HEAVY TAILED ANALYSIS
EURANDOM
SUMMER 2005

SIDNEY RESNICK

School of Operations Research and Industrial Engineering
Cornell University
Ithaca, NY 14853 USA
sir1@cornell.edu
http://www.orie.cornell.edu/∼sid

&
Eurandom
http://www.eurandom.tue.nl/people/EURANDOM_chair/eurandom_chair.htm

1. Course Abstract

This is a survey of some of the mathematical, probabilistic and statistical tools used in heavy tail analysis. Heavy tails are characteristic of phenomena where the probability of a huge value is relatively big. Record breaking insurance losses, financial log-returns, file sizes stored on a server, transmission rates of files are all examples of heavy tailed phenomena. The modeling and statistics of such phenomena are tail dependent and much different than classical modeling and statistical analysis which give primacy to central moments, averages and the normal density, which has a wimpy, light tail. An organizing theme is that many limit relations giving approximations can be viewed as applications of almost surely continuous maps.

2. Introduction

Heavy tail analysis is an interesting and useful blend of mathematical analysis, probability and stochastic processes and statistics. Heavy tail analysis is the study of systems whose behavior is governed by large values which shock the system periodically. This is in contrast to many stable systems whose behavior is determined largely by an averaging effect. In heavy tailed analysis, typically the asymptotic behavior of descriptor variables is determined by the large values or merely a single large value.

Roughly speaking, a random variable X has a heavy (right) tail if there exists a positive parameter $\alpha > 0$ such that

$$P[X > x] \sim x^{-\alpha}, \quad x \to \infty. \tag{2.1}$$

(Note here we use the notation

$$f(x) \sim g(x), \quad x \to \infty$$

as shorthand for

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1,$$

for two real functions f, g.) Examples of such random variables are those with Cauchy, Pareto, t, F or stable distributions. Stationary stochastic processes, such as the ARCH, GARCH, EGARCH etc, which have been proposed as models for financial returns have marginal distributions satisfying (2.1). It turns out that (2.1) is not quite the right mathematical setting for discussing heavy tails (that pride of place belongs to regular variation of real functions) but we will get to that in due course.

Note the elementary observation that a heavy tailed random variable has a relatively large probability of exhibiting a really large value, compared to random variables which have exponentially bounded tails such as normal, Weibull, exponential or gamma random variables. For a $N(0,1)$ normal random variable N, with density $n(x)$, we have by Mill's ratio that

$$P[N > x] \sim \frac{n(x)}{x} \sim \frac{1}{x\sqrt{2\pi}} e^{-x^2/2}, \quad x \to \infty,$$

which has much weaker tail weight than suggested by (2.1).

There is a tendency to sometimes confuse the concept of a heavy tail distribution with the concept of a distribution with infinite right support. (For a probability distribution F, the support is the smallest closed set C such that $F(C) = 1$. For the exponential distribution with no translation, the support is $[0, \infty)$ and for the normal distribution, the support is $\mathbb{R}$.) The distinction is simple and exemplified by comparing a normally distributed random variable with one whose distribution is Pareto. Both have positive probability of achieving a value bigger than any pre-assigned threshold. However, the Pareto random variable has, for large thresholds, a much bigger probability of exceeding the threshold. One cannot rule out heavy tailed distributions by using the argument that everything in the world is bounded unless one agrees to rule out all distributions with unbounded support.

Much of classical statistics is often based on averages and moments. Try to imagine a statistical world where you do not rely on moments since if (2.1) holds, moments above the α-th do not exist! This follows since

$$\int_0^\infty x^{\beta-1} P[X > x] dx \approx \int_0^\infty x^{\beta-1} x^{-\alpha} dx \begin{cases} < \infty, & \text{if } \beta < \alpha, \\ = \infty, & \text{if } \beta \geq \alpha, \end{cases}$$

where (in this case)

$$\int f \approx \int g$$

means both integrals either converge or diverge together. Much stability theory in stochastic modeling is expressed in terms of mean drifts but what if the means do not exist. Descriptor variables in queueing theory are often in terms of means such as mean waiting time, mean queue lengths and so on. What if such expectations are infinite?

Consider the following scenarios where heavy tailed analysis is used.

(i) Finance. It is empirically observed that "returns" possess several notable features, sometimes called *stylized facts*. What is a "return"? Suppose $\{S_i\}$ is the stochastic process representing the price of a speculative asset (stock, currency, derivative, commodity (corn, coffee, etc)) at the ith measurement time. The return process is

$$\tilde{R}_i := (S_i - S_{i-1})/S_{i-1};$$

that is, the process giving the relative difference of prices. If the returns are small then the differenced log-Price process approximates the return process

$$\begin{aligned} R_i := \log S_i - \log S_{i-1} &= \log \frac{S_i}{S_{i-1}} = \log\left(1 + \left(\frac{S_i}{S_{i-1}} - 1\right)\right) \\ &\sim \frac{S_i}{S_{i-1}} - 1 = \tilde{R}_i \end{aligned}$$

since for $|x|$ small,

$$\log(1 + x) \sim x$$

by, say, L'Hospital's rule. So instead of studying the returns $\{\tilde{R}_i\}$, the differenced log-Price process $\{R_i\}$ is studied and henceforth we refer to $\{R_i\}$ as the *returns*.

Empirically, either process is often seen to exhibit notable properties:

(1) Heavy tailed marginal distributions (but usually $2 < \alpha$ so the mean and variance exist);
(2) Little or no correlation. However by squaring or taking absolute values of the process one gets a highly correlated, even long range dependent process.
(3) The process is dependent. (If the random variables were independent, so would the squares be independent but squares are typically correlated.)

Hence one needs to model the data with a process which is stationary, has heavy tailed marginal distributions and a dependence structure. This leads to the study of specialized models in economics with lots of acronyms like ARCH and GARCH. Estimation of, say, the marginal distribution's shape parameter α are made more complex due to the fact that the observations are not independent.

Classical Extreme Value Theory which subsumes heavy tail analysis uses techniques to estimate *value-at-risk* or (VaR), which is an extreme quantile of the profit/loss density, once the density is estimated.

Note, that given S_0, there is a one-to-one correspondence between

$$\{S_0, S_1, \ldots, S_T\} \text{ and } \{S_0, R_1, \ldots, R_T\}$$

since

$$\begin{aligned}\sum_{t=1}^{T} R_t =&(\log S_1 - \log S_0) + (\log S_2 - \log S_1)\\ &+ \cdots + (\log S_T - \log S_{T-1})\\ =&\log S_T - \log S_0 = \log \frac{S_T}{S_0},\end{aligned}$$

so that

$$S_T = S_0 e^{\sum_{t=1}^{T} R_t}. \tag{2.2}$$

Why deal with returns rather than the price process?

(1) The returns are scale free and thus independent of the size of the investment.
(2) Returns have more attractive statistical properties than prices such as stationarity. Econometric models sometimes yield non-stationary price models but stationary returns.

To convince you this might make a difference to somebody, note that from 1970-1995, the two worst losses world wide were Hurrricane Andrew (my wife's cousin's yacht in Miami wound up on somebody's roof 30 miles to the north) and the Northridge earthquake in California. Losses in 1992 dollars were \$16,000 and \$11,838 million dollars respectively. (Note the unit is "millions of dollars".)

Why deal with log-returns rather than returns?

(1) Log returns are nicely additive over time. It is easier to construct models for additive phenomena than multiplicative ones (such as $1 + \tilde{R}_t = S_t/S_{t-1}$). One

can recover S_T from log-returns by what is essentially an additive formula (2.2). (Additive is good!) Also, the T-day return process

$$R_T - R_1 = \log S_T - \log S_0$$

is additive. (Additive is good!)

(2) Daily returns satisfy

$$\frac{S_t}{S_{t-1}} - 1 \geq -1,$$

and for statistical modeling, it is a bit unnatural to have the variable bounded below by -1. For instance one could not model such a process using a normal or two-sided stable density.

(3) Certain economic facts are easily expressed by means of log-returns. For example, if S_t is the exchange rate of the US dollar against the British pound and $R_t = \log(S_t/S_{t-1})$, then $1/S_t$ is the exchange rate of pounds to dollars and the return from the point of view of the British investor is

$$\log \frac{1/S_t}{1/S_{t-1}} = \log \frac{S_{t-1}}{S_t} = -\log \frac{S_t}{S_{t-1}}$$

which is minus the return for the American investor.

(4) The operations of taking logarithms and differencing are standard time series tools for coercing a data set into looking stationary. Both operations, as indicated, are easily undone. So there is a high degree of comfort with these operations.

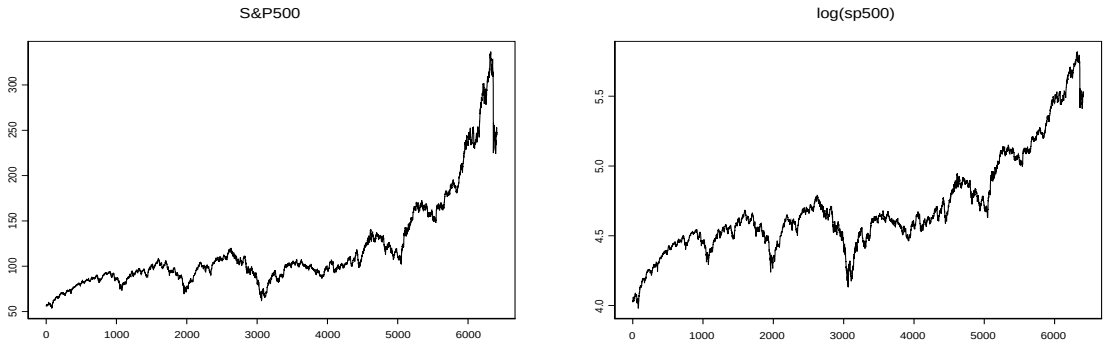

FIGURE 1. Time series plot of S&P 500 data (left) and log(S&P500) (right).

Example 1 (Standard & Poors 500)**.** We consider the data set *fm-poors.dat* in the package Xtremes which gives the Standard & Poors 500 stock market index. The data is daily data from July 1962 to December 1987 but of course does not include days when the market is closed. In Figure 1 we display the time series plots of the actual data for the index and the log of the data. Only a lunatic would conclude these two series were stationary. In the left side of Figure 2 we exhibit the 6410 returns $\{R_t\}$ of the data by differencing at lag 1 the log(S&P) data. On the right side is the sample autocorrelation function. There is a biggish lag 1 correlation but otherwise few spikes are outside the magic window.

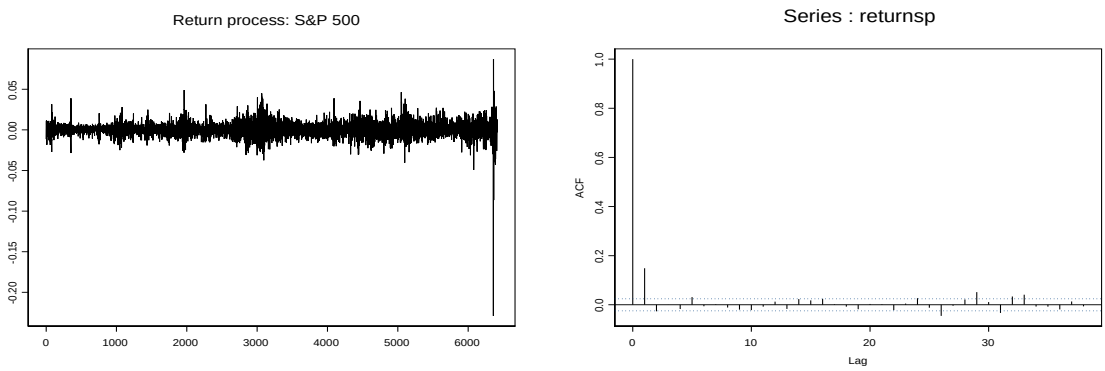

FIGURE 2. Time series plot of S&P 500 return data (left) and the autocorrelation function (right).

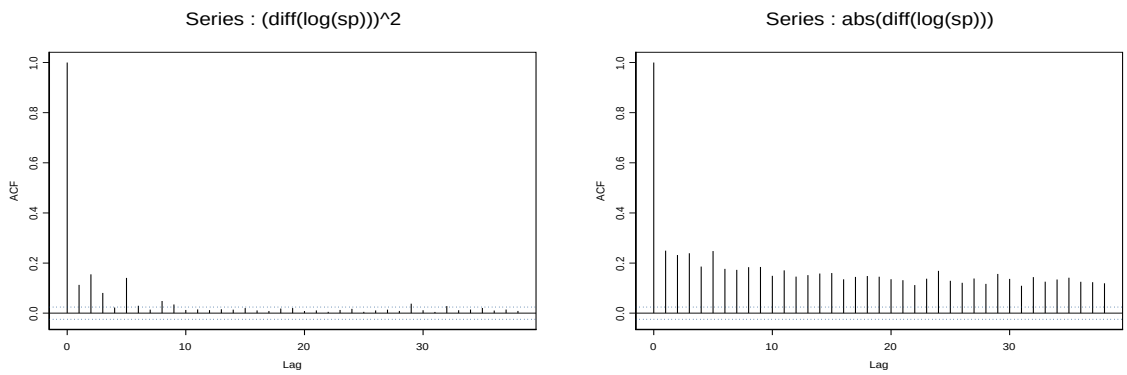

FIGURE 3. i) The autocorrelation function of the squared returns (left). (ii) The autocorrelation function of the absolute values of the returns. (right)

For a view of the *stylized facts* about these data, and to indicate the complexities of the dependence structure, we exhibit the autocorrelation function of the squared returns in Figure 3 (left) and on the right the autocorrelation function for the absolute value of the returns. Though there is little correlation in the original series, the iid hypothesis is obviously false.

One can compare the heaviness of the right and left tail of the marginal distribution of the process $\{R_t\}$ even if we do not believe that the process is iid. A reasonable assumption seems to be that the data can be modelled by a stationary, uncorrelated process and we hope the standard exploratory extreme value and heavy tailed methods developed for iid processes still apply. We apply the QQ-plotting technique to the data. After playing a bit with the number of upper order statistics used, we settled on $k = 200$ order statistics for the positive values (upper tail) which gives the slope estimate of $\hat{\alpha} = 3.61$. This is shown in the left side of Figure 4. On the right side of Figure 4 is the comparable plot for the left tail; here we applied the routine to abs(returns[returns¡0]); that is, to the absolute value of

the negative data points in the log-return sample. After some experimentation, we obtained an estimate $\hat{\alpha} = 3.138$ using $k = 150$. Are the two tails symmetric which is a common theoretical assumption? Unlikely!

(ii) Insurance and reinsurance. The general theme here is to model insurance claim sizes and frequencies so that premium rates may be set intelligently and risk to the insurance company quantified.

Smaller insurance companies sometimes pay for reinsurance or excess-of-loss (XL) insurance to a bigger company like Lloyd's of London. The excess claims over a certain contractually agreed threshhold is covered by the big insurance company. Such excess claims are by definition very large so heavy tail analysis is a natural tool to apply. What premium should the big insurance company charge to cover potential losses?

As an example of data you might encounter, consider the Danish data on large fire insurance losses McNeil (1997), Resnick (1997). Figure 5 gives a time series plot of the 2156 Danish data consisting of losses over one million Danish Krone (DKK) and the right hand plot is the QQ plot of this data yielding a remarkably straight plot. The straight line plot indicates the appropriateness of heavy tail analysis.

(iii) Data networks. A popular idealized data transmission model of a source destination pair is an *on/off* model where constant rate transmissions alternate with *off* periods. The *on* periods are random in length with a *heavy tailed distribution* and this leads to occasional large transmission lengths. The model offers an explanation of perceived *long range dependence* in measured traffic rates. A competing model which is marginally more elegant in our eyes is the infinite source Poisson model to be discussed later along with all its warts.

Example 2. The Boston University study (Crovella and Bestavros (1995), Crovella and Bestavros (1996a), Cunha et al. (1995)) suggests self-similarity of web traffic stems from heavy tailed file sizes. This means that we treat files as being randomly selected from a population and if X represents a randomly selected file size then the heavy tail hypothesis

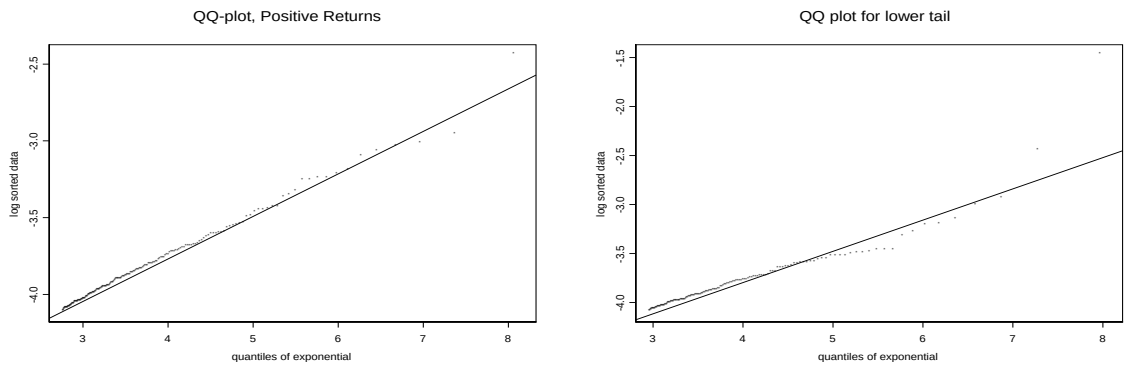

FIGURE 4. Left: QQ-plot and parfit estimate of α for the right tail using $k = 200$ upper order statistics. Right: QQ-plot and parfit estimate of α for the left tail using the absolute value of the negative values in the log-returns.

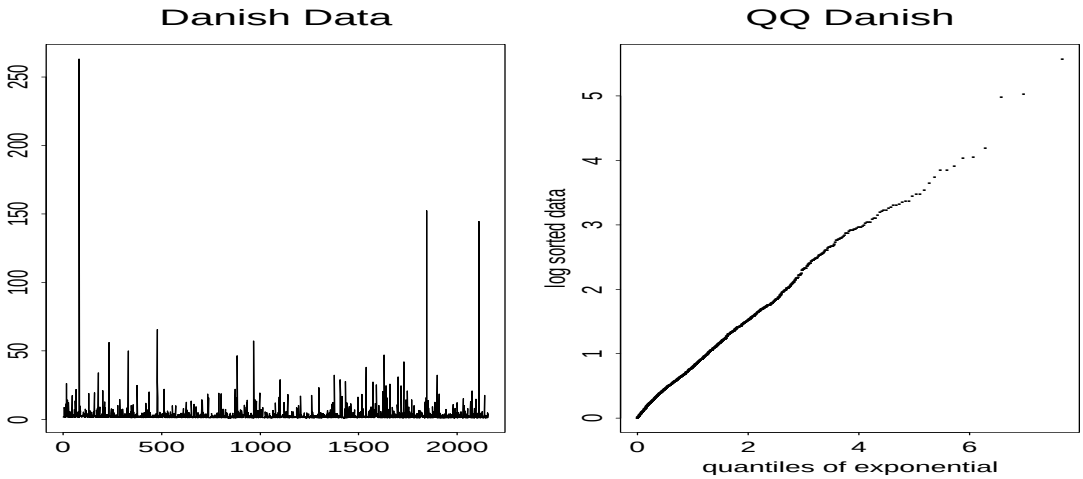

FIGURE 5. Danish Data (left) and QQ-plot.

means for large $x > 0$

$$P[X > x] \sim x^{-\alpha}, \quad \alpha > 0, \tag{2.3}$$

where α is a shape parameter that must be statistically estimated. The BU study reports an overall estimate for a five month measurement period (see Cunha et al. (1995)) of $\alpha = 1.05$. However, there is considerable month-to-month variation in these estimates and, for instance, the estimate for November 1994 in room 272 places α in the neighborhood of 0.66. Figure 6 gives the QQ and Hill plots (Beirlant et al. (1996), Hill (1975), Kratz and Resnick (1996), Resnick and Stărică (1997)) of the file size data for the month of November in the Boston University study. These are two graphical methods for estimating α and will be discussed in more detail later.

Extensive traffic measurements of *on* periods are reported in Willinger et al. (1995) where measured values of α were usually in the interval $(1, 2)$. Studies of sizes of files accessed on various servers by the Calgary study (Arlitt and Williamson (1996)), report estimates of α from 0.4 to 0.6. So accumulating evidence already exists which suggests values of α outside the range $(1, 2)$ should be considered. Also, as user demands on the web grow and access speeds increase, there may be a drift toward heavier file size distribution tails. However, this is a hypothesis that is currently untested.

REFERENCES

M. Arlitt and C. Williamson. Web server workload characterization: The search for invariants (extended version). In *Proceedings of the ACM Sigmetrics Conference, Philadelphia, Pa*, 1996. Available from {mfa126,carey}@cs.usask.ca.

J. Beirlant, P. Vynckier, and J. Teugels. Tail index estimation, Pareto quantile plots, and regression diagnostics. *J. Amer. Statist. Assoc.*, 91(436):1659–1667, 1996. ISSN 0162-1459.

M. Crovella and A. Bestavros. Explaining world wide web traffic self–similarity. Preprint available as TR-95-015 from {crovella,best}@cs.bu.edu, 1995.

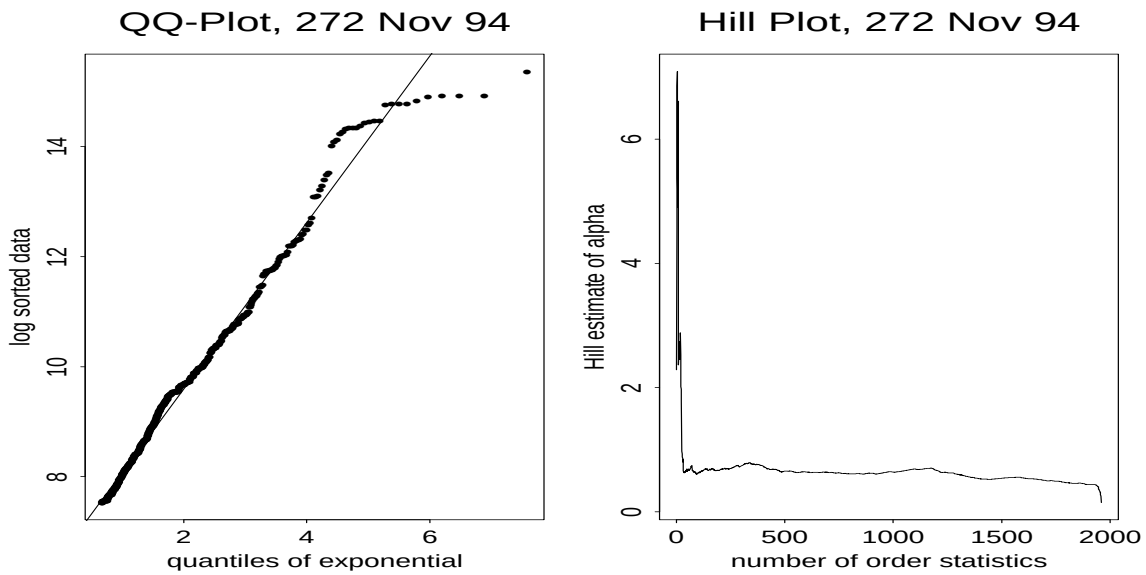

FIGURE 6. QQ and Hill plots of November 1994 file lengths.

M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. In *Proceedings of the ACM SIGMETRICS '96 International Conference on Measurement and Modeling of Computer Systems*, volume 24, pages 160–169, 1996.

C. Cunha, A. Bestavros, and M. Crovella. Characteristics of www client–based traces. Preprint available as BU-CS-95-010 from {crovella,best}@cs.bu.edu, 1995.

B. Hill. A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, 3:1163–1174, 1975.

M. Kratz and S. Resnick. The qq–estimator and heavy tails. *Stochastic Models*, 12:699–724, 1996.

A. McNeil. Estimating the tails of loss severity distributions using extreme value theory. *Astin Bulletin*, 27:117–137, 1997.

S. Resnick. Discussion of the danish data on large fire insurance losses. *Astin Bulletin*, 27: 139–151, 1997.

S. Resnick and C. Stărică. Smoothing the Hill estimator. *Adv. Applied Probability*, 29: 271–293, 1997.

W. Willinger, M. Taqqu, M. Leland, and D. Wilson. Self–similarity in high–speed packet traffic: analysis and modelling of ethernet traffic measurements. *Statistical Science*, 10: 67–85, 1995.

3. A Crash Course on Regular Variation

The theory of regularly varying functions is the appropriate mathematical analysis tool for proper discussion of heavy tail phenomena. We begin by reviewing some results from analysis starting with uniform convergence.

3.1. **Preliminaries from analysis.**

3.1.1. *Uniform convergence.* If $\{f_n, n \geq 0\}$ are real valued functions on $\mathbb{R}$ (or, in fact, any metric space) then f_n converges uniformly on $A \subset \mathbb{R}$ to f if

$$\sup_A |f_0(x) - f_n(x)| \to 0 \tag{3.1}$$

as $n \to \infty$. The definition would still make sense of the range of $f_n, n \geq 0$ were a metric space but then $|f_0(x) - f_n(x)|$ would need to be replaced by $d(f_0, f_n)$, where $d(\cdot, \cdot)$ is the metric. For functions on $\mathbb{R}$, the phrase *local uniform convergence* means that (3.1) holds for any compact interval A.

If $U_n, n \geq 0$ are non-decreasing real valued functions on $\mathbb{R}$, then a useful fact is that if U_0 is continuous and $U_n(x) \to U_0(x)$ as $n \to \infty$ for all x, then $U_n \to U$ locally uniformly; i.e. for any $a < b$

$$\sup_{x \epsilon [a,b]} |U_n(x) - U_0(x)| \to 0.$$

(See (Resnick, 1987, page 1).) One proof of this fact is outlined as follows: If U_0 is continuous on $[a, b]$, then it is uniformly continuous. From the uniform convergence, for any x, there is an interval-neighborhood O_x on which $U_0(\cdot)$ oscillates by less than a given ϵ. This gives an open cover of $[a, b]$. Compactness of $[a, b]$ allows us to prune $\{O_x,\ x \in [a, b]\}$ to obtain a finite subcover $\{(a_i, b_i), i = 1, \ldots, K\}$. Using this finite collection and the monotonicity of the functions leads to the result: Given $\epsilon > 0$, there exists some large N such that if $n \geq N$ then

$$\max_{1 \leq i \leq N} |U_n(a_i) - U_0(a_i)| \bigvee U_n(b_i) - U_0(b_i)| < \epsilon,$$

(by pointwise convergence). Observe that

$$\sup_{x \in [a,b]} |U_n(x) - U_0(x)| \leq \max_{1 \leq i \leq N} \sup_{[a_i, b_i]} |U_n(x) - U_0(x)|. \tag{3.2}$$

For any $x \in [a_i, b_i]$, we have by monotonicity

$$\begin{aligned} U_n(x) - U_0(x) \leq & U_n(b_i) - U_0(a_i) \\ \leq & U_0(b_i) + \epsilon - U_0(a_i), \quad \text{(by (3.2))} \\ \leq & 2\epsilon, \end{aligned}$$

with a similar lower bound. This is true for all i and hence we get uniform convergence on $[a, b]$.

3.1.2. *Inverses of monotone functions.* Suppose $H : \mathbb{R} \mapsto (a, b)$ is a non-decreasing function on $\mathbb{R}$ with range (a, b), $-\infty \le a < b \le \infty$. With the convention that the infimum of an empty set is $+\infty$, we define the (left continuous) inverse $H^{\leftarrow} : (a, b) \mapsto \mathbb{R}$ of H as

$$H^{\leftarrow}(y) = \inf\{s : H(s) \ge y\}.$$

In case the function H is right continuous we have the following interesting properties:

$$A(y) := \{s : H(s) \ge y\} \text{ is closed,} \tag{3.3}$$

$$H(H^{\leftarrow}(y)) \ge y \tag{3.4}$$

$$H^{\leftarrow}(y) \le t \text{ iff } y \le H(t). \tag{3.5}$$

For (3.3), observe that if $s_n \in A(y)$ and $s_n \downarrow s$, then $y \le H(s_n) \downarrow H(s)$ so $H(s) \ge y$ and $s \in A(y)$. If $s_n \uparrow s$ and $s_n \in A(y)$, then $y \le H(s_n) \uparrow H(s-) \le H(s)$ and $H(s) \ge y$ so $s \in A(y)$ again and $A(y)$ is closed. Since $A(y)$ is closed, $\inf A(y) \in A(y)$; that is, $H^{\leftarrow}(y) \in A(y)$ which means $H(H^{\leftarrow}(y)) \ge y$. This gives (3.4). Lastly, (3.5) follows from the definition of $H^{\leftarrow}$.

3.1.3. *Convergence of monotone functions.* For any function H denote

$$\mathcal{C}(H) = \{x \in \mathbb{R} : H \text{ is finite and continuous at } x\}.$$

A sequence $\{H_n, n \ge 0\}$ of non-decreasing functions on $\mathbb{R}$ converges weakly to H_0 if as $n \to \infty$ we have

$$H_n(x) \to H_0(x),$$

for all $x \in \mathcal{C}(H_0)$. We will denote this by $H_n \to H_0$ and no other form of convergence for monotone functions will be relevant. If $F_n, n \ge 0$ are non-defective distributions, then a myriad of names give equivalent concepts: complete convergence, vague convergence, weak* convergence, narrow convergence. If $X_n, n \ge 0$ are random variables and X_n has distribution function $F_n, n \ge 0$, then $X_n \Rightarrow X_0$ means $F_n \to F_0$. For the proof of the following, see (Billingsley, 1986, page 343), (Resnick, 1987, page 5), (Resnick, 1998, page 259).

Proposition 1. *If $H_n, n \ge 0$ are non-decreasing functions on $\mathbb{R}$ with range (a, b) and $H_n \to H_0$, then $H_n^{\leftarrow} \to H_0^{\leftarrow}$ in the sense that for $t \in (a, b) \cap \mathcal{C}(H_0^{\leftarrow})$*

$$H_n^{\leftarrow}(t) \to H_0^{\leftarrow}(t).$$

3.1.4. *Cauchy's functional equation.* Let $k(x), x \in \mathbb{R}$ be a function which satisfies

$$k(x + y) = k(x) + k(y), x, y \in R.$$

If k is measurable and bounded on a set of positive measure, then $k(x) = cx$ for some $c \in \mathbb{R}$. (See Seneta (1976), (Bingham et al., 1987, page 4).)

3.2. **Regular variation: definition and first properties.** An essential analytical tool for dealing with heavy tails, long range dependence and domains of attraction is the theory of regularly varying functions. This theory provides the correct mathematical framework for considering things like Pareto tails and algebraic decay.

Roughly speaking, *regularly varying functions* are those functions which behave asymptotically like power functions. We will deal currently only with real functions of a real variable. Consideration of multivariate cases and probability concepts suggests recasting definitions in terms of vague convergence of measures but we will consider this reformulation later.

Definition 1. A measurable function $U : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is regularly varying at ∞ with index $\rho \in \mathbb{R}$ (written $U \in RV_\rho$) if for $x > 0$

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\rho.$$

We call ρ the *exponent of variation.*

If $\rho = 0$ we call U *slowly varying.* Slowly varying functions are generically denoted by $L(x)$. If $U \in RV_\rho$, then $U(x)/x^\rho \in RV_0$ and setting $L(x) = U(x)/x^\rho$ we see it is always possible to represent a ρ-varying function as $x^\rho L(x)$.

Examples. The canonical ρ-varying function is x^ρ. The functions $\log(1+x),\ \log\log(e+x)$ are slowly varying, as is $\exp\{(\log x)^\alpha\}, 0 < \alpha < 1$. Any function U such that $\lim_{x\to\infty} U(x) =: U(\infty)$ exists finite is slowly varying. The following functions are not regularly varying: $e^x, \sin(x+2)$. Note $[\log x]$ is slowly varying, but $\exp\{[\log x]\}$ is not regularly varying.

In probability applications we are concerned with distributions whose tails are regularly varying. Examples are

$$1 - F(x) = x^{-\alpha},\ x \geq 1,\ \alpha > 0,$$

and the extreme value distribution

$$\Phi_\alpha(x) = \exp\{-x^{-\alpha}\},\ x \geq 0.$$

$\Phi_\alpha(x)$ has the property

$$1 - \Phi_\alpha(x) \sim x^{-\alpha} \text{ as } x \to \infty.$$

A stable law (to be discussed later) with index $\alpha, 0 < \alpha < 2$ has the property

$$1 - G(x) \sim cx^{-\alpha},\ x \to \infty,\ c > 0.$$

The Cauchy density $f(x) = (\pi(1+x^2))^{-1}$ has a distribution function F with the property

$$1 - F(x) \sim (\pi x)^{-1}.$$

If $N(x)$ is the standard normal df then $1 - N(x)$ is not regularly varying nor is the tail of the Gumbel extreme value distribution $1 - \exp\{-e^{-x}\}$.

The definition of regular variation can be weakened slightly (cf Feller (1971), de Haan (1970), Resnick (1987)).

Proposition 2. *(i) A measurable function $U : \mathbb{R}_+ \mapsto \mathbb{R}_+$ varies regularly if there exists a function h such that for all $x > 0$*

$$\lim_{t\to\infty} U(tx)/U(t) = h(x).$$

In this case $h(x) = x^\rho$ *for some* $\rho \in \mathbb{R}$ *and* $U \in RV_\rho$.

(ii) A monotone function $U : \mathbb{R}_+ \mapsto \mathbb{R}_+$ *varies regularly provided there are two sequences* $\{\lambda_n\}, \{a_n\}$ *of positive numbers satisfying*

$$a_n \to \infty, \qquad \lambda_n \sim \lambda_{n+1}, \quad n \to \infty, \tag{3.6}$$

and for all $x > 0$

$$\lim_{n\to\infty} \lambda_n U(a_n x) =: \chi(x) \text{ exists positive and finite.} \tag{3.7}$$

In this case $\chi(x)/\chi(1) = x^\rho$ *and* $U \in RV_\rho$ *for some* $\rho \in \mathbb{R}$.

We frequently refer to (3.7) as the *sequential form of regular variation*. For probability purposes, it is the most useful. Typically U is a distribution tail, $\lambda_n = n$ and a_n is a distribution quantile.

Proof. (i) The function h is measurable since it is a limit of measurable functions. Then for $x > 0, y > 0$

$$\frac{U(txy)}{U(t)} = \frac{U(txy)}{U(tx)} \cdot \frac{U(tx)}{U(t)}$$

and letting $t \to \infty$ gives

$$h(xy) = h(y)h(x).$$

So h satisfies the Hamel equation, which by change of variable can be converted to the Cauchy equation. Therefore, the form of h is $h(x) = x^\rho$ for some $\rho \in \mathbb{R}$.

(ii) For concreteness assume U is nondecreasing. Assume (3.6) and (3.7) and we show regular variation. Since $a_n \to \infty$, for each t there is a finite $n(t)$ defined by

$$n(t) = \inf\{m : a_{m+1} > t\}$$

so that

$$a_{n(t)} \le t < a_{n(t)+1}.$$

Therefore by monotonicity for $x > 0$

$$\left(\frac{\lambda_{n(t)+1}}{\lambda_{n(t)}}\right)\left(\frac{\lambda_{n(t)}U(a_{n(t)}x)}{\lambda_{n(t)+1}U(a_{n(t)+1})}\right) \le \frac{U(tx)}{U(t)} \le \left(\frac{\lambda_{n(t)}}{\lambda_{n(t)+1}}\right)\left(\frac{\lambda_{n(t)+1}U(a_{n(t)+1}x)}{\lambda_{n(t)}U(a_{n(t)})}\right).$$

Now let $t \to \infty$ and use (3.6) and (3.7) to get $\lim_{t\to\infty} \frac{U(tx)}{U(t)} = 1\frac{\chi(x)}{\chi(1)}$. Regular variation follows from part (i). □

Remark 1. Proposition 2 (ii) remains true if we only assume (3.7) holds on a dense set. This is relevant to the case where U is nondecreasing and $\lambda_n U(a_n x)$ converges weakly.

3.2.1. *A maximal domain of attraction.* Suppose $\{X_n, n \geq 1\}$ are iid with common distribution function $F(x)$. The extreme is

$$M_n = \bigvee_{i=1}^{n} X_i = \max\{X_1, \ldots, X_n\}.$$

One of the extreme value distributions is

$$\Phi_\alpha(x) := \exp\{-x^{-\alpha}\}, \quad x > 0,\ \alpha > 0.$$

What are conditions on F, called *domain of attraction conditions*, so that there exists $a_n > 0$ such that

$$P[a_n^{-1} M_n \leq x] = F^n(a_n x) \to \Phi_\alpha(x) \tag{3.8}$$

weakly. How do you characterize the normalization sequence $\{a_n\}$?

Set $x_0 = \sup\{x : F(x) < 1\}$ which is called the right end point of F. We first check (3.8) implies $x_0 = \infty$. Otherwise if $x_0 < \infty$ we get from (3.8) that for $x > 0, a_n x \to x_0$; i.e. $a_n \to x_0 x^{-1}$. Since $x > 0$ is arbitrary we get $a_n \to 0$ whence $x_0 = 0$. But then for $x > 0, F^n(a_n x) = 1$, which violates (3.8). Hence $x_0 = \infty$.

Furthermore $a_n \to \infty$ since otherwise on a subsequence $n', a_{n'} \leq K$ for some $K < \infty$ and

$$0 < \Phi_\alpha(1) = \lim_{n' \to \infty} F^{n'}(a_{n'}) \leq \lim_{n' \to \infty} F^{n'}(K) = 0$$

since $F(K) < 1$ which is a contradiction.

In (3.8), take logarithms to get for $x > 0, \lim_{n\to\infty} n(-\log F(a_n x)) = x^{-\alpha}$. Now use the relation $-\log(1-z) \sim z$ as $z \to 0$ and (7) is equivalent to

$$\lim_{n\to\infty} n(1 - F(a_n x)) = x^{-\alpha}, \quad x > 0. \tag{3.9}$$

From (3.9) and Proposition 2 we get

$$1 - F(x) \sim x^{-\alpha} L(x), \quad x \to \infty, \tag{3.10}$$

for some $\alpha > 0$. To characterize $\{a_n\}$ set $U(x) = 1/(1 - F(x))$ and (3.9) is the same as

$$U(a_n x)/n \to x^\alpha, x > 0$$

and inverting we find via Proposition 1 that

$$\frac{U^{\leftarrow}(ny)}{a_n} \to y^{1/\alpha}, y > 0.$$

So $U^{\leftarrow}(n) = (1/(1-F))^{\leftarrow}(n) \sim a_n$ and this determines a_n by the convergence to types theorem.(See Feller (1971), Resnick (1998, 1987).)

Conversely if (3.10) holds, define $a_n = U^{\leftarrow}(n)$ as previously. Then

$$\lim_{n\to\infty} \frac{1 - F(a_n x)}{1 - F(a_n)} = x^{-\alpha}$$

and we recover (3.9) provided $1 - F(a_n) \sim n^{-1}$ or what is the same provided $U(a_n) \sim n$ i.e., $U(U^{\leftarrow}(n)) \sim n$. Recall from (3.5), that $z < U^{\leftarrow}(n)$ iff $U(z) < n$ and setting $z = U^{\leftarrow}(n)(1-\varepsilon)$ and then $z = U^{\leftarrow}(n)(1+\varepsilon)$ we get

$$\frac{U(U^{\leftarrow}(n))}{U(U^{\leftarrow}(n)(1+\varepsilon))} \leq \frac{U(U^{\leftarrow}(n))}{n} \leq \frac{U(U^{\leftarrow}(n))}{U(U^{\leftarrow}(n)(1-\varepsilon))}.$$

Let $n \to \infty$, remembering $U = 1/(1-F) \in RV_\alpha$. Then

$$(1+\varepsilon)^{-\alpha} \leq \liminf_{n\to\infty} n^{-1}U(U^{\leftarrow}(n)) \leq \limsup_{n\to\infty} U(U^{\leftarrow}(n)) \leq (1-\varepsilon)^{-\alpha}$$

and since $\varepsilon > 0$ is arbitrary the desired result follows.

3.3. **Regular variation: Deeper Results; Karamata's Theorem.** There are several deeper results which give the theory power and utility: uniform convergence, Karamata's theorem which says a regularly varying function integrates the way you expect a power function to integrate, and finally the Karamata representation theorem.

3.3.1. *Uniform convergence.* The first useful result is the uniform convergence theorem.

Proposition 3. *If $U \in RV_\rho$ for $\rho \in \mathbb{R}$, then*

$$\lim_{t\to\infty} U(tx)/U(t) = x^\rho$$

locally uniformly in x on $(0,\infty)$. If $\rho < 0$, then uniform convergence holds on intervals of the form $(b,\infty), b > 0$. If $\rho > 0$ uniform convergence holds on intervals $(0,b]$ provided U is bounded on $(0,b]$ for all $b > 0$.

If U is monotone the result already follows from the discussion in Subsubsection 3.1.1, since we have a family of monotone functions converging to a continuous limit. For detailed discussion see Bingham et al. (1987), de Haan (1970), Geluk and de Haan (1987), Seneta (1976).

3.3.2. *Integration and Karamata's theorem.* The next set of results examines the integral properties of regularly varying functions. For purposes of integration, a ρ-varying function behaves roughly like x^ρ. We assume all functions are locally integrable and since we are interested in behavior at ∞ we assume integrability on intervals including 0 as well.

Theorem 1 (Karamata's Theorem)**.** *(a) Suppose $\rho \geq -1$ and $U \in RV_\rho$. Then $\int_0^x U(t)dt \in RV_{\rho+1}$ and*

$$\lim_{x\to\infty} \frac{xU(x)}{\int_0^x U(t)dt} = \rho + 1. \tag{3.11}$$

If $\rho < -1$ (or if $\rho = -1$ and $\int_x^\infty U(s)ds < \infty$) then $U \in RV_\rho$ implies $\int_x^\infty U(t)dt$ is finite, $\int_x^\infty U(t)dt \in RV_{\rho+1}$ and

$$\lim_{x\to\infty} \frac{xU(x)}{\int_x^\infty U(t)dt} = -\rho - 1. \tag{3.12}$$

(b) If U satisfies

$$\lim_{x\to\infty} \frac{xU(x)}{\int_0^x U(t)dt} = \lambda \in (0,\infty) \tag{3.13}$$

then $U \in RV_{\lambda-1}$. If $\int_x^\infty U(t)dt < \infty$ and

$$\lim_{x\to\infty} \frac{xU(x)}{\int_x^\infty U(t)dt} = \lambda \in (0,\infty) \tag{3.14}$$

then $U \in RV_{-\lambda-1}$.

What Theorem 1 emphasizes is that for the purposes of integration, the slowly varying function can be passed from inside to outside the integral. For example the way to remember and interpret (3.11) is to write $U(x) = x^\rho L(x)$ and then observe

$$\int_0^x U(t)dt = \int_0^x t^\rho L(t)dt$$

and pass the $L(t)$ in the integrand outside as a factor $L(x)$ to get

$$\begin{aligned} &\sim L(x)\int_0^x t^\rho dt = L(x)x^{\rho+1}/(\rho+1) \\ &= xx^\rho L(x)/(\rho+1) = xU(x)/(\rho+1), \end{aligned}$$

which is equivalent to the assertion (3.11).

Proof. (a). For certain values of ρ, uniform convergence suffices after writing say

$$\frac{\int_0^x U(s)ds}{xU(x)} = \int_0^x \frac{U(sx)}{U(x)}ds.$$

If we wish to proceed, using elementary concepts, consider the following approach, which follows de Haan (1970).

If $\rho > -1$ we show $\int_0^\infty U(t)dt = \infty$. From $U \in RV_\rho$ we have

$$\lim_{s\to\infty} U(2s)/U(s) = 2^\rho > 2^{-1}$$

since $\rho > -1$. Therefore there exists s_0 such that $s > s_0$ necessitates $U(2s) > 2^{-1}U(s)$. For n with $2^n > s_0$ we have

$$\int_{2^{n+1}}^{2^{n+2}} U(s)ds = 2\int_{2^n}^{2^{n+1}} U(2s)ds > \int_{2^n}^{2^{n+1}} U(s)ds$$

and so setting $n_0 = \inf\{n : 2^n > s_0\}$ gives

$$\int_{s_0}^\infty U(s)ds \geq \sum_{n:2^n>s_0} \int_{2^{n+1}}^{2^{n+2}} U(s)ds > \sum_{n\geq n_0} \int_{2^{n_0+1}}^{2^{n_0+2}} U(s)ds = \infty.$$

Thus for $\rho > -1, x > 0$, and any $N < \infty$ we have

$$\int_0^t U(sx)ds \sim \int_N^t U(sx)ds, t \to \infty,$$

since $U(sx)$ is a ρ-varying function of s. For fixed x and given ε, there exists N such that for $s > N$

$$(1-\varepsilon)x^\rho U(s) \le U(sx) \le (1+\varepsilon)x^\rho U(s)$$

and thus

$$\begin{aligned}\limsup_{t\to\infty} \frac{\int_0^{tx} U(s)ds}{\int_0^t U(s)ds} =& \limsup_{t\to\infty} \frac{x\int_0^t U(sx)ds}{\int_0^t U(s)ds}\\ =& \limsup_{t\to\infty} \frac{x\int_N^t U(sx)ds}{\int_N^t U(s)ds}\\ \le& \limsup_{t\to\infty} x^{\rho+1}(1+\varepsilon)\frac{\int_N^t U(s)ds}{\int_N^t U(s)ds}\\ =&(1+\varepsilon)x^{\rho+1}.\end{aligned}$$

An analogous argument applies for $\liminf$ and thus we have proved

$$\int_0^x U(s)ds \in RV_{\rho+1}$$

when $\rho > -1$.

In case $\rho = -1$ then either $\int_0^\infty U(s)ds < \infty$ in which case $\int_0^x U(s)ds \in RV_{-1+1} = RV_0$ or else $\int_0^\infty U(s)ds = \infty$ and the previous argument is applicable. So we have checked that for $\rho \ge -1$, $\int_0^x U(s)ds \in RV_{\rho+1}$.

We now focus on proving (3.11) when $U \in RV_\rho$, $\rho \ge -1$. As in the development leading to (3.22), set

$$b(x) = xU(x)/\int_0^x U(t)dt$$

so that integrating $b(x)/x$ leads to the representations

$$\begin{aligned}\int_0^x U(s)ds =& c\exp\left\{\int_1^x t^{-1}b(t)dt\right\}\\ U(x) =& cx^{-1}b(x)\exp\left\{\int_1^x t^{-1}b(t)dt\right\}.\end{aligned} \tag{3.15}$$

We must show $b(x) \to \rho+1$. Observe first that

$$\begin{aligned}\liminf_{x\to\infty} 1/b(x) =& \liminf_{x\to\infty} \frac{\int_0^x U(t)dt}{xU(x)}\\ =& \liminf_{x\to\infty} \int_0^1 \frac{U(sx)}{U(x)}ds.\end{aligned}$$

Now make a change of variable $s = x^{-1}t$ and and by Fatou's lemma this is

$$\ge \int_0^1 \liminf_{x\to\infty}(U(sx)/U(x))ds$$

$$= \int_0^1 s^\rho ds = \frac{1}{\rho+1}$$

and we conclude

$$\limsup_{x\to\infty} b(x) \le \rho + 1. \tag{3.16}$$

If $\rho = 1$ then $b(x) \to 0$ as desired, so now suppose $\rho > -1$.

We observe the following properties of $b(x)$:

(i) $b(x)$ is bounded on a semi-infinite neighborhood of ∞ (by (3.16)).
(ii) b is slowly varying since $xU(x) \in RV_{\rho+1}$ and $\int_0^x U(s)ds \in RV_{\rho+1}$.
(iii) We have

$$b(xt) - b(x) \to 0$$

boundedly as $x \to \infty$.

The last statement follows since by slow variation

$$\lim_{x\to\infty} (b(xt) - b(x))/b(x) = 0$$

and the denominator is ultimately bounded.

From (iii) and dominated convergence

$$\lim_{x\to\infty} \int_1^s t^{-1}(b(xt) - b(x))dt = 0$$

and the left side may be rewritten to obtain

$$\lim_{x\to\infty} \left\{ \int_1^s t^{-1} b(xt)dt - b(x)\log s \right\} = 0. \tag{3.17}$$

From (3.15)

$$c\exp\left\{ \int_1^x t^{-1} b(t)dt \right\} = \int_0^x U(s)ds \in RV_{\rho+1}$$

and from the regular variation property

$$\begin{aligned}(\rho+1)\log s &= \lim_{x\to\infty} \log\left\{ \frac{\int_0^{xs} U(t)dt}{\int_0^x U(t)dt} \right\} \\ &= \lim_{x\to\infty} \int_x^{xs} t^{-1} b(t)dt = \lim_{x\to\infty} \int_1^s t^{-1} b(xt)dt\end{aligned}$$

and combining this with (3.17) leads to the desired conclusion that $b(x) \to \rho + 1$.

(b). We suppose (3.13) holds and check $U \in RV_{\lambda-1}$. Set

$$b(x) = xU(x)/\int_0^x U(t)dt$$

so that $b(x) \to \lambda$. From (3.15)

$$U(x) = cx^{-1} b(x) \exp \int_1^x t^{-1} b(t)dt \Big\}$$

$$= cb(x)\exp\int_1^x t^{-1}(b(t)-1)dt\Big\}$$

and since $b(t)-1 \to \lambda - 1$ we see that U satisfies the representation of a $(\lambda-1)$ varying function. □

3.3.3. *Karamata's representation.* Theorem 1 leads in a straightforward way to what has been called the *Karamata representation* of a regularly varying function.

Corollary 1 (The Karamata Representation)**.** *(i) The function L is slowly varying iff L can be represented as*

$$L(x) = c(x)\exp\left\{\int_1^x t^{-1}\varepsilon(t)dt\right\}, \quad x > 0, \tag{3.18}$$

where $c : \mathbb{R}_+ \mapsto \mathbb{R}_+$, $\varepsilon : \mathbb{R}_+ \mapsto \mathbb{R}_+$ and

$$\lim_{x\to\infty} c(x) = c \in (0,\infty), \tag{3.19}$$

$$\lim_{t\to\infty} \varepsilon(t) = 0. \tag{3.20}$$

(ii) A function $U : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is regularly varying with index ρ iff U has the representation

$$U(x) = c(x)\exp\left\{\int_1^x t^{-1}\rho(t)dt\right\} \tag{3.21}$$

where $c(\cdot)$ satisfies (3.19) *and $\lim_{t\to\infty}\rho(t) = \rho$. (This is obtained from (i) by writing $U(x) = x^\rho L(x)$ and using the representation for L.)*

Proof. If L has a representation (3.18) then it must be slowly varying since for $x > 1$

$$\lim_{t\to\infty} L(tx)/L(t) = \lim_{t\to\infty}(c(tx)/c(t))\exp\left\{\int_t^{tx} s^{-1}\varepsilon(s)ds\right\}.$$

Given ε, there exists t_0 by (3.20) such that

$$-\varepsilon < \varepsilon(t) < \varepsilon, \quad t \geq t_0,$$

so that

$$-\varepsilon\log x = -\varepsilon\int_t^{tx} s^{-1}ds \leq \int_t^{tx} s^{-1}\varepsilon(s)ds \leq \varepsilon\int_t^{tx} s^{-1}ds = \varepsilon\log x.$$

Therefore $\lim_{t\to\infty}\int_t^{tx} s^{-1}\varepsilon(s)ds = 0$ and $\lim_{t\to\infty} L(tx)/L(t) = 1$.

Conversely suppose $L \in RV_0$. By Karamata's theorem

$$b(x) := xL(x)/\int_0^x L(s)ds \to 1$$

and $x \to \infty$. Note

$$L(x) = x^{-1}b(x)\int_0^x L(s)ds.$$

Set $\varepsilon(x) = b(x) - 1$ so $\varepsilon(x) \to 0$ and

$$\int_1^x t^{-1}\varepsilon(t)dt = \int_1^x \left(L(t)/\int_0^t L(s)ds\right)dt - \log x$$

$$= \int_1^x d\left(\log \int_0^t L(s)ds\right) - \log x$$
$$= \log\left(x^{-1}\int_0^x L(s)ds / \int_0^1 L(s)ds\right)$$

whence

$$\exp\left\{\int_1^x t^{-1}\varepsilon(t)dt\right\} = x^{-1}\int_0^x L(s)ds / \int_0^1 L(s)ds$$
$$= L(x) / \left(b(x)\int_0^1 L(s)ds\right), \tag{3.22}$$

and the representation follows with

$$c(x) = b(x)\int_0^1 L(s)ds.$$

□

3.3.4. *Differentiation.* The previous results describe the asymptotic properties of the indefinite integral of a regularly varying function. We now describe what happens when a ρ-varying function is differentiated.

Proposition 4. *Suppose $U : R_+ \mapsto R_+$ is absolutely continuous with density u so that*

$$U(x) = \int_0^x u(t)dt.$$

(a) (Von Mises) If

$$\lim_{x\to\infty} xu(x)/U(x) = \rho, \tag{3.23}$$

then $U \in RV_\rho$.

(b) (Landau, 1916) See also (de Haan, 1970, page 23, 109), Seneta (1976),Resnick (1987). If $U \in RV_\rho, \rho \in \mathbb{R}$, and u is monotone then (3.23) *holds and if $\rho \neq 0$ then $|u|(x) \in RV_{\rho-1}$.*

Proof. (a) Set

$$b(x) = xu(x)/U(x)$$

and as before we find

$$U(x) = U(1)\exp\left\{\int_1^x t^{-1}b(t)dt\right\}$$

so that U satisfies the representation theorem for a ρ-varying function.

(b) Suppose u is nondecreasing. An analogous proof works in the case u is nonincreasing. Let $0 < a < b$ and observe

$$(U(xb) - U(xa))/U(x) = \int_{xa}^{xb} u(y)dy/U(x).$$

By monotonicity we get

$$u(xb)x(b-a)/U(x) \geq (U(xb) - U(xa))/U(x) \geq u(xa)x(b-a)/U(x). \tag{3.24}$$

From (22) and the fact that $U \in RV_\rho$ we conclude

$$\limsup_{x\to\infty} xu(xa)/U(x) \le (b^\rho - a^\rho)/(b-a) \tag{3.25}$$

for any $b > a > 0$. So let $b \downarrow a$, which is tantamount to taking a derivative. Then (3.25) becomes

$$\limsup_{x\to\infty} xu(xa)/U(x) \le \rho a^{\rho-1} \tag{3.26}$$

for any $a > 0$. Similarly from the left-hand equality in (3.24) after letting $a \uparrow b$ we get

$$\liminf_{x\to\infty} xu(xb)/U(x) \ge \rho b^{\rho-1} \tag{3.27}$$

for any $b > 0$. Then (3.23) results by setting $a = 1$ in (3.26) and $b = 1$ in (3.27). □

3.4. **Regular variation: Further properties.** For the following list of properties, it is convenient to define *rapid variation* or regular variation with index ∞. Say $U : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is regularly varying with index ∞ ($U \in RV_\infty$) if for every $x > 0$

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\infty := \begin{cases} 0, & \text{if } x < 1, \\ 1, & \text{if } x = 1, \\ \infty, & \text{if } x > 1. \end{cases}$$

Similarly $U \in RV_{-\infty}$ if

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^{-\infty} := \begin{cases} \infty, & \text{if } x < 1, \\ 1, & \text{if } x = 1, \\ 0, & \text{if } x > 1. \end{cases}$$

The following proposition collects useful properties of regularly varying functions. (See de Haan (1970).)

Proposition 5. *(i) If $U \in RV_\rho$, $-\infty \le \rho \le \infty$, then*

$$\lim_{x\to\infty} \log U(x)/\log x = \rho$$

so that

$$\lim_{x\to\infty} U(x) = \begin{cases} 0, & \text{if } \rho < 0, \\ \infty, & \text{if } \rho > 0. \end{cases}$$

(ii) (Potter bounds.) Suppose $U \in RV_\rho$, $\rho \in \mathbb{R}$. Take $\varepsilon > 0$. Then there exists t_0 such that for $x \ge 1$ and $t \ge t_0$

$$(1-\varepsilon)x^{\rho-\varepsilon} < \frac{U(tx)}{U(t)} < (1+\varepsilon)x^{\rho+\varepsilon}. \tag{3.28}$$

(iii) If $U \in RV_\rho$, $\rho \in \mathbb{R}$, and $\{a_n\}$, $\{a'_n\}$ satisfy, $0 < a_n \to \infty$, $0 < a_{n'} \to \infty$, and $a_n \sim ca_{n'}$, for $0 < c < \infty$, then $U(a_n) \sim c^\rho U(a_{n'})$. If $\rho \ne 0$ the result also holds for $c = 0$ or ∞. Analogous results hold with sequences replaced by functions.

(iv) If $U_1 \in RV_{\rho_1}$ and $U_2 \in RV_{\rho_2}$ and $\lim_{x\to\infty} U_2(x) = \infty$ then

$$U_1 \circ U_2 \in RV_{\rho_1\rho_2}.$$

(v) Suppose U is nondecreasing, $U(\infty) = \infty$, and $U \in RV_\rho, 0 \le \rho \le \infty$. Then

$$U^{\leftarrow} \in RV_{\rho^{-1}}.$$

(vi) Suppose U_1, U_2 are nondecreasing and ρ-varying, $0 < \rho < \infty$. Then for $0 \le c \le \infty$

$$U_1(x) \sim cU_2(x), \quad x \to \infty$$

iff

$$U_1^{\leftarrow}(x) \sim c^{-\rho^{-1}} U_2^{\leftarrow}(x), \quad x \to \infty.$$

(vii) If $U \in RV_\rho$, $\rho \ne 0$, then there exists a function U^* which is absolutely continuous, strictly monotone, and

$$U(x) \sim U(x)^*, \quad x \to \infty.$$

Proof. (i) We give the proof for the case $0 < \rho < \infty$. Suppose U has Karamata representation

$$U(x) = c(x) \exp \int_1^x t^{-1} \rho(t) dt \Big\}$$

where $c(x) \to c > 0$ and $\rho(t) \to \rho$. Then

$$\log U(x)/\log x = o(1) + \int_1^x t^{-1}\rho(t)dt / \int_1^x t^{-1} dt \to \rho.$$

(ii) Using the Karamata representation

$$U(tx)/U(t) = (c(tx)/c(t)) \exp \left\{ \int_1^x s^{-1} \rho(ts) ds \right\}$$

and the result is apparent since we may pick t_0 so that $t > t_0$ implies $\rho - \varepsilon < \rho(ts) < \rho + \varepsilon$ for $s > 1$.

(iii) If $c > 0$ then from the uniform convergence property in Proposition 3

$$\lim_{n\to\infty} \frac{U(a_n)}{U(a_{n'})} = \lim_{n\to\infty} \frac{U(a_{n'}(a_n/a_{n'}))}{U(a_{n'})} = \lim_{t\to\infty} \frac{U(tc)}{U(t)} = c^\rho.$$

(iv) Again by uniform convergence, for $x > 0$

$$\begin{aligned} \lim_{t\to\infty} \frac{U_1(U_2(tx))}{U_1(U_2(t))} &= \lim_{t\to\infty} \frac{U_1(U_2(t)(U_2(tx)/U_2(t)))}{U_1(U_2(t))} \\ &= \lim_{y\to\infty} \frac{U_1(yx^{\rho_2})}{U_1(y)} = x^{\rho_2\rho_1}. \end{aligned}$$

(v) Let $U_t(x) = U(tx)/U(t)$ so that if $U \in RV_\rho$ and U is nondecreasing then $(0 < \rho < \infty)$

$$U_t(x) \to x^\rho, \quad t \to \infty,$$

which implies by Proposition 1

$$U_t^{\leftarrow}(x) \to x^{\rho^{-1}}, \quad t \to \infty;$$

that it,

$$\lim_{t\to\infty} U^{\leftarrow}(xU(t))/t = x^{\rho^{-1}}.$$

Therefore

$$\lim_{t\to\infty} U^{\leftarrow}(xU(U^{\leftarrow}(t)))/U^{\leftarrow}(t) = x^{\rho^{-1}}.$$

This limit holds locally uniformly since monotone functions are converging to a continuous limit. Now $U \circ U^{\leftarrow}(t) \sim t$ as $t \to \infty$, and if we replace x by $xt/U \circ U^{\leftarrow}(t)$ and use uniform convergence we get

$$\begin{aligned}\lim_{t\to\infty} \frac{U^{\leftarrow}(tx)}{U^{\leftarrow}(t)} &= \lim_{t\to\infty} \frac{U^{\leftarrow}((xt/U \circ U^{\leftarrow}(t))U \circ U^{\leftarrow}(t))}{U^{\leftarrow}(t)} \\ &= \lim_{t\to\infty} \frac{U^{\leftarrow}(xU \circ U^{\leftarrow}(t))}{U^{\leftarrow}(t)} = x^{\rho^{-1}}\end{aligned}$$

which makes $U^{\leftarrow} \in RV_{\rho^{-1}}$.

(vi) If $c > 0, 0 < \rho < \infty$ we have for $x > 0$

$$\lim_{t\to\infty} \frac{U_1(tx)}{U_2(t)} = \lim_{t\to\infty} \frac{U_1(tx)U_2(tx)}{U_2(tx)U_2(t)} = cx^{\rho}.$$

Inverting we find for $y > 0$

$$\lim_{t\to\infty} U_1^{\leftarrow}(yU_2(t))/t = (c^{-1}y)^{\rho^{-1}}$$

and so

$$\lim_{t\to\infty} U_1^{\leftarrow}(yU_2 \circ U_2^{\leftarrow}(t))/U_2^{\leftarrow}(t) = (c^{-1}y)^{\rho^{-1}}$$

and since $U_2 \circ U_2^{\leftarrow}(t) \sim t$

$$\lim_{t\to\infty} U_1^{\leftarrow}(yt)/U_2^{\leftarrow}(t) = (c^{-1}y)^{\rho^{-1}}.$$

Set $y = 1$ to obtain the result.

(vii) For instance if $U \in RV_{\rho}, \rho > 0$ define

$$U^*(t) = \int_1^t s^{-1}U(s)ds.$$

Then $s^{-1}U(s) \in RV_{\rho-1}$ and by Karamata's theorem

$$U(x)/U^*(x) \to \rho.$$

U^* is absolutely continuous and since $U(x) \to \infty$ when $\rho > 0, U^*$ is ultimately strictly increasing. □

References

P. Billingsley. *Probability and Measure.* Wiley, New York, 2nd edition, 1986.

N. Bingham, C. Goldie, and J. Teugels. *Regular Variation.* Cambridge University Press, 1987.

L. de Haan. *On Regular Variation and Its Application to the Weak Convergence of Sample Extremes.* Mathematisch Centrum Amsterdam, 1970.

W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, New York, 2nd edition, 1971.

J. L. Geluk and L. de Haan. *Regular Variation, Extensions and Tauberian Theorems*, volume 40 of *CWI Tract*. Stichting Mathematisch Centrum, Centrum voor Wiskunde en Informatica, Amsterdam, 1987. ISBN 90-6196-324-9.

S. Resnick. *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag, New York, 1987.

S. Resnick. *A Probability Path*. Birkhäuser, Boston, 1998.

E. Seneta. *Regularly Varying Functions*. Springer-Verlag, New York, 1976. Lecture Notes in Mathematics, 508.

4. A Crash Course in Weak Convergence.

Many asymptotic properties of statistics in heavy tailed analysis are clearly understood with a fairly high level interpretation which comes from the modern theory of weak convergence of probability measures on metric spaces as originally promoted in Billingsley (1968) and updated in Billingsley (1999).

4.1. **Definitions.** Let $\mathbb{S}$ be a complete, separable metric space with metric d and let $\mathcal{S}$ be the Borel σ- algebra of subsets of $\mathbb{S}$ generated by open sets. Suppose $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space. A random element X in $\mathbb{S}$ is a measurable map from such a space $(\Omega, \mathcal{A})$ into $(\mathbb{S}, \mathcal{S})$.

With a random variable, a point $\omega \in \Omega$ is mapped into a real valued member of $\mathbb{R}$. With a random element, a point $\omega \in \Omega$ is mapped into a an element of the metric space $\mathbb{S}$. Here are some common examples of this paradigm.

Metric space $\mathbb{S}$	**Random element X is a:**
$\mathbb{R}$	random variable
$\mathbb{R}^d$	random vector
$C[0,\infty)$, the space of real valued, continuous functions on $[0,\infty)$	random process with continuous paths
$D[0,\infty)$, the space of real valued, right continuous functions on $[0,\infty)$ with finite left limits existing on $(0,\infty)$	right continuous random process with jump discontinuities
$M_p(\mathbb{E})$, the space of point measures on a nice space $\mathbb{E}$	stochastic point process on $\mathbb{E}$
$M_+(\mathbb{E})$, the space of Radon measures on a nice space $\mathbb{E}$.	random measure on $\mathbb{E}$

Table 1. Various metric spaces and random elements.

Given a sequence $\{X_n, n \geq 0\}$ of random elements of $\mathbb{S}$, there is a corresponding sequence of distributions on $\mathcal{S}$,

$$P_n = \mathbb{P} \circ X_n^{-1} = \mathbb{P}[X_n \in \cdot], \quad n \geq 0.$$

P_n is called the *distribution* of X_n. Then X_n converges weakly to X_0 (written $X_n \Rightarrow X_0$ or $P_n \Rightarrow P_0$) if whenever $f \in C(\mathbb{S})$, the class of bounded, continuous real valued functions on $\mathbb{S}$, we have

$$\mathbf{E}f(X_n) = \int_{\mathbb{S}} f(x) P_n(dx) \to \mathbf{E}f(X_0) = \int_{\mathbb{S}} f(x) P_0(dx).$$

Recall that the definition of weak convergence of random variables in $\mathbb{R}$ is given in terms of one dimensional distribution functions which does not generalize nicely to higher dimensions. The definition in terms of integrals of test functions $f \in C(\mathbb{S})$ is very flexible and well defined for any metric space $\mathbb{S}$.

4.2. **Basic properties of weak convergence.**

4.2.1. *Portmanteau Theorem.* The basic *Portmanteau Theorem* ((Billingsley, 1968, page 11), Billingsley (1999)) says the following are equivalent:

$$X_n \Rightarrow X_0. \tag{4.1}$$

$$\lim_{n\to\infty} \mathbb{P}[X_n \in A] = \mathbb{P}[X_0 \in A],\ \forall A \in \mathcal{S} \text{ such that } \mathbb{P}[X_0 \in \partial A] = 0. \tag{4.2}$$

Here ∂A denotes the boundary of the set A.

$$\limsup_{n\to\infty} \mathbb{P}[X_n \in F] \le \mathbb{P}[X_0 \in F],\ \forall \text{ closed } F \in \mathcal{S}. \tag{4.3}$$

$$\liminf_{n\to\infty} \mathbb{P}[X_n \in G] \ge \mathbb{P}[X_0 \in G],\ \forall \text{ open } G \in \mathcal{S}. \tag{4.4}$$

$$\mathbf{E}f(X_n) \to \mathbf{E}f(X_0), \quad \text{for all } f \text{ which are bounded and } \underline{\text{uniformly}} \text{ continuous.} \tag{4.5}$$

Although it may seem comfortable to express weak convergence of probability measures in terms of sets, it is mathematically simplist to rely on integrals with respect to test functions as given, for instance, in (4.5).

4.2.2. *Skorohod's theorem.* A nice way to think about weak convergence is using *Skorohod's theorem* ((Billingsley, 1971, Proposition 0.2)) which, for certain purposes, allows one to replace convergence in distribution with almost sure convergence. In a theory which relies heavily on continuity, this is a big advantage. Almost sure convergence, being pointwise, is very well suited to continuity arguments.

Let $\{X_n, n \ge 0\}$ be random elements of the metric space $(\mathbb{S}, \mathcal{S})$ and suppose the domain of each X_n is $(\Omega, \mathcal{A}, \mathbb{P})$. Let

$$([0,1], \mathcal{B}[0,1], \mathbb{LEB}(\cdot))$$

be the usual probability space on $[0,1]$, where $\mathbb{LEB}(\cdot)$ is Lebesgue measure or length. Skorohod's theorem says that $X_n \Rightarrow X_0$ iff there exists random elements $\{X_n^*, n \ge 0\}$ in $\mathbb{S}$ defined on the uniform probability space such that

$$X_n \stackrel{d}{=} X_n^*, \quad \text{for each } n \ge 0,$$

and

$$X_n^* \to X_0^* \text{ a.s.}$$

The second statement means

$$\mathbb{LEB}\left\{t \in [0,1] : \lim_{n\to\infty} d(X_n^*(t), X_0^*(t)) = 0\right\} = 1.$$

Almost sure convergence always implies convergence in distribution so Skorohod's theorem provides a partial converse. To see why almost sure convergence implies weak convergence is easy. With $d(\cdot,\cdot)$ as the metric on $\mathbb{S}$ we have $d(X_n, X_0) \to 0$ almost surely and for any $f \in C(\mathbb{S})$ we get by continuity that $f(X_n) \to f(X_0)$, almost surely. Since f is bounded, by dominated convergence we get $\mathbf{E}f(X_n) \to \mathbf{E}f(X_0)$.

Recall that in one dimension, Skorohod's theorem has an easy proof. If $X_n \Rightarrow X_0$ and X_n has distribution function F_n then

$$F_n \to F_0, \quad n \to \infty.$$

Thus, by Proposition 1, $F_n^{\leftarrow} \to F_0^{\leftarrow}$. Then with U, the identity function on $[0,1]$, (so that U is uniformly distributed)

$$X_n \stackrel{d}{=} F_n^{\leftarrow}(U) =: X_n^*, \quad n \geq 0,$$

and

$$\begin{aligned}\mathbb{LEB}[X_n^* \to X_0^*] =& \mathbb{LEB}\{t \in [0,1] : F_n^{\leftarrow}(t) \to F_n^{\leftarrow}(t)] \\ \geq& \mathbb{LEB}\big(\mathcal{C}(F_0^{\leftarrow})\big) = 1,\end{aligned}$$

since the set of discontinuities of the monotone function $F_0^{\leftarrow}(\cdot)$ is countable, and hence has Lebesgue measure 0.

The power of weak convergence theory comes from the fact that once a basic convergence result has been proved, many corollaries emerge with little effort, often using only continuity. Suppose $(\mathbb{S}_i, d_i), i = 1, 2$, are two metric spaces and $h : \mathbb{S}_1 \mapsto \mathbb{S}_2$ is continuous. If $\{X_n, n \geq 0\}$ are random elements in $(\mathbb{S}_1, \mathcal{S}_1)$ and $X_n \Rightarrow X_0$ then $h(X_n) \Rightarrow h(X_0)$ as random elements in $(\mathbb{S}_2, \mathcal{S}_2)$.

To check this is easy: Let $f_2 \in C(\mathbb{S}_2)$ and we must show that $\mathbf{E}f_2(h(X_n)) \to \mathbf{E}f_2(h(X_0))$. But $f_2(h(X_n)) = f_2 \circ h(X_n)$ and since $f_2 \circ h \in C(\mathbb{S}_1)$, the result follows from the definition of $X_n \Rightarrow X_0$ in $\mathbb{S}_1$.

If $\{X_n\}$ are random variables which converge, then letting $h(x) = x^2$ or $\arctan x$ or ... yields additional convergences for free.

4.2.3. *Continuous mapping theorem.* In fact, the function h used in the previous paragraphs, need not be continuous everywhere and in fact, many of the maps h we will wish to use are definitely not continuous everywhere.

Theorem 2 (Continuous Mapping Theorem.)**.** *Let $(\mathbb{S}_i, d_i), i = 1, 2$, be two metric spaces and suppose $\{X_n, n \geq 0\}$ are random elements of $(\mathbb{S}_1, \mathcal{S}_1)$ and $X_n \Rightarrow X_0$. For a function $h : S_1 \mapsto S_2$, define the discontinuity set of h as*

$$D_h := \{s_1 \in S_1 : h \textit{ is discontinuous at } s_1\}.$$

If h satisfies

$$\mathbb{P}[X_0 \in D_h] = \mathbb{P}[X_0 \in \{s_1 \in S_1 : h \textit{ is discontinuous at } s_1\}] = 0$$

then

$$h(X_n) \Rightarrow h(X_0)$$

in $\mathbb{S}_2$.

Proof. For a traditional proof, see (Billingsley, 1968, page 30). This result is an immediate consequence of Skorohod's theorem. If $X_n \Rightarrow X_0$ then there exist almost surely convergent random elements of $\mathbb{S}_1$ defined on the unit interval, denoted X_n^*, such that

$$X_n^* \stackrel{d}{=} X_n, \quad n \geq 0.$$

Then it follows that

$$\mathbb{LEB}[h(X_n^*) \to h(X_0^*)] \geq \mathbb{LEB}[X_0^* \notin D_h]$$

where we denote by disc(h) the discontinuity set of h; that is, the complement of $\mathcal{C}(h)$. Since $X_0 \stackrel{d}{=} X_0^*$ we get the previous probability equal to

$$=\mathbb{P}[X_0 \notin \text{ disc}(h)] = 1$$

and therefore $h(X_n^*) \to h(X_0^*)$ almost surely. Since almost sure convergence implies convergence in distribution $h(X_n^*) \Rightarrow h(X_0^*)$. Since $h(X_n) \stackrel{d}{=} h(X_n^*),\ n \geq 0$, the result follows. □

4.2.4. *Subsequences and Prohorov's theorem.* Often to prove weak convergence, subsequence arguments are used and the following is useful. A family Π of probability measures on a complete, separable metric space is *relatively compact* if every sequence $\{P_n\} \subset \Pi$ contains a weakly convergent subsequence. Relative compactness is theoretically useful but hard to check in practice so we need a workable criterion. Call the family Π tight (and by abuse of language we will refer to the corresponding random elements also as a tight family) if for any ε, there exists a compact $K_\varepsilon \in \mathcal{S}$ such that

$$P(K_\varepsilon) > 1 - \varepsilon, \quad \text{for all } P \in \Pi.$$

This is the sort of condition that precludes probability mass from escaping from the state space. Prohorov's theorem (Billingsley (1968)) assures us that when S is separable and complete, tightness of Π is the same as relative compactness. Tightness is checkable although it is seldom easy.

4.3. **Some useful metric spaces.** It pays to spend a bit of time remembering details of examples of metric spaces that will be useful. To standardize notation we set

$$\begin{aligned}\mathcal{F}(\mathbb{S}) &= \text{ closed subsets of } \mathbb{S},\\ \mathcal{G}(\mathbb{S}) &= \text{ open subsets of } \mathbb{S},\\ \mathcal{K}(\mathbb{S}) &= \text{ compact subsets of } \mathbb{S}.\end{aligned}$$

4.3.1. $\mathbb{R}^d$*, finite dimensional Euclidean space.* We set

$$\mathbb{R}^d := \{(x_1, \ldots, x_d) :\ x_i \in \mathbb{R},\ i = 1, \ldots, d\} = \mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}.$$

The metric is defined by

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2},$$

for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. Convergence of a sequence in this space is equivalent to componentwise convergence.

Define an interval

$$(\boldsymbol{a}, \boldsymbol{b}] = \{\boldsymbol{x} \in \mathbb{R}^d : a_i < x_i \leq b_i,\ i = 1, \ldots, d\}$$

A probability measure P on $\mathbb{R}^d$ is determined by its distribution function

$$F(\boldsymbol{x}) := P(-\boldsymbol{\infty}, \boldsymbol{x}]$$

and a sequence of probability measures $\{P_n, n \geq 0\}$ on $\mathbb{R}^d$ converges to P_0 iff

$$F_n(\boldsymbol{x}) \to F_0(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathcal{C}(F_0).$$

Note this says that a sequence of random vectors converges in distribution iff their distribution functions converge weakly. While this is concrete, it is seldom useful since multivariate distribution functions are usually awkward to deal with in practice.

Also, recall $K \in \mathcal{K}(\mathbb{R}^d)$ iff K is closed and bounded.

4.3.2. $\mathbb{R}^\infty$, *sequence space.* Define

$$\mathbb{R}^\infty := \{(x_1, x_2, \dots) : \ x_i \in \mathbb{R}, \ i \geq 1\} = \mathbb{R} \times \mathbb{R} \times \dots.$$

The metric can be defined by

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{\infty} (|x_i - y_i| \wedge 1) 2^{-i},$$

for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. This gives a complete, separable metric space where convergence of a family of sequences means coordinatewise convergence which means

$$\boldsymbol{x}(n) \to \boldsymbol{x}(0) \text{ iff } x_i(n) \to x_i(0), \ \forall i \geq 1.$$

The topology $\mathcal{G}(\mathbb{R}^\infty)$ can be generated by basic neighborhoods of the form

$$N_k(\boldsymbol{x}) = \{\boldsymbol{y} : \bigvee_{i=1}^{d} |x_i - y_i| < \epsilon\},$$

as we vary d, the center $\boldsymbol{x}$ and ϵ.

A set $A \subset \mathbb{R}^\infty$ is relatively compact iff every one-dimensional section is bounded; that is iff for any $i \geq 1$

$$\{x_i : \boldsymbol{x} \in A\} \text{ is bounded.}$$

4.3.3. $C[0,1]$ *and* $C[0,\infty)$, *continuous functions.* The metric on $C[0,M]$, the space of real valued continuous functions with domain $[0,M]$ is the uniform metric

$$d_M(x(\cdot), y(\cdot)) = \sup_{0 \leq t \leq M} |x(t) - y(t)| =: \|x(\cdot) - y(\cdot)\|_M.$$

and the metric on $C[0,\infty)$ is

$$d(x(\cdot), y(\cdot)) = \sum_{n=1}^{\infty} \frac{d_n(x,y) \wedge 1}{2^n}$$

where we interpret $d_n(x,y)$ as the $C[0,n]$ distance of x and y restricted to $[0,n]$. The metric on $C[0,\infty)$ induces the topology of local uniform convergence.

For $C[0,1]$ (or $C[0,M]$), we have that every function is uniformly continuous since a continuous function on a compact set is always uniformly continuous. Uniform continuity can be expressed by the modulus of continuity which is defined for $x \in C[0,1]$ by

$$\omega_x(\delta) = \sup_{|t-s|<\delta} |x(t) - x(s)|, \quad 0 < \delta < 1.$$

Then, uniform continuity means

$$\lim_{\delta \to 0} \omega_x(\delta) = 0.$$

The Arzela-Ascoli theorem says a uniformly bounded equicontinuous family of functions in $C[0,1]$ has a uniformly convergent subsequence; that is, this family is relatively compact or has compact closure. Thus a set $A \subset C[0,1]$ is relatively compact iff

(i) A is uniformly bounded; that is,

$$\sup_{0\le t\le 1}\sup_{x\in K}|x(t)| < \infty, \tag{4.6}$$

and

(ii) A is equicontinuous; that is

$$\lim_{\delta\downarrow 0}\sup_{x\in K}\omega_x(\delta) = 0.$$

Since the functions in a compact family vary in a controlled way, (4.6) can be replaced by

$$\sup_{x\in K}|x(0)| < \infty. \tag{4.7}$$

Compare this result with the compactness characterization in $\mathbb{R}^\infty$ where compactness meant each one-dimensional section was compact. Here, a continuous function is compact if each one dimensional section is compact in a uniform way AND equicontinuity is present.

4.3.4. *$D[0,1]$ and $D[0,\infty)$.* Start by considering $D[0,1]$, the space of right continuous functions on $[0,1)$ which have finite left limits on $(0,1]$. Minor changes allow us to consider $D[0,M]$ for any $M>0$.

In the uniform topology, two functions $x(\cdot)$ and $y(\cdot)$ are close if their graphs are uniformly close. In the Skorohod topology on $D[0,1]$, we consider x and y close if after deforming the time scale of one of them, say y, the resulting graphs are close. Consider the following simple example:

$$x_n(t) = 1_{[0,\frac{1}{2}+\frac{1}{n}]}(t), \quad x(t) = 1_{[0,\frac{1}{2}]}(t). \tag{4.8}$$

The uniform distance is always 1 but a time deformation allows us to consider the functions to be close. (Various metrics and their applications to functions with jumps are considered in detail in Whitt (2002).)

Define time deformations

$$\begin{aligned}\Lambda = \{\lambda :&[0,1]\mapsto[0,1] : \lambda(0)=0,\ \lambda(1)=1,\\ &\lambda(\cdot) \text{ is continuous, strictly increasing, 1-1, onto.}\}\end{aligned} \tag{4.9}$$

Let $e(t)\in\Lambda$ be the identity transformation and denote the uniform distance between x,y as

$$\|x-y\| := \bigvee_{t=0}^{1}|x(t)-y(t)|.$$

The Skorohod metric $d(x,y)$ between two functions $x,y\in D[0,1]$ is

$$\begin{aligned} d(x,y) =& \inf\{\epsilon>0 : \exists\lambda\in\Lambda, \text{ such that } \|\lambda-e\|\vee\|x-y\circ\lambda\|\le\epsilon\},\\ =& \inf_{\lambda\in\Lambda}\|\lambda-e\|\vee\|x-y\circ\lambda\|.\end{aligned}$$

Simple consequences of the definitions:

(1) Given a sequence $\{x_n\}$ of functions in $D[0,1]$, we have $d(x_n, x_0) \to 0$, iff there exist $\lambda_n \in \Lambda$ and

$$\|\lambda - e\| \to 0, \quad \|x_n \circ \lambda_n - x_0\| \to 0. \tag{4.10}$$

(2) From the definition, we always have

$$d(x,y) \le \|x - y\|, \quad x, y \in D[0,1]$$

since one choice of λ is the identity but this may not give the infimum. Therefore, uniform convergence always implies Skorohod convergence. The converse is very false; see (4.8).

(3) If $d(x_n, x_0) \to 0$ for $x_n, \in D[0,1]$, $n \ge 0$, then for all $t \in \mathcal{C}(x_0)$, we have pointwise convergence

$$x_n(t) \to x_0(t).$$

To see this suppose (4.10) holds. Then

$$\|\lambda_n - e\| = \|\lambda_n^{\leftarrow} - e\| \to 0.$$

Thus

$$\begin{aligned} |x_n(t) - x_0(t)| \le & |x_n(t) - x_0 \circ \lambda_n^{\leftarrow}(t)| + |x_0 \circ \lambda_n^{\leftarrow}(t) - x_0(t)| \\ \le & \|x_n \circ \lambda_n - x_0\| + o(1), \end{aligned}$$

since x is continuous at t and $\lambda_n^{\leftarrow} \to e$.

(4) If $d(x_n, x_0) \to 0$ and $x \in C[0,1]$, then uniform convergence holds.
If (4.10) holds then as in item 3 we have for each $t \in [0,1]$

$$|x_n(t) - x_0(t)| \le \|x_n \circ \lambda_n - x_0\| + \|x_0 - x_0 \circ \lambda_n\| \to 0$$

and hence

$$\|x_n(t) - x_0(t)\| \to 0.$$

The space $D[0,\infty)$. Now we extend this metric to $D[0,\infty)$. For a function $x \in D[0,\infty)$ write

$$r_s x(t) = x(t), \quad 0 \le t \le s,$$

for the restriction of x to the interval $[0,s]$ and write

$$\|x\|_s = \bigvee_{t=0}^{s} |x(t)|.$$

Let d_s be the Skorohod metric on $D[0,s]$ and define d_∞, the Skorohod metric on $D[0,\infty)$ by

$$d_\infty(x,y) = \int_0^\infty e^{-s}\Big(d_s(r_s x, r_s y) \wedge 1\Big) ds.$$

The impact of this is that Skorohod convergence on $D[0,\infty)$ reduces to convergence on finite intervals since $d_\infty(x_n, x_0) \to 0$ iff for any $s \in \mathcal{C}(x_0)$ we have $d_s(r_s x_n, r_s x_0) \to 0$.

4.3.5. *Radon measures and point measures; vague convergence.* Suppose $\mathbb{E}$ is a nice space. The technical meaning of *nice* is that $\mathbb{E}$ should be a locally compact topological space with countable base but it is safe to think of $\mathbb{E}$ as a finite dimensional Euclidean space and we may think of $\mathbb{E}$ as a subset of $\mathbb{R}^d$. The case $d = 1$ is important but $d > 1$ is also very useful. When it comes time to construct point processes, $\mathbb{E}$ will be the space where our points live. We assume $\mathbb{E}$ comes with a σ-field $\mathcal{E}$ which can be the σ-field generated by the open sets or, equivalently, the rectangles of E. So the important sets in $\mathcal{E}$ are built up from rectangles.

How can we model a random distribution of points in $\mathbb{E}$? One way is to specify random elements in $\mathbb{E}$, say $\{X_n\}$, and then to say that a stochastic point process is the counting function whose value at the region $A \in \mathcal{E}$ is the number of random elements $\{X_n\}$ which fall in A. This is intuitively appealing but has some technical drawbacks and it mathematically preferable to focus on counting functions rather than points.

A measure $\mu : \mathcal{E} \mapsto [0, \infty]$ is an assignment of positive numbers to sets in $\mathcal{E}$ such that

(1) $\mu(\emptyset) = 0$ and $\mu(A) \geq 0$, for all $A \in \mathcal{E}$,
(2) If $\{A_n, n \geq 1\}$ are mutually disjoint sets in $\mathcal{E}$, then the σ-additivity property holds

$$\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i).$$

The measure μ is called *Radon*, if

$$\mu(K) < \infty, \quad \forall K \in \mathcal{K}(\mathbb{E});$$

so that compact sets are known to have finite μ-mass. Knowing where the measure is required to be finite helps us to keep track of infinities in a useful way and prevents illegal operations like $\infty - \infty$.

Define

$$M_+(\mathbb{E}) = \{\mu : \mu \text{ is a non-negative measure on } \mathcal{E} \text{and } \mu \text{ is Radon.}\} \tag{4.11}$$

The space $M_+(\mathbb{E})$ can be made into a complete separable metric space under what is called the *vague* metric. For now, instead of describing the metric, we will describe the notion of convergence the metric engenders.

The way we defined convergence of *probability* measures was by means of test functions. We integrate a test function which is bounded and continuous on the metric space and if the resulting sequence of numbers converges, then we have weak convergence. However, with infinite measures in $M_+(\mathbb{E})$, we cannot just integrate a bounded function to get something finite. However, we know our measures are also Radon and this suggests using functions which vanish off of compact sets. So define

$$C_K^+(\mathbb{E}) = \{f : \mathbb{E} \mapsto \mathbb{R}_+ : f \text{ is continuous with compact support.}\}$$

For a function to have compact support means that it vanishes off a compact set.

The notion of convergence in $M_+(\mathbb{E})$: If $\mu_n \in M_+(\mathbb{E})$, for $n \geq 0$, then μ_n converges vaguely to μ_0, written $\mu_n \xrightarrow{v} \mu_0$, provided for all $f \in C_K^+(\mathbb{E})$ we have

$$\mu_n(f) := \int_{\mathbb{E}} f(x)\mu_n(dx) \to \mu_0(f) := \int_{\mathbb{E}} f(x)\mu_0(dx),$$

as $n \to \infty$.

Example 3 (Trivial but mildly illuminating example)**.** Suppose $\mathbb{E}$ is some finite dimensional Euclidean space and define for $\boldsymbol{x} \in \mathbb{E}$, and $A \in \mathcal{E}$

$$\epsilon_{\boldsymbol{x}}(A) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{if } x \in A^c. \end{cases}$$

Then

$$\mu_n := \epsilon_{\boldsymbol{x}_n} \xrightarrow{v} \mu_n := \epsilon_{\boldsymbol{x}_0},$$

in $M_+(\mathbb{E})$, iff

$$\boldsymbol{x}_n \to \boldsymbol{x}_0$$

in the metric on $\mathbb{E}$.

To see this, suppose that $\boldsymbol{x}_n \to \boldsymbol{x}_0$ and $f \in C_K^+(\mathbb{E})$. Then

$$\mu_n(f) = f(\boldsymbol{x}_n) \to f(\boldsymbol{x}_0) = \mu_0(f),$$

since f is continuous and the points are converging. Conversely, suppose $\boldsymbol{x}_n \not\to \boldsymbol{x}_0$. Define $\phi : \mathbb{R} \mapsto [0, 1]$ by

$$\phi(t) = \begin{cases} 1, & \text{if } t < 0, \\ 1 - t, & \text{if } 0 \le t \le 1, \\ 0, & \text{if } t > 1. \end{cases}$$

There exists $\{n'\}$ such that $d(\boldsymbol{x}_{n'}, \boldsymbol{x}_0) > \epsilon$. Define

$$f(\boldsymbol{y}) = \phi(d(\boldsymbol{x}_0, \boldsymbol{y})/\epsilon)$$

so that $f \in C_K(\mathbb{E})$. Then

$$|f(\boldsymbol{x}_{n'}) - f(\boldsymbol{x}_0)| = |0 - 1| \not\to 0$$

and then we have $\mu_n(f) \not\to \mu_0(f)$.

Point measures. A point measure m is an element of $M_+(\mathbb{E})$ of the form

$$m = \sum_i \epsilon_{x_i}. \tag{4.12}$$

Built into this definition is the understanding that $m(\cdot)$ is Radon: $m(K) < \infty$, for $K \in \mathcal{K}(\mathbb{E})$. Think of $\{x_i\}$ as the atoms and m as the function which counts how many atoms fall in a set. The set $M_p(\mathbb{E})$ is the set of all Radon point measures satisfying (4.12). This turns out to be a closed subset of $M_+(\mathbb{E})$.

More on $M_+(\mathbb{E})$ (and hence, more on $M_p(\mathbb{E})$): THE VAGUE TOPOLOGY ON $M_+(\mathbb{E})$, OPEN SETS: We can specify open sets, a topology (a system of open sets satisfying closure properties) and then a notion of "distance" in $M_+(\mathbb{E})$. Define a *basis* set to be a subset of $M_+(\mathbb{E})$ of the form

$$\{\mu \in M_+(\mathbb{E}) : \mu(f_i) \in (a_i, b_i), i = 1, \ldots, k\} \tag{4.13}$$

where $f_i \in C_K^+(\mathbb{E})$ and $0 \le a_i \le b_i$. Now imagine varying the choices of integer k, functions $f_1, \ldots, f_k$, and endpoints $a_1, \ldots, a_k; b_1, \ldots, b_k$. Unions of basis sets form the class of open sets constituting the vague topology.

The topology is metrizable as a complete, separable metric space and we can put a metric $d(\cdot,\cdot)$ on the space which yields the same open sets. The metric "d" can be specified as follows: There is some sequence of functions $f_i \in C_K^+(\mathbb{E})$ and for $\mu_1, \mu_2 \in M_+(\mathbb{E})$

$$d(\mu_1,\mu_2) = \sum_{i=1}^{\infty} \frac{|\mu_1(f_i) - \mu_2(f_i)| \wedge 1}{2^i}. \tag{4.14}$$

An interpretation: If $\mu \in M_+(\mathbb{E})$, μ is determined by knowledge of $\{\mu(f), f \in C_K^+(\mathbb{E}))$. This may seem reasonable and we will see why this is true shortly. Think of μ as an object with components $\{\mu(f), f \in C_K^+(\mathbb{E})$. Think of $\mu(f)$ as the fth-component of μ. Then (4.14) indicates, in fact, it is enough to have a countable set of components to determine μ and we can think about μ being represented

$$\mu = \{\mu(f_i), i \geq 1\}. \tag{4.15}$$

So there is an analogy with $\mathbb{R}^\infty$.

This analogy makes plausible the following characterization of compactness: A subset $M \subset M_+(\mathbb{E})$ is vaguely relatively compact iff

$$\sup_{\mu \epsilon M} \mu(f) < \infty, \quad \forall f \in C_K^+(\mathbb{E}). \tag{4.16}$$

To show compactness implies (4.16) is easy and helps digest the concepts. Suppose M is relatively compact. For $f \in C_K^+(\mathbb{E})$, define the projection onto the f-th component $T_f : M_+(\mathbb{E}) \mapsto [0,\infty)$ by

$$T_f(\mu) = \mu(f).$$

Then T_f is continuous since $\mu_n \xrightarrow{v} \mu$ implies

$$T_f(\mu_n) = \mu_n(f) \to \mu(f) = T_f(\mu).$$

For fixed $f \in C_K^+(\mathbb{E})$, we note

$$\sup_{\mu \in M} \mu(f) = \sup_{\mu \in M} T_f(\mu) = \sup_{\mu \in M^-} T_f(\mu),$$

since the supremum of a continuous function on M must be the same as the supremum on the closure M^-.

If M is relatively compact, then the closure M^- is compact. Since T_f is continuous on $M_+(\mathbb{E})$, $T_f(M^-)$ is a compact subset of $[0,\infty)$. (Continuous images of compact sets are compact.) Compact sets in $[0,\infty)$ are bounded so

$$\infty > \sup T_f(M^-) = \sup\{T_f(\mu), \mu \in M^-\} = \sup_{\mu \in M^-} \{\mu(f)\}.$$

Why emphasize integrals of test functions rather than measures of sets? Proofs are a bit easier with this formulation and it is easier to capitilize on continuity arguments. One can always formulate parallel definitions and concepts with sets using a variant of Urysohn's lemma. See (Dudley, 1989, page 47), (Simmons, 1963, page 135), Kallenberg (1983), (Resnick, 1987, page 141).

Lemma 1. *(a) Suppose $K \in \mathcal{K}(\mathbb{E})$. There exists $K_n \in \mathcal{K}(\mathbb{E})$, $K_n \downarrow K$ and there exist $f_n \in C_K^+(\mathbb{E})$, with $\{f_n\}$ non-increasing such that*

$$1_K \le f_n \le 1_{K_n} \downarrow 1_K. \tag{4.17}$$

(b) Suppose $G \in \mathcal{G}(\mathbb{E})$, and G is relatively compact. There exist open, relatively compact $G_n \uparrow G$ and $f_n \in C_K^+(\mathbb{E})$, with $\{f_n\}$ non-decreasing such that

$$1_G \ge f_n \ge 1_{G_n} \uparrow 1_G \tag{4.18}$$

From Lemma 1, comes a Portmanteau Theorem.

Theorem 3. *Let $\mu, \mu_n \in M_+(\mathbb{E})$. The following are equivalent.*

(i) $\mu_n \xrightarrow{v} \mu$.
(ii) $\mu_n(B) \to \mu(B)$ *for all relatively compact B satisfying $\mu(\partial B) = 0$.*
(iii) *For all $K \in \mathcal{K}(\mathbb{E})$ we have*

$$\limsup_{n\to\infty} \mu_n(K) \le \mu(K)$$

and for all $G \in \mathcal{G}$ which are relatively compact, we have

$$\liminf_{n\to\infty} \mu_n(G) \ge \mu(G).$$

5. Vague Convergence, Regular Variation and the Multivariate Case.

Regular variation of distribution tails can be reformulated in terms of vague convergence and with this reformulation, the generalization to higher dimensions is effortless. We begin by discussing the reformulation in one dimension.

5.0.6. *Vague convergence on $(0, \infty]$.*

Theorem 4. *Suppose X_1 is a non-negative random variable with distribution function $F(x)$. Set $\bar{F} = 1 - F$. The following are equivalent:*

(i) $\bar{F} \in RV_{-\alpha}$, $\alpha > 0$.
(ii) *There exists a sequence $\{b_n\}$, with $b_n \to \infty$ such that*

$$\lim_{n\to\infty} n\bar{F}(b_n x) = x^{-\alpha}, \quad x > 0.$$

(iii) *There exists a sequence $\{b_n\}$ with $b_n \to \infty$ such that*

$$\nu_n(\cdot) := nP\left[\frac{X_1}{b_n} \in \cdot\right] \xrightarrow{v} \nu(\cdot) \tag{5.1}$$

in $M_+\big((0,\infty]\big)$, where $\nu(x, \infty] = x^{-\alpha}$.

Remark 2. (a) If any of (i), (ii) or (iii) is true we may always define

$$b(t) = \left(\frac{1}{1-F}\right)^{\leftarrow}(t) \tag{5.2}$$

and set $b_n = b(n)$. Note if (i) holds, then

$$\bar{F} \in RV_{-\alpha} \text{ implies } \frac{1}{1-F} \in RV_\alpha \text{ implies } b(\cdot) = \left(\frac{1}{1-F}\right)^{\leftarrow}(\cdot) \in RV_{1/\alpha}.$$

(b) Note in (iii) that the space $\mathbb{E} = (0, \infty]$ has 0 excluded and ∞ included. This is required since we need neighborhoods of ∞ to be relatively compact. Vague convergence only controls set wise convergence on relatively compact sets (with no mass on the boundary). With the usual topology on $[0, \infty)$ sets of the form (x, ∞) are not bounded; yet consideration of $n\bar{F}(b_n x) = nP[X_1/b_n > x]$ requires considering exactly such sets. We need some topology which makes semi-infinite intervals compact. More on this later. If it helps, think of $(0, \infty]$ as the homeomorphic stretching of $(0, 1]$ or as the homeomorphic image of $[0, \infty)$ under the map $x \mapsto 1/x$ which takes $[0, \infty) \mapsto (0, \infty]$. A convenient way to handle this is by using the *one point uncompactification* method to be discussed soon in Subsection 5.0.7.

(c) Preview of things to come: Note that if $\{X_j, j \geq 1\}$ is an iid sequence of non-negative random variables with common distribution F, then the measure ν_n defined in (5.1)

$$\nu_n(\cdot) = \mathbf{E}\left(\sum_{i=1}^{n} \epsilon_{X_i/b(n)}(\cdot)\right)$$

is the mean measure of the empirical measure of the scaled sample. The convergence of ν_n is equivalent to convergence of the sequence of empirical measures to a limiting Poisson process.

Proof. The equivalence of (i) and (ii) is Part (ii) of Proposition 2.

$(ii) \to (iii)$. Let $f \in C_K^+((0, \infty])$ and we must show

$$\nu_n(f) := nEf\left(\frac{X_1}{b_n}\right) = \int f(x) nP[\frac{X_1}{b_n} \in dx] \to \nu(f).$$

Since f has compact support, the support of f is contained in $(\delta, \infty]$ for some $\delta > 0$. We know

$$\nu_n(x, \infty] \to x^{-\alpha} = \nu(x, \infty], \quad \forall x > 0. \tag{5.3}$$

On $(\delta, \infty]$ define

$$P_n(\cdot) = \frac{\nu_n}{\nu_n(\delta, \infty]} \tag{5.4}$$

so that P_n is a probability measure on $(\delta, \infty]$. Then for $y \in (\delta, \infty]$

$$P_n(y, \infty) \to P(y, \infty] = \frac{y^{-\alpha}}{\delta^{-\alpha}}.$$

In $\mathbb{R}$, convergence of distribution functions (or tails) is equivalent to weak convergence so $\{P_n\}$ converges weakly to P. Since f is bounded and continuous on $(\delta, \infty]$, we get from weak convergence:

$$P_n(f) \to P(f);$$

that is,

$$\frac{\nu_n(f)}{\nu_n(\delta, \infty]} \to \frac{\nu(f)}{\delta^{-\alpha}}$$

and in light of (5.3), this implies

$$\nu_n(f) \to \nu(f)$$

as required.

$(iii) \to (ii)$. Since

$$\nu_n \xrightarrow{v} \nu,$$

we have

$$\nu_n(x, \infty] \to \nu(x, \infty], \quad \forall x > 0,$$

since $(x, \infty]$ is relatively compact and

$$\nu(\partial(x, \infty]) = \nu(\{x\}) = 0.$$

□

5.0.7. *Topological clarification: The one point uncompactification.* In reformulating the function theory concept of regularly varying functions into a measure theory concept, there is continual need to deal with sets which are bounded away from 0. Such sets need to be regarded as "bounded" in an appropriate topology so sequences of measures of such sets can converge non-trivially. A convenient way to think about this is by means of something we will call the one point un-compactification.

Let $(\mathbb{X}, \mathcal{T})$ be a nice topological space; $\mathbb{X}$ is the set and $\mathcal{T}$ is the topology, that is a collection of subsets of $\mathbb{X}$ designated as *open* satisfying

(i) Both $\emptyset \in \mathcal{T}$ and $\mathbb{X} \in \mathcal{T}$.
(ii) The collection $\mathcal{T}$ is closed under finite intersections and arbitrary unions.

(For our purposes, $\mathbb{X}$ would be a subset of Euclidean space.) Consider a subset $\mathbb{D} \subset \mathbb{X}$ and define

$$\mathbb{X}^{\#} = \mathbb{X} \setminus \mathbb{D}$$

and give $\mathbb{X}^{\#}$ the relative topology

$$\mathcal{T}^{\#} = \mathcal{T} \bigcap \mathbb{D}^c = \mathcal{T} \bigcap \mathbb{X}^{\#}.$$

So a set is open in $\mathbb{X}^{\#}$ if it is an open subset of $\mathbb{X}$ intersected with $\mathbb{X}^{\#}$.

What we need to understand is what are the compact sets of $\mathbb{X}^{\#}$.

Proposition 6. *Suppose, as usual, the compact subsets of $\mathbb{X}$ are denoted by $\mathcal{K}(\mathbb{X})$. Then*

$$\mathcal{K}(\mathbb{X}^{\#}) = \{K \in \mathcal{K}(\mathbb{X}) : K \bigcap \mathbb{D} = \emptyset\}$$

are the compact subsets of $\mathbb{X}^{\#}$.

So the compact sets of $\mathbb{X}^{\#}$ are the original compact sets of $\mathbb{X}$, provided they do not intersect the piece D chopped away from $\mathbb{X}$ to form $\mathbb{X}^{\#}$.

Specialize this to the *one point un-compactification*: Suppose $\mathbb{E}$ is a compact set and $e \in \mathbb{E}$. Give $\mathbb{E}\backslash\{e\}$ the relative topology consisting of sets in $\mathbb{E}\backslash\{e\}$ of the form $G\backslash\{e\}$, where $G \in \mathcal{G}(\mathbb{E})$. The compact sets of $\mathbb{E}\backslash\{e\}$ are those compact subsets $K \subset \mathbb{E}$ such that $e \notin K$.

So the *one point un-compactification* describes what are the compact sets when a compact space is punctured by the removal of a point.

Consider the following special cases:

(1) Suppose $\mathbb{E}$ is the compact set $[0,\infty]^d = [\mathbf{0},\boldsymbol{\infty}]$, which we may consider as the stretching of $[0,1]^d = [\mathbf{0},\mathbf{1}]$ onto $[\mathbf{0},\boldsymbol{\infty}]$. The compact sets of $[\mathbf{0},\boldsymbol{\infty}]$ consist of any closed set. The compact subsets of $[\mathbf{0},\boldsymbol{\infty}] \setminus \{\mathbf{0}\}$ are closed subsets of $[\mathbf{0},\boldsymbol{\infty}]$ bounded away from $\mathbf{0}$.

(2) Suppose $\mathbb{E}$ is the compact set $[-\boldsymbol{\infty},\boldsymbol{\infty}]$. The compact sets of $[-\boldsymbol{\infty},\boldsymbol{\infty}]$ consist of any closed set. The compact subsets of $[-\boldsymbol{\infty},\boldsymbol{\infty}]\setminus\{\mathbf{0}\}$ are closed subsets of $[-\boldsymbol{\infty},\boldsymbol{\infty}]$ bounded away from $\mathbf{0}$. This choice of $\mathbb{E}$, and the associated space of Radon measures $M_+(\mathbb{E})$ is needed for considering weak convergence of partial sums to multivariate Lévy processes and for analyzing multivariate problems related to *value at risk*.

(3) As a last example of the use of Proposition 6 suppose $\mathbb{E} = [\mathbf{0},\boldsymbol{\infty}] \setminus \{\mathbf{0}\}$ and define the cone $\mathbb{E}^0$ by

$$\mathbb{E}^0 := \{\boldsymbol{s} \in \mathbb{E} : \text{ For some } 1 \le i < j \le d, s^{(i)} \wedge s^{(j)} > 0\},$$

where we wrote the vector $\boldsymbol{s} = (s^{(1)},\ldots,s^{(d)})$. Here is an alternative description of $\mathbb{E}^0$: For $i = 1,\ldots,d$, define the basis vectors

$$\boldsymbol{e}_i = (0,\ldots,0,1,0,\ldots,0)$$

so that the axes originating at $\mathbf{0}$ are $\mathbb{L}_i := \{t\boldsymbol{e}_i, t > 0\}$, $i = 1,\ldots,d$. Then we also have

$$\mathbb{E}^0 = \mathbb{E} \setminus \bigcup_{i=1}^d L_i.$$

If $d = 2$, we have $\mathbb{E}^0 = (0,\infty]^2$.

The relatively compact subsets of $\mathbb{E}^0$ are those sets bounded away from the axes originating at $\mathbf{0}$. So G is relatively compact in $\mathbb{E}^0$ if for some $\delta > 0$ we have that for every $\boldsymbol{x} \in G$, for some $1 \le i < j \le d$, that $x^{(i)} \wedge x^{(j)} > \delta$.

Such a space is useful in consideration of asymptotic independence.

Proof. (Proposition 6.) Begin by supposing

$$K \in \mathcal{K}(\mathbb{X}), \quad K \cap \mathcal{D} = \emptyset,$$

and we show $K \in \mathcal{K}(\mathbb{X}^\#)$. Let

$$\{G_\gamma^\# = G_\gamma \cap \mathbb{X}^\#, \gamma \in \Lambda\}$$

be some arbitrary cover of K by open subsets of $\mathbb{X}^\#$ where $G_\gamma \in \mathcal{G}(\mathbb{X})$ and Λ is some index set. So

$$K \subset \bigcup_{\gamma\in\Lambda} G_\gamma \bigcap \mathbb{X}^\# \subset \bigcup_{\gamma\in\Lambda} G_\gamma.$$

Since $K \in \mathcal{K}(\mathbb{X})$, there is a finite subcollection indexed by $\Lambda' \subset \Lambda$ such that $K \subset \bigcup_{\gamma\in\Lambda'} G_\gamma$. Since $K \cap \mathbb{D} = \emptyset$,

$$K \subset \bigcup_{\gamma\in\Lambda'} G_\gamma \bigcap \mathbb{X}^\#.$$

So, any cover of K by open subsets of $\mathbb{X}^{\#}$ has a finite subcover and thus K is compact in $\mathbb{X}^{\#}$. Thus

$$\{K \in \mathcal{K}(\mathbb{X}) : K \bigcap \mathbb{D} = \emptyset\} \subset \mathcal{K}(\mathbb{X}^{\#}).$$

The converse is quite similar. □

References

P. Billingsley. *Convergence of Probability Measures.* Wiley, New York, 1968.

P. Billingsley. *Weak convergence of measures: Applications in probability.* Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1971. Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 5.

P. Billingsley. *Convergence of probability measures.* John Wiley & Sons Inc., New York, second edition, 1999. ISBN 0-471-19745-9. A Wiley-Interscience Publication.

R. Dudley. *Real Analysis and Probability.* Wadsworth and Brook/Cole, 1989.

O. Kallenberg. *Random Measures.* Akademie-Verlag, Berlin, third edition, 1983. ISBN 0-12-394960-2.

S. Resnick. *Extreme Values, Regular Variation and Point Processes.* Springer-Verlag, New York, 1987.

G. Simmons. *Topology and Modern Analysis.* McGraw-Hill, New York, 1963.

W. Whitt. *Stochastic Processs Limits: An Introduction to Stochastic-Process Limits And their Application to Queues.* Springer-Verlag, New York, 2002. ISBN 0-387-95358-2.

6. How to prove weak convergence?

Here is an outline of what it takes to prove weak convergence in 3 spaces of immediate interest.

(1) In $\mathbb{R}^d$, we can show random vectors $\{\boldsymbol{X}_n, n \geq 0\}$ converge weakly
$$\boldsymbol{X}_n \Rightarrow \boldsymbol{X}_0$$
by any of the following methods.
   (a) Show convergence of the finite dimensional distributions
$$\mathbb{P}[\boldsymbol{X}_n \leq x] \to \mathbb{P}[\boldsymbol{X}_0 \leq x]$$
   at continuity points of the limit. Sometimes this can even be done by showing convergence of the joint densities when they exist.
   (b) Show convergence of the characteristic functions
$$\mathbf{E}e^{i\boldsymbol{t}\cdot\boldsymbol{X}_n} \to \mathbf{E}e^{i\boldsymbol{t}\cdot\boldsymbol{X}_0}$$
   for $\boldsymbol{t} \in \mathbb{R}^d$.
   (c) Reduce the problem to one-dimension and prove
$$\boldsymbol{t}\cdot\boldsymbol{X}_n \Rightarrow \boldsymbol{t}\cdot\boldsymbol{X}_0,$$
   which works because of the previous item.
   (d) If $\boldsymbol{X}_n \geq \mathbf{0}$, show Laplace transforms converge,
$$\mathbf{E}e^{-\boldsymbol{\lambda}\cdot\boldsymbol{X}_n} \to \mathbf{E}e^{-\boldsymbol{\lambda}\cdot\boldsymbol{X}_0}$$
   for $\boldsymbol{\lambda} > \mathbf{0}$.

(2) In $\mathbb{R}^\infty$, random sequences $\{\boldsymbol{X}_n, n \geq 0\}$ of the form
$$\boldsymbol{X}_n = (X_n^{(1)}, X_n^{(2)}, \dots)$$
satisfy
$$\boldsymbol{X}_n \Rightarrow \boldsymbol{X}_0$$
if we show for any $d > 0$ that
$$(X_n^{(1)}, X_n^{(2)}, \dots X_n^{(d)}) \Rightarrow (X_0^{(1)}, X_0^{(2)}, \dots X_0^{(d)})$$
in $\mathbb{R}^d$.

(3) In $M_+(\mathbb{E})$, random measures $\{\xi(\cdot), n \geq 0\}$ converge weakly
$$\xi_n \Rightarrow \xi_0$$
iff for a family $\{h_j\}$, with $h_j \in C_K^+(\mathbb{E})$ we have
$$(\xi_n(h_j), j \geq 1) \Rightarrow (\xi_0(h_j), j \geq 1)$$
in $\mathbb{R}^\infty$.

7. The Tail Empirical Process and Hill's Estimator

The following describes a one-dimensional result but after converting regular variation to vague convergence, the result is really dimensionless.

7.1. **Tail empirical process.** Suppose $\{X_j, j \geq 1\}$ is a stationary sequence. Suppose the one-dimensional marginals F have regularly varying tails

$$\bar{F}(x) := 1 - F(x) = \mathbb{P}[X_1 > x] = x^{-\alpha} L(x), \quad x \to \infty, \ \alpha > 0, \tag{7.1}$$

and for convenience, assume the variables are non-negative. A useful scaling quantity is the *quantile function* $b(t)$ defined by

$$b(t) = \Big(\frac{1}{1-F}\Big)^{\leftarrow}(t) = F^{\leftarrow}(1 - \frac{1}{t}). \tag{7.2}$$

The *tail empirical measure* is defined as a random element of $M_+((0,\infty])$ by

$$\nu_n := \frac{1}{k} \sum_{i=1} \epsilon_{X_i/b(\frac{n}{k})}. \tag{7.3}$$

The new feature here is the presence of k, which represents the number of upper order statistics that we think (guess) are relevent for estimating tail probabilities. Note that the notation ν_n suppresses the dependence on k but that the k is critical. The tail empirical measure is used in a variety of inference contexts but note that, as defined, it's statistical use needs to overcome the fact that in a data context where F is unknown, $b(\cdot)$ is also unknown.

Theorem 5. *Suppose $\{X_j, j \geq 1\}$ are iid, non-negative random variables whose common distribution has a tail which is regularly varying which implies*

$$\frac{n}{k}\mathbb{P}[\frac{X_1}{b(n/k)} \in \cdot] \xrightarrow{v} \nu_\alpha \tag{7.4}$$

in $M_+((0,\infty])$ as $n \to \infty$ and $k = k(n) \to \infty$ with $n/k \to \infty$. Then in $M_+((0,\infty])$

$$\nu_n \Rightarrow \nu_\alpha \tag{7.5}$$

where

$$\nu_\alpha(x,\infty] = x^{-\alpha}, \quad x > 0, \ \alpha > 0.$$

Proof. It suffices to show for a sequence $h_j \in C_K^+(0,\infty]$ that in $\mathbb{R}^\infty$

$$(\nu_n(h_j), j \geq 1) \Rightarrow \nu_\alpha(h_j), j \geq 1\}, \quad (n \to \infty).$$

Convergence in $\mathbb{R}^\infty$ reduces to convergence in $\mathbb{R}^k$ for any k so it suffices to show

$$(\nu_n(h_j), 1 \leq j \leq k) \Rightarrow (\nu_\alpha(h_j), 1 \leq j \leq k), \quad (n \to \infty).$$

To show this, we can show the joint Laplace transforms converge: For $\lambda_j > 0, \ j = 1, \ldots, k$, we must show

$$\mathbf{E}e^{-\sum_{j=1}^k \lambda_j \nu_n(h_j)} \to \mathbf{E}e^{-\sum_{j=1}^k \lambda_j \nu_\alpha(h_j)}.$$

However,

$$\sum_{j=1}^k \lambda_j \nu_n(h_j) = \nu_n(\sum_{j=1}^k \lambda_j h_j)$$

and similarly for ν_α substituted for ν_n. Note $\sum_{j=1}^k \lambda_j h_j \in C_K^+(0,\infty]$ so it suffices to show for any $h \in C_K^+(0,\infty]$ that

$$\mathbf{E}e^{-\nu_n(h)} \to \mathbf{E}e^{-\nu_\alpha(h)}. \tag{7.6}$$

The left side of (7.6) is

$$\begin{aligned}\mathbf{E}e^{-\frac{1}{k}\sum_{j=1}^n h\left(X_j/b(n/k)\right)} &= \left(\mathbf{E}e^{-\frac{1}{k}h\left(X_1/b(n/k)\right)}\right)^n \\ &= \left(1-\int_{(0,\infty]}\left(1-e^{-\frac{1}{k}h(x)}\right)\mathbb{P}[\frac{X_1}{b(n/k)}\in dx]\right)^n \\ &= \left(1-\frac{\int_{(0,\infty]}\left(1-e^{-\frac{1}{k}h(x)}\right)n\mathbb{P}[\frac{X_1}{b(n/k)}\in dx]}{n}\right)^n \\ &\to e^{-\nu_\alpha(h)},\end{aligned}$$

since

$$\int_{(0,\infty]}\left(1-e^{-\frac{1}{k}h(x)}\right)n\mathbb{P}[\frac{X_1}{b(n/k)}\in dx] \sim \int_{(0,\infty]} h(x)\frac{n}{k}\mathbb{P}[\frac{X_1}{b(n/k)}\in dx] \to \nu_\alpha(h).$$

□

7.2. **The Hill Estimator.**

7.2.1. *Introduction.* The first steps in heavy tailed analysis are

- Decide that a heavy tailed model is appropriate, and then
- Estimate the tail index α.

Various graphical and estimation techniques exist to help accomplish these steps: QQ estimation and plotting, Hill estimation and plotting, Pickands estimation, extreme value techniques, etc.

Consider the problem of estimating the index of regular variation when the tail probability $P[X > x]$ of a random variable X has a distribution with regularly varying tail. A popular estimator is the *Hill estimator* defined as follows. Assume for simplicity that observations $X_1, \ldots, X_n$ are non-negative. For $1 \le i \le n$, write $X_{(i)}$ for the ith largest value of $X_1, X_2, \ldots, X_n$. and then Hill's estimator based on k upper order statistics is defined as:

$$H_{k,n} := \frac{1}{k}\sum_{i=1}^k \log\frac{X_{(i)}}{X_{(k+1)}}. \tag{7.7}$$

The theory is most developed for the case that $\{X_j, j \ge 1\}$ is iid but applications often do not provide us with independent observations but rather with dependent, stationary data. So attention needs to be paid to applying the Hill estimator in non-iid cases.

Suppose at a minimum that $\{X_n\}$ is a sequence of random variables having the same marginal distribution function F and where $\bar{F}$:= 1 - F is regularly varying at ∞ and satisfies (7.1). The quantile function (7.2) is $b(t)$. The random measure ν_n given in (7.3) is a random element of $M_+(0,\infty]$ and is *assumed* to be a consistent estimator (in the vague topology) of the measure $\nu_\alpha \in M_+(0,\infty]$, provided $n \to \infty$, and $k/n \to 0$. However, because $b(\cdot)$ is unknown, $b(n/k)$ will be estimated by a consistent estimator, $\hat{b}(n/k)$ to be specified, and we will denote

$$\hat{\nu}_n =: \frac{1}{k}\sum_{i=1}^n \epsilon_{X_i/\hat{b}(n/k)}, \tag{7.8}$$

the estimator when $b(n/k)$ is replaced by an estimator $\hat{b}(n/k)$.

We know from Theorems 5 that (7.5) is satisfied if $\{X_j\}$ is iid with common distribution F where $1-F \in \mathrm{RV}_{-\alpha}$.

7.2.2. *Random measures and the consistency of the Hill estimator.* Consistency of the tail empirical measure given in (7.5) imply consistency of the Hill estimator for $1/\alpha$.

Theorem 6. *If* (7.5) *holds then, as* $n \to \infty$, $k \to \infty$ *and* $k/n \to 0$,

$$H_{k,n} \xrightarrow{P} \frac{1}{\alpha}.$$

Proof. The proof proceeds by a series of steps.

Step 1. Consistency of the empirical measure given in (7.5) implies

$$\frac{X_{(k)}}{b(n/k)} \xrightarrow{P} 1, \tag{7.9}$$

as $n \to \infty$, $k \to \infty$ and $k/n \to 0$. This allows us to consider $X_{(k)}$ as a consistent estimator of $b(n/k)$.

To see this, write

$$\begin{aligned} P[|\frac{X_{(k)}}{b(\frac{n}{k})} - 1| > \varepsilon] =& P[X_{(k)} > (1+\varepsilon)b(\frac{n}{k})] + P[X_{(k)} < (1-\varepsilon)b(\frac{n}{k}))] \\ \leq& P[\frac{1}{k}\sum_{i=1}^{n} \epsilon_{X_i/b(\frac{n}{k})}(1+\varepsilon,\infty] \geq 1) + P[\frac{1}{k}\sum_{i=1}^{n} \epsilon_{X_i/b(\frac{n}{k})}\,[1-\varepsilon,\infty] < 1]. \end{aligned}$$

But (7.5) implies that

$$\frac{1}{k}\sum_{i=1}^{k} \epsilon_{X_i/b(\frac{n}{k})}(1+\varepsilon,\infty] \xrightarrow{P} (1+\varepsilon)^{-\alpha} < 1,$$

and

$$\frac{1}{k}\sum_{i=1}^{k} \epsilon_{X_i/b(\frac{n}{k})}[1-\varepsilon,\infty] \xrightarrow{P} (1-\varepsilon)^{-\alpha} > 1$$

and therefore (7.9) follows. □

Appendix: In fact, more is true. We have that (7.5) implies

$$\frac{X_{(\lceil kt \rceil)}}{b(n/k)} \xrightarrow{P} t^{-1/\alpha}, \quad \text{in } D(0,\infty], \tag{7.10}$$

where $\lceil kt \rceil$ is the smallest integer greater than or equal to kt. To prove this more muscular version (7.10), proceed as follows: The map from $M_+\big((0,\infty]\big) \mapsto D[0,\infty)$ defined by

$$\mu \mapsto \mu(t^{-1},\infty], \quad t \geq 0,$$

is continuous at measures μ such that $\mu(t,\infty]$ is continuous, strictly decreasing in t. So we have from (7.5)

$$\nu_n(t^{-1},\infty] \xrightarrow{P} t^{\alpha}, \quad t \geq 0, \tag{7.11}$$

in $D[0,\infty)$. This implies (see the Lemma below) that functional inverses also converge in probability

$$\Big(\nu_n\big((\cdot)^{-1},\infty]\Big)^{\leftarrow}(t) \xrightarrow{P} t^{1/\alpha}, \quad t \geq 0, \tag{7.12}$$

in $D[0,\infty)$. We now unpack the inverse and see what we get:

$$\begin{aligned}
\Big(\nu_n(()^{-1},\infty]\Big)^{\leftarrow}(t) &= \inf\{s : \nu_n(s^{-1},\infty] \geq t\} \\
&= \inf\{s : \sum_{i=1}^{n} \epsilon_{X_i/b(n/k)}(s^{-1},\infty] \geq kt\} \\
&= \inf\{y^{-1} : \sum_{i=1}^{n} \epsilon_{X_i/b(n/k)}(y,\infty] \geq kt\} \\
&= \Big(\sup\{y : \sum_{i=1}^{n} \epsilon_{X_i/b(n/k)}(y,\infty] \geq kt\}\Big)^{-1} \\
&= \Big(\frac{X_{(\lceil kt\rceil)}}{b(n/k)}\Big)^{-1}.
\end{aligned}$$

So

$$\Big(\frac{X_{(\lceil kt\rceil)}}{b(n/k)}\Big)^{-1} \Rightarrow t^{1/\alpha}$$

in $D(0,\infty]$ and therefore, we conclude

$$\frac{X_{(\lceil kt\rceil)}}{b(n/k)} \xrightarrow{P} t^{-1/\alpha}$$

in $D(0,\infty]$.

The justification for the inversion step comes from the following simple lemma.

Lemma 2. *(a) If $x_n \in D[0,\infty)$ is non-decreasing, $x_n(0) = 0$ and $x_n \to x_0$ in $D[0,\infty)$ where x_0 is continuous,strictly increasing, then*

$$x_n^{\leftarrow} \to x_0^{\leftarrow}$$

locally uniformly and in $D[0,\infty)$.

(b) Suppose ξ_n is a stochastic process with non-decreasing paths in $D[0,\infty)$ such that $\xi_n(0) = 0$, and

$$\xi_n \xrightarrow{P} \xi_0, \tag{7.13}$$

in $D[0,\infty)$, and almost all paths of ξ_0 are continuous, strictly increasing, then

$$\xi_n^{\leftarrow} \xrightarrow{P} \xi_0^{\leftarrow}. \tag{7.14}$$

Proof of Lemma 2. (a) We have

$$x_n^{\leftarrow}(t) \to x_0^{\leftarrow}(t)$$

pointwise by inversion. This gives monotone functions converging to a continuous limit and hence convergence is locally uniform. Local uniform convergence implies convergence in the Skorohod metric.

(b) Let $d(\cdot,\cdot)$ be the Skorohod metric on $D[0,\infty)$ and (7.13) says

$$d(\xi_n, \xi_0) \xrightarrow{P} 0 \tag{7.15}$$

and we need to show

$$d(\xi_n^{\leftarrow}, \xi_0^{\leftarrow}) \xrightarrow{P} 0. \tag{7.16}$$

Use the subsequence characterization of convergence in probability: given a subsequence $\{n''\}$, it suffices to find a further subsequence $\{n'\} \subset \{n''\}$ such that

$$d(\xi_{n'}^{\leftarrow}, \xi_0^{\leftarrow}) \xrightarrow{\text{a.s.}} 0.$$

From (7.15), pick $\{n'\}$ such that

$$d(\xi_{n'}, \xi_0) \xrightarrow{\text{a.s.}} 0.$$

Then for almost all ω

$$\xi_{n'}(t,\omega) \to \xi_0(t,\omega), \quad \forall t \geq 0,$$

and so by inverting of the monotone functions

$$\xi_{n'}^{\leftarrow}(t,\omega) \to \xi_0^{\leftarrow}(t,\omega), \quad \forall t \geq 0.$$

Since $\xi_{n'}^{\leftarrow}(t,\omega)$ is monotone in t and $\xi_0^{\leftarrow}(t,\omega)$ is continuous in t, the convergence is locally uniform in t as required. □

Henceforth, set

$$\hat{b}(n/k) = X_{(k)}.$$

Step 2. The following results from (7.5): In $M_+(0,\infty]$,

$$\hat{\nu}_n \xrightarrow{P} \nu, \tag{7.17}$$

as $n \to \infty$, $k \to \infty$ and $k/n \to 0$. This is proven with a scaling argument. We need the following simple lemma from the theory of weak convergence.

Lemma 3. *Let $\mathbb{E}$ and $\mathbb{E}'$ be two complete separable metric spaces and suppose $\{\xi_n, n \geq 0\}$ and $\{\eta_n, n \geq 1\}$ are random elements of $\mathbb{E}$ and $\mathbb{E}'$ respectively defined on the same probability space. Suppose*

$$\xi_n \Rightarrow \xi_0$$

in $\mathbb{E}$ and

$$\eta_n \xrightarrow{P} e_0'$$

where e_0' is a fixed point of $\mathbb{E}'$; that is, e_0' is non-random. Then jointly in $\mathbb{E} \times \mathbb{E}'$ we have

$$(\xi_n, \eta_n) \Rightarrow (\xi_0, e_0'),$$

as $n \to \infty$.

Proof. Let $f : \mathbb{E} \times \mathbb{E}'$ be bounded and continuous and without loss of generality suppose f is uniformly continuous. Write

$$\begin{aligned}|E(f(\xi_n,\eta_n)-&Ef(\xi_0,e_0')| \\ &\le |E(f(\xi_n,\eta_n) - Ef(\xi_n,e_0')| + |E(f(\xi_n,e_0') - Ef(\xi_0,e_0')| \\ &= I + II.\end{aligned}$$

Now $II \to 0$ since $f(\cdot, e_0')$ is bounded and continuous on $\mathbb{E}$ and $\xi_n \Rightarrow \xi_0$. For I we note for any $\delta > 0$,

$$\begin{aligned}I \le& E|(f(\xi_n,\eta_n) - f(\xi_n,e_0')|1_{[|\eta_n - e_0'|\le\delta]} + E|(f(\xi_n,\eta_n) - f(\xi_n,e_0')|1_{[|\eta_n - e_0'|\ge\delta]} \\ =& Ia + Ib.\end{aligned}$$

We have

$$Ia \le \sup_{\|\boldsymbol{x}-\boldsymbol{y}\|\le\delta} |f(\boldsymbol{x}) - f(\boldsymbol{y})|$$

(where $\|\cdot\|$ should be interpreted as the metric on $\mathbb{E}\times\mathbb{E}'$) which is small by uniform continuity of f and

$$Ib \le (\text{const})P[|\eta_n - e_0'| > \delta] \to 0.$$

□

We now proceed with the scaling argument to confirm (7.17). Define the operator:

$$T : M_+(E) \times (0,\infty) \mapsto M_+(E)$$

by

$$T(\mu, x)(A) = \mu(xA).$$

From (7.5) and Lemma 3 we get joint weak convergence

$$(\nu_n, \frac{X_{(k)}}{b(\frac{n}{k})}) \Rightarrow (\nu_\alpha, 1) \tag{7.18}$$

in $M_+(0,\infty] \times (0,\infty)$. Since

$$\hat{\nu}_n(\cdot) = \nu_n\left(\frac{X_{(k)}}{b(\frac{n}{k})}\cdot\right) = T\left(\nu_n, \frac{X_{(k)}}{b(\frac{n}{k})}\right),$$

the conclusion will follow by the continuous mapping theorem, provided we prove the continuity of the operator T at $(\nu_\alpha, 1)$. In fact, we prove the continuity of the operator at (ν_α, x) where $x > 0$. Towards this goal, let $\mu_n \xrightarrow{v} \nu_\alpha$ and $x_n \to x$, where $\mu_n \in M_+(0,\infty]$, x_n, $x \in (0,\infty)$. It suffices to show for any $f \in C_K^+(0,\infty]$ that

$$\int_{(0,\infty]} f(t)\mu_n(x_n dt) = \int_{(0,\infty]} f(y/x_n)\mu_n(dy) \to \int_{(0,\infty]} f(y/x)\nu_\alpha(dy). \tag{7.19}$$

Write

$$|\int_{(0,\infty]} f(y/x_n)\mu_n(dy) - \int_{(0,\infty]} f(y/x)\nu_\alpha(dy)|$$

$$\begin{aligned}
&\leq |\int_{(0,\infty]} f(y/x_n)\mu_n(dy) - \int_{(0,\infty]} f(y/x)\mu_n(dy)| \\
&\qquad + |\int_{(0,\infty]} f(y/x)\mu_n(dy) - \int_{(0,\infty]} f(y/x)\nu_\alpha(dy)| \\
&\leq \int_{(0,\infty]} |f(y/x_n) - f(y/x)|\mu_n(dy) + o(1)
\end{aligned}$$

where the second difference goes to 0 because $f(\frac{\cdot}{x}) \in C_K^+(0,\infty]$. To see that the first difference can be made small, note the supports of $f(\frac{\cdot}{x})$ and $f(\frac{\cdot}{x_n})$ for large n are contained in $[\delta_0, \infty]$ for some δ_0. Since f is continuous with compact support, f is uniformly continuous on $(0,\infty]$. To get an idea what this means, metrize $(0,\infty]$ by the metric $(s,t \in (0,\infty])$

$$d(s,t) = |s^{-1} - t^{-1}|$$

and then uniform continuity means

$$\sup_{d(u,v)<\delta} |f(u) - f(v)| \stackrel{\delta\downarrow 0}{\to} 0.$$

Then

$$d(y/x_n, y/x) = y^{-1}|x_n - x| < \delta$$

if $y > \delta_0$ and n is large and therefore for any $\epsilon > 0$ we can make

$$\sup_{y\geq\delta_0} |f(y/x_n) - f(y/x)| < \epsilon.$$

Since $\mu_n(\delta_0,\infty]$ is bounded, this completes the proof of continuity of the scaling map. □

Step 3. Integrate the tails of the measures against $x^{-1}dx$. The integral functional is continuous on $[1,M]$ for any M and so it is only on $[M,\infty]$ that care must be exercised. By the second converging together theorem (a variant of Slutsky's theorem–see below) we must show

$$\lim_{M\to\infty} \limsup_{n\to\infty} P[\int_M^\infty \hat{\nu}_n(x,\infty]x^{-1}dx > \delta] = 0. \tag{7.20}$$

Set $\hat{b}(n/k) = X_{(k)}$. Decompose the probability as

$$\begin{aligned}
P[\int_M^\infty \hat{\nu}_n(x,\infty]x^{-1}dx > \delta] \leq & P[\int_M^\infty \hat{\nu}_n(x,\infty]x^{-1}dx > \delta, \frac{\hat{b}(n/k)}{b(n/k)} \in (1-\eta, 1+\eta)] \\
& + P[\int_M^\infty \hat{\nu}_n(x,\infty]x^{-1}dx > \delta, \frac{\hat{b}(n/k)}{b(n/k)} \notin (1-\eta, 1+\eta)] \\
= & I + II.
\end{aligned}$$

Note

$$II \leq P[|\frac{\hat{b}(n/k)}{b(n/k)} - 1| > \eta] \to 0$$

by (7.9). We have that I is bounded above by

$$P[\int_M^\infty \nu_n((1-\eta)x,\infty]x^{-1}dx > \delta] = P[\int_{M(1-\eta)}^\infty \nu_n(x,\infty]x^{-1}dx > \delta]$$

and the above probability has a Chebychev bound

$$\begin{aligned}
&\delta^{-1}E\Big(\int_{M(1-\eta)}^\infty \nu_n(x,\infty]x^{-1}dx\Big)\\
&= \delta^{-1}\int_{M(1-\eta)}^\infty \frac{n}{k}P[X_1 > b(n/k)x]x^{-1}dx\\
&\stackrel{n\to\infty}{\to} \delta^{-1}\int_{M(1-\eta)}^\infty x^{-\alpha-1}dx = (\text{const})M^{-\alpha}\\
&\stackrel{M\to\infty}{\to} 0.
\end{aligned}$$

□

Step 4. We have proven that

$$\int_1^\infty \hat{\nu}_n(x,\infty]x^{-1}dx \stackrel{P}{\to} \int_1^\infty \nu_\alpha(x,\infty]x^{-1}dx = 1/\alpha.$$

So $\int_1^\infty \hat{\nu}_n(x,\infty]x^{-1}dx$ is a consistent estimator of $1/\alpha$ and we just need to see that this is indeed the Hill estimator as defined in (7.7). This is done as follows:

$$\begin{aligned}
\int_1^\infty \hat{\nu}_n(x,\infty]x^{-1}dx &= \int_1^\infty \frac{1}{k}\sum_{i=1}^n \epsilon_{X_i/\hat{b}(n/k)}(x,\infty]x^{-1}dx\\
&= \frac{1}{k}\sum_{i=1}^n \int_1^{X_i/\hat{b}(n/k)\vee 1} x^{-1}dx
\end{aligned}$$

which is equivalent to $H_{k,n}$ defined in (7.7). □

7.2.3. *Appendix: The second converging together theorem.* This simple result provides a framework when approximations or truncations must be employed.

Proposition 7. *Supppose $\{X_{n,M}\}$ are random elements of the metric space $(S,\mathcal{S})$ with metric $d(\cdot,\cdot)$ and defined on the same probability space. The family $\{X_{n,M}\}$ satisfies*

(1) *As $n\to\infty$*
$$X_{n,M} \Rightarrow X_{\infty,M},$$
for each fixed M.

(2) *As $M\to\infty$*
$$X_{\infty,M} \Rightarrow X_{\infty,\infty}.$$

(3) *For some random elements $\{Y_n\}$ we have*
$$\lim_{M\to\infty}\limsup_{n\to\infty}\mathbb{P}[d(X_{n,M},Y_n) > \eta] = 0, \quad \forall \eta > 0.$$

Then

$$Y_n \Rightarrow X_{\infty,\infty}.$$

7.2.4. *The Hill estimator in practice.* In practice, the Hill estimator is used as follows: We make the *Hill plot*, of

$$\{(k, H_{k,n}^{-1}), 1 \le k \le n\}$$

and hope the graph looks stable so you can pick out a value of α.

Sometimes this works beautifully and sometimes the plots are not very revealing. Consider Figure 7 which shows two cases where the procedure is heart-warming. The top row are time series plots. The top left plot is 4045 simulated observations from a Pareto distribution with $\alpha = 1$ and the top right plot is 4045 telephone call holding times indexed according to the time of initiation of the call. Both plots are scaled by division by 1000. The range of the Pareto data is (1.0001, 10206.477) and the range of the call holding data is (2288,11714735). The bottom two plots are Hill plots $\{(k, H_{k,n}^{-1}), 1 \le k \le 4045\}$, the bottom left plot being for the Pareto sample and the bottom right plot for the call holding times. After settling down, both Hill plots are gratifyingly stable and are in a tight neighborhood. The Hill plot for the Pareto seems to estimate $\alpha = 1$ correctly and the estimate in the call holding example seems to be between .9 and 1. (So in this case, not only does the variance not exist but the mean appears to be infinite as well.) The Hill plots could be modified to include a confidence interval based on the asymptotic normality of the Hill estimator. McNeil's Hillplot function does just this.

The Hill plot is not always so revealing. Consider Figure 8, one of many Hill Horror Plots. The left plot is for a simulation of size 10,000 from a symmetric α–stable distribution with $\alpha = 1.7$. One would have to be paranormal to discern the correct answer of 1.7 from the

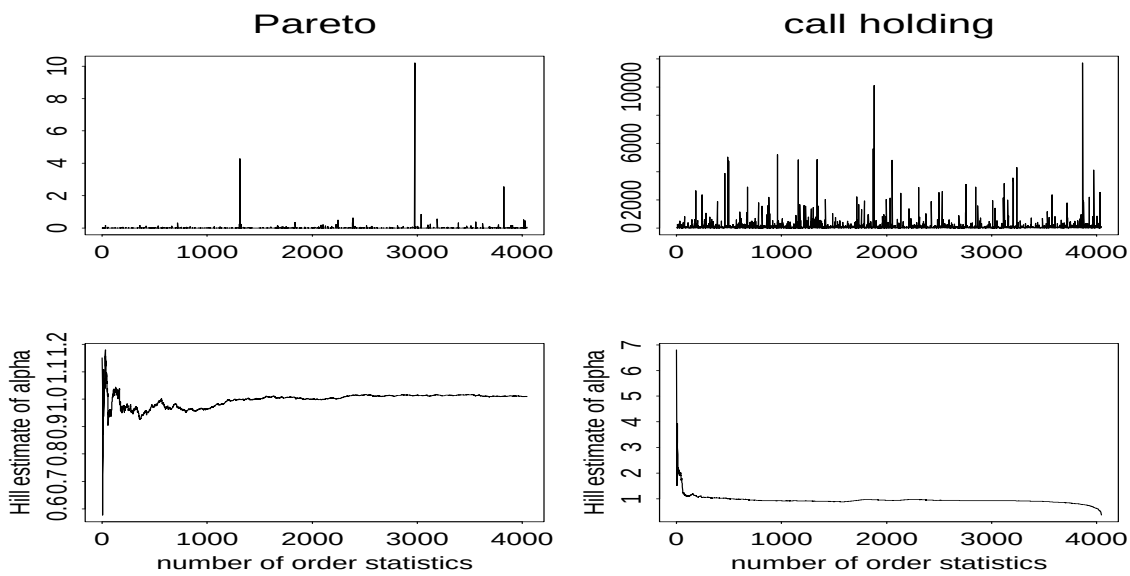

FIGURE 7. Time series and Hill plots for Pareto (left) and call holding (right) data.

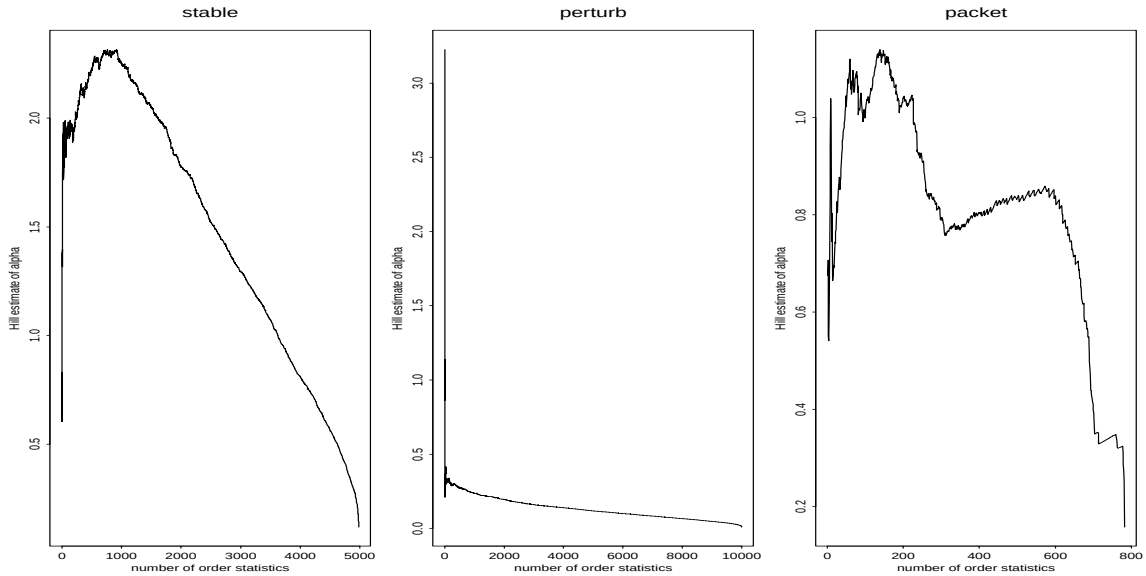

FIGURE 8. A Hill Horror Plot.

plot. The middle plot is for a simulated iid sample of size 10,000, called *perturb*, from the distribution tail

$$1 - F(x) \sim x^{-1}(\log x)^{10}, \quad x \to \infty,$$

so that $\alpha = 1$. The plot exhibits extreme bias and comes nowhere close to indicating the correct answer of 1. The problem, of course, is that the Hill estimator is designed for the Pareto distribution and thus does not know how to interpret information correctly from the factor $(\log x)^{10}$ and merely readjusts its estimate of α based on this factor rather than identifying the logarithmic perturbation. The third plot is 783 real data called *packet* representing inter-arrival times of packets to a server in a network. The problem here is that the graph is volatile and it is not easy to decide what the estimate should be. The sample size may just be too small.

A summary of difficulties when using the Hill estimator include:

(1) One must get a point estimate from a graph? What value of k should one use?
(2) The graph may exhibit considerable volatility and/or the true answer may be hidden in the graph.
(3) The Hill estimate has optimality properties only when the underlying distribution is close to Pareto. If the distribution is far from Pareto, there may be outrageous error, even for sample sizes like 1,000,000.
(4) The Hill estimator is not location invariant. A shift in location does not theoretically affect the tail index but may throw the Hill estimate way off.

The lack of location invariance means the Hill estimator can be surprisingly sensitive to changes in location. Figure 9 illustrates this. The top plots are time series plots of 5000 iid Pareto observations where the true $\alpha = 1$. The two right plots on top have the Pareto

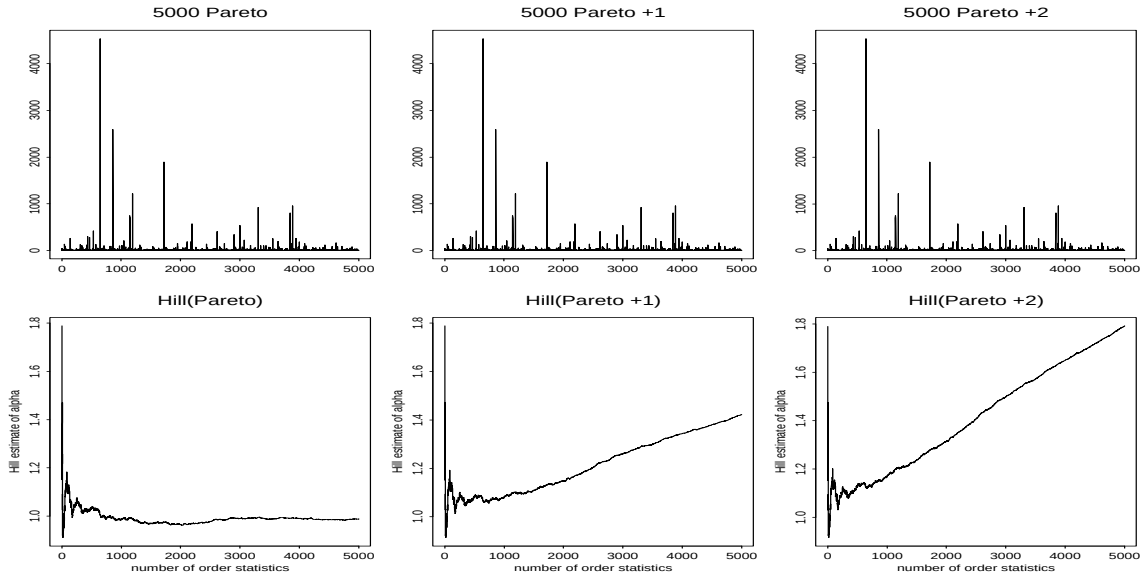

FIGURE 9. Lack of location invariance.

observations shifted by 1 and then 2. The bottom two plots are the corresponding Hill plots. Shifting by larger and larger amounts soon produces a completely useless plot.

For point 1, several previous studies advocate choosing k to minimize the asymptotic mean squared error of Hill's estimator (Hall (1982), Peng (1998)). In certain cases, the asymptotic form of this optimal k can be expressed but such a form requires one to know the distribution rather explicitly and it is not clear how much value one gets from an asymptotic formula. There are adaptive methods and bootstrap techniques which try to overcome these problems; it remains to be seen if they will enter the research community's toolbox.

For point 2, there are simple smoothing techniques which always help to overcome the volatility of the plot and plotting on a different scale frequently overcomes the difficulty associated with the stable example. These techniques are discussed in the next paragraph.

7.2.5. *Variant 1. The smooHill plot.* The Hill plot often exhibits extreme volatility which makes finding a stable regime in the plot more guesswork than science and to counteract this, Resnick and Stărică (1997) developed a smoothing technique yielding the smooHill plot: Pick an integer u (usually 2 or 3) and define

$$smooH_{k,n} = \frac{1}{(u-1)k} \sum_{j=k+1}^{uk} H_{j,n}.$$

In the iid case, when a second order regular variation condition holds, the asymptotic variance of $smooH_{k,n}$ is less than that of the Hill estimator, namely:

$$\frac{1}{\alpha^2}\frac{2}{u}(1 - \frac{\log u}{u}).$$

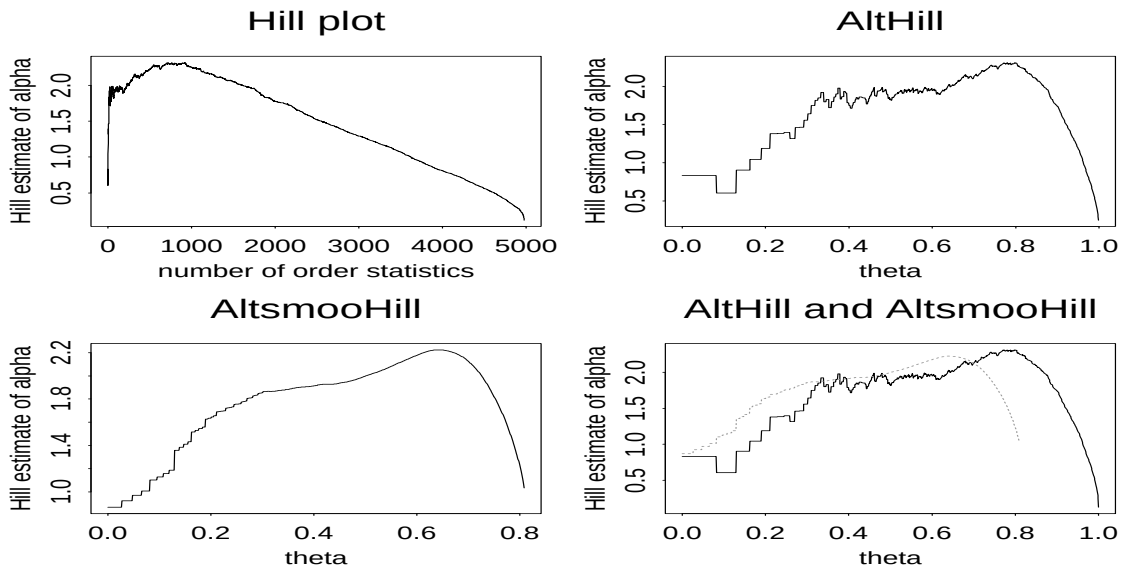

FIGURE 10. Stable, $\alpha = 1.7$

7.2.6. *Variant 2: Alt plotting; Changing the scale.* As an alternative to the Hill plot, it is sometimes useful to display the information provided by the Hill or smooHill estimation as

$$\{(\theta, H^{-1}_{\lceil n^\theta \rceil, n}), 0 \le \theta \le 1, \}$$

and similarly for the smooHill plot where we write $\lceil y \rceil$ for the smallest integer greater or equal to $y \ge 0$. We call such plots the *alternative Hill plot* abbreviated AltHill and the *alternative smoothed Hill plot* abbreviated AltsmooHill. The alternative display is sometimes revealing since the initial order statistics get shown more clearly and cover a bigger portion of the displayed space. Unless the distribution is Pareto, the AltHill plot spends more of the display space in a small neighborhood of α than the conventional Hill plot (Drees et al. (2000)).

Figure 10 compares several Hill plots for 5000 observations from a stable distribution with $\alpha = 1.7$. Plotting on the usual scale is not revealing and the alt plot is more informative.

A Hill plot was given on page 9 for file lengths downloaded in BU web sessions in November 1994 in a particular lab under study. The Danish fire insurance data was introduced on page 8. In Figure 11, we have a Hill and SmooHill plot of the Danish data.

7.2.7. *Software routines.* Splus or R is a convenient environment for graphical analysis. We will give some routines that have been found useful for heavy tailed analysis. These have been written in the Splus environment but should run in R.

Other software that may be helpful comes from the extreme value world. Alexander McNeil (www.math.ethz.ch/∼mcneil/software.html) has a very professional compilation of Splus routines for performing extreme value analysis which can be adapted to heavy tailed analysis. A version will now marketed as part of the *Finmetrics* module by Mathsoft but I believe McNeil will continue to offer the software on his web site.

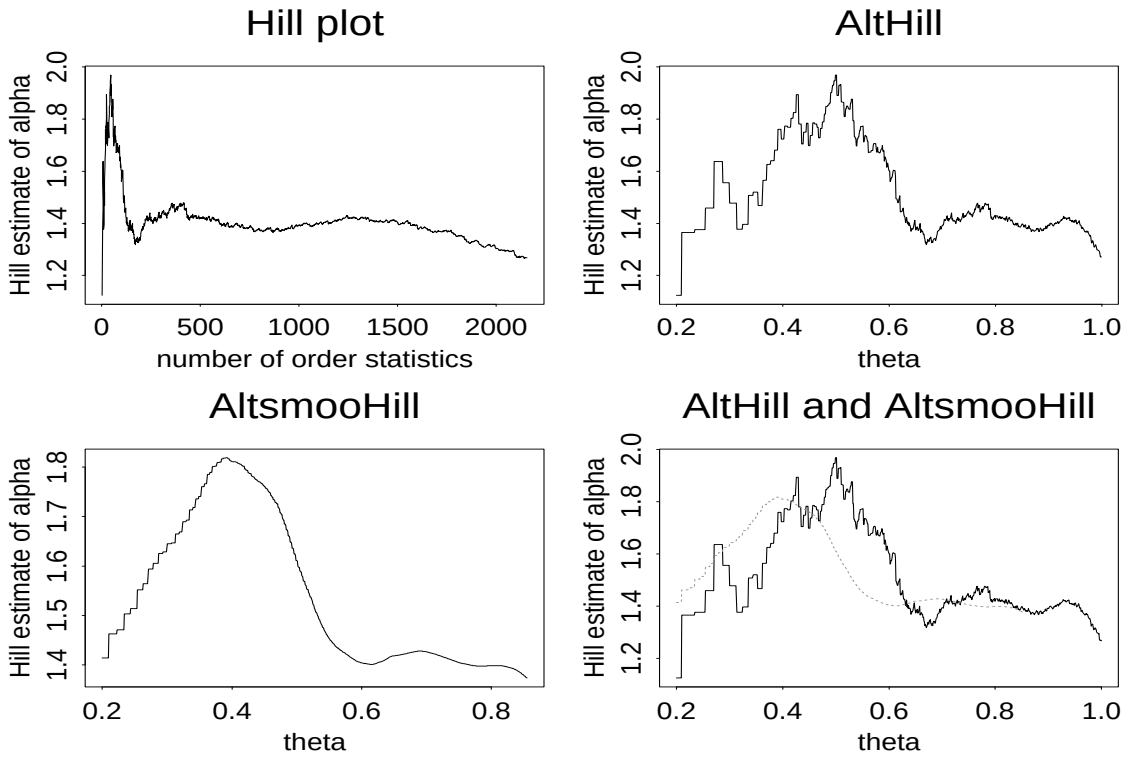

FIGURE 11. Hill and smooHill plots for Danish data.

Stuart Coles also has notes explaining his Splus routines linked to http://www.stats.bris.ac.uk/∼masgc/. His routines are used in his new extreme values book Coles (2001). There is an excellent menu driven program called XTREMES constructed under the leadership of Rolf Reiss, Siegen, with an accompanying book (Reiss and Thomas (2001)) and more information is available at www.xtremes.math.uni-siegen.de.

1. Hillalpha(x) where x=name of dataset produces the Hill plot for estimating α.

```
function(x)
{
ordered <- rev(sort(x))
ordered <- ordered[ordered[] > 0]
n <- length(ordered)
loggs <- log(ordered)
hill <- cumsum(loggs[1:(n - 1)])/(1:(n - 1)) - loggs[2:n]
hill <- 1/(hill)
plot(1:length(hill), hill, type = "l",
        xlab = "number of order statistics",
        ylab = "Hill estimate of alpha")
}
```

2. twoHill(x) where x=name of dataset; produces the Hill and altHill plots side by side.

```
function(x)
{
par(mfrow = c(1, 2))
ordered <- rev(sort(x))
ordered <- ordered[ordered[] > 0]
n <- length(ordered)
loggs <- log(ordered)
hill <- cumsum(loggs[1:(n - 1)])/(1:(n - 1)) - loggs[2:n]
hill <- 1/hill
plot(1:length(hill), hill, type = "l", xlab =
"number of order statistics", ylab =
"Hill estimate of alpha", main = "Hill plot")
s <- log(1:length(hill))/log(length(x))
plot(s, hill, type = "l", xlab = "theta", ylab =
"Hill estimate of alpha", main = "AltHill")
par(mfrow = c(1, 1))
}
```

3. threeHill(x) where x=name of the data set, l1=percent of the plots to cut off on the left to make the graph scale in an informative way; 0 is a possible value. u1=percent of the plots to cut off on the left to make the graph scale in an informative way; 1 is a possible value. r=amount of smoothing.

```
function(x, l1, u1, r)
{
par(mfrow = c(1, 3))
ordered <- rev(sort(x))
ordered <- ordered[ordered[] > 0]
n <- length(ordered)
loggs <- log(ordered)
hill <- cumsum(loggs[1:(n - 1)])/(1:(n - 1)) - loggs[2:n]
hill <- 1/hill
plot((n^l1):(n^u1), hill[(n^l1):(n^u1)], type = "l",
            xlab="number of order statistics",
            ylab="Hill estimate of alpha",
            main = "Hill plot")
s <- vector(mode = "numeric", length = (u1 - l1) * 100 + 1)
f <- vector(mode = "numeric", length = (u1 - l1) * 100 + 1)
for(i in (l1 * 1000):(u1 * 1000)) {
a <- n^(i/1000)
f[i - l1 * 1000 + 1] <- hill[a]
s[i - l1 * 1000 + 1] <- i/1000
}
plot(s, f, type = "l", xlab = "theta", ylab =
"Hill estimate of alpha", main = "AltHill")
```

```
hill1 <- 1/hill
u2 <- (log(n))^(-1) * (u1 * log(n) - log(r))
s <- vector(mode = "numeric", length = (u1 - l1) * 100 + 1)
f <- vector(mode = "numeric", length = (u1 - l1) * 100 + 1)
f1 <- vector(mode = "numeric", length = (u2 - l1) * 100 + 1
)
for(i in (l1 * 1000):(u2 * 1000)) {
a <- n^(i/1000)
f[i - l1 * 1000 + 1] <- hill[a]
s[i - l1 * 1000 + 1] <- i/1000
f1[i - l1 * 1000 + 1] <- 1/mean(hill1[a:(a * r)])
}
for(i in (u2 * 1000):(u1 * 1000)) {
a <- n^(i/1000)
f[i - l1 * 1000 + 1] <- hill[a]
f1[i - l1 * 1000 + 1] <- NA
s[i - l1 * 1000 + 1] <- i/1000
}
plot(s, f1, type = "l", xlab = "theta",
            ylab ="Hill estimate of alpha",
            main = "AltsmooHill")
par(mfrow = c(1, 1))
}
```

7.3. **Some alternative estimators from extreme value theory.** Suppose $\{Z_n, n \geq 1\}$ is iid with common distribution F. The distribution F is in the domain of attraction of the extreme value distribution G_γ, written $F \in \mathcal{D}(G_\gamma)$, if there exist $a(n) > 0,\ b(n) \in \mathbb{R}$ such that

$$n\mathbb{P}[Z_1 > a(n)x + b(n)] \to -\log G_\gamma(x) = (1+\gamma x)^{-1/\gamma}, \quad \gamma \in \mathbb{R},\ 1+\gamma x > 0. \tag{7.21}$$

Call

$$\mathbb{E}_\gamma = \{x : 1 + \gamma x > 0\}.$$

For $\gamma = 0$, we interpret $-\log G_\gamma(x) = e^{-x}$. Note

$$\mathbb{E}_\gamma = \begin{cases} (-\frac{1}{\gamma}, \infty], & \text{if } \gamma > 0, \\ (-\infty, \infty], & \text{if } \gamma = 0, \\ (-\infty, \frac{1}{|\gamma|}, & \text{if } \gamma < 0. \end{cases}$$

The heavy tailed case corresponds to $\gamma > 0$ and then $\gamma = 1/\alpha$.

Note (7.21) is a vague convergence statement about mean measures converging and we therefore have on $M_+(\mathbb{E}_\gamma)$

$$\frac{1}{k}\sum_{i=1}^{n} \epsilon_{\frac{Z_i - b(n/k)}{a(n/k)}} \Rightarrow \nu^{(\gamma)}, \tag{7.22}$$

where $\nu^{(\gamma)}(x,\infty] = -\log G_\gamma(x)$. Repeating the procedure that yielded (7.10) gives the equivalent statement

$$\frac{Z_{(\lceil k/y \rceil)} - b(n/k)}{a(n/k)} \to \frac{y^\gamma - 1}{\gamma}, \quad 0 \le y < \infty, \tag{7.23}$$

in $D[0,\infty)$.

7.3.1. *The Pickands estimator.* The Pickands estimator (Dekkers and de Haan (1989), Pickands (1975), Peng (1998)) like the de Haan moment estimator, is a semiparametric estimator of γ when we only assume $F \in \mathcal{D}(G_\gamma)$. It uses differences of quantiles. The Pickands estimator of γ based on using k upper order statistics from a sample of size n is

$$\hat{\gamma}_{k,n}^{(\text{Pickands})} = \left(\frac{1}{\log 2}\right) \log\left(\frac{Z_{(k)} - Z_{(2k)}}{Z_{(2k)} - Z_{(4k)}}\right). \tag{7.24}$$

Properties of the Pickands estimator.

(1) The Pickands estimator is a consistent estimator for $\gamma \in \mathbb{R}$ and does not require the assumption $\gamma > 0$ as does the Hill estimator. The consistency holds as $n \to \infty$, $k \to \infty$ and $n/k \to \infty$.

We can check consistency easily using (7.23). We have

$$\begin{aligned}\frac{Z_{(k)} - Z_{(2k)}}{Z_{(2k)} - Z_{(4k)}} &= \frac{\frac{(Z_{(k)} - b(n/k))}{a(n/k)} - \frac{(Z_{(2k)} - b(n/k))}{a(n/k)}}{\frac{(Z_{(2k)} - b(n/k))}{a(n/k)} - \frac{(Z_{(4k)} - b(n/k))}{a(n/k)}} \\ &\xrightarrow{P} \left(\frac{0 - \gamma^{-1}\left((\frac{1}{2})^\gamma - 1\right)}{\gamma^{-1}\left((\frac{1}{2})^\gamma - 1\right) - \gamma^{-1}\left((\frac{1}{4})^\gamma - 1\right)}\right) \\ &= 2^\gamma.\end{aligned}$$

Taking logarithms and dividing by $\log 2$ gives convergence in probability of the estimator to γ.

(2) Usually (under conditions which are uncheckable in practice), if $k \to \infty$, and $k/n \to 0$ we have asymptotic normality

$$\sqrt{k}(\hat{\gamma}_{k,n}^{(\text{Pickands})} - \gamma) \Rightarrow N(0, v(\gamma))$$

where

$$v(\gamma) = \frac{\gamma^2(2^{2\gamma+1} + 1)}{(2(2^\gamma - 1)\log 2)^2}.$$

More on asymptotic normality later.

(3) Unlike the Hill estimator, the Pickands estimator is location invariant. It is also scale invariant.

(4) Good plots may require a large sample of the order of several thousand.

(5) In terms of asymptotic mean squared error, the Pickands estimator sometimes is preferred over the moment estimator and Hill estimator (where comparable because you know $\gamma > 0$) and sometimes not. See Peng (1998).

(6) One can make a *Pickands plot* consisting of the points $\{(k, \hat{\gamma}_{k,n}^{(\text{Pickands})}), 1 \leq k < n/4\}$. Choice of k and volatility of the plots are issues as they were with the Hill and moments estimators.

Consider Figure 12 which is the Pickands estimator applied to 10,000 simulated Pareto random variables with $\alpha = 1$. The Pickands plot on the left picks up the correct value of $\alpha = 1$ quite well. In contrast to the degradation in the Hill plots when the data was shifted (recall Figure 9), the Pickands plot is unaffected.

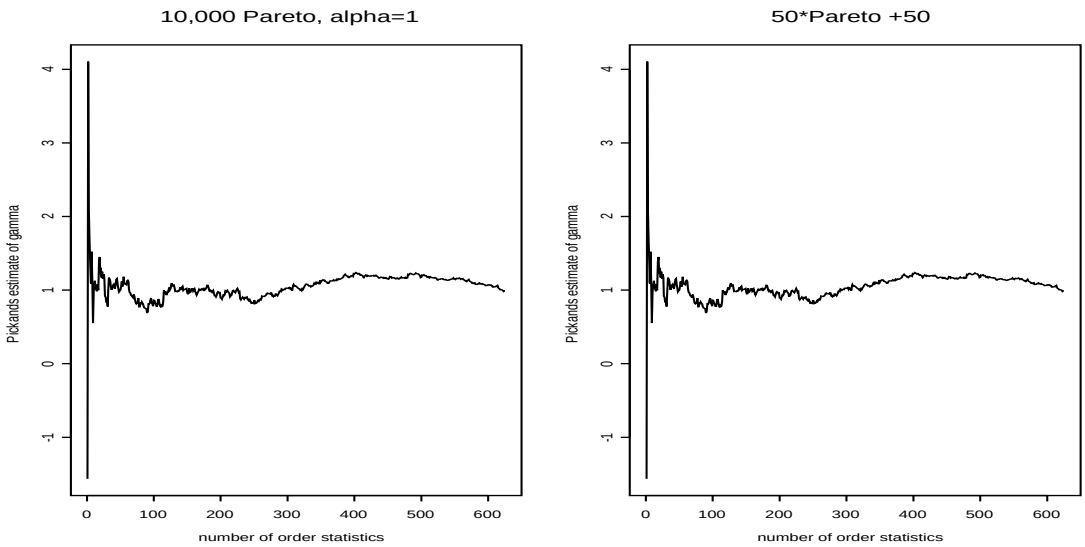

FIGURE 12. Pickands plots of 10,000 simulated Pareto random variables with $\alpha = 1$ (left) and with the same data but multiplied by 50 and shifted by 50.

Earlier, we had found $\alpha \approx 1.4$ for the Danish data. The plot in Figure 13 is not very informative even after accounting for the relation between α and γ and taking reciprocals $1/1.4 = 0.71$.

One more example where the Pickands plot does better. Consider returns on the BMW share prices. Recall from Section 2 that to get returns, we take the time series of the BMW share prices, take logarithms and difference at lag 1. On the left in Figure 14 is the full set of BMW returns and on the right are the positive returns.

The Pickands estimator does pretty well with the positive returns as shown in Figure 15, the Pickands plot of the positive BMW returns.

The plots were produced with the Splus function *Pickands*:

```
function(x)
{
ordered <- rev(sort(x))
n <- length(ordered)
ordered2k <- ordered[seq(2, (n/4), by = 2)]
ordered4k <- ordered[seq(4, (n/4), by = 4)]
```

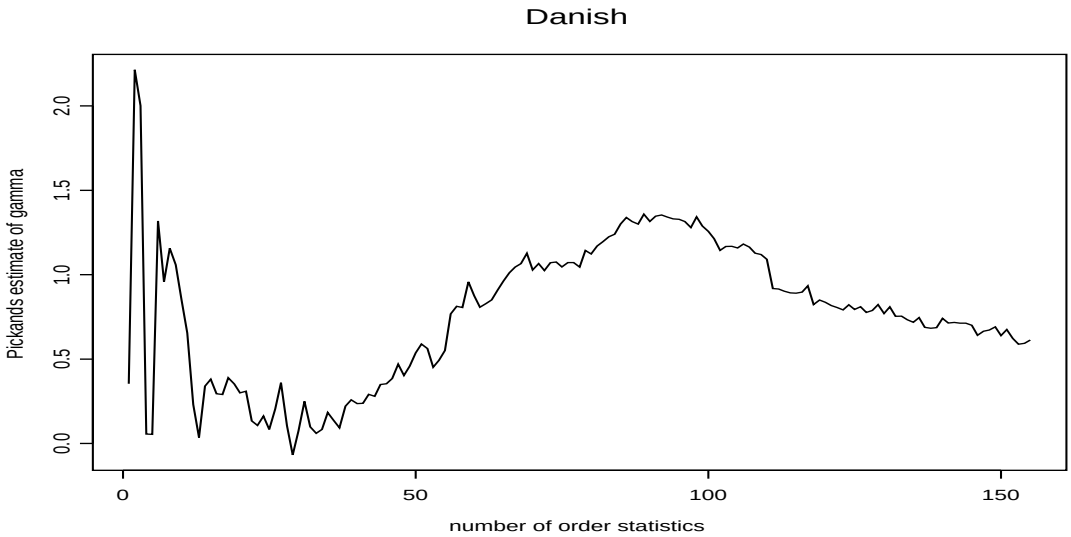

FIGURE 13. Pickands plots of the Danish.all data (left) where the estimate of $\gamma \approx 0.71$ was obtained from other methods.

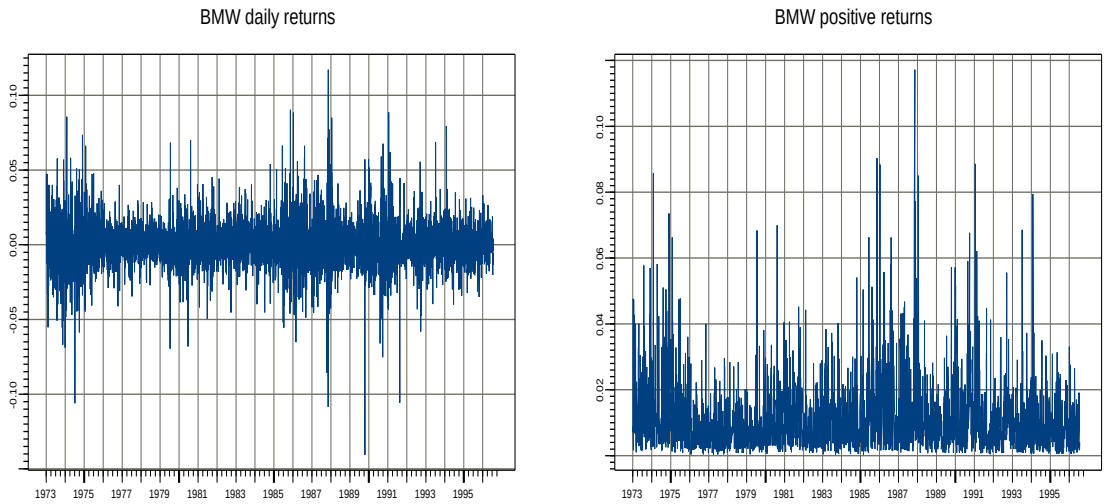

FIGURE 14. BMW returns (left) and positive returns (right).

```
l <- length(ordered4k)
gammak <- (1/log(2)) * log((ordered[1:l] - ordered2k[1:l])/(
ordered2k[1:l] - ordered4k[1:l]))
plot(1:length(gammak), gammak, type = "l", xlab =
"number of order statistics", ylab =
```

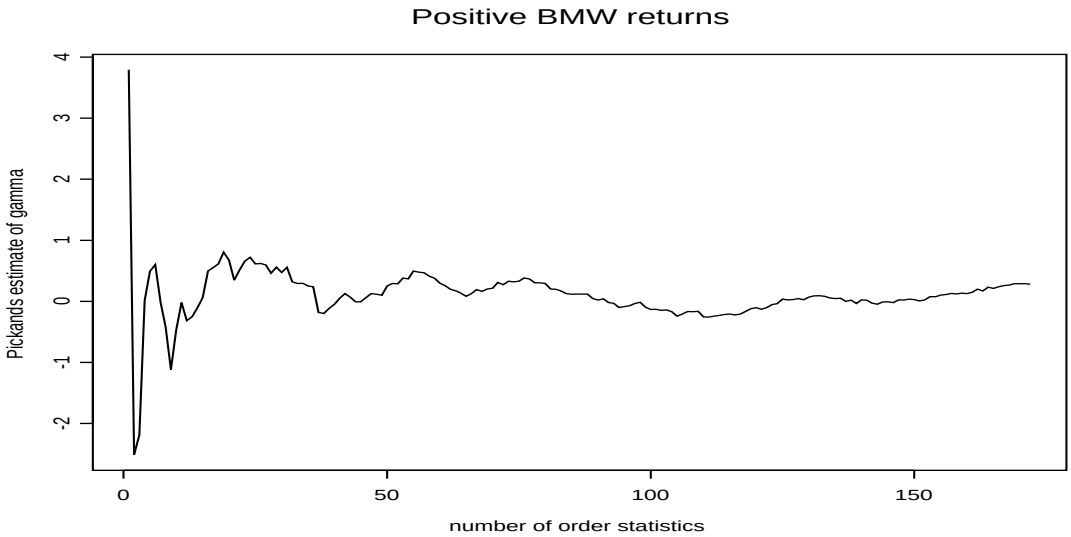

FIGURE 15. Pickands plot of the positive BMW returns.

```
"Pickands estimate of gamma")
}
```

REFERENCES

S. Coles. *An Introduction to Statistical Modeling of Extreme Values.* Springer Series in Statistics. London: Springer. xiv, 210 p. , 2001.

A. Dekkers and L. de Haan. On the estimation of the extreme-value index and large quantile estimation. *Ann. Statist.*, 17:1795–1832, 1989.

H. Drees, L. de Haan, and S. Resnick. How to make a Hill plot. *Ann. Statistics*, 28(1): 254–274, 2000.

P. Hall. On some simple estimates of an exponent of regular variation. *J. Roy. Stat. Assoc.*, 44:37–42, 1982. Series B.

J. Pickands. Statistical inference using extreme order statistics. *Ann. Stat.*, 3:119–131, 1975.

L. Peng. *Second Order Condition and Extreme Value Theory.* PhD thesis, Tinbergen Institute, Erasmus University, Rotterdam, 1998.

R.-D. Reiss and M. Thomas. *Statistical Analysis of Extreme Values.* Birkhäuser Verlag, Basel, second edition, 2001. ISBN 3-7643-6487-4.

S. Resnick and C. Stărică. Smoothing the Hill estimator. *Adv. Applied Probability*, 29: 271–293, 1997.

8. Asymptotic Normality of the Tail Empirical Measure

As in Section 7.1, suppose $\{X_j, j \geq 1\}$ are iid, non-negative random variables with common distribution $F(x)$, where $\bar{F} \in RV_{-\alpha}$ for $\alpha > 0$. Continue with the notation in (7.1), (7.2), and (7.3). Define the tail empirical process,

$$\begin{aligned} W_n(y) =& \sqrt{k}\Big(\frac{1}{k}\sum_{i=1}^{n} \epsilon_{X_i/b(n/k)}(y^{-1/\alpha}, \infty] - \frac{n}{k}\bar{F}(b(n/k)y^{-1/\alpha})\Big) \\ =& \sqrt{k}\Big(\nu_n(y^{-1/\alpha}, \infty] - \mathbf{E}(\nu_n(y^{-1/\alpha}, \infty])\Big). \end{aligned} \tag{8.1}$$

Theorem 7. *If* (7.1), (7.2), *and* (7.3) *hold, then in* $D[0, \infty)$

$$W_n \Rightarrow W$$

where W *is Brownian motion on* $[0, \infty)$.

Remark 3. Note, because of regular variation, as $n \to \infty$, $k/n \to 0$,

$$E\nu_n(y^{-1/\alpha}, \infty] = \frac{n}{k}\bar{F}(b(n/k)y^{-1/\alpha}) \to \left(y^{-1/\alpha}\right)^{-\alpha} = y. \tag{8.2}$$

For applications to such things as the asymptotic normality of the Hill estimator and other estimators derived from the tail empirical measure, we would prefer the centering in (8.1) be y. However, to make this substitution in (8.1) requires knowing or assuming

$$\lim_{n\to\infty} \sqrt{k}\Big(\frac{n}{k}\bar{F}(b(n/k)y^{-1/\alpha}) - y\Big) \tag{8.3}$$

exists finite. This is one of the origins for the need of *second order regular variation*.

Proof. The proof requires several steps:

Step 1: Donsker theorem. Suppose $\{\xi_j, j \geq 1\}$ are iid, $\mathbf{E}(\xi_j) = 0$, and $\mathrm{Var}(\xi_j) = 1$. Then in $D[0, \infty)$,

$$\sum_{i=1}^{[n\cdot]} \frac{\xi_i}{\sqrt{n}} \Rightarrow W.$$

Step 2: Vervaat's lemma. See Vervaat (1972b) or Vervaat (1972a). Suppose $x_n \in D[0, \infty)$ and $x_\infty \in C[0, \infty)$ and x_n are non-decreasing. If $c_n \to \infty$ and

$$c_n(x_n(t) - t) \to x_\infty(t), \quad (n \to \infty)$$

locally uniformly, then also

$$c_n(x_n^{\leftarrow}(t) - t) \to -x_\infty(t), \quad (n \to \infty)$$

locally uniformly.

From Skorohod's theorem, we get the following stochastic processes version: Suppose X_n is a sequence of $D[0, \infty)$ valued random elements and X_∞ has continuous paths. If X_n has non-decreasing paths and if $c_n \to \infty$, then

$$c_n(X_n(t) - t) \Rightarrow X_\infty(t), \quad (n \to \infty)$$

in $D[0, \infty)$ implies

$$c_n(X_n^{\leftarrow}(t) - t) \Rightarrow -X_\infty(t), \quad (n \to \infty)$$

in $D[0,\infty)$. In fact, we also have $(e(t)=t)$

$$\Big(c_n(X_n(\cdot)-e, c_n(X_n^{\leftarrow}(\cdot)-e\Big) \Rightarrow (X_\infty(\cdot), X_\infty^{\leftarrow}(\cdot)), \tag{8.4}$$

in $D[0,\infty)\times D[0,\infty)$.

Step 3: Renewal theory. Suppose $\{Y_n, n\ge 1\}$ are iid, non-negative random variables with $\mathbf{E}(Y_j)=\mu$, and $\mathrm{Var}(Y_j)=\sigma^2$. Set $S_n=\sum_{i=1}^n Y_i$. Then Donsker's theorem says

$$\frac{S_{[nt]}-[nt]\mu}{\sigma\sqrt{n}} \Rightarrow W(t)$$

in $D[0,\infty)$. Since for any $M>0$

$$\sup_{0\le t\le M}\frac{|nt\mu-[nt]\mu|}{\sqrt{n}}\to 0,$$

it is also true that in $D[0,\infty)$

$$\frac{S_{[nt]}-nt\mu}{\sigma\sqrt{n}} \Rightarrow W(t)$$

or by dividing numerator and denominator by $n\mu$

$$c_n(X_n(t)-t) := \frac{\Big(\frac{S_{[nt]}}{n\mu}-t\Big)}{\sigma n^{-1/2}/\mu} \Rightarrow W(t).$$

This implies the result for $X_n^{\leftarrow}$ and we need to evaluate this process:

$$\begin{aligned} X_n^{\leftarrow}(t) &= \inf\{s: X_n(s)\ge t\} \\ &= \inf\{s: S_{[ns]}/n\mu \ge t\} = \inf\{s: S_{[ns]}\ge tn\mu\} \\ &= \inf\{\frac{j}{n}: S_j\ge tn\mu\} = \frac{1}{n}N(tn\mu), \end{aligned}$$

some version of the renewal function. For later use, note that $N(t)$ differs at most by 1 from

$$\sum_{i=1}^\infty 1_{[S_j\le t]}. \tag{8.5}$$

The conclusion from Vervaat is

$$\sqrt{n}\frac{\mu}{\sigma}\Big(\frac{1}{n}N(n\mu t)-t\Big) \Rightarrow W(t)$$

or changing variables $s=\mu t$

$$\sqrt{n}\frac{\mu}{\sigma}\Big(\frac{1}{n}N(ns)-\frac{s}{\mu}\Big) \Rightarrow W(\frac{s}{\mu}) \stackrel{d}{=} \frac{1}{\sqrt{\mu}}W(s).$$

The conclusion:

$$\sqrt{n}\frac{\mu^{3/2}}{\sigma}\Big(\frac{1}{n}N(ns)-\frac{s}{\mu}\Big) \Rightarrow W(s), \tag{8.6}$$

in $D[0,\infty)$.

Special case: The Poisson process. Let

$$\Gamma_n = E_1 + \cdots + E_n$$

be a sum of n iid standard exponential random variables. In this case $\mu = \sigma = 1$ and

$$\sqrt{k}\Big(\frac{1}{k}N(ks) - s\Big) \Rightarrow W(s), \quad (k \to \infty), \tag{8.7}$$

in $D[0,\infty)$.

Step 4: Approximation. Review the relation of $N(t)$ with the quantity in (8.5). We claim, as $n \to \infty$, and $k/n \to 0$, for any $T > 0$,

$$\sup_{0\le s\le T} \sqrt{k}\Big|\frac{1}{k}\sum_{i=1}^{\infty} 1_{[\Gamma_i \le ks]} - \frac{1}{k}\sum_{i=1}^{n} 1_{[\Gamma_i \le ks]}\Big| \xrightarrow{P} 0. \tag{8.8}$$

The idea is that Γ_i is localized about its mean and any term with i too far from k is unlikely. The result says that $i > n$ gives a term which is negligible. More formally, the difference in (8.8) is

$$\begin{aligned}\sup_{0\le s\le T} \frac{1}{\sqrt{k}}\sum_{i=n+1}^{\infty} 1_{[\Gamma_i \le ks]} \le & \frac{1}{\sqrt{k}}\sum_{i=n+1}^{\infty} 1_{[\Gamma_i \le kT]} \\ = & \frac{1}{\sqrt{k}}\sum_{i=1}^{\infty} 1_{[\Gamma_n + \Gamma_i' \le kT]} \qquad (\Gamma_i' = \sum_{l=1}^{i} E_{l+n}).\end{aligned}$$

Now for any $\delta > 0$,

$$P[\frac{1}{\sqrt{k}}\sum_{i=1}^{\infty} 1_{[\Gamma_n + \Gamma_i' \le kT]} > \delta] \le P[\Gamma_n \le kT] = P[\frac{\Gamma_n}{n} \le \frac{k}{n}T]$$

and since $k/n \to 0$, for any $\eta > 0$ we ultimately have this last term bounded by

$$\le P[\frac{\Gamma_n}{n} \le 1 - \eta] \to 0,$$

by the weak law of large numbers.

Combining (8.8), the definition of N, and (8.7) we get the <u>conclusion</u>

$$\sqrt{k}\Big(\frac{1}{k}\sum_{i=1}^{n} 1_{[\Gamma_i \le ks]} - s\Big) \Rightarrow W(s) \qquad (k \to \infty,\ k/n \to 0), \tag{8.9}$$

in $D[0,\infty)$.

Step 5: Time change. For $s \ge 0$, define

$$\phi_n(s) = \frac{n}{k}\bar{F}\big(b(n/k)s^{-1/\alpha}\big)\frac{\Gamma_{n+1}}{n}$$

so that from regular variation and the weak law of large numbers,

$$\sup_{0\le s\le T} |\phi_n(s) - s| \xrightarrow{P} 0 \tag{8.10}$$

for any $T > 0$. Therefore, joint convergence ensues

$$\Big(\sqrt{k}\Big(\frac{1}{k}\sum_{i=1}^{n} 1_{[\Gamma_i \le k\cdot]} - (\cdot), \phi_n(\cdot)\Big) \Rightarrow (W, e), \qquad (e(t) = t)$$

in $D[0,\infty) \times D[0,\infty)$ and applying composition we arrive at

$$\sqrt{k}\Big(\frac{1}{k}\sum_{i=1}^{n} 1_{[\Gamma_i \le k\phi_n(s)]} - \phi_n(s)\Big) \Rightarrow W(s) \tag{8.11}$$

in $D[0,\infty)$.

Step 6: Probability integral transform. The Γ's have the property that

$$\Big(\frac{\Gamma_1}{\Gamma_{n+1}}, \dots \frac{\Gamma_n}{\Gamma_{n+1}}\Big) \stackrel{d}{=} \Big(1 - \frac{\Gamma_n}{\Gamma_{n+1}}, \dots, 1 - \frac{\Gamma_1}{\Gamma_{n+1}}\Big)\Big) \stackrel{d}{=} (U_{(1:n)}, \dots, U_{(n:n)})$$

where

$$U_{(1:n)} \le \cdots \le U_{(n:n)})$$

are the order statistics in increasing order of n iid $U(0,1)$ random variables $U_1, \dots, U_n$.

Observe from (8.11),

$$\begin{aligned}
\frac{1}{k}\sum_{i=1}^{n} 1_{[\Gamma_i \le k\phi_n(s)]} =& \frac{1}{k}\sum_{i=1}^{n} 1_{[\frac{\Gamma_i}{k} \le \frac{1}{k}\bar{F}(b(n/k)s^{-1/\alpha})\Gamma_{n+1}]} = \frac{1}{k}\sum_{i=1}^{n} 1_{[\frac{\Gamma_i}{\Gamma_{n+1}} \le \bar{F}(b(n/k)s^{-1/\alpha})]} \\
=& \frac{1}{k}\sum_{i=1}^{n} 1_{[F(b(n/k)s^{-1/\alpha}) \le 1 - \frac{\Gamma_i}{\Gamma_{n+1}}]} = \frac{1}{k}\sum_{i=1}^{n} 1_{[b(n/k)s^{-1/\alpha} \le F^{\leftarrow}\left(1 - \frac{\Gamma_i}{\Gamma_{n+1}}\right)]} \\
\stackrel{d}{=}& \frac{1}{k}\sum_{i=1}^{n} 1_{[b(n/k)s^{-1/\alpha} \le F^{\leftarrow}\left(U_{(i:n)}\right)]} = \frac{1}{k}\sum_{i=1}^{n} 1_{[b(n/k)s^{-1/\alpha} \le F^{\leftarrow}(U_i)]} \\
\stackrel{d}{=}& \frac{1}{k}\sum_{i=1}^{n} 1_{[b(n/k)s^{-1/\alpha} \le X_i)]} = \frac{1}{k}\sum_{i=1}^{n} 1_{[\frac{X_i}{b(n/k)} \ge s^{-1/\alpha}]} = \nu_n[s^{-1/\alpha}, \infty].
\end{aligned}$$

Also,

$$\begin{aligned}
\sqrt{k} \sup_{0 \le s \le T} \Big| \frac{n}{k}\bar{F}\big(b(n/k)s^{-1/\alpha}\big)\frac{\Gamma_{n+1}}{n} - \frac{n}{k}\bar{F}(b(n/k)s^{-1/\alpha}\Big| \\
= \sup_{0 \le s \le T} \frac{n}{k}\bar{F}(b(n/k)s^{-1/\alpha})\sqrt{k}\Big|\frac{\Gamma_{n+1}}{n} - 1\Big| \\
= O(1)\sqrt{\frac{k}{n}}\Big|\frac{\Gamma_{n+1} - n}{\sqrt{n}}\Big| = O(1)o(1)O_p(1),
\end{aligned}$$

from the central limit theorem, and this $\stackrel{P}{\to} 0$.

This proves the result since the last statement removes the difference between $\phi_n(s)$ and $\mathbf{E}\big(\nu_n[s^{-1/\alpha}, \infty]\big)$. □

From this result we can recover Theorem 5 page 41 and its consequences.

8.1. **Asymptotic normality of the Hill estimator.** For this section it is convenient to assume (8.3) and, in fact, we assume for simplicity

$$\lim_{n\to\infty} \sqrt{k}\Big(\frac{n}{k}\bar{F}(b(n/k)y^{-1/\alpha}) - y\Big) = 0 \tag{8.12}$$

uniformly for $x > x_0$. (The uniformity is not an extra assumption but this requires proof.)

With (8.12), we can modify the result of Theorem 7 to

$$W_n^{\#}(y) = \sqrt{k}\Big(\frac{1}{k}\sum_{i=1}^{n} 1_{[X_i/b(n/k)>y^{-1/\alpha}]} - y\Big) =: \sqrt{k}(V_n(y) - y) \Rightarrow W(y), \tag{8.13}$$

in $D[0,\infty)$. Therefore, from Vervaat's lemma,

$$\sqrt{k}(V_n^{\leftarrow}(y) - y) = \sqrt{k}\bigg(\Big(\frac{X_{\lceil ky\rceil}}{b(n/k)}\Big)^{-\alpha} - y\bigg) \Rightarrow -W(y). \tag{8.14}$$

In fact we have joint convergence

$$\bigg(\sqrt{k}\big(V_n(\cdot) - e\big), \sqrt{k}\big(V_n^{\leftarrow}(\cdot) - e\big), \Big(\frac{X_{(k)}}{b(n/k)}\Big)^{-\alpha}\bigg) \Rightarrow (W, -W, 1) \tag{8.15}$$

in $D[0,\infty) \times D[0,\infty) \times \mathbb{R}$. Apply the map,

$$(x_1(\cdot), x_2(\cdot), k) \mapsto (x_1(k\cdot), x_2(1)e)$$

to get

$$\bigg(\sqrt{k}\Big(\frac{1}{k}\sum_{i=1}^{n} 1_{[\frac{X_i}{b(n/k)} > \frac{X_{(k)}}{b(n/k)}y^{-1/\alpha}]} - \big(\frac{X_{(k)}}{b(n/k)}\big)^{-\alpha}y\Big), \sqrt{k}\Big(\frac{X_{(k)}}{b(n/k)}\big)^{-\alpha}y - y\Big)\bigg) \Rightarrow \big(W(y), -yW(1)\big).$$

Add the components to get,

$$\sqrt{k}\Big(\frac{1}{k}\sum_{i=1}^{n} 1_{[\frac{X_i}{b(n/k)} > \frac{X_{(k)}}{b(n/k)}y^{-1/\alpha}]} - y\Big) \Rightarrow W(y) - yW(1). \tag{8.16}$$

This removes the unknown function $b(n/k)$. The limit is Brownian bridge.

Apply the map (needs justification by means of the Second converging together theorem, Proposition 7 on page 48)

$$x(\cdot) \mapsto \int_0^1 x(s)\frac{ds}{\alpha s}.$$

The result is

$$\sqrt{k}(H_{k,n} - \frac{1}{\alpha}) \Rightarrow \frac{1}{\alpha}[\int_0^1 W(s)\frac{ds}{s} - W(1)]. \tag{8.17}$$

Lemma 4. *The random variable*

$$\int_0^1 W(s)\frac{ds}{s} - W(1)$$

is $N(0,1)$.

Proof. The integral is a Gaussian random variable (it is a limit of linear combinations of Gaussian random variables) so we just calculate the variance: We use

$$\mathbf{E}(W(s)W(t)) = s \wedge t.$$

Then

$$\begin{aligned}\mathbf{E}[\int_0^1 W(s)\frac{ds}{s} - W(1)]^2 =& \mathbf{E}\Big(\int_0^1 \cdot ds \int_0^1 du - 2\int_0^1 W(s)W(1)\frac{ds}{s} + W(1)^2\Big)\\ =& 2\iint_{0\le s<u\le 1}(s\wedge u\frac{ds}{s}\frac{du}{u} - 2\int_0^1 s\frac{ds}{s} + 1\\ =& 2\int_{u=0}^1\Big(\int_{s=0}^u s\frac{ds}{s}\Big)\frac{du}{u} - 2 + 1 = 1.\end{aligned}$$

□

Conclusion: With regular variation and the 2nd order condition,

$$\sqrt{(}H_{k,n} - \frac{1}{\alpha}) \Rightarrow N(0, \frac{1}{\alpha^2}).$$

9. More on Point Processes.

We continue to explore the role that point processes plays both as modeling elements and as a transform that allows us to understand regular variation in a dimensionless way.

9.1. **The Poisson process.** Let N be a point process with nice (locally compact, countable base) state space $\mathbb{E}$. Suppose $\mathcal{E}$ is a class of reasonable subsets of $\mathbb{E}$; that is, the Borel σ-algebra of subsets of $\mathbb{E}$ generated by the open sets in $\mathbb{E}$.

Definition 2. N is a *Poisson process with mean measure* μ or synonomously a *Poisson random measure (PRM(μ))* if

(1) For $A \in \mathcal{E}$

$$P[N(A) = k] = \begin{cases}\frac{e^{-\mu(A)}(\mu(A))^k}{k!}, & \text{if } \mu(A) < \infty\\ 0, & \text{if } \mu(A) = \infty.\end{cases}$$

(2) If $A_1, \dots, A_k$ are disjoint subsets of $\mathbb{E}$ in $\mathcal{E}$, then $N(A_1), \dots, N(A_k)$ are independent random variables.

So N is Poisson if the random number of points in a set A is Poisson distributed with parameter $\mu(A)$ and the number of points in disjoint regions are independent random variables.

Property 2 is called *complete randomness*. When $\mathbb{E} = \mathbb{R}$ it is called the *independent increments* property since we have for any $t_1 < t_2 < \cdots < t_k$ that $(N((t_i, t_{i+1}]), 1 = 1, \dots, k-1)$ are independent random variables. When the mean measure is a multiple of Lebesgue measure (ie, length when $\mathbb{E} = [0, \infty)$ or $\mathbb{R}$, area when $\mathbb{E} = \mathbb{R}^2$, volume when $\mathbb{E} = \mathbb{R}^3$, etc) we call the process *homogeneous.* Thus in the homogeneous case, there is a parameter $\lambda > 0$ such that for any A we have $N(A)$ Poisson distributed with mean $EN(A) = \lambda|A|$ where $|A|$

is the Lebesgue measure of A. When $\mathbb{E} = [0, \infty)$ the parameter λ is called the *rate* of the (homogeneous) Poisson process.

9.1.1. *Transformations of Poisson processes.* Useful results are connected with a circle of ideas about what happens to a Poisson process under various types of transformations. The first result, though very elementary, is enormously useful in understanding inhomogeneity. To prepare for this result, suppose $\sum_n \epsilon_{X_n}$ is a Poisson process with state space $\mathbb{E}$ and mean measure μ. Suppose T is some transformation with domain $\mathbb{E}$ and range $\mathbb{E}'$, where $\mathbb{E}'$ is another nice space; that is,

$$T : \mathbb{E} \mapsto \mathbb{E}'.$$

The function T defines a set mapping of subsets of $\mathbb{E}'$ to subsets of $\mathbb{E}$, defined for $A' \subset \mathbb{E}'$ by

$$T^{-1}(A') = \{e \in E : T(e) \in A'\}.$$

Thus $T^{-1}(A')$ is the pre-image of A' under T; that is, it is the set of points of $\mathbb{E}$ which T maps into A'.

As an example, suppose $\mathbb{E} = (0, \infty)$, $\mathbb{E}' = (-\infty, \infty), T(x) = \log x$. If $a < b$ and $A' = (a, b)$ we have

$$\begin{aligned} T^{-1}((a,b)) =& \{x > 0 : T(x) \in (a,b)\} = \{x > 0 : \log x \in (a,b)\} \\ & \{x > 0 : x \in (e^a, e^b)\}. \end{aligned}$$

Given the measures N, μ defined on subsets of $\mathbb{E}$, we may use T to define induced measures N', μ' on subsets of $\mathbb{E}'$. For $A' \subset E'$ define

$$N'(A') = N(T^{-1}(A')), \quad \mu'(A') = \mu(T^{-1}(A')).$$

So to get the measure of A', we map A' back into $\mathbb{E}$ and take the measure of the pre-image under T. Also, if N has points $\{X_n\}$, then N' has points $\{X'_n\} = \{T(X_n)\}$ since for $A' \subset \mathbb{E}'$

$$\begin{aligned} N'(A') =& N(T^{-1}(A')) = \sum_n \epsilon_{X_n}(T^{-1}(A')) \\ =& \sum_n 1_{[X_n \in T^{-1}(A')]} = \sum_n 1_{[T(X_n) \in A']} = \sum_n \epsilon_{T(X_n)}(A'). \end{aligned}$$

The next result asserts that if N is a Poisson process with mean measure μ and with points $\{X_n\}$ living in the state space $\mathbb{E}$, then $N' = N(T^{-1}(\cdot))$ is a Poisson process process with mean measure μ' and with points $\{T(X_n)\}$ living in the state space $\mathbb{E}'$.

Proposition 8. *Suppose*

$$T : \mathbb{E} \mapsto \mathbb{E}'$$

is a mapping of one Euclidean space $\mathbb{E}$ into another $\mathbb{E}'$ such that if $K' \in \mathcal{K}(\mathbb{E}')$ is compact in $\mathbb{E}'$ then so is $T^{-1}K' := \{e \in E : Te \in K'\} \in \mathcal{K}(\mathbb{E})$. If N is PRM(μ) on $\mathbb{E}$ then $N' := N \circ T^{-1}$ is PRM(μ') on $\mathbb{E}'$ where $\mu' := \mu \circ T^{-1}$.

Remember that if N has the representation

$$N = \sum_n \epsilon_{X_n}$$

then

$$N' = \sum_n \epsilon_{TX_n}$$

so this result says that if you shift the points of a Poisson process around you still get a Poisson process.

Proof. We have

$$P[N'(B') = k] = P[N(T^{-1}(B')) = k] = p(k, \mu(T^{-1}(B')))$$

so N' has Poisson distributions. It is easy to check the independence property since if $B'_1, \ldots, B'_m$ are disjoint then so are $T^{-1}(B'_1), \ldots, T^{-1}(B_m)$ whence

$$(N'(B'_1), \ldots, N'(B'_m)) = \big(N(T^{-1}(B'_1), \ldots, N(T^{-1}(B'_m)\big)$$

are independent. Thus the postulates 1 and 2 of a Poisson process are satisfied. □

Examples: Consider 3 easy three examples. For each, let $N = \sum_{n=1}^{\infty} \epsilon_{\Gamma_n}$ be a homogeneous Poisson process with rate $\lambda = 1$ on the state space $\mathbb{E} = [0, \infty)$. The mean measure μ is Lebesgue measure so that $\mu(A) = |A|$ and in particular $\mu([0, t]) = t$.

(1) If $Tx = x^2$ then $\sum_n \epsilon_{\Gamma_n^2}$ is PRM and the mean measure μ' is given by

$$\mu'[0, t] = \mu\{x : Tx \leq t\} = \mu[0, \sqrt{t}] = \sqrt{t}.$$

Note that μ' has a density

$$\alpha(t) = \frac{d}{dt}\sqrt{t} = \frac{1}{2}t^{-1/2}.$$

(2) If $T : \mathbb{E} \mapsto \mathbb{E} \times \mathbb{E}$ via $Tx = (x, x^2)$ then $\sum_n \epsilon_{T\Gamma_n} = \sum_n \epsilon_{(\Gamma_n, \Gamma_n^2)}$ is Poisson on $\mathbb{E} \times \mathbb{E}$. The mean measure concentrates on the graph $\{(x, x^2) : x \geq 0\}$.

(3) If $\sum_n \epsilon_{\Gamma_n}$ is homogeneous Poisson on $[0, \infty)$ then $\sum_n \epsilon_{\Gamma_n^{-1}}$ is Poisson on $(0, \infty]$ with mean measure μ' given by $(x > 0)$

$$\mu'(x, \infty] = \mu\{t \geq 0 : t^{-1} \geq x\} = \mu[0, x^{-1}) = x^{-1}.$$

Note that the bounded sets of $\mathbb{E}'$ are those sets bounded away from 0; ie the bounded sets are neighborhoods of ∞. μ' has a density

$$\alpha(t) = -\frac{d}{dt}t^{-1} = t^{-2}.$$

Poisson processes with this mean measure μ' are particularly important in extreme value theory and in the theory of stable processes. It is easy to see the connection with extreme value theory: Given $N' = \sum_n \epsilon_{X'_n}$, the Poisson process on $(0, \infty]$ with mean measure μ', define

$$Y = \bigvee_n X'_n.$$

So Y is simply the biggest point of the Poisson process. The distribution of Y is easily computed: For $x > 0$

$$P[Y \leq x] = P[\bigvee_n X'_n \leq x]$$

$$
\begin{aligned}
&= P[N'((x,\infty]) = 0] \\
&= e^{-\mu'((x,\infty])} \\
&= e^{-x^{-1}}
\end{aligned}
$$

which is one of the classical extreme value distributions.

To see the connection with stable laws, it is simplest to look at a variant of the above construction: Let $0 < \alpha < 1$ and define the Poisson process

$$N^{\#} = \sum_{n=1}^{\infty} \epsilon_{\Gamma_n^{-1/\alpha}}.$$

The mean measure of $N^{\#}$ is $\mu^{\#}$ and since $N^{\#}$ is constructed from the transformation $T(x) = x^{-1/\alpha}$ we have that

$$\mu^{\#}((x,\infty]) = x^{-\alpha}.$$

Define the random variable

$$X = \sum_{n=1}^{\infty} \Gamma^{-1/\alpha}.$$

This is a *stable* random variable with index α. It is easy to see the series converges because $\Gamma_n \sim n$ as $n \to \infty$ by the law of large numbers. Since $1/\alpha > 1$, convergence of series defining X follows by comparison with the series $\sum_n n^{-1/\alpha}$.

A similar construction can be used to define stable random variables with indices $1 \le \alpha < 2$ but the terms $\Gamma_n^{-1/\alpha}$ must be centered before summing in order to guarantee convergence of the infinite series of random variables. This will be considered in more detail later when we discuss Lévy processes.

9.1.2. *Augmentation.* Given a Poisson process, under certain circumstances it is possible to enlarge the dimension of the points and retain the Poisson structure. One way to do this was given in example 2 of the previous section, but the enlargement of dimension was illusory since the points concentrated on a graph $\{(x, x^2) : x > 0\}$. The result presented here allows independent components to be added to the points of the Poisson process. This proves very useful in a variety of applications.

Proposition 9. *Suppose $\{X_n\}$ are random elements of a Euclidean space $\mathbb{E}_1$ such that*

$$\sum_n \epsilon_{X_n}$$

is PRM(μ). Suppose $\{J_n\}$ are iid random elements of a second Euclidean space $\mathbb{E}_2$ with common probability distribution $F(\cdot)$ and suppose the Poisson process and the sequence $\{J_n\}$ are defined on the same probability space and are independent. Then the point process on $\mathbb{E}_1 \times \mathbb{E}_2$

$$\sum_n \epsilon_{(X_n, J_n)}$$

is PRM with mean measure $\mu \times F(\cdot)$ meaning that if $A_i \subset \mathbb{E}_i,\ i = 1, 2$ are Borel sets, then $\mu \times F(A_1 \times A_2) = \mu \times F(\{(e_1, e_2) : e_1 \in A_1, e_2 \in A_2\}) = \mu(A_1)F(A_2)$.

Often this procedure is described by saying we give to point X_n the *mark* J_n. Think about a picture where the points of the original Poisson process $\{X_n\}$ appear on the horizontal axis and the marked points appear in the $\mathbb{E}_1 \times \mathbb{E}_2$ plane.

The proof is deferred. For now, note the mean measure is correct since for a rectangle set of the form $A_1 \times A_2 = \{(e_1, e_2) : e_1 \in A_1 \subset \mathbb{E}_1, e_2 \in A_2 \subset \mathbb{E}_2\}$ we have

$$\begin{aligned} E\sum_n \epsilon_{(X_n,J_n)}(A_1 \times A_2) &= \sum_n P[(X_n, J_n) \in A_1 \times A_2] \\ &= \sum_n P[X_n \in A_1]P[J_n \in A_2] \end{aligned}$$

since $\{J_n\}$ is independent of the Poisson process. Since $\{J_n\}$ are iid random variables this is the same as

$$\begin{aligned} &= \sum_n P[X_n \in A_1]P[J_1 \in A_2] \\ &= E(\sum_n \epsilon_{X_n}(A_1))P[J_1 \in A_2] \\ &= \mu(A_1)P[J_1 \in A_2]. \end{aligned}$$

10. The Infinite Node Poisson Model.

The infinite node Poisson model is a simple (probably too simple) model which enjoyed early success in explaining long range dependence in measured internet traffic. The explanation is simple probability based on properties of a Poisson process.

10.1. **Background.** The story begins around 1993 with the publication of what is now known as the Bellcore study (Duffy et al. (1993), Leland et al. (1993b), Willinger et al. (1995)). Traditional queueing models had thrived on assumptions of exponentially bounded tails, Poisson inputs and lots of independence. Collected network data studied at what was then Bellcore (now Telcordia) exhibited properties which were inconsistent with traditional queueing models. These anomalies were also found in world wide web downloads in the Boston University study (Crovella and Bestavros (1995, 1996a,b, 1997), Crovella et al. (1999, 1996), Cunha et al. (1995)). The unusual properties found in the data traces included:

- self-similarity (ss) and long-range dependence (LRD) of various transmission rates:
  - packet counts per unit time,
  - www bits/time.
- heavy tails of quantities such as
  - file sizes,
  - transmission rates,
  - transmission durations,
  - CPU job completion times,
  - call lengths

The Bellcore study in early 90's resulted in a paradigm shift worthy of a sociological study to understand the frenzy to jump on and off various bandwagons but after some resistence

to the presence of long range dependence, there was widespread acceptance of the statement that *packet counts per unit time exhibit self similarity and long range dependence.* Research goals then shifted from detection of the phenomena to greater understanding of the causes. The challenges were:

- Explain the origins and effects of long-range dependence and self-similarity.
- Understand some connections between self-similarity, long range dependence and heavy tails. Use these connections for finding an explanation for the perceived long range dependence in traffic measurements.
- Begin to understand the effect of network protocols and architecture on traffic. The simplist models, such as the featured infinite source Poisson model, pretend protocols and controls are absent. This is an ambitious goal.
- Say something useful for the purposes of capacity planning.

10.2. **The infinite node Poisson model.** Attempts to explain long range dependence and self-similarity in traffic rates centered around the paradigm: *heavy tailed file sizes cause LRD in network traffic.* Specific models must be used to explain this and the two most effective and simple models were:

- Superposition of on/off processes (Heath et al. (1997, 1998), Jelenković and Lazar (1998), Mikosch et al. (2002), Mikosch and Stegeman (1999), Parulekar and Makowski (1996a), Stegeman (1998), Taqqu et al. (1997), Willinger et al. (1995)). This is described as follows: imagine a source/destination pair. The source sends at unit rate for a random length of time to the destination and then is silent or inactive for a random period. Then the source sends again and when finished is silent. And so on. So the transmission schedule of the source follows an alternating renewal or on/off structure. Now imagine the traffic generated by many source/destination pairs being superimposed and this yields the overall traffic.
- The infinite source Poisson model, sometimes called the M/G/∞ input model (Guerin et al. (2003), Heath et al. (1999), Jelenković and Lazar (1996), Jelenković and Lazar (1999), Mikosch et al. (2002), Parulekar and Makowski (1996b), Resnick and Rootzén (2000), Resnick and van den Berg (2000)). Imagine infinitely many potential users connected to a single server which processes work at constant rate r. At a Poisson time point, some user begins transmitting work to the server at constant (ugh!) rate which, without loss of generality, we take to be rate 1. The length of the transmission is random with heavy tailed distribution. The length of the transmission may be considered to be the size of the file needing transmission.

Both models have their adherents and the two models are asymptotically equivalent in a manner nobody (to date) has made fully transparent. We will focus on the infinite source Poisson model.

Some good news about the model:

- It is somewhat flexible and certainly simple.
- Since each node transmits at unit rate, the overall transmission rate at time t is simply the number of active users $N(t)$ at t. From classical M/G/∞ queueing theory, we

know $N(t)$ is a Poisson random variable with mean $\lambda\mu_{\text{on}}$ where λ is the rate parameter of the Poisson process and μ_{on} is the mean file size or mean transmission length.
- The length of each transmission is random and heavy tailed.
- The model offers a very simple explanation of long range dependence being caused by heavy tailed file sizes.
- The model predicts traffic aggregated over users and accumulated over time $[0, T]$ is approximated by either a Gaussian process (fractional Brownian motion or FBM) or a heavy tailed stable Lévy motion (Mikosch et al. (2002)). Thus the two approximations are very different in character but at least both are self-similar.

Some less good news about the model:

- The model does not fit collected data traces all that well.
  - The constant transmission rate assumption is clearly wrong. Each of us knows from personal experience that downloads and uploads do not proceed at constant rate.
  - Not all times of transmissions are Poisson. Identifying Poisson time points in the data can be problematic. Some are machine triggered and these will certainly not be Poisson. While network engineers rightly believe in the *invariant* that behavior associated with humans acting independently can be modeled as a Poisson process, it is highly unlikely that, for example, subsidiary downloads triggered by going to the CNN website (imagine the calls to DoubleClick's ads) would follow a Poisson pattern.
- There is no hope that this simple model can successfully match fine time scale behavior observed below, say, 100 milliseconds. Below this time scale threshhold, observational studies speculate that traffic exhibits multifractal characteristics.
- The model does not take into account admission and congestion controls such as TCP. How can one incorporate a complex object like a control mechanism into an informative probability model?

10.3. **Long range dependence.** A stationary L_2 sequence $\{\xi_n, n \geq 1\}$ possesses long-range dependence (lrd) if

$$\text{Cov}(\xi_n, \xi_{n+h}) \sim h^{-\beta} L(h), \quad h \to \infty \tag{10.1}$$

for $0 < \beta < 1$ and $L(\cdot)$ slowly varying (Beran (1992)). Set $\gamma(h) = \text{Cov}(\xi_n, \xi_{n+h})$ and $\rho(h) = \gamma(h)/\gamma(0)$ for the covariance and correlation functions of the stationary process $\{\xi_n\}$. There is no universal agreement about terminology and sometimes long range dependence is taken to mean that covariances are not summable: $\sum_h |\gamma(h)| = \infty$, whereas short range dependence means $\sum_h |\gamma(h)| < \infty$. Traditional time series models such as ARMA models have covariances which go to zero geometrically fast as a function of the lag h.

Long-range dependence, like the property of heavy tails, has acquired a mystical, almost religious, significance and generated controversy. Researchers argue over whether it exists, whether it matters if it exists or not, or whether analysts have been fooled into mistaking some other phenomena like shifting levels, undetected trend (Künsch (1986)) or nonstationarity for long range dependence. Discussions about this have been going on since (at

least) the mid 70's in hydrology (Bhattacharya et al. (1983), Boes (1988), Boes and Salas-La Cruz (1973), Brockwell et al. (1982), Salas and Boes (1974, 1978), Salas et al. (1979)), in finance (Mikosch and Stărică (2003), Mikosch and Stărică (1999)) and in data network modeling (Cao et al. (2001a,b), Duffield et al. (1994), Garrett and Willinger (1994), Heyman and Lakshman (1996), Leland et al. (1993a), Park and Willinger (2000)). Think of it as one more modeling decision that needs to be made. Since long range dependence is an asymptotic property, models that possess long range dependence presumably have different asymptotic properties than those models where long range dependence is absent although even this is sometimes disputed.

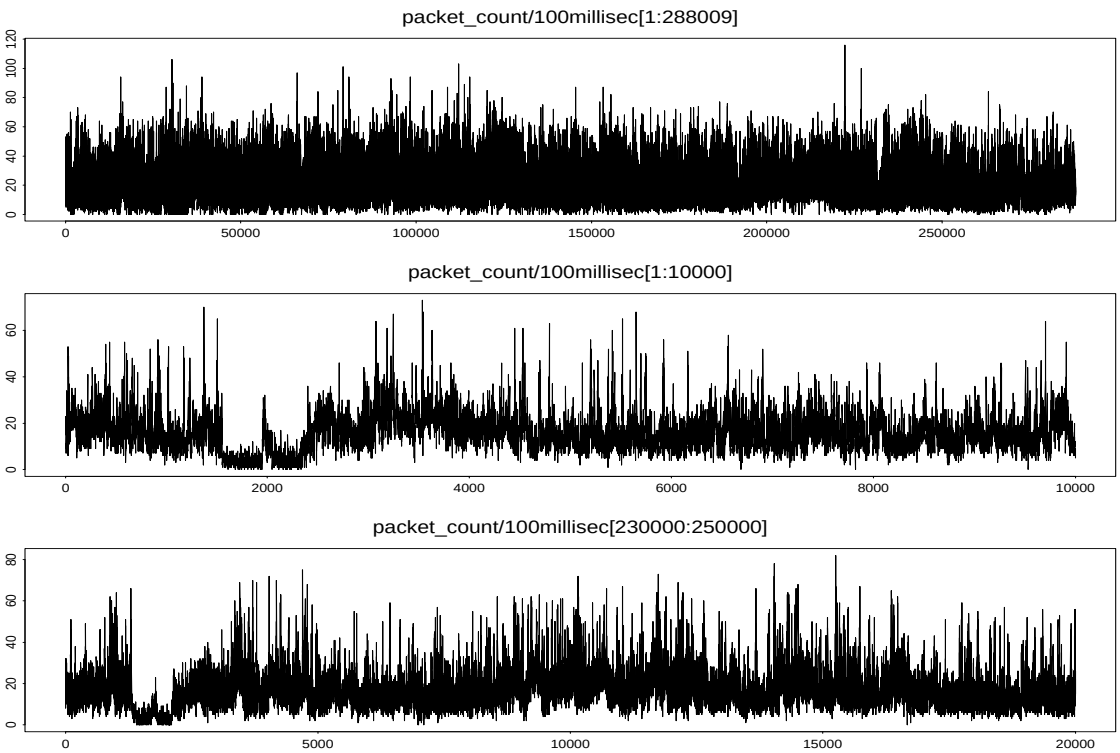

FIGURE 16. Time series plots for Company X data. Top: full data set. Middle: First 10,000 data. Bottom: last 20,000 observations.

10.3.1. *Simple minded detection of long range dependence using the sample acf.* The most common, ubiquitous, quick and dirty method to detect long range dependence (assuming you are convinced the data comes from a stationary process) is to graph the sample acf $\{\rho(h), h = 1, 2, \ldots, N\}$ where N is a large number but not a significant proportion of the whole sample size.

The plot should not decline rapidly. Classical time series data that one encounters in ARMA (Box-Jenkins) modeling exercises has a sample acf which is essentially zero after a few lags and acf plots of financial or teletraffic data are often in stark contrast.

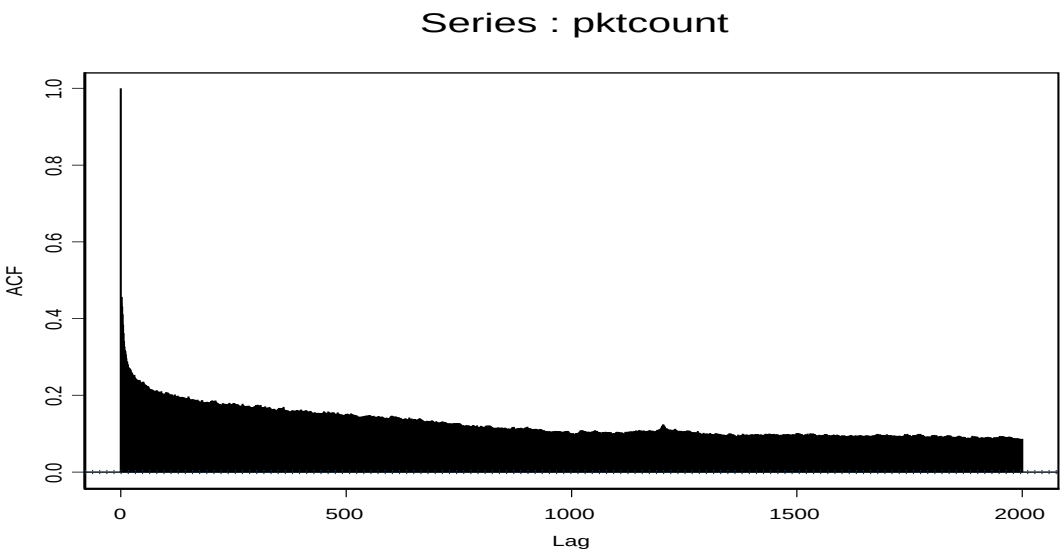

FIGURE 17. Sample autocorrelation plot for Company X data for 2000 lags.

Example 4 (Company X)**.** This trace is packet counts per 100 milliseconds=1/10 second for Financial Company X's wide area network link including USA–UK traffic. It consists of 288,009 observations corresponding to 8 hours of collection from 9am–5pm. Figure 16 shows time series plots. The top plot of the whole data set does not raise any alarms about lack of stationarity but this is partly due to the muddy plotting resulting from the abundance of data. The middle plot shows the first 10,000 and the bottom plot displays the last 20,000. With reduced data size, the last two plots raise some question whether stationarity is appropriate but this has not been pursued.

Figure 17 shows the acf plot for 2000 lags. There is little hurry for the plot to approach zero. (Don't try to model this with ARMA.)

10.4. **The infinite node Poisson model.** Understanding the connection between heavy tails and long range dependence requires a context. For the simplist explanations one can choose either the superposition of on/off processes or the infinite node Poisson model and our preference is for the latter. which is sometimes called the M/G/∞ input model.

In this model, there are potentially an infinite number of sources capable of sending work to the server. Imagine that transmission sources turn on and initiate *sessions* or *connections* at Poisson time points $\{\Gamma_k\}$ with rate λ. The lengths of sessions $\{L_n\}$ are iid non-negative random variables with common distribution F_{on} and during a session, work is transmitted to the server at constant rate. As a normalization, we assume the transmission rate is 1. Assume

$$1 - F_{\text{on}}(t) := \bar{F}_{\text{on}}(t) = t^{-\alpha} L(t), \quad t \to \infty; \tag{10.2}$$

that is,

$$\lim_{t\to\infty} \frac{\bar{F}_{\text{on}}(tx)}{\bar{F}_{\text{on}}(t)} = x^{-\alpha}, \quad x > 0.$$

In practice, empirical estimates of α usually range between 1 and 2 (Leland et al. (1994), Willinger et al. (1995)). However, studies of file sizes sometimes report measurements of $\alpha < 1$ (Arlitt and Williamson (1996), Resnick and Rootzén (2000)). The assumption of a fixed unit transmission rate is one of the more unrealistic aspects of the model which accords neither with anyones personal experience nor with measurement studies. For the present, in the interests of simplicity and for tractability, the fixed transmission rate will be assumed.

We will assume $1 < \alpha < 2$, so that the variance of F_{on} is infinite but

$$\mu_{\text{on}} = E(L_1) = \int_0^\infty \bar{F}_{\text{on}}(t)dt < \infty.$$

The processes of primary interest for describing this system are the following:

$$\begin{aligned} N(t) &= \text{number of sessions in progress at } t \\ &= \text{number of busy servers in the M/G/}\infty \text{ model} \\ &= \sum_{k=1}^{\infty} 1_{[\Gamma_k \le t < \Gamma_k + L_k]} \end{aligned} \tag{10.3}$$

and

$$\begin{aligned} A(t) &= \int_0^t N(s)ds = \text{cumulative input in } [0,t], \\ r &= \text{release rate or the rate at which the server works off the offered load.} \end{aligned} \tag{10.4}$$

Note that expressing $A(t)$ as an integral gives $N(t)$ the interpretation of "instantaneous input rate at time t". So realizations of $N(t)$ correspond to data traces of "packet counts per unit time". So we seek within the model an explanation of why $\{N(t)\}$ possesses long range dependence.

Stability requires us to assume that the long term input rate should be less that the output rate so we require

$$\lambda\mu_{\text{on}} < r.$$

This means the content or buffer level process $\{X(t), t \ge 0\}$ which satisfies

$$dX(t) = N(t)dt - r1_{[X(t)>0]}dt,$$

is regenerative with finite mean regeneration times and achieves a stationary distribution.

10.5. **Connection between heavy tails and long range dependence.** The common explanation for long range dependence in the total transmission rate by the system is that *high variability causes long range dependence* where we understand high variability means heavy tails. The long range dependence resulting from the heavy tailed distribution F_{on} can be easily seen for the infinite node Poisson model.

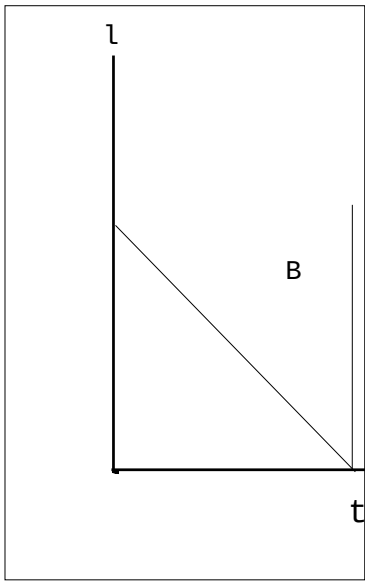

FIGURE 18. The region B

Assume that $1 < \alpha < 2$. To make our argument transparent, we consider the following background. For each t, $N(t)$ is a Poisson random variable. Why? When $1 < \alpha < 2$, $N(\cdot)$ has a stationary version on $\mathbb{R}$, the whole real line. Assume

$$\sum_k \epsilon_{\Gamma_k} = \mathrm{PRM}(\lambda dt)$$

is a homogeneous Poisson random measure on $\mathbb{R}$, with rate λ. Then using augmentation

$$M := \sum_k \epsilon_{(\Gamma_k, L_k)} = \mathrm{PRM}(\lambda dt \times F_{\mathrm{on}}) \tag{10.5}$$

is a two dimensional Poisson random measure on $\mathbb{R} \times [0, \infty)$ with mean measure $\lambda dt \times F_{\mathrm{on}}(dx)$ and

$$\begin{aligned} N(t) =& \sum_k 1_{[\Gamma_k \le t < \Gamma_k + L_k]} \\ =& M(\{(s,l) : s \le t < s + l\} = M(B) \end{aligned}$$

is Poisson because it is the two dimensional Poisson process M evaluated on the region B. See the gorgeous Figure 18. Note B is the region in the (s,l)-plane to the left of the vertical line through t and above the -45 degree line through $(t, 0)$. The mean of $M(B)$ is

$$\begin{aligned} E\Big(M(\{(s,l) : s \le t < s + l\}\Big) &= \iint_{\{(s,l): s \le t < s+l\}} \lambda ds F_{\mathrm{on}}(dl) \\ &= \int_{s=-\infty}^{t} \bar{F}_{\mathrm{on}}(t-s) \lambda ds = \lambda \mu_{\mathrm{on}}. \end{aligned} \tag{10.6}$$

Understanding the relation between $\{N(t)\}$ and the random measure M allows us to easily compute the covariance function. Refer to Figure 2. Recall that $N(t)$ corresponds to points to the left of the vertical through $(t,0)$ and above the -45-degree line through $(t,0)$ with a similar interpretation for $N(t+s)$. The process $\{N(t), t \in \mathbb{R}\}$ is stationary with covariance function

$$\begin{aligned}&\mathrm{Cov}(N(t), N(t+s))\\ &\quad=\mathrm{Cov}\big(M(A_1)+M(A_2), M(A_2)+M(A_3)\big)\end{aligned}$$

and because $M(A_1)$ and $M(A_3)$ are independent, the previous expression reduces to

$$=\mathrm{Cov}\big(M(A_2), M(A_2)\big) = \mathrm{Var}\big(M(A_2)\big).$$

For a Poisson random variable, the mean and the variance are equal and therefore the above equals

$$\begin{aligned}&=E\big(M(A_2)\big) = \iint_{\substack{u\le t\\ u+l>t+s}} \lambda du F_{\mathrm{on}}(dl)\\ &=\int_{u=-\infty}^{t} \lambda \bar{F}_{\mathrm{on}}(t+s-u)\, du\\ &=\lambda\int_s^{\infty} \bar{F}_{\mathrm{on}}(v)dv \sim cs\bar{F}_{\mathrm{on}}(s) \sim cs^{-(\alpha-1)}L(s).\end{aligned}$$

Note the use of Karamata's theorem to evaluate the asymptotic form of the integral of the regularly varying tail.

To summarize, we find that

$$\begin{aligned}\mathrm{Cov}(N(t), N(t+s)) =&\lambda\int_s^{\infty} \bar{F}_{\mathrm{on}}(v)dv\\ =&(\mathrm{const})s^{-(\alpha-1)}L(s)\\ =&(\mathrm{const})s\bar{F}_{\mathrm{on}}(s), \quad s\to\infty.\end{aligned} \tag{10.7}$$

The slow decay of the covariance as a function of the lag s characterizes *long range dependence*.

References

M. Arlitt and C. Williamson. Web server workload characterization: the search for invariants. Master's thesis, University of Saskatchewan, 1996.

J. Beran. Statistical methods for data with long-range dependence. *Statistical Science*, 7(4): 404–416, 1992. With discussions and rejoinder, pages 404-427.

R. Bhattacharya, V. Gupta, and E. Waymire. The Hurst effect under trends. *J. Applied Probability*, 20:649–662, 1983.

D. Boes. Schemes exhibiting Hurst behavior. In J. Srivastava, editor, *Essays in Honor of Franklin Graybill*, pages 21–42, Amsterdam, 1988. North Holland.

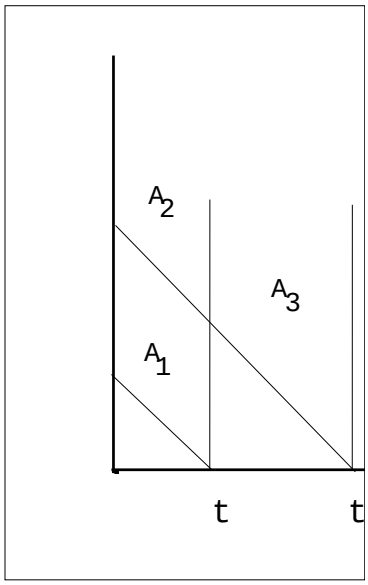

FIGURE 19. The regions A_1, A_2, A_3

D. Boes and J. D. Salas-La Cruz. On the expected range and expected adjusted range of partial sums of exchangeable random variables. *J. Applied Probability*, 10:671–677, 1973.

P. Brockwell, N. Pacheco-Santiago, and S. Resnick. Weak convergence and range analysis for dams with Markovian input rate. *J. Applied Probability*, 19:272–289, 1982.

J. Cao, W. Cleveland, D. Lin, and D. X. Sun. The effect of statistical multiplexing on internet packet traffic: Theory and empirical study. http://cm.bell-labs. com/cm/ms/departments /sia/InternetTraffic/ webpapers.html, 2001a.

J. Cao, W. Cleveland, D. Lin, and D. X. Sun. On the nonstationarity of internet traffic. *Proceedings of the ACM SIGMETRICS '01*, pages 102–112, 2001b. http://cm.bell-labs.com /cm/ms/departments/sia/ InternetTraffic/webpapers.html.

M. Crovella and A. Bestavros. Explaining world wide web traffic self–similarity. Preprint available as TR-95-015 from {crovella,best}@cs.bu.edu, 1995.

M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. In *Proceedings of the ACM SIGMETRICS '96 International Conference on Measurement and Modeling of Computer Systems*, volume 24, pages 160–169, 1996a.

M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. *Performance Evaluation Review*, 24:160–169, 1996b.

M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, 1997.

M. Crovella, A. Bestavros, and M. Taqqu. Heavy-tailed probability distributions in the world wide web. In M. T. R. Adler, R. Feldman, editor, *A Practical Guide to Heavy Tails: Statistical Techniques for Analysing Heavy Tailed Distributions*. Birkhäuser, Boston, 1999.

M. Crovella, G. Kim, and K. Park. On the relationship between file sizes, transport protocols, and self-similar network traffic. In *Proceedings of the Fourth International Conference on Network Protocols (ICNP'96)*, pages 171–180, 1996.

C. Cunha, A. Bestavros, and M. Crovella. Characteristics of www client–based traces. Preprint available as BU-CS-95-010 from {crovella,best}@cs.bu.edu, 1995.

N. Duffield, J. Lewis, N. O'Connell, R. Russell, and F. Toomey. Statistical issues raised by the bellcore data. In *Proceedings of 11th IEE UK Teletraffic Symposium, Cambridge UK*, pages 23–25, London, 1994. IEE.

D. Duffy, A. McIntosh, M. Rosenstein, and W. Willinger. Analyzing telecommunications traffic data from working common channel signaling subnetworks. In M. Tarter and M. Lock, editors, *Computing Science and Statistics Interface, Proceedings of the 25th Symposium on the Interface*, volume 25, pages 156– 165, San Diego, California, 1993.

M. Garrett and W. Willinger. Analysis, modeling and generation of self similar vbr video traffic. In *Proceedings of the ACM SigComm*, London, 1994.

C. Guerin, H. Nyberg, O. Perrin, S. Resnick, H. Rootzen, and C. Stărică. Empirical testing of the infinite source poisson data traffic model. *Stochastic Models*, 19(2):151–200, 2003.

D. Heath, S. Resnick, and G. Samorodnitsky. Patterns of buffer overflow in a class of queues with long memory in the input stream. *Ann. Applied Probability*, 7(4):1021–1057, 1997. ISSN 1050-5164.

D. Heath, S. Resnick, and G. Samorodnitsky. Heavy tails and long range dependence in on/off processes and associated fluid models. *Math. Oper. Res.*, 23(1):145–165, 1998. ISSN 0364-765X.

D. Heath, S. Resnick, and G. Samorodnitsky. How system performance is affected by the interplay of averages in a fluid queue with long range dependence induced by heavy tails. *Ann. Appl. Probab.*, 9:352–375, 1999.

D. Heyman and T. Lakshman. What are the implications of long-range dependence for vbr-video traffic engineering? *IEEE-ACM Transactions on Networking*, 4(3):301–317, 1996.

P. Jelenković and A. Lazar. A network multiplexer with multiple time scale and subexponential arrivals. In G. P., K. Sigman, and D. Yao, editors, *Stochastic Networks*, volume 117 of *Lecture Notes in Statistics*, pages 215–235. Springer, New York, 1996.

P. Jelenković and A. Lazar. Subexponential asymptotics of a Markov-modulated random walk with queueing applications. *J. Applied Probability*, 35(2):325–347, 1998. ISSN 0021-9002.

P. Jelenković and A. Lazar. Asymptotic results for multiplexing subexponential on-off processes. *Advances in Applied Probability*, 31:394–421, 1999.

H. Künsch. Discrimination between monotonic trends and long range dependence. *J. Applied Probability*, 23:1025–1030, 1986.

W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic. *ACM/SIGCOMM Computer Communications Review*, pages 183–193, 1993a.

W. Leland, M. Taqqu, W. Willinger, and D. Wilson. Statistical analysis of high time-resolution ethernet Lan traffic measurements. In *Proceedings of the 25th Symposium on the Interface between Statistics and Computer Science*, pages 146– 155, 1993b.

W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2:1–15, 1994.

T. Mikosch, S. Resnick, H. Rootzén, and A. Stegeman. Is network traffic approximated by stable Lévy motion or fractional Brownian motion? *Ann. Applied Probability*, 12(1):23–68, 2002.

T. Mikosch and C. Stărică. Long range dependence effects and ARCH modeling. In P. Doukhan, G. Oppenheim, and M. Taqqu, editors, *Theory and applications of long-range dependence*, pages xii+719. Birkhäuser Boston Inc., Boston, MA, 2003. ISBN 0-8176-4168-8.

T. Mikosch and A. Stegeman. The interplay between heavy tails and rates in self-similar network traffic. Technical Report. University of Groningen, Department of Mathematics, Available via www.cs.rug.nl/˜eke/iwi/preprints/, 1999.

T. Mikosch and C. Stărică. Change of structure in financial data. Available at: http://www.math.ku.dk/˜mikosch/preprint.html; to appear "Review of Economics and Statistics", 2004, 1999.

K. Park and W. Willinger. Self-similar network traffic: An overview. In K. Park and W. Willinger, editors, *Self-Similar Network Traffic and Performance Evaluation*, pages 1–38. Wiley-Interscience, New York, 2000.

M. Parulekar and A. Makowski. Tail probabilities for a multiplexer with a self-similar traffic. *Proceedings of the 15th Annual IEEE INFOCOM*, pages 1452–1459, 1996a.

M. Parulekar and A. Makowski. Tail probabilities for M/G/∞ input process (i): Preliminary asymptotics. University of Maryland Institute for Systems Research Technical Report TR 96-41, 1996b.

S. Resnick and H. Rootzén. Self-similar communication models and very heavy tails. *Ann. Applied Probability*, 10:753–778, 2000.

S. Resnick and E. van den Berg. Weak convergence of high-speed network traffic models. *J. Applied Probability*, 37(2):575–597, 2000.

J. Salas and D. Boes. Expected range and adjusted range of hydrologic sequences. *Water Resoures Research*, 10(3):457–463, 1974.

J. Salas and D. Boes. Nonstationarity of mean and Hurst phenomenon. *Water Resources Research*, 14((1)):135–143, 1978.

J. Salas, D. Boes, and V. Yevjevich. Hurst phenomenon as a pre-asymptotic behavior. *J. Hydrology*, 44((1-2)):1–15, 1979.

A. Stegeman. Modeling traffic in high-speed networks by on/off models. Master's thesis, University of Groningen, Department of Mathematics, 1998.

M. Taqqu, W. Willinger, and R. Sherman. Proof of a fundamental result in self-similar traffic modeling. *Computer Communications Review*, 27:5–23, 1997.

W. Vervaat. Functional central limit theorems for processes with positive drift and their inverses. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 23:245–253, 1972a.

W. Vervaat. *Success Epochs in Bernoulli Trials (with applications in number theory)*. Mathematisch Centrum, Amsterdam, 1972b. Mathematical Centre Tracts, No. 42.

W. Willinger, M. Taqqu, M. Leland, and D. Wilson. Self–similarity through high variability: statistical analysis of ethernet lan traffic at the source level. *Computer Communications Review*, 25:100–113, 1995. Proceedings of the ACM/SIGCOMM'95, Cambridge, MA.

11. The Laplace functional.

The Laplace functional is a transform technique which is useful for manipulating distributions of random measures and point processes. When applied to Poisson processes and empirical measures, algebraic manipulations become familiar to ones used with either characteristic functions or Laplace transforms applied to sums of iid random variables.

Continue to assume the state space of the random measures or point processes is the nice space $\mathbb{E}$. For a non-negative, bounded measureable function $f : \mathbb{E} \mapsto \mathbb{R}_+$ and for $\mu \in M_+(\mathbb{E})$, we continue to use the notation

$$\mu(f) = \int_{x \in E} f(x) d\mu(x).$$

For a point measure $m = \sum_i \epsilon_{x_i}$, this is

$$m(f) = \int_{x \in \mathbb{E}} f(x) m(dx) = \sum_i f(x_i).$$

If we think of f ranging over all the non-negative bounded functions, $m(f)$ yields all the information contained in m; certainly we learn about the value of m on each set $A \in \mathcal{E}$ since we can always set $f = 1_A$. So integrals of measures with respect to arbitrary test functions contain as much information as evaluating the measures on arbitrary sets.

Definition 3 (Laplace functional)**.** Suppose $\mathcal{B}_+$ are the non-negative, bounded, measurable functions from $\mathbb{E} \mapsto \mathbb{R}_+$ and let

$$\xi : (\Omega, \mathcal{A}, \mathbb{P}) \mapsto (M_+(\mathbb{E}), \mathcal{M}_+(\mathbb{E}))$$

be a random measure. The *Laplace functional* of ξ is the non-negative function on $\mathcal{B}_+$ given by

$$\Psi_\xi(f) = \mathbf{E} \exp\{-\xi(f)\} = \int_\Omega \exp\{-N(\omega, f) d\mathbb{P}(\omega) = \int_{M_+(\mathbb{E})} \exp\{-m(f) \mathbb{P} \circ N^{-1}(dm).$$

Note if P is a probability measure on $\mathcal{M}_p(\mathbb{E})$, its Laplace functional is

$$\int_{M_p(\mathbb{E})} \exp\{-m(f) P(dm), \quad f \in \mathcal{B}_+.$$

Proposition 10. *The Laplace functional of ξ uniquely determines the distribution of ξ, namely $\mathbb{P} \circ \xi^{-1}$ on $\mathcal{M}_+(\mathbb{E})$).*

Proof. The topology (and hence the Borel σ-algebra on $M_+(\mathbb{E})$ is generated by basic open sets of the form

$$\{\mu \in M_+(\mathbb{E}) : \mu(f_i) \in I_i;\ i = 1, \ldots, k\},$$

where $f_i \in C_K^+(\mathbb{E})$ and I_i are bounded open intervals. (Review page 33.) This class of sets is closed under finite intersection and hence is a Π-system generating the σ-algebra. So from (Resnick, 1998, page 38) is suffices to show that $\mathbb{P} \circ \xi^{-1}$ is uniquely determined on this class by the Laplace functional.

For $f = \sum_{i=1}^k \lambda_i f_i$, with $\lambda_i > 0$, we have

$$\psi_\xi(f) = \mathbf{E}e^{-\xi(f)} = \mathbf{E}e^{-\sum_{i=1}^k \lambda_i \xi(f_i)}.$$

So ψ_ξ determines the joint distribution of the random vector

$$(\xi(f_1), \ldots, \xi(f_k))$$

and hence

$$\mathbb{P} \circ \xi^{-1}\{\mu : \mu(f_i) \in I_i,\ i = 1, \ldots, k\} = \mathbb{P}[(\xi(f_1), \ldots, \xi(f_k)) \in \cdot]$$

is determined, as required. □

11.1. **The Laplace functional of the Poisson process.** Recall Definition 2 of the Poisson process and the two parts 1 and 2 of the definition.

The Poisson process can be identified by the characteristic form of its Laplace functional as discussed next.

Theorem 8 (Laplace functional of PRM)**.** *The distribution of PRM(μ) is uniquely determined by 1 and 2 in Definition 2. Furthermore, the point process N is PRM(μ) iff its Laplace functional is of the form*

$$\Psi_N(f) = \exp\{-\int_{\mathbb{E}} (1 - e^{-f(x)})\mu(dx)\}, \quad f \in \mathcal{B}_+. \tag{11.1}$$

So PRM(μ) can be identified by the characteristic form of its Laplace functional.

Proof. We first show (1) and (2) imply (11.1). If $f = \lambda 1_A$ where $\lambda > 0$ then because $N(f) = \lambda N(A)$ and $N(A)$ is Poisson with parameter $\mu(A)$ we get

$$\begin{aligned}\Psi_N(f) =& E\Big(e^{-\lambda N(A)}\Big) = \exp\{(e^{-\lambda} - 1)\mu(A)\} \\ =& \exp\{-\int_E (1 - e^{-f(x)})\mu(dx)\}\end{aligned}$$

which is the correct form given in (11.1).

Next suppose f has a somewhat more complex form

$$f = \sum_{i=1}^k \lambda_i 1_{A_i}$$

where $\lambda_i \geq 0, A_i \in \mathcal{E}, 1 \leq i \leq k$ and $A_1, \ldots, A_k$ are disjoint. Then

$$\begin{aligned}\Psi_N(f) =& E\left(\exp\left\{-\sum_{i=1}^k \lambda_i N(A_i)\right\}\right) \\ =& \prod_{i=1}^k E\Big(\exp\{-\lambda_i N(A_i)\}\Big) \quad \text{from independence} \\ =& \prod_{i=1}^k \exp\left\{-\int_{\mathbb{E}} (1 - e^{-\lambda_i 1_{A_i}(x)})\mu(dx)\right\} \quad \text{from the previous step}\end{aligned}$$

$$
\begin{aligned}
&= \exp\left\{\int_{\mathbb{E}} \sum_{i=1}^{k} (1 - e^{-\lambda_i 1_{A_i}(x)})\mu(dx)\right\} \\
&= \exp\left\{\int_{\mathbb{E}} (1 - e^{-\sum_{i=1}^{k} \lambda_i 1_{A_i}(x)})\mu(dx)\right\} \\
&= \exp\left\{\int_{\mathbb{E}} (1 - e^{-f(x)})\mu(dx)\right\}
\end{aligned}
$$

which again verifies (11.1).

Now the last step is to take general $f \in \mathcal{B}_+$ and verify (11.1) for such f. We may approximate f from below by simple f_n of the form just considered. We may take for instance

$$
f_n(x) = \sum_{i=1}^{n2^n} \frac{i-1}{2^n} 1_{[\frac{i-1}{2^n}, \frac{i}{2^n})}(f(x)) + n 1_{[n,\infty)}(f(x))
$$

so that

$$
0 \le f_n(x) \uparrow f(x).
$$

By monotone convergence $N(f_n) \uparrow N(f)$ and since $e^{-f} \le 1$ we get by dominated convergence that

$$
\Psi_N(f) = \lim_{n\to\infty} \Psi_N(f_n).
$$

We have from the previous step that

$$
\Psi_N(f_n) = \exp\left\{-\int_{\mathbb{E}} (1 - e^{-f_n(x)})\mu(dx)\right\}.
$$

Since

$$
1 - e^{-f_n} \uparrow 1 - e^{-f}
$$

we conclude by monotone convergence that

$$
\int_{\mathbb{E}} (1 - e^{-f_n(x)}\mu(dx) \uparrow \int_{\mathbb{E}} (1 - e^{-f(x)}\mu(dx)
$$

and thus we conclude (11.1) holds for any $f \in \mathcal{B}_+$. Since the distribution of N is uniquely determined by Ψ_N we have shown that 1 and 2 in Definition 2 determine the distribution of N.

Conversely if the Laplace functional of N is given by (11.1) then $N(A)$ must be Poisson distributed with parameter $\mu(A)$ for any $A \in \mathcal{E}$ which is readily checked by substituting $f = \lambda 1_A$ in (11.1) to get a Laplace transform of a Poisson distribution. Furthermore if $A_1, \ldots, A_k$ are disjoint sets in $\mathcal{E}$ and $f = \sum_{i=1}^{k} \lambda_i 1_{A_i}$ then substituting in (11.1) gives

$$
\begin{aligned}
Ee^{-\sum_{i=1}^{k} \lambda_i N(A_i)} &= \exp\left\{-\int_{\mathbb{E}} (1 - e^{-\sum_{i=1}^{k} \lambda_i 1_{A_i}})d\mu\right\} \\
&= \exp\left\{-\int_{\mathbb{E}} \sum_{i=1}^{k} (1 - e^{-\lambda_i 1_{A_i}})d\mu\right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \prod_{i=1}^{k} \exp\{-(1-e^{-\lambda_i})\mu(A_i)\} \\
&= \prod_{i=1}^{k} E e^{-\lambda_i N(A_i)}
\end{aligned}
$$

and so the joint Laplace transform of $(N(A_i), 1 \le i \le k)$ factors into a product of Laplace transforms and this shows independence. □

11.1.1. *Application: A general construction of the Poisson process.* When $\mathbb{E} = [0, \infty)$, we know the renewal theory construction yields a Poisson process with Lebesgue measure as the mean measure. Here is a general scheme for constructing a Poisson process with mean measure μ.

Start by supposing that $\mu(\mathbb{E}) < \infty$. Define the probability measure F

$$F(dx) = \mu(dx)/\mu(\mathbb{E})$$

on $\mathcal{E}$. Let $\{X_n, n \ge 1\}$ be iid random elements of $\mathbb{E}$ with common distribution F and let ν be independent of $\{X_n\}$ with a Poisson distribution with parameter $\mu(\mathbb{E})$. Define

$$N = \begin{cases} \sum_{i=1}^{\nu} \epsilon_{X_i}, & \text{if } \nu \ge 1 \\ 0, & \text{if } \nu = 0. \end{cases}$$

We claim N is PRM(μ). Verify this by computing the Laplace functional of N:

$$
\begin{aligned}
\Psi_N(f) &= E e^{-\sum_{i=1}^{\nu} f(X_i)} \\
&= \sum_{j=0}^{\infty} E e^{-\sum_{i=1}^{j} f(X_i)} P[\nu = j] \\
&= \sum_{j=0}^{\infty} \left(E e^{-f(X_1)}\right)^j P[\nu = j] \\
&= \exp\left\{\mu(\mathbb{E})(E e^{-f(X_1)} - 1)\right\} \\
&= \exp\left\{-\mu(\mathbb{E})\left(1 - \int_{\mathbb{E}} e^{-f(x)} \frac{\mu(dx)}{\mu(\mathbb{E})}\right)\right\} \\
&= \exp\left\{-\int_{\mathbb{E}} \left(1 - e^{-f(x)}\right) \mu(dx)\right\}.
\end{aligned}
$$

The Laplace functional has the correct form and hence N is indeed PRM(μ). Note that what the construction does is to toss points at random into $\mathbb{E}$ according to distribution $\mu(dx)/\mu(\mathbb{E})$; the number of points tossed is Poisson with parameter $\mu(\mathbb{E})$.

When the condition $\mu(\mathbb{E}) < \infty$ fails, we proceed as follows to make a minor modification in the foregoing construction: Decompose $\mathbb{E}$ into disjoint bits $\mathbb{E}_1, \mathbb{E}_2, \ldots$ so that $\mathbb{E} = \cup_i \mathbb{E}_i$ and $\mu(\mathbb{E}_i) < \infty$ for each i. Let $\mu_i(dx) = \mu(dx) 1_{\mathbb{E}_i}(x)$ and let N_i be PRM(μ_i) on $\mathbb{E}$ (do the

construction just outlined) and arrange things so the collection $\{N_i\}$ is independent. Define $N := \sum_i N_i$ and N is PRM(μ) since

$$\begin{aligned}\Psi_N(f) &= \prod_i \Psi_{N_i}(f)\\ &= \prod_i \exp\left\{-\int_{\mathbb{E}_i}\left(1-e^{-f(x)}\right)\mu_i(dx)\right\}\\ &= \exp\left\{-\sum_i\int_{\mathbb{E}}\left(1-e^{-f(x)}\right)\mu_i(dx)\right\}\\ &= \exp\left\{-\int_{\mathbb{E}}\left(1-e^{-f(x)}\right)\sum_i\mu_i(dx)\right\}\\ &= \exp\left\{-\int_{\mathbb{E}}\left(1-e^{-f(x)}\right)\mu(dx)\right\}\end{aligned}$$

since $\sum_i \mu_i = \mu$. This completes the construction.

11.2. **Weak convergence of point processes and random measures.** We now discuss weak convergence in $M_+(\mathbb{E})$ and specialize to give criterion for weak convergence to a Poisson process.

11.2.1. *Basic criterion via Laplace functionals.* Since $M_+(\mathbb{E})$ can be metrized as a complete separable metric space, the theory of weak convergence applies. A useful criterion for our purposes is in terms of Laplace functionals.

Theorem 9 (Convergence criterion)**.** *Let $\{\xi_n, n \geq 0\}$ be random elements of $M_+(\mathbb{E})$. Then*

$$\xi_n \Rightarrow \xi_0 \quad \text{in } M_p(\mathbb{E}),$$

iff

$$\psi_{\xi_n}(f) = \mathbf{E}e^{-\xi_n(f)} \to \mathbf{E}e^{-\xi_0(f)} = \psi_{\xi_0}(f), \quad \forall f \in C_K^+(\mathbb{E}). \tag{11.2}$$

So weak convergence is characterized by convergence of Laplace functionals on $C_K^+(\mathbb{E})$.

Proof. Here is the proof of the easy half of the equivalence. Suppose $\xi_n \Rightarrow \xi_0$ in $M_+(\mathbb{E})$. The map $M_+(\mathbb{E}) \mapsto [0,\infty)$ defined by $\mu \mapsto \mu(f)$ is continuous, so the continuous mapping theorem gives $\xi_n(f) \Rightarrow \xi_0(f)$ in $\mathbb{R}$. Thus

$$e^{-\xi_n(f)} \Rightarrow e^{-\xi_0(f)},$$

and by Lebesque's dominated convergence theorem

$$\mathbf{E}e^{-\xi_n(f)} \to \mathbf{E}e^{-\xi_0(f)}$$

as required.

To prove weak convergence, it suffices (see page 40) to prove for a given sequence $h_j \in C_K^+(\mathbb{E})$ that

$$(\xi_n(h_j), j \geq 1) \Rightarrow (\xi_0(h_j), j \geq 1)$$

in $\mathbb{R}^\infty$, for which it suffices to show for any k

$$(\xi_n(h_j), 1 \le j \le k) \Rightarrow (\xi_0(h_j), 1 \le j \le k)$$

in $\mathbb{R}^k$. Taking Laplace transforms, it suffices to show

$$\mathbf{E}e^{-\sum_{i=1}^k \lambda_i \xi_n(h_i)} = \mathbf{E}e^{-\xi_n(\sum_{i=1}^k \lambda_i h_i)} \to \mathbf{E}e^{-\sum_{i=1}^k \lambda_i \xi_0(h_i)} = \mathbf{E}e^{-\xi(\sum_{i=1}^k \lambda_i h_i)}.$$

Hence showing

$$\psi_{\xi_n}(h) \to \psi_{\xi_0}(h)$$

for any $h \in C_K^+(\mathbb{E})$ is sufficient. □

11.2.2. *Example: Convergence of Poisson processes.* As a further application of the convergence criterion in Theorem 9 we have the following.

Corollary 2 (Convergence of PRM's)**.** *Suppose for each $n \ge 0$ that N_n is PRM(μ_n) on $\mathbb{E}$. Then as $n \to \infty$,*

$$N_n \Rightarrow N_0,$$

iff

$$\mu_n \xrightarrow{v} \mu_0$$

in $M_+(\mathbb{E})$.

Proof. Use the convergence criterion in terms of convergence of Laplace functionals given in Theorem 9. We have from the form of the Laplace functional of a PRM that $N_n \Rightarrow N_0$ iff for any $f \in C_K^+(\mathbb{E})$ that

$$\exp\{-\int_{\mathbb{E}} \Big(1 - e^{-f(e)}\Big)\mu_n(de)\} \to \exp\{-\int_{\mathbb{E}} \Big(1 - e^{-f(e)}\Big)\mu_0(de)\}.$$

This means that for any $g \in C_K^+(\mathbb{E})$ which is bounded by 1 we have

$$\mu_n(g) \to \mu_0(g).$$

Since any function $h \in C_K^+(\mathbb{E})$ is bounded, dividing by $\sup_{e\in\mathbb{E}} h(e)$ gives a function bounded by 1 and the desired result follows. □

11.3. **Basic convergences of empirical measures.** We now give two criteria for convergence. One gives necessary and sufficient conditions for empirical measures of scaled observations to converge to a Poisson random measure limit and the other discusses convergence to a constant limit measure. The first is the basis for manipulating iid random variables with regularly varying tails by means of the Poisson transform and the second is the basis for consistency of estimates of heavy tailed parameters.

Theorem 10 (Basic convergence)**.** *Suppose that for each $n \ge 1$ we have $\{X_{n,j}, j \ge 1\}$ is a sequence of iid random elements of $(\mathbb{E}, \mathcal{E})$. Let ξ be PRM(μ) on $M_p(\mathbb{E})$.*

(i) We have

$$\sum_{j=1}^n \epsilon_{X_{n,j}} \Rightarrow \xi = PRM(\mu) \tag{11.3}$$

on $M_p(\mathbb{E})$*, iff*

$$n\mathbb{P}[X_{n,1} \in \cdot] = \mathbf{E}\left(\sum_{j=1}^{n} \epsilon_{X_{n,j}}(\cdot)\right) \xrightarrow{v} \mu \tag{11.4}$$

in $M_+(\mathbb{E})$*.*

(ii) Suppose additionally that $0 < a_n \uparrow \infty$*. Then for a measure* $\mu \in M_+(\mathbb{E})$ *we have*

$$\frac{1}{a_n}\sum_{j=1}^{n} \epsilon_{X_{n,j}} \Rightarrow \mu \tag{11.5}$$

on $M_+(\mathbb{E})$ *iff*

$$\frac{n}{a_n}P[X_{n,1} \in \cdot] = \mathbf{E}\left(\frac{1}{a_n}\sum_{j=1}^{n} \epsilon_{X_{n,j}}(\cdot)\right) \xrightarrow{v} \mu \tag{11.6}$$

in $M_+(\mathbb{E})$*.*

Remark 4. Note that part (ii) has already been used in Section 7.1 and proven in a superficially more specific context on page 41 in Theorem 5 in connection with the tail empirical measure.

Proof. (i) We compute Laplace functionals and show they converge: For $f \in C_K^+(\mathbb{E})$,

$$\begin{aligned}
\mathbf{E}e^{-\sum_{j=1}^{n}\epsilon_{X_{n,j}}(f)} &= \mathbf{E}e^{-\sum_{j=1}^{n} f(X_{n,j})} = \left(\mathbf{E}e^{-f(X_{n,1})}\right)^n \\
&= \left(1 - \frac{\mathbf{E}(n(1-e^{-f(X_{n,1})}))}{n}\right)^n \\
&= \left(1 - \frac{\int_{\mathbb{E}}(1-e^{-f(x)})nP[X_{n,1}\in dx]}{n}\right)^n
\end{aligned}$$

and this converges to

$$\exp\{\int_{\mathbb{E}}(1-e^{-f(x)})\mu(dx)\},$$

the Laplace functional of PRM (μ), iff

$$\int_{\mathbb{E}}(1-e^{-f(x)})nP[X_{n,1}\in dx] \to \int_{\mathbb{E}}(1-e^{-f(x)})\mu(dx).$$

and this last statement is equivalent to vague convergence in (11.4).

(ii) Here again we prove the result by showing Laplace functionals converge. We compute the Laplace functional for the quantity on the left side of (11.5):

$$\begin{aligned}
\mathbf{E}e^{-\frac{1}{a_n}\sum_{i=1}^{n}\epsilon_{X_{n,1}}(f)} &= \left(\mathbf{E}e^{-\frac{1}{a_n}f(X_{n,1})}\right)^n \\
&= \left(1 - \frac{\int_{\mathbb{E}}\left(1-e^{-\frac{1}{a_n}f(x)}\right)n\mathbb{P}[X_{n,1}\in dx]}{n}\right)^n
\end{aligned}$$

and we claim this converges to $e^{-\mu(f)}$, the Laplace functional of μ, iff

$$\int_{\mathbb{E}}(1-e^{-\frac{1}{a_n}f(x)})nP[X_{n,1}\in dx]\to\mu(f). \tag{11.7}$$

We show (11.7) is equivalent to (11.6) as follows: Suppose (11.6) holds. On the one hand,

$$\int_{\mathbb{E}}(1-e^{-f(x)/a_n})n\mathbb{P}[X_{n,1}\in dx]\le\int_{\mathbb{E}}f(x)\frac{n}{a_n}P[X_{n,1}\in dx]\to\mu(f)$$

so

$$\limsup_{n\to\infty}\int_{\mathbb{E}}(1-e^{-f(x)/a_n})nP[X_{n,1}\in dx]\le\mu(f).$$

On the other hand

$$\begin{aligned}\int_{\mathbb{E}}&(1-e^{-f(x)/a_n})nP[X_{n,1}\in dx]\\ &\ge\int_{\mathbb{E}}f(x)\frac{n}{a_n}P[X_{n,1}\in dx]-\int_{\mathbb{E}}\frac{f^2(x)}{2a_n}\frac{n}{a_n}P[X_{n,1}\in dx]\\ &=I+II.\end{aligned}$$

Now $I\to\mu(f)$ from (11.6) and since $f^2\in C_K^+(\mathbb{E})$ we have

$$II\sim\frac{\mu(f^2)}{2a_n}\to 0$$

since $a_n\uparrow\infty$. So

$$\liminf_{n\to\infty}\int_{\mathbb{E}}(1-e^{-f(x)/a_n})nP[X_{n,1}\in dx]\ge\mu(f)$$

providing the other half of the sandwich.

Conversely, let $g\in C_K^+(\mathbb{E})$, and suppose $g\le 1$. Supposing (11.7) is true, we get

$$g/a_n\ge 1-e^{-g/a_n}$$

leading to

$$\liminf_{n\to\infty}\int_{\mathbb{E}}g(x)\frac{n}{a_n}P[X_{n,1}\in dx]\ge\mu(f)$$

and

$$\frac{g}{a_n}-\frac{g^2}{2a_n^2}\le 1-e^{-g/a_n}$$

so

$$\limsup_{n\to\infty}\int_{\mathbb{E}}\left(\frac{g(x)}{a_n}-\frac{g^2(x)}{2a_n^2}\right)n\mathbb{P}[X_{n,1}\in dx]\le\mu(f).$$

As before, we may show

$$\int_{\mathbb{E}}\frac{g^2(x)}{2a_n^2}nP[X_{n,1}\in dx]\to 0.$$

□

11.4. **Regular variation and Poisson convergence.** Recall from Theorem 4, that regular variation of a distribution tail is equivalent to vague convergence of induced measures. Using the Basic Convergence Theorem 10, we will see that regular variation of distribution tails is equivalent to weak convergence of a sequence of associated empirical measures with proper scaling to a limiting Poisson random measure. This, in effect, provides a transform method for dealing with regular variation.

The specialization of Theorem 10 to regular variation is given next.

Theorem 11. *Suppose that $\{Z_n, n \geq 1\}$ are iid, non-negative random variables with common distribution F. As usual, we define*

$$(\frac{1}{1-F})^{\leftarrow}(t) = F^{\leftarrow}(1 - \frac{1}{t}) =: b(t).$$

The following are equivalent.

(1) $\bar{F} \in RV_{-\alpha}$.

(2) *There exists $a_n \to \infty$ such that*

$$\lim_{n\to\infty} n\bar{F}(a_n x) = x^{-\alpha}, \quad x > 0.$$

(3) *There exists $a_n \to \infty$ such that*

$$n\mathbb{P}\left[\frac{Z_1}{a_n} \in \cdot\right] \xrightarrow{v} \nu \tag{11.8}$$

on $M_+\big((0,\infty]\big)$, where $\nu(x,\infty] = x^{-\alpha}$, $x > 0$.

(4) *There exists a sequence $a_n \to \infty$ such that $M_p\big((0,\infty]\big)$ we have*

$$\sum_{i=1}^{n} \epsilon_{Z_i/a_n} \Rightarrow PRM(\nu) \tag{11.9}$$

If any of 2, 3 or 4 hold, we may always choose $a_n = b(n)$.

Any of 1,2, 3, 4 imply that for any sequence $k = k(n) \to \infty$ such that $n/k \to \infty$ we have

(5) *In $M_+\big((0,\infty]\big)$*

$$\frac{1}{k}\sum_{i=1}^{n} \epsilon_{Z_i/b\left(\frac{n}{k}\right)} \Rightarrow \nu \tag{11.10}$$

and 5 is equivalent to any of 1–4, provided $k(\cdot)$ satisfies

$$k(n) \sim k(n+1).$$

As we will see, this result is basically dimensionless and indicates what a good definition of multivariate regular variation should be.

Proof. The equivalence of 1–3 is exactly Theorem 4. The equivalence of 3 and 4 is the Basic Convergence Theorem 10 with

$$X_{n,i} = Z_i/a_n.$$

Consider 5. From the Basic Convergence Theorem 10 part (ii), (11.10) equivalent to

$$\frac{n}{k}\mathbb{P}\left[\frac{Z_1}{b(n/k)} \in \cdot\right] \xrightarrow{v} \nu. \tag{11.11}$$

The relation (11.11) certainly follows from, say, 1, and if $k(n) \sim k(n+1)$ we may mimic the proof of the equivalence of the sequential version of regular variation to the actual definition in Part (ii) of Proposition 2. □

We will refer to the equivalence of regular variation of $\bar{F}$ to (11.9) as the *Poisson transform* or the *point process method*.

11.4.1. *Preservation of weak convergence under mappings.* Consider two state spaces $\mathbb{E}_1$ and $\mathbb{E}_2$ with a mapping $T : \mathbb{E}_1 \mapsto \mathbb{E}_2$ taking one into the other. A measure $\mu \in M_+(\mathbb{E}_1)$ has an image $\hat{T}(\mu) \in M_+(\mathbb{E}_2)$ given by the map

$$\hat{T}(\mu) = \mu \circ T^{-1}.$$

If T is a continuous point transformation, is $\hat{T} : M_+(\mathbb{E}_1) \mapsto M_+(\mathbb{E}_2)$ continuous? (Maybe!)

Note, if $m \in M_p(\mathbb{E})$ is a point measure of the form $\sum_i \epsilon_{x_i}$, then

$$\hat{T}(m) = m \circ T^{-1} = \sum_i \epsilon_{T(x_i)}.$$

So if T operates on the state space where the points live then $\hat{T}m$ is the induced point measure after mapping the points of the original point measure. A restriction is needed without which continuity of T does not guarantee continuity of $\hat{T}$.

Proposition 11. *Suppose $T : \mathbb{E}_1 \mapsto \mathbb{E}_2$ is a continuous function such that*

$$T^{-1}(K_2) \in \mathcal{K}(\mathbb{E}_1), \quad \forall K_2 \in \mathcal{K}(\mathbb{E}_2), \tag{11.12}$$

that is,

$$T^{-1}\Big(\mathcal{K}(\mathbb{E}_2)\Big) \subset \mathcal{K}(\mathbb{E}_1).$$

Then

(a) *If $\mu_n \xrightarrow{v} \mu_0$ in $M_+(\mathbb{E}_1)$, we have also that*

$$\hat{T}\mu_n = \mu_n \circ T^{-1} \xrightarrow{v} \mu_n \circ T^{-1} = \hat{T}\mu_n.$$

(b) *If for each $n \geq 0$ we have*

$$N_n = \sum_j \epsilon_{X_n^{(n)}}$$

are random elements of $M_+(\mathbb{E}_1)$ such that

$$N_n \Rightarrow N_0,$$

in $M_p(\mathbb{E}_1)$, then also

$$\hat{T}N_n = \sum_j \epsilon_{T(X_n^{(n)})} \Rightarrow \sum_j \epsilon_{T(X_n^{(0)})} = \hat{T}N_0$$

in $M_p(\mathbb{E}_2)$.

Proof. (a) Suppose $\mu_n \xrightarrow{v} \mu_0$. Let $f_2 \in C_K^+(\mathbb{E}_2)$ and we must show

$$\mu_n \circ T^{-1}(f_2) \to \mu_0 \circ T^{-1}(f_2). \tag{11.13}$$

Unpack the notation:

$$\mu_n \circ T^{-1}(f_2) = \int_{\mathbb{E}_2} f_2(e_2)\mu_n \circ T^{-1}(de_2)$$

and using the change of variable formula or transformation theorem for integrals ((Resnick, 1998, page 135)), this is

$$= \int_{\mathbb{E}_1} f_2(T(e_1))\mu_n(de_1).$$

Now f_2 and T are both continuous so $f_2 \circ T$ is continuous. Since $f_2 \in C_K^+(\mathbb{E}_2)$, there exists $K_2 \in \mathcal{K}(\mathbb{E}_2)$ such that $f_2(e_2) = 0$, if $e_2 \notin K_2$. So

$$f(T(e_1)) = 0, \text{ if } T(e_1) \notin K_2,$$

that is,

$$f(T(e_1)) = 0, \text{ if } e_1 \notin T^{-1}(K_2).$$

From the hypothesis (11.12), $T^{-1}(K_2) \in \mathcal{K}(\mathbb{E}_1)$. So, this says $f_2 \circ T$ is null off a compact set. Thus $f_2 \circ T \in C_K^+(\mathbb{E}_1)$, and since $\mu_n \xrightarrow{v} \mu_0$ in $M_+(\mathbb{E}_1)$, we have

$$\int_{\mathbb{E}_1} f_2(T(e_1))\mu_n(de_1) \to \int_{\mathbb{E}_1} f_2(T(e_1))\mu_0(de_1),$$

which gives (11.13).

(b) For $f_2 \in C_K^+(\mathbb{E}_2)$, it is enough to show

$$\mathbf{E}\Big(e^{-N_n \circ T^{-1}(f_2)}\Big) \to \mathbf{E}\Big(e^{-N_0 \circ T^{-1}(f_2)}\Big). \tag{11.14}$$

Again, we unpack the notation

$$\mathbf{E}\Big(e^{-N_n \circ T^{-1}(f_2)}\Big) = \mathbf{E}\Big(e^{-N_n(f_2 \circ T)}\Big)$$

by the transformation theorem. As in (a), $f_2 \circ T \in C_K^+(\mathbb{E}_1)$ and $N_n \Rightarrow N_0$ imply convergence on $C_K^+(\mathbb{E}_1)$ of the Laplace functionals. Thus

$$\mathbf{E}(e^{-N_n(f_2 \circ T)} \to \mathbf{E}(e^{-N_0(f_2 \circ T)} = \mathbf{E}(e^{-N_0 \circ T^{-1}(f_2)}$$

which is (11.14). $\square$

11.4.2. *Example: Sums of components of a random vector.* Given a random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_d$ of iid non-negative components whose common distribution has a regularly varying tail F,

$$1 - F(x) \sim x^{-\alpha} L(x), \quad x \to \infty. \tag{11.15}$$

What happens if we sum the components of $\boldsymbol{Z}$? For $\boldsymbol{t} \in \mathbb{R}^d$, define as usual

$$\boldsymbol{t} \cdot \boldsymbol{Z} = \sum_{i=1}^{d} t^{(i)} Z^{(i)}.$$

Proposition 12. *Suppose* (11.15) *holds. Then for* $\boldsymbol{t} > \boldsymbol{0}$,

$$\mathbb{P}[\boldsymbol{t} \cdot \boldsymbol{Z} > x] \sim \sum_{i=1}^{d} (t^{(i)})^{\alpha} x^{-\alpha} \sim \sum_{i=1}^{d} (t^{(i)})^{\alpha} P[Z^{(1)} > x], \quad x \to \infty.$$

Proof. Suppose $d = 2$ and $\boldsymbol{t} = \boldsymbol{1}$. Let

$$b(t) = \left(\frac{1}{1 - F} \right)^{\leftarrow} (t)$$

and let ν_α be the measure on $(0, \infty]$ such that $\nu_\alpha(x, \infty] = x^{-\alpha}$. First we claim

$$nP[\Big(\frac{Z_1}{b(n)}, \frac{Z_2}{b(n)} \Big) \in \cdot] \xrightarrow{v} \nu_\alpha \times \epsilon_{\{0\}} + \epsilon_{\{0\}} \times \nu_\alpha =: \mu^{(0)},$$

on $\mathbb{E} := [0, \infty]^2 \setminus \{\boldsymbol{0}\}$. (The limit measure on the punctured plane is concentrating all mass on the axes.) To see this note for $x > 0,\ y > 0$,

$$nP[\frac{Z_1}{b(n)} > x, \frac{Z_2}{b(n)} > y] = n\bar{F}(b(n)x)\bar{F}(b(n)y) \sim x^{-\alpha} \cdot 0 = 0,$$

so the limit has no mass in the interior. Likewise,

$$nP[\frac{Z_1}{b(n)} > x, \frac{Z_2}{b(n)} \geq 0] \to x^{-\alpha},$$

so while there is no mass in the interior, there must certainly be mass on the axes.

Now let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ be iid copies of the vector (Z_1, Z_2). From Basic Convergence Theorem 10, we get

$$\sum_{i=1}^{n} \epsilon_{\boldsymbol{Z}_i / b(n)}(\cdot) \Rightarrow \mathrm{PRM}(\mu^{(0)}).$$

Define $T : \mathbb{E} \mapsto (0, \infty]$ by $T(x, y) = x + y$. Then applying Proposition 11 (the compactness condition must be checked)

$$\sum_{i=1}^{n} \epsilon_{T(\boldsymbol{Z}_i / b(n))}(\cdot) = \sum_{i=1}^{n} \epsilon_{(Z_i(1) + Z_i(2))/b(n))}(\cdot) \Rightarrow \mathrm{PRM}(\mu^{(0)} \circ T^{-1}) = \mathrm{PRM}(2\nu_\alpha).$$

Again applying Basic Convergence Theorem 10, we get the regular variation of the distribution of the tail sum. □

For a traditional proof, see Feller (1971).

11.4.3. *Application:Products; Breiman's theorem.* Here we take a random variable Z and multiply by a scalar random variable with a relatively thin tail and prove a result by Breiman (1965).

Proposition 13. *Suppose Z is a non-negative random variable satisfying* (11.15) *or*

$$n\mathbb{P}[\frac{Z}{b_n} \in \cdot] \xrightarrow{v} \nu_\alpha.$$

We further suppose that $Y \geq 0$ is a random variable with a moment greater than α. This is equivalent to the existence of $\epsilon > 0$, such that

$$\mathbf{E}(Y^{\alpha(1+2\epsilon)}) < \infty. \tag{11.16}$$

Then

$$n\mathbb{P}[\frac{YZ}{b_n} \in \cdot] \xrightarrow{v} \mathbf{E}(Y^\alpha)\nu_\alpha$$

so that

$$\lim_{x\to\infty} \frac{\mathbb{P}[YZ > x]}{\mathbb{P}[Z > x]} = E(Y^\alpha).$$

Proof. Suppose $\{Z_n, n \geq 1\}$ are iid copies of Z. The regular variation condition is equivalent to

$$\sum_{i=1}^n \epsilon_{Z_i/b_n} \Rightarrow \sum_k \epsilon_{j_k} = \text{ PRM}(\nu_\alpha),$$

in $M_p((0,\infty])$. Now let $\{Y_j\}$ be iid copies of Y which are independent of $\{Z_n\}$ as well as independent of $\{j_k\}$. It follows that

$$\sum_{i=1}^n \epsilon_{(Z_i/b_n, Y_i)} \Rightarrow \sum_k \epsilon_{(j_k, Y_k)} = \text{ PRM}(\nu_\alpha \times \mathbb{P}[Y \in \cdot]) \tag{11.17}$$

in $M_p((0,\infty] \times (0,\infty))$. The reason (11.17) is true is that by Theorem 10 it suffices to check that the mean measures converge which easily follows from independence:

$$n\mathbb{P}[(\frac{Z_1}{b_n}, Y_1) \in \cdot] = n\mathbb{P}[\frac{Z_1}{b_n} \in \cdot] \times \mathbb{P}[Y_1 \in \cdot] \xrightarrow{v} \nu_\alpha \times \mathbb{P}[Y_1 \in \cdot].$$

Define the product map $T_p : (0,\infty] \times (0,\infty) \mapsto (0,\infty]$ by $T_p(z,y) = yz$. The compactness condition fails! However, after a truncation of the state space and a Slutsky type argument, we get Breiman's result. Pretending all is well and applying T_p to (11.17) we get

$$\sum_{i=1}^n \epsilon_{T_p(Z_i/b_n, Y_i)} \Rightarrow \sum_k \epsilon_{T_p(j_k, Y_k)} = \text{ PRM}(\nu_\alpha \times \mathbb{P}[Y \in \cdot]) \circ T_p^{-1}).$$

Check that

$$\begin{aligned}\nu_\alpha \times \mathbb{P}[Y \in \cdot]) \circ T_p^{-1}(x,\infty] =& \nu_\alpha \times \mathbb{P}[Y \in \cdot])\{(u,y) : uy > x\} \\ =& \iint_{\{(u,y):uy>x\}} \nu_\alpha(du)\mathbb{P}[Y \in dy]\end{aligned}$$

$$= \int_{y>0} \Big(\int_{u>x/y} \nu_\alpha(du) \Big) \mathbb{P}[Y \in dy] = \int_{y>0} (x/y)^{-\alpha} \mathbb{P}[Y \in dy]$$
$$= EY^\alpha \cdot x^{-\alpha}.$$

Finish with another application of the Basic Convergence Theorem. □

For more detail, see Resnick (1986).

11.4.4. *Application: Karamata tauberian theorem for Laplace transforms.* Suppose $U : [0, \infty) \mapsto [0, \infty)$ is non-decreasing. We define the Laplace transform of U, or the measure associated with U, by

$$\hat{U}(\lambda) := \int_0^\infty e^{-\lambda x} U(dx). \tag{11.18}$$

Theorem 12. *For $0 < \rho < \infty$, we have*

$$U(t) \in RV_\rho, \quad t \to \infty,$$

iff

$$\hat{U}(1/t) \in RV_\rho, \quad t \to \infty,$$

and then

$$\hat{U}(1/t) \sim U(t)\Gamma(\rho + 1), \quad t \to \infty.$$

Proof. Let $b(t) = U^{\leftarrow}(t)$ and suppose $U \in RV_\rho,\ \rho > 0$. Then define $U_t(x) = t^{-1}U(b(t)x)$ and $U_0(x) = x^\rho,\ x \geq 0$ and

$$U_t \to U_0, \quad n \to \infty. \tag{11.19}$$

For $t \geq 0$, let $N_t = \sum_k \epsilon_{j_k^{(t)}}$ be PRM(U_t) on $[0, \infty)$, and from (11.19) we have

$$N_t \Rightarrow N_0, \quad \text{in } M_p([0, \infty)).$$

If $\{E_k, k \geq 1\}$ are iid unit exponential random variables independent of all the $\{N_t\}$, then for each t,

$$N_t^+ = \sum_k \epsilon_{(j_k^{(t)}, E_k)} = \text{ PRM } (U_t(dy)e^{-s}ds)$$

(from augmentation) on the state space $[0, \infty) \times (0, \infty)$. Then

$$N_t^+ \Rightarrow N_0^+, \quad \text{in } M_p([0, \infty) \times (0, \infty)). \tag{11.20}$$

Now define the map $T_r : [0, \infty) \times (0, \infty) \mapsto [0, \infty)$ by $T_r(y, s) = y/s$. If we did not have to worry (but we do!) about whether inverse images of compact sets are compact, we could proceed as follows: Apply the map to the points of the convergence of (11.20) and we get a new sequence of Poisson processes converge and hence so do their mean measures. In particular we get

$$E\Big(N_t^+ \circ T_r^{-1}([0, x])\Big) \to E\Big(N_0^+ \circ T_r^{-1}([0, x])\Big),$$

since the Poisson random measures converge. Now

$$E\Big(N_t^+ \circ T_r^{-1}([0, x])\Big) = \iint_{\{(y,s):0 \leq \frac{y}{s} \leq x\}} U_t(dy)e^{-s}dx$$

$$
\begin{aligned}
&= \int_{y=0}^{\infty} \Big(\int_{s \geq y/x} e^{-s} ds \Big) U_t(dy) \\
&= \int_{y=0}^{\infty} e^{-y/x} U_t(dy) = \hat{U}_t(\frac{1}{x}) \\
&= t^{-1} \hat{U}(\frac{1}{b(t)x}).
\end{aligned}
$$

Likewise

$$
\begin{aligned}
E\Big(N_0^+ \circ T_r^{-1}([0,x]) \Big) &= \int_0^{\infty} e^{-y/x} \rho y^{\rho-1} dy \\
&= \rho x^{\rho} \int_0^{\infty} e^{-s} s^{\rho-1} ds = \rho x^{\rho} \Gamma(\rho) \\
&= x^{\rho} \Gamma(\rho+1).
\end{aligned}
$$

Therefore setting $x = 1$,

$$
\lim_{t \to \infty} t^{-1} \hat{U}(\frac{1}{U^{\leftarrow}(t)}) = \lim_{s \to \infty} \frac{\hat{U}(\frac{1}{s})}{U(s)} \to \Gamma(\rho+1).
$$

□

For more detail, see Resnick (1991).

11.5. **Did it achieve total heaviosity? (Alvie (Woody Allen) to Annie (Diane Keaton) in** *Annie Hall*). This was a brief introduction and there are almost infinitely other topics that could be discussed. A sampling:

(1) The multivariate case. If $\boldsymbol{Z} \in \mathbb{R}^d$ is a d-dimensional random vector, which for convenience we suppose has non-negative components, we say that the distribution of $\boldsymbol{Z}$ has a multivariate regularly varying tail (in standard form) if there exists $b(t) \to \infty$ such that

$$
t\mathbb{P}[\frac{\boldsymbol{Z}}{b(t)} \in \cdot] \xrightarrow{v} \nu
$$

in $M_+\big([0,\infty]^d \setminus \{\mathbf{0}\}\big)$. Here ν is a limit measure. After a polar coordinate transformation

$$
\nu\{\boldsymbol{x} : \|\boldsymbol{x}\| > r, \ \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \in \Lambda\} = r^{-\alpha} S(\Lambda),
$$

for $r > 0$, $\alpha > 0$, and S a finite measure on the unit sphere with respect to the selected norm.

Some questions:

(a) How do you specify the dependence structure of the random vector? (Correlations are no good and may not even exist.)
(b) The stated definition assumes components of the random vector are tail equivalent; that is, they have essentially the same tail. This doesn't happen in practice. How to we broaden the theory so it has applicability?
(c) How is the limit measure ν or the angular measure S estimated?

(d) How do you account for the effect of asymptotic independence on inference procedures?

(2) Limit theory for reasonably conventional functionals.

(a) The Poisson transform (11.9) has an easy extension where a time component is added: Suppose $\{Z_i\}$ are iid, non-negative random variables with common distribution F with a regularly varying tail. Then (11.9) extends to

$$\sum_{i=1}^{\infty} \epsilon_{\left(\frac{k}{n}, \frac{Z_i}{b_n}\right)} \Rightarrow N_\infty := \mathrm{PRM}(\mathrm{Leb} \times \nu_\alpha). \tag{11.21}$$

(b) Define

$$T_\vee : M_p([0,\infty) \times (0,\infty]) \mapsto D[0,\infty)$$

by

$$T_\vee\Big(\sum_i \epsilon_{(t_i, j_i)}\Big) = \bigvee_{t_i \le t} j_i =: Y(t).$$

Apply $T_\vee$ to (11.21) and use the continuous mapping theorem to get in $D[0,\infty)$

$$\bigvee_{i=1}^{[nt]} \frac{Z_i}{b(n)} \Rightarrow Y(t),$$

and for $x > 0$

$$\begin{aligned} P[Y(t) \le x] =& P[N_\infty([0,t] \times (x,\infty]) = 0] \\ =& \exp\{-\mathrm{Leb} \times \nu_\alpha[0,t] \times (x,\infty]\} \\ =& e^{-tx^{-\alpha}} = (\Phi_\alpha(x))^t. \end{aligned}$$

(c) Define the map

$$T_{\Sigma,\delta} : M_p([0,\infty) \times (0,\infty] \mapsto D[0,\infty)$$

by

$$T_{\Sigma,\delta}\big(\sum_k \epsilon_{(t_k, j_k)}\big) \sum_{t_k \le t} j_k 1_{|j_k| > \delta}.$$

Apply this to (11.21). The limit is a compound Poisson process. If $0 < \alpha < 2$, then after centering and letting $\delta \to 0$, we get the invariance principle for partial sum processes converging to a stable Lévy motion. The result is dimensionless and easily extends to higher dimensional sums.

(3) Queueing and Data network applications.

(a) The infinite source Poisson model, or the superposition of on/off processes give cumulative input that is approximated either by stable Lévy motion or fractional Brownian motion or .... This assumes heavy tailed durations for unit rate inputs. Generalizations to where input rates are random depending on the session are possible.

(b) Heavy traffic limit theorems are possible when the n-th customer arrives with a heavy tailed amount of work.

References

L. Breiman. On some limit theorems similar to the arc-sin law. *Theory Probab. Appl.*, 10: 323–331, 1965.

W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, New York, 2nd edition, 1971.

S. Resnick. Point processes, regular variation and weak convergence. *Adv. Applied Probability*, 18:66–138, 1986.

S. Resnick. Point processes and Tauberian theory. *Math. Sci.*, 16(2):83–106, 1991. ISSN 0312-3685.

S. Resnick. *A Probability Path*. Birkhäuser, Boston, 1998.