Quantile Bandits for Best Arms Identification

Mengyan Zhang [1 2] **Cheng Soon Ong** [2 1]

Abstract

We consider a variant of the best arm identification task in stochastic multi-armed bandits. Motivated by risk-averse decision-making problems, our goal is to identify a set of m arms with the highest τ-quantile values within a fixed budget. We prove asymmetric two-sided concentration inequalities for order statistics and quantiles of random variables that have non-decreasing hazard rate, which may be of independent interest. With these inequalities, we analyse a quantile version of Successive Accepts and Rejects (Q-SAR). We derive an upper bound for the probability of arm misidentification, the first justification of a quantile based algorithm for fixed budget multiple best arms identification. We show illustrative experiments for best arm identification.

	Mean	0.5-Quantile		Mean	0.8-Quantile
A	3.50	**3.50**	C	1.45	2.33
B	**4.00**	2.80	D	**2.50**	**4.00**
OptArm	B	A	Gap	1.05	1.67

Table 1. Summary statistics for toy example rewards

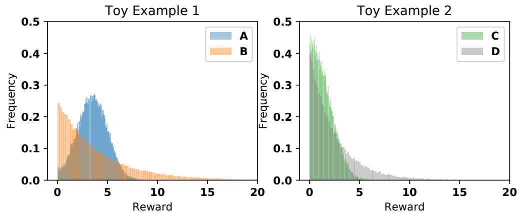

Figure 1. Toy example reward histograms.

1. Introduction

Multi-Armed Bandits (MAB) are sequential experimental design problems where an agent adaptively chooses one (or multiple) option(s) among a set of choices based on certain policies. We refer to "options" as "arms" in MAB problems. In contrast with *full feedback* online decision-making problems where sample rewards for all arms are fully observable to agents in each round, in MAB tasks the agent only observes the sample reward from the selected arm in each round, with no information about other arms.

One of the key steps in the theoretical analysis for bandit algorithms is concentration inequalities, which provides bounds on how a random variable deviates from some statistical summary (typically its expected value). Inspired by the approach in Boucheron & Thomas (2012), we propose in Section 2 new concentration inequalities for order statistics and quantiles of distributions with non-decreasing hazard rates (Definition 1). Previous work derived concentration bounds of quantiles via the empirical cumulative distribution function (c.d.f.). Our proof uses a new approach based on the extended Exponential Efron-Stein inequality (Theorem 3), and non-trivially extends the concentration of order statistics (Boucheron & Thomas, 2012). Our proposed concentration inequality can be useful for various applications, for example, the multi-armed bandits problem as illustrated in this work, learning theory, A/B-testing (Howard & Ramdas, 2019), and model selection (Massart, 2000).

We apply the proposed concentration inequality to the *Best Arm Identification* (BAI) task with *fixed budget* (Audibert et al., 2010; Bubeck et al., 2013). The goal of BAI is to select the best arm (in our case the top $m \geq 1$ arms) after the exploration phase (i.e. budget has run out). The agent can explore the environment and perform actions during the exploration phase without penalty. In contrast to the majority of previous work which identifies optimal arms by summarising a distribution by its mean, we address risk-averse bandits by evaluating the quality of arms by a quantile value of the reward distribution. We consider the bandit problem of quantile based m best arms identification, where the goal is to identify a set of m arms with the highest τ-quantile values. To the best of our knowledge, existing quantile work focuses on single optimal arm identification, and we are the first work that addresses multiple best arms identification for fixed budget setting with respect to the τ-quantile. Our proposed algorithm is in Section 4.

Studying quantile concentrations and identifying arms with optimal quantiles have been shown to be beneficial for many cases, such as when the rewards are qualitative (Szörényi

[1] The Australian National University [2] Data61, CSIRO. Correspondence to: Cheng Soon Ong <chengsoon.ong@anu.edu.au>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

et al., 2015), when the decision-making is risk-averse (Yu & Nikolova, 2013; David et al., 2018), when the rewards contain outliers (Altschuler et al., 2019), or when the reward distributions are highly skewed (Howard & Ramdas, 2019).

We motivate the use of quantiles as summary statistics with two toy examples, with simulated reward histograms of 2 arms shown in Figure 1 and corresponding summary statistics shown in Table 1. The first example illustrates when risk-averse agents should prefer quantiles. Consider a vaccine testing problem (Cunningham et al., 2016), where the goal is to identify the most reliable vaccine after the exploration phase. The reward is the efficacy of the vaccine. Risk-averse agents tend to exclude vaccine candidates which return a large number of small rewards even though they may have a larger expected value (e.g. B in Figure 1). In such case, a policy guided by a fixed level of quantiles (e.g. 0.5-quantile, the median) will choose a less risky arm (i.e. one with less low rewards). The second example shows when the distributions are skewed, the quantile can provide a bigger gap between arms, which turns out to produce a smaller probability of error (Definition 2). As shown by toy example 2 in Figure 1, the quantile and mean reflect the same preference, but the difference between arms is larger for the 0.8-quantile. The choice of quantile level τ provides an extra degree of modelling freedom, that the practitioner may use to capture domain requirements or to achieve a smaller error probability.

Our **contributions** are: **(i)** Two-sided exponential concentration inequalities for order statistics of rank k (w.r.t its expectation) for a general family (with non-decreasing hazard rate) of random variables. **(ii)** Two-sided exponential concentration inequalities for estimations of τ-quantile (w.r.t. population quantile) based on our results on order statistics. **(iii)** The first τ-quantile based multiple ($m \geq 1$) arms identification algorithm (Q-SAR) for the fixed budget setting. **(iv)** Theoretical analysis for the proposed Q-SAR algorithm, showing an exponential rate upper bound on the probability of error. **(v)** Empirical illustrations for the Q-SAR algorithm, which indicates that Q-SAR outperforms baseline algorithms for the best arms identifications task.

2. Concentration Inequalities

In this section, we show our results for concentration inequalities on order statistics and quantiles. We apply these results to prove error bounds for bandits in Section 4. Order statistics have been used and studied in various areas, such as robust statistics and extreme value theory. The non-asymptotic convergence analysis for order statistics provides a way to understand the probability of order statistics deviates from its expectation, and it is useful to support the decision-making with limited samples under uncertainty.

Let $\{X_t\}_{t=1}^n$ be n i.i.d samples drawn from the distribution of X, and let the $\{X_{(t)}\}_{t=1}^n$ be the order statistics of $\{X_t\}_{t=1}^n$ written in decreasing order, i.e.

$$X_{(1)} \geq X_{(2)} \cdots \geq X_{(n)}. \tag{1}$$

We call $X_{(k)}$ the k rank order statistic, and $X_{(1)}$ and $X_{(n)}$ the maximum and minimum respectively. Denote the (left-continuous) quantile with $\tau \in (0,1)$ of a random variable X by

$$Q^\tau(X) := \inf\{x : \mathbb{P}(X \leq x) \geq \tau\}. \tag{2}$$

We will refer $Q^\tau(X)$ as Q^τ whenever X is clear from the context. With the empirical c.d.f. defined as $\hat{F}_n(x) = \frac{1}{n}\sum_{s=1}^n \mathbb{I}\{X_s \leq x\}$, the empirical τ-quantile with n samples is defined as

$$\hat{Q}_n^\tau := \inf\{x : \hat{F}_n(x) \geq \tau\} = X_{(\lfloor n(1-\tau)\rfloor)}. \tag{3}$$

2.1. Problem Setting

We now introduce the family of reward distributions we consider in this work. We consider continuous non-negative reward random variables X with p.d.f. f and c.d.f. F which satisfy Assumption 1. Note that we are considering distributions that are unbounded on the right.

Definition 1 (Hazard rate)**.** *The hazard rate of a random variable X evaluated at the point x is defined as (assuming density $f(x)$ exists)*

$$h(x) := \lim_{\theta \to 0} \frac{\mathbb{P}\left(x \leq X \leq x+\theta | X \geq x\right)}{\theta} = \frac{f\left(x\right)}{1-F\left(x\right)}.$$

Assumption 1 (IHR)**.** *We consider reward distributions with non-negative support $[0, \infty)$ having* non-decreasing hazard rate (IHR)*, i.e. for all $x_1 \geq x_2 \geq 0$, the hazard rate satisfies $h(x_1) \geq h(x_2)$. We further suppose that the lower bound of the hazard rate $L := \inf_x h(x) > 0$.*

The IHR assumption is useful in survival analysis. If the hazard rate $h(x)$ increases as x increases, $\mathbb{P}\left(x \leq X \leq x+\theta | X \geq x\right)$ will increase as well. For example, a man is more likely to die within the next month when he is 88 years old than when he is 18 years old. Common examples of IHR distributions include the absolute Gaussian, exponential, and Gumbel distributions. Log-concave distributions are IHR and have widely been applied to economics, search theory, monopoly theory (Bagnoli & Bergstrom, 2005).

In the following sections, we show our main results about the concentration bounds for order statistics and quantiles, and details are provided in Appendix B.

2.2. Order Statistics

Our goal is to derive two exponential rate concentration bounds in terms of rank k order statistics out of n samples. A roadmap of the technical derivations needed is deferred to Section 3, and we present only the lemmas needed for analysing BAI in this section. For $\gamma \geq 0$, the right and left tail are respectively,

$$\mathbb{P}\left(X_{(k)} - \mathbb{E}[X_{(k)}] \geq d^r_{k,\gamma}\right) \leq \exp(-\gamma), \tag{4}$$
$$\mathbb{P}\left(\mathbb{E}[X_{(k)}] - X_{(k)} \geq d^l_{k,\gamma}\right) \leq \exp(-\gamma). \tag{5}$$

where $d^r_{k,\gamma}, d^l_{k,\gamma}$ are the right and left confidence intervals.

To derive such bounds, we consider the entropy method and the Cramér-Chernoff method (Boucheron et al., 2013). These results are used to derive the following lemmas for deviation of order statistics. Recall L is the lower bound of the hazard rate, k is chosen from the positive integers $\mathbb{N}^*$.

Lemma 1 (Right Tail Concentration Bounds for Order Statistics)**.** *Define $v^r := \frac{2}{kL^2}$, $c^r := \frac{2}{kL}$. Under Assumption 1 , for all $\lambda \in [0, 1/c^r)$, and all $k \in [1, n) \wedge \mathbb{N}^*$, we have*

$$\log \mathbb{E}[\exp\left(\lambda\left(X_{(k)} - \mathbb{E}[X_{(k)}]\right)\right)] \leq \frac{\lambda^2 v^r}{2(1 - c^r\lambda)}. \tag{6}$$

For all $\gamma \geq 0$, we obtain the concentration inequality

$$\mathbb{P}\left(X_{(k)} - \mathbb{E}[X_{(k)}] \geq \sqrt{2v^r\gamma} + c^r\gamma\right) \leq \exp(-\gamma). \tag{7}$$

Lemma 2 (Left Tail Concentration Bounds for Order Statistics)**.** *Define $v^l := \frac{2(n-k+1)}{(k-1)^2L^2}$. Under Assumption 1, for all $\lambda \geq 0$, and all $k \in (1, n] \wedge \mathbb{N}^*$, we have*

$$\log \mathbb{E}[\exp\left(\lambda\left(\mathbb{E}[X_{(k)}] - X_{(k)}\right)\right)] \leq \frac{\lambda^2 v^l}{2}. \tag{8}$$

For all $\gamma \geq 0$, we obtain the concentration inequality

$$\mathbb{P}\left(\mathbb{E}[X_{(k)}] - X_{(k)} \geq \sqrt{2v^l\gamma}\right) \leq \exp(-\gamma). \tag{9}$$

The above results imply $X_{(k)}$ is sub-gamma on the right tail with v^r and c^r, and sub-Gaussian on the left tail with v^l. The two different rates of tail bounds reflect the nature of the asymmetric (non-negative) random variable assumption.

Comparison with related work: The result in Boucheron & Thomas (2012) is a special case of Lemma 1, i.e. when $X_{(k)}$ is the order statistics of absolute Gaussian random variable and $k = n/2$ or 1 only (i.e. median and maximum). We extended their results as follows: **(i)** Our results work for a general family of distributions under Assumption 1; **(ii)** We provide a new left tail concentration bound in Lemma 2; **(iii)** The concentration result in Boucheron & Thomas (2012) only covered the cases $K = n/2$ or $k = 1$, while their results can be trivially extended to $k \in [1, n/2]$. We show a non-trivial further extension to $k \in [1, n)$ on right tail and $k \in (1, n]$ on left tail. **(iv)** While we follow similar proof technique (i.e. entropy method) as shown in Boucheron & Thomas (2012), we claim novelty of several propositions and lemmas which enables us to derive the new results, which can be independent interest, see Remark 1 in Section 3 for details.

Kandasamy et al. (2018) extended the result from Boucheron & Thomas (2012) to exponential random variables, but we have a tighter left tail bound and a more general analysis in terms of distributions and ranks. To our best knowledge, we are the first work studying the two-side order statistic concentration for general IHR distributions.

2.3. Quantiles

Now we convert the concentration results for order statistics to quantiles, namely our goal is to derive two concentration bounds, for $\gamma \geq 0$,

$$\mathbb{P}\left(\hat{Q}^\tau_n - Q^\tau \geq d^r_{n,\tau,\gamma}\right) \leq \exp(-\gamma), \tag{10}$$
$$\mathbb{P}\left(Q^\tau - \hat{Q}^\tau_n \geq d^l_{n,\tau,\gamma}\right) \leq \exp(-\gamma). \tag{11}$$

By definition of empirical quantile in Eq. (3), the empirical quantile is the order statistic with the rank expressed as a function of quantile level, i.e. $\hat{Q}^\tau_n = X_{(\lfloor n(1-\tau)\rfloor)}$. David (2003) studied the relationship between the expected order statistics and the population quantile under Assumption 2, we use their results (Theorem 1) to convert the concentration results of order statistics to quantiles. The constant b depends on the density around τ-quantile. Linking Theorem 1 and Lemma 1 or 2 gives the concentration of quantiles (Theorem 2).

Assumption 2. *Assume the probability density function of random variable X is continuously differentiable.*

Theorem 1 (Link expected order statistics and population quantile (David, 2003))**.** *Under Assumption 2, there exists constant $b \geq 0$ and scalars w_n such that $w_n = \frac{b}{n}$, then $|\mathbb{E}[X_{(\lfloor n(1-\tau)\rfloor)}] - Q^\tau| \leq w_n$.*

Theorem 2 (Two-side Concentration Inequality for Quantiles)**.** *Recall $v^r = \frac{2}{kL^2}$, $v^l = \frac{2(n-k+1)}{(k-1)^2L^2}$, $c^r = \frac{2}{kL}$, $w_n = \frac{b}{n}$. For quantile level $\tau \in (0, 1)$, let rank $k = \lfloor n(1-\tau)\rfloor$. Under Assumption 1 and 2, we have*

$$\mathbb{P}\left(\hat{Q}^\tau_n - Q^\tau \geq \sqrt{2v^r\gamma} + c^r\gamma + w_n\right) \leq \exp(-\gamma).$$
$$\mathbb{P}\left(Q^\tau - \hat{Q}^\tau_n \geq \sqrt{2v^l\gamma} + w_n\right) \leq \exp(-\gamma).$$

Our confidence intervals depend on the number of samples n, the quantile level τ and the lower bound of hazard rate L.

Our bound is tighter when L is larger or τ is smaller. Our methods also provide a way to understand the two-sided asymmetric concentration for quantiles when the distributions are asymmetric.

Bias term: Compared with the concentration results of order statistics (Lemma 1 and 2), there is an extra term w_n in Theorem 2, which has the rate $\mathcal{O}(\frac{1}{n})$ and comes from the gap between the expected order statistics and population quantile (Theorem 1). Our results show that although the quantile estimations based on single order statistics with finite samples are biased, the concentration of empirical quantiles to population quantiles has the same convergence rate as the concentration of order statistics to its expectation, i.e. both with convergence rate $\mathcal{O}(\frac{1}{\sqrt{n}})$. One could potentially consider more than one expected order statistic around the true quantile value to obtain a better estimate, which is beyond the scope of this work. But as we will see in Corollary 1 the bias term does not affect our error bound for best arms identification.

Comparison to related work: The main difference of our approach is that we directly analyse the object of interest (the random variable itself) instead of the the value of its distribution. In constrast to proof techniques shown in the literature, our approach is based on the entropy method and does not convert empirical quantiles to empirical c.d.f. Instead, we study the concentration bound based on the spacing between consecutive order statistics (See Appendix 3 for details). We provide concentration inequalities to two distinct quantities, the expected order statistics and the true quantile. Apart from Boucheron & Thomas (2012) we are not aware of any other work on order statistics.

There are two types of concentration inequalities for quantile estimations with the exponential rate in the literature. Because the empirical quantile is non-linear, the two types of concentration inequalities are not interchangable. Most of the literature focuses on the concentration of empirical quantile at level $\tau \pm \delta$ ($\forall \delta \geq 0$ s.t. $\tau + \delta \leq 1$ and $\tau - \delta \geq 0$) to the population quantile at level τ, i.e.

$$\mathbb{P}\left(\hat{Q}_t^{\tau-\delta} \geq Q^\tau\right) \leq \exp(-\gamma), \tag{12}$$

$$\mathbb{P}\left(\hat{Q}_t^{\tau+\delta} \leq Q^\tau\right) \leq \exp(-\gamma). \tag{13}$$

Note that (by comparing with Eq. (10) and (11)) this paper considers a deviation in the quantile, and not τ. Based on assuming the c.d.f. is continuous and strictly increasing, this type of concentration can benefit from directly converting the concentration to quantiles to the concentration to c.d.f.. For example, Szörényi et al. (2015); Torossian et al. (2019) showed concentration inequalities for quantiles with rate $\mathcal{O}(\sqrt{\frac{\log n}{n}})$; Howard & Ramdas (2019) improved previous work and proved confidence sequences for quantile estimations with the confidence width shrinks in rate $\mathcal{O}(\sqrt{\frac{\log \log n}{n}})$.

Our results are about a different aspect of deviation, where the concentration is between empirical quantile at level τ and the population quantile at level τ, as shown in Eq. (10) and (11). Tran-Thanh & Yu (2014); Yu & Nikolova (2013) proposed concentration for quantile estimations based on the concentration of order statistics under Chebyshev's inequality, while their concentration inequality is not in exponential rate (in terms of γ) thus their bounds decrease much slower when γ increases. A different set of assumptions are needed for this type of concentration to achieve an exponential rate. For example, Cassel et al. (2018) assumed Lipschitz continuity of c.d.f. and derive bounds with rate $\mathcal{O}(\sqrt{\frac{\log n}{n}})$. By assuming that the c.d.f. is continuous and strictly increasing, and knowledge of the density around quantiles, Kolla et al. (2019) provided an exponential concentration inequality with $\mathcal{O}(\frac{1}{\sqrt{n}})$ by using a generalized notion of an inverse empirical c.d.f. to be able to apply the DKW inequality (Dvoretzky et al., 1956). Their confidence intervals on both two sides are decreasing in rate $\mathcal{O}(\sqrt{\frac{1}{n}})$, which is comparable to ours. Our bound can further benefit from the case where quantile level τ is small or the lower bound of hazard rate L is big.

3. Roadmap for Concentration Proofs

In this section, we provide the roadmap to the technical results behind Section 2.2, which may be of independent interest for other applications. Figure 2 summarises the roadmap of the theorems, and the detailed proof is shown in Appendix B. This section is useful to readers interested in techical aspects of concentration inequalities, but may be skipped by others. We briefly introduce the *entropy method* here, and we refer the reader to Boucheron et al. (2013) Chapter 6 for a comprehensive review. The logarithmic moment generating function of the random variable X is defined as

$$\psi_X(\lambda) := \log \mathbb{E}[\exp(\lambda X)]. \tag{14}$$

Define the *entropy* (different from Shannon entropy) of a non-negative random variable X as

$$\operatorname{Ent}[X] := \mathbb{E}[X \log X] - \mathbb{E}[X] \log \mathbb{E}[X]. \tag{15}$$

Then normalising $\operatorname{Ent}[\exp(\lambda X)]$ by $\mathbb{E}[\exp(\lambda X)]$ gives us an expression in terms of the logarithmic moment generating function $\psi_{X-\mathbb{E}[X]}(\lambda)$ (refer to (14)), i.e.

$$\frac{\operatorname{Ent}\left(e^{\lambda X}\right)}{\mathbb{E}e^{\lambda X}} = \lambda \psi'_{X-\mathbb{E}[X]}(\lambda) - \psi_{X-\mathbb{E}[X]}(\lambda). \tag{16}$$

One can derive an upper bound for $\operatorname{Ent}[\exp(\lambda X)]$ by the modified logarithm Sobolev inequality (Ledoux, 2001) (see

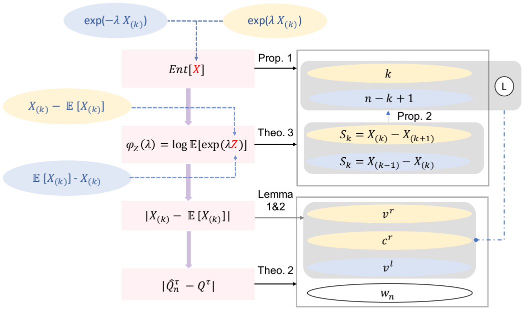
$\exp(-\lambda X_{(k)})$
$\exp(\lambda X_{(k)})$
$Ent[X]$
Prop. 1
k
$n-k+1$
L
$X_{(k)} - \mathbb{E}[X_{(k)}]$
Prop. 2
Theo. 3
$\varphi_Z(\lambda) = \log \mathbb{E}[\exp(\lambda Z)]$
$S_k = X_{(k)} - X_{(k+1)}$
$\mathbb{E}[X_{(k)}] - X_{(k)}$
$S_k = X_{(k-1)} - X_{(k)}$
Lemma 1&2
$|X_{(k)} - \mathbb{E}[X_{(k)}]|$
v^r
c^r
v^l
Theo. 2
$|\hat{Q}_n^\tau - Q^\tau|$
w_n

Figure 2. Roadmap of concentration proof. Variables related to left tail bound and right tail bound are specified by blue and yellow respectively. We upper bound the variables highlighted in pink in the order indicated by arrows between them. The upper bounds are functions of the blocks of variables pointed by solid arrows, with the corresponding theorem names on the arrow.

Theorem 5 in Appendix B). Then by solving a differential inequality, tail bounds can be obtained via Chernoff's bound.

In the following, we apply the entropy method to order statistics and focus on our contributions in terms of the proof technique. We derive a technical result to bound the entropy in Proposition 1. The bound on entropy, along with another technical result on the spacing between consecutive order statistics allows us to derive an exponential Efron-Stein inequality (Theorem 3). Recall $X_{(1)} \geq \ldots \geq X_{(n)}$ are the order statistics of $X_1, \ldots, X_n$. Define the spacing between order statistics of order k and $k-1$ as

$$S_k := X_{(k)} - X_{(k+1)}; \quad S_{k-1} := X_{(k-1)} - X_{(k)}. \quad (17)$$

We first show the upper bounds of entropy in terms of the spacing between order statistics in Proposition 1.

Proposition 1 (Entropy upper bounds)**.** *Define* $\phi(x) := \exp(x) - x - 1$ *and* $\zeta(x) := \exp(x)\phi(-x) = 1 + (x-1)\exp(x)$*. For all* $\lambda \geq 0$*, and for* $k \in [1, n) \wedge \mathbb{N}^*$*,*

$$\text{Ent}\left[\exp(\lambda X_{(k)})\right] \leq k\mathbb{E}\left[\exp(\lambda X_{(k+1)})\zeta\left(\lambda S_k\right)\right]. \quad (18)$$

For $k \in (1, n] \wedge \mathbb{N}^*$*,*

$$\text{Ent}\left[\exp(-\lambda X_{(k)})\right] \leq (n-k+1)\mathbb{E}\left[\exp(-\lambda X_{(k)})\phi\left(-\lambda S_{k-1}\right)\right]. \quad (19)$$

The upper bounds in Proposition 1 are expressed in terms of the corresponding order statistics and the spacing for consecutive order statistics. From Proposition 1, and by normalising entropy as shown in Eq. (16), we show the upper bound of the logarithmic moment generating function of $Z_k := X_{(k)} - \mathbb{E}[X_{(k)}]$ and $Z'_k := \mathbb{E}[X_{(k)}] - X_{(k)}$.

Theorem 3 (Extended Exponential Efron-Stein inequality)**.** *With the logarithmic moment generating function defined in Eq. 14, for* $\lambda \geq 0$ *and* $k \in [1, n) \wedge \mathbb{N}^*$*,*

$$\psi_{Z_k}(\lambda) \leq \lambda \frac{k}{2} \mathbb{E}\left[S_k\left(\exp(\lambda S_k) - 1\right)\right]. \quad (20)$$

For $k \in (1, n] \wedge \mathbb{N}^*$*,*

$$\psi_{Z'_k}(\lambda) \leq \frac{\lambda^2(n-k+1)}{2}\mathbb{E}[S_{k-1}^2]. \quad (21)$$

Observe that the upper bounds in Theorem 3 depend on the order statistics spacings S_k, S_{k-1} in expectation.

The non-decreasing hazard rate assumption (Assumption 1) allows us to upper bound the spacings in expectation. We show the upper bound of expected spacing in Proposition 2. Based on a similar proof technique of Proposition 2, we can further bound Theorem 3 and the results are shown in Lemma 1 and Lemma 2.

Proposition 2. *For any* $k \in [1, n) \wedge \mathbb{N}^*$*, the expectation of spacing* S_k *defined in Eq. (17) can be bounded under Assumption 1,* $\mathbb{E}[S_k] \leq \frac{1}{kL}$*.*

Remark 1 (Novelty of our proof techinique)**.** *The results shown in this section may be of independent interest.* ***(i)*** *In Proposition 1, we show the upper bounds of both* $\text{Ent}\left[\exp(\lambda X_{(k)})\right]$ *and* $\text{Ent}\left[\exp(-\lambda X_{(k)})\right]$ *for all rank* k *except extremes. This allows two-sided tail bounds to hold for all ranks except extremes.* ***(ii)*** *In Theorem 3, we show upper bounds of logarithmic moment generating function for both* $X_{(k)} - \mathbb{E}[X_{(k)}]$ *and* $\mathbb{E}[X_{(k)}] - X_{(k)}$ *w.r.t the order statistics spacing in expectation. The upper bound of* $\mathbb{E}[X_{(k)}] - X_{(k)}$ *is tighter (sub-Gaussian).* ***(iii)*** *We propose an upper bound for the expected order statistics spacing* $S_k = X_{(k)} - X_{(k+1)}$ *in Proposition 2.*

4. Quantile Bandits Policy: Q-SAR

We consider the setting of multi-armed bandits with a finite number of arms K. For each arm $i \in \mathcal{K} = \{1, ..., K\}$, the rewards are sampled from an unknown stationary reward distribution F_i. We assume arms are independent. The environment ν consists of the set of all reward distributions F_i, i.e. $\nu := \{F_i : i \in \mathcal{K}\}$. The agent makes a sequence of decisions based on a policy π for N rounds, where each round is denoted by $t \in \{1, \ldots, N\}$. We denote the arm chosen at round t as A_t, and $T_i(t)$ as the number of times for arm i was chosen at the end of round t, i.e. $T_i(t) := \sum_{s=1}^{t} \mathbb{I}\{A_s = i\}$. At round t, the agent observes reward $X^{A_t}_{T_{A_t}(t)}$ sampled from F_{A_t}.

The quality of an arm is determined by the τ-quantile of its reward distribution. Arms with higher τ-quantile values are better. We order the arms according to optimality as $o_1, \ldots, o_K$ s.t. $Q^\tau_{o_1} \geq \cdots \geq Q^\tau_{o_K}$. The optimal arm set of size m is $\mathcal{S}^*_m = \{o_1, \ldots, o_m\}$. Without loss of generality, we assume $\mathcal{S}^*_m$ is unique. Following Audibert et al. (2010); Bubeck et al. (2013), we formulate our objective by the probability of error.

Definition 2 (Probability of error/misidentification)**.** *We denote* $\mathcal{S}^N_m \subset \mathcal{K}$ *as the set of* m *arms returned by the policy*

Algorithm 1 Q-SAR

```
Let the active set $\mathcal{A}_1 = \{1, ..., K\}$, the accepted set $\mathcal{M}_1 = \emptyset$, the number of arms left to find $l_1 = m$, $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}, n_0 = 0$, and for $p \in \{1, ..., K-1\}$, $n_p = \left\lceil \frac{1}{\overline{\log}(K)} \frac{N-K}{K+1-p} \right\rceil$.
for each phase $p = 1, 2, ..., K-1$ do
    (1) For each $i \in \mathcal{A}_p$, sample arm $i$ for $n_p - n_{p-1}$ times.
    (2) For each $i \in \mathcal{A}_p$, sort the empirical quantile of arm $i$ in a non-increasing order, denote the sorted arms as $a_{best}, a_2, ..., a_{l_p}, a_{l_p+1}, ..., a_{worst}$ (with ties broken arbitrarily).
    (3) Calculate $\widehat{\Delta}_{best} = \hat{Q}^{\tau}_{a_{best}, n_p} - \hat{Q}^{\tau}_{a_{l_p+1}, n_p}$, $\widehat{\Delta}_{worst} = \hat{Q}^{\tau}_{a_{l_p}, n_p} - \hat{Q}^{\tau}_{a_{worst}, n_p}$.
    if $\widehat{\Delta}_{best} > \widehat{\Delta}_{worst}$ then
        $\mathcal{A}_{p+1} = \mathcal{A}_p \backslash \{a_{best}\}$, $\mathcal{M}_{p+1} = \mathcal{M}_p \cup \{a_{best}\}$,
        $l_{p+1} = l_p - 1$.
    else
        $\mathcal{A}_{p+1} = \mathcal{A}_p \backslash \{a_{worst}\}$, $\mathcal{M}_{p+1} = \mathcal{M}_p$,
        $l_{p+1} = l_p$.
    end if
end for
Return the $m$ arms $\mathcal{S}^N_m = \mathcal{A}_K \cup \mathcal{M}_K$.
```

at the end of the exploration phase. Define the probability of error as

$$e_N := \mathbb{P}\left(\mathcal{S}^N_m \neq \mathcal{S}^*_m\right). \tag{22}$$

The goal is, with a fixed budget of N rounds, to design a policy which returns a set of arms of size $m \geq 1$ so that probability of error e_N is minimised.

We propose a Quantile-based Successive Accepts and Rejects (Q-SAR) algorithm (Algorithm 1), which adapts the *Successive Accepts and Rejects* (SAR) algorithm (Bubeck et al., 2013). We divide the budget N into $K-1$ phases. The number of samples drawn for each arm in each phase remains the same as in the Bubeck et al. (2013). At each phase $p \in \{1, 2, \dots, K-1\}$, we maintain two sets (refer to Figure 3): i) the active set $\mathcal{A}_p$, which contains all arms that are actively drawn in phase p; ii) the accepted set $\mathcal{M}_p$, which contains arms that have been accepted. In each phase p, an arm is removed from the active set, and it is either accepted or rejected. The accepted arm is added into the accepted set. At the end of phase $K-1$, only one arm remains in the active set. The last remaining arm together with and the accepted set (containing $m-1$ arms) form the returned recommendation $\mathcal{S}^N_m$.

We can consider the task of identifying m best arms as grouping arms into the optimal set $\mathcal{S}^*_m$ and non-optimal set. Then intuitively it is easier to firstly group arms which are farther away from the boundary of the two groups, since the estimation error is less likely to influence the grouping (Refer to Figure 3). Q-SAR follows this intuition and determines to accept or reject the arm which is the farthest (based on estimates) from the boundary in each phase.

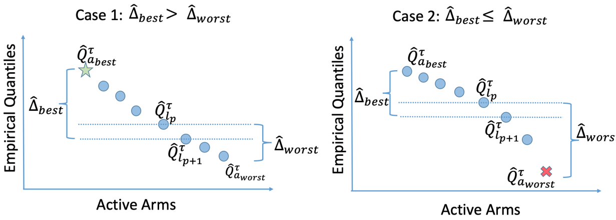
Case 1: $\hat{\Delta}_{best} > \hat{\Delta}_{worst}$
Empirical Quantiles
$\hat{Q}^{\tau}_{a_{best}}$
$\hat{\Delta}_{best}$
$\hat{Q}^{\tau}_{l_p}$
$\hat{Q}^{\tau}_{l_{p+1}}$
$\hat{\Delta}_{worst}$
$\hat{Q}^{\tau}_{a_{worst}}$
Active Arms
Case 2: $\hat{\Delta}_{best} \leq \hat{\Delta}_{worst}$
Empirical Quantiles
$\hat{Q}^{\tau}_{a_{best}}$
$\hat{\Delta}_{best}$
$\hat{Q}^{\tau}_{l_p}$
$\hat{Q}^{\tau}_{l_{p+1}}$
$\hat{\Delta}_{worst}$
$\hat{Q}^{\tau}_{a_{worst}}$
Active Arms

Figure 3. Phase p of Q-SAR illustration. Star indicates accepted arm, cross indicates rejected arm.

We introduce a simplified version of the SAR algorithm. Instead of considering all empirical gaps as SAR algorithm (Bubeck et al., 2013) proposed, Q-SAR decides whether to accept or reject an arm by only comparing two empirical gaps: $\widehat{\Delta}_{best}$ and $\widehat{\Delta}_{worst}$ (defined in Algorithm 1 step (3)). This simplification also applies when using the mean as a summary statistic, and results in an equivalent algorithm to the original SAR in (Bubeck et al., 2013). If $\widehat{\Delta}_{best} > \widehat{\Delta}_{worst}$, the arm with maximum empirical quantile is accepted; otherwise, the arm with minimum empirical quantile is rejected.

When $m = 1$, $\widehat{\Delta}_{worst} = \hat{Q}^{\tau}_{a_{best}} - \hat{Q}^{\tau}_{a_{worst}}$, and $\widehat{\Delta}_{best} = \hat{Q}^{\tau}_{a_{best}} - \hat{Q}^{\tau}_{a_2}$. For all phases, $\widehat{\Delta}_{worst} \geq \widehat{\Delta}_{best}$, thus Q-SAR will keep rejecting arms. In this case, Q-SAR is the same as Q-SR (Algorithm 2, shown in Appendix A).

4.1. Theoretical Analysis

Recall optimality of arms are denoted as $o_1, \dots, o_K$ s.t. $Q^{\tau}_{o_1} \geq \dots \geq Q^{\tau}_{o_K}$. The optimal arm set of size m is $\mathcal{S}^*_m = \{o_1, \dots, o_m\}$. For each arm $i \in \mathcal{K}$, we define the *gap* $\Delta_i \geq 0$ by

$$\Delta_i := \begin{cases} Q^{\tau}_i - Q^{\tau}_{o_{m+1}} & \text{if } i \in \mathcal{S}^*_m; \\ Q^{\tau}_{o_m} - Q^{\tau}_i & \text{if } i \notin \mathcal{S}^*_m. \end{cases}$$

We sort the gaps in a non-decreasing order and denote the i^{th} gap as $\Delta_{(i)}$, i.e. $\Delta_{(1)} \leq \Delta_{(2)} \leq \dots \leq \Delta_{(K)}$. The gaps characterise how separate the arms are and reflect the hardness of the problem. The smaller the (minimum) gaps of the arms are, the harder the BAI task is. We define the *problem complexity* H^{τ} as

$$H^{\tau} = \max_{\{i,j \in \mathcal{K}\}} \frac{8j}{1-\tau} \left(\frac{4\alpha}{L_i^2 \Delta^2_{(j)}} + \frac{\beta_i}{L_i^2 \Delta_{(j)}}\right). \tag{23}$$

where $\alpha = \frac{4(1+\tau)}{1-\tau}, \beta_i = \frac{4}{3}(2L_i + b_i(1-\tau)L_i^2)$, with L_i as the lower bound of hazard rate of arm i.

To bound the probability of error under Q-SAR policy, it is convenient to re-express Theorem 2 such that the deviation between the empirical and true quantile is given by ϵ. Observe that for Q-SAR, we are interested in events of small probability, that is for large values of γ in Theorem 2. In the corollary below, we focus on such events of small probability by considering $\gamma \geq 1$ (i.e. error less than $\frac{1}{e} \approx 0.37$), which allows a simpler expression.

Corollary 1 (Representation of Concentration inequalities for Quantiles)**.** *For $\epsilon > 0$, v^r, v^l, c^r, w_n stay the same as stated in Theorem 2. With $\gamma \geq 1$, Theorem 2 can be represented as*

$$\mathbb{P}\left(\hat{Q}_n^\tau - Q^\tau \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2(v^r + (c^r + w_n)\epsilon)}\right),$$
$$\mathbb{P}\left(Q^\tau - \hat{Q}_n^\tau \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2(v^l + w_n\epsilon)}\right).$$

Applying Corollary 1, we show an upper bound of the probability of error in Theorem 4. Note we assume the total budget is at least $\frac{4}{1-\tau}\overline{\log}(K) + K$, which guarantees that after the initial round each arm has enough samples (See Lemma 3 in Appendix for details). We also present an error bound without assuming the lower bound of the budget in Theorem 8, based on concentration results in Lemma 4.

Theorem 4 (Q-SAR Probability of Error Upper Bound)**.** *For the problem of identifying m best arms out of K arms, with budget $N \geq \frac{4}{1-\tau}\overline{\log}(K) + K$, the probability of error (Definition 2) for Q-SAR satisfies*

$$e_N \leq 2K^2 \exp\left(-\frac{N-K}{\overline{\log}(K)H^\tau}\right),$$

where problem complexity H^τ is defined in Eq. (23).

Observe the error bound depends on the problem complexity and has rate $\mathcal{O}(K^2 \exp(\frac{-N+K}{\log K}))$ w.r.t the number of arms K and budget N. The smaller H^τ is, the smaller the upper bound of error probability is. In the following, we show a sketch of the proof. The detailed proof is provided in Appendix C.

Sketch of Proof. Define the event ξ,

$$\xi := \{\forall i \in \{1, \dots, K\}, p \in \{1, \dots, K-1\}, \left|\hat{Q}_{i,n_p}^\tau - Q_i^\tau\right| < \frac{1}{4}\Delta_{(K+1-p)}\}, \tag{24}$$

where n_p is the number of samples at phase p for arm i. One can upper bound $\mathbb{P}(\bar{\xi})$, i.e. the probability that complementary event of ξ happens, by the union bound and our concentration results (Corollary 1). Then it suffices to show Q-SAR does not make any error on event ξ, which implies (I) no arm from the optimal set is rejected and (II) no arm from non-optimal set is accepted.

To show Q-SAR does not make any error on event ξ, we prove by induction on $p \geq 1$. In phase p, there are $K+1-p$ arms inside of the active set $\mathcal{A}_p$, we sort the arms inside of $\mathcal{A}_p$ and denote them as $\ell_1, \ell_2, \dots, \ell_{K+1-p}$ such that $Q_{\ell_1}^\tau \geq Q_{\ell_2}^\tau \geq \cdots \geq Q_{\ell_{K+1-p}}^\tau$. We assume there is no wrong decision made in all previous $p-1$ phases and prove by contradiction. We assume that one arm from non-optimal set is accepted, or one arm from the optimal set is rejected in phase p. These two assumptions give us

$$\Delta_{(K+1-p)} > \max\left\{Q_{\ell_1}^\tau - Q_{o_{m+1}}^\tau, Q_{o_m}^\tau - Q_{\ell_{K+1-p}}^\tau\right\},$$

which contradicts with the fact that $\Delta_{(K+1-p)} \leq \max\{Q_{\ell_1}^\tau - Q_{o_{m+1}}^\tau, Q_{o_m}^\tau - Q_{\ell_{K+1-p}}^\tau\}$, since there are only $p-1$ arms that have been accepted or rejected at phase p. This concludes the proof. □

Comparison to related work: There are two problems in the standard MAB, namely the *regret minimisation* problems (Auer et al., 2002) and the *Best Arm Identification* (BAI) problems (Audibert et al., 2010). The goal of regret minimisation problems in bandit setting (Auer et al., 2002) is to maximise the cumulative rewards, i.e. minimise the cumulative regret. Best arm identification has been studied for fixed budget (Audibert et al., 2010) and fixed confidence (Even-Dar et al., 2006) settings. The difference between the two settings is how the exploration phase is terminated (when the budget runs out or when the quality of recommendations is at a fixed confidence level). We focus on the fixed budget setting for this work. For a comprehensive review of bandits we refer to Lattimore & Szepesvári (2020).

Previous quantile related BAI work (Szörényi et al., 2015; David & Shimkin, 2016; Yu & Nikolova, 2013; Howard & Ramdas, 2019) mostly focused on another setting of BAI, the fixed confidence setting. Literature concerning quantile bandits with the fixed budget is scarce. The most related work is Tran-Thanh & Yu (2014), which studied functional bandits, with quantiles as one example. They proposed the quantile based batch elimination algorithm. Since their concentration inequality is based on Chebyshev's inequality (and hence not exponential), the upper bound on the probability of error has a rate $\mathcal{O}(K^2/N)$, which is slower than ours. Torossian et al. (2019) considered the fixed budget setting but focused on quantile optimization on stochastic black-box functions, which is different from our setting. As far as we know, Q-SAR is the first policy designed to identify multiple arms with highest τ-quantile values.

The upper bound of error probability of Q-SAR and SAR (Bubeck et al., 2013) have the same rate (in terms of the budget N and number of arms K) up to constant factors. Our constant term is smaller when the minimum lower bound of hazard rate takes a larger value. Unlike the

	Mean	0.5-Quantile	0.8-Quantile
A	1.60	1.35	2.55
B	3.60	**3.50**	5.21
C	**4.00**	2.76	**6.42**
Opt Arm	C	B	C
Min Gap	0.40	0.74	1.21

Table 2. Summary Statistics of Reward Distributions

mean-based algorithm, H^τ depends on the quantile level τ as well. The smaller τ is, the smaller the H^τ is. This can be intuitively explained as needing more samples to estimate higher level quantiles for IHR distributions.

5. Experiments

In this section, we illustrate how the proposed Q-SAR algorithm works on a toy example (Section 5.1) and demonstrate the empirical performance on a vaccine simulation (Section 5.2)[1].

5.1. Illustrative Example

We set up simulated environments by constructing three arms with absolute Gaussian distribution or exponential distribution. The summary statistics of reward distributions are shown in Table 2. We expand the size of environments by replicating arms. Details about the experimental setting are shown in Appendix A.

We design two environments (sets of reward distributions) to show how one can benefit from considering quantiles as summary statistics with our algorithm. The design of the two environments reflects the two motivations in Section 1. Let K be the total number of arms and m be the number of arms to recommend. For each environment, we choose $m = 1$ (single best arm identification) and $m = 5$. We evaluate the probability of errors defined in Eq. (2). As a comparison, we introduce a Quantile-based Successive Rejects (Q-SR) algorithm in Algorithm 2 (Appendix A.3), which is adapted from Successive Rejects algorithm (Audibert et al., 2010) and we modify it to recommend multiple arms.

Environment I: We consider $K = 20 + m$ arms with 15 A arms, m B (optimal) arms, and 5 C arms. The goal is to identify m arms with largest 0.5-quantile (i.e. median). We compare our algorithms with the quantile-based baseline algorithms: **(i)** Quantile uniform sampling (Q-Uniform), where each arm is sampled uniformly and we select the arm with the maximum 0.5-quantile; **(ii)** Quantile Batch Elimination (Q-BE) proposed in Tran-Thanh & Yu (2014) **(iii)** Quantile-based Successive Rejects (Q-SR).

Environment II: We consider $K = 20 + m$ arms with 15 A arms, 5 B arms, and m C (optimal) arms. The goal is to identify m arms with the maximum 0.8-quantile. Both mean and 0.8-quantile provide the same order of arms, while 0.8-quantile can provide a larger gap compared with the mean. According to Theorem 4, the environment with a larger minimum gap has a smaller probability complexity and thus smaller upper bound of the probability of error (it holds for the mean-based algorithm (Bubeck et al., 2013) as well). We compare our algorithms with the baseline algorithms: **(i)** mean-based Successive Accepts and Rejects (SAR) (Bubeck et al., 2013). **(ii)** mean-based Successive Rejects (SR) (Audibert et al., 2010). **(iii)** Quantile-based Successive Rejects (Q-SR).

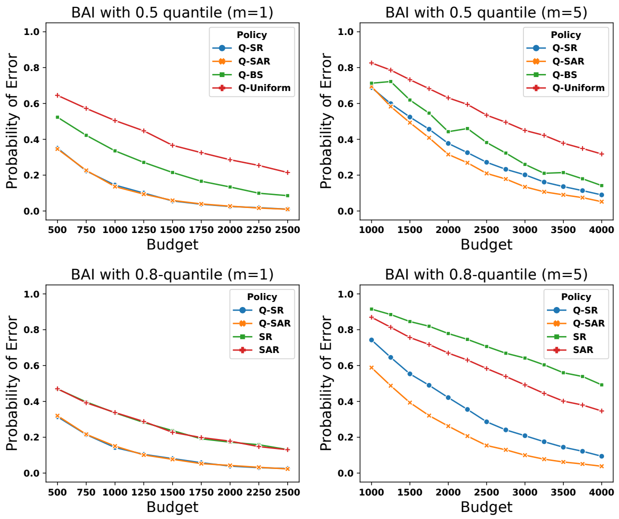

Figure 4. Illustrative examples. The first row shows the Environment I simulation where arms are evaluated by 0.5-quantiles. The second row shows the Environment II simulation where arms are evaluated by 0.8-quantiles (mean provides the same order of arms). We consider the task of recommending a single arm (left column) and recommending multiple arms ($m = 5$, right column). The performance is evaluated in terms of the probability of error (with 5000 independent runs).

Results: We show the empirical probability of error as a function of budget in Figure 4. Q-SAR has the best performance under all settings. Q-SAR and Q-SR has the same performance for the single best arm identification ($m = 1$) task in both environments, while Q-SAR outperforms Q-SR for multiple identifications ($m = 5$). For Environment II, Q-SAR and Q-SR outperform SAR and SR since the gap between the optimal and suboptimal arms is bigger when evaluating arms by 0.8-quantiles than by means, and Q-SAR has a clear lower probability of error than Q-SR when the sample size is small.

5.2. Vaccine Simulation

We consider the problem of identifying optimal strategies for allocating an influenza vaccine. Following Libin et al.

[1] https://github.com/Mengyanz/QSAR

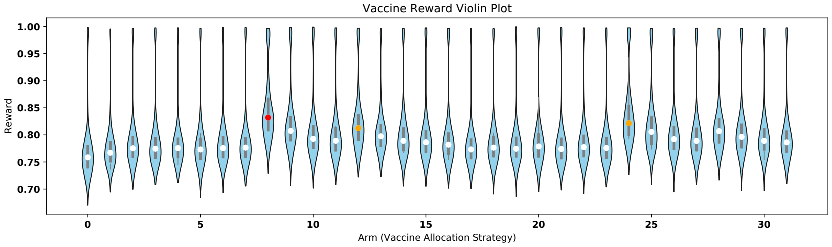

Figure 5. Vaccine Reward Violin Plot. The X-axis represents arms (vaccine allocation strategies), Y-axis represents rewards, which is the proportion of individuals that did not experience symptomatic infection. The circle in each violin represents the median, where the red one is the highest and the orange ones are the second and the third highest. The black line in each violin shows the range of 0.25-quantile to 0.75-quantile.

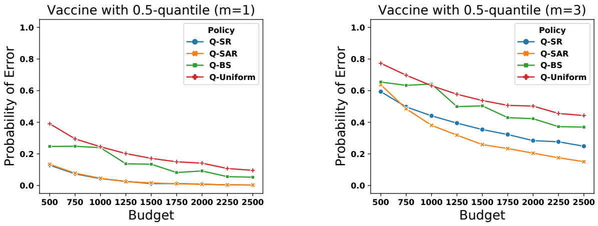

Figure 6. Vaccine BAI Experiments (with 5000 independent runs).

(2017), we format this problem as an instance of the BAI where each vaccine allocation strategy is an arm. Details of allocation strategies are available in the Appendix A.2. The reward of a strategy is defined as the proportion of individuals that did not experience symptomatic infection. We generate 1000 rewards for each strategy by simulating the epidemic for 180 days using FluTE [2] (with basic reproduction number $R_0 = 1.3$). The violin plot of reward samples is shown in Figure 5. The empirical reward distribution of arms are IHR, but with outliers close to 1 (which violate IHR). These outliers are due to the fact that the pathogen does not result in an epidemic in some simulation runs, which does not reflect the efficacy of the vaccine.

We use the median ($\tau = 0.5$) as a robust summary statistic for each strategy. We apply our Q-SAR algorithm on two tasks. (1) the task of identifying $m = 1$ best arm (index 8) with a fixed budget ranging from 500 to 2,500, and (2) the task of identifying $m = 3$ best arms (index 8, 24 and 12) with a fixed budget ranging from 1,000 to 4,000. The performance for the BAI task is shown in Figure 6. We compare our algorithm other quantile-based algorithms only, since $0.5-$quantile and means results in different optimal arms. The empirical evidence shows Q-SAR is the best for multiple best arms identification (when $m = 3$) and is robust to outliers. We leave the theoretical analysis for the outlier robustness of our approach as future work.

6. Conclusion and Discussion

Building on Boucheron & Thomas (2012), we prove a new concentration inequality of order statistics w.r.t its expectation, with which we prove a new concentration inequality for quantile estimation w.r.t. the population quantile. The new concentration inequalities are two-sided, and work for all distributions with non-decreasing hazard rate (IHR). Our assumption of positive (and unbounded) rewards results in asymmetric left and right tail bounds. The concentration inequalities for both order statistics and quantiles have convergence rate $\mathcal{O}(\sqrt{\frac{1}{n}})$. A larger value of L, the lower bound of the hazard rate, results in faster convergence. The proposed inequalities may be of independent interest.

In this paper, we consider the m best arms identification problem with fixed budget. Motivated by risk-averse decision-making, the optimal arm set is determined by the τ-quantiles of reward distributions instead of the mean. The quantile level τ provides an additional level of flexibility for modelling, depending on the risk preference. We proposed the quantile-based successive accepts and rejects (Q-SAR), the first quantile based bandit algorithm for the fixed budget setting. We apply our concentration inequality to prove an upper bound on error probability, which is characterised by the problem complexity. Empirical results show Q-SAR outperforms baseline algorithms for identifying multiple arms. One extension of this work is to allow different sample sizes in Q-SAR, that takes different quantile levels or lower bounds of hazard rate into consideration. Another future work is to derive the matching lower bound of error probability. We hope that this work opens the door towards new bandit approaches for other summary statistics.

Acknowledgements

The authors would like to thank Sébastien Bubeck for constructive suggestions about valid assumptions, Dawei Chen for generating the vaccine data, and Russell Tsuchida and Michael Yang for helpful comments.

[2] https://github.com/dlchao/FluTE

References

Altschuler, J., Brunel, V.-E., and Malek, A. Best arm identification for contaminated bandits. *Journal of Machine Learning Research*, 20(91):1–39, 2019.

Audibert, J., Bubeck, S., and Munos, R. Best arm identification in multi-armed bandits. In *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pp. 41–53. Omnipress, 2010.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

Bagnoli, M. and Bergstrom, T. Log-concave probability and its applications. *Economic Theory*, 26(2):445–469, 2005. ISSN 0938-2259, 1432-0479.

Bernstein, S. On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

Boucheron, S. and Thomas, M. Concentration inequalities for order statistics. *Electron. Commun. Probab.*, 17:12 pp., 2012.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Bubeck, S., Wang, T., and Viswanathan, N. Multiple identifications in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 258–265. JMLR.org, 2013.

Cassel, A., Mannor, S., and Zeevi, A. A general approach to multi-armed bandits under risk criteria. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1295–1306. PMLR, 2018.

Cunningham, A. L., Garçon, N., Leo, O., Friedland, L. R., Strugnell, R., Laupèze, B., Doherty, M., and Stern, P. Vaccine development: From concept to early clinical testing. *Vaccine*, 34(52):6655–6664, 2016. ISSN 0264-410X. doi: 10.1016/j.vaccine.2016.10.016.

David, H. A. *Order statistics.* J. Wiley, 2003. ISBN 978-0-471-02723-2. OCLC: 301102741.

David, Y. and Shimkin, N. Pure Exploration for Max-Quantile Bandits. In *Machine Learning and Knowledge Discovery in Databases*, volume 9851, pp. 556–571. Springer International Publishing, 2016.

David, Y., Szörényi, B., Ghavamzadeh, M., Mannor, S., and Shimkin, N. PAC bandits with risk constraints. In *ISAIM*, 2018.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27(3):642–669, 1956. doi: 10.1214/aoms/1177728174.

Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105, 2006.

Hoeffding, W. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pp. 409–426. Springer, 1994.

Howard, S. R. and Ramdas, A. Sequential estimation of quantiles with applications to A/B-testing and best-arm identification. *arXiv:1906.09712*, 2019.

Kandasamy, K., Krishnamurthy, A., Schneider, J., and Póczos, B. Parallelised bayesian optimisation via thompson sampling. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pp. 133–142. PMLR, 2018.

Kolla, R. K., L.a., P., P. Bhat, S., and Jagannathan, K. Concentration bounds for empirical conditional value-at-risk: The unbounded case. *Operations Research Letters*, 47(1): 16–20, 2019.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Ledoux, M. *The concentration of measure phenomenon*. American Mathematical Soc., 2001.

Libin, P., Verstraeten, T., Roijers, D. M., Grujic, J., Theys, K., Lemey, P., and Nowé, A. Bayesian best-arm identification for selecting influenza mitigation strategies. *CoRR*, abs/1711.06299, 2017.

Massart, P. Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, 2000.

Szörényi, B., Busa-Fekete, R., Weng, P., and Hüllermeier, E. Qualitative multi-armed bandits: A quantile-based approach. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1660–1668. JMLR.org, 2015.

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

Tran-Thanh, L. and Yu, J. Y. Functional Bandits. *arXiv:1405.2432 [cs, stat]*, 2014. arXiv: 1405.2432.

Yu, J. Y. and Nikolova, E. Sample complexity of risk-averse bandit-arm selection. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, pp. 2576–2582. IJCAI/AAAI, 2013.