design confidence sequences—sequences of confidence intervals which are valid uniformly over time—for quantiles of any distribution over a complete, fully-ordered set, based on a stream of i.i.d. observations. We give methods both for tracking a fixed quantile and for tracking all quantiles simultaneously. Specifically, we provide explicit expressions with small constants for intervals whose widths shrink at the fastest possible $\sqrt{t^{-1}\log\log t}$ rate, along with a nonasymptotic concentration inequality for the empirical distribution function which holds uniformly over time with the same rate. The latter strengthens Smirnov's empirical process law of the iterated logarithm and extends the Dvoretzky-Kiefer-Wolfowitz inequality to hold uniformly over time. We give a new algorithm and sample complexity bound for selecting an arm with an approximately best quantile in a multi-armed bandit framework. In simulations, our method needs fewer samples than existing methods by a factor of five to fifty.

Keywords: Quantile estimation; confidence sequences; empirical process; Dvoretzky-Kiefer-Wolfowitz inequality; best-arm identification

1. Introduction

A fundamental problem in statistics is the estimation of the location of a distribution based on independent and identically distributed samples. While the mean is the most common measure of location, the median and other quantiles are important alternatives. Quantiles are more robust to outliers and are well-defined for ordinal variables, and sample quantiles exhibit favorable concentration properties, which allow for strong estimation guarantees with minimal assumptions. Beyond estimation, one may choose to actively seek a distribution which maximizes a particular quantile, as in a multi-armed bandit setup, in contrast to the usual setting of finding an arm with maximal mean. In such problems, we wish to find an arm having an approximately best quantile with high probability, while minimizing the total number of samples drawn.

In this paper, we consider the sequential estimation of quantiles and its application to quantile best-arm identification. Specifically, given a stream of i.i.d. observations, we wish to form an estimate of a population quantile, or of all population quantiles, and to continuously update this estimate as more samples are observed to reflect our decreasing uncertainty. Our key tool is the *confidence sequence*: a sequence of confidence intervals which are guaranteed to contain the desired quantile uniformly over an unbounded time horizon, with the desired coverage probability. For example, if $Q(p)$ denotes the true quantile function and $\widehat{Q}_t(p)$ the sample quantile function after having observed t samples (see Section 3 for precise definitions), then for any desired coverage level $\alpha \in (0,1)$, Theorem 1(a) yields the following confidence sequence for the true median, using as confidence bounds a pair of sample quantiles at each time t:

$$\mathbb{P}\left(\forall t \in \mathbb{N} : \widehat{Q}_t(1/2 - u_t) \le Q(1/2) \le \widehat{Q}_t(1/2 + u_t)\right) \ge 1 - \alpha,$$

$$\text{where } u_t := 0.72\sqrt{t^{-1}[1.4 \log\log(2.04t) + \log(9.97/\alpha)]}. \quad (1)$$

Informally, with high probability, the (unknown) population median lies between (observed) sample quantiles slightly above and below the sample median, where "slightly" is determined by a decreasing sequence $u_t = O(\sqrt{t^{-1} \log\log t})$, and moreover, this sequence of upper and lower bounds *never* fails to contain the true median. In addition to confidence sequences for a fixed quantile, we also derive families of confidence sequences which hold uniformly both over time and over all quantiles. As an example, for any $\alpha \in (0, 0.25)$, Corollary 2 yields

$$\mathbb{P}\left(\forall t \in \mathbb{N}, p \in (0,1) : \widehat{Q}_t(p - u_t) \le Q(p) \le \widehat{Q}_t(p + u_t)\right) \ge 1 - \alpha,$$

$$\text{where } u_t := 0.85\sqrt{t^{-1}[\log\log(et) + 0.8 \log(1612/\alpha)]}. \quad (2)$$

The above closed form for u_t is one of many possibilities, but Corollary 2 offers better constants, and permits any $\alpha \in (0,1)$, if one is willing to perform numerical root-finding. For example, with $\alpha = 0.05$, we can take $u_t := 0.85\sqrt{t^{-1}(\log\log(et) + 8.12)}$ in (2).

Confidence sequences of the form (1) are critical for quantile best-arm algorithms, while those of the form (2) are highly useful for proving corresponding sample complexity bounds. We demonstrate these applications by proving a state-of-the-art sample complexity bound for a new, LUCB-style algorithm. This algorithm outperforms existing algorithms by a large margin in simulation, while the corresponding sample complexity bound matches the best-known rates and requires considerably more technical work than analogous proofs for successive elimination algorithms previously considered.

For a fixed sample size, the celebrated Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Dvoretzky, Kiefer and Wolfowitz, 1956, Massart, 1990) bounds the uniform-norm deviation of the empirical CDF from the truth with high probability. Corollary 2 follows from Theorem 2, which gives an extension of the DKW inequality that holds uniformly over time. From a theoretical point of view, Theorem 2 gives a non-asymptotic strengthening of the empirical process law of the iterated logarithm (LIL) by Smirnov (1944). From a practical point of view, as Figure 2 illustrates, our time-uniform DKW inequality of Theorem 2 is only about a factor of about two wider in the radius of the high-probability bound, relative to the fixed-sample DKW inequality. This factor grows at a slow $\sqrt{\log\log t}$ rate, so holds over a very long time horizon. Figure 1 illustrates our confidence sequences both for a fixed quantile and for the entire CDF.

Our quantile confidence sequences provide strong guarantees under minimal assumptions while granting the decision-maker a great deal of flexibility. We emphasize the following specific benefits of our confidence sequences:

(P1) **Non-asymptotic and distribution-free**: our confidence sequences offer coverage guarantees for all sample sizes in any i.i.d. sampling scenario, regardless of the underlying distribution on any totally ordered space.

(P2) **Unbounded sample size**: our methods do not require a final sample size to be chosen ahead of time. Nevertheless, they may be tuned for a planned sample size, but always permit additional sampling.

(P3) **Arbitrary stopping rules**: we make no assumptions on the rule used to decide when to stop collecting data and act on given inferences. A user may even perform inference in hindsight based on a previously-seen sample size. That is, the "stopping rule" can be any random time and does not need to be a formal stopping time.

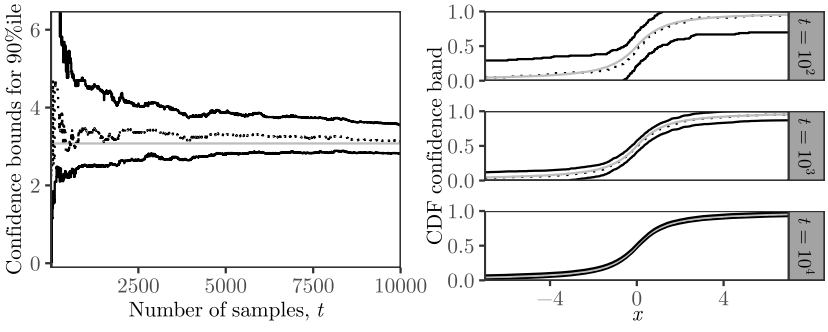


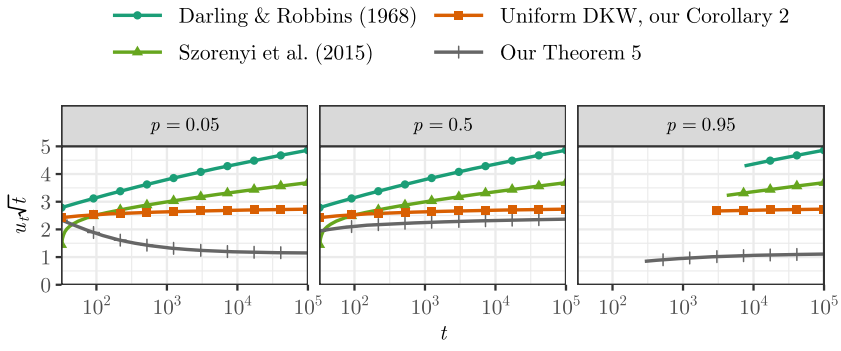
Figure 1. *Left*: solid lines show upper and lower 95%-confidence sequences using Theorem 1 for the 90%ile of a Cauchy distribution based on one sequence of i.i.d. draws. Grey line shows the true quantile, which lies between the bounds uniformly over all time $t \in \mathbb{N}$ with probability 0.95. Dotted line shows point estimates. *Right*: solid lines show 95%-confidence bands for the CDF of a Cauchy distribution at three times, $t = 100$, 1,000, and 10,000, based on one sequence of i.i.d. draws. True CDF, grey, lies between the upper and lower bounds uniformly over all $x \in \mathbb{R}$ and $t \in \mathbb{N}$ with probability 0.95. Dotted line shows empirical CDF.

(P4) **Asymptotically zero width**: our confidence bounds for the p-quantile are based on $p \pm O(t^{-1/2})$ sample quantiles, ignoring log factors. In this sense, our confidence intervals shrink in width at nearly the same rate as pointwise confidence intervals (see supplement Section 8 (Howard and Ramdas, 2021) for a simple example of pointwise confidence intervals based on the central limit theorem).

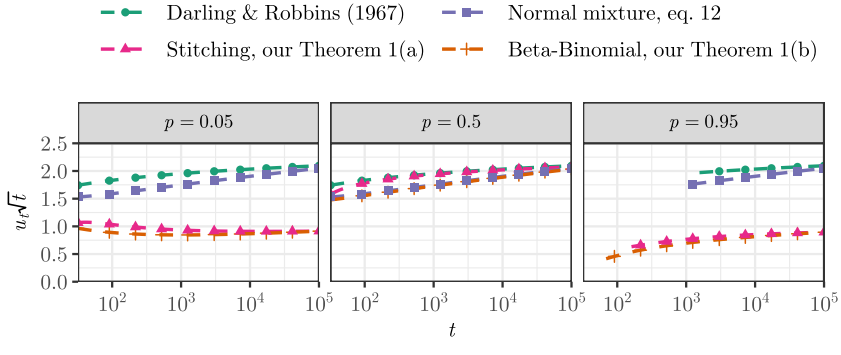
1.1. Related work

The pioneering work of Darling and Robbins (1967a) introduced the idea of a confidence sequence, as far as we are aware, and gave a confidence sequence for the median. Their method exploits a standard connection between concentration of quantiles and concentration of the empirical CDF, as does our work, and their method extends trivially to estimating any other fixed quantile. Their confidence sequence was based on the iterated-logarithm, time-uniform bound derived in Darling and Robbins (1967b), and so shrinks in width at the fastest possible $\sqrt{t^{-1} \log \log t}$ rate, like our Theorem 1(a). For the median, their constants are excellent, but the lack of dependence on which quantile is being estimated leads to looseness for tail quantiles, as illustrated in Figure 2. Our results for fixed-quantile estimation yield significantly tighter confidence sequences for tail quantiles (and are also slightly tighter for the median). Schreuder, Brunel and Dalalyan (2020) give another iterated-logarithm-rate confidence sequence for quantiles, a special case of their general method for M-estimators.

Our methods for deriving time-uniform, iterated-logarithm CDF and quantile bounds are closely related to the class of methods known as "chaining" in probability theory (Boucheron, Lugosi and Massart, 2013, Dudley, 1967, Giné and Nickl, 2016, Talagrand, 2005), and similar bounds can be derived using existing chaining techniques. We emphasize our focus on practical constants; our Theorem 2, for example, extends the fixed-sample DKW bound of Massart (1990) to hold uniformly over time at a price of roughly doubling the bound width over many orders of magnitude of time (see Figure S2 in the appendix). Our work is also related to the vast literature on extreme value theory, which contains many results on concentration of extreme sample quantiles (Anderson, 1984, Dekkers and de Haan, 1989, Drees, 1998, Drees, de Haan and Li, 2003), though not typically with our focus on time-uniform estimation. Our results can be used to estimate any population quantile, but we place no



(a) Confidence sequences uniform over both time and quantiles.



(b) Confidence sequences uniform over time for a fixed quantile.

Figure 2. Plot of upper confidence bound radii u_t, normalized by $\sqrt{t}$ to facilitate comparison. Each panel shows estimation radius for a different quantile, $p = 0.05$, 0.5, and 0.95, respectively. All bounds correspond to two-sided $\alpha = 0.05$. Upper row (a) shows confidence sequences valid uniformly over both time and quantiles. Lower row (b) shows confidence sequences valid uniformly over either time for a fixed quantile. In rightmost panels, lines start at the sample size for which the upper confidence bound becomes nontrivial. See supplement Section 5 for details of each bound shown.

particular emphasis on the behavior of extreme sample quantiles. If one were particularly interested in extreme tail behavior, e.g., in the distributional properties of the sample maximum, then such references would prove more useful. In addition, general distributional theory of order statistics (empirical quantiles) is well established (Arnold, Balakrishnan and Nagaraja, 2008), and specific variance and concentration bounds for order statistics are available (Boucheron and Thomas, 2012). Our methods are rather different in that we always bound population quantiles using sample quantiles, an approach which fits naturally into applications and which yields methods that apply universally without concern for specifics of the underlying distribution. Finally, in contrast to our focus on sequential inference for the CDF and quantiles, a line of work uses series expansions for improved point estimation of the CDF and quantile functions; see Stephanou, Varughese and Macdonald (2017) and references therein.

Shorack and Wellner (1986) give an extensive survey of results for the empirical process $(\widehat{F}_t - F)_{t=1}^{\infty}$ for uniform observations, and by extension, the empirical distribution function for any sequence of i.i.d. observations. Of particular relevance is the LIL proved by Smirnov (1944), and the proof given by Shorack and Wellner (1986), based on an improvement of a maximal inequality due to James (1975).

This maximal inequality is the key to our sophisticated non-asymptotic empirical process iterated logarithm inequality, Theorem 2. The latter leads to new quantile confidence sequences that are uniform over both quantiles and time which are significantly tighter than the earlier such bounds used for quantile best-arm identification (Szörényi et al., 2015).

The problem of selecting an approximately best arm, as measured by the largest mean, was studied by Even-Dar, Mannor and Mansour (2002) and Mannor and Tsitsiklis (2003/04), who gave an algorithm and sample complexity upper and lower bounds within a logarithmic factor of each other. The best-arm identification or pure exploration problem has received a great deal of attention since then; we mention the influential work of Bubeck, Munos and Stoltz (2009) and the proposals of Jamieson et al. (2014), Kaufmann, Cappé and Garivier (2016), and Zhao et al. (2016), whose methods included iterated-logarithm confidence sequences for means.

The problem of seeking an arm with the largest median (or other quantile), rather than mean, was first considered by Yu and Nikolova (2013), as far as we are aware. Szörényi et al. (2015) proposed the (ϵ, δ)-PAC problem formulation that we use, and gave an algorithm with a sample complexity upper bound mirroring that of Even-Dar, Mannor and Mansour, including the logarithmic factor. Szörényi et al. include a confidence sequence valid over quantiles and time, derived via a union bound applied to the DKW inequality (Dvoretzky, Kiefer and Wolfowitz, 1956, Massart, 1990), similar to the bound used by Darling and Robbins (1968, Theorem 4). Szörényi et al. also analyzed a quantile-based regret-minimization problem, recently studied by Torossian, Garivier and Picheny (2019) as well. David and Shimkin (2016) extended the sample complexity of Szörényi et al. to include dependence on the quantile being optimized, while Kalogerias et al. (2020) discuss the $\epsilon = 0$ case and give careful consideration to the gap definition appearing in the sample complexity bound. Our procedure is a variant of the LUCB algorithm by Kalyanakrishnan et al. (2012), unlike previous quantile best-arm algorithms; our analysis covers both the $\epsilon = 0$ and $\epsilon > 0$ cases; we improve the upper bounds of Szörényi et al. by replacing the logarithmic factor by an iterated-logarithm one and removing unnecessary dependence on a unique best arm's gap; and we achieve considerably better performance than prior algorithms in simulations.

1.2. Paper outline

After an introduction to the conceptual ideas of the paper in Section 2, we present our confidence sequences for estimation of a fixed quantile in Section 3, while Section 4 gives a confidence sequence for all quantiles simultaneously. Section 5 offers a graphical comparison of our bounds with each other and with existing bounds from the literature, as well as advice for tuning bounds in practice. In Section 6, we analyze a new algorithm for quantile ϵ-best-arm identification in a multi-armed bandit, with a state-of-the-art sample complexity bound. We gather proofs in Section 7. Implementations are available online for all confidence sequences presented here (https://github.com/gostevehoward/confseq), along with code to reproduce all plots and simulations (https://github.com/gostevehoward/quantilecs).

2. Warmup: Linear boundaries and quantile confidence sequences

Before stating our main results, we first walk through the derivation of a simple confidence sequence for quantiles to illustrate techniques. In effect, we spell out the less-known duality between sequential tests and confidence sequences (Howard et al., 2021), analogous to the well-known duality between fixed-time hypothesis tests and confidence intervals.

Let $(X_t)_{t=1}^{\infty}$ be a sequence of i.i.d., real-valued observations from an unknown distribution, which we assume is continuous for this section only. For a given $p \in (0,1)$, let $q \in \mathbb{R}$ be such that $\mathbb{P}(X_1 \leq q) = p$.

We wish to sequentially estimate this p-quantile, q, based on the observations (X_t). At a high level, our strategy is as follows:

1. We first imagine testing a specific hypothesis $H_{0,x} : q = x$ for some $x \in \mathbb{R}$ at a fixed sample size. Using the aforementioned duality between tests and intervals, we could construct a fixed-sample confidence interval for q consisting of all those values of $x \in \mathbb{R}$ for which we fail to reject $H_{0,x}$.
2. To test $H_{0,x}$ for some fixed x, we observe that $H_{0,x}$ is true if and only if the random variables $(1_{X_t \le x})_{t=1}^{\infty}$ are i.i.d. draws from a Bernoulli(p) distribution. Hence, if the number of samples were fixed in advance, testing $H_{0,x}$ would be equivalent to a standard parametric test: we observe a set of coin flips $(1_{X_t \le x})$, and the null hypothesis states that the bias of this coin is p. Inverting this test, as mentioned in the previous point, yields a fixed-sample confidence interval for q.
3. Instead of a fixed-sample test, we could apply a sequential hypothesis test, one which can be repeatedly conducted after each new sample X_t is observed, with the guarantee that, with the desired, high probability, we will *never* reject $H_{0,x}$ when it is true. For example, appropriate variants of Wald's Sequential Probability Ratio Test (SPRT) would suffice. Inverting such a sequential test, we upgrade our fixed-sample confidence interval to a *confidence sequence*, a sequence of confidence intervals $(\mathrm{CI}_t)_{t=1}^{\infty}$ which is guaranteed to contain q uniformly over time with high probability: $\mathbb{P}(\forall t : q \in \mathrm{CI}_t) \ge 1 - \alpha$.

To give a rigorous example, consider the random variables $\xi_t := 1_{X_t \le q}$ for $t \in \mathbb{N}$. We cannot observe ξ_t since q is unknown, but we know (ξ_t) is a sequence of i.i.d. Bernoulli(p) random variables. A standard (suboptimal, but sufficient for our current exposition) result due to Hoeffding (1963) implies that the centered random variable $\xi_1 - p$ is sub-Gaussian with variance parameter $1/4$, i.e., $\mathbb{E}e^{\lambda(\xi_1 - p)} \le e^{\lambda^2/8}$ for any $\lambda \in \mathbb{R}$. Writing $L_0 := 1$ and, for $t \in \mathbb{N}$, defining

$$L_t := \exp\left\{\lambda \sum_{i=1}^{t} (\xi_i - p) - \frac{\lambda^2 t}{8}\right\}, \tag{3}$$

we observe that $(L_t)_{t=0}^{\infty}$ is a positive supermartingale for any $\lambda \in \mathbb{R}$ (Darling and Robbins, 1967a, Howard et al., 2020). Then, for any $\alpha \in (0,1)$, Ville's inequality (Ville, 1939) yields $\mathbb{P}(\exists t \ge 1 : L_t \ge 1/\alpha) \le \alpha$, or equivalently,

$$\mathbb{P}\left(\exists t \ge 1 : \sum_{i=1}^{t} \xi_i \ge tp + \frac{\log \alpha^{-1}}{\lambda} + \frac{\lambda t}{8}\right) \le \alpha. \tag{4}$$

The sequence $\left(\frac{\log \alpha^{-1}}{\lambda} + \frac{\lambda t}{8}\right)_{t=1}^{\infty}$ gives a boundary, linear in t, which the centered process $(\sum_{i=1}^{t}(\xi_i - p))_{t=1}^{\infty}$ is unlikely to ever cross. For $\lambda > 0$, this bounds the upper deviations of the partial sums $\left(\sum_{i=1}^{t} \xi_i\right)_{t=1}^{\infty}$ above their expectations, while for $\lambda < 0$, this bounds the lower deviations. Thus by simple rearrangement, and writing

$$u_t := \frac{\log \alpha^{-1}}{\lambda t} + \frac{\lambda}{8},$$

we infer that $t(p - u_t) < \sum_{i=1}^{t} \xi_i < t(p + u_t)$ uniformly over all $t \in \mathbb{N}$ with probability at least $1 - \alpha$. Observe that $\sum_{i=1}^{t} \xi_i$ equals $|\{i \in [t] : X_i \le q\}|$, the number of observations up to time t which lie below q. So if $\sum_{i=1}^{t} \xi_i < t(p + u_t)$, then we must have $q < X^t_{(\lceil t(p+u_t) \rceil)}$, where $X^t_{(k)}$ is the k^{th} order statistic of $X_1, \ldots, X_t$. Likewise, $\sum_{i=1}^{t} \xi_i > t(p - u_t)$ implies $q > X^t_{(\lfloor t(p-u_t) \rfloor)}$. In other words, with probability at

least $1-\alpha$,

$$q \in \left(X^t_{(\lfloor t(p-u_t)\rfloor)}, X^t_{(\lceil t(p+u_t)\rceil)}\right) \quad \text{simultaneously for all } t \in \mathbb{N}, \tag{5}$$

yielding a confidence sequence for the p-quantile, q. The main drawback of this confidence sequence is that u_t does not decrease to zero as $t \uparrow \infty$, so that we do not, in general, expect the confidence sequence to approach zero width as our sample size grows without bound. In other words, the precision of this estimation strategy is unnecessarily limited. The confidence sequences of Section 3 remove this restriction by replacing the $O(t)$ boundary of (4) with a curved boundary growing at the rate $O(\sqrt{t \log t})$ or $O(\sqrt{t \log\log t})$.

3. Confidence sequences for a fixed quantile

We now state our general problem formulation, which removes the assumption that observations are real-valued or from a continuous distribution. Let $(X_i)_{i=1}^\infty$ be a sequence of i.i.d. observations taking values in some complete, totally-ordered set $(\mathcal{X}, \leq)$. We shall also make use of the corresponding relations $\geq$, $<$ and $>$ on $\mathcal{X}$. Write $F(x) \coloneqq \mathbb{P}(X_1 \leq x)$ for the cumulative distribution function (CDF), $F^-(x) \coloneqq \mathbb{P}(X_1 < x)$, and define the empirical versions of these functions $\widehat{F}_t(x) \coloneqq t^{-1}\sum_{i=1}^t 1_{X_i \leq x}$ and $\widehat{F}_t^-(x) \coloneqq t^{-1}\sum_{i=1}^t 1_{X_i < x}$. Define the (standard) upper quantile function as

$$Q(p) \coloneqq \sup\{x \in \mathcal{X} : F(x) \leq p\}$$

and the lower quantile function

$$Q^-(p) \coloneqq \sup\{x \in \mathcal{X} : F(x) < p\}.$$

Finally, define the corresponding (plug-in) upper and lower empirical quantile functions $\widehat{Q}_t(p) \coloneqq \sup\{x \in \mathcal{X} : \widehat{F}_t(x) \leq p\}$ and $\widehat{Q}_t^-(p) \coloneqq \sup\{x \in \mathcal{X} : \widehat{F}_t(x) < p\}$. We extend the empirical quantile functions to hold over domain $p \in \mathbb{R}$ by taking the convention that the supremum of the empty set is $\inf \mathcal{X}$, so that $\widehat{Q}_t(p) = \widehat{Q}_t^-(p) = \inf \mathcal{X}$ for $p < 0$ while $\widehat{Q}_t(p) = \widehat{Q}_t^-(p) = \sup \mathcal{X}$ for $p > 1$.

Fixing any $p \in (0,1)$ and $\alpha \in (0,1)$, our goal in this section is to give a $(1-\alpha)$-confidence sequence for the true quantiles $Q^-(p), Q(p)$ in terms of sample quantiles. In particular, we propose positive, real-valued sequences $l_t(p)$ and $u_t(p)$ for $t \in \mathbb{N}$, each decreasing to zero as $t \uparrow \infty$, satisfying

$$\mathbb{P}\left(\exists t \in \mathbb{N} : Q^-(p) < \widehat{Q}_t(p - l_t(p)) \text{ or } Q(p) > \widehat{Q}_t^-(p + u_t(p))\right) \leq \alpha. \tag{6}$$

Stated differently, for any $q \in [Q^-(p), Q(p)]$, we would have

$$\mathbb{P}\left(\forall t \in \mathbb{N} : q \in [\widehat{Q}_t(p - l_t(p)), \widehat{Q}_t^-(p + u_t(p))]\right) \geq 1 - \alpha. \tag{7}$$

The sequences $(l_t(p), u_t(p))_{t=1}^\infty$ characterize the lower and upper radii of the confidence intervals in "p-space", before passing through the sample quantile functions $\widehat{Q}_t$ and $\widehat{Q}_t^-$ to obtain final confidence bounds in $\mathcal{X}$. In what follows, we characterize the asymptotic rates of our confidence interval widths in terms of these "p-space" widths.

Before stating our confidence sequences, we observe the following lower bound, a straightforward consequence of the law of the iterated logarithm.

Proposition 1 (Quantile confidence sequence lower bound). *Consider any* $p \in (0,1)$ *such that* $F(Q(p)) = p$. *If*

$$\limsup_{t\to\infty} \frac{u_t}{\sqrt{2p(1-p)t^{-1}\log\log t}} < 1, \tag{8}$$

then $\mathbb{P}(\exists t \in \mathbb{N} : Q(p) \geq \widehat{Q}_t(p+u_t)) = 1$.

This result is proved in supplement Section 4.2. Note that the condition $F(Q(p)) = p$ holds for all $p \in (0,1)$ when F is continuous, and holds for at least some p otherwise; more technical effort can be expended to remove this restriction, but we do not do this since the takeaway message is already transparent.

We now propose two confidence sequences. The first has radii given by the function

$$f_t(p) \coloneqq 1.5\sqrt{p(1-p)\ell(t)} + 0.8\ell(t) \quad \text{where} \quad \ell(t) \coloneqq \frac{1.4\log\log(2.1t) + \log(10/\alpha)}{t}. \tag{9}$$

This method has the advantage of a closed-form expression with small constants, and evidently $f_t(p) \sim \sqrt{3.15p(1-p)t^{-1}\log\log t}$ as $t \to \infty$, matching the lower bound given in Proposition 1 up to the leading constant. Section 7.1 gives a more general version of $f_t(p)$ involving several hyperparameters, showing that the leading constant may in fact be brought arbitrarily close to the optimal value of two appearing in Proposition 1, though doing so tends to yield inferior performance in practice. The derivation of $f_t(p)$ relies on a method that goes by different names — chaining, "peeling", or "stitching" — in which we divide time into geometrically-spaced epochs $[\eta^k, \eta^{k+1})$, and bound the miscoverage event within the k^{th} epoch by a probability which decays like k^{-s}, for hyperparameters $\eta, s > 1$ described in Section 7.1.

Our second method uses a function $\widetilde{f}_t(p)$ which requires numerical root-finding to compute exactly, but has the asymptotic expansion

$$\widetilde{f}_t(p) = \sqrt{\frac{p(1-p)}{t}\left[\log\left(\frac{p(1-p)t}{C_{p,r}^2\alpha^2}\right) + o(1)\right]} \quad \text{where} \quad C_{p,r} \coloneqq \sqrt{2\pi}p(1-p)f_\beta\left(p; \frac{r}{1-p}, \frac{r}{p}\right), \tag{10}$$

as $t \to \infty$; here $f_\beta(x; a, b)$ denotes the density of the Beta distribution with parameters a, b, and $r > 0$ is a tuning parameter. The function $\widetilde{f}_t(p)$ is described fully in Section 7.1, while we discuss the choice of the tuning parameter r in Section 5 and derive the asymptotic expansion (10) in supplement Section 4.1. We note here that as p approaches zero or one, the constant $C_{p,r}$ approaches a constant depending only on r, so it does not contribute to dependence on p for tail quantiles. Compared to $f_t(p)$, $\widetilde{f}_t(p)$ yields confidence interval widths with a slightly worse asymptotic rate of $O(\sqrt{t^{-1}\log t})$. Even though neither of our methods uniformly dominates the other, the worse rate is usually preferable in practice, as we explore in Section 5. Informally, the reason is that any method with asymptotically optimal rates must be looser at practically relevant sample sizes in order to gain this later tightness, since the overall probability of error of both envelopes can be made arbitrarily close to α. The following result shows that both the above methods yield valid confidence sequences for any fixed p.

Theorem 1 (Confidence sequence for a fixed quantile). *Taking* f_t *from* (9), *for any* $p \in (0,1)$ *and any* $\alpha \in (0,1)$, *we have*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : Q^-(p) < \widehat{Q}_t\left(p - f_t(1-p)\right) \text{ or } Q(p) > \widehat{Q}_t^-\left(p + f_t(p)\right)\right) \leq \alpha. \tag{11}$$

The same holds with $\widetilde{f}_t$ *from* (45) *(asymptotically,* (10)*) in place of* f_t.

The proof, given in Section 7.1, involves constructing a martingale having bounded increments as a function of the true quantiles $Q^-(p)$ and $Q(p)$. Then uniform concentration arguments show that $f_t(p)$ and $\widetilde{f}_t(p)$ bound the deviations of this martingale from zero, uniformly over time, with high probability. We deduce plausible values for the true quantiles from this high-probability restriction on the values of the martingale.

We could derive a simpler boundary from a sub-Gaussian bound, like that presented in the previous section. For example, one can replace $f_t(p)$ or $\widetilde{f}_t(p)$ with

$$\sqrt{\frac{t+r}{t^2}\log\left(\frac{t+r}{\alpha^2 r}\right)} \tag{12}$$

for any $r > 0$ (e.g., Howard et al., 2021, eq. 3.7). This bound lacks the appropriate dependence on $\sqrt{p(1-p)}$ indicated in Proposition 1. Our method instead uses "sub-gamma" (for f_t) and "sub-Bernoulli" (for $\widetilde{f}_t$) bounds (Howard et al., 2020) to achieve the correct dependence. The presented bounds are never looser than those obtained by a sub-Gaussian argument, and will be much tighter when p is close to zero or one, as we later illustrate in Figure 2(b).

Having presented our confidence sequences for a fixed quantile, we next present bounds that are uniform over both quantiles and time.

4. Confidence sequences for all quantiles simultaneously

Theorem 1 is useful when the experimenter has decided ahead of time to focus attention on a particular quantile, or perhaps a small number of quantiles (via a union bound). In some cases, however, it may be preferable to estimate all quantiles simultaneously, so that the experimenter may adaptively choose which quantiles to estimate after seeing the data. Equivalently, one may wish to bound the deviations of all sample quantiles uniformly over time, as in our proof of Theorem 3 in Section 6. Recall that for a fixed time t and $\alpha \in (0,1)$, the DKW inequality (Dvoretzky, Kiefer and Wolfowitz, 1956, Massart, 1990) states

$$\mathbb{P}\left(\left\|\widehat{F}_t - F\right\|_\infty > \sqrt{\frac{\log(2/\alpha)}{2t}}\right) \le \alpha. \tag{13}$$

In tandem with the implications in (34) of Section 7, the DKW inequality yields

$$\mathbb{P}\left(\exists p \in (0,1) : Q^-(p) < \widehat{Q}_t^-(p - l_t) \text{ or } Q(p) > \widehat{Q}_t(p + u_t)\right) \le \alpha, \tag{14}$$

where $l_t = u_t = \sqrt{\log \alpha^{-1}/(2t)}$. In this section, we derive a $(1-\alpha)$-confidence sequence which is valid uniformly over both quantiles and time, based on a function sequence $l_t(p), u_t(p)$ decreasing to zero pointwise as $t \uparrow \infty$:

$$\mathbb{P}\left(\exists t \in \mathbb{N}, p \in (0,1) : Q^-(p) < \widehat{Q}_t^-(p - l_t(p)) \text{ or } Q(p) > \widehat{Q}_t(p + u_t(p))\right) \le \alpha. \tag{15}$$

Our method is based on the following non-asymptotic iterated logarithm inequality for the empirical process $(\widehat{F}_t - F)_{t=1}^\infty$, which may be of independent interest.

Theorem 2 (Empirical process finite LIL bound). *For any initial time $m \geq 1$ and $C \geq 7$, we have*

$$\mathbb{P}\left(\exists t \geq m : \left\|\widehat{F}_t - F\right\|_\infty > 0.85\sqrt{\frac{\log\log(et/m) + C}{t}}\right) \leq 1612e^{-1.25C}. \tag{16}$$

Furthermore, for any $A > 1/\sqrt{2}$, $C > 0$, and $m \geq 1$, we have

$$\mathbb{P}\left(\left\|\widehat{F}_t - F\right\|_\infty > A\sqrt{\frac{\log\log(et/m) + C}{t}} \text{ infinitely often}\right) = 0. \tag{17}$$

We give a more detailed result and proof in Section 7.2, based on a maximal inequality due to James (1975) and Shorack and Wellner (1986) combined with a union bound over exponentially-spaced epochs. Taking A arbitrarily close to $1/\sqrt{2}$ immediately implies the following asymptotic upper LIL.

Corollary 1 (Smirnov, 1944). *For any (possibly discontinuous) F, we have*

$$\limsup_{t\to\infty} \frac{\|\widehat{F}_t - F\|_\infty}{\sqrt{(1/2)t^{-1}\log\log t}} \leq 1 \text{ almost surely}. \tag{18}$$

A comprehensive overview of results for the empirical process $\sqrt{t}(\widehat{F}_t - F)$ can be found in Shorack and Wellner (1986). We mention in particular the law of the iterated logarithm derived by Smirnov (1944) (cf. Shorack and Wellner, 1986, page 12, equation (11)), which says that for continuous F, the bound (18) holds with equality, seeing as the lower bound on the lim sup follows directly from the original LIL (Khintchine, 1924) applied to $\widehat{F}_t(Q(1/2))$, an average of i.i.d. Bernoulli(1/2) random variables. Theorem 2 strengthens Smirnov's asymptotic upper bound to one holding uniformly over time, without costing constant factors in the resulting asymptotic implication.

The following confidence sequence follows from Theorem 2, as detailed in supplement Section 4.4.

Corollary 2 (Quantile-uniform confidence sequence I). *For any initial time $m \geq 1$ and $C \geq 7$, letting $g_t := 0.85\sqrt{t^{-1}(\log\log(et/m) + C)}$, we have*

$$\mathbb{P}\left(\exists t \geq m, p \in (0,1) : Q^-(p) < \widehat{Q}_t^-(p - g_t) \text{ or } Q(p) > \widehat{Q}_t(p + g_t)\right) \leq 1612e^{-1.25C}. \tag{19}$$

For example, take $m = 1$ and $C = 8.3$, so that $g_t = 0.85\sqrt{t^{-1}(\log\log(et) + 8.3)}$ and

$$\mathbb{P}\left(\exists t \geq 1, p \in (0,1) : Q^-(p) < \widehat{Q}_t^-(p - g_t) \text{ or } Q(p) > \widehat{Q}_t(p + g_t)\right) \leq 0.05. \tag{20}$$

Figure 2(a) shows that Corollary 2 yields a other published methods based on the fixed-time DKW inequality combined with a more naive union bound over time.

Note that g_t does not depend on p, like the DKW-based fixed-time inequality (14), but this is not ideal for tail quantiles. We describe an alternative "double stitching" method in Theorem S1 of supplement Section 2 which does include such dependence, and yields improved performance for p near zero or one. We demonstrate this performance in Figure 2 of the following section, graphically comparing all of our bounds.

5. Graphical comparison of bounds

Figure 2 compares our four quantile confidence sequences with a variety of alternatives from the literature. In each case, we show the upper confidence bound radius u_t which satisfies $\widehat{Q}_t(p+u_t) \geq Q(p)$ with high probability, uniformly over t, p, or both. Figure S2 in supplement Section 5 includes an additional plot with all bounds together, along with details on all bounds displayed.

Among bounds holding uniformly over both time and quantiles, Corollary 2 and Theorem S1 yield the tightest bounds outside of a brief time window near the start. The bound of Theorem S1 gives u_t growing at an $O(\sqrt{t^{-1}\log t})$ rate for all $p \neq 1/2$, which is worse than that of Corollary 2, but the superior constants of Theorem S1 and its dependence on p give it the advantage in the plotted range. Szörényi et al. (2015) also give a bound which grows as $O(\sqrt{t^{-1}\log t})$, but with worse constants due to the application of a union bound over individual time steps $t \in \mathbb{N}$. A similar technique was employed by Darling and Robbins (1968, Theorem 4), but using worse constants in the DKW bound, as their work preceded Massart (1990). Finally, Corollary 2 gives an $O(\sqrt{t^{-1}\log\log t})$ bound which is especially useful for theoretical work, as in our proof of Theorem 3.

Among bounds holding uniformly over time for a fixed quantile, the beta-binomial confidence sequence of Theorem 1(b) performs best over the plotted range, slightly outperforming its iterated-logarithm counterpart from Theorem 1(a). It is evident, though, that the iterated-logarithm bound will become tighter for large enough t, thanks to its smaller asymptotic rate. Darling and Robbins (1967a, Section 2) give a similar bound based on a sub-Gaussian uniform boundary, which is only slightly worse than Theorem 1(a) for the median, but substantially worse for p near zero and one.

Figure 2 starts at $t = 32$ and all bounds have been tuned to optimize for, or start at, $t = 32$, in order to ensure a fair comparison. For Theorem 1(a), Corollary 2, and Theorem S1, we simply set $m = 32$. For Theorem 1(b), we suggest setting r as follows to optimize for time $t = m$:

$$\frac{r}{p(1-p)} = \frac{m}{-W_{-1}(-\alpha^2/e) - 1} - 1 \approx \frac{m}{2\log(\alpha^{-1}) + \log\log(e\alpha^{-2})} - 1, \tag{21}$$

where $W_{-1}(x)$ is the lower branch of the Lambert W function, the most negative real-valued solution in z to $ze^z = x$, and the second expression uses the asymptotic expansion of W_{-1} near the origin (Corless et al., 1996). See Howard et al. (2021, Proposition 3, Proposition 7, and discussion therein) for details on this choice. Figure 3 illustrates the effect of this choice. The confidence radius u_t gets loose very quickly for values of t lower than about $m/2$, but grows quite slowly for values of $t > m$. For this reason, we suggest setting m around the smallest sample size at which inferences are desired.

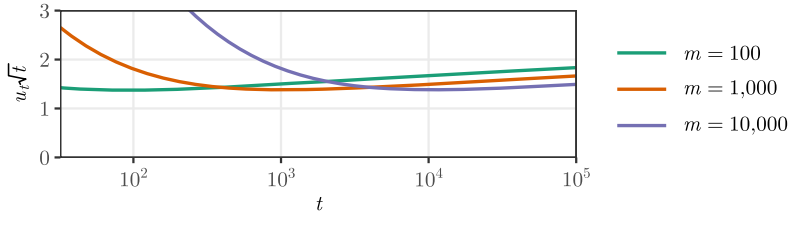


Figure 3. Plot of upper confidence bound radii u_t, normalized by $\sqrt{t}$ to facilitate comparison, for the confidence sequence of Theorem 1(b) optimized for three different times $m = 100$, 1,000, and 10,000, according to (21).

6. Quantile ϵ-best-arm identification

As an application of our quantile confidence sequences, we present and analyze a novel algorithm for identifying an arm with an approximately optimal quantile in a multi-armed bandit setting. Our problem setup matches that of Szörényi et al. (2015), a slight modification of the standard stochastic multi-armed bandit setting. We assume K arms are available, numbered $k = 1, \ldots, K$, each corresponding to a distribution F_k over the sample space $\mathcal{X}$. At each round, the algorithm chooses any arm k to pull, receiving an independent sample from the distribution F_k. Write Q_k for the upper quantile function on arm k, $Q_k(p) \coloneqq \sup\{x \in \mathcal{X} : F_k(x) \le p\}$, and Q_k^- for the lower quantile function. Fixing some $\pi \in (0,1)$, the goal is to stop as soon as possible and, with probability at least $1 - \delta$, select an ϵ-optimal arm according to the following definition:

Definition 1. *For $\epsilon \in [0, 1-\pi)$, we say arm k is* ϵ-optimal *if*

$$Q_k^-(\pi + \epsilon) \ge \max_{j \in [K]} Q_j^-(\pi - \epsilon). \tag{22}$$

Denote the set of ϵ-optimal arms by

$$\mathcal{A}_\epsilon \coloneqq \left\{ k \in [K] : Q_k^-(\pi + \epsilon) \ge \max_{j \in [K]} Q_j^-(\pi - \epsilon) \right\}.$$

Kalyanakrishnan et al. (2012) introduced the LUCB algorithm for highest mean identification, for which Jamieson and Nowak (2014) gave a simplified analysis in the $\epsilon = 0$ case. Both are key inspirations for our QLUCB (Quantile LUCB) algorithm and following sample complexity analysis. Despite being conceptually similar, our analysis faces significantly harder technical hurdles due to the nonlinearity and nonsmoothness of quantiles compared to the (sample and population) mean.

QLUCB proceeds in rounds indexed by t. At the start of round t, $N_{k,t}$ denotes the number of observations from arm k. Write $X_{k,i}$ for the i^{th} observation from arm k, and let $\widehat{Q}_{k,t}(p)$ denote the upper sample quantile function for arm k at round t,

$$\widehat{F}_{k,t}(x) \coloneqq N_{k,t}^{-1} \sum_{i=1}^{N_{k,t}} 1_{X_{k,i} \le x}, \qquad \widehat{Q}_{k,t}(p) \coloneqq \sup\left\{ x \in \mathcal{X} : \widehat{F}_{k,t}(x) \le p \right\}, \tag{23}$$

with an analogous definition of $\widehat{Q}_{k,t}^-$. QLUCB requires a sequence $(l_n(p), u_n(p))$ which yields fixed-quantile confidence sequences, as in (6). Our analysis is based on confidence sequences given by (9), by using $\alpha \equiv 2\delta/K$; the factor of two gives us one-sided instead of two-sided coverage at level δ/K, which is all that is needed. Let

$$f_t(p) = 1.5\sqrt{p(1-p)\ell(t)} + 0.8\ell(t), \text{ where } \ell(t) = \frac{1.4 \log\log(2.1t) + \log(5K/\delta)}{t}, \tag{24}$$

similar to (9), and let $l_t(p) \coloneqq f_t(1-p)$ and $u_t(p) \coloneqq f_t(p)$. We write $L_{k,t}^{\pi+\epsilon}$ and $U_{k,t}^{\pi-\epsilon}$ for the lower and upper confidence sequences on $Q_k(\pi + \epsilon)$ and $Q_k(\pi - \epsilon)$, respectively:

$$L_{k,t}^{\pi+\epsilon} \coloneqq \widehat{Q}_{k,t}\left(\pi + \epsilon - l_{N_{k,t}}(\pi + \epsilon)\right), \tag{25}$$

$$U_{k,t}^{\pi-\epsilon} \coloneqq \widehat{Q}_{k,t}^-\left(\pi - \epsilon + u_{N_{k,t}}(\pi - \epsilon)\right). \tag{26}$$

Input target quantile $\pi \in (0,1)$, approximation error $\epsilon \in [0, \pi \wedge (1-\pi))$, and error probability $\delta \in (0,1)$.
Sample each arm once, set $N_{k,1} = 1$ for all $k \in [K]$ and set $t = 1$.
while $L_{k,t}^{\pi+\epsilon} < \max_{j \neq k} U_{j,t}^{\pi-\epsilon}$ for all $k \in [K]$ **do**,
  Set $h_t \in \arg\max_{k \in [K]} L_{k,t}^{\pi+\epsilon}$ and $\mathcal{L}_t = \arg\max_{k \in [K] \setminus h_t} U_{k,t}^{\pi-\epsilon} \subseteq [K]$.
  Sample all arms in $\{h_t\} \cup \mathcal{L}_t$.
  Set $N_{k,t+1} = N_{k,t} + 1$ if $k \in \{h_t\} \cup \mathcal{L}_t$, and $N_{k,t+1} = N_{k,t}$ otherwise.
  Increment $t \leftarrow t + 1$.
end while
Output any k such that $L_{k,t}^{\pi+\epsilon} \geq \max_{j \neq k} U_{j,t}^{\pi-\epsilon}$.

Figure 4. The QLUCB algorithm samples an arm with highest LCB (time-uniform lower confidence bound) for the $(\pi + \epsilon)$-quantile (called h_t) and the arm(s) with highest UCB (time-uniform upper confidence bound) for the π-quantile excluding the former (called $\mathcal{L}_t$), as long as the aforementioned LCB and UCB overlap.

QLUCB is described in Figure 4. Its sample complexity is determined by the following quantities, which capture how difficult the problem is based on the sub-optimality of the π-quantiles of each arm:

$$\Delta_k := \begin{cases} \sup\left\{\Delta \in [0, \pi \wedge (1-\pi)] : Q_k^-(\pi + \Delta) \leq \max\limits_{j \in [K]} Q_j^-(\pi - \Delta)\right\}, & |\mathcal{A}_\epsilon| > 1 \text{ or } k \notin \mathcal{A}_\epsilon, \\ \sup\left\{\Delta \in [0, \pi] : Q_k^-(\pi - \Delta) > \max\limits_{j \neq k} Q_j^-(\pi + \Delta_j)\right\}, & \mathcal{A}_\epsilon = \{k\}. \end{cases} \tag{27}$$

To understand (27), it is helpful to consider three cases in turn:

- Consider first a suboptimal arm $k \notin \mathcal{A}_\epsilon$. Then Δ_k is given by the first case and captures (informally) how much worse arm k is than some better arm. When arm k is sufficiently sampled relative to Δ_k, then with high probabiltiy, the upper confidence bound on $Q_k^-(\pi - \epsilon)$ will be given by a sample quantile which lies below $Q_k^-(\pi + \Delta_k)$, and by the gap definition, this will be smaller than the lower confidence bound on $Q_j^-(\pi + \epsilon)$ for some other sufficiently-sampled arm j. Thus we will be confident that $Q_j^-(\pi + \epsilon) \geq Q_k^-(\pi - \epsilon)$, a necessary step to conclude that $j \in \mathcal{A}_\epsilon$.
- Suppose there is a unique optimal arm, $\mathcal{A}_\epsilon = \{k^\star\}$. Then $\Delta_{k^\star}$ is given by the second case and captures (again informally) how much *better* arm $k^\star$ is than some "best" suboptimal arm. When arm $k^\star$ is sufficiently sampled relative to $\Delta_{k^\star}$, then with high probability, the lower confidence bound on $Q_{k^\star}^-(\pi + \epsilon)$ will be given by a sample quantile which lies above $Q_{k^\star}^-(\pi - \Delta_{k^\star})$, and by the gap definition, this will be larger than upper confidence bound on $Q_j^-(\pi - \epsilon)$ for any other (suboptimal) sufficiently-sampled arm j. So when all arms are sufficiently sampled, we will be able to conclude that $Q_{k^\star}^-(\pi + \epsilon) \geq Q_j^-(\pi - \epsilon)$ for all suboptimal arms $j \neq k^\star$.
- Suppose there are multiple optimal arms, $|\mathcal{A}_\epsilon| > 1$. Then Δ_k is given by the first case and must be no larger than ϵ. Because the gap only appears as $\epsilon \vee \Delta_k$ in our sample complexity bound, the gap is irrelevant in this case. Informally, we must sample both arms sufficiently that we can determine they are "within ϵ of each other", regardless of the actual distance between their quantile functions.

Below, Theorem 3 bounds the sample complexity of QLUCB and shows that it successfully selects an ϵ-optimal arm, both with high probability.

Theorem 3. *For any $\pi \in (0,1)$, $\epsilon \in [0, \pi \wedge (1-\pi))$, and $\delta \in (0,1)$, QLUCB stops with probability one, and chooses an ϵ-optimal arm with probability at least $1 - \delta$. Furthermore, with probability at least*

$1 - 3\delta$, *the total number of samples* T *taken by QLUCB satisfies*

$$T = O\left(\sum_{k=1}^{K} (\epsilon \vee \Delta_k)^{-2} \log\left(\frac{K\,|\log(\epsilon \vee \Delta_k)|}{\delta}\right)\right). \tag{28}$$

A recent preprint by Kalogerias et al. (2020, Theorem 8) gave a lower bound for the expected sample complexiy when $\epsilon = 0$ of the form $O(\Delta^{-2} \log \delta^{-1})$, where Δ is the minimum gap among suboptimal arms. Our bound matches the dependence on Δ up to a doubly-logarithmic factor, and includes an extra factor of $\log K$. We are not aware of a better upper or lower bound, thus removing the (small) $\log K$ gap remains open. David and Shimkin (2016, Theorem 1) give a lower bound when $\epsilon > 0$ of the form $O(\sum_k (\epsilon \vee \widetilde{\Delta}_k)^{-2} \log \delta^{-1})$ using a slightly different gap definition $\widetilde{\Delta}_k$. Our bound holds at $\epsilon = 0$ in addition to $\epsilon > 0$. Our QLUCB algorithm performs considerably better than existing algorithms in our experiments, including the correct scaling with π, and we hope that will motivate others to work towards fully matching upper and lower bounds.

Theorem 3 is proved in Section 7.3. In brief, the algorithm can only stop with a sub-optimal arm if one of the confidence sequences $L_{k,t}^{\pi+\epsilon}$ or $U_{k,t}^{\pi-\epsilon}$ fails to correctly cover its target quantile, and Theorem 1 bounds the probability of such an error. Furthermore, Theorem 2 ensures that the confidence bounds converge towards their target quantiles at an $O(\sqrt{t^{-1} \log\log t})$ rate, with high probability, so that the algorithm must stop after all arms have been sufficiently sampled, and the allocation strategy given in the algorithm ensures we achieve sufficient sampling with the desired sample complexity. While our proof is inspired by Kalyanakrishnan et al. (2012) and Jamieson and Nowak (2014) but significantly extends them. The fact that quantile confidence bounds are determined by the random sample quantile function, rather than simply as deterministic offsets from the sample mean, introduces new difficulties which require novel techniques to overcome.

As an alternative to (24), one may use a one-sided variant of $\widetilde{f}_t$ from (45) (Howard et al., 2020, Proposition 7). As seen below, this alternative performs well in practice, though the rate of the sample complexity bound suffers slightly, replacing the $\log|\log(\epsilon \vee \Delta_k)|$ term with $|\log(\epsilon \vee \Delta_k)|$.

Figure 5 shows mean sample size from simulations of the quantile ϵ-best-arm identification problem, for variants of QLUCB as well as the QPAC algorithm of Szörényi et al. (2015) and the Doubled Max-Q algorithm of David and Shimkin (2016). In all cases, we have $K = 10$ arms and set $\epsilon = 0.025$, while π ranges between 0.05 and 0.95. In the left panel, nine arms have a uniform distribution on $[0, 1]$, while one arm is uniform on $[2\epsilon, 1 + 2\epsilon]$. In the middle panel, nine arms have Cauchy distributions with location zero and unit scale, while one arm has location $2(Q(\pi + \epsilon) - Q(\pi))$, where $Q(\cdot)$ is the Cauchy quantile function. This choice ensures that the one exceptional arm is the only ϵ-optimal arm. In the right panel, nine arms have $\mathcal{N}(0, 1)$ distributions, while one arm has a $\mathcal{N}(0, 2^2)$ distribution. In this case, the exceptional arm is the only ϵ-optimal arm for π larger than approximately 0.53, while it is the only non-ϵ-optimal arm for π smaller than approximately 0.45. Between these values, all ten arms are ϵ-optimal.

The results show that QLUCB provides a substantial improvement on QPAC and Doubled Max-Q, reducing mean sample size by a factor of at least five among the cases considered, and often much more, when using the one-sided beta-binomial confidence sequence. As Figure S3 in supplement Section 7 shows, most of the improvement appears to be due to the tighter confidence sequence given by Theorem 1, although the QLUCB sampling procedure also gives a noticeable improvement. The stitched confidence sequence in QLUCB performs similarly to the beta-binomial one, staying within a factor of three across all scenarios and usually within a factor of 1.5.

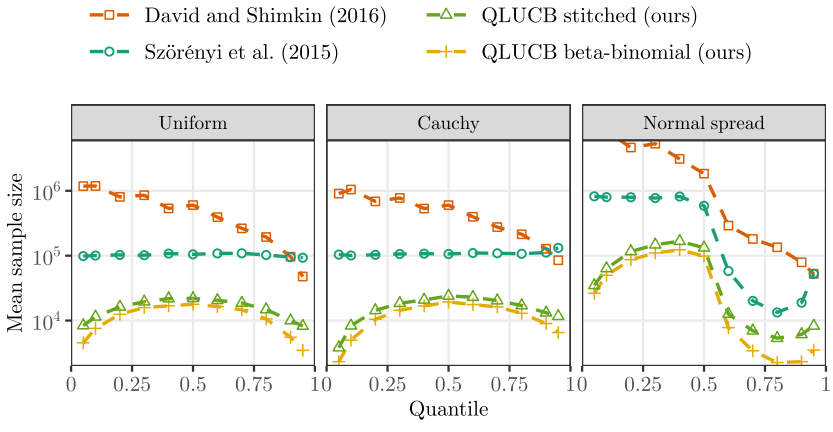


Figure 5. Average sample size for various quantile best-arm identification algorithms based on 64 simulation runs, with $\epsilon = 0.025$ and $\pi = 0.05, 0.1, 0.2, \ldots, 0.8, 0.9, 0.95$. Left panel shows results for arms with uniform distributions on intervals of length one; middle panel shows arms with Cauchy distributions having unit scale; and right panel shows arms with standard normal distributions except for one, which has a standard deviation of two instead of one. In this last case, the exceptional arm is best for quantiles above 0.53, while for quantiles below 0.45, the other arms are all ϵ-optimal. Plot includes Algorithm 2 of David and Shimkin (2016), Algorithm 1 of Szörényi et al. (2015), and our QLUCB algorithm based on two choices of confidence sequence: the stitched confidence sequence (24) based on Theorem 1(a) and a one-sided variant of the beta-binomial confidence sequence, Theorem 1(b).

7. Proofs

We make use of some results from Howard et al. (2020, 2021). We begin with the definitions of sub-Bernoulli, sub-gamma, and sub-Gaussian processes and uniform boundaries:

Definition 2 (Sub-ψ condition). *Let $(S_t)_{t=0}^{\infty}, (V_t)_{t=0}^{\infty}$ be real-valued processes adapted to an underlying filtration $(\mathcal{F}_t)_{t=0}^{\infty}$ with $S_0 = V_0 = 0$ and $V_t \geq 0$ for all t. For a function $\psi : [0, \lambda_{\max}) \to \mathbb{R}$, we say (S_t) is* sub-ψ with variance process (V_t) *if, for each $\lambda \in [0, \lambda_{\max})$, there exists a supermartingale $(L_t(\lambda))_{t=0}^{\infty}$ w.r.t. $(\mathcal{F}_t)$ such that $\mathbb{E}L_0(\lambda) \leq 1$ and*

$$\exp\{\lambda S_t - \psi(\lambda) V_t\} \leq L_t(\lambda) \text{ a.s. for all } t. \tag{29}$$

Definition 3. *Given $\psi : [0, \lambda_{\max}) \to \mathbb{R}$ and $l_0 \geq 1$, a function $u : \mathbb{R} \to \mathbb{R}$ is called a* sub-ψ uniform boundary *with crossing probability α if*

$$\mathbb{P}(\exists t \geq 1 : S_t \geq u(V_t)) \leq \alpha \tag{30}$$

whenever (S_t) is sub-ψ with variance process V_t.

Definition 4. *We use the following ψ functions in what follows.*

1. *A* sub-Bernoulli *process or boundary is sub-ψ with*

$$\psi_{B,g,h}(\lambda) := \frac{1}{gh} \log\left(\frac{ge^{h\lambda} + he^{-g\lambda}}{g+h}\right) \tag{31}$$

on $0 \leq \lambda < \infty$ for some parameters $g, h > 0$.

2. *A* sub-Gaussian *process or boundary is sub-ψ with*

$$\psi_N(\lambda) \coloneqq \lambda^2/2 \tag{32}$$

on $0 \le \lambda < \infty$.

3. *A* sub-gamma *process or boundary is sub-ψ with*

$$\psi_{G,c}(\lambda) \coloneqq \lambda^2/(2(1-c\lambda)) \tag{33}$$

on $0 \le \lambda < 1/(c \vee 0)$ *(taking* $1/0 = \infty$*) for some scale parameter* $c \in \mathbb{R}$.

The following facts will aid intuition for the true and empirical quantile functions:

- $Q(p)$ and $\widehat{Q}_t(p)$ are right-continuous, while $Q^-(p)$ and $\widehat{Q}_t^-(p)$ are left-continuous.
- $\widehat{Q}_t(p)$ is the $\lfloor tp \rfloor + 1$ order statistic of $X_1, \dots, X_t$, and $\widehat{Q}_t^-(p)$ is the $\lceil tp \rceil$ order statistic.
- $Q^-(p) \le Q(p)$, and $Q^-(p) = Q(p)$ unless the p-quantile is ambiguous, that is, $F(x) = F(x') = p$ for some $x \ne x'$.
- $\widehat{Q}_t^-(p) \le \widehat{Q}_t(p)$, and $\widehat{Q}_t^-(p) = \widehat{Q}_t(p)$ for all $p \notin \{1/t, 2/t, \dots, (t-1)/t\}$.
- Q^- is sometimes denoted F^{-1} (e.g., Shorack and Wellner, 1986, p. 3, equation (13)). Our notation seems to improve clarity in the case of ambiguous quantiles.

The functions $\widehat{Q}_t^-$ and $\widehat{Q}_t$ act as "inverses" for $\widehat{F}_t$ and $\widehat{F}_t^-$ in the following sense: for any $x \in \mathcal{X}$ and any $p \in \mathbb{R}$, we have

$$\widehat{F}_t(x) \ge p \iff x \ge \widehat{Q}_t^-(p), \qquad \widehat{F}_t(x) \le p \implies x \le \widehat{Q}_t(p), \tag{34}$$

$$\widehat{F}_t(x) > p \implies x \ge \widehat{Q}_t(p), \quad \text{and} \quad \widehat{F}_t^-(x) < p \implies x \le \widehat{Q}_t^-(p). \tag{35}$$

Our strategy in the proof of Theorem 1 will be to construct a martingale $(S_t(p))_{t=1}^\infty$ which almost surely satisfies

$$\widehat{F}_t^-(Q(p)) \le p + S_t(p)/t \le \widehat{F}_t(Q^-(p)) \tag{36}$$

for all $t \in \mathbb{N}$. Applying a time-uniform concentration inequality to bound the deviations of $(S_t(p))$, we obtain a time-uniform lower bound $\widehat{F}_t(Q^-(p)) > p - l_t(p)$ and upper bound $\widehat{F}_t^-(Q(p)) < p + u_t(p)$, both of which hold with high probability. We then invoke the implications in (35) to obtain a confidence sequence for $Q^-(p), Q(p)$ of the form (6).

The martingale $(S_t(p))$ is defined as follows. Let

$$\pi(p) \coloneqq \begin{cases} 0, & F(Q(p)) = F^-(Q(p)), \\ \dfrac{p - F^-(Q(p))}{F(Q(p)) - F^-(Q(p))}, & F(Q(p)) > F^-(Q(p)), \end{cases} \tag{37}$$

noting that $\pi(p) \in [0,1]$ since $F^-(Q(p)) \le p \le F(Q(p))$. Now define $S_0(p) = 0$ and

$$S_t(p) \coloneqq \sum_{i=1}^{t} \left[1_{X_i < Q(p)} + \pi(p) 1_{X_i = Q(p)} - p \right] \tag{38}$$

for $t \in \mathbb{N}$. When $F(Q(p)) = F^-(Q(p))$, so that $\mathbb{P}(X_1 = Q(p)) = 0$, we have $\widehat{F}_t^-(Q(p)) = p + S_t(p)/t = \widehat{F}_t(Q(p))$ for all $t \in \mathbb{N}$ a.s. When $F(Q(p)) > F^-(Q(p))$, we are still assured $\widehat{F}_t^-(Q(p)) \le p + S_t(p)/t \le \widehat{F}_t(Q(p))$ for all $t \in \mathbb{N}$, as desired. In either case, the increments $\Delta S_t(p) \coloneqq S_t(p) - S_{t-1}(p)$ are i.i.d., mean-zero, and bounded in $[-p, 1-p]$ for all $t \in \mathbb{N}$. This key fact allows us to bound the deviations of $S_t(p)$ using time-uniform concentration inequalities for Bernoulli random walks.

7.1. Proof of Theorem 1

As defined in (38), the i.i.d. increments of the process $(S_t(p))_{t=1}^{\infty}$,

$$S_t(p) - S_{t-1}(p) = 1_{X_i < Q(p)} + \pi(p) 1_{X_i = Q(p)} - p, \tag{39}$$

are mean-zero and bounded in $[-p, 1-p]$. Fact 1(b) and Lemma 2 of Howard et al. (2020) verify that the process $(S_t(p))$ is a sub-Bernoulli process (31) with range parameters $g = p, h = 1-p$. Then, defining the intrinsic variance process $V_t := p(1-p)t$ and

$$\psi(\lambda) := \frac{1}{p(1-p)} \log\left(pe^{(1-p)\lambda} + (1-p)e^{-p\lambda}\right), \tag{40}$$

it is straightforward to verify that the process $(\exp\{\lambda S_t(p) - \psi(\lambda)V_t\})_{t=1}^{\infty}$ is a supermartingale for all $\lambda \geq 0$. We now construct time-uniform bounds for the process $(S_t(p))$ based on the above property:

- Using the fact that a sub-Bernoulli process with range parameters $g = p$ and $h = 1-p$ is also sub-gamma with scale $c = (1-2p)/3$, the sequence $f_t(p)$ is based on the "polynomial stitched boundary" (Howard et al., 2021, Proposition 1, equation 6, and Theorem 1). That result allows us to fix any $\eta > 1$, $s > 1$, which control the shape of the confidence radius over time, and $m \geq 1$, the time at which the confidence sequence starts to be tight, and obtain $f_t(p) = \mathcal{S}_p(t \vee m)/t$ with

$$\mathcal{S}_p(t) := \sqrt{k_1^2 p(1-p)t\ell(t) + k_2^2 c_p^2 \ell^2(t)} + c_p k_2 \ell(t),$$
$$\text{where } \begin{cases} \ell(t) := s \log\log\left(\frac{\eta t}{m}\right) + \log\left(\frac{2\zeta(s)}{\alpha \log^s \eta}\right) \\ k_1 := (\eta^{1/4} + \eta^{-1/4})/\sqrt{2} \\ k_2 := (\sqrt{\eta} + 1)/2 \\ c_p := (1-2p)/3. \end{cases} \tag{41}$$

  The special case given in eq. (9) follows from the choices $\eta = 2.04$, $s = 1.4$, and $m = 1$. Then

$$\mathbb{P}(\exists t \in \mathbb{N} : S_t(p) \geq t f_t(p)) \leq \alpha/2. \tag{42}$$

  If we replace $(S_t(p))$ with $(-S_t(p))$, which is sub-Bernoulli with range parameters $g = 1-p$ and $h = p$ and therefore sub-gamma with scale $c = 2p-1$, we obtain

$$\mathbb{P}(\exists t \in \mathbb{N} : S_t(p) \leq -t f_t(1-p)) \leq \alpha/2. \tag{43}$$

  A union bound yields the two-sided result

$$\mathbb{P}\left(\exists t \in \mathbb{N} : t^{-1} S_t(p) \notin (-f_t(1-p), f_t(p))\right) \leq \alpha. \tag{44}$$

- The sequence $\widetilde{f}_t(p)$ is based on a two-sided beta-binomial mixture boundary drawn from Proposition 7 of Howard et al. (2021). Below, we denote the beta function by $B(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1}\,\mathrm{d}u$. Fix any $r > 0$, a tuning parameter, and define

$$\widetilde{f}_t(p) := \frac{1}{t} \sup\left\{s \in \left[0, \frac{r + p(1-p)t}{p}\right) : M_{p,r}(s, p(1-p)t) < \frac{1}{\alpha}\right\}, \tag{45}$$

$$\text{where} \quad M_{p,r}(s,v) \coloneqq \frac{1}{p^{v/(1-p)+s}(1-p)^{v/p-s}} \cdot \frac{B\left(\frac{r+v}{p} - s, \frac{r+v}{1-p} + s\right)}{B\left(\frac{r}{p}, \frac{r}{1-p}\right)}. \tag{46}$$

Then we have

$$\mathbb{P}\left(\exists t \in \mathbb{N} : t^{-1} S_t(p) \notin \left(-\widetilde{f}_t(1-p), \widetilde{f}_t(p)\right)\right) \leq 1 - \alpha. \tag{47}$$

By construction, $\widehat{F}_t^-(Q(p)) \leq p + S_t(p)/t \leq \widehat{F}_t(Q^-(p))$ for all t, so with (44) we have

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \widehat{F}_t(Q^-(p)) \leq p - f_t(1-p) \text{ or } \widehat{F}_t^-(Q(p)) \geq p + f_t(p)\right) \leq \alpha. \tag{48}$$

We now use the implications in (35) to conclude

$$\mathbb{P}\left(\exists t \in \mathbb{N} : Q^-(p) < \widehat{Q}_t\left(p - f_t(1-p)\right) \text{ or } Q(p) > \widehat{Q}_t^-\left(p + f_t(p)\right)\right) \leq \alpha, \tag{49}$$

as desired. The same conclusion follows for $\widetilde{f}$ by using (47) in place of (44). □

We remark that (49) implies that the running intersection of confidence intervals also yields a valid confidence sequence: for any $q \in [Q^-(p), Q(p)]$, we have

$$\mathbb{P}\left(\forall t \in \mathbb{N} : q \in \left[\max_{s \leq t} \widehat{Q}_s\left(p - f_s(1-p)\right), \min_{s \leq t} \widehat{Q}_s^-\left(p + f_s(p)\right)\right]\right) \geq 1 - \alpha. \tag{50}$$

This intersection yields smaller confidence intervals. However, on the miscoverage event of probability α, or if the assumption of i.i.d. observations is violated, then the intersection method may lead to an empty confidence interval. This can be viewed as a benefit, as an empty confidence interval is evidence of problematic assumptions. In such cases, however, it may also lead to misleadingly small, but not empty, confidence intervals, which may be harder to detect.

7.2. Proof of Theorem 2

We prove the following more general result:

Theorem 4. *For any $m \geq 1$, $A > 1/\sqrt{2}$, and $C > 0$, we have*

$$\mathbb{P}\left(\exists t \geq m : \left\|\widehat{F}_t - F\right\|_\infty > A\sqrt{\frac{\log\log(et/m) + C}{t}}\right)$$
$$\leq \alpha_{A,C} \coloneqq \inf_{\substack{\eta \in (1, 2A^2), \\ \gamma(A,C,\eta) > 1}} 4e^{-\gamma^2(A,C,\eta)C}\left(1 + \frac{1}{(\gamma^2(A,C,\eta) - 1)\log\eta}\right), \tag{51}$$

where $\gamma(A,C,\eta) \coloneqq \sqrt{2/\eta}\left(A - \sqrt{2(\eta-1)/C}\right)$. Furthermore,

$$\mathbb{P}\left(\left\|\widehat{F}_t - F\right\|_\infty > A\sqrt{t^{-1}(\log\log(et/m) + C)} \text{ infinitely often}\right) = 0. \tag{52}$$

To better understand the quantity $\alpha_{A,C}$, note that any value of $\eta \in (1, 2A^2)$ satisfying $\gamma(A,C,\eta)$ gives an upper bound for $\alpha_{A,C}$. For fixed A, any value $\eta \in (1, 2A^2)$ is feasible for sufficiently large C, while for fixed C, any value $\eta > 1$ is feasible for sufficiently large A. In either case, $\gamma^2(A,C,\eta) \sim 2A^2/\eta$ as $A \to \infty$ or $C \to \infty$, which yields $\log \alpha_{A,C} = O(-A^2 C)$, as may be expected from a typical exponential concentration bound.

To obtain the special case stated in in Theorem 2, take $A = 0.85$ and any $C \geq 7$, and observe that the value $\eta = 1.01$ ensures that $\gamma^2(0.85, C, 1.01) \geq 1.25$ and is thus feasible for the right-hand side of (51).

Our proof is based on inequality 13.2.1 of Shorack and Wellner (1986, p. 511) (cf. James, 1975). We repeat the following special case; here $(\cdot)_\pm$ denotes that we may take either the positive part of $(\cdot)$ throughout, or the negative part throughout.

Lemma 1 (Shorack and Wellner, 1986, Inequality 13.2.1). *Fix $\lambda > 0$, $\beta \in (0,1)$, and $\eta > 1$ satisfying $(1-\beta)^2\lambda^2 \geq 2(\eta - 1)$. Then for all integers $n' \leq n''$ having $n''/n' \leq \eta$, we have*

$$\mathbb{P}\left(\max_{n' \leq t \leq n''} \left\| \sqrt{t}(\widehat{F}_t - F)_\pm \right\|_\infty > \lambda \right) \leq 2\mathbb{P}\left(\left\| \sqrt{n''}(\widehat{F}_{n''} - F)_\pm \right\|_\infty > \frac{\beta\lambda}{\sqrt{\eta}} \right). \tag{53}$$

Now fix any $\eta \in (1, 2A^2)$ satisfying $\gamma(A,C,\eta) > 1$, and for $k = 0, 1, \ldots$, define the event

$$\mathcal{A}_k^\pm := \left\{ \exists t \in [m\eta^k, m\eta^{k+1}) : \left\| (\widehat{F}_t - F)_\pm \right\|_\infty > A\sqrt{\frac{\log\log(e\eta^k) + C}{t}} \right\}. \tag{54}$$

On the one hand, we have

$$\left\{ \exists t \geq m : \left\| \widehat{F}_t - F \right\|_\infty > \frac{g_t}{t} \right\} = \bigcup_{k \in \mathbb{Z}_{\geq 0}} \left\{ \exists t \in [m\eta^k, m\eta^{k+1}) : \left\| \widehat{F}_t - F \right\|_\infty > \frac{g_t}{t} \right\} \tag{55}$$

$$\subseteq \bigcup_{k \in \mathbb{Z}_{\geq 0}} \left(\mathcal{A}_k^+ \cup \mathcal{A}_k^- \right). \tag{56}$$

On the other hand, we will show that, for each $k \geq 0$, the conditions of Lemma 1 are satisfied with $\lambda := A\sqrt{\log\log(e\eta^k) + C}$ and $\beta := 1 - \sqrt{2(\eta-1)/(A^2 C)} = \gamma(A,C,\eta)\sqrt{\eta/(2A^2)}$. It is clear that $\beta \in (0,1)$ since A, C, η, and $\gamma(A,C,\eta)$ must all be positive. Also,

$$2(\eta - 1) = (1-\beta)^2 A^2 C \leq (1-\beta)^2 A^2 (\log\log(e\eta^k) + C) = (1-\beta)^2\lambda^2, \quad \forall k \geq 0. \tag{57}$$

Hence, for each k, Lemma 1 implies

$$\mathbb{P}(\mathcal{A}_k^\pm) \leq 2\mathbb{P}\left(\left\| \sqrt{\lfloor \eta^{k+1} \rfloor}(\widehat{F}_{\lfloor \eta^{k+1} \rfloor} - F)_\pm \right\|_\infty > \frac{\beta A\sqrt{\log\log(e\eta^k) + C}}{\sqrt{\eta}} \right). \tag{58}$$

Applying the one-sided DKW inequality (Massart, 1990, Theorem 1) then yields

$$\mathbb{P}(\mathcal{A}_k^\pm) \leq 2\exp\left\{ -\frac{2c^2 A^2 (\log\log(et_k) + C)}{\eta} \right\} = \frac{2e^{-\gamma^2(A,C,\eta)C}}{(1 + k\log\eta)^{\gamma^2(A,C,\eta)}}. \tag{59}$$

Since $\gamma(A,C,\eta) > 1$, a union bound yields

$$\mathbb{P}\left(\bigcup_{k \in \mathbb{N}} \left(\mathcal{A}_k^+ \cup \mathcal{A}_k^- \right) \right) \leq 4e^{-\gamma^2(A,C,\eta)C} \sum_{k=0}^{\infty} \frac{1}{(1 + k\log\eta)^{\gamma^2(A,C,\eta)}} \tag{60}$$

$$\leq 4e^{-\gamma^2(A,C,\eta)C}\left(1+\frac{1}{(\gamma^2(A,C,\eta)-1)\log\eta}\right), \tag{61}$$

after bounding the sum by an integral. Combining (56) with (61), we conclude

$$\mathbb{P}\left(\exists t\geq m:\left\|\widehat{F}_t-F\right\|_\infty>\frac{g_t}{t}\right)\leq 4e^{-\gamma^2(A,C,\eta)C}\left(1+\frac{1}{(\gamma^2(A,C,\eta)-1)\log\eta}\right). \tag{62}$$

Note Theorem 1 of Massart (1990) requires the tail probability bound in (59) to be less than $1/2$. But if this is not true, then our final tail probability will be at least one, so the result holds vacuously. Hence the first part of the theorem is proved.

To obtain the final claim, (52), note that the calculations in (59) and (61), together with the first Borel-Cantelli lemma, imply $\mathbb{P}(A_k^+ \text{ or } A_k^- \text{ infinitely often}) = 0$. □

7.3. Proof of Theorem 3

Recall that the set of ϵ-optimal arms is denoted by

$$\mathcal{A}_\epsilon := \left\{k\in[K]: Q_k^-(\pi+\epsilon)\geq \max_{j\in[K]} Q_j^-(\pi-\epsilon)\right\}.$$

First, we prove that if QLUCB stops, it selects an ϵ-optimal arm with probability at least $1-\delta$. Choose any $k^\star\in\arg\max_{k\in[K]} Q_k^-(\pi-\epsilon)$, an arm with optimal $(\pi-\epsilon)$-quantile, and write $q^\star := Q_{k^\star}^-(\pi-\epsilon)$ for the corresponding optimum quantile value. By our choice of u_n and l_n to give one-sided coverage at level δ/K, the proof of Theorem 1 and a union bound show that

$$\mathbb{P}\left(\exists t\in\mathbb{N} \text{ and } k\neq k^\star: U_{k^\star,t}^{\pi-\epsilon}<q^\star \text{ or } L_{k,t}^{\pi+\epsilon}>Q_k^-(\pi+\epsilon)\right)\leq\delta. \tag{63}$$

Suppose QLUCB stops at time T with some arm $k\in\mathcal{A}_\epsilon^c$, so that $Q_k^-(\pi+\epsilon)<q^\star$. Then it must be true that $L_{k,T}^{\pi+\epsilon}\geq U_{k^\star,T}^{\pi-\epsilon}$, which implies that $L_{k,T}^{\pi+\epsilon}>Q_k^-(\pi+\epsilon)$ or $U_{k^\star,T}^{\pi-\epsilon}<q^\star$ must hold. But (63) shows that this can only occur on an event of probability at most δ. So with probability at least $1-\delta$, QLUCB can only stop with an ϵ-optimal arm.

Next, we prove that QLUCB stops with probability one and obeys the sample complexity bound (28) with probability at least $1-3\delta$. We first address the case when $|\mathcal{A}_\epsilon|>1$ so that Δ_k is given by (27) for all k; we consider the case $|\mathcal{A}_\epsilon|=1$ at the end. Let

$$g_n := 0.85\sqrt{n^{-1}\left(\log\log(en)+0.8\log\left(\frac{1612K}{\delta}\right)\right)}, \tag{64}$$

for $n\in N$. We choose this quantity to eventually control the deviations of $\widehat{Q}_{k,t}(p)$ and $\widehat{Q}^-k,t(p)$ from $Q_k^-(p)$ and $Q_k(p)$ uniformly over k, t and p, via Corollary 2. For each $k\in[K]$, define

$$\tau_k := \min\left\{n\in\mathbb{N}: g_n+[u_n(\pi)\vee l_n(\pi+\epsilon)]<\Delta_k\vee\epsilon\right\}. \tag{65}$$

We will show that, once each arm has been sampled in $\mathcal{L}_t$ at least $2\tau_k$ times, the confidence bounds are sufficiently well-behaved to ensure that QLUCB must stop, on a "good" event with probability at least $1-3\delta$. This will imply that QLUCB stops after no more than $2\sum_{k=1}^K\tau_k$ rounds on the "good" event, and this sum has the desired rate.

Define the "bad" event at time t, $\mathcal{B}_t = \mathcal{B}_t^1 \cup \mathcal{B}_t^2$, where

$$\mathcal{B}_t^1 := \left\{\exists k \in [K] : U_{k,t}^{\pi-\epsilon} < Q_k(\pi-\epsilon) \text{ or } L_{k,t}^{\pi+\epsilon} > Q_k^-(\pi+\epsilon)\right\}, \text{ and} \tag{66}$$

$$\mathcal{B}_t^2 := \left\{\exists k \in [K], p \in (0,1) : \widehat{Q}_{k,t}(p) < Q_k(p - g_{N_{k,t}}) \text{ or } \widehat{Q}_{k,t}^-(p) > Q_k^-(p + g_{N_{k,t}})\right\}. \tag{67}$$

We exploit our previous results to bound the probability that $\mathcal{B}_t$ ever occurs:

Lemma 2. $\mathbb{P}\left(\cup_{t=1}^{\infty}\mathcal{B}_t\right) \le 3\delta$.

Proof. First, by the definitions of $U_{k,t}^{\pi-\epsilon}, L_{k,t}^{\pi+\epsilon}$ and our choices of u_n, l_n, the proof of Theorem 1(a) yields

$$\mathbb{P}\left(\bigcup_{t=1}^{\infty}\mathcal{B}_t^1\right) \le 2\delta. \tag{68}$$

For $\mathcal{B}_t^2$, we invoke Corollary 2. Our choice of $C = 0.8\log(1612K^2/(\delta(K-1)))$ ensures that $\alpha_{0.85,C} \le (K-1)\delta/K^2$, noting that $K \ge 2$ implies $C > 7$ as required in (2). Hence, by a union bound,

$$\mathbb{P}\left(\bigcup_{t=1}^{\infty}\mathcal{B}_t^2\right) \le \delta. \tag{69}$$

Combining (68) with (69) via a union bound, we have $\mathbb{P}(\cup_{t=1}^{\infty}\mathcal{B}_t) \le 3\delta$ as desired. □

The following lemma verifies that an arm's confidence bounds are well-behaved, in a specific sense, once the arm has been sampled τ_j times and $\mathcal{B}_t^2$ does not occur. We use the notation $a_+ := \max(0,a)$.

Lemma 3. *For any $t \in \mathbb{N}$ and $j,k \in [K]$, on $(B_t^2)^c$, if $N_{k,t} \ge \tau_j$, then*

$$U_{k,t}^{\pi-\epsilon} \le Q_k^-(\pi + (\Delta_j - \epsilon)_+), \quad \textit{and} \tag{70}$$

$$L_{k,t}^{\pi+\epsilon} \ge Q_k(\pi - (\Delta_j - \epsilon)_+). \tag{71}$$

Proof. From the definition of $U_{k,t}^{\pi-\epsilon}$,

$$U_{k,t}^{\pi-\epsilon} = \widehat{Q}_{k,t}^-(\pi - \epsilon + u_{N_{k,t}}(\pi-\epsilon)) \le Q_k^-(\pi - \epsilon + u_{N_{k,t}}(\pi-\epsilon) + g_{N_{k,t}}), \tag{72}$$

since we are on $(\mathcal{B}_t^2)^c$. Then since $N_{k,t} \ge \tau_j$,

$$\begin{aligned} Q_k^-(\pi - \epsilon + u_{N_{k,t}}(\pi-\epsilon) + g_{N_{k,t}}) &\le Q_k^-(\pi - \epsilon + (\Delta_j \vee \epsilon)) \\ &= Q_k^-(\pi + (\Delta_j - \epsilon)_+). \end{aligned} \tag{73}$$

An analogous argument shows the second conclusion:

$$L_{k,t}^{\pi+\epsilon} = \widehat{Q}_{k,t}(\pi + \epsilon - l_{N_{k,t}}(\pi+\epsilon)) \ge Q_k(\pi + \epsilon - l_{N_{k,t}}(\pi+\epsilon) - g_{N_{k,t}}) \tag{74}$$

$$\ge Q_k(\pi + \epsilon - (\Delta_j \vee \epsilon)) = Q_k(\pi - (\Delta_j - \epsilon)_+). \tag{75}$$

□

The next three lemmas will show that, once an arm in $\mathcal{L}_t$ has been sufficiently sampled, QLUCB must stop. The easier case is when an arm's gap is small, $\Delta_k < \epsilon$.

Lemma 4. *For any $t \in \mathbb{N}$ and $k \in [K]$ with $\Delta_k < \epsilon$, on $(B_t^2)^c$, if $N_{k,t} \geq \tau_k$, then $L_{h_t,t}^{\pi+\epsilon} \geq U_{k,t}^{\pi-\epsilon}$.*

Proof. Our choice of h_t ensures $L_{h_t,t}^{\pi+\epsilon} \geq L_{k,t}^{\pi+\epsilon}$, while (70) and (71) show that

$$L_{k,t}^{\pi+\epsilon} \geq Q_k(\pi) \geq Q_k^-(\pi) \geq U_{k,t}^{\pi-\epsilon}, \tag{76}$$

as claimed. □

To handle arms with $\Delta_k \geq \epsilon$, we associate with each arm k an arm $g(k)$ which satisfies $Q_k^-(\pi + \Delta_k) \leq Q_{g(k)}(\pi - \Delta_k)$. Some such arm must exist by the definition of Δ_k and the fact that Q^- is left-continuous while Q is right-continuous. We first show that, when an arm $k \in \mathcal{L}_t$ with $\Delta_k \geq \epsilon$ has been sufficiently sampled, we must also sample $g(k)$:

Lemma 5. *For any $t \in \mathbb{N}$ and $k \in [K]$ with $\Delta_k \geq \epsilon$, on B_t^c, if $N_{k,t} \geq \tau_k$, then $U_{g(k),t}^{\pi-\epsilon} \geq U_{k,t}^{\pi-\epsilon}$.*

Proof. Bound (70) and our choice of $g(k)$ ensure

$$U_{k,t}^{\pi-\epsilon} \leq Q_k^-(\pi + \Delta_k) \leq Q_{g(k)}(\pi - \Delta_k). \tag{77}$$

But $\Delta_k \geq \epsilon$, so $Q_{g(k)}(\pi - \Delta_k) \leq Q_{g(k)}(\pi - \epsilon)$, and the latter is upper bounded by $U_{g(k),t}^{\pi-\epsilon}$ since we are on $(B_t^1)^c$. □

Finally, we show that once arms $k \in \mathcal{L}_t$ and $g(k)$ have both been sufficiently sampled, we must stop.

Lemma 6. *Consider any $t \in \mathbb{N}$ and $k \in [K]$ with $\Delta_k \geq \epsilon$. On B_t^c, if $N_{k,t} \geq \tau_k$ and $N_{g(k),t} \geq \tau_k$, then $L_{h_t,t}^{\pi+\epsilon} \geq U_{k,t}^{\pi-\epsilon}$.*

Proof. As in (77), we have

$$U_{k,t}^{\pi-\epsilon} \leq Q_{g(k)}(\pi - \Delta_k) \leq Q_{g(k)}(\pi - (\Delta_k - \epsilon)_+). \tag{78}$$

But since $N_{g(k),t} \geq \tau_k$, (71) and our choice of h_t imply

$$Q_{g(k)}(\pi - (\Delta_k - \epsilon)_+) \leq L_{g(k),t}^{\pi+\epsilon} \leq L_{h_t,t}^{\pi+\epsilon}. \tag{79}$$

□

We combine the preceding lemmas in the following key result. Write $M_{k,t} = \sum_{s=1}^t 1_{k \in \mathcal{L}_s}$ and note that $N_{k,t} \geq M_{k,t}$ since we sample every arm in $\mathcal{L}_t$ at time t.

Lemma 7. *For any $t \in \mathbb{N}$, on $\mathcal{B}_t^c$, if $M_{k,t} \geq 2\tau_k$ for any $k \in \mathcal{L}_t$, then QLUCB must stop at time t.*

Proof. If $\Delta_k < \epsilon$ then the conclusion follows immediately from Lemma 4. If $\Delta_k \geq \epsilon$, then Lemma 5 implies $N_{g(k),t} \geq M_{k,t} - \tau_k$, since once $M_{k,t} \geq \tau_k$, we must have $U_{g(k),t} \geq U_{k,t}$ so that either $g(k) = h_t$ or $g(k) \in \mathcal{L}_t$ whenever $k \in \mathcal{L}_t$. Thus when $M_{k,t} \geq 2\tau_k$, we must have $N_{g(k),t} \geq \tau_k$ and the conclusion follows from Lemma 6. □

We can now show that QLUCB stops after no more than $4\sum_{k=1}^{K}\tau_k$ samples with probability at least $1-3\delta$. On $\mathcal{B}_t^c$, Lemma 7 allows us to write

$$T \le \sum_{t=1}^{\infty}(1+|\mathcal{L}_t|)1\{M_{k,t} < 2\tau_k \text{ for all } k \in \mathcal{L}_t\} \tag{80}$$

$$\le 2\sum_{t=1}^{\infty}\sum_{k=1}^{K} 1\{k \in \mathcal{L}_t \text{ and } M_{k,t} < 2\tau_k\} \tag{81}$$

$$\le 4\sum_{k=1}^{K}\tau_k, \tag{82}$$

by the definition of $M_{k,t}$. Hence $\mathbb{P}(T \le 4\sum_{k=1}^{K}\tau_k) \ge 1-\mathbb{P}(\cup_{t=1}^{\infty}\mathcal{B}_t) \ge 1-3\delta$ using Lemma 2. It remains to show that $T<\infty$ a.s. and that $\sum_{k=1}^{K}\tau_k$ has the desired rate.

First, Corollary 1 of Howard et al. (2021) implies that $\mathbb{P}(\mathcal{B}_t^1 \text{ infinitely often}) = 0$, while Theorem 2 implies $\mathbb{P}(\mathcal{B}_t^2 \text{ infinitely often}) = 0$. So, with probability one, there exists t_0 such that $\mathcal{B}_t$ occurs for no $t \ge t_0$, and the above calculations show that $T \le t_0 + 4\sum_{k=1}^{K}\tau_k$. We conclude $T<\infty$ almost surely.

Second, to show that $\sum_{k=1}^{K}\tau_k$ has the rate in (28), we use the following bound on the time for an iterated-logarithm confidence sequence radius to shrink to a desired size, proved in supplement Section 1.2.

Lemma 8. *Suppose* $(a_n(C))_{n\in\mathbb{N}}$ *is a real-valued sequence for each* $C>0$ *satisfying* $a_n = O(\sqrt{n^{-1}(\log\log n + C)})$ *as* $n, C \uparrow \infty$. *Then*

$$\min\{n\in\mathbb{N} : a_n(C) \le x\} = O\left(\frac{\log\log x^{-1} + C}{x}\right) \quad \text{as } x \downarrow 0, C \uparrow \infty. \tag{83}$$

Examining the form of u_n and l_n given in (41) along with the definition of g_n, we see that $a_n(C) = g_n + [u_n(\pi) \vee l_n(\pi+\epsilon)]$ satisfies the condition of Lemma 8 with $C = \log(K/\delta)$, which implies

$$\tau_k = O\left((\epsilon\vee\Delta_k)^{-2}\log\left(\frac{K\,|\log(\epsilon\vee\Delta_k)|}{\delta}\right)\right). \tag{84}$$

Summing over k yields the desired sample complexity (28), completing the proof. □

Acknowledgements

We thank Jon McAuliffe for helpful comments.

Funding

Howard thanks Office of Naval Research (ONR) Grant N00014-15-1-2367.

Supplementary Material

Additional proofs and discussion (DOI: 10.3150/21-BEJ1388SUPP; .pdf). Proofs of some minor results as well as additional results, plots, and discussion.

References

Anderson, C.W. (1984). Large deviations of extremes. In *Statistical Extremes and Applications (Vimeiro, 1983). NATO Adv. Sci. Inst. Ser. C: Math. Phys. Sci.* **131** 325–340. Dordrecht: Reidel. MR0784827

Arnold, B.C., Balakrishnan, N. and Nagaraja, H.N. (2008). *A First Course in Order Statistics*. *Classics in Applied Mathematics* **54**. Philadelphia, PA: SIAM. MR2399836 https://doi.org/10.1137/1.9780898719062

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford: Oxford Univ. Press. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

Boucheron, S. and Thomas, M. (2012). Concentration inequalities for order statistics. *Electron. Commun. Probab.* **17** no. 51, 12. MR2994876 https://doi.org/10.1214/ECP.v17-2210

Bubeck, S., Munos, R. and Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory*. *Lecture Notes in Computer Science* **5809** 23–37. Berlin: Springer. MR2564216 https://doi.org/10.1007/978-3-642-04414-4_7

Corless, R.M., Gonnet, G.H., Hare, D.E.G., Jeffrey, D.J. and Knuth, D.E. (1996). On the Lambert *W* function. *Adv. Comput. Math.* **5** 329–359. MR1414285 https://doi.org/10.1007/BF02124750

Darling, D.A. and Robbins, H. (1967a). Confidence sequences for mean, variance, and median. *Proc. Natl. Acad. Sci. USA* **58** 66–68. MR0215406 https://doi.org/10.1073/pnas.58.1.66

Darling, D.A. and Robbins, H. (1967b). Iterated logarithm inequalities. *Proc. Natl. Acad. Sci. USA* **57** 1188–1192. MR0211441 https://doi.org/10.1073/pnas.57.5.1188

Darling, D.A. and Robbins, H. (1968). Some nonparametric sequential tests with power one. *Proc. Natl. Acad. Sci. USA* **61** 804–809. MR0238437 https://doi.org/10.1073/pnas.61.3.804

David, Y. and Shimkin, N. (2016). Pure exploration for max-quantile bandits. In *Machine Learning and Knowledge Discovery in Databases*. *Lecture Notes in Computer Science* 556–571. Springer.

Dekkers, A.L.M. and de Haan, L. (1989). On the estimation of the extreme-value index and large quantile estimation. *Ann. Statist.* **17** 1795–1832. MR1026314 https://doi.org/10.1214/aos/1176347396

Drees, H. (1998). On smooth statistical tail functionals. *Scand. J. Stat.* **25** 187–210. MR1614276 https://doi.org/10.1111/1467-9469.00097

Drees, H., de Haan, L. and Li, D. (2003). On large deviation for extremes. *Statist. Probab. Lett.* **64** 51–62. MR1995809 https://doi.org/10.1016/S0167-7152(03)00130-5

Dudley, R.M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Funct. Anal.* **1** 290–330. MR0220340 https://doi.org/10.1016/0022-1236(67)90017-1

Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Stat.* **27** 642–669. MR0083864 https://doi.org/10.1214/aoms/1177728174

Even-Dar, E., Mannor, S. and Mansour, Y. (2002). PAC bounds for multi-armed bandit and Markov decision processes. In *Computational Learning Theory (Sydney, 2002)*. *Lecture Notes in Computer Science* **2375** 255–270. Berlin: Springer. MR2040418 https://doi.org/10.1007/3-540-45435-7_18

Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics, [40]*. New York: Cambridge Univ. Press. MR3588285 https://doi.org/10.1017/CBO9781107337862

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

Howard, S.R. and Ramdas, A. (2022). Supplement to "Sequential estimation of quantiles with applications to A/B testing and best-arm identification." https://doi.org/10.3150/21-BEJ1388SUPP

Howard, S.R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2020). Time-uniform Chernoff bounds via nonnegative supermartingales. *Probab. Surv.* **17** 257–317. MR4100718 https://doi.org/10.1214/18-PS321

Howard, S.R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *Ann. Statist.* **49** 1055–1080. MR4255119 https://doi.org/10.1214/20-aos1991

James, B.R. (1975). A functional law of the iterated logarithm for weighted empirical distributions. *Ann. Probab.* **3** 762–772. MR0402881 https://doi.org/10.1214/aop/1176996263

Jamieson, K. and Nowak, R. (2014). Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *48th Annual Conference on Information Sciences and Systems (CISS)* 1–6.

Jamieson, K., Malloy, M., Nowak, R. and Bubeck, S. (2014). Lil' UCB: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the 27th Conference on Learning Theory*. *Proceedings of Machine Learning Research* **35** 423–439.

Kalogerias, D.S., Nikolakakis, K.E., Sarwate, A.D. and Sheffet, O. (2020). Best-Arm Identification for Quantile Bandits with Privacy. arXiv:2006.06792.

Kalyanakrishnan, S., Tewari, A., Auer, P. and Stone, P. (2012). PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning* 655–662.

Kaufmann, E., Cappé, O. and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *J. Mach. Learn. Res.* **17** Paper No. 1, 42. MR3482921

Mannor, S. and Tsitsiklis, J.N. (2003/04). The sample complexity of exploration in the multi-armed bandit problem. *J. Mach. Learn. Res.* **5** 623–648. MR2247994 https://doi.org/10.1007/978-3-540-45167-9_31

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* **18** 1269–1283. MR1062069

Schreuder, N., Brunel, V.-E. and Dalalyan, A. (2020). A nonasymptotic law of iterated logarithm for general M-estimators. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. *Proceedings of Machine Learning Research, PMLR* **108** 1331–1341.

Shorack, G.R. and Wellner, J.A. (1986). *Empirical Processes with Applications to Statistics*. *Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. New York: Wiley. MR0838963

Smirnov, N.V. (1944). Approximate laws of distribution of random variables from empirical data. *Uspekhi Mat. Nauk* **10** 179–206. MR0012387

Stephanou, M., Varughese, M. and Macdonald, I. (2017). Sequential quantiles via Hermite series density estimation. *Electron. J. Stat.* **11** 570–607. MR3619317 https://doi.org/10.1214/17-EJS1245

Szörényi, B., Busa-Fekete, R., Weng, E. and Hüllermeier, P. (2015). Qualitative multi-armed bandits: A quantile-based approach. In *Proceedings of the 32nd International Conference on Machine Learning* 1660–1668.

Talagrand, M. (2005). *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*. *Springer Monographs in Mathematics*. Berlin: Springer. MR2133757

Torossian, L., Garivier, A. and Picheny, V. (2019). X-armed bandits: Optimizing quantiles, CVaR and other risks. In *'Asian Conference on Machine Learning', PMLR* 252–267.

Khintchine, A. (1924). Über einen Satz der Wahrscheinlichkeitsrechnung. *Fund. Math.* **6** 9–20.

Ville, J. (1939). *Étude Critique de la Notion de Collectif*. Gauthier-Villars, Paris. MR3533075

Yu, J.Y. and Nikolova, E. (2013). Sample complexity of risk-averse bandit-arm selection. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Zhao, S., Zhou, E., Sabharwal, A. and Ermon, S. (2016). Adaptive concentration inequalities for sequential decision problems. In *30th Conference on Neural Information Processing Systems*.

Received August 2019 and revised April 2021