Huber-Robust Confidence Sequences

Hongjian Wang
Machine Learning Department
Carnegie Mellon University

Aaditya Ramdas
Department of Statistics and Data Science
Machine Learning Department
Carnegie Mellon University

Abstract

Confidence sequences are confidence intervals that can be sequentially tracked, and are valid at arbitrary data-dependent stopping times. This paper presents confidence sequences for a univariate mean of an unknown distribution with a known upper bound on the p-th central moment ($p > 1$), but allowing for (at most) ε fraction of arbitrary distribution corruption, as in Huber's contamination model. We do this by designing new robust exponential supermartingales, and show that the resulting confidence sequences attain the optimal width achieved in the nonsequential setting. Perhaps surprisingly, the constant margin between our sequential result and the lower bound is smaller than even fixed-time robust confidence intervals based on the trimmed mean, for example. Since confidence sequences are a common tool used within A/B/n testing and bandits, these results open the door to sequential experimentation that is robust to outliers and adversarial corruptions.

1 INTRODUCTION

In this paper, we study the problem of robust, sequential mean estimation; we are not just interested in producing a point estimator of the mean (of which there are many robust ones), but in quantifying uncertainty in a manner that is both theoretically optimal and practically tight. Let P be an unknown distribution over $\mathbb{R}$ with a known upper bound σ^2 on the variance, and unknown mean μ that we want to estimate. Let Q be another distribution, seen as a "corruption" of P, such that its total variation (TV) distance from P is at most ε. We assume that an infinite stream of data $X_1, X_2, \dots$ is generated i.i.d. according to the corrupted distribution Q. The task that this paper shall focus on, is the derivation of a robust *confidence sequence* (CS) (Darling and Robbins, 1967) for μ, which is a sequence of confidence intervals $\{\mathrm{CI}_t\}_{t\in\mathbb{N}^+}$ that guarantees that

$$\forall \text{ stopping times } \tau > 0,\ \mathbb{P}[\mu \in \mathrm{CI}_\tau] \geqslant 1-\alpha, \quad (1)$$

where α is a predefined miscoverage tolerance. Howard et al. (2021) showed that condition (1) is equivalent to the time-uniform coverage condition

$$\mathbb{P}[\forall t \in \mathbb{N}^+,\ \mu \in \mathrm{CI}_t] \geqslant 1-\alpha. \quad (2)$$

Both (equivalent) guarantees are of course much stronger than fixed-time confidence intervals for μ, for which the sample size must be fixed in advance. Instead a CS allows for sequentially tracking the mean and stopping at any data-dependent time, while still having correct inference. Note that the covered parameter μ is the mean of the uncorrupted P, while the mean of the corrupted, data-generating Q may be arbitrarily away from μ, or even undefined. Also, the variance of Q need not exist.

Confidence sequences are increasingly common in sequential experimentation in the IT industry. Even though they are cast in terms of estimation, they can be used to define anytime-valid p-values (Johari et al., 2021; Howard et al., 2021) for composite hypotheses like testing if $\mu \leqslant 0$. CSs are useful for A/B testing or multi-armed bandit testing (Yang et al., 2017; Howard and Ramdas, 2022). In fact, they are explicitly used in internal tools or external services of Adobe (Analytics for Target), Amazon (Evidently) and Netflix (Lindon et al., 2022), for example. CSs are an integral part of multi-armed bandit algorithms, right from the original papers on regret minimization (Lai and Robbins, 1985, Section 4) and best-arm identification (Jamieson et al., 2014). Thus robust CSs, that are theoretically tight and not practically loose, could have immediate downstream applications.

1.1 Our Contributions

The CS that we shall present in this paper is partly inspired by the non-sequential + non-robust study by Catoni (2012),

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

and the recent sequential + non-robust extension due to Wang and Ramdas (2022). Like the latter paper, both the upper and lower endpoints of the CS can be seen as M-estimators. In fact, setting $\varepsilon = 0$ in this work recovers the latter paper's results, which in turn recovers the former paper when interested only in fixed-time CIs under heavy-tailed settings. In this sense, the current paper strengthens the ties between heavy-tailed mean estimation and robust mean estimation, two goals that are often separately pursued but closely related (Prasad et al., 2019).

Our CS not only works nonparametrically for any distribution P and its corruption Q that satisfy the succinct assumptions we stated in the beginning of this section, but also enjoys near-optimal tightness. To elaborate, the width of our robust CS provably matches the lower bound $\sigma\sqrt{\varepsilon}$ for *any* robust estimation methods, sequential or not. Mathematically, we prove that its width is optimal up to a constant factor (currently 28) with high probability. Perhaps both lower and upper bound can be improved by constant factors in future work, but the important point is that one does not appear to pay even this constant factor price in practice. This is to be contrasted with the non-sequential recent work by Lugosi and Mendelson (2021), whose analysis only yields robust confidence intervals with widths at least $96\sqrt{2}$ times the lower bound. The reason behind it is simple: most works, including the above, are not aiming at tight or practical uncertainty quantification — they typically design point estimators and prove mathematically that these are "close" to the true mean achieving the optimal rate in terms of σ, ε and sample size; these rate-optimal theorems can be translated to confidence intervals, but these are loose. It is true that these authors often do not to optimize their constants, but the reality is that if we desire a $(1-\alpha)$-CI using their techniques, we must pay a practical price for their loose analysis. This is because their constants appear in the actual construction of the CI. In contrast, our CS does not have large constants in the actual construction: the factor of 28 only arises in the theoretical analysis. This means that our CS is much tighter in practice (than their method, and than our bounds); see Figure 4 for details.

1.2 Related Work

Robust Statistics. The coinage of the term "robust" in the statistical literature was first due to Box (1953), albeit more in today's sense of nonparametric statistics. Early pioneers in the development of the concept include Tukey (1960); Huber (1964, 1965, 1968, 1973); Bickel (1965), and Hampel (1968, 1971). While historically "loaded with many, sometimes inconsistent, connotations" (Ronchetti and Huber, 2009), the concept of statistical robustness in recent years refers almost exclusively to the one pioneered by Huber (1964), i.e. the setting where data are subject to a certain level of contamination — which may either be on the underlying distribution (Huber, 1964; Maronna et al., 2019; Diakonikolas et al., 2019), or directly on the data (Lecué and Lerasle, 2020; Lugosi and Mendelson, 2021; Minsker and Ndaoud, 2021). In either case, valid inference on functionals of the original, uncontaminated distribution is of interest. The definition of robustness (Definition 1) that we shall use in this paper also follows this recent convention. Also, we discuss the relation to another notion of robustness in Section 4.1, as the premise for a comparative study.

Prominent ideas in robust estimation include, among others, M-estimators and trimming. The use of M-estimators in robust statistical procedures dates back to Huber (1964), which achieve robustness by curbing and bounding the influence that individual data points can make on the statistics. A recent work by Bhatt et al. (2022a), concurrent with ours, shows that the fixed sample size M-estimator due to Catoni (2012) is robustly minimax. On the other hand, trimming refers to the practice of directly discarding outliers (Anscombe, 1960), and trimmed means have long been known to be robust (Bickel, 1965). Recently, a sample splitting variant of the trimmed mean due to Lugosi and Mendelson (2021) was shown to be robust nonparametrically over all finite variance distributions, which we shall discuss in Section 4.2. Other recent techniques in obtaining robust estimators include median of means (Lerasle and Oliveira, 2011; Depersin and Lecué, 2022), self-normalization (Minsker and Ndaoud, 2021), and filtering (Diakonikolas et al., 2019).

Sequential Statistics. Wald (1945) first formulated the problem of sequential statistical testing, as deciding to reject the null or not *every time* a new data point is seen. The sequential type I error is the probability of *ever* rejecting the null when it is true. Apart from sequential tests, the issue of sequential validity leads to a tapestry of concepts including the anytime valid p-value (Johari et al., 2021; Howard et al., 2021), and the e-process (Howard et al., 2020; Grünwald et al., 2019; Ramdas et al., 2022b). The confidence sequence (1) (Darling and Robbins, 1967) is the only concept that focuses on sequential estimation rather than testing, and from which the other tools can be derived. Since Howard et al. (2020), the use of nonnegative (super)martingales in conjunction with Ville's inequality (see Section 2.4) has been increasingly common practice in sequential inference, generalizing Wald's (1945) likelihood ratios favored by the earlier works, and yielding new tools in nonparametric settings (Howard et al., 2021). Sequentially valid methods provide a theoretical safeguard against the potential peril of p-hacking. We refer the reader to a recent survey (Ramdas et al., 2022a) for more details.

There have been a few studies in the literature that address the feasibility of robustifying sequential tests, mostly notably the works by Huber (1965) and by Quang (1985), both of them robustifying the likelihood ratio by censoring. Hence they only apply to relatively simple parametric settings, and it is hard to generalize them into our nonparametric setting — indeed our class of distributions consists of both continuous and discrete distributions with any support, and so there is no single common reference measure with respect to which one can define likelihood ratios between pairs of distributions in our class, and so we will need to entirely avoid likelihood ratio style methods.

In summary, we believe this is the first work to design a robust confidence sequence in any setting.

2 PRELIMINARIES

2.1 Notation and Problem Setup

Throughout the paper, $\mathbb{R}$ denotes the set of real numbers and $\mathcal{B}(\mathbb{R})$ its Borel σ-algebra. The diameter of any $I \in \mathcal{B}(\mathbb{R})$ is denoted by $\operatorname{diam}(I)$. $\mathbb{N}$ and $\mathbb{N}^+$ denotes respectively the set of nonnegative and positive integers. We fix a universal probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$ when discussing randomness.

Denote by $\mathcal{M}$ the set of all probability measures over $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Its elements, understood as the distributions of real-valued data, are denoted by upper-case italic letters P, Q and so on. The following important nonparametric subsets of $\mathcal{M}$ are of interest. We denote by $\mathcal{M}^1$ the set of all distributions with finite mean on $\mathbb{R}$, where $\mu : \mathcal{M}^1 \to \mathbb{R}$ denotes the mean functional

$$\mu(P) = \int x \, \mathrm{d}P. \tag{3}$$

For $p > 1$, denote by $\mathcal{M}^p$ the subset of $\mathcal{M}^1$ of those distributions with that have both means and finite p-th moments on $\mathbb{R}$, while $v_p : \mathcal{M}^p \to \mathbb{R}$ denotes the p-th absolute central moment functional

$$v_p(P) = \int |x - \mu(P)|^p \, \mathrm{d}P. \tag{4}$$

Finally, we use $\mathcal{M}^p_\kappa = \{P \in \mathcal{M}^p : v_p(P) \leqslant \kappa\}$ to denote distributions with p-th absolute central moment bounded by κ. The familiar class of distributions with variance bounded by σ^2 is hence denoted by $\mathcal{M}^2_{\sigma^2}$.

Recall that the total variation (TV) distance D_{TV} is defined via $D_{\mathrm{TV}}(P,Q) = \sup_{A\in\mathcal{B}(\mathbb{R})} |P(A) - Q(A)|$ and is a metric on $\mathcal{M}$. Further, it is an *integral probability metric* in the sense that for any pair of real numbers $c_1 < c_2$,

$$D_{\mathrm{TV}}(P,Q) = \frac{1}{c_2 - c_1} \sup_{c_1 \leqslant f \leqslant c_2} \left| \mathop{\mathbb{E}}_{X\sim P} f(X) - \mathop{\mathbb{E}}_{X\sim Q} f(X) \right|, \tag{5}$$

where the supremum is taken over all measurable functions from $\mathbb{R}$ to $[c_1, c_2]$. For any $P \in \mathcal{M}$, we denote by $\mathbb{B}_{\mathrm{TV}}(P, \varepsilon)$ the *closed TV ball* of radius ε around P,

$$\mathbb{B}_{\mathrm{TV}}(P, \varepsilon) = \{Q \in \mathcal{M} : D_{\mathrm{TV}}(P,Q) \leqslant \varepsilon\}. \tag{6}$$

We are now ready to define robust CIs and CSs (applicable beyond the scope of this paper).

2.2 Definition of a Robust CS

Definition 1 (Robust CI and CS)**.** *Let $\mathcal{P} \subseteq \mathcal{M}$ and $\chi : \mathcal{P} \to \mathbb{R}$ be a functional. A sequence of measurable interval-valued functions $\{\mathrm{CI}_t\}$*

$$\mathrm{CI}_t : \mathbb{R}^t \to \Big\{ [l, u] : -\infty \leqslant l \leqslant u \leqslant \infty \Big\} \tag{7}$$

is called a sequence of ε-robust $(1-\alpha)$-confidence intervals over $\mathcal{P}$ for the functional χ, or $(\varepsilon, 1-\alpha)$-RCIs for $(\mathcal{P}, \chi)$ for short, if

$$\forall P \in \mathcal{P},\ \forall Q \in \mathbb{B}_{\mathrm{TV}}(P, \varepsilon),\ \forall t \in \mathbb{N}^+, \\ \mathop{\mathbb{P}}_{X_i \overset{\mathrm{iid}}{\sim} Q} [\chi(P) \in \mathrm{CI}_t(X_1, \dots, X_t)] \geqslant 1 - \alpha. \tag{8}$$

$\{\mathrm{CI}_t\}$ is called an ε-robust $(1-\alpha)$-confidence sequence over $\mathcal{P}$ for χ, or an $(\varepsilon, 1-\alpha)$-RCS for $(\mathcal{P}, \chi)$, if

$$\forall P \in \mathcal{P},\ \forall Q \in \mathbb{B}_{\mathrm{TV}}(P, \varepsilon), \\ \mathop{\mathbb{P}}_{X_i \overset{\mathrm{iid}}{\sim} Q} \left[\forall t \in \mathbb{N}^+,\ \chi(P) \in \mathrm{CI}_t(X_1, \dots, X_t) \right] \geqslant 1 - \alpha. \tag{9}$$

The central question that the current paper seeks to answer is the possibility of constructing a tight RCS for $(\mathcal{M}^2_{\sigma^2}, \mu)$. While our method actually produces RCS for $(\mathcal{M}^p_\kappa, \mu)$ where p may be any real number > 1, we shall focus primarily on the $p = 2$ case for simplicity until Section 5.2.

Remark 1. *Our formulation of robustness via the TV-ball allows for a more general class of contaminations compared to the notion of "gross error model" by Huber (1964, 1965, 1968, 1973) where data are drawn with ε chance from an arbitrary other distribution. To wit, the ε-*contamination neighborhood *in Huber's cited works, is*

$$\mathcal{C}(P, \varepsilon) := \{(1-\varepsilon)P + \varepsilon P' : P' \in \mathcal{M}\}, \tag{10}$$

and it is easy to see that $\mathcal{C}(P, \varepsilon) \subseteq \mathbb{B}_{\mathrm{TV}}(P, \varepsilon)$.

Remark 2. *Trivially, any RCS yields an RCI at a fixed time; and with any RCI, for example the one by Lugosi and Mendelson (2021), one may define an RCS by simply applying a union bound to it (e.g. defining CI_t to be the $\left(\varepsilon, 1 - \frac{\alpha}{T}\right)$-RCI for $t \leqslant T$, and repeating CI_T for $t > T$, or alternatively defining it to be the entire real line before time T and repeating CI_T after T). Our supermartingale techniques lead to much better performance than such trivial constructions, as shown at the end of Section 4.2.*

Remark 3. *We note that in this paper σ, ε are assumed known. As discussed in Wang and Ramdas (2022) which deals with the contamination-free case, not knowing a bound on σ makes nonasymptotic inference impossible. Indeed, the class $\mathcal{M}^2$ (or any $\mathcal{M}^p$) is one that satisfies the impossibility result of Bahadur and Savage (1956, Theorem 1). Hence by Bahadur and Savage (1956, Corollary*

2), no nontrivial CI exists for μ over these classes. The smaller class $\mathcal{M}^2_{\sigma^2}$ (or any $\mathcal{M}^p_\kappa$) with known bounds σ^2 or κ suffers from no such limitation. Catoni (2012) and Bhatt et al. (2022b) use the well-known "Lepskii's method" to obtain point estimators *that adapt to unknown variance (whose dominant terms of the mean squared error depend on the true variance which is smaller than σ^2), but it is important to note that no adaptive CI exists (or can exist). However, we do not know if robust CIs require the contamination parameter ε to be known, and no prior work, to the best of our knowledge, either bypasses this requirement or proves it necessary. While Minsker and Ndaoud (2021, Section 4.1) have a robust mean estimator that "adapts" to ε, the corresponding CI does not (the coverage probability depends on the unknown number of outliers). We leave the problem to future work.*

2.3 Width Lower Bound

It is well-known that robust mean estimators cannot attain consistency (meaning the width of robust CIs cannot shrink to zero), and various information-theoretic lower bounds state that an error scaling as a function (depending on the distribution class) of ε is unavoidable for ε-robust methods (Chen et al., 2018; Lugosi and Mendelson, 2021). For the sake of completeness, we state (and prove in Appendix A) the following lower bound for the diameter of RCIs and RCSs for $(\mathcal{M}^2_{\sigma^2}, \mu)$.

Lemma 1 (Lower bound for RCIs and RCSs for $(\mathcal{M}^2_{\sigma^2}, \mu)$). *Suppose $\{\mathrm{CI}_t\}$ is a sequence of $(\varepsilon, 1-\alpha)$-RCIs for $(\mathcal{M}^2_{\sigma^2}, \mu)$. Then, there is some $P \in \mathcal{M}^2_{\sigma^2}$ such that*

$$\forall t \in \mathbb{N}^+, \quad \mathbb{P}_{X_i \overset{\text{iid}}{\sim} P}\left[\operatorname{diam}(\mathrm{CI}_t) \geqslant \sigma\sqrt{\varepsilon}\right] \geqslant 1 - 2\alpha. \tag{11}$$

Further, if $\{\mathrm{CI}_t\}$ is an $(\varepsilon, 1-\alpha)$-RCS for $(\mathcal{M}^2_{\sigma^2}, \mu)$, then

$$\mathbb{P}_{X_i \overset{\text{iid}}{\sim} P}\left[\forall t \in \mathbb{N}^+,\ \operatorname{diam}(\mathrm{CI}_t) \geqslant \sigma\sqrt{\varepsilon}\right] \geqslant 1 - 2\alpha. \tag{12}$$

2.4 Supermartingales and Ville's inequality

A stochastic process $\{M_t\}_{t\in\mathbb{N}}$ adapted to the filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$ is called a supermartingale if $\mathbb{E}[M_t|\mathcal{F}_{t-1}] \leqslant M_{t-1}$ for all $t \in \mathbb{N}^+$. Nonnegative supermartingales are widely used to construct CSs (Howard et al., 2021; Waudby-Smith and Ramdas, 2023; Wang and Ramdas, 2022) thanks to the following theorem by Ville (1939).

Lemma 2 (Ville's inequality). *Let $\{M_t\}_{t\in\mathbb{N}}$ be a nonnegative supermartingale and $\alpha \in (0,1)$. Then, with probability at least $1-\alpha$, $\sup_{t\in\mathbb{N}^+} M_t \leqslant \alpha^{-1}\mathbb{E}[M_0]$.*

We refer the reader to Howard et al. (2020) for a short, modern proof of Lemma 2.

3 MAIN RESULTS

We now present our RCS construction. It is based on the design of new robust supermartingales, inspired by recent Catoni-style supermartingales (Wang and Ramdas, 2022).

3.1 Robust Nonnegative Supermartingales

Consider the "narrowest possible" influence function in Catoni (2012, Equation (2.3)), given by

$$\phi(x) = \begin{cases} \log 2, & x \geqslant 1, \\ -\log(1 - x + x^2/2), & 0 \leqslant x < 1, \\ \log(1 + x + x^2/2), & -1 \leqslant x < 0, \\ -\log 2, & x < -1. \end{cases} \tag{13}$$

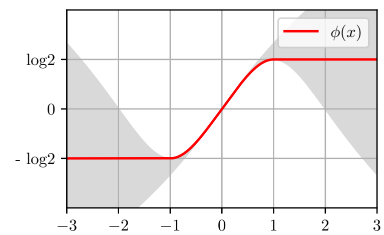

Figure 1: The influence function (13). The shaded region refers to Catoni's condition (14).

It is easy to check (see Figure 1) that it satisfies

$$-\log(1 - x + x^2/2) \leqslant \phi(x) \leqslant \log(1 + x + x^2/2) \tag{14}$$

and $|\phi(x)| \leqslant \log 2$. Catoni (2012) uses any function that satisfies (14) to construct sub-Gaussian M-estimators of mean over $\mathcal{M}^2_{\sigma^2}$, explicitly stating that their objective "should not be confused with robust statistics". However, in a concurrent work, Bhatt et al. (2022a) show that (14) does lead to (fixed-time) robustness if boundedness is further assumed. We shall show that *time-uniform* robustness is also viable by this idea of logarithmic influence function. The key is to obtain the following pair of "robust nonnegative supermartingales" over the entire TV ball around P.

Lemma 3 (Robust Catoni Supermartingales). *For any $P \in \mathcal{M}^2_{\sigma^2}$ and $Q \in \mathbb{B}_{\mathrm{TV}}(P, \varepsilon)$, let*

$$X_1, X_2, \ldots \overset{\text{iid}}{\sim} Q, \quad \text{each } X_t \text{ being } \mathcal{F}_t\text{-measurable}. \tag{15}$$

Let $\{\lambda_t\}_{t\in\mathbb{N}^+}$ be an $\{\mathcal{F}_t\}$-predictable process. Then, the following processes $\{M_t^{\mathsf{RC}}\}_{t\in\mathbb{N}}$, $\{N_t^{\mathsf{RC}}\}_{t\in\mathbb{N}}$, with $M_0^{\mathsf{RC}} = N_0^{\mathsf{RC}} = 1$, are nonnegative supermartingales:

$$M_t^{\mathsf{RC}} = \prod_{i=1}^t \frac{\exp\{\phi(\lambda_i(X_i - \mu(P)))\}}{1 + \lambda_i^2\sigma^2/2 + 1.5\varepsilon}, \tag{16}$$

$$N_t^{\mathsf{RC}} = \prod_{i=1}^t \frac{\exp\{-\phi(\lambda_i(X_i - \mu(P)))\}}{1 + \lambda_i^2\sigma^2/2 + 1.5\varepsilon}. \tag{17}$$

Note in particular that by setting $\varepsilon = 0$ and applying $1 + x \leqslant e^x$ to the denominator, one recovers the (non-robust) Catoni supermartingales of Wang and Ramdas (2022, Lemma 8).

The constant 1.5 appearing above is not magical: it arises because the maximum and minimum values of $\exp(\phi(x))$ are 1/2 and 2, differing by 3/2. Indeed, the proof of Lemma 3 is based on the boundedness of $\exp(\phi(x))$, with the property (5) of TV that leads to the translation of expectation from under Q to under P. Given the centrality of the above lemma, and its uniqueness as being the first robust nonnegative supermartingale we are aware of, we present the proof immediately.

Proof of Lemma 3. Since $|\phi(x)| \leqslant \log 2$, we have $1/2 \leqslant \exp\{\phi(\lambda_t(x - \mu(P)))\} \leqslant 2$. Note that

$$\mathbb{E}\left[\frac{\exp\{\phi(\lambda_t(X_t-\mu(P)))\}}{1+\lambda_t^2\sigma^2/2+1.5\varepsilon}\middle|\mathcal{F}_{t-1}\right] \tag{18}$$

$$=\frac{\mathbb{E}_{X_t\sim Q}\left[\exp\{\phi(\lambda_t(X_t-\mu(P)))\}\right]}{1+\lambda_t^2\sigma^2/2+1.5\varepsilon} \tag{19}$$

$$\leqslant\frac{\mathbb{E}_{X_t\sim P}\left[\exp\{\phi(\lambda_t(X_t-\mu(P)))\}\right]+1.5\varepsilon}{1+\lambda_t^2\sigma^2/2+1.5\varepsilon} \tag{20}$$

$$\leqslant\frac{\mathbb{E}_{X_t\sim P}\left[1+\lambda_t(X_t-\mu(P))+\lambda_t^2(X_t-\mu(P))^2/2\right]+1.5\varepsilon}{1+\lambda_t^2\sigma^2/2+1.5\varepsilon} \tag{21}$$

$$=\frac{\mathbb{E}_{X_t\sim P}\left[1+\lambda_t^2(X_t-\mu(P))^2/2\right]+1.5\varepsilon}{1+\lambda_t^2\sigma^2/2+1.5\varepsilon} \tag{22}$$

$$=\frac{(1+\lambda_t^2 v_2(P)/2)+1.5\varepsilon}{1+\lambda_t^2\sigma^2/2+1.5\varepsilon} \leqslant 1. \tag{23}$$

The first inequality above is due to (5) and $Q \in \mathbb{B}_{\mathrm{TV}}(P, \varepsilon)$. Hence $\{M_t^{\mathsf{RC}}\}$ is a supermartingale. The proof for $\{N_t^{\mathsf{RC}}\}$ is analogous. □

It is clear from the proof above that the i.i.d. assumption $X_i \sim Q \in \mathbb{B}_{\mathrm{TV}}(P, \varepsilon)$ can be relaxed; it suffices to assume that each X_i is generated from *some* distribution (or even a random distribution measurable w.r.t. $\mathcal{F}_{t-1}$) in $\mathbb{B}_{\mathrm{TV}}(P, \varepsilon)$. We keep the i.i.d. assumption (and notation) throughout the paper for simplicity.

3.2 Our Huber-Robust CS

Applying Ville's inequality (Lemma 2) on (16) and (17) leads to our robust confidence sequence, stated next and proved soon after.

Theorem 1 (Huber-robust CS)**.** *Define* $f_t(m) := \sum_{i=1}^t \phi(\lambda_i(X_i - m))$. *We define* $\mathrm{CI}_t^{\mathsf{R}}(X_1, \dots, X_t)$ *to be*

$$\Big\{m \in \mathbb{R} : |f_t(m)| \leqslant \log(2/\alpha) + \sum_{i=1}^t \log\left(1+\lambda_i^2\sigma^2/2+1.5\varepsilon\right)\Big\}. \tag{24}$$

Then, $\{\mathrm{CI}_t^{\mathsf{R}}\}$ *is an* $(\varepsilon, 1-\alpha)$*-RCS for* $(\mathcal{M}^2_{\sigma^2}, \mu)$.

The confidence sequence is defined like M-estimators are; indeed, both end points of the interval $\mathrm{CI}_t^{\mathsf{R}}$ are level points of $f_t(m)$ (where the inequality holds with equality). If a point estimate of the mean is desired, we may define $\widehat{\mu}_t^{\mathsf{R}}$ as the solution to the estimating equation $f_t(m) = 0$. While it is always true that $\widehat{\mu}_t^{\mathsf{R}} \in \mathrm{CI}_t^{\mathsf{R}}$, note that $\widehat{\mu}_t^{\mathsf{R}}$ is not the center of $\mathrm{CI}_t^{\mathsf{R}}$ (despite its apparently symmetric definition).

Proof of Theorem 1. For any $Q \in \mathbb{B}_{\mathrm{TV}}(P, \varepsilon)$, applying Ville's inequality (Lemma 2) to the nonnegative supermartingale $\{M_t^{\mathsf{RC}}\}$ under Q, we infer the following: with probability at least $1 - \alpha/2$, we have that $\forall t \in \mathbb{N}^+$,

$$f_t(\mu(P)) \leqslant \sum_{i=1}^t \log\{1+\lambda_i^2\sigma^2/2+1.5\varepsilon\}+\log(2/\alpha). \tag{25}$$

Applying the same logic to $\{N_t^{\mathsf{RC}}\}$: with probability at least $1 - \alpha/2$, we have that $\forall t \in \mathbb{N}^+$,

$$-f_t(\mu(P)) \leqslant \sum_{i=1}^t \log\{1+\lambda_i^2\sigma^2/2+1.5\varepsilon\}+\log(2/\alpha). \tag{26}$$

Combining them with a union bound concludes the proof of the theorem. □

The confidence sequence, as many previous ones in the literature, is tunable up to a sequence of parameters $\{\lambda_t\}$, which could be seen as weights put on the data sequence $\{X_t\}$. Past works that involve such sequence parameters have oftentimes set it to be a decreasing sequence with a rate of decay approximately $t^{-1/2}$ (Waudby-Smith and Ramdas, 2023; Wang and Ramdas, 2022), which led to a $t^{-1/2}\,\mathrm{polylog}(t)$ rate of shrinkage in the width of the CS. However, as we have seen in Lemma 1, shrinkage of any RCIs (and thus any RCS) to zero at *any* rate is impossible, so intuitive a decaying $\{\lambda_t\}$ would no longer be preferred. Indeed, as we shall soon see in Section 3.3, the CS would unboundedly inflate if $\{\lambda_t\}$ is set either to decrease, or increase, at a polynomial rate. Instead, a constant sequence $\lambda_1 = \lambda_2 = \cdots = \lambda$, where $\lambda \propto \varepsilon^{1/2}\sigma^{-1}$, leads to near-optimal width of the CS.

Under the choice $\lambda = 0.5\varepsilon^{1/2}\sigma^{-1}$, we conduct a simulation with true distribution P being $\mathcal{N}(0, 9)$, along with a $1/9$ chance of contamination from an asymmetric 0.75-Lévy stable distribution with location parameter 0 and skewness parameter $\beta = 0.5$. Our RCS with $\sigma^2 = 9$, $\varepsilon = 1/9$ is compared to a non-robust CS for σ^2-subGaussian distributions of Howard et al. (2021, Eq. (11)) in Figure 2. In practice, our RCS is much better than the concentration bound we will prove in Theorem 2, presented next.

We finally remark on the computational aspects of our RCS in Theorem 1. As the function $f_t(m)$ is continuous and monotonic, root-finding algorithms including the bisection method, secant method, and Brent's method can efficiently calculate the upper and lower endpoints according to (24).

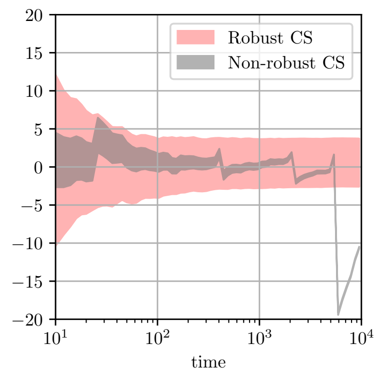

Figure 2: Our robust confidence sequence versus a non-robust confidence sequence by Howard et al. (2021), under contaminated Gaussian data. Our RCS always cover the true mean $\mu(P) = 0$ while non-robust CS does not. The y-axis of the plot scales one-to-one with the lower bound $\sigma\sqrt{\varepsilon} = 1$. Our RCS is tighter than our Theorem 2 would predict.

We can further accelerate the root-finding via *warm starting*, viz., initializing the iterate for finding $\max(\mathrm{CI}_t^{\mathsf{R}})$ with the previous solution $\max(\mathrm{CI}_{t-1}^{\mathsf{R}})$.

3.3 Tuning and Width Analysis

Our tightness bound for the RCS in Theorem 1 is stated in the form of a concentration bound, since the width of (24), unlike traditional confidence intervals, is random. Similar width concentration results have been used in random-width CSs by, for example, Wang and Ramdas (2022). Supposing that the weight sequence is taken constant, $\lambda_1 = \lambda_2 = \cdots = \lambda$, we have the following bound.

Theorem 2. *For any $0 < \delta, \alpha < 1$ and $0 < \varepsilon \leqslant 1/7$, suppose that $t > 4\varepsilon^{-1}\log(4/\alpha\delta)$ and $\lambda^2 = \frac{\varepsilon}{4\sigma^2}$. Under any corrupted distribution $Q \in \mathbb{B}_{\mathrm{TV}}(P, \varepsilon)$, the $(\varepsilon, 1-\alpha)$-RCS for $(\mathcal{M}^2_{\sigma^2}, \mu)$ stated in Theorem 1 satisfies*

$$\underset{X_i \overset{\text{iid}}{\sim} Q}{\mathbb{P}}\left[\mathrm{diam}(\mathrm{CI}_t^{\mathsf{R}}) \leqslant 28\sigma\sqrt{\varepsilon}\right] \geqslant 1-\delta, \tag{27}$$

matching the lower bound $\sigma\sqrt{\varepsilon}$.

Despite its centrality, the proof of Theorem 2 is a bit long, and is thus deferred to Appendix A.

We remark that Theorem 2 answers the question of tuning in a satisfactory manner, since we have shown that setting $\lambda \propto \varepsilon^{1/2}\sigma^{-1}$ leads to an optimal width. We further demonstrate the point by following experiments: we set the true distribution P to be $\mathcal{N}(0, 9)$, with $1/9$ chance of contamination from an asymmetric 0.3-Lévy stable distribution with location parameter 1000 and skewness parameter $\beta = 0.5$. RCS with $\sigma^2 = 9$, $\varepsilon = 1/9$ and $\lambda_t = 0.5\varepsilon^{1/2}\sigma^{-1}t^u$, where $u \in \{-0.5, -0.25, 0, 0.25\}$. The comparison is shown in Figure 3. We can observe from the plots that only setting $\{\lambda_t\}$ to a constant can keep the RCS within constant width.

4 AN EMPIRICAL COMPARISON

Robust uncertainty quantification (i.e. valid $(1-\alpha)$-RCIs or RCSs) has not been a key focus of past work, which has been dominated by point estimators. Thus, we must in some sense extract an RCI from related work as a point of comparison, and we do this below.

4.1 Relationship between Robustness Models

Apart from the robustness model in Definition 1 we consider (which is itself a generalization of the model of Huber (1964)), there is in the literature another model of robustness, where, instead of an ε chance of random contamination to the distribution, there is an adversary who can *replace* up to ε fraction of the data *after* they are generated from the true distribution. This fully adversarial set-up has been considered, among others, by Diakonikolas et al. (2019); Cheng et al. (2019), and Lugosi and Mendelson (2021) whose distribution assumption closely matches ours (both being $\mathcal{M}^2_{\sigma^2}$). To compare our results to those by Lugosi and Mendelson (2021), we first formally relate the robustness model they operate on to ours.

Let $\varepsilon \in (0,1)$ and $t \in \mathbb{N}^+$. Define $\mathcal{R}_{\varepsilon,t}$ as the class of all functions $\mathbb{R}^t \to \mathbb{R}^t$ that change at most $\lfloor \varepsilon t \rfloor$ of the t coordinates into constants. For example, the function $c(x,y,z) = (x, 100, z)$ satisfies $c \in \mathcal{R}_{\varepsilon,3}$ for all $\varepsilon \geqslant 1/3$. We call $\mathcal{R}_{\varepsilon,t}$ the set of *ε-replacements*.

Definition 2 (Replacement robust CI (RRCI))**.** *Let $\mathcal{P} \subseteq \mathcal{M}$ and $\chi : \mathcal{P} \to \mathbb{R}$ be a functional. A sequence of measurable interval-valued functions $\{\mathrm{CI}_t\}$ is called a sequence of ε-replacement robust $(1-\alpha)$-confidence intervals over $\mathcal{P}$ for the functional χ, or $(\varepsilon, 1-\alpha)$-RRCIs for $(\mathcal{P}, \chi)$, if*

$$\forall P \in \mathcal{P},\ \forall t \in \mathbb{N}^+,$$
$$\underset{X_i \overset{\text{iid}}{\sim} P}{\mathbb{P}}\left[\forall c \in \mathcal{R}_{\varepsilon,t},\ \chi(P) \in \mathrm{CI}_t\left(c(X_1, \ldots, X_t)\right)\right] \geqslant 1-\alpha. \tag{28}$$

We emphasize that the probability bound is uniform over $\forall c \in \mathcal{R}_{\varepsilon,t}$ because the adversary may examine the data before deciding on how to corrupt them with a replacement $c \in \mathcal{R}_{\varepsilon,t}$. While it is tempting to sequentialize Definition 2 by, say, putting "$\forall t \in \mathbb{N}^+$" inside the probability bound, such ostensible generalization does not make practical sense. As we have pointed out, Definition 2 is motivated by the setting where an adversary corrupts the data *offline*, after they are entirely generated, which is an inherently fixed-time scenario. By contrast, sequential statistics is concerned with the scenario where we face an infinite stream of data, and inference is conducted *online* along the

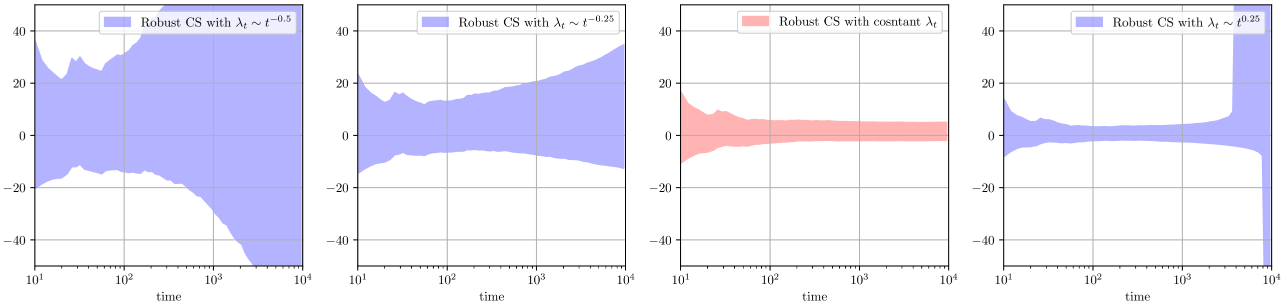

Figure 3: Our robust confidence sequences with different choice of $\{\lambda_t\}$. Indeed as discussed earlier, setting $\{\lambda_t\}$ to a constant multiple of $\sqrt{\varepsilon}/\sigma$ (the third plot in red) controls the width, but other choices do not.

way, instead of post-hoc. This is why we hold ourselves back from defining "RRCS". Still, we define RRCIs largely due to its following relation with RCIs in Definition 1.

Lemma 4. *Any sequence of $(2\varepsilon, 1-\alpha)$-RRCIs for $(\mathcal{P}, \chi)$ is a sequence of $(\varepsilon, 1-(\alpha + \mathrm{e}^{-2t\varepsilon^2}))$-RCIs for $(\mathcal{P}, \chi)$.*

Lemma 4 states that RRCIs can simulate RCIs up to a constant. While similar statements have been known, for example in Li (2019, Corollary 2.1), we prove Lemma 4 in full in Appendix A. We now have the necessary terminology to introduce the trimmed mean RCI.

4.2 Comparison with the Trimmed Mean RCI

A variant of the trimmed mean based on sample splitting is shown by Lugosi and Mendelson (2021) to be robustly concentrated in the replacement sense, which can be used to construct a sequence of RRCIs. We rephrase as follows the main finding of Lugosi and Mendelson (2021).

Theorem 3 (Theorem 1 of Lugosi and Mendelson (2021), rephrased)**.** *Let $\widehat{\mu}_t(x_1, \dots, x_t)$ be (a slight variant of) the trimmed mean of $x_1, \dots, x_t$. Suppose $0 < \alpha < 1$, and $t/2 \geqslant \log(1/4\alpha)$. Then, for all $P \in \mathcal{M}^2_{\sigma^2}$, with probability at least $1-\alpha$ over $X_1, X_2, \overset{\text{iid}}{\sim} P$,*

$$|\widehat{\mu}_t(c(X_1, \dots, X_t)) - \mu(P)| \leqslant 12\sqrt{2\varepsilon'}\sigma + 2\sigma\sqrt{\frac{\log(4/\alpha)}{t/2}} \tag{29}$$

uniformly over all $c \in \mathcal{R}_{\varepsilon,t}$, where

$$\varepsilon' = 8\varepsilon + \frac{12\log(4/\alpha)}{t/2}. \tag{30}$$

In particular, if we are to construct $(\varepsilon, 1-\alpha)$-RRCIs from the concentration bound (29), the width is strictly and deterministically greater than $24\sqrt{2\varepsilon'}\sigma$ which is in turn greater than $96\sqrt{\varepsilon}$.[1] For example, if we assume

$$t \geqslant \max\left\{ \frac{\varepsilon^{-1}\log(4/\alpha)}{0.09}, 2\log(1/4\alpha) \right\}. \tag{31}$$

Then, $[\widehat{\mu}_t \pm 49\sqrt{\varepsilon}\sigma]$ is a valid RRCI. Now recall Lemma 4. To simulate a sequence of $(\varepsilon, 1-\alpha)$-RCIs, one needs a sequence of $(2\varepsilon, 1-\alpha')$-RRCIs where α' is slightly smaller than α. Hence, the RCIs simulated by Theorem 3 are at least $96\sqrt{2\varepsilon}$ wide. Compare this with our RC*S* which is only $28\sqrt{\varepsilon}\sigma$ wide, with high probability (Theorem 2).

As mentioned before, Theorem 2 is quite conservative about the tightness of our RCS. In practice, the advantage of our approach is even more pronounced. The comparison shown in Figure 4 is conducted under $\sigma^2 = 1/\varepsilon = 36$, with contamination from an asymmetric 0.3-Lévy stable distribution with location parameter 1000 and skewness parameter $\beta = 0.5$. When t is small, Lugosi and Mendelson's (2021) split-sample trimming is even undefined (all of the data trimmed); and when their RCIs are defined, they are in a cosmic distance compared to our RCS. Other robust mean estimators over $\mathcal{M}^2_{\sigma^2}$ (e.g. Depersin and Lecué, 2022) suffer from even larger, "galactic" concentration constants.

We remark that the advantage of our approach lies in the fact that it directly solves for the interval, instead of hinging on some concentration inequality on a point estimator (which causes many other works to have large constants, in turn causing very loose CIs). We do use concentration bounds to *analyze* our widths in Theorem 2, but not to *construct* them; thus constant-factor bottlenecks in theoretical analysis do hurt practical performance of other estimators, but not ours. Even if concentration constants are improved in other methods, they are unlikely to beat ours in practice as Ville's inequality is known for its tightness.

Another important advantage of our M-estimation approach lies in its large *break-down point*. That is, it tolerates large amount of corruption. The advantage is also mentioned by Bhatt et al. (2022a) in their fixed-time study, their robust M-estimator breaking down at a higher ε (36% for $\mathcal{M}^2_{\sigma^2}$) compared to the trimmed mean of Lugosi and Mendelson (2021) (which breaks down at $\varepsilon = 1/16$). In our Theorem 1, taking each $\lambda_i = \lambda$, the interval in (24) will not span the entire $\mathbb{R}$ if $t\log 2 > \log(2/\alpha) + t\log(1 + \lambda^2\sigma^2/2 + 1.5\varepsilon)$, which will happen for large t as long as $2 > 1 + \lambda^2\sigma^2/2 + 1.5\varepsilon$. Under our choice $\lambda^2 = \frac{\varepsilon}{4\sigma^2}$ this

[1] In Lugosi and Mendelson (2021, Theorem 1), the first displayed equation is correct (which we quote in (29)), while their second displayed equation is off by a constant.

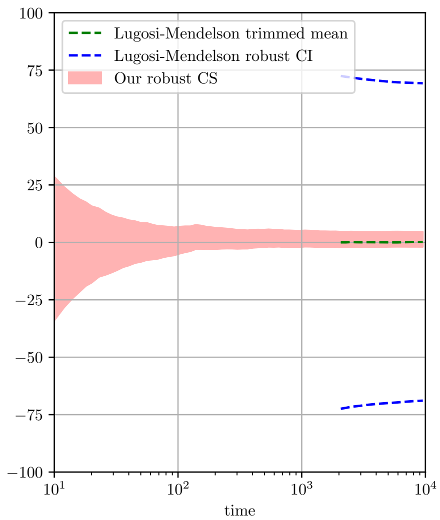

Figure 4: Comparison of our RCS with Lugosi and Mendelson's (2021) RCIs. The lack of dashed lines before $t \approx 2 \times 10^3$ is because all the data are trimmed using the latter approach. Again, the y-axis of the plot scales one-to-one with the lower bound $\sigma\sqrt{\varepsilon} = 1$.

amounts to $\varepsilon < 8/13$.

Finally, we return to the trivial RCS construction mentioned in Remark 2. First, as we can observe in Figure 4, even the best fixed-time $(1-\alpha)$-RCIs are already very loose as they suffer from impractical large constants. Second, the use of union bounds will lead to even poorer results as union bounds can only be tight when the underlying events are nearly independent; here the miscoverage events at subsequent times are highly dependent. Nonnegative supermartingales and Ville's inequality, on the other hand, pay no such (indeed, constant-level) price. It is worth stressing here that *constants matter* in confidence sequences, especially because they are widely used in deployed IT services.

5 EXTENSIONS

5.1 Robust Test Supermartingales

The well-known duality between confidence intervals and hypothesis testing also extends to the robust sequential setting. For some set S in the range of a functional χ, consider the null hypothesis:

$$\mathcal{H}_0 = \{P \in \mathcal{P} : \chi(P) \in S\}. \tag{32}$$

If $\{\mathrm{CI}_t\}$ is an $(\varepsilon, 1-\alpha)$-RCS for $(\mathcal{P}, \chi)$, then the sequential decision to reject $\mathcal{H}_0$ whenever S and CI_t do not intersect attains the following robustified control of type I error:

$$\forall Q \in \mathbb{B}_{\mathrm{TV}}(\mathcal{H}_0, \varepsilon), \quad \mathbb{P}_{X_i \overset{\mathrm{iid}}{\sim} Q}[\text{ever rejecting } \mathcal{H}_0] \leqslant \alpha. \tag{33}$$

Here $\mathbb{B}_{\mathrm{TV}}(\mathcal{H}_0, \varepsilon)$ is the closed ε-neighorhood of the null set $\mathcal{H}_0$ under the TV metric, i.e., $\bigcup_{P \in \mathcal{H}_0} \mathbb{B}_{\mathrm{TV}}(P, \varepsilon)$. Robust tests thus enlarge $\mathcal{H}_0$ to its neighbor $\mathbb{B}_{\mathrm{TV}}(\mathcal{H}_0, \varepsilon)$.

In particular, our RCS in Theorem 1 can be used to robustly test if $\mu(P) = \mu_0$. Equivalently, $\mu(P) = \mu_0$ is rejected when both of the supermartingales in Lemma 3 (replacing $\mu(P)$ with μ_0) surpass $2/\alpha$. Moreover, similar to the discussion of Wang and Ramdas (2022, Section 10.4), each one of the supermartingale pair in Lemma 3 works implicitly to test a one-sided composite null. To wit, whenever

$$M_t^{\mathsf{RC}} = \prod_{i=1}^{t} \frac{\exp\{\phi(\lambda_i(X_i - \mu_0))\}}{1 + \lambda_i^2\sigma^2/2 + 1.5\varepsilon} \tag{34}$$

is greater than $1/\alpha$, reject the one-sided composite

$$\mathcal{H}_0 = \{P \in \mathcal{M}^2_{\sigma^2} : \mu(P) \leqslant \mu_0\}.$$

Then, the sequential type I error is controlled at level α over the enlarged null $\mathbb{B}_{\mathrm{TV}}(\mathcal{H}_0, \varepsilon)$. While it is natural to think that the powers of these tests, characterized by the growth of the test processes under the alternative, are reduced due to the trade-off with robustification (i.e. preventing false rejections due to corrupted data), an exponential rate of growth, similar to the nonrobust tests, can still be achieved under sufficient separation between the null and the alternative. We formalize this with the following dual to Theorem 2, proved in Appendix A:

Corollary 1. *Under the assumptions of Theorem 2 except $t \geqslant 4\varepsilon^{-1}\log(4/\alpha\delta)$, when $\mu(P) > \mu_0 + 14\sigma\sqrt{\varepsilon}$, the process $\{M_t^{\mathsf{RC}}\}$ in (34) grows exponentially with*

$$\mathbb{P}_{X_i \overset{\mathrm{iid}}{\sim} Q}\left[M_t^{\mathsf{RC}} > \frac{\delta}{4}\exp\left(\frac{t\varepsilon}{4}\right)\right] \geqslant 1 - \delta/2. \tag{35}$$

5.2 Infinite Variance

Recalling the notation around (4), we now turn to the question of constructing RCS for $(\mathcal{M}^p_\kappa, \mu)$, for $1 < p < 2$.

Now, instead of (13), we define

$$\phi_p(x) = \begin{cases} \log p, & x \geqslant 1, \\ -\log(1 - x + x^p/p), & 0 \leqslant x < 1, \\ \log(1 + x + |x|^p/p), & -1 \leqslant x < 0, \\ -\log p, & x < -1. \end{cases} \tag{36}$$

Then, akin to Lemma 3, we have,

Lemma 5. *For any $P \in \mathcal{M}^p_\kappa$ and $Q \in \mathbb{B}_{\mathrm{TV}}(P, \varepsilon)$, let $X_1, X_2, \ldots \overset{\mathrm{iid}}{\sim} Q$, each X_t being $\mathcal{F}_t$-measurable. Let $\{\lambda_t\}_{t \in \mathbb{N}^+}$ be a $\{\mathcal{F}_t\}$-predictable process. Then, the following processes $\{M_t^{\mathsf{RC}p}\}_{t \in \mathbb{N}}$, $\{N_t^{\mathsf{RC}p}\}_{t \in \mathbb{N}}$ are nonnega-*

tive supermartingales adapted to $\{\mathcal{F}_t\}$*:*

$$M_0^{\mathsf{RC}^p} = N_0^{\mathsf{RC}^p} = 1, \tag{37}$$

$$M_t^{\mathsf{RC}^p} = \prod_{i=1}^t \frac{\exp\{\phi_p(\lambda_i(X_i - \mu(P)))\}}{1 + \lambda_i^p \kappa/p + (p-1/p)\varepsilon}, \tag{38}$$

$$N_t^{\mathsf{RC}^p} = \prod_{i=1}^t \frac{\exp\{-\phi_p(\lambda_i(X_i - \mu(P)))\}}{1 + \lambda_i^p \kappa/p + (p-1/p)\varepsilon}. \tag{39}$$

Applying Ville's inequality, we can extend Theorem 1.

Theorem 4. *Define* $f_{pt}(m) := \sum_{i=1}^t \phi_p(\lambda_i(X_i - m))$*. We define* $\mathrm{CI}_t^{\mathsf{R}^p}(X_1, \dots, X_t)$ *to be*

$$\Big\{ m \in \mathbb{R} : |f_{pt}(m)| \leqslant \log(2/\alpha) + \sum_{i=1}^t \log\left\{1 + \lambda_i^p \kappa/p + (p-1/p)\varepsilon\right\} \Big\}, \tag{40}$$

Then, $\{\mathrm{CI}_t^{\mathsf{R}^p}\}$ *is an* $(\varepsilon, 1-\alpha)$*-RCS for* $(\mathcal{M}_\kappa^p, \mu)$*.*

This RCS, again, has random widths and satisfies a high probability width bound:

Theorem 5. *For any* $0 < \alpha, \delta < 1$ *and* $0 < \varepsilon \leqslant \frac{p-1}{7p}$*, suppose* $t \geqslant \varepsilon^{-1}\log(4/\alpha\delta)$*, under any corrupted distribution* $Q \in \mathbb{B}_{\mathrm{TV}}(P, \varepsilon)$*, the* $(\varepsilon, 1-\alpha)$*-RCS* $\{\mathrm{CI}_t^{\mathsf{R}^p}\}$ *with* $\lambda_t = (\varepsilon/\kappa)^{1/p}$ *for all t satisfies*

$$\mathbb{P}_{X_i \overset{\text{iid}}{\sim} Q}\left[\mathrm{diam}(\mathrm{CI}_t^{\mathsf{R}^p}) \leqslant \frac{14p}{p-1}\kappa^{1/p}\varepsilon^{(p-1)/p}\right] \geqslant 1 - \delta. \tag{41}$$

We defer the above proofs to Appendix A. The lower bound from Bhatt et al. (2022a, Theorem 4.4) implies that RCIs for $(\mathcal{M}_\kappa^p, \mu)$ must, with high probability, have widths at least $\frac{1}{8}\kappa^{1/p}\varepsilon^{(p-1)/p}$. Thus, our width is optimal up to constant factors.

We remark that in (36), we set the coefficient of x^p and $|x|^p$ to be $1/p$, as this leads to a succinct, aesthetically pleasing bound (41). One may also follow the tuning technique by Bhatt et al. (2022b) in their fixed-time, non-robust setting, to set the coefficient unknown and then optimize over it. This may lead to constant-level improvement.

5.3 Robust Betting

We finally demonstrate a sibling case of potential interest, the robust extension of the Kelly betting scheme by Waudby-Smith and Ramdas (2023) for bounded data. Let $\mathcal{M}_{[0,1]}$ be the set of all distributions on $[0,1]$; the original and corrupted distribution both belong to this class. The following theorem (proved in Appendix A) establishes a robust supermartingale analogous to the capital process when betting on $\mu(P)$ (cf. Waudby-Smith and Ramdas (2023, Section 4)).

Theorem 6. *Let* $P \in \mathcal{M}_{[0,1]}$ *and* $Q \in \mathbb{B}_{\mathrm{TV}}(P,\varepsilon) \cap \mathcal{M}_{[0,1]}$*. Suppose* $X_1, X_2, \dots \overset{\text{iid}}{\sim} Q$ *and each* X_t *is* $\mathcal{F}_t$*-measurable. The process* $\{L_t\}$ *defined as follows is a supermartingale:*

$$L_t = \prod_{i=1}^t (1 + \lambda_i(X_i - \mu(P)) - \varepsilon|\lambda_i|), \ |\lambda_i| \leqslant \frac{1}{1+\varepsilon}. \tag{42}$$

Unlike (16), the ε-dependent term here (which is like an "insurance cost" against corruptions) scales linearly with λ_i, and does not discount the process if λ_i is 0. When testing $\mu(P) = \mu_0$ (replacing $\mu(P)$ with μ_0 in (42)), the tradeoff between rate of growth and safeguard against falsely rejecting the null due to corrupted observations is smaller if the "bet" λ_i is smaller.

6 CONCLUSION

In this paper, we derive Huber-robust confidence sequences for the class of distributions with bounded p-th central moments, the first Huber-robust CSs we are aware of for any class. These are based on the design of new robust nonnegative supermartingales.

Our CS matches the width lower bound (up to a constant) and it performs even better than robust nonsequential (fixed-time) CIs in the literature. As referenced earlier, our methods will enable immediate robustification of downstream applications of confidence sequences, e.g., multi-armed bandits with contaminated distributions, and A/B/n testing within existing experimentation pipelines in multiple companies in the IT industry.

We hope that followup work can extend our ideas to multidimensional settings, perhaps utilizing the recent reductions presented in Prasad et al. (2020) ("A robust univariate mean estimator is all you need"). This does not appear to be straightforward because the aforementioned paper often hides constants (which do not matter for the rate-optimality results they are interested in, but which do matter for achieving the desired level α), and it operates in a fixed-sample size setting. It appears that many theoretical and practical details need to be worked out for an exact and practical $(1-\alpha)$-RCIs or RCS to be feasible.

Other problems of potential future interest that this paper brings up include the necessity (or lack thereof) of foreknowledge of the parameter ε, as well as a further understanding of the betting scheme and especially its λ_i-proportional robustification cost in Section 5.3. Last, given the well known connections between robust statistics and differential privacy (Dwork and Lei, 2009), the privacy implications of this work may be interesting to pursue. Indeed, large-scale private and robust sequential experimentation is certainly of increasing interest in the IT industry.

References

Analytics for Target. Experimentation panel. URL https://experienceleague.adobe.com/docs/analytics-platform/using/cja-workspace/panels/experimentation.html?lang=en. Accessed: 13 Oct, 2022.

F. J. Anscombe. Rejection of outliers. *Technometrics*, 2(2): 123–146, 1960.

R. R. Bahadur and L. J. Savage. The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics*, 27(4): 1115–1122, 1956.

S. Bhatt, G. Fang, P. Li, and G. Samorodnitsky. Minimax M-estimation under adversarial contamination. In *International Conference on Machine Learning*, pages 1906–1924. PMLR, 2022a.

S. Bhatt, G. Fang, P. Li, and G. Samorodnitsky. Nearly optimal Catoni's M-estimator for infinite variance. In *International Conference on Machine Learning*, pages 1925–1944. PMLR, 2022b.

P. J. Bickel. On some robust estimates of location. *The Annals of Mathematical Statistics*, pages 847–858, 1965.

G. E. Box. Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335, 1953.

O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et Statistiques*, volume 48, pages 1148–1185, 2012.

M. Chen, C. Gao, and Z. Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

Y. Cheng, I. Diakonikolas, and R. Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2755–2771. SIAM, 2019.

D. A. Darling and H. Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences of the United States of America*, 58(1):66, 1967.

J. Depersin and G. Lecué. Robust sub-Gaussian estimation of a mean vector in nearly linear time. *The Annals of Statistics*, 50(1):511–536, 2022.

I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

C. Dwork and J. Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009.

Evidently. How evidently calculates results. URL https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/CloudWatch-Evidently-calculate-results.html. Accessed: 13 Oct, 2022.

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International statistical review*, 70(3): 419–435, 2002.

P. Grünwald, R. de Heide, and W. M. Koolen. Safe testing. *arXiv:1906.07801*, 2019.

F. R. Hampel. *Contributions to the theory of robust estimation*. University of California, Berkeley, 1968.

F. R. Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42(6): 1887–1896, 1971.

S. R. Howard and A. Ramdas. Sequential estimation of quantiles with applications to A/B testing and best-arm identification. *Bernoulli*, 28(3):1704–1728, 2022.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

P. J. Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758, 1965.

P. J. Huber. Robust confidence limits. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 10 (4):269–278, 1968.

P. J. Huber. Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, pages 799–821, 1973.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

R. Johari, L. Pekelis, and D. J. Walsh. Always valid inference: Bringing sequential analysis to A/B testing. *Operations Research*, 70(3):1806–1821, 2021.

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

G. Lecué and M. Lerasle. Robust machine learning by median-of-means: theory and practice. *The Annals of Statistics*, 48(2):906–931, 2020.

M. Lerasle and R. I. Oliveira. Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*, 2011.

J. Li. Total variation, statistical models, and lower bounds, 2019. URL https://jerryzli.github.io/robust-ml-fall19/lec2.pdf. Lecture 2 in the Lecture notes of Robustness in Machine Learning (CSE 599-M) taught at the University of Washington.

M. Lindon, C. Sanden, and V. Shirikian. Rapid regression detection in software deployments through sequential testing. *arXiv preprint arXiv:2205.14762*, 2022.

G. Lugosi and S. Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021.

R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

S. Minsker and M. Ndaoud. Robust and efficient mean estimation: an approach based on the properties of self-normalized sums. *Electronic Journal of Statistics*, 15 (2):6036–6070, 2021.

A. Prasad, S. Balakrishnan, and P. Ravikumar. A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*, 2019.

A. Prasad, S. Balakrishnan, and P. Ravikumar. A robust univariate mean estimator is all you need. In *International Conference on Artificial Intelligence and Statistics*, pages 4034–4044. PMLR, 2020.

P. X. Quang. Robust sequential testing. *The Annals of Statistics*, pages 638–649, 1985.

A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *arXiv preprint arXiv:2210.01948*, 2022a.

A. Ramdas, J. Ruf, M. Larsson, and W. M. Koolen. Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 141:83–109, 2022b.

E. M. Ronchetti and P. J. Huber. *Robust statistics*. John Wiley & Sons, 2009.

J. W. Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

J. Ville. Etude critique de la notion de collectif. *Bull. Amer. Math. Soc*, 45(11):824, 1939.

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

H. Wang and A. Ramdas. Catoni-style confidence sequences for heavy-tailed mean estimation. *arXiv preprint arXiv:2202.01250*, 2022.

H. Wang, M. Gurbuzbalaban, L. Zhu, U. Simsekli, and M. A. Erdogdu. Convergence rates of stochastic gradient descent under infinite noise variance. *Advances in Neural Information Processing Systems*, 34:18866–18877, 2021.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B (Methodological), to appear with discussion*, 2023.

F. Yang, A. Ramdas, K. G. Jamieson, and M. J. Wainwright. A framework for multi-A(rmed)/B(andit) testing with online FDR control. *Advances in Neural Information Processing Systems*, 30, 2017.

A ADDITIONAL THEORETICAL RESULTS AND PROOFS

Lemma 6. *For every $\varepsilon \in (0, 1/2)$ and $\sigma > 0$, there exist $P_1, P_2 \in \mathcal{M}^2_{\sigma^2}$ such that*

1. *$P_2 = (1-\varepsilon)P_1 + \varepsilon N$ for some $N \in \mathcal{M}$, and*
2. *$|\mu(P_1) - \mu(P_2)| = \sigma\sqrt{\varepsilon}$.*

Proof of Lemma 6. We take $P_1 = \delta_0$ and $N = \delta_{\sigma\varepsilon^{-1/2}}$, the Dirac point mass at 0 and $\sigma\varepsilon^{-1/2}$ respectively. The fact that $P_1 \in \mathcal{M}^2_{\sigma^2}$ is obvious. Note that

$$\mu(P_2) = (1-\varepsilon)\mu(P_1) + \varepsilon\mu(N) = \sigma\varepsilon^{1/2}, \tag{43}$$

and

$$v_2(P_2) = (1-\varepsilon)\int (x - \sigma\varepsilon^{1/2})^2 \,\mathrm{d}P_1 + \varepsilon \int (x - \sigma\varepsilon^{1/2})^2 \,\mathrm{d}P_2 = (1-\varepsilon)\sigma^2. \tag{44}$$

So $P_2 \in \mathcal{M}^2_{\sigma^2}$, concluding the proof. $\square$

Proof of Lemma 1. Let P_1 and P_2 be the two distributions in Lemma 6. First applying the definition of RCS with $P_2 \in \mathbb{B}_{\mathrm{TV}}(P_1, \varepsilon)$, we have

$$\mathop{\mathbb{P}}_{X_i \overset{\text{iid}}{\sim} P_2} \left[\forall t \in \mathbb{N}^+,\ \mu(P_1) \in \mathrm{CI}_t(X_1, \dots, X_t)\right] \geqslant 1 - \alpha. \tag{45}$$

Then applying the definition of an RCS with $P_2 \in \mathbb{B}_{\mathrm{TV}}(P_2, \varepsilon)$, we get

$$\mathop{\mathbb{P}}_{X_i \overset{\text{iid}}{\sim} P_2} \left[\forall t \in \mathbb{N}^+,\ \mu(P_2) \in \mathrm{CI}_t(X_1, \dots, X_t)\right] \geqslant 1 - \alpha. \tag{46}$$

We see that, by a union bound,

$$\mathop{\mathbb{P}}_{X_i \overset{\text{iid}}{\sim} P_2} \left[\forall t \in \mathbb{N}^+,\ \{\mu(P_1), \mu(P_2)\} \subseteq \mathrm{CI}_t(X_1, \dots, X_t)\right] \geqslant 1 - 2\alpha. \tag{47}$$

This implies that

$$\mathop{\mathbb{P}}_{X_i \overset{\text{iid}}{\sim} P_2} \left[\forall t \in \mathbb{N}^+,\ \mathrm{diam}(\mathrm{CI}_t(X_1, \dots, X_t)) \geqslant \sigma\sqrt{\varepsilon}\right] \geqslant 1 - 2\alpha, \tag{48}$$

as claimed. The case for RCIs is entirely analogous, by taking "$\forall t \in \mathbb{N}^+$" outside of "$\mathbb{P}[\,\cdot\,]$". $\square$

Lemma 7. *On $x \in [0, 1/7]$, we have*

$$0 \leqslant \underbrace{\frac{x}{4} - 2\left(\frac{1}{1 + x/8 + 1.5x} - (1.5x + 1)\right)}_{A(x)} \leqslant \underbrace{\frac{x}{4} - 2\left(\frac{\exp(-x/4)}{1 + x/8 + 1.5x} - (1.5x + 1)\right)}_{B(x)} \leqslant 7x \leqslant 1. \tag{49}$$

Proof of Lemma 7. The first inequality is straightforward when writing $A(x)$ as $\frac{3.25x(1.625x+2)}{1.625x+1}$. The second inequality is trivial. To prove the third inequality, first note that $B(x) = 3.25x - \frac{2\exp(-x/4)}{1.625x+1} + 2$; then note that

$$2\exp(-x/4) \geqslant 2 - 0.5x \geqslant 2 - 0.5x - 3.75 \times 1.625x^2 = (2 - 3.75x)(1.625x + 1), \tag{50}$$

which clearly implies $B(x) \leqslant 7x$. The fourth inequality is trivial. $\square$

Proof of Theorem 2. Let $X_1, X_2, \dots \overset{\text{iid}}{\sim} Q \in \mathbb{B}_{\mathrm{TV}}(P, \varepsilon)$.

Define

$$M_t(m) = \prod_{i=1}^{t} \frac{\exp\{\phi(\lambda_i(X_i - m))\}}{1 + \lambda_i(\mu(P) - m) + \frac{\lambda_i^2}{2}(\sigma^2 + (\mu(P) - m)^2) + 1.5\varepsilon} \tag{51}$$

$$= \frac{\prod_{i=1}^{t} \exp\{\phi(\lambda(X_i - m))\}}{\left(1 + \lambda(\mu(P) - m) + \frac{\lambda^2}{2}(\sigma^2 + (\mu(P) - m)^2) + 1.5\varepsilon\right)^t}. \tag{52}$$

Note that

$$\mathbb{E}\left[\frac{\exp\{\phi(\lambda_t(X_t - m))\}}{1+\lambda_t(\mu(P)-m)+\frac{\lambda_t^2}{2}\left(\sigma^2+(\mu(P)-m)^2\right)+1.5\varepsilon}\middle|\mathcal{F}_{t-1}\right] \tag{53}$$

$$=\frac{\mathbb{E}_{X_t\sim Q}\left[\exp\{\phi(\lambda_t(X_t - m))\}\right]}{1+\lambda_t(\mu(P)-m)+\frac{\lambda_t^2}{2}\left(\sigma^2+(\mu(P)-m)^2\right)+1.5\varepsilon} \tag{54}$$

$$\leqslant\frac{\mathbb{E}_{X_t\sim P}\left[\exp\{\phi(\lambda_t(X_t - m))\}\right]+1.5\varepsilon}{1+\lambda_t(\mu(P)-m)+\frac{\lambda_t^2}{2}\left(\sigma^2+(\mu(P)-m)^2\right)+1.5\varepsilon} \tag{55}$$

$$\leqslant\frac{\mathbb{E}_{X_t\sim P}\left[1+\lambda_t(X_t-m)+\lambda_t^2(X_t-m)^2/2\right]+1.5\varepsilon}{1+\lambda_t(\mu(P)-m)+\frac{\lambda_t^2}{2}\left(\sigma^2+(\mu(P)-m)^2\right)+1.5\varepsilon} \tag{56}$$

$$=\frac{(1+\lambda_t(\mu(P)-m)+\frac{\lambda_t^2}{2}\left(v_2(P)+(\mu(P)-m)^2)\right)+1.5\varepsilon}{1+\lambda_t(\mu(P)-m)+\frac{\lambda_t^2}{2}\left(\sigma^2+(\mu(P)-m)^2\right)+1.5\varepsilon} \tag{57}$$

$$\leqslant\frac{(1+\lambda_t(\mu(P)-m)+\frac{\lambda_t^2}{2}\left(\sigma^2+(\mu(P)-m)^2)\right)+1.5\varepsilon}{1+\lambda_t(\mu(P)-m)+\frac{\lambda_t^2}{2}\left(\sigma^2+(\mu(P)-m)^2\right)+1.5\varepsilon}=1. \tag{58}$$

Hence, $\{M_t(m)\}$ is a supermartingale under $X_1, X_2, \dots \overset{\text{iid}}{\sim} Q$. We remark that $M_t(\mu(P))$ is just M_t^{RC} and when $\varepsilon = 0$ it is reduced to the non-robust case in our last paper.

Now that $\{M_t(m)\}$ is a supermartingale with $M_0(m) = 1$, we see that $\mathbb{E}M_t(m) \leqslant 1$; that is,

$$\mathbb{E}\exp(f_t(m)) \leqslant \left(1+\lambda(\mu(P)-m)+\frac{\lambda^2}{2}\left(\sigma^2+(\mu(P)-m)^2\right)+1.5\varepsilon\right)^t. \tag{59}$$

Define the function

$$B_t^+(m) = t\log\left(1+\lambda(\mu(P)-m)+\frac{\lambda^2}{2}\left(\sigma^2+(\mu(P)-m)^2\right)+1.5\varepsilon\right)+\log(2/\delta). \tag{60}$$

By Markov's inequality,

$$\forall m\in\mathbb{R},\ \mathbb{P}[f_t(m)\leqslant B_t^+(m)]\geqslant 1-\delta/2. \tag{61}$$

Let $m = \pi_t$ be the smaller solution (whose existence is discussed soon) to the following *quadratic* equation

$$B_t^+(m) = -\log(2/\alpha) - t\log(1+\lambda^2\sigma^2/2+1.5\varepsilon). \tag{62}$$

Then

$$\mathbb{P}[f_t(\pi_t)\leqslant -\log(2/\alpha)-t\log(1+\lambda^2\sigma^2/2+1.5\varepsilon)]\geqslant 1-\delta/2, \tag{63}$$

$$\mathbb{P}[\max(\mathrm{CI}_t^{\mathsf{R}})\leqslant \pi_t]\geqslant 1-\delta/2. \tag{64}$$

Let us see how π_t can exist. π_t is a solution to a *quadratic* equation. Hence it has closed-form expressions. Consider the equation (62),

$$-\lambda(m-\mu(P))+\frac{\lambda^2}{2}\left(\sigma^2+(m-\mu(P))^2)\right)=\left(\frac{4}{\alpha\delta}\right)^{-1/t}\frac{1}{1+\lambda^2\sigma^2/2+1.5\varepsilon}-(1+1.5\varepsilon). \tag{65}$$

It has roots if and only if

$$1-\left(\lambda^2\sigma^2-2\left(\left(\frac{4}{\alpha\delta}\right)^{-1/t}\frac{1}{1+\lambda^2\sigma^2/2+1.5\varepsilon}-(1+1.5\varepsilon)\right)\right)\geqslant 0. \tag{66}$$

Recall that we assume the following:

1. $t > 4\varepsilon^{-1}\log(4/\alpha\delta)$, consequently $\mathrm{e}^{-\varepsilon/4} < \left(\frac{4}{\alpha\delta}\right)^{-1/t} < 1$,

2. $\lambda^2 = \frac{\varepsilon}{4\sigma^2}$,
3. $0 < \varepsilon \leqslant 1/7$.

Hence we can apply Lemma 7 on $\varepsilon \in [0, 1/7]$ to get

$$0 \leqslant \lambda^2\sigma^2 - 2\left(\left(\frac{4}{\alpha\delta}\right)^{-1/t} \frac{1}{1+\lambda^2\sigma^2/2+1.5\varepsilon} - (1+1.5\varepsilon)\right) \leqslant 7\varepsilon \leqslant 1. \tag{67}$$

So the smaller root exists and it satisfies

$$\begin{aligned}
\pi_t &= \mu(P) + \frac{1 - \sqrt{1 - \left(\lambda^2\sigma^2 - 2\left(\left(\frac{4}{\alpha\delta}\right)^{-1/t} \frac{1}{1+\lambda^2\sigma^2/2+1.5\varepsilon} - (1+1.5\varepsilon)\right)\right)}}{\lambda} &(68)\\
&\leqslant \mu(P) + \frac{1 - \left[1 - \left(\lambda^2\sigma^2 - 2\left(\left(\frac{4}{\alpha\delta}\right)^{-1/t} \frac{1}{1+\lambda^2\sigma^2/2+1.5\varepsilon} - (1+1.5\varepsilon)\right)\right)\right]}{\lambda} &(69)\\
&= \mu(P) + \frac{\lambda^2\sigma^2 - 2\left(\left(\frac{4}{\alpha\delta}\right)^{-1/t} \frac{1}{1+\lambda^2\sigma^2/2+1.5\varepsilon} - (1+1.5\varepsilon)\right)}{\lambda} &(70)\\
&\leqslant \mu(P) + \frac{7\varepsilon}{\lambda} &(71)\\
&= \mu(P) + 14\sigma\sqrt{\varepsilon}. &(72)
\end{aligned}$$

We see that

$$\boxed{\mathbb{P}[\max(\mathrm{CI}_t^{\mathsf{R}}) \leqslant \mu(P) + 14\sigma\sqrt{\varepsilon}] \geqslant 1 - \delta/2.} \tag{73}$$

Similarly, $\{N_t(m)\}$ is a supermartingale under $X_1, X_2, \ldots \overset{\text{iid}}{\sim} Q$, where

$$\begin{aligned}
N_t(m) &= \prod_{i=1}^{t} \frac{\exp\{-\phi(\lambda_i(X_i - m))\}}{1 - \lambda_i(\mu(P) - m) + \frac{\lambda_i^2}{2}(\sigma^2 + (\mu(P) - m)^2)) + 1.5\varepsilon} &(74)\\
&= \frac{\prod_{i=1}^{t} \exp\{-\phi(\lambda(X_i - m))\}}{\left(1 - \lambda(\mu(P) - m) + \frac{\lambda^2}{2}(\sigma^2 + (\mu(P) - m)^2)) + 1.5\varepsilon\right)^t}. &(75)
\end{aligned}$$

Define the function

$$B_t^-(m) = -t\log\left(1 - \lambda(\mu(P) - m) + \frac{\lambda^2}{2}(\sigma^2 + (\mu(P) - m)^2)) + 1.5\varepsilon\right) - \log(2/\delta). \tag{76}$$

By Markov's inequality,

$$\forall m \in \mathbb{R},\ \mathbb{P}[f_t(m) \geqslant B_t^-(m)] \geqslant 1 - \delta/2. \tag{77}$$

Let $m = \rho_t$ be the larger solution to the following equation

$$B_t^-(m) = \log(2/\alpha) + t\log\{1 + \lambda^2\sigma^2/2 + 1.5\varepsilon\}. \tag{78}$$

Then

$$\mathbb{P}[\min(\mathrm{CI}_t^{\mathsf{R}}) \geqslant \rho_t] \geqslant 1 - \delta/2. \tag{79}$$

Now, consider the equation (76),

$$-\lambda(\mu(P) - m) + \frac{\lambda^2}{2}(\sigma^2 + (\mu(P) - m)^2) = \left(\frac{4}{\alpha\delta}\right)^{-1/t} \frac{1}{1+\lambda^2\sigma^2/2+1.5\varepsilon} - (1+1.5\varepsilon). \tag{80}$$

The obvious isometry $(m - \mu(P) \leftrightarrow \mu(P) - m)$ between (65) and (80) yields that $\rho_t \geqslant \mu(P) - 16\sigma\sqrt{\varepsilon}$. Therefore, we see that

$$\boxed{\mathbb{P}[\min(\mathrm{CI}_t^{\mathsf{R}}) \geqslant \mu(P) - 14\sigma\sqrt{\varepsilon}] \geqslant 1 - \delta/2.} \tag{81}$$

Combining two boxed inequalities via a union bound, we complete the proof. $\square$

Proof of Lemma 4. For any $P \in \mathcal{P}$ and any $Q \in \mathbb{B}_{\mathrm{TV}}(P,\varepsilon)$, consider $Z_1,\dots,Z_t \overset{\text{iid}}{\sim} Q$. Recall that the TV metric between P and Q always equals the minimum of $\mathbb{P}_{(X,Y)\sim(P,Q)}[X \neq Y]$ over all coupling (X,Y) with marginals P and Q respectively (Gibbs and Su, 2002). Take the coupling (P,Q) such that $\mathbb{P}_{(X,Y)\sim(P,Q)}[X \neq Y] = D_{\mathrm{TV}}(P,Q) \leqslant \varepsilon$. Let $(X_1,Y_1),\dots,(X_t,Y_t) \overset{\text{iid}}{\sim} (P,Q)$. Then,

$$\mathbb{P}_{Z_i \overset{\text{iid}}{\sim} Q}[\chi(P) \in \mathrm{CI}_t(Z_1,\dots,Z_t)] \tag{82}$$

$$= \mathbb{P}_{(X_i,Y_i) \overset{\text{iid}}{\sim} (P,Q)}[\chi(P) \in \mathrm{CI}_t(Y_1,\dots,Y_t)] \tag{83}$$

$$\geqslant \mathbb{P}_{(X_i,Y_i) \overset{\text{iid}}{\sim} (P,Q)}[\exists c_0 \in \mathcal{R}_{2\varepsilon,t}, (Y_1,\dots,Y_t) = c_0(X_1,\dots,X_t) \text{ and } \chi(P) \in \mathrm{CI}_t(c_0(X_1,\dots,X_t))] \tag{84}$$

$$\geqslant \mathbb{P}_{(X_i,Y_i) \overset{\text{iid}}{\sim} (P,Q)}[\exists c_0 \in \mathcal{R}_{2\varepsilon,t}, (Y_1,\dots,Y_t) = c_0(X_1,\dots,X_t)] + \mathbb{P}_{X_i \overset{\text{iid}}{\sim} P}[\forall c \in \mathcal{R}_{2\varepsilon,t},\ \chi(P) \in \mathrm{CI}_t(c(X_1,\dots,X_t))] - 1 \tag{85}$$

(bounding the first term by Hoeffding's inequality) (86)

$$\geqslant (1 - \mathrm{e}^{-2t\varepsilon^2}) + (1-\alpha) - 1 = 1 - (\alpha + \mathrm{e}^{-2t\varepsilon^2}). \tag{87}$$

This completes the proof. □

Proof of Corollary 1. Recall the definition of $\{M_t(m)\}$ back in (51). Note that $\min(\mathrm{CI}_t^{\mathsf{R}})$ is the solution for m of the equation $M_t(m) = 1/\alpha$.

Under the assumptions of Theorem 2, inequality (81) holds. Therefore, if $\mu_0 < \mu(P) - 14\sigma\sqrt{\varepsilon}$, $\mathbb{P}[\min(\mathrm{CI}_t^{\mathsf{R}}) > \mu_0] \geqslant 1 - \delta/2$. Note that $M_t(m)$ is a decreasing function of m. This implies that $\mathbb{P}[M_t(\mu_0) > 1/\alpha] \geqslant 1 - \delta/2$.

Let us use the smallest α that satisfies the assumption $t \geqslant 4\varepsilon^{-1}\log(4/\alpha\delta)$ of Theorem 2. This would give

$$\mathbb{P}_{X_i \sim Q}\left[M_t(\mu_0) > \frac{\delta}{4}\exp\left(\frac{t\varepsilon}{4}\right)\right] \geqslant 1 - \delta/2. \tag{88}$$

The proof is complete as $M_t(\mu_0)$ is just the M_t^{RC} in Corollary 1. □

Proof of Lemma 5 and Theorem 4. Since $|\phi_p(x)| \leqslant \log p$, we have $1/p \leqslant \exp\{\phi(\lambda_t(x - \mu(P)))\} \leqslant p$. Note that

$$\mathbb{E}\left[\frac{\exp\{\phi_p(\lambda_t(X_t - \mu(P)))\}}{1 + \lambda_t^p\kappa/p + (p-1/p)\varepsilon}\middle|\mathcal{F}_{t-1}\right] \tag{89}$$

$$= \frac{\mathbb{E}_{X_t\sim Q}[\exp\{\phi_p(\lambda_t(X_t - \mu(P)))\}]}{1 + \lambda_t^p\kappa/p + (p-1/p)\varepsilon} \tag{90}$$

$$\leqslant \frac{\mathbb{E}_{X_t\sim P}[\exp\{\phi_p(\lambda_t(X_t - \mu(P)))\}] + (p-1/p)\varepsilon}{1 + \lambda_t^p\kappa/p + (p-1/p)\varepsilon} \tag{91}$$

$$\leqslant \frac{\mathbb{E}_{X_t\sim P}[1 + \lambda_t(X_t - \mu(P)) + \lambda_t^p|X_t - \mu(P)|^p/p] + (p-1/p)\varepsilon}{1 + \lambda_t^p\kappa/p + (p-1/p)\varepsilon} \tag{92}$$

$$= \frac{\mathbb{E}_{X_t\sim P}[1 + \lambda_t^p|X_t - \mu(P)|^p/p] + (p-1/p)\varepsilon}{1 + \lambda_t^p\kappa/2 + (p-1/p)\varepsilon} \tag{93}$$

$$= \frac{(1 + \lambda_t^p v_p(P)/p) + (p-1/p)\varepsilon}{1 + \lambda_t^p\kappa/p + (p-1/p)\varepsilon} \leqslant 1. \tag{94}$$

The first inequality above is due to (5) and $Q \in \mathbb{B}_{\mathrm{TV}}(P,\varepsilon)$. Hence $\{M_t^{\mathsf{RC}p}\}$ is a supermartingale. The proof for $\{N_t^{\mathsf{RC}p}\}$ is analogous. Now apply Ville's inequality (Lemma 2) to $\{M_t^{\mathsf{RC}p}\}$. With probability at least $1 - \alpha/2$, we have that $\forall t \in \mathbb{N}^+$,

$$f_{pt}(\mu(P)) \leqslant \sum_{i=1}^t \log\{1 + \lambda_i^p\kappa/p + (p-1/p)\varepsilon\} + \log(2/\alpha). \tag{95}$$

And to $\{N_t^{\mathsf{RC}p}\}$,

$$-f_{pt}(\mu(P)) \leqslant \sum_{i=1}^t \log\{1 + \lambda_i^p\kappa/p + (p-1/p)\varepsilon\} + \log(2/\alpha). \tag{96}$$

A union of the above two bounds concludes the proof. $\square$

Lemma 8. *Let $p \in (1,2]$, $C > 0$, and $g(y) = y^p - y + C$. Suppose there is a $c > 0$ such that $C = \left(\frac{c}{(1+c)^p}\right)^{1/(p-1)}$. Then $g((1+c)C) = 0$.*

The above lemma is checked by direct substitution, so the proof is omitted.

Lemma 9. *Let $p \in (1,2]$, $A, B, C > 0$, and $g(x) = Ax^p - Bx + C$. Suppose there is a $c > 0$ such that $A^{1/(p-1)} B^{-p/(p-1)} C = \left(\frac{c}{(1+c)^p}\right)^{1/(p-1)}$. Then g has a positive zero $(1+c)B^{-1}C$.*

Proof of Lemma 9. Substituting $x = (B/A)^{1/(p-1)} y$, the equation reads

$$y^p - y + A^{1/(p-1)} B^{-p/(p-1)} C = 0. \tag{97}$$

So there is a solution $y_0 = (1+c)A^{1/(p-1)} B^{-p/(p-1)} C$ due to Lemma 8. Correspondingly $x_0 = (B/A)^{1/(p-1)} \cdot (1+c)A^{1/(p-1)} B^{-p/(p-1)} C = (1+c)B^{-1}C$. $\square$

Lemma 10 (One-dimensional special case of Lemma 7 of Wang et al. (2021), Appendix A)**.** *Let $p \in (1,2]$. For any real x and y, $|x+y|^p \leqslant |x|^p + 4|y|^p + py|x|^{p-1}\operatorname{sgn}(x)$.*

Lemma 11. *For any $p \in (1,2]$ and $x > 0$, we have*

$$0 < \underbrace{1 + (p+3/p)x - \frac{1}{1+px}}_{A(p,x)} < 1 + (p+3/p)x - \mathrm{e}^{-x}\frac{1}{1+px} < 7x. \tag{98}$$

Proof of Lemma 11. The first inequality is straightforward when writing $A(p,x)$ as $\frac{x(p^3x+2p^2+3px+3)}{p(px+1)}$. The second inequality is trivial. The third inequality is equivalent to $\mathrm{e}^{-x} > (1+px)(1+(p+3/p-7)x)$. Since $p+3/p-7 < 0$ and $2p+3/p < 6$, we have $(1+px)(1+(p+3/p-7)x) < 1+(2p+3/p-7)x < 1-x < \mathrm{e}^{-x}$. $\square$

Proof of Theorem 5. Define

$$M_t^p(m) = \frac{\prod_{i=1}^t \exp\{\phi_p(\lambda_i(X_i - m))\}}{\prod_{i=1}^t \left(1 + \lambda_i(\mu(P) - m) + \frac{\lambda_i^p}{p}(4\kappa + |\mu(P) - m|^p) + (p - 1/p)\varepsilon\right)}. \tag{99}$$

Then

$$\mathbb{E}\left[\frac{\exp\{\phi_p(\lambda_t(X_t - m))\}}{1 + \lambda_t(\mu(P) - m) + \frac{\lambda_t^p}{p}(4\kappa + |\mu(P) - m|^p) + (p - 1/p)\varepsilon}\middle|\mathcal{F}_{t-1}\right] \tag{100}$$

$$= \frac{\mathbb{E}_{X_t \sim Q}[\exp\{\phi_p(\lambda_t(X_t - m))\}]}{1 + \lambda_t(\mu(P) - m) + \frac{\lambda_t^p}{p}(4\kappa + |\mu(P) - m|^p) + (p - 1/p)\varepsilon} \tag{101}$$

$$\leqslant \frac{\mathbb{E}_{X_t \sim P}[\exp\{\phi_p(\lambda_t(X_t - m))\}] + (p - 1/p)\varepsilon}{1 + \lambda_t(\mu(P) - m) + \frac{\lambda_t^p}{p}(4\kappa + |\mu(P) - m|^p) + (p - 1/p)\varepsilon} \tag{102}$$

$$\leqslant \frac{\mathbb{E}_{X_t \sim P}[1 + \lambda_t(X_t - m) + \lambda_t^p|X_t - m|^p/p] + (p - 1/p)\varepsilon}{1 + \lambda_t(\mu(P) - m) + \frac{\lambda_t^p}{p}(4\kappa + |\mu(P) - m|^p) + (p - 1/p)\varepsilon} \tag{103}$$

$$\text{(by Lemma 10)} \tag{104}$$

$$\leqslant \frac{\mathbb{E}_{X_t \sim P}[1 + \lambda_t(X_t - m) + (\lambda_t^p/p)(|\mu(P) - m|^p + 4|X_t - \mu(P)|^p + p(X_t - \mu(P))|\mu(P) - m|^{p-1}\operatorname{sgn}(\mu(P) - m))] + (p - 1/p)\varepsilon}{1 + \lambda_t(\mu(P) - m) + \frac{\lambda_t^p}{p}(4\kappa + |\mu(P) - m|^p) + (p - 1/p)\varepsilon} \tag{105}$$

$$\leqslant \frac{1 + \lambda_t(\mu(P) - m) + (\lambda_t^p/p)(|\mu(P) - m|^p + 4\kappa) + (p - 1/p)\varepsilon}{1 + \lambda_t(\mu(P) - m) + \frac{\lambda_t^p}{p}(4\kappa + |\mu(P) - m|^p) + (p - 1/p)\varepsilon} = 1. \tag{106}$$

Thus $\{M_t^p(m)\}$ is a nonnegative supermartingale issued at 1.

When $\lambda_1 = \cdots = \lambda$, we see that

$$\mathbb{E}\exp(f_{pt}(m)) \leqslant \left(1 + \lambda(\mu(P) - m) + \frac{\lambda^p}{p}\left(4\kappa + |\mu(P) - m|^p\right) + (p - 1/p)\varepsilon\right)^t. \tag{107}$$

Define the function

$$B_{pt}^+(m) = t\log\left(1 + \lambda(\mu(P) - m) + \frac{\lambda^p}{p}\left(4\kappa + |\mu(P) - m|^p\right) + (p - 1/p)\varepsilon\right) + \log(2/\delta). \tag{108}$$

Markov's inequality yields

$$\forall m \in \mathbb{R}, \mathbb{P}[f_{pt}(m) \leqslant B_{pt}^+(m)] \geqslant 1 - \delta/2. \tag{109}$$

Suppose $m = \pi_t$ is a solution in (μ, ∞) (existence to be discussed soon) to the equation

$$B_{pt}^+(m) = -\log(2/\alpha) - t\log(1 + \lambda^p\kappa/p + (p - 1/p)\varepsilon). \tag{110}$$

Then

$$\mathbb{P}[f_{pt}(\pi_t) \leqslant -\log(2/\alpha) - t\log(1 + \lambda^p\kappa/p + (p - 1/p)\varepsilon)] \geqslant 1 - \delta/2 \tag{111}$$

$$\mathbb{P}[\max(\mathrm{CI}_t^{\mathsf{R}p}) \leqslant \pi_t] \geqslant 1 - \delta/2. \tag{112}$$

Let us see how π_t can exist. The equation that π_t would satisfy, (110), expands into

$$1 - \lambda(m - \mu(P)) + \frac{\lambda^p}{p}\left(4\kappa + |m - \mu(P)|^p\right) + (p - 1/p)\varepsilon = \left(\frac{4}{\alpha\delta}\right)^{-1/t} \frac{1}{1 + \lambda^p\kappa/p + (p - 1/p)\varepsilon}; \tag{113}$$

$$\frac{\lambda^p}{p}|m - \mu(P)|^p - \lambda(m - \mu(P)) + \underbrace{1 + 4\lambda^p\kappa/p + (p - 1/p)\varepsilon - \left(\frac{4}{\alpha\delta}\right)^{-1/t} \frac{1}{1 + \lambda^p\kappa/p + (p - 1/p)\varepsilon}}_{\mathfrak{C}} = 0 \tag{114}$$

So, if we can make sure that there is a $c > 0$ such that

$$\underbrace{(\lambda^p/p)^{1/(p-1)}\lambda^{-p/(p-1)}}_{p^{-1/(p-1)}}\mathfrak{C} = \left(\frac{c}{(1+c)^p}\right)^{1/(p-1)} \tag{115}$$

then according to Lemma 9 such π_t exists and

$$\pi_t = \mu(P) + (1 + c)\lambda^{-1}\mathfrak{C}. \tag{116}$$

To achieve this, recall that we have the following assumptions:

1. $t \geqslant \varepsilon^{-1}\log(4/\alpha\delta)$. So $\mathrm{e}^{-\varepsilon} \leqslant \left(\frac{4}{\alpha\delta}\right)^{-1/t} < 1$.
2. $\lambda^p\kappa = \varepsilon$.
3. $\varepsilon \in (0, \frac{p-1}{7p}]$.

Then, we can bound $\mathfrak{C} = 1 + 4\lambda^p\kappa/p + (p - 1/p)\varepsilon - \left(\frac{4}{\alpha\delta}\right)^{-1/t} \frac{1}{1 + \lambda^p\kappa/p + (p-1/p)\varepsilon}$ by functions of ε:

$$1 + (p + 3/p)\varepsilon - \frac{1}{1 + p\varepsilon} \leqslant \mathfrak{C} \leqslant 1 + (p + 3/p)\varepsilon - \mathrm{e}^{-\varepsilon}\frac{1}{1 + p\varepsilon}. \tag{117}$$

Applying Lemma 11, we have $0 < \mathfrak{C} < 7\varepsilon \leqslant \frac{p-1}{p}$.

By elementary calculus, the function $\frac{x}{(1+x)^p}$ $(x > 0)$ takes its maximum $J = \frac{(p-1)^{p-1}}{p^p}$ at $x_p = \frac{1}{p-1}$. Since $C \leqslant \frac{p-1}{p} = (pJ)^{1/(p-1)}$, there is a $c \in (0, x_p]$ that satisfies $\mathfrak{C} = \left(\frac{pc}{(1+c)^p}\right)^{1/(p-1)}$, which is exactly the condition (115).

Therefore, π_t exists and $\pi_t = \mu(p) + (1+c)\lambda^{-1}\mathfrak{C} \leqslant \mu(P) + 7(1+x_p)\kappa^{1/p}\varepsilon^{(p-1)/p}$. It then follows that

$$\mathbb{P}[\max(\mathrm{CI}_t^{\mathsf{R}p}) \leqslant \mu(P) + 7(1+x_p)\kappa^{1/p}\varepsilon^{(p-1)/p}] \geqslant 1 - \delta/2. \tag{118}$$

The other direction, $\mathbb{P}[\min(\mathrm{CI}_t^{\mathsf{R}p}) \geqslant \mu(P) - 7(1+x_p)\kappa^{1/p}\varepsilon^{(p-1)/p}] \geqslant 1 - \delta/2$, is entirely analogous. So we have

$$\mathbb{P}[\mathrm{diam}(\mathrm{CI}_t^{\mathsf{R}p}) \leqslant 14(1+x_p)\kappa^{1/p}\varepsilon^{(p-1)/p}] \geqslant 1 - \delta. \tag{119}$$

Finally, recall that $x_p = \frac{1}{p-1}$. So the bound (119) is exactly the one stated in the Theorem 5. $\square$

Proof of Theorem 6. First, each $1 + \lambda_i(X_t - \mu(P)) - \varepsilon|\lambda_t|$ is nonnegative since

$$\varepsilon \leqslant \frac{1}{|\lambda_t|} - 1 \leqslant \frac{1}{|\lambda_t|} + X_t - \mu(P) = \frac{1 + \lambda_t(X_t - \mu(P))}{|\lambda_t|}. \tag{120}$$

Further,

$$\mathop{\mathbb{E}}_{X_t\sim Q}[1 + \lambda_t(X_t - \mu(P)) - \varepsilon|\lambda_t|] = 1 + \lambda_t\left(\mathop{\mathbb{E}}_{X_t\sim Q}[X_t] - \mu(P)\right) - \varepsilon|\lambda_t| \tag{121}$$

$$\leqslant 1 + \lambda_t\left(\mathop{\mathbb{E}}_{X_t\sim P}[X_t] + \varepsilon - \mu(P)\right) - \varepsilon|\lambda_t| \tag{122}$$

$$= 1 + \lambda_t\varepsilon - \varepsilon|\lambda_t| \leqslant 1. \tag{123}$$

Therefore $\{L_t\}$ is a nonnegative supermartingale. $\square$