Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

DATA-DRIVEN ESTIMATION OF TAIL PROBABILITIES UNDER VARYING DISTRIBUTIONAL ASSUMPTIONS

Dohyun Ahn[1], Sandeep Juneja[2], Tejas Pagare[2], and Shreyas Samudra[2]

[1]Department of SEEM, The Chinese University of Hong Kong, Shatin, NT, HONG KONG
[2]Safexpress Center for Data, Learning and Decision Sciences, Ashoka University, INDIA

ABSTRACT

We consider estimating $p_x = \mathbb{P}(X > x)$ in a data-driven manner or through simulation, when x is large and when independent samples of X are available. Naively, this involves generating $O(1/p_x)$ samples. Making distributional assumptions on X reduces the sample complexity under commonly used distributional parameter estimators. It equals $O(\log(1/p_x))$ for Gaussian distribution with unknown mean and known variance, and $O(\log^2(1/p_x))$ when the variance is also unknown and when the distribution is either exponential or Pareto. We also critically examine the more sophisticated assumption that the data belong to the domain of attraction of the Fréchet distribution allowing estimation methods from extreme value theory (EVT) . Our sobering and practically important conclusion based on sample complexity analysis and numerical experiments is that under these settings errors from estimation can be significant so that for probabilities as low as 10^{-6}, naive methods may be preferable to those based on EVT.

1 INTRODUCTION

Estimating rare event probabilities is crucial in a variety of applications, including finance, insurance, communications networks, traffic modeling, and climate modeling; see Glasserman (2003), Rubino and Tuffin (2009), and Asmussen and Glynn (2007) for comprehensive overviews. A fundamental problem is to estimate the probability $p_x = \mathbb{P}(X > x)$ for a critical random variable X that exceeds a large threshold x. For example, X may denote daily rainfall, and the interest might be in the possibility of observing an extreme precipitation event (de Vries et al. 2024). In finance, one may seek to estimate the probability of a severe market downturn within a single day. Many variance reduction techniques have been developed to address this problem, such as importance sampling, conditional Monte Carlo, splitting, and stratification; see, for example, Glasserman et al. (2000), Juneja and Shahabuddin (2002), L'Ecuyer et al. (2006), Dean and Dupuis (2009), Botev et al. (2016), Bai et al. (2022), Deo and Murthy (2025), and Ahn and Zheng (2025). All of these methods typically rely on the knowledge of specific underlying distributions, which is often not available in practice. In contrast, we focus on completely data-driven settings where such distributional knowledge is incomplete or entirely absent.

A naive, yet distributionally robust, data-driven approach is simply to observe a large number of independent samples of X and develop an empirical estimator of p_x guaranteed to be within a small percentage (e.g., δ) of p_x with high probability (e.g., $1-\delta_0$). Such an approach requires $O(1/p_x)$ samples of X; for instance, achieving a desired relative accuracy is 10% (i.e., $\delta = 0.1$) with a confidence level of 95% (i.e., $\delta_0 = 0.05$) roughly necessitates $480/p_x$ number of samples. This could be a large number when x is large, motivating more approximate approaches to estimate p_x. One pragmatic approach then is to assume that X belongs to a parametric family of distributions and to estimate its underlying parameters from data using commonly accepted techniques, in the hope that this parameter estimation will require fewer samples than the naive approach in achieving the same accuracy. For example, past experience may suggest that the data-generating distribution adheres to the Gaussian family, either with unknown mean and known variance, or with both unknown mean and variance.

In this paper, we analyze the number of samples required under these assumptions to estimate parameters such that the resulting error in the estimator for p_x matches that of the naive estimator. For the Gaussian distribution, our analysis shows that $O(\log(1/p_x))$ samples are needed when the mean is known but the variance is known, and a larger $O(\log^2(1/p_x))$ samples when both are unknown. We leverage and refine known concentration equalities for estimators of Gaussian parameters to derive lower bounds on the sample sizes that ensure the desired accuracy. We perform a similar analysis for the exponential distribution, which exhibits a heavier tail than the Gaussian, and again find that $O(\log^2(1/p_x))$ samples assure the desired accuracy. Since the Pareto distribution is fat-tailed, one might naively expect that estimating its parameters from the data would require a significantly larger sample size than for Gaussian and exponential distributions. However, we observe, somewhat surprisingly, that $O(\log^2(1/p_x))$ samples are also sufficient here. This represents a substantial reduction in sample size compared to the native estimator at least asymptotically (as $x \to \infty$). We confirm all these theoretical findings through numerical experiments.

Our key observation, which we believe is underappreciated by both theoreticians and practitioners, carries significant practical implications. Extreme Value Theory (EVT) establishes that when X belongs to the domain of attraction of a Fréchet distribution (a large class of heavy-tailed random variables), the distribution of $X-u$ conditioned on $X > u$ converges to the generalized Pareto distribution (GPD) as u increases; see McNeil et al. (2015) for an overview of EVT and a detailed discussion on this convergence. This theory has motivated a pragmatic approach: selecting a reasonably large value of u under the assumption that the distribution of $X-u$ conditioned on $X > u$ is GPD, and subsequently estimating the parameters for this GPD from generated data. We observe that when the underlying distribution is indeed GPD (thereby precluding any EVT-based approximation), parameter estimation via a widely used method requires more samples than a naive estimator for the same accuracy when the target probability is of order 10^{-6} or larger. This threshold can increase when the underlying distribution is regularly varying but not precisely a GPD. This is a sobering observation because, in practice, extreme event probabilities of order $10^{-4} - 10^{-2}$ are often of interest, a range where researchers commonly apply EVT approximations. Our experiments suggest that, within this range, the naive estimator might prove to be the most reliable option. It is noteworthy that while the existing EVT literature focuses extensively on a variety of estimators for estimating the polynomial decay rate of the tail probability and associated concentration inequalities, the errors introduced by these approximations in the estimated tail probabilities appear to have limited discussion.

The remainder of this paper is organized as follows. Section 2 presents the background on estimator efficiency and discusses the naive estimator. In Section 3, we discuss the sample complexity under common distributional assumptions on X. In Section 4, we address scenarios where the underlying distribution family is not explicitly known, but the distribution is assumed to lie in the domain of attraction of the Fréchet distribution. The numerical experiments are conducted in Section 5. Finally, Section 6 provides a brief conclusion.

2 BACKGROUND

Given an estimator $\hat{p}_n$ for the tail probability $p_x := \mathbb{P}(X > x)$ under various distributional assumptions, our objective is to characterize the minimal sample size n so that for sufficiently large x the following is satisfied for any $\delta, \delta_0 \in (0,1)$

$$\mathbb{P}(|\hat{p}_n - p_x| > \delta p_x) \leq \delta_0. \tag{1}$$

To establish a benchmark, we first construct a naive estimator for the tail probability p_x as $\hat{p}_n^{\mathtt{N}} := (1/n)\sum_{i=1}^{n} \mathbb{I}\{X_i > x\}$, where X_i is the i-th i.i.d. sample of X, and $\mathbb{I}\{A\}$ is the indicator function of A. Using the multiplicative Chernoff bound on the centered Binomial random variable $n\hat{p}_n^{\mathtt{N}} - np_x$ we get

$$\mathbb{P}\left(|\hat{p}_n^{\mathtt{N}}/p_x - 1| > \delta\right) \leq 2\exp\left(-\frac{np_x\delta^2}{3}\right).$$

It follows that for for any $\delta, \delta_0 \in (0,1)$, if the sample size n satisfies $n \geq p_x^{-1}(3/\delta^2)\log(2/\delta_0)$, then $\mathbb{P}(|\hat{p}_n^{\mathtt{N}}/p_x - 1| \geq \delta) \leq \delta_0$.

Remark 1 (Selecting δ_0 in high-impact rare-event setting)**.** The rare event community is often satisfied with confidence intervals for a rare event quantity p with a small relative width δ, say 10%, and a confidence level that allows an error probability δ_0 to be, for example, around 5%, which is significantly higher than p. This relatively high value of δ_0, however, may distort catastrophic risk assessment when a conservative view is adopted. For instance, in a conservative worst-case scenario, one needs to set the estimator of p to 1 with probability δ_0. Then, the conservative estimate of the rare event probability is estimated to be $\delta_0 + (1-\delta_0)(1+\delta)p$. For small p, δ_0 dominates this assessment. One way to handle this issue is to keep δ_0 of a similar or lower order than p. As we observe, the sample complexity in our reaults depends on δ_0 only through the term $\log(1/\delta_0)$. Thus, keeping δ_0 of such an order does not lead to a dramatic increase in sample complexity, but may be essential in conservative risk management settings.

3 PARAMETRIC ESTIMATORS

We now demonstrate the significant reduction in sample complexity achieved under the assumption that the distribution family of X is known and is Gaussian, exponential, or Pareto, and we use data to estimate the distributional parameters using commonly used algorithms.

3.1 Gaussian Distribution

In this section, we consider a situation where the decision-maker is aware that the underlying distribution is Gaussian. We first consider a Gaussian distribution with unknown mean μ and known variance σ^2 and then study the unknown variance case later. In the first case, estimating the tail probability thus reduces to estimating the mean. Accordingly, we define an estimator of $p_x = 1 - \Phi((x-\mu)/\sigma)$ as $\hat{p}_n^{\mathtt{G1}} = 1 - \Phi((x-\hat{\mu}_n)/\sigma)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and $\hat{\mu}_n = \sum_{i=1}^n X_i/n$ represents the sample mean of X when the sample size is n.

Proposition 1. Assume that X follows a Gaussian distribution with unknown mean μ and known variance σ^2. For any $\delta, \delta_0 \in (0,1)$, when the sample size satisfies

$$n \geq \frac{2\log(2/\delta_0)}{\log^2(1+\delta)} \left(\frac{x-\mu}{\sigma}\right)^2 \overset{x\to\infty}{\sim} \frac{4\log(2/\delta_0)}{\log^2(1+\delta)} \log(1/p_x),$$

we have $\mathbb{P}(|\hat{p}_n^{\mathtt{G1}}/p_x - 1| > \delta) \leq \delta_0$ for all sufficiently large x.

Proof. Recall that $\hat{\mu}_n - \mu$ is Gaussian with mean 0 and variance σ^2/n. Then, by the concentration inequality for sub-Gaussian random variables (Boucheron et al. 2013), we have

$$\mathbb{P}(|\hat{\mu}_n - \mu| > \varepsilon) \leq 2\exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right) \quad \forall \varepsilon > 0 \tag{2}$$

Since $\hat{\mu}_n \geq \mu$ if and only if $\hat{p}_n^{\mathtt{G1}} \geq p_x$, we observe that $\mathbb{P}\left(\left|\hat{p}_n^{\mathtt{G1}}/p_x - 1\right| > \delta\right)$ is bounded from above by

$$\mathbb{P}\left(\frac{\hat{p}_n^{\mathtt{G1}}}{p_x} > 1+\delta, \hat{\mu}_n \in [\mu, \mu+\varepsilon)\right) + \mathbb{P}\left(\frac{\hat{p}_n^{\mathtt{G1}}}{p_x} < 1-\delta, \hat{\mu}_n \in (\mu-\varepsilon, \mu)\right) + \mathbb{P}(|\hat{\mu}_n - \mu| \geq \varepsilon) \quad \forall \varepsilon > 0. \tag{3}$$

Therefore, if ε satisfies

$$p_{x-\varepsilon} := 1 - \Phi\left(\frac{x-\mu-\varepsilon}{\sigma}\right) \leq (1+\delta)p_x \quad \text{and} \quad p_{x+\varepsilon} := 1 - \Phi\left(\frac{x-\mu+\varepsilon}{\sigma}\right) \geq (1-\delta)p_x, \tag{4}$$

then the first two probabilities in (3) are zero. To achieve this, we use the following asymptotic equivalence as $x \to \infty$:

$$\frac{p_{x-\varepsilon}}{p_x} \sim \exp\left(-\frac{(x-\mu-\varepsilon)^2}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^2}\right).$$

Thus, the first and second inequalities in (4) are asymptotically equivalent to

$$\varepsilon(2x-2\mu-\varepsilon) \le 2\sigma^2\log(1+\delta) \quad \text{and} \quad \varepsilon(2x-2\mu+\varepsilon) \le 2\sigma^2\log\frac{1}{1-\delta}.$$

Given this analysis, one possible choice of ε that asymptotically satisfies the above two inequalities for large x is $\varepsilon = \sigma^2(x-\mu)^{-1}\log(1+\delta)$. Accordingly, using this value of ε and the inequality in (3), we obtain the following asymptotic relationship for large x:

$$\mathbb{P}\left(\left|\frac{\hat{p}_n^{\mathtt{G1}}}{p_x}-1\right| > \delta\right) \le 2\exp\left(-\frac{n\sigma^2\log^2(1+\delta)}{2(x-\mu)^2}\right).$$

Hence, bounding the right-hand side by δ_0 leads to the desired result. □

We now delve into the case where both μ and σ^2 are unknown. In this case, we use the sample mean and sample variance to construct an estimator for p_x as $\hat{p}_n^{\mathtt{G2}} = 1-\Phi\left((x-\hat{\mu}_n)/\hat{\sigma}_n\right)$, where $\hat{\sigma}_n = \sqrt{\sum_{i=1}^n (X_i-\hat{\mu}_n)^2/(n-1)}$ represents the corrected sample standard deviation of X for the sample size of n.

Proposition 2. Assume that X follows a Gaussian distribution with unknown mean μ and unknown variance σ^2. For any $\delta, \delta_0 \in (0,1)$, if the sample size n satisfies

$$n \ge \max\left\{2\sigma^2, \frac{1}{2}\left(1+\sqrt{1+\frac{4\log(1+\delta)}{\log(1/p_x)}}\right)^2\right\}\frac{\log(4/\delta_0)}{\log^2(1+\delta)}\log^2(1/p_x), \tag{5}$$

then $\mathbb{P}(|\hat{p}_n^{\mathtt{G2}}/p_x - 1| > \delta) \le \delta_0$ for all sufficiently large x.

Proof. Using the fact that $(n-1)\hat{\sigma}_n^2/\sigma^2 \sim \chi^2_{n-1}$ and using the concentration inequality for a chi-squared random variable with k degrees of freedom (Laurent and Massart 2000, Lemma 1), we have

$$\mathbb{P}\left(|\hat{\sigma}_n^2-\sigma^2| > \sigma^2\left(\sqrt{\frac{2t}{n-1}}+\frac{2t}{n-1}\right)\right) \le 2e^{-t} \tag{6}$$

For all $\varepsilon > 0$, it is easy to see that $\mathbb{P}\left(\left|\hat{p}_n^{\mathtt{G2}}/p_x - 1\right| > \delta\right)$ is bounded from above by

$$\begin{aligned}&\mathbb{P}\left(\frac{\hat{p}_n^{\mathtt{G2}}}{p_x} > 1+\delta, |\hat{\mu}_n-\mu| \le \varepsilon, |\hat{\sigma}_n^2-\sigma^2| \le \sigma^2\varepsilon\right) + \mathbb{P}\left(\frac{\hat{p}_n^{\mathtt{G2}}}{p_x} < 1-\delta, |\hat{\mu}_n-\mu| \le \varepsilon, |\hat{\sigma}_n^2-\sigma^2| \le \sigma^2\varepsilon\right) \\ &+\mathbb{P}(|\hat{\mu}_n-\mu| > \varepsilon) + \mathbb{P}\left(|\hat{\sigma}_n^2-\sigma^2| > \sigma^2\varepsilon\right)\end{aligned} \tag{7}$$

Let $p_x^{\mu+\varepsilon_1,\sigma^2(1+\varepsilon_2)} = 1-\Phi\left((x-\mu-\varepsilon_1)/\sqrt{\sigma^2(1+\varepsilon_2)}\right)$ for any $\varepsilon_1, \varepsilon_2 > 0$. This increases in both ε_1 and ε_2. Then, if $p_x^{\mu+\varepsilon,\sigma^2(1+\varepsilon)} < (1+\delta)p_x$, the first term in (7) is zero. Similarly, if $p^{\mu-\varepsilon,\sigma^2(1-\varepsilon)} > (1-\delta)p_x$, the second term in (7) is zero. To satisfy these two conditions in the regime where $x \to \infty$, we use the following asymptotic relationship that holds for any $\varepsilon, \varepsilon > 0$:

$$\log\left(\frac{p_x^{\mu+\varepsilon,\sigma^2(1+\varepsilon)}}{p_x}\right) \overset{x\to\infty}{\sim} \frac{2\varepsilon(x-\mu)+\varepsilon(x-\mu)^2}{2\sigma^2(1+\varepsilon)}.$$

Thus, the said two conditions can be asymptotically achieved by choosing ε and ε such that

$$2\varepsilon(x-\mu)+\varepsilon(x-\mu)^2 \le \min\{2\sigma^2(1+\varepsilon)\log(1+\delta), 2\sigma^2(1-\varepsilon)\log(1/(1-\delta))\}.$$

Let $\varepsilon = \log(1+\delta)/\log(1/p_x)$. Then, one can show that the above inequality is satisfied asymptotically as $x \to \infty$. Given this value of ε, we now aim to achieve $\mathbb{P}(|\hat{\mu}_n - \mu| > \varepsilon) \le \delta_0/2$ and $\mathbb{P}(|\hat{\sigma}_n^2 - \sigma^2| > \sigma^2\varepsilon) \le \delta_0/2$. For the former, it suffices to find n_1 such that the right-hand side of (2) is bounded from above by $\delta_0/2$ for all $n \ge n_1$. A simple calculation leads to $n_1 = (2\sigma^2/\varepsilon^2)\log(4/\delta_0)$. To ensure the latter, we set $t = \log(4/\delta_0)$ in (6). Then, the right-hand side of (6) is equal to $\delta_0/2$, and therefore, it remains to find n_2 such that $\varepsilon \ge \sqrt{2(n-1)^{-1}\log(4/\delta_0)} + 2(n-1)^{-1}\log(4/\delta_0)$ for all $n \ge n_2$. With some arithmetic manipulation, we get $n_2 = (2\varepsilon^2)^{-1}\left(1+\sqrt{1+4\varepsilon}\right)^2 \log(4/\delta_0)$. Hence, for any n larger than $\max\{n_1, n_2\}$, the result follows. □

3.2 Exponential Distribution

Assume that the data-generating distribution is an exponential distribution with rate parameter $\lambda > 0$, i.e., $p_x = \exp(-\lambda x)$ for all $x \ge 0$ and $\mathbb{P}(X > x) = 1$ for all $x < 0$. We use the estimator $\hat{p}_n^{\texttt{Exp}} = \exp(-\hat{\lambda}_n x)$ where $\hat{\lambda}_n := n/\sum_i X_i$ is the maximum likelihood estimator of λ.

Proposition 3. Assume that X follows an exponential distribution with unknown rate parameter λ. For any $\delta, \delta_0 \in (0,1)$, when the sample size n satisfies

$$n \ge \frac{1}{2}\left(\frac{1+\sqrt{1+2\log(1+\delta)/\log(1/(p_x(1+\delta)))}}{\log(1+\delta)/\log(1/(p_x(1+\delta))}\right)^2 \log\left(\frac{2}{\delta_0}\right) \overset{x\to\infty}{\sim} \frac{2\log^2(1/(p_x(1+\delta)))}{\log^2(1+\delta)} \log\left(\frac{2}{\delta_0}\right),$$

we have $\mathbb{P}(|\hat{p}_n^{\texttt{Exp}}/p_x - 1| > \delta) \le \delta_0$ for all sufficiently large x.

Proof. By the monotonicity of p_x in λ, it can be easily shown that $1/\hat{\lambda}_n > (1+\varepsilon)/\lambda$ if and only if $\hat{p}_n^{\texttt{Exp}} = e^{-\hat{\lambda}_n x} > e^{-\lambda x/(1+\varepsilon)} = p_x^{1/(1+\varepsilon)}$. Fix $\varepsilon = \log(1+\delta)/\log(1/(p_x(1+\delta)))$. Then, $p_x^{1/(1+\varepsilon)} = p_x(1+\delta)$, and thus, $1/\hat{\lambda}_n > (1+\varepsilon)/\lambda$ is equivalent to $\hat{p}_n^{\texttt{Exp}}/p_x > 1+\delta$. Similarly, we have $1/\hat{\lambda}_n < (1-\varepsilon)/\lambda$ if and only if $\hat{p}_n^{\texttt{Exp}} < p_x^{1/(1-\varepsilon)}$, and moreover, it is easy to check that $p_x^{1/(1-\varepsilon)} > p_x(1-\delta)$. Thus, we obtain that $1/\hat{\lambda}_n < (1-\varepsilon)/\lambda$ if $\hat{p}_n^{\texttt{Exp}}/p_x < 1-\delta$. Therefore, $\mathbb{P}(|\hat{p}_n^{\texttt{Exp}}/p_x - 1| > \delta) \le \mathbb{P}(|1/\hat{\lambda}_n - 1/\lambda| > \varepsilon/\lambda)$.

Next, it is well known that $1/\hat{\lambda}_n$ follows a gamma distribution with the shape parameter n and the rate parameter $n\lambda$. Thus, by applying Theorem 2.3 in Boucheron et al. (2013), we get

$$\mathbb{P}\left(|1/\hat{\lambda}_n - 1/\lambda| > (\sqrt{2nt}+t)/(n\lambda)\right) \le 2e^{-t} \text{ for any } t > 0. \tag{8}$$

Hence, by substituting $t = \log(2/\delta_0)$ in (8) and using the above argument, we have $\mathbb{P}(|\hat{p}_n^{\texttt{Exp}}/p_x - 1| > \delta) \le \delta_0$ if $\sqrt{2n\log(2/\delta_0)} + \log(2/\delta_0) \le n\varepsilon$. This condition can be satisfied when n satisfies

$$n \ge \frac{1}{2}\left(\frac{1+\sqrt{1+2\varepsilon}}{\varepsilon}\right)^2 \log\left(\frac{2}{\delta_0}\right),$$

substituting the value of ε completes the proof. □

3.3 Pareto Distribution

Suppose now that samples are generated from a Pareto distribution with parameters $\xi, u > 0$ such that $p_x = (u/x)^{1/\xi}$ for all $x \ge u$ and $p_x = 1$ for all $x < u$. We first consider the case where u is known. Then, we construct an estimator for p_x as $\hat{p}_n^{\texttt{P1}} := (u/x)^{1/\hat{\xi}_n}$, where $\hat{\xi}_n := n^{-1}\sum_{i=1}^n \log(X_i/u)$ is the maximum likelihood estimator for the parameter ξ.

Proposition 4. Assume that X follows a Pareto distribution with unknown ξ and known u. For any $\delta, \delta_0 \in (0,1)$, when the sample size n satisfies

$$n \ge \frac{1}{2}\left(\frac{1+\sqrt{1+2\log(1+\delta)/\log(1/(p_x(1+\delta)))}}{\log(1+\delta)/\log(1/(p_x(1+\delta))}\right)^2 \log\left(\frac{2}{\delta_0}\right) \overset{x\to\infty}{\sim} \frac{2\log^2(1/(p_x(1+\delta)))}{\log^2(1+\delta)} \log\left(\frac{2}{\delta_0}\right),$$

then $\mathbb{P}(|\hat{p}_n^{\mathsf{P1}}/p_x - 1| > \delta) \le \delta_0$ for all sufficiently large x.

Proof. Since p_x increases in ξ, it can be easily shown that $\hat{\xi}_n > \xi(1+\varepsilon)$ if and only if $\hat{p}_n^{\mathsf{P1}} > p_x^{1/(1+\varepsilon)}$. Fix $\varepsilon = \log(1+\delta)/\log(1/(p_x(1+\delta)))$. Then, $p_x^{1/(1+\varepsilon)} = p_x(1+\delta)$, and thus, $\hat{\xi}_n > \xi(1+\varepsilon)$ is equivalent to $\hat{p}_n^{\mathsf{P1}}/p_x > 1+\delta$. Similarly, we have $\hat{\lambda}_n < \xi(1-\varepsilon)$ if and only if $\hat{p}_n^{\mathsf{P1}} < p_x^{1/(1-\varepsilon)}$, and moreover, it is easy to check that $p_x^{1/(1-\varepsilon)} > p_x(1-\delta)$. Thus, we obtain that $\hat{\xi}_n < \xi(1-\varepsilon)$ if $\hat{p}_n^{\mathsf{P1}}/p_x < 1-\delta$. Therefore, $\mathbb{P}(|\hat{p}_n^{\mathsf{P1}}/p_x - 1| > \delta) \le \mathbb{P}(|\hat{\xi}_n - \xi| > \xi\varepsilon)$.

It can be easily shown that $\log(X/u)$ follows an exponential distribution with rate $1/\xi$, implying that $\hat{\xi}_n$ follows a gamma distribution with the shape parameter n and the rate parameter n/ξ. Thus, by replacing $1/\hat{\lambda}_n$ and $1/\lambda$ with $\hat{\xi}_n$ and ξ, respectively, we have

$$\mathbb{P}\left(|\hat{\xi} - \xi| > \xi(\sqrt{2nt} + t)/n\right) \le 2e^{-t} \text{ for any } t > 0.$$

Hence, using the same arguments as in the proof of Proposition 3, we obtain the desired result. □

When both u and ξ are unknown, their maximum likelihood estimators are given by $\hat{u}_n = \min_{i=1,\dots,n} X_i$ and $\hat{\xi}_n' = n^{-1}\sum_{i=1}^n \log(X_i/\hat{u}_n)$, respectively. Accordingly, we define an estimator for p_x as $\hat{p}_n^{\mathsf{P2}} := (\hat{u}_n/x)^{1/\hat{\xi}_n'}$.

Proposition 5. Assume that X follows a Pareto distribution with unknown ξ and a. We define $\varepsilon_{x,\delta} = \log(1+\delta)/(\log(1/p_x) + 1 - \log(1+\delta))$. Then, for any $\delta, \delta_0 \in (0,1)$, when the sample size n satisfies

$$n \ge \max\left\{\frac{2\log(3/\delta_0)}{\varepsilon_{x,\delta}^2/(1+\varepsilon_{x,\delta})}, \frac{\xi\log(3/\delta_0)}{\log(1+\varepsilon_{x,\delta})}\right\} \overset{x\to\infty}{\sim} \max\left\{\frac{2\log^2(1/p_x)}{\log^2(1+\delta)}, \frac{\xi\log(1/p_x)}{\log(1+\delta)}\right\}\log\left(\frac{3}{\delta_0}\right),$$

then $\mathbb{P}(|\hat{p}_n^{\mathsf{P2}}/p_x - 1| > \delta) \le \delta_0$ for all sufficiently large x.

Proof.

According to Malik (1970), we know that (i) $\hat{\xi}_n'$ and $\hat{u}_n$ are independent; (ii) $\hat{\xi}_n'$ follows a gamma distribution with the shape parameter $n-1$ and the rate parameter n/ξ; and (iii) $\hat{u}_n$ has a Pareto distribution such that $\mathbb{P}(\hat{u}_n > x) = (u/x)^{n/\xi}$. Thus, by applying Theorem 2.3 in Boucheron et al. (2013), we get

$$\mathbb{P}\left(|\hat{\xi}_n' - \xi| > \xi(\sqrt{2(n-1)t} + t)/n\right) \le 2e^{-t}. \tag{9}$$

Also, we obtain that for any $\varepsilon > 0$,

$$\mathbb{P}(|\hat{u}_n - u| > u\varepsilon) = \mathbb{P}\big(\hat{u}_n \ge u(1+\varepsilon)\big) = (1+\varepsilon)^{-n/\xi}. \tag{10}$$

Then, this proof basically follows the proof of Proposition 2. Specifically, for all $\varepsilon > 0$, the probability $\mathbb{P}(|\hat{p}_n^{\mathsf{P2}}/p_x - 1| > \delta)$ is bounded from above by

$$\begin{aligned}&\mathbb{P}\left(\frac{\hat{p}_n^{\mathsf{P2}}}{p_x} > 1+\delta, |\hat{\xi}_n' - \xi| \le \xi\varepsilon, |\hat{u}_n - u| \le u\varepsilon\right) + \mathbb{P}\left(\frac{\hat{p}_n^{\mathsf{P2}}}{p_x} < 1-\delta, |\hat{\xi}_n' - \xi| \le \xi\varepsilon, |\hat{u}_n - u| \le u\varepsilon\right)\\ &+ \mathbb{P}\left(|\hat{\xi}_n' - \xi| > \xi\varepsilon\right) + \mathbb{P}\left(|\hat{u}_n - u| > u\varepsilon\right)\end{aligned} \tag{11}$$

We define $p_x^{\xi(1+\varepsilon_1),u(1+\varepsilon_2)} = p_x^{1/(1+\varepsilon_1)}(1+\varepsilon_2)^{1/(1+\varepsilon_1)}$ for any $\varepsilon_1 > 0$ and for any $\varepsilon_2 > 0$ such that $p_x(1+\varepsilon_2) < 1$. This value increases in both ε_1 and ε_2. Then, the first and second terms in (11) are zero if $p_x^{\xi(1+\varepsilon),u(1+\varepsilon)} < (1+\delta)p_x$ and $p_x^{\xi(1-\varepsilon),u(1-\varepsilon)} > (1-\delta)p_x$. Let $\varepsilon = \varepsilon_{x,\delta}$ defined in the theorem. Then, all the above conditions are satisfied for large x since $\log(1+\varepsilon) < \varepsilon$, $\log(1-\varepsilon) > -\varepsilon$, and $\log(1+\delta) < -\log(1-\delta)$.

Finally, it remains to achieve $\mathbb{P}(|\hat{\xi}_n' - \xi| > \xi\varepsilon) \le 2\delta_0/3$ and $\mathbb{P}(|\hat{u}_n - u| > u\varepsilon) \le \delta_0/3$. For the former, by plugging $t = \log(3/\delta_0)$ in (9), it suffices to find n_1 such that $\sqrt{2(n-1)\log(3/\delta_0)} + \log(3/\delta_0) \le \varepsilon n$ for all $n \ge n_1$. One can easily check that it is satisfied by the following choice of $n_1 = 2(1+\varepsilon)\log(3/\delta_0)/\varepsilon^2$. For the latter, by (10), we need $n \ge n_2 := \xi\log(3/\delta_0)/\log(1+\varepsilon)$. Therefore, the result follows for any n larger than $\max\{n_1, n_2\}$. □

3.4 Generalized Pareto Distribution

We now assume that a random variable X follows a Generalized Pareto distribution (GPD) with parameters $x_0 \in \mathbb{R}$ and $\xi, \beta > 0$, whose tail distribution is defined as

$$G_{x_0,\xi,\beta}(x) := \mathbb{P}(X > x) = \left(1 + \xi \frac{(x - x_0)}{\beta}\right)^{-1/\xi}, \tag{12}$$

where the support of X is $x \geq x_0$ and x_0 is assumed to be known throughout this subsection. The Hill estimator is one of the widely studied estimators for the parameter ξ, which uses the top k order statistics $X_{\cdot,n}$. It is formally given by $\hat{\xi}(k) := \frac{1}{k}\sum_{j=1}^{k} \log\left(\frac{X_{j,n}}{X_{k+1,n}}\right)$. As noted in Boucheron and Thomas (2015), the variance of the Hill estimator scales as ξ^2/k, while its bias is upper bounded by a von Mises function that increases with k. Consequently, a larger k reduces variance but simultaneously increases bias, leading to a bias-variance tradeoff. To strike the balance between them, the adaptive Hill estimator is proposed in Boucheron and Thomas (2015), which selects $k = \hat{k}_n$ given by:

$$\hat{k}_n = \max\left\{k \in \{\ell_n, \ldots, n\} : \forall i \in \{\ell_n, \ldots, n\},\ |\hat{\xi}(i) - \hat{\xi}(k)| \leq \frac{\hat{\xi}(i) r_n(\delta_0)}{\sqrt{i}}\right\}, \tag{13}$$

where $\ell_n \asymp \lceil 2\log n \rceil$ and $r_n(\delta_0) \asymp \sqrt{\log((2/\delta_0)\log n)}$. Then, the adaptive Hill estimator, defined as $\hat{\xi}(\hat{k}_n)$, is shown to achieve the minimax lower bound across distributions satisfying the von Mises condition, which is formalized in the following result. Interested readers are also referred to Boucheron and Thomas (2015) for its superior performance over existing tail index estimators.

Lemma 6 (Corollary 3.12 of Boucheron and Thomas (2015))**.** Consider the adaptive Hill Estimator $\hat{\xi}(\hat{k}_n)$ with n independent and identically distributed (i.i.d.) samples from a certain distribution. Let $\bar{\eta}(t) := \sup_{s \geq t} |\eta(s)|$, where $\eta(\cdot)$ is the von Mises function of the distribution. Assume that there exist $C, \xi > 0$ and $t_0 \geq 1$ satisfying $\bar{\eta}(t) \leq C t^{-\xi}\ \forall t > t_0$. Then, the following holds for sufficiently large n and $\delta_0 \in (0,2)$:

$$\mathbb{P}\left(|\hat{\xi}(\hat{k}_n)/\xi - 1| > \kappa_{\delta_0}\left(\frac{\log((16/\delta_0)\log n)}{n}\right)^{\xi/(1+2\xi)}\right) \leq \frac{\delta_0}{2}$$

where κ_{δ_0} is a constant dependent on δ_0 and the distribution parameters.

We note that if the underlying distribution is a GPD with the tail distribution in (12), then we have $\bar{\eta}(t) = \xi(1 - x_0\xi/\beta)/(t^{\xi} + x_0\xi/\beta - 1) \leq C(t_0)t^{-\xi}\ \forall t > t_0$ for some constants $C > 0$ and $t_0 \geq 1$. Hence, the result in the above lemma holds in this case.

We now focus on the estimation of the parameter β in (12). We particularly use the Method of Moments estimator (MME) $\hat{\beta}_n = (\hat{S}_n - x_0)(1 - \hat{\xi}_n'')$, where $\hat{S}_n$ is the median-of-means (MoM) estimator defined as $\hat{S}_n := \texttt{med}(|\mathscr{B}_1|^{-1}\sum_{i\in\mathscr{B}_1} X_i, \ldots, |\mathscr{B}_m|^{-1}\sum_{i\in\mathscr{B}_m} X_i)$, where m is the number of sample blocks with with each block containing at least $\lfloor n/m \rfloor$ samples, $\texttt{med}(\texttt{a}_1, \ldots, \texttt{a}_\texttt{m})$ denotes the median of $a_1, \ldots, a_m$, and we write $\hat{\xi}_n'' = \hat{\xi}(\hat{k}_n)$ for brevity. The next lemma presents a concentration inequality for the MoM estimator $\hat{S}_n$, which, in conjunction with Lemma 6, will be utilized to construct our main result for the GPD case.

Lemma 7 (Lemma 2 of Bubeck et al. (2013))**.** Let $\hat{S}_n$ denote the MoM estimator with n i.i.d. samples with $n \geq 2 + 16\log(1/\delta)$ and $m \asymp 8\log(1/\delta)$. Then, we have $\mathbb{P}\left(|\mathbb{E}[X] - \hat{S}_n| > \varepsilon\right) \leq \exp\left(-(n/8)\varepsilon^{(1+a)/a}/(12M)^{1/a}\right)$, where $M := \mathbb{E}[|X - \mathbb{E}[X]|^{1+a}] < \infty$ for $a \in (0,1]$.

Based on the estimators discussed above, we define an estimator for p_x as $\hat{p}_n^{\texttt{GPD}} := \left(1 + \hat{\xi}_n''(x - x_0)/\hat{\beta}_n\right)^{-1/\hat{\xi}_n''}$. This construction leads to our main result as follows:

Proposition 8. Assume that X follows a Generalized Pareto distribution (GPD) with a known parameter $x_0 \in \mathbb{R}$ and unknown parameters $\xi, \beta > 0$ and that $\mathbb{E}[|X - \mathbb{E}[X]|^2] < \infty$, or equivalently, $\xi < 1/2$. Then, for any $\delta, \delta_0 \in (0,1)$, $\mathbb{P}(|\hat{p}_n^{\mathtt{GPD}}/p_x - 1| > \delta) \le \delta_0$ for all sufficiently large x if the sample size n satisfies

$$n \ge \max\left\{ \kappa_1 \left(\frac{\log(1/p_x)}{\delta}\right)^{(1+2\xi)/\xi},\ \kappa_2 \left(\frac{\log(1/p_x)}{\delta}\right)^2 \right\} \log\left(\frac{9}{\delta_0}\right), \tag{14}$$

where κ_1 and κ_2 are positive constants dependent on x_0, ξ, β, and δ_0.

Proof Sketch. Let $p_{x,x_0}^{\tilde{\xi},\tilde{\beta}} := \left(1 + \tilde{\xi}(x - x_0)/\tilde{\beta}\right)^{-1/\tilde{\xi}}$ for $\tilde{\xi}, \tilde{\beta} > 0$. Note that $p_x = p_{x,x_0}^{\xi,\beta}$ and $\hat{p}_n^{\mathtt{GPD}} = p_{x,x_0}^{\hat{\xi}_n'',\hat{\beta}_n}$. Then, by the first-order Taylor approximation, we get

$$p_{x,x_0}^{\tilde{\xi},\tilde{\beta}}/p_x - 1 \approx \left(\log(1/p_x) - \frac{1 - (p_x)^{\xi}}{\xi}\right)\left(\tilde{\xi}/\xi - 1\right) + \frac{1 - (p_x)^{\xi}}{\xi}\left(\tilde{\beta}/\beta - 1\right).$$

The rest of this proof sketch is built upon the assumption that the above approximation is exact.

Let $\varepsilon_x = \delta/\log(1/p_x)$. Then, using the same argument as in the proofs of Propositions 2 and 5, it is easy to see that

$$\mathbb{P}(|\hat{p}_n^{\mathtt{GPD}}/p_x - 1| > \delta, |\hat{\xi}_n''/\xi - 1| \le \varepsilon_x, |\hat{\beta}_n/\beta - 1| \le \varepsilon_x) = 0.$$

Hence, it remains to achieve $\mathbb{P}(|\hat{\xi}_n''/\xi - 1| > \varepsilon_x) \le 8\delta_0/9$ and $\mathbb{P}(|\hat{\xi}_n''/\xi - 1| \le \varepsilon_x, |\hat{\beta}_n/\beta - 1| > \varepsilon_x) \le \delta_0/9$. By Lemma 6, the former is straightforward if

$$\frac{n}{\log n} \ge n_1 := \left(\frac{\kappa_{\delta_0}}{\varepsilon_x}\right)^{(1+2\xi)/\xi} \log\left(\frac{9}{\delta_0}\right). \tag{15}$$

For the latter, recall that $\hat{\beta}_n = (\hat{S}_n - x_0)(1 - \hat{\xi}_n'')$ and $\hat{S}_n \ge x_0$ almost surely. Also, since X has finite variance, we have $\xi < 1/2$, and thus, $0 < \xi(1-\varepsilon_x),\ \xi(1+\varepsilon_x) < 1$ for all sufficiently large x. Then, we observe

$$\begin{aligned}
&\mathbb{P}(|\hat{\xi}_n''/\xi - 1| \le \varepsilon_x, |\hat{\beta}_n/\beta - 1| > \varepsilon_x) \\
&= \mathbb{P}(|\hat{\xi}_n''/\xi - 1| \le \varepsilon_x, \hat{\beta}_n/\beta > 1 + \varepsilon_x) + \mathbb{P}(|\hat{\xi}_n''/\xi - 1| \le \varepsilon_x, (\hat{S}_n - x_0)(1 - \hat{\xi}_n'')/\beta < 1 - \varepsilon_x) \\
&\le \mathbb{P}((\hat{S}_n - x_0)(1 - \xi + \xi\varepsilon_x)/\beta > 1 + \varepsilon_x) + \mathbb{P}((\hat{S}_n - x_0)(1 - \xi - \xi\varepsilon_x)/\beta < 1 - \varepsilon_x) \\
&= \mathbb{P}\left(\hat{S}_n - \mathbb{E}[X] > \frac{(\mathbb{E}[X] - x_0)(1-2\xi)\varepsilon_x}{1 - \xi + \xi\varepsilon_x}\right) + \mathbb{P}\left(\hat{S}_n - \mathbb{E}[X] < -\frac{(\mathbb{E}[X] - x_0)(1-2\xi)\varepsilon_x}{1 - \xi - \xi\varepsilon_x}\right) \\
&\le \mathbb{P}\left(|\hat{S}_n - \mathbb{E}[X]| > (\mathbb{E}[X] - x_0)(1-2\xi)\varepsilon_x\right),
\end{aligned}$$

where the second equality holds since $\beta = (\mathbb{E}[X] - x_0)(1 - \xi)$. Finally, by applying Lemma 7, the above probability is bounded by $\delta_0/9$ if

$$n \ge n_2 := \{96M\log(9/\delta_0)\}/\{(\mathbb{E}[X] - x_0)^2(1-2\xi)^2\varepsilon_x^2\}.$$

Thus, the result follows for any n larger than $\max\{n_1, n_2\}$, ignoring the logarithmic term $\log n$ in (15). □

4 FRÉCHET MAXIMUM DOMAIN OF ATTRACTION

In this section, we consider a situation where the parametric representation of the data-generating distribution F is unknown but it is known to be in the maximum domain of attraction (MDA) of the Fréchet distribution. To be more specific, there exist sequences of constants $\{d_n\}$ and $\{c_n\}$ such that $c_n > 0$ for all n and $\lim_{n\to\infty} F^n(c_n x + d_n) = H_\xi(x) := \exp(-(1+\xi x)^{-1/\xi})$ for all $x \in \mathbb{R}$ for $\xi > 0$. In this case, we say that

$F \in \mathsf{MDA}(H_\xi)$ with $\xi > 0$. To tackle the issue of distributional uncertainty, we use the peak-over-threshold (POT) approach in extreme value theory (McNeil et al. 2015, Chapter 5). This approach is also introduced in Bai et al. (2022). In their work, the authors conduct several numerical experiments to validate this approach. In contrast, we aim to theoretically analyze the sample complexity based on the POT approach.

In what follows, we assume that there exist $v, \xi, \beta(\cdot)$ satisfying $F_v(x) := \mathbb{P}(X > x - v | X > v) = G_{v,\xi,\beta(v)}(x)$ for all $x \in \mathbb{R}$. This assumption asymptotically holds for all $F \in \mathsf{MDA}(H_\xi)$ with $\xi > 0$. Then, it follows that $p_x = \bar{F}(u) G_{u,\xi,\beta(u)}(x)$ for any $u > v$ and $x \in \mathbb{R}$. This allows us to construct an estimator for p_x as $\hat{p}_n^{\mathtt{POT}} = (n_u/n)\hat{p}_{n_u}^{\mathtt{G}}$, where

$$\hat{p}_{n_u}^{\mathtt{G}} := \left(1 + \hat{\xi}_{n_u} \frac{(x-u)}{\hat{\beta}_{n_u}}\right)^{-1/\hat{\xi}_{n_u}}, \quad \hat{\xi}_{n_u} = \frac{1}{\hat{k}_{n_u}} \sum_{i=1}^{\hat{k}_{n_u}} \log\left(\frac{X_{i,n}}{X_{\hat{k}_{n_u}+1,n}}\right), \quad \hat{\beta}_{n_u} = \left(\hat{S}_{n_u} - u\right)\left(1 - \hat{\xi}_{n_u}\right),$$

n_u is the number of samples exceeding u, $\hat{k}_{n_u}$ is defined as in (13) with n replaced by n_u, and $\hat{S}_{n_u}$ is the MoM estimator based on the samples exceeding u. Let $p_x^{\mathtt{G}} := G_{u,\xi,\beta(u)}(x) = p_x/\bar{F}(u)$ for all $x \in \mathbb{R}$.

Proposition 9. Assume that X follows a distribution $F \in \mathsf{MDA}(H_\xi)$ with $\xi > 0$, which satisfies the von Mises condition (GPD) with a known parameter $\mu \in \mathbb{R}$ and unknown parameters $\xi, \beta > 0$ and that $\mathbb{E}[|X - \mathbb{E}[X]|^2 | X > v] < \infty$. Then, for any $\delta, \delta_0 \in (0,1)$, $\mathbb{P}(|\hat{p}_n^{\mathtt{POT}}/p_x - 1| > \delta) \le \delta_0$ for all sufficiently large x if the sample size n satisfies

$$n \ge \max\left\{\tilde{\kappa}_1 \left(\frac{\log(\bar{F}(u)/p_x)}{\delta}\right)^{(1+2\xi)/\xi}, \ \tilde{\kappa}_2 \left(\frac{\log(\bar{F}(u)/p_x)}{\delta}\right)^2\right\} \frac{\log(11/\delta_0)}{\bar{F}(u)}, \tag{16}$$

where $\tilde{\kappa}_1$ and $\tilde{\kappa}_2$ are positive constants dependent on μ, ξ, β, and δ_0.

Proof sketch. We first observe that

$$|\hat{p}_n^{\mathtt{POT}}/p_x - 1| = \frac{|(n_u/n)\hat{p}_{n_u}^{\mathtt{G}} - \bar{F}(u)p_x^{\mathtt{G}}|}{\bar{F}(u)p_x^{\mathtt{G}}} \le \left|\hat{p}_{n_u}^{\mathtt{G}}/p_x^{\mathtt{G}} - 1\right| + \left|\frac{n_u/n}{\bar{F}(u)} - 1\right| \left|\hat{p}_{n_u}^{\mathtt{G}}/p_x^{\mathtt{G}} - 1\right| + \left|\frac{n_u/n}{\bar{F}(u)} - 1\right|.$$

Hence, for any $\varepsilon > 0$, if $|n_u/n - \bar{F}(u)| < \varepsilon \bar{F}(u)$, then the right-hand side is bounded from above by $(1+\varepsilon)|\hat{p}_{n_u}^{\mathtt{G}}/p_x^{\mathtt{G}} - 1| + \varepsilon$. Accordingly, we have

$$\mathbb{P}\left(|\hat{p}_n^{\mathtt{POT}}/p_x - 1| > \delta\right) \le \mathbb{P}\left(\left|\frac{n_u/n}{\bar{F}(u)} - 1\right| > \varepsilon\right) + \mathbb{P}\left(|\hat{p}_{n_u}^{\mathtt{G}}/p_x^{\mathtt{G}} - 1| > \frac{\delta - \varepsilon}{1+\varepsilon}\right). \tag{17}$$

We aim to bound the first term on the right-hand side by $2\delta_0/11$ and the second term by $9\delta_0/11$. Let $\varepsilon = \delta/\log(1/p_x^{\mathtt{G}})$. Then, by Section 2, $\mathbb{P}(|(n_u/n)/\bar{F}(u) - 1| > \varepsilon) \le 2\delta_0/11$ if

$$n \ge \frac{3}{\bar{F}(u)} \left(\frac{\log(1/p_x^{\mathtt{G}})}{\delta}\right)^2 \log\left(\frac{11}{\delta_0}\right).$$

Furthermore, by Proposition 8, the last term in (17) is bounded by $9\delta_0/11$ if for all sufficiently large x,

$$n_u \approx n\bar{F}(u) \ge \max\left\{\kappa_1' \left(\frac{\log(1/p_x^{\mathtt{G}})}{\delta}\right)^{(1+2\xi)/\xi}, \ \kappa_2' \left(\frac{\log(1/p_x^{\mathtt{G}})}{\delta}\right)^2\right\} \log\left(\frac{11}{\delta_0}\right),$$

where κ_1' and κ_2' are positive constants dependent on μ, ξ, β, and δ_0. Setting $\tilde{\kappa}_1 = \kappa_1'$, $\tilde{\kappa}_2 = \max\{\kappa_2, 3\}$ and using $\log(\bar{F}(u))/\log(1/p_x) \approx 0$ leads to the desired result. □

5 NUMERICAL RESULTS

We analyze the behavior of the lower bound on sample complexity that we derived in earlier sections, as p_x decreases, and compare it to the lower bound that follows from simulation experiments. Theoretical lower bounds are based on concentration inequalities and may be higher than the sample size required in the experiments if these inequalities are not tight.

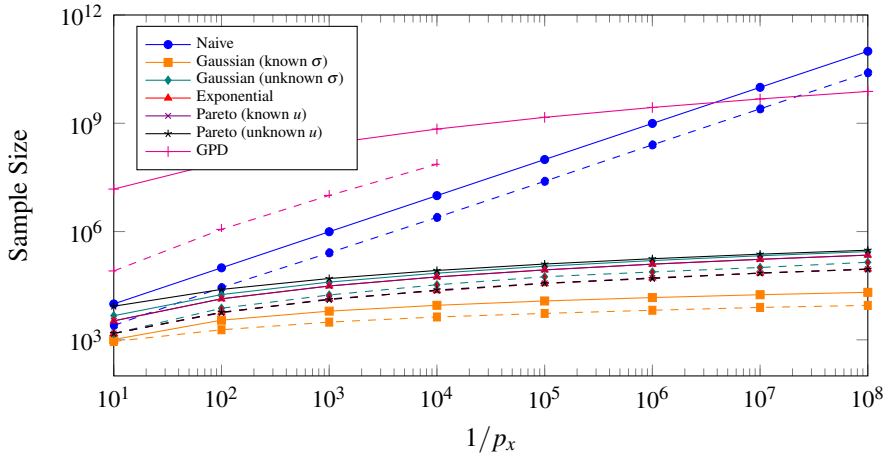
Simulation experiments are conducted as follows. We generate n samples from a given distribution and estimate the underlying parameters. The tail probability p_x is estimated by plugging in the estimated parameters in the tail CDF. This is compared with the true probability p_x, and we estimate the fraction of times the relative error in the probability estimate is within δ, by repeating this experiment 10,000 times; exceptionally for the Generalized Pareto Distribution (GPD), we repeat it 1,000 times due to a significant computational burden. We vary the number of samples n from the given distribution until the acceptable relative error fraction is equal to $1-\delta_0$. The results are shown in Figure 1a as dashed lines. Here, we have used $\texttt{Gaussian}(\mu=0,\sigma=0.5), \texttt{Exponential}(\lambda=0.5), \texttt{Pareto}(\xi=0.5,u=1), \texttt{GPD}(\xi=0.5,\beta=1,x_0=0)$ distributions. For the Type 1 Pareto distribution, the lower bounds in the cases where u is known and unknown tend to match almost exactly. This is because the threshold u can be estimated with relative ease—most of the probability mass is concentrated near u, making its estimation stable. For the GPD, significantly more samples are required to accurately estimate the probabilities of the low tail (10^{-1} to 10^{-4}). This is primarily due to the need for sufficient samples to obtain a stable estimate of the tail index (via the Hill estimator). In Figure 1a, the theoretical lower bounds are plotted as solid lines for the parametric estimators defined in Section 3 for the same distributions. To obtain precise lower bounds, we numerically solve some steps in the derivation instead of relying on approximations. We observe that for higher values of p_x (of order 10^{-6}), the naive estimator outperforms the estimators for GPD. Observe that the theoretical values are always above the experimental values; this is due to loose concentration inequalities and non-sharp analysis.

Next, in Figure 1b, we analyze the effectiveness of the approach in Section 4. We assume that the underlying distribution is $\texttt{GPD}(\xi=0.5,\beta=1,x_0=3)$. The red curve with square markers represents the lower bound in (14), whereas the other plots exhibit the lower bound in (16) under varying thresholds $u=x^{\alpha}$ with $\alpha\in\{0.1,0.3,0.5,0.7\}$. Figure 1b suggests that when the underlying distribution is GPD, the lower the threshold, the lower the sample complexity. To see if this observation is consistent across different underlying distributions, we separately conducted a similar experiment where the underlying data density is proportional to $(\log(1+x)/(1+x))^3$, $x\geq 0$, and we fit a GPD distribution to the data above varying thresholds of u. While the corresponding numerical results are not presented in this paper due to space constraints, we find that for a probability of about 10^{-6}, even after generating over 10^9 samples, our estimator based on fitting the GPD above the threshold does not lie within 10% of the true probability for a comprehensive set of thresholds. Both simulations and our theoretical analysis suggest that EVT-based estimators could show performance degradation compared to the naive estimator, depending on problem instances and the choice of thresholds u. A detailed theoretical investigation into this issue is left for future research. Further numerical experiments and their corresponding results are provided in an online appendix, available on the authors' websites.

6 CONCLUSION

In this paper, we studied the classical problem of estimating a rare event probability of a single random variable exceeding a large threshold in a data-driven manner. Naive empirical estimators are distributionally robust; however, they can have high sample complexity, which motivates faster approximate methods. We observed that under the assumption that the underlying random variable belongs to a known family, for many commonly occurring distribution families, lower bounds on sample complexity can be developed based on standard parameter estimators. While these approaches can lead to lower sample complexity compared to the naive estimator, the corresponding results are very sensitive to the underlying distribution.

(a) Sample complexities from theoretical analyses (solid) and simulation (dashed).



(b) Sample complexities based on different values of the threshold u in the POT approach.

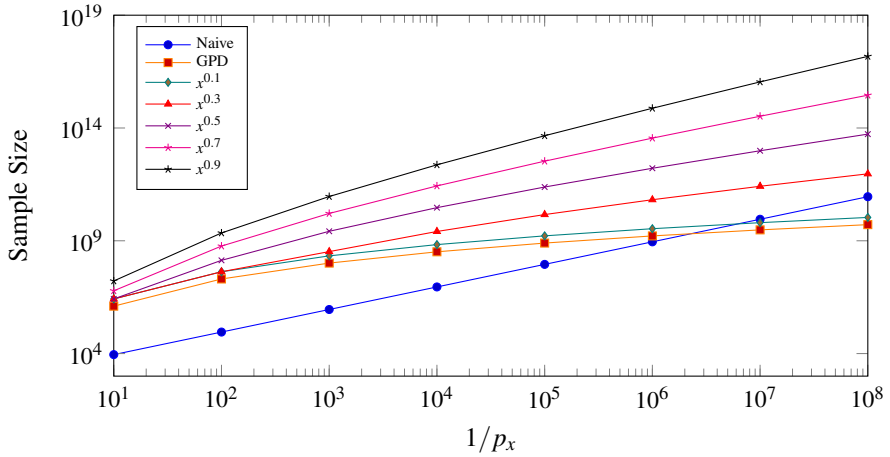


Figure 1: Sample complexity is plotted as a function of $1/p_x$ in a log-log scale.

For instance, for a given dataset with sample mean 10 and variance 2, when we fit a Gaussian distribution to the data, the tail probability of exceeding a threshold of 16.3 is roughly 4.2×10^{-6}, while fitting a GPD to the same data gives a probability of exceeding that threshold about 1,000 times higher, underscoring the need to be extremely careful in selecting the underlying distribution family for tail probability estimation. Extreme value theory offers a principled way to arrive at tail probability estimates. We, however, observe that it may be substantially less effective compared to a naive estimator for probabilities even as low in order as one in a million. This observation may have sobering implications on the use of probabilistic methods in the estimation of rare events.

REFERENCES

Ahn, D., and L. Zheng. 2025. "Efficient Simulation of Polyhedral Expectations with Applications to Finance". *Mathematics of Operations Research* Articles in Advance https://doi.org/10.1287/moor.2023.0145.

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*. New York, NY: Springer.

Bai, Y., Z. Huang, H. Lam, and D. Zhao. 2022. "Rare-Event Simulation for Neural Network and Random Forest Predictors". *ACM Transactions on Modeling and Computer Simulation* 32(3):Article 18.

Bai, Y., H. Lam, and S. Engelke. 2022. "Rare-Event Simulation Without Variance Reduction: An Extreme Value Theory Approach". In *2022 Winter Simulation Conference (WSC)*, 133–144 https://doi.org/10.1109/WSC57314.2022.10015403.

Botev, Z. I., A. Ridder, and L. Rojas-Nandayapa. 2016. "Semiparametric Cross Entropy for Rare-Event Simulation". *Journal of Applied Probability* 53(3):633–649.

Boucheron, S., G. Lugosi, and P. Massart. 2013. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, UK: Oxford University Press.

Boucheron, S., and M. Thomas. 2015. "Tail Index Estimation, Concentration and Adaptivity". *Electronic Journal of Statistics* 9(2):2751 – 2792.

Bubeck, S., N. Cesa-Bianchi, and G. Lugosi. 2013. "Bandits with Heavy Tail". *IEEE Transactions on Information Theory* 59(11):7711–7717.

de Vries, I., S. Sippel, J. Zeder, E. Fischer, and R. Knutti. 2024. "Increasing Extreme Precipitation Variability Plays a Key Role in Future Record-Shattering Event Probability". *Communications Earth & Environment* 5(1):482.

Dean, T., and P. Dupuis. 2009. "Splitting for Rare Event Simulation: A Large Deviation Approach to Design and Analysis". *Stochastic Processes and their Applications* 119(2):562–587.

Deo, A., and K. Murthy. 2025. "Achieving Efficiency in Black-Box Simulation of Distribution Tails with Self-Structuring Importance Samplers". *Operations Research* 73(1):325–343.

Glasserman, P. 2003. *Monte Carlo Methods in Financial Engineering*. New York, NY: Springer.

Glasserman, P., P. Heidelberger, and P. Shahabuddin. 2000. "Variance Reduction Techniques for Estimating Value-at-Risk". *Management Science* 46(10):1349–1364.

Juneja, S., and P. Shahabuddin. 2002. "Simulating Heavy Tailed Processes Using Delayed Hazard Rate Twisting". *ACM Transactions on Modeling and Computer Simulation* 12(2):94–118.

Laurent, B., and P. Massart. 2000. "Adaptive Estimation of a Quadratic Functional by Model Selection". *The Annals of Statistics* 28(5):1302–1338.

L'Ecuyer, P., V. Demers, and B. Tuffin. 2006. "Splitting for Rare-Event Simulation". In *2006 Winter Simulation Conference (WSC)*, 137–148 https://doi.org/10.1109/WSC.2006.323046.

Malik, H. J. 1970. "Estimation of the Parameters of the Pareto Distribution". *Metrika* 15(1):126–132.

McNeil, A. J., R. Frey, and P. Embrechts. 2015. *Quantitative Risk Management: Concepts, Techniques and Tools*. Revised ed. Princeton, NJ: Princeton University Press.

Rubino, G., and B. Tuffin. 2009. *Rare Event Simulation using Monte Carlo Methods*. Chichester, UK: John Wiley & Sons, Ltd.

AUTHOR BIOGRAPHIES

DOHYUN AHN is an Associate Professor in the Department of Systems Engineering and Engineering Management at the Chinese University of Hong Kong. He received his B.S., M.S., and Ph.D. degrees in Industrial & Systems Engineering, all from KAIST. His research interests include, but are not limited to, quantitative risk management, rare-event simulation, optimization via simulation, decision-making under model risk, and analysis of network effects in finance and operations. His email address is dohyun.ahn@cuhk.edu.hk and his website is https://sites.google.com/view/dohyun/.

SANDEEP JUNEJA is a Professor of computer science and the founding director for the Safexpress Centre for Data, Learning and Decision Sciences at Ashoka University. Formerly, he was a senior professor at the School of Technology and Computer Science (STCS) at TIFR Mumbai. His research interests lie in applied probability including in sequential learning, mathematical finance, Monte Carlo methods, game theoretic analysis of queues, epidemiological modeling as well as in use of machine learning techniques for monsoon weather modelling in India. He is currently the area editor for Operations Research in simulation. sandeep.juneja@ashoka.edu.in; https://www.tcs.tifr.res.in/~sandeepj/.

TEJAS PAGARE is an AI Researcher at the Honda R&D Lab in Tokyo. He obtained his B.Tech and M.Tech degrees in Electrical Engineering from Indian Institute of Technology Bombay. His research interest is in decision making under uncertainty with applications to economics in general. His email address is tejaspagare2002@gmail.com and his website is https://tejassp2002.github.io/.

SHREYAS SAMUDRA is a Research Associate at the Safexpress Centre for Data, Learning and Decision Sciences at Ashoka University. He completed his Bachelor's and Master's degrees in Aerospace Engineering, specializing in Applied Mechanics and Engineering Systems Reliability, from the Indian Institute of Technology, Kharagpur. His email address is sssamudra98@gmail.com