ON DISTRIBUTIONALLY ROBUST EXTREME VALUE ANALYSIS

BLANCHET, J. AND MURTHY, K.

Abstract. We study distributional robustness in the context of Extreme Value Theory (EVT). As an application, we provide a data-driven method for estimating extreme quantiles in a manner that is robust against incorrect model assumptions underlying the application of the standard Extremal Types Theorem. Typical studies in distributional robustness involve computing worst case estimates over a model uncertainty region expressed in terms of the Kullback-Leibler discrepancy. We go beyond standard distributional robustness in that we investigate different forms of discrepancies, and prove rigorous results which are helpful for understanding the role of a putative model uncertainty region in the context of extreme quantile estimation. Finally, we illustrate our data-driven method in various settings, including examples showing how standard EVT can significantly underestimate quantiles of interest.

1. Introduction

Extreme Value Theory (EVT) provides reasonable statistical principles which can be used to extrapolate tail distributions, and, consequently, estimate extreme quantiles. However, as with any form for extrapolation, extreme value analysis rests on assumptions that are rather difficult (or impossible) to verify. Therefore, it makes sense to provide a mechanism to robustify the inference obtained via EVT.

The goal of this paper is to study non-parametric distributional robustness (i.e. finding the worst case distribution within some discrepancy of a natural baseline model) in the context of EVT. We ultimately provide a data-driven method for estimating extreme quantiles in a manner that is robust against possibly incorrect model assumptions. Our objective here is different from standard statistical robustness which is concerned with data contamination only (not model error); see, for example, [22], for this type of analysis in the setting of EVT.

Our focus in this paper is closer in spirit to distributionally robust optimization as in, for instance, [10, 2, 3]. However, in contrast to the literature on robust optimization, the emphasis here is on understanding the implications of distributional uncertainty regions in the context of EVT. As far as we know this is the first paper that studies distributional robustness in the context of EVT.

We now describe the content of the paper, following the logic which motivates the use of EVT.

1.1. Motivation and Standard Approach.

In order to provide a more detailed description of the content of this paper, its motivations, the specific contributions, and the methods involved, let us invoke a couple of typical examples which motivate the use of extreme value theory. As a first example consider wind speed data observed over the last 200 years in New York City, engineers need to forecast the necessary strength that is required for a skyscraper to withstand a wind speed that gets exceeded only about once in 1,000 years. In another instance, given the losses observed during the last few decades, a reinsurance firm may want

to compute, as required by Solvency II standard, a capital requirement that is needed to withstand all but about one loss in 200 yrs.

These tasks, and many others in practice, present a common challenge of extrapolating tail distributions over regions involving unobserved evidence from available observations. There are many reasonable ways of doing these types of extrapolations. One might take advantage of physical principles and additional information, if available, in the windspeed setting; or use economic principles in the reinsurance setting. In the absence of any fundamental principles which inform tail extrapolation, one might opt to use purely statistical considerations. This is the motivation behind EVT.

We shall provide a quick review of EVT in Section 2, here we discuss intuitively its principles. In its basic form, the most fundamental result in extreme value theory, known as the Fisher-Tippett-Gnedenko (FTG) Theorem or the Extremal types Theorem, postulates that n independent and identically distributed (iid) measurements, $X_1, ..., X_n$, representing, for instance, wind speeds, can be summarized by an affine transformation (depending on n) of a single random source. In other words, the FTG Theorem postulates the approximation in distribution

$$\max(X_1, ..., X_n) \overset{D}{\approx} b_n + a_n Z(\gamma), \tag{1}$$

where $(a_n, b_n)_{n\geq 1}$ are suitable deterministic sequences, and $Z(\gamma)$ is a fixed random variable parameterized by $\gamma \in \mathbb{R}$. The FTG Theorem concludes that $Z(\gamma)$ belongs to a specific parametric family, which is given the name of Generalized Extreme Value (GEV) distribution. We shall write $G_\gamma(z) := \Pr(Z(\gamma) \leq z)$.

The parameter γ ranges over the regions $\gamma < 0$, $\gamma = 0$, and $\gamma > 0$. These regions correspond to to the following cases. The case $\gamma < 0$ roughly corresponds to the case in which the variable X_i is bounded to the right (i.e. $\Pr(X_i \leq m_0) = 1$ for some $m_0 < \infty$; the case $\gamma = 0$ roughly corresponds to semiexponentially decaying tails (i.e. $\Pr(X_i > t) = \exp(-t^\alpha + o(t^\alpha))$ for some $\alpha > 0$), and the case $\gamma > 0$ corresponds to power-law-type decaying tails, basically, $\Pr(X_i > t) = ct^{-\alpha}(1 + o(1))$. These distinctions inform the risk perception of the users of extreme value theory, making fat tails, for instance, sources of high perceived risk relative to light tails.

Due to its simplicity, the affine form (1) can be easily used for extrapolation. Given N independent samples of a random variable X, let us say our objective is to compute the quantile x_p such that $\Pr\{X > x_p\} = 1 - p$, for some p close to 1. One could divide the data into blocks of size n and compute the maxima in each block. Using these $m = \lfloor N/n \rfloor$ samples of block-maxima M_n, say $M_n^1, ..., M_n^m$ one uses the approximation (1). In other words, one can use these m data points, combined with maximum likelihood to estimate the corresponding shape, scale, and location parameters (γ, a, and b, respectively) of the distribution $G_\gamma(a\times\cdot + b)$. At this point one is basically taking approximation (1) as an exact distributional identity. This is the main source of model error and quantifying this error is difficult because the rate of convergence in (1) depends on the behavior of the underlying density of X. Once we have a satisfactory model for block-maxima M_n, computing the desired quantile or value at risk (VaR) of X is simple, because of the relation $G_\gamma(ax + b) \approx \Pr\{M_n \leq x\} = \Pr\{X \leq x\}^n$.

Dividing the data into blocks, although asymptotically correct, is not the most common technique. A more standard approach is the use of the peaks-over-threshold technique. Regardless of the technique used, the point is that various assumptions behind the FTG Theorem, including approximation (1), might be subject to model error. Consequently, it has been widely accepted that tail risk measures, particularly for high confidence levels, can only be estimated with considerable statistical as well as model uncertainty (see, for example,

[13]). Moreover, the following remark due to [4] holds significance in this discussion: "Though the GEV model is supported by mathematical argument, its use in extrapolation is based on unverifiable assumptions, and measures of uncertainty on return levels should properly be regarded as *lower bounds* that could be much greater if uncertainty due to model correctness were taken into account."

Despite these difficulties, however, EVT is widely used (see, for example, [6]) and regarded as a reasonable way of extrapolation to estimate extreme quantiles.

1.2. **Proposed Approach Based on Infinite Dimensional Optimization.** We share the point of view that EVT is a reasonable approach, so we propose a procedure that builds on the use of EVT to provide upper bounds which address the types of errors discussed in the remark above from [4]. For large values of n, under the assumptions of EVT, the distribution of M_n lies close to, and appears like, a GEV distribution. Therefore, instead of considering only the GEV distribution as a candidate model, we propose a non-parametric approach. In particular, we consider a family of probability models, all of which lie in a "neighborhood" of a GEV model, and compute a conservative worst-case estimate of VaR over all of these candidate models.

Mathematically, given a reference model, P_{ref}, which we consider to be obtained using EVT (using a procedure such as the one outlined in the previous subsection), we consider the optimization problem

$$\sup\Big\{P\{X > x\} :\ d(P, P_{ref}) \le \delta\Big\}. \tag{2}$$

Note that the previous problem proposes optimizing over all probability measures that are within a tolerance level δ (in terms of a suitable discrepancy measure d) from the chosen baseline reference model P_{ref}.

There is a wealth of literature that pursues this line of thought (see [10, 2, 3, 23, 8, 11]), but, no study has been carried out in the context of EVT. Moreover, while the solvability of problems as in (2) have understandably received a great deal of attention, the qualitative differences that arise by using various choices of discrepancy measures, d, has not been explored, and this is an important contribution of this paper. For tractability reasons, the usual choice for discrepancy d in the literature has been KL-divergence. In Section 3 we study the solution to infinite dimensional optimization problems such as (2) for a large class of discrepancies that include KL-divergence, and discuss how such problems can be solved at no significant computational cost.

1.3. **Choosing Discrepancy and Consistency Results.** One of our main contributions in this paper is to systematically demonstrate the qualitative differences that arise by using different choices of discrepancy measures d in (2). Since our interest in the paper is limited to robust tail modeling via EVT, this narrow scope, in turn, lets us analyse the qualitative differences that may arise because of different choices of d.

As mentioned earlier, the KL-divergence[1] is the most popular choice for d. In Section 4 we show that for any divergence neighborhood $\mathcal{P}$, defined using $d =$ KL-divergence around a baseline reference P_{ref}, there exists a probability measure P in $\mathcal{P}$ that has tails as heavy as

$$P(x, \infty) \ge c \log^{-2} P_{ref}(x, \infty),$$

[1]KL-divergence, and all other relevant divergence measures, are defined in Section 3.1

for a suitable constant c, and all large enough x. This means, irrespective of how small δ is (smaller δ corresponds to smaller neighborhood $\mathcal{P}$), a KL-divergence neighborhood around a commonly used distribution (such as exponential, (or) Weibull (or) Pareto) typically contains tail distributions that have infinite mean or variance, and whose tail probabilities decay at an unrealistically slow rate (even logarithmically slow, like $\log^{-2} x$, in the case of reference models that behave like a power-law or Pareto distribution). As a result, computations such as worst-case expected short-fall[2] may turn out to be infinite. Such worst-case analyses are neither useful nor interesting.

For our purposes, we also consider a general family of divergence measures D_α that includes KL-divergence as a special case (when $\alpha = 1$). It turns out that for any $\alpha > 1$, the divergence neighborhoods defined as in $\{P : D_\alpha(P, P_{_{ref}}) \leq \delta\}$ consists of tails that are heavier than $P_{_{ref}}$, but not prohibitively heavy. More importantly, we prove a "consistency" result in the sense that if the baseline reference model belongs to the domain of attraction of a GEV distribution with shape parameter $\gamma_{_{ref}}$, then the corresponding worst-case tail distribution,

$$\bar{F}_\alpha(x) := \sup\{P(x, \infty) : D_\alpha(P, P_{_{ref}}) \leq \delta\}, \tag{3}$$

belongs to the domain of attraction of a GEV distribution with shape parameter $\gamma^* = (1 - \alpha^{-1})^{-1}\gamma_{_{ref}}$ (if it exists).

Since our robustification approach is built resting on EVT principles, we see this consistency result as desirable. If a modeler who is familiar with certain type of data expects the EVT inference to result in an estimated shape parameter which is positive, then the robustification procedure should preserve this qualitative property. An analysis of the domain of attraction of the distribution $\bar{F}_\alpha(x)$, depending on α and $\gamma_{_{ref}}$, is presented in Section 4, along with a summary of the results in Table 1.

Note that the smaller the value of α, the larger the absolute value of shape parameter γ^*, and consecutively, heavier the corresponding worst-case tail is. This indicates a gradation in the rate of decay of worst-case tail probabilities as parameter α decreases to 1, with the case $\alpha = 1$ (corresponding to KL-divergence) representing the extreme heavy-tailed behaviour. This gradation, as we shall see, offers a great deal of flexibility in modeling by letting us incorporate domain knowledge (or) expert opinions on the tail behaviour. If a modeler is suspicious about the EVT inference he/she could opt to select $\alpha = 1$, but, as we have mentioned earlier, this selection may result in pessimistic estimates.

The relevance of these results shall become more evident as we introduce the required terminology in the forthcoming sections. Meanwhile, Table 1 and Figure 1 offer illustrative comparisons of $\bar{F}_\alpha(x)$ for various choices of α.

1.4. **The Final Estimation Procedure.** The framework outlined in the previous subsections yields a data driven procedure for estimating VaR which is presented in Section 5. A summary of the overall procedure is given in Algorithm 2. The procedure is applied to various data sets, resulting in different reference models, and we emphasize the choice of different discrepancy measures via the parameter α. The numerical studies expose the salient points discussed in the previous subsections and rigorously studied via our theorems. For instance, Example 3 shows how the use of the KL divergence might lead to rather pessimistic estimates. Moreover, Example 4 illustrates how the direct application of EVT can severely

[2]Similar to VaR, expected shortfall (or) conditional value at risk (referred as CVaR) is another widely recognized risk measure.

underestimate the quantile of interest, while the procedure that we advocate provides correct coverage for the extreme quantile of interest.

The very last section of the paper, Section 6, contains technical proofs of various results invoked in our development.

2. The extremal types theorem and its application to the estimation of extreme quantiles

Recall that the Central Limit Theorem characterizes the limiting distribution that may arise for normalized sums of iid random variables. On similar lines, the Extremal types theorem identifies the non-trivial limiting distributions that may result for maxima of random variables normalized as in

$$\lim_{n\to\infty} \frac{M_n - b_n}{a_n}.$$

Here, M_n represents the maxima of n independent copies of a random variable X, and a_n, b_n are suitable scaling constants. If we let $F(x) = \Pr\{X \le x\}$ be the distribution function of x, then extreme value theory identifies all non-degenerate distributions $G(\cdot)$ that may occur in the limiting relationship,

$$\lim_{n\to\infty} P\left\{\frac{M_n - b_n}{a_n} \le x\right\} = \lim_{n\to\infty} F^n\left(a_n x + b_n\right) = G(x), \tag{4}$$

for every continuity point x of $G(\cdot)$. All such distributions $G(x)$ that occur in the right-hand side of (4) are called *extreme value distributions.*

Extremal types theorem (Fisher and Tippet (1928), Gnedenko(1943)). The class of extreme value distributions is $G_\gamma(ax+b)$ with $a > 0, b, \gamma \in \mathbb{R}$, and

$$G_\gamma(x) := \exp\left(-\left(1+\gamma x\right)^{-1/\gamma}\right), \qquad 1+\gamma x > 0. \tag{5}$$

If $\gamma = 0$, the right-hand side is interpreted as $\exp(-\exp(-x))$.

The extremal types theorem asserts that any $G(x)$ that occurs in the right-hand side of (4) must be of the form $G_\gamma(ax+b)$. As a convention, any probability distribution $F(x)$ that gives rise to the limiting distribution $G(x) = G_\gamma(ax+b)$ in (4) is said to belong to the domain of attraction of $G_\gamma(x)$. In short, it is written as $F \in \mathcal{D}(G_\gamma)$. The parameters γ, a and b are, respectively, called the shape, scale and location parameters.

2.1. The different domains of attraction.

Though the limiting distributions $G_\gamma(ax+b)$ seem to constitute a simple parametric family, they include a wide-range of tail behaviours in their domains of attraction, as discussed below: For a distribution F, let $\bar{F}(x) = 1 - F(x)$ denote the corresponding tail probabilities, and $x_F^* = \sup\{x : F(x) < 1\}$ denote the right endpoint of its support.

1) **The Frechet Case** ($\gamma > 0$). A distribution $F \in \mathcal{D}(G_\gamma)$ for some $\gamma > 0$, if and only if right endpoint x_F^* is unbounded, and its tail probabilities satisfy

$$\bar{F}(x) = \frac{L(x)}{x^{1/\gamma}}, \qquad x > 0 \tag{6}$$

for a function $L(\cdot)$ slowly varying at ∞[3]. As a consequence, moments greater than or equal to $1/\gamma$ do not exist. Any distribution $F(x)$ that lies in $\mathcal{D}(G_\gamma)$ for some $\gamma > 0$ is also said to belong to the domain of attraction of a Frechet distribution with parameter $1/\gamma$. The Pareto distribution $1 - F(x) = x^{-\alpha} \wedge 1$ is an example for a distribution that belongs to the domain of attraction of $G_{1/\alpha}(x)$.

2) **The Weibull case** ($\gamma < 0$). Unlike the Frechet case, a distribution $F \in \mathcal{D}(G_\gamma)$ for some $\gamma < 0$, if and only if its right endpoint x_F^* is finite, and its tail probabilities satisfy

$$\bar{F}(x_F^* - \epsilon) = \epsilon^{-1/\gamma} L\left(\frac{1}{\epsilon}\right), \qquad \epsilon > 0 \tag{7}$$

for a function $L(\cdot)$ slowly varying at ∞. A distribution that belongs to the domain of attraction of $G_\gamma(x)$ for some $\gamma < 0$ is also said to belong to the domain of attraction of Weibull family. The uniform distribution on the interval $[0, 1]$ is an example that belongs to this class of extreme value distributions.

3) **The Gumbel case** ($\gamma = 0$). A distribution $F \in \mathcal{D}(G_0)$ if and only if

$$\lim_{t \uparrow x_F^*} \frac{\bar{F}(t + xf(t))}{\bar{F}(t)} = \exp(-x), \qquad x \in \mathbb{R} \tag{8}$$

for a suitable positive function $f(\cdot)$. In general, the members of G_0 have exponentially decaying tails, and consequently, all moments exist. Probability distributions $F(\cdot)$ that give rise to limiting distributions $G_0(ax + b)$ are also said to belong to the Gumbel domain of attraction. Common examples that belong to the Gumbel domain of attraction include exponential and normal distributions.

Given a distribution function F, Proposition 1 is an useful to test to determine its domain of attraction:

Proposition 1. Suppose $F''(x)$ exists and $F'(x)$ is positive for all x in some left neighborhood of x_F^*. If

$$\lim_{x \uparrow x_F^*} \left(\frac{1 - F}{F'}\right)'(x) = \gamma, \tag{9}$$

then F belongs to the domain of attraction of G_γ.

Further details on the classification of extreme value distributions into Frechet, Gumbel and Weibull cases, and proofs of the above statements can be found in a standard text on extreme value theory (see, for example, [14] or [6]).

2.2. **A model for the maxima of random variables.** The family of extreme value distributions $\{G_\gamma(ax + b) : a > 0, b, \gamma \in \mathbb{R}\}$ is also popularly referred as the GEV (generalized extreme value) family. One can perhaps interpret the limit in (4) as an approximation for large values of n as below:

$$\Pr\{M_n \leq z\} \approx G\left(\frac{z - b_n}{a_n}\right).$$

The difficulty that the normalizing constants a_n and b_n are not known can be eradicated by identifying that $H(x) := G(a_n^{-1}x - a_n^{-1}b_n)$ also belongs to the GEV family. Therefore, for

[3]A function $L : \mathbb{R} \to \mathbb{R}$ is said to be slowly varying at infinity if $\lim_{x\to\infty} L(tx)/L(x) = 1$ for every $t > 0$. Common examples of slowly varying function include $\log x, \log\log x, 1 - \exp(-x)$, constants, etc.

large values of n, the GEV family consists of a candidate distribution that serves as a good proxy for the distribution of the maximum M_n.

2.3. **Estimation of extreme quantiles.** Let X be a random variable with unknown distribution function $F(\cdot)$. Given N independent observations of a random variable X and a level p such that $1-p$ is comparable to (or) smaller than $1/N$, our objective is to estimate the quantile

$$\mathrm{VaR}_p(X) = F^{\leftarrow}(p) := \inf\{x : P\{X \leq x\} \geq p\}.$$

Since the tail regions of the probability distribution of X are not observed enough in the data, one has to to resort to additional assumptions (or) extrapolation techniques. As mentioned in the Introduction, a commonly used extrapolation method involves the use of extremal types theorem: Recall from Section 2.2 that for sufficiently large values of n, the GEV family contains a candidate distribution that well-approximates the distribution of the maximum M_n. As a result, a standard practice to estimate $\mathrm{VaR}_p(X)$ is to first calibrate a GEV model for the maxima, and compute the corresponding quantile in the calibrated GEV distribution. A rough sketch of the procedure is outlined below in Algorithm 1.

Algorithm 1 To estimate $\mathrm{VaR}_p(X)$ for values of p close to 1
Given: N independent samples $X_1, \ldots, X_N$ of X, and a level p close to 1

Initialize $n < N$ and let $m = \lfloor \frac{N}{n} \rfloor$.

Step 1 (Compute block-maxima): Partition $X_1, \ldots, X_N$ into blocks of size n, and compute the block maxima for each block to obtain samples $M_{n,1}, \ldots, M_{n,m}$ of maxima M_n.

Step 2 (Calibrate a GEV model): Treat the samples $M_{n,1}, \ldots, M_{n,m}$ as independent samples coming from a member of the GEV family and use a parameter estimation technique (for example, maximum-likelihood) to estimate the parameters a_0, b_0 and γ_0.

Step 3 (Compute the p^n-th quantile of the GEV model): Solve for x such that $G_{\gamma_0}(a_0 x + b_0) = p^n$, and let x_p be the corresponding solution.

RETURN x_p.

2.4. **On model errors and robustness.** For brevity, let P_{GEV} denote the probability measure corresponding to the distribution $G_{\gamma_0}(a_0 x + b_0)$ calibrated in Step 2 of Algorithm 1. If the maxima M_n is indeed distributed according to P_{GEV}, then

$$P_{GEV}(-\infty, x] = P\{M_n \leq x\} = P\{X \leq x\}^n,$$

in which case, the estimate returned by Algorithm 1 is the desired quantile $F^{\leftarrow}(p)$. However, $P_{GEV}(-\infty, x]$ is only an approximation of $P\{M_n \leq x\}$, and the quality of the approximation is, in turn, dependent on the unknown distribution function F (see [21, 6]). Therefore, in practice, one does not know the block-size n for which the GEV model P_{GEV} well-approximates the distribution of M_n. Even if a good choice of n is known, one cannot often employ it in practice, because larger n means smaller m, and consequentially, the inferential errors could be large. In addition, as discussed in the Introduction, the estimate x_p returned by Algorithm 1 can only be taken as a lower bound of the actual quantile estimate due to the uncertainties in the choice of the model. Due to the arbitrariness in the estimation

procedure and the nature of applications (calculating wind speeds for building sky-scrapers, building dykes for preventing floods, etc.), it is desirable to have, in addition, a data-driven procedure that yields a conservative upper bound for x_p that is robust against model errors. To accomplish this, one can form a collection of competing probability models $\mathcal{P}$, all of which appear plausible as the distribution of M_n, and compute the maximum of p^n-th quantile over all the plausible models in $\mathcal{P}$. This is indeed the objective of the forthcoming sections.

3. A non-parametric framework for addressing model errors

Let $(\Omega, \mathcal{F})$ be a measurable space and $M_1(\mathcal{F})$ denote the set of probability measures on $(\Omega, \mathcal{F})$. Let us assume that a reference probability model $P_{_{ref}} \in M_1(\mathcal{F})$ is inferred by stochastic modelling and standard estimation procedures from historical data. Naturally, this model is not the same as the true model that generates the data, and is expected only to be close to the true model. In the context of Section 2, the model $P_{_{ref}}$ corresponds to $P_{_{GEV}}$, and the data generating model corresponds to the true distribution of M_n. With slight perturbations in data, we would, in turn, be working with a slightly different reference model. Therefore, it has been of recent interest to consider a family of probability models $\mathcal{P}$, all of which are plausible, and perform computations over all the models in that family. Following the rich literature of robust optimization, where it is common to describe the set of plausible models using distance measures (see [2, 12]), we consider the set of plausible models to be of the form

$$\mathcal{P} = \left\{P \in M_1(\mathcal{F}) : d\left(P, P_{_{ref}}\right) \leq \delta\right\}$$

for some distance functional $d : M_1(\mathcal{F}) \times M_1(\mathcal{F}) \to \mathbb{R}$, and a suitable $\delta > 0$. Since $d(P_{_{ref}}, P_{_{ref}}) = 0$ for any reasonable distance functional, $P_{_{ref}}$ lies in $\mathcal{P}$. Therefore, for any random variable X, along with the conventional computation of $E_{P_{_{ref}}}[X]$, one aims to provide "robust" bounds,

$$\inf_{P\in\mathcal{P}} E_P[X] \leq E_{P_{_{ref}}}[X] \leq \sup_{P\in\mathcal{P}} E_P[X].$$

Here, we follow the notation that $E_P[X] = \int X dP$ for any $P \in M_1(\mathcal{F})$. Since the state-space Ω is uncountable, evaluation of the above sup and inf-bounds, in general, are infinite-dimensional problems. However, as it has been shown in the recent works [3, 8], it is indeed possible to evaluate these robust bounds for carefully chosen distance functionals d.

3.1. Divergence measures.

Consider two probability measures P and Q on $(\Omega, \mathcal{F})$ such that P is absolutely continuous with respect to Q. The Radon-Nikodym derivative dP/dQ is then well-defined. The Kullback-Liebler divergence (or KL-divergence) of P from Q is defined as

$$D_1(P, Q) := E_Q\left[\frac{dP}{dQ}\log\left(\frac{dP}{dQ}\right)\right]. \tag{10}$$

This quantity, also referred to as relative entropy (or) information divergence, arises in various contexts in probability theory. For our purposes, it will be useful to consider a general class of divergence measures that includes KL-divergence as a special case. For any $\alpha > 1$, the Rényi divergence of degree α is defined as:

$$D_\alpha(P, Q) := \frac{1}{\alpha - 1}\log E_Q\left[\left(\frac{dP}{dQ}\right)^\alpha\right]. \tag{11}$$

It is easy to verify that for every α, $D_\alpha(P, Q) = 0$, if and only if $P = Q$. Additionally, the map $\alpha \mapsto D_\alpha$ is nondecreasing, and continuous from the left. Letting $\alpha \to 1$ in (11)

yields the formula for KL-divergence $D_1(P,Q)$. Thus KL-divergence is a special case of the family of Rényi divergences, when the parameter $\alpha = 1$. If the probability measure P is not absolutely continuous with respect to Q, then $D_\alpha(P,Q)$ is taken as ∞. Though none of these divergence measures form a metric on the space of probability measures, they have been used in a variety of scientific disciplines to discriminate between probability measures. For more details on the divergences D_α, see [20, 15].

3.2. **Robust bounds via maximization of convex integral functionals.** Recall that $P_{_{ref}}$ is the reference probability measure arrived via standard estimation procedures. Since the model $P_{_{ref}}$ could be misspecified, we consider all models that are not far from $P_{_{ref}}$ in the sense quantified by divergence D_α, for any fixed $\alpha \geq 1$. Given a random variable X, we consider optimization problems of form

$$V_\alpha(\delta) := \sup\left\{E_{_P}[X] : D_\alpha(P, P_{_{ref}}) \leq \delta\right\}. \tag{12}$$

Though KL-divergence has been a popular choice in defining sets of plausible probability measures as above, use of divergences $D_\alpha,\ \alpha \neq 1$ is not new altogether: see [1, 8]. Due to Radon-Nikodym theorem, $V_\alpha(\delta)$ can be alternatively written as,

$$V_\alpha(\delta) = \sup\left\{E_{_{P_{ref}}}[LX] : E_{_{P_{ref}}}[\phi_\alpha(L)] \leq \bar{\delta}, E_{_{P_{ref}}}[L] = 1, L \geq 0\right\}, \tag{13}$$

where

$$\phi_\alpha(x) = \begin{cases} x^\alpha & \text{if } \alpha > 1, \\ x\log x & \text{if } \alpha = 1 \end{cases} \quad \text{and} \quad \bar{\delta} = \begin{cases} \exp\left((\alpha-1)\delta\right) & \text{if } \alpha > 1, \\ \delta & \text{if } \alpha = 1. \end{cases} \tag{14}$$

A standard approach for solving optimization problems of the above form is to write the corresponding dual problem as below:

$$V_\alpha(\delta) \leq \inf_{\substack{\lambda>0,\\ \mu}} \sup_{L\geq 0} E_{_{P_{ref}}}\left[LX - \lambda\left(\phi_\alpha(L) - \bar{\delta}\right) + \mu(L-1)\right].$$

The above dual problem can, in turn, be relaxed by taking the sup inside the expectation:

$$V_\alpha(\delta) \leq \inf_{\substack{\lambda>0,\\ \mu}} \left\{\lambda\bar{\delta} - \mu + \lambda E_{_{P_{ref}}}\left[\sup_{L\geq 0}\left\{\frac{(X+\mu)}{\lambda}L - \phi_\alpha(L)\right\}\right]\right\}.$$

It can be easily verified that the inner supremum is solved by

$$L^*_\alpha(c_1,c_2) := \begin{cases} c_1\exp(c_2 X), & \text{if } \alpha = 1, \\ (c_1 + c_2 X)_+^{1/(\alpha-1)}, & \text{if } \alpha > 1, \end{cases} \tag{15}$$

for suitable constants c_1 and c_2. Then the following theorem is intuitive:

Theorem 2. *Fix any $\alpha \geq 1$. For $L^*_\alpha(c_1,c_2)$ defined as in (15), if there exists constants c_1 and c_2 such that*

$$L^*_\alpha(c_1,c_2) \geq 0,\ E_{_{P_{ref}}}[L^*_\alpha(c_1,c_2)] = 1 \text{ and } E_{_{P_{ref}}}\left[\phi_\alpha\left(L^*_\alpha(c_1,c_2)\right)\right] = \bar{\delta},$$

then $L^*_\alpha(c_1,c_2)$ solves the optimization problem (13). The corresponding optimal value is

$$V_\alpha(\delta) = E_{_{P_{ref}}}[L^*_\alpha(c_1,c_2)X]. \tag{16}$$

Remark 3. Let us say one can determine constants c_1 and c_2 for given X, α and δ. Then, as a consequence of Theorem 2, the optimization problem (12) involving uncountably many measures can, in turn, be solved by simply simulating X from the original reference measure $P_{_{ref}}$, and multiplying by corresponding $L^*_\alpha(c_1, c_2)$ to compute the expectation as in (16).

A general theory for optimizing convex integral functionals of form (13), that includes a bigger class of general divergence measures, can be found in [3]. Theorem 2 is a simply a corollary of Theorem 4.2 of [3]. If the random variable X above is an indicator function, then computation of bounds $V_\alpha(\delta)$ turns out to be even simpler, as illustrated in the example below:

Example 1. Let $P_{_{ref}}$ be a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For a given $\delta > 0$ and $\alpha \geq 1$, let us say we are interested in evaluating the worst-case tail probabilities

$$\bar{F}_\alpha(x) := \sup\{P(x, \infty) : D_\alpha(P, P_{_{ref}}) \leq \delta\}.$$

Consider the canonical mapping $Z(\omega) = \omega,\ \omega \in \mathbb{R}$. Then

$$\bar{F}_\alpha(x) = \sup\left\{E_{P_{_{ref}}}[L\mathbf{1}(Z > x)] : E_{P_{_{ref}}}\left[\phi_\alpha(L)\right] \leq \bar{\delta}, E_{P_{_{ref}}}[L] = 1, L \geq 0\right\}.$$

is an optimization problem of the form(12). Therefore, due to Theorem 2, the optimal L^* is easily verified to be of the form $\theta\mathbf{1}(x, \infty) + \tilde{\theta}\mathbf{1}(-\infty, x)$ for some constants $\theta > 1$ and $\tilde{\theta} \in (0, 1)$. Substituting for $L^* = \theta\mathbf{1}(x, \infty) + \tilde{\theta}\mathbf{1}(-\infty, x)$ in the constraints $E_{P_{_{ref}}}[\phi_\alpha(L^*)] = \bar{\delta}$ and $E_{P_{_{ref}}}[L^*] = 1$, we obtain the following conclusion: Given $x > 0$, if there exists a $\theta_x > 1$ such that

$$P_{_{ref}}(x, \infty)\phi_\alpha(\theta_x) + P_{_{ref}}(-\infty, x)\phi_\alpha\left(\frac{1 - \theta_x P_{_{ref}}(x, \infty)}{P_{_{ref}}(-\infty, x)}\right) = \bar{\delta}, \tag{17}$$

then $\bar{F}_\alpha(x) = \theta_x P_{_{GEV}}(x, \infty)$.

4. Asymptotic analysis of robust estimates of tail probabilities

In this section we study the asymptotic behaviour of the tail distribution functions

$$\bar{F}_\alpha(x) := \sup\{P(x, \infty) : D_\alpha(P, P_{_{ref}}) < \delta\},$$

for every $\alpha \geq 1$, as $x \to \infty$. From here onwards, we shall call $\bar{F}_\alpha(\cdot)$ as the *α-family worst-case tail distribution*. All the probability measures involved, unless explicitly specified, are taken to be defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Since $D_\alpha(P_{_{ref}}, P_{_{ref}}) = 0$, it is evident that the worst-case tail estimate $\bar{F}_\alpha(x)$ is at least as large as $P_{_{ref}}(x, \infty)$. While the overall objective has been to provide robust estimates that account for model perturbations, it is certainly not desirable that the worst-case tail distribution $\bar{F}_\alpha(\cdot)$, for example, has unrealistically slow logarithmic decaying tails. Seeing this, our interest in this section is to quantify how heavier the tails of $\bar{F}_\alpha(\cdot)$ are, when compared to that of the reference model.

The bigger the plausible family of measures $\left\{P : D_\alpha(P, P_{_{ref}}) \leq \delta\right\}$, the slower the decay of tail $\bar{F}_\alpha(x)$ is, and vice versa. Hence it is conceivable that the parameter δ is influential in determining the rates of decay of tail distributions $\bar{F}_\alpha(\cdot)$. However, as we shall see below in Theorem 5, it is the parameter α (along with the tail properties of the reference model $P_{_{ref}}$) that solely determines the domain of attraction, and hence the heaviness of tail, of $\bar{F}_\alpha(\cdot)$.

Since our primary interest in the paper is with respect to reference model $P_{_{ref}}$ being a GEV model, we first state the result in this context:

Theorem 4. *Let the reference GEV model P_{GEV} has shape parameter γ_{ref}. Fix any $\alpha \geq 1$, and let $\bar{F}_\alpha(x) := \sup\{P(x,\infty) : D_\alpha(P, P_{ref}) < \delta\}$. If γ^* defined as in*

$$\gamma^* := \frac{\alpha}{\alpha - 1}\gamma_{ref}$$

exists, then the distribution function $F_\alpha(x) = 1 - \bar{F}_\alpha(x)$ belongs to the domain of attraction of G_{γ^*}.

Theorem 4 is, however, a corollary of Theorem 5 below.

Theorem 5. *Let the reference model P_{ref} belong to the domain of attraction of $G_{\gamma_{\text{ref}}}$. In addition, let P_{ref} induce a distribution F that satisfy the regularity assumptions of Proposition 1 with $\gamma = \gamma_{ref}$. Fix any $\alpha \geq 1$, and let $\bar{F}_\alpha(x) := \sup\{P(x,\infty) : D_\alpha(P, P_{ref}) < \delta\}$. If γ^* defined as in*

$$\gamma^* := \frac{\alpha}{\alpha - 1}\gamma_{ref}$$

exists, then the distribution function $F_\alpha(x) = 1 - \bar{F}_\alpha(x)$ belongs to the domain of attraction of G_{γ^*}.

Remark 6. First, observe that $P(x,\infty) \leq \bar{F}_\alpha(x)$, for every P in the neighborhood set of measures $\mathcal{P}_{\alpha,\delta} := \{P : D_\alpha(P, P_{ref}) \leq \delta\}$. Therefore, apart from characterizing the domain of attraction of $\bar{F}_\alpha$, Theorem 5 offers the following insights on the neighborhood $\mathcal{P}_{\alpha,\delta}$:

1) If the reference model belongs to the domain of attraction of a Frechet distribution (that is, $\gamma_{ref} > 0$), and if P is a probability measure that lies in its neighborhood $\mathcal{P}_{\alpha,\delta}$, then P must satisfy that

$$P(x,\infty) = O\left(x^{-\frac{\alpha-1}{\alpha\gamma_{ref}}+\epsilon}\right),$$

as $x \to \infty$, for every $\epsilon > 0$. This conclusion is a direct consequence of (6) and the observation that $P(x,\infty) \leq \bar{F}_\alpha(x)$. In addition, as in the proof of Theorem 5, one can exhibit a measure $P \in \mathcal{P}_{\alpha,\delta}$ such that $P(x,\infty) = \Omega(x^{-(\alpha-1)/\alpha\gamma_{ref}})$.

3) On the other hand, if the reference model belongs to the Gumbel domain of attraction ($\gamma_{ref} = 0$), then every $P \in \mathcal{P}_{\alpha,\delta}$ satisfies $P(x,\infty) = o(x^{-\epsilon})$, as $x \to \infty$, for every $\epsilon > 0$.

3) Now consider the case where $P_{ref} \in \mathcal{D}(G_{\gamma_{\text{ref}}})$ for some $\gamma_{ref} < 0$ (that is, the reference model belongs to the domain of attraction of a Weibull distribution). Let $x_F^* < \infty$ denote the supremum of its bounded support. In that case, any probability measure P that belongs to the neighborhood $\mathcal{P}_{\alpha,\delta}$ must satisfy that $P(-\infty, x_F^*) = 1$ and

$$P(x_F^* - \epsilon, x_F^*) = O\left(\epsilon^{-\frac{\alpha-1}{\alpha\gamma_{ref}}-\epsilon'}\right),$$

as $\epsilon \to 0$, for every $\epsilon' > 0$. In addition, one can exhibit a measure $P \in \mathcal{P}_{\alpha,\delta}$ such that $P(x_F^* - \epsilon, x_F^*) = \Omega(\epsilon^{-(\alpha-1)/\alpha\gamma_{ref}})$.

Verification of the above observations is straightforward once we recall the characterizations of Frechet, Gumbel and Weibull domains of attraction in Section 2. It is important to remember that the above properties hold for all $\alpha > 1$, and is not dependent on δ.

For a fixed reference model P_{ref}, it is evident from Remark 6 that the neighborhoods $\mathcal{P}_{\alpha,\delta} = \{P : D_\alpha(P, P_{ref}) < \delta\}$ include probability distributions with heavier and heavier tails as α approaches 1 from above. This is in line with the observation that $D_\alpha(P, P_{ref})$ is a non-decreasing function in α, and hence larger neighborhoods $\mathcal{P}_{\alpha,\delta}$ for smaller values of α.

In particular, when $\alpha = 1$ and shape parameter $\gamma_{ref} = 0$, the quantity $\gamma^* = \gamma_{ref}\alpha/(\alpha-1)$ as in Theorem 4 is not well-defined. This corresponds to the set of plausible measures $\{P : D_1(P, G_0) \le \delta\}$ defined using KL-divergence around the reference Gumbel model G_0. The following result describes the tail behaviour of $\bar{F}_\alpha$ in this case:

Proposition 7. *Recall the definition of extreme value distributions G_γ in (2). Let $\bar{F}_1(x) = \sup\{P(x,\infty) : D_1(P, G_0) \le \delta\}$, and $F_1(x) = 1 - \bar{F}_1(x)$. Then F_1 belongs to the domain of attraction of G_1.*

The following result, when contrasted with Remark 6, better illustrates the difference between the cases $\alpha > 1$ and $\alpha = 1$.

Proposition 8. *Recall the definition of G_γ as in (5). For every $\delta > 0$, one can find a probability measure P in the neighborhood $\{P : D_1(P, G_{\gamma_{\text{ref}}}) \le \delta\}$ such that*

a) $P(x,\infty) = \Omega(\log^{-3} x)$, *if* $\gamma_{ref} > 0$,
b) $P(x,\infty) = \Omega(x^{-1})$, *if* $\gamma_{ref} = 0$, *and*
c) $P(-\infty, x^*_G) = 1$ *and* $P(x^*_G - \epsilon, x^*_G) = \Omega(\log^{-3}\frac{1}{\epsilon})$, *if* $\gamma_{ref} < 0$. *Here, the right endpoint* $x^*_G = \sup\{x : G_{\gamma_{ref}}(x) < 1\}$ *is finite because* $\gamma_{ref} < 0$.

While the asymptotic bounds in a) and b) are with respect to $x \to \infty$, the corresponding bound in c) holds as $\epsilon \to 0$.

In addition, it is useful to contrast these tail decay results for neighboring measures with that of the corresponding reference measure $G_{\gamma_{\text{ref}}}$ for which

a) $1 - G_{\gamma_{\text{ref}}}(x) = \Omega(x^{-1/\gamma_{ref}})$, if $\gamma_{ref} > 0$,
b) $1 - G_{\gamma_{\text{ref}}} = \Omega(e^{-x})$, if $\gamma_{ref} = 0$, and
c) $G_{\gamma_{\text{ref}}}(x^*) = 1$ and $G_{\gamma_{\text{ref}}}(x^* - \epsilon, x^*) = \Omega(\epsilon^{-1/\gamma_{ref}})$, if $\gamma_{ref} < 0$.

It is evident from the above comparison that the worst-case tail probabilities $\bar{F}_\alpha(x)$ decay at a significantly slower rate than the reference measure when $\alpha = 1$ (the KL-divergence case). Table 1 below summarizes the rates of decay of worst-case tail probabilities $\bar{F}_\alpha(\cdot)$ over different choices of α when the reference model is a GEV distribution. In addition, Figure 1, which compares the worst-case tail distributions $\bar{F}_\alpha(x)$ for three different GEV example models, is illustrative. Proofs of Theorems 4 and 5, Propositions 7 and 8 are presented in Section 6.

5. Robust estimation of VaR

Given independent samples $X_1, \ldots, X_N$ from an unknown distribution F, our objective has been to compute the quantile $F^{\leftarrow}(p)$ for values of p close to 1. In this section, we develop a data-driven algorithm for estimating these extreme quantiles by employing the traditional extreme value theory in tandem with the insights derived in Sections 3 and 4. Our motivation has been to provide conservative estimates for $F^{\leftarrow}(p)$ that are robust against incorrect model assumptions as well as calibration errors. Naturally, the first step in the estimation procedure is to arrive at a reference measure $P_{GEV}(-\infty, x) = G_{\gamma_0}(a_0 x + b_0)$ for the distribution of block-maxima M_n as in Algorithm 1. Once we have a candidate model P_{GEV} for M_n, the p^n-th quantile of the distribution P_{GEV} serves as an estimator for $F^{\leftarrow}(p)$. Instead, if we have a family of candidate models (as in Sections 3 and 4) for M_n, a corresponding robust alternative to this estimator is to compute the worst-case quantile estimate over all the candidate models as below:

$$\hat{x}_p := \sup\left\{G^{\leftarrow}(p^n) : D_\alpha(G, P_{GEV}) \le \delta\right\}. \tag{18}$$

TABLE 1. A summary of domains of attraction of $F_\alpha(x) = 1 - \bar{F}_\alpha(x)$ for GEV models. Throughout the paper, $\gamma^* := \frac{\alpha}{\alpha-1}\gamma_{ref}$

Reference model	Domain of attraction of Worst-case tail $\bar{F}_\alpha(\cdot)$, $\alpha > 1$	Domain of attraction of Worst-case tail $\bar{F}_\alpha(\cdot)$, $\alpha = 1$ (the KL-divergence case)
G_0 (Gumbel light tails)	G_0 (Gumbel light tails)	G_1 (Frechet heavy tails)
$G_{\gamma_{\text{ref}}}, \gamma_{ref} > 0$ (Frechet heavy tails)	G_{γ^*} (Frechet heavy tails)	– (slow logarithmic decay of $\bar{F}_\alpha(x)$ as $x \to \infty$)
$G_{\gamma_{\text{ref}}}$, $\gamma_{ref} < 0$ (Weibull)	G_{γ^*} (Weibull)	– (slow logarithmic decay of $\bar{F}_\alpha(x)$ to 0 at a finite right endpoint x^*)

Here $G^{\leftarrow}$ denotes the usual inverse function $G^{\leftarrow}(u) = \inf\{x : G(x) \geq u\}$ with respect to distribution G. Since the framework of Section 3 is limited to optimization over objective functionals in the form of expectations (as in (12)), it is immediately not clear whether the supremum in (18) can be evaluated using tools developed in Section 3. Therefore, let us proceed with the following alternative: First, compute the worst-case tail distribution

$$\bar{F}_\alpha(x) := \sup\left\{G(x,\infty) : D_\alpha(G, P_{GEV}) \leq \delta\right\}, \quad x \in \mathbb{R}$$

over all candidate models, and compute the corresponding inverse

$$F_\alpha^{\leftarrow}(p^n) := \inf\{x : 1 - \bar{F}_\alpha(x) \geq p^n\}.$$

The estimate $\hat{x}_p$ (defined as in (18)) is indeed equal to $F_\alpha^{\leftarrow}(p^n)$, and this is the content of Lemma 9.

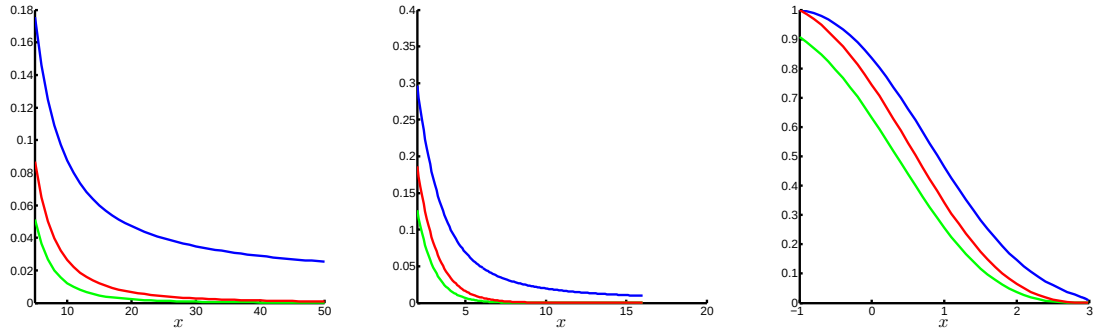
Lemma 9. *For every $u \in (0,1)$, $F_\alpha^{\leftarrow}(u) = \sup\left\{G^{\leftarrow}(u) : D_\alpha(G, P_{GEV}) \leq \delta\right\}$.*

Proof. For brevity, let $\mathcal{P} = \{G : D_\alpha(G, P_{GEV}) \leq \delta\}$. Then, it follows from the definition of $\bar{F}_\alpha$ and $F_\alpha^{\leftarrow}$ that

$$\begin{aligned} F_\alpha^{\leftarrow}(u) &= \inf\left\{x : \sup_{G\in\mathcal{P}} G(x,\infty) \leq 1-u\right\} \\ &= \inf \bigcap_{G\in\mathcal{P}} \left\{x : G(x,\infty) \leq 1-u\right\} \\ &= \inf \bigcap_{G\in\mathcal{P}} \left[G^{\leftarrow}(p),\infty\right) = \sup_{G\in\mathcal{P}} G^{\leftarrow}(p). \end{aligned}$$

Hence proved. □

FIGURE 1. Comparison of $\bar{F}_\alpha(x)$ for different GEV models: The green curves represents the reference model $G_{\gamma_{\text{ref}}}(x)$ for $\gamma_{ref} = 1/3$ (left figure), $\gamma_{ref} = 0$ (middle figure) and $\gamma_{ref} = -1/3$ (right figure). Computations of corresponding $\bar{F}_\alpha(x)$ are done for $\alpha = 1$ (the blue curves), and $\alpha = 5$ (the red curves) with δ fixed at 0.1. The blue curves (corresponding to $\alpha = 1$, the KL-divergence case) conform with our reasoning that $\bar{F}_\alpha(x)$ have vastly different tail behaviours from the reference models when KL-divergence is used.



(a) $G_{\frac{1}{3}}(x)$, a Frechet example (b) $G_0(x)$, a Gumbel example (c) $G_{-\frac{1}{3}}(x)$, a Weibull example

Now that we know $\hat{x}_p = F_\alpha^{\leftarrow}(p^n)$ is the desired estimate, let us recall from Example 1 how to evaluate $\bar{F}_\alpha(x)$ for any x of interest. If $\theta_x > 1$ solves

$$P_{GEV}(x,\infty)\phi_\alpha(\theta_x) + P_{GEV}(-\infty,x)\phi_\alpha\left(\frac{1-\theta_x P_{GEV}(x,\infty)}{P_{GEV}(-\infty,x)}\right) = \bar{\delta},$$

then $\bar{F}_\alpha(x) = \theta_x P_{GEV}(x,\infty)$. Though θ_x cannot be obtained in closed-form, given any $x > 0$, it is rather straight-forward to solve for θ_x numerically and compute $\bar{F}_\alpha(x)$ to a desired level of precision. On the other hand, given a level $u \in (0,1)$, it is also easy to compute $F_\alpha^{\leftarrow}(u)$ by solving for x that satisfies $P_{GEV}(x,\infty) < u$ and

$$P_{GEV}(x,\infty)\phi_\alpha\left(\frac{1-u}{P_{GEV}(x,\infty)}\right) + P_{GEV}(-\infty,x)\phi_\alpha\left(\frac{u}{P_{GEV}(-\infty,x)}\right) = \bar{\delta} \tag{19}$$

Therefore, given α and δ, it is computationally not any more demanding to evaluate the robust estimates $F_\alpha^{\leftarrow}(p^n)$ for $F^{\leftarrow}(p)$. One can perhaps choose α so that the corresponding $\gamma^* = \gamma_0\alpha/(\alpha-1)$ matches with an appropriate confidence interval for the estimate γ_0 : For example, if $\gamma_0 > 0$ and the confidence interval for γ_0 is given by $(\gamma_0 - \epsilon, \gamma_0 + \epsilon)$, then we choose α satisfying

$$\gamma\frac{\alpha}{\alpha-1} = \gamma_0 + \epsilon. \tag{20}$$

Otherwise, α can be chosen based on domain knowledge as well: For example, consider the case where one uses Gaussian distribution to model returns of a portfolio. In this instance, if a financial expert identifies the returns are instead heavy-tailed, then one can take $\alpha = 1$ to account for the imperfect assumption of Gaussian tails. Once the parameter α is identified, it

is straightforward to obtain an estimate for δ using any divergence estimation procedure. For our examples, we use the k-nearest neighbor (k-NN) algorithm of [19]. See also [18, 17, 9] for similar divergence estimators. These divergence estimation procedures provide an empirical estimate of the divergence of the data samples (samples representing the true model) from the calibrated GEV model P_{GEV}. Algorithm 2 summarizes the whole procedure.

Algorithm 2 To compute an estimate for $\mathrm{VaR}_p(X)$ that are robust to model errors
Given: N independent samples $X_1, \ldots, X_N$ of X, and a level p close to 1

Initialize $n < N$, and let $m = \lfloor \frac{N}{n} \rfloor$.

Step 1 (Compute block-maxima): Partition $X_1, \ldots, X_N$ into blocks of size n, and compute the block maxima for each block to obtain samples $M_{n,1}, \ldots, M_{n,m}$ of maxima M_n.

Step 2 (Calibrate a reference GEV model): Treat the samples $M_{n,1}, \ldots, M_{n,m}$ as independent samples coming from a member of the GEV family and use a parameter estimation technique (for example, maximum-likelihood) to estimate the parameters a_0, b_0 and γ_0, along with suitable confidence intervals.

Step 3 (Determine the family of candidate models): Choose an appropriate α (either based on domain knowledge, (or) to match the desired confidence interval of shape parameter γ_0, as in (20)). Determine δ using a divergence estimation procedure. Then the set $\{P : D_\alpha(P, P_{GEV}) \leq \delta\}$ represents the family of candidate models.

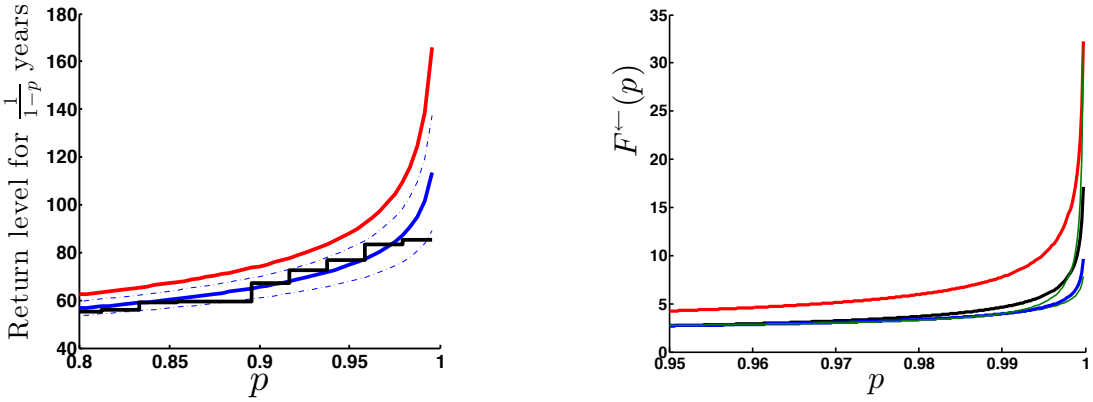
Step 4 (Compute the p^n-th quantile over the reference model and all candidate models): Solve for x such that $G_{\gamma_0}(a_0 x + b_0) = p^n$, and let x_p be the corresponding solution. Solve for $x > x_p$ in (19) and let the solution be $\hat{x}_p$.

RETURN x_p and $\hat{x}_p$

Example 2. For a demonstration of the ideas introduced, we consider the rainfall accumulation data, due to the study of [5], from a location in south-west England (see also [4] for further extreme value analysis with the dataset). Given annual maxima of daily rainfall accumulations over a period of 48 years (1914-1962), we attempt to compute, for example, the 100-year return level for the daily rainfall data. In other words, we aim to estimate the daily rainfall accumulation level that is exceeded about only once in 100 years. As a first step, we calibrate a GEV model for the annual maxima. Maximum-likelihood estimation of parameters results in the following values for shape, scale and location parameters: $\gamma_0 = 0.1072$, $a_0 = 9.7284$ and $b_0 = 40.7830$. The 100-year return level due to this model yields a point estimate 98.63mm with a standard error of $\pm$17.67mm (for 95% confidence interval). It is instructive to compare this with the corresponding estimate 106.3 ± 40.7mm obtained by fitting a generalized Pareto distribution (GPD) to the large exceedances (see Example 4.4.1 of [4]). To illustrate our methodology, we pick $\alpha = 2$, as suggested in (20). Next, we obtain $\delta = 0.05$ as an empirical estimate of divergence D_α between the data points representing annual maxima and the calibrated GEV model $P_{GEV} = G_{\gamma_0}(a_0 x + b_0)$. This step is accomplished using a simple k-nearest neighbor estimator (see [19]). Consequently, the worst-case quantile estimate over all probability measures satisfying $D_\alpha(P, P_{GEV}) \leq \delta$

is computed to be $F_\alpha^{\leftarrow}(1-1/100) = 132.24$mm. While not being overly conservative, this worst-case 100 year return level of 132.44mm also acts as an upper bound to estimates obtained due to different modelling assumptions (GEV vs GPD assumptions). To demonstrate the quality of estimates throughout the tail, we plot the return levels for every $1/(1-p)$ years, for values of p close to 1, in Figure 5(a). While the return levels predicted by the GEV reference model is plotted in blue (with the dashed blue lines representing 95% confidence intervals), the red curve represents the worst-case estimates $F_\alpha^{\leftarrow}(p)$. The empirical quantiles are drawn in black.

FIGURE 2. Plots for Examples 2 and 3



(a) Quantile plots for rainfall data, Eg. 2

(b) Quantile plots for Pareto data, Eg. 3

Example 3. In this example, we are provided with 100 independent samples of a Pareto random variable satisfying $P\{X > x\} = 1 - F(x) = 1 \wedge x^{-3}$. As before, the objective is to compute quantiles $F^{\leftarrow}(p)$ for values of p close to 1. As the entire probability distribution is known beforehand, this offers an opportunity to compare the quantile estimates returned by our algorithm with the actual quantiles. Unlike Example 2, the data in this example does not present a natural means to choose block sizes. As a first choice, we choose block size $n = 5$ and perform routine computations as in Algorithm 2 to obtain a reference GEV model P_{GEV} with parameters $\gamma_0 = 0.11, a_0 = 0.58, b_0 = 1.88$, and corresponding tolerance parameters $\alpha = 1.5$ and $\delta = 0.8$. Then the worst-case quantile estimate $F_\alpha^{\leftarrow}(p^n) = \sup\{G^{\leftarrow}(p^n) : D_\alpha(G, P_{GEV}) \leq \delta\}$ is immediately calculated for various values of p close to 1, and the result is plotted (in red) against the true quantiles $F^{\leftarrow}(p) = (1-p)^{-1/3}$ (in black) in Figure 5(b). These can, in turn, be compared with the quantile estimates x_p (in blue) due to traditional GEV extrapolation with reference model P_{GEV}. Recall that the initial choice for block size, $n = 5$, was arbitrary. One can perhaps choose a different block size, which will result in a different model for corresponding block-maximum M_n. For example, if we choose $n = 10$, the respective GEV model for M_{10} has parameters $\gamma_0 = 0.22, a_0 = 0.55$ and $b_0 = 2.3$. Whereas, if we choose $n = 15$, the GEV model for M_{15} has parameters $\gamma_0 = 0.72, a_0 = 0.32$ and $b_0 = 2.66$. These models are different, and subsequently, the corresponding quantile estimates (plotted using green lines in Figure 5(b)) are also different. However, as it can be inferred from Figure 5(b), the robust quantile estimates (in red) obtained by running Algorithm 2 forms a good upper

bound to the actual quantiles $F^{\leftarrow}(p)$, as well as to the quantile estimates due to different GEV extrapolations from different block sizes $n = 10$ and 15.

Example 4. The objective of this example is to demonstrate the applicability of Algorithm 2 in an instance where the traditional extrapolation techniques tend to not yield stable estimates. For this purpose, we use N = 2000 independent samples of the random variable $Y = X + 50\mathbf{1}(X > 5)$ as input to Algorithm 1, with the aim of calculating the extreme quantile $F^{\leftarrow}(0.999)$. Here, F denotes the distribution function of random variable Y, and X is a Pareto random variable with distribution $\max(1 - x^{-1.1}, 0)$. The quantile estimates (and the corresponding 95% confidence intervals) output by this traditional GEV estimation procedure, for various choices of block sizes, is displayed in blue in Figure 3(a). For a majority of block size choices, it can be observed that the 95% confidence regions obtained from the calibrated GEV models are far below the true quantile drawn in black. This underestimation is perhaps because of the sudden shift of samples of block-maxima M_n from a value less than 5 to a value larger than 55 (recall that the distribution F assigns zero probability to the interval $(5, 55)$). As the true value at risk is significantly underestimated even for modest choices of block sizes, we use Algorithm 2 to yield an upper bound that is robust against model errors. The curve in red in Figure 3(a) corresponds to the upper bound on $F^{\leftarrow}(0.999)$ output by Algorithm 2. We note the following observations: First, the estimates output by Algorithm 2 indeed act as an upper bound for the true quantile (drawn in black), irrespective of the block-size chosen and the baseline GEV model used. Second, for block-sizes smaller than $n = 45$, it appears that the calibrated GEV models are not representative enough of the distribution of M_n, and hence more divergence from the calibrated GEV distribution. Understandably, this results in a conservative upper bound when $n < 45$. To illustrate that problems concerning instabilities in parameters estimation cannot be alleviated by simply choosing an alternate extrapolation technique, we consider another popular procedure where a generalized Pareto distribution (abbreviated as GPD) is fit to observations above a certain threshold (see [16] for an explanation). The point estimates and 95% confidence intervals obtained from this estimation procedure, for different choices of threshold levels, are displayed in Figure 3(b) in blue. As in the GEV extrapolation case, the additional robustification procedure yields an upper bound (drawn in red) that is conservative. One of the future research objectives is to include meaningful constraints in the formulation that makes the bound less conservative.

6. Proofs of main results

In this section, we provide proofs of Theorems 4 and 5, Propositions 7 and 8.

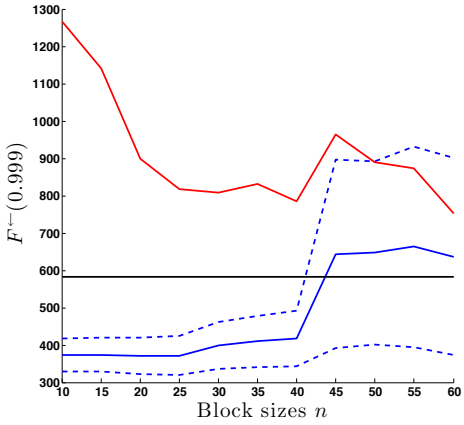
Proof of Theorem 5. Our goal is to determine the domain of attraction of $\bar{F}_\alpha(x) = \sup\{P(x, \infty) : D_\alpha(P, P_{ref}) \le \delta\}$. As in Example 1, we have that $\bar{F}_\alpha(x) = \theta_x P_{ref}(x, \infty)$, where θ_x solves (17). Since θ_x satisfies $P_{ref}(x, \infty)\phi_\alpha(\theta_x) \le \bar{\delta}$, it follows that

$$\bar{F}_\alpha(x) \le \phi_\alpha^{-1}\left(\frac{\bar{\delta}}{P_{ref}(x, \infty)}\right) P_{ref}(x, \infty). \tag{21}$$

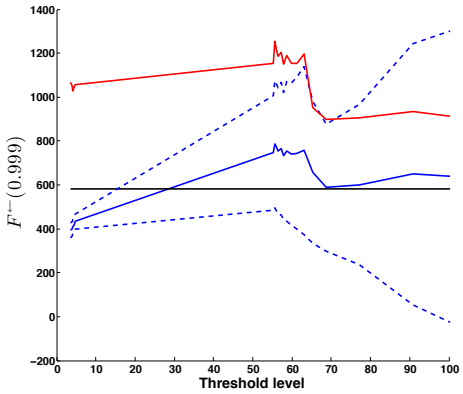
Here $\phi_\alpha^{-1}(\cdot)$ denotes the inverse function of $\phi_\alpha(x)$. Similarly, to obtain a lower bound for $\bar{F}_\alpha(x)$, first consider a probability measure Q defined by

$$\frac{dQ}{dP_{ref}}(x) = \phi_\alpha^{-1}\left(\frac{c}{P_{ref}(x, \infty)(1 - \log P_{ref}(x, \infty))^2}\right),$$

FIGURE 3. Plots for Example 4



(a) Instability in quantile $F^{\leftarrow}(0.999)$ estimated via GEV extrapolation, Eg. 4



(b) Instability in quantile $F^{\leftarrow}(0.999)$ estimated via GPD extrapolation, Eg. 4

for a suitable positive constant c. Then $D_\alpha(Q, P_{ref}) < \infty$ because of a simple change of variables $u = P_{ref}(x, \infty)$ in the integration

$$\int \phi_\alpha \left(\frac{dQ}{dP_{ref}} \right) dP_{ref} = \int_0^1 \frac{c}{u(1 - \log u)^2} du < \infty.$$

Consequently, due to a continuity argument, one can demonstrate a constant $a \in (0, 1)$ such that $D_\alpha(aQ + (1-a)P_{ref}, P_{ref}) \leq \delta$. Then, it follows from the definition of $\bar{F}_\alpha(x)$ that

$$\bar{F}_\alpha(x) \geq \big(aQ + (1-a)P_{ref}\big)(x, \infty) = \int_x^\infty \left(a \frac{dQ}{dP_{ref}}(t) + 1 - a \right) P_{ref}(dt)$$

Since $dQ/dP_{ref}(t)$ is eventually increasing, as $t \to \infty$, we have that,

$$\bar{F}_\alpha(x) \geq a\phi_\alpha^{-1} \left(\frac{c}{P_{ref}(x, \infty)(1 - \log P_{ref}(x, \infty))^2} \right) P_{ref}(x, \infty),$$

for sufficiently large values of x. For brevity, let

$$A(x) := P_{ref}(x, \infty), \; g(x) := a\phi_\alpha^{-1}(c(1 - \log x)^{-2}/x) \text{ and } h(x) := \phi_\alpha^{-1}(\bar{\delta}/x).$$

Then, combining the above lower bound with the upper bound in (21), we obtain

$$\bar{F}_{low}(x) := g(A(x))A(x) \leq \bar{F}_\alpha(x) \leq h(A(x))A(x) =: \bar{F}_{up}(x), \tag{22}$$

for large values of x. Recall that the reference measure P_{ref} belongs to the domain of attraction of $G_{\gamma_{\text{ref}}}$. The following lemma characterizes the extreme value distributions corresponding to the upper and lower bounds $\bar{F}_{up}$ and $\bar{F}_{low}$.

Lemma 10. *Suppose that the quantity $\gamma^* = \frac{\alpha}{\alpha - 1}\gamma_{ref}$ is well-defined. Additionally, let $x^* = \sup\{x : A(x) > 0\}$. Then the following are true:*

$$(a) \lim_{x \uparrow x^*} - \left(\frac{\bar{F}_{up}}{\bar{F}'_{up}} \right)'(x) = \gamma^*, \text{ and } (b) \lim_{x \uparrow x^*} - \left(\frac{\bar{F}_{low}}{\bar{F}'_{low}} \right)'(x) = -\gamma^*.$$

As a consequence of Proposition 1 and Lemma 10, if γ^* is finite, both $\bar{F}_{low}$ and $\bar{F}_{up}$ lie in the domain of attraction of G_{γ^*}. As $\bar{F}_\alpha(x)$ is sandwiched between $\bar{F}_{low}(x)$ and $\bar{F}_{up}(x)$ as in (22), if at all $\bar{F}_\alpha$ belongs to the domain of attraction of G_γ for some $\gamma \in \mathbb{R}$, then γ must equal γ^*. Since $\bar{F}_\alpha(x) \sim \bar{F}_\alpha(x^-)$ as $x \uparrow x^*$, due to Theorem 1.7.13 of [14], this is indeed the case. Therefore, the α-family worst-case tail distribution $\bar{F}_\alpha$ belongs to the domain of attraction of G_{γ^*}. □

Proof of Lemma 10(a). Recall that $\bar{F}_{up}(x) = h(A(x))A(x)$. By repeatedly applying elementary rules of differentiation, it is obtained that

$$-\left(\frac{\bar{F}_{up}}{\bar{F}'_{up}}\right)'(x) = -\left(\frac{A}{A'}\right)'(x)\left(1+\frac{A(x)h'(A(x))}{h(A(x))}\right)^{-1} + \left(\frac{A(x)h'(A(x))}{h(A(x))} + A^2(x)\left(\frac{h'}{h}\right)'(A(x))\right)\left(1+\frac{A(x)h'(A(x))}{h(A(x))}\right)^{-2} \tag{23}$$

Case $\alpha > 1$: It is easily verified that $h(x) = (\bar{\delta}/x)^{1/\alpha}$ and $h'(x)/h(x) = -(\alpha x)^{-1}$. As a result, we obtain

$$-\left(\frac{\bar{F}_{up}}{\bar{F}'_{up}}\right)'(x) = -\left(\frac{A}{A'}\right)'(x)\left(1-\frac{1}{\alpha}\right)^{-1} + \left(-\frac{1}{\alpha}+\frac{1}{\alpha}\right)\left(1-\frac{1}{\alpha}\right)^{-2}.$$

In addition, as required in the statement of Theorem 5, $A(x) := P_{ref}(x,\infty)$ satisfies $-(A/A')'(x) \to \gamma_{ref}$, as x approaches its right endpoint $x^* = \sup\{x : A(x) > 0\}$. Therefore,

$$\lim_{x\uparrow x^*} -\left(\frac{\bar{F}_{up}}{\bar{F}'_{up}}\right)'(x) = \frac{\alpha}{\alpha-1}\lim_{x\uparrow x^*}\left[-\left(\frac{A}{A'}\right)'(x)\right] = \frac{\alpha}{\alpha-1}\gamma_{ref}.$$

Case $\alpha = 1$: When α equals 1, $\phi_\alpha^{-1}(x) = x/W(x)$, where $W(x)$ is the product log function[4]. Then the following calculations are simply algebraic:

$$\frac{xh'(x)}{h(x)} = -\left(1+\frac{1}{W\left(\frac{\bar{\delta}}{x}\right)}\right)^{-1} \text{ and } x^2\left(\frac{h'}{h}\right)'(x) = \left[1+\left(1+W\left(\frac{\bar{\delta}}{x}\right)\right)^{-1}\right]\left(1+\frac{1}{W\left(\frac{\bar{\delta}}{x}\right)}\right)^{-2}.$$

Substituting these in (23), we obtain

$$-\left(\frac{\bar{F}_{up}}{\bar{F}'_{up}}\right)'(x) = \left[-\left(\frac{A}{A'}\right)'(x)W\left(\frac{\bar{\delta}}{A(x)}\right)-1\right]\left(1+\frac{1}{W\left(\frac{\bar{\delta}}{A(x)}\right)}\right)^{-1}. \tag{24}$$

Recall that $-(A/A')'(x)$ converges to γ_{ref}, as $x \uparrow x^*$. Letting $x \to x^*$ in the above expression, we obtain

$$-\left(\frac{\bar{F}_{up}}{\bar{F}'_{up}}\right)'(x) = \begin{cases} \infty, & \text{if } \gamma_{ref} > 0, \\ -\infty, & \text{if } \gamma_{ref} < 0, \end{cases}$$

which indeed equals $\frac{\alpha}{\alpha-1}\gamma_{ref}$. This completes the proof of Part (a) of Lemma 10. □

[4] W is the inverse function of $f(x) = xe^x$

Proof of Lemma 10(b). First, an expression for $(\bar{F}_{low}/\bar{F}'_{low})'$ similar to (23) can be obtained by simply substituting g in place of h in (23). Again, the cases $\alpha > 1$ and $\alpha = 1$ are calculated separately:
Case $\alpha > 1$: When $\alpha > 1, \phi_\alpha^{-1}(x) = x^{1/\alpha}$. By applying elementary rules of differentiation, we obtain

$$\frac{xg'(x)}{g(x)} = \frac{1}{\alpha}\frac{1+\log x}{1-\log x} \quad \text{and} \quad x^2\left(\frac{g'}{g}\right)'(x) = \frac{1}{\alpha}\frac{1+\log^2 x}{(1-\log x)^2}.$$

Letting $x \uparrow x^*$, we obtain $A(x)g'(A(x))/g(A(x)) \to -1/\alpha$ and $A(x)^2(g'/g)'(A(x)) \to 1/\alpha$. Subsequently,

$$\lim_{x\uparrow x^*}\left(\frac{\bar{F}_{low}}{\bar{F}'_{low}}\right)'(x) = \left(1-\frac{1}{\alpha}\right)^{-1}\lim_{x\uparrow x^*}\left(\frac{A}{A'}\right)'(x) + \left(-\frac{1}{\alpha}+\frac{1}{\alpha}\right)\left(1-\frac{1}{\alpha}\right)^2,$$

which equals $\frac{\alpha}{\alpha-1}\gamma_{ref}$, as in the proof of Part (a) of Lemma 10. The case $\alpha = 1$ is similar to that of proof of Part (a), but more tedious, and is not presented here in the interest of space and readability. □

Proof of Theorem 4. Theorem 4 follows as a simple corollary of Theorem 5, once we verify that any GEV model $G(x) := P_{GEV}(-\infty, x)$ satisfies $G'(x) > 0$ in a left neighborhood of $x^*_G = \sup\{x : G(x) < 1\}$, and

$$\lim_{x\uparrow x^*_G}\left(\frac{1-G}{G'}\right)'(x) = \gamma_{ref},$$

where γ_{ref} is the shape parameter of G. Such a GEV model $G(x) = G_{\gamma_{\rm ref}}(ax+b)$ for some scaling and translation constants a and b. Therefore, it is enough to verify these properties only for $G(x) = G_{\gamma_{\rm ref}}(x)$. Once we recall the definition of G_γ in (5), the desired properties are elementary exercises in calculus. □

Proof of Proposition 7. First, we derive a lower bound for $\bar{F}_1(x) = \sup\{P(x,\infty) : D_1(P, G_0) \le \delta\}$. Consider the probability density function $f(x) = c(x\log x)^{-2}\mathbf{1}(x \ge 2)$, where c is a normalizing constant that makes $\int f(x)dx = 1$. In addition, let $g(x) = G_0'(x)$ denote the probability density function corresponding to the distribution G_0. Clearly,

$$\begin{aligned} D_1(f,g) &= \int f(x)\log\left(\frac{f(x)}{g(x)}\right)dx \\ &= c\int_2^\infty (x\log x)^{-2}\log\left(\frac{c(x\log x)^{-2}}{\exp(-\exp(-x)\exp(-x))}\right)dx \\ &\le \int_2^\infty \frac{x+\exp(-x)+\log c}{x^2\log^2 x}dx < \infty. \end{aligned}$$

Now, as in the proof of Theorem 5, consider a family of densities $\{af+(1-a)G_0' : a \in (0,1)\}$. Due to the continuity of $D_1(af+(1-a)G_0', G_0')$ with respect to a, there exists an $\bar{a} \in (0,1)$ such that $D_1(\bar{a}f+(1-\bar{a})G_0', G_0') \le \delta$. Then, according to the definition of $\bar{F}_1$,

$$\begin{aligned} \bar{F}_1(x) &\ge \int_x^\infty \left(\bar{a}f+(1-\bar{a})G_0'\right)(u)du \\ &\ge \bar{a}\int_x^\infty \frac{c}{u^2\log^2 u}du = \frac{\bar{a}c+o(1)}{x\log^2 x}, \end{aligned}$$

as $x \to \infty$. The asymptotic equivalence used above is due to Karamata's theorem (see Theorem 1 in Chapter VIII.9 of [7]). Combining this lower bound with the upper bound in (21), we obtain, for large enough x,

$$\frac{\bar{a}c}{2x\log^2 x} \leq \bar{F}_1(x) \leq h\big(1-G_0(x)\big)\big(1-G_0(x)\big),$$

where $h(x) = \phi_\alpha^{-1}(\bar{\delta}/x)$. For convenience, let us write $\bar{F}_{up}(x) := h\big(1-G_0(x)\big)\big(1-G_0(x)\big)$ and $\bar{F}_{low}(x) := \bar{a}c/(2x\log^2 x)$. Due to the characterization in (6), we have that $\bar{F}_{low} \in \mathcal{D}(G_1)$. On the other hand, following the lines of Proof of Lemma 10(a), from (24), we obtain that

$$-\left(\frac{\bar{F}_{up}}{\bar{F}'_{up}}\right)'(x) = \left[\left(\frac{1-G_0}{G'_0}\right)'(x) W\left(\frac{\bar{\delta}}{1-G_0(x)}\right)+1\right]\left(1+\frac{1}{W\left(\frac{\bar{\delta}}{1-G_0(x)}\right)}\right)^{-1}.$$

Since $G_0(x) = \exp(-e^{-x})$, we obtain

$$\left(\frac{1-G_0}{G'_0}\right)'(x) = e^{e^{-x}}\left(e^x\left(1-e^{-e^{-x}}\right)-1\right) = \frac{e^{-x}}{2}(1+o(1)),$$

as $x \to \infty$. Therefore,

$$-\left(\frac{\bar{F}_{up}}{\bar{F}'_{up}}\right)'(x) \sim \frac{e^{-x}}{2}(1+o(1)) W\left(\frac{\bar{\delta}}{e^{-x}(1+o(1))}\right)+1$$

as $x \to \infty$. Since $tW(1/t) \to 0$ as $t \to 0$, it follows that $-(\bar{F}_{up}/\bar{F}'_{up})(x)$ converges to 1 as $x \to \infty$. Then, due to Proposition 1, we have that $\bar{F}_{up}$ also belong to the domain of attraction of G_1. Since both $\bar{F}_{low}$ and $\bar{F}_{up}$ lie in the domain of attraction of G_1, following the same line of reasoning as in the proof of Theorem 5, we obtain that $\bar{F}_1(x) \in \mathcal{D}(G_1)$. This completes the proof. □

Proof of Proposition 8. First, let us consider the case $\gamma_{ref} \neq 0$: Recall the probability measure $aQ + (1-a)P_{ref}$ exhibited for establishing the lower bound in the proof of Theorem 5. For proving Proposition 8, we take the reference measure P_{ref} as $G_{\gamma_{\text{ref}}}$. Further, if we let $g(t) = a\phi_1^{-1}(c(1-\log t)^{-2}/t)$ and $A(x) := 1 - G_{\gamma_{\text{ref}}}(x)$, then as in the proof of Theorem 5, the measure $P := aQ + (1-a)P_{ref}$

1) satisfies $D_1(P, G_{\gamma_{\text{ref}}}) \leq \delta$, and
2) admits a lower bound $P(x,\infty) \geq g(A(x))A(x)$.

To proceed further, observe that $A(x) = 1 - G_{\gamma_{\text{ref}}}(x) \geq \bar{c}(1+\gamma_{ref}x)^{-1/\gamma_{ref}}$ for some constant $\bar{c} < 1$ and all x close enough to the right endpoint $x^*_G := \sup\{x : G_{\gamma_{\text{ref}}}(x) < 1\}$. In addition, $tg(t)$ strictly decreases to 0 as t decreases to 0. Therefore, for all x close to the right endpoint $x^*_G := \sup\{x : G_{\gamma_{\text{ref}}}(x) < 1\}$, it follows that

$$P(x,\infty) \geq g\left(\bar{c}(1+\gamma_{ref}x)^{-1/\gamma_{ref}}\right)\bar{c}(1+\gamma_{ref}x)^{-1/\gamma_{ref}}.$$

Since $\phi_1^{-1}(u) \geq u/\log u$ for large enough u, $g(t) \geq act^{-1}(1-\log t)^{-2}\log^{-1}(c/t)$ for all t close to 0. Therefore,

$$\begin{aligned} P(x,\infty) &\geq ac\Big(1-\log\left(\bar{c}\left(1+\gamma_{ref}x\right)^{-1/\gamma_{ref}}\right)\Big)^{-2}\log^{-1}\left(c(1+\gamma_{ref}x)^{1/\gamma_{ref}}/\bar{c}\right) \\ &= \Omega\left(\frac{1}{\gamma_{ref}}\log^{-3}(1+\gamma_{ref}x)\right), \text{ as } x \to x^*_G. \end{aligned}$$

This verifies the statement in cases (a) and (b) where $\gamma_{ref} \neq 0$. When $\gamma_{ref} = 0$, see the proof of Proposition 7 where we exhibit a measure P such that $D_1(P, G_0) \leq \delta$ and $P(x, \infty) = \Omega\left(x^{-1} \log^{-2} x\right)$. This completes the proof.

References

[1] R. Atar, K. Chowdhary, and P. Dupuis. Robust bounds on risk-sensitive functionals via Rényi divergence. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):18–33, 2015.

[2] A. Ben-Tal, D. den Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[3] T. Breuer and I. Csiszár. Measuring distribution model risk. *Mathematical Finance*, 2013.

[4] S. Coles. *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer-Verlag London, Ltd., London, 2001.

[5] S. G. Coles and J. A. Tawn. A bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 45(4):pp. 463–478, 1996.

[6] L. de Haan and A. Ferreira. *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006. An introduction.

[7] W. Feller. *An introduction to probability theory and its applications. Vol. II*. John Wiley & Sons, Inc., New York-London-Sydney, 1966.

[8] P. Glasserman and X. Xu. Robust risk measurement and model risk. *Quantitative Finance*, 14(1):29–58, 2014.

[9] M. Gupta and S. Srivastava. Parametric bayesian estimation of differential entropy and relative entropy. *Entropy*, 12(4):818, 2010.

[10] L. P. Hansen and T. J. Sargent. Robust control and model uncertainty. *The American Economic Review*, 91(2):pp. 60–66, 2001.

[11] Z. Hu and L. J. Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available on optimization online*, 2012.

[12] R. Jiang and Y. Guan. Data-driven chance constrained stochastic program. *Available at Optimization Online*, 2013.

[13] P. Jorion. *Value at Risk, 3rd Ed.: The New Benchmark for Managing Financial Risk*. McGraw-Hill Education, 2006.

[14] M. R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and related properties of random sequences and processes*. Springer Series in Statistics. Springer-Verlag, New York-Berlin, 1983.

[15] F. Liese and I. Vajda. *Convex statistical distances*. Teubner Texts in Mathematics. BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1987.

[16] A. J. McNeil and T. Saladin. The peaks over thresholds method for estimating high quantiles of loss distributions. In *Proceedings of 28th International ASTIN Colloquium*, pages 23–43, 1997.

[17] X. Nguyen, M. Wainwright, and M. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *Information Theory, IEEE Transactions on*, 56(11):5847–5861, Nov 2010.

[18] X. Nguyen, M. J. Wainwright, and M. I. Jordan. On surrogate loss functions and f-divergences. *Ann. Statist.*, 37(2):876–904, 04 2009.

[19] B. Póczos and J. Schneider. On the estimation of alpha-divergences. *In AISTATS*, 2011.

[20] A. Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 547–561. Univ. California Press, Berkeley, Calif., 1961.

[21] S. I. Resnick. *Extreme values, regular variation and point processes*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2008. Reprint of the 1987 original.

[22] Y.-L. Tsai, D. J. Murdoch, and D. J. Dupuis. Influence measures and robust estimators of dependence in multivariate extremes. *Extremes*, 14(4):343–363, 2010.

[23] Z. Wang, P. Glynn, and Y. Ye. Likelihood robust optimization for data-driven problems. *arXiv preprint arXiv:1307.6279*, 2013.

Columbia University, Department of Industrial Engineering & Operations Research, 340 S. W. Mudd Building, 500 W. 120 Street, New York, NY 10027, United States.

E-mail address: {jose.blanchet, karthyek.murthy}@columbia.edu