OPTIMAL NONPARAMETRIC ESTIMATION OF THE EXPECTED SHORTFALL RISK

DANIEL BARTL AND STEPHAN ECKSTEIN

Abstract. We address the problem of estimating the expected shortfall risk of a financial loss using a finite number of i.i.d. data. It is well known that the classical plug-in estimator suffers from poor statistical performance when faced with (heavy-tailed) distributions that are commonly used in financial contexts. Further, it lacks robustness, as the modification of even a single data point can cause a significant distortion. We propose a novel procedure for the estimation of the expected shortfall and prove that it recovers the best possible statistical properties (dictated by the central limit theorem) under minimal assumptions and for all finite numbers of data. Further, this estimator is adversarially robust: even if a (small) proportion of the data is maliciously modified, the procedure continuous to optimally estimate the true expected shortfall risk. We demonstrate that our estimator outperforms the classical plug-in estimator through a variety of numerical experiments across a range of standard loss distributions.

1. Introduction

A central task in risk management involves accurately quantifying the level of risk associated with a given financial position. A rigorous framework addressing this question is provided by the theory of risk measures (see, e.g., [4, 12, 24]), and the most widely employed method in practice involves using the *expected shortfall* risk measure (cf. [15, 33, 34]). For a random financial loss X, the expected shortfall at the level $\alpha \in (0, \frac{1}{2})$ is defined as

$$\mathrm{ES}_\alpha(X) := \frac{1}{\alpha} \int_{1-\alpha}^{1} \mathrm{VaR}_u(X)\, du,$$

where $\mathrm{VaR}_u(X) := \inf\{t \in \mathbb{R} : \mathbb{P}(X \leq t) \geq u\}$ is the value at risk at the level $u \in [1-\alpha, 1)$ of the loss X (cf. [24, 13]).

However, calculating $\mathrm{ES}_\alpha(X)$ requires knowledge of the distribution function $F_X(t) = \mathbb{P}(X \leq t)$ of X, which is rarely known precisely. Instead, one commonly relies on *historical data*, denoted as $X_1, \ldots, X_N$. While these data may typically be subject to some mild correlation, in this article, we adopt the conventional approach found in many papers

Date: May 2, 2024.

Key words and phrases. Sub-Gaussian estimators, adversarial robustness, heavy tails, non-asymptotic statistics, risk measures.

addressing the estimation of the expected shortfall (see, e.g., [1, 7, 11, 18, 22, 42]) and assume the idealized scenario where the data are independent and identically distributed.

1.1. **The plug-in estimator.** The arguably most natural estimator for $\mathrm{ES}_\alpha(X)$ using the data $X_1, \dots, X_N$ is the *plug-in estimator*. Indeed, noting that $\mathrm{ES}_\alpha(X)$ is a quantity that only depends on the distribution function F_X of X, the plug-in estimator $\widehat{T}_N$ consists in taking the empirical distribution function $\widehat{F}_N(t) := \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{(-\infty,t]}(X_i)$ as a proxy for F_X and computing

$$\widehat{T}_N := \frac{1}{\alpha}\int_{1-\alpha}^1 \mathrm{VaR}_u\left(\widehat{F}_N^{-1}\right)\, du,$$

where $\mathrm{VaR}_u(\widehat{F}_N) = \widehat{F}_N^{-1}(u) := \inf\{t : \widehat{F}_N(t) \geq u\}$.

The central question of interest in the present context is then the extent to which the estimator $\widehat{T}_N$ approximates $\mathrm{ES}_\alpha(X)$. More accurately, since $\widehat{T}_N$ is a random quantity, one is interested in obtaining bounds on the error $\widehat{T}_N - \mathrm{ES}_\alpha(X)$ that hold with *high confidence*. As a first step towards that goal, it often turns out helpful to analyse the *asymptotic* behavior as $N \to \infty$. In fact, the latter has been studied in detail and is well understood by now. To formulate the results, denote by

$$\sigma^2_{\mathrm{ES}_\alpha(X)} := \frac{1}{\alpha^2}\int_{\mathrm{VaR}_{1-\alpha}(X)}^\infty \int_{\mathrm{VaR}_{1-\alpha}(X)}^\infty F_X(\min\{t,s\}) - F_X(t)F_X(s)\, dtds$$

the variance associated to the estimation problem of the expected shortfall. Note that $\sigma^2_{\mathrm{ES}_\alpha(X)}$ is always well-defined taking values in $[0,\infty]$, and it is finite whenever the positive part of X is square integrable. With this notation set in place, the following central limit theorem for the convergence of $\widehat{T}_N$ towards $\mathrm{ES}_\alpha(X)$ is known.

Theorem 1.1 ([7, 32])**.** *Assume that $\sigma^2_{\mathrm{ES}_\alpha(X)}$ is finite and that $u \mapsto \mathrm{VaR}_u(X)$ is continuous in $1-\alpha$. Then, as $N \to \infty$, we have the weak convergence*

$$\sqrt{N}\left(\widehat{T}_N - \mathrm{ES}_\alpha(X)\right) \to \mathcal{N}\left(0, \sigma^2_{\mathrm{ES}_\alpha(X)}\right).$$

Here, $\mathcal{N}(0,\sigma^2)$ denotes the normal distribution with mean zero and variance σ^2. While it is obvious that $\sigma^2_{\mathrm{ES}_\alpha(X)} < \infty$ is a necessary assumption in Theorem 1.1, it is worthwhile to emphasise that the assumption regarding the continuity of $u \mapsto \mathrm{VaR}_u(X)$ happens to be necessary as well, see [32, Example 3.4] and Section 3.2 for more details.

1.2. **Finite sample performance of the plug-in estimator.** Upon initial examination, it may seem that Theorem 1.1 offers a satisfactory answer to the question of how well $\widehat{T}_N$ approximates $\mathrm{ES}_\alpha(X)$. Indeed, an overly optimistic interpretation of the normal approximation presented in Theorem 1.1 is that

$$\mathbb{P}\left(\left|\widehat{T}_N - \mathrm{ES}_\alpha(X)\right| \geq \varepsilon\sigma_{\mathrm{ES}_\alpha(X)}\right) \approx 2\exp\left(-\frac{1}{2}N\varepsilon^2\right); \tag{1.1}$$

or, equivalently and as is commonly adopted in terms of confidence levels, that with probability at least $1-\delta$,

$$\left|\widehat{T}_N - \mathrm{ES}_\alpha(X)\right| \lesssim \sigma_{\mathrm{ES}_\alpha(X)} \sqrt{\frac{2}{N}\log\left(\frac{2}{\delta}\right)}.$$

However, it is crucial to emphasize that this interpretation is in general *incorrect*.

The foremost reason is that an exponential rate as in (1.1) is simply false unless X is very light-tailed, in the sense that it satisfies a Gaussian-like tail decay. Indeed, while it can be established that the estimate in (1.1) is the best that one can hope for (see Section 3.3 for more details), the actual statistical behavior is significantly worse for distributions typically considered in finance and insurance.

To illustrate this point, let us consider the case where X follows a Pareto distribution with a scale parameter $\lambda > 2$ (ensuring a finite variance). This distribution class is known for its heavy tails and is commonly used to model losses in insurance (see, e.g., [5, 40]). Focusing on the dependence on N for simplicity, we shall show in Section 4 that

$$\mathbb{P}\left(\left|\widehat{T}_N - \mathrm{ES}_\alpha(X)\right| \geq \varepsilon\sigma_{\mathrm{ES}_\alpha(X)}\right) \geq \frac{C_\varepsilon}{N^{\lambda-1}}, \tag{1.2}$$

which is far from the behavior suggested in (1.1). In fact, for values of λ close to 2, the right-hand side in (1.2) scales almost as $\frac{1}{N}$ — in other words, linearly as opposed to exponentially as in (1.1). In this regard, Section 3.1 establishes that for more general classes of distributions, the linear rate

$$\mathbb{P}\left(\left|\widehat{T}_N - \mathrm{ES}_\alpha(X)\right| \geq \varepsilon\sigma_{\mathrm{ES}_\alpha(X)}\right) \gtrsim \frac{1}{N\varepsilon^2}$$

is indeed minimax optimal.

An additional reason why (1.1) is false stems from a distinct aspect, one that remains unaffected by the assumed tail behavior of X. Indeed, Theorem 1.1 asserts that if $u \mapsto \mathrm{VaR}_u(X)$ is continuous, then the typical behavior of the error is

$$\left|\widehat{T}_N - \mathrm{ES}_\alpha(X)\right| = \mathcal{O}\left(\frac{\sigma_{\mathrm{ES}_\alpha(X)}}{\sqrt{N}}\right) \tag{1.3}$$

for *sufficiently large* N, but it does not provide a characterization of when N falls within this regime. It turns out that such a characterization is intricately tied to the quantitative continuity of the function $u \mapsto \mathrm{VaR}_u(X)$. Specifically, assuming L-Lipschitz continuity of the latter function (and, say, that $|X| \leq 1$ to alleviate concerns related to the tail behavior of X), we show that the typical behavior of the error has an additional 'second order' term:

$$\left|\widehat{T}_N - \mathrm{ES}_\alpha(X)\right| = \mathcal{O}\left(\frac{\sigma_{\mathrm{ES}_\alpha(X)}}{\sqrt{N}}\right) + \mathcal{O}\left(\frac{L}{N}\right) \tag{1.4}$$

for all N, see Section 3.2 for more details and in particular Proposition 3.4 for a precise minimax result in this regard. This shows that the regime suggested in (1.3) is the right one *only* for $N \gtrsim \sigma^2_{\mathrm{ES}_\alpha(X)} L^{-2}$.

In summary, the most optimistic estimate that one can aspire to achieve is

$$\mathbb{P}\left(\left|\widehat{T}_N - \mathrm{ES}_\alpha(X)\right| \gtrsim \varepsilon \sigma_{\mathrm{ES}_\alpha(X)} + \varepsilon^2 L\right) \leq 2\exp\left(-\frac{1}{2}N\varepsilon^2\right), \tag{1.5}$$

but crucially this estimate can only be true if X has very light tails, comparable to a Gaussian distribution; otherwise the performance of $\widehat{T}_N$ tends to be significantly worse.

1.3. **An improved estimator.** Given that the tails of distributions used in finance and insurance are usually significantly heavier than those of a Gaussian distribution, the poor statistical performance of the plug-in estimator was addressed by many researchers. A natural course of action going back to the early days of robust statistics is to explore *alternative* estimators — i.e. functions $\widehat{R}_N \colon \mathbb{R}^N \to \mathbb{R}$ that assign to the data $(X_i)_{i=1}^N$ an estimate for $\mathrm{ES}_\alpha(X)$, and we refer to [9, 18, 28, 41, 42] for several works in this direction and the survey [27]. Broadly speaking, the proposed estimators were usually demonstrated to enhance performance in statistical experiments, and/or their asymptotic statistical behavior was studied (with results similar to those in Theorem 1.1). However, to the best of our knowledge, the statistical performance for finite samples has not been addressed so far — in particular the presence of the second order term in (1.4) remained unnoticed.

In this article, we construct an estimator that exhibits the *optimal* statistical behavior: irrespective of the tail-decay of X, if $u \mapsto \mathrm{VaR}_u(X)$ is L-Lipschitz and $\sigma^2_{\mathrm{ES}_\alpha(X)}$ is finite, then the optimal estimate (1.5) is true once we replace $\widehat{T}_N$ by our proposed estimator.

We start with the formal definition of the estimator and explain the intuition behind it in Remark 1.5. To that end, for $x > 0$, denote by $\lfloor x \rfloor$ the largest integer smaller than x and by $\lceil x \rceil$ the smallest integer larger than x.

Definition 1.2. Set $m := \lceil \frac{11}{\varepsilon^2} \rceil$, $n := \lfloor \frac{N}{m} \rfloor$ and

$$I_j := \{(j-1)m + 1, \dots, jm\}, \quad \text{for } j = 1, \dots, n,$$

so that $nm \leq N$ and $(I_j)_{j=1}^n$ forms a disjoint partition of $\{1, \dots, mn\}$. For each j, denote by $\widehat{T}_{I_j}$ the plug-in estimator using only the sample $(X_i)_{i \in I_j}$, i.e.,

$$\widehat{T}_{I_j} := \frac{1}{\alpha} \int_{1-\alpha}^1 \mathrm{VaR}_u\left(\widehat{F}_{I_j}\right) du, \text{ where } \widehat{F}_{I_j}(t) := \frac{1}{m} \sum_{i \in I_j} \mathbb{1}_{(-\infty, t]}(X_i).$$

Denote by $\hat{Q}(\beta)$ the linearly interpolated empirical quantile function of $(\widehat{T}_{(j)})_{j=1}^n$.[1] Then, for some hyperparameters $0.35 \le \beta_1 \le \beta_2 \le 0.65$, we set

$$\widehat{S}_N := \min\left\{\max\left\{\widehat{T}_N, \hat{Q}(\beta_1)\right\}, \hat{Q}(\beta_2)\right\}.$$

We will explain the reasoning behind the estimator $\widehat{S}_N$ and discuss the importance of the hyperparameters β_1 and β_2 in Remark 1.5. Before doing so, let us present our first main result pertaining to the statistical behaviour of $\widehat{S}_N$.

Theorem 1.3. *There are absolute constants $c_0, c_1, c_2, c_3 > 0$ such that the following holds. Assume that $\sigma^2_{\mathrm{ES}_\alpha(X)}$ is finite and that $u \mapsto \mathrm{VaR}_u(X)$ is L-Lipschitz continuous on $[1-2\alpha, 1-\frac{1}{2}\alpha]$. Then, for every $\varepsilon < c_0\sqrt{\alpha}$ and $N \ge c_1\varepsilon^{-2}$,*

$$\mathbb{P}\left(\left|\widehat{S}_N - \mathrm{ES}_\alpha(X)\right| \ge \varepsilon\sigma_{\mathrm{ES}_\alpha(X)} + c_2 L\varepsilon^2\right) \le \exp\left(-c_3 N\varepsilon^2\right).$$

Thus, following up on the previous discussion, the estimator $\widehat{S}_N$ achieves the optimal statistical behavior (1.5) (up to the specific choice of the multiplicative constants $c_0, \dots, c_3$) under minimal and necessary assumptions.

Remark 1.4. Our proof shows that one may choose $c_0 = \frac{1}{6}$, $c_1 = 140$, $c_2 = 10$, and $c_3 = \frac{1}{700}$ though we believe that it is possible to obtain considerably better estimates for these constants. Moreover, in Definition 1.2, the constant 11 when setting $m := \lceil \frac{11}{\varepsilon^2} \rceil$ should be regarded suggestively: a different constant would change the statement of Theorem 1.3 only via the potential need for different multiplicative constants.

Remark 1.5. Let us shortly discuss the intuition behind the proposed estimator. While Theorem 1.3 is valid for all choices of $0.35 \le \beta_1 \le \beta_2 \le 0.65$, our numerical experiments reveal that these values significantly affect the empirical *bias* of $\widehat{S}_N$. This happens because the block estimators $\widehat{T}_{I_j}$ tend to be right-skewed. Thus, setting e.g. $\beta_1 = \beta_2 = 0.5$ (resulting in $\widehat{S}_N$ being the median of blocks estimator, see Section 1.5 for a discussion) often leads to a negative bias. Conversely, choosing $\beta_1 = \beta_2 > 0.5$ would induce a positive bias in situations where the central limit approximation for $\widehat{T}_{I_j}$ takes its effects (i.e. when m is large). Thus, the optimal choice of β_1, β_2 should depend on whether one believes the underlying distribution to be skewed and on the size of m. We present some numerical examples in Section 5 which showcase the impact of the choice of β_1 and β_2.

For choices $\beta_1 < 0.5 < \beta_2$ the estimator $\widehat{S}_N$ can be interpreted in terms of adaptive truncation: if the data $(X_i)_{i=1}^N$ has no outliers, the plug-in estimator $\widehat{T}_N$ is likely to fall

[1] Following [20, Definition 7], denote by $(\widehat{T}_{(j)})_{j=1}^n$ the order statistics of $(\widehat{T}_{I_j})_{j=1}^n$, set $\hat{Q}(0) = \widehat{T}_{(1)}, \hat{Q}(1) = \widehat{T}_{(n)}$,

$$\hat{Q}\left(\frac{j-1}{n-1}\right) = \widehat{T}_{(j)}, \quad \text{for } j = 2, \dots, n-1,$$

and let $\hat{Q}$ be the function that linearly interpolates between these values.

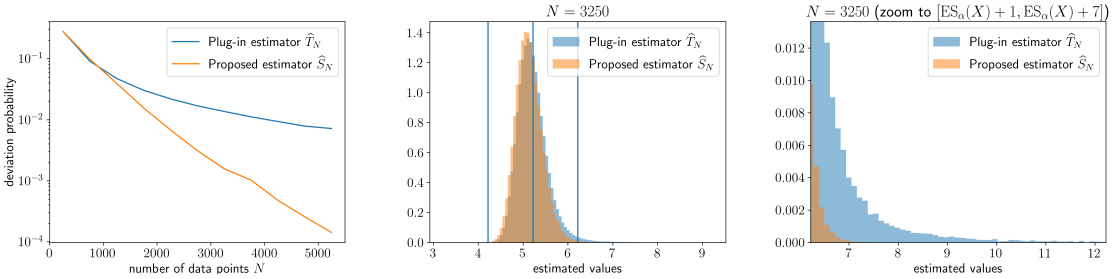

FIGURE 1. The left image depicts $\mathbb{P}(|\widehat{T}_N - \mathrm{ES}_\alpha(X)| \geq 1)$ (plug-in estimator) and $\mathbb{P}(|\widehat{S}_N - \mathrm{ES}_\alpha(X)| \geq 1)$ (the proposed estimator uses $\beta_1 = 0.5, \beta_2 = 0.6$ and $m = 250$) for varying values of N, showcasing the exponential rate for $\widehat{S}_N$ and the lack thereof for $\widehat{T}_N$. Here $\alpha = 0.1$, X is Pareto distributed with $\lambda = 2.2$, and the deviation probabilities are estimated using 10^6 many simulations. The middle image showcases the histogram of estimated values across simulations for $N = 3250$, with the blue vertical lines depicting the values $\mathrm{ES}_\alpha(X) - 1, \mathrm{ES}_\alpha(X), \mathrm{ES}_\alpha(X) + 1$. The right hand image shows the tail behavior of the estimators. Among the 10^6 simulations for $N = 3250$, there are 13637 errors larger than one for $\widehat{T}_N$ (largest realized value 98.98), and 1568 for $\widehat{S}_N$ (largest realized value 7.83).

between $\widehat{Q}(\beta_1)$ and $\widehat{Q}(\beta_2)$; thus $\widehat{S}_N = \widehat{T}_N$ and since the latter averages all data points, it should have a small bias. Conversely, if $\widehat{T}_N$ lies outside that range, one suspects the presence of outliers and our estimator truncates $\widehat{T}_N$ at $\widehat{Q}(\beta_1)$ and $\widehat{Q}(\beta_2)$.

A simple illustration of the performance of our estimator $\widehat{S}_N$ compared with the plug-in estimator $\widehat{T}_N$ is provided in Figure 1. Beyond the rate of estimation, another interesting aspect which can be seen from the illustrated example is that the proposed estimator never fails drastically. Indeed, while the plug-in estimator $\widehat{T}_N$ sometimes yields extreme values (e.g. a larger than 1000% deviation from the true value even with $N = 3250$), the proposed estimator $\widehat{S}_N$ is much more robust in this regard, never deviating more than 45% from the true value across all 10^6 many simulations.

Figure 1 further underlines the intuition given in Remark 1.5: whenever the values of the plug-in estimator $\widehat{T}_N$ lie within the typical regions, then $\widehat{S}_N \approx \widehat{T}_N$ (middle image). Conversely, the right image showcases that in the presence of outliers in the data, $\widehat{S}_N$ outperforms $\widehat{T}_N$.

For completeness, let us note that in the regime when ε is small, the 'second order' term $\mathcal{O}(L\varepsilon^2)$ may be neglected and the estimate in Theorem 1.3 is of the same order as the (in

general too) optimistic guess from the central limit theorem in (1.1). To state the result formally, let $c_0, \dots, c_3$ be the constants appearing in Theorem 1.3.

Corollary 1.6. *In the setting of Theorem 1.3, if in addition* $\varepsilon \leq \frac{\sigma_{\mathrm{ES}_\alpha(X)}}{c_2 L}$, *then*

$$\mathbb{P}\left(\left|\widehat{S}_N - \mathrm{ES}_\alpha(X)\right| \geq 2\varepsilon\sigma_{\mathrm{ES}_\alpha(X)}\right) \leq \exp\left(-c_3 N\varepsilon^2\right). \tag{1.6}$$

Combining the conditions on ε from Theorem 1.3 and Corollary 1.6 implies that $\varepsilon \lesssim \min\{\sqrt{\alpha}, \frac{\sigma_{\mathrm{ES}_\alpha(X)}}{L}\}$. While the value of $\frac{\sigma_{\mathrm{ES}_\alpha(X)}}{L}$ heavily depends on the underlying distribution of X and can be significantly smaller than $\sqrt{\alpha}$ in general, we shall show in Section 4 that if X has a Pareto distribution, then $\frac{\sigma_{\mathrm{ES}_\alpha(X)}}{L} \lesssim \sqrt{\alpha}$. Thus, for the Pareto distribution, the additional restriction on ε in Corollary 1.6 is (essentially) satisfied automatically.

1.4. **Adversarial robustness.** In addition to its favourable statistical behavior, our estimator $\widehat{S}_N$ possesses another advantage: it is *adversarially robust*. The latter means that even if an adversary can modify a certain fraction of the data $X_1, \dots, X_N$ arbitrarily (even under knowledge of the real data and knowledge of the estimation procedure), one still has guarantees that the estimator $\widehat{S}_N$ performs well. It is worth noting that this is in stark contrast to the behavior of the classical plug-in estimator, where (regardless of the distribution of X) the modification of even a *single* sample X_i can completely distort the resulting estimate.

Theorem 1.7. *Let the assumptions in Theorem 1.3 hold. Then there are absolute constants* C_0, C_1, C_2 *such that even if at most* $C_0 N\varepsilon^2$ *of the* N *samples are maliciously modified,*

$$\mathbb{P}\left(\left|\widehat{S}_N - \mathrm{ES}_\alpha(X)\right| \geq \varepsilon\sigma_{\mathrm{ES}_\alpha(X)} + C_1 L\varepsilon^2\right) \leq \exp\left(-C_2 N\varepsilon^2\right).$$

We show that $C_0 := \frac{1}{140}$, $C_1 = 10$, and $C_2 := \frac{1}{2800}$ are valid choices, but believe that the optimal values of these constants is considerably better.

Finally, let us note that if a sufficiently large number of data points are corrupted (say, all of them), then clearly there can be no procedure that estimates the expected shortfall in a satisfactory manner. In our context, denoting by K the number of corrupted data points, we require that $K \leq C_0 N\varepsilon^2$; or, in other words, if we expect that K data points are corrupted, the lower bound on the level of accuracy of the estimator is $\varepsilon \geq \sqrt{\frac{K}{C_0 N}}$. As it happens, this dependence of the error on the number of corrupted data turns out to be minimax-optimal in the class of distributions satisfying the assumption of Theorem 1.7: We show in Proposition 3.7 that no estimator can have a better error rate than $\mathcal{O}(\sigma_{\mathrm{ES}_\alpha(X)}\sqrt{K/N})$, where K is the number of corrupted data points.

1.5. **Related literature.** For the particular choice of $\beta_1 = \beta_2 = 1/2$, the estimator $\widehat{S}_N$ is the median of the block estimators $(\widehat{T}_{I_j})_{j=1}^n$, which we call the median of blocks estimator for short. In the context of estimating the mean of a random variable, this median of blocks

estimator constitutes the analogue to the so-called median-of-means estimator, whose optimal statistical performance is well understood (see, e.g. [2, 29, 21]). Contrary to mean estimation, properties for estimating the expected shortfall have only been obtained more recently. For instance, proving a central limit theorem for the empirical plug-in estimator $\widehat{T}_N$ was obtained in [7, 32], and, to the best of the authors' knowledge, no statistically optimal finite sample guarantees for any expected shortfall estimator are available in the literature.

Particularly in relation to practical tools employed in finance, it is worth mentioning that the proposed estimator in Definition 1.2 and the median of blocks estimator can be understood in terms of resampling methods, particularly non-parametric bootstrapping (cf. [19]). Indeed, the splitting method for the intervals in Definition 1.2 can be regarded as resampling n many samples of size m (not independently as usually done in bootstrapping, but non-overlapping). This resampling perspective is still in line with recent improvements proposed to the median-of-mean estimator using overlapping intervals (cf. [25, 26]). In this direction, further practical improvements to the estimator proposed in Definition 1.2 may be possible, which we believe is an interesting topic for future research.

Another topic this paper relates to is that of stability and (adversarial) robustness of estimators. Adversarial robustness of data-driven systems (see, e.g., [8, 17, 30]) can be regarded as a strong form of robustness against outliers (see, e.g., [23, 35, 36]), and has recently gathered increasing attention, particularly in fields adjacent to machine learning. In line with this, Theorem 1.7 provides statistical guarantees even when a small fraction of the data may be erroneous or even manipulated. We believe such guarantees are particularly helpful if growing amounts of data are used which make strict quality control harder to enforce. Finally, we emphasize that adversarial robustness can also be seen as a stability property of estimators with respect to perturbations in the data (cf. [3, 10, 14]), particularly since modifying a part of the data corresponds to a perturbation of the data distribution with respect to the total variation norm. In this respect, Theorem 1.7 may also be interpreted not just as robustness with respect to errors in the data, but also as robustness with respect to rare structural breaks in the data-generating distribution.

1.6. **Structure of the paper.** The remainder of the paper is structured as follows. The following Section 2 contains the proofs of Theorems 1.3 and 1.7. In Section 3 we focus on the results regarding optimality of the estimates. The focus of Section 4 is on Pareto distributions, in which case an explicit lower bound for the plug-in estimator is also established. In Section 5 we present several numerical experiments.

2. Proofs of Theorem 1.3 and Theorem 1.7

For shorthand notation, we write $\sigma^2_{\mathrm{ES}_\alpha}$ instead of $\sigma^2_{\mathrm{ES}_\alpha(X)}$ and F instead of F_X. Moreover, recall that $m = \lceil \frac{11}{\varepsilon^2} \rceil$, that $\varepsilon \leq \frac{1}{6}\sqrt{\alpha}$ where $\alpha < \frac{1}{2}$, and that $n = \lfloor \frac{N}{m} \rfloor$. In particular, the

assumptions that $\varepsilon \le 1$ and that $N\varepsilon^2 \ge 140$ immediately imply that

$$\frac{12}{\varepsilon^2} \ge m \ge 11, \quad \text{and} \quad n \ge \frac{N\varepsilon^2}{14} \ge 10. \tag{2.1}$$

We shall use these facts many times.

To make the analysis more transparent, we will assume (in this section only) that F is continuous and strictly increasing. This can be done without loss of generality: Otherwise, instead of X consider $Y^\eta := X + \eta G$ where G is a standard Gaussian random variable and $\eta > 0$ is arbitrarily small. Clearly

$$\mathrm{ES}_\alpha(Y^\eta) \to \mathrm{ES}_\alpha(X) \quad \text{and} \quad \sigma^2_{\mathrm{ES}_\alpha(Y^\eta)} \to \sigma^2_{\mathrm{ES}_\alpha(X)}$$

as $\eta \to 0$. In a similar manner, for every arbitrarily small $\gamma > 0$, if η is sufficiently small, then the samples from X and from Y^η are almost the same, i.e., $|Y_i^\eta - X_i| \le \gamma$ for all i with arbitrarily high probability. Thus computing the estimator $\widehat{S}_N$ using the samples from X or from Y will results in an error of at most γ, which is arbitrarily small.

The proof of Theorem 1.3 is based on the following standard fact about Binomial concentration. For the sake of completeness, we provide the short proof.

Lemma 2.1. *Let $y_0 \in \mathbb{R}$, let $\delta \ge 0$, and set $(Y_j)_{j=1}^n$ to be independent random variables satisfying $\mathbb{P}(|Y_j - y_0| > \delta) \le \frac{1}{10}$ for every $j = 1, \dots, n$. Then,*

$$\mathbb{P}\left(|\{j \in \{1,\dots,n\} : |Y_j - y_0| > \delta|\} \ge 0.3n\right) \le \exp\left(-\frac{n}{50}\right).$$

Proof. An application of Hoeffding's inequality (see, e.g., [6, Theorem 2.8]) implies that for every $\lambda > 0$, with probability at least $1 - \exp(-\frac{\lambda^2}{2n})$,

$$\sum_{j=1}^n \mathbb{1}_{\{|Y_j - y_0| > \delta\}} < \sum_{j=1}^n \mathbb{E}[\mathbb{1}_{\{|Y_j - y_0| > \delta\}}] + \lambda.$$

To complete the proof, set $\lambda = \frac{2}{10}n$. □

Let us recall that

$$\widehat{T}_{I_j} = \mathrm{ES}_\alpha\left(\widehat{F}_{I_j}\right) = \frac{1}{\alpha}\int_{1-\alpha}^1 \mathrm{VaR}_u\left(\widehat{F}_{I_j}\right) du$$

is the plug-in estimator of the sample restricted to the block I_j. Moreover, as the blocks $(I_j)_{j=1}^n$ are disjoint, the family $(\widehat{T}_{I_j})_{j=1}^n$ is independent. Hence, with Lemma 2.1 in mind, if we can show that

$$\mathbb{P}\left(\left|\widehat{T}_{I_j} - \mathrm{ES}_\alpha(X)\right| > \varepsilon\sigma_{\mathrm{ES}} + 10L\varepsilon^2\right) \le \frac{1}{10}, \tag{2.2}$$

then letting $0.3n \le k \le \ell \le 0.7n$, Lemma 2.1 implies that with exponentially high probability,

$$\widehat{T}_{(k)}, \widehat{T}_{(\ell)} \in \left[\mathrm{ES}_\alpha(X) - \varepsilon\sigma_{\mathrm{ES}} - 10L\varepsilon^2\,,\, \mathrm{ES}_\alpha(X) + \varepsilon\sigma_{\mathrm{ES}} + 10L\varepsilon^2\right], \tag{2.3}$$

where we recall that $(\widehat{T}_{(j)})_{j=1}^n$ are the order statistics of $(\widehat{T}_{I_j})_{j=1}^n$. In particular, since $\widehat{S}_N$ is sandwiched between $\widehat{T}_{(k)}$ and $\widehat{T}_{(\ell)}$, this will imply the statement in Theorem 1.3.

In order to prove (2.2), we make use of the following representation of the expected shortfall as a so-called distortion risk measure (see, e.g., [37, 39]): Setting

$$\psi\colon [0,1] \to [0,1], \qquad x \mapsto \min\left\{\frac{1-x}{\alpha}, 1\right\},$$

then, for every random variable Y,

$$\mathrm{ES}_\alpha(Y) = \int_{-\infty}^0 \psi(F_Y(t)) - 1\, dt + \int_0^\infty \psi(F_Y(t))\, dt. \tag{2.4}$$

As our arguments for showing that (2.2) holds crucially rely on (2.4), we give the short proof of (2.4) for the convenience of the reader.

Proof of (2.4). In a first step, assume that $F_Y(0) \leq 1-\alpha$, i.e. $\mathbb{P}(Y > 0) \geq \alpha$. Then $\psi(F_Y(t)) = 1$ for every $t < 0$ and the first integral in (2.4) is equal to zero. An application of Fubini's theorem shows that

$$\begin{aligned}\mathrm{ES}_\alpha(Y) = \frac{1}{\alpha}\int_{1-\alpha}^1 F_Y^{-1}(u)\, du &= \frac{1}{\alpha}\int_{1-\alpha}^1 \int_0^\infty \mathbb{1}_{[0,F_Y^{-1}(u)]}(t)\, dt\, du \\ &= \int_0^\infty \frac{1}{\alpha}\int_{1-\alpha}^1 \mathbb{1}_{(F_Y(t),1]}(u)\, du\, dt \\ &= \int_0^\infty \frac{1}{\alpha}\min\{1-F_Y(t), \alpha\}\, dt = \int_0^\infty \psi(F_Y(t))\, dt,\end{aligned}$$

which proves (2.4) in case that $\mathbb{P}(Y > 0) \geq \alpha$.

As for the general case, let y be a constant that satisfies $\mathbb{P}(Y + y > 0) \geq \alpha$ and set $Y' := Y + y$. The proof then follows from a short computation using the fact that (2.4) holds for Y' (by the first step), that $\mathrm{ES}_\alpha(Y') = \mathrm{ES}_\alpha(Y) + y$, and that $F_{Y'}(t) = F_Y(t-y)$. □

To control the difference between $\widehat{T}_{I_j}$ and $\mathrm{ES}_\alpha(X)$, note that ψ is almost everywhere differentiable with $\psi' = -\frac{1}{\alpha}\mathbb{1}_{[1-\alpha,1]}$. Therefore, a first order Taylor expansion yields a decomposition of the estimation error

$$\begin{aligned}\widehat{T}_{I_j} - \mathrm{ES}_\alpha(X) = \mathrm{ES}_\alpha\left(\widehat{F}_{I_j}\right) - \mathrm{ES}_\alpha(X) &= \int_{\mathbb{R}} \psi\left(\widehat{F}_{I_j}(t)\right) - \psi(F(t))\, dt \\ &= \mathcal{L}_j + \mathcal{E}_j\end{aligned}$$

into a *linearised term* and an *error term*

$$\begin{aligned}\mathcal{L}_j &:= \int_{\mathbb{R}} \psi'(F(t))\left(\widehat{F}_{I_j}(t) - F(t)\right)\, dt, \\ \mathcal{E}_j &:= \int_{\mathbb{R}} \left(\psi'\big(F(t) + \xi(\widehat{F}_{I_j}(t) - F(t))\big) - \psi'(F(t))\right)\left(\widehat{F}_{I_j}(t) - F(t)\right)\, dt,\end{aligned}$$

where $\xi \in (0,1)$ is some midpoint (that depends on t, $\widehat{F}_{I_j}$ and F).

In the following we analyse the linearised term and the error term separately.

Lemma 2.2. *We have that*

$$\mathbb{P}\left(|\mathcal{L}_j| > \varepsilon\sigma_{\mathrm{ES}_\alpha}\right) \leq \frac{1}{11}.$$

Proof. We first claim that $\mathbb{E}[\mathcal{L}_j^2] = \sigma_{\mathrm{ES}}^2/m$. To that end, note that

$$\mathcal{L}_j^2 = \int_{\mathbb{R}}\int_{\mathbb{R}} \psi'(F(t))\psi'(F(s)) \left(\widehat{F}_{I_j}(t) - F(t)\right)\left(\widehat{F}_{I_j}(s) - F(s)\right) dtds.$$

Moreover, since $\mathbb{E}[\widehat{F}_{I_j}(t)] = F(t)$ for every $t \in \mathbb{R}$, an application of Fubini's theorem shows that

$$\begin{aligned}\mathbb{E}[\mathcal{L}_j^2] &= \int_{\mathbb{R}}\int_{\mathbb{R}} \psi'(F(t))\psi'(F(s))\mathbb{C}\mathrm{ov}\left[\widehat{F}_{I_j}(t), \widehat{F}_{I_j}(s)\right] dtds \\ &= \frac{1}{\alpha^2}\int_{F^{-1}(1-\alpha)}^{\infty}\int_{F^{-1}(1-\alpha)}^{\infty} \mathbb{C}\mathrm{ov}\left[\widehat{F}_{I_j}(t), \widehat{F}_{I_j}(s)\right] dtds,\end{aligned}$$

where the second equality holds because $\psi' = -\frac{1}{\alpha}\mathbb{1}_{[1-\alpha,1]}$ and $F(t) < 1-\alpha$ for every $t < F^{-1}(1-\alpha)$. Next, by independence of the sample $(X_i)_{i\in I_j}$,

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left[\widehat{F}_{I_j}(t), \widehat{F}_{I_j}(s)\right] &= \frac{1}{m^2}\sum_{i\in I_j} \mathbb{E}\left[(\mathbb{1}_{(-\infty,t]}(X_i) - F(t))(\mathbb{1}_{(-\infty,s]}(X_i) - F(s))\right] \\ &= \frac{F(\min\{t,s\}) - F(t)F(s)}{m}.\end{aligned}$$

Therefore, it follows from the definition of $\sigma_{\mathrm{ES}_\alpha}^2$ that $\mathbb{E}[\mathcal{L}_j^2] = \sigma_{\mathrm{ES}_\alpha}^2/m$.

To complete the proof, we use Chebychev's inequality which guarantees that

$$\mathbb{P}\left(|\mathcal{L}_j| > \varepsilon\sigma_{\mathrm{ES}}\right) \leq \frac{\mathbb{E}[\mathcal{L}_j^2]}{(\varepsilon\sigma_{\mathrm{ES}})^2} = \frac{\sigma_{\mathrm{ES}}^2}{m\varepsilon^2\sigma_{\mathrm{ES}}^2} \leq \frac{1}{11},$$

where the last inequality holds since $m \geq \frac{11}{\varepsilon^2}$ (see (2.1)). □

Lemma 2.3. *Set $\gamma := \frac{5}{4}\varepsilon\sqrt{\alpha}$. Then we have that*

$$\mathbb{P}\left(|\mathcal{E}_j| > \frac{3\gamma}{\alpha}\left(F^{-1}(1-\alpha+\gamma) - F^{-1}(1-\alpha-\gamma)\right)\right) \leq \frac{1}{110}.$$

Proof. Note that $\gamma \leq \frac{5}{24}\alpha$ by the assumption that $\varepsilon < \frac{1}{6}\sqrt{\alpha}$, and let $t_0, t_1 \in \mathbb{R}$ satisfy that

$$F(t_0) = 1-\alpha-\gamma, \quad F(t_1) = 1-\alpha+\gamma.$$

Such t_0, t_1 exist since F is strictly increasing and continuous in this section.

Step 1: We claim that

$$\mathcal{A} := \left\{\left|\widehat{F}_{I_j}(s) - F(s)\right| \leq \gamma \text{ for } s = t_0, t_1\right\}$$

satisfies $\mathbb{P}(\mathcal{A}) \leq \frac{1}{110}$. To that end, recall that Bernstein's inequality (see, e.g., [38, Theorem 6.12]) guarantees that if $Y_1, \dots, Y_m$ are i.i.d. random variables which take values between 0 and 1, then for every $\lambda \geq 0$, with probability at least $1 - 2\exp(-\lambda)$,

$$\left| \frac{1}{m} \sum_{i=1}^{m} Y_i - \mathbb{E}[Y_1] \right| \leq \sqrt{\frac{2\mathbb{V}\mathrm{ar}[Y_1]\lambda}{m}} + \frac{2\lambda}{3m}.$$

Applied to $Y_i = 1_{(-\infty,s]}(X_i)$ (for $s = t_0, t_1$) and $\lambda := \ln(440)$, it follows that with probability at least $1 - \frac{1}{220}$,

$$\left| \widehat{F}_{I_j}(s) - F(s) \right| \leq \sqrt{\frac{2F(s)(1-F(s))\ln(440)}{m}} + \frac{2\ln(440)}{3m} =: \varphi(s).$$

Thus, it remains to show that $\varphi(s) \leq \gamma$ for $s = t_0, t_1$.

To that end, fix $s \in \{t_0, t_1\}$ and note that since $\gamma \leq \frac{5}{24}\alpha$,

$$F(s)(1-F(s)) \leq \alpha + \gamma \leq \frac{29}{24}\alpha.$$

Therefore, using that $m \geq \frac{11}{\varepsilon^2}$ and that $\varepsilon^2 \leq \frac{1}{6}\varepsilon\sqrt{\alpha}$,

$$\begin{aligned} \varphi(s) &\leq \sqrt{\frac{2 \cdot 29 \cdot \alpha\varepsilon^2 \cdot \ln(440)}{24 \cdot 11}} + \frac{2 \cdot \ln(440)\varepsilon^2}{33} \\ &\leq \varepsilon\sqrt{\alpha} \cdot \left(\sqrt{\frac{2 \cdot 29 \cdot \ln(440)}{24 \cdot 11}} + \frac{2 \cdot \ln(440)}{33 \cdot 6} \right) \leq \varepsilon\sqrt{\alpha} \cdot \frac{5}{4} = \gamma. \end{aligned}$$

Step 2: We claim that in the event $\mathcal{A}$,

$$|\mathcal{E}_j| \leq \frac{1}{\alpha} \int_{t_0}^{t_1} \left| F(t) - \widehat{F}_{I_j}(t) \right| \, dt. \tag{2.5}$$

Indeed, recall that

$$\mathcal{E}_j = \int_{\mathbb{R}} \Big(\psi'\big(F(t) + \xi(\widehat{F}_{I_j}(t) - F(t))\big) - \psi'(F(t)) \Big) \Big(\widehat{F}_{I_j}(t) - F(t) \Big) \, dt$$

and observe that for $t < t_0$ or $t > t_1$, the term in the first bracket is equal to zero. Indeed, consider $t < t_0$ and note that by the strict monotonicity of F and the definition of t_0, $F(t) < 1 - \alpha - \gamma$. Moreover, by the monotonicity of $\widehat{F}_{I_j}$,

$$\widehat{F}_{I_j}(t) \leq \widehat{F}_{I_j}(t_0) \leq F(t_0) + \gamma \leq 1 - \alpha,$$

where the second inequality follows from the definition of $\mathcal{A}$. Since $\psi' = -\frac{1}{\alpha}\mathbb{1}_{[1-\alpha,1]}$, it therefore follows that for $\xi \in (0,1)$,

$$\psi'\big(F(t) + \xi(\widehat{F}_{I_j}(t) - F(t))\big) - \psi'(F(t)) = 0 - 0 = 0.$$

In a similar manner, for every $t > t_1$, $F(t) > 1-\alpha+\gamma$ and $\widehat{F}_{I_j}(t) \geq 1-\alpha$, and thus for $\xi \in (0,1)$,

$$\psi'\big(F(t)+\xi(\widehat{F}_{I_j}(t)-F(t))\big) - \psi'(F(t)) = 1-1 = 0$$

and (2.5) follows.

Step 3: First observe that in the event $\mathcal{A}$, for every $t \in [t_0, t_1]$,

$$\Big|F(t) - \widehat{F}_{I_j}(t)\Big| \leq 3\gamma. \tag{2.6}$$

Indeed, fix $t \in [t_0, t_1]$. Then, by the monotonicity of $\widehat{F}_{I_j}$ and the definition of $\mathcal{A}$,

$$\widehat{F}_{I_j}(t) \leq \widehat{F}_{I_j}(t_1) \leq F(t_1) + \gamma \leq F(t) + 2\gamma + \gamma.$$

The same argument implies that $\widehat{F}_{I_j}(t) \geq F(t) - 3\gamma$ which shows (2.6).

Finally, using (2.5) and (2.6), we conclude that in the event $\mathcal{A}$,

$$|\mathcal{E}_j| \leq \frac{1}{\alpha}\int_{t_0}^{t_1} 3\gamma\, dt = \frac{3(t_1-t_0)\gamma}{\alpha},$$

and the proof is completed by the definition of t_0 and t_1. □

Proof of Theorem 1.3. Recall that

$$\widehat{T}_{I_j} - \mathrm{ES}_\alpha(X) = \mathcal{L}_j + \mathcal{E}_j$$

and that by Lemma 2.2, $\mathbb{P}(|\mathcal{L}_j| \geq \varepsilon\sigma_{\mathrm{ES}_\alpha}) \leq \frac{1}{11}$. To control the term $\mathcal{E}_j$, set $\gamma := \frac{5}{4}\varepsilon\sqrt{\alpha}$ so that Lemma 2.3 guarantees that with probability at least $1 - \frac{1}{110}$,

$$|\mathcal{E}_j| \leq \frac{3\gamma}{\alpha}\left(F^{-1}(1-\alpha+\gamma) - F^{-1}(1-\alpha-\gamma)\right) =: (*).$$

Since $\gamma \leq \frac{1}{2}\alpha$ (by the assumption that $\varepsilon \leq \frac{1}{6}\sqrt{\alpha}$) and F^{-1} is L-Lipschitz on $[1-2\alpha, 1-\frac{1}{2}\alpha]$, it follows that

$$(*) \leq \frac{6\gamma^2 L}{\alpha} \leq 10\varepsilon^2 L.$$

Combining the estimates on $\mathcal{L}_j$ and $\mathcal{E}_j$ shows that

$$\mathbb{P}\left(\Big|\widehat{T}_{I_j} - \mathrm{ES}_\alpha(X)\Big| \geq \varepsilon\sigma_{\mathrm{ES}_\alpha} + 10L\varepsilon^2\right) \leq \frac{1}{10}.$$

Therefore, (2.2) holds, and the proof is completed by an application of Lemma 2.1. □

Proof of Theorem 1.7. The proof requires only minor modification to the proof of Theorem 1.3. For shorthand notation, set $\delta := \varepsilon\sigma_{\mathrm{ES}_\alpha(X)} + 10L\varepsilon^2$.

In a first step, suppose that the data $X_1, \ldots, X_N$ are not modified, i.e. we work in the setting of Theorem 1.3. Then, by Lemma 2.2 and Lemma 2.3, $\mathbb{P}(|\widehat{T}_{I_j} - \mathrm{ES}_\alpha(X)| > \delta) \leq \frac{1}{10}$ for all j. Hence, the same arguments as presented in the proof of Lemma 2.1

together with the estimates on n and m (see (2.1)) imply that with probability at least $1-\exp(-\frac{1}{14\cdot 200}N\varepsilon^2)$,

$$|\{j : |\widehat{T}_{I_j} - \mathrm{ES}_\alpha(X)| > \delta\}| < \sum_{j=1}^{n} \mathbb{P}(|\widehat{T}_{I_j} - \mathrm{ES}_\alpha(X)| > \delta) + \frac{n}{10} \leq \frac{2n}{10}. \tag{2.7}$$

In a next step, recall that at most $K \leq \frac{1}{140}N\varepsilon^2$ samples are modified and that $\frac{1}{140}N\varepsilon^2 \leq \frac{n}{10}$, see (2.1). Therefore, if at most K of the samples $X_1, \dots, X_N$ are modified, this can yield at most a modification of $\widehat{T}_{I_j}$ on $\frac{n}{10}$ of the blocks $I_1, \dots, I_n$. Hence, in the high probability event in which (2.7) holds, we still have that for more than $\frac{7}{10}n$ of the j's, $|\widehat{T}_{I_j} - \mathrm{ES}_\alpha(X)| \leq \delta$ and thus (2.3) holds, which yields the claim. □

For completeness, let us state the following immediate consequence of Theorem 1.7 on the consistency of $\widehat{S}_N$ (as $N \to \infty$) in the presence of corrupted data.

Corollary 2.4. *Suppose that $\sigma^2_{\mathrm{ES}_\alpha}$ is finite, that $u \mapsto \mathrm{VaR}_u(X)$ is Lipschitz continuous on a neighbourhood of $1-\alpha$, that at most $K_N < N$ data points are maliciously modified, and that $\lim_{N\to\infty} K_N/N = 0$. Then, setting $\varepsilon_N := 12\sqrt{K_N/N}$, it follows that*

$$\left|\widehat{S}_N - \mathrm{ES}_\alpha(X)\right| \xrightarrow{\mathbb{P}} 0 \quad \text{as } N \to \infty,$$

where $\xrightarrow{\mathbb{P}} 0$ denotes convergence in probability.

3. On the optimality

In this section, we rigorously prove the optimality claims made in the introduction. To that end, let us start with some preliminaries. We shall often make use of scaled Bernoulli distributions: For $x > 0$, write

$$X \sim \mathrm{B}(p, x) \quad \text{if } \mathbb{P}(X = 0) = 1 - p \text{ and } \mathbb{P}(X = x) = p.$$

In particular, if $X \sim \mathrm{B}(p, x)$, then

$$F_X = (1-p)\mathbb{1}_{[0,x)} + \mathbb{1}_{[x,\infty)} \quad \text{and} \quad F_X^{-1} = -\infty\mathbb{1}_{\{0\}} + 0\mathbb{1}_{(0,1-p]} + x\mathbb{1}_{(1-p,1]}. \tag{3.1}$$

The reason to consider the Bernoulli distribution steams from the fact that its empirical distribution remains a Bernoulli distribution (with a random success parameter), and that explicit calculations are simpler.

Lemma 3.1. *Let $p \in [0,1]$, $x > 0$, and $X \sim \mathrm{B}(p, x)$. Then*

$$\mathrm{ES}_\alpha(X) = x \min\left\{1, \frac{p}{\alpha}\right\},$$
$$\sigma^2_{\mathrm{ES}_\alpha(X)} = \begin{cases} \frac{x^2(p-p^2)}{\alpha^2} & \text{if } p \leq \alpha, \\ 0 & \text{else.} \end{cases}$$

Proof. By definition of the expected shortfall and by (3.1),

$$\mathrm{ES}_\alpha(X) = \frac{1}{\alpha}\int_{1-\alpha}^1 x\mathbb{1}_{(1-p,1]}(u)\,du = \begin{cases} \frac{1}{\alpha}x\alpha & \text{if } p \geq \alpha, \\ \frac{1}{\alpha}xp & \text{if } p < \alpha \end{cases}$$

from which the first claim readily follows.

As for the second claim, assume first that $p \leq \alpha$. Then $F_X^{-1}(1-\alpha) = 0$ (see (3.1)) and therefore

$$\sigma^2_{\mathrm{ES}_\alpha(X)} = \frac{1}{\alpha^2}\int_0^\infty \int_0^\infty F_X(\min\{t,s\}) - F_X(t)F_X(s)\,dtds.$$

Moreover, note that for $t \geq x$, $F_X(t) = 1$, hence, if either $t \geq x$ or $s \geq x$, then $F_X(\min\{t,s\}) - F_X(t)F_X(s) = 0$. On the other hand, for $t, s \in [0,x)$, the latter term is equal to $1 - p - (1-p)^2 = p - p^2$ from which the claim follows.

The proof in the case $p > \alpha$ follows because $F_X^{-1}(1-\alpha) = x$ (see (3.1)) and hence

$$\begin{aligned}\sigma^2_{\mathrm{ES}_\alpha(X)} &= \frac{1}{\alpha^2}\int_x^\infty \int_x^\infty F_X(\min\{t,s\}) - F_X(t)F_X(s)\,dtds \\ &= \frac{1}{\alpha^2}\int_x^\infty \int_x^\infty 1 - 1\,dtds = 0.\end{aligned}$$

□

In what follows, we will often use the following standard estimate, the proof of which we provide for completeness.

Lemma 3.2. *Let $N \geq 1$, let $p \in [0, \frac{1}{N}]$, and let $(\mathcal{A}_i)_{i=1}^N$ be independent events that satisfy $\mathbb{P}(\mathcal{A}_i) \geq p$ for all $i = 1, \ldots, N$. Then*

$$\mathbb{P}\left(\bigcup_{i=1}^N \mathcal{A}_i\right) \geq \frac{1}{4}Np.$$

Proof. Clearly

$$\mathbb{P}\left(\bigcup_{i=1}^N \mathcal{A}_i\right) = 1 - \mathbb{P}\left(\bigcap_{i=1}^N \mathcal{A}_i^c\right) \geq 1 - (1-p)^N.$$

Moreover, using that $\log(1-x) \leq -x$ for $x \in (0,1]$,

$$1 - (1-p)^N = 1 - \exp(N\log(1-p)) \geq 1 - \exp\left(-Np\right).$$

Finally, since $\exp(-x) \geq 1 - \frac{1}{4}x$ for $x \in [0,1]$, we conclude that

$$\mathbb{P}\left(\bigcup_{i=1}^N \mathcal{A}_i\right) \geq 1 - \exp\left(-Np\right) \geq \frac{1}{4}Np,$$

as claimed. □

3.1. **The minimax rate.** Let us recall that $\widehat{T}_N$ is the plug-in estimator for $\mathrm{ES}_\alpha(X)$. In this section we prove that the linear rate $\mathcal{O}(\frac{1}{N\varepsilon^2})$ for the rate of convergence of $\mathbb{P}(|\widehat{T}_N - \mathrm{ES}_\alpha(X)| \geq \varepsilon\sigma_{\mathrm{ES}_\alpha(X)})$ is optimal in a minimax sense. The alternative result using only Pareto-distributions is presented in Section 4.

Proposition 3.3. *Let $\alpha \leq \frac{1}{10}$, let $\varepsilon \leq \frac{1}{2}\sqrt{\alpha}$ and assume that $N\varepsilon^2 \geq 5$. Then there exists a random variable X such that $\sigma_{\mathrm{ES}_\alpha(X)} = 1$, F_X^{-1} is 1-Lipschitz on $[1-2\alpha, 1-\frac{1}{2}\alpha]$ and*

$$\mathbb{P}\left(\left|\widehat{T}_N - \mathrm{ES}_\alpha(X)\right| \geq \varepsilon\sigma_{\mathrm{ES}_\alpha(X)}\right) \geq \frac{1}{16N\varepsilon^2}.$$

Proof. For shorthand notation, put $\delta = 2\varepsilon$. Fix $N \geq 1$ and set

$$p := \frac{1}{\delta^2 N^2} \quad \text{and} \quad x := \frac{\alpha}{\sqrt{p-p^2}}.$$

Let $X \sim \mathrm{B}(p,x)$ so that $\sigma^2_{\mathrm{ES}_\alpha(X)} = 1$ by Lemma 3.1.

Let us start by collecting some trivial estimates on p. By assumption $N\delta^2 \geq 20$ and $\delta^2 \leq \alpha$, hence $N \geq \frac{20}{\alpha}$ and thus $p \leq \frac{1}{N} \leq \frac{\alpha}{20} \leq \frac{1}{100}$. Moreover, since $N\delta^2 \geq 20$, $\sqrt{p} \leq \frac{1}{20}\delta$. We will use these facts later. In particular, let us note that since $p \leq \frac{\alpha}{2}$, F_X^{-1} is constant on $(1-2\alpha, 1-\alpha/2)$ (see (3.1)) and thus also 1-Lipschitz.

Define the empirical success probability $\widehat{p}_N = \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{X_i=x}$ so that the empirical distribution $\widehat{F}_N$ of X is equal to that of $\mathrm{B}(\widehat{p}_N, x)$. Consider the event

$$\mathcal{A} := \left\{\widehat{p}_N \geq \frac{1}{N}\right\} = \bigcup_{i=1}^N \{X_i = x\}.$$

Since $\mathbb{P}(X_i = x) = p$ and $p \leq \frac{1}{N}$, Lemma 3.2 guarantees that

$$\mathbb{P}(\mathcal{A}) \geq \frac{1}{4}Np = \frac{1}{16N\varepsilon^2}.$$

It remains to show that in the event $\mathcal{A}$, $|\widehat{T}_N - \mathrm{ES}_\alpha(X)| \geq \frac{1}{2}\delta = \varepsilon\sigma_{\mathrm{ES}_\alpha(X)}$.

To that end, we first compute $\mathrm{ES}_\alpha(X)$. By Lemma 3.1 and since $p \leq \alpha$,

$$\mathrm{ES}_\alpha(X) = \frac{xp}{\alpha} = \frac{p}{\sqrt{p-p^2}} \leq \frac{11}{10}\sqrt{p},$$

where the inequality follows because $p \leq \frac{1}{100}$ and therefore $\sqrt{p-p^2} \geq \frac{10}{11}\sqrt{p}$.

On the other hand, in the event $\mathcal{A}$ we have $\widehat{p}_N \geq \frac{1}{N}$. Since $N\alpha \geq N\varepsilon^2 \geq 1$, it follows again from Lemma 3.1 that

$$\widehat{T}_N = x\min\left\{1, \frac{\widehat{p}_N}{\alpha}\right\} \geq x\min\left\{1, \frac{1}{N\alpha}\right\} = \frac{x}{N\alpha}.$$

Moreover, by definition of p and x,

$$\frac{x}{N\alpha} = \frac{1}{N\sqrt{p-p^2}} \geq \frac{1}{N\sqrt{p}} = \delta.$$

Therefore, using that $\sqrt{p} \leq \frac{1}{20}\delta$, in the event $\mathcal{A}$,

$$\widehat{T}_N - \mathrm{ES}_\alpha(X) \geq \delta - \frac{11}{10}\sqrt{p} \geq \frac{\delta}{2}$$

which completes the proof. □

3.2. **On the continuity of the quantile functions.** In this section we make rigorous the claim in the introduction that if $u \mapsto \mathrm{VaR}_u(X)$ is L-Lipschitz continuous, then the typical behavior of the estimation error scales like

$$\left|\widehat{T}_N - \mathrm{ES}_\alpha(X)\right| = \mathcal{O}\left(\frac{\sigma_{\mathrm{ES}_\alpha(X)}}{\sqrt{N}}\right) + \mathcal{O}\left(\frac{L}{N}\right)$$

in general. In the next proposition, we focus on the (perhaps more surprising) $\mathcal{O}(\frac{L}{N})$-term; the $\mathcal{O}(\frac{\sigma_{\mathrm{ES}_\alpha(X)}}{\sqrt{N}})$-term will be dealt with in Proposition 3.6.

Proposition 3.4. *There are absolute constants $c_0, c_1, c_2, c_3 > 0$ such that the following holds. Let $L \geq 1$ and set $\varepsilon \leq \min\{c_0\sqrt{\alpha}, \frac{1}{L}\}$. Then there exists a random variable X which satisfies*

(a) $|X| \leq 1$ almost surly,
(b) F_X^{-1} is L-Lipschitz on $(0,1)$,
(c) $\sigma^2_{\mathrm{ES}_\alpha(X)} = 0$,

and for every $N \geq \frac{c_1}{\varepsilon^2}$, we have that

$$\mathbb{P}\left(\left|\widehat{T}_N - \mathrm{ES}_\alpha(X)\right| \geq c_2\varepsilon^2 L\right) \geq \exp\left(-c_3 N\varepsilon^2\right).$$

Remark 3.5. In the setting and notation of Proposition 3.4: Suppose that $L \geq \frac{1}{c_0\sqrt{\alpha}}$, that $N \geq c_1 L^2$, and set $\varepsilon = \sqrt{c_1/N}$. Then, with probability at least $\exp(-c_1c_3)$,

$$\left|\widehat{T}_N - \mathrm{ES}_\alpha(X)\right| \geq c_1c_2\frac{L}{N};$$

and in particular $\mathbb{E}[|\widehat{T}_N - \mathrm{ES}_\alpha(X)|] \geq c\frac{L}{N}$ for $c = c_1c_2\exp(-c_1c_3) > 0$.

Proof of Proposition 3.4. Step 1: Let $\delta := \frac{1}{40}\varepsilon\sqrt{\alpha}$, put $x_0 := -\delta L$, and set X to have the following distribution:

$$F_X(t) = \begin{cases} 0 & \text{if } t < x_0, \\ 1 - \alpha - \delta + \frac{x_0 - t}{x_0}\delta & \text{if } t \in [x_0, 0), \\ 1 & \text{if } t \geq 0. \end{cases}$$

In other words, $\mathbb{P}(X = x_0) = 1 - \alpha - \delta$, $\mathbb{P}(X = 0) = \alpha$ and for $A \subset (x_0, 0)$ we have that $\mathbb{P}(X \in A) = \frac{\delta}{|x_0|} \operatorname{Leb}(A)$. Clearly

$$F_X^{-1}(u) = \begin{cases} x_0 & \text{if } u \in (0, 1-\alpha-\delta), \\ x_0 - \frac{x_0}{\delta}(u - (1-\alpha-\delta)) & \text{if } u \in [1-\alpha-\delta, 1-\alpha), \\ 0 & \text{if } u \in [1-\alpha, 1), \end{cases}$$

and in particular F_X^{-1} is Lipschitz continuous on $(0,1)$ with constant $L = \frac{|x_0|}{\delta}$. Moreover, since $\varepsilon \leq \frac{1}{L}$, it follows that $|x_0| \leq \varepsilon L \leq 1$, thus $|X| \leq 1$ almost surely.

Finally, since $F_X^{-1}(u) = 0$ for $u \in (1-\alpha, 1)$, it follows from the definitions of ES_α and $\sigma^2_{\mathrm{ES}_\alpha}$ that $\mathrm{ES}_\alpha(X) = \sigma^2_{\mathrm{ES}_\alpha(X)} = 0$.

Step 2: We claim that there is an absolute constant c such that

$$\mathbb{P}\left(\widehat{F}_N(x_0) \geq 1 - \alpha + \delta\right) \geq c \exp\left(-\frac{1}{2}\varepsilon^2 N\right).$$

To that end, we shall use the following Binomial anti-concentration result due to Feller. It is a consequence of [16, Theorem 1] that if $(Z_i)_{i=1}^N$ are i.i.d. random variables satisfying that $|Z_i| \leq 1$ and $N \geq \frac{400}{\mathbb{V}\mathrm{ar}[Z]}$, then for every $\lambda \in (0, \frac{\mathbb{V}\mathrm{ar}[Z]}{100})$,

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^N Z_i \geq \mathbb{E}[Z] + \lambda\right) \geq c \exp\left(-\frac{\lambda^2 N}{3\mathbb{V}\mathrm{ar}[Z]}\right). \tag{3.2}$$

We apply Feller's result to $Z_i = \mathbb{1}_{(-\infty, x_0]}(X_i)$. If we assume that $\varepsilon \leq \sqrt{\alpha}$, then

$$\mathbb{V}\mathrm{ar}[Z] = (1-\alpha-\delta)(\alpha+\delta) \in \left[\frac{\alpha}{2}, 2\alpha\right]$$

by the definition of $\delta = \frac{1}{40}\varepsilon\sqrt{\alpha}$. Moreover, if $\varepsilon \leq \frac{1}{40}\sqrt{\alpha}$, $\lambda := 2\delta$ satisfies that $\lambda \leq \frac{1}{100}\mathbb{V}\mathrm{ar}[Z]$. Therefore, by (3.2),

$$\mathbb{P}\left(\widehat{F}_N(x_0) \geq F(x_0) + 2\delta\right) \geq c \exp\left(-\frac{8\delta^2 N}{3\alpha}\right) = c \exp\left(-\tilde{c}\varepsilon^2 N\right)$$

for $\tilde{c} := \frac{8}{3 \cdot 40^2}$. Finally, if $N\varepsilon^2 \geq \frac{2}{\tilde{c}} \log(\frac{1}{c+1})$, then $c \exp\left(-\tilde{c}\varepsilon^2 N\right) \geq \exp(-2\tilde{c}\varepsilon^2 N)$.

Step 3: Fix a realization in the high probability event of step 2, i.e., such that $\widehat{F}_N(x_0) \geq 1 - \alpha + \delta$. Then, by definition of the inverse function,

$$\widehat{F}_N^{-1}(u) \leq x_0 \quad \text{for all } u < 1 - \alpha + \delta.$$

Moreover, clearly $\widehat{F}_N^{-1}(\cdot) \leq 0$, and therefore

$$\widehat{T}_N = \frac{1}{\alpha}\int_{1-\alpha}^1 \widehat{F}_N^{-1}(u)\, du \leq \frac{1}{\alpha}\int_{1-\alpha}^{1-\alpha+\delta} \widehat{F}_N^{-1}(u)\, du \leq \frac{x_0 \delta}{\alpha}.$$

In particular, since $\mathrm{ES}_\alpha(X) = 0$,

$$|\widehat{T}_N - \mathrm{ES}_\alpha(X)| \geq \frac{|x_0|\delta}{\alpha} = \frac{L\delta^2}{\alpha} = \frac{L\varepsilon^2}{1600},$$

which concludes the claim. □

3.3. **No estimator can outperform the CLT-rates.** We prove the claim from the introduction that

$$\mathbb{P}\left(\left|\widehat{T}_N - \mathrm{ES}_\alpha(X)\right| \gtrsim \varepsilon\sigma_{\mathrm{ES}_\alpha(X)}\right) \geq \exp\left(-N\varepsilon^2\right).$$

As it happens, such a lower bound is valid not only for the plug-in estimator $\widehat{T}_N$, but for *any* estimator:

Proposition 3.6. *Let $\varepsilon \leq \frac{1}{2}\sqrt{\alpha}$ and let $\widehat{R}_N$ be any estimator. Then there exists a random variable X for which $\sigma_{\mathrm{ES}_\alpha(X)} = 1$ and F_X^{-1} is 1-Lipschitz on $[1-2\alpha, 1-\frac{1}{2}\alpha]$ and*

$$\mathbb{P}\left(\left|\widehat{R}_N - \mathrm{ES}_\alpha(X)\right| \geq \frac{1}{10}\varepsilon\sigma_{\mathrm{ES}_\alpha(X)}\right) \geq \exp\left(-N\varepsilon^2\right).$$

Proof. Set

$$(p, x) := \left(\varepsilon^2, \frac{\alpha}{\sqrt{\varepsilon^2 - \varepsilon^4}}\right), \quad \text{and} \quad (q, y) := \left(\frac{\varepsilon^2}{2}, \frac{\alpha}{\sqrt{\varepsilon^2/2 - \varepsilon^4/4}}\right).$$

Moreover, let $X \sim \mathrm{B}(p, x)$ and $Y \sim \mathrm{B}(q, y)$. It follows from Lemma 3.1 that $\sigma^2_{\mathrm{ES}_\alpha(X)} = \sigma^2_{\mathrm{ES}_\alpha(Y)} = 1$. By the assumption that $\varepsilon \leq \frac{1}{2}\sqrt{\alpha}$, we have that $p, q \leq \frac{1}{4}\alpha$. In particular, the two functions F_X^{-1} and F_Y^{-1} are constant on $(1-2\alpha, 1-\frac{1}{2}\alpha)$, and thus 1-Lipschitz. Moreover, by Lemma 3.1,

$$\mathrm{ES}_\alpha(X) = \min\left\{\frac{xp}{\alpha}, x\right\} = \frac{xp}{\alpha} = \frac{p}{\sqrt{p - p^2}} \geq \frac{p}{\sqrt{p}} = \varepsilon.$$

In a similar manner,

$$\mathrm{ES}_\alpha(Y) = \min\left\{\frac{yq}{\alpha}, y\right\} = \frac{q}{\sqrt{q - q^2}}$$

and since $q = \frac{\varepsilon^2}{2} \leq \frac{1}{8}$ it holds that $q - q^2 \geq \frac{7}{8}q$ and therefore

$$\mathrm{ES}_\alpha(X) - \mathrm{ES}_\alpha(Y) \geq \varepsilon - \sqrt{\frac{8}{7}}\sqrt{q} \geq \frac{1}{5}\varepsilon.$$

We couple the random variables X and Y such that $\mathbb{P}(X \neq Y) = \frac{1}{2}\varepsilon^2$, and let $(X_i, Y_i)_{i=1}^N$ be an i.i.d. sample of the random vector (X, Y). Then, setting

$$\mathcal{A} := \{X_i = Y_i \text{ for all } i = 1, \ldots, N\},$$

it follows that

$$\mathbb{P}(\mathcal{A}) = \left(1 - \frac{\varepsilon^2}{2}\right)^N = \exp\left(N \log\left(1 - \frac{\varepsilon^2}{2}\right)\right) \geq \exp(-N\varepsilon^2),$$

where the inequality holds because $\varepsilon^2 \leq \frac{1}{4}$ and $\log(1-x) \geq -2x$ for $x \in [0, \frac{1}{8}]$.

Now let $\widehat{R}_N \colon \mathbb{R}^N \to \mathbb{R}$ be any estimator. Then, in the event $\mathcal{A}$,

$$\widehat{R}_N\left((X_i)_{i=1}^N\right) = \widehat{R}_N\left((Y_i)_{i=1}^N\right).$$

On the other hand, since $|\mathrm{ES}_\alpha(X) - \mathrm{ES}_\alpha(Y)| \geq \frac{1}{5}\varepsilon$, the (reverse) triangle inequality implies that it must either hold that

$$\left|\widehat{R}_N\left((X_i)_{i=1}^N\right) - \mathrm{ES}_\alpha(X)\right| \geq \frac{\varepsilon}{10} \text{ or } \left|\widehat{R}_N\left((Y_i)_{i=1}^N\right) - \mathrm{ES}_\alpha(Y)\right| \geq \frac{\varepsilon}{10},$$

which completes the proof. □

3.4. **The level of adversarial corruption.** Suppose that an adversary can modify at most K out of the N samples, where $K \lesssim \alpha N$. In Theorem 1.7 (setting $\varepsilon \sim \sqrt{K/N}$) we have shown that the estimator $\widehat{S}_N$ satisfies

$$\mathbb{P}\left(\left|\widehat{S}_N - \mathrm{ES}_\alpha(X)\right| \gtrsim \sqrt{\frac{K}{N}}\sigma_{\mathrm{ES}_\alpha(X)} + \frac{K}{N}L\right) \leq \exp(-cK).$$

In the next proposition we focus on the $\sqrt{\frac{K}{N}}$ term and show that the given dependence is minimax optimal.

Proposition 3.7. *Let $\alpha \leq \frac{2}{10}$, let $N \geq 10$, let $K \leq \frac{1}{2}\alpha N$, and let $\widehat{R}_N$ be any estimator. Then there exists a random variable X for which $\sigma_{\mathrm{ES}_\alpha(X)} = 1$ and F_X^{-1} is 1-Lipschitz on $[1-2\alpha, 1-\frac{1}{2}\alpha]$ and the following holds: with probability at least $1 - 2\exp(-\frac{K}{4})$, an adversary can modify K of the N samples in a way such that*

$$\left|\widehat{R}_N - \mathrm{ES}_\alpha(X)\right| \geq \frac{1}{20}\sqrt{\frac{K}{N}}\sigma_{\mathrm{ES}_\alpha(X)}.$$

Proof. Set $\varepsilon := \sqrt{\frac{K}{4N}}$. We use the same notation as the proof of Proposition 3.6; in particular X and Y are the Bernoulli random variables that appear in the proof of Proposition 3.6 so that $\mathbb{P}(X \neq Y) = \frac{1}{2}\varepsilon^2$.

Set

$$U := |\{i \in \{1, \ldots, N\} : X_i \neq Y_i\}|$$

so that $\mathbb{E}[U] = \frac{1}{2}\varepsilon^2 N$. It follows from Bernstein's inequality that for every $\lambda \geq 0$, with probability at least $1 - 2\exp(-\lambda)$,

$$\frac{U}{N} - \frac{\varepsilon^2}{2} = \frac{1}{N}\sum_{i=1}^N (\mathbb{1}_{X_i \neq Y_i} - \mathbb{E}[\mathbb{1}_{X_i \neq Y_i}]) \leq \sqrt{\frac{2\mathbb{V}\mathrm{ar}[\mathbb{1}_{X \neq Y}]\lambda}{N}} + \frac{2\lambda}{3N}.$$

In particular, noting that $\mathbb{V}\mathrm{ar}[\mathbb{1}_{X\neq Y}] \leq 2\varepsilon^2$ and applying Bernstein's inequality to $\lambda = \varepsilon^2 N$, it follows that with probability $1 - 2\exp(-\varepsilon^2 N)$, one has that $U \leq 4\varepsilon^2 N = K$.

Further, on the event $\{U \leq K\}$, the adversary can modify the sample of X in a way that it is equal to the sample of Y. Finally, we have seen in the proof of Proposition 3.6 that $|\mathrm{ES}_\alpha(X) - \mathrm{ES}_\alpha(Y)| \geq \frac{1}{5}\varepsilon$, from which the statement readily follows. □

4. Explicit computations in case of the Pareto distribution

Recall that X is said to have a Pareto distribution with parameters $\lambda > 0$ and $x_0 \in \mathbb{R}$ if $\mathbb{P}(X \leq t) = 1 - (\frac{x_0}{t})^\lambda 1_{[x_0,\infty)}(t)$ for $t \in \mathbb{R}$. Let us first prove the finite sample rate mentioned in the introduction.

Lemma 4.1. *Assume that X has a Pareto distribution with parameters $\lambda > 2$ and $x_0 > 0$. Then, for all $N \geq \frac{1}{\alpha}(\frac{\lambda-1}{\lambda})^{\frac{\lambda}{\lambda-1}}$, it holds that*

$$\mathbb{P}\left(\left|\widehat{T}_N - \mathrm{ES}_\alpha(X)\right| \geq \varepsilon\sigma_{\mathrm{ES}_\alpha}\right) \geq \frac{1}{2}\left(\frac{x_0}{\alpha\left(\mathrm{ES}_\alpha(X) + \varepsilon\sigma_{\mathrm{ES}_\alpha}\right)}\right)^\lambda \frac{1}{N^{\lambda-1}}.$$

Proof. Observe that for all $t > 0$,

$$\mathbb{P}\left(\widehat{T}_N \geq t\right) \geq \mathbb{P}\left(X_i \geq \alpha N t \text{ for some } 1 \leq i \leq N\right) =: (*).$$

Moreover, by the definition of the Pareto distribution, $\mathbb{P}(X \geq \alpha N t) = (\frac{x_0}{\alpha t N})^\lambda$. If $t > 0$ is such that $N^{\lambda-1} \geq (\frac{x_0}{\alpha t})^\lambda$, then $(\frac{x_0}{\alpha t N})^\lambda \leq \frac{1}{N}$ and we may apply Lemma 3.2, which yields

$$(*) \geq \frac{1}{2}\frac{x_0^\lambda}{\alpha^\lambda t^\lambda N^{\lambda-1}}.$$

Therefore, the proof is completed if we may apply the previous estimate to $t = \mathrm{ES}_\alpha(X) + \varepsilon\sigma_{\mathrm{ES}_\alpha}$ – in other words, if this choice of t satisfies that $N^{\lambda-1} \geq \left(\frac{x_0}{\alpha t}\right)^\lambda$. To that end, it suffices to note that

$$\mathrm{ES}_\alpha(X) = \frac{x_0\lambda}{\alpha^{1/\lambda}(\lambda-1)},$$

see, e.g, [31, Proposition 2]; thus the desired estimate $N^{\lambda-1} \geq \left(\frac{x_0}{\alpha t}\right)^\lambda$ follows from the assumption that $N^{\lambda-1} \geq \frac{1}{\alpha^{\lambda-1}}\left(\frac{\lambda-1}{\lambda}\right)^\lambda$. □

Lemma 4.2. *Assume that X has a Pareto distribution with parameters $\lambda > 2$ and $x_0 > 0$. Further set $r := \frac{\lambda+1}{\lambda}$ and let $\alpha \leq \frac{1}{2}$. Then the following hold.*

(i) F_X^{-1} is Lipschitz continuous on $[1 - 2\alpha, 1 - \alpha/2]$ with constant

$$L = \frac{x_0 2^r}{\lambda\alpha^r}.$$

(ii) We have that

$$\frac{x_0^2\alpha^{1-2r}}{2\lambda^2(3-2r)} \leq \sigma^2_{\mathrm{ES}_\alpha(X)} \leq \frac{2x_0^2\alpha^{1-2r}}{\lambda^2(r-1)(3-2r)}.$$

Remark 4.3. Recall that the error for the estimation of the expected shortfall risk scales like

$$|\widehat{S}_N - \mathrm{ES}_\alpha(X)| = \mathcal{O}(\varepsilon \sigma_{\mathrm{ES}_\alpha(X)}) + \mathcal{O}(\varepsilon^2 L), \tag{4.1}$$

see Theorem 1.3. Hence, the ratio $\frac{\sigma_{\mathrm{ES}_\alpha}}{L}$ is the threshold such that for all $\varepsilon > \frac{\sigma_{\mathrm{ES}_\alpha}}{L}$ the second error term in (4.1) is the dominant one, and for all $\varepsilon \leq \frac{\sigma_{\mathrm{ES}_\alpha}}{L}$ the first term in (4.1) is the dominant one; see also the discussion proceeding Corollary 1.6.

Lemma 4.2 shows that for the Pareto distribution, $\frac{\sigma_{\mathrm{ES}_\alpha}}{L} \geq \frac{\sqrt{\alpha}}{2^r\sqrt{2}} \geq \frac{1}{4}\sqrt{\alpha}$, i.e. for all $\varepsilon \leq \frac{1}{4}\sqrt{\alpha}$, the $\mathcal{O}(\varepsilon\sigma_{\mathrm{ES}_\alpha(X)})$ term in (4.1) is the dominant one.

Proof of Lemma 4.2. We start by proving the first claim. To that end, write $F = F_X$, denote by $f = F'$ the density of X, and note that

$$\frac{d}{du}F^{-1}(u) = \frac{1}{f(F^{-1}(u))}. \tag{4.2}$$

It is straightforward to verify that $f(t) = \lambda x_0^\lambda t^{-\lambda-1}$ for $t \geq x_0$ and that $F^{-1}(u) = x_0(1-u)^{-1/\lambda}$, from which the first claim readily follows.

As for the second claim, we assume that $x_0 = 1$ for simpler notation. By definition of $\sigma^2_{\mathrm{ES}_\alpha}$ and by a change of variables

$$\begin{aligned}\sigma^2_{\mathrm{ES}_\alpha(X)} &= \frac{1}{\alpha^2}\int_{F^{-1}(1-\alpha)}^\infty \int_{F^{-1}(1-\alpha)}^\infty F(\min\{t,s\}) - F(t)F(s)\,dtds \\ &= \frac{1}{\alpha^2}\int_{1-\alpha}^1\int_{1-\alpha}^1 \frac{\min\{u,v\}-uv}{\lambda^2(1-u)^{\frac{\lambda+1}{\lambda}}(1-v)^{\frac{\lambda+1}{\lambda}}}\,dudv.\end{aligned}$$

Indeed, we used the substitution $u = F(t)$ (hence $t = F^{-1}(u)$ and $dt = \frac{du}{f(F^{-1}(u))}$) and similarly $v = F(s)$. Recall that $r = \frac{\lambda+1}{\lambda}$ and set, for every $v \in [1-\alpha, 1)$,

$$A_v := \int_{1-\alpha}^v \frac{u(1-v)}{(1-u)^r}\,du \quad \text{and} \quad B_v := \int_v^1 \frac{v(1-u)}{(1-u)^r}\,du.$$

Then we have that

$$\sigma^2_{\mathrm{ES}_\alpha(X)} = \frac{1}{\alpha^2\lambda^2}\int_{1-\alpha}^1 \frac{A_v + B_v}{(1-v)^r}\,dv. \tag{4.3}$$

We start by proving the lower bound on $\sigma^2_{\mathrm{ES}_\alpha(X)}$. To that end, note that

$$B_v = \int_v^1 \frac{v}{(1-u)^{r-1}}\,du = \frac{-v}{(2-r)(1-u)^{r-2}}\Big|_v^1 = \frac{v}{(2-r)(1-v)^{r-2}}.$$

Moreover, for every $v \in [1-\alpha, 1]$ we have that $A_v \geq 0$ and $v \geq \frac{1}{2}$; thus by (4.3),

$$\sigma^2_{\mathrm{ES}_\alpha(X)} \geq \frac{1}{\alpha^2\lambda^2}\int_{1-\alpha}^1 \frac{1}{2(2-r)(1-v)^{2r-2}}\,dv = \frac{1}{\alpha^2\lambda^2 2(2-r)(3-2r)\alpha^{2r-3}}.$$

The claimed lower bound on $\sigma^2_{\mathrm{ES}_\alpha(X)}$ follows by noting that $r \in (1, \frac{3}{2})$ since $\lambda > 2$. The upper estimate follows from the same arguments, additionally noting that

$$A_v \leq (1-v) \int_{1-\alpha}^{v} \frac{1}{(1-u)^r}\, du \leq \frac{1}{(r-1)(1-v)^{r-2}} = \frac{2-r}{(r-1)v} B_v. \qquad \square$$

5. Numerical experiments

This section aims to shortly extend the numerical illustration of the proposed estimator from the introduction in a few more examples. For computational reasons, we always choose $\alpha = 0.1$ so that a significant proportion of the data lies in the tails and simulations are hence less costly. The proposed estimator always uses $\beta_1 = 0.5$ and $\beta_2 = 0.6$. This is chosen such that it works well for block sizes m around 250. As discussed in Remark 1.5, for smaller values of m, one would need to widen the interval $[\beta_1, \beta_2]$, while for larger values of m, one can narrow the interval.

First, Figure 2 aims to illustrate the intuition of $\widehat{S}_N$ given in Remark 1.5, in that it can harness the benefits of the median of blocks estimator ($\beta_1 = \beta_2 = 0.5$) for large deviations and high confidence, and the benefits of the plug-in estimator for small deviations and moderate confidence. Figure 3 shows that this behavior occurs for other heavy tailed distributions as well, and not just Pareto distributions.

Figure 4 showcases the behavior for more light-tailed distributions like the normal or (slightly less so) the log-normal distributions. In both cases, we can see that the plug-in estimator performs very well, and the proposed estimator almost matches its performance, while the median of blocks estimator performs worse in these well-behaved cases.

Figure 5 illustrates Theorem 1.7 by showcasing the robustness of the proposed estimator against adversarial manipulation of the data, whereas the plug-in estimator is heavily distorted even by manipulating just three data points.

Finally, Figure 6 illustrates the behavior of the proposed estimator in a setting with infinite variance $\sigma^2_{\mathrm{ES}_\alpha} = \infty$. While the theoretical guarantees from Theorem 1.3 fail, Figure 6 shows that even in this situation, our estimator performs well, and in particular much better than the plug-in estimator.

Acknowledgements: Daniel Bartl is grateful for financial support through the Austrian Science Fund (grant doi: 10.55776/ESP31 and 10.55776/P34743). The authors are grateful to Felix Liebrich and Gilles Stupfler for helpful comments and suggestions.

References

[1] C. Acerbi. Coherent measures of risk in everyday market practice. 2007.

[2] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29, 1996.

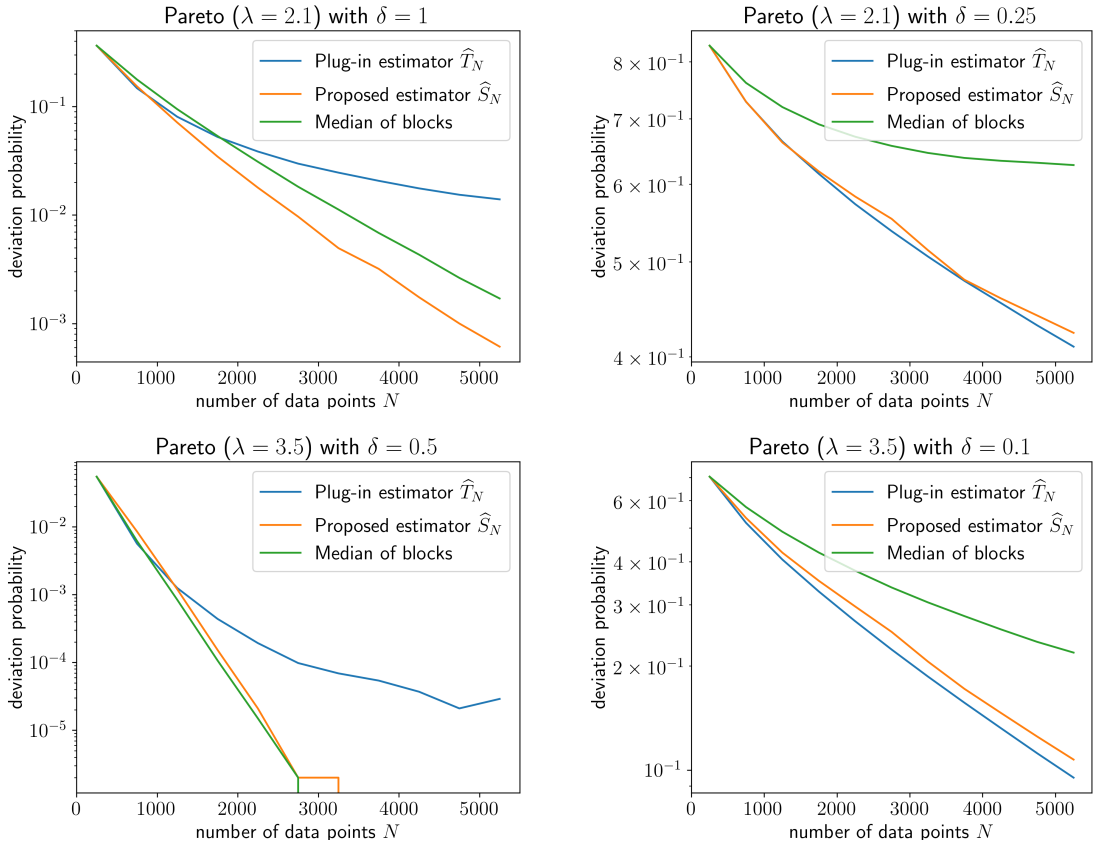

FIGURE 2. The figures show $\mathbb{P}(|\widehat{R}_N - \mathrm{ES}_\alpha(X)| \geq \delta)$ (estimated using 10^6 many experiments) for different estimators $\widehat{R}_N$, where X is either Pareto distributed with $\lambda = 2.1$ or $\lambda = 3.5$. The proposed estimator and the median of blocks estimator use sub-intervals of size $m = 250$. The left hand side illustrates that the confidence bands for $\hat{S}_N$ and the median of blocks estimator behave like $\exp(-c_\delta N)$ as shown in Theorem 1.3, which is in contrast to the rate $\tilde{c}_\delta N^{-(\lambda-1)}$ for $\widehat{T}_N$ as shown in Lemma 4.1. The figure further illustrates the regime switch for the proposed estimator. For large values of δ (left hand images, notice the scale on the y-axis), the proposed estimator behaves similarly to the median of blocks estimator; that is, statistically optimal with high confidence. For regimes with smaller δ (right hand images, notice the different scale on the y-axis), the proposed estimator manages to obtain an accuracy similar to the plug-in estimator.

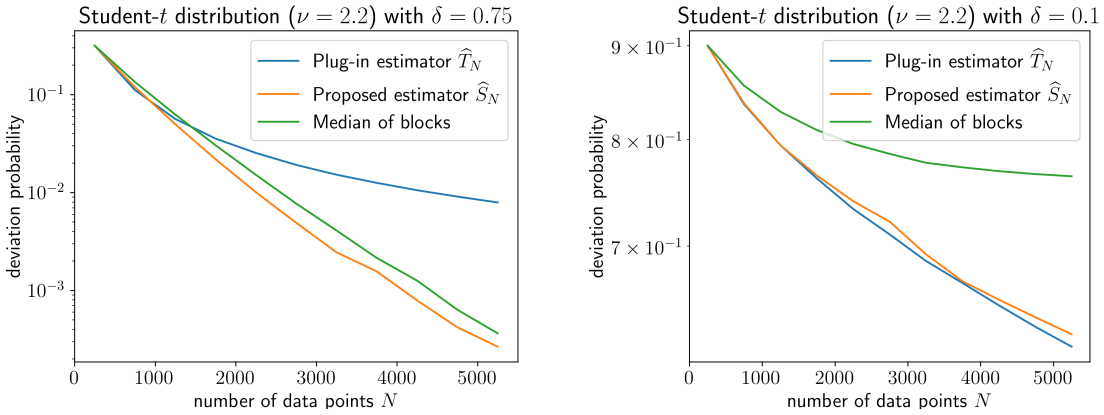

FIGURE 3. The figures show $\mathbb{P}(|\widehat{R}_N - \mathrm{ES}_\alpha(X)| \geq \delta)$ (estimated using 10^6 many experiments) for different estimators $\widehat{R}_N$, where X follows a student-t distribution with parameter $\nu = 2.2$. The proposed estimator and the median of blocks estimator use sub-intervals of size $m = 250$.

[3] D. W. Andrews. Stability comparison of estimators. *Econometrica: Journal of the Econometric Society*, pages 1207–1235, 1986.

[4] P. Artzner, F. Delbaen, J. M. Eber, and D. Heath. Coherent measures of risk. *Math. Finance*, 9:203–228, 1999.

[5] J. Beirlant and J. L. Teugels. Modeling large claims in non-life insurance. *Insurance: Mathematics and Economics*, 11(1):17–29, 1992.

[6] S. Boucheron, G. Lugosi, and O. Bousquet. Concentration inequalities. In *Summer school on machine learning*, pages 208–240. Springer, 2003.

[7] V. Brazauskas, B. L. Jones, M. L. Puri, and R. Zitikis. Estimating conditional tail expectation with actuarial applications in view. *Journal of Statistical Planning and Inference*, 138(11):3590–3604, 2008.

[8] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

[9] S. X. Chen. Nonparametric estimation of expected shortfall. *Journal of financial econometrics*, 6(1):87–107, 2008.

[10] A. Christmann, D. Xiang, and D.-X. Zhou. Total stability of kernel methods. *Neurocomputing*, 289:101–118, 2018.

[11] R. Cont, R. Deguest, and G. Scandolo. Robustness and sensitivity analysis of risk measurement procedures. *Quantitative finance*, 10(6):593–606, 2010.

[12] J. Dhaene, S. Vanduffel, M. J. Goovaerts, R. Kaas, Q. Tang, and D. Vyncke. Risk measures and comonotonicity: a review. *Stochastic models*, 22(4):573–606, 2006.

[13] D. Duffie and J. Pan. An overview of value at risk. *Journal of derivatives*, 4(3):7–49, 1997.

[14] S. Eckstein, A. Iske, and M. Trabs. Dimensionality reduction and wasserstein stability for kernel regression. *Journal of Machine Learning Research*, 24(334):1–35, 2023.

[15] P. Embrechts, G. Puccetti, L. Rüschendorf, R. Wang, and A. Beleraj. An academic response to basel 3.5. *Risks*, 2(1):25–48, 2014.

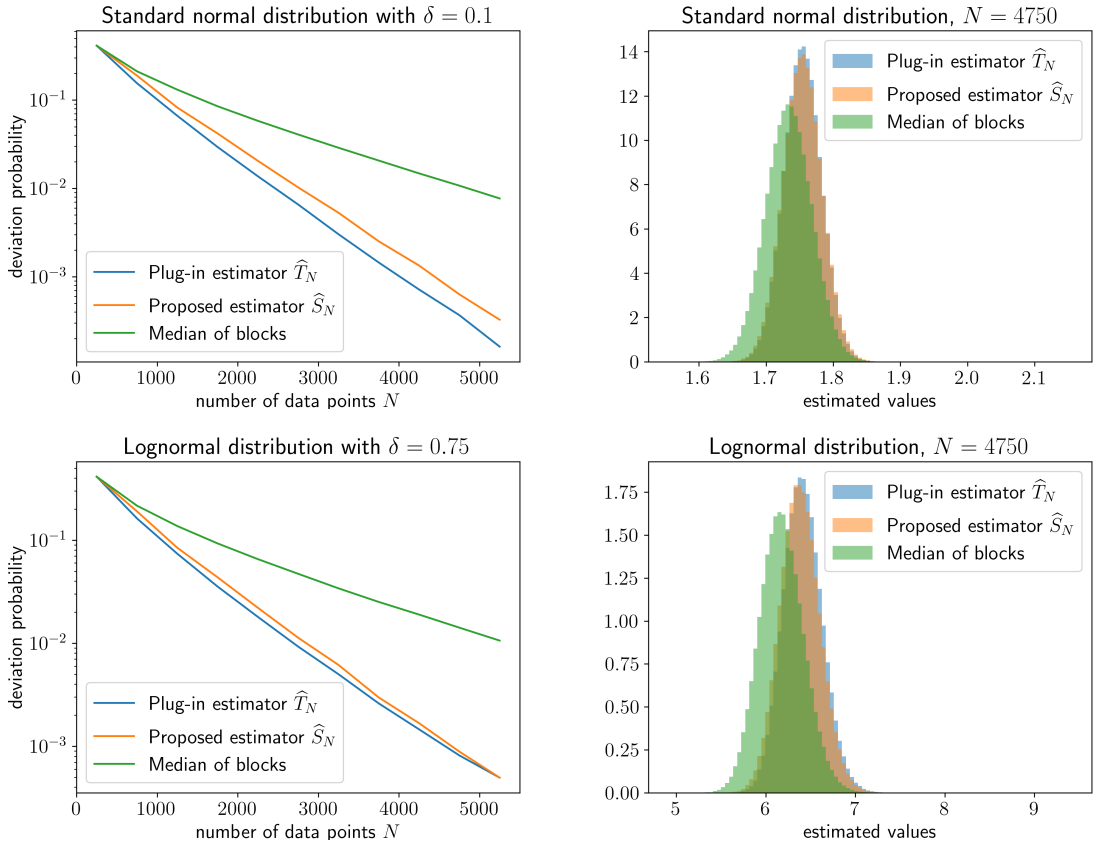

FIGURE 4. The left hand side shows $\mathbb{P}(|\widehat{R}_N - \mathrm{ES}_\alpha(X)| \geq \delta)$ (estimated using 10^6 many experiments) for different estimators $\widehat{R}_N$, where X is either standard normally distributed (top) and or standard log-normally distributed (bottom). The proposed estimator and the median of blocks estimator use sub-intervals of size $m = 125$. We see that the median of blocks estimator exhibits a noticeable negative bias, as shown by the histograms on the right hand side. The proposed estimator mitigates this downside almost completely and performs similarly to the plug-in estimator in these examples.

[16] W. Feller. Generalization of a probability limit theorem of cramér. *Transactions of the American Mathematical Society*, 54(3):361–372, 1943.

[17] I. Goodfellow, P. McDaniel, and N. Papernot. Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7):56–66, 2018.

[18] J. B. Hill. Expected shortfall estimation and gaussian inference for infinite variance time series. *Journal of Financial Econometrics*, 13(1):1–44, 2015.

[19] J. L. Horowitz. Bootstrap methods in econometrics. *Annual Review of Economics*, 11:193–224, 2019.

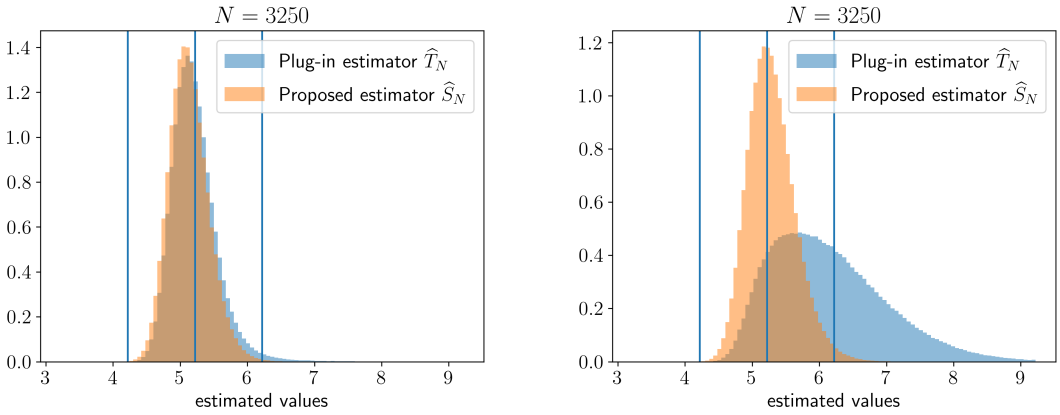

FIGURE 5. The figures show histograms across estimated values for $\mathrm{ES}_\alpha(X)$, where X is Pareto distributed with $\lambda = 2.2$, with the blue vertical lines depicting the values $\mathrm{ES}_\alpha(X) - 1, \mathrm{ES}_\alpha(X), \mathrm{ES}_\alpha(X) + 1$. The left hand side shows the standard case without any adversarial manipulation, as reported in the introduction. The right hand side shows the case in which three of the 3250 data points, w.l.o.g. X_1, X_2, X_3, are modified to $\tilde{X}_i = \max\{X_i, U_i\}$ with $U_i \sim \mathcal{N}(5, 250^2)$ independent of all other variables, $i = 1, 2, 3$. We see that the plug-in estimator is heavily distorted through those modified data points, while the proposed estimator is much more robust and its histogram remains close to the one using the uncorrupted data.

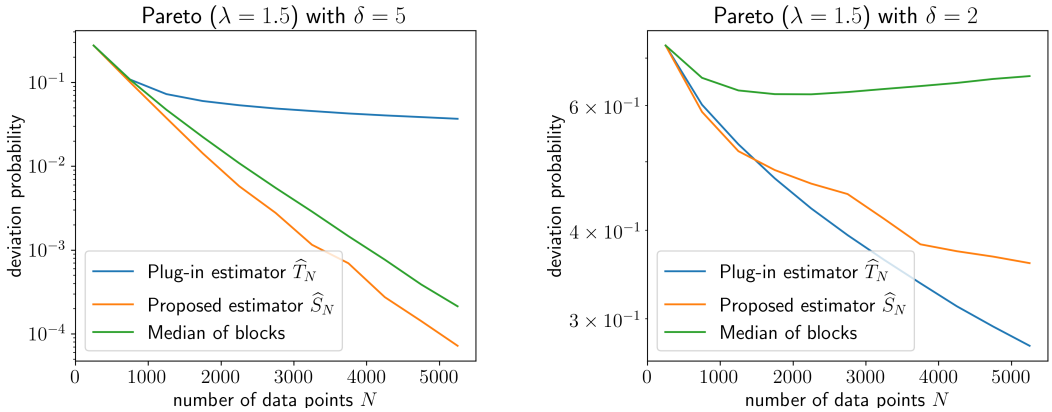

FIGURE 6. The figures show $\mathbb{P}(|\widehat{R}_N - \mathrm{ES}_\alpha(X)| \geq \delta)$ (estimated using 10^6 many experiments) for different estimators $\widehat{R}_N$, where X is Pareto distributed with parameter $\lambda = 1.5$. The proposed estimator and the median of blocks estimator use sub-intervals of size $m = 250$.

[20] Y. Hyndman, Rob J.; Fan. Sample quantiles in statistical packages. *The American Statistician 1996-nov vol. 50 iss. 4*, 50, nov 1996.
[21] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical computer science*, 43:169–188, 1986.
[22] B. L. Jones and R. Zitikis. Empirical estimation of risk measures and related quantities. *North American Actuarial Journal*, 7(4):44–54, 2003.
[23] J. Kim and C. D. Scott. Robust kernel density estimation. *The Journal of Machine Learning Research*, 13(1):2529–2565, 2012.
[24] A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press, 2015.
[25] S. Minsker. U-statistics of growing order and sub-gaussian mean estimators with sharp constants. *arXiv preprint arXiv:2202.11842*, 2022.
[26] S. Minsker. Efficient median of means estimator. *arXiv preprint arXiv:2305.18681*, 2023.
[27] S. Nadarajah, B. Zhang, and S. Chan. Estimation methods for expected shortfall. *Quantitative Finance*, 14(2):271–291, 2014.
[28] A. Necir, A. Rassoul, and R. Zitikis. Estimating the conditional tail expectation in the case of heavy-tailed losses. *Journal of Probability and Statistics*, 2010, 2010.
[29] A. S. Nemirovskij and D. B. Yudin. Problem complexity and method efficiency in optimization. 1983.
[30] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. Molloy, and B. Edwards. Adversarial robustness toolbox v1.2.0. *CoRR*, 1807.01069, 2018.
[31] M. Norton, V. Khokhlov, and S. Uryasev. Calculating cvar and bpoe for common probability distributions with application to portfolio optimization and density estimation. *Annals of Operations Research*, 299:1281–1315, 2021.
[32] G. Pflug and N. Wozabal. Asymptotic distribution of law-invariant risk functionals. *Finance and Stochastics*, 14(3):397–418, 2010.
[33] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Jounal of Risk*, 2:21–42, 2000.
[34] R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.
[35] P. J. Rousseeuw and M. Hubert. Robust statistics for outlier detection. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 1(1):73–79, 2011.
[36] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. John wiley & sons, 2005.
[37] E. N. Sereda, E. M. Bronshtein, S. T. Rachev, F. J. Fabozzi, W. Sun, and S. V. Stoyanov. Distortion risk measures in portfolio optimization. *Handbook of portfolio construction*, pages 649–673, 2010.
[38] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.
[39] H. Tsukahara. Estimation of distortion risk measures. *Journal of Financial Econometrics*, 12(1):213–235, 2014.
[40] M. V. Wüthrich and M. Merz. *Statistical foundations of actuarial learning and its applications*. Springer Nature, 2023.
[41] Y. Yamai and T. Yoshiba. Value-at-risk versus expected shortfall: A practical perspective. *Journal of Banking & Finance*, 29(4):997–1015, 2005.
[42] Y. Yamai, T. Yoshiba, et al. Comparative analyses of expected shortfall and value-at-risk: their estimation error, decomposition, and optimization. *Monetary and economic studies*, 20(1):87–121, 2002.

University of Vienna, faculty of mathematics
Email address: daniel.bartl@univie.ac.at

University of Tübingen, department of mathematics
Email address: stephan.eckstein@uni-tuebingen.de