Statistica Sinica Preprint No: SS-2024-0249	
Title	Catoni-type Confidence Sequences under Infinite Variance
Manuscript ID	SS-2024-0249
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0249
Complete List of Authors	Guanhua Fang, Sujay Bhatt, Ping Li and Gennady Samorodnitsky
Corresponding Authors	Guanhua Fang
E-mails	fanggh@fudan.edu.cn
Notice: Accepted author version.	

Statistica Sinica

Catoni-type Confidence Sequences under Infinite Variance

Guanhua Fang[1], Sujay Bhatt[2], Ping Li[3], Gennady Samorodnitsky[4]

[1] *Fudan University* [2] *J.P. Morgan* [3] *VecML Inc.* [4] *Cornell University*

Abstract: In this paper, we provide an extension of confidence sequences for settings where the variance of the data-generating distribution does not exist or is infinite. Confidence sequences furnish confidence intervals that are valid at arbitrary data-dependent stopping times, naturally having a wide range of applications. We first derive the Catoni-style confidence sequences for data distributions having a bounded p^{th} moment, where $p \in (1, 2)$, using Ville's inequality, and strengthen the existing upper bound results. The derived results are shown to be better than confidence sequences obtained using vanilla Dubins-Savage inequality. We next establish a lower bound for the width of the Catoni-style confidence sequences for $p \in (1, 2]$, and establish the statistical limitation of applying Ville's inequality based techniques to Catoni-style confidence sequence estimation. To close this gap, we further establish the tighter confidence sequences using the stitching methods. Our new methodology can be easily applied to risk control and parameter estimation problems.

Key words and phrases: Catoni estimator, Heavy tail, Confidence sequence, Law of iterated logarithm

1. Introduction

Sequential design of experiments is a classical framework in statistical sampling theory, in which the size and the composition of samples are not fixed in advance and are allowed to be functions of the observations themselves (Robbins, 1952). Confidence sequence (CS) is one particular tool in sequential design that facilitates anytime-valid inference (Darling and Robbins, 1967; Jamieson and Jain, 2018; Howard et al., 2021). In particular, a confidence sequence is a sequence of confidence intervals that is valid at data-dependent stopping times.

Formally, let $X_1, X_2, \cdots$ be an independent and identically distributed (i.i.d.) stream drawn from distribution P. The basic object of interest is the unknown mean of this distribution, namely, $\mu = \mathbf{E}_P[X_1]$. A crude way to quantify the uncertainty associated with the mean estimation problem is via *confidence intervals*. Here one constructs a $\sigma(X_1, \cdots, X_t)-$measurable interval $\texttt{CI}_t$ for each $t \in \mathbb{N}^+$ such that $\forall\ t \in \mathbb{N}^+$, the following holds: $\mathbb{P}(\mu \in \texttt{CI}_t) \geq 1-\alpha$, for some coverage probability $1-\alpha \in (0,1)$. However, as argued in Wang and Ramdas (2023), confidence intervals may undercover at stopping times. Also, it is well known that the data-dependent peeking at confidence intervals inflates the Type-1 error (Johari et al., 2017; Howard et al., 2021). This motivates confidence sequences, which provide a universal

quantification over t. For a confidence parameter $\alpha \in (0,1)$ and $t \in \mathbb{N}^+$, the sequence of random intervals $\{\texttt{CI}_t\}$ that satisfy $\mathbb{P}(\forall\, t \in \mathbb{N}^+,\ \mu \in \texttt{CI}_t) \geq 1-\alpha$ are called a $(1-\alpha)-$*confidence sequence*.

Confidence sequences are instrumental in modern application tools like multi-armed bandits (Zhan et al., 2021), A/B testing (Howard et al., 2021), causal inference (Malek and Chiappa, 2021), etc. Given the importance, it is not surprising that a significant research effort has been allocated to construct confidence sequences under various distributional assumptions on the data. Darling and Robbins (1967); Jennison and Turnbull (1989) consider P as normal distributed and construct confidence sequences, while Johari et al. (2015, 2017) consider P belonging to an exponential family. Waudby-Smith and Ramdas (2024) consider arbitrary but bounded P, while Howard et al. (2021) consider P having a bounded moment generating function.

Recently, Wang and Ramdas (2023) make a substantial contribution to the literature on confidence sequences by relaxing the distributional assumptions to requiring only the existence of a bounded second moment. This is made possible by using a robust mean estimator developed in Catoni (2012), which uses influence functions to stabilize the effect of the outliers. While the previous works (Howard et al., 2021) required a Chernoff-type assumption on the distribution resulting in $O(\sqrt{\log t/t})$ shrinkage rates

for the confidence sequences, Waudby-Smith and Ramdas (2024) construct an LIL type confidence sequence for bounded means using stitching arguments, and Wang and Ramdas (2023) show that employing Catoni's estimator improves the rate to $O(\sqrt{\log\log 2t/t})$ under weaker assumptions on the distribution. The significance of this result is that there is no excess compromise in considering confidence sequences or weakening the distributional assumptions. A careful reading of Wang and Ramdas (2023) shows however that there are inefficiencies compared to the lower bound that can be improved. While Wang and Ramdas (2023) proposes an extension to deal with the relaxed case of $p \in (1, 2]$ using the ideas for $p = 2$ and recent results on M-estimation in Chen et al. (2021); Bhatt et al. (2022a,b), the inefficiencies in both methods contribute towards only a näive extension.

1.1 Main Results

The key improvements over the existing results are clarified as follows:

- We sharpen the analysis for the case of $p \in (1, 2)$, thereby relaxing the distributional assumptions on the data. This allows sequential design for a larger class of distributions while achieving good control over the width of the CS.

- We derive lower bounds for the width of Catoni-style confidence se-

quences (CS) for both finite and infinite variance cases in Sec. 5. This shows that Wang and Ramdas (2023) almost - *but not quite* - matches the lower bound, leaving room for improvement.

- We improve on the näive CS derived using generalized Dubins-Savage inequality in terms of the growth of the width of the confidence interval (CI) w.r.t. the confidence parameter α in Sec. 3. We improve on the Catoni-style CS by using Ville's inequality to obtain a smaller CI width in Sec. 4.

- We finally apply stitching methods to achieve even tighter confidence sequences to close the gap with the lower bound in Sec. 6. We further derive the lower bounds for the stitching methods thereby providing the LIL type results for this new class of stitching methods. A summary of results is described in Table 1.

- We give two application scenarios to show the usefulness of the proposed methodology. In particular, our method can provide guarantees for any-time valid risk control and any-time valid confidence set construction for heavy-tailed data.

Notation. In this paper, we use $\mathbb{E}$ and $\mathbb{P}$ to denote the generic expectation and probability. We say $a_n = O(b_n)$(or $\Theta(b_n)$) if there exists a

Method	Width
Dubins-Savage	$O(\frac{\log t}{t^{\frac{p-1}{p}}}\alpha^{-1/p})$
Catoni + Ville's ineq	$O(\frac{\log t}{t^{\frac{p-1}{p}}}(\log \frac{1}{\alpha})^{\frac{p-1}{p}})$
Stitch + Catoni + Ville's ineq	$O((\frac{\log\log t+\log(1/\alpha)}{t})^{\frac{p-1}{p}})$

Table 1: The asymptotic width of confidence sequences via different techniques.

constant c that $a_n \leq cb_n$ (or $\frac{1}{c}a_n \leq b_n \leq ca_n$). Symbol $\tilde{O}$ hides all logarithmic terms. We also write $a_n \gtrsim b_n$ if $\liminf_n a_n/b_n > 0$. We use $X[i]$ to represent the i^{th} element of vector X and $X[i,j]$ to represent the element of matrix X on i^{th} row and j^{th} column.

1.2 Deriving Confidence Sequences

As discussed in Howard et.al [2021] and the references therein, there are different ways of constructing confidence sequences. We discuss a few of them here highlighting the challenges. (i) One could construct confidence sequences by inverting a suitably formulated sequential probability test (Sec. 3.6, Howard et.al [2020]). This typically uses a sub-Gaussianity assumption and is not always guaranteed to shrink towards zero width as $n \to \infty$. Two common approaches to address the last issue is to either replace the like-

lihood ratio with a mixture forming a martingale or choose a sequence of point alternatives approaching null while their corresponding error probabilities go to 0, so that a union bound yields the desired confidence bound. (ii) One could use a method of self-normalized bounds (Sec. 4.4 Howard et.al [2021], Sec.5 Wang & Ramdas [2023]), which bound a mixture sequence using intrinsic time. (iii) One could use conjugate mixtures for certain family of distributions, while these obtain worse rates, they are useful in certain practical applications (Sec.3.5 Howard et.al [2021]). (iv) One could use inverted stitching method which gives numerical upper bounds on the crossing probability of any increasing, strictly concave boundary over a limited horizon. (v) One could also use the stitching method that uses carefully designed scaling parameters over geometrically spaced intervals to design confidence sequences, so that a union bound yields the desired confidence bound.

Most of these methods design boundaries avoiding unnecessary approximations and are unimprovable in the sub-Gaussian case. In this paper, we extend some of the above techniques to work in the infinite variance case (improving in the finite variance case over Wang & Ramdas [2023]) using a supermartingale construction of the Catoni estimator, and provide the confidence sequences using traditional Ville's inequality and an improved

one using the stitching method.

2. Problem Formulation

In this section, we first introduce the notation and describe the main problem considered in this paper. Recall that $X_1, X_2, \cdots$ are a sequence of independent and identically distributed random variables with mean $\mathbb{E}[X_1] = \mu$ and the p^{th} moment $\mathbb{E}|X_1 - \mu|^p \leq \upsilon_p$ for $p \in (1, 2]$. A level $1 - \alpha \in (0, 1)$ confidence sequence for μ is a sequence of real numbers $L_n(X_1, \cdots, X_n)$ and $U_n(X_1, \cdots, X_n)$, where $L_n, U_n : \mathbb{R}^n \to \mathbb{R}$ with $n = 1, 2, \cdots$ such that $L_n \leq U_n$ point-wise and

$$\mathbb{P}\Big(L_n(X_1, ..., X_n) \leq \mu \leq U_n(X_1, ..., X_n), \forall n \geq 1\Big) \geq 1 - \alpha.$$

Let the width of the n^{th} confidence interval be

$$W_n(X_1, \cdots, X_n) := U_n(X_1, \cdots, X_n) - L_n(X_1, \cdots, X_n).$$

Objective. We are interested in how fast this width W_n can shrink as n increases, when the $p^{th}-$moment of the data distribution P is bounded with $p \in (1, 2]$.

To study the general form of upper and lower confidence thresholds – $U_n(X_1, \ldots, X_n)$ and $L_n(X_1, \ldots, X_n)$ – has its own challenges (see Sec.1.2), and we only consider the following specific type of construction for confi-

dence sequence, following Wang and Ramdas (2023). Let $\psi(x) : \mathbb{R} \to \mathbb{R}$ be an arbitrary non-decreasing continuously differentiable function. Let (λ_i) be a sequence of scaling parameters. Then we can construct a confidence sequence as

$$\begin{aligned} L_n(X_1, \cdots, X_n) &:= \text{solution to } \sum_{i=1}^{n} \psi(\lambda_i(X_i - x)) = b_n, \\ U_n(X_1, \cdots, X_n) &:= \text{solution to } \sum_{i=1}^{n} \psi(\lambda_i(X_i - x)) = a_n. \end{aligned} \tag{2.1}$$

Here (b_n) and (a_n) are real-number sequences such that $a_n \leq b_n$. Under mild conditions on function ψ, we can show that the width $W_n(X_1, \ldots, X_n) \propto \frac{b_n - a_n}{\sum_{i=1}^{n} \lambda_i}$. Therefore, our main task it make sequence $\{(b_n - a_n)/\sum_{i=1}^{n} \lambda_i\}$ is as tight as possible.

Remark 1. Note that we consider i.i.d. data as it has much wider applicability and has simpler notation, however, the results carry over for stochastic processes with constant conditional expectation using the influence function and the standard supermartingale arguments.

3. Confidence Sequence for Infinite Variance via Dubins-Savage

In this section, we take the first steps to derive confidence sequences for the infinite variance case. A generalization of the classical Dubins-Savage inequality Dubins and Savage (1965) first appeared in Kallenberg (1975),

and most recently in Khan (2009). Lemma 1 is a restatement and can be derived using an approach of Doob Khan (2009).

Lemma 1. *Let $\{S_t\}$ be a martingale with $V_t = \mathbb{E}[|S_t - S_{t-1}|^p|\mathcal{F}_{t-1}]$. Then for all $a \geq 0,\ b > 0$, we have*

$$\mathbb{P}\Big(S_t \geq a + b\sum_{i=1}^{t} V_i\Big) \leq \frac{1}{(1 + m_p a b^{\frac{1}{p-1}})^{p-1}},$$

where $m_p = \left(\frac{p-1}{2^{2-p}}\right)^{\frac{1}{p-1}}$.

Let $\{X_t\}$ be any real-valued stochastic process adapted to the filtration $\{\mathcal{F}_t\}$, where $\mathcal{F}_0$ is the trivial sigma-algebra. To make use of Lemma 1, we can take $S_t = \sum_{i=1}^{t} \psi(\lambda_i(X_i - \mu)) = \sum_{i=1}^{t} \lambda_i(X_i - \mu)$ (i.e. $\psi(x) \equiv x$ is an identity function). By taking $a_n = -a - b\sum_{i=1}^{t} V_i, b_n = a + b\sum_{i=1}^{t} V_i$ and applying Lemma 1, the corresponding confidence sequence will have the following properties. We borrow the width optimization ideas from Howard et al. (2021); Wang and Ramdas (2023) and make use of the L_p version of the classical Dubins-Savage.

Theorem 1 (Dubins-Savage)**.** *Let $a = \frac{1}{m_p b^{\frac{1}{p-1}}} \cdot \left(\left(\frac{2}{\alpha}\right)^{\frac{1}{p-1}} - 1\right)$. The width of the confidence interval using Dubins-Savage inequality is given as*

$$W_n(X_1, \cdots, X_n) = \frac{2a + 2b\upsilon_p \sum_{i=1}^{n} \lambda_i^p}{\sum_{i=1}^{n} \lambda_i},$$

where the coefficients λ_i are chosen as $\lambda_i = \left(\frac{a}{ib\upsilon_p(p-1)}\right)^{1/p}$ for $1 \leq i \leq n$.

From Theorem 1, we can obtain $W_n(X_1, \ldots, X_n) = O(\frac{\log t}{t^{\frac{p-1}{p}}} \alpha^{-1/p})$. Therefore, the width of the confidence sequence shrinks as $t \to \infty$. When $p = 2$, we obtain $\widetilde{O}(t^{-1/2})$ which is known to be unimprovable Howard et al. (2021). However, on the other hand, the dependence on α is $O(\alpha^{-1/p})$, which may lead to a very wide confidence interval when α is very small. In next section, we focus on the improvement in terms of α.

4. Catoni-style Confidence Sequence for $p \in (1, 2]$

Recently, Wang and Ramdas (2023) discuss one possible way to construct a Catoni-style confidence sequence which is as follows. Suppose the observations X_i have a finite second moment and $\texttt{Var}(X_1) \leq \sigma^2$ for a known σ^2. Consider an increasing Catoni-type influence function ψ such that for $x \in \mathbb{R}$ as in Catoni (2012),

$$\psi(x) = \begin{cases} -\log(1 - x + x^2/2), & x < 0 \\ \log(1 + x + x^2/2), & x \geq 0. \end{cases} \tag{4.2}$$

Theorems 9 & 10 in Wang and Ramdas (2023) show that, for a specific choice of the sequences (a_n) and (b_n), (2.1) together with (4.2) defines a Catoni-style confidence sequence for μ. Furthermore, under certain condi-

tions on the scaling parameters (λ_i), for all n large enough,

$$W_n(X_1, \cdots, X_n) \leq 4\frac{\sigma^2 \sum_{i=1}^n \lambda_i^2 + G(\alpha, \varepsilon)}{\sum_{i=1}^n \lambda_i} \tag{4.3}$$

for some $n-$independent constant $G(\alpha, \varepsilon)$ and $\varepsilon \in (0, 1)$.

4.1 Improvement over Wang and Ramdas (2023)

Note that the constant $G(\alpha, \varepsilon)$ in (4.3) depends on an additional parameter $\varepsilon \in (0, 1)$ (which is unnecessary), with the width holding with probability $1 - \alpha$. We **sharpen** the results in Wang and Ramdas (2023), even in the case of finite variance, where the constant $G(\alpha, \varepsilon)$ is now *only* a function of α. Using similar line of arguments, in addition to addressing the challenge of roots characterization of polynomials of degree smaller than 2 in the infinite variance case, leads to the following width of the confidence sequence

$$\mathbb{P}\Big(W_n(X1, \cdots, X_n) \leq 2(1 + \tau_n)\frac{C_p v_p \sum \lambda_i^p (1 + t_i^{-(p-1)}) + \log(2/\varepsilon_n) + \log(2/\alpha)}{\sum \lambda_i}\Big)$$
$$\geq 1 - \sum \varepsilon_n.$$

Now, by choosing $\varepsilon_n = \alpha \exp\{-C_p v_p \sum \lambda_i^p (1 + t_i^{-(p-1)})\}$, the width of the sequence results as

$$\mathbb{P}\Big(W_n(X1, \cdots, X_n) \leq 4(1+\tau_n)\frac{C_p v_p \sum \lambda_i^p (1 + t_i^{-(p-1)}) + \log(2/\alpha)}{\sum \lambda_i}\Big) \geq 1 - \alpha \sum \exp\{-C_p v_p \sum \lambda_i^p (1 + t_i^{-(p-1)})\},$$

where the sum on the right is finite owing to the assumption

$$\sum_i \lambda_i^p = \infty \quad \text{and} \quad \lim_{n\to\infty} \lambda_n = 0.$$

The right hand term can hence be made small by choosing a small α. For the special case $p = 2$, $t_i \to 1$ and $\tau_i \to 0$, we have that

$$\mathbb{P}\Big(W_n(X1, \cdots, X_n) \leq 4\sigma^2 \frac{\sum \lambda_i^p + \log(2/\alpha)}{\sum \lambda_i}\Big) \geq 1 - \alpha \sum_{n=1}^{\infty} \exp\{-\sigma^2 \sum_i^n \lambda_i^p\},$$

resulting in a more tighter width compared to Wang and Ramdas (2023).

4.2 Confidence Sequences for Infinite Variance

Following Bhatt et al. (2022b), for general $1 < p < 2$, with an appropriate $C_p > 0$, we define ψ to be a non-decreasing function $\mathbb{R} \to \mathbb{R}$ such that

$$-\log(1 - x + C_p|x|^p) \leq \psi(x) \leq \log(1 + x + C_p|x|^p) \tag{4.4}$$

for all $x \in \mathbb{R}$. One way to chose $C_p = \left(\frac{p-1}{p}\right)^{p/2} \left(\frac{2-p}{p-1}\right)^{(2-p)/2}$, whence $C_2 = 1/2$ as in Catoni (2012). Let (λ_n) be a sequence of positive numbers such

that

$$\lim_{n\to\infty} \lambda_n = 0, \quad \sum_{n=1}^{\infty} \lambda_n^p = \infty. \tag{4.5}$$

It follows immediately that the processes $M_n^+ = \prod_{i=1}^n \exp\{\psi\big(\lambda_i(X_i-\mu)\big) - C_p v_p \lambda_i^p\}$ and $M_n^- = \prod_{i=1}^n \exp\{-\psi\big(\lambda_i(X_i-\mu)\big) - C_p v_p \lambda_i^p\}$ are non-negative supermartingales with respect to the natural filtration of the sequence $X_1, X_2, \ldots$. Let $0 < \alpha < 1$ be a confidence level. By the **Ville's inequality** Ville (1939) for non-negative supermartingales, the following sequence of sets forms a $(1-\alpha)$-confidence sequence for μ:

$$I_n(\alpha) = \Big\{ x \in \mathbb{R} : -\log\frac{2}{\alpha} - C_p v_p \sum_{i=1}^n \lambda_i^p \le \sum_{i=1}^n \psi\big(\lambda_i(X_i - x)\big) \le \log\frac{2}{\alpha} + C_p v_p \sum_{i=1}^n \lambda_i^p \Big\}, \tag{4.6}$$

$n = 1, 2, \ldots$. In this case, we can simply take $a_n = -\log\frac{2}{\alpha} - C_p v_p \sum_{i=1}^n \lambda_i^p$ and $b_n = \log\frac{2}{\alpha} + C_p v_p \sum_{i=1}^n \lambda_i^p$. Note that by (4.5), the sum $\sum_{i=1}^n \lambda_i^p$ grows slower than linearly fast with n. Therefore, at least for large n, the equations

$$\sum_{i=1}^n \psi\big(\lambda_i(X_i - x)\big) = \pm\left(\log\frac{2}{\alpha} + C_p v_p \sum_{i=1}^n \lambda_i^p\right) \tag{4.7}$$

have unique real roots, which we will denote by $x_{+,n} (:= L_n(X_1, \ldots, X_n))$ and $x_{-,n} (:= U_n(X_1, \ldots, X_n))$ respectively, in which case the set $I_n(\alpha)$ is an interval of the finite length $\big|I_n(\alpha)\big| = x_{-,n} - x_{+,n}$. The next result characterizes how fast these lengths grow as n increases.

Theorem 2. *Suppose the sequence* (λ_n) *is non-random,* $0 < t_n < 1$ *and* $\tau_n > 0$. *Suppose* $\varepsilon_n = \alpha \exp\left\{-C_p v_p \sum_{i=1}^n \lambda_i^p (1+t_i^{-(p-1)})\right\}$ *for* $n = 1, 2, \cdots$. *Consider the condition*

$$C_p v_p \sum_{i=1}^n \lambda_i^p (1 + t_i^{-(p-1)}) + \log 2/\alpha + \log 2/\varepsilon_n \leq \frac{\tau_n^{1/(p-1)}}{(1+\tau_n)^{p/(p-1)}} \frac{(\sum_{i=1}^n \lambda_i)^{p/(p-1)}}{(C_p \sum_{i=1}^n \lambda_i^p (1-t_i)^{-(p-1)})^{1/(p-1)}}. \tag{4.8}$$

The following holds for the width of the confidence sequence

$$P\Bigg(\big|I_n(\alpha)\big| \leq 4(1+\tau_n)\frac{C_p v_p \sum_{i=1}^n \lambda_i^p \big(1 + t_i^{-(p-1)}\big)}{\sum_{i=1}^n \lambda_i} + \frac{\log 2/\alpha}{\sum_{i=1}^n \lambda_i},\ \forall n \text{ such that (4.8) holds}\Bigg) \geq 1 - \alpha \sum_{i=1}^\infty \varepsilon_i. \tag{4.9}$$

Note that with ε_n chosen as in the statement, the condition (4.8) holds for all large n, at least if (t_n) are bounded away from 0, and (τ_n) are not too small (we could, in fact, keep τ_n a small positive constant). The right-hand side can be made small if α is small.

Note that $W_n(X_1, \cdots, X_n) := \big|I_n(\alpha)\big|$, so comparing with (4.3) for $p = 2$, we note the following differences: (i) The width in (4.3) depends on both α and another confidence parameter δ implying a compromise in the width, while the width in Theorem 2 depends only on the confidence parameter α. (ii) Abusing the notation, let $W_{n,\lambda} = W_n(X_1, \cdots, X_n) \cdot \sum_{i=1}^n \lambda_i$. For special case, $p = 2$, $t_i \to 1$, and $\tau_i \to 0$, we find $W_{n,\lambda} \leq 4\sigma^2 \sum_{i=1}^n \lambda_i^2 +$

$\log(2/\alpha)$ using Theorem 2. The coefficient $4\sigma^2$ matches that in inequality (4.3). This confirms the tightness of our theoretical analysis.

In particular, we choose $\lambda_t = \Theta((\log(1/\alpha)/t)^{1/p})$ from Catoni (2012); Chen et al. (2021) implying that the Catoni-style confidence sequence enjoys $O(\frac{\log t(\log(1/\alpha))^{1-1/p}}{t^{\frac{p-1}{p}}})$ shrinkage rate. The dependence on α is $O(\log(1/\alpha)^{1-1/p})$ improving over Theorem 1, as the width increases slowly in case of Catoni-style sequence as $\alpha \downarrow 0$.

5. Lower Bounds

In this section, we establish the lower bounds for the width of Catoni-style confidence sequences. Our results indicates there is a gap between the lower bounds and upper bounds of width W_n. In other words, there exists a statistical limitation in directly applying Ville's inequality technique to Catoni-style estimation method.

5.1 Finite Variance

We are interested in establishing a lower bound nearly matching (4.3). We will allow any choice of the sequences (a_n) and (b_n) and a larger flexibility in the choice of the scaling sequences (λ_n).

Theorem 3. *Let $\{X_i\}$ be the sequence of i.i.d random variables with finite*

variance σ^2. *Assume that one of the following conditions holds.*

S1. There exists $0 < \vartheta \leq 1$ *such that* $\mathbb{E}|X_1|^{2+\vartheta} < \infty$. *Let the scale sequence* (λ_i) *be such that* $\lambda_i \downarrow 0$ *and* $\sum_{i=1}^{\infty} \lambda_i^2 = \infty$.

S2. The scale sequence (λ_i) *satisfies that* $\lambda_i \downarrow 0$, $\sum_{i=1}^{\infty} \lambda_i^2 = \infty$ *and* $\sum_n (\sum_{i=1}^n \lambda_i^2)^{-1-\vartheta/2} \lambda_n^2 < \infty$

For any $1-$*Lipschitz Catoni-style influence function* ψ, *the width of confidence sequence* $W_n(X_1, \cdots, X_n)$ *holds that*

$$\mathbb{P}(W_n(X_1, \cdots, X_n) \geq \frac{a\Big(\sum_{i=1}^n \lambda_i^2 \log\log \sum_{i=1}^n \lambda_i^2\Big)^{1/2}}{\sum_{i=1}^n \lambda_i} \text{ infinitely often}) = 1$$

for any $a < 2\sigma\sqrt{2}$.

The result provides the minimum width of the confidence sequence in the finite variance case. Below, we sketch the broad ideas employed in establishing the result. Notice that since the influence function ψ is $1-$Lipschitz, we have $\left|\frac{d}{dx}\sum_{i=1}^n \psi(\lambda_i(X_i - x))\right| \leq \sum_{i=1}^n \lambda_i$. Therefore, we have the following result, whose proof is given in the appendix.

Proposition 1. *The confidence interval defined in* (2.1) *satisfies*

$$W_n(X_1, \cdots, X_n) \geq \frac{b_n - a_n}{\sum_{i=1}^n \lambda_i}.$$

Next, if (2.1) defines the confidence sequence for μ we need to show that $b_n - a_n$ cannot be too small. Indeed, from (2.1) we have

$$\mathbb{P}\Big(a_n \le \sum_{i=1}^{n} \psi(\lambda_i(X_i - m)) \le b_n \text{ for all } n \ge 1\Big) \ge 1 - \alpha. \tag{5.10}$$

Consider the transformation $Y_i = \psi(\lambda_i(X_i - m))$, whence $Y_1, Y_2, \cdots$ is a sequence of independent random variables with all finite moments. For $n \ge 1$, let $\widetilde{a}_n = a_n - \sum_{i=1}^{n} \mathbb{E}Y_i$, $\widetilde{b}_n = b_n - \sum_{i=1}^{n} \mathbb{E}Y_i$. Then (5.10) implies that

$$\mathbb{P}\Big(\widetilde{a}_n \le \sum_{i=1}^{n} (Y_i - \mathbb{E}Y_i) \le \widetilde{b}_n \text{ for all } n \ge 1\Big) \ge 1 - \alpha. \tag{5.11}$$

Clearly, $\widetilde{b}_n - \widetilde{a}_n = b_n - a_n$, and we will show that these differences cannot be too small using (5.10) and the general law of iterated logarithm in Wittmann (1985).

Here we make some additional comments on conditions *S1* and *S2* in Theorem 3. *S1* assumes an existence of higher-order $(2 + \vartheta)$ moment of X_i's, while *S2* does not. Then *S2* requires one more restriction on the choice of scaling parameter λ_i's. In the estimation problem Catoni (2012), the optimal λ_i is $\Theta(i^{-1/2})$. Then requirements $\lambda_i \downarrow 0$, $\sum_i \lambda_i^2 = \infty$ and $\sum_n (\sum_{i=1}^{n} \lambda_i^2)^{-1-\vartheta/2} \lambda_n^2 < \infty$ are met automatically. Unlike in the cases of dealing with sub-Gaussian random variables X_i's which does not require using Catoni-style functions, the requirements $\lambda_i \downarrow 0$ here cannot be removed. Otherwise, it will lead to a biased estimation of μ and make the confidence

sequence no longer useful.

5.2 Infinite Variance

For $1 < p < 2$, we establish a similar lower bound for confidence width. Although the analysis is more thorny, the technical idea is similar to that in the finite variance case.

Theorem 4. *Let symmetric random variable X_1 satisfy that $\mathbb{E}|X_1|^p < \infty$ for some $1 < p < 2$. Let influence function ψ be $sgn(x)\log(1+|x|+C_p|x|^p)$. Then the width of confidence sequence $W_n(X_1, \cdots, X_n)$ is such that*

$$\mathbb{P}(W_n(X_1, \cdots, X_n) \gtrsim \frac{\left(\sum_{i=1}^n \lambda_i^{p+\vartheta'} \log\log \sum_{i=1}^n \lambda_i^{p+\vartheta'}\right)^{1/2}}{\sum_{i=1}^n \lambda_i} \ i.o.) = 1$$

for any $\vartheta' > 0$ as long as the scale sequence $\{\lambda_i\}$ satisfies that $\lambda_i \downarrow 0$, $\sum_{i=1}^\infty \lambda_i^{p+\vartheta'} = \infty$, $\sum_{n=1}^\infty \left(\sum_{i=1}^n \lambda_i^{p+\vartheta'}\right)^{-1-\vartheta/2} \lambda_n^p < \infty$ for some $\vartheta > 0$.

By comparing Theorems 3 and 4 to Theorem 2, the lower bound is of order $\tilde{\Theta}(\frac{\sqrt{\sum_{i=1}^n \lambda_i^p}}{\sum_{i=1}^n \lambda_i})$ while the upper bounds of order $O(\frac{\sum_{i=1}^n \lambda_i^p}{\sum_i \lambda_i})$. Therefore, there is a gap of order $O(\sqrt{\sum_{i=1}^n \lambda_i^p})$ between upper and lower bounds. This gap cannot be ignored when $\sum_{i=1}^n \lambda_i^p \to \infty$ as $n \to \infty$. The reason behind the gap is that the Ville's inequality is a very general technique that can be applied to any non-negative supermartingale. One way for improvement is to apply Ville's inequality multiple times and get the final confidence

sequence via the stitching method, whose details are described in the next section.

6. Improvement By Stitching Techniques

As explained in the previous section, it is clear that the lower bounds are sharper than the width obtained (see (4.3)) in Wang and Ramdas (2023) or our Theorem 2. We can make the confidence sequence even tighter by considering the so-called **stitching** method (Howard et al., 2021). The very high-level idea can be described as follows. In all previous sections, we keep using the same λ_i for computing $L_n(X_1, \ldots, X_n)$ and $U_n(X_1, \ldots, X_n)$ for any time $n \geq i$. An alternative way is to choose λ_i in a triangular format, that is,

$$\begin{aligned} L_n(X_1, \cdots, X_n) &:= \text{solution of } \sum_{i=1}^{n} \psi(\lambda_i^{(n)}(X_i - x)) = b_n, \\ U_n(X_1, \cdots, X_n) &:= \text{solution of } \sum_{i=1}^{n} \psi(\lambda_i^{(n)}(X_i - x)) = a_n. \end{aligned} \tag{6.12}$$

In other words, the choice of i-th scaling parameter $\lambda_i^{(n)}$ also depends on the time index n. In the following of section, we provide a choice of $\lambda_i^{(n)}$ which leads to a sharper result.

In particular, we consider the following time grids,

$$t_j = a^j. \tag{6.13}$$

For any time t in j-th time interval ($t_{j-1} < t \leq t_j$, $t_0 = 0$ by default), we choose $\lambda_i^{(t)} \equiv \Lambda_j$ for $1 \leq i \leq t$, where

$$\begin{aligned} \Lambda_j &:= \big(\frac{1}{v_p} \cdot \log(\frac{2}{\alpha_j}) a^{-j}\big)^{1/p}; \\ \text{with } \alpha_j &:= \frac{\alpha j^{-q}}{\sum_{l=1}^{\infty} l^{-q}}. \end{aligned} \tag{6.14}$$

Here both tuning parameters a and q are greater than 1. Therefore, we can apply Ville's inequality to each time interval $(t_{j-1}, t_j]$ and have that

$$\begin{aligned} &\mathbb{P}\big(-\log(\frac{2}{\alpha_{\lceil \log_a t \rceil}}) - C_p v_p \sum_{i=1}^{t} \lambda_i^{(t)p} \leq \sum_{i=1}^{t} \psi(\lambda_i^{(t)}(X_i - \mu)) \\ &\leq C_p v_p \sum_{i=1}^{t} \lambda_i^{(t)p} + \log(\frac{2}{\alpha_{\lceil \log_a t \rceil}}); \forall t \in (t_{j-1}, t_j]\big) \geq 1 - \alpha_j. \end{aligned} \tag{6.15}$$

By stitching all time intervals together, it holds that

$$\begin{aligned} &\mathbb{P}\big(-\log(\frac{2}{\alpha_{\lceil \log_a t \rceil}}) - C_p v_p \sum_{i=1}^{t} \lambda_i^{(t)p} \leq \sum_{i=1}^{t} \psi(\lambda_i^{(t)}(X_i - \mu)) \\ &\leq C_p v_p \sum_{i=1}^{t} \lambda_i^{(t)p} + \log(\frac{2}{\alpha_{\lceil \log_a t \rceil}}); \forall t \geq 1)\big) \\ \geq\ & 1 - \sum_j \alpha_j = 1 - \alpha. \end{aligned} \tag{6.16}$$

With some calculations, we will arrive at the following theorem.

Theorem 5. *With probability $1 - \alpha - \alpha'$, it holds*

$$|I_t| \leq 2 v_p^{1/p} \frac{\big((C_p + 1) q \log(2(\log_a t + 1) \sum_{l=1}^{\infty} l^{-q}/\alpha)\big)^{1-1/p}}{c_{t,\alpha'} \cdot t(1/at)^{1/p}} \tag{6.17}$$

uniformly for any $t \geq 1$, where $0 < \alpha' < 1$ is any fixed constant and $c_{t,\alpha'}$'s is a sequence of positive constants such that $c_{t,\alpha'} \to 1$ as $t \to \infty$.

To be self-complete, we also provide the lower bound results of the width for the stitching method.

Theorem 6. *Suppose the scaling parameters $\lambda_i^{(n)}$'s $(i = 1, ..., n)$ have the same order of $\lambda^{(n)} = n^{-1/q}$ with $(q > p)$. Let influence function ψ be $sgn(x)\log(1 + |x| + C_p|x|^p)$.*

Then the width of confidence sequence $W_n(X_1, \cdots, X_n)$ is such that

$$\mathbb{P}(W_n(X_1, \cdots, X_n) \gtrsim \frac{\left(\sum_{i=1}^n (\lambda_i^{(n)})^{p+\vartheta'} \log\log \sum_{i=1}^n (\lambda_i^{(n)})^{p+\vartheta'}\right)^{1/2}}{\sum_{i=1}^n \lambda_i^{(n)}} \ i.o.) = 1$$

for $1 < p < 2$ and any $\vartheta' \in (0, (q-p)/3)$, and

$$\mathbb{P}(W_n(X_1, \cdots, X_n) \gtrsim \frac{\left(\sum_{i=1}^n \lambda_i^{(n)2} \log\log \sum_{i=1}^n \lambda_i^{(n)2}\right)^{1/2}}{\sum_{i=1}^n \lambda_i^{(n)}} \ i.o.) = 1$$

for $p = 2$.

Theorem 5 implies that the confidence sequence shrinks in order of $O((\frac{\log\log t + \log(1/\alpha)}{t})^{\frac{p-1}{p}})$ which is faster than the bound given in Theorem 2. Especially when $p = 2$, the order of width reduces to $O((\frac{\log\log t + \log(1/\alpha)}{t})^{\frac{1}{2}})$, which matches the lower bound given by Theorem 6. When $1 < p < 2$, there is still a small gap between the lower bound and upper bounds. It remains an open question whether we could further improve it.

7. Applications

We apply our methods described in the previous sections to two settings, any-time valid risk control and any-time valid confidence set construction.

7.1 Risk Control

We define $R(X, \beta)$ to be the risk of observation X, where $\beta \in \mathbb{R}$ is a user-specified threshold. Without loss of generality, we can assume that the function $R(X, \beta)$ is a decreasing function of β. That is, higher threshold β leads to a smaller risk value. We define $R(\beta) := \mathbb{E}[R(X, \beta)]$ to be the expected risk. We aim to choose a reasonable β to control the risk under a certain level r^*. To be more specific, we want to choose a sequence of $\{\beta_n, n = 1, ...\}$ such that

$$\mathbb{P}(R(\beta_n) \leq r^* \text{ for } n = 1, 2, ...) \geq 1 - \alpha. \tag{7.18}$$

To construct $\{\hat{\beta}_t\}$ that satisfies (7.18), we consider the following functional sequence,

$$M_n(\beta) := \prod_{i=1}^{n} \exp\{\phi(-\lambda_i(R_i(\beta)) - r^*)\} \cdot \exp\{-C_p v_p \lambda_i^p\}.$$

It can be checked that $M_n(\beta^*)$ is a non-negative supermartingale, where β^* is the solution to $R(\beta^*) = r^*$.

Define

$$\hat{\beta}_n := \min\{\beta : M_n(\beta') \geq 1/\alpha \quad \text{for} \quad \beta' > \beta\}. \tag{7.19}$$

Therefore, $\{\hat{\beta}_n, n = 1, 2, ...\}$ is the desired sequence and we have the following theorem.

Theorem 7. *The sequence* $\{\hat{\beta}_n; n = 1, 2, ...\}$ *defined in* (7.19) *provides any-time risk level guarantee, i.e.,*

$$\mathbb{P}(R(\hat{\beta}_n) \leq r^*, \ \textit{for } n = 1, 2, ...) \geq 1 - \alpha.$$

One practical application of Theorem 7 is the conditional value at risk (Pflug, 2000; Rockafellar and Uryasev, 2002), where we can treat $R(X, \beta) = X\mathbf{1}\{X > \beta\}$. In other words, we want to dynamically choose β so that the expected tail loss is controlled under a certain level.

7.2 Parameter Confidence Set

We consider a regression problem under the heavy tail setting, that is,

$$Y = X^T\boldsymbol{\beta} + \epsilon, \tag{7.20}$$

where $Y \in \mathbb{R}$, $X \in \mathbb{R}^d$, and $\boldsymbol{\beta}$ is the parameter vector to be estimated. Both X and Y are possibly heavy-tailed. Moreover, the noise term ϵ is assumed to be independent of X and satisfy $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon^{2+\vartheta}] < \infty$ with

$\vartheta > 0$. Let $p = 1 + \vartheta/2$ and also suppose $\mathbb{E}[|XX^T|^p]$ and $\mathbb{E}[|YX|^p]$ exist, where A^p represents the elememt-wise p-th power of matrix/vector A. We define the loss for the sample (X, Y) to be $L(X, Y; \boldsymbol{\beta}) := (Y - X^T\boldsymbol{\beta})^2$ and its population version to be $l(\boldsymbol{\beta}) := \mathbb{E}[(Y - X^T\boldsymbol{\beta})^2]$. Then it is easy to see that $\mathbb{E}[L(X, Y; \boldsymbol{\beta})^p]$ exists for any fixed $\boldsymbol{\beta}$. We let $L_i(\boldsymbol{\beta}) = L(X_i, Y_i; \boldsymbol{\beta})$ be the loss of the i-th observation.

We define

$$M_n(a) = \prod_{i=1}^{n} \exp\{\phi(\lambda_i(Z_i - a)\} \cdot \exp\{-C_p v_p \lambda_i^p\}, \tag{7.21}$$

where Z_i could take form of $Y_i X_i[j]$ or $X_i[j_1]X_i[j_2]$ $(j, j_1, j_2 \in [d])$ and $v_p := \max\{\mathbb{E}[|X[j_1]X[j_2] - \mathbb{E}[X[j_1]X[j_2]]|]^p, \max\{\mathbb{E}[|X[j]Y - \mathbb{E}[X[j]Y|]]^p\}$. It is easy to check that $M_n(a)$ is a non-negative supermartingale for $a = \mathbb{E}[Z_i]$. At each round n, we denote the solution to

$$0 = \prod_{i=1}^{n} \exp\{\phi(\lambda_i(Y_i X_i[j] - a)\} \cdot \exp\{-C_p v_p \lambda_i^p\}, \tag{7.22}$$

as $\widehat{XY}[j]$ and the solution to

$$0 = \prod_{i=1}^{n} \exp\{\phi(\lambda_i(X_i[j_1]X_i[j_2] - a)\} \cdot \exp\{-C_p v_p \lambda_i^p\}, \tag{7.23}$$

as $\widehat{XX^T}[j_1, j_2]$. Therefore, at time n, the estimates for $\mathbb{E}[XY]$, $\mathbb{E}[XX^T]$ are $\widehat{XY}$ and $\widehat{XX^T}$, respectively. The parameter estimate $\hat{\boldsymbol{\beta}}_n$ is then constructed as follows,

$$\hat{\boldsymbol{\beta}}_n := \arg\min_{\boldsymbol{\beta}} \hat{l}_n(\boldsymbol{\beta}), \tag{7.24}$$

with

$$\hat{l}_n(\boldsymbol{\beta}) := -2\widehat{XY}\boldsymbol{\beta} + \boldsymbol{\beta}^T\widehat{XX^T}\boldsymbol{\beta}. \tag{7.25}$$

With these preparations, we can construct a sequence of parameter confidence sets which simultaneously contain the true parameter $\boldsymbol{\beta}^*$ with probability at least $1-\alpha$.

Theorem 8. *Given the choices of λ_i's in Theorem 5, we define the confidence region*

$$\mathcal{C}_{n,\boldsymbol{\beta}} := \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n\| \le 4(d\|\boldsymbol{\beta}^*\| + \sqrt{d})\delta_n/\lambda_{min}\},$$

with

$$\delta_n := 2\frac{\left((C_p v_p + 1)q\log(2(\log_a n + 1)\sum_{l=1}^{\infty} l^{-q}/\alpha')\right)^{1-1/p}}{n(1/an)^{1/p}},$$

and $\alpha' := \alpha/(d^2+d)$. Then it holds

$$\mathbb{P}(\boldsymbol{\beta}^* \in \mathcal{C}_{n,\boldsymbol{\beta}},\ \textit{for all}\ \ n \ge \max\{t_0, n_0\}) \ge 1-\alpha. \tag{7.26}$$

Here n_0 is the largest integer n such that $\delta_n \ge \lambda_{min}/2d$ and t_0 is the same as in Theorem 5.

It is not hard to check that n_0 is of order $\tilde{O}((d/\lambda_{min})^{\frac{p}{p-1}})$. By Theorem 8, the radius of confidence set is asymptotically $O(\frac{d(\log\log n + \log d)^{1-1/p}}{n^{1-1/p}})$. In other words, Catoni-based approach gives the law of iterated logarithm

result for parameter estimation. When $p = 2$, we find that the radius is of order $\frac{d\sqrt{\log\log n + \log d}}{n^{1/2}}$, which is $\tilde{O}(\sqrt{d})$ worse than the classical rate given the fixed number of observations n. This is the price we pay for estimating $\mathbb{E}[XX^T]$ via using the Catoni influence function. If the second moment of XX^T exists, we can use the average estimator $\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$ instead of $\widehat{XX^T}$. Then the rate $O(\frac{d(\log\log n + \log d)^{1-1/p}}{n^{1-1/p}})$ can reduce to $O(\frac{\sqrt{d}(\log\log n + \log d)^{1-1/p}}{n^{1-1/p}})$ which is nearly optimal.

Remark 2. It will be interesting to study parameter confidence set construction in more complex models, such as generalized linear models and multi-layer perceptrons, other than (7.20). It remains an open question whether we can also achieve a similar law of iterated logarithm rate under the general model.

8. Numerical Illustration

In this section, we examine the empirical performance of the three methods. The choice of tuning parameters and scaling sequences are given below and are used throughout the entire section.

1. Dubins-Savage method: $b = 1$, $v = 1$, $\alpha = 0.05$, $a = \frac{1}{m_p b^{\frac{1}{p-1}}} \cdot \left(\left(\frac{2}{\alpha}\right)^{\frac{1}{p-1}} - 1\right)$ with $m_p = (\frac{p-1}{2^{2-p}})^{1/(p-1)}$. $\lambda_t = (\frac{a}{t(p-1)})^{1/p}$ for $t = 1, 2, \ldots$.

2. Catoni-Ville method: $v = 1, \alpha = 0.05$, $\lambda_t = \frac{1}{t^{1/p}}$ for $t = 1, 2, \ldots$.

3. Stitch-Catoni-Ville method: $v = 1, \alpha = 0.05$, $\lambda_i^{(t)} \equiv \tau \cdot (\frac{\log(\frac{2}{\alpha_t})}{t})^{1/p}$, where $\alpha_t = \alpha \frac{\lceil \log_a(t) \rceil^q}{\sum_{q=1}^{\infty} j^q}$. We further set $a = 2, q = 1.4, \tau = (C_p)^{-1/p}$ with $C_p = \left(\frac{p-1}{p}\right)^{p/2} \left(\frac{2-p}{p-1}\right)^{(2-p)/2}$. We also write $S_q = \sum_{q=1}^{\infty} j^q$.

The empirical performances of confidence width under three methods are shown in Figure 1.

From upper two plots in Figure 1, we can see that "Dubins-Savage" method is much worse. The width decreases to zero at a slower rate. Moreover, when $p \leq 1.7$, we find that the width of "Dubins-Savage" method is more than 1000 times wider than the other two methods. It indicates that "Dubins-Savage" method is not useful in very heavy-tailed situations.

From bottom four plots in Figure 1, we can see that Stitch-Catoni-Ville method is clearly better than Catoni-Ville method when p is close to 2. As p decreases from 2 to 1, Catoni-Ville method in fact performs well especially when the number of samples n is small. (For example, the confidence width of Catoni-Ville method is always tighter for $1 \leq n \leq 100,000$ when $p = 1.2$.) This suggests the practical advantage of Catoni-Ville method, while Stitch-Catoni-Ville method enjoys a theoretically better rate.

According to this observation, we propose to construct the following

confidence sequence (CS),

$$L_n(X_1,\ldots,X_n) = \max\{L_n^{Cat+Ville}(X_1,\ldots,X_n), L_n^{Stitch}(X_1,\ldots,X_n)\} \text{ and}$$
$$U_n(X_1,\ldots,X_n) = \min\{U_n^{Cat+Ville}(X_1,\ldots,X_n), U_n^{Stitch}(X_1,\ldots,X_n)\},$$

which is the tightest possible CS under our current technique and is the best result known so far in the literature.

Moreover, by calculations, we find that

$$\begin{aligned} U_n^{Cat+Ville}(X_1,\ldots,X_n) - L_n^{Cat+Ville}(X_1,\ldots,X_n) &= 2\frac{\log(n)\cdot C_p + \log(\frac{2}{\alpha})}{\frac{p}{p-1}n^{1/p}}\cdot(1+o_p(1)) \\ U_n^{Stitch}(X_1,\ldots,X_n) - L_n^{Stitch}(X_1,\ldots,X_n) &= 2\frac{(q\log\log n + \log(\frac{2S_q}{\alpha}))^{1-1/p}\cdot(2C_p^{1/p})}{n^{1/p}}\cdot(1+o_p(1)). \end{aligned}$$

There exists a threshold n_c such that

$$\log(n)\cdot C_p + \log(\frac{2}{\alpha}) - (q\log\log n + \log(\frac{2S_q}{\alpha}))^{1-1/p}\cdot(2C_p^{1/p})\cdot\frac{p}{p-1} \quad (8.27)$$

changes its sign! If the value in (8.27) is negative, then the Catoni-Ville method gives a tighter width. Otherwise, Stitch-Catoni-Ville method generates a tighter confidence sequence. To summarize, it happens that

- $L_n^{Cat+Ville}(X_1,\ldots,X_n) > L_n^{Stitch}(X_1,\ldots,X_n)$ (or $U_n^{Cat+Ville}(X_1,\ldots,X_n) < U_n^{Stitch}(X_1,\ldots,X_n)$) for $n < n_c$;
- $L_n^{Cat+Ville}(X_1,\ldots,X_n) < L_n^{Stitch}(X_1,\ldots,X_n)$ (or $U_n^{Cat+Ville}(X_1,\ldots,X_n) > U_n^{Stitch}(X_1,\ldots,X_n)$) for $n \geq n_c$.

To help readers gain more intuitions, we provide a table of thresholds size n_c; see Table 2.

p	2	1.8	1.6	1.4	1.3	1.25	1.2	1.15
n_c	4288	1819	1018	2137	9121	33165	245972	$>$ 7,286,000

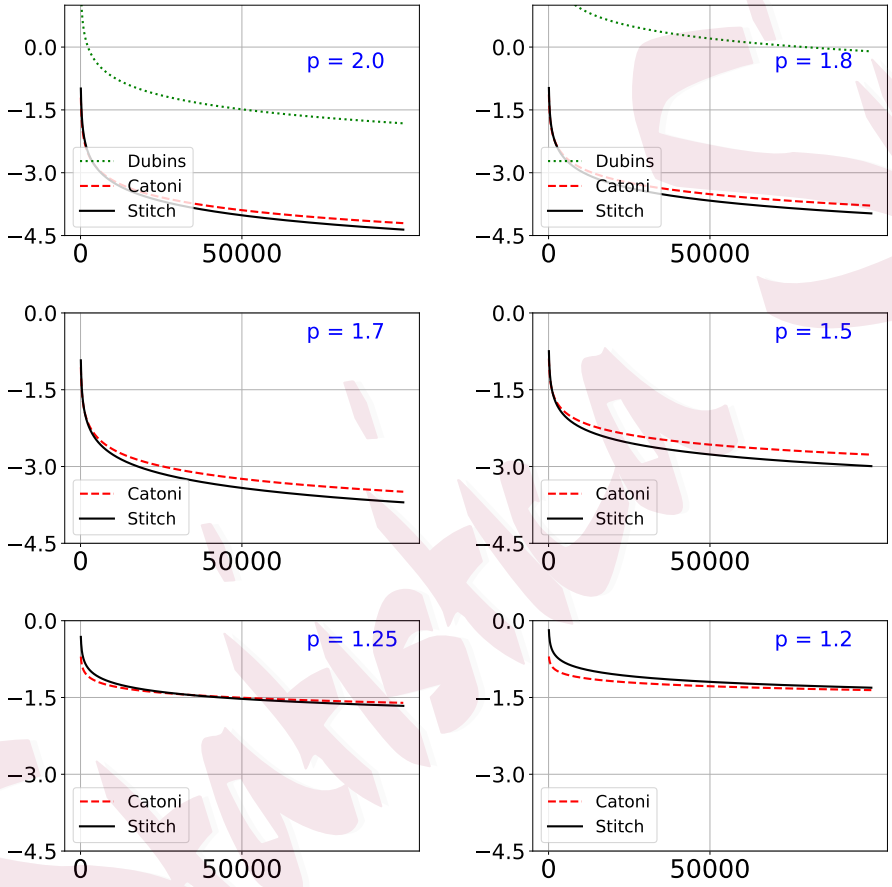
Table 2: Threshold n_c under different p's. Larger n_c means "Stitch-Catoni-Ville" method performs less efficiently.

9. Real Data Examples

We compare the performances of confidence sequences given by two methods, "Catoni + Ville's ineq" and "Stitch + Catoni + Ville's ineq" in the two real data applications.

A/B test. In the first application, we consider an A/B testing problem. We choose the air pollution data, which is publicly available at R package "texmex". This data set is from Leeds (U.K.) city center, collected from 1994 to 1998. It is split into two parts, "summer" and "winter". The summer part corresponds to the months of April to July. The winter part corresponds to the months of November to February. Some outliers have been removed, as discussed by Heffernan and Tawn, 2004. We want to study whether there exists a significant difference between NO air-quality

Figure 1: Plots of Confidence Widths for Three Methods, "Dubins-Savage", "Catoni+Ville" and "Stitch+Catoni+Ville". The X-axis shows the number of samples (i.e., $1 \leq n \leq 100,000$). The Y-axis shows the logarithm of confidence width.



indices of summer and that of winter. The A/B setting can be formulated as follows. We denote $y_t^{(s)}$ as the observed NO index for t-th day in the "summer" subset and $y_t^{(w)}$ as the observed NO index for t-th day in the

"winter" subset. Suppose $y_t^{(s)}$'s and $y_t^{(w)}$ are i.i.d. respectively. Let $\mu^{(s)}$ and $\mu^{(w)}$ be the mean NO index for the summer and winter. We want to know whether there is a siginificant difference between the mean values, $\mu^{(s)}$ and $\mu^{(w)}$. If so, how many days of observations do we need to test this difference. We denote $\{I_t^{(s),CV}\}$ and $\{I_t^{(w),CV}\}$ as the confidence sequences for summer and winter data, respectively, via "Catoni + Ville's ineq" (CV) method. We also let $t_{stop}^{CV} := \min\{t : I_t^{(s),CV} \cap I_t^{(w),CV} = \emptyset\}$. $\{I_t^{(s),SCV}\}$, $\{I_t^{(w),SCV}\}$, and t_{stop}^{SCV} are similarly defined for "Stitch + Catoni + Ville's inseq" (SCV) method. We compare between t_{stop}^{CV} and t_{stop}^{SCV} to see which method could give the smaller data requirement.

We first provide some exploratory analysis. A histogram of NO index for summer and winter parts is given in Figure 2. We can clearly see the data distribution is right-skewed, indicating the heavy-tail phenomenon. We also calculate the Hill estimator of tail index for this dataset and obtain $\hat{p} \approx 3.01$. Therefore, we can safely take the moment $p = 2$ for the following analysis. In the implementation, we take $\alpha = 0.05$, $a = 2$, $q = 2$, and v_p to be the empirical estimate based on the data. The confidence sequences of the two methods are shown in Figure 3. As we can see, the confidence width of the SCV method is a little bit wider than that of CV method. It is also computed that $t_{stop}^{CV} = 171$ and $t_{stop}^{SCV} = 187$, indicating that CV can

stop the A/B test earlier compared with SCV. This phenomenon suggests that confidence sequences based on LIL could be conservative in practical applications.

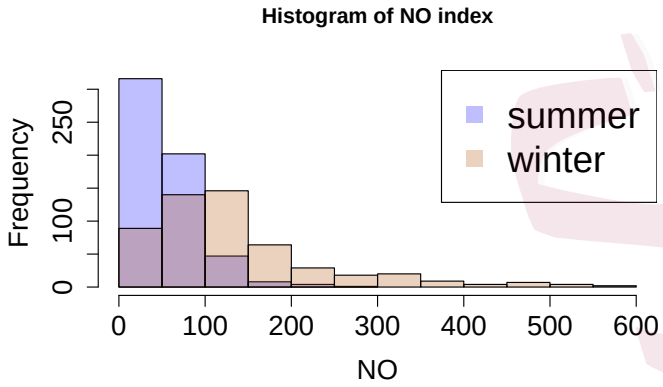


Figure 2: Histogram of NO index for summer and winter parts.

Cost Control. In the second application, we consider a cost control problem. We chose the artauction dataset, which is available at https://github.com/jasonshi10/art_auction_valuation/tree/master. The dataset consists of 37,638 art pieces sold at a total valuation of $ 9.47 billion. Sold prices include a maximum of $119.92 million and a minimum of $3. A quantile table is given in Table 3. We can clearly see that the sold prices are heavy-tailed. We treat ourselves as bidders and want to control the prices we bid on the auctions. We let β be the strategic threshold, i.e., we only bid on the art piece when its price is below β. Otherwise, we give up this art piece. Therefore, for the t-th art piece, our cost is $X_t = Y_t \cdot \mathbf{1}\{Y_t \leq \beta\}$,

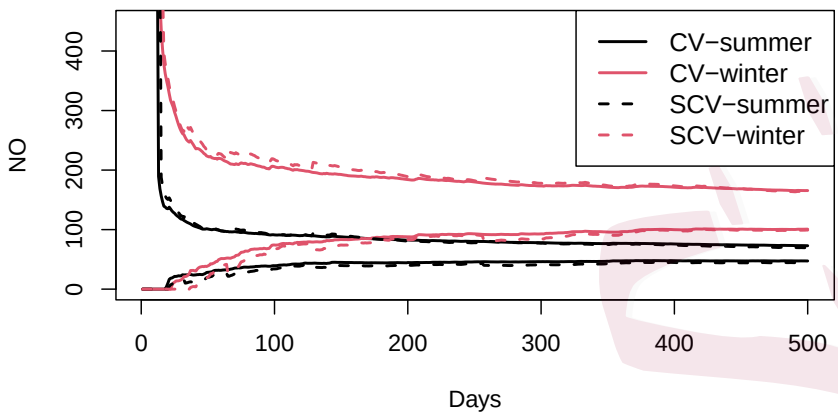


Figure 3: Confidence sequences of NO index given by the two methods for the air pollution data. Black lines stand for the summer part and red lines stand for the winter part. Solid lines are for "Catoni + Ville's ineq" and dashed lines are for "Stitch + Catoni + Ville's ineq".

where Y_t is the observed price of the t-th art piece. For simplicity, we suppose Y_t's are i.i.d.. Our goal is to choose suitable β such that the average cost does not exceed a pre-determined value r^*, that is, $\mathbb{E}[X_t] \leq r^*$. We still apply CV and SCV methods to this cost control problem. The better method could return a higher threshold β so that the bidder can buy more art pieces while keeping the cost below r^*.

In the implementation, we take $r^* = 50,000$ dollars, $\alpha = 0.05$, $a = 2$, $q = 2$. In the dataset, some art pieces' sold dates are missing. We remove

Prices of Art Pieces (unit: $)					
Quantile	0%	25%	50%	75%	100%
Value	3	1714	7021	25000	119922500

Table 3: The quantile table of art pieces.

those and have 17678 art pieces remaining. We randomly select 26,78 art pieces to calculate the moment order p and the moment v_p, and use the rest, 15,000 to construct the confidence sequences. By calculation, we have $\hat{p} = 1.78$. The estimated threshold curves are plotted in Figure 4. We find that the "Catoni + Ville's ineq" method can return higher bidding threshold than that of the "Stitch + Catoni + Ville's ineq" method. This again indicates that the LIL-based method is conservative in bidding the price so that the average cost can be controlled below r^*.

10. Conclusion

We provided an extension of confidence sequences for settings where the variance of the data-generating distribution need not exist. Dealing with such challenging heavy-tail settings required using robust estimation methods to obtain acceptable deviation bounds. We made use of the influence functions inspired by Catoni (2012) to obtain Catoni-style confidence se-

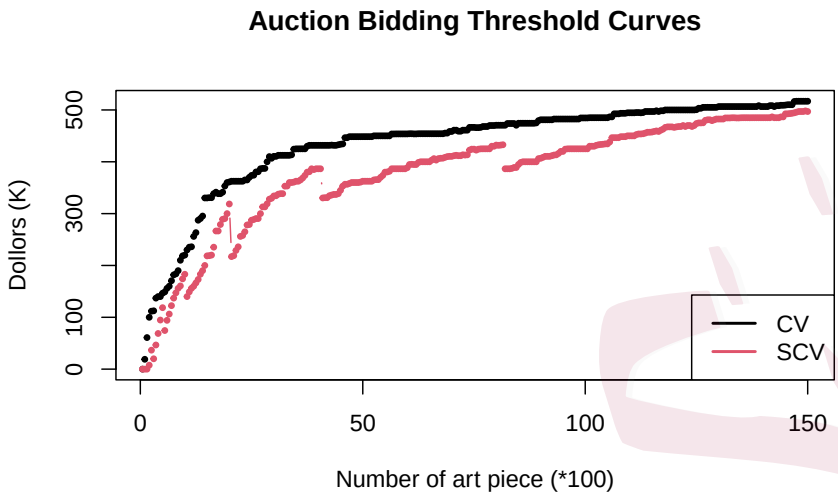


Figure 4: The curves of bidding thresholds returned by two methods.

quences. We established lower bounds on the widths of the Catoni-style confidence sequences for both finite and infinite variance cases using a general law of iterated logarithm. We provided a thorough theoretical analysis of constructing the confidence sequence using the Dubins-Savage inequality, Ville's inequality, and stitching methods. The theoretical results add to the scant literature on online decision-making under heavy-tail settings. Applications of risk control and confidence set construction are given to show the practical importance and usefulness of our method. In the future, it will be interesting to further expand the applicability of the proposed method to different real-world sequential decision problems.

Acknowledgment. The authors thank the Associate Editor and the anonymous referee for their constructive suggestions and comments on improving the quality of our paper. The earlier version of this work was done when the authors were employed by Baidu USA. Guanhua Fang is partly supported by the National Natural Science Foundation of China (nos. 12301376) and Shanghai Educational Development Foundation (23CGA02). Gennady Samorodnitsky is partially supported by U.S. National Science Foundation grant (DMS-2310974) at Cornell University.

References

Bhatt, S., G. Fang, P. Li, and G. Samorodnitsky (2022a). Minimax M-estimation under adversarial corruption. In *International Conference on Machine Learning*. PMLR.

Bhatt, S., G. Fang, P. Li, and G. Samorodnitsky (2022b). Nearly optimal catoni's M-estimator for infinite variance. In *International Conference on Machine Learning*. PMLR.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, Volume 48, pp. 1148–1185.

Chen, P., X. Jin, X. Li, and L. Xu (2021). A generalized catoni's m-estimator under finite α-th moment assumption with $\alpha \in (1, 2)$. *Electronic Journal of Statistics 15*(2), 5523–5544.

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

Dubins, L. E. and L. J. Savage (1965). A tchebycheff-like inequality for stochastic processes. *Proceedings of the National Academy of Sciences of the United States of America 53*(2), 274.

Howard, S. R., A. Ramdas, J. McAuliffe, and J. Sekhon (2021). Time-uniform, nonparametric,

nonasymptotic confidence sequences. *The Annals of Statistics 49*(2), 1055–1080.

Jamieson, K. G. and L. Jain (2018). A bandit approach to sequential experimental design with false discovery control. *Advances in Neural Information Processing Systems 31*.

Jennison, C. and B. W. Turnbull (1989). Interim analyses: the repeated confidence interval approach. *Journal of the Royal Statistical Society: Series B (Methodological) 51*(3), 305–334.

Johari, R., P. Koomen, L. Pekelis, and D. Walsh (2017). Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1517–1525.

Johari, R., L. Pekelis, and D. J. Walsh (2015). Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922*.

Kallenberg, O. (1975). On the existence and path properties of stochastic integrals. *The Annals of Probability*, 262–280.

Khan, R. A. (2009). l_p-version of the dubins–savage inequality and some exponential inequalities. *Journal of Theoretical Probability 22*(2), 348–364.

Malek, A. and S. Chiappa (2021). Asymptotically best causal effect identification with multi-armed bandits. *Advances in Neural Information Processing Systems 34*.

Pflug, G. C. (2000). Some remarks on the value-at-risk and the conditional value-at-risk. *Probabilistic constrained optimization: Methodology and applications*, 272–281.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society 58*(5), 527–535.

Rockafellar, R. T. and S. Uryasev (2002). Conditional value-at-risk for general loss distributions. *Journal of banking & finance 26*(7), 1443–1471.

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

Wang, H. and A. Ramdas (2023). Catoni-style confidence sequences for heavy-tailed mean

estimation. *Stochastic Processes and Their Applications 163*, 168–202.

Waudby-Smith, I. and A. Ramdas (2024). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology 86*(1), 1–27.

Wittmann, R. (1985). A general law of iterated logarithm. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 68*(4), 521–543.

Zhan, R., V. Hadad, D. A. Hirshberg, and S. Athey (2021). Off-policy evaluation via adaptive weighting with data from contextual bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2125–2135.