

# Tail Similarity

Vali Asimit

*Bayes Business School, City, University of London, UK*

Zhongyi Yuan\*

*Smeal College of Business, The Pennsylvania State University, USA*

Feng Zhou

*University of Cambridge, UK and Bayes Business School, City, University of London, UK*

---

## Abstract

Simple tail similarity measures are investigated in this paper so that the overarching tail similarity between two distributions is captured. We develop some theoretical results to support our novel measures, where the focus is on asymptotic approximations of our similarity measures for Fréchet-type tails. A simulation study is provided to validate the effectiveness of our proposed measures and demonstrate their great potential in capturing the intricate tail similarity. We conclude that our measure and the standard comparisons between the (first-order) extreme index estimates provide complementary information, and one should analyse them in tandem rather than in isolation. We also provide some very simple guiding principles of good practice when using the two sources of information; these are recommended to be complemented further by domain knowledge to validate and clarify the conclusions of our guiding principles, especially in situations when there is no clear-cut conclusion as it is often the case in real-life applications.

*Keywords:* Tail similarity; Divergence measures; Extreme value theory; Probability distance; Regular variation.

---

## 1. Introduction

We study in this paper the (dis)similarity between the tails of two distributions, which lies in the intersection of two research areas that have not seen much overlap in the literature: characterization of the (dis)similarity between distributions and asymptotic analysis of tails. Both study important questions that have attracted growing research attention and seen broad applications in risk management, insurance, economics, machine learning, etc.

First of all, understanding the (dis)similarities between random quantities and their distributions has been a crucial question encountered across diverse research domains. The ability to distinguish between quantities or distributions based on these (dis)similarities is essential for many scientific discoveries and decision-making processes. For instance, within risk management, practitioners frequently deal with the task of comparing different risk scenarios and

---

\*Corresponding author.

*Email addresses:* [asimit@city.ac.uk](mailto:asimit@city.ac.uk) (Vali Asimit), [zuy11@psu.edu](mailto:zuy11@psu.edu) (Zhongyi Yuan), [feng.zhou@city.ac.uk](mailto:feng.zhou@city.ac.uk) and [feng.zhou@mrc-bsu.cam.ac.uk](mailto:feng.zhou@mrc-bsu.cam.ac.uk) (Feng Zhou)

assessing how the addition of certain positions reshapes the loss distribution. Robust risk management requires decisions robust enough to withstand changes in the loss distribution that are within close proximity to a benchmark distribution (Birghila and Pflug, 2019; Blanchet et al., 2019; Lux and Papapantoleon, 2019; Tang and Yang, 2023). Similarly, within causal analysis that is ubiquitous in the economics literature, one may opt to quantify the (dis)similarity between the conditional distribution of a response variable given its historical context and that given the history of both the response variable and a potential causal variable. Detecting substantial dissimilarity provides a crucial basis for concluding upgraded predictability and, therefore, a causal relationship (Chen et al., 2014; Granger, 1969; Hong et al., 2009; Mazzarisi et al., 2020; Sims, 1972). Within the domain of machine learning, diverse applications hinge on similarity calculations. Anomaly detection algorithms may identify anomalies in a set of data based on the (dis)similarity between its distribution and how the distribution usually appears (Nassif et al., 2021). In fact, monitoring machine learning models' performance after their deployment into production also requires detecting structural changes of input data distribution, which could impair the model performance. Moreover, the training of various types of neural networks, such as Generative Adversarial Networks and many of their variants, often amounts to training the generator to minimize a loss that is represented by the dissimilarity between the data distribution and the model distribution (Arjovsky et al., 2017; Bellemare et al., 2017; Goodfellow et al., 2014). All of these applications require measuring or approximating the (dis)similarities between distributions as a crucial step. Since the quantifications of both similarity and dissimilarity are fundamentally equivalent, in this paper, we may refer to either term depending on the context.

Second, tail analysis is a crucial component of tail risk management and has been a focus of risk management and insurance research for decades (Asimit et al., 2011; Embrechts et al., 1997; Hua and Joe, 2011; Ji et al., 2021; Kelly and Jiang, 2014; Mao et al., 2012, 2023; Qin and Zhou, 2021; Sun et al., 2022). Over the years, tail risk management is becoming increasingly important in light of the extreme events that are shocking regional or global economy with increasing frequency—some recent ones include the COVID-19 pandemic, the financial market turmoils ensuing the pandemic, and the natural catastrophes that flooded or scorched numerous regions across continents and caused record-level economic losses. Evidently, the catastrophic consequences of such events render understanding the tails of the potential loss incredibly important.

While the tail risk research area is perhaps too broad for a thorough yet concise literature review—it has spanned various fields including actuarial science, economics, insurance, and finance—we highlight a few examples that employ the same asymptotic approach as ours to examine tail risk-related quantities. Asimit et al. (2011) approximate a multi-line business's risk capital allocations to each line when a tail risk measure is utilized. In doing so, the authors investigate how individual lines' tail behaviors and their tail dependence structures influence the allocations. Hua and Joe (2011) conduct asymptotic analysis on the *Conditional Tail Expectation (CTE)* risk measure under a condition of second-order regular variation, which is also our standing assumption in this paper. They obtain close-form second-order approximations for both univariate risks and multivariate risks, where for the latter the authors focus on a

variant of *Marginal Expected Shortfall*. [Mao et al. \(2012\)](#) study the diversification effect via the ratio of the CTE of an aggregate risk to the sum of the individual risks' CTEs and obtain an asymptotic approximation of the ratio under a second-order regular variation assumption. [Sun et al. \(2022\)](#) propose a *Tail-based Cumulative Residual Entropy* as a measure of tail risk variability, study its asymptotic approximation under a tail scenario, also assuming second-order regular variation for the risk variable, and investigate how the tail dependence structure and individual risk severity influence the approximations. [Qin and Zhou \(2021\)](#) propose the concept of *Asymptotic Marginal Expected Shortfall (AMES)*, which is the limit of the share of Marginal Expected Shortfall in the total system-wide risk, find its value under an assumption of multivariate regular variation for the risk variables, and employ the estimated AMES to allocate systematic risk to individual institutions.

Even though all of the aforementioned areas have been extensively studied, no prior work has probed probabilistic similarities from the specific perspective of tail analysis. The previous studies on probability distances and divergences have predominately concentrated on evaluating the distributions as a whole and have treated tails no differently from the bodies of the distributions. This results in a methodological difference between our study and the existing ones on similarity measures: ours employs asymptotic analysis for large values of the variables while the methods in the literature depend on the specific research goal and range from optimization to large sample statistical analysis. Methodologically, the most closely related works in the literature are those reviewed above on asymptotic analysis of tails. Nonetheless, despite the methodological similarities, the research objectives are very different between their works and ours.

The main contributions of our paper are as follows: First, we explore for the first time divergence-based similarities within the tails of risks through asymptotic analysis. In doing so, we derive novel asymptotic approximations for wide classes of probability distance measures with a focus on the  $\phi$ -divergence and Wasserstein distance, chosen in view of their broad applicability and versatility ([Arjovsky et al., 2017](#); [Ben-Tal et al., 2013](#); [Jager and Wellner, 2007](#); [Tang and Yang, 2023](#)). The approximations are fine enough to capture the subtle dissimilarity between close tails that have the same first-order tail index. Second, we propose a tail similarity measure and identify practical applications for the measure in tail risk budgeting through real-data analysis. Finally, on a more fundamental level, we pave the way into further research within the intersection of two primary strands of literature: on probability distance and on tail risk analysis.

The remainder of our paper is structured as follows: in [Section 2](#), we comprehensively formulate the problem and provide foundational insights into probability distance and divergence measures. [Section 3](#) then presents our main results, while [Section 4](#) illustrates our results through numerical tests and real-world applications. Lastly, in [Section 5](#), we provide concluding remarks. Preliminaries on risk budgeting and risk parity, description of the data generation processes used for [Section 4](#), and all detailed proofs are relegated to the Appendices.

## 2. Problem Formulation and Preliminaries

### 2.1. Problem formulation

In this section we specify the main problem of interest. Consider an atomless probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , where all random variables considered in this paper reside. Suppose that  $X$  and  $Y$  are random variables with distribution functions  $F$  and  $G$ , respectively. We shall build our theory and tail similarity measure on the probability distances and divergence measures between the tails of two distributions; specifically, between the conditional distributions

$$F_t(x) = \mathbb{P}(X \leq x | X > t) \quad \text{and} \quad G_t(x) = \mathbb{P}(Y \leq x | Y > t), \quad x > t,$$

for large  $t$ . Let  $\mathbb{P}_{X,t}$  and  $\mathbb{P}_{Y,t}$  be the two probability measures on  $((t, \infty), \mathcal{B}(t, \infty))$  induced by  $F_t$  and  $G_t$ , where  $\mathcal{B}(t, \infty)$  is the collection of Borel sets on  $(t, \infty)$ . That is,  $\mathbb{P}_{X,t}$  and  $\mathbb{P}_{Y,t}$  are defined by, respectively,

$$\mathbb{P}_{X,t}(\cdot) = \mathbb{P}(X \in \cdot | X > t) \quad \text{and} \quad \mathbb{P}_{Y,t}(\cdot) = \mathbb{P}(Y \in \cdot | Y > t).$$

Below we provide some background on probability distances and divergence measures, as well as tail characterization via regular variation. Our study will be undertaken within the latter framework.

### 2.2. Probability distances and divergence measures

In this section, we provide a brief introduction to the notions of *probability distance* and *divergence*. Simply speaking, a probability distance is a proper metric over a metric space of probability measures, which is popular in statistical theory, probability theory and general measure theory. On the other hand, a divergence is a measure of dissimilarity that is a popular in information theory, machine learning and data science. The general definition of divergence is given in terms of general Radon-Nikodym derivatives, but for presentation purposes, we restrict ourselves to a simpler form that assumes densities with respect to Lebesgue measure.

To provide the definition of a probability distance, we first define a probability semidistance, as follows.

**Definition 1** (Probability semidistance). *Let  $\mathcal{X}$  be the set of univariate real-valued random variables. A mapping  $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is called a (probability) semidistance if the following conditions are satisfied:*

- i)  $d(X_1, X_2) \geq 0$  for any  $X_1, X_2 \in \mathcal{X}$ ;*
- ii)  $d(X_1, X_2) = 0$  if  $X_1$  and  $X_2$  share the same distribution;*
- iii)  $d(X_1, X_2) = d(X_2, X_1)$  for any  $X_1, X_2 \in \mathcal{X}$ ;*
- iv)  $d(X_1, X_2) + d(X_2, X_3) \geq d(X_1, X_3)$  for any  $X_1, X_2, X_3 \in \mathcal{X}$ .*

When there is no risk of confusion, we simply call  $d(\cdot, \cdot)$  a probability distance. Technically, for  $d(\cdot, \cdot)$  to define a probability distance, instead of condition *ii)* above, we need a stronger condition:  $d(X_1, X_2) = 0$  if and only if  $X_1$  and  $X_2$  share the same distribution. Note that the probability semidistance (resp., distance) between  $(X_1, X_2)$  only depends on the marginal

distributions of  $X_1$  and  $X_2$  and is sometimes called a simple semidistance (resp., distance). By contrast, compound probability semidistances/distances between  $(X_1, X_2)$  are determined by its joint distribution, though not considered in the sequel. A more comprehensive discussion on probability distances can be found in [Rachev et al. \(2013\)](#).

Below we list some well-known probability distances for two probability measures associated with  $X_1, X_2 \in \mathcal{X}$ :

1) *Total Variation*

$$d_{TV}(X_1, X_2) := \sup_{A \in \mathcal{F}} |\mathbb{P}(X_1 \in A) - \mathbb{P}(X_2 \in A)|;$$

2) *Kolmogorov–Smirnov*

$$d_{KS}(X_1, X_2) := \sup_{x \in \mathbb{R}} |\mathbb{P}(X_1 \leq x) - \mathbb{P}(X_2 \leq x)|;$$

clearly,  $d_{KS}(X_1, X_2) \leq d_{TV}(X_1, X_2)$ ;

3) *Wasserstein*

$$d_{W_p}(X_1, X_2) := \left( \int_0^1 |q_{X_1}(s) - q_{X_2}(s)|^p \, ds \right)^{1/p} \text{ for any } p \geq 1,$$

where  $q_X(s) := \inf_{x \in \mathbb{R}} \{\mathbb{P}(X \leq x) \geq s\}$  is the  $s^{\text{th}}$  quantile of  $X$ ; note that the special case with  $p = 1$  satisfies

$$d_{W_1}(X_1, X_2) = \int_{\mathbb{R}} |\mathbb{P}(X_1 > x) - \mathbb{P}(X_2 > x)| \, dx;$$

4) *Lévy–Prokhorov*

$$d_L(X_1, X_2) := \inf_{\epsilon > 0} \left\{ \mathbb{P}(X_1 \leq x - \epsilon) - \epsilon \leq \mathbb{P}(X_2 \leq x) \leq \mathbb{P}(X_1 \leq x + \epsilon) + \epsilon, \text{ for all } x \in \mathbb{R} \right\}.$$

The notion of divergence is less strict than a probability distance, and requires only the first two conditions in [Definition 1](#). The literature on divergence measures is quite rich, but the class of  $\phi$ -divergence captures the most attention; a quick review appears in [Ben-Tal et al. \(2013\)](#), while a wider discussion is provided in [Pardo \(2005\)](#). A formal definition of the  $\phi$ -divergence is given as follows.

**Definition 2** ( $\phi$ -divergence). *Let  $p, q : \Omega \rightarrow \mathbb{R}_+$  be the densities of two univariate real-valued random variables. The divergence measure of  $(p, q)$  corresponding to the function  $\phi : \mathbb{R}_+ \rightarrow \mathbb{R}$  is defined as*

$$d_\phi(p, q) := \int_{\mathbb{R}} q(x) \phi\left(\frac{p(x)}{q(x)}\right) \, dx,$$

where  $\phi(\cdot)$  is convex on  $\mathbb{R}_+$  and  $\phi(1) = 0$ . By convention,  $0^+ \phi\left(\frac{a}{0^+}\right) := a \lim_{t \rightarrow \infty} \frac{\phi(t)}{t}$  for any  $a > 0$ , and  $0^+ \phi\left(\frac{a}{0^+}\right) := 0$  if  $a = 0$ .

Denote by  $\Phi$  the set of functions  $\phi$  that satisfy the conditions in [Definition 2](#); further, let  $\Phi^* \subseteq \Phi$  be the set of functions  $\phi$  that are differentiable at 1 with derivative equal to 0. Now,

if  $\phi \in \Phi$  is differentiable at 1—since  $\phi$  is convex, it is differentiable almost everywhere—then  $\psi(\cdot) := \phi(\cdot) - \phi'(1)(t - 1)$  on  $\mathbb{R}_+$  satisfies

$$\psi \in \Phi^*, d_\phi(p, q) = d_\psi(p, q), \psi'(1) = 0, \psi(\cdot) \geq 0 \text{ on } \mathbb{R}_+,$$

and thus, from now on, for  $\phi \in \Phi$  differentiable at 1, we assume without loss of generality that  $\phi \in \Phi^*$ ; further details can be found in [Pardo \(2005\)](#). We prefer to use divergence measures with  $\phi$  functions in  $\Phi^*$ , since these  $\phi$ -divergences are non-negative on the entire domain.

Definition 2 ensures only the first condition in Definition 1 holds, i.e.,  $d_\phi(p, q) \geq 0$  for any  $\Phi \in \Phi$ . Proposition 1.1 of [Pardo \(2005\)](#) shows that for any  $\phi \in \Phi^*$  that is strictly convex in a neighbourhood of 1,  $d_\phi(p, q) = 0$  if and only if  $p(\cdot) = q(\cdot)$  on  $\Omega$ . Therefore,  $\phi \in \Phi^*$  is a large class of divergence measures that satisfy the first two conditions of a proper probability distance as explained in Definition 1. The symmetry property is tackled in the next paragraph, but the triangle inequality is hardly satisfied by  $\phi$ -divergences, and thus most  $\phi$ -divergence measures are not proper probability distances.

We now explain how to construct a symmetric  $\phi$ -divergence. First, by definition, if  $\phi \in \Phi$  or  $\phi \in \Phi^*$ , then  $\tilde{\phi} \in \Phi$  and  $\tilde{\phi} \in \Phi^*$ , respectively, where  $\tilde{\phi}$  is known as the *adjoint divergence* function corresponding to  $\phi$  with  $\tilde{\phi}(\cdot) := \cdot\phi(1/\cdot)$  on  $\mathbb{R}_+$ ; further, if  $\phi(\cdot) = \tilde{\phi}(\cdot)$  on  $\mathbb{R}_+$ , then  $\phi$  is called *self-adjoint divergence*. One can show that  $d_\phi(p, q) = d_{\tilde{\phi}}(q, p)$  (even if  $\Phi$  is not self-adjoint); for details, see ([Ben-Tal et al., 1991, 2013](#)). Therefore, a straightforward way to construct a symmetric  $\phi$ -divergence for any  $\phi \in \Phi$  is to consider  $\frac{\phi + \tilde{\phi}}{2} \in \Phi$  since

$$d_{\frac{\phi + \tilde{\phi}}{2}}(p, q) = \frac{1}{2}d_\phi(p, q) + \frac{1}{2}d_{\tilde{\phi}}(p, q) = \frac{1}{2}d_\phi(p, q) + \frac{1}{2}d_\phi(q, p). \quad (2.1)$$

Clearly,  $\frac{\phi + \tilde{\phi}}{2} \in \Phi^*$  if  $\phi \in \Phi^*$ . Secondly, any  $\phi \in \Phi$  (or  $\phi \in \Phi^*$ ) has another corresponding symmetric  $\phi$ -divergence given by

$$\frac{1}{2}d_\phi\left(p, \frac{p+q}{2}\right) + \frac{1}{2}d_\phi\left(q, \frac{p+q}{2}\right) = d_{\hat{\phi}}(p, q), \quad (2.2)$$

where  $\hat{\phi}(x) := (x+1)\left(\phi\left(\frac{2x}{x+1}\right) + \phi\left(\frac{2}{x+1}\right)\right)$  on  $\mathbb{R}_+$ , and  $\hat{\phi} \in \Phi$  (or  $\hat{\phi} \in \Phi^*$ ).

In summary, almost any  $\phi$ -divergence from Definition 2 satisfies the first two conditions of a probability distance as given in Definition 1, and it can be transformed into a symmetric  $\phi$ -divergence as in (2.1) and (2.2). Below we enunciate some well-known  $\phi$ -divergences belonging to  $\Phi^*$  (see Table A.1 and Figure A.4 in [Appendix A](#) for a summary) with some interesting relationships; a more exhaustive list can be found in e.g. [Pardo \(2005\)](#). Additionally, some transformations of well-known  $\phi$ -divergences become proper probability distances in the sense of Definition 1. A detailed discussion of this matter is provided in [Endres and Schindelin \(2003\)](#) and the list of divergences below provides such transformations when applicable.

- 1) *Kullback-Leibler divergence* with  $\phi_{KL}(x) := x \log x - x + 1$ ; note that  $\phi_{KL} \in \Phi^*$  and is the transformation of the standard formulation of the Kullback-Leibler divergence with  $\tilde{\phi}_{KL}(t) = x \log x$  that satisfies  $\tilde{\phi}_{KL} \in \Phi \setminus \Phi^*$ ;
- 2) *Burg divergence* with  $\phi_B(x) := -\log x + x - 1$ ; note that  $\tilde{\phi}_B(\cdot) = \phi_{KL}(\cdot)$  on  $\mathbb{R}_+$ , i.e., the

Burg divergence is the ad-joint transformation of the Kullback-Leibler divergence;

- 3) *J-divergence* with  $\phi_J(x) := (x - 1) \log x$ ; note that  $\phi_J = \frac{\phi_{KL} + \tilde{\phi}_{KL}}{2}$ , i.e., the J-divergence is the symmetric Kullback-Leibler divergence through (2.1), and thus, condition iii) in Definition 1 is satisfied;
- 4) *Jensen-Shannon divergence* with  $\phi_{JS}(x) = x \log \left( \frac{2x}{x+1} \right) + \log \left( \frac{2}{x+1} \right)$ ; note that  $\phi_{JS} \in \Phi^*$ , and that  $\sqrt{d_{\phi_{JS}}(\cdot)}$  and  $\sqrt{d_{\tilde{\phi}_{JS}}(\cdot)}$  satisfy all conditions in Definition 1; finally,  $\phi_{JS}$  is the symmetric Kullback-Leibler divergence through (2.2);
- 5)  $\chi^2$ -divergence with  $\phi_C(x) := \frac{(x-1)^2}{x}$ ;
- 6) *Modified  $\chi^2$ -divergence* with  $\phi_{MC}(x) := (x - 1)^2$ ; note that  $\tilde{\phi}_{MC}(\cdot) = \phi_C(\cdot)$  on  $\mathbb{R}_+$ , i.e., the Modified  $\chi^2$ -divergence is the ad-joint transformation of the  $\chi^2$ -divergence;
- 7) *Hellinger-divergence* with  $\phi_H(x) := (\sqrt{x} - 1)^2$ ; note that the Hellinger-divergence is self-adjoint, i.e.  $\tilde{\phi}_H(\cdot) = \phi_H(\cdot)$ , and thus, condition iii) in Definition 1 is satisfied; this appears in the literature as the *Squared Hellinger-divergence*, and one can show that  $\sqrt{d_{\phi_H}(\cdot)}$  satisfies all conditions in Definition 1;
- 8) *Variation-divergence* with  $\phi_V(x) := |x - 1|$ ; the Variation-divergence is self-adjoint, i.e.  $\tilde{\phi}_V = \phi_V$ , and thus, symmetric; one can show that  $\sqrt{d_{\phi_V}(\cdot)}$  satisfies all conditions in Definition 1 as it is closely related to the Total Variation probability distance;
- 9)  $\chi$ -divergence of order  $\theta > 1$  with  $\phi_\theta(x) := |x - 1|^\theta$  and  $\theta > 1$ ;
- 10) *Le Cam divergence* with  $\phi_{LC}(x) := \frac{(x-1)^2}{4(x+1)}$ ; for details, see [Le Cam \(1986\)](#); one can show that  $\sqrt{d_{\phi_{LC}}(\cdot)}$  satisfies all conditions in Definition 1;
- 11) *Cressie-Read divergence* with  $\phi_{CR}(x; \theta) := \frac{1-\theta+x\theta-x^\theta}{\theta(1-\theta)}$  where  $\theta \in \mathbb{R} \setminus \{0, 1\}$ , which is also known as the *power divergence*; for details, see [Cressie and Read \(1984\)](#); note that  $\phi_{CR}(\cdot; -1) = \frac{1}{2}\phi_C(\cdot)$ ,  $\phi_{CR}(\cdot; 1/2) = 2\phi_H(\cdot)$  and  $\phi_{CR}(\cdot; 2) = \frac{1}{2}\phi_{MC}(\cdot)$  on  $(0, \infty)$ .

We should mention that the Cressie-Read divergence class is used in the statistical literature for goodness-of-fit testing; e.g., [Cressie and Read \(1984\)](#) uses it for testing multinomial data, and [Jager and Wellner \(2007\)](#) uses two novel statistics in a more general context that rely on the general  $\phi$ -divergence test statistics in ([Ali and Silvey, 1966](#); [Csiszár, 1966](#)).

There are many inequalities amongst the  $\phi$ -divergences, and two notable ones are

$$\frac{1}{2}d_{\phi_H}(p, q) \leq d_{\phi_V}(p, q) \leq \sqrt{d_{\phi_H}(p, q) \left( 1 - \frac{1}{4}d_{\phi_H}(p, q) \right)} \leq 1 \quad (2.3)$$

and

$$d_{\phi_V}(p, q) \leq \sqrt{\frac{1}{2}d_{\phi_{KL}}(p, q)} \quad (2.4)$$

for any two densities  $p$  and  $q$ . Note that (2.4) is the well-known Pinsker's inequality. Useful inequalities, such (2.3) and (2.4), appear in the literature, and comprehensive collections of results are provided in ([Sason and Verdu, 2016](#); [Tsybakov, 2009](#)). Elegant proofs for showing that such inequalities are tight is possible via the so-called *joint range* of a given pair of  $\phi$ -divergences; for details, see [Harremoës and Vajda \(2011\)](#).

### 2.3. Regular variation

For our tail analysis, we will focus on distributions with heavy tails; specifically, tails that are regularly varying. Regular variation has long been used to model a large class of heavy-tailed distributions, that is, every distribution in the Fréchet max-domain of attraction. In this section, we provide some preliminaries on first- and second-order regular variation.

**Definition 3** (Regular variation). *A positive measurable function  $h(\cdot)$  is said to be regularly varying at  $\infty$  with index  $-\alpha \in \mathbb{R}$ , written as  $h \in \text{RV}_{-\alpha}$ , if for every  $x > 0$*

$$\lim_{t \rightarrow \infty} \frac{h(xt)}{h(t)} = x^{-\alpha}.$$

Hereafter, all limiting relations are according to  $t \rightarrow \infty$  unless otherwise stated.

We are particularly interested in the case where the distributions have close tails and distinction is challenging, especially when the distributions share the same tail index. For the case with distinct tail indexes, the difference between their tail behaviors are easier to quantify and our approach readily extends to derive analogous results. Yet for the equal-index case, we require a more nuanced description of the tails to obtain non-trivial results. Specifically, we impose a second-order regular variation assumption on both tails being compared.

**Definition 4** (Second-order regular variation). *A positive measurable function  $h(\cdot)$  is said to be second-order regularly varying with indexes  $-\alpha \in \mathbb{R}$  and  $\rho \leq 0$ , written as  $h \in 2\text{RV}_{-\alpha, \rho}$ , if for some constant  $k \neq 0$  and some function  $A(\cdot)$  with  $\lim_{t \rightarrow \infty} A(t) = 0$  that ultimately has a constant sign, it holds that*

$$\lim_{t \rightarrow \infty} \frac{h(xt)/h(t) - x^{-\alpha}}{A(t)} = kx^{-\alpha} \int_1^x u^{\rho-1} du, \quad x > 0. \quad (2.5)$$

Here  $A(t)$  is called the auxiliary function and known to possess regular variation. Specifically, we have  $|A(\cdot)| \in \text{RV}_{\rho}$ . With a slight abuse of notation, we denote by  $\text{sgn}(A(\infty^-))$  the constant sign of  $A(t)$  for  $t$  large, which is equal to 1 if  $A(t)$  is positive and  $-1$  if it is negative.

By carefully choosing the auxiliary function, one can ensure  $k = 1$  in (2.5). Thus, in the sequel we assume without loss of generality that  $k = 1$  in the definition of  $2\text{RV}$ . Note that the  $2\text{RV}$  condition in (2.5) implies that

$$\frac{h(xt)}{h(t)} = x^{-\alpha} + x^{-\alpha} \int_1^x u^{\rho-1} du A(t) + o(A(t)). \quad (2.6)$$

Also note that the convergence in relation (2.5) holds locally uniformly and that we have certain Drees-type inequalities. The following proposition is a version from Mao (2013).

**Proposition 5** (Dree-type inequality; Mao (2013)). *If  $h \in 2\text{RV}_{-\alpha, \rho}$  with indexes  $-\alpha \in \mathbb{R}$  and  $\rho \leq 0$ , then, for any fixed  $\varepsilon, \delta > 0$ , there exists  $t_0 = t_0(\varepsilon, \delta, \alpha, \rho) > 0$ , such that, for all  $\min(t, xt) > t_0$ , we have*

$$\left| \frac{h(xt)/h(t) - x^{-\alpha}}{A(t)} - x^{-\alpha} \int_1^x u^{\rho-1} du \right| \leq \varepsilon x^{-\alpha} \left( \left| \int_1^x u^{\rho-1} du \right| + x^{\rho} \max\{x^{\delta}, x^{-\delta}\} \right).$$

While some Drees-type inequalities require careful choices of auxiliary function, the version above uses precisely the  $A(t)$  function in (2.5); see also, e.g., Theorem 2.3.9 of [de Haan and Ferreira \(2007\)](#). Note that Proposition 5 implies that, if  $\alpha > 0$ , then equations (2.5) and (2.6) hold uniformly over  $[x_0, \infty)$  for every  $x_0 > 0$ , in the sense that

$$\lim_{t \rightarrow \infty} \sup_{x \geq x_0} \left| \frac{h(xt)/h(t) - x^{-\alpha}}{A(t)} - x^{-\alpha} \int_1^x u^{\rho-1} du \right| = 0. \quad (2.7)$$

In asymptotic analysis for regularly varying functions, uniform asymptotic relations usually hold locally on bounded intervals. Below we show a useful result stating that under certain second-order conditions, the property holds on intervals of form  $[x_0, \infty)$  for any  $x_0 > 0$ . The proof is relegated to [Appendix D](#).

**Lemma 6** (Uniformity). *If  $h \in 2RV_{-\alpha, \rho}$  with  $\alpha > 0$  and  $\rho < 0$ , then  $h(xt) \sim x^{-\alpha}h(t)$  holds uniformly over  $[x_0, \infty)$  for any  $x_0 > 0$ , that is*

$$\lim_{t \rightarrow \infty} \sup_{x \geq x_0} \left| \frac{h(xt)}{x^{-\alpha}h(t)} - 1 \right| = 0.$$

Hereafter, for a non-decreasing function  $h$  on  $\mathbb{R}$ , we write its left-continuous inverse function as  $h^{\leftarrow}(\cdot) = \inf\{x \in \mathbb{R} : h(x) \geq \cdot\}$ . For a distribution function  $H$ , write its tail quantile function as  $U_H(\cdot) = H^{\leftarrow}(1 - 1/\cdot) = (1/\bar{H})^{\leftarrow}(\cdot)$  and write  $\bar{H}^{\leftarrow}(\cdot) = H^{\leftarrow}(1 - \cdot)$ .

The following lemma, proved in [Appendix D](#), provides properties related to the inverse function of a 2RV distribution function.

**Lemma 7** (Properties of  $U_H$  and  $\bar{H}$ ). *Let  $\bar{H} \in 2RV_{-\nu, \rho}$  for some  $\nu > 0$  and  $\rho \leq 0$  with auxiliary function  $A(\cdot)$ . Then*

- a)  $U_H \in 2RV_{1/\nu, \rho/\nu}$  and a possible choice of auxiliary function is  $a(\cdot) = \nu^{-2}A(U_H(\cdot))$ ;
- b)  $\bar{H}^{\leftarrow}(\bar{H}(t)) \sim t$  and  $\bar{H}^{\leftarrow}(\bar{H}(t))/t - 1 = o(A(t))$ .

We are now ready to present our main results.

### 3. Main Results

Our first main result is an asymptotic characterization of the  $p$ -Wasserstein distance and second main result is one of the  $\phi$  divergence. The Wasserstein distance and  $\phi$  divergence both depend only on the tails of the marginal distributions, and hereafter, are denoted by  $d_{W,p}(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})$  and  $d_\phi(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})$ , respectively.

#### 3.1. Wasserstein distance

Recall that the  $p$ -Wasserstein distance between  $\mathbb{P}_{X,t}$  and  $\mathbb{P}_{Y,t}$  with  $p \geq 1$  is given by

$$d_{W,p}(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t}) = \left( \int_0^1 |\bar{F}_t^{\leftarrow}(q) - \bar{G}_t^{\leftarrow}(q)|^p dq \right)^{1/p}. \quad (3.1)$$

Some technical set of assumptions are needed in this section, namely, some standard second-order regular variation for the survival functions, summarized as follows.

**Assumption 3.1** (Second-order RV of tails). Assume that  $\bar{F} \in 2\text{RV}_{-\nu, \rho_X}$  and  $\bar{G} \in 2\text{RV}_{-\nu, \rho_Y}$ , with  $\nu > 1$ ,  $\rho_X \neq \rho_Y$ <sup>1</sup>, and auxiliary functions  $A_X$  and  $A_Y$ , respectively.

With the help of Assumption 3.1, we obtain the following asymptotic result that establishes the rate of convergence for the Wasserstein distance.

**Theorem 8** (Wasserstein distance asymptotic). Suppose Assumption 3.1 holds with  $\nu > p$  and  $\rho_{XY} := \max\{\rho_X, \rho_Y\} \leq 0$ . Then, for  $p \in [1, \infty)$ , it holds that

$$d_{W,p}(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t}) \sim C(\nu, \rho_{XY}, p) t |A_{XY}(t)|, \quad (3.2)$$

where

$$A_{XY} = A_X I_{\{\rho_X > \rho_Y\}} + A_Y I_{\{\rho_X < \rho_Y\}},$$

and the constant  $C(\nu, \rho_{XY}, p) \in (0, \infty)$  is given by

$$C(\nu, \rho_{XY}, p) = \frac{\left( \int_1^\infty (z^{1/\nu-2} - z^{(\rho_{XY}+1)/\nu-2})^p dz \right)^{1/p}}{\nu |\rho_{XY}|}.$$

When  $\rho_{XY} = 0$ , the constant  $C(\nu, \rho_{XY}, p)$  is interpreted as

$$C(\nu, 0, p) = \lim_{\rho_{XY} \rightarrow 0} C(\nu, \rho_{XY}, p) = \nu^{-2} \left( \int_1^\infty z^{p/\nu-2p} (\log z)^p dz \right)^{1/p}.$$

It is straightforward to verify that for  $p = 1$ , the constant reduces to

$$C(\nu, \rho_{XY}, 1) = \frac{1}{(\nu-1)(\nu-\rho_{XY}-1)}.$$

We see that under the conditions above, the Wasserstein distance is regularly varying with index  $\rho_{XY} + 1$ . The second-order difference between the two distributions determines whether the Wasserstein distance vanishes or explodes.

### 3.2. Phi-divergence

We now study the behavior of the  $\phi$ -divergence, given by

$$d_\phi(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t}) = \int_t^\infty \phi\left(\frac{f_t(x)}{g_t(x)}\right) g_t(x) dx, \quad (3.3)$$

where  $f_t$  and  $g_t$  are the densities of  $\mathbb{P}_{X,t}$  and  $\mathbb{P}_{Y,t}$ . That is,

$$f_t(x) = \frac{f(x)}{\bar{F}(t)}, \quad g_t(x) = \frac{g(x)}{\bar{G}(t)}, \quad x > t,$$

where  $f$  and  $g$  are the densities of  $F$  and  $G$ , respectively.

---

<sup>1</sup>The condition  $\rho_X \neq \rho_Y$  ensures that the two distributions are at least distinguishable at the second order. The case with equal second-order indexes would require a cumbersome higher-order expansion to solve. We restrict our attention to the current setup.

We assume that the function  $\phi$  behaves like a power function around 1, with possibly different behavior on the left and right of 1. The mathematical formulation is provided as follows.

**Assumption 3.2** (Local behavior). *As  $h \rightarrow 0$ , the function  $\phi$  satisfies*

$$\phi(1+h) - ah \sim \zeta(h) := k^+ h_+^\lambda + k^- h_-^\lambda, \quad (3.4)$$

for some  $a \in \mathbb{R}$  and some non-degenerate convex function  $\zeta(\cdot)$  on  $\mathbb{R}$  given above, where  $h_+ = h1_{(h>0)}$  and  $h_- = -h1_{(h<0)}$  denote the positive and negative parts of  $h$ , respectively.

Here, the constant  $a$  in equation (3.4) can be 0. When one of  $k^+$  and  $k^-$  is zero, we follow the convention that a quantity asymptotically equivalent to zero is zero.

By Assumption 3.2, we are assuming  $\phi(1+h)$  can be approximated  $ah + \zeta(h)$  for small  $h$ . Note that the decomposition into a linear term plus a convex function may not be unique; for details, see Example 1.

**Remark 9.** *Because the functions  $\phi(t)$  and  $\phi(t) + c(t-1)$ —provided that they are proper generating functions for the corresponding divergence—result in the same  $\phi$ -divergence for any pair of distributions and for any  $c \in \mathbb{R}$ , our anticipated asymptotic approximation should be invariant with respect to the value of  $a$  in Assumption 3.2. As it turns out, this is indeed the case.*

**Example 1.** *All  $\phi$  functions mentioned in Section 2 satisfy Assumption 3.2. Here are some examples:*

- *For variation divergence (or,  $\chi$ -divergence with  $\theta = 1$ ), (3.4) is satisfied with  $\lambda = 1$  and infinitely many choices of  $a$ ,  $k^+$ , and  $k^-$ ; that is,  $(a, k^+, k^-) \in \{(c, 1-c, 1+c) : c \in \mathbb{R}\}$ .*
- *For Kullback-Leibler divergence, the function  $\phi(x) = x \log x$  satisfies (3.4) with*

$$(a, k^+, k^-) \in \{(c, 1-c, c-1) : c \in \mathbb{R}\}, \lambda = 1 \quad \text{or} \quad a = 1, k^+ = k^- = \frac{1}{2}, \lambda = 2$$

*and the function  $\phi(x) = x \log x - x + 1$  satisfies (3.4) with*

$$(a, k^+, k^-) \in \{(c, -c, c) : c \in \mathbb{R} \setminus \{0\}\}, \lambda = 1 \quad \text{or} \quad a = 0, k^+ = k^- = \frac{1}{2}, \lambda = 2.$$

Note that all previously-mentioned examples except of the variation divergence satisfy  $\phi'(1^+) = \phi'(1^-) = 0$ ; further, the variation divergence satisfies  $\phi'(1^+) = 1$  and  $\phi'(1^-) = -1$ . Secondly, all previously-mentioned examples (except the variation divergence and  $\chi$ -divergence with  $\theta \in [1, \infty) \setminus \{2\}$ ) satisfy  $\phi''(1^+) = \phi''(1^-) > 0$ ; specifically,  $\phi''(1) = 1$  for Kullback-Leibler and Burg divergences,  $\phi''(1) = 1/2$  for Jensen-Shannon and Hellinger divergences,  $\phi''(1) = 2$  for J,  $\chi^2$  and Modified- $\chi^2$  divergences,  $\phi''(1) = 1/4$  for Le Cam divergence,  $\phi''(1) = 1$  for Cressie-Read divergences.

In addition to Assumption 3.2, we require some knowledge of the tail behavior of  $X$  and  $Y$  that mirrors Assumption 3.1. That is, we assume 2RV conditions on the densities, which are given as follows.

**Assumption 3.3** (Second-order RV of densities). *The density functions  $f$  and  $g$  satisfy  $f \in 2RV_{-\alpha, \rho_X}$  and  $g \in 2RV_{-\alpha, \rho_Y}$  for some  $\alpha > 1$ , some  $\rho_X \neq \rho_Y$ , both negative, and respective auxiliary functions  $A_X$  and  $A_Y$ .*

A link between Assumptions 3.1 and 3.3 is now provided. Essentially, Assumption 3.3 implies Assumption 3.1, and their tail indexes are linked through  $\nu = \alpha - 1$ , as is shown below. That is, a sufficient condition for 2RV to hold for the densities can be found in Theorem 2.1 of de Haan and Resnick (1996).<sup>2</sup> Further, Assumption 3.3 implies that  $\bar{F} \in 2RV_{-\nu, \rho_X}$  and that  $\bar{G} \in 2RV_{-\nu, \rho_Y}$  with  $\nu = \alpha - 1$ , as stated in the following lemma, which is built on Proposition 6 of Hua and Joe (2011) and will be useful for our asymptotic approximations.

**Lemma 10** (Transfer of 2RV). *Let  $h \in 2RV_{-\alpha, \rho}$ ,  $\alpha > 1$ ,  $\rho < 0$ , with auxiliary function  $A(\cdot)$ . Then*

$$\frac{th(t)}{\int_t^\infty h(x) dx} - \nu = -\frac{\nu}{\nu - \rho}A(t) + o(A(t)),$$

where  $\nu = \alpha - 1$ . Moreover,  $\int_t^\infty h(x) dx \in 2RV_{-\nu, \rho}$  with auxiliary function  $A^*(t) = A(t)\nu/(\nu - \rho)$ .

From this point on, our results will be presented in terms of the tail index of the survival functions,  $\nu$ . We are now ready to provide the main result of this section, given as follows.

**Theorem 11** ( $\phi$ -divergence asymptotic). *Suppose that Assumptions 3.2 and 3.3 hold with  $\rho_X > \rho_Y$ . Then we have*

$$\lim_{t \rightarrow \infty} \frac{d_\phi(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})}{|A_X(t)|^\lambda} = C(\nu, \rho_X, \lambda), \quad (3.5)$$

where  $\nu = \alpha - 1$  and  $C(\nu, \rho_X, \lambda)$  is a finite non-negative constant given by

$$\frac{\nu}{(-\rho_X)^\lambda} \int_1^\infty \zeta \left( \text{sgn}(A_X(\infty^-)) \left( \frac{\nu}{\nu - \rho_X} - z^{\rho_X} \right) \right) z^{-(\nu+1)} dz. \quad (3.6)$$

**Remark 12.** *Following a similar proof, we can show that if  $\rho_Y > \rho_X$ , then, similarly to (3.5), we have*

$$\lim_{t \rightarrow \infty} \frac{d_\phi(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})}{|A_Y(t)|^\lambda} = C(\nu, \rho_Y, \lambda)$$

holds with  $C(\nu, \rho_Y, \lambda)$  given by

$$\frac{\nu}{(-\rho_Y)^\lambda} \int_1^\infty \zeta \left( \text{sgn}(A_Y(\infty^-)) \left( z^{\rho_Y} - \frac{\nu}{\nu - \rho_Y} \right) \right) z^{-(\nu+1)} dz.$$

---

<sup>2</sup>Specifically, if the functions

$$\tilde{A}_X(t) := \frac{tf'(t)}{f(t)} - \alpha \quad \text{and} \quad \tilde{A}_Y(t) := \frac{tg'(t)}{g(t)} - \alpha$$

have constant sign near  $\infty$  and satisfy, as  $t \rightarrow \infty$ ,

$$\tilde{A}_X(t) \rightarrow 0 \quad \text{with} \quad \left| \tilde{A}_X(t) \right| \in RV_{\rho_X}, \quad \text{and} \quad \tilde{A}_Y(t) \rightarrow 0 \quad \text{with} \quad \left| \tilde{A}_Y(t) \right| \in RV_{\rho_Y}$$

for some  $\rho_X \leq 0$  and  $\rho_Y \leq 0$ , then both the density functions  $f$  and  $g$  and the survival functions  $\bar{F}$  and  $\bar{G}$  are 2RV.

**Remark 13.** Recall that  $\phi'(1^+)$  and  $\phi'(1^-)$  exist, and  $\phi'(1^+) \geq \phi'(1^-)$  holds due to the convexity of  $\phi$ , and thus, for any  $\phi \in \Phi$ , (3.4) holds with

$$(a, k^+, k^-) \in \{(c, \phi'(1^+) - c, -\phi'(1^-) + c) : c \in \mathbb{R}\}, \lambda = 1.$$

Further, if  $\phi'(1^+) = \phi'(1^-)$  denoted as  $\phi'(1)$ , and  $\phi''(1^+)$  and  $\phi''(1^-)$  exist, then another possible choice is

$$a = \phi'(1), k^+ = \frac{\phi''(1^+)}{2}, k^- = \frac{\phi''(1^-)}{2}, \lambda = 2.$$

Since all previously-mentioned divergences (except variation divergence and  $\chi$ -divergence with  $\theta \in [1, \infty) \setminus \{2\}$ ), can be formulated to satisfy  $\phi'(1) = 0$  and  $\phi''(1) > 0$ , all these divergences satisfy Assumption 3.2 with  $a = 0$ ,  $k^+ = k^- > 0$  and  $\lambda = 2$ , and in turn,  $C(\nu, \rho_X, \lambda)$  in (3.5) is simplified to

$$C(\nu, \rho_X, 2) = \frac{\phi''(1)}{2(-\rho_X)^2} \mathbb{E} \left[ \left( Z^{\rho_X} - \frac{\nu}{\nu - \rho_X} \right)^2 \right] \in (0, \infty), \quad (3.7)$$

where  $Z$  is Pareto distributed with survival function  $\mathbb{P}(Z > z) = z^{-\nu}$  on  $[1, \infty)$ .

Similarly, these divergences satisfy Assumption 3.2 for some  $(a, k^+, k^-)$  and  $\lambda = 1$  such that  $k^+ = -k^-$ , and the corresponding constant in (3.5) is  $C(\nu, \rho_X, 1) = 0$ , regardless of the sign of  $A_X$ . In fact, this can also be seen as a consequence of the conclusion for  $\lambda = 2$  preceding (3.7).

**Remark 14.** The function  $\zeta(\cdot)$  is convex and non-degenerate in a neighbourhood of 0, and thus,  $\lambda \geq 1$ , but we must either have  $\lambda > 1$  and  $k^+, k^- \geq 0$  with  $k^+ + k^- > 0$ , or  $\lambda = 1$  and  $k^+ + k^- \geq 0$ . Strict convexity of  $\zeta(\cdot)$  in a neighbourhood of 0 holds if and only if  $\lambda > 1$  and  $k^+, k^- > 0$ . Therefore, strict convexity of  $\zeta(\cdot)$  in a neighbourhood of 0 ensures (3.5) holds with positive asymptotic constants; further,  $C(\nu, \rho_X, \lambda)$  in (3.5) is the same regardless of the sign of  $A_X$  if  $k^+ = k^-$ , and one may find a similar expression (to (3.7), with  $Z$  defined in the same way) as follows:

$$C(\nu, \rho_X, \lambda) = \frac{k^+}{(-\rho_X)^\lambda} \mathbb{E} \left[ \left| Z^{\rho_X} - \frac{\nu}{\nu - \rho_X} \right|^\lambda \right] \in (0, \infty).$$

Theorem 11 and its two corollaries below help us establish some asymptotic results useful for constructing the tail similarity measure.

**Corollary 15** (Ratio of  $\phi$ -divergences). Suppose that  $\phi_1$  and  $\phi_2$  satisfy Assumption 3.2 with  $k_1^+ = k_1^- = k_1 > 0$ ,  $k_2^+ = k_2^- = k_2 > 0$ , and  $\lambda_1 = \lambda_2$ . Further, assume that Assumption 3.3 holds. Then, we have

$$\frac{d_{\phi_1}(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})}{d_{\phi_2}(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})} \sim \frac{k_1}{k_2}. \quad (3.8)$$

We have already shown that many  $\phi$ -divergences satisfy the condition in Corollary 15; namely, all previously-mentioned divergence examples, except for variation divergence and  $\chi$ -divergence with some choices of  $\theta$ .

A similar result to Corollary 15 is given to asymptotically compare the Wasserstein divergence with  $p = 1$  and a bespoke  $\phi$ -divergence, namely  $\phi^*(x) := \phi_{KL}(x)I_{\{x < 1\}} + \phi_V(x)I_{\{x \geq 1\}}$  that satisfies Assumption 3.2 with any tuple

$$(a, k^+, k^-) \in \{(c, 1 - c, c) : c \in \mathbb{R}\}, \lambda = 1.$$

Now, let  $z_0 = (\nu / (\nu - \rho_X))^{1/\rho_X} \in (1, \infty)$ ; that is,  $z_0$  is the unique solution of  $z^{\rho_X} - \frac{\nu}{\nu - \rho_X} = 0$ . Some algebraic manipulations show that, for  $A_X(\infty^-) > 0$ , the constant in (3.6) is given by

$$\begin{aligned} & \frac{\nu}{(-\rho_X)^\lambda} \int_1^\infty \zeta \left( \frac{\nu}{\nu - \rho_X} - z^{\rho_X} \right) z^{-(\nu+1)} dz \\ &= \frac{\nu k^-}{(-\rho_X)^\lambda} \left( \int_1^{z_0} \left( z^{\rho_X} - \frac{\nu}{\nu - \rho_X} \right) z^{-(\nu+1)} dz \right) + \frac{\nu k^+}{(-\rho_X)^\lambda} \left( \int_{z_0}^\infty \left( \frac{\nu}{\nu - \rho_X} - z^{\rho_X} \right) z^{-(\nu+1)} dz \right) \\ &= \frac{\nu (k^+ + k^-)}{(-\rho_X)^\lambda} \left( \int_{z_0}^\infty \left( \frac{\nu}{\nu - \rho_X} - z^{\rho_X} \right) z^{-(\nu+1)} dz \right) \\ &= \frac{1}{(-\rho_X)^\lambda} \mathbb{E} \left[ \max \left\{ \frac{\nu}{\nu - \rho_X} - Z^{\rho_X}, 0 \right\} \right], \end{aligned}$$

which a positive and finite constant since  $\rho_X < 0$ , where  $Z$  is Pareto distributed random variable defined as before. Note that  $\mathbb{E} \left[ Z^{\rho_X} - \frac{\nu}{\nu - \rho_X} \right] = 0$  is used in the second equality from above. The exact same expression can be derived analogously for  $A_X(\infty^-) < 0$ . Going through similar calculations for  $C(\nu, \rho_Y, \lambda)$  in Remark 12 for the case of  $\rho_Y > \rho_X$ . Hence, (3.5) is given by

$$\lim_{t \rightarrow \infty} \frac{d_{\phi^*}(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})}{|A_{XY}(t)|} = \frac{1}{(-\rho_{XY})^\lambda} \mathbb{E} \left[ \max \left\{ \frac{\nu}{\nu - \rho_{XY}} - Z^{\rho_{XY}}, 0 \right\} \right], \quad (3.9)$$

where

$$\rho_{XY} = \max\{\rho_X, \rho_Y\}, \quad \text{and} \quad A_{XY} = A_X I_{\{\rho_X > \rho_Y\}} + A_Y I_{\{\rho_X < \rho_Y\}}.$$

Now, if Assumption 3.3 holds, then, by Lemma 10, Assumption 3.1 holds with  $\nu = \alpha - 1$  and auxiliary functions

$$A_X^* = \nu A_X / (\nu - \rho_X) \quad \text{and} \quad A_Y^* = \nu A_Y / (\nu - \rho_Y).$$

The asymptotic result for the Wasserstein distance in (3.2) states that

$$\lim_{t \rightarrow \infty} \frac{d_W(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})}{t |A_{XY}^*(t)|} = C(\nu, \rho_{XY}, p) = \frac{\left( \int_1^\infty (z^{1/\nu-2} - z^{(\rho_{XY}+1)/\nu-2})^p dz \right)^{1/p}}{\nu |\rho_{XY}|}, \quad (3.10)$$

where  $A_{XY}^* = A_X^* I_{\{\rho_X > \rho_Y\}} + A_Y^* I_{\{\rho_X < \rho_Y\}}$ . Combining (3.9) and (3.10), we could conclude the following result:

**Corollary 16** (Wasserstein and  $\phi$ -divergence ratio). *Let  $\phi^*(t) = \phi_{KL}(t)I_{\{t < 1\}} + \phi_V(t)I_{\{t \geq 1\}}$ . Suppose that Assumption 3.3 holds. Then, we have*

$$\lim_{t \rightarrow \infty} \frac{d_W(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})}{t d_{\phi^*}(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})} = \frac{(-\rho_{XY})^{\lambda-1} \left( \int_1^\infty (z^{1/\nu-2} - z^{(\rho_{XY}+1)/\nu-2})^p dz \right)^{1/p}}{(\nu - \rho_{XY}) \mathbb{E} \left[ \max \left\{ \nu / (\nu - \rho_{XY}) - Z^{\rho_{XY}}, 0 \right\} \right]}, \quad (3.11)$$

where  $\mathbb{P}(Z > z) = z^{-\nu}$  on  $[1, \infty)$ .

The main difference between Corollaries 15 and 16 is that the asymptotic constant in (3.8) is not dependent on the unobserved tail indexes ( $\nu$ ,  $\rho_X$  and  $\rho_Y$ ), which is instead the case in (3.11). Although constructing a tail similarity measure based on the asymptotic approximation in Corollary 16 is possible, constructing one based on Corollary 15 is more straightforward. It suggests that the relative difference between the two sides of equation (3.8) serves as a measurement of the overarching tail similarity for two distributions with close tails, that is, they have the same first-order tail index but different second-order tail indexes. Specifically, for  $\phi_1$  and  $\phi_2$  that satisfy Assumption 3.2 with  $k_1^+ = k_1^- = k_1 > 0$ ,  $k_2^+ = k_2^- = k_2 > 0$ , and  $\lambda_1 = \lambda_2$ , tail similarity is measured by the relative difference:

$$r_{\phi_1; \phi_2}(X, Y) := \left| \frac{d_{\phi_1}(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t}) / d_{\phi_2}(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t}) - k_1/k_2}{k_1/k_2} \right| = \frac{k_2}{k_1} \left| \frac{d_{\phi_1}(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})}{d_{\phi_2}(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})} - \frac{k_1}{k_2} \right|. \quad (3.12)$$

Obviously, the measure depends on the choice of  $\phi_1$  and  $\phi_2$ . We shall use the  $\phi$  functions of Hellinger-divergence and  $\chi^2$ -divergence as an example to illustrate the performance and applications of this tail similarity measure in Section 4 below.

#### 4. Numerical Experiments

We present two sets of numerical experiments in this section: a simulation study in Section 4.1, followed by a real-life data analysis in Section 4.2. To this end, we first estimate  $d_{\phi_1}(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})$  and  $d_{\phi_2}(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})$  in Corollary 15 for the following  $\phi$ -divergence functions:

- *Hellinger-divergence* with  $\phi_H(x) := (\sqrt{x} - 1)^2$ ,  $k^+ = k^- = \frac{1}{4}$  and  $\lambda = 2$ ;
- $\chi^2$ -divergence with  $\phi_C(x) := \frac{(x-1)^2}{x}$ ,  $k^+ = k^- = 1$  and  $\lambda = 2$ .

For  $\phi \in \{\phi_H, \phi_C\}$ , a plug-in estimator of (3.3) is given by

$$d_{\phi}(\hat{\mathbb{P}}_{X,t}, \hat{\mathbb{P}}_{Y,t}) = \int_t^{\infty} \phi \left( \frac{\hat{f}_t(x)}{\hat{g}_t(x)} \right) \hat{g}_t(x) dx, \quad (4.1)$$

where  $\hat{f}_t$  and  $\hat{g}_t$  are suitable estimators of the densities of  $\mathbb{P}_{X,t}$  and  $\mathbb{P}_{Y,t}$ . These densities could be estimated via the *Generalised Pareto Distribution (GPD)* though not without computational issues via Maximum Likelihood Estimation (MLE) (see [Del Castillo and Serra \(2015\)](#)), and thus, we use a simpler approach. Namely we consider

$$\hat{f}_t(z) = \frac{1}{\hat{\xi}_X} \frac{1}{t} (z/t)^{-1/\hat{\xi}_X - 1}, \quad \hat{g}_t(z) = \frac{1}{\hat{\xi}_Y} \frac{1}{t} (z/t)^{-1/\hat{\xi}_Y - 1} \quad \text{for all } z > t,$$

where  $\hat{\xi}_X$  and  $\hat{\xi}_Y$  are suitable *extreme value index (EVI)* estimates. That is, if we observe a sample of size  $n$  that is decreasingly ordered,  $x_1 \geq x_2 \geq \dots \geq x_n$  and the optimal threshold is  $k$ , then  $\hat{f}_t(z) = \frac{1}{\hat{\xi}_X} \frac{1}{x_k} (z/x_k)^{-1/\hat{\xi}_X - 1}$  for all  $z > x_k$  with the Hill estimator (see [Hill \(1975\)](#) and also [McNeil et al. \(2015\)](#)) given by:

$$\hat{\xi}_{k,n}^H = \frac{1}{k} \sum_{i=1}^k \log x_i - \log x_{k+1}. \quad (4.2)$$

Thus, if  $n_X$  and  $n_Y$  are the sample sizes for  $X$  and  $Y$ , respectively, then (4.1) is estimated by

$$\begin{aligned} d_\phi \left( \hat{\mathbb{P}}_{X,k,n_X}, \hat{\mathbb{P}}_{Y,k,n_Y} \right) \\ = \int_{t_k}^{\infty} \phi \left( \hat{\xi}_{Y,k,n_Y} / \hat{\xi}_{X,k,n_X} (z/t_k)^{1/\hat{\xi}_{Y,k,n_Y} - 1/\hat{\xi}_{X,k,n_X}} \right) \frac{1}{\hat{\xi}_{Y,k,n_Y}} \frac{1}{t_k} (z/t_k)^{-1/\hat{\xi}_{Y,k,n_Y} - 1} dz, \end{aligned} \quad (4.3)$$

where  $t_k := x_k \vee y_k$  for every  $k$ . Note that both observed samples for  $X$  and  $Y$  are assumed to be decreasingly ordered samples. Taking  $\phi \in \{\phi_H, \phi_C\}$  in the above, we find that

$$d_{\phi_H} \left( \hat{\mathbb{P}}_{X,k,n_X}, \hat{\mathbb{P}}_{Y,k,n_Y} \right) = 2 - 4 \frac{\sqrt{\hat{\xi}_{X,k,n_X} \hat{\xi}_{Y,k,n_Y}}}{\hat{\xi}_{X,k,n_X} + \hat{\xi}_{Y,k,n_Y}}$$

and

$$d_{\phi_C} \left( \hat{\mathbb{P}}_{X,k,n_X}, \hat{\mathbb{P}}_{Y,k,n_Y} \right) = \frac{\left( \hat{\xi}_{X,k,n_X} - \hat{\xi}_{Y,k,n_Y} \right)^2}{\hat{\xi}_{Y,k,n_Y} \left( 2\hat{\xi}_{X,k,n_X} - \hat{\xi}_{Y,k,n_Y} \right)}. \quad (4.4)$$

Clearly, both vanish when the tail index estimators coincide. Note that for  $d_{\phi_C}$  to exist (i.e., for (4.3) to be integrable for  $\phi_C$ ), the tail index estimators need to satisfy  $\hat{\xi}_{X,k,n_X} / \hat{\xi}_{Y,k,n_Y} > 1/2$ , which also guarantees (4.4) is non-negative. If  $\hat{\xi}_{X,k,n_X} / \hat{\xi}_{Y,k,n_Y} \leq 1/2$ , then one could compare the pair  $(Y, X)$  instead of  $(X, Y)$ , but the dissimilarity measure may differ, since it is not always a commutative measure. In fact, Definition 1 tells us that all probability semidistances are symmetric, which is not the case for the class of phi-divergences; self-adjoint divergences are explained earlier to possess the symmetry property, and therefore, the Hellinger-divergence is symmetric, while  $\chi^2$ -divergence is not symmetric.

For simplicity, we assume from now on that the samples have equal sizes,  $n_X = n_Y = n$ , and compute the following tail similarity measure

$$\hat{r}_{\phi_H; \phi_C}(X, Y) = 4 \left| \frac{d_{\phi_H} \left( \hat{\mathbb{P}}_{X,k,n}, \hat{\mathbb{P}}_{Y,k,n} \right)}{d_{\phi_C} \left( \hat{\mathbb{P}}_{X,k,n}, \hat{\mathbb{P}}_{Y,k,n} \right)} - \frac{1}{4} \right|. \quad (4.5)$$

The measure is computed for various values of  $k$  and denoted by  $\hat{r}_{H;C}(k, n)$  for each choice of  $k$ . Since we want to understand how well our main result help distinguish similar tails, it is natural to compare  $\hat{r}_{H;C}(k, n)$  with the equivalent EVI ratio

$$\hat{r}_{EVI}(k, n) := \left| \frac{\hat{\xi}_{Y,k,n}}{\hat{\xi}_{X,k,n}} - 1 \right|. \quad (4.6)$$

#### 4.1. Simulation study

A small simulation study is performed in this section so that we understand the performance of our asymptotic approximations and tail similarity measure. Specifically, we apply Corollary 15 and calculate (4.5) and (4.6) for six pairs of distributions, most of which satisfy Assumption 3.3. That is, the second-order indexes ( $\rho$ ) are negative and unequal for all pairs, except of one pair—when comparing Distributions 1–4 listed below—where those indexes are

equal, a case where we do not know whether Corollary 15 applies. We believe that it is worth comparing that pair of distributions even though there is no theoretical support, though any extrapolation from comparing 1) and 4) should be taken with caution.

We simulate  $N = 5,000$  samples of size  $n = 2,500$  from the following distributions with survival function being  $2RV_{-\nu,\rho}$  such that  $\nu = 4$  (and in turn,  $\xi = 1/4$ ):

- 1) *Pareto* with  $\bar{F}(\cdot) := (1 + \cdot)^{-4}$  on  $\mathbb{R}_+$ ; that is,  $\rho = -1$ ;
- 2) *Paralogistic* with  $\bar{F}(\cdot) := (1 + \cdot^2)^{-2}$  on  $\mathbb{R}_+$ ; that is,  $\rho = -2$ ;
- 3) *Fréchet* with  $F(\cdot) := \exp\{-\cdot^{-4}\}$  on  $\mathbb{R}_+$ ; that is,  $\rho = -4$ ;
- 4) *Inverse-gamma* with  $f(\cdot) := f_{IG}(\cdot; 1)$  on  $\mathbb{R}_+$ ; that is,  $\rho = -1$ .

These four distributions are chosen to have the same first-order index ( $\nu$ ), and all pairs are compared in Figure 1. It would be good to know how the average errors for  $\hat{r}_{H;C}(k, n)$  and  $\hat{r}_{EVI}(k, n)$ , given by (4.5) and (4.6) behave when the first-order indexes are different, which is precisely what we plot in Figure 2. Therefore, two additional sampling distributions are considered as follows:

- 5) *Pareto* with  $\bar{F}(\cdot) := (1 + \cdot)^{-3}$  on  $\mathbb{R}_+$ ; that is,  $\nu = 3$  (hence  $\xi = 1/3$ ), and  $\rho = -1$ ;
- 6) *LogNormal* with  $\mu = 0$  and  $\sigma = 1$ ; that is,  $\nu = \infty$  (and hence  $\xi = 0$ );

Further details about the *Data Generation Process (DGP)* are provided in Appendix C.

We compute the average errors for  $\hat{r}_{H;C}(k, n)$  and  $\hat{r}_{EVI}(k, n)$ , given by (4.5) and (4.6), respectively, and provide the ratios based on the  $\chi^2$  and Hellinger divergences in Figures 1 and 2. Recall that the average errors are computed via the Hill estimator—see (4.2)—and we plot the ratios for various  $125 \leq k \leq 750$ , i.e., by considering 5% to 30% upper order statistics.

Among the twelve comparisons we run between Distributions 1–6, all but one produce EVI estimators that satisfy the integrability condition  $\hat{\xi}_{X,k,n_X}/\hat{\xi}_{Y,k,n_Y} > 1/2$ , with the comparison between Distributions 3 and 4 being the only exception. For the latter case, 94% of the  $N$  sample lead to EVI estimators that satisfy the integrability condition; the 6% samples that violate the condition are discarded, which presumably has a negligible effect on the comparison. For comparisons between Distributions 1–6, it is not surprising that the condition is mostly satisfied, because the distributions have the same or close EVIs. What is somewhat surprising is that the integrability condition is satisfied for all samples generated for comparisons between Distribution 6 — which has an EVI of 0 — and Distributions 2–4 (shown in the bottom panel of Figure 2). This emphasizes the limitation of using EVI only for measuring tail similarity.

The overall conclusion in Figures 1 and 2 is that (4.5) provides more accurate estimates than (4.6) except of the Fréchet vs Inverse-Gamma pair in Figure 1, where the differences between  $\hat{r}_{H;C}(k, n)$  and  $\hat{r}_{EVI}(k, n)$  are extremely small. Therefore, we could clearly argue that our proposed tail similarity in Corollary 15 would be preferred instead of using the vanilla comparison between the tail indexes that are measured via the Hill estimator in this paper.

Note that  $N = 5,000$  samples are obtained in Figures 1 and 2 by generating  $M = 50$  samples of size  $n = 2,500$ , and for each of the  $M$  samples, we bootstrap with replacement  $m = 100$  samples of the same size. We choose this bootstrapping procedure since this procedure is closer to the blockwise bootstrapping needed for the temporal dependent data discussed in the real data analysis provided in Section 4.2.

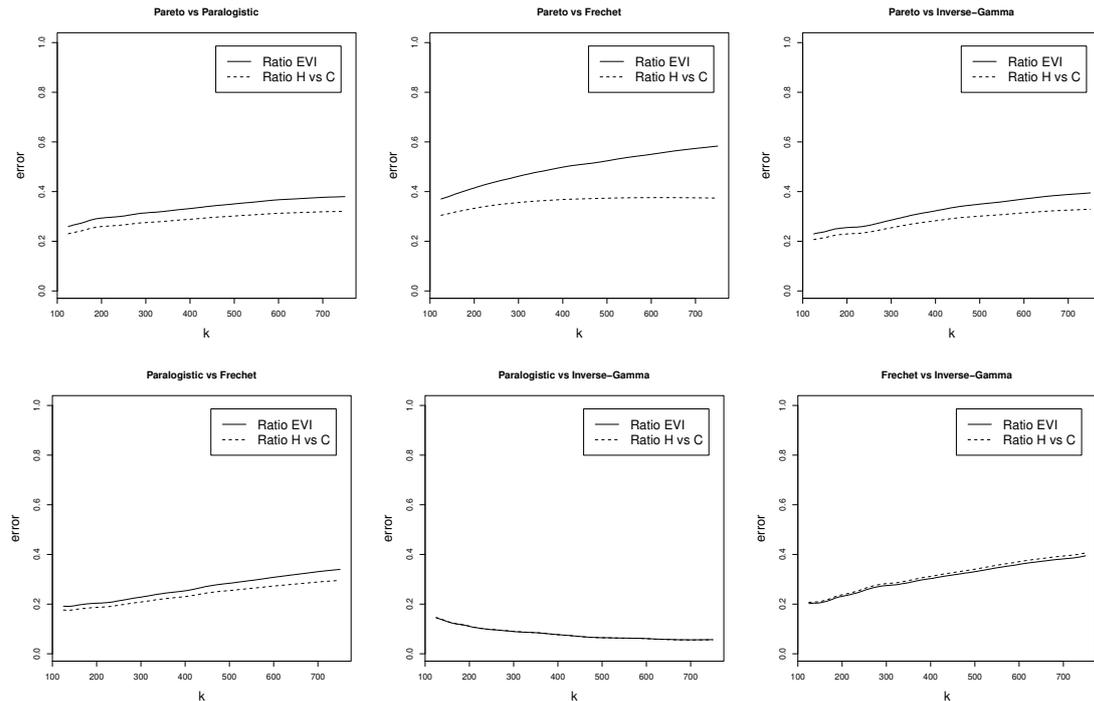


Figure 1: Averaged  $\hat{r}_{H;C}(k, n)$  and  $\hat{r}_{EVI}(k, n)$  ratios across  $N = 5,000$  samples of size  $n = 2,500$ , and relying on the estimators shown in (4.5) and (4.6), respectively. Comparisons are made for pairs with the same first-order index ( $\nu$ ), namely, all possible pairs of Distributions 1)–4) that are described before.

We notice in Figure 1 that the second-order index ( $\rho$ ) has an impact on detecting the degree of tail dissimilarity, and our estimator in (4.5) is more sensitive—than the Hill estimator in (4.6)—in capturing fine differences between second-order indexes. The next question is how the comparisons in Figure 1 would change if the first-order index of the distributions compared in each pair would be different; Figure 2 answers this question. This like-for-like comparison between Figures 2 and 1 show that ratios in (4.5) and (4.6) increase when the first-order indexes are different, which is very much expected. In addition, changing from a heavy tailed distribution (Distribution 5) in top plots of Figure 2) to a moderately heavy tailed distribution (Distribution 6) in bottom plots of Figure 2) the degree of tail dissimilarity is tamed. We have not included light tailed distributions (such as Exponential) or very light tailed distributions (such as some Weibull parameterisations) since these are very unrealistic comparisons, and thus, we compare either heavy tailed distributions or a heavy tailed distribution with a moderately heavy tailed distribution. Note that the concept of moderately heavy tailed distributions is in line with the description in Embrechts et al. (1997).

In a nutshell, the simulation study tells us that small differences between the estimators in (4.5) and (4.6) for the same pair of distributions would imply that the two distributions are either having i) the same first-order index or ii) one distribution has a heavy tail and the other one has a moderately heavy tailed distribution. These guiding principles are not expected to be universally true, and should be viewed with caution. A lengthy simulation study would clarify this point, but this would be beyond the scope of this paper. Therefore, we strongly recommend practitioners to use domain knowledge to validate the choice between i) and ii), which we do in Section 4.2.

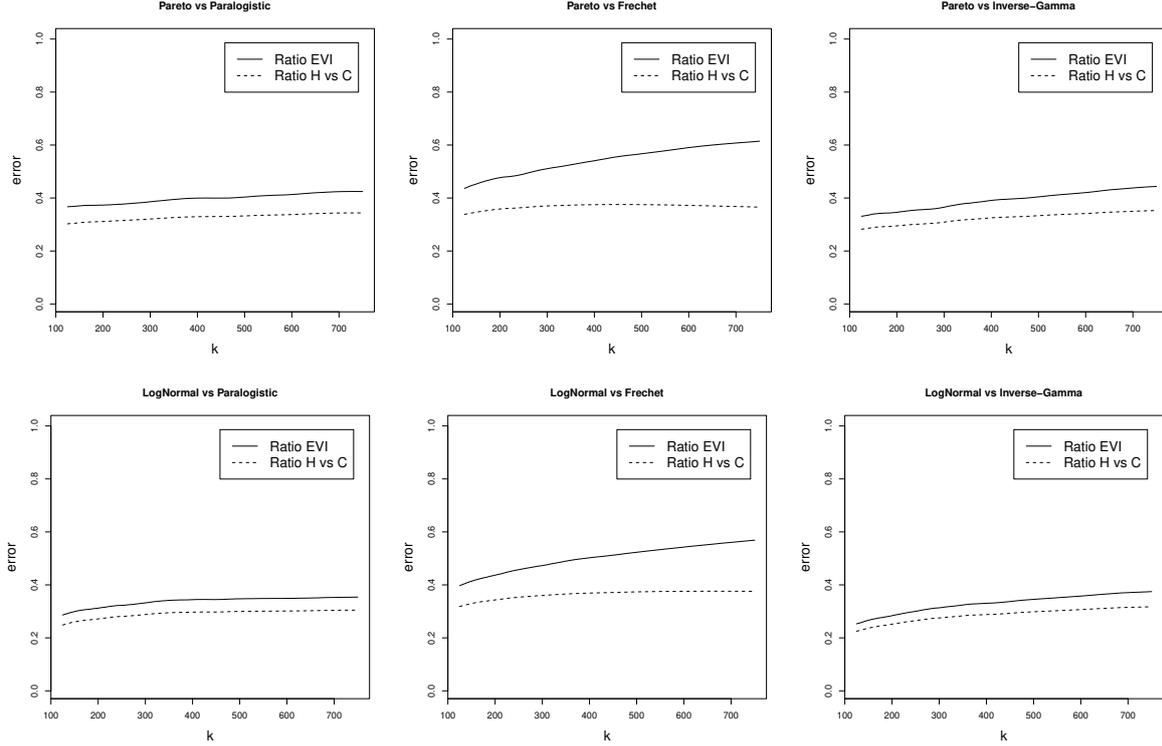


Figure 2: Averaged  $\hat{r}_{H;C}(k, n)$  and  $\hat{r}_{EVI}(k, n)$  ratios across  $N = 5,000$  samples of size  $n = 2,500$ , and relying on the estimators shown in (4.5) and (4.6), respectively. The top plots compare Distribution 5) with each of the Distributions 2)–4), while bottom plots compare Distribution 6) with each of the Distributions 2)–4).

#### 4.2. Real data analysis

Our real-life analysis from this section supplements the previous simulation study in Section 4.1, and we are interested in comparing different investment strategies (or portfolios), which usually result in differing first-order tail behaviour. This section relies on a very robust (to the market risk changes) investment strategy known as *Risk Parity (RP)* portfolios, which was introduced in Qian (2005) and further explored in the empirical finance literature (Asimit et al., 2024b; Roncalli, 2013; Roncalli and Weisang, 2016). A detailed description of such investment strategies can be found in Roncalli (2013), while Asimit et al. (2024b) and Asimit et al. (2024a) provide a more theoretical approach of such strategies that could be extended to risk sharing problems. For completeness, a brief background on the four RP portfolios discussed in this section is provided in Appendix B.

Hallerbach et al. (2004) developed a framework for constructing an investment portfolio with social responsible elements that is measured by SRI scores. The SRI scores of these companies were constructed from the questionnaires gathered by the SiRi research group in Year 2000.<sup>3</sup> The raw dataset in their study contains 590 firms from Europe, US and other countries in the Rest of the World.<sup>4</sup> We select 100 US firms in this dataset and collect their daily stock returns from Year 2001 to 2010 (i.e., a ten-year period covers the 2008 global financial/banking crisis).

<sup>3</sup>SiRi is a cooperation of 12 European Social research companies that developed an identical research questionnaire for analysing companies. A full description of the questionnaires is provided in Hallerbach et al. (2004).

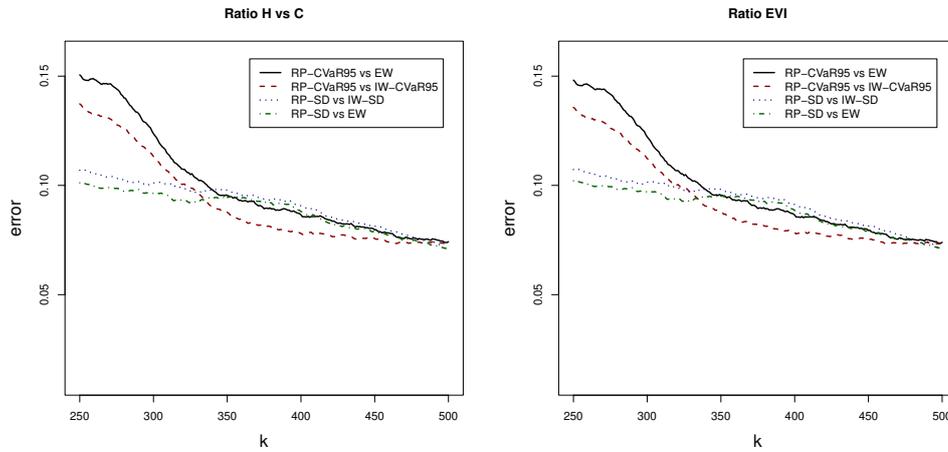
<sup>4</sup>We would like to thank Aloy Soppe for making the original raw dataset available to us.

Then, based on the 100 US firms' daily stock returns of size  $n = 2,500$  (i.e. 250 stock trading days per year during these 10 years), we create the following five portfolios:

- EW: equal weighted portfolio;
- RP-SD: RP portfolio constructed on the standard deviation (SD) risk measure;
- IW-SD: inverse weighted portfolio constructed on the SD risk measure;
- RP-CVaR<sub>95%</sub>: RP portfolio constructed on the Conditional Value at Risk (CVaR) risk measure at 95% level;
- IW-CVaR<sub>95%</sub>: inverse weighted portfolio constructed on the CVaR risk measure at 95% level.

For detailed descriptions of the portfolios, see [Appendix B](#).

The computations in [Figure 3](#) are made by bootstrapping  $N = 1,000$  samples from our observational data and for various  $250 \leq k \leq 500$ , i.e., by considering 10% to 20% upper order statistics. The averaged  $\hat{r}_{H;C}(k, n)$  and  $\hat{r}_{EVI}(k, n)$  ratios are plotted in [Figure 3](#) in a similar way to those provided in the simulation study in [Figures 1](#) and [2](#). As anticipated in [Section 4.1](#), we require in [Section 4.2](#) to use a bespoke resampling technique that is more suitable for time series data. Therefore, we use the standard blockwise bootstrapping in [Figure 3](#).



[Figure 3](#): Averaged  $\hat{r}_{H;C}(k, n)$  (left-hand-side plot computed via [\(4.5\)](#)) and  $\hat{r}_{EVI}(k, n)$  (right-hand-side computed via [\(4.6\)](#)) across  $N = 1,000$  bootstrapped samples of size  $n = 2,500$  based on the real-life daily stock returns of the 100 selected US companies.

We observe that the four comparisons in [Figure 3](#) are quite different though the averaged  $\hat{r}_{H;C}(k, n)$  and  $\hat{r}_{EVI}(k, n)$  for each of the four pairs are extremely negligible. These plots are not provided in this paper so as to avoid redundant information, but they are available upon request. Therefore, we choose this concise pictorial representation in [Figure 3](#). This scenario is discussed at the end of [Section 4.1](#), and we may conclude the four pairs of portfolios are either very tail similar or one portfolio distribution has a heavy tail and the other one has a moderately heavy tailed distribution.

First, note that the observation period (01/01/2001 to 31/12/2010) contains periods of bull and bear markets with very different behaviour, and a good measure of tail dissimilarity should exhibit an oscillatory behaviour when comparing these portfolios. Therefore, good practice reasoning would refute the possibility of having tail similarity for the two pairs with a flat behaviour, i.e., RP-SD vs IW-SD and RP-SD vs EW. Further, the same good practice reasoning would suggest only RP-CVaR<sub>95%</sub> vs EW or RP-CVaR<sub>95%</sub> vs IW-CVaR<sub>95%</sub> may exhibit some degree of tail similarity based on what we observe in Figure 3.

Second, we provide some domain knowledge to validate and clarify the claims from the previous paragraph. The tail discrepancy between the RP-SD vs EW portfolios is confirmed by the right-panel plot of Figure B.5, where the EW portfolio is expected to show a more pronounced tail risk exposure; a theoretical foundation of why any RP portfolios are less risky than EW portfolios, and in turn, tail discrepancy is highly expected can be found in Theorem 1 c) of Asimit et al. (2024b). The left-panel plot of Figure B.5 may allude tail similarity between the RP-SD vs IW-SD portfolios due to similar risk positions of the two portfolios, which would not be true; in fact, more empirical evidence in that sense is available in Asimit et al. (2024b). The oscillatory shape in Figure 3 is just another source of soft validation of such tail dissimilarity. These two comparisons (RP-SD vs IW-SD and RP-SD vs EW) are also influenced by the risk measure choice, namely SD, which is not tail sensitive by design, and thus, RP-SD and IW-SD portfolios are unlikely to be tail similar; recall that RP portfolios only allocate the overall risk equally amongst the individual assets, but the overall perception of risk is measured by a non-tail risk measure.

Third, we explain the last two comparisons, namely RP-CVaR<sub>95%</sub> vs IW-CVaR<sub>95%</sub> and RP-CVaR<sub>95%</sub> vs EW for which our good practice reasoning explained on our first point. Theorem 1 c) of Asimit et al. (2024b) suggests that RP-CVaR<sub>p</sub> and EW should be tail dissimilar for any  $0 \leq p < 1$ , and thus, such oscillatory behavior in Figure 3 should not imply tail similarity. Portfolios RP-CVaR<sub>95%</sub> vs IW-CVaR<sub>95%</sub> show a non-oscillatory behavior (irrespective of the chosen ratio estimator, either by (4.5) or (4.6)) after a certain threshold—namely, when  $k/n \geq 325/2,500 = 13\%$  – and thus, tail similarity is inferred in this case.

As a conclusion, we found that our proposed ratio estimator in (4.5) is more sensible than the standard Hill ratio estimator in (4.6), but both provide meaningful information that could complement each other. We recommend using both ratio estimators for a more informative comparisons by keeping in mind the guiding principles of good practice provided in the last paragraph of Section 4.1. Adding domain knowledge is highly recommended to validate and clarify the conclusions of our guiding principles, especially in situations with no clear-cut conclusion.

## 5. Conclusions

This paper introduces novel measures of tail similarity designed to comprehensively assess the overarching tail behaviour between two probability distributions. The theoretical foundation of these measures is broad, offering a versatile framework for quantifying tail similarity. Our approach supplements the traditional practice of comparing the (first-order) extreme index estimates in finite samples by incorporating additional sources of validation.

Empirical evidence presented in this study showcases the core focus of our research, in particular providing a wide range of new EVI dissimilarity estimators, and paves the way for novel lines of research. Additionally, our analysis of real-life data illustrates the conclusions drawn from our simulation study, and shows how the proposed measures of tail similarity may be used in practice. The conclusion is that our tail similarity measure and the very basic comparisons between the (first-order) extreme index estimates provide complementary information, and one should analyse them in tandem rather than in isolation. We provide some very simple guiding principles of good practice when using the two sources of information, which are recommended to be used with domain knowledge to validate and clarify the conclusions of our guiding principles especially in situations with no clear-cut conclusion as it is often the case in real-life applications.

### **Acknowledgements**

We would like to thank the participants at the *Actuarial Science and Quantitative Risk Management Seminar* at Purdue University and those at the 26<sup>th</sup> *International Congress on Insurance: Mathematics and Economics* for their helpful comments.

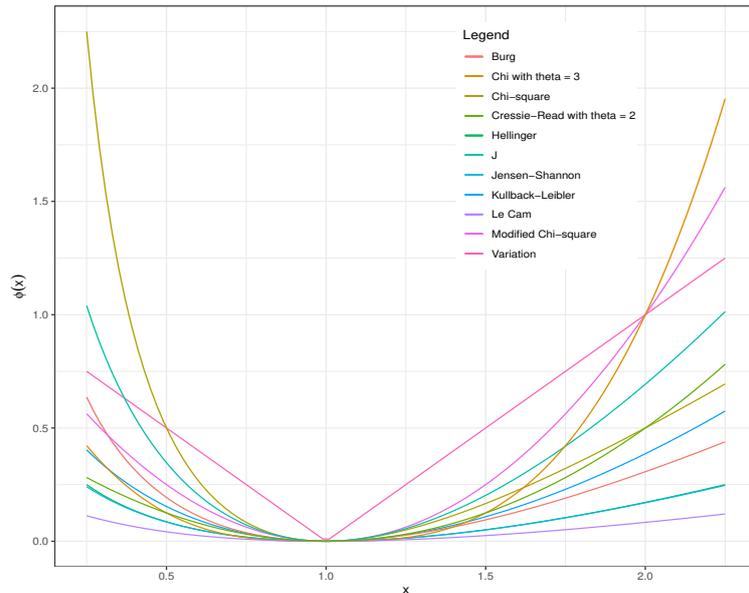
## Appendix A. Phi-divergences Examples

A summary of commonly used  $\phi$ -divergences is provided in Table A.1, and a pictorial representing of these functions is given in Figure A.4. We see that the variation divergence penalises more than any other  $\phi$ -divergence at small dissimilarities around 1, which makes the variation divergence a very conservative  $\phi$ -divergence; we also notice in Figure A.4 that  $\chi$ -divergence with  $\theta = 3$  and Modified  $\chi^2$ -divergence penalise more than any other  $\phi$ -divergence at large dissimilarities. Therefore, we expect goodness-of-fit tests to behave differently for different  $\phi$ -divergences.

Table A.1: Commonly used  $\phi$ -divergences

$\phi$ function	Statistical distance
$\phi_{KL}(x) = x \log x - x + 1$	Kullback-Leibler divergence
$\phi_B(x) = -\log x + x - 1$	Burg divergence
$\phi_J(x) = (x - 1) \log x$	$J$ -divergence
$\phi_{JS}(x) = -(x + 1) \log((1 + x)/2) + x \log x$	Jensen-Shannon divergence
$\phi_C(x) = (x - 1)^2 / x$	$\chi^2$ -divergence
$\phi_{MC}(x) = (x - 1)^2$	Modified $\chi^2$ -divergence
$\phi_H(x) = (\sqrt{x} - 1)^2$	Hellinger divergence
$\phi_V(x) =  x - 1 $	Variation divergence
$\phi_\theta(x) =  x - 1 ^\theta$ with $\theta > 1$	$\chi$ -divergence
$\phi_{LC}(x) = \frac{(x-1)^2}{4(x+1)}$	Le Cam divergence
$\phi_{CR}(x) = \frac{1-\theta+x\theta}{\theta(1-\theta)}$ with $\theta \in \mathbb{R} \setminus \{0, 1\}$	Cressie-Read divergence

Figure A.4: Plot of commonly used  $\phi$  functions.



## Appendix B. Risk Parity

A short description of RP strategies is now provided. The decision-maker aims to invest in a given opportunity set of  $d$  assets with losses (i.e. negative returns) of  $\mathbf{X} = (X_1, \dots, X_d)$ . The investment strategy is given by a vector of proportions  $\boldsymbol{\alpha} \in \Delta_d$ , where  $\Delta_d$  is the unit  $d$ -simplex  $\Delta_d := \{\mathbf{x} \in \mathbb{R}_+^d : \mathbf{1}^T \mathbf{x} = 1\}$  for any positive integer  $d$ ; recall that  $\mathbb{R}_+^d := \{\mathbf{x} \in \mathbb{R}^d : \mathbf{x} \geq 0\}$  is the standard polyhedral cone of the positive quadrant of  $\mathbb{R}^d$ . We also use the notation  $\mathbb{R}_{++}^d := \{\mathbf{x} \in \mathbb{R}^d : \mathbf{x} > 0\}$ . This setting leads to a portfolio loss of  $\boldsymbol{\alpha}^T \mathbf{X}$ .

The financial assets are defined on  $(\Omega, \mathcal{F}, \mathbb{P})$ , the same atomless probability space defined earlier. We assume that the risk preferences of an investor are represented by the risk measure  $\varphi$ , which is a function that maps an element of  $\mathcal{X}$  to the real set, i.e.  $\varphi : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$ . Therefore, the investor's perception of risk is given by  $\mathcal{R}(\boldsymbol{\alpha}) := \varphi(\boldsymbol{\alpha}^T \mathbf{X})$ .

For a risk measure  $\varphi$  that is homogeneous of order  $\tau > 0$  (i.e.,  $\varphi(cY) = c^\tau \varphi(Y)$  for any  $Y \in \mathcal{X}$  and  $c \geq 0$ ), the Euler's Homogeneous Function Theorem implies that

$$\mathcal{R}(\boldsymbol{\alpha}) = \frac{1}{\tau} \sum_{k=1}^d \alpha_k \frac{\partial \mathcal{R}(\boldsymbol{\alpha})}{\partial \alpha_k} = \sum_{k=1}^d \mathcal{RC}_k(\boldsymbol{\alpha}), \quad \text{where} \quad \mathcal{RC}_k(\boldsymbol{\alpha}) := \frac{\alpha_k}{\tau} \frac{\partial \varphi(\boldsymbol{\alpha}^T \mathbf{X})}{\partial \alpha_k}. \quad (\text{B.1})$$

By definition,  $\mathcal{RC}_k(\boldsymbol{\alpha})$  is the *risk contribution* of the  $k^{\text{th}}$  individual risk. An investment strategy  $\boldsymbol{\alpha} \in \Delta_d \cap \mathbb{R}_{++}^d$  is a *RP* portfolio if  $\mathcal{RC}_k(\boldsymbol{\alpha}) = \frac{1}{d} \varphi(\boldsymbol{\alpha}^T \mathbf{X})$  for all  $k \in \{1, 2, \dots, d\}$ , where  $\mathcal{RC}_k(\boldsymbol{\alpha})$  is given in (B.1). When  $\varphi = SD$  (or  $\varphi = \text{CVaR}_{95\%}$ ), this RP strategy with risk preferences ordered by the standard deviation and is denoted as  $\boldsymbol{\alpha}^{*SD}$  (or  $\boldsymbol{\alpha}^{*CVaR_{95\%}}$ ), is an RP strategy with  $\varphi = SD$  (or  $\varphi = \text{CVaR}_{95\%}$ ). Recall that  $SD(\cdot) := \sqrt{\mathbb{E}(\cdot^2) - (\mathbb{E}(\cdot))^2}$  and  $\text{CVaR}_p(\cdot) = \min_{t \in \mathbb{R}} \{t + \frac{1}{1-p} \mathbb{E} \max(\cdot - t, 0)\}$  on  $L^0$  for any  $0 \leq p < 1$ .

We next provide some numerical evidence that helps us to understand the performance of RP portfolios. An RP portfolio constructed on the standard deviation (SD) risk measure (denoted as RP-SD) does not distinguish between independent and comonotonic risks; e.g., see [Roncalli \(2013\)](#). Therefore, the risk management literature tends to compare the performance of RP portfolios with the so-called *Inverse Weighted (IW)* portfolio, which was originally coined by the Bridgewater asset management firm in the 1990s. Specifically, a portfolio is said to be *in parity* when weights are inverse proportional to the asset-class risk position, and therefore, such IW portfolios are an earlier version of RP portfolios, but different than the formal RP portfolios defined here. The mathematical formulation of IW portfolios with SD (resp., CVaR) risk preferences is given in (B.2) (resp., (B.3)) is denoted as *IW-SD* (resp., *IW-CVaR*):

$$\boldsymbol{\alpha}^{**SD} = \left( \frac{1/SD(X_1)}{\sum_{l=1}^d 1/SD(X_l)}, \dots, \frac{1/SD(X_d)}{\sum_{l=1}^d 1/SD(X_l)} \right)^T \quad (\text{B.2})$$

and

$$\boldsymbol{\alpha}^{**CVaR} = \left( \frac{1/\text{CVaR}_{95\%}(X_1)}{\sum_{l=1}^d 1/\text{CVaR}_{95\%}(X_l)}, \dots, \frac{1/\text{CVaR}_{95\%}(X_d)}{\sum_{l=1}^d 1/\text{CVaR}_{95\%}(X_l)} \right). \quad (\text{B.3})$$

where  $X_l$  is asset daily losses (i.e. negative returns) of an individual asset  $l$  in a portfolio of  $d$  assets. Further details about IW portfolios and how to compare with RP portfolios are available in [Clarke et al. \(2013\)](#), while a more detailed analysis is provided in [Asimit et al. \(2024b\)](#).

Beside IW portfolios, *Equal Weighted (EW)* portfolios (when  $\alpha = \frac{1}{d}\mathbf{1}$ ) are another well-known benchmark portfolios that are introduced in the seminal paper of DeMiguel et al. (2009) for which the same weight is assumed to every asset. Such portfolios have been praised for their blindness to historical trends and have been proved to be insensitive to market changes, a property shared with IW and RP portfolios. There is an extensive literature that compares the performance of EW vs RP portfolios; e.g., see (Asimit et al., 2024b; Fisher et al., 2015; Lassance et al., 2022).

We illustrate this section with a small simulation study (as shown in Figures B.5 and B.6) in order to familiarise the reader with the behaviour of various investment strategies that we provide in Figures 3. Clearly, EW vs IW portfolios are easily computable, which is not the case for RP portfolios. Bespoke algorithms for RP-SD and RP-CVaR portfolios are available in (Asimit et al., 2024a; Bai et al., 2016; Spinu, 2013) and (Asimit et al., 2024b; Mausser and Romanko, 2018), respectively.

Our first point is about the anecdotal claim alluding that the RP estimation is less sensitive with respect to the underlying dependence among risks and thus, the covariance matrix estimation error may be less detrimental when constructing RP portfolios than any other investment strategies base on risk optimisation such as mean-variance; see Section 3.2.3.4 in Roncalli (2013) or Maillard et al. (2010). More empirical evidence is provided in Figure B.5, where we generate  $N = 1,000$  random (positive definite) covariance matrices  $\Sigma_d^{(k)}, k \in \{1, \dots, N\}$  for various  $d$ . Then, we compute the ratios between the risk position of RP-SD portfolio and its corresponding IW and EW competitive portfolios. That is, we compute

$$\sqrt{\frac{(\alpha^{*SD;(k)})^T \Sigma_d^{(k)} \alpha^{*SD;(k)}}{(\alpha^{**SD;(k)})^T \Sigma_d^{(k)} \alpha^{**SD;(k)}}} \quad \text{and} \quad \sqrt{\frac{(\alpha^{*SD;(k)})^T \Sigma_d^{(k)} \alpha^{*SD;(k)}}{\frac{1}{d^2} (\mathbf{1}^T \Sigma_d^{(k)} \mathbf{1})}},$$

where  $\alpha^{*SD;(k)}$  is the RP-SD portfolio risk proportions and  $\alpha^{**SD;(k)}$  is the IW-SD portfolio risk proportions (as in (B.2)) for the  $k^{th}$  replicate.

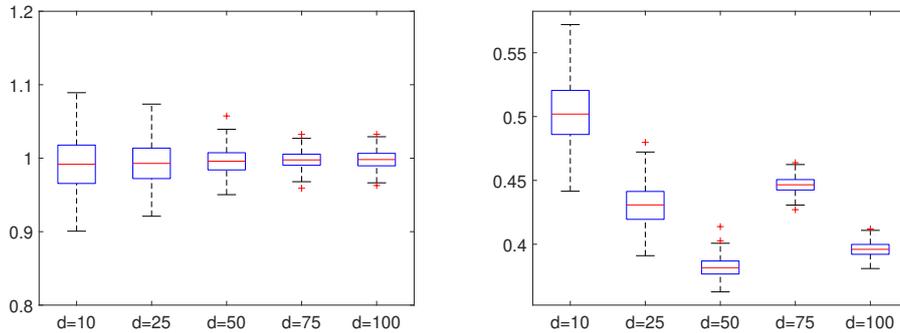


Figure B.5: Boxplots of ratios between the standard deviation of RP-SD and IW-SD portfolios (based on various  $d$  assets) are shown in the left-panel plot, while the corresponding ratios between RP-SD and EW portfolios are displayed in the right plot; computations are based on  $N = 1,000$  random (positive definite) covariance matrices.

In the right panel of Figure B.5, RP portfolios are shown to be less risky than their corresponding EW portfolio, since the ratios between RP-SD and EW are shown to be less than 1,

which confirms the result in Theorem 1 c) of [Asimit et al. \(2024b\)](#). The left panel of Figure [B.5](#) shows that the risk of RP-SD and IW-SD portfolios are not significantly different, since the ratios are close to 1. The RP literature discusses such pattern when the risk is measured via SD, and one may believe that this is true for risk preferences measured by other risk measures. We show that such behaviour is different in Figure [B.6](#), where tail risk measures (specifically, CVaR) is considered.

We redo the previous computations in Figure [B.6](#) when the risk preferences are ordered by  $\text{CVaR}_{95\%}$  instead of SD, while  $\text{IW-CVaR}_{95\%}$  is the equivalent portfolio in [\(B.3\)](#). Note that we only generate random covariance matrices for computations shown in Figure [B.5](#), since variance based computations (for RP-SD and IW-SD) require only information about the sample covariance matrix, and thus, we have not simulated the  $N = 1,000$  asset loss/return samples in the SD examples. This is not the case for  $\text{CVaR}_{95\%}$  computations, where the sample losses (i.e.  $\mathbf{X}$ ) are needed. That is, we compute

$$\frac{\text{CVaR}_{95\%} \left( \left( \boldsymbol{\alpha}^{*\text{CVaR};(k)} \right)^T \mathbf{X}^{(k)} \right)}{\text{CVaR}_{95\%} \left( \left( \boldsymbol{\alpha}^{**\text{CVaR};(k)} \right)^T \mathbf{X}^{(k)} \right)} \quad \text{and} \quad \frac{\text{CVaR}_{95\%} \left( \left( \boldsymbol{\alpha}^{*\text{CVaR};(k)} \right)^T \mathbf{X}^{(k)} \right)}{\text{CVaR}_{95\%} \left( \frac{1}{d} \mathbf{1}^T \mathbf{X}^{(k)} \right)},$$

where  $\boldsymbol{\alpha}^{*\text{CVaR};(k)}$  and  $\boldsymbol{\alpha}^{**\text{CVaR};(k)}$  are the RP- $\text{CVaR}_{95\%}$  and IW- $\text{CVaR}_{95\%}$  portfolio risk proportions for the  $k^{\text{th}}$  replicate, respectively; note that  $\boldsymbol{\alpha}^{**\text{CVaR};(k)}$  are computed as in [\(B.3\)](#).

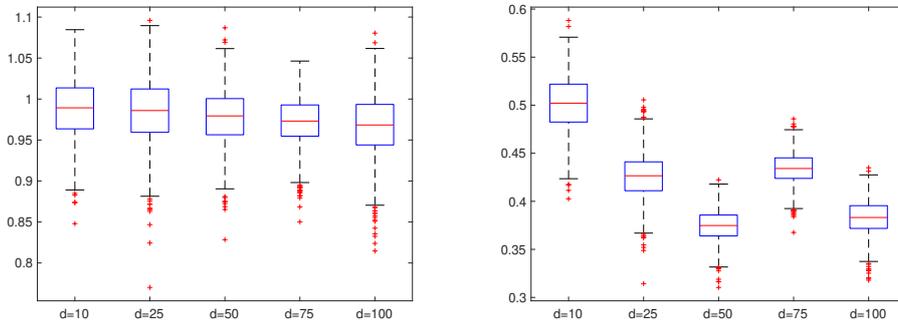


Figure B.6: Boxplots of ratios between the  $\text{CVaR}_{95\%}$  of RP- $\text{CVaR}_{95\%}$  and IW- $\text{CVaR}_{95\%}$  portfolios (based on various  $d$  assets) are shown in the left-panel plot, while the corresponding ratios between RP- $\text{CVaR}_{95\%}$  and EW portfolios are displayed in the right plot; computations are based on  $N = 1,000$  samples of size  $n = 1,000$  from multivariate normal returns with zero mean vectors and random (positive definite) covariance matrices.

Figure [B.6](#) computations are made by generating  $N = 1,000$  samples of size  $n = 1,000$  multivariate normal distributed (for various  $d$ -asset portfolios) with zero mean vectors and random covariance matrices as in SD examples earlier. We note that the right-panel plots in Figures [B.5](#) and [B.6](#) show similar conclusions, which are in accordance with Theorem 1 c) of [Asimit et al. \(2024b\)](#). The left-panel plots are different and show that the risk position of RP- $\text{CVaR}_{95\%}$  portfolio is smaller than the IW- $\text{CVaR}_{95\%}$  portfolio (as in Figure [B.6](#)), which is different with our findings in Figure [B.5](#). This is a noteworthy empirical result and explains the importance of tail risk measures (e.g., CVaR) compared to moment-like risk measures (e.g., SD).

## Appendix C. Data Generation Process (DGP)

The DGP processes used in this paper are detailed in this section. First, we describe the specific parametric family used in Section 4.1. Note that  $f_{IG}(\cdot; \alpha) := \frac{\gamma^\alpha}{\Gamma(\alpha)} \cdot^{-\alpha-1} \exp\{-\gamma/\cdot\}$  on  $\mathbb{R}_+$  is the pdf of the Inverse-gamma distribution as defined in the *actuar* R package for any  $\gamma > 0$ ; in fact, all distributional class parameterisations considered in this section are exactly as in the *actuar* R package, except of the Fréchet that is simulated from first principles. The precise R commands are as follows:

- 1) *Pareto*: `rpareto(n=2500, shape=4, scale=1)`.
- 2) *Paralogistic*: `rparalogis(n=2500, shape=2, rate=1, scale=1)`.
- 3) *Fréchet*:  $(-\log U)^{-1/4}$ , where  $U \sim U[0, 1]$ .
- 4) *Inverse-gamma*: `rinvgamma(n=2500, shape=4, rate=1, scale=1)`.

Distribution 5) is the same as Distribution 1), but with `shape=3`, while Distribution 6) is fairly standard and no additional information would be needed.

Second, we briefly describe the steps of our DGP in Section 4.2. That is, the DGP in Figure B.5 is a repetition (for  $N = 1,000$  times) of the following steps:

- **Step 1**: The variance of each set of  $d$  assets is randomly generated from  $U[0, 10]$  so that we achieve a wide range of individual asset's variance when computing the IW-SD portfolio risk proportions.
- **Step 2**: Generate random positive definite correlation matrices, which are further transformed so that the individual asset's variances are as generated in **Step 1**.

Further, the DGP in Figure B.6 is a repetition (for  $N = 1,000$  times) of the following steps:

- **Steps 1 & 2**: Same as the DGP in Figures B.5.
- **Step 3**: Generate  $n = 1,000$  individual asset daily losses (i.e. negative returns) from a multivariate normal distribution with a zero mean vector and covariance matrix that is generated as in **Steps 1 & 2**, in order to perform the necessary sample based calculations for the IW-CVaR and RP-CVaR portfolios.

## Appendix D. Proofs

### Appendix D.1. Proof of Lemma 6

By Lemma 3 of [Hua and Joe \(2011\)](#),

$$h(t) = k_h t^{-\alpha} l_h(t)$$

for some  $k_h > 0$  and slowly varying function  $l_h(\cdot)$  such that  $\lim_{t \rightarrow \infty} l_h(t) = 1$ . Therefore, for fixed  $\varepsilon > 0$ , there exists a large  $t_h(\varepsilon)$ , such that, for all  $\min\{x_0 t, t\} > t_h$  and all  $x \geq x_0$ ,

$$1 - \varepsilon \leq \frac{l_h(xt)}{l_h(t)} \leq 1 + \varepsilon.$$

Hence, for  $t$  large such that  $\min\{x_0 t, t\} > t_h$ ,

$$\sup_{x \geq x_0} \left| \frac{h(xt)}{x^{-\alpha} h(t)} - 1 \right| = \sup_{x \geq x_0} \left| \frac{k_h(xt)^{-\alpha} l_h(xt)}{k_h x^{-\alpha} t^{-\alpha} l_h(t)} - 1 \right| \leq \varepsilon.$$

This completes the proof.

#### Appendix D.2. Proof of Lemma 7

Lemma 7a) is a result known in the literature; see, e.g., Remark 2.4 of [Mao et al. \(2012\)](#).

To prove Lemma 7b), write  $s = 1/\bar{H}(t)$ . We have  $\bar{H}^{\leftarrow}(\bar{H}(t)) = (1/\bar{H})^{\leftarrow}(1/\bar{H}(t)) = U_H(s)$  and, for any  $\epsilon(s) > 0$ ,

$$U_H(s) \leq t \leq U_H(s(1 + \epsilon(s))), \quad (\text{D.1})$$

where we applied Proposition A.3(v) of [McNeil et al. \(2015\)](#) to the right continuous function  $1/\bar{H}$  and applied equation (0.6c) of [Resnick \(2008\)](#). For arbitrarily fixed  $\varepsilon \in (0, 1)$ , letting  $\epsilon(s) = \varepsilon$ , we have from (D.1) and the regular variation of  $U_H$  that

$$1 \geq \frac{U_H(s)}{t} \geq \frac{U_H(s)}{U_H(s(1 + \varepsilon))} \rightarrow (1 + \varepsilon)^{-1/\nu}.$$

The first asymptotic relation in b) follows from the arbitrariness of  $\varepsilon$ .

Now we prove the second relation. Obviously,

$$\left| \frac{\bar{H}^{\leftarrow}(\bar{H}(t))}{t} - 1 \right| = \left| \frac{U_H(s)}{t} - 1 \right| \leq \left| \frac{U_H(s)}{U_H(s(1 + \epsilon(s)))} - 1 \right|. \quad (\text{D.2})$$

Note that  $U_H \in 2\text{RV}_{1/\nu, \rho/\nu}$  with an auxiliary function that satisfies

$$a(s) = \nu^{-2} A(U_H(s)) = \nu^{-2} A(\bar{H}^{\leftarrow}(\bar{H}(t))) \sim \nu^{-2} A(t). \quad (\text{D.3})$$

By Proposition 5, we have, for any fixed  $\varepsilon, \delta > 0$ , there exists  $s_0 = s_0(\varepsilon, \delta, \nu, \rho)$ , such that, for all  $t$  large with  $s > s_0$ , it holds that

$$\begin{aligned} & \left| \frac{\frac{U_H(s)}{U_H(s(1 + \epsilon(s)))} - (1 + \epsilon(s))^{-1/\nu}}{a(s)} - (1 + \epsilon(s))^{-1/\nu} \frac{(1 + \epsilon(s))^{-\rho/\nu} - 1}{\rho/\nu} \right| \\ & \leq \varepsilon (1 + \epsilon(s))^{-1/\nu} \left( \left| \frac{(1 + \epsilon(s))^{-\rho/\nu} - 1}{\rho/\nu} \right| + (1 + \epsilon(s))^{-\rho/\nu} (1 + \epsilon(s))^\delta \right), \end{aligned}$$

where, by convention,  $(\cdot)^\rho - 1 / \rho$  is interpreted as  $\log(\cdot)$ . With  $\epsilon(s) \downarrow 0$ , the right-hand side above can be made close enough to  $\varepsilon$ . Hence, for  $s$  sufficiently large,

$$\left| \frac{\frac{U_H(s)}{U_H(s(1 + \epsilon(s)))} - (1 + \epsilon(s))^{-1/\nu}}{a(s)} \right| \leq 2\varepsilon. \quad (\text{D.4})$$

By (D.2), for  $t$  large,

$$\begin{aligned}
\left| \frac{\frac{\bar{H}^\leftarrow(\bar{H}(t))}{t} - 1}{A(t)} \right| &\leq \left| \frac{\frac{U_H(s)}{U_H(s(1+\epsilon(s)))} - 1}{A(t)} \right| \\
&\leq \left| \frac{\frac{U_H(s)}{U_H(s(1+\epsilon(s)))} - (1 + \epsilon(s))^{-1/\nu}}{A(t)} \right| + \left| \frac{(1 + \epsilon(s))^{-1/\nu} - 1}{A(t)} \right| \\
&\leq \left| \frac{\frac{U_H(s)}{U_H(s(1+\epsilon(s)))} - (1 + \epsilon(s))^{-1/\nu}}{a(s)} \right| \left| \frac{a(s)}{A(t)} \right| + (1 + \epsilon) \left| \frac{\epsilon(s)}{\nu A(t)} \right| \\
&\leq (1 + \epsilon) \left( 2\epsilon\nu^{-2} + \left| \frac{\epsilon(s)}{\nu A(t)} \right| \right),
\end{aligned}$$

where in the last step we applied (D.3) and (D.4). Letting  $\epsilon(s) = o(|A(t)|)$ , we conclude that

$$\left| \frac{\frac{\bar{H}^\leftarrow(\bar{H}(t))}{t} - 1}{A(t)} \right| \rightarrow 0,$$

which completes the proof.

### Appendix D.3. Proof of Theorem 8

We only prove for the case with  $\rho_X > \rho_Y$  because the other case with  $\rho_X < \rho_Y$  can be proved in a similar way. In this case,  $\rho_{XY} = \rho_X$  and  $A_{XY} = A_X$ .

Write  $U_F(\cdot) = F^\leftarrow(1 - 1/\cdot)$  and  $U_G(\cdot) = G^\leftarrow(1 - 1/\cdot)$ . We have that, for  $q \in (0, 1)$ ,

$$\bar{F}_t^\leftarrow(q) = \bar{F}^\leftarrow(\bar{F}(t)q) = U_F\left(\frac{1}{\bar{F}(t)q}\right) \quad \text{and} \quad \bar{G}_t^\leftarrow(q) = \bar{G}^\leftarrow(\bar{G}(t)q) = U_G\left(\frac{1}{\bar{G}(t)q}\right).$$

Consider the ratio of

$$\begin{aligned}
\frac{d_{W,p}(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})^p}{t^p |A_X(t)|^p} &= \int_0^1 \left| \frac{\bar{F}_t^\leftarrow(q) - \bar{G}_t^\leftarrow(q)}{t |A_X(t)|} \right|^p dq \\
&= \int_0^1 \left| \frac{\bar{F}_t^\leftarrow(q) - \bar{G}_t^\leftarrow(q)}{t A_X(t)} \right|^p dq \\
&= \int_0^1 \left| \frac{\bar{F}_t^\leftarrow(q) - tq^{-1/\nu}}{t A_X(t)} - \frac{\bar{G}_t^\leftarrow(q) - tq^{-1/\nu}}{t A_Y(t)} \frac{A_Y(t)}{A_X(t)} \right|^p dq. \quad (\text{D.5})
\end{aligned}$$

Note that, by the  $C_r$  inequality, there exists some constant  $C_p > 0$ , such that the integrand above is bounded by

$$C_p \left| \frac{\bar{F}_t^\leftarrow(q) - tq^{-1/\nu}}{t A_X(t)} \right|^p + C_p \left| \frac{\bar{G}_t^\leftarrow(q) - tq^{-1/\nu}}{t A_Y(t)} \frac{A_Y(t)}{A_X(t)} \right|^p. \quad (\text{D.6})$$

Further note that, with  $s := 1/\bar{F}(t)$  and  $z = 1/q$ , there exists some constant  $C'_p > 0$ , such that

$$\begin{aligned} \left| \frac{\bar{F}_t^{\leftarrow}(q) - tq^{-1/\nu}}{tA_X(t)} \right|^p &= \left| \frac{\frac{U_F(sz)}{U_F(s)} \frac{U_F(s)}{t} - z^{1/\nu}}{A_X(t)} \right|^p \\ &= \left| \frac{\frac{U_F(s)}{t} \left( \frac{U_F(sz)}{U_F(s)} - z^{1/\nu} \right) + \left( \frac{U_F(s)}{t} - 1 \right) z^{1/\nu}}{A_X(t)} \right|^p \\ &\leq C'_p \left( \left| \frac{\frac{U_F(s)}{t} \left( \frac{U_F(sz)}{U_F(s)} - z^{1/\nu} \right)}{A_X(t)} \right|^p + \left| \frac{U_F(s) - 1}{A_X(t)} z^{1/\nu} \right|^p \right). \end{aligned}$$

Fix  $\varepsilon, \delta > 0$  such that  $p/\nu + \delta p < 1$ . By Lemma 7b), for  $t$  large enough, the second term in the parenthesis above is bounded by  $\varepsilon z^{p/\nu} = \varepsilon q^{-p/\nu}$ , which is integrable over  $q \in (0, 1]$  because  $\nu > p$ . Moreover, write  $a_X(\cdot) = \nu^{-2} A_X(U_F(\cdot))$ . By Lemma 7a) and Proposition 5, there exists  $s_0 = s_0(\varepsilon, \delta, \nu, \rho_X)$ , such that, for all  $t$  large with  $s > s_0$ , it holds that

$$\begin{aligned} \left| \frac{\frac{U_F(s)}{t} \left( \frac{U_F(sz)}{U_F(s)} - z^{1/\nu} \right)}{A_X(t)} \right| &\leq (1 + \varepsilon) \left| \frac{\frac{U_F(sz)}{U_F(s)} - z^{1/\nu}}{a_X(s)} \right| \left| \frac{a_X(s)}{A_X(t)} \right| \\ &\leq (1 + \varepsilon)^2 \nu^{-2} \left| \frac{\frac{U_F(sz)}{U_F(s)} - z^{1/\nu}}{a_X(s)} \right| \\ &\leq (1 + \varepsilon)^2 \nu^{-2} \left( (1 + \varepsilon) \left| z^{1/\nu} \frac{z^{\rho_X/\nu} - 1}{\rho_X/\nu} \right| + \varepsilon z^{1/\nu} z^\delta \right), \end{aligned}$$

where in the second step we used the fact that  $U_F(s) \sim t$  and in the third step we used (D.3). Therefore, using  $C_r$  inequality again, we have

$$\left| \frac{\frac{U_F(s)}{t} \left( \frac{U_F(sz)}{U_F(s)} - z^{1/\nu} \right)}{A_X(t)} \right|^p \leq (1 + \varepsilon)^{2p} \nu^{-2p} C_p'' \left( \frac{(1 + \varepsilon)^p}{(\rho_X/\nu)^p} \left| z^{1/\nu} (z^{\rho_X/\nu} - 1) \right|^p + \varepsilon z^{p/\nu + \delta p} \right)$$

holds for some constant  $C_p'' > 0$ . Since

$$\frac{(1 + \rho_X)p}{\nu} < 1, \quad \frac{p}{\nu} < 1, \quad \text{and} \quad \frac{p}{\nu} + \delta p < 1,$$

the right-hand side above as a function of  $q = 1/z$  is integrable over  $q \in (0, 1]$ . Similar arguments can be made for the second term in (D.6) by further using the fact that  $A_Y(t)/A_X(t) \rightarrow 0$ . Hence, by (D.5) and the Dominated Convergence Theorem, we have

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{d_{W,p}(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})^p}{t^p |A_X(t)|^p} &= \int_0^1 \lim_{t \rightarrow \infty} \left| \frac{\bar{F}_t^{\leftarrow}(q) - tq^{-1/\nu}}{tA_X(t)} - \frac{\bar{G}_t^{\leftarrow}(q) - tq^{-1/\nu}}{tA_Y(t)} \frac{A_Y(t)}{A_X(t)} \right|^p dq \\ &= \int_0^1 \left| \lim_{t \rightarrow \infty} \frac{\bar{F}_t^{\leftarrow}(q) - tq^{-1/\nu}}{tA_X(t)} \right|^p dq \end{aligned}$$

$$\begin{aligned}
&= \int_0^1 \left| \lim_{t \rightarrow \infty} \frac{\frac{U_F(1/(\bar{F}(t)q))}{t} - q^{-1/\nu}}{A_X(U_F(1/\bar{F}(t)))} \right|^p dq \\
&= \int_0^1 \left| \lim_{t \rightarrow \infty} \nu^{-2} \frac{\frac{U_F(1/(\bar{F}(t)q))}{U_F(1/\bar{F}(t))} - q^{-1/\nu}}{a_X(1/\bar{F}(t))} \right|^p dq \\
&= \int_0^1 \left| \nu^{-2} q^{-1/\nu} \frac{q^{-\rho_X/\nu} - 1}{\rho_X/\nu} \right|^p dq \\
&= \frac{1}{\nu^p |\rho_X|^p} \int_1^\infty \left( z^{1/\nu-2} - z^{(1+\rho_X)/\nu-2} \right)^p dz.
\end{aligned}$$

It is straightforward to see that the integration on the right-hand side above is finite. Taking  $1/p$ -th power on both sides completes the proof.

#### Appendix D.4. Proof of Lemma 10

Write  $h^*(t) = \int_t^\infty h(x) dx$ . By Proposition 6 of [Hua and Joe \(2011\)](#),  $h^*(t) \in 2\text{RV}_{-\alpha+1,\rho}$  with the auxiliary function defined as above, and it holds that

$$h^*(t) = th(t) \left( \frac{1}{\nu} + \frac{A(t)}{\rho} \left( \frac{1}{\nu-\rho} - \frac{1}{\nu} \right) + o(A(t)) \right),$$

which implies

$$h^*(t) = \frac{th(t)}{\nu} \left( 1 + \frac{A(t)}{\rho} \left( \frac{\nu}{\nu-\rho} - 1 \right) + o(A(t)) \right) = \frac{th(t)}{\nu} \left( 1 + \frac{A(t)}{\nu-\rho} + o(A(t)) \right).$$

Therefore,

$$\begin{aligned}
\frac{th(t) - \nu h^*(t)}{h^*(t)} &= \frac{th(t) - th(t) \left( 1 + \frac{A(t)}{\nu-\rho} + o(A(t)) \right)}{\frac{th(t)}{\nu} \left( 1 + \frac{A(t)}{\nu-\rho} + o(A(t)) \right)} \\
&= -\frac{\nu \left( \frac{A(t)}{\nu-\rho} + o(A(t)) \right)}{1 + \frac{A(t)}{\nu-\rho} + o(A(t))} \\
&= -\frac{\nu}{\nu-\rho} A(t) + o(A(t)),
\end{aligned}$$

which completes the proof.

#### Appendix D.5. Proof of Theorem 11

Recall equation (3.3). We have

$$\begin{aligned}
d_\phi(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t}) &= \int_t^\infty \phi \left( \frac{f(x) \bar{G}(t)}{g(x) \bar{F}(t)} \right) \frac{g(x)}{\bar{G}(t)} dx \\
&= \int_1^\infty \phi \left( \frac{f(zt)/\bar{F}(t)}{g(zt)/\bar{G}(t)} \right) \frac{tg(zt)}{\bar{G}(t)} dz \\
&= \int_1^\infty \phi(1 + h(z, t)) \frac{tg(zt)}{\bar{G}(t)} dz,
\end{aligned}$$

where

$$h(z, t) = \frac{f(zt)/\bar{F}(t) - g(zt)/\bar{G}(t)}{g(zt)/\bar{G}(t)}.$$

It is straightforward to see that for every  $t$ ,

$$\int_1^\infty h(z, t) \frac{tg(zt)}{\bar{G}(t)} dz = 0. \quad (\text{D.7})$$

Hence, we have

$$\begin{aligned} \frac{d_\phi(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})}{|A_X(t)|^\lambda} &= \frac{1}{|A_X(t)|^\lambda} \int_1^\infty \phi(1 + h(z, t)) \frac{tg(zt)}{\bar{G}(t)} dz \\ &= \frac{1}{|A_X(t)|^\lambda} \int_1^\infty (\phi(1 + h(z, t)) - ah(z, t)) \frac{tg(zt)}{\bar{G}(t)} dz \\ &= \int_1^\infty \frac{\phi(1 + h(z, t)) - ah(z, t)}{\zeta(h(z, t))} \frac{\zeta(h(z, t))}{|A_X(t)|^\lambda} \frac{tg(zt)}{\bar{G}(t)} dz \\ &= \frac{|A_X^*(t)|^\lambda}{|A_X(t)|^\lambda} \int_1^\infty \frac{\phi(1 + h(z, t)) - ah(z, t)}{\zeta(h(z, t))} \zeta\left(\frac{h(z, t)}{|A_X^*(t)|}\right) \frac{tg(zt)}{\bar{G}(t)} dz, \end{aligned} \quad (\text{D.8})$$

where  $A_X^*(t) = A_X(t)\nu/(\nu - \rho_X)$ , in the second step we used (D.7), and in the last step we used the fact that

$$\zeta(ch) = \zeta(h)c^\lambda, \quad h \in \mathbb{R}, c > 0. \quad (\text{D.9})$$

By Karamata's Theorem (see Resnick, 2007, Theorem 2.1) and Lemma 6,

$$\frac{tg(zt)}{\bar{G}(t)} = \frac{ztg(zt)}{\bar{G}(zt)} \frac{\bar{G}(zt)}{z\bar{G}(t)} \sim \nu z^{-(\nu+1)}, \quad (\text{D.10})$$

where the asymptotic relation holds uniformly over  $z \in (1, \infty)$ . A similar conclusion can be made for  $tf(zt)/\bar{F}(t)$ .

Recall  $A_X^*(t)$  given above and write  $A_Y^*(t) = A_Y(t)\nu/(\nu - \rho_Y)$ . By (2.7) and Lemma 10, it holds uniformly over  $z \in (1, \infty)$  that

$$\begin{aligned} &\frac{f(zt)}{\bar{F}(t)} - \frac{g(zt)}{\bar{G}(t)} \\ &= \frac{f(zt)}{\bar{F}(zt)} \frac{\bar{F}(zt)}{\bar{F}(t)} - \frac{g(zt)}{\bar{G}(zt)} \frac{\bar{G}(zt)}{\bar{G}(t)} \\ &= \frac{1}{tz} \left( \nu - \frac{\nu}{\nu - \rho_X} A_X(tz) + o(A_X(tz)) \right) \left( z^{-\nu} + z^{-\nu} \int_1^z u^{\rho_X-1} du A_X^*(t) + o(A_X^*(t)) \right) \\ &\quad - \frac{1}{tz} \left( \nu - \frac{\nu}{\nu - \rho_Y} A_Y(tz) + o(A_Y(tz)) \right) \left( z^{-\nu} + z^{-\nu} \int_1^z u^{\rho_Y-1} du A_Y^*(t) + o(A_Y^*(t)) \right) \\ &= \frac{1}{tz} \left( \nu - A_X^*(tz) + o(A_X(tz)) \right) \left( z^{-\nu} + z^{-\nu} \frac{z^{\rho_X} - 1}{\rho_X} A_X^*(t) + o(A_X^*(t)) \right) \\ &\quad - \frac{1}{tz} \left( \nu - A_Y^*(tz) + o(A_Y(tz)) \right) \left( z^{-\nu} + z^{-\nu} \frac{z^{\rho_Y} - 1}{\rho_Y} A_Y^*(t) + o(A_Y^*(t)) \right) \\ &= \frac{1}{tz} \left( \nu z^{-\nu} \frac{z^{\rho_X} - 1}{\rho_X} A_X^*(t) - z^{-\nu} A_X^*(tz) + o(A_X^*(t)) \right), \end{aligned} \quad (\text{D.11})$$

which is interpreted as

$$\limsup_{t \rightarrow \infty} \sup_{z > 1} \left| \frac{f(zt)/\bar{F}(t) - g(zt)/\bar{G}(t) - \frac{1}{tz} \left( \nu z^{-\nu} \frac{z^{\rho_X} - 1}{\rho_X} A_X^*(t) - z^{-\nu} A_X^*(tz) \right)}{A_X^*(t)} \right| = 0. \quad (\text{D.12})$$

Dividing both the numerator and denominator of the left-hand side of (D.12) by  $g(zt)/\bar{G}(t)$ , we obtain

$$\limsup_{t \rightarrow \infty} \sup_{z > 1} \left| \frac{h(z, t)}{A_X^*(t)} - \frac{\nu z^{-\nu-1} \frac{z^{\rho_X} - 1}{\rho_X} A_X^*(t) - z^{-\nu-1} A_X^*(tz)}{A_X^*(t) t g(zt)/\bar{G}(t)} \right| = 0.$$

It follows that for fixed  $\varepsilon > 0$ , there exists a  $t_1 = t_1(\varepsilon)$ , such that, for all large  $t > t_1(\varepsilon)$  and all  $z > 1$ ,

$$\begin{aligned} \left| \frac{h(z, t)}{A_X^*(t)} \right| &\leq \left| \frac{\nu z^{-\nu-1} \frac{z^{\rho_X} - 1}{\rho_X} A_X^*(t) - z^{-\nu-1} A_X^*(tz)}{A_X^*(t) t g(zt)/\bar{G}(t)} \right| + \varepsilon \\ &\leq (1 + \varepsilon) \left| \frac{\nu z^{-\nu-1} \frac{z^{\rho_X} - 1}{\rho_X} - z^{-\nu-1} A_X^*(tz)/A_X^*(t)}{\nu z^{-\nu-1}} \right| + \varepsilon \\ &\leq (1 + \varepsilon) \left| \frac{z^{\rho_X} - 1}{\rho_X} - \frac{A_X^*(tz)}{\nu A_X^*(t)} \right| + \varepsilon \\ &\leq (1 + \varepsilon) \frac{z^{\rho_X} - 1}{\rho_X} + (1 + \varepsilon) \frac{A_X^*(tz)}{\nu A_X^*(t)} + \varepsilon \\ &\leq (1 + \varepsilon) \frac{-1}{\rho_X} + \frac{1 + \varepsilon}{\nu} + \varepsilon \\ &:= C_\varepsilon \in (0, \infty), \end{aligned} \quad (\text{D.13})$$

where in the second step we used (D.10) and in the last step used the fact that  $|A_X^*(\cdot)| \in \text{RV}_{\rho_X}$ . Note that (D.13) implies

$$\limsup_{t \rightarrow \infty} \sup_{z > 1} h(z, t) = 0. \quad (\text{D.14})$$

Using (D.14), (3.4), (D.13), and (D.10), we see that, for large  $t$ , the integrand in (D.8) is bounded above by

$$(1 + \varepsilon) \max\{|k^+|, |k^-|\} C_\varepsilon^\lambda \nu z^{-(\nu+1)},$$

which is integrable over  $z \in (1, \infty)$ .

Applying the Dominated Convergence Theorem to (D.8), we have

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{d_\phi(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})}{|A_X(t)|^\lambda} &= \lim_{t \rightarrow \infty} \frac{|A_X^*(t)|^\lambda}{|A_X(t)|^\lambda} \int_1^\infty \frac{\phi(1 + h(z, t)) - ah(z, t)}{\zeta(h(z, t))} \zeta\left(\frac{h(z, t)}{|A_X^*(t)|}\right) \frac{tg(zt)}{\bar{G}(t)} dz \\ &= \frac{\nu^\lambda}{(\nu - \rho_X)^\lambda} \int_1^\infty \lim_{t \rightarrow \infty} \frac{\phi(1 + h(z, t)) - ah(z, t)}{\zeta(h(z, t))} \zeta\left(\frac{h(z, t)}{|A_X^*(t)|}\right) \frac{tg(zt)}{\bar{G}(t)} dz \\ &= \frac{\nu^\lambda}{(\nu - \rho_X)^\lambda} \int_1^\infty \zeta\left(\lim_{t \rightarrow \infty} \frac{h(z, t)}{|A_X^*(t)|}\right) \nu z^{-(\nu+1)} dz \end{aligned} \quad (\text{D.15})$$

where in the last step we used the continuity of  $\zeta(\cdot)$ . Moreover, it is easy to derive from (D.11)

and (D.10) that, for every fixed  $z > 1$ ,

$$\frac{h(z, t)}{A_X^*(t)} \rightarrow I(z),$$

where

$$I(z) = \frac{z^{\rho_X} - 1}{\rho_X} - \frac{z^{\rho_X}}{\nu} = \frac{\nu - \rho_X}{-\nu\rho_X} \left( \frac{\nu}{\nu - \rho_X} - z^{\rho_X} \right). \quad (\text{D.16})$$

It follows from (D.15) and (D.16) that

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{d_\phi(\mathbb{P}_{X,t}, \mathbb{P}_{Y,t})}{|A_X(t)|^\lambda} &= \frac{\nu^\lambda}{(\nu - \rho_X)^\lambda} \int_1^\infty \zeta \left( \lim_{t \rightarrow \infty} \frac{h(z, t)}{|A_X^*(t)|} \right) \nu z^{-(\nu+1)} dz \\ &= \frac{\nu^{\lambda+1}}{(\nu - \rho_X)^\lambda} \int_1^\infty \zeta \left( I(z) \text{sgn}(A_X^*(\infty^-)) \right) z^{-(\nu+1)} dz \\ &= \frac{\nu}{(-\rho_X)^\lambda} \int_1^\infty \zeta \left( \text{sgn}(A_X^*(\infty^-)) \left( \frac{\nu}{\nu - \rho_X} - z^{\rho_X} \right) \right) z^{-(\nu+1)} dz \\ &< \infty, \end{aligned}$$

where the second last step is due to (D.9). It is not difficult to verify that  $\int_1^\infty I(z) z^{-(\nu+1)} dz = 0$ , and hence, by Jensen's inequality we can show that the constant  $C(\nu, \rho_X, \lambda)$  is non-negative regardless of the sign of  $A_X$ . This completes the proof.

## References

- Ali, S. and Silvey, S. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of Royal Statistical Society, Series B*, 28(1):131–142.
- Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.
- Asimit, A. V., Furman, E., Tang, Q., and Vernic, R. (2011). Asymptotics for risk capital allocations based on conditional tail expectation. *Insurance: Mathematics and Economics*, 49(3):310–324.
- Asimit, V., Chong, W. F., Tunaru, R., and Zhou, F. (2024a). Portfolio selection and risk sharing via risk budgeting. City Research Online <https://tinyurl.com/7zv3frn>. Working paper.
- Asimit, V., Peng, L., Tunaru, R., and Zhou, F. (2024b). Constructing optimal portfolio under risk budgeting. City Research Online <https://tinyurl.com/ys76cpvu>. Working paper.
- Bai, X., Scheinberg, K., and Tutuncu, R. (2016). Least-square approach to risk parity in portfolio selection. *Quantitative Finance*, 16(3):357–376.
- Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., and Munos, R. (2017). The cramer distance as a solution to biased wasserstein gradients. *CoRR*, abs/1705.10743.

- Ben-Tal, A., Ben-Israel, A., and Teboulle, M. (1991). Certainty equivalents and information measures: Duality and extremal principles. *Journal of Mathematical Analysis and Application*, 157(1):211–236.
- Ben-Tal, A., den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357.
- Birghila, C. and Pflug, G. C. (2019). Optimal xl-insurance under wasserstein-type ambiguity. *Insurance: Mathematics and Economics*, 88:30–43.
- Blanchet, J., Lam, H., Tang, Q., and Yuan, Z. (2019). Robust actuarial risk analysis. *North American Actuarial Journal*, 23(1):33–63.
- Chen, H., Cummins, J. D., Viswanathan, K. S., and Weiss, M. A. (2014). Systemic risk and the interconnectedness between banks and insurers: An econometric analysis. *Journal of Risk and Insurance*, 81(3):623–652.
- Clarke, R., De Silva, H., and Thorley, S. (2013). Risk parity, maximum diversification, and minimum variance: An analytic perspective. *The Journal of Portfolio Management*, 39(3):39–53.
- Cressie, N. and Read, T. (1984). Multinomial goodness-of-fit tests. *Journal of Royal Statistical Society, Series B*, 46(3):440–464.
- Csiszár, I. (1966). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2(1):299–318.
- de Haan, L. and Ferreira, A. (2007). *Extreme value theory: an introduction*. Springer Science & Business Media.
- de Haan, L. and Resnick, S. (1996). Second-order regular variation and rates of convergence in extreme-value theory. *Annals of Probability*, 24(1):97–124.
- Del Castillo, J. and Serra, I. (2015). Likelihood inference for generalized Pareto distribution. *Computational Statistics & Data Analysis*, 83:116–128.
- DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of Financial studies*, 22(5):1915–1953.
- Embrechts, P., Mikosch, T., and Klüppelberg, C. (1997). *Modelling extremal events: for insurance and finance*. Springer-Verlag.
- Endres, D. and Schindelin, J. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860.
- Fisher, G. S., Maymin, P. Z., and Maymin, Z. G. (2015). Risk parity optimality. *The Journal of Portfolio Management*, 41(2):42–56.

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.
- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 37(3):424–438.
- Hallerbach, W., Ning, H., Spronk, J., and Soppe, A. (2004). A framework for managing a portfolio of socially responsible investments. *European Journal of Operational Research*, 153(2):517–529.
- Harremoës, P. and Vajda, I. (2011). On pairs of f-divergences and their joint range. *IEEE Transactions on Information Theory*, 57(6):3230–3235.
- Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution.
- Hong, Y., Liu, Y., and Wang, S. (2009). Granger causality in risk and detection of extreme risk spillover between financial markets. *Journal of Econometrics*, 150(2):271–287.
- Hua, L. and Joe, H. (2011). Second order regular variation and conditional tail expectation of multiple risks. *Insurance: Mathematics and Economics*, 49(3):537–546.
- Jager, L. and Wellner, J. (2007). Goodness-of-fit tests via phi-divergences. *Annals of Statistics*, 35(5):2018–2053.
- Ji, L., Tan, K. S., and Yang, F. (2021). Tail dependence and heavy tailedness in extreme risks. *Insurance: Mathematics and Economics*, 99:282–293.
- Kelly, B. and Jiang, H. (2014). Tail risk and asset prices. *The Review of Financial Studies*, 27(10):2841–2871.
- Lassance, N., DeMiguel, V., and Vrins, F. (2022). Optimal portfolio diversification via independent component analysis. *Operations Research*, 70:55–72.
- Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer-Verlag, New York, NY.
- Lux, T. and Papapantoleon, A. (2019). Model-free bounds on value-at-risk using extreme value information and statistical distances. *Insurance: Mathematics and Economics*, 86:73–83.
- Maillard, S., Roncalli, T., and Teïletche, J. (2010). The properties of equally weighted risk contribution portfolios. *The Journal of Portfolio Management*, 36(4):60–70.
- Mao, T. (2013). Second-order conditions of regular variation and drees-type inequalities. In *Stochastic Orders in Reliability and Risk: In Honor of Professor Moshe Shaked*, pages 313–330. Springer.
- Mao, T., Lv, W., and Hu, T. (2012). Second-order expansions of the risk concentration based on cte. *Insurance: Mathematics and Economics*, 51(2):449–456.

- Mao, T., Stupfler, G., and Yang, F. (2023). Asymptotic properties of generalized shortfall risk measures for heavy-tailed risks. *Insurance: Mathematics and Economics*, 111:173–192.
- Mausser, H. and Romanko, O. (2018). Long-only equal risk contribution portfolios for cvar under discrete distributions. *Quantitative Finance*, 18(11):1927–1945.
- Mazzarisi, P., Zaoli, S., Campajola, C., and Lillo, F. (2020). Tail granger causalities and where to find them: Extreme risk spillovers vs spurious linkages. *Journal of Economic Dynamics and Control*, 121:104022.
- McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press.
- Nassif, A. B., Talib, M. A., Nasir, Q., and Dakalbab, F. M. (2021). Machine learning for anomaly detection: A systematic review. *Ieee Access*, 9:78658–78700.
- Pardo, L. (2005). *Statistical Inference Based on Divergence Measures (1st ed.)*. Chapman and Hall/CRC.
- Qian, E. E. (2005). On the financial interpretation of risk contribution: Risk budgets do add up. Working paper.
- Qin, X. and Zhou, C. (2021). Systemic risk allocation using the asymptotic marginal expected shortfall. *Journal of Banking & Finance*, 126:106099.
- Rachev, S., Klebanov, L., Stoyanov, S., and Fabozzi, F. (2013). *The Methods of Distances in the Theory of Probability and Statistics*. Springer.
- Resnick, S. I. (2007). *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media.
- Resnick, S. I. (2008). *Extreme values, regular variation, and point processes*, volume 4. Springer Science & Business Media.
- Roncalli, T. (2013). *Introduction to risk parity and budgeting*. CRC Press.
- Roncalli, T. and Weisang, G. (2016). Risk parity portfolio with risk factors? *Quantitative Finance*, 16:377–388.
- Sason, I. and Verdu, S. (2016). f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006.
- Sims, C. A. (1972). Money, income, and causality. *The American economic review*, 62(4):540–552.
- Spinu, F. (2013). An algorithm for computing risk parity weights. SSRN.
- Sun, H., Chen, Y., and Hu, T. (2022). Statistical inference for tail-based cumulative residual entropy. *Insurance: Mathematics and Economics*, 103:66–95.

Tang, Q. and Yang, Y. (2023). Worst-case moments under partial ambiguity. *ASTIN Bulletin: The Journal of the IAA*, 53(2):443–465.

Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*. Springer Verlag, New York, NY.