The Fundamentals of Heavy Tails: Properties, Emergence, and Estimation

Jayakrishnan Nair, Adam Wierman, and Bert Zwart

Table of Contents

Chapter 1

Introduction

> "The top 1% of the population controls 35% of the wealth. On Twitter, the top 2% of users send 60% of the messages. In the health care system, the treatment for the most expensive fifth of patients create four-fifths of the overall cost. These figures are always reported as shocking, as if the normal order of things has been disrupted, as if [it] is a surprise of the highest order. It's not. Or rather, it shouldn't be." – Clay Shirky, in response to the question "What scientific concept would improve everybody's cognitive toolkit?" [172]

Introductory probability courses often leave the impression that the Gaussian distribution is what we should expect to see in the world around us. It is referred to as the "Normal" distribution after all! As a result, statistics like the ones in the quote above tend to be treated as aberrations, since they would never happen if the world were Gaussian. The Gaussian distribution has a "scale", a typical value (the mean) around which individual measurements are centered and do not deviate from by too much. For example, if we consider human heights, which are approximately Gaussian, the average height of an adult male in the US is 5'9" and most people's heights do not differ by more than 10" from this. In contrast, there are order-of-magnitude differences between individuals in terms of wealth, twitter followers, health care costs, etc.

However, statistics like those above are not new and should not be surprising. Over a century ago, Italian economist Vilfredo Pareto discovered that the richest 20% of the population controlled 80% of the property in Italy. This is now termed the "Pareto Principle," a.k.a. the "80-20", rule and variations of this principle have shown up repeatedly in widely disparate areas in the time since Pareto's discovery. For example, in 2002 Microsoft reported that 80% of the errors in Windows are caused by 20% of the bugs and 20% of customers [167] and similar versions of the Pareto principle apply (though not always with 80/20) to many aspects of business, e.g., most of the profit is made from a small percentage of the customers and most of the sales are made by a small percentage of the sales team.

Statistics related to the Pareto principle make for compelling headlines, but they are typically an indication of something deeper. When we see such figures, it is likely that there is not a Gaussian distribution underlying them, but rather a heavy-tailed distribution is the reason for the "surprising" statistics. The most celebrated such distribution again carries Vilfredo Pareto's name: *the Pareto distribution*.

Because of the prominent role of the central limit theorem in introductory probability and statistics courses, such courses leave the impression that we should expect to see the Gaussian distribution everywhere in the world around us. Indeed, the Gaussian distribution is common; however, heavy-tailed distributions such as the Pareto distribution are just as (if not more) prominent. The Pareto distribution, in particular, has been observed in hundreds of applications in physics, biology, computer science, the social sciences, and beyond over the past century. Some examples include the sizes of cities [79, 146], the file sizes in computer systems and networks [49, 130], the size of avalanches and earthquakes [95, 128], the length of protein sequences in genomes [115, 129], the size of meteorites [11, 145], the degree distribution of the web graph [33, 101], the returns of stocks [46, 81], the number of copies of books sold [12, 96], the number of households affected during blackouts in power grids [100], the frequency of word use in natural language [66, 200], and many more.

Given the breadth of areas where heavy-tailed phenomena have been observed, one might guess that, by now, observations of heavy-tailed phenomena in new areas are expected – that heavy tails are treated as "more normal than the Normal." After all, it has been more than a century since Pareto's work. However, despite a century of history, statistics related to the Pareto Principle and, more broadly, heavy-tailed distributions are still typically presented as surprising curiosities – anomalies that could not have been anticipated. Even in scientific communities, observations of heavy-tailed phenomena are often viewed as mysteries to be explained rather than something to be expected a priori. In many cases, there is even a significant amount of controversy and debate that follows the identification of heavy-tailed phenomena in data.

Surprising? Mysterious? Controversial? Given the century of mathematical and statistical work around heavy tails, it certainly should not be the case that heavy tails are still surprising, mysterious, and controversial. In fact, there are many reasons why one should *expect* to see heavy-tailed distributions arise. Perhaps the main reason why they are still viewed as surprising is that the version of the central limit theorem taught in introductory probability courses gives the impression that the Gaussian will occur everywhere. However, this introductory version of the central limit theorem does not tell the whole story. There is a "generalized" version of the central limit theorem that states that either the Gaussian *or a heavy-tailed distribution* will emerge as the limit of sums of random variables. Unfortunately, the technical nature of this result makes it so that it is typically not presented in introductory courses, which leads to unnecessary surprises about the presence of heavy-tailed distributions. Going beyond sums of random variables, when random variables are combined in other natural ways (e.g., products, or max/min) then heavy tails are even more likely to emerge, whereas the Gaussian distribution is no longer prominent.

So heavy-tailed phenomena should not be considered surprising. What about mysterious? The view of heavy tails as mysterious is, to some extent, a consequence of unfamiliarity. People are familiar with the Gaussian distribution because of its importance in introductory probability courses, and when something emerges that has qualitatively and quantitatively different properties it seems mysterious and counter-intuitive. The Pareto Principle is one illustration of the counter-intuitive properties that make heavy-tailed distributions seem mysterious, but there are many others. For example, while the Gaussian distribution has a clear "scale" – most samples will be close to the mean – samples from heavy-tailed distributions frequently differ by orders of magnitude and may even be "scale free," e.g., in the case of the Pareto distribution. Another example is that, while the moments (i.e., the mean, variance, etc.) of the Gaussian distribution are all finite, it is not uncommon to see data that fits a heavy-tailed distribution having an infinite variance, or even an infinite mean! For example, the degree distribution of many complex networks tends to have a tail that matches that of a Pareto with infinite variance (see, for example, [24]). This can potentially lead to mind-bending challenges when trying to apply statistical tools, which often depend on averages and variances.

The combination of surprise and mystery that accompanies heavy-tailed phenomena means that there is often considerable excitement that follows the discovery of data that fits a heavy-tailed distribution in a new field. Unfortunately, this excitement often leads to debate and controversy. At this point, there is an unfortunate pattern that has emerged. A heavy-tailed phenomenon is discovered in a new field. The excitement over the discovery leads researchers to search for heavy tails in other parts of the field. Heavy tails are then discovered in many places and are claimed to be a universal property. However, the initial excitement of discovery and lack of previous background in statistics related to heavy tails means that the first wave of research identifying heavy tails uses intuitive, but flawed statistical tools. As a result, a controversy emerges – which places where heavy tails have been observed really have heavy tails? Are they really universal? Over time, more careful statistical analyses are used, showing that some places really do exhibit heavy tails while others were false discoveries. By the end, a mature view of heavy tails emerges, but the whole process can take decades.

At this point, the pattern has been replicated in many areas, including computer science [60], biology [105], chemistry [143], ecology [9], astronomy [189]. Maybe the most prominent example of this story is still ongoing in the area of *network science*. Near the turn of the century, the study of complex networks began to explode in popularity due to the growing importance of networks in our lives and the increasing ease of gathering data about large networks. Initial results in the area were widely celebrated and drove an enormous amount of research to look at the universality of scale free networks. However, as the field matured and the statistical tools became more sophisticated, it became clear that many of the initial results were flawed. For example, claims that the internet graph [68] and the power network [21] are heavy-tailed were refuted [3, 195], among others. This led to a controversy in the area that continues to this day, twenty years later [34, 186].

Demystifying heavy tails. The goal of this book is to demystify heavy-tailed phenomena. Heavy tails are not mysterious anomalies – and their emergence should not be surprising or controversial either! Heavy tails are an unavoidable part of our lives and viewing statistics like the ones that started this chapter as anomalies prevents us from thinking clearly about the world around us. Further, while properties of heavy-tailed phenomena like the Pareto Principle may initially make heavy-tailed distributions seem counter-intuitive, they need not be. This book strives to provide tools and techniques that can make heavy tails as easy and intuitive to reason about as the Gaussian, to highlight when one should expect the emergence of heavy-tailed phenomena, and to help avoid controversy when identifying heavy tails in data.

Because of the ubiquitousness and seductive nature of heavy-tailed phenomena, they are a topic that has permeated wide ranging fields, from astronomy and physics to biology and physiology to social science and economics. However, despite their ubiquity, they are also, perhaps, one of the most misused and misunderstood mathematical areas, shrouded in both excitement and controversy. It is easy to get excited about heavy-tailed phenomena as you start to realize the important role they play in the world around us and become exposed to the beautiful and counter-intuitive properties they possess. However, as you start to dig into the topic, it quickly becomes difficult. The mathematics that underlie the analysis of heavy-tailed distributions are technical and advanced, often requiring prerequisites of graduate level probability and statistics courses. This is the reason why introductory probability courses typically do not present much, if any, material related to heavy-tailed distributions. If they are mentioned, they are typically used as examples illustrating that "strange" things can happen, e.g., distributions can have an infinite mean. Thus, a scientist or researcher in a field outside of mathematics who is interested in learning more about heavy tails may find it difficult, if not impossible, to learn from the classical texts on the topic.

It is exactly this difficulty that led us to write this book. In this book we hope to introduce the fundamentals of heavy-tailed distributions using only tools that one learns in an introductory probability course. The book intentionally does not spend much time on describing everywhere heavy tails come up – there are simply too many different areas to do justice to even a small subset of them. Instead, we assume that if you have found your way to this book, then heavy tails are important to you. Given that, our goal is to provide an introduction to how to think about heavy tails both intuitively and mathematically.

The book is divided into three parts, which focus on three of the most foundational questions about heavy tails.

- **Part I: Properties.** *What leads to the counter-intuitive properties of heavy-tailed phenomena?*
- **Part II: Emergence.** *Why do heavy-tailed phenomena occur so frequently in the world around us?*
- **Part III: Estimation** *How can we identify and estimate heavy-tailed phenomena using data?*

In Part I of the book we provide insight into some of most mysterious and elegant properties of heavy-tailed distributions, connecting these properties to formal definitions of subclasses of heavy-tailed distributions. We focus on three foundational properties: "scale-invariance" (a.k.a., scale-free), the "catastrophe principle", and "increasing residual life." We illustrate that these properties provide qualitatively different behaviors than what is seen under light-tailed distributions like the Gaussian, and provide intuition underlying the properties. The three chapters that make up Part I strive to demystify some of the particularly exotic properties of heavy-tailed distributions, and to provide a clear view of how these properties interact with each other and with the broader class of heavy-tailed distributions.

In Part II of the book we seek to explore simple laws that can "explain" the emergence of heavy-tailed distributions similarly to the way in which the central limit theorem "explains" the prominence of the Gaussian distribution. We study three foundational stochastic processes in order to understand when one should expect the emergence of heavy-tailed distributions as opposed to light-tailed distributions. Our discussions in the three chapters that make up Part II highlight that heavy-tailed distributions should not be viewed as anomalies. In fact, heavy tails should not be surprising at all, in many cases they should treated as something as natural as, if not more natural than, the emergence of the Gaussian distribution.

In Part III of this book we focus on providing an introduction to the statistical tools used for the estimation of heavy-tailed phenomena. Unfortunately, there is no perfect recipe for how to "properly" detect and estimate heavy-tailed distributions in data. Thus, our treatment seeks to highlight a handful of important approaches, and to provide insight into when each approach is appropriate and when each may be misleading. Combined, the three chapters that make up Part III highlight a crucial point: one must proceed carefully when seeking to estimate heavy-tailed phenomena in real-world data. It is typically naive to seek to estimate *exact* heavy-tailed distributions in data. Instead, the focus should be on estimating the *tail* of heavy-tailed phenomena. However, even in doing this, one should not rely on a single method for estimation. Instead, it is a necessity to build confidence through the use of multiple, complementary estimation approaches.

1.1 Defining heavy-tailed distributions

Before we can move to study of the questions discussed above, the starting point for a book that seeks to demystify heavy-tailed phenomena has to be the basic question: *What is a heavy-tailed distribution?*

One of the reasons for the mystique that surrounds heavy-tailed distributions is that if you ask five people from different communities this question, you are likely to get five different answers. Depending on the community, the term heavy-tailed may be used interchangeably with terms like scale-free, power-law, fat-tailed, long-tailed, subexponential, self-similar, stable, etc. Or, each of these terms may be used to mean something different. Further, the names may mean different things to different communities!

Sometimes the term "heavy-tailed" may be used to refer to specific distributions such as Pareto or Zipf distributions. Other times it may be used to refer to particular properties of a distribution, such as the fact that a distribution is scale-free, has an infinite (or very large) variance, a decreasing failure rate, etc. As a result, there is often a language barrier when discussing heavy-tailed distributions that stems from different associations with the same terms across communities.

Hopefully, by reading this book, the zoo of terminology related to heavy-tailed distributions will not be as difficult for you to navigate. Each of the terms mentioned above does have a concrete, precise, established mathematical definition. It is just that these terms are often used carelessly, which leads to confusion. It will take us most of the book to get through the definitions of all the terms mentioned in the previous paragraph, but we start in this section by laying the foundation – defining the term "heavy-tailed" and discussing some of the most celebrated examples.

The term "heavy-tailed" is inherently relative – heavier than what? A Gaussian distribution has a heavier tail than a Uniform distribution and an Exponential distribution has a heavier tail than a Gaussian distribution, but none of these are considered "heavy-tailed". Thus, the key feature of the definition is the comparison point chosen.

The comparison point that is used to define the class of heavy-tailed distributions is the Exponential distribution. That is, a distribution is considered to be heavy-tailed if it has a heavier tail than any Exponential distribution. Formally, this is stated in terms of the cumulative distribution function (c.d.f.) F of a random variable X, i.e., $F(x) = \Pr(X > x)$, and the complementary cumulative density function (c.c.d.f) $\bar{F}$, i.e., $\bar{F}(x) = 1 - F(x)$.

Definition 1.1. *A distribution function F is said to be heavy-tailed if and only if, for all $\mu > 0$,*

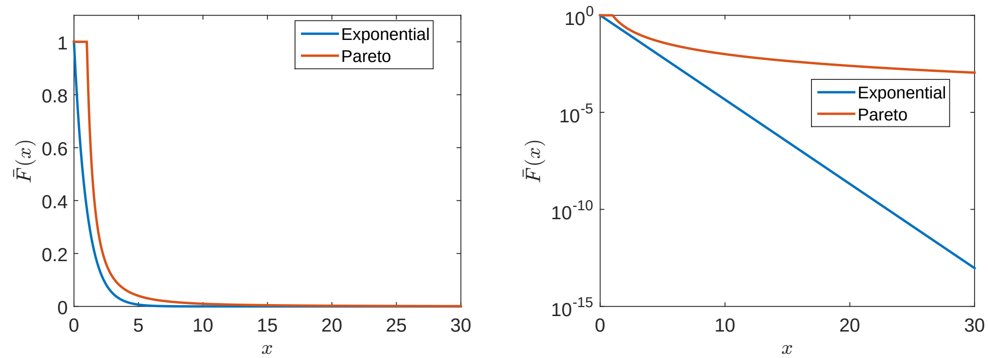
$$\limsup_{x\to\infty} \frac{1 - F(x)}{e^{-\mu x}} = \limsup_{x\to\infty} \frac{\bar{F}(x)}{e^{-\mu x}} = \infty.$$

Otherwise, F is light-tailed. A random variable X is said to be heavy-tailed (light-tailed) if its distribution function is heavy-tailed (light-tailed).

Note that the definition of heavy-tailed distributions given above applies to *right* tail of the distribution, i.e., it is concerned with the behavior of the probability of taking values larger than x as $x \to \infty$. In some applications, one might also be interested in the *left* tail. In such cases, the definition of heavy-tailed can be applied to both the right tail (without change) and the left tail (by considering the right tail of $-X$).

The definition of heavy-tailed is, in some sense, quite natural. It explicitly looks at the "tail" of the distribution, i.e., the complementary cumulative distribution function (c.c.d.f.) $\bar{F}(x) = 1 - F(x) = \Pr(X > x)$ for large x, and it is easy to see from the definition that the tails of distributions that are heavy-tailed are "heavier", i.e., decay more slowly, than the tails of distributions that are light-tailed; see Figure 1.1.

The particular choice of the Exponential distribution as the boundary between heavy-tailed and light-tailed may, at first, seem arbitrary. In fact, without detailed study of the class of heavy-tailed distributions it



(a) Linear scale (b) Log scale

Figure 1.1: Contrasting heavy-tailed and light-tailed distributions: *The plots show the c.c.d.f. of the exponential distribution (with mean 1) and a heavy-tailed Pareto distribution (with $x_m = 1$, scale parameter $\alpha = 2$). While the contrast in tail behavior is difficult to discern on a linear scale (Fig (a)), it is quite evident when the probabilities are plotted on a logarithmic scale (Fig (b)).*

is difficult to justify this particular choice. But, as we will see throughout this book, the Exponential distribution serves to separate two classes of distributions that have qualitatively different behavioral properties and require fundamentally different mathematical tools to work with.

To begin to highlight the distinction between heavy-tailed and light tailed distributions, it turns out to be useful to consider two alternative, but equivalent, definitions of "heavy-tailed".

Lemma 1.1. *Consider a random variable X. The following statements are equivalent.*

(i) *X is heavy-tailed.*

(ii) $M(s) := \mathbb{E}\left[e^{sX}\right] = \infty$ *for all* $s > 0$.

(iii) $\liminf_{x\to\infty} -\frac{\log Pr(X>x)}{x} = 0$.

The proof of this lemma provides useful intuition about heavy-tailed distribution; however, before proving this result, let us interpret the two new, equivalent definitions of heavy-tailed that it provides.

First, consider (ii), which states that a random variable is heavy-tailed if and only if its moment generating function $M(s)$ is infinite for all $s > 0$. This definition highlights that heavy-tailed distributions require a different analytic approach than light-tailed distributions. For light-tailed distributions the moment generating function often provides an important tool for characterizing the distribution. It can be used to derive the moments of the distribution, but it also can be inverted to characterize the distribution itself. Further, it is a crucial tool for analysis because of the simplicity of handling convolutions, etc., via the moment generating function, for example, when deriving concentration inequalities such as Chernoff bounds. In contrast, the definition given by (ii) highlights that such techniques are not possible for heavy-tailed distributions.

Next, consider (iii), which states that a random variable X is heavy-tailed if and only if the log of its tail, $\log Pr(X > x)$, decays sub-linearly. This again highlights that heavy-tailed distributions require a different analytic approach than light-tailed distributions. In particular, when studying the tail of light-tailed distributions it is typical to use concentration inequalities such as Chernoff bounds, which inherently have an exponential decay. As a result, such bounds focus on determining the optimal decay rate, which is characterized by deriving a maximal μ such that $Pr(X > x) \leq Ce^{-\mu x}$. However, the definition given by (iii) highlights that the maximum possible μ for heavy-tailed distributions is zero, and so fundamentally different analytic approaches must be used.

To build more intuition for the relationship between these three equivalent definitions of "heavy-tailed", as well as to get practice working the definition of heavy-tailed distribution, it is useful to consider the proof of Lemma 1.1.

Proof of Lemma 1.1. To prove Lemma 1.1 we need to show the equivalence of each of the three definitions of heavy-tailed. We do this by showing that (i) implies (ii), that (ii) implies (iii), and finally that (iii) implies (i).

$\boldsymbol{(i) \Rightarrow (ii)}$**.** Suppose that X is heavy-tailed, with distribution F. By definition, this implies that for any $s > 0$, there exists a strictly increasing sequence $(x_k)_{k\geq 1}$ satisfying $\lim_{k\to\infty} x_k = \infty$, such that

$$\lim_{k\to\infty} e^{sx_k}\bar{F}(x_k) = \infty. \tag{1.1}$$

We can now bound $\mathbb{E}\left[e^{sX}\right]$ as follows.

$$\begin{aligned}\mathbb{E}\left[e^{sX}\right] &= \int_0^\infty e^{sx}dF(x)\\ &\geq \int_{x_k}^\infty e^{sx}dF(x)\\ &\geq e^{sx_k}\bar{F}(x_k).\end{aligned}$$

Since the above inequality holds for all k, it now follows from (1.1) that $\mathbb{E}\left[e^{sX}\right] = \infty$. Therefore, Condition (i) implies Condition (ii).

$\boldsymbol{(ii) \Rightarrow (iii)}$**.** Suppose that X satisfies Condition (ii). For the purpose of obtaining a contradiction, let us assume that Condition (iii) does not hold. Since $-\frac{\log \Pr(X>x)}{x} \geq 0$, this means that

$$\liminf_{x\to\infty} -\frac{\log \Pr\left(X > x\right)}{x} > 0.$$

The above statement implies that there exists $\mu > 0$ and $x_0 > 0$ such that

$$-\frac{\log \Pr\left(X > x\right)}{x} \geq \mu \iff \Pr\left(X > x\right) \leq e^{-\mu x} \quad \forall\, x \geq x_0. \tag{1.2}$$

Now, pick s such that $0 < s < \mu$. We may now bound the moment generating function of X at s as follows:

$$\begin{aligned} M(s) = \mathbb{E}\left[e^{sX}\right] &= \int_0^\infty \Pr\left(e^{sX} > x\right) dx \\ &= \int_0^{e^{sx_0}} \Pr\left(e^{sX} > x\right) dx + \int_{e^{sx_0}}^\infty \Pr\left(X > \frac{\log(x)}{s}\right) dx. \end{aligned}$$

Here, we have used the following representation for the expectation of any non-negative random variable Y: $\mathbb{E}[Y] = \int_0^\infty \Pr(Y > y)\, dy$. While the first term above can be bounded from above by e^{sx_0}, we may bound the second using (1.2), since $x \geq e^{sx_0}$ is equivalent to $\log(x)/s \geq x_0$.

$$\begin{aligned} M(s) &\leq e^{sx_0} + \int_{e^{sx_0}}^\infty e^{-\mu\frac{\log(x)}{s}} dx \\ &= e^{sx_0} + \int_{e^{sx_0}}^\infty x^{-\mu/s} dx. \end{aligned}$$

Since $\mu/s > 1$, we have $\int_1^\infty x^{-\mu/s} dx < \infty$, which implies that $M(s) < \infty$, giving us a contradiction. Therefore, Condition (ii) implies Condition (iii).

$\boldsymbol{(iii) \Rightarrow (i)}$**.** Suppose that the random variable X, having distribution F, satisfies Condition (iii). Thus, there exists a strictly increasing sequence $(x_k)_{k\geq 1}$ satisfying $\lim_{k\to\infty} x_k = \infty$, such that

$$\lim_{k\to\infty} -\frac{\log \bar{F}(x_k)}{x_k} = 0.$$

Given $\mu > 0$, this in turn implies that there exists $k_0 \in \mathbb{N}$ such that

$$\begin{aligned} & -\tfrac{\log \bar{F}(x_k)}{x_k} < e^{-\frac{\mu}{2}} && \forall\, k > k_0 \\ \iff \quad & \bar{F}(x_k) > e^{-\frac{\mu x_k}{2}} && \forall\, k > k_0 \\ \iff \quad & \tfrac{\bar{F}(x_k)}{e^{-\mu x_k}} > e^{\frac{\mu x_k}{2}} && \forall\, k > k_0. \end{aligned}$$

The last assertion above implies that $\lim_{k\to\infty} \frac{\bar{F}(x_k)}{e^{-\mu x_k}} = \infty$, which implies $\limsup_{x\to\infty} \frac{\bar{F}(x)}{e^{-\mu x}} = \infty$. Since this is true for any $\mu > 0$, we conclude that Condition (iii) implies Condition (i).

$\square$

1.2 Examples of heavy-tailed distributions

At this point we have three equivalent definitions of heavy-tailed distributions and, through the proof, we understand how these three definitions are related. But, even with these restatements, the definition of heavy-tailed is still opaque. It is difficult to get behavioral intuition about the properties of heavy-tailed distributions from any of the definitions. Further, it is very hard to see much about what makes heavy-tailed distributions have the mysterious properties that are associated with them using these definitions alone.

In part, this is due to the breadth of the definition of heavy-tailed. The important properties of heavy-tailed distributions that are most often used and cited are not easily seen to be consequences of this definition because, in fact, many properties such as scale invariance, infinite variance, the Pareto principle, etc., do not hold for all heavy-tailed distributions; they hold only for subclasses of heavy-tailed distributions.

As a result, it is important to build intuition for the class of heavy-tailed distributions by looking specific examples. That is the goal of the remainder of this chapter. In particular, we focus in detail on the Pareto distribution, the Weibull distribution, and the LogNormal distribution with the goal of providing both the mathematical formalism for these distributions and some insight in their important properties and applications. Additionally, we briefly introduce some of the other important examples of heavy-tailed distributions that come up frequently in applications, including the Cauchy, Fréchet, Lévy, Burr, and Zipf distributions.

Perhaps the most important thing to keep in mind as you read these sections is the contrast between the properties of the heavy-tailed distributions that we discuss and the properties of light-tailed distributions, such as the Gaussian and Exponential distributions, with which you are likely more familiar. For the ease of the reader, we summarize the important formulas for these two distributions below and illustrate the contrast between Gaussian and Exponential distributions and the Pareto distribution in Figure 1.2 before moving to the heavy-tailed distributions that are our focus.

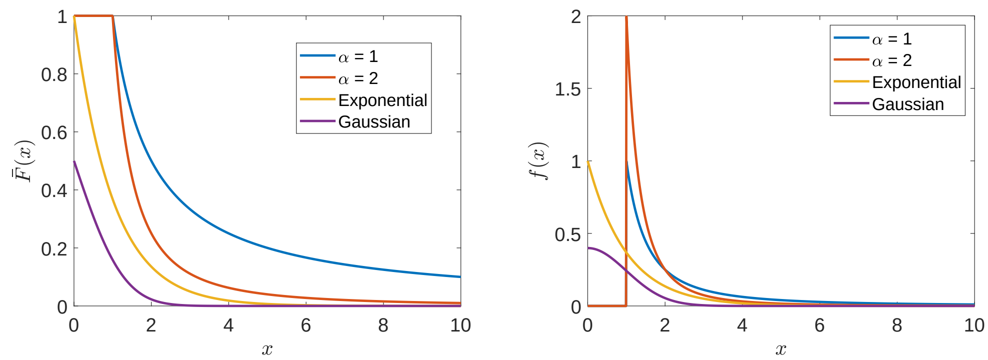
The Gaussian Distribution. The Gaussian distribution, a.k.a., the Normal distribution or the bell curve, is perhaps the most widely recognized distribution and is extremely important in statistics and beyond. It is defined using two parameters: the mean μ and the variance σ^2 and is expressed most conveniently through its probability density function (p.d.f.), $f(x)$, or its moment generating function (m.g.f.), $M(s)$. Given a random variable $Z \sim$ Gaussian(μ, σ), we have

$$f_Z(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$
$$M_Z(s) = E[e^{sZ}] = e^{\mu s + \frac{1}{2}\sigma^2 s^2}.$$

Since $M_Z(s) < \infty$ for all $s > 0$, it follows that the Gaussian distribution is light-tailed. The light-tailedness of the Gaussian distribution can also be deduced directly by bounding its c.c.d.f. (see Exercise 2).

The Exponential Distribution. The Exponential distribution is a widely known and broadly applicable distribution that serves as the light-tailed distribution on the boundary between light-tailed and heavy-tailed distributions. It is a nonnegative distribution defined in terms of one parameter: λ, which is referred to as the "rate" since the mean of the distribution is $1/\lambda$. Given a random variable $X \sim$ Exponential(λ), the p.d.f., c.c.d.f., and m.g.f., can be expressed as:

$$f_X(x) = \lambda e^{-\lambda x} \qquad (x \geq 0),$$
$$\bar{F}(x) = e^{-\lambda x} \qquad (x \geq 0),$$
$$M_X(s) = \frac{1}{(1 - s/\lambda)} \qquad (s < \lambda).$$



(a) c.c.d.f. (b) p.d.f.

Figure 1.2: Contrasting Pareto distribution with the Exponential and the Gaussian: *The plots show the c.c.d.f. (Fig. (a)) and p.d.f. (Fig. (b)) corresponding to different Pareto distributions with $x_m = 1$ with different values of α, alongside the Exponential distribution (with unit mean) and the standard Gaussian.*

1.2.1 The Pareto distribution

Vilfredo Pareto originally presented the Pareto distribution, and coined the idea of the Pareto Principle, in the study of the allocation of wealth, but since then it has been used as a model for numerous other settings including the sizes of cities, the file sizes in computer systems and networks, the price returns of stocks, the size of meteorites, causalities and damages due to natural disasters, frequency of works, and many more. It is perhaps the most celebrated example of a heavy-tailed distribution, and as a result, it is sometimes, unfortunately, used interchangeably with the term heavy-tailed.

Formally, a random variable X follows a Pareto(x_m, α) distribution if

$$\Pr(X \geq x) = \bar{F}(x) = \left(\frac{x}{x_m}\right)^{-\alpha}, \text{ for } \alpha > 0, x \geq x_m > 0.$$

Here, α is the shape parameter of the distribution and is also commonly referred to as the *tail index*, while x_m is the minimum value of the distribution, i.e., $X \geq x_m$. Given the c.c.d.f. above, it is straightforward to differentiate and obtain the p.d.f.

$$f(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, \quad x \geq x_m.$$

It is easy to see from the c.c.d.f. that the Pareto is heavy-tailed. In particular, using Definition 1.1, we can compute

$$\limsup_{x\to\infty} \frac{\bar{F}(x)}{e^{-\mu x}} = \left(\frac{x_m}{x}\right)^\alpha e^{\mu x} = \infty, \tag{1.3}$$

since the exponential $e^{\mu x}$ grows more quickly than the polynomial x^α.

This highlights the key contrast between the Pareto distribution and common light-tailed distributions like the Gaussian and Exponential distributions: the distribution decays *polynomially*, as $x^{-\alpha}$, instead of *exponentially* (as $e^{-\mu x}$) in the case of the Exponential distribution, or *superexponentially* (as $e^{-x^2/2\sigma^2}$) in

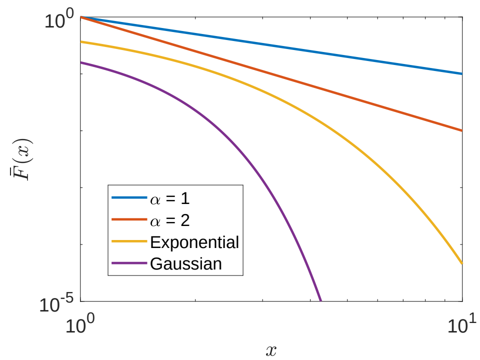


Figure 1.3: A clearer contrast between the Pareto distribution and the Exponential/Gaussian: *The plots show the c.c.d.f. corresponding to different Pareto distributions with $x_m = 1$ with different values of α, alongside the Exponential distribution (with unit mean) and the standard Gaussian, on a log-log scale. Notice how this scaling demonstrates clearly how the Pareto tail (linear on a log-log plot) is heavier than that of the Exponential and the Gaussian.*

the case of the Gaussian distribution. This means that large values are much more likely to occur under a Pareto distribution than under a Gaussian or Exponential distribution. For example, you are much more likely to meet someone who's income is 10 times the average than someone who's height is ten times the average.

This contrast is present visually too. Figure 1.2 highlights that the tail of the Pareto is considerably heavier. The figure illustrates the p.d.f. and c.c.d.f. of the Pareto for different values of the tail index α. Note that the tail index is typically the parameter of interest since it controls the degree of the polynomial decay in the Pareto, and thus determines the "weight" of the tail. As α decreases the tail becomes heavier, while as $\alpha \to \infty$ the Pareto distribution approaches the Dirac delta function centered around x_m.

While Figure 1.2 highlights a contrast between the Pareto, Gaussian, and Exponential distributions; we can present the figure in a different way to emphasize the contrast much more dramatically. By scaling the axes of the figure in different ways important properties and contrasts between the distributions emerge. In particular, Figure 1.3 shows the same c.c.d.f.s but presents the data on a log-log scale, i.e., with logarithmic horizontal and vertical axes. With this change, a remarkable pattern emerges – the Pareto c.c.d.f. becomes a straight line, while the Gaussian and Exponential distributions quickly drop off a cliff and disappear. This viscerally highlights the heaviness of the Pareto's tail as compared to the tails of the Exponential and the Gaussian.

To understand why the Pareto is linear when viewed on a log-log scale, let us do a quick calculation. Letting $C_1 = x_m^\alpha$ we can write

$$\bar{F}(x) = \left(\frac{x}{x_m}\right)^{-\alpha} = C_1 x^{-\alpha}.$$

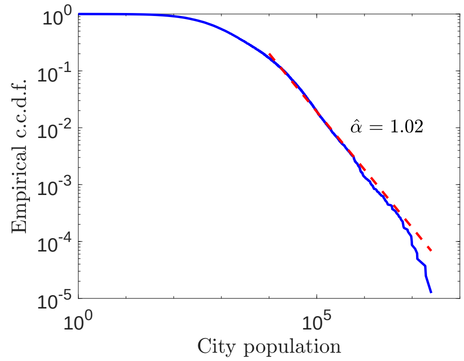


Figure 1.4: Visualizing data on a log-log plot: *The figure shows the empirical c.c.d.f. of the populations of U.S. cities as per the 2010 census (data sourced from [2]). Note that on a log-log scale, the data beyond population* 10^4 *looks approximately linear. The least squares regression line on this data yields a estimate* $\hat{\alpha} = 1.02$ *of the power law exponent.*

Taking logarithms of both sides then gives

$$\underbrace{\log \bar{F}(x)}_{'y'} = \underbrace{\log C_1}_{y\text{-intercept}} + \underbrace{(-\alpha)}_{\text{slope}} \underbrace{\log x}_{'x'},$$

which highlights that, in log-log scale, the c.c.d.f. is simply a linear function with y-intercept $\log C_1$ and slope $-\alpha$. Not only that, the p.d.f. is also of the same form, i.e., $f(x) = C_2 x^{-(\alpha+1)}$ where $C_2 = \alpha x_m^\alpha$ and so it also is linear in the log-log scaling.

This property – being (approximately) linear in log-log scale – is important enough that it has received a few different names from different communities. Distributions of the form $\bar{F}(x) = Cx^{-\alpha}$ for some constant C are referred to as *power law* distributions. A quite related set of distributions are *fat-tailed distributions*, which are distributions with $\bar{F}(x) \sim x^{-\alpha}$ as $x \to \infty$, where we use $a(x) \sim b(x)$ as $x \to \infty$ as shorthand for $\lim_{x\to\infty} a(x)/b(x) = 1$. Finally, the class of *regularly varying* distributions we introduce in Chapter 2 generalizes both power law and fat-tailed distributions and has strong connections to the concept of scale-invariance.

The fact that the Pareto distribution, and more generally power law distributions, are approximately linear on a log-log plot leads to a number of important properties. Maybe the most prominent is that it provides an intuitive exploratory tool for use in the identification of power laws in data. Specifically, when presented with data, it is possible to look at the empirical p.d.f. and c.c.d.f. on a log-log scale and see if it is approximately linear. If so, then there is the potential that the data is coming from a power-law distribution. One can even go further and hope to use such an approach to estimate the tail index α using linear regression on the empirical p.d.f. and c.c.d.f. This is a common approach across fields, and we illustrate it in Figure 1.4 using population data for US cities as per the 2010 census. Notice that the empirical c.c.d.f. (on a log-log scale) looks roughly linear for large populations. It is therefore tempting to postulate that the distribution of city populations (asymptotically) follows a power law, and further to estimate the tail index by fitting a least squares regression line to the empirical c.c.d.f. beyond, say 10^4 (since the tail 'looks linear' beyond this point), as shown in Figure 1.4. However, as we discuss in Chapter 8, this approach is not statistically sound and may lead to incorrect conclusions. In fact, the temptation to make conclusions based on such simple approaches is one of the most common reasons for the controversy that often surrounds the identification of heavy-tailed phenomena.

Moments. One of the biggest contrasts between the Pareto distribution and light-tailed distributions such as the Gaussian and Exponential distribution is the fact that the Pareto distribution can have infinite moments. In fact, for $X \sim \text{Pareto}(x_m, \alpha)$, $\mathbb{E}[X^n] = \infty$ if $n \geq \alpha$. More specifically, the mean of the Pareto distribution is:

$$\mathbb{E}[X] = \begin{cases} \infty, & \alpha \leq 1; \\ \frac{\alpha x_m}{\alpha - 1}, & \alpha > 1. \end{cases}$$

The variance is:

$$\text{Var}[X] = \begin{cases} \infty, & \alpha \in (1, 2]; \\ \left(\frac{x_m}{\alpha - 1}\right)^2 \frac{\alpha}{\alpha - 2}, & \alpha > 2. \end{cases}$$

And, in general, the nth moment is:

$$\mathbb{E}[X^n] = \begin{cases} \frac{\alpha x_m^n}{\alpha - n}, & n < \alpha; \\ \infty, & n \geq \alpha. \end{cases}$$

Importantly, it is not just a curiosity that the Pareto distribution can have infinite moments. In many cases where data has been modeled using the Pareto distribution, the distribution that is fit has an infinite variance and/or mean. For example, file sizes in computer systems and networks [49] and the degree distributions of complex networks such as the web [60] suggest infinite variance. Additionally, the logarithmic returns on stocks in finance tend to have a finite variance, but infinite fourth moment [64], leading to values of α in the range $(2, 4)$.

The Pareto principle. We began this chapter with a quote about the Pareto principle, so it is important to return to it now that we have formally introduced the Pareto distribution. The classical version of the Pareto principle is that the wealthiest 20% of the population holds 80% of the wealth. Mathematically, we can ask a more general question about what fraction of the wealth the largest P fraction of the population holds.

To compute an analytic version of the Pareto principle, we consider the fraction of the population whose wealth exceeds x. Call this fraction $P(x)$, and then we can calculate $P(x)$ in the case of a Pareto distribution as follows:

$$P(x) = \int_x^\infty f(t)dt = \alpha x_m^\alpha \int_x^\infty t^{-(\alpha+1)} dt = \left(\frac{x}{x_m}\right)^{-\alpha}.$$

Then, the fraction of wealth that is in the hands of such people, which we denote by $W(x)$, is

$$W(x) = \frac{\int_x^\infty t f(t) dt}{\int_{x_m}^\infty t f(t) dt} = \frac{\alpha x_m^\alpha \int_x^\infty t^{-\alpha} dt}{\alpha x_m^\alpha \int_{x_m}^\infty t^{-\alpha} dt} = \left(\frac{x}{x_m}\right)^{-\alpha+1},$$

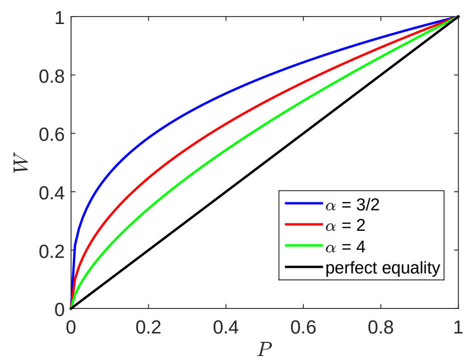


Figure 1.5: Lorenz curves for the Pareto distribution: *Lorenz curves for different values of α. Note that the smaller the value of α, the more pronounced the concentration of wealth within a small fraction of the population. The black line represents perfect equality, i.e., the utopian scenario where all individuals have exactly the same wealth.*

assuming that $\alpha > 1$. Combining the above equations then gives that, regardless of x, the fraction of wealth W owned by the richest P fraction of the population is

$$W = P^{(\alpha-1)/\alpha}.$$

We illustrate the curve of W as a function of P in Figure 1.5. It is always concave and increasing, and when α is close to 1, it highlights that wealth is concentrated in a very small fraction of the population. This is an example of a more general phenomenon called the "catastrophe principle", which we discuss in detail in Chapter 3.

Curves like those in Figure 1.5 are referred to as Lorenz curves, after Max Lorenz, who developed them in 1905 as a way to represent the inequality of wealth distribution. The Gini coefficient, which is typically used to quantify income inequality today, is the ratio of the area between the line of perfect equality (the 45 degree line) and the Lorenz curve, and the area above the line of perfect equality. The greater the value of the Gini coefficient, the more pronounced the asymmetry in wealth distribution.

Relationship to the Exponential distribution. While heavy-tailed distributions often behave qualitatively differently than light-tailed distributions, there are still some connections between the two that can be useful. In particular, often, a heavy-tailed distribution can be viewed as an exponential transformation of a light-tailed distribution. In the case of the Pareto, this connection is to the Exponential distribution. Specifically,

$$X \sim \text{Pareto}(x_m, \alpha) \quad \Longleftrightarrow \quad \log(X/x_m) \sim \text{Exponential}(\alpha)$$

Or, equivalently,

$$Y \sim \text{Exponential}(\alpha) \quad \Longleftrightarrow \quad x_m e^Y \sim \text{Pareto}(x_m, \alpha).$$

To see why this is true requires a simple change of variables. In particular, let $Y = \log(X/x_m)$ where

$X \sim \text{Pareto}(x_m, \alpha)$. Then,

$$\Pr(Y > x) = \Pr(\log(X/x_m) > y) = \Pr(X > x_m e^y) = \left(\frac{x_m e^y}{x_m}\right)^{-\alpha} = e^{-\alpha y},$$

where the last expression is the c.c.d.f. of an Exponential distribution with rate α. This transformation turns out to be a powerful analytic tool, and we make use of it on multiple occasions in this book, e.g., Chapter 6, to study multiplicative processes, and Chapter 8, to derive properties of the maximum likelihood estimator for data from a Pareto distribution.

1.2.2 The Weibull distribution

We just saw that the Pareto distribution has an intimate connection to the Exponential distribution – it is an *exponential* of the Exponential. The second heavy-tailed distribution we introduce has a similar connection to the Exponential distribution – it is a *polynomial* of the Exponential. Specifically, for $\alpha, \beta > 0$,

$$X \sim \text{Exponential}(1) \quad \Longleftrightarrow \quad \frac{1}{\beta} X^{1/\alpha} \sim \text{Weibull}(\alpha, \beta).$$

From the above relationship, one would expect that when $0 < \alpha < 1$, the Weibull distribution has a heavier tail than the Exponential (though lighter than the Pareto), making it heavy-tailed. On the other hand, when $\alpha > 1$, one would expect that the Weibull has a lighter tail than the Exponential.

It is straightforward to see what this transformation means for the c.c.d.f. of the Weibull distribution. In particular, a random variable follows a Weibull(α, β) distribution if,

$$\bar{F}(x) = e^{-(\beta x)^\alpha}, \text{for } x \geq 0.$$

Differentiating the c.c.d.f. gives that the p.d.f. is

$$f(x) = \alpha\beta\,(\beta x)^{\alpha-1}\,e^{-(\beta x)^\alpha}, \text{for } x \geq 0.$$

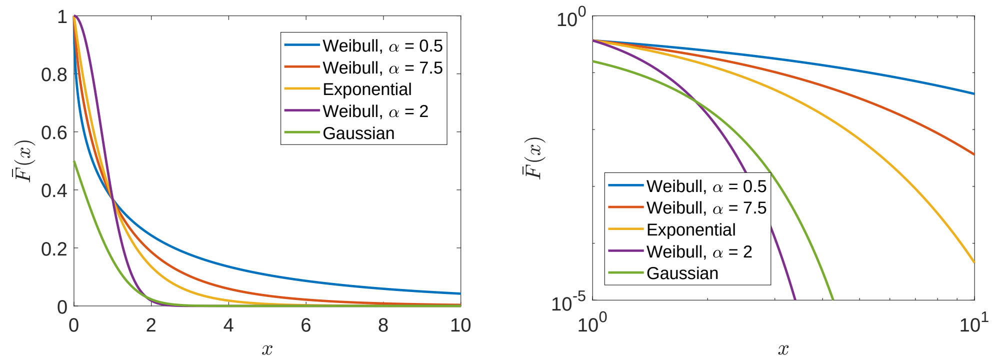
In these, α is referred to as the *shape parameter* of the distribution and β is the *scale parameter*. Note that when $\alpha = 1$ the Weibull is equivalent to an Exponential(β).

In fact, the Weibull distribution is an especially helpful distribution when seeking to contrast heavy tails with light-tails because it can be either heavy-tailed or light-tailed depending on the shape parameter α. If $\alpha < 1$ then the Weibull is heavy-tailed, while if $\alpha \geq 1$ the Weibull is light-tailed. Mathematically, this can be seen using a quick calculation based on the definition of heavy-tailed distributions.

$$\limsup_{x\to\infty} \frac{\bar{F}(x)}{e^{-\mu x}} = \limsup_{x\to\infty} e^{\mu x - (\beta x)^\alpha}.$$

If $\alpha < 1$ this limit equals ∞ for any $\mu > 0$, while if $\alpha < 1$ it is 0 for any $\mu > 0$. (If $\alpha = 1$ the Weibull is equivalent to an Exponential distribution, which if of course light-tailed.)

Figure 1.6(a) illustrates the tail of the Weibull distribution for different values of α, contrasting the c.c.d.f. with that of the Gaussian and Exponential distributions. Like for the Pareto, the heaviness of the tail is clearly visible when we look at the distribution of a log-log plot; see Figure 1.6(b). While the Weibull



(a) c.c.d.f. on linear scale (b) c.c.d.f. on log-log scale

Figure 1.6: Illustration of Weibull c.c.d.f.: *Plots show the c.c.d.f. of the Weibull distribution with scale parameter β = 1 and different values of shape parameter α, alongside the Exponential distribution (with unit mean, which corresponds to $\alpha = \beta = 1$) and the standard Gaussian. Fig. (a) shows the c.c.d.f.s on a linear scale, while Fig. (b) plots them on a log-log scale.*

looks nearly linear on a log-log plot when α is small, i.e., when the tail is heaviest, it is not perfectly linear like the Pareto distribution. To see why, we can take logarithms of both sides of (1.2.2) to obtain

$$\log \bar{F}(x) = -(\beta x)^{\alpha}.$$

While x^{α} gets close to $\log x$ as α shrinks to zero, it never entirely matches. However, if we move the negative to the other side and take logarithms again, we see that the Weibull looks linear according to a different scaling:

$$\underbrace{\log(-\log \bar{F}(x))}_{'y'} = \underbrace{\alpha \log \beta}_{y\text{-intercept}} + \underbrace{\alpha}_{\text{slope}} \underbrace{\log x}_{'x'}.$$

This highlights that the Weibull c.c.d.f. is linear on a $\log(-\log \bar{F}(x))$ vs. $\log x$ plot. Like with the Pareto distribution, this is useful tool for exploratory analysis of data, but one which needs to be used with care and should not be relied on for estimation.

The hazard rate. Not only does varying α change the tail behavior, varying α impacts other important properties of the Weibull as well. One that is of particular interest is the *hazard rate*, a.k.a., failure rate, of the distribution. We study the hazard rate in detail in Chapter 4, and the Weibull is a particularly important distribution for that chapter because its hazard rate can have widely varying properties.

The hazard rate is defined as $q(t) = f(t)/\bar{F}(t)$ and has the following interpretation. Thinking of the distribution as capturing the *time to failure* (lifetime) of a component, $q(t)$ captures the instantaneous likelihood of a failure at time t of a component that entered into use at time 0, given that failure has not occurred until time t. Interestingly, when $\alpha > 1$, the hazard rate of the Weibull is increasing, meaning the likelihood of an impending failure increases with the age of the component; when $\alpha < 1$, the hazard rate is decreasing, meaning the likelihood of an impending failure actually decreases with the age of the

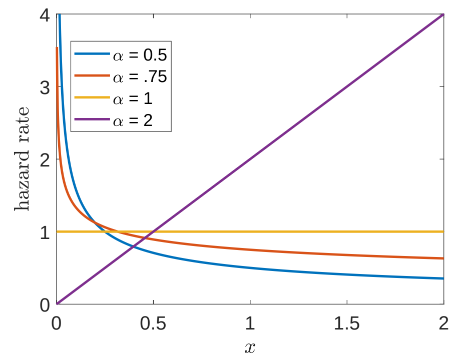


Figure 1.7: *Weibull hazard rate for scale parameter $\beta = 1$ and different values of shape parameter α*

component; and when $\alpha = 1$ the hazard rate is constant. We illustrate this in Figure 1.7.

The properties of the Weibull with respect to its hazard rate make it an extremely important distribution for survival analysis, reliability analysis, and failure analysis in a variety of areas. Additionally, the Weibull plays an important role in weather forecasting, specifically related to wind speed distributions and rainfalls. As we discuss in Chapter 7, the Weibull is an "extreme value distribution", which means that it is deeply connected to extreme events, such as the maximal rainfall in a day or year, the maximal overvoltage in an electrical system, or the maximal size of insurance claims. However, its first usage was in the context of describing the particle size distribution by milling and crushing operations in the 1930s [168]. Interestingly, though the distribution is named after Waloddi Weibull, who studied it in detail in the 1950s, it was introduced much earlier by Fréchet in 1927 in the context of extreme value theory [78].

Moments. An important contrast between the Weibull distribution and the Pareto distribution is that all the moments of the Weibull distribution are finite. They can be large, especially when α is small, but they are not infinite.

To express the moments, we need to use the gamma function, Γ, which is a continuous extension of the factorial function. Specifically, $\Gamma(n) = (n-1)!$, for integer n. More generally,

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx, \text{for } z > 0.$$

Using the gamma function, we can write the mean and variance of the Weibull as

$$\mathbb{E}[X] = \left(\frac{1}{\beta}\right) \Gamma(1 + 1/\alpha),$$
$$\text{Var}[X] = \left(\frac{1}{\beta}\right)^2 \left[\Gamma(1 + 2/\alpha) - (\Gamma(1 + 1/\alpha))^2\right].$$

Notice that the mean grows quickly as $\alpha \to 0$: it grows like the factorial of $1/\alpha$. More generally, the raw moments of the Weibull are given by

$$\mathbb{E}\left[X^n\right] = \left(\frac{1}{\beta}\right)^n \Gamma(1 + n/\alpha).$$

1.2.3 The LogNormal distribution

While both the Pareto and the Weibull can be viewed as transformations of the Exponential distribution, as the name would suggest, the LogNormal distribution is a transformation of the Normal, a.k.a., Gaussian distribution. In fact, the transformation of the Gaussian distribution that leads to the logNormal distribution is the same transformation that creates the Pareto distribution from the Exponential distribution – the LogNormal distribution is an *exponential* of the Gaussian distribution. Specifically,

$$X \sim \text{LogNormal}(\mu, \sigma^2) \quad \Longleftrightarrow \quad \log(X) \sim \text{Gaussian}(\mu, \sigma^2).$$

Or, equivalently,

$$Z \sim \text{Gaussian}(\mu, \sigma^2) \quad \Longleftrightarrow \quad e^Z \sim \text{LogNormal}(\mu, \sigma^2).$$

This means that the LogNormal distribution can be specified in terms of the Gaussian distribution via a logarithmic transformation. For example, the p.d.f. of a LogNormal(μ, σ^2) distribution is

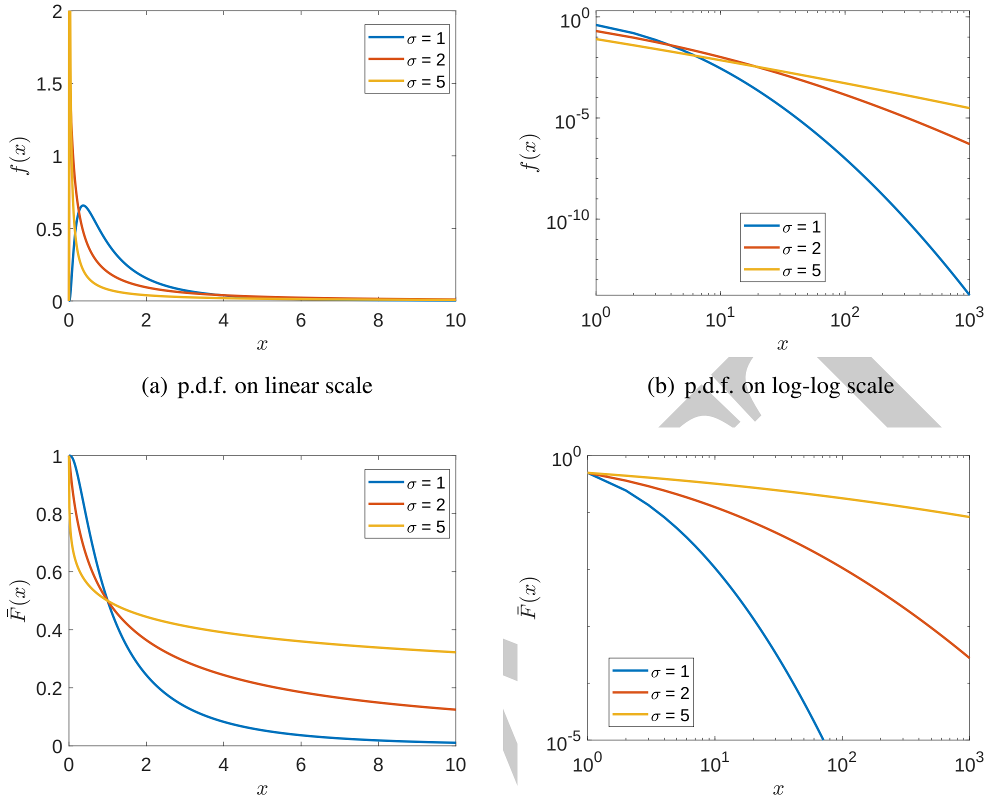
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-(\log x - \mu)^2/(2\sigma^2)}.$$

Note that the change of variables of $\log x$ introduces a $1/x$ term outside of the exponential in the p.d.f. as compared to the Gaussian distribution. This connection with the Gaussian distribution can also be used to show that the LogNormal distribution is heavy-tailed; this is left as an exercise for the reader (see Exercise 3).

While it is not evident from the form of the p.d.f., the LogNormal distribution has a shape that it extremely similar to that of the Pareto distribution. We illustrate the p.d.f. and c.c.d.f. in Figure 1.8. In fact, even when viewed on a log-log plot the LogNormal and the Pareto look similar. Specifically, when the variance parameter σ^2 is large, the LogNormal looks linear on the log-log plot. To see why, let us take logarithms of both sides of (1.2.3):

$$\begin{aligned}
\underbrace{\log f(x)}_{'y'} &= -\log x - \log(\sigma\sqrt{2\pi}) - \frac{(\log x - \mu)^2}{2\sigma^2} \\
&= -\frac{(\log x)^2}{2\sigma^2} + \left(\frac{\mu}{\sigma^2} - 1\right)\underbrace{\log x}_{'x'} - \log(\sigma\sqrt{2\pi}) - \frac{\mu^2}{2\sigma^2}.
\end{aligned}$$

This highlights that, when σ is sufficiently large, the quadratic term above will be small for a large range of x and so the log-log plot will look nearly linear. Consequently, it is nearly impossible to distinguish the LogNormal from a Pareto using the log-log plot. Hence, one should be very careful when using the log-log plot as a statistical tool. We emphasize this point further in Chapter 8.



(a) p.d.f. on linear scale (b) p.d.f. on log-log scale

(c) c.c.d.f. on linear scale (d) c.c.d.f. on log-log scale

Figure 1.8: Illustration of LogNormal distribution: *The c.c.d.f. and p.d.f. of LogNormal distributions with $\mu = 0$ and different values of σ are depicted. The p.d.f.s are plotted on a linear scale in Fig. (a), and on a log-log scale in Fig. (b). The corresponding c.c.d.f.s are plotted on a linear scale in Fig. (c), and on a log-log scale in Fig. (d). Note that the p.d.f. as well as the c.c.d.f. appears nearly linear on a log-log plot when σ is large.*

Properties. The LogNormal maintains many of the useful properties of the Gaussian distribution, they are just adjusted a bit due to the exponential transformation between the distributions.

Perhaps the most important property of the Gaussian is that the sum of independent Gaussians is a Gaussian. Because of the exponential transformation, for the LogNormal, this property holds for the product rather than the sum. In particular, suppose $Y_i \sim \text{LogNormal}(\mu_i, \sigma_i^2)$ are n independent random variables, Then

$$Y = \prod_{i=1}^{n} Y_i \quad \Longrightarrow \quad Y \sim \text{LogNormal}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

This suggests that the LogNormal distribution is intimately tied to the growth of *multiplicative processes*. In particular, if a process grows multiplicatively then it is additive on a logarithmic scale and, by the central limit theorem, it is likely to be Gaussian on the logarithmic scale. This, in turn, means that it is a LogNormal in the original scale. As a consequence, the LogNormal is a very common distribution in nature and human behavior. It has been used to model phenomena in finance, computer networks, hydrology, biology, medicine and more. In fact, the LogNormal distribution was first studied by Robert Gibrat in the context of deriving a multiplicative version of the central limit theorem, which is sometimes termed "Gibrat's Law". Gibrat formulated this law during his study of the dynamics of firm sizes and industry structure [180]. We devote Chapter 6 to a discussion of multiplicative versions of the central limit theorem, and their connections to heavy-tailed distributions.

Beyond products, LogNormal distributions also behave pleasantly with respect to other transformations. An important example is that

$$X \sim \text{LogNormal}(\mu, \sigma^2) \quad \Longrightarrow \quad X^a \sim \text{LogNormal}(a\mu, a^2\sigma^2) \text{ for } a \neq 0. \tag{1.4}$$

Moments. Like the Weibull distribution, the moments of the LogNormal are always finite. They can be quite large, but are never infinite. Perhaps the most counterintuitive thing about the moments of the LogNormal distribution is that, while we adopt the same parameter names as for the Gaussian, μ and σ^2 do not refer to the mean and variance of the LogNormal. Instead, they refer to the mean and variance of the Gaussian that is obtained by taking the $\log$ of the LogNormal. The mean and variance of the LogNormal are as follows:

$$\begin{aligned}\mathbb{E}[X] &= e^{\mu+\sigma^2/2},\\ \text{Var}[X] &= e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right).\end{aligned}$$

The fact that mean and variance are exponentials of the distribution's parameters emphasizes that one should expect them to be large. More generally, the raw moments of the LogNormal distribution are given by

$$\mathbb{E}[X^n] = e^{n\mu+\frac{1}{2}n^2\sigma^2}.$$

1.2.4 Other heavy-tailed distributions

The Pareto, Weibull, and LogNormal are the most commonly used heavy-tailed distributions, but there are many other heavy-tailed distributions that appear frequently. We end this chapter by briefly introducing a few other distributions that come up later in the book as important examples of the concepts we discuss.

The Cauchy distribution. The Cauchy distribution is an important distribution in statistics, and is strongly connected to the central limit theorem, as we discuss in Chapter 5. However, it is most often used as a pathological example as a result of the fact that it does not have a well defined mean (or variance). In fact, though it is named after Cauchy, the first explicit analysis of it was conducted by Poisson in 1824 in order to provide a counterexample showing that the variance condition in the central limit theorem cannot be dropped.

The c.d.f. and p.d.f. of a Cauchy(x_0, γ) distribution are given, for $x \in \mathbb{R}$, by

$$F(x) = \frac{1}{\pi} \arctan\left(\frac{x - x_0}{\gamma}\right) + \frac{1}{2},$$
$$f(x) = \frac{1}{\pi\gamma}\left(\frac{\gamma^2}{(x - x_0)^2 + \gamma^2}\right),$$

with location parameter $x_0 \in \Re$ and scale parameter $\gamma > 0$. The distribution is plotted in Figure 1.9.

While the distribution function looks complicated, the Cauchy has a simple representation as the ratio of two Gaussian random variables. Specifically, if U and V are independent Gaussian random variables with mean 0 and variance 1, then $U/V \sim$ Cauchy$(0, 1)$ (see Exercise 12). A Cauchy$(0, 1)$ is referred to as the *standard Cauchy* and is important its own right because it coincides with the *Student's t-distribution*, which is crucially important for estimating the mean and variance of a Gaussian distribution from data.

The Cauchy distribution's emergence in the context of the central limit theorem is a result of the fact that sums of Cauchy distributions have a similar property as sums of Gaussian distributions: if $X_1, \ldots, X_n$ are i.i.d. Cauchy$(0, 1)$ random variables, then the sum is also a Cauchy. Specifically, $\frac{1}{n}\sum_{i=1}^{n} X_i \sim$ Cauchy$(0, 1)$; we prove this property in Chapter 5 using characteristic functions.

Finally, a related distribution to the Cauchy is the LogCauchy, which has the same relationship to the Cauchy distribution that the LogNormal has to the Gaussian distribution, i.e.,

$$X \sim \text{Cauchy}(x_0, \gamma) \quad \Longleftrightarrow \quad e^X \sim \text{LogCauchy}(x_0, \gamma),$$

or, equivalently,

$$Y \sim \text{LogCauchy}(x_0, \gamma) \quad \Longleftrightarrow \quad \log(Y) \sim \text{Cauchy}(x_0, \gamma).$$

The LogCauchy is important because it is one of the few common distributions that has a heavier tail that the Pareto distribution—it has a logarithmically decaying tail. For this reason, it is sometimes referred to as a *super-heavy-tailed distribution*.

The Fréchet distribution. The Fréchet distribution is an important distribution for extreme value theory, as we discuss in Chapter 7. It is commonly used in hydrology when studying the extremes of rainfall distributions.

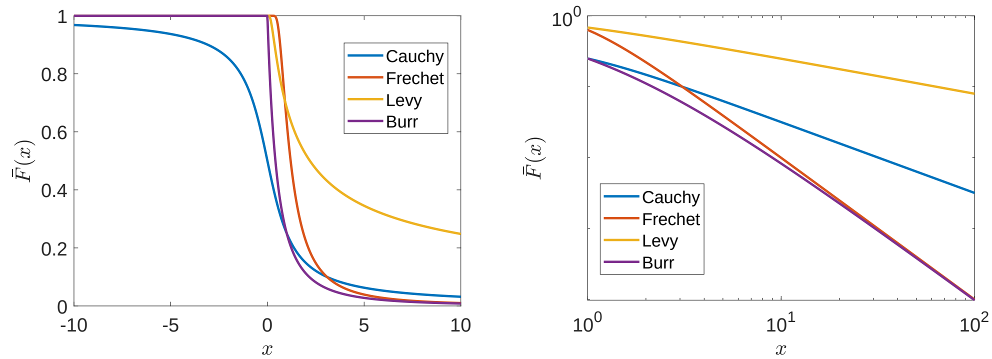
The distribution is named after Maurice Fréchet, who introduced it in 1927; however it is also referred to as the inverse Weibull distribution, which is a much more descriptive name since it is defined as exactly that. Specifically,

$$X \sim \text{Weibull}(\alpha, \beta) \quad \Longleftrightarrow \quad 1/X \sim \text{Fréchet}(\alpha, \beta, 0).$$

More generally, the c.d.f. and p.d.f. of the Fréchet(α, β, x_m) distribution are given, for $x > x_m$, by

$$F(x) = e^{-(\beta(x - x_m))^{-\alpha}},$$
$$f(x) = \alpha\beta\left(\beta(x - x_m)\right)^{-1-\alpha} e^{-(\beta(x - x_m))^{-\alpha}}.$$

Here, $\alpha > 0$ is the shape parameter, $\beta > 0$ is the scale parameter, and x_m is the minimum value taken by the distribution. The distribution is plotted in Figure 1.9.



(a) c.c.d.f.s (b) c.c.d.f.s on log-log scale

Figure 1.9: *The plots show the c.c.d.f. of the standard Cauchy ($x_0 = 0$, $\gamma = 1$), the Fréchet (with $x_m = 0$, $\beta = 1$, $\alpha = 2$), the Lévy (with $\mu = 0$, $c = 1$), and the Burr distribution (with $c = \lambda = 1$, $k = 2$). Fig. (a) shows the plots on a linear scale, and Fig. (b) on a log-log scale. Note that all c.c.d.f.s look (asymptotically) linear on a log-log scale; we will formalize this property in Chapter 2.*

The Lévy distribution. The Lévy distribution is used most prominently in the study of financial models to explain stylized phenomena such as volatility clustering. Within mathematics and physics, it also plays an important role in the study of Brownian motion: the hitting time of a single point at a fixed distance from the starting point of a Brownian motion has a Lévy distribution. But, perhaps its most prominent use is in the context of the generalized central limit theorem, which we discuss in Chapter 5.

Like the Cauchy distribution and the LogNormal distribution, the Lévy distribution is most conveniently defined as a transformation of the Gaussian distribution. In particular, a Lévy distribution coincides with the square of the inverse of a Gaussian distribution (and is therefore sometimes also called the inverse Gaussian):

$$Z \sim \text{Gaussian}(\mu, \sigma^2) \quad \Longrightarrow \quad \frac{1}{(Z-\mu)^2} \sim \text{Lévy}(0, 1/\sigma^2).$$

More directly, the p.d.f. of the a Lévy(μ, c) distribution is given, for $x \in \mathbb{R}$, by

$$f(x) = \sqrt{\frac{c}{2\pi}} \frac{e^{\frac{-c}{2(x-\mu)}}}{(x-\mu)^{3/2}}. \tag{1.5}$$

From this equation, it is straightforward to see that the Lévy distribution is not just heavy-tailed, it is more specifically a power law distribution, since $f(x) \sim \sqrt{\frac{c}{2\pi}} x^{-3/2}$. A plot of the distribution is shown in Figure 1.9.

The Burr distribution. The Burr distribution is a generalization of the Pareto distribution that often appears in statistics and econometrics. It is most frequently used in the study of household incomes and related wealth distributions. It was introduced by Irving Burr in 1942 as one of a family of 12 distributions, of which it is the *Burr Type XII distribution*. The c.c.d.f. and p.d.f. of a Burr(c, k, λ) distribution are given, for $x > 0$, by

$$\bar{F}(x) = (1 + \lambda x^c)^{-k} \tag{1.6}$$

$$f(x) = \frac{ckx^{c-1}}{(1 + x^c)^{k+1}}, \tag{1.7}$$

where $c, k, \lambda > 0$. The distribution is illustrated in Figure 1.9. When $c = 1$ the Burr corresponds to a so-called "Type II" Pareto distribution, and it is easy to see that it is a power law distribution, like the Cauchy, Fréchet, and Lévy distributions. Interestingly, in Chapter 7 we discuss a connection between the Burr distribution and the residual life of the Pareto distribution. The hazard rate of the Burr distribution itself serves as an important counterexample in the same chapter.

The Zipf distribution. We conclude by mentioning the Zipf distribution, which is a discrete version of the Pareto distribution. The Zipf distribution rose to prominence because of “Zipf’s law”, which states that, given a natural language corpus, the frequency of any word is inversely proportional to its rank in the frequency table of the corpus. That is, the most common word occurs twices as often as the second most common word, three times more common than the third most common word, and so on. This law is named after George Zipf, who popularized it in 1935; however the observation of the phenomenon predated his work by more than fifty years.

The Zipf(s, N) distribution is one example of a distribution that would explain Zipf’s law, and is defined in terms of its probability mass function (p.m.f.):

$$p(n; s, N) = \frac{1/n^s}{\sum_{i=1}^{N} 1/i^s}, \tag{1.8}$$

where N can be thought of as the number of elements in the corpus and s is the exponent characterizing the power law.

While the Zipf distribution is not heavy-tailed, given that it has a finite support, its generalization to the case $N = \infty$, which is called the Zeta distribution, is heavy-tailed (for $s > 1$).

1.3 How to use this book

In this chapter we have introduced the definition of the class of heavy-tailed distributions, along with a few examples of common heavy-tailed distributions. Through these examples, you have already seen some illustrations of how the behavior of heavy-tailed distributions contrasts with light-tailed distributions, but we have not yet sought to build intuition for these differences or to explain why heavy-tailed distributions are so common in the world around us. We have mentioned that controversy often surrounds heavy-tailed distributions because intuitive statistical approaches for identifying heavy tails in data are flawed, but we have not yet provided tools for correct identification and estimation of heavy-tailed phenomena.

The remainder of this book is organized to first provide intuition, both qualitative and mathematical, for the defining properties of heavy-tailed distributions (Part I: Properties), then explain why heavy-tailed distributions are so common in the world around us (Part II: Emergence), and finally develop the statistical tools for the estimation of heavy-tailed distributions (Part III: Estimation).

Given the mystique and excitement that surrounds the discovery of heavy-tailed phenomena, the detection and estimation of heavy tails in data is a task that is often (over)zealously pursued. While reading this book, you may be tempted to skip directly to Part III on estimation. However, the book is written so that the tools used in Part II are developed in Part I, and the tools used in Part III are developed in Parts I and II. Thus, we encourage readers to work through the book in order. That said, we have organized the material in each chapter so that there is a main body that presents the core ideas that are important for later chapters, followed by sections that present examples and/or variations of the main topic. These later sections can be viewed as enrichment opportunities that can be skipped as desired if the goal is to move quickly to Part III. Additionally, if one is looking for the quickest path to understand the background needed before digging into Part III, then we recommend focusing on Chapter 2 from Part I, then Chapters 5 and 7 from Part II before moving to Part III.

Finally, we would like to emphasize that the aim of this book is to present the fundamentals of heavy-tailed distributions in a way that is accessible to readers who have an introductory undergraduate background in probability. Given that the theory of heavy tails uses advanced mathematical tools and is typically presented in a way that is accessible only to upper-level graduate students in mathematics, our target has required us to rethink and reprove many classical results in the area with the goal of providing a simple, intuitive presentation. This often means presenting theorems that are more restrictive in terms of their assumptions than the most general results known and presenting proofs that convey the key ideas but gloss over some of the more technical details via either an assumption or a reference to a technical lemma in another source. To provide interested readers with references to the full generality of the results we discuss, each chapter ends with an "Additional Notes" section that includes references to more detail on the topics presented in the chapter. We encourage readers to follow up on the references in these sections. Additionally, many of the exercises at the end of the chapters ask the reader to derive extensions or fill in technical details that we have left out of the main body of the chapters, so we encourage readers to work through the exercises.

Our goal for this book is that, through reading it, heavy-tailed distributions will be demystified for you. That their properties will be intuitive, not mysterious. That their emergence will be expected, not surprising. And, that you will have the proper statistical tools for studying heavy-tailed phenomena and so will you will help resolve (or avoid) controversies rather than feed them. Happy reading!

1.4 Exercises

1. For a standard Gaussian random variable N, show that for $x > 0$,

$$\Pr(N > x) \leq \frac{e^{-x^2/x}}{\sqrt{2\pi x}}.$$

   Note: In fact, the above bound can be shown to be asymptotically tight, i.e., it can be shown that Pr $(N > x) \sim \frac{e^{-x^2/x}}{\sqrt{2\pi x}}$*; see [69, Chapter 7].*

2. Use the bound of Exercise 1 to prove that the Gaussian(μ, σ^2) distribution is light-tailed.

3. Prove that the LogNormal(μ, σ^2) distribution is heavy-tailed.

4. Consider a distribution F over $\mathbb{R}_+$ with finite mean μ. The *excess* distribution corresponding to F is defined as

$$\bar{F}_e(x) = \frac{1}{\mu}\int_x^\infty \bar{F}(y)dy.$$

Prove that F is heavy-tailed if and only if F_e is heavy-tailed.

5. Let $X \sim$ Exponential(μ) and $Y = 1/X$. Prove that Y is heavy-tailed.

6. The random variable N takes values in $\mathbb{N}$. The distribution of N, conditioned on a uniformly distributed random variable U taking values in $(0, 1)$, is given by Pr $(N > n \mid U) = U^n$. Assuming U is uniformly distributed, show that N is heavy-tailed.

Note: Even though the conditional distribution of N given the value of U is light-tailed (in fact, Geometrically distributed), N itself is heavy-tailed!

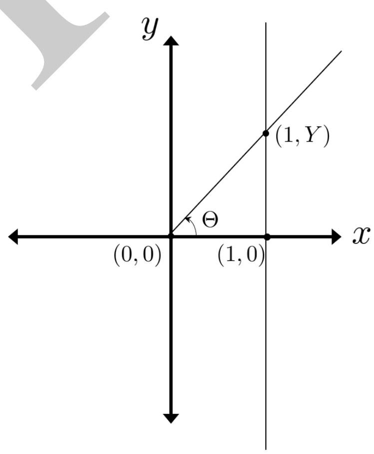
7. In the above problem, you do not need U to be uniformly distributed for N to be heavy-tailed. Prove that, so long as Pr $(U > x) > 0$ for all $x \in (0, 1)$, N is heavy-tailed.

8. Derive an expression for the Gini coefficient corresponding to the Pareto distribution. Show that the Gini coefficient converges to 1 as tail index $\alpha \downarrow 1$.

9. Compute the Lorenz curve corresponding to the Exponential distribution. Prove that the Gini coefficient in this case equals 1/2.

10. Prove the property (1.4) of the LogNormal distribution.

11. The goal of this exercise is to prove the following geometric interpretation of the standard Cauchy distribution: On the Cartesian plane, draw a random line passing through the origin making an angle Θ with the x-axis as shown in the following figure, where Θ is uniformly distributed over $(-\pi/2, \pi/2)$. Let $(1, Y)$ denote the point where this random line intersects the vertical line $x = 1$. Prove that Y is a standard Cauchy random variable.



12. Prove that if U and V are independent, standard Gaussian random variables, then U/V is a standard Cauchy.

Hint: The geometric interpretation from Exercise 11 might help. Interpreting (U, V) as the Cartesian coordinates of a random point on the Cartesian plane, what is the joint distribution of the polar coordinates (R, Θ)?

13. The goal of this exercise is to compare the 'heaviness' of the tails of the Pareto, Weibull, and Log-Normal distributions. Let $X \sim \text{Pareto}(x_m, \alpha_1)$, $Y \sim \text{LogNormal}(\mu, \sigma^2)$, and $Z \sim \text{Weibull}(\alpha_2, \beta_2)$. Prove that

$$\lim_{x\to\infty} \frac{\Pr(Z > x)}{\Pr(Y > x)} = 0, \quad \lim_{x\to\infty} \frac{\Pr(Y > x)}{\Pr(X > x)} = 0.$$

Note: This exercise shows that the Pareto has a heavier tail than the LogNormal, which has a heavier tail than the Weibull.

Part I

Properties

The mystique that surrounds heavy-tailed distributions follows, in part, from the ambiguous nature of the term “heavy-tailed”. As we have seen, the precise mathematical definition is too opaque to provide insight and, often, too broad to correspond to the properties that are of interest in specific applications. Further, terms such as power-law, scale-free, self-similar, fat-tailed, long-tailed, stable, subexponential, etc. are often used synonymously with “heavy-tailed” when in fact they refer to particular properties of *some* heavy-tailed distributions. The result is a confusing and, at times, conflicting zoo of informal and formal terminology.

In Part I of this book we provide insight into some of most mysterious and elegant properties of heavy-tailed distributions, connecting these properties to formal definitions of subclasses of heavy-tailed distributions. We focus on three illustrative properties: “scale-invariance” (a.k.a., scale-free), the “catastrophe principle”, and “increasing residual life.” We illustrate that these properties provide qualitatively different behaviors than what is seen under light-tailed distributions, and we describe how to formalize these properties mathematically as subclasses of heavy-tailed distributions. Combined, the following chapters help to demystify some of the particularly exotic properties of heavy-tailed distributions, and to provide a clear view of how these properties interact with each other and with the broader class of heavy-tailed distributions.

Specifically, we introduce the classes of regularly varying distributions, subexponential distributions, and long-tailed distributions, which are perhaps the three most prominent classes of heavy-tailed distributions. These formalizations are particularly noteworthy because, while the general class of heavy-tailed distribution is difficult to work with, each of these subclasses has properties that make them appealing to work with analytically. These three classes also form the building blocks that allow us to study the emergence and estimation of heavy-tailed distributions in Parts II and III of this book.

Chapter 2

Scale invariance, power laws, and regular variation

In our daily lives, many things that we come across have a size, or scale, that we associate with them. For example, people's heights and weights differ, but they are not *that* different – they rarely differ by more than a factor of two or three and do not differ much from the population average. In contrast, the incomes of people we encounter in our daily lives do not have a typical size or scale – they may differ by a factor of *100 or more* and can be very far from the population average! This contrast is a consequence of the fact that many heavy-tailed phenomena, such as incomes, are *scale invariant*, a.k.a., *scale-free*, while light-tailed phenomena, such as heights and weights, are not.

Scale invariance is a property that seems particularly magical the first time you observe it. An object is scale invariant if it looks the same regardless of what scale you look at it. Perhaps the easiest way to understand scale invariance is using fractals, like the one shown in Figure 2.1. If you zoom in or out, the picture will look the same. It turns out that Pareto distributions have the same property (see Figure 2.2). But, Pareto distributions are even more special than fractals. With fractals, you have to zoom in or out a specific, discrete amount for the picture to look the same, with Pareto distributions, no matter how much you change the scale, the picture does not change.

Scale invariance is a particularly mysterious aspect of heavy-tailed phenomena. It is natural to think of the average of a distribution as a good predictor of what samples will occur, but for scale invariant distributions the average is actually a very poor predictor. This fact leads to many of the counter-intuitive properties of heavy-tailed distributions. For example, consider the old economics joke: "If Bill Gates walks into a bar... on average, everybody in the bar is a millionaire."

Though initially mysterious and counter-intuitive, scale invariance is a beautiful and widely observed phenomenon that has received attention broadly beyond mathematics and statistics, e.g., in physics, computer science, economics, and even art. For example, scale invariance is an important phenomena in both classical and quantum field theory as well as statistical mechanics. In fact, it is closely tied to the notion of "universality" in physics, which relates to the fact that widely different systems can be described by the same underlying theory. Further, in the context of network science, scale invariance has received considerable attention. Widely varying networks have been found to have scale invariant degree distributions (and are thus termed "scale-free networks" [23, 38]) and this observation has had dramatic impacts for our understanding of the structural properties of networks. For a discussion of scale invariance broadly, we recommend

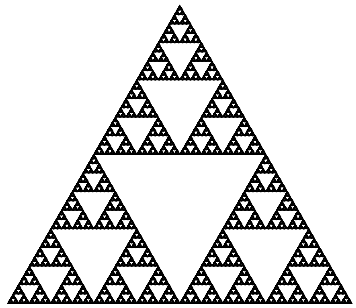


Figure 2.1: *Illustration of the Sierpinski fractal [135].*

referring to [193]. Here, we focus on scale invariance in the context of probability and statistics.

In particular, in this chapter we explore the mathematics of the property of scale invariance and its connections with Pareto distributions and so-called *power law distributions*. Note that both "scale invariance" and "power law" are often used synonymously with "heavy-tailed", and thus it is important to point out that not all heavy-tailed distributions are scale invariant or have a power law (though all scale invariant distributions are heavy-tailed, as are all power law distributions). The goal of this chapter is to describe how to formalize and generalize the notions of scale invariance and power laws as a subclass of heavy-tailed distributions termed "regularly varying distributions" that is particularly appealing from a mathematical perspective. The properties of this class shed light on many of the counter-intuitive properties of heavy-tailed distributions; highlighting what properties of heavy-tailed distributions can be viewed as simple consequences of scale invariance. Further, in order to illustrate the usefulness of the class, we highlight how to apply properties of regular variation in order analyze heavy-tailed phenomenon more broadly. These examples illustrate that it is not much more challenging to analyze the entire class of regularly varying distributions than it is to work with the specific case of the Pareto distribution.

2.1 Scale invariance and power laws

To this point we have only introduced scale invariance informally as the property that something looks the "same" regardless of the scale it is observed. Given that our focus is on probability distributions, we can rephrase this idea as follows: if the scale (or units) with which the samples from the distribution are measured is changed, then the shape of the distribution is unchanged. This leads to the following formal definition.

Definition 2.1. *A distribution function F is scale invariant if there exists an $x_0 > 0$ and a continuous positive function g such that*

$$\bar{F}(\lambda x) = g(\lambda)\bar{F}(x),$$

for all x, λ satisfying $x, \lambda x \geq x_0$.

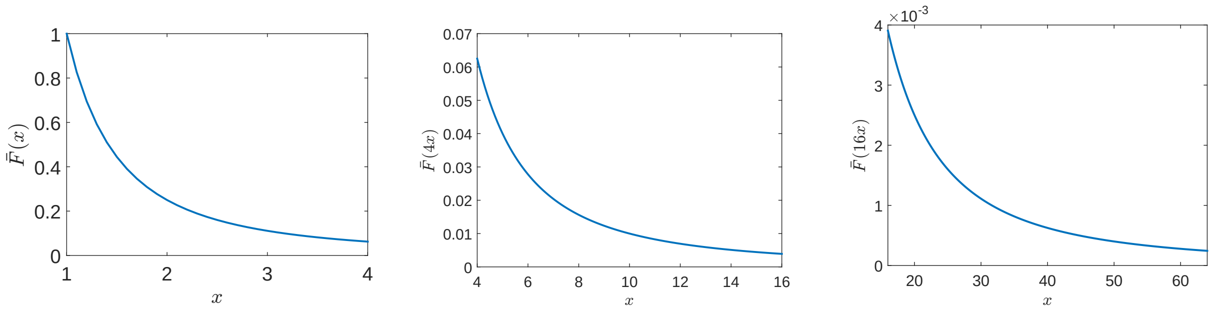


Figure 2.2: *The c.c.d.f. corresponding to the Pareto distribution ($\alpha = 2$, $x_m = 1$) plotted at different scales of the independent variable. Note that the shape of the curve is preserved upto a multiplicative scaling, consistent with scale invariance.*

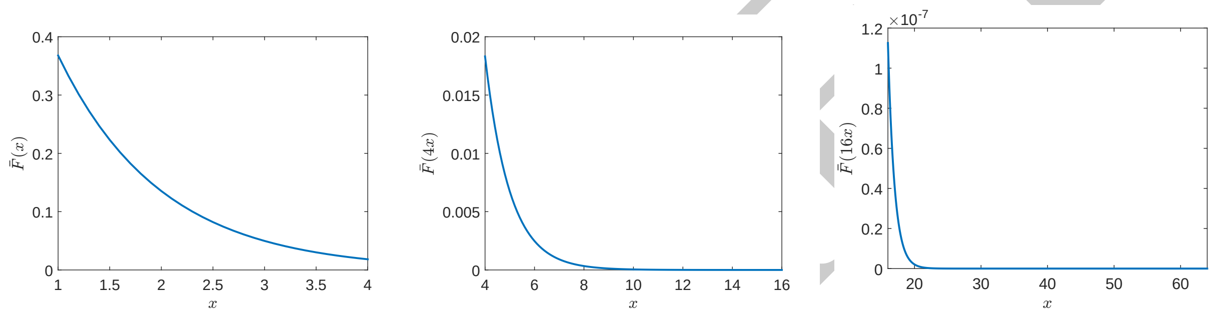


Figure 2.3: *The c.c.d.f. corresponding to the Exponential distribution (with mean 1) plotted at different scales of the independent variable. Note that the shape of the curve looks fundamentally altered at different scales.*

To interpret the definition of scale invariant, one can think of λ as the "change of scale" for the units being used. With this interpretation, the definition says that the shape of the distribution $\bar{F}$ remains unchanged, up to a multiplicative factor $g(\lambda)$ if the measurements are scaled by λ. This is exactly what is shown in Figure 2.2 for the Pareto distribution.

More formally, to see that the Pareto is scale invariant, recall that a Pareto distribution has $\bar{F}(x) = (x/x_m)^{-\alpha}$ for $x > x_m$. Thus,

$$\bar{F}(\lambda x) = \left(\frac{\lambda x}{x_m}\right)^{-\alpha} = \bar{F}(x)\lambda^{-\alpha}, \text{ whenever } x, \lambda x > x_m.$$

Scale invariance is a very elegant property, but it is also a fragile one. In particular, it does not hold for most probability distributions. For example, it is easy to see that the Exponential distribution is not scale invariant. Recall that an Exponential distribution has $\bar{F}(x) = e^{-\mu x}$ for $x \geq 0$. Therefore,

$$\bar{F}(\lambda x) = e^{-\mu\lambda x} = \bar{F}(x)e^{-\mu(\lambda-1)x}.$$

Thus, there is not a choice for g that is independent of x. This is also illustrated in Figure 2.3.

One may initially think that the lack of scale invariance of the Exponential distribution is a consequence of the fact that it is a light-tailed distribution. But, that is not the case. For example, let us generalize the Exponential distribution to the the Weibull distribution, $\bar{F}(x) = e^{-\beta x^\alpha}$ for $x \geq 0$, which is equivalent to the Exponential distribution when the shape parameter $\alpha = 1$ and is heavy-tailed when $\alpha < 1$. As the following calculation shows, the Weibull is also not scale invariant:

$$\bar{F}(\lambda x) = e^{-\beta(\lambda x)^\alpha} = \bar{F}(x) e^{-\beta x^\alpha(\lambda^\alpha - 1)}.$$

The examples above start to give some intuition about scale invariance, but they leave open a fundamental, natural question:

Which distributions are scale invariant?

From the examples above, we know that there is at least one scale invariant distribution (the Pareto distribution), but we also know that not all common distributions are scale invariant – not even all common heavy-tailed distributions. Perhaps surprisingly, it turns out that scale invariance is an extremely special property: distributions with “power law tails”, i.e., tails that match the Pareto distribution up to a multiplicative constant, are the *only* scale invariant distributions. That is, “scale invariance” can be thought of interchangeably with “power law”. The following theorem states this formally.

Theorem 2.1. *A distribution function F is scale invariant if and only if F has a power law tail, i.e., there exists $x_0 > 0$, $c \geq 0$, and $\alpha > 0$ such that $\bar{F}(x) = cx^{-\alpha}$ for $x \geq x_0$.*

Proof. Note that the case where $\bar{F}$ is identically zero over $[x_0, \infty)$ trivially satisfies the conditions of the lemma (this corresponds to the case $c = 0$.)

Excluding the above trivial case from consideration, it is easy to see that $\bar{F}(x)$ must be non-zero for all $x \geq x_0$. Indeed, if $\bar{F}(x') = 0$ for some $x' \geq x_0$, then for any $x \geq x_0$, $\bar{F}(x) = \bar{F}(x')g(x/x') = 0$.

Fix $x, y > 0$. We may then pick z large enough such that $z, zx, zxy \geq x_0$. From the scale invariant property of $\bar{F}$, $\bar{F}(xyz) = \bar{F}(z)g(xy)$. Of course, we may also write $\bar{F}(xyz) = \bar{F}(xz)g(y) = \bar{F}(z)g(x)g(y)$. Since $\bar{F}(z) \neq 0$, we can immediately see that the function g satisfies the following property

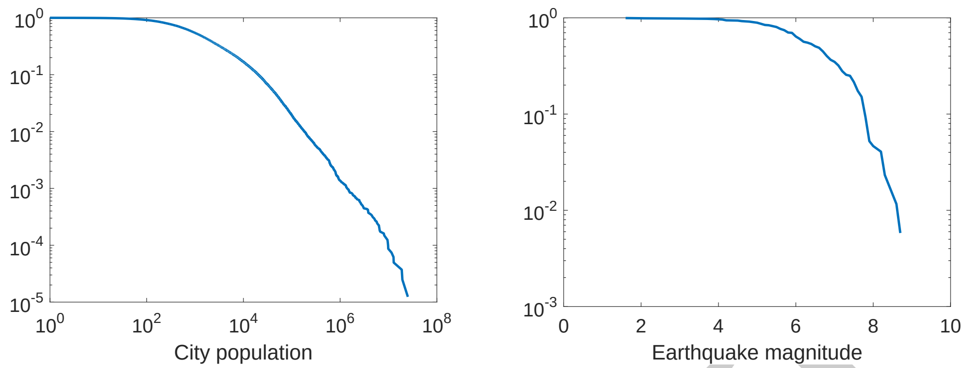
$$g(xy) = g(x)g(y) \quad \text{for all } x, y > 0. \tag{2.1}$$

This is a very special property, and the only continuous positive functions satisfying the above condition are $g(x) = x^{-\alpha}$ for some $\alpha \in \mathbb{R}$.[1] Noting that $\bar{F}(x) = \bar{F}(x_0)g(x/x_0)$ for all $x \geq x_0$, we conclude that $\alpha > 0$ (since $\bar{F}$ must be monotonically decreasing, with $\lim_{x\to\infty} \bar{F}(x) = 0$). Therefore, $\bar{F}(x) = cx^{-\alpha}$ for $x \geq x_0$ for some $c, \alpha > 0$. □

2.2 Approximate scale invariance and regular variation

We have just seen that all scale invariant distributions are power law distributions, a.k.a., distributions with tails matching a Pareto distribution up to a multiplicative constant. This makes scale invariance a very

[1] Defining $f(x) := \log g(e^x)$, the condition (2.1) is equivalent to $f(x + y) = f(x) + f(y)$ for $x, y \in \mathbb{R}$. This is known as *Cauchy's functional equation*. The stated claim now follows from the fact that the only continuous solutions of Cauchy's functional equation are of the form $f(x) = \alpha x$ for $\alpha \in \mathbb{R}$ (see, for example, [4]).



(a) Populations of U.S. cities as per the 2010 census (data sourced from [2])

(b) Intensities of earthquakes in the U.S. between 1900 and 2017 (sourced from [1]). Earthquake intensity is measured on the Richter scale, which is inherently logarithmic. Thus, the values of the x-axis of should be interpreted as being proportional to the logarithm of the intensity of the earthquake.

Figure 2.4: *Empirical c.c.d.f. (a.k.a. rank plot) corresponding to two real world datasets on a log-log scale.*

special property, or a very fragile property depending on how you look at it. In fact, one interpretation of the theorem in the last section is that we should not expect to see scale invariance in reality since it is so fragile.

In the strictest sense, that interpretation is correct. It is quite unusual for the distribution of an observed phenomenon to *exactly* match a power law distribution, and thus be scale invariant. Instead, what tends to be observed in reality is that the *body* of the distribution is not an exact power law, but the *tail* of the distribution is *approximately* power law. Consider the examples in Figure 2.4, which depicts the empirical c.c.d.f. corresponding to two real-world datasets on a log-log scale. Notice that the body of the empirical c.c.d.f. (on the log-log scale) does not look linear, which it would if the data were sampled from a power law distribution, but instead seems to approach a straight line *asymptotically*, which suggests that the c.c.d.f. behaves *asymptotically* like a power law.

Given that we should not expect to see precise scale invariance in real observations, it is natural to move our focus from precise scale invariance to notions of approximate or asymptotic scale invariance; and it is natural not to focus on the whole distribution, but rather to focus on just the tail of the distribution. In particular, the relevant formalism becomes *asymptotic scale invariance*, which we define as follows.

Definition 2.2. *A distribution F is asymptotically scale invariant if there exists a continuous positive function g such that for any $\lambda > 0$,*

$$\lim_{x\to\infty} \frac{\bar{F}(\lambda x)}{\bar{F}(x)} = g(\lambda).$$

The notion of asymptotic scale invariance almost exactly parallels the notion of scale invariance, except that it only requires the property to hold in the limit as $x \to \infty$, i.e., it only requires the property to approximately hold for the tail, i.e., $\bar{F}(\lambda x) \sim g(\lambda)\bar{F}(x)$ as $x \to \infty$.[2]

[2]Throughout the book we use $a(x) \sim b(x)$ as $x \to \infty$ to mean $\lim_{x\to\infty} a(x)/b(x) = 1$.

As a result, it is immediate to see that Pareto distributions are asymptotically scale invariant:

$$\bar{F}(\lambda x)/\bar{F}(x) = \lambda^{-\alpha}.$$

Similarly, it is easy to see that asymptotic scale invariance is still quite a special property that is not satisfied by most distributions. For example the Weibull and Exponential (Weibull with $\alpha = 1$) distributions are not asymptotically scale invariant since, as $x \to \infty$,

$$\frac{\bar{F}(\lambda x)}{\bar{F}(x)} = e^{-\beta(\lambda^\alpha - 1)x^\alpha} \to \begin{cases} \infty, & \lambda < 1, \\ 1, & \lambda = 1, \\ 0, & \lambda > 1. \end{cases}$$

However, asymptotic scale invariance is significantly broader than scale invariance, and it is easy to see that other distributions besides power law distributions are asymptotically scale invariant. For example, it follows from Exercise 8 that the convolution of a Pareto and an Exponential distribution is asymptotically scale invariant (though it is clearly not scale invariant). Similarly, consider the Fréchet distribution, which we encounter in our analysis of extremal processes in Chapter 7. This distribution, which is supported over the non-negative reals, is defined by the distribution function $F(x) = e^{-x^{-\alpha}}$ for $x \geq 0$, where the parameter $\alpha > 0$. While this distribution is clearly not a power law, it is not hard to see that $\bar{F}(x) \sim x^{-\alpha}$ (see Exercise 1). In other words, the Fréchet distribution has an asymptotically power law tail, which in turn implies asymptotic scale invariance:

$$\lim_{x \to \infty} \frac{\bar{F}(\lambda x)}{\bar{F}(x)} = \lambda^{-\alpha}.$$

In general, since asymptotic scale invariance only focuses on the tail of the distribution, the body of such a distribution may behave in an arbitrary manner as long as the tail is approximately scale invariant.

As in the case of scale invariance, given the examples above, the natural question becomes:

Which distributions are asymptotically scale invariant?

It is clear that the class of asymptotically scale invariant distributions includes a variety of heavy-tailed distributions beyond the Pareto distribution, but it is also clear that it does not include all heavy-tailed distributions since it does not include the Weibull distribution. However, the fact that “scale invariant” can be thought of equivalently to “power law”, leads to the suggestion that “asymptotically scale invariant” should correspond to some notion of “approximately power law”, and this turns out to be true. In particular, it turns out that asymptotically scale invariant distributions have tails that are approximately power law in a rigorous sense that can be formalized via the class of *regularly varying distributions*.

Definition 2.3. *A function $f : \mathbb{R}_+ \to \mathbb{R}_+$ is said to be regularly varying of index $\rho \in \mathbb{R}$ if for all $y > 0$,*

$$\lim_{x \to \infty} \frac{f(xy)}{f(x)} = y^\rho.$$

Further, for $\rho \leq 0$, a distribution F is regularly varying of index ρ, denoted as $F \in \mathcal{RV}(\rho)$, if $\bar{F}(x) = 1 - F(x)$ is a regularly varying function of index ρ.[3]

[3]It is common practice to write the domain of regularly varying functions as $\mathbb{R}_+$. That said, it is important to understand that

The form of the definition immediately highlights that regularly varying distributions are asymptotically scale invariant. Further, since $\lim_{x\to\infty} \bar{F}(xy)/\bar{F}(x) = y^\rho$, they seem to mimic the behavior of power law distributions, such as Pareto distribution, asymptotically. This intuitively suggests that, all asymptotically scale invariant distributions are regularly varying distributions – which turns out to be true.

Theorem 2.2. *A distribution F is asymptotically scale invariant if and only if it is regularly varying.*

Proof. It is immediate that regularly varying distributions are asymptotically scale invariant, and so we need only prove the other direction. Fix $x, y > 0$. The asymptotic scale-free property implies that

$$\lim_{z\to\infty} \frac{\bar{F}(xyz)}{\bar{F}(z)} = g(xy).$$

We can also compute the same limit by writing $\frac{\bar{F}(xyz)}{\bar{F}(z)} = \frac{\bar{F}(xyz)}{\bar{F}(xz)}\frac{\bar{F}(xz)}{\bar{F}(z)}$. Note that $\frac{\bar{F}(xyz)}{\bar{F}(xz)} \to g(y)$ and $\frac{\bar{F}(xz)}{\bar{F}(z)} \to g(x)$ as $z \to \infty$, implying that

$$\lim_{z\to\infty} \frac{\bar{F}(xyz)}{\bar{F}(z)} = g(x)g(y).$$

From this, we can conclude that the function g satisfies

$$g(xy) = g(x)g(y) \quad \text{for all } x, y > 0.$$

This is the same relationship we used in the proof of Theorem 2.1 and, like in that case, it follows that there exists $\theta \in \mathbb{R}$ such that $g(x) = x^\theta$. Of course, by definition, this means that $\bar{F}$ is a regularly varying function, and F is a regularly varying distribution. $\square$

The above result highlights that the class of regularly varying distributions characterizes precisely those distributions that are asymptotically scale invariant, and from it we can immediately see that a number of common heavy-tailed distributions are scale invariant. In particular, with a little effort, it is possible to verify that the Student's t-distribution, the Cauchy distribution, the Burr distribution, the Lévy, and also the Zipf distribution distribution are all regularly varying and thus, asymptotically scale invariant (see Exercise 1). This is summarized in Figure 2.5. Note that have not yet proven that regularly varying distributions are heavy-tailed; this follows from the analytic properties of regularly varying functions discussed next (see Lemma 2.2).

2.3 Analytic properties of regularly varying functions

The fact that regularly varying distributions are exactly those distributions that are asymptotically scale invariant suggests that they should be able to be analyzed (at least asymptotically) as if they are simply

regular variation is an *asymptotic* property of a function as its argument tends to ∞. Thus, for a function f to be regularly varying, we only require that its domain include $[x_0, \infty)$ for some $x_0 > 0$. Similarly, for a distribution to be regularly varying, we only require that its support include $[x_0, \infty)$ for some $x_0 > 0$. Specifically, regularly varying distributions need not be supported on the non-negative reals. For example, the Cauchy distribution is regularly varying (see Exercise 1). Finally, why the index ρ must be non-positive for regularly varying distributions will become apparent soon (see Lemma 2.1).

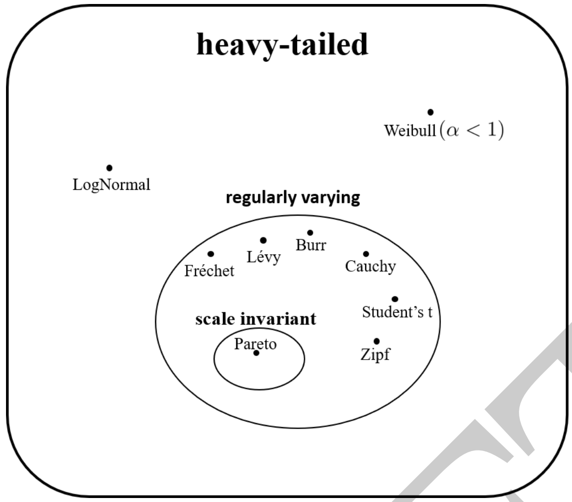


Figure 2.5: *Scale invariant and regularly varying distributions.*

Pareto distributions. In fact, this intuition is correct and can be formalized explicitly, as we show in this section. Concretely, the properties we outline in this section provide the tools to enable regularly varying distributions to be analyzed nearly "as if" they are polynomials, as far as the tail is concerned. This makes them remarkably easy to work with and highlights that the added generality that comes from working with the class of regularly varying distributions, as opposed to working specifically with Pareto distributions, comes without too much added technical complexity.

To begin, it is important to formalize exactly what we mean when we say that regularly varying distributions have tails that are approximately power law. To do this, we need to first introduce the concept of a *slowly varying function*.

Definition 2.4. *A function $L : \mathbb{R}_+ \to \mathbb{R}_+$ is said to be slowly varying if $\lim_{x\to\infty} \frac{L(xy)}{L(x)} = 1$ for all $y > 0$.*

Slowly varying functions are simply regularly varying functions of index zero. So, intuitively, they can be thought of as functions that grow/decay asymptotically slower than any polynomial. For example $\log x$, $\log \log x$, etc. This can be formalized as follows.

Lemma 2.1. *If the function $L : \mathbb{R}_+ \to \mathbb{R}_+$ is slowly varying, then*

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

We prove this lemma later in this section using properties of regularly varying functions. But, we state it now in order to highlight an equivalent definition of regularly varying distributions as distributions that are "asymptotically power law". The following representation theorem for regularly varying functions makes this precise.

Theorem 2.3. *A function $f : \mathbb{R}_+ \to \mathbb{R}_+$ is a regularly varying function with index ρ if and only if $f(x) = x^\rho L(x)$, where $L(x)$ is a slowly varying function.*

Proof. We start by proving the "if" direction. Suppose that $f \in \mathcal{RV}(\rho)$. Define $L(x) = f(x)/x^\rho$. To prove the result, it is enough to show that L is slowly varying, which can be argued as follows:

$$\lim_{x\to\infty} \frac{L(xy)}{L(x)} = \lim_{x\to\infty} \frac{f(xy)}{f(x)} \frac{x^\rho}{(xy)^\rho} = 1.$$

To prove the other direction, we need to show that, given $f(x) = x^\rho L(x)$, where $L(x)$ is a slowly varying function, $f \in \mathcal{RV}(\rho)$. For $y > 0$,

$$\lim_{x\to\infty} \frac{f(xy)}{f(x)} = \lim_{x\to\infty} \frac{(xy)^\rho}{x^\rho} \frac{L(xy)}{L(x)} = y^\rho,$$

which implies, by definition, that $f \in \mathcal{RV}(\rho)$. $\square$

It is important to remember when applying this theorem that regularly varying *functions* can have arbitrary index ρ, however regularly varying *distributions* should must have index $\rho \le 0$.

The key implication of Theorem 2.3 in the context of heavy-tailed distributions is that regularly varying distributions can be thought of as distributions with approximately power law tails in a rigorous sense. That is, they differ from a power law tail only by a slowly varying function $L(x)$, which can intuitively be treated as a constant when doing analysis. This intuition leads to many of the analytic properties that we discuss in the remainder of this section.

However, before we move to the analytic properties of regularly varying distributions, it is useful to illustrate how powerful the representation theorem is by itself. To illustrate this, we use it in the following to argue that regularly varying distributions are heavy-tailed. Of course, this is not a surprising result, given the tie to Pareto distributions, but it is an important foundational result and it provides a simple illustration of how to work with the representation theorem.

Lemma 2.2. *All regularly varying distributions are heavy-tailed.*

Proof. Suppose that the distribution F is regularly varying. We know then that $\bar{F}(x) = x^{-\alpha}L(x)$, where $\alpha \ge 0$, and $L(x)$ is a slowly varying function. Consider $\mu > 0$ and $\beta > \alpha$. Now,

$$\frac{\bar{F}(x)}{e^{-\mu x}} = (x^{-\beta}e^{\mu x})(x^{\beta-\alpha}L(x)).$$

Since $x^{-1}e^{\mu x} \to \infty$ as $x \to \infty$, and $x^{\beta-\alpha}L(x) \to \infty$ as $x \to \infty$ (via Lemma 2.1), we can conclude that $\lim_{x\to\infty} \frac{\bar{F}(x)}{e^{-\mu x}} = \infty$, which proves that F is heavy-tailed. $\square$

The above provides an example of using the fact that regular varying distributions have tails that are approximately power law; however this representation of regularly varying distributions has much broader implications as well. In particular, in the remainder of this section we illustrate a variety of analytic properties of regularly varying distributions that highlight how regularly varying distributions can be analyzed nearly "as if" they are Pareto distributions, as far as the tail is concerned.

In the following, we focus on two crucial analytic properties: (i) integration/differentiation of regularly varying functions, which is formalized via Karamata's Theorem and (ii) inverting the integral transforms of regularly varying functions to understand properties of the tail of the distribution, which is formalized via

Karamata's Tauberian Theorem. In each case, we include a simple example application of the result. Then, in the following section (Section 2.4), we prove closure properties of regularly varying distributions with respect to various algebraic operations in order to provide further illustrations of how to apply these analytic properties.

2.3.1 Integration and differentiation of regularly varying distributions

Perhaps one of the most appealing aspects of working with power law and Pareto distributions is that when one needs to manipulate them to calculate moments, conditional probabilities, convolutions, and other such things, all that is required is the integration or differentiation of polynomials, which is quite straightforward. This is in stark contrast to distributions such as the Gaussian and LogNormal which can be very difficult to work with in this way.

One of the nicest properties of regularly varying distributions is that, in a sense, you can treat them as if they are simply polynomials when integrating or differentiating them – as long as you only care about the tail – and so they are not much more difficult to work with than Pareto distributions. This is especially useful when calculating things like moments, conditional probabilities, convolutions, etc., as we show see repeatedly in the remainder of this chapter and throughout the book.

The foundational properties of regularly varying functions with respect to integration and differentiation are typically referred to as Karamata's Theorem. This result provides the building block for working with regularly varying distributions.

Karamata's Theorem. Karamata's Theorem is perhaps the most important result in the study of regular variation. We start our discussion of it by stating Karamata's Theorem for integration of regularly varying functions, since its statement is a bit cleaner than that of differentiation.

It is useful to begin by anticipating what we should expect Karamata's Theorem to say. To do this, begin by considering what would happen if $f(t) = x^\rho$. In that case,

$$\int_0^x f(t)dt = \frac{x^{\rho+1}}{\rho+1} = \frac{xf(x)}{\rho+1} \text{ if } \rho > -1, \text{ and}$$
$$\int_x^\infty f(t)dt = \frac{x^{\rho+1}}{-(\rho+1)} = \frac{xf(x)}{-(\rho+1)} \text{ if } \rho < -1.$$

Thus, we may expect that Karamata's Theorem would say that, asymptotically, the integrals of regularly varying functions should behave as if the function is a polynomial as far as the tail is concerned, i.e., the $=$ above should be replaced by a $\sim$. In fact, this is exactly what Karamata's theorem states.

Theorem 2.4 (Karamata's Theorem)**.**

(a) For $\rho > -1$, $f \in \mathcal{RV}(\rho)$ if and only if

$$\int_0^x f(t)dt \sim \frac{xf(x)}{\rho+1}.$$

(b) For $\rho < -1$, $f \in \mathcal{RV}(\rho)$ if and only if

$$\int_x^\infty f(t)dt \sim \frac{xf(x)}{-(\rho+1)}.$$

Not surprisingly, regularly varying distributions also asymptotically behave as if they were polynomials with respect to differentiation. In particular, if $f(x) = x^\alpha$, then $f'(x) = \alpha x^{\alpha-1}$, and so $\alpha f(x) = xf'(x)$. The following result, which is commonly referred to as the *monotone density theorem*, shows that exactly this relationship holds for regularly varying distributions with $=$ replaced by $\sim$, modulo some technical conditions.

Theorem 2.5 (Monotone Density Theorem)**.** *Suppose that the function f is absolutely continuous with derivative f'. If $f \in \mathcal{RV}(\rho)$ and f' is eventually monotone, then $xf'(x) \sim \rho f(x)$. Moreover, if $\rho \neq 0$, then $|f'(x)| \in \mathcal{RV}(\rho - 1)$.*[4]

Below, we give the proof of Theorem 2.5. The proof of Theorem 2.4 is a bit more cumbersome and we refer the interested reader to [163, Section 2.3.2] for the proof.

Proof of Theorem 2.5. For simplicity, we assume that $f'(x)$ is non-decreasing over $x \geq x_0$ (the proof for the case of eventually non-increasing f' is along similar lines). Fixing a, b such that $0 < a < b$, we may write

$$\int_{ax}^{bx} \frac{f'(t)}{f(x)}dt = \frac{f(bx) - f(ax)}{f(x)}.$$

For $x > x_0/a$, the monotonicity of f' implies that

$$\frac{f'(ax)x(b-a)}{f(x)} \leq \frac{f(bx) - f(ax)}{f(x)} \leq \frac{f'(bx)x(b-a)}{f(x)}. \tag{2.2}$$

Noting that $f \in \mathcal{RV}(\rho)$, the first inequality above implies that

$$\limsup_{x\to\infty} \frac{f'(ax)x}{f(x)} \leq \frac{b^\rho - a^\rho}{b-a}.$$

Next, letting $b \downarrow a$ on the right side of the above inequality, and noting that this corresponds to taking the derivative of the function x^ρ at $x = a$, we obtain

$$\limsup_{x\to\infty} \frac{f'(ax)x}{f(x)} \leq \rho a^{\rho-1}. \tag{2.3}$$

[4] A function f is said to be *absolutely continuous* if it has a derivative f' almost everywhere that is integrable, such that

$$f(x) = f(0) + \int_0^x f'(t)dt \quad \forall x.$$

A function g is *eventually monotone* if there exists $x_0 > 0$ such that g is monotone over $[x_0, \infty)$.

Similarly, using the second inequality in (2.2), and letting $a \uparrow b$, we obtain

$$\liminf_{x\to\infty} \frac{f'(bx)x}{f(x)} \geq \rho b^{\rho-1}. \tag{2.4}$$

Setting $a = 1$ in (2.3) and $b = 1$ in (2.4), we conclude that $xf'(x) \sim \rho f(x)$. Finally, when $\rho \neq 0$, it is easy to see that $f'(x) \sim \rho \left(\frac{f(x)}{x}\right)$ implies that $|f'(x)| \in \mathcal{RV}(\rho - 1)$.

$\square$

Hopefully, it is already clear that Karamata's Theorem is a particularly appealing and powerful property of regularly varying functions. But, it is worth considering a few examples in order to highlight this further. Perhaps the most powerful example of the use of Karamata's Theorem is in deriving the so-called "Karamata Representation Theorem" for regularly varying functions.

Karamata's Representation Theorem. We have already discussed one representation theorem for regularly varying functions (Theorem 2.3), which allows us to write any regularly varying function as $\bar{F}(x) = x^\rho L(x)$ for some slowly varying function $L(x)$. This is a particularly nice form since it highlights the view of regularly varying distributions as asymptotically power law; however, it is also a fairly implicit view of regularly varying functions since the form of $L(x)$ is not defined. Using Karamata's Theorem, we can derive a much more precise representation theorem for regularly varying functions.

Theorem 2.6 (Karamata's Representation Theorem)**.** *$f \in \mathcal{RV}(\rho)$ if and only if it can be represented as*

$$f(x) = c(x)\exp\left\{\int_1^x \frac{\beta(t)}{t}dt\right\} \tag{2.5}$$

for $x > 0$, where $\lim_{x\to\infty} c(x) = c \in (0, \infty)$ and $\lim_{x\to\infty} \beta(x) = \rho$.

The representation of regularly varying distributions in Karamata's Representation Theorem may initially seem surprising since it does not superficially look like a power law. However, note that if one treats $\beta(t)$ as if it is a constant ρ (which it converges to in the limit), then the exponent becomes $\rho \log x$ and so the power law form appears (since $e^{\rho \log x} = x^\rho$).

Proof. To begin, let us first check that if a function f can be represented via (2.5), then $f \in \mathcal{RV}(\rho)$. Note that for $y > 0$,

$$\frac{f(xy)}{f(x)} = \frac{c(xy)}{c(x)}\exp\left\{\int_x^{xy} \frac{\beta(t)}{t}dt\right\}.$$

Now, since $\beta(x) \to \rho$ as $x \to \infty$, given $\epsilon > 0$, there exists $x_0 > 0$ such that $\rho - \epsilon < \beta(x) < \rho + \epsilon$ for all $x > x_0$. Therefore, for x large enough so that $x, xy > x_0$,

$$\frac{c(xy)}{c(x)}\exp\left\{\int_x^{xy} \frac{\rho-\epsilon}{t}dt\right\} < \frac{f(xy)}{f(x)} < \frac{c(xy)}{c(x)}\exp\left\{\int_x^{xy} \frac{\rho+\epsilon}{t}dt\right\},$$

which is equivalent to

$$\frac{c(xy)}{c(x)}y^{\rho-\epsilon} < \frac{f(xy)}{f(x)} < \frac{c(xy)}{c(x)}y^{\rho+\epsilon}.$$

Now, since $\frac{c(xy)}{c(x)} \to 1$ as $x \to \infty$, we obtain

$$y^{\rho-\epsilon} \leq \liminf_{x\to\infty} \frac{f(xy)}{f(x)} \leq \limsup_{x\to\infty} \frac{f(xy)}{f(x)} \leq y^{\rho+\epsilon}.$$

Letting $\epsilon \downarrow 0$, we finally conclude that $\lim_{x\to\infty} \frac{f(xy)}{f(x)} = y^\rho$, which of course implies by definition that $f \in \mathcal{RV}(\rho)$.

Next, we prove that if $f \in \mathcal{RV}(\rho)$, then f has a representation of the form (2.5). We prove this first for the slowly varying case (i.e., $\rho = 0$), and then consider the case of general ρ.

Accordingly, suppose that $L \in \mathcal{RV}(0)$. Define $b(x) = \frac{xL(x)}{\int_0^x L(t)dt}$. Note that Karamata's theorem implies that $b(x) \to 1$ as $x \to \infty$. Let $\beta_L(x) = b(x) - 1$. Now,

$$\begin{aligned}
\int_1^x \frac{\beta_L(t)}{t} dt &= \int_1^x \frac{L(t)}{\int_0^t L(y)dy} dt - \log(x) \\
&= \log\left(\int_0^x L(y)dy\right) - \log\left(\int_0^1 L(y)dy\right) - \log(x).
\end{aligned}$$

Now, using $\int_0^x L(y)dy = \frac{xL(x)}{b(x)}$, we obtain

$$\int_1^x \frac{\beta_L(t)}{t} dt = \log\left(\frac{L(x)}{b(x)\int_0^1 L(y)dy}\right),$$

which finally gives us

$$L(x) = c_L(x)\exp\left\{\int_1^x \frac{\beta_L(t)}{t} dt\right\}, \tag{2.6}$$

where $c_L(x) = b(x)\int_0^1 L(y)dy$. Noting that $c_L(x) \to \int_0^1 L(y)dy$ and $\beta_L(x) \to 0$ as $x \to \infty$, we have proved that L has the postulated representation.

Finally, moving to the case of general ρ, suppose that $f \in \mathcal{RV}(\rho)$. We know then that $f(x) = x^\rho L(x)$, where $L(x)$ is a slowly varying function. Since we have already established that $L(x)$ can be represented as (2.6), with $\beta_L(x) \to 0$ and $c_L(x) \to c \in (0, \infty)$ as $x \to \infty$. It then follows immediately that

$$f(x) = c_L(x)\exp\left\{\int_1^x \frac{(\beta_L(t) + \rho)}{t} dt\right\},$$

which gives us the desired representation. □

Karamata's Representation Theorem is an extremely powerful tool for working with regularly varying distributions. To highlight this, note that it is straightforward to prove Lemma 2.1 using Karamata's representation theorem (see Exercise 4). Additionally, Karamata's representation theorem can be used to show a number of other properties of regularly varying distributions that connect them to power law and Pareto distributions. We illustrate two of these here: (i) the observation that regularly varying distributions appear approximately linear on a log-log plot, and (ii) properties of the moments of regularly varying distributions.

Let us start by considering the behavior of regularly varying distributions in logarithmic scale. We have seen earlier that one distinguishing property of Pareto distributions is that they are exactly linear when viewed on a log-log scale. Specifically, recall that for Pareto distributions $\bar{F}(x) = (x/x_m)^{-\alpha}$ for $x > x_m$, and so

$$\log \bar{F}(x) = -\alpha \log(x) + \alpha \log(x_m).$$

Thus, $\log \bar{F}(x)$ is exactly linear in terms of $\log x$, with slope α. This is a property that allows for easy preliminary identification of them in data, as we have seen in Chapter 1 and explore in detail in Chapter 8. Note that one must be very cautious using this approach for estimation, as we illustrate in Chapter 8.

Using Karamata's Representation Theorem, we can easily obtain the corresponding property for the tail of regularly varying distributions. In particular, we have the following result, which shows that the tail of regularly varying distributions with index α is asymptotically linear with slope α when viewed on a log-log plot.

Lemma 2.3. *If f is a regularly varying function with index ρ, then*

$$\lim_{x\to\infty} \frac{\log f(x)}{\log(x)} = \rho.$$

Proof. From the Karamata's Representation Theorem, we know that

$$f(x) = c(x)\exp\left\{\int_1^x \frac{\beta(t)}{t}dt\right\},$$

where $\lim_{x\to\infty} c(x) = c \in (0,\infty)$ and $\lim_{x\to\infty} \beta(x) = \rho$.

Given $\epsilon > 0$, there exists $x_0 > 1$ such that for all $x \geq x_0$, $\rho - \epsilon < \beta(x) < \rho + \epsilon$. Therefore, for $x > x_0$,

$$\begin{aligned}\log f(x) &\leq \log c(x) + \int_1^{x_0} \frac{\beta(t)}{t}dt + \int_{x_0}^x \frac{\rho+\epsilon}{t}dt \\ &= \log c(x) + \int_1^{x_0} \frac{\beta(t)}{t}dt + (\rho+\epsilon)(\log(x) - \log(x_0)).\end{aligned}$$

From the above inequality, it follows that

$$\limsup_{x\to\infty} \frac{\log f(x)}{\log(x)} \leq \rho + \epsilon.$$

Using similar arguments, it can be shown that

$$\liminf_{x\to\infty} \frac{\log f(x)}{\log(x)} \geq \rho - \epsilon.$$

Letting ϵ approach zero completes the proof. □

Next, let us move to studying the moments of regularly varying distributions. Recall that the moments of Pareto distributions are a bit peculiar: for Pareto(x_m, ρ) distributions, the ith moment is finite if $i < \rho$ and infinite if $i > \rho$. The fact that moments can be infinite is, as we have seen in Chapter 1, not just of theoretical interest. Data from a variety of situations has been shown to exhibit power law tails with ρ around 1.2—2.1, and so is often well-approximated by distributions with infinite variance.

Using the above result, it is not hard to show that regularly varying distributions have moments that parallel those of Pareto distributions. In particular, we have the following result.

Theorem 2.7. *Suppose that a non-negative random variable X is regularly varying of index $-\rho$. Then*

$$\mathbb{E}\left[X^i\right] < \infty \text{ for } 0 \leq i < \rho,$$
$$\mathbb{E}\left[X^i\right] = \infty \text{ for } i > \rho.$$

The moment conditions in the theorem above should not be particularly surprising at this point since computing moments has to do with integration and the finiteness of moments has to do mainly with the tail, which means that Karamata's theorem should ensure that regularly varying distributions behave like power laws. This intuition serves as a good guide for the proof, which makes use of Lemma 2.3, which was a consequence of Karamta's theorem. We leave the proof of the result as an exercise for the reader (see Exercise 5).

2.3.2 Integral transforms of regularly varying distributions

Integral transforms like the moment generating function, the characteristic function, and the Laplace-Stieltjes transform are of fundamental importance in probability, as well as many physical problems in applied mathematics. It is often easier to study probabilistic and stochastic models using transforms than it is to study them directly as a result of the ease of computing convolutions, moments, time scalings, performing integration of the distribution, etc. Thus, one can typically complete the analysis in "transform space" and then invert the transform to understand the distribution itself, taking advantage of the uniqueness of the representation.

In the context of this book, we have already seen the importance of transforms in the definition of heavy-tailed distributions. Recall that the definition of heavy-tailed distributions explicitly uses the moment generating function (m.g.f.) and defines heavy-tailed distributions as those distributions for which $M_X(t) := \mathbb{E}\left[e^{tX}\right] = \infty$ for all $t > 0$. This highlights that, while moment generating functions are often a powerful analytic tool, working with the m.g.f. of heavy-tailed distributions is problematic. Thus, one needs to consider other integral transforms in the case of heavy-tailed distributions. This section provides the tools for working with transforms in the heavy-tailed setting.

Though the m.g.f. is not appropriate for heavy-tailed distributions, one can instead use other transforms. When the distribution is non-negative, the Laplace-Stieltjes transform (LST) is appropriate and, more generally, the characteristic function is the appropriate tool. The *Laplace-Stieltjes transform* of a function f is defined as

$$\psi_f(s) := \int_{-\infty}^{\infty} e^{-sx} df(x).$$

Specializing to probability distributions, given a random variable X following distribution F, the *Laplace-Stieltjes transform* of F (or X) is defined as

$$\psi_X(s) := \int_{-\infty}^{\infty} e^{-sx} dF(x) = \mathbb{E}\left[e^{-sX}\right].$$

Notice that the LST of a distribution is related to the m.g.f. via a change of variable: we replace the argument t in the definition $M(t)$ by $-s$. Similarly, the characteristic function can be obtained by replacing t with it, where i is the imaginary unit. So, given a random variable X following distribution F, the *characteristic function* of F (or X) is defined as:

$$\phi_X(t) := \mathbb{E}\left[e^{itX}\right] = \mathbb{E}\left[\cos(tX) + i\sin(tX)\right].$$

Note that if X is non-negative, the LST $\psi_X(s)$ is well defined and finite for $s \geq 0$. On the other hand, the characteristic function $\phi(t)$ associated with any distribution is well defined and finite for all t.

To get a feel for the behavior of transforms in the case of heavy-tailed distributions it is useful to look at the specific case of a power law function. To keep things simple, we consider power laws of the form

$$f(x) = \begin{cases} x^\rho & x \geq 0, \\ 0 & x < 0, \end{cases}$$

where $\rho > 0$. Note that with $\rho > 0$, f cannot capture a probability distribution, but limiting our attention positive indices makes things simpler, so we do that for now and then extend the analysis to probability distributions later in the section. We do this because the case of $\rho > 0$ turns out to be quite instructive. In this case, the LST of f can be written as follows:

$$\begin{aligned}
\psi_f(s) &= \int_0^\infty e^{-sx} df(x) \\
&= \rho \int_0^\infty e^{-sx} x^{\rho-1} dx \\
&= \rho s^{-\rho} \int_0^\infty e^{-sx} (sx)^{\rho-1} d(sx) \\
&= \rho s^{-\rho}\Gamma(\rho) = s^{-\rho}\Gamma(\rho+1),
\end{aligned}$$

where the last line uses the Gamma function Γ, which is a continuous extension of the factorial function to the the real numbers defined as $\Gamma(z) := \int_0^\infty e^{-t} t^{z-1} dt$ and satisfying $\Gamma(n) = (n-1)!$ for $n \in \mathbb{Z}$ and $z\Gamma(z) = \Gamma(z+1)$.

The calculation in this example reveals something exciting. The LST of a function f with a power law (as $x \to \infty$) also behaves like a power law (as $s \downarrow 0$). This is exciting because it highlights that one can potentially understand properties of the tail of a distribution by studying properties of the LST near zero. Of course, one could potentially obtain this information by inverting the LST, but that is typically very involved and cannot be done in closed-form apart from a few special cases. In contrast, in this example, the tail behavior of f can be obtained with a simple observation about ψ.

However, before getting too excited, it is important to remember that, so far, we have only seen this behavior in the case of specific f following a power law with a positive index. Thus, the question becomes:

Do regularly varying distributions have regularly varying transforms?

If so, it would be quite powerful since it is often the case that one can derive the LST in situations where it is not tractable to work directly with the distribution.

Fortunately. the answer is "yes". Results of this form are called *Abelian* and *Tauberian* theorems and there are a wide variety of these theorems for the LST and the characteristic function. As a first example, we present Karamata's Tauberian theorem, which is the direct extension of the power law example above for the case of increasing functions.

Theorem 2.8. *(Karamata's Tauberian theorem) Let f be a non-negative right-continuous increasing function such that $f(x) = 0$ for $x < 0$, and let $\rho \geq 0$. Then, for slowly varying $L(x)$, the following are equivalent.*

$$f(x) \sim L(x)x^{\rho} \qquad (x \to \infty), \tag{2.7}$$

$$\psi_f(s) \sim \Gamma(\rho+1)L(1/s)s^{-\rho} \qquad (s \downarrow 0). \tag{2.8}$$

This result says something very powerful. Informally, it says that if the behavior of the LST as $s \downarrow 0$ is approximately a power law, then the corresponding function is also a power law *with the same index*. Further, given the representation of regularly varying functions in Theorem 2.3, this can be interpreted in a different light as well. In particular, since (2.7) characterizes regularly varying functions, the theorem states that regularly varying functions are exactly those that have LSTs that are regularly varying around zero.

Though Theorem 2.8 is commonly called a Tauberian theorem, it actually includes both a Tauberian theorem and an Abelian theorem. In particular the direction showing that (2.7) implies (2.8) is called an Abelian theorem and the reverse direction is a Tauberian theorem. The Tauberian direction is typically harder to prove, which is why such theorems are typically referred to as Tauberian theorems. We omit the proof of Theorem 2.8 here; the interested reader is referred to Theorem 1.7.1 in [28].

Theorem 2.8 is powerful, but does not yet give us exactly what we would like since it still assumes that the index ρ is positive. Thus, it does not apply to regularly varying *distributions* directly. However, it is possible to remedy this. In particular, the following is a more general version of Karamata's Tauberian theorem that uses a Taylor expansion of the LST in terms of the moments of the distribution.

Theorem 2.9. *Consider a non-negative random variable X with distribution F. For $n \in \mathbb{Z}_+$, suppose that $\mathbb{E}[X^n] < \infty$. Then for slowly varying $L(x)$ and $\alpha = n + \beta$ where $\beta \in (0,1)$, the following are equivalent:*

$$\bar{F}(x) \sim \frac{(-1)^n}{\Gamma(1-\alpha)} x^{-\alpha} L(x) \qquad (x \to \infty), \tag{2.9}$$

$$(-1)^{n+1}\left[\psi_X(s) - \sum_{i=0}^{n} \frac{\mathbb{E}\left[X^i\right](-s)^i}{i!}\right] \sim s^{\alpha} L(1/s) \qquad (s \downarrow 0). \tag{2.10}$$

To interpret the statement of Theorem 2.9, note that if a non-negative random variable X satisfies $\mathbb{E}[X^n] < \infty$, then its LST can be expressed via a Taylor expansion as follows:

$$\psi_X(s) = \sum_{i=0}^{n} \frac{\mathbb{E}\left[X^i\right](-s)^i}{i!} + o(s^n) \tag{2.11}$$

The Abelian component of Theorem 2.9 states that if X is regularly varying with index $-\alpha$, where $\alpha \in (n, n+1)$, then the $o(s^n)$ correction term in (2.11) is of the order of s^{α}. The Tauberian component makes the converse implication.

To understand this connection better, let us first consider the case $n = 0$. For this case, Theorem 2.9

states that for $\alpha \in (0,1)$,

$$\bar{F}(x) \sim \frac{1}{\Gamma(1-\alpha)} x^{-\alpha} L(x) \iff \psi_X(s) - 1 \sim -s^{\alpha} L(1/s).$$

Clearly, the above applies to distributions with infinite mean. Consider, for example, the Lévy distribution, which is parameterized by $c > 0$, and has LST $\psi(s) = e^{-\sqrt{2sc}}$ for $s \geq 0$ [106]. Noting that as $s \downarrow 0$, $1-\psi(s) \sim -\sqrt{2cs}$, Theorem 2.9 (specifically, the Tauberian part) implies that the Lévy tail satisfies $\bar{F}(x) \sim \frac{\sqrt{2c}}{\Gamma(1/2)} x^{-1/2}$, i.e., $\bar{F}(x) \sim \sqrt{\frac{2c}{\pi}}\, x^{-1/2}$. The same conclusion can be arrived at by applying Karamata's theorem to the Lévy density function; see Exercise 1.

Next, consider the case $n = 1$. In this case, Theorem 2.9 states that for $\alpha \in (1,2)$,

$$\bar{F}(x) \sim \frac{-1}{\Gamma(1-\alpha)} x^{-\alpha} L(x) \iff \psi_X(s) - 1 + \mathbb{E}[X]\, s \sim s^{\alpha} L(1/s).$$

The above statement in turn is applicable to distributions with a finite first moment, but an infinite second moment.

Proof sketch of Theorem 2.9. We present here the proof of Theorem 2.9 for the case $n = 0$ to illustrate how Theorem 2.9 actually follows from Theorem 2.8. The case $n \geq 1$ is slightly more cumbersome, but follows along similar lines (see Exercise 10).

Let us first consider the Abelian direction. Accordingly, suppose that $\bar{F}(x) \sim \frac{1}{\Gamma(1-\alpha)} x^{-\alpha} L(x)$ for $\alpha \in (0,1)$. Consider now the function $g(x) = \int_0^x \bar{F}(y)dy$, which has LST $\psi_g(s) = \frac{1-\psi_X(s)}{s}$ (checking this claim is left as an exercise for the reader). From Karamata's theorem (Theorem 2.4), note that

$$g(x) \sim \frac{1}{(1-\alpha)\Gamma(1-\alpha)} x^{1-\alpha} L(x).$$

Since $1-\alpha > 0$, we can now invoke Theorem 2.8 (specifically, the Abelian part) to conclude that

$$\psi_g(s) = \frac{1-\psi_X(s)}{s} \sim \frac{\Gamma(2-\alpha)}{(1-\alpha)\Gamma(1-\alpha)} s^{\alpha-1} L(1/s),$$

which in turn implies that $\psi_X(s) - 1 \sim -s^{\alpha} L(1/s)$ (note that $\Gamma(2-\alpha) = (1-\alpha)\Gamma(1-\alpha)$).

Next, consider the Tauberian direction, i.e., suppose that $\psi_X(s) - 1 \sim -s^{\alpha} L(1/s)$, which is equivalent to $\psi_g(s) \sim s^{\alpha-1} L(1/s)$. Invoking Theorem 2.8 (specifically, the Tauberian part), we conclude that $g(x) \sim \frac{1}{\Gamma(2-\alpha)} x^{1-\alpha} L(s)$. Finally, the monotone density theorem (Theorem 2.5) implies that

$$\bar{F}(x) \sim \frac{1-\alpha}{\Gamma(2-\alpha)} x^{-\alpha} L(x) = \frac{1}{\Gamma(1-\alpha)} x^{-\alpha} L(x).$$

This completes the proof.

The proof for general n follows along similar lines; the Abelian direction involves applying Karamata's theorem $n+1$ times, while the Tauberian direction involves applying the monotone density theorem $n+1$ times (see Exercise 10). $\square$

Karamata's Tauberian theorem is only one of many Tauberian theorems that are useful when studying heavy-tailed distributions. In particular, given that it relies on the LST, the versions we have stated above are only relevant for non-negative distributions. For other distributions, one needs Tauberian theorems for the characteristic function. An example of such a Tauberian theorem is the following, which is due to Pitman [158] (see also Page 336 of [28]). Note that this Tauberian theorem uses only the real component of the characteristic function, $U_X(t)$, i.e.,

$$U_X(t) := \mathrm{Re}(\phi_X(t)) = \int_{-\infty}^{\infty} \cos(tx)dF(x).$$

Theorem 2.10 (Pitman's Tauberian theorem)**.** *For slowly varying $L(x)$, and $\alpha \in (0,2)$, the following are equivalent:*

$$\begin{aligned} &Pr(|X| > x) \sim x^{-\alpha} L(x) \text{ as } x \to \infty, \\ &1 - U_X(t) \sim \frac{\pi}{2\Gamma(\alpha)\sin(\pi\alpha/2)} t^{\alpha} L(1/t) \text{ as } t \downarrow 0. \end{aligned}$$

While there are many versions of Abelian and Tauberian theorems for the characteristic function, we choose to highlight this one because we make use of it later in the book in Chapter 5 when introducing and proving the generalized central limit theorem. Like Karamata's Tauberian theorem, this result connects the tail of the distribution to the behavior of a "transform" around zero, only in this case the "transform" considered is the characteristic function. Note that, because this Tauberian theorem applies to the tail of $|X|$ rather than X, it cannot be used to distinguish the behavior of the right and left tails of the distribution. Rather, it provides information only about the sum of the two tails. However, because of this fact, it deals only with the real part of the characteristic function, which makes it much simpler to work with analytically. The interested reader can find more general Abelian and Tauberian theorems in [28].

We have not focused on examples in this section, however there are a number of illustrative examples of how to apply the theorems in this section scattered throughout the book. Two particularly important ones are in Chapter 5, where we apply the Abelian part of Theorem 2.10 to prove the generalized central limit theorem and the Tauberian part of Theorem 2.9 to study the return time of a one-dimensional random walk.

2.4 An example: Closure properties of regularly varying distributions

Regularly varying distributions play a central role in this book, showing up in nearly every chapter. So, as you work through the book you will encounter a variety of applications of the properties and theorems discussed in the previous sections of this chapter. For example, regularly varying distributions play a foundational role in the generalized central limit theorem discussed in Chapter 5, the analysis of a multiplicative processes in Chapter 6, and the discussion of the extremal central limit theorem in Chapter 7.

Here, as a "warm-up" to those applications we provide some simple illustrations of the properties we have studied so far in order to prove some important *closure* properties about the set of regularly varying distributions. These closure properties, while intuitive, should not be taken for granted. In fact, these closure properties do not always hold for the more general classes of heavy-tailed distributions we study in the next two chapters.

Lemma 2.4. *Suppose that the random variables X and Y are independent, and regularly varying of index $-\alpha_X$ and $-\alpha_Y$ respectively.*

(i) $\min(X, Y)$ *is regularly varying with index* $-(\alpha_X + \alpha_Y)$.

(ii) $\max(X, Y)$ *is regularly varying with index* $-\min\{\alpha_X, \alpha_Y\}$.

(iii) $X+Y$ *is regularly varying with index* $-\min\{\alpha_X, \alpha_Y\}$. *Moreover,* $\Pr(X + Y > t) \sim \Pr(\max(X, Y) > t)$.

Lemma 2.4 highlights that the class of regularly varying distributions is closed with respect to min, max, and convolution. These properties should be exactly what you should expect given that intuition that regularly varying distributions are generalizations of Pareto distributions. For example, the sum of two Pareto distributions does not yield another Pareto distribution, of course, but when one considers only the tail the resulting convolution will certainly continue to have a tail that decays like a polynomial, and thus be regularly varying. A similar intuition holds for both the $\min$ and $\max$ of two Pareto distributions. Though the resulting distributions are certainly not Pareto distributions, they still have a tail that decays like a polynomial, and thus are regularly varying.

While simple and intuitive, these closure properties often turn out to be powerful. For example, the third property in Lemma 2.4 can be extended to the case of n i.i.d. regularly varying random variables easily, i.e., $\Pr(Y_1 + Y_2 + \cdots + Y_n > t) \sim n\Pr(Y_1 > t)$ (see Exercise 7). This fact is used crucially in our analysis of random walks in Chapter 7, specifically in the proof of Theorem 7.4.

Proof. We begin by using the representation of regularly varying distributions given by Theorem 2.3. Since X and Y are regularly varying, there exist slowly varying functions L_X and L_Y such that $\Pr(X > t) = t^{-\alpha_X} L_X(t)$ and $\Pr(Y > t) = t^{-\alpha_Y} L_Y(t)$. Now, using these representations, we can prove each closure property in turn.

(i) Note that $\Pr(\min(X, Y) > t) = \Pr(X > t)\Pr(Y > t) = t^{-(\alpha_X+\alpha_Y)} L_X(t) L_Y(t)$. Since the product of slowly varying functions is also slowly varying, Claim (i) of the lemma follows.

(ii) Since $\{\max(X, Y) > t\} = \{X > t\} \cup \{Y > t\}$, we have

$$\Pr(\max(X, Y) > t) = \Pr(X > t) + \Pr(Y > t) - \Pr(X > t)\Pr(Y > t). \tag{2.12}$$

Without loss of generality, we can consider the following cases separately: $\alpha_X < \alpha_Y$, and $\alpha_X = \alpha_Y$.

If $\alpha_X < \alpha_Y$, it follows from (2.12) that $\Pr(\max(X, Y) > t) \sim \Pr(X > t)$, which then implies that $\max(X, Y)$ is regularly varying with index $-\alpha_X$.

If $\alpha_X = \alpha_Y$, then it follows from (2.12) that $\Pr(\max(X, Y) > t) \sim \Pr(X > t) + \Pr(Y > t)$, i.e., $\Pr(\max(X, Y) > t) \sim t^{-\alpha_X}(L_X(t) + L_Y(t))$. Since the sum of slowly varying functions is also slowly varying, it follows that $\max(X, Y)$ is regularly varying with index $-\alpha_X$.

This completes the proof of Claim (ii).

(iii) This final claim is the most involved. The first step in our proof is to establish an upper bound and a lower bound on the probability of the event $\{X + Y > t\}$. Then we analyze those bounds to obtain the result.

To begin, note that the event $\{X > t\} \cup \{Y > t\}$ implies $\{X + Y > t\}$. This gives us the following lower bound.

$$\Pr(X + Y > t) \geq \Pr(X > t) + \Pr(Y > t) - \Pr(X > t)\Pr(Y > t) \tag{2.13}$$

Next, let us fix $\delta \in (0, 1/2)$. It is easy to see that the event $\{X + Y > t\}$ implies the event $\{X > (1-\delta)t\} \cup \{Y > (1-\delta)t\} \cup \{X > \delta t, Y > \delta t\}$. This implication, along with the union bound leads to the following upper bound.

$$\Pr(X + Y > t) \leq \Pr(X > (1-\delta)t) + \Pr(Y > (1-\delta)t) + \Pr(X > \delta t)\Pr(Y > \delta t) \tag{2.14}$$

Now, to complete the proof we consider the following two cases separately: $\alpha_X < \alpha_Y$, and $\alpha_X = \alpha_Y$.

Let us first consider the case $\alpha_X < \alpha_Y$. It follows from (2.13) that

$$\liminf_{t\to\infty} \frac{\Pr(X + Y > t)}{\Pr(X > t)} \geq 1.$$

Similarly, it follows from (2.14) that

$$\limsup_{t\to\infty} \frac{\Pr(X + Y > t)}{\Pr(X > t)} \leq \lim_{t\to\infty} \frac{\Pr(X > (1-\delta)t)}{\Pr(X > t)} = (1-\delta)^{-\alpha_X}.$$

Letting δ approach zero, we conclude that $\Pr(X + Y > t) \sim \Pr(X > t)$. This implies that $X + Y$ is regularly varying with index $-\alpha_X$, and also that $\Pr(X + Y > t) \sim \Pr(\max(X, Y) > t)$ (since we have established in the proof of Claim (ii) that $\Pr(\max(X, Y) > t) \sim \Pr(X > t)$).

Finally, we consider the case $\alpha_X = \alpha_Y$. In this case, using the same steps as above, it can be shown that

$$\Pr(X + Y > t) \sim \Pr(X > t) + \Pr(Y > t).$$

This of course implies that $X+Y$ is regularly varying with index $-\alpha_X$, and also that $\Pr(X + Y > t) \sim \Pr(\max(X, Y) > t)$ (we have established in the proof of Claim (ii) that $\Pr(\max(X, Y) > t) \sim \Pr(X > t) + \Pr(Y > t)$).

This completes the proof of Claim (iii).

□

2.5 An example: Branching processes

Branching processes are a fundamental and widely applicable area of stochastic processes. While they were born from the study of surnames in genealogy, at this point they have found applications broadly in the study of reproduction, epidemiology, queueing theory, statistics, and many other areas. Here we use one of the first, and most famous branching process models – the Galton-Watson process – as an illustrative example of the power of the properties of regularly variation that we have explored in this chapter.

Not only is the Galton-Watson model one of the most prominent examples of a branching process, it has an interesting story behind it. As the story goes, Victorian aristocrats were concerned about keeping their surnames from going extinct and wanted to understand how many children they needed to have to ensure the survival of their name. This prompted Sir Francis Galton to pose the following question in the Educational Times in 1873 [82]:

How many children (on average) must each generation of a family have
in order for the family name to continue in perpetuity?

Just a year later, Reverend Henry William Watson came up with the answer and the two wrote a paper [188]. By now, the model named after them has become the canonical model of branching processes and has been used in wide-reaching areas from biology (see [10]), to the analysis of algorithms (see [57]), to the spread of epidemics (see, for example, [32, 147]).

The modern version of this model is defined formally as follows. In particular, a Galton-Watson process $\{X_n\}_{n\geq 0}$ is defined by:

$$\begin{aligned} X_0 &= 1, \\ X_{n+1} &= \sum_{j=1}^{X_n} N_j^i \quad (n \geq 0), \end{aligned}$$

where $\{N_j^i\}$ are i.i.d. random variables taking non-negative integer values, distributed as N. In the Victorian context, N was interpreted as the number of male children (since the woman took the man's surname at marriage) in a family, and X_n as the number of men in the $n+1$st generation. Given this model, the question asked by Victorian aristocrats can be studied by asking, given the distribution of N, will the process go on forever (i.e., $X_n > 0$ for all n) or will it go extinct (i.e., for some n_o, $X_n = 0$ for all $n \geq n_0$)? And, if it goes extinct, how many total distinct descendents (across all generations) would exist?

It turns out that the answers to these questions depends on the expected number of male children each person has, i.e., $\mu := \mathbb{E}[N]$. It is not hard to see that the probability of extinction, η, is given by $\eta = \lim_{n\to\infty} \Pr(X_n = 0)$. The foundational theorem for Galton-Watson processes illustrates that there are three cases, depending on whether μ is greater than, less than, or equal to 1. Basically, to have a positive probability of avoiding extinction, the expected number of children of each generation needs to be strictly greater than one.

Theorem 2.11. *The probability of extinction, η, in a Galton-Watson branching process satisfies the following:*

(i) Subcritical case: If $\mu < 1$ then $\eta = 1$.

(ii) Critical case: If $\mu = 1$ and N has positive variance, then $\eta = 1$.

(iii) Supercritical case: If $\mu > 1$ then $\eta \in (0, 1)$.

Note that in the subcritical case and the critical case, extinction is guaranteed (except in the trivial case where N equals 1 with probability 1.) On the other hand, in the supercritical case, the lineage has a positive probability of surviving in perpetuity. Theorem 2.11 is classically proven using an approach based on probability generation functions; and we refer the interested reader to [91, Section 5.4] for the proof.

While the subcritical and critical cases are identical from the standpoint of extinction probability, they differ in terms of the distribution of the total number of distinct male descendants $Z := \sum_{n\geq 0} X_n$ as well as the time to extinction $\tau := \min\{n : X_n = 0\}$ (see [18,90]). Here, our goal in studying branching processes is to illustrate the power of the Tauberian theorems we have introduced in this chapter; thus we study the tail of Z in the critical case. Specifically, we prove the following result.

Theorem 2.12. *Suppose that $\mu = 1$ and that N has finite, positive variance. Then, the total number of distinct male descendants, Z, is regularly varying with*

$$Pr\,(Z > t) \sim \frac{1}{\sqrt{\pi(\mathbb{E}\left[N^2\right] - \mathbb{E}\left[N\right])}} t^{-1/2}.$$

Before moving to the proof, it is important to notice that the total number of distinct male descendants has the following recursive structure.

$$Z \overset{d}{=} 1 + \sum_{i=1}^{N} Z_i, \tag{2.15}$$

where $\{Z_i\}$ are i.i.d. random variables with the same distribution as Z and independent of N. To see this, think of N as the number of descendants of the 'first' individual, and Z_i as the number of descendants of his ith child. That each Z_i has the same distribution as Z is referred to as the *branching property*. In the proof below, we exploit this branching property to characterize the tail of Z.

Proof of Theorem 2.12. Since we are going to apply a Tauberian theorem, the transform of Z is important. Here, we use the LST of Z, denoted by ψ_Z. Let $G_N(t) := \mathbb{E}\left[t^N\right] = \sum_{i=0}^{\infty} t^i \Pr\left(N = i\right)$ denote the *probability generating function* of N. It is now not hard to show, using 2.15, that ψ_Z satisfies the following functional equation:

$$\psi_Z(s) = e^{-s} G_N(\psi_Z(s)); \tag{2.16}$$

see Exercise 13.

Analogously to the LST, the probability generating function admits the following Taylor expansion around $t = 1$:

$$G_N(t) = 1 + m_1(t-1) + m_2(t-1)^2(1 + o(1)) \quad \text{as } t \uparrow 1,$$

where $m_1 = \mathbb{E}\left[N\right]$, $m_2 = \mathbb{E}\left[N^2\right] - \mathbb{E}\left[N\right]$. Given that N has finite, positive variance, $m_2 > 0$ (see Exercise 14) and using the fact that $m_1 = \mu = 1$ in the critical case that we are studying, the above expansion simplifies to

$$G_N(t) = t + m_2(t-1)^2(1 + o(1)).$$

Now, we combine this with the functional equation for $\psi_Z(s)$, to obtain

$$\psi_Z(s) = e^{-s}[\psi_Z(s) + m_2(1 - \psi_Z(s))^2(1 + o(1)) \quad \text{as } s \downarrow 0. \tag{2.17}$$

Our goal is to apply Theorem 2.9, but we must first simplify the above expression. To do so, we use a few Taylor expansions. First, note that $e^{-s} = 1 - s(1 + o(1))$. Similarly, we also know that $\psi_Z(s) = 1 - o(1)$, since the total size of Z is finite with probability 1 by Theorem 2.11. Now, if we first move the term $e^{-s}\psi(s)$

to the left hand side of 2.16 and then use the two expansions, we get

$$s = m_2(1-\psi_Z(s))^2(1+o(1)) \quad \text{as } s \downarrow 0.$$

It follows that

$$1-\psi(s) \sim \sqrt{s/m_2} \quad \text{as } s \downarrow 0.$$

Finally, we are ready to apply Theorem 2.9 with $n = 0$, $\alpha = 1/2$, and $L(x) = 1/\sqrt{m_2}$. Using the identity $\Gamma(1/2) = \sqrt{\pi}$, we get

$$\Pr(Z > x) \sim \frac{1}{\sqrt{\pi m_2}} x^{-1/2} \quad \text{as } x \to \infty.$$

□

2.6 Additional notes

In the chapter we gave an overview of several properties of regularly varying *distributions*. While we did not focus much on regularly varying *functions*, the properties we described also apply more broadly. However, the interested reader can find much more on regularly varying *functions* in [28]. That book also contains results for sums that are discrete anologues of Section 2.3.1. That section is the only section in this chapter that assumes continuity.

While we described many analytic and closure properties of regularly varying distributions within this chapter, there are many other useful properties we did not have space to cover. For example, with respect to closure properties, an additional important property is the closure of products of random variables, for which we refer to one of the exercises below; see also [55]. In addition, it can be shown that certain generalized inverses of regularly varying distributions are still regularly varying. On the analytical side, an important property that should be mentioned is the *uniform convergence theorem*, stating that the convergence of $L(at)/L(t)$ in the definition of slowly varying functions is necessarily uniform on any interval $a \in [g, d]$ for $0 < g < d < \infty$. For an overview of these, and many other properties, we refer to the landmark monograph on regular variation [28].

We have focused on the most classical version of regularly varying distributions in this chapter, but it is important to be aware that there are several important extensions of regular variation, some of which will appear in later chapters. Two particularly useful extensions are (i) *intermediate regular variation* and (ii) *dominated variation*. A function f is of intermediate regular variation if

$$\lim_{\epsilon \downarrow 0} \limsup_{x\to\infty} \frac{f(x(1+\epsilon))}{f(x)} = \lim_{\epsilon \downarrow 0} \liminf_{x\to\infty} \frac{f(x(1+\epsilon))}{f(x)} = 1; \tag{2.18}$$

and a function f is of dominated variation if

$$\limsup_{x\to\infty} \frac{f(xa)}{f(x)} < \infty, \liminf_{x\to\infty} \frac{f(xa)}{f(x)} > 0. \tag{2.19}$$

for every $a > 0$.

It is straightforward to see that regularly varying functions of index $\neq 0$ satisfy both intermediate regular variation and dominated variation. In some particular situations, for example in queueing theory the assumption of intermediate regular variation is the most general possible assumption for particular approximations of the tail to hold, or for particular proof methods to work, see [30, 148] for examples.

In statistical applications such as establishing asymptotic normality of estimators, it convenient to consider a subclass of slowly varying functions, which is the class of second-order slowly varying functions. In particular a function L is second-order slowly varying of index γ if there exists a function g which is regularly varying with index γ such that

$$\lim_{t\to\infty} \frac{\frac{L(tx)}{L(t)} - 1}{g(t)} = K\frac{x^\gamma - 1}{\gamma} \tag{2.20}$$

See e.g. Chapter 2 in [28].

As we have mentioned already, regular variation plays an important role in queueing theory and statistics. In addition to its usefulness in these areas, the concept of regular variation is also paramount in financial and insurance mathematics [64], [15]. Another application area where the concept of scale-freeness and regular variation is important is that of complex networks, where the definition of power laws versus the more flexible class of regularly varying distributions sometimes seems to cause some confusion, cf. the discussion in [187]. For an introduction to the field of complex networks, see [25]. In Chapter 6 of this book, we come back to this application, when we discuss the mechanism of preferential attachment as a classical example of the emergence of heavy-tailed phenomena.

Regular variation will reappear at many other places in this book. A non-exhaustive list of examples is

- Chapter 3, where we use regularly varying distributions to investigate the behavior of random sums.
- Chapter 4, where we apply regularly varying distributions to study residual life and illustrate a connection between slowly varying functions and long tailed-distributions.
- Chapter 5, where we use Tauberian Theorems to derive the generalized Central Limit Theorem.
- Chapter 6, where we use regular variation to understand variations of the multiplicative Central Limit Theorem.
- Chapter 7, where characterizing the classes of distributions that admit a limit law for their maxima rely on analytic tools of regular variation, as does the analysis of the all-time maximum of a random walk with negative drift.
- Chapters 8 and 9, where regular variation plays a key role in the development of statistic tools for estimating heavy-tailed distributions from data.

2.7 Exercises

1. Show that the following distributions are asymptotically scale invariant (i.e., regularly varying).

   (a) The Cauchy distribution.

   (b) The Burr distribution.

   (c) The Lévy distribution.

The definitions of these distributions can be found in Chapter 1.

2. Consider a distribution F over $\mathbb{R}_+$ with finite mean μ. The *excess* distribution corresponding to F, denoted by F_e, is defined as
$$\bar{F}_e(x) = \frac{1}{\mu}\int_x^\infty \bar{F}(y)dy.$$
If $F \in \mathcal{RV}(-\alpha)$ for $\alpha > 1$, show that $F_e \in \mathcal{RV}(-(\alpha-1))$. Specifically, show that
$$\bar{F}_e(x) \sim \frac{x}{\alpha - 1}\bar{F}(x).$$

3. Prove that the LogNormal distribution is *not* regularly varying.

4. Prove Lemma 2.1.

5. Prove Theorem 2.7.

6. Prove that if the function f satisfies $xf'(x) \sim \rho f(x)$, then $f \in \mathcal{RV}(\rho)$.
Hint: Use the Karamata representation theorem (Theorem 2.6).

7. Suppose that $X_1, X_2, \cdots, X_n$ are i.i.d. regularly varying random variables with index $-\alpha$, where $n \geq 2$. Prove that
$$\Pr(X_1 + X_2 + \cdots + X_n > x) \sim n\Pr(X_1 > x).$$

8. Suppose that the random variables X and Y are independent, with $X \in \mathcal{RV}(-\alpha_X)$ and
$$\Pr(Y > t) = o(\Pr(X > t)) \quad (t \to \infty).$$
Prove that $X + Y \in \mathcal{RV}(-\alpha_X)$.

9. Suppose that that the random variable $X \in \mathcal{RV}(-\alpha)$. Show that the same property holds for its integer part $\lfloor X \rfloor$.

10. Prove Theorem 2.9 for the case $n = 1$. Specifically, for a non-negative random variable with finite mean, prove that for $\alpha \in (1, 2)$,
$$\bar{F}(x) \sim \frac{-1}{\Gamma(1-\alpha)}x^{-\alpha}L(x) \iff \psi_X(s) - 1 + \mathbb{E}[X]\, s \sim s^\alpha L(1/s).$$
Note: The above exercise should give the reader an idea of how to prove Theorem 2.9 for general n.

11. Let X be a non-negative random variable of which the distribution function F is regularly varying with index $-\alpha$, and let Y be a random variable independent of X for which $E[Y^{\alpha+\epsilon}] < \infty$ for some $\epsilon > 0$. Show that
$$\frac{\Pr(XY > t)}{\Pr(X > t)} \to \mathbb{E}[Y^\alpha] \tag{2.21}$$
as $t \to \infty$. [Hint, condition on the value of Y and use the condition $E[Y^{\alpha+\epsilon}] < \infty$ and a useful property of slowly varying functions to justify the interchange of limit and integral.]

12. Suppose that $f(t)$ is regularly varying of index $\alpha > 0$. Define $f^{\leftarrow}(x) = \inf\{t : f(t) = x\}$. Prove that $f^{\leftarrow}(x)$ is regularly varying of index $1/\alpha$.

13. Suppose that $\{X_i\}_{i \geq 1}$ is a sequence of i.i.d. random variables. The random variable N takes non-negative integer values, and is independent of $\{X_i\}_{i \geq 1}$. Let $G_N(\cdot)$ denote the probability generating function corresponding to N. Define $S_N = \sum_{i=1}^{N} X_i$. Prove that

$$\psi_{S_N}(s) = G_N(\psi_X(s)).$$

Here, $\psi_Y(\cdot)$ denotes the LST corresponding to random variable Y.

14. Suppose that the random variable N takes non-negative integer values. If the variance of N is positive and finite, show that $\mathbb{E}\left[N^2\right] > \mathbb{E}\left[N\right]$.

Chapter 3

Catastrophes, conspiracies, and subexponential distributions

Suppose you are in a class with fifty other students and the professor does an experiment. She records the heights and the number of Twitter followers of every student in the class. It turns out that both the sum of the heights and the sum of the numbers of followers are unexpectedly large. The sum of the student heights comes to more than 310 feet and the sum of the number of twitter followers comes to just over half a million! The average height of a male in the US is 5 feet 9" and the average number of Twitter followers is 700, so these two totals are quite surprising. Since they are significantly larger than the average number one would expect for the sums, the professor asks the class:

What do you think led to the unexpectedly large sums?

Of course there are many possible explanations, but the explanations fall into two general categories: either there are a lot of students in the class that have slightly larger values than average or there are a few people in the class that have extremely large values and everyone else is nearly average.

Let us first think about the case of heights. In this case, the more intuitive of these explanations is clearly the first: that the sum is large because a lot of students are slightly taller than average. For example, maybe the basketball and volleyball teams are taking the class. It is certainly not very likely that the sum of the heights is large because a few students are 20 foot tall giants!

However, in the case of Twitter followers, it is the other way around – the most likely explanation for why the sum is large is that there is one person in the class that is a twitter celebrity, and that person has nearly half a million followers all by themselves.

The contrast between these two explanations is what leads to the title of this chapter: the most likely explanation for a large sum of heights is a "conspiracy" where many people are slightly taller than average and the combination leads to a large sum, while the most likely explanation for a large sum of twitter followers is a "catastrophe" where one person has an extremely large number of followers. The difference between these explanations is quite jarring and, of course, it is not limited to the comparison of heights and number of twitter followers. The fundamental reason for this difference is that the distribution of heights is light-tailed and the distribution of twitter followers is heavy-tailed. Light-tailed distributions tend to follow a "conspiracy principle", while heavy-tailed distributions tend to follow a "catastrophe principle."

To drive the point home, consider a classical example of a heavy-tailed phenomenon – earthquakes. A priori, one might expect that if there are a lot of deaths due to earthquakes in a given year it is most likely the result of there being a lot of earthquakes. However, the catastrophe principle for heavy-tailed distributions tells us that we should expect something different. As in the case of Twitter followers, we should expect that the most likely reason that a year has an unexpectedly large number of deaths due to earthquakes is that there was one earthquake that was extremely deadly. In fact, we should expect that there was one earthquake that led to nearly all of the deaths during the year, i.e., an extreme catastrophe. To make this concrete, let us look back at the deaths caused by earthquakes in recent years. 2010 and 2011 were particularly bad years for earthquakes. It is estimated that there were $320,627$ deaths worldwide from earthquakes in 2010 and $22,053$ deaths worldwide from earthquakes in 2011. Both of these years have considerably larger death tolls than any other years in the last recent past [191, 192]. In both cases, just as the catastrophe principle predicts, the unusually large death toll is primarily the result of one catastrophic event. In 2010 a 7.0 magnitude earthquake in Lèogâne, Haiti led to an estimated $316,000$ deaths, and in 2011 a 9.1 magnitude earthquake in Tōhoku, Japan led to an estimated $20,896$ deaths.

This example highlights that the catastrophe principle can be viewed as a form of *Occam's razor* – the simplest explanation for a large sum is that one large event happened, not that a collection of many slightly larger than expected events conspired together to make the sum large. It is this simplicity that leads to the power of the catastrophe principle's insight about the behavior of heavy-tailed distributions. However, because of its counterintuitive nature, the catastrophe principle is one of the most mysterious properties associated with heavy-tailed distributions. It is completely contrary to what happens under the "typical" light-tailed distributions we learn about in introductory probability courses, e.g., the Gaussian and Exponential distributions.

To understand a bit more about why the conspiracy principle often feels more intuitive, it is useful to think a bit more mathematically. In particular, an overly simplistic, but still useful characterization of light-tailed distributions is that the samples never differ too much from the mean of the distribution, i.e., samples are concentrated around the mean. This means that all the samples from a light-tailed distribution are of a similar size, i.e., on the same scale. The conspiracy principle is quite natural given this view of light-tailed distributions – since all samples are similar, any large sum is most likely to be a combination of many slightly larger than average samples rather than the result of one big outlier, which would be extremely unlikely.

In contrast, a simplistic, but useful view of heavy-tailed distributions is that they are made of "many mice and a few elephants." That is, heavy-tailed distributions yield a lot of small samples (many mice), and when there are large samples, they are *very* large (a few elephants).[1] This view of heavy-tailed distributions provides a very tangible explanation of the catastrophe principle – if the sum is unexpectedly large it is most likely because an elephant arrived. Further, since there are only a few elephants, it is unlikely that two elephants arrived, so the sum is probably dominated by a single elephant. This is sometimes referred to as "the principle of a single big jump," which we formalize in Section 3.4.

While we have been casual in our description of conspiracies and catastrophes so far, these concepts are precise, formal properties. They are not only useful for intuitively reasoning about heavy-tailed and light-tailed distributions; they are also important analytic tools, as we show in this chapter. In particular, in this

[1] The view of heavy-tailed distributions in terms of mice and elephants came out of the computer networks community, where it provided useful intuition for guiding protocol design, e.g., [65, 94, 153]. The perspective highlights that one often needs to handle small events in a different way than large events, or at least one needs to separate the mice from the elephants when designing systems. Given that computing workloads are typically heavy-tailed, this insight has played a key role in the design of routing and scheduling policies within computer networks and distributed systems, e.g., [29, 98, 170].

chapter we formally present the conspiracy and catastrophe principles and connect the catastrophe principle to a general class of heavy-tailed distributions termed "subexponential distributions."

3.1 Conspiracies and catastrophes

To this point we have only described the conspiracy and catastrophe principles informally. However, these principles can be made rigorous, and can serve as powerful analytics tools when studying heavy-tailed and light-tailed distributions. It is important to note that there is not really one catastrophe principle and one conspiracy principle. Instead, there are many variations of these principles that can be defined and used, each with varying strengths and generality. In this section we introduce the simplest statements of each in order to highlight how these properties can be formalized. Later in the chapter, in Section 3.4, we give examples of variations of both catastrophe and conspiracy principles.

3.1.1 A catastrophe principle

As we have already described, the idea behind the catastrophe principle is that an unexpectedly large sum of random variables is likely the result of a catastrophe, i.e, a result of one unexpectedly large event (sample). We can formalize this idea in terms of the tail of a sum of random variables. In particular, if a large sum is most likely the result of a single "catastrophic event", that means that the tail of the sum is *on the same order as* the tail of the maximum element in the sum. Formally, we state this property as follows.

Definition 3.1. *A distribution F over the non-negative reals is said to satisfy the catastrophe principle if, for $n \geq 2$ independent random variables $X_1, X_2, \ldots, X_n$ with distribution F,*

$$Pr\left(\max(X_1, X_2, \ldots, X_n) > t\right) \sim Pr\left(X_1 + X_2 + \ldots + X_n > t\right) \qquad \text{as } t \to \infty.$$

The catastrophe principle is a particularly powerful property because, a priori, there are many things that could have led to the sum being large but the catastrophe principle specifies precisely how it happened. That is, a priori the sum could have been large because every sample was slightly bigger than average, but the catastrophe principle specifies that the sum is large because of exactly one very large sample.

This catastrophe principle can be rephrased in many different ways, and another appealing one is the following. Assuming that the random variables are positive, i.e., $X_i \geq 0$, the catastrophe principle can be reformulated in terms of a conditional probability as follows:

$$\begin{aligned}
\Pr\left(\max(X_1, \ldots, X_n) > t | X_1 + \ldots + X_n > t\right) &= \frac{\Pr\left(\max(X_1, \ldots, X_n) > t,\ X_1 + \ldots + X_n > t\right)}{\Pr\left(X_1 + \ldots + X_n > t\right)} \\
&= \frac{\Pr\left(\max(X_1, \ldots, X_n) > t\right)}{\Pr\left(X_1 + \ldots + X_n > t\right)} \to 1 \text{ as } t \to \infty.
\end{aligned}$$

This highlights that if the catastrophe principle holds, then the maximum of n samples is very likely to be bigger than t given that the sum of n samples is bigger than t. So, the event that the sum is unusually large is, with very high probability, the result of precisely one large "catastrophic" event occurring.

The definition of this catastrophe principle is simple and general enough that it is satisfied by almost all common heavy-tailed distributions. For example, the class of regularly varying distributions satisfies this catastrophe principle. Indeed, Lemma 2.4 states that for i.i.d. regularly varying random variables X_1 and X_2, $\Pr(\max(X_1, X_2) > t) \sim \Pr(X_1 + X_2 > t)$. It is easy to see that the same argument extends to the case of n i.i.d. regularly varying random variables (see Exercise 1). Additionally, many other common heavy-tailed distributions satisfy this catastrophe principle. For example, the Weibull (with $\alpha < 1$) and the LogNormal distribution, although it is more difficult to verify these.[2]

The generality of this catastrophe principle has further led to the definition of a formal class of heavy-tailed distributions – the class of subexponential distributions, which contains the class of regularly varying distributions we study in the previous chapter. We discuss the class of subexponential distributions in depth in Section 3.2.

Though most common heavy-tailed distributions satisfy this catastrophe principle, it is important to note that not all heavy-tailed distributions do. Thus, as is the case with scale-invariance, one needs to be careful to separate the catastrophe principle from the notion of heavy-tailed distributions as a whole. However, it is worth noting that distributions that are heavy-tailed but do not satisfy the catastrophe principle are rather pathological; we comment on constructing such examples in Chapter 4.

3.1.2 A conspiracy principle

In contrast to the catastrophe principle, the conspiracy principle aligns with our intuition about how unexpectedly large events happen – as the combination of a large number of factors. This makes the conspiracy principle more intuitive than the catastrophe principle for most people. However, it is also less powerful in many cases because it does not provide as "simple" an explanation for the rare event.

As in the case of the catastrophe principle, formalizing the notion of the conspiracy principle is most naturally done in terms of the tail of sums of random variables. In particular, the conspiracy principle implies that the tail of a sum of random variables *dominates* the tail of the maximum element in the sum.

Definition 3.2. *A distribution F over the non-negative reals is said to satisfy the conspiracy principle if, for $n \geq 2$ independent random variables $X_1, X_2, \ldots, X_n$ with distribution F,*

$$Pr(\max(X_1, X_2, \ldots, X_n) > t) = o(Pr(X_1 + X_2 + \ldots + X_n > t)) \quad \textit{as } t \to \infty.^3$$

As in the case of the catastrophe principle, this definition can also be interpreted in terms of the conditional probability when the random variables are positive, i.e., $X_i \geq 0$.

$$\begin{aligned}\Pr(\max(X_1, \ldots, X_n) > t | X_1 + \ldots + X_n > t) &= \frac{\Pr(\max(X_1, \ldots, X_n) > t \cap X_1 + \ldots + X_n > t)}{\Pr(X_1 + \ldots + X_n > t)} \\ &= \frac{\Pr(\max(X_1, \ldots, X_n) > t)}{\Pr(X_1 + \ldots + X_n > t)} \to 0 \text{ as } t \to \infty.\end{aligned}$$

The above definition of a conspiracy principle is simple and broad enough that it is easy to show that it is satisfied by all common light-tailed distributions. For example, consider the exponential distribution. For simplicity, we just consider the case of two independent Exponential(μ) random variables X_1 and X_2.

[2] We prove that the Weibull distribution satisfies the catastrophe principle in Definition 3.1 using properties of the class of subexponential distributions in Section 3.2. We leave the case of the LogNormal as Exercise 2.

[3] Recall that $f(t) = o(g(t))$ as $t \to \infty$ if $\lim_{t\to\infty} \frac{f(t)}{g(t)} = 0$.

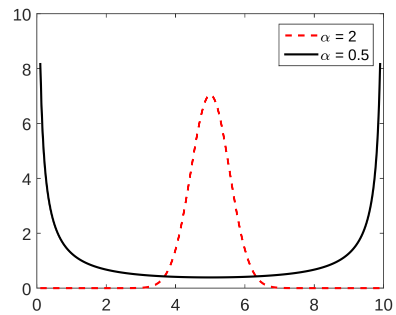


Figure 3.1: ***A visualization of the conspiracy and catastrophe principles under the Weibull distribution.*** *The conditional density function of X_1 given that $X_1 + X_2 = 20$ is shown, where X_1 and X_2 are i.i.d. Weibull random variables that are either light-tailed ($\alpha = 2$) or heavy-tailed ($\alpha = 0.5$).*

In this case we have $\Pr\left(X_1 + X_2 > t\right) = e^{-\mu t}\left(1 + \mu t\right)$. On the other hand, $\Pr\left(\max(X_1, X_2) > t\right) \sim 2e^{-\mu t}$. Therefore, $\Pr\left(\max(X_1, X_2) > t\right) = o(\Pr\left(X_1 + X_2 > t\right))$, which shows that the conspiracy principle holds. It is easy to extend this to the sum of n i.i.d. exponentials (see Exercise 4). Similarly, the Gaussian distribution and the Weibull distribution ($\alpha > 1$) can easily be shown to satisfy the conspiracy principle too (see Exercises 5 and 6).

In fact, the Weibull distribution is a particularly nice example to use to contrast the conspiracy principle with the catastrophe principle since, depending on α, it can satisfy either. Figure 3.1 illustrates the conspiracy and catastrophe principles under different Weibull distributions. For i.i.d. Weibull random variables X_1 and X_2, Figure 3.1 plots the conditional density function of X_1 given that $X_1 + X_2 = 10$, for different values of α. This sum is unexpectedly large since the $\mathbb{E}\left[X_1\right] = \mathbb{E}\left[X_2\right] = 1$. For $\alpha = 0.5$ (X_i are heavy-tailed), we see the catastrophe principle in action. Indeed, given that the sum of X_1 and X_2 is large, then with high probability, X_1 either contributes almost the entire sum or almost nothing. On the other hand, for $\alpha = 2$ (X_i are light-tailed), we see that the density peaks at the center of its support, implying that with high probability, X_2 and X_2 contribute comparably to the sum.

An important insight from Figure 3.1 is that the definition of a conspiracy principle we have stated is, in fact, fairly weak – there is a much stronger notion of a "conspiracy" that is happening under the Weibull distribution with $\alpha > 1$. In particular, not only is $\Pr\left(\max(X_1, X_2) > t\right) = o(\Pr\left(X_1 + X_2 > t\right))$, but neither X_1 or X_2 is contributing much more than the other with respect to the sum. That is, both X_1 and X_2 are equal partners in the conspiracy. It turns out that this can be proven formally in the case of the Weibull as well as other light-tailed distributions, e.g., the Gaussian (see Exercise 7). To highlight this point, the following result shows that, if a sum of two light-tailed Weibull random variables is unexpectedly large, then it is most likely that each of the random variables contributes equally to the sum, i.e., they have "conspired" to make the sum large. This is in stark contrast to the case of heavy-tailed Weibull random variables, where it is most likely that one of the random variables contributes most of the sum, i.e., a "catastrophe" occurs.

Proposition 3.1. *Suppose X_1 and X_2 are independent and identically distributed light-tailed Weibull ran-*

dom variables with shape parameter $\alpha > 1$. Then, for any $\delta \in (1/2, 1)$,

$$Pr\left(X_1 + X_2 > t,\ X_1 > \delta t\right) = o(Pr\left(X_1 + X_2 > t\right)) \quad \text{as } t \to \infty.^4$$

Proof. Recall that the cumulative distribution function of X_i is given by $\Pr\left(X_i > t\right) = e^{-\lambda t^\alpha}$.

To begin, we establish a lower bound on $\Pr\left(X_1 + X_2 > t\right)$, since we need to compare this quantity in order to establish the lemma. To bound this quantity we consider one specific way in which the sum could be large – both X_1 and X_2 could have been bigger than $t/2$. Since there are many other ways the sum could have been large, this is a lower bound. This yields

$$\begin{aligned}\Pr\left(X_1 + X_2 > t\right) &\geq \Pr\left(X_1 > t/2,\ X_2 > t/2\right)\\ &= \Pr\left(X_1 > t/2\right)\ \Pr\left(X_2 > t/2\right)\\ &= e^{-\lambda 2^{1-\alpha} t^\alpha},\end{aligned} \tag{3.1}$$

where the second step follows from independence of X_1 and X_2.

Now, let us move to bounding $\Pr\left(X_1 + X_2 > t,\ X_1 > \delta t\right)$. We can rewrite this probability as

$$\Pr\left(X_1 + X_2 > t,\ X_1 > \delta t\right) = \Pr\left(X_1 > t\right) + \Pr\left(X_1 + X_2 > t,\ \delta t < X_1 \leq t\right).$$

Thus, to prove the lemma, it suffices to show that

$$\Pr\left(X_1 > t\right) = o(\Pr\left(X_1 + X_2 > t\right)), \tag{3.2}$$

and

$$\Pr\left(X_1 + X_2 > t,\ \delta t < X_1 \leq t\right) = o(\Pr\left(X_1 + X_2 > t\right)). \tag{3.3}$$

Our bound of $\Pr\left(X_1 + X_2 > t\right)$ in (3.1) is useful for both of these steps. In particular (3.2) follows easily from (3.1):

$$\frac{\Pr\left(X_1 > t\right)}{\Pr\left(X_1 + X_2 > t\right)} \leq \exp\{-\lambda(1 - 2^{1-\alpha})t^\alpha\},$$

which implies (3.2) since $\alpha > 1$.

All that remains is to prove (3.3). In this case we can directly compute the probability of interest:

$$\begin{aligned}\Pr\left(X_1 + X_2 > t,\ \delta t < X_1 \leq t\right) &= \int_{\delta t}^{t} f_{X_1}(x)\bar{F}_{X_2}(t-x)dx\\ &= \int_{\delta t}^{t} \lambda\alpha x^{\alpha-1} e^{-\lambda x^\alpha} e^{-\lambda(t-x)^\alpha} dx\\ &= \lambda\alpha \int_{\delta t}^{t} x^{\alpha-1} \exp\left\{-\lambda t^\alpha \left[\left(\frac{x}{t}\right)^\alpha + \left(1 - \frac{x}{t}\right)^\alpha\right]\right\} dx.\end{aligned}$$

The form in the last step is chosen in order to facilitate rewriting the equation using the change of variables $g(t) = t^\alpha + (1-t)^\alpha$, where $g(t)$ is defined over the interval $[1/2, 1]$. It is easy to show that the function g

[4] Interestingly, the exponential distribution ($\alpha = 1$), which satisfies the conspiracy principle of Definition 3.2, does not satisfy this stronger conspiracy principle; see Exercise 8.

is strictly increasing over $[1/2, 1]$, with $g(1/2) = 2^{1-\alpha}$. With this definition, the calculation yields

$$\begin{aligned}\Pr\left(X_1 + X_2 > t,\ \delta t < X_1 \le t\right) &= \lambda\alpha \int_{\delta t}^{t} x^{\alpha-1} \exp\left\{-\lambda t^{\alpha} g(x/t)\right\} dx \\ &\le \lambda\alpha \exp\left\{-\lambda t^{\alpha} g(\delta)\right\} \int_{\delta t}^{t} x^{\alpha-1} dx \\ &\le \lambda \exp\left\{-\lambda t^{\alpha} g(\delta)\right\} t^{\alpha}.\end{aligned}$$

Finally, combining the above bound with (3.1), we obtain

$$\frac{\Pr\left(X_1 + X_2 > t,\ \delta t < X_1 \le t\right)}{\Pr\left(X_1 + X_2 > t\right)} \le \lambda t^{\alpha} \exp\left\{-\lambda t^{\alpha}(g(\delta) - g(1/2))\right\},$$

which implies (3.3) and completes the proof. □

The above result provides both an example of how to apply the formal statement of the conspiracy principle in a proof *and* an example of how to formalize a stronger version of the conspiracy principle in the specific case of the Weibull distribution. Note that even stronger versions of the conspiracy principle exist for many light-tailed distributions. We discuss an example of one such variation in detail in Section 3.4.

A final remark about this first conspiracy principle is that, although all common light-tailed distributions satisfy the conspiracy principle, not all light-tailed distributions do (as was the case for the catastrophe principle and heavy-tailed distributions). To illustrate this, consider the following light-tailed distribution, created by mixing common heavy-tailed and light-tailed distributions. Let $X = \min(Y, Z)$, where Y and Z are independent, $Y \sim$ Exponential(μ), and $Z \sim$ Pareto(x_m, α) with $\alpha > 1$. It is not hard to show that X is light-tailed, but does not satisfy the conspiracy principle (see Exercise 10).

3.2 Subexponential distributions

Since our focus in this book is on heavy-tailed distributions, we will dwell a little longer on the catastrophe principle. The importance (and usefulness) of the catastrophe principle has led to the definition of a formal subclass of heavy-tailed distributions called "subexponential distributions" that correspond to those distributions for which a catastrophe principle holds, though the connection between the definition of subexponential distributions and the catastrophe principle will not be obvious initially. The most classical definition of the class of subexponential distributions is the following.

Definition 3.3. *A distribution F with support $\mathbb{R}_+$ is subexponential ($F \in \mathcal{S}$) if, for all $n \ge 2$ independent random variables $X_1, X_2, \cdots, X_n$ with distribution F,*

$$Pr\left(X_1 + X_2 + \cdots + X_n > t\right) \sim n Pr\left(X_1 > t\right),$$

which we can state more compactly as $\bar{F}^{n*}(t) \sim n\bar{F}(t)$.[5]

[5]Recall that $\bar{F}^{n*}(t) := \Pr\left(X_1 + X_2 + \cdots + X_n > t\right)$ denotes the c.c.d.f. of the sum of n i.i.d. random variables having distribution F.

This classical definition does not immediately make it clear how subexponential distributions are related to the catastrophe principle. However, with a simple calculation it is easy to see that they are intimately related. In particular, with a little effort, it is possible to see that $n\text{Pr}\left(X_1 > t\right)$ is asymptotically equivalent to the $\text{Pr}\left(\max(X_1, \ldots, X_n) > t\right)$. To see this, we simply need to expand $\text{Pr}\left(\max(X_1, \ldots, X_n) > t\right)$ as follows

$$\begin{aligned}
\lim_{t\to\infty} \frac{\text{Pr}\left(\max(X_1,\ldots,X_n) > t\right)}{\text{Pr}\left(X_1 > t\right)} &= \lim_{t\to\infty} \frac{1-(1-\bar{F}(t))^n}{\bar{F}(t)} \\
&= \lim_{t\to\infty} \frac{1-(1-n\bar{F}(t)+\binom{n}{2}\bar{F}(t)^2-\ldots)}{\bar{F}(t)} \\
&= \lim_{t\to\infty} \frac{n\bar{F}(t)+o(\bar{F}(t))}{\bar{F}(t)} = n.
\end{aligned}$$

The above highlights that the tail of the max of n random variables is proportional to n times the tail of a single random variable, i.e.,

$$\text{Pr}\left(\max(X_1,\ldots,X_n) > t\right) \sim n\text{Pr}\left(X_1 > t\right), \tag{3.4}$$

and so the definition of subexponential distributions *exactly* matches the catastrophe principle. We state this formally in the following lemma.

Lemma 3.1. *Consider $X_1, X_2, \ldots$ independent random variables with distribution F having support $\mathbb{R}_+$. The following statements are equivalent.*

(i) F is subexponential, i.e., Pr $\left(X_1 + X_2 + \cdots + X_n > t\right) \sim n$Pr $\left(X_1 > t\right)$ for all $n \geq 2$.

(ii) F satisfies the catastrophe principle, i.e., Pr $\left(X_1 + X_2 + \cdots + X_n > t\right) \sim$ Pr $\left(\max(X_1, X_2, \ldots, X_n) > t\right)$ for all $n \geq 2$.

We have already pointed out that most common heavy-tailed distributions satisfy the catastrophe principle, and so a consequence of the above is that they are subexponential distributions. That is, the class of subexponential distributions includes all regularly varying distributions (including the Pareto), the LogNormal, the Weibull (with $\alpha < 1$), and many others; see Figure 3.2.

Though we have already proven that regularly varying distributions satisfy the catastrophe principle (see Lemma 2.4), we have not actually proven that the same is true for the LogNormal, the Weibull, or any other distribution. This is not an accident, it is not straightforward to prove inclusion of these distributions directly from the definitions of the class we have seen so far. In particular, though the definitions we have given so far have intuitive forms and provide useful structure for the analysis of sums of subexponential distributions, they do not provide an easy approach for proving that distributions are subexponential in the first place.

However, there are a number of more easily verifiable conditions that can be used to show that distributions are subexponential. The first such condition is yet another equivalent definition of the class of subexponential distributions. Specifically, it turns out that it is not required for the definition of subexponentiality to hold for *all* $n \geq 2$, if it holds for $n = 2$ then it necessarily holds for all $n \geq 2$. Further, if it can be shown that if it holds for *some* $n \geq 2$ it necessarily holds for $n = 2$ and, consequently, for all $n \geq 2$. This is all summarized in the following lemma.[6]

[6]We state Lemma 3.2 only for the classical notion of subexponentiality $\text{Pr}\left(X_1 + X_2 + \cdots + X_n > t\right) \sim n\text{Pr}\left(X_1 > t\right)$, but

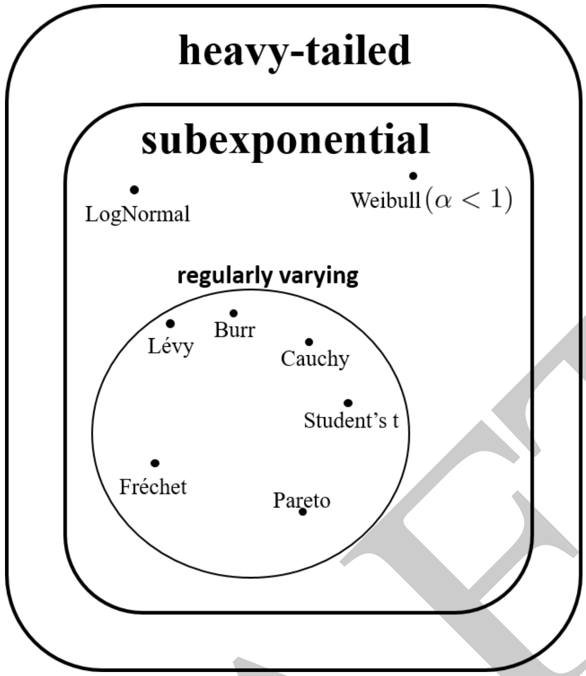


Figure 3.2: *Illustration of the relationship of the class of subexponential distributions to other heavy-tailed distributions.*

Lemma 3.2. *Consider $X_1, X_2, \ldots$ independent random variables with distribution F having support $\mathbb{R}_+$. The following statements are equivalent.*

(i) F is subexponential, i.e., $Pr(X_1 + X_2 + \cdots + X_n > t) \sim nPr(X_1 > t)$ for all $n \geq 2$.

(ii) $Pr(X_1 + X_2 > t) \sim 2Pr(X_1 > t)$.

(iii) $Pr(X_1 + X_2 + \cdots + X_n > t) \sim nPr(X_1 > t)$ for some $n \geq 2$.

We do not give the proof of this lemma here. That (i) and (ii) are equivalent was originally proved in [43] and is the goal of Exercise 13. A proof of the equivalence of (ii) and (iii) can be found in [64, Appendix A].

Clearly, Lemma 3.2 makes the task of verifying subexponentiality easier; however it can still be difficult to work with. Often, the most natural approach for verifying subexponentiality comes through the use of the hazard rate. Recall that the hazard rate, a.k.a., the failure rate, of a distribution F with density f is defined as $q(x) = f(x)/\bar{F}(x)$. Heavy-tailed distributions often have a decreasing hazard rate. In fact, this is such an important property of heavy-tailed distributions that it is the focus of Chapter 4. However, for now, the role of the hazard rate is simply as a tool for checking subexponentiality – as long as the hazard rate decays to zero quickly enough then the distribution is subexponential.

Lemma 3.3. *Suppose that $q(x)$ is eventually decreasing, with $q(x) \to 0$ as $x \to \infty$. If $\int_0^\infty e^{xq(x)} f(x)dx < \infty$, then X is subexponential.*

---

the parallel equivalence holds for $\Pr(X_1 + X_2 + \cdots + X_n > t) \sim \Pr(\max(X_1, X_2, \ldots, X_n) > t)$ as well.

This result is due to Pitman [159], and allows one to exploit a representation of the convolution in terms of hazard rates. The condition in the lemma makes a certain crucial interchange of limit and integration permissible. We do not go into the details, and instead refer to a proof which can be found in [159] or [64, Proposition A3.16].

The form of the condition above is not particularly intuitive; however it provides a clear approach for verifying that a distribution is subexponential. In particular, it is quite effective for showing that the Weibull (with $\alpha < 1$) and the LogNormal distributions are subexponential. For example, recall that the c.c.d.f. of the Weibull distribution is given by $\bar{F}(x) = e^{-\beta x^\alpha}$, and so its corresponding hazard rate is given by

$$q(x) = \frac{\alpha\beta x^{\alpha-1} e^{-\beta x^\alpha}}{e^{-\beta x^\alpha}} = \alpha\beta x^{\alpha-1}.$$

Clearly, this hazard rate is decreasing to zero monotonically. So, plugging this hazard rate into the condition of Lemma 3.3, we can simply calculate the integral to show that the Weibull with $\alpha < 1$ is subexponential.

$$\int_0^\infty e^{xq(x)} f(x)dx = \alpha\beta \int_0^\infty e^{x\alpha\beta x^{\alpha-1}} \alpha\beta x^{\alpha-1} e^{-\beta x^\alpha} dx = \frac{1}{1-\alpha} < \infty.$$

Note that if $\alpha \geq 1$ then the integral is infinite.

As one would expect from the name “subexponential”, all subexponential distributions have a tail that is “sub” exponential, i.e., their tail decays more slowly than an Exponential and thus the tail is heavier than an exponential. However, to this point, we have not actually proven that this is true. In fact, it is not obvious from any of the definitions we have given of subexponential distributions that they are *necessarily* heavy-tailed.

Of course, all subexponential distributions are indeed heavy-tailed, as the name suggests. However, proving this statement is not as easy as one might expect. The difficulty of the proof of the fact that subexponential distributions are heavy-tailed highlights that it is not always easy to work with subexponential distributions. They are well-suited for analysis of random sums and extrema, where the definition provides useful analytic structure; however they can be difficult to work with in general. To end this section, we prove the seemingly obvious statement that subexponential distributions are heavy-tailed and, in the process, highlight some common techniques for working with subexponential distributions. The proof highlights a strong connection between the class of subexponential distributions and the class of long-tailed distributions, which we discuss in Chapter 4.

Lemma 3.4. *Subexponential distributions are heavy-tailed.*

Proof. Our proof has two steps. First, we show that if a distribution $F \in \mathcal{S}$, then it satisfies the following property

$$\lim_{t\to\infty} \frac{\bar{F}(t-y)}{\bar{F}(t)} = 1 \qquad \forall\, y > 0. \tag{3.5}$$

While this property may seem mysterious now, it is actually the defining property of the class of long-tailed distributions, which are the focus of Chapter 4. However, that will not be important for us at this point. At this point, Condition (3.5) is important for us because it implies that F is heavy-tailed, and the second step in the proof is showing that implication.

To begin, suppose that the distribution $F \in \mathcal{S}$, and let $X_1,\ X_2$ be independent random variables with distribution F. Recall that $\bar{F}^{2*}(t) := \Pr(X_1 + X_2 > t)$. From the definition of subexponentiality we have

$\lim_{t\to\infty} \frac{\bar{F}^{2*}(t)}{\bar{F}(t)} = 2$. Combining this with the fact that we can write

$$\frac{\bar{F}^{2*}(t)}{\bar{F}(t)} = \frac{\bar{F}(t) + \int_0^t \bar{F}(t-u)dF(u)}{\bar{F}(t)},$$

it follows that

$$\lim_{t\to\infty} \int_0^t \frac{\bar{F}(t-u)}{\bar{F}(t)} dF(u) = 1.$$

Let us denote the function in the above limit by $I(t)$. Fixing $y > 0$, we may bound $I(t)$ as follows:

$$\begin{aligned} I(t) &= \int_0^y \frac{\bar{F}(t-u)}{\bar{F}(t)} dF(u) + \int_y^t \frac{\bar{F}(t-u)}{\bar{F}(t)} dF(u) \\ &\geq \int_0^y dF(u) + \frac{\bar{F}(t-y)}{\bar{F}(t)} \int_y^t dF(u) \\ &= F(y) + \frac{\bar{F}(t-y)}{\bar{F}(t)} (F(t) - F(y)). \end{aligned}$$

The above bound implies that

$$\frac{\bar{F}(t-y)}{\bar{F}(t)} \leq \frac{I(t) - F(y)}{F(t) - F(y)}.$$

Now, since $\lim_{t\to\infty} I(t) = 1$, we obtain

$$\limsup_{t\to\infty} \frac{\bar{F}(t-y)}{\bar{F}(t)} \leq 1.$$

Of course, since $\bar{F}(t-y) \geq \bar{F}(t)$, we also have

$$\liminf_{t\to\infty} \frac{\bar{F}(t-y)}{\bar{F}(t)} \geq 1.$$

This proves that F satisfies Condition (3.5), completing the first step of the proof.

We now move to the second step of the proof: proving that Condition (3.5) implies that F is heavy-tailed. To show this we will prove that F is heavy-tailed using Lemma 1.1, which says that it is enough to prove that $\liminf_{x\to\infty} -\frac{\log \Pr(X>x)}{x} = 0$.

To this end, define $\Psi(t) = -\log \Pr(X > t)$. It is easy to see that Condition (3.5) implies that $\lim_{t\to\infty}(\Psi(t) - \Psi(t-1)) = 0$. Therefore, for $\epsilon > 0$, there exists $t_0 > 0$ such that $\Psi(t) - \Psi(t-1) < \epsilon$ for all $t \geq t_0$. Since Ψ is a non-decreasing function, it follows that for $t \geq t_0$,

$$\Psi(t) \leq \Psi(t_0) + \lceil t - t_0 \rceil \epsilon,$$

which implies that $\limsup_{t\to\infty} \frac{\Psi(t)}{t} \leq \epsilon$. Letting $\epsilon \downarrow 0$, we get that

$$\limsup_{t\to\infty} \frac{\Psi(t)}{t} \leq 0.$$

Moreover, since $\Psi(t) \geq 0$, we can conclude that

$$\lim_{t\to\infty} \frac{\Psi(t)}{t} = 0.$$

From Lemma 1.1, it follows now that F is heavy-tailed, which completes the proof. $\square$

3.3 An example: Random sums

Given the overview of subexponential distributions provided in the previous section, the goal of this section is to provide an example of how properties of subexponential distributions can be powerful analytic tools, enabling analysis in a wide array of applications. Since the class of subexponential distributions serves as a formalization of the catastrophe principle, it is natural that it finds application most readily in settings that are fundamentally related to some form of random sum. Of course, such applications are common in domains such as finance, insurance, scheduling, and queueing (among many others).

In many such settings the core of the analysis relies on understanding a very simple process – a sum of a random number of independent and identically distributed random variables. For example, if one considers the number of deaths from earthquakes in a given year, which we discussed at the beginning of this chapter, it is reasonable to consider that there are a random number of earthquakes in a year and that the number of deaths from each one is independent and identically distributed. Of course, both the distribution of the number of earthquakes and the number of deaths from an earthquake are heavy-tailed [114, 157]. A similar model could capture the amount of money paid out in insurance claims in a given year, the amount of load arriving to a cloud service in a given day, and many other situations.

In this section, we illustrate the power of the class of subexponential distributions by considering a simple and classical model of random sums that underlies such situations. Specifically, suppose $\{X_i\}_{i\geq 1}$ is a sequence of independent and identically distributed random variables with mean $\mathbb{E}[X]$ and the random variable N takes values in $\mathbb{N}$ and is independent of $\{X_i\}_{i\geq 1}$. Our goal will be to characterize

$$S_N = \sum_{i=1}^{N} X_i.$$

You have likely studied the expectation of this random sum in an introductory probability course. In particular, you have likely heard of Wald's equation, which gives a simple and pleasing formula for $\mathbb{E}[S_N]$:

$$\mathbb{E}[S_N] = \mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}[N]\,\mathbb{E}[X].$$

Wald's equation is particularly pleasing because it tells us that, with respect to the expectation, we can basically ignore the fact that N is random. That is, if we had just considered S_n for some fixed constant n, then $E[S_n] = nE[X]$, and Wald's equation simply replaces n with $E[N]$.

While Wald's equation is a very useful result, it is not always enough to have a characterization of the expectation S_N. We often want to understand the variance of S_N, or even the distribution of S_N. Luckily, it is not hard to generalize Wald's equation. For example, the variance of the random sum S_N still has a pleasing form:

$$\text{Var}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}[N]\,\text{Var}[X] + (\mathbb{E}[X])^2\text{Var}[N].$$

It is even possible to go much further than just the variance and to derive Wald-like inequalities for the tail of random sums. However, results about the tail of S_N are not as general as Wald's equation and rely on using particular properties of distributions. In fact, as we show below, the tail of random sums can behave very differently depending on whether X_i and/or N are heavy-tailed or light-tailed. It is in the derivation of these results that the class of subexponential distributions shows its value (as will the class of regularly varying distributions).

The tail of random sums

Before talking about formal results characterizing the tail of random sums it is useful to think about how we should expect that tail to behave. One natural suggestion is that we should expect something like we saw with Wald's equation – it should be "as if" the random variable N was simply a constant n. If this were the case, and if the X_i are subexponential, we would have very simple equation for the tail of the sum:

$$\text{Pr}\left(\sum_{i=1}^{n} X_i > t\right) \sim n\text{Pr}(X_1 > t).$$

Thus, a natural guess is that this formula should hold for the tail of the random sum as well, by simply replacing n with $\mathbb{E}[N]$.

The above guess is indeed correct when N is light-tailed. In this case, since the X_i are heavy-tailed they "dominate" the behavior of the tail of the random sum, and so only the expectation of N plays a role. We formalize this in the following theorem.

Theorem 3.1. *Consider an infinite i.i.d. sequence of subexponential random variables $X_1, X_2, \ldots$, and a light-tailed random variable $N \in \mathbb{N}$ that is independent of $\{X_i\}_{i\geq 1}$. Then,*

$$Pr\left(\sum_{i=1}^{N} X_i > t\right) \sim \mathbb{E}[N]\, Pr(X_1 > t).$$

Proof. The idea behind this proof is simple. We rewrite the tail of our random sum taking advantage of the definition of subexponential distributions. Denoting the distribution of X by F, conditioning on N gives

$$\text{Pr}\left(\sum_{i=1}^{N} X_i > t\right) = \sum_{i\in\mathbb{N}} \text{Pr}(N = i)\, \bar{F}^{i*}(t).$$

Looking at the limit we can rewrite this in a form where we can apply the definition of subexponential distributions as follows:

$$\lim_{t\to\infty} \frac{\text{Pr}\left(\sum_{i=1}^{N} X_i > t\right)}{\text{Pr}(X_1 > t)} = \lim_{t\to\infty} \sum_{i\in\mathbb{N}} \frac{\bar{F}^{i*}(t)}{\bar{F}_1(t)} \text{Pr}(N = i).$$

Applying the definition of subexponential distributions, specifically that $\lim_{t\to\infty} \frac{\bar{F}^{i*}(t)}{\bar{F}_1(t)} = i$, gives the following, where we assume for the moment that we can interchange the order of the limit and the summation:

$$\lim_{t\to\infty} \sum_{i\in\mathbb{N}} \frac{\bar{F}^{i*}(t)}{\bar{F}_1(t)} \Pr(N=i) = \sum_{i\in\mathbb{N}} i\Pr(N=i) = \mathbb{E}[N].$$

All that remains is to justify the interchange of the limit and the summation. This can be done via an application of the dominated convergence theorem, which states that if we can find a t-independent upper bound β_i of $\frac{\bar{F}^{i*}(t)}{\bar{F}_1(t)}\Pr(N=i)$ such that $\sum_{i\in\mathbb{N}} \beta_i < \infty$, then the limit and the summation can be interchanged. The upper bound β_i can be obtained as follows. It can be proved that for any $\epsilon > 0$, there exists $K > 0$ such that $\frac{\bar{F}^{i*}(t)}{\bar{F}_1(t)} \leq K(1+\epsilon)^i$ (see Exercise 11). Taking $\beta_i = K(1+\epsilon)^i \Pr(N=i)$, we get $\sum_{i\in\mathbb{N}} \beta_i = \mathbb{E}\left[(1+\epsilon)^N\right]$, which is finite for small enough ϵ since N is light-tailed.

□

The characterization of random sums in the previous theorem relies on the fact that the distribution of X_i is dominant, i.e., X_i are heavy-tailed and N is light-tailed. In this case, we obtained a simple form for the tail of the random sum that follows the intuition we derived from Wald's equation. What remains now is to understand what happens when things are reversed and the distribution of N is dominant, i.e., when N is heavy-tailed and X_i is light-tailed.

Intuitively, in this case, we should expect the tail of the random sum to be determined by the tail of N. To get intuition, let us consider what would happen if the X_i were deterministically x. In such a case, the behavior of the sum would be simple:

$$\Pr\left(\sum_{i=1}^{N} x > t\right) = \Pr(Nx > t) = \Pr(N > t/x).$$

Thus, one might guess that this formula should continue to hold for the tail of the random sum as well, simply by replacing x with $\mathbb{E}[X]$.

This guess is again correct. In particular, when X_i is light-tailed and N is heavy-tailed, specifically when N is regularly varying, we can characterize the random sum as follows. The proof of this result serves as a good example of the usefulness of the properties of regularly varying distributions that we discussed in Chapter 2.

Theorem 3.2. *If the X_i are light-tailed and N is regularly varying with index $-\alpha$ ($\alpha > 0$), then*

$$Pr\left(\sum_{i=1}^{N} X_i > t\right) \sim Pr\left(N > \frac{t}{\mathbb{E}[X]}\right).$$

Before we prove this result, it is important to point out that this characterization does not hold in general for heavy-tailed N. To get intuition as to why not, consider the case when N is deterministically n. In this case, the central limit theorem implies that for large n, $S_n \approx n\mathbb{E}[X_1] + \sqrt{n}Z$ for a Gaussian random variable Z. Thus, intuitively one could guess that

$$\Pr(S_N > t) \approx \Pr\left(N\mathbb{E}[X_1] + \sqrt{N}Z > t\right) \approx \Pr\left(N > t/\mathbb{E}[X_1] - O(\sqrt{t})\right). \tag{3.6}$$

This heuristic turns out to be correct and it can be shown that, if the X_i are not deterministic, a necessary condition for $\Pr\left(\sum_{i=1}^N X_i > t\right) \sim \Pr\left(N > \frac{t}{\mathbb{E}[X]}\right)$ to hold is that

$$\Pr(N > x) \sim \Pr\left(N > x - \sqrt{x}\right) \tag{3.7}$$

This property is referred to as *square-root insensitivity* and plays a prominent role in the analysis of heavy-tailed distributions. Many common heavy-tailed distributions are square-root insensitive, including the Pareto, the LogNormal, and the Weibull (if $\alpha < 0.5$) (see Exercise 9). For a more in depth discussion of square-root insensitivity and its applications, see [54, 138].

Proof of Theorem 3.2. To prove the theorem we establish asymptotically matching upper and lower bounds on $\Pr\left(\sum_{i=1}^N X_i > t\right)$. In both cases we rely on properties of regularly varying distributions introduced in Chapter 2.

We start with the lower bound. To obtain a lower bound we characterize the likelihood that the sum is large as a result of N being large. The intuition behind this is that, since N is heavy-tailed, this is likely to be the dominant event. Specifically, let $\epsilon > 0$, and let $n(t) = \frac{t(1+\epsilon)}{\mathbb{E}[X_1]}$. We now condition on $N > n(t)$ as follows.

$$\begin{aligned}\Pr\left(\sum_{i=1}^N X_i > t\right) &\geq \Pr\left(\sum_{i=1}^N X_i > t,\ N > n(t)\right) \\ &\geq \Pr\left(\sum_{i=1}^{\lceil n(t)\rceil} X_i > t,\ N > n(t)\right).\end{aligned}$$

In the second step above we used the fact that the event $\{\sum_{i=1}^{\lceil n(t)\rceil} X_i > t,\ N > n(t)\}$ implies the event $\{\sum_{i=1}^N X_i > t,\ N > n(t)\}$ in order to eliminate N from the sum. This gives independence between the two events in the probability and allows us to simplify the expression as follows

$$\begin{aligned}\Pr\left(\sum_{i=1}^N X_i > t\right) &\geq \Pr\left(\sum_{i=1}^{\lceil n(t)\rceil} X_i > t\right)\Pr(N > n(t)) \\ &\geq \Pr\left(\sum_{i=1}^{\lceil n(t)\rceil} X_i > \frac{\mathbb{E}[X_1]}{1+\epsilon}\lceil n(t)\rceil\right)\Pr(N > n(t)),\end{aligned}$$

where the second step follows from the definition of $n(t)$. Now, the weak law of large numbers allows us to conclude that the first probability above approaches 1 as $t \to \infty$. Therefore, we have

$$\liminf_{t\to\infty} \frac{\Pr\left(\sum_{i=1}^{N} X_i > t\right)}{\Pr\left(N > \frac{t}{\mathbb{E}[X_1]}\right)} \geq \lim_{t\to\infty} \frac{\Pr\left(N > \frac{t(1+\epsilon)}{\mathbb{E}[X_1]}\right)}{\Pr\left(N > \frac{t}{\mathbb{E}[X_1]}\right)} = (1+\epsilon)^{-\alpha},$$

where the last step follows from the fact that N is regularly varying with index $-\alpha$. Finally, letting $\epsilon \downarrow 0$, gives the desired lower bound

$$\liminf_{t\to\infty} \frac{\Pr\left(\sum_{i=1}^{N} X_i > t\right)}{\Pr\left(N > \frac{t}{\mathbb{E}[X_1]}\right)} \geq 1.$$

We now turn to the upper bound. We partition the event of interest based on the size of N. In particular, we pick $\epsilon \in (0,1)$, and let $m(t) = \frac{t(1-\epsilon)}{\mathbb{E}[X_1]}$. Then, we partition the event that the random sum exceeds t based on whether or not N exceeds $m(t)$:

$$\begin{aligned}\Pr\left(\sum_{i=1}^{N} X_i > t\right) &= \Pr\left(\sum_{i=1}^{N} X_i > t,\ N > m(t)\right) + \Pr\left(\sum_{i=1}^{N} X_i > t,\ N \leq m(t)\right)\\ &\leq \Pr\left(N > m(t)\right) + \Pr\left(\sum_{i=1}^{\lfloor m(t)\rfloor} X_i > t\right)\end{aligned}$$

In the section step we have simplified the expression by using the fact that the event $\{\sum_{i=1}^{N} X_i > t,\ N \leq m(t)\}$ implies the event $\{\sum_{i=1}^{\lfloor m(t)\rfloor} X_i > t\}$. Next, plugging in the definition of $m(t)$ yields

$$\Pr\left(\sum_{i=1}^{N} X_i > t\right) \leq \Pr\left(N > m(t)\right) + \Pr\left(\sum_{i=1}^{\lfloor m(t)\rfloor} X_i > \frac{\mathbb{E}\left[X_1\right]}{1-\epsilon} \lfloor m(t)\rfloor\right)$$

Now, to bound the second term we apply a Chernoff bound. Specifically, we use the following result: Given light-tailed i.i.d. random variables $\{Y_i\}_{i\geq 1}$ having positive mean, for $\alpha > \mathbb{E}\left[Y_1\right]$, there exists $\phi > 0$ such that $\Pr\left(Y_1 + Y_2 + \cdots Y_n > n\alpha\right) \leq e^{n\phi}$ for all $n \geq 1$. Applying this result to $\{X_i\}_{i\geq 1}$, we conclude that there exists a positive constant ϕ such that

$$\Pr\left(\sum_{i=1}^{\lfloor m(t)\rfloor} X_i > \frac{\mathbb{E}\left[X_1\right]}{1-\epsilon} \lfloor m(t)\rfloor\right) \leq e^{-\phi\lfloor m(t)\rfloor} \leq e^{-\phi(m(t)-1)} = ce^{-\mu t},$$

where $c = e^{\phi}$ and $\mu = \frac{\phi(1-\epsilon)}{\mathbb{E}[X_1]}$. Therefore, we have

$$\frac{\Pr\left(\sum_{i=1}^{N} X_i > t\right)}{\Pr\left(N > \frac{t}{\mathbb{E}[X_1]}\right)} \leq \frac{\Pr\left(N > \frac{t(1-\epsilon)}{\mathbb{E}[X_1]}\right)}{\Pr\left(N > \frac{t}{\mathbb{E}[X_1]}\right)} + \frac{ce^{-\mu t}}{\Pr\left(N > \frac{t}{\mathbb{E}[X_1]}\right)}.$$

Since N is regularly varying, its tail decays approximately like a polynomial and thus it is easy to see that

$$\lim_{t\to\infty} \frac{ce^{-\mu t}}{\Pr\left(N > \frac{t}{\mathbb{E}[X_1]}\right)} = 0,$$

which implies that

$$\begin{aligned}\limsup_{t\to\infty} \frac{\Pr\left(\sum_{i=1}^N X_i > t\right)}{\Pr\left(N > \frac{t}{\mathbb{E}[X_1]}\right)} &\le \lim_{t\to\infty} \frac{\Pr\left(N > \frac{t(1-\epsilon)}{\mathbb{E}[X_1]}\right)}{\Pr\left(N > \frac{t}{\mathbb{E}[X_1]}\right)} \\ &= (1-\epsilon)^{-\alpha},\end{aligned}$$

where the last step follows from the fact that N is regularly varying with index $-\alpha$. Letting $\epsilon \downarrow 0$, we get our desired upper bound

$$\limsup_{t\to\infty} \frac{\Pr\left(\sum_{i=1}^N X_i > t\right)}{\Pr\left(N > \frac{t}{\mathbb{E}[X_1]}\right)} \le 1.$$

□

3.4 An example: Conspiracies and catastrophes in random walks

In order to introduce the catastrophe principle and the conspiracy principle as fundamental properties of many heavy-tailed and light-tailed distributions, we have focused on simple, intuitive forms of these properties up until this point. The goal of this section is to highlight that more precise versions of these principles can be obtained, at the cost of technical complexity. To illustrate this we highlight variations of both the catastrophe principle and the conspiracy principle in the context of *random walks*.

Random walks are one of the most classical examples of a stochastic process and have found application in remarkably diverse settings, including finance, computer science, physics, biology, and more. In this section, we study a generic random walk of the following form

$$S_n = X_1 + X_2 + \ldots + X_n,$$

for i.i.d. X_i with mean μ. Here, S_n is the position of a walker who is taking steps of size X_i at each stage.

Most typically, analyses of random walks focus on the expected behavior of the random walk; however, that is not our goal here. The power of the conspiracy and catastrophe principle comes from providing explanations for rare events, situations where an unexpectedly large deviation from the "normal" behavior happens. Thus, the goal of the results here is to understand the large deviations of random walks, i.e., $\Pr(S_n > t)$ for large t. But, it is not enough to consider a fixed t since what qualifies as a large deviation should depend on n. Thus, more precisely, in this section we study

$$\Pr(S_n > an) \text{ as } n \to \infty \text{ for } a > \mu.$$

Events where $S_n > an$ for $a > \mu$ truly are "large" in the sense that they cannot be characterized by the central limit theorem, which focuses on "small" deviations, on the order of $\sqrt{n}$, around the mean μn. This means that understanding $\Pr(S_n > an)$ requires understanding *rare events*. To this end, the results we present in this section both characterize how likely such rare events are *and* what leads to such rare events.

In particular, in the setting where X_i are heavy-tailed (specifically regularly varying), we strengthen the catastrophe principle to provide the "principle of a single big jump", which has seen application in a variety of applications from scheduling and queueing to insurance and finance. In the setting where X_i are light-tailed, we strengthen the conspiracy principle to obtain Cramér's theorem, which is a fundamental concentration inequality in the theory of large deviations whose applications cannot be overstated. Cramér's theorem can be thought of as a third fundamental limit theorem, alongside the law of large numbers and the central limit theorem.

Of course these are only two such examples of stronger catastrophe and conspiracy principles and there are many others available, depending on the assumptions and requirements of the desired setting. We discuss some other extensions in the additional notes at the end of the chapter.

3.4.1 The principle of a single big jump

The principle of a single big jump is a variation of the catastrophe principle that adapts the ideas we have discussed so far to the setting of random walks. To begin, let us consider what insight we can already obtain for random walks given the catastrophe principle in Definition 3.1. If the steps of the walk, X_i, are subexponential, the catastrophe principle already gives us a powerful insight into the behavior of the random walk. In particular, if n is fixed, then

$$\frac{\Pr(S_n > t)}{n\Pr(X_1 > t)} \to 1, \text{ as } t \to \infty. \tag{3.8}$$

The intuition behind this statement should be very familiar by this point in the chapter: it is highly likely that, when random walk has a large deviation from the mean, a single large X_i is responsible.

However, in the context of random walks, the statement in (3.8) is not completely satisfying since it considers only the case when n is fixed. When studying random walks, the goal is typically to understand the behavior of the random walk in the long run as n grows large. Thus, what we really want is a more powerful statement that bounds $\Pr(S_n > an)$ as $n \to \infty$ for $a > \mu$.

If you have internalized the ideas in this chapter you, of course, expect the intuition from the catastrophe principle to still hold in this case. This intuition allows us to heuristically work out what we should expect a bound on $\Pr(S_n > an)$ to look like. In particular, if we expect that exactly one X_i will be large and the others will be approximately the expected value μ, we obtain

$$\begin{aligned}\Pr(S_n > an) &\approx n\Pr(X_1 > (an - \mu(n-1)), X_2 + \ldots + X_n \approx \mu(n-1)) \\ &\approx n\Pr(X_1 > (a-\mu)n), \text{ for large } n.\end{aligned}$$

The form suggested by this intuition turns out to hold, and is termed the "principle of a single big jump". Not only does the intuition bound the likelihood of such a rare event, it also suggests that a large deviation of a random walk is most likely the result of one, and not more than one, big jump in the walk.

In the rest of this section we state and prove the principle of a single big jump formally. Note that the principle does not hold generally for subexponential distributions. We prove it here in the case of regularly varying distributions, though some generalizations can be found in [54].

Theorem 3.3 (The principle of a single big jump)**.** *Suppose X_i are i.i.d. with mean μ and follow a regularly varying distribution with index $-\alpha$, where $\alpha > 1$. Then for $a > \mu$, the random walk $S_n = X_1 + \ldots + X_n$ satisfies*

$$\lim_{n\to\infty} \frac{Pr\left(S_n > an\right)}{nPr\left(X_1 > (a-\mu)n\right)} = 1. \tag{3.9}$$

Proof. We prove this result by constructing matching upper and lower bounds. Each of the bounds provides important intuition for the result. In particular, we prove the lower bound by constructing a simple event that is sufficient to cause the event $\{S_n > an\}$ to happen, and at the same time has a probability large enough to match the desired asymptotic behavior. This event is exactly the one provided by the intuition behind the catastrophe principle – that a single big jump occurred. In contrast, the upper bound shows that, without a big jump, the random walk could not have been so large with such a high probability.

Throughout, without loss of generality we let $\mu = 0$ and $a > 0$. [7]

Lower bound: Our first goal is to show that an asymptotic lower bound holds, i.e., to show that

$$\liminf_{n\to\infty} \frac{\Pr\left(S_n > an\right)}{n\Pr\left(X_1 > an\right)} \geq 1. \tag{3.10}$$

To do this, we formalize the event that the random walk is large because of a single big jump. This is one way a large deviation could have occurred and so it provides a lower bound on the probability. Formally, pick an auxiliary constant $\delta > a$ and observe that

$$\Pr\left(S_n > an\right) \geq \Pr\left(\cup_{i=1}^n B_i^n\right), \tag{3.11}$$

with $B_i^n = \{X_i > \delta n, \sum_{j=1, j\neq i}^n X_j > (a-\delta)n\}$. Observe that $\Pr\left(B_i^n\right)$ is constant in i for fixed n. Thus,

$$\begin{aligned}
\Pr\left(S_n > an\right) &\geq \Pr\left(\cup_{i=1}^n B_i^n\right) \\
&\geq \sum_{i=1}^n \Pr\left(B_i^n\right) - \sum_{i,j:i\neq j} \Pr\left(B_i^n \cap B_j^n\right) \\
&= n\Pr\left(B_1^n\right) - \frac{n(n-1)}{2}\Pr\left(B_1^n \cap B_2^n\right).
\end{aligned}$$

Since

$$\Pr\left(B_1^n \cap B_2^n\right) \leq P(X_1 > \delta n; X_2 > \delta n) = P(X_1 > \delta n)^2,$$

we see that, by decomposing $\Pr\left(B_1\right)$,

$$\Pr\left(S_n > an\right) \geq n\Pr\left(X_1 > \delta n\right)\Pr\left(C_1^n\right) - \frac{n(n-1)}{2}\Pr\left(X_1 > \delta n\right)^2, \tag{3.12}$$

with $C_1^n = \{\sum_{j=2}^n X_j > (a-\delta)n\}$. Since $a < \delta$,

$$\Pr\left(C_1^n\right) \to 1$$

[7] Otherwise, replace X_i with $X_i - \mu, i \geq 1$ and a with $a - \mu$.

by the weak law of large numbers. In addition, we see that

$$\lim_{n\to\infty} \frac{n(n-1)\Pr(X_1 > \delta n)^2}{n\Pr(X_1 > an)} = 0, \tag{3.13}$$

since the numerator is regularly varying with index $2 - 2\alpha$ and the denominator is regularly varying with index $1 - \alpha$. Combining the last three displays we obtain

$$\liminf_{n\to\infty} \frac{\Pr(S_n > an)}{n\Pr(X_1 > an)} \geq \liminf_{n\to\infty} \frac{n\Pr(X_1 > \delta n)}{n\Pr(X_1 > an)} = (\delta/a)^{-\alpha}.$$

As this conclusion holds for any $\delta > a$, (3.10) follows by having $\delta \downarrow a$.

Upper bound: We now turn to the corresponding asymptotic upper bound, i.e., we aim to show that

$$\limsup_{n\to\infty} \frac{\Pr(S_n > an)}{n\Pr(X_1 > an)} \leq 1. \tag{3.14}$$

Let us denote $A_n = \{S_n > an\}$. To prove (3.14), we have to show that the event A_n is most likely caused by a single big jump. Our strategy will be to partition A_n into two events, one in which a single jump contributes significantly to the sum S_n, and another in which no single jump contributes significantly to S_n. The crux of the proof is then to show that the latter event has an asymptotically negligible probability relative to the former.

Let $\tau \in (0, a)$. We partition A_n as follows.

$$\begin{aligned}\Pr(A_n) &= \Pr\left(A_n, \max_i X_i > \tau n\right) + \Pr\left(A_n, \max_i X_i \leq \tau n\right) \\ &:= \Pr(A_{1,n}) + \Pr(A_{2,n}).\end{aligned}$$

We now deal with both terms separately.

The first term is easy to handle. Take $\delta \in (\tau, a)$ and observe that, since $\delta > \tau$,

$$\begin{aligned}\Pr(A_{1,n}) &= \Pr\left(A_n, \max_i X_i > \delta n\right) + \Pr\left(A_n, \tau n < \max_i X_i \leq \delta n\right) \\ &\leq \Pr\left(\cup_{i=1}^n \{X_i > \delta n\}\right) + \Pr\left(\cup_{i=1}^n \{X_i > \tau n, \sum_{j=1, j\neq i}^n X_j > (a-\delta)n\}\right) \\ &\leq n\Pr(X_1 > \delta n) + n\Pr(X_1 > \tau n)\Pr\left(\sum_{j=2}^n X_j > n(a-\delta)\right).\end{aligned}$$

By the weak law of large numbers we see that $\Pr\left(\sum_{j=2}^n X_j > n(a-\delta)\right) \to 0$. Since $\Pr(X_1 > \tau n)/\Pr(X_1 > an)$

stays bounded in n for any $\tau > 0$, we have

$$\limsup_{n\to\infty} \frac{\Pr(A_{1n})}{n\Pr(X_1 > an)} \leq (\delta/a)^{-\alpha}. \tag{3.15}$$

This can be made arbitrarily close to 1 by having δ approach a.

Thus, the proof of (3.14) is be complete once we show there exists a $\tau \in (0, a)$ such that

$$\lim_{n\to\infty} \frac{\Pr(A_{2,n})}{n\Pr(X_1 > an)} = 0. \tag{3.16}$$

Note that

$$\Pr(A_{2n}) \leq \Pr(S_n > an \mid X_i \leq \tau n,\ i \leq n).$$

Thus, it suffices to have an upper bound on the probability that the sum of n i.i.d. random variables exceeds a thresold, given that each random variable is bounded from above by a different threshold. Such a bound is provided by the following *concentration inequality*, which dates back to Prokhorov [161]: Let $Y_i, i \geq 1$ be an i.i.d. sequence such that $E[Y_i] = 0$ and $\Pr(Y_i \leq c) = 1$ for some $c \in (0, \infty)$. For any $t > 0$ and $n \geq 1$:

$$\Pr(Y_1 + \ldots + Y_n > t) \leq \left(\frac{ct}{nVar(Y_1)}\right)^{-t/2c} \tag{3.17}$$

Note that the random variables Y_i in the above inequality are bounded from above by c. We apply the concentration inequality to $\Pr(A_{2n})$ as follows. Let $X_i^*, i \geq 1$ be an i.i.d. sequence such that

$$\Pr(X_1^* \leq y) = \Pr(X_i \leq y \mid X_i \leq \tau n).$$

Observe that the random variables X_i^* are bounded from above by τn, and $\mu^* = E[X_1^*] \leq E[X_1] = 0$. Therefore,

$$\Pr(A_{2n}) \leq \Pr(S_n > an \mid X_i \leq \tau n i \leq n) = \Pr\left(\sum_{i=1}^{n} X_i^* > an\right).$$

We can now apply (3.17) with $Y_i = X_i - \mu^*$, $c = \tau n$, $t = an$ to obtain

$$\Pr\left(\sum_{i=1}^{n} X_i^* > an\right) \leq \Pr\left(\sum_{i=1}^{n} Y_i > an\right) \leq \left(\frac{\tau an}{Var(Y_1)}\right)^{-a/2\tau}. \tag{3.18}$$

To simplify bounding $Var(Y_1)$, we assume that $\alpha > 2$. In this case, $Var(Y_1) \leq \mathbb{E}\left[(X_1^*)^2\right] \leq \mathbb{E}\left[X_1^2\right] < \infty$. Thus, if we choose τ small enough to satisfy $a/2\tau > \alpha - 1$, we arrive at the desired conclusion (3.16), completing the proof.

The case $\alpha \in (1, 2]$ requires a more delicate analysis, since $Var(Y_1)$ can grow with n in this case; see [203] for the details. □

3.4.2 Cramér's theorem

Cramér's theorem strengthens and refines the conspiracy principle for random walks in a parallel way to how the principle of a single big jump strengthens and refines the catastrophe principle. As we have already highlighted, the conspiracy principle in Definition 3.2 is too general to provide a precise characterization. Cramér's provides a much more powerful characterization.

Like the principle of a single big jump, Cramér's theorem provides a bound on Pr $(S_n > an)$ though, of course, the bound has a very different form since it applies to light-tailed distributions instead of heavy-tailed distributions. Cramér's theorem is tightly connected to concentration inequalities, specifically the Chernoff bound. In fact, a crisp statement of Cramér's theorem is that it proves that the Chernoff bound is tight. Thus, to motivate the form of Cramér's theorem we must start with the Chernoff bound.

The Chernoff bound itself is best viewed in the context of another fundamental result – Markov's inequality. Markov's inequality is the simplest, and most fundamental concentration inequality. It provides a bound on the tail of a probability distribution in terms of the mean of the distribution. Specifically, Markov's inequality states that

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a},$$

for any nonnegative random variable X and constant $a > 0$. Though simple, Markov's inequality is extremely powerful and is the building block for many other more sophisticated concentration inequalities, such as the Chernoff bound. In fact, the Chernoff bound is a simple extension of Markov's inequality that focuses on obtaining a tighter bound by applying Markov's inequality to e^{sX} instead of X. In the case of the random walk S_n this gives

$$\Pr(S_n \geq an) = \Pr\left(e^{sS_n} \geq e^{ans}\right) \leq \frac{\mathbb{E}\left[e^{sS_n}\right]}{e^{ans}}.$$

Since $S_n = X_1 + \ldots + X_n$, this can be expanded further to yield

$$\Pr(S_n \geq an) \leq e^{-san}\mathbb{E}\left[e^{sX_1} \cdots e^{sX_n}\right] = e^{-ans}\left(\mathbb{E}\left[e^{sX_i}\right]\right)^n.$$

Moving the expectation into the exponent then yields

$$\Pr(S_n \geq an) \leq e^{-n(as - \log \mathbb{E}[e^{sX_i}])}.$$

Finally, since the bound holds for all $s \geq 0$, we can optimize over s. This gives a Chernoff bound for S_n:

$$\Pr(S_n \geq an) \leq e^{-n \sup_{s\geq 0}\left(as - \log \mathbb{E}[e^{sX_i}]\right)}. \tag{3.19}$$

The key to understanding why this Chernoff bound is so much more powerful than Markov's inequality is to look at what information about the distribution is considered by each. Markov's inequality uses only the mean to determine a bound, but the transformation to the Chernoff bound brings the whole of the moment generating function into the bound. This means that the $\sup$ in (3.19) chooses an optimal bound that depends on the whole distribution; thus yielding a much tighter bound on the tail.

Chernoff's bound has proven to be a powerful tool across a wide variety of applications. However, it only provides insight into the likelihood of rare events, it provides no information about what the rare events "look" like. This is because Chernoff's bound provides only an *upper bound* on $\Pr(S_n \geq an)$. This is enough if all that you want to do is show that rare events are unlikely; but often it is also important to understand the cause of rare events. To get insight what leads to the rare events requires understanding the events that lead to a tight *lower bound*, as in the case of the principle of a single big jump.

Cramér's theorem provides this insight or, more specifically, the proof of Cramér's theorem does. Specifically, Cramér's theorem shows that the Chernoff bound is tight and the proof provides insight into what events led to the rare event. The intuition is crisp and serves as a refinement of the conspiracy principle in Definition 3.2. In particular, the proof highlights that during rare events, it is as if each X_i is sampled i.i.d. from a slightly different distribution – a "twisted" distribution – which has mean slightly larger than that of the original distribution, just large enough to make the previously rare event become likely. Thus, the X_i truly conspired together to create the rare event.

Theorem 3.4 (Cramér's theorem)**.** *Consider $S_n = X_1 + \ldots + X_n$, where X_i are i.i.d. and light-tailed. Then for $a > \mathbb{E}[X_1]$,*

$$\lim_{n\to\infty} \frac{-\log Pr(S_n > an)}{n} = \sup_{s>0}[as - \log \mathbb{E}\left[e^{sX_i}\right]]. \tag{3.20}$$

Proof. Chernoff's bound already implies the upper bound on the tail in the theorem, i.e., that

$$\liminf_{n\to\infty} \frac{-\log \Pr(S_n > an)}{n} \geq \sup_{s>0}[as - \log \mathbb{E}\left[e^{sX_1}\right]].$$

In fact, Chernoff's bound implies that this inequality holds for every finite n.

Thus, we can focus on the lower bound on the tail. We assume, for convenience, that X_i has a density f. Additionally, to eliminate some technical complexities, we make the simplifying assumption that the moment generating function $\mathbb{E}\left[e^{sX_1}\right]$ is "steep". Specifically, setting $\bar{s} = \sup\{s : \mathbb{E}\left[e^{sX_1}\right] < \infty\}$, we assume that $\lim_{s\uparrow\bar{s}} \mathbb{E}\left[e^{sX_1}\right] = \infty$. Under this assumption the optimization problem $\sup_{s>0}[as - \log \mathbb{E}\left[e^{sX_1}\right]]$ has a unique solution s_a for which $s_a < \bar{s}$. This means that s_a solves the first order optimality condition

$$a = \frac{\mathbb{E}\left[X_1 e^{s_a X_1}\right]}{\mathbb{E}\left[e^{s_a X_1}\right]}$$

The key to the proof is the "twisted" distribution with density function $\tilde{f}(x_i) = \frac{e^{s_a x_i} f(x_i)}{\mathbb{E}\left[e^{s_a X_1}\right]}$, which characterizes the distribution from which X_i are sampled in the rare event. This twisted distribution, which stochastically dominates the distribution of X_i, can be shown to have mean a. To see this, let $\tilde{X}$ denote a random variable with the twisted density $\tilde{f}(\cdot)$. Note that $\mathbb{E}\left[e^{s\tilde{X}}\right] = \mathbb{E}\left[e^{(s_a+s)X_1}\right] / \mathbb{E}\left[e^{s_a X_1}\right]$. This implies that

$$\mathbb{E}\left[\tilde{X}\right] = \frac{d}{ds}\mathbb{E}\left[e^{s\tilde{X}_i}\right]|_{s=0} = \frac{\mathbb{E}\left[X_1 e^{s_a X_1}\right]}{\mathbb{E}\left[e^{s_a X_1}\right]} = a.$$

Let $\tilde{X}_i$ denote an i.i.d. sequence of random variable following the twisted distribution. We bound the desired probability as follows. To begin, note that

$$\Pr\left(S_n > an\right) = \int_{(x_i):x_1+\ldots+x_n>an} \prod_{i=1}^{n} [f(x_i)dx_i].$$

The main step is to rewrite the above integral in terms of the twisted densities as follows.

$$\begin{aligned}
\Pr\left(S_n > an\right) &= \int_{(x_i):\ x_1+\ldots+x_n>an} e^{-s_a \sum_i x_i} \prod_{i=1}^{n} [e^{s_a x_i} f(x_i)dx_i] \\
&= \mathbb{E}\left[e^{s_a X_1}\right]^n \int_{(x_i):\ x_1+\ldots+x_n>an} e^{-s_a \sum_i x_i} \prod_{i=1}^{n} \left[\tilde{f}(x_i)dx_i\right] \\
&\geq \mathbb{E}\left[e^{s_a X_1}\right]^n \int_{(x_i):\ an+\sqrt{n}>x_1+\ldots+x_n>an} e^{-s_a \sum_i x_i} \prod_{i=1}^{n} \left[\tilde{f}(x_i)dx_i\right] \\
&\geq e^{-s_a(an+\sqrt{n})} \mathbb{E}\left[e^{s_a X_1}\right]^n \int_{(x_i):\ an+\sqrt{n}>x_1+\ldots+x_n>an} \prod_{i=1}^{n} \left[\tilde{f}(x_i)dx_i\right] \\
&= e^{-s_a(an+\sqrt{n})} \mathbb{E}\left[e^{s_a X_1}\right]^n \Pr\left(an < \tilde{X}_1 + \tilde{X}_2 + \cdots \tilde{X}_n < an + \sqrt{n}\right). \qquad (3.21)
\end{aligned}$$

Next, we take the log and look at the limit to complete the proof. In doing so, the first two terms remain, but the third term disappears. To show this, we apply central limit theorem to obtain

$$\Pr\left(an < \tilde{X}_1 + \tilde{X}_2 + \cdots \tilde{X}_n < an + \sqrt{n}\right) \to \Pr\left(0 < N < 1\right), \qquad (3.22)$$

where N is a Gaussian random variable with mean zero and variance equal to $\mathrm{Var}(\tilde{X}_i)$. Note that this application requires that $\mathrm{Var}(\tilde{X}_i)$ is finite. To verify this, note that since $s_a < \bar{s}$, the moment generating function of $\tilde{X}_i$ is finite in a neighborhood of the origin, which implies the finiteness of the variance of $\tilde{X}_i$.

Now, taking the log and looking at the limit of (3.21) gives

$$\begin{aligned}
\limsup_{n\to\infty} \frac{-\log \Pr\left(S_n > an\right)}{n} &\leq s_a a - \log \mathbb{E}\left[e^{s_a X_1}\right] \\
&= \sup_{s>0}[sa - \log \mathbb{E}\left[e^{s X_1}\right]],
\end{aligned}$$

which is the desired lower bound. □

3.5 Additional notes

This chapter has introduced the basics of the class of subexponential distributions, and their connection to the catastrophe and conspiracy principles. Our treatment only scratches the surface of these topics, and so we point the interested reader to more in depth treatments of each below.

Subexponential distributions: The class of subexponential distributions was introduced in Chistyakov [43] in 1964, and originally found applications to problems in branching processes [19]. Its significance for applications in insurance and queueing theory was recognized in [150, 183, 185]. Readers in search of more technical details about the class of properties and examples of modern applications can find excellent surveys in [64, 76, 88].

Subclasses of subexponential distributions have also proven interesting and useful. In particular, technical issues with the full class of subexponential distributions has led to many variations. For example, the fact that the class of subexponential distributions is not closed under addition (i.e. if X and Y are independent nonnegative subexponential random variables, then $X + Y$ may not be subexponential, see [122]) led to the important subclass termed S^*, introduced by [112]. A distribution function F of a non-negative random variable X belongs to S^* if it has finite mean μ and satisfies

$$\lim_{x\to\infty} \frac{\int_0^x \bar{F}(x-y)\bar{F}(y)dy}{\mu\bar{F}(x)} = 2. \tag{3.23}$$

Essentially, this entails that the density $\bar{F}/\mu$ can be seen as a subexponential distribution. This property implies that F, as well as $\int_0^x F(y)dy/\mu$ are subexponential and that the class is closed under addition. Other subclasses of subexponential distributions, e.g., strong subexponential distributions and local subexponential distributions, have received attention too and are discussed in detail in [76].

Another important, related class of distributions is the class of *subexponentially concave* distributions, which are distributions for which $-\log \bar{F}$ is concave. The motivation for this class is the existence of powerful concentration inequalities, e.g., see [107] and the survey of Sergey Nagaev [142]. Note that this class of random variables is particular useful when extending Theorems 3.1 and 3.2 from the regularly varying case towards more general subexponential distributions. The most general properties for Theorem 3.2 can be found in [54, 138]. These references also have discussions on extensions of square-root insensitivity and its connection with extreme value theory, which is connected to the material we discuss in Chapter 7.

Stronger conspiracy and catastrophe principles: The bulk of the chapter focused on simple, general versions of conspiracy and catastrophe principles. We gave examples of stronger versions of these principles in Section 3.4 (Cramér's Theorem and the principle of a single big jump); however there are many other such generalizations.

In 1938, Cramer [48] developed the result that, eventually, was named after him. The initial work was motivated by a problem in insurance however, after that, versions of his theorem found applications in a wide variety of fields, such as statistics, information theory, and communication networks. For an overview of these applications, as well as extensions of Cramér's theorem, we refer interested readers to the literature on large deviations. A concise introduction to both theory and applications of large deviations is [53]. Monographs that are focused on applications in communication and computer networks are [35, 83, 173]. Other valuable resources with more technical presentations are [52, 56, 61, 71].

Though the phrase itself has become popular relatively recently, the principle of a single big jump appeared first in [43, 141]. Variations of the principle are many and varied, and can be found in [17, 20, 74, 77, 199, 202]. These are often motivated by various applications in insurance mathematics and queueing theory as well as applications to communication networks [75, 99, 205].

One interesting variation that we use later in this book is the following. In the version we presented here (Theorem 3.3) it was shown that the limit

$$\frac{\Pr(S_n > t)}{n\Pr(X_1 > t)} \to 1, \text{ as } t \to \infty \tag{3.24}$$

holds if n is growing proportional with t. The result can actually be extended to cases where n is growing faster than t. To be precise, it can be shown (see [138]), for $\alpha > 2$ that there exists a constant $C > 0$ such that

$$\lim_{n\to\infty} \sup_{t:t>C\sqrt{n\log n}} |\frac{\Pr(S_n > t)}{n\Pr(X_1 > t)} - 1| = 0. \tag{3.25}$$

Another important set of variations are extensions to cases where there is more than one big jump. In particular, in some cases it is likely that there are multiple big jumps and recent research has focused on developing a generic mathematical framework to handle such extensions, e.g., [126, 166].

3.6 Exercises

1. Prove that regularly varying random variables are subexponential. Specifically, in Chapter 2, we proved that for i.i.d. regularly varying random variables X_1 and X_2, $\Pr(X_1 + X_2 > t) \sim \Pr(\max(X_1, X_2) > t)$ (see Lemma 2.4). Your task is to extend the same argument to n random variables, i.e., prove that

   $$\Pr(X_1 + X_2 + \cdots + X_n > t) \sim \Pr(\max(X_1, X_2, \cdots, X_n) > t),$$

   where $n \geq 2$, and $X_1, X_2, \cdots, X_n$ are i.i.d., regularly varying random variables.

2. Prove that the LogNormal distribution is subexponential.

   Hint: One approach is to use Lemma 3.3 in conjunction with the statement of Exercise 12 below. You may also find the following asymptotic approximation of the tail of the standard Gaussian useful: For a standard Gaussian random variable N,

   $$Pr(N > t) \sim \frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t} \quad \text{as } t \to \infty. \tag{3.26}$$

3. Consider a distribution F with support $[0, \infty)$ satisfying

   $$\bar{F}(x) \sim e^{-cx/log(x)} \quad \text{as } x \to \infty.$$

   Prove that F is subexponential.

4. Prove that the exponential distribution satisfies the conspiracy principle in Definition 3.2.

5. Prove that the light-tailed Weibull distribution (with shape parameter $\alpha > 1$) satisfies the conspiracy principle in Definition 3.2.

   Hint: One approach is to develop a simple lower bound on $Pr(X_1 + X_2 + \cdots + X_n > t)$, where $X_1, X_2, \cdots, X_n$ are i.i.d. Weibull with $\alpha > 1$.

6. Prove that the Gaussian distribution satisfies the conspiracy principle in Definition 3.2.

   Hint: You may find the asymptotic approximation (3.26) *of the tail of the standard Gaussian useful.*

7. Prove the following stronger conspiracy principle for the standard Gaussian (similar to the result of Proposition 3.1 for the light-tailed Weibull). Given i.i.d. standard Gaussian random variables X_1 and X_2, and $\delta \in (1/2, 1)$, prove that

$$\Pr\left(X_1 + X_2 > t, X_1 > \delta t\right) = o(\Pr\left(X_1 + X_2 > t\right)) \quad \text{as } t \to \infty.$$

8. Prove that the exponential distribution does not satisfy the stronger conspiracy principle in the statement of Proposition 3.1. Specifically, prove that for $X_1,\ X_2$ are i.i.d. exponential random variables and $\delta \in (0, 1)$,

$$\lim_{t\to\infty} \Pr\left(X_1 > \delta t \mid X_1 + X_2 > t\right) = 1 - \delta.$$

9. A distribution with support over $[0, \infty)$ is said to be *square-root insensitive* if

$$\bar{F}(x) \sim \bar{F}(x - \sqrt{x}) \quad \text{as } x \to \infty.$$

Prove that the following distributions are square-root insensitive:

a regularly varying distributions

b the LogNormal distribution

c the Weibull distribution with shape parameter $\alpha < 1/2$.

10. Let $X = \min(Y, Z)$, where Y and Z are independent, $Y \sim$ Exponential(μ), and $Z \sim$ Pareto(x_m, α) with $\alpha > 1$.

a Prove that X is light-tailed.

b Prove that X does not satisfy the conspiracy principle in Definition 3.2. Specifically, prove that

$$\limsup_{x\to\infty} \frac{\bar{F}^{2*}(x)}{\bar{F}(x)} < \infty,$$

where F denotes the distribution of X.

11. *** Your task in this exercise is to prove a non-asymptotic bound on the ratio of $\bar{F}^{n*}(x)$ to $\bar{F}(x)$ for subexponential F called *Kesten's bound*. This bound is a technical result that is often useful in proofs; for example, it is used in the proof of Theorem 3.1.

Suppose that F is the distribution function corresponding to a non-negative subexponential random variable. Prove that for any $\epsilon > 0$, there exists $K > 0$ such that for any $n \geq 2$ and $x \geq 0$,

$$\frac{\bar{F}^{*n}(x)}{\bar{F}(x)} \leq K(1+\epsilon)^n.$$

Hint: Let $\alpha_n := \sup_{x\geq 0} \frac{\bar{F}^{*n}(x)}{\bar{F}(x)}$. *Prove that for large enough* T,

$$\alpha_{n+1} \leq 1 + \frac{1}{\bar{F}(T)} + \alpha_n(1+\epsilon).$$

From this, it follows inductively that

$$\alpha_n \leq \frac{1}{\epsilon}\left(1 + \frac{1}{\bar{F}(T)}\right)(1+\epsilon)^n,$$

which completes the proof taking $K = \frac{1}{\epsilon}\left(1 + \frac{1}{\bar{F}(T)}\right)$.

12. *** Prove that the class of subexponential distributions is closed under tail equivalence. Specifically, prove that if the distribution F is subexponential, and the distribution G satisfies

$$\bar{G}(x) \sim C\bar{F}(x)$$

for some $C > 0$ (i.e., G is tail-equivalent to F), then G is also subexponential.

Hint: Let Y_1 and Y_2 denote two i.i.d. random variables with distribution G. For a constant, but large enough v, partition the event $\{Y_1 + Y_2 > x\}$ as,

$$\{Y_1 + Y_2 > x\} = \{Y_1 \leq v,\ Y_1 + Y_2 > x\} \cup \{Y_2 \leq v,\ Y_1 + Y_2 > x\} \\ \cup \{v < Y_2 \leq x - v,\ Y_2 + Y_2 > x\} \cup \{Y_1 > v,\ Y_2 > x - v\}$$

so that

$$\begin{aligned}\frac{\bar{G}^{*2}(x)}{\bar{G}}(x) &= \frac{2}{\bar{G}(x)}\int_0^v \bar{G}(x-t)dG(t) + \frac{1}{\bar{G}(x)}\int_v^{x-v} \bar{G}(x-t)dG(t) + \frac{\bar{G}(v)\bar{G}(x-v)}{\bar{G}(x)} \\ &=: T_1 + T_2 + T_3.\end{aligned}$$

Use tail equivalence and the fact that $\bar{F}(x) \sim \bar{F}(x-y)$ for constant y (see the proof of Lemma 3.4) to argue that

$$\lim_{x\to\infty} T_1 = 2G(v), \quad \lim_{x\to\infty} T_3 = \bar{G}(v).$$

Next, and this is the tricky part of the proof, use tail equivalence and the subexponentiality of F to show that $\limsup_{x\to\infty} T_2$ can be made arbitrarily small for large enough v.

13. *** Prove the equivalence of statements (i) and (ii) of Lemma 3.2. Specifically, let $\{X_i\}_{i\geq 1}$ denote a sequence of i.i.d. subexponential random variables. For $n \geq 2$, recall that the defining property of subexponential distributions is

$$\Pr(X_1 + X_2 + \cdots + X_n > t) \sim n\Pr(X_1 > t) \qquad \text{as } t \to \infty.$$

Prove that, if the above property holds for $n = 2$, then it holds for all $n \geq 2$.

Hint: Of course, this result is proved inductively. Note that it is easy to see that

$$\liminf_{t\to\infty} \frac{Pr(X_1 + X_2 + \cdots + X_n > t)}{Pr(X_1 > t)} \geq \liminf_{t\to\infty} \frac{Pr(\max(X_1, X_2, \cdots, X_n > t)}{Pr(X_1 > t)} = n.$$

The challenge in the proof is therefore to show the upper bound, i.e.,

$$\limsup_{t\to\infty} \frac{\boldsymbol{Pr}\left(X_1 + X_2 + \cdots + X_n > t\right)}{\boldsymbol{Pr}\left(X_1 > t\right)} \leq n.$$

Chapter 4

Residual lives, hazard rates, and long tails

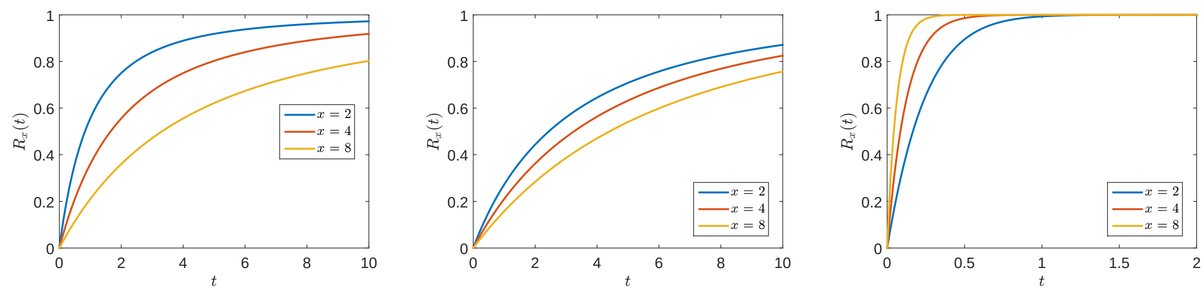
Over the course of our days we spend a lot of our time waiting for things – we wait for a table at restaurants, we wait for a subway train or a bus to show up, we wait for people to respond to our emails, etc. In such scenarios, we hold on to the belief that, as we wait, the likely amount of remaining time we will need to wait is getting smaller. For example, we believe that, if we have waited ten minutes for a table at a restaurant, the expected time we have left to wait should be smaller than it was when we arrived and that, if we have waited five minutes for the subway, then our expected remaining wait time should be less than it was when we arrived.

In many cases this belief holds true. For example, as other diners finish eating, our expected waiting time for a table at a restaurant drops. Similarly, subway trains follow a schedule with (nearly) deterministic gaps between trains and thus, as long as the train is on schedule, our expected remaining waiting time decreases as we wait. However, a startling aspect of heavy-tailed distributions is that this is not always true. For example, if you have been waiting for a long time after the scheduled arrival time for a subway train, then it is very likely that there was some failure and the train may take an extremely long time to arrive, and so your expected remaining waiting time has actually increased while you waited. Similarly, if you are waiting for a response to an email and have not heard for a few days, it is likely to be a very long time until a response comes (if it ever does).

The above examples highlight another fundamental distinction between light-tailed distributions (e.g. restaurant waiting times) and heavy-tailed distributions (e.g. email waiting times). To make the contrast even clearer, we can illustrate the same distinction using our classic examples of heavy-tailed and light-tailed distributions: incomes and heights. If we know someone is taller than 6 feet tall, then it is most likely that they are only a few inches taller; but if we know someone has more than \$1 million, then it is much more likely that they are multi-millionaires than that they are just barely millionaires.

All these examples highlight that, as with scale-invariance and the catastrophe principle, the behavior we expect to see is aligned with what happens under light-tailed distributions and so, upon first encounter, the behavior of heavy-tailed distributions is mysterious. We expect that if we have waited a long time the remaining waiting time, i.e., the *residual life*, should have decreased, and so it is particularly jarring that under heavy-tailed distributions the residual life will likely have increased dramatically.

In this chapter we explore the residual life of heavy-tailed distributions in order to build intuition for the counterintuitive phenomena described above. To do this, we start by exploring the distribution of residual life via two common measures: the hazard rate function and the mean residual life function. We then



(a) Pareto distribution: $x_0 = 1, \alpha = 2$ (b) heavy-tailed Weibull: $\beta = 1, \alpha = 1/2$ (c) light-tailed Weibull: $\beta = 1, \alpha = 2$

Figure 4.1: *Illustration of the residual life distribution for different choices of F. Note that when F is heavy-tailed (see (a) and (b)), the residual life distribution 'grows' stochastically with increasing age x; whereas the residual life distribution 'shrinks' stochastically with increasing x in the light-tailed example (see (c)).*

study the relationship between heavy-tailed distributions and properties of the hazard rate and the mean residual life, which leads us to the formalization of a subclass of heavy-tailed distributions, termed *long-tailed distributions*, that we explore in depth.

4.1 Residual lives and hazard rates

The foundation of this chapter is the concept of the residual life of a distribution. The term "residual life" refers to the remaining waiting time given that you have already been waiting for some amount of time. Clearly, the residual life crucially depends on how long you have waited already; and so it is a conditional concept. Formally, we define the residual life distribution as follows.

Definition 4.1. *For a nonnegative random variable X with distribution function F, the residual life distribution $R_x(t)$ is defined such that*

$$R_x(t) = 1 - Pr\,(X > x + t | X > x) = 1 - \frac{\bar{F}(x+t)}{\bar{F}(x)} \qquad (\bar{F}(x) > 0,\ t \geq 0).$$

The complementary residual life distribution $\bar{R}_x(t)$ is defined as $\bar{R}_x(t) = 1 - R_x(t)$.

The residual life $R_x(t)$ is the distribution of the waiting time given that you have already waited for x time, or in different terminology this size of the *excess* beyond the threshold t, given that the random variable exceeds the threshold. The residual life distribution is a foundational concept that is found in widely varying applications, ranging from the insurance assessment and reliability theory to the social sciences, where it has found use in studying the lifetimes of everything from the length of wars to human life expectancies. In this book, we make use of properties of the residual life distribution when studying the emergence of heavy-tailed distributions under multiplicative and extremal processes in Chapters 6 and 7; and also in the design of statistical tools for estimating heavy-tailed phenomena in chapter 9.

To begin to get a feel for the residual life distribution, let us consider some examples. Conveniently, it tends to be quite straightforward to calculate $\bar{R}_x(t)$ for common distributions. For example, $\bar{R}_x(t)$ for the Pareto distribution can be computed as follows

$$\text{Pareto: } \bar{R}_x(t) = \frac{\left(\frac{x_m}{x+t}\right)^\alpha}{\left(\frac{x_m}{x}\right)^\alpha} = \left(1 + \frac{t}{x}\right)^{-\alpha}. \tag{4.1}$$

This is shown in Figure 4.1. Interestingly, for any x, the residual life distribution under the Pareto distribution, $\bar{R}_x(t)$, follows a Burr distribution. Thus, the residual life distribution has a regularly varying tail.

Similarly, the residual life distribution of the exponential can be calculated easily.

$$\text{Exponential: } \bar{R}_x(t) = \frac{e^{-\mu(x+t)}}{e^{-\mu x}} = e^{-\mu t} \tag{4.2}$$

A striking aspect of the above is that $\bar{R}_x(t) = \bar{F}(t)$. This is a restatement of the "memoryless" property of the exponential distribution, which says that regardless of how long you have waited so far, the distribution of the remaining time you have to wait is exactly the same as if you just arrived. This is a particularly special property of exponential distributions, which are the only continuous distributions that are memoryless (see Exercise 1).

Finally, let us consider the residual life of the Weibull distribution. Again, it is simple to calculate.

$$\text{Weibull: } \bar{R}_x(t) = \frac{e^{-\beta(x+t)^\alpha}}{e^{-\beta x^\alpha}} = e^{-\beta[(x+t)^\alpha - x^\alpha]} \tag{4.3}$$

However, though it is simple to calculate $\bar{R}_x(t)$ for the Weibull, the resulting form is not particularly informative. It is hard to get much insight from it directly. However, Figure 4.1 highlights the shape of the residual life distribution for both light-tailed and heavy-tailed Weibull distributions.

More generally, despite the fact that the residual life distribution is easy to derive for many distributions, it is often useful (or even necessary) to look at statistics of the residual life distribution in order to obtain insight into its behavior. There are two statistics that are most commonly used: the *mean residual life* and the *hazard rate*. These are the focus of the next two sections.

4.1.1 The mean residual life

Whenever we consider a distribution, it is natural to use its mean in order to obtain insight. In this case, because residual life is a conditional concept that depends on the how long you have waited so far, the mean of the residual life distribution is a function of the time you have waited. More specifically, the mean residual life function is defined as follows.

Definition 4.2. *For a nonnegative random variable X with distribution function F, define the mean residual life (MRL) function $m(x) = \mathbb{E}[X - x \mid X > x]$ (for $x \geq 0$ satisfying $\bar{F}(x) > 0$). Equivalently*

$$m(x) = \int_0^\infty \bar{R}_x(t)dt = \int_0^\infty \frac{\bar{F}(x+t)}{\bar{F}(x)}dt. \tag{4.4}$$

Note that the MRL function $m(x)$ is well defined over the interval $\{x \geq 0 : \bar{F}(x) > 0.\}$ Over this interval, the MRL is bounded if and only if the distribution has finite mean.

The MRL function $m(x)$ appears far less frequently than the density function $f(x)$, the distribution function $F(x)$, or moment generating function $M(s)$; however $m(x)$ also completely determines the distribution when the distribution has a finite mean. Thus, for example, it is possible to "invert" $m(x)$ to calculate the distribution function $F(x)$ [92]. Further, as we show in Chapter 9, the mean residual life also can be useful in statistical exploration of heavy-tailed phenomena.

It is typically not hard to compute $m(x)$ for common distributions. To illustrate the behavior of $m(x)$ let us return to the examples of the Pareto, Exponential, and Weibull distributions. For the Pareto distribution, it is quite straightforward to calculate $m(x)$. In particular, assuming that the mean is finite ($\alpha > 1$), we have

$$\text{Pareto: } m(x) = \int_0^\infty \bar{R}_x(t)dt = \int_0^\infty \left(1 + \frac{t}{x}\right)^{-\alpha} dt = \frac{x}{\alpha - 1}. \tag{4.5}$$

Interestingly, from the above we see that the mean residual life of the Pareto distribution is increasing, and grows unboundedly with x. In particular, under a Pareto distribution, the expected remaining waiting time grows linearly with the amount of time you have waited so far.

The calculation of the mean residual life is also straightforward for the Exponential distribution:

$$\text{Exponential: } m(x) = \int_0^\infty \bar{R}_x(t)dt = \frac{1}{\mu}\int_0^\infty \mu e^{-\mu t} dt = \frac{1}{\mu}.$$

This derivation highlights a consequence of the memoryless property of Exponential distributions – the mean residual life is constant with respect to x, specifically $m(x) = 1/\mu = \mathbb{E}[X]$. That is, the expected remaining waiting time is the same as when you first arrived, regardless of how long you have waited.

Though it is straightforward to calculate the mean residual life function $m(x)$ for the Pareto and the Exponential, it is not always so easy. In fact, it is difficult to derive an explicit formula for $m(x)$ for the Weibull distribution. However, we can numerically compute the mean residual life; see Figures 4.2(a)–(c). The figure highlights that when $\alpha < 1$ the mean residual life is increasing, and thus one should expect the remaining waiting time to grow as you wait longer, but that the opposite is true when $\alpha > 1$. In that case, the mean residual life is decreasing, which means the expected remaining waiting time decreases as you wait.

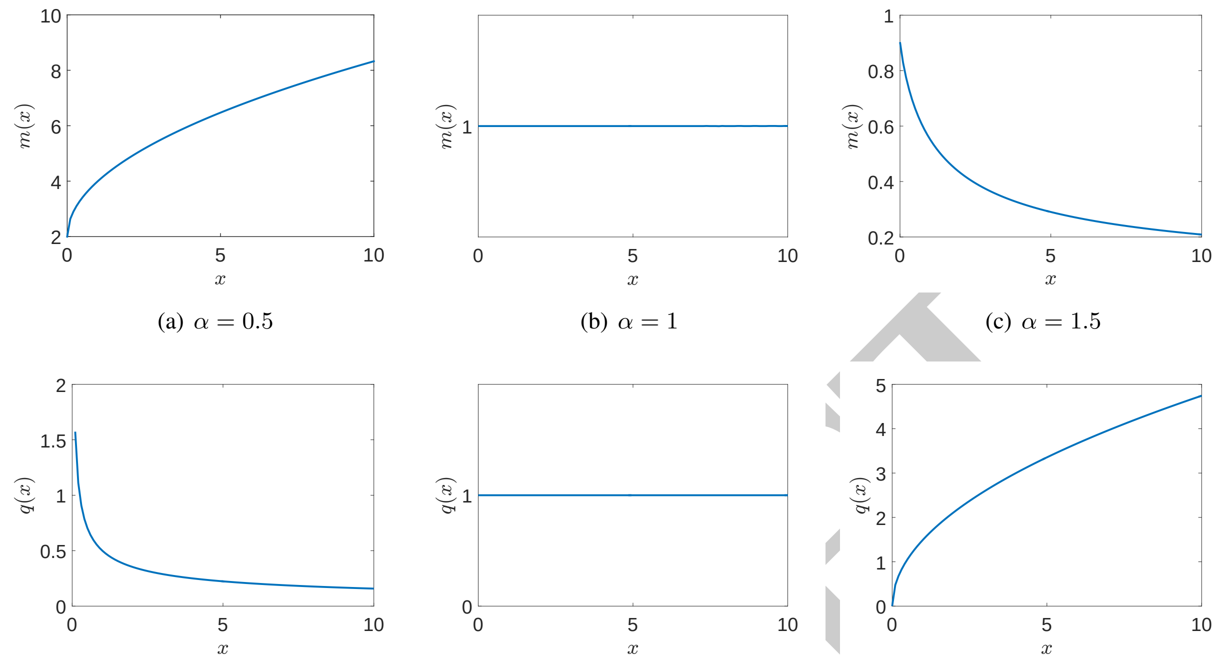
4.1.2 The hazard rate

A second important statistic of the residual life distribution is the hazard rate. We have already mentioned the hazard rate a few times in the book, but here we introduce it formally and study it in detail, since it is fundamentally related to the residual life distribution.

The residual life distribution $\bar{R}_x(t)$ looks at the remaining waiting time given that you have already waited for a certain amount of time x, while the hazard rate can be thought of as the likelihood of the wait time ending now, given that you have waited x time already. Formally, the hazard rate is defined as follows.

Definition 4.3. *For a nonnegative random variable X with distribution function F and density function f, define the hazard rate, a.k.a., the failure rate, as $q(x) = f(x)/\bar{F}(x)$ (for $x \geq 0$ satisfying $\bar{F}(x) > 0$).*[1]

[1] Unlike the MRL, the hazard rate is only defined for continuous random variables. Moreover, the domain of the hazard rate function is the same as that of the residual life distribution and the MRL, i.e., $\{x \geq 0 : \bar{F}(x) > 0\}$.



(a) $\alpha = 0.5$ (b) $\alpha = 1$ (c) $\alpha = 1.5$

(d) $\alpha = 0.5$ (e) $\alpha = 1$ (f) $\alpha = 1.5$

Figure 4.2: *Mean residual life and hazard rate of the Weibull distribution with $\beta = 1$ and different values of α. Recall that $\alpha = 1$ corresponds to the Exponential distribution.*

Further, define the cumulative hazard as $Q(y) = \int_0^y q(x)dx$.

It is easy to see that the hazard rate and the residual life are intimately related. The hazard rate is the density of the residual life distribution evaluated at zero.

$$R'_x(0) = \frac{d}{dt}\left(1 - \frac{\bar{F}(x+t)}{\bar{F}(x)}\right)|_{t=0} = \frac{f(x+t)}{\bar{F}(x)}|_{t=0} = \frac{f(x)}{\bar{F}(x)} = q(x)$$

Further, the hazard rate is intrinsically tied to the tail of the distribution. To see this, note that

$$q(t) = \frac{f(t)}{\bar{F}(t)} = -\frac{d}{dt}\log \bar{F}(t),$$

and so

$$Q(x) = -\log \bar{F}(x),$$

which gives

$$\bar{F}(x) = e^{-Q(x)} = e^{-\int_0^x q(t)dt}. \tag{4.6}$$

As a consequence of the above, it is easy to see that the hazard rate and the mean residual life are also closely related. In particular, as long as both exist, we have

$$m(x) = \int_0^\infty e^{-\int_x^{x+t} q(y)dy} dt, \tag{4.7}$$

which highlights that if the hazard rate is monotonically decreasing (increasing) then the mean residual life will be monotonically increasing (decreasing). Further, it is possible to show (see Exercise 3) that

$$m'(x) = m(x)q(x) - 1. \tag{4.8}$$

The interested reader is referred to [92] and the references therein for more details.

To get a feeling for the behavior of the hazard rate, let us return again to our examples of the Pareto, Exponential, and Weibull distributions. Either by computing directly or by using the above relationships, it is straightforward to see that the hazard rates of the Pareto and the Exponential are as follows.

$$\text{Pareto: } q(t) = \frac{\frac{\alpha}{x_m}\left(\frac{x_m}{t}\right)^{\alpha+1}}{\left(\frac{x_m}{t}\right)^{\alpha}} = \frac{\alpha}{t},$$

$$\text{Exponential: } q(t) = \frac{\mu e^{-\mu t}}{e^{-\mu t}} = \mu.$$

Thus, the Pareto has a hazard rate that decreases to zero, while the Exponential has a constant hazard rate. This contrast is interesting: the memoryless property of the Exponential distribution means that the likelihood that your waiting time ends is unchanging as you wait, while under the Pareto distribution the likelihood that your waiting time ends decreases to zero as you wait.

It is also straightforward to compute the hazard rate of the Weibull distribution, which is in contrast to the difficulty of computing $m(x)$ in this case. In particular, the hazard rate of the Weibull is:

$$\text{Weibull: } q(t) = \frac{\alpha\beta t^{\alpha-1} e^{-\beta t^\alpha}}{e^{-\beta t^\alpha}} = \alpha\beta t^{\alpha-1}$$

The form of the hazard rate under the Weibull distribution highlights a similar contrast to what we saw between the Pareto and the Exponential, only more extreme (see Figures 4.2(d)–(f)). When $\alpha > 1$ the hazard rate is increasing (and thus the mean residual life is decreasing), which means that the likelihood your wait ends increases as you wait, while when $\alpha < 1$ the hazard rate is decreasing (and thus the mean residual life is increasing), similarly to that of the Pareto distribution. Of course, $\alpha = 1$ corresponds to the case of the Exponential distribution, and so the hazard rate is constant.

4.2 Heavy tails and residual lives

The simple examples of the Pareto, Exponential, and Weibull that we have used so far in the chapter illustrate the contrast between light-tailed and heavy-tailed distributions that we have discussed informally in the introduction to the chapter: if we have waited a long time, then under light-tailed Weibull distributions the expected remaining waiting time will have decreased, while under the heavy-tailed Weibull and Pareto distributions the expected remaining waiting time will have increased dramatically.

In particular, we have seen that under the light-tailed Weibull the mean residual life is decreasing and the hazard rate is increasing, while under the heavy-tailed Weibull and Pareto distributions the mean residual life is increasing unboundedly and the hazard rate is decreasing to zero. The fact that these three distributions all have monotonic hazard rates and mean residual lives points us toward the importance of this property, and in particular, motivates the definition of the following four classes of distributions.

Definition 4.4. *A nonnegative distribution F with mean residual life function m is said to have increasing/decreasing mean residual life (IMRL/DMRL) if $m(x)$ is increasing/decreasing in x for all x such that $\bar{F}(x) \in (0,1)$.*[2]

Definition 4.5. *A nonnegative distribution F with hazard rate q is said to have increasing/decreasing hazard rate (IHR/DHR) if $q(x)$ is increasing/decreasing in x for all x such that $F(x) \in (0,1)$.*[3]

Clearly, heavy-tailed Weibull and Pareto distributions are IMRL and DHR; while light-tailed Weibull distributions are DMRL and IHR. Given these examples and the relationship between the hazard rate and the mean residual life, one would expect a strong connection between the DMRL/IMRL and IHR/DHR, and this is indeed the case. In fact, it follows immediately from (4.7) that the IHR class is contained within the DMRL class and the DHR class is contained within the IMRL class.

Theorem 4.1. *All distributions with an increasing (decreasing) hazard rate have a decreasing (increasing) mean residual life, i.e., IHR⊆DMRL and DHR⊆IMRL.*

At this point it is natural to notice that, because the Exponential distribution has constant mean residual life and hazard rate, it is, in some sense, the boundary between the IHR and DHR classes and between the IMRL and DMRL class. Of course, Exponential distributions also serve as the boundary between light-tailed and heavy-tailed distributions, and so it is quite tempting to think of IMRL/DHR distributions has "heavy-tailed" and DMRL/IHR distributions as "light-tailed". In fact, the temptation is so strong that in some disciplines, IMRL is used as a defining property of "heavy-tailed" distributions. However, heavy-tailed and IMRL/DHR are actually quite different concepts, as are light-tailed and DMRL/IHR.

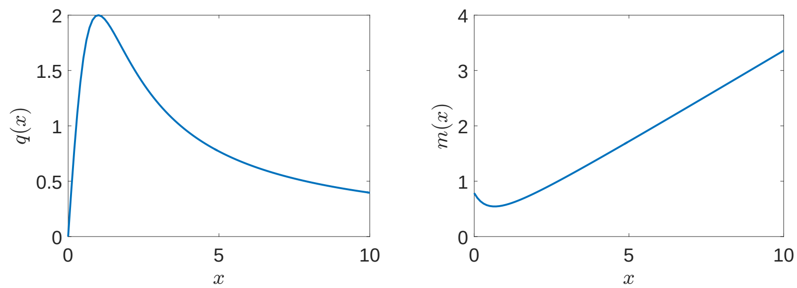
Specifically, it is easy to construct examples of IMRL and DHR distributions that are not heavy-tailed and it is easy to construct examples of heavy-tailed distributions that are not IMRL or DHR. For example, a heavy-tailed distribution that is not IMRL or DHR is the Burr distribution. Recall that the Burr distribution is defined by $\bar{F}(x) = (1 + \lambda x^c)^{-k}$ for $x \geq 0$ where λ, c, k are positive parameters. It is easy to check that the Burr is regularly varying (with index $-ck$), and thus heavy-tailed. A simple calculation shows that the hazard rate function for the Burr distribution is given by

$$q(x) = \frac{k\lambda c x^{c-1}}{1 + \lambda x^c}.$$

Clearly, q eventually decreases, with $\lim_{x\to\infty} q(x) = 0$. However, for $c > 1$, note that $q(0) = 0$. This means that when $c > 1$, the Burr distribution is not DHR. Next, we turn to the mean residual life. Note that the expectation of the Burr is finite if $ck > 1$. While it is difficult to come up with an explicit formula

[2]A random variable X with distribution F takes values in the interval $\{x : \bar{F}(x) < 1\}$ with probability 1. Moreover, that the mean residual life is defined for x such that $\bar{F}(x) > 0$. Thus, for defining whether or not a distribution is IMRL/DMRL, we check for monotonicity of $m(\cdot)$ over only over the interval $\{x : \bar{F}(x) \in (0,1)\}$.

[3]Note that a continuous random variable X with distribution F takes values in the interval $\{x : \bar{F}(x) \in (0,1)\}$ with probability 1. Thus, to define a distribution as IHR/DHR, we only check for monotonicity of $q(\cdot)$ over this interval.



(a) Hazard rate (b) Mean residual life

Figure 4.3: *Hazard rate and mean residual life for the Burr distribution, with $\lambda = 1$, $c = k = 2$. Note that the distribution, though heavy-tailed, is neither IMRL not DHR.*

for the mean residual life of the Burr distribution, we can invoke (4.8) to argue that $m'(0) = -1$ whenever $c > 1$. Thus, we may conclude that the Burr distribution is not IMRL for $c > 1$. The above highlights that, though $m(x)$ is eventually increasing and $q(x)$ is eventually decreasing for the Burr distribution, they are not monotonic over the entire support of the distribution; see Figure 4.3.

Similarly, it is easy to construct examples of light-tailed distributions that are not DMRL and IHR. The Hyperexponential distribution is one such example. The Hyperexponential distribution is a mixture of n exponential distributions where, with probability p_i, a sample is drawn from an Exponential distribution with rate μ_i, for $i = 1, \ldots, n$ with $\sum p_i = 1$. Note that the Hyperexponential distribution is defined by the c.c.d.f. $\bar{F}(x) = \sum_{i=1}^n p_i e^{-\mu_i x}$ for $x \geq 0$. The hazard rate is now easily seen to be, for $x \geq 0$,

$$q(x) = \frac{\sum_{i=1}^n p_i \mu_i e^{-\mu_i x}}{\sum_{i=1}^n p_i e^{-\mu_i x}}.$$

We now differentiate q to see that the Hyperexponential is DHR.

$$q'(x) = \frac{\left(\sum_{i=1}^n p_i \mu_i e^{-\mu_i x}\right)^2 - \left(\sum_{i=1}^n p_i e^{-\mu_i x}\right)\left(\sum_{i=1}^n p_i \mu_i^2 e^{-\mu_i x}\right)}{\left(\sum_{i=1}^n p_i e^{-\mu_i x}\right)^2} \leq 0,$$

where the inequality above is a consequence of the Cauchy-Schwartz inequality.[4]. Since the Hyperexponential is DHR, it follows from Theorem 4.1 that it is also IMRL.

4.3 Long-tailed distributions

The discussion in the previous section highlights that one must be careful in connecting "heavy-tailed" with the concepts of "increasing mean residual life" and "decreasing hazard rate". However, if we think again about the informal examples of waiting times that we discussed at the beginning of the chapter, it becomes clear that IMRL and DHR are too "precise" to capture the phenomena we were describing. For example, consider the case of waiting for a response to an email. It is not that we expect our average remaining

[4]The Cauchy-Schwartz inequality states that for real n-dimensional vectors x and y, $(\sum_{i=1}^n x_i y_i)^2 \leq (\sum_{i=1}^n x_i^2)(\sum_{i=1}^n y_i^2)$.

waiting time to monotonically increase as we wait. In fact, we are very likely to get a response quickly, and so the expected waiting time should drop initially (and the hazard rate should increase initially). It is only after we have waited a "long" time already, in this case a few days, that we expect to see a dramatic increase in our residual life. Further, in the extreme, if we have not received a response in a month, we can reasonably expect that we may never receive a response, and so the mean residual life is, in some sense, growing unboundedly, or equivalently the hazard rate is decreasing to zero. The example of waiting for a subway train highlights the same issues. Initially, we expect that the mean residual life should decrease, because if the train is on schedule, things are very predictable. However, once we have waited a long time beyond when the train was supposed to arrive, it likely means something went wrong, and could mean the train has had some sort of mechanical problem and will never arrive.

These examples illustrate two important aspects that need to be captured in a formalization of this phenomena. First, they highlight that strict monotonicity of the residual life is not crucial (or desirable), and that we should instead focus on the behavior of the tail. This is similar to the way we relaxed scale-invariance to asymptotic scale-invariance in Chapter 2. Second, they highlight that the phenomena we would like to capture includes the fact that the residual life distribution "blows up", in the sense that if we have waited a very long time we should expect to wait forever. Note that this property is true of heavy-tailed Weibull and Pareto distributions, but is not true of the light-tailed distributions that are a part of the IMRL and DHR classes. For example, the Hyperexponential distribution with rate parameters $\mu_1, \dots, \mu_n$ has a mean residual life that is upper bounded by $\max_i(1/\mu_i)$ (see Exercise 4).

These two observations lead us to the following definition of the class of long-tailed distributions.

Definition 4.6. *A distribution F over the nonnegative reals is said to be long-tailed, denoted by $F \in \mathcal{L}$, if*

$$\lim_{x\to\infty} \bar{R}_x(t) = \lim_{x\to\infty} \frac{\bar{F}(x+t)}{\bar{F}(x)} = 1 \tag{4.9}$$

for all $t > 0$, i.e., $\bar{F}(x+t) \sim \bar{F}(x)$ as $x \to \infty$. A non-negative random variable is said to be long-tailed if its distribution function is long-tailed.

The definition of long-tailed distributions exactly parallels our discussion above. In particular, long-tailed distributions are those where the distribution of residual life "blows up," i.e., for any finite t, the probability the residual life is larger than t goes to 1 as $x \to \infty$. Naturally, this leads to the immediate consequence that the mean residual life grows unboundedly. We leave the proof of this result to the reader (see Exercise 11).

Lemma 4.1. *Suppose that the distribution F is long-tailed. Then, $\lim_{x\to\infty} m(x) = \infty$.*[5]

The connection between long-tailed distributions and the asymptotic behavior of the mean residual life and the hazard rate can be made more precise if we restrict attention to continuous distributions that are "well-behaved"; specifically, distributions having the property that the limit of the hazard rate function $q(x)$ exists as $x \to \infty$. Note that this assumption is not limiting, and captures all non-pathological continuous distributions.

[5] Note that if the expectation corresponding to the distribution F is ∞, then $m(x) = \infty$ identically. In this case, we follow the convention that $\lim_{x\to\infty} m(x) = \infty$ holds.

Theorem 4.2. *Suppose that the distribution F over the non-negative reals is associated with a density function f. Assuming $\lim_{x\to\infty} q(x)$ exists,*

$$F \in \mathcal{L} \iff \lim_{x\to\infty} q(x) = 0 \iff \lim_{x\to\infty} m(x) = \infty.$$[5]

Proof. Note that we have made the assumption that the hazard rate $q(x)$ has a limit as $x \to \infty$. Let $c = \lim_{x\to\infty} q(x)$. Clearly, $c \geq 0$.

We first prove that $F \in \mathcal{L} \iff c = 0$. Recall that the c.c.d.f. can be represented in terms of the hazard rate as follows: $\bar{F}(x) = e^{-\int_0^x q(s)ds}$. It follows from the above representation that for $t \geq 0$,

$$\frac{\bar{F}(x+t)}{\bar{F}(x)} = e^{-\int_x^{x+t} q(s)ds}.$$

It is now not hard to see that

$$\lim_{x\to\infty} \frac{\bar{F}(x+t)}{\bar{F}(x)} = 1 \iff \lim_{x\to\infty} \int_x^{x+t} q(s)ds = 0 \iff c = 0.$$

Next, we show that $c = 0 \iff \lim_{x\to\infty} m(x) = \infty$. If $c = 0$, we have already proved that $F \in \mathcal{L}$, which implies, from Lemma 4.1, that $\lim_{x\to\infty} m(x) = \infty$. It thus suffices to show that if $c > 0$, then $m(x)$ is bounded. To see this, recall that

$$m(x) = \int_0^\infty e^{-\int_x^{x+t} q(s)ds} dt.$$

Assuming $c > 0$, there exists $x_0 > 0$ such that $q(x) \geq c/2$ for $x \geq x_0$. Therefore, for $x \geq x_0$,

$$m(x) \leq \int_0^\infty e^{-ct/2} dt = \frac{2}{c} < \infty.$$

This completes the proof. □

These two results highlight that the definition of long-tailed is both "weaker" than IMRL/DHR in that it focuses only on the tail, and "stronger" than IMRL/DHR in that it requires the mean residual life to grow unboundedly and the hazard rate to decrease to zero asymptotically.

Though the name "long-tailed" may initially seem strange, since it has no connection to the idea of residual life, it is actually natural to see from the definition why the name "long-tailed" is appropriate. In particular, $\bar{F}(x+t) \sim \bar{F}(x)$ as $x \to \infty$ for all t means that the tail stretches out with seemingly no decay over any finite range t. Thus, the tail of the distribution is indeed quite long. As this suggests, long-tailed distributions are a subclass of heavy-tailed distributions. In fact, if you recall, the definition of long-tailed distributions came up organically in our proof that subexponential distributions are heavy-tailed (Lemma 3.4) which, in retrospect, consisted of first showing that subexponential distributions are a subclass of long-tailed distributions and then showing that long-tailed distributions are heavy-tailed. Thus, we have already proven the following theorem in Chapter 3.

Theorem 4.3. *All long-tailed distributions are heavy-tailed.*

The class of long-tailed distributions is an extremely broad subclass of heavy-tailed distributions. As we just mentioned, the class of long-tailed distributions contains the class of subexponential distributions,

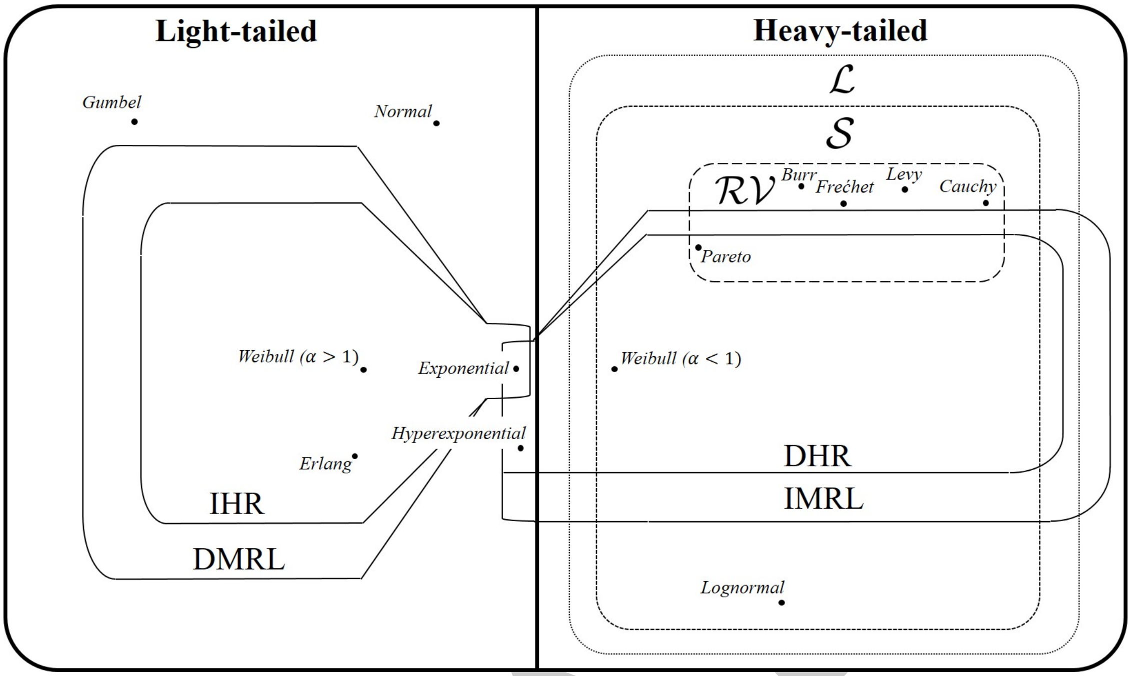


Figure 4.4: *Illustration of various classes of heavy-tailed distributions, along with common distributions.*

which in turn contains the class of regularly varying distributions. As a result, all common heavy-tailed distributions are long-tailed, e.g., the Pareto, the Weibull (with $\alpha < 1$), the LogNormal, the Burr, etc. Figure 4.4 illustrates how the classes we discuss in this chapter fit into those we have discussed previously. As the figure shows, there is a complex zoo of terminology surrounding sub-classes of heavy-tailed distributions but, by this point in the book, each of the terms should mean something precise to you.[6]

In particular, to distinguish long-tailed distributions from heavy-tailed distributions, we recall one of our equivalent definitions of heavy-tailed distributions (Lemma 1.1): A distribution F over the non-negative reals is heavy-tailed if

$$\liminf_{x\to\infty} \frac{-\log \bar{F}(x)}{x} = 0.$$

In contrast, for long tailed distributions, we have the following, which we proved in the process of proving that subexponential distributions are heavy-tailed (Lemma 3.4).

Lemma 4.2. *If $F \in \mathcal{L}$, then* $\lim_{x\to\infty} \frac{-\log \bar{F}(x)}{x} = 0$.

Thus, in order to construct a heavy-tailed distribution that is not long tailed, all that is necessary is to

[6]Since our descriptions of residual life, hazard rate, and MRL are essentially applicable to only non-negative distributions, we mark the Gaussian and the Gumbel distributions as light-tailed but not DMRL in Figure 4.4. Similarly, we mark the Cauchy distribution as regularly varying, but not IMRL. Proving the placement of the Erlang, the LogNormal, the Fréchet, and the Lévy distributions in Figure 4.4 is the goal of Exercises 5, 6, 7, and 8, respectively. Finally, that all DMRL distributions are light-tailed follows from the following property of DMRL distributions (see Prop. 6.1.2 in [179]: If a random variable X having mean $1/\mu$ is DMRL, then for $Y \sim \mathrm{Exp}(\mu)$, $\mathbb{E}\left[g(X)\right] \leq \mathbb{E}\left[g(Y)\right]$ for all convex functions $g(\cdot)$. Taking $g(x) = e^{\epsilon x}$ for $\epsilon \in (0, \mu)$, it then follows that $\mathbb{E}\left[e^{\epsilon X}\right] < \infty$, which implies that X is light-tailed.

ensure that the limit of $\frac{-\log \bar{F}(x)}{x}$ does not exist as $x \to \infty$, but $\liminf_{x\to\infty} \frac{-\log \bar{F}(x)}{x} = 0$ (see Exercise 14).

To understand the contrast between long-tailed distributions and subexponential distributions, let us go back to our proof of Lemma 3.4 in Chapter 3. In particular, in that proof, we showed that F is subexponential if and only if

$$\lim_{t\to\infty} \int_0^t \frac{\bar{F}(t-u)}{\bar{F}(t)} dF(u) = 1.$$

Correspondingly, if F is long-tailed, then it satisfies the above condition so long as we may interchange the limit and the integral. This highlights that the class of subexponential distributions corresponds to those long-tailed distributions that are "well-behaved" enough for this interchange to be valid.

The distinction between long-tailed and subexponential distributions can also be understood in terms of the hazard rate. We have already seen a sufficient condition in terms of the hazard rate for a distribution to be subexponential in Lemma 3.3. In fact, Pitman [159] proved the following (stronger) necessary and sufficient condition for a distribution to be subexponential.

Lemma 4.3. *Suppose that the hazard rate $q(\cdot)$ associated with the distribution F (over the non-negative reals) is eventually decreasing, with $\lim_{x\to\infty} q(x) = 0$. Then $F \in \mathcal{S}$ if and only if*

$$\lim_{x\to\infty} \int_0^x e^{yq(x)} f(y) dy = 1. \tag{4.10}$$

Now, from Theorem 4.2, we see that a distribution F having a hazard rate eventually decreasing to zero, but does not satisfy (4.10), is long-tailed but not subexponential. The interested reader is referred to [159] for an example of such a construction.

4.4 An example: Random extrema

Since the class of long-tailed distributions can be defined in terms of the limiting behavior of the hazard rate and the residual life, it is natural that it finds application most readily in the study of extremes, i.e., the maximum and minimum of a set of random variables. Of course, the study of extremes is crucial to a wide variety of areas such identifying outliers in statistics; determining the likelihood of extreme events in risk management; and many more, including applications in both the physical and social sciences. In this book, we focus on extremal processes in Chapter 7 when discussing the emergence of heavy-tailed distributions and then again in Part III, where extreme value theory plays a crucial role in the statistical tools we develop for estimating heavy-tailed distributions from data.

In settings where extremes are of interest, the core of the analysis often relies on understanding a very simple process – the maximum or minimum of a random number of independently and identically distributed random variables. For example, if one wants to predict the size of the maximal earthquake damage in the US in a particular year, then there is a random number of earthquakes in a year, each of which may be assumed to be independent and identically distributed. Of course, either (or both) the distribution of the number of events and the payout of each event could be heavy-tailed.

More formally, let us consider the following setting, which parallels the example of random sums considered in Chapter 3. Suppose $\{X_i\}_{i\geq 1}$ is a sequence of independent and identically distributed random variables with mean $\mathbb{E}[X]$ and the random variable N takes values in $\mathbb{N}$ and is independent of $\{X_i\}_{i\geq 1}$. Our goal is to characterize

$$M_N = \max(X_1, \ldots, X_N) \text{ and } m_N = \min(X_1, \ldots, X_N).$$

While you may not have studied the random extrema M_N and m_N before, you have almost certainly studied the version where the number of random variables is fixed at n, i.e., M_n and m_n. In particular, for each of these, it is simple to characterize the distribution function

$$\bar{F}_{m_n}(x) = \Pr(X_1 > x, \ldots, X_n > x) = \bar{F}(x)^n$$
$$F_{M_n}(x) = \Pr(X_i < x, \ldots, X_n < x) = F(x)^n.$$

Using this, it is not difficult to see that the class of long-tailed distributions is well-behaved with respect to extrema. For example, we show in the following that the class is closed with respect to max and min.

Lemma 4.4. *For $n \geq 2$, suppose that $X_1, X_2, \cdots, X_n$ are i.i.d., long-tailed random variables. Then*

(i) $\max(X_1, X_2, \cdots, X_n)$ *is long-tailed, and*

(ii) $\min(X_1, X_2, \cdots, X_n)$ *is long-tailed.*

Proof. Let us denote the distribution of each i.i.d. X_i by F. We start by focusing on the first claim: that $M_n = \max(X_1, X_2, \cdots, X_n)$ is long-tailed. To begin, note that the tail of M_n satisfies $\Pr(M_n > x) \sim n\bar{F}(x)$, since

$$\begin{aligned}
\lim_{x\to\infty} \frac{\Pr(M_n > x)}{\bar{F}(x)} &= \lim_{x\to\infty} \frac{1 - (1 - \bar{F}(x))^n}{\bar{F}(x)} \\
&= \lim_{x\to\infty} \frac{1 - (1 - n\bar{F}(x) + \binom{n}{2}\bar{F}(x)^2 - \ldots)}{\bar{F}(x)} \\
&= \lim_{x\to\infty} \frac{n\bar{F}(x) + o(\bar{F}(x))}{\bar{F}(x)} = n.
\end{aligned}$$

Next, for any $y > 0$, we can write

$$\frac{\Pr(M_n > x + y)}{\Pr(M_n > x)} = \left(\frac{\Pr(M_n > x + y)}{n\bar{F}(x + y)}\right) \cdot \left(\frac{n\bar{F}(x)}{\Pr(M_n > x)}\right) \cdot \left(\frac{\bar{F}(x + y)}{\bar{F}(x)}\right).$$

Each of the three functions in brackets above approaches 1 as $x \to \infty$, where the third is a result of the fact that X_i is long-tailed. Thus, it follows that

$$\lim_{x\to\infty} \frac{\Pr(M_n > x + y)}{\Pr(M_n > x)} = 1,$$

which completes the proof of the first claim in the lemma.

The second claim, that $m_n = \min(X_1, X_2, \cdots, X_n)$ is long-tailed, is easy to verify too. Given that X_i

is long-tailed, we have $\lim_{x\to\infty} \frac{\bar{F}(x+y)}{\bar{F}(x)} = 1$. Thus, for any $y > 0$,

$$\lim_{x\to\infty} \frac{\Pr(m_n > x + y)}{\Pr(m_n > x)} = \lim_{x\to\infty} \frac{\bar{F}(x+y)^n}{\bar{F}(x)^n} = 1,$$

which completes the proof. □

While the lemma above is useful, it is concerned only with the minimum and maximum of a fixed number of random variables. Our goal in this section is to study *random extrema*. The fact that long-tailed distributions are closed with respect to max/min suggest that the same may be true for random extrema. In fact, it is possible to say quite a bit more about the behavior of the tail of random extrema. In particular, it is possible to derive precise characterizations of the tail behavior, in quite general settings, with elementary analytic techniques.

Theorem 4.4. *Consider an infinite i.i.d. sequence of random variables $X_1, X_2, \ldots$, and a random variable $N \in \mathbb{N}$ that is independent of $\{X_i\}_{i\geq 1}$ and has $\mathbb{E}[N] < \infty$. Define n_0 such that Pr $(N \geq n_0) = 1$ and Pr $(N = n_0) > 0$. Then,*

$$\textit{Pr}(\max(X_1, X_2, \cdots, X_N) > t) \sim \mathbb{E}[N]\, \textit{Pr}(X_1 > t)$$

and

$$\textit{Pr}(\min(X_1, X_2, \cdots, X_N) > t) \sim \textit{Pr}(N = n_0)\, \textit{Pr}(X_1 > t)^{n_0}.$$

Thus, if X_i are long-tailed then both $\max(X_1, X_2, \cdots, X_N)$ and $\min(X_1, X_2, \cdots, X_N)$ are also long-tailed.

To get intuition for this theorem, an interesting special case to consider is when $N \sim$ Geometric(p). In this case, $\Pr(N = i) = p(1-p)^{i-1}$ for $i \geq 1$ and $\mathbb{E}[N] = 1/p$. So, to apply Theorem 4.4 we set $n = 1$ and $p_n = p$, and thus we obtain

$$\Pr(\max(X_1, X_2, \cdots, X_N) > t) \sim \frac{1}{p}\Pr(X_1 > t)$$
$$\Pr(\min(X_1, X_2, \cdots, X_N) > t) \sim p\Pr(X_1 > t),$$

which gives

$$\lim_{t\to\infty} \frac{\Pr(\max(X_1, X_2, \cdots, X_N) > t)}{\Pr(\min(X_1, X_2, \cdots, X_N) > t)} = \frac{1}{p^2} = \mathbb{E}[N]^2.$$

Interestingly, this does not depend on the distribution of X_i, just on the distribution of N.

Beyond the example of a Geometric distribution, note that the form of Theorem 4.4 for the random maximum exactly parallels the form of the maximum of a deterministic number of samples, i.e., $N = n$. In particular, the tail of the max of n random variables satisfies

$$\Pr(\max(X_1, \ldots, X_n) > t) \sim n\Pr(X_1 > t). \tag{4.11}$$

The parallel form of the above to that of Theorem 4.4 highlights that, with respect to the tail, we can basically ignore the fact that N is random when studying random maxima. Interestingly, this is similar to the insight provided by Wald's equation for random sums, as we discussed in Section 3.3.

Given the above observation, it is natural to further contrast behavior of random extrema (Theorem 4.4) with the behavior of random sums (Theorems 3.1 and 3.2). For random sums, the behavior of the tail depends crucially on whether the tail of N is heavier or lighter than the tail of the X_i. In contrast, for random extrema, the tail of N is unimportant. Further, note that combining Theorem 4.4 with Theorem 3.1 we see that, when X_i are subexponential,

$$\Pr\left(\sum_{i=1}^{N} X_i > t\right) \sim \mathbb{E}[N] \Pr(X_1 > t) \sim \Pr(\max(X_1, X_2, \cdots, X_N) > t).$$

Thus, the random sum is tightly coupled to the random max in this case, as should be expected for subexponential distributions given the catastrophe principle.

Now, with this intuition in hand, let us move to the proof of Theorem 4.4.

Proof. We begin with the following representation of the tail of $M_N = \max(X_1, X_2, \cdots, X_N)$.

$$\Pr(M_N > t) = \sum_{i \in \mathbb{N}} \Pr(N = i) \Pr(\max(X_1, X_2, \cdots, X_i) > t).$$

Therefore,

$$\lim_{t\to\infty} \frac{\Pr(M_N > t)}{\Pr(X_1 > t)} = \lim_{t\to\infty} \sum_{i \in \mathbb{N}} \Pr(N = i) \frac{\Pr(\max(X_1, X_2, \cdots, X_i) > t)}{\Pr(X_1 > t)}.$$

Now, since $\lim_{t\to\infty} \frac{\Pr(\max(X_1, X_2, \cdots, X_i) > t)}{\Pr(X_1 > t)} = i$, the statement of the theorem follows once we justify interchanging the order of the limit and the summation in the above equation. As we show below, this interchange can be justified using the dominated convergence theorem (DCT).

The dominated convergence theorem states that, if we can find an upper bound β_i of $\frac{\Pr(\max(X_1, X_2, \cdots, X_i) > t)}{\Pr(X_1 > t)}$ that is independent of t, such that $\sum_{i \in \mathbb{N}} \Pr(N = i) \beta_i < \infty$, then the order of the limit and the summation can be interchanged. In this case, we may take $\beta_i = i$ since

$$\frac{\Pr(\max(X_1, X_2, \cdots, X_i) > t)}{\Pr(X_1 > t)} = \frac{1 - (1 - \Pr(X_1 > t))^i}{\Pr(X_1 > t)} \leq i,$$

where we have used the inequality $(1 - x)^i \geq 1 - ix$ for $x \geq 0$. The application of the DCT is now justified noting that $\sum_{i \in \mathbb{N}} \Pr(N = i) \beta_i = \mathbb{E}[N] < \infty$. This completes the proof of the tail characterization of M_n.

We now turn our attention to the tail characterization of $m_N = \min(X_1, X_2, \cdots, X_N)$. As before, we may represent the tail of m_N by conditioning with respect to N.

$$\Pr(m_N > t) = \sum_{i \in \mathbb{N}} \Pr(N = i) \Pr(\min(X_1, X_2, \cdots, X_i) > t) = \sum_{i \in \mathbb{N}} \Pr(N = i) \Pr(X_1 > t)^i.$$

Noting that n_0 is the smallest value taken by N with positive probability, we have

$$\begin{aligned}\text{Pr}\,(m_N > t) &= \sum_{i \geq n_o} \text{Pr}\,(N = i)\,\text{Pr}\,(X_1 > t)^i \\ &= \text{Pr}\,(X_1 > t)^{n_0} \Big(\text{Pr}\,(N = n_0) + \text{Pr}\,(N = n_0 + 1)\,\text{Pr}\,(X_1 > t) \\ &\quad + \text{Pr}\,(N = n_0 + 2)\,\text{Pr}\,(X_1 > t)^2 + \cdots \Big) \\ &= \text{Pr}\,(X_1 > t)^{n_0}\, g(\text{Pr}\,(X_1 > t)), \end{aligned} \tag{4.12}$$

where the function g is defined via the power series

$$g(x) = \sum_{i=0}^{\infty} \text{Pr}\,(N = n_0 + i)\, x^i.$$

Suppose that we can show that g is continuous at 0, i.e., $\lim_{x \to 0} g(x) = g(0) = \text{Pr}\,(N = n_0)$. Then it follows from (4.12) that

$$\lim_{t \to \infty} \frac{\text{Pr}\,(m_N > t)}{\text{Pr}\,(X_i > t)^{n_0}} = \lim_{t \to \infty} g(\text{Pr}\,(X_1 > t)) = \text{Pr}\,(N = n_0)\,,$$

which is our desired characterization of the tail of m_N.

It remains then to argue that g is continuous at 0. Note that $g(1) = \sum_{i \geq n_o} \text{Pr}\,(N = i) = 1$, i.e., the power series defined by $g(x)$ converges at $x = 1$. It therefore follows that the power series is continuous over $|x| < 1$, which of course implies the continuity of g at 0. □

4.5 Additional notes

In this chapter we introduced IHR/DHR and IMRL/DMRL in the context of contrasting heavy-tailed and light-tailed distributions; however it is important to note that these classes are useful in much broader contexts. In particular, monotonicity properties of the hazard rate and the mean residual life yield crucial tools in the area of stochastic comparisons and stochastic orderings, which are fundamental tools for applied probability. For in an introduction into the connections between these tools and stochastic orderings we point the interested reader to [42, 139, 171].

Our introduction to long-tailed distributions focused providing intuitive proofs of some important properties of the class. However, readers who want to go into more depth in the study of long-tailed distributions should note that most elementary properties of long-tailed distributions can be shown to quickly follow from their connection with slowly varying functions (as introduced in Chapter 2). This connection is highlighted in [76], which also includes a comprehensive treatment of long-tailed distributions, including properties of long-tailed distributions on the real line, and so-called locally long-tailed distributions.

Unlike the the classes of regularly varying and subexponential distributions it is usually too ambitious to expect asymptotic approximations to hold for the full class of long-tailed distributions. This makes them more difficult to work with in applications. However, in several contexts it is possible to develop asymptotic lower bounds without additional assumptions, thus making the class more easily applicable. For example, in [199] the following is shown. If $S_n = X_1 + ... + X_n$ is a random walk with negative drift, with $P(X_1 > x + y) \sim P(X_1 > x)$, then $\liminf_{x\to\infty} \frac{\Pr(\sup_n S_n > x)}{\int_x^\infty \Pr(X_1 > u)du} \geq -E[X_1]$.

A second, more advanced, application where the class of long-tailed is used is in the analysis of Brownian motion. Specifically, let $B(t)$ be a Brownian motion with strictly positive drift, $M(t) = \sup_{s<t} B(s)$ and T an independent long-tailed random variable. Then, $P(M(T) > x) \sim P(B(T) > x)$. For a proof and a discussion of some applications of this result, see [194] and [204].

4.6 Exercises

1. The goal of this exercise is to prove that the exponential is the only continuous distribution that is memoryless. Specifically, suppose that the distribution F corresponding to a continuous non-negative random variable satisfies the following property:

$$\frac{\bar{F}(x+t)}{\bar{F}(x)} = \bar{F}(t) \qquad \forall\, s, t \geq 0.$$

Prove that F is necessarily exponential.

2. Consider a distribution F corresponding to a non-negative, continuous random variable having finite mean. Show that if the hazard rate function $q(\cdot)$ corresponding to F is eventually decreasing, then the mean residual life function $m(\cdot)$ corresponding to F is eventually increasing.

3. Consider a distribution F corresponding to a non-negative, continuous random variable having finite mean. Let $m(\cdot)$ and $q(\cdot)$ denote, respectively, the MRL and hazard rate corresponding to F.

   Prove that $m'(x) = m(x)q(x) - 1$.

4. Recall that the hyperexponential distribution is the mixture of independent exponentials. Specifically, consider the hyperexponential distribution F defined by the c.c.d.f.

$$\bar{F}(x) = \sum_{i=1}^{n} p_i e^{-\mu_i x} \qquad (x \geq 0).$$

Here, $\mu_i > 0$ and $p_i > 0$ for $1 \leq i \leq n$, with $\sum_{i=1}^n p_i = 1$. Let $m(\cdot)$ denote the mean residual life function corresponding to F.

Prove that $\lim_{x\to\infty} m(x) = 1/\hat{\mu}$, where $\hat{\mu} = \min_i \mu_i$.

Note: Since the hyperexponential is IMRL, the above statement also implies that $m(x) \leq 1/\hat{\mu}$ for all x.

5. Recall that the Erlang distribution with parameters (k, μ), where $k \in \mathbb{N}$ and $\mu > 0$, is associated with the c.c.d.f.

$$\bar{F}(x) = \begin{cases} e^{-\mu x} \displaystyle\sum_{i=0}^{k-1} \frac{(\mu x)^i}{i!} & \text{for } x \geq 0 \\ 1 & \text{for } x < 0 \end{cases}.$$

Prove that the Erlang distribution is IHR.

6. Prove that the LogNormal distribution is neither IMRL nor DMRL.

   Hint: It suffices to establish the following properties of the LogNormal hazard rate function $q(\cdot)$*: (i)* $\lim_{x\downarrow 0} q(x) = 0$*, and (ii)* $\lim_{x\to\infty} q(x) = 0$*. While Property (i) implies that* $\lim_{x\downarrow 0} m'(x) < 0$*, Property (ii) implies that* $\lim_{x\to\infty} m(x) = \infty$*. To prove Property (ii), you may use the following property of the hazard rate function* q_N *of the standard Gaussian:* $q_N(x) \sim x$ *as* $x \to \infty$ *(this follows from Theorem 1.2.3 in [62].*

7. Recall that the Fréchet distribution with shape parameter $\alpha > 0$ is characterized by the c.d.f.

$$F(x) = \begin{cases} e^{-x^{-\alpha}} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}.$$

   Prove that the Fréchet distribution is neither DHR nor IHR. For $\alpha > 1$, prove that the Fréchet distribution is neither DMRL nor IMRL.

   Hint: Once again, it suffices to show that the hazard rate function $q(\cdot)$ *of the Fréchet distribution satisfies: (i)* $\lim_{x\downarrow 0} q(x) = 0$*, and (ii)* $\lim_{x\to\infty} q(x) = 0$*.*

8. Recall that the Lévy distibution with scale parameter $c > 0$ is characterized by the density function

$$f(x) = \begin{cases} \sqrt{\frac{c}{2\pi}} \frac{e^{-\frac{c}{2x}}}{x^{\frac{3}{2}}} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}.$$

   Prove that the Lévy distibution is neither DMRL nor IMRL.

9. Prove that, to check that a distribution F is long-tailed, it suffices to check that (4.9) holds for *some* $t > 0$.

10. Prove that a non-negative random variable X is long-tailed if and only if $\lfloor X \rfloor$ is long-tailed.

11. Let $m(\cdot)$ denote the mean residual life function corresponding to distribution F over the non-negative reals. If $F \in \mathcal{L}$, prove that $\lim_{x\to\infty} m(x) = \infty$.

12. Consider a distribution F over $\mathbb{R}_+$ with finite mean μ. Recall that the *excess* distribution corresponding to F is defined as

$$\bar{F}_e(x) = \frac{1}{\mu} \int_x^\infty \bar{F}(y) dy.$$

   Prove that if F is long-tailed, then F_e is long-tailed.

13. The goal of this exercise is to come up with a representation theorem for long-tailed distributions, analogous to the Karamata representation theorem for regularly varying functions (Theorem 2.6 in Chapter 2). Indeed, the representation theorem for long-tailed distributions turns out to be a direct consequence of the Karamata representation theorem for slowly varying functions.

    Consider a distribution F over the non-negative reals.

(a) Show that $F \in \mathcal{L}$ if and only if $\bar{F}(\log(\cdot))$ is slowly varying.

(b) Use the representation theorem for slowly varying functions (Theorem 2.6) to establish the following representation theorem for long-tailed distributions.

F is long-tailed if and only if $\bar{F}$ can be represented as a monotonically decreasing function of the form

$$\bar{F}(x) = \bar{c}(x)e^{\int_0^x \bar{\beta}(t)dt},$$

where $\lim_{x\to\infty} \bar{c}(x) = c \in (0, \infty)$ *and* $\lim_{x\to\infty} \bar{\beta}(x) = 0$.

(c) Deduce from the above representation theorem that long-tailed distributions are heavy-tailed distributions. Specifically, show that if $F \in \mathcal{L}$, then

$$\lim_{x\to\infty} \frac{\bar{F}(x)}{e^{-\mu x}} = \infty \qquad \forall\, \mu > 0.$$

14. Construct a distribution that is heavy-tailed but not long-tailed.

15. Recall that a non-negative random variable X is square root insensitive if Pr$\left(X > x\right) \sim$ Pr$\left(X > x - \sqrt{x}\right)$. We discussed these distributions briefly in Chapter 3. Prove that if X is square root insensitive, then $\sqrt{X}$ is long-tailed.

Part II

Emergence

When heavy-tailed distributions are found in the world around us they are often reported as surprising curiosities, something in stark contrast to the Gaussian distribution that the central limit theorem teaches us to expect. Of course, the Gaussian distribution *is* prominent in the world around us; but it is also true that heavy-tailed distributions are more than a curiosity – they arise frequently and in many disparate contexts. This begs the question: *Why?* Are there simple laws that can "explain" the emergence of heavy-tailed distributions similarly to the way in which the central limit theorem "explains" the prominence of the Gaussian distribution?

In Part II of this book we focus on this question. We study three generic, foundational stochastic processes in order to understand when one should expect the emergence of heavy-tailed distributions as opposed to light-tailed distributions. In particular, we study additive processes (Chapter 5), multiplicative processes (Chapter 6), and extremal processes (Chapter 7). Additive process are likely the most familiar, given the prominence of the central limit theorem, but multiplicative and extremal processes are nearly as wide-spread. Multiplicative processes arise in situations where growth happens proportionally to the current size, e.g., investments on incomes, population growth, etc.; whereas extremal processes are crucial for characterizing extreme events such as large earthquakes, floods, world records, etc.

Our discussions in the chapters that follow highlight that heavy-tailed distributions should *not* be viewed as anomalies. In fact, heavy-tails should not be surprising at all, in many cases they should be *expected*. Both additive processes and extremal processes can lead to both light-tailed and heavy-tailed distributions, in some cases *creating* heavy-tailed distributions from light-tailed inputs. Further, multiplicative processes are almost guaranteed to lead to heavy-tailed distributions, regardless of the inputs to the process. Taken together, these chapters highlight that the emergence of heavy-tailed distributions should be treated as something as natural as, if not more natural than, the emergence of the Gaussian distribution.

Chapter 5

Additive processes

Discoveries of heavy-tailed phenomena are often viewed with surprise, as if heavy-tailed distributions are merely a probabilistic curiosity. In large part, this is a consequence of the prominence and beauty of the central limit theorem. The central limit theorem leads to the idea that the emergence of the Gaussian (Normal) distribution is almost a "rule of nature," since processes in nature can often be viewed as being formed by a sum of many random events, i.e., via some *additive process*. Consequently, the Gaussian distribution becomes our expectation about how the world around us will behave.

Of course, the Gaussian distribution *is* prominent in our lives. Many aspects of human growth and behavior are approximately Gaussian – heights, weights, test scores, etc. However, heavy-tailed distributions are also very prominent in the world around us, as we have seen throughout this book. The prominence of heavy-tailed phenomena seems to fly in the face of the "explanation" provided by the central limit theorem for the prominence of the Gaussian distribution. Thus, the following natural question emerges:

Given that the central limit theorem predicts the emergence of the Gaussian distribution, why are heavy-tailed distributions so common?

To answer this question, it is natural to first think about where the central limit theorem applies, and where it does not. The most obvious limitation of the central limit theorem is that it only applies to additive processes. Of course, there are many other ways things can evolve besides additive processes and other processes could certainly lead to other, possibly heavy-tailed, distributions. In fact, other types of processes do lead to distributions besides the Gaussian, as we highlight in the following two chapters where we consider two other general processes – multiplicative processes and extremal processes.

But, even in the case when things evolve according to an additive process, it turns out that there is more to the story than the version of the central limit theorem that is typically taught in introductory probability and statistics courses. In particular, the typical statement of the central limit theorem that we learn in these courses is not the full version of the central limit theorem. There is a generalized central limit theorem that highlights that one should not expect additive processes to *always* yield distributions that are approximately Gaussian. The generalized central limit theorem tells us that a broad class of *stable distributions* can emerge from additive processes, and that these distributions often have heavy tails, specifically regularly varying tails.

Thus, in some sense, intuitive expectations about the prominence of the Gaussian distribution are skewed because people only have a partial view of the central limit theorem. In this chapter we seek to remedy this by introducing the generalized central limit theorem and the class of stable distributions. Note that the typical treatment of the generalized central limit theorem and the class of stable distributions is extremely technical, but we do our best to provide an elementary treatment here. Of course, this means that we do not prove results in their full generality. However, throughout the chapter, we point interested readers to references where the full details can be found.

5.1 The central limit theorem

Our focus in this chapter is on additive processes. As the name implies, additive processes are processes that evolve as the sum of random events. A simple, general class of additive processes, which is our focus in this chapter is the following:

$$S_n = X_1 + X_2 + \ldots + X_n, \text{ where } X_i \text{ are i.i.d.}$$

The study of additive processes is a classical and important area, and two of the most celebrated results in probability provide us insight into the behavior of S_n: (i) *the law of large numbers* and (ii) *the central limit theorem*.

While you have almost certainly seen both of these results before, we present them here in a non-standard way in order to highlight a perspective on these results that is not often taught in introductory probability courses. In particular, in the treatment below, we show that the law of large numbers and the central limit theorem can be interpreted as the *first and second order approximations* of the additive process S_n. This "engineers" view is particularly useful for applications in algorithms and computer networking, where random sums are often key to analysis.

To begin to understand this view, let us first recall the statement of the law of large numbers.

Theorem 5.1 (The strong law of large numbers)**.** *Consider an infinite sequence of i.i.d. random variables $X_1, X_2, \ldots$ having $\mathbb{E}[X_i] = \mathbb{E}[X] \in (-\infty, \infty)$. Then,*

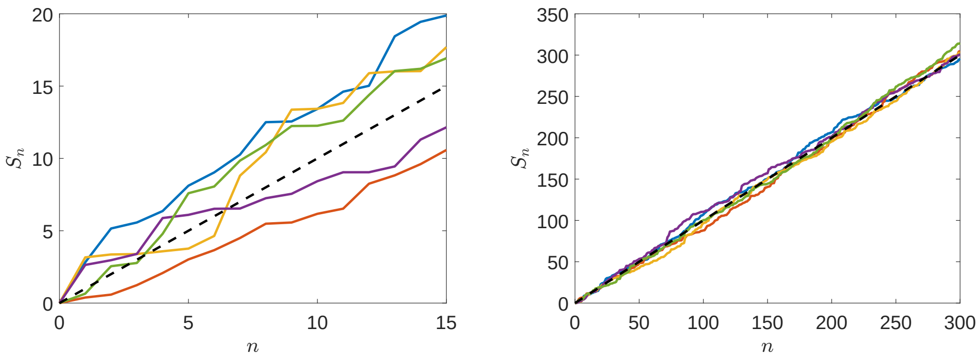
$$\frac{S_n}{n} \xrightarrow{a.s.} \mathbb{E}[X] \text{ as } n \to \infty.[1]$$

Informally, the law of large numbers says that, up to a first-order approximation, S_n is well-approximated by its mean, and is thus linear in n, i.e.,

$$S_n \overset{a.s.}{=} \mathbb{E}[X]\, n + o(n).$$

This is a particularly nice approximation because it is deterministic; all the randomness of the S_n has disappeared and the approximation only depends on n and $\mathbb{E}[X]$. Figure 5.1 highlights this view of the law of large numbers, and the accuracy of the first order approximation the law of large numbers provides. In particular, Figure 5.1 shows that it does not take very large n before the linear, deterministic, first-order approximation of S_n becomes accurate: (a) shows that the approximation is not particularly useful at $n = 15$, but by $n = 300$ the approximation is quite predictive.

[1] Almost sure (a.s.) convergence of a sequence of random variables $\{X_n\}_{n\geq 1}$ to a (possibly random) limit X implies that the with probability 1, the sequence of values taken by the random variables $\{X_n\}_{n\geq 1}$ converge to the value taken by the random variable X [97, Chapter 2]. Almost sure convergence is also commonly referred to as convergence *with probability 1* (w.p.1).



(a) $n = 15$ (b) $n = 300$

Figure 5.1: *Illustration of the first-order approximation of S_n provided by the law of large numbers. The plots illustrate 5 realizations of S_n with $X_i \sim \mathrm{Exp}(1)$. The dotted line shows the law of large numbers approximation.*

Given that the first order approximation of S_n from the law of large numbers is deterministic, it is natural to ask if it can be made more precise, i.e., what is the second order correction term? One would hope to have a correction term that captures the randomness of S_n, and this is exactly what is provided by the central limit theorem.

Theorem 5.2 (The central limit theorem)**.** *Consider an infinite sequence of i.i.d. random variables $X_1, X_2, \ldots$ having $\mathbb{E}[X_i] = \mathbb{E}[X] < \infty$ and Var $[X_i] = \sigma^2 < \infty$. Then,*

$$\frac{S_n - n\mathbb{E}[X]}{\sqrt{n}} \xrightarrow{d} Z \text{ as } n \to \infty, \text{ where } Z \sim \text{Gaussian}(0, \sigma^2).$$ [2]

The form of the central limit theorem highlights that it can be viewed as a correction term for the law of large numbers. Specifically, the term $S_n - n\mathbb{E}[X]$ is the error of the approximation provided by the law of large numbers, and so the central limit theorem highlights that this error is on the order of $\sqrt{n}$. Thus, by combining the central limit theorem and the law of large numbers, we get a second order approximation of the additive process S_n, where the second term is of order $\sqrt{n}$:

$$S_n \stackrel{d}{=} \mathbb{E}[X]\, n + Z\sqrt{n} + o(\sqrt{n}), \text{ where } Z \sim \text{Gaussian}(0, \sigma^2).$$ [3]

Of course, since the central limit theorem does not have a deterministic limit, the approximation is no longer deterministic. However, this is natural (and even desirable) given the randomness of S_n.

[2] A sequence $\{X_n\}_{n \geq 1}$ random variables converges *in distribution* to a (possibly random) limit X (denoted $X_n \xrightarrow{d} X$) if the distributions of X_n converge to the distribution of X as $n \to \infty$, i.e., $\lim_{n\to\infty} \Pr(X_n \leq x) = \Pr(X \leq x)$ for all x. To be more precise, if the c.d.f. $\Pr(X \leq x)$ has discontinuities, then we require that the preceding convergence hold at all points x where $\Pr(X \leq x)$ is continuous [97, Chapter 2].

[3] Recall that $X \stackrel{d}{=} Y$ if the random variables X and Y have the same distribution, i.e., $\Pr(X \leq x) = \Pr(Y \leq x)$ for all x. Thus, the central limit gives us a distributional description of S_n for large n.

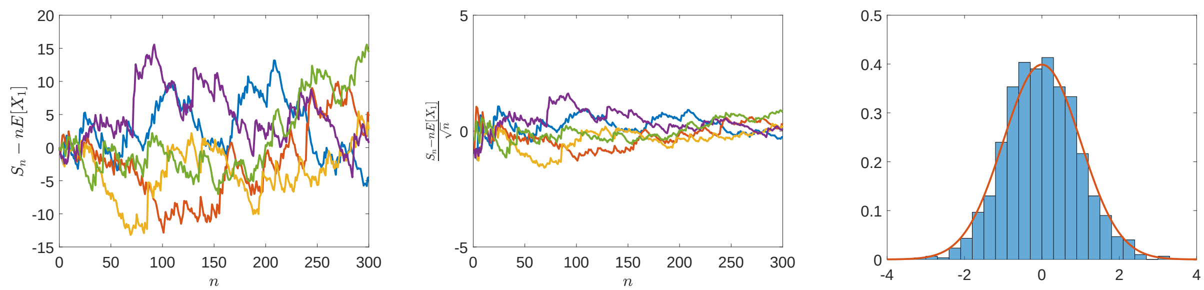


Figure 5.2: *Illustration of the second-order approximation of S_n provided by the central limit theorem. The plots illustrate 5 realizations of S_n with $X_i \sim \text{Exp}(1)$ for $n = 300$. (a) shows the error of the first-order approximation provided by the law of large numbers, i.e., $S_n - n\mathbb{E}[X_i]$. (b) shows the normalized error of the first-order approximation, i.e., $(S_n - n\mathbb{E}[X_i])/\sqrt{n}$. (c) shows the histogram of the normalized error of the first-order approximation for $n = 300$ over 1000 sample paths. The red curve is the probability density function of the standard Gaussian distribution.*

Figure 5.2 highlights the view of the central limit theorem as a "correction term" for the approximation of S_n. In particular, the y-axis of these plots can be interpreted as the error of the first-order approximation provided by the law of large numbers. Figure 5.2 (a) shows the error without normalization, which highlights that the error is growing with n on the order of $\sqrt{n}$. Figure 5.2 (b) shows the error normalized by $\sqrt{n}$. The normalization prevents the correction term from blowing up but, as the figure highlights, significant randomness remains (see the contrast between Figure 5.2 (b) and Figure 5.1 (b)). However, the central limit theorem highlights that this randomness has structure – it follows a Gaussian distribution. This is illustrated in Figure 5.1 (c).

Since understanding and generalizing the central limit theorem is core to our goals in this chapter, it is important for us to consider the proof. In fact, when we move to the generalized central limit theorem, the proof is more technical, but the structure mimics the approach we use in the proof of the classical version of the central limit theorem below.

Our proof of the central limit theorem relies on manipulating the characteristic function of the sum, and so let us first recall some important properties of the characteristic function. The characteristic function of a random variable X is defined as

$$\phi_X(t) = \mathbb{E}\left[e^{itX}\right] = \mathbb{E}\left[\cos(tX) + i\sin(tX)\right].$$

The characteristic function is a particularly useful quantity because it completely determines the distribution of the random variable, i.e., it can be inverted to find the distribution function, and it is often easier than the distribution function to work with analytically. For example, an important property that highlights the value of the characteristic function is the fact that characteristic function of a linear combination of independent random variables can be expressed as the product of the characteristic functions of the random variables, i.e.,

$$\phi_{a_1X_1+a_2X_2+\ldots+a_nX_n}(t) = \phi_{X_1}(a_1t)\phi_{X_2}(a_2t)\ldots\phi_{X_n}(a_nt).$$

As you can imagine, this property makes the characteristic function an appealing tool for studying additive processes in general and, specifically, for proving the central theorem.

Proof of Theorem 5.2. To begin, we shift the X_i to obtain a zero mean random variable $Y_i = (X_i - \mathbb{E}[X])$ and denote the characteristic function of the i.i.d. Y_i by ϕ_Y. Though we do not know much about the form of ϕ_Y, we do know that $\phi_Y(0) = 1$ and that $\phi'_Y(0) = i\mathbb{E}[Y] = 0$. Further, we know that the variance of X_i is finite, and thus the variance of Y is finite as well. It follows that the characteristic function of Y is twice differentiable and that and $\phi''_Y(0) = -\mathbb{E}\left[Y^2\right] = -\mathrm{Var}[Y] = -\sigma^2$. This allows us to write a Taylor expansion of ϕ_Y around 0 as follows:

$$\begin{aligned}\phi_Y(t) &= \phi_Y(0) + \phi'_Y(0)t + \phi''_Y(0)\frac{t^2}{2} + o(t^2) \text{ as } t \downarrow 0\\ &= 1 - \frac{\sigma^2 t^2}{2} + o(t^2), \text{ as } t \downarrow 0.\end{aligned}$$

Next, we use $Z_n = \frac{\sum_{i=1}^n Y_i}{\sqrt{n}}$ to denote normalized additive process we are interested in characterizing. Using the above, we can write the characteristic function of Z_n as follows

$$\phi_{Z_n}(t) = \left(\phi_Y(t/\sqrt{n})\right)^n = \left(1 - \frac{\sigma^2 t^2}{2n} + o(t^2/n)\right)^n \to e^{-\sigma^2 t^2/2} \text{ as } n \to \infty.$$

The key of the proof is to recognize that this limit is exactly the characteristic function of a Gaussian distribution with mean zero and variance σ^2. This completes the proof because, by Lévy's continuity theorem, convergence of the characteristic functions implies convergence in distribution. □

While the central limit theorem is likely second nature for anyone who has taken an introductory probability course, there are some surprises when considering it in the context of heavy-tailed distributions. In particular, note that the central limit theorem applies even when the X_is are heavy-tailed, as long as the variance is finite. So, for example, if one considers the Pareto distribution (or more generally a regularly varying distribution) with index $\alpha > 2$, the variance is finite and so $\frac{1}{\sqrt{n}}(S_n - nE[X])$ converges to a Gaussian distribution.

This initially seems natural, since the central limit theorem is so familiar. However, upon more careful examination it is a bit surprising. Recall that if X_1 and X_2 are regularly varying with index $-\alpha$, then $X_1 + X_2$ is also regularly varying with index $-\alpha$ (see Lemma 2.4). Thus, by induction, any finite sum $X_1 + \ldots + X_n$ is regularly varying with index $-\alpha$. So, for all n, $S_n = X_1 + \ldots + X_n$ is heavy-tailed, yet as $n \to \infty$ the central limit theorem gives a limiting distribution that is light-tailed – the Gaussian. Even more surprisingly, while all moments of the Gaussian distribution are finite, S_n will have infinite higher-order moments for all finite n! Specifically, $E[S_n^k] = \infty$ for all n and $k > \alpha$.

This seems contradictory at first; however, the reason for the apparent contradiction is simple. For any large but finite n, the approximation of S_n provided by the Gaussian distribution is only accurate up to a particular point in the tail, say for $t < t_n$. Since the Gaussian is light-tailed, it cannot approximate the full tail of a heavy-tailed distribution but, as $n \to \infty$, $t_n \to \infty$, and so convergence in distribution is achieved.[4]

[4] In fact, for subexponential X_i, there is a precise understanding of the range of t over which the probability $\Pr(S_n > t)$ is well approximated by a Gaussian tail, and the range over which the 'catastrophe' approximation $n\Pr(X_1 > t)$ is more accurate;

5.2 Generalizing the central limit theorem

The classical statement of the central limit theorem gives a convincing explanation for the emergence of the Gaussian distribution in the world around us: if a process grows additively then it can be approximated via the Gaussian distribution. But, of course, there are conditions on when this may happen. It is the relaxing of these conditions that highlights that the Gaussian is not as special as the classical statement of the central limit theorem makes it appear.

In particular, there are four key assumptions in the classical central limit theorem: (i) X_i are identically distributed; (ii) X_i are independent; (iii) X_i have finite mean; and (iv) X_i have finite variance. It turns out that (i) and (ii) are not particularly crucial. Specifically, if X_i are not identically distributed, then it is still possible to obtain versions of the central limit theorem where S_n converges to a Gaussian distribution, e.g., the Lyapunov central limit theorem [27, Section 27]. Similarly, if X_i are dependent it is still possible to prove versions of the central limit theorem where S_n converges to a Gaussian distribution, as long as the dependence is "local", e.g., see Theorem 27.4 of [27]. Of course, if the dependence is extreme, then it can be constructed to create an arbitrary limiting distribution for S_n.

Thus, we are led to considering assumptions (iii) and (iv) – that the X_i have finite mean and variance. In both of these cases, we can immediately see that S_n cannot have a limit that is Gaussian since the Gaussian distribution is defined via a finite mean and variance. However, this does not mean that a version of the central limit theorem does not apply here.

In particular, we might still hope to have a result of the form

$$\frac{(X_1 + X_2 + \cdots + X_n) - b_n}{a_n} \xrightarrow{d} G, \tag{5.1}$$

for some sequences of scaling parameters a_n and translation parameters b_n, and some random variable G. This form naturally generalizes both the law of large numbers and the classical central limit theorem. In the case of the law of large numbers $a_n = n$ and $b_n = 0$ and in the case of the standard central limit theorem $a_n = \sqrt{n}$ and $b_n = n\mathbb{E}[X]$.

If we hope to obtain a generalized form of the central limit theorem such as the one in (5.1), then a first step is to understand what distributions may serve as limits, i.e.,

What distributions might the limiting distribution, G, follow?

Clearly, the Gaussian distribution is a candidate. In fact, the Gaussian distribution has some very useful properties that make it natural in the context of the central limit theorem. The most important is that, if X_1, X_2 are independent and Gaussian with mean μ and variance σ^2, then $X_1 + X_2$ is Gaussian with mean 2μ and variance $2\sigma^2$. More generally, for any $a_1, a_2 > 0$, $a_1X_1 + a_2X_2$ is Gaussian distributed with mean $(a_1 + a_2)\mu$ and variance $(a_1^2 + a_2^2)\sigma^2$, which implies that

$$a_1X_1 + a_2X_2 \stackrel{d}{=} \left(\sqrt{a_1^2 + a_2^2}\right) X_1 + \left(a_1 + a_2 - \sqrt{a_1^2 + a_2^2}\right) \mu$$

Thus, the distribution of the sum of two Gaussian random variables yields a simple linear scaling of one of the original random variables. So, the Gaussian is, in a sense, "stable" with respect to addition and scalar multiplication.

the interested reader is referred to [138] for a survey of such results.

This notion of "stability" turns out to be interesting beyond the Gaussian distribution. In particular, it is the foundation of the class of *stable distributions*, which is defined as follows.

Definition 5.1. *A distribution F is stable if, for any $n \geq 2$ i.i.d. random variables $X_1, X_2, \cdots, X_n$ with distribution F, there exist constants $c_n > 0$, $d_n \in \mathbb{R}$ such that*

$$X_1 + X_2 + \cdots + X_n \overset{d}{=} c_n X_1 + d_n.$$

We have already seen that the Gaussian distribution is an example of a stable distribution. Another (trivial) example of a stable distribution is a degenerate point mass where $X_i = c$ with probability 1. The Cauchy and Lévy distributions provide two other common examples of a stable distribution. However, beyond these examples, it is difficult to understand how broad the class of stable distributions is and what properties stable distributions have from the definition alone.

To better understand the definition of stable distributions, note that it is tightly coupled to the form of the central limit theorem, and to its generalized form in (5.1). In particular, the class of stable distributions is of foundational importance because stable distributions are precisely the distributions that can serve as the limiting distribution G in (5.1).

Theorem 5.3. *A random variable Z has a stable distribution if and only if there exists an infinite sequence of i.i.d. random variables $X_1, X_2, \ldots$ and deterministic sequences $\{a_n\}, \{b_n\}$ $(a_n > 0)$, such that*

$$\frac{(X_1 + X_2 + \cdots + X_n) - b_n}{a_n} \xrightarrow{d} Z.$$

In a sense, the above theorem provides a first-cut at a generalized central limit theorem. In particular, Theorem 5.3 highlights that there is an entire class of distributions, stable distributions, that serve as candidate limiting distributions for additive processes. So, the Gaussian distribution is not as "special" as the standard statement of the central limit theorem makes it appear, and the fact that it is the limiting distribution can be "explained" simply by the fact that it is a stable distribution. However, Theorem 5.3 is not particularly satisfying when interpreted as a generalized central limit theorem since it says very little about the form of the distributions that can emerge as limiting distributions or about what the limiting distribution will be for specific X_i. The next two sections fill in these holes as we continue to move toward the full statement of the generalized central limit theorem.

Proof of Theorem 5.3. First, we show that if F is a stable distribution, then it is the limit, in distribution, of a centered, normalized additive process. Let $\{X_i\}_{i \geq 1}$ denote an i.i.d. sequence of random variables with distribution F. By Definition 5.1, for any $n \geq 2$, there exist constants $c_n > 0$, $d_n \in \mathbb{R}$ such that

$$X_1 + X_2 + \cdots + X_n \overset{d}{=} c_n X_1 + d_n.$$

In other words, for any $n \geq 2$,

$$\frac{(X_1 + X_2 + \cdots + X_n) - d_n}{c_n} \overset{d}{=} X_1.$$

It therefore follows trivially that as $n \to \infty$,

$$\frac{(X_1 + X_2 + \cdots + X_n) - d_n}{c_n} \xrightarrow{d} F.$$

Next, we show that if the distribution F is the limit in distribution of a centered, normalized additive process, then F is stable. Accordingly, suppose that

$$\frac{(X_1 + X_2 + \cdots + X_n) - b_n}{a_n} \xrightarrow{d} F,$$

where $\{X_i\}_{i\geq 1}$ is an i.i.d. sequence of random variable, and $\{a_n\}$, $\{b_n\}$ are deterministic sequences satisfying $a_n > 0$. Now, fix integer $k \geq 2$, and define, for $m = jk$, $j \in \mathbb{N}$,

$$Y_m = \frac{(X_1 + X_2 + \cdots X_m) - b_m}{a_m},$$
$$Z_m = \frac{(X_1 + X_2 + \cdots X_m) - kb_j}{a_j}.$$

Consider the limit now as $m \to \infty$ by taking $j \to \infty$. Clearly, $Y_m \xrightarrow{d} F$. On the other hand, note that Z_m is the sum of k i.i.d. random variables, each distributed as $\frac{(X_1+X_2+\cdots+X_j)-b_j}{a_j}$. Therefore, $Z_m \xrightarrow{d} F^{*k}$, where F^{*k} is the distribution corresponding to the sum of k i.i.d. random variables having distribution F. Moreover, since Y_m and Z_m differ only via translation and scaling parameters, it can be shown that their limiting distributions also only differ via translation and scaling parameters. In other words, F^{*k} and F differ only via translation and scaling parameters. Since this is true for all $k \geq 2$, it then follows from Definition 5.1 that F is stable.

The above argument can be formalized by invoking the following classical result: Suppose that a random sequence $\{Z_n\}_{n\geq 1}$ converges in distribution to a (non-degenerate) limit Z. Then for a deterministic real sequence $\{\beta_n\}_{n\geq 1}$ and a deterministic positive sequence $\{\alpha_n\}_{n\geq 1}$, if the sequence $\frac{Z_n - \beta_n}{\alpha_n}$ has a (non-degenerate) limit, then this limit must be a translated and scaled version of Z. (see [87, Section 10] or [64, Appendix A1.5]). $\square$

5.3 Understanding stable distributions

The discussion in the previous section highlights the importance of the class of stable distributions in the context of generalizing the central limit theorem; however the results we have discussed so far are not particularly satisfying because they provides little information about the form of the limiting distributions or about when different limiting distributions will emerge, e.g., Theorem 5.3. To address these holes it is necessary to develop a deeper understanding of the class of stable distributions.

To this point in the chapter, we have noted that the class of stable distributions includes a few common distributions: the Gaussian distribution, deterministic distributions, the Lévy distribution, and the Cauchy distribution. But, beyond these distributions we have said very little. The reason is that it is difficult to understand the generality of the class from the definition directly. However, it turns out that the class of stable distributions can be characterized quite concisely. In particular, the following "representation theorem" characterizes all stable distributions via their characteristic functions.[5]

Theorem 5.4 (Representation theorem)**.** *A non-degenerate random variable X is α-stable if and only if $X \overset{d}{=} aZ + b$, where $a > 0$, $b \in \mathbb{R}$, and the random variable Z has a characteristic function of the following form, parameterized by $\alpha \in (0, 2]$, and $\beta \in [-1, 1]$.*

$$\phi_Z(t) = \exp\left\{-|t|^\alpha \left(1 - i\beta \mathrm{sign}(t)\gamma(t,\alpha)\right)\right\}, \tag{5.2}$$

where

$$\gamma(t,\alpha) = \begin{cases} \tan\left(\frac{\pi\alpha}{2}\right) & \text{for } \alpha \neq 1 \\ -\frac{2}{\pi}\log|t| & \text{for } \alpha = 1 \end{cases}$$

and

$$\mathrm{sign}(x) = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}.$$

This representation theorem gives a concrete characterization of stable distributions, which is crucial for understanding the generality of the class. In particular, the representation theorem highlights that the class of stable distributions can be parameterized via two parameters, $\alpha \in (0, 2]$, and $\beta \in [-1, 1]$. Typically, α is referred to as the stability parameter and stable distributions with stability parameter α are referred to as $\alpha-$stable distributions. It turns out that α is closely tied to the tail of the distribution. On the other hand, β is the skewness parameter: $\beta = 0$ yields distributions symmetric around zero, while $\beta > 0$ (β < 0) yields distributions skewed to the right (left) of zero. Note that $\beta = 0$ corresponds to a real-valued ϕ, which implies a symmetric distribution.

While the representation theorem provides a concise characterization of the class of stable distributions, the fact that it is stated in terms of the characteristic function makes it difficult to derive insight about the properties of stable distribution functions from it directly. It would be more convenient to have a characterization of the class in terms of the distribution function itself, as we had for the class of regularly varying distributions (Chapter 2), subexponential distributions (Chapter 3), and long-tailed distributions (Chapter 4). However, while the characteristic functions of stable distributions have closed forms, only in a handful of cases can these distributions be described via their density function or distribution function in closed form.

We have already discussed a few of these. The case of $\alpha = 2$, $\beta = 0$ corresponds to the Gaussian distribution. This is easy to observe by substituting into (5.2), which yields $\phi_Z(t) = e^{-t^2}$, which of course means Z is Guassian distributed with mean 0 and variance 2. Similarly, it is easy to see that $\alpha = 1$, $\beta = 0$ corresponds to the Cauchy distribution. Another well-known stable distribution is the Lévy distribution, which corresponds to $\alpha = 1/2$, $\beta = 1$.

In these three cases, the distribution function can be written in closed form, but it is difficult to explicitly write the distribution and density functions for stable distributions in general. However, there are some properties of the distribution functions of stable distributions that can be characterized. One that is of particular interest in the context of this book is the tail of stable distributions. Specifically, the following theorem characterizes the tail behavior of α-stable distributions, for $\alpha \in (0, 2)$. Given that the case of $\alpha = 2$ corresponds to the Gaussian distribution, this gives a complete characterization of the tail behavior of non-degenerate stable distributions.

[5]Proofs of this result tend to be quite technical, and so we do note prove them here. The interested reader can refer to [70, Chapter 17] for the proofs.

Theorem 5.5. *If X is α-stable for $\alpha \in (0,2)$, there exist $p, q \geq 0$ with $p + q > 0$ such that, as $x \to \infty$,*

$$Pr(X > x) = (p + o(1))x^{-\alpha}, \quad Pr(X < -x) = (q + o(1))x^{-\alpha}.$$

Thus, either the left tail, the right tail, or both tails of X are regularly varying with index $-\alpha$.

This theorem has deep consequences. It highlights that, within the class of stable distributions, the only non-degenerate light-tailed distribution is the Normal distribution. Further, it highlights that the only stable distribution with finite variance is the Gaussian. Thus, it is natural that the Gaussian distribution emerges as the limit in the standard central limit theorem. However, at the same time, Theorem 5.5 highlights that the Gaussian distribution is, in some sense, a corner case since all other limiting distributions of additive processes are heavy-tailed with infinite variance.

The proof of Theorem 5.5 is quite technical, but it is possible to give a proof of a restricted version of the result using elementary techniques. This proof is a good illustration of the use of a Tauberian theorem; recall that we introduced two Tauberian theorems in Chapter 2: Theorem 2.8 for Laplace-Stieltjes transforms and Theorem 2.10 for characteristic functions. Here we make use of the Tauberian theorem in Theorem 2.10 for characteristic functions, which is due to Pitman [158] (see also Page 336 of [28]). Recall that this theorem uses only the real component of the characteristic function, $U_X(t)$, i.e.,

$$U_X(t) := \mathrm{Re}(\phi_X(t)) = \int_{-\infty}^{\infty} \cos(tx) dF(x).$$

We restate the result here for the convenience of the reader.

Theorem 5.6 (Pitman's Tauberian theorem)**.** *For slowly varying $L(x)$, and $\alpha \in (0,2)$, the following are equivalent:*

$$Pr(|X| > x) \sim x^{-\alpha} L(x) \text{ as } x \to \infty,$$
$$1 - U_X(t) \sim \frac{\pi}{2\Gamma(\alpha)\sin(\pi\alpha/2)} t^{\alpha} L(1/t) \text{ as } t \downarrow 0.$$

The power of this result is that it connects the tail of the distribution to the behavior of the characteristic function around zero. However, this particular Tauberian theorem applies only to the tail of $|X|$, not the tail of X. Thus, it cannot be used to distinguish the behavior of the right and left tails of the distribution. Rather, it provides information about the *sum* of the two tails. But, we choose to use this particular Tauberian theorem here because it deals only with the real part of the characteristic function, which makes it much simpler to work with analytically, which allows the proof to be more instructive. Naturally, using the Tauberian theorem in Theorem 5.6, we cannot hope to prove the entirety of Theorem 5.5, instead we prove the following restricted version.

Theorem 5.7 (Restricted version of Theorem 5.5)**.** *If X is α-stable for $\alpha \in (0,2)$, then $|X| \in \mathcal{RV}(\alpha)$.*

Though not as precise as Theorem 5.5, this restricted result already highlights the most important aspect of Theorem 5.5: that α-stable distributions with $\alpha \in (0,2)$ are heavy-tailed. Note the key difference between the proof of the restricted version and the proof of the full result is the use of a more detailed Tauberian theorem than Theorem 5.6, and the added analytic complexity that comes along with the need to work with the complex-valued characteristic function $\phi_X(t)$ instead of just the real-valued component $U_X(t)$.

Proof of Theorem 5.7. Let Z denote an α-stable random variable. Thus, the characteristic function of Z is given by the representation theorem as (5.2).

For simplicity, we treat only the case of $\alpha \in (0,2)$, $\alpha \neq 1$, and leave the case of $\alpha = 1$ as an exercise for the reader. In this case, the representation theorem gives, for $t > 0$, $\phi_Z(t) = \exp\{-t^\alpha(1 - i\beta\tan\left(\frac{\pi\alpha}{2}\right))\}$. This means that

$$U_Z(t) = \exp\{-t^\alpha\}\cos(\delta t^\alpha),$$

where $\delta = \beta\tan\left(\frac{\pi\alpha}{2}\right)$.

To prove the theorem, it is sufficient to show that

$$1 - U_Z(t) \sim t^\alpha \quad (t \downarrow 0). \tag{5.3}$$

The theorem follows from the above because we can apply the Tauberian theorem in Theorem 5.6 to conclude that

$$\Pr(|Z| > x) = \Pr(Z > x) + \Pr(Z < -x) \sim \frac{\pi}{2\Gamma(\alpha)\sin(\pi\alpha/2)}x^{-\alpha}.$$

To show (5.3), we first consider the case when $\delta = 0$. In this case,

$$U_Z(t) = \exp\{-t^\alpha\} = 1 - t^\alpha + o(t^\alpha),$$

which implies (5.3).

All that remains is the case when $\delta \neq 0$. In this case,

$$U_Z(t) = [1 - t^\alpha + o(t^\alpha)]\cos(\delta t^\alpha).$$

Therefore,

$$\begin{aligned}1 - U_Z(t) &= (1 - \cos(\delta t^\alpha)) + \cos(\delta t^\alpha)t^\alpha - \cos(\delta t^\alpha)o(t^\alpha)\\ &= 2\sin^2(\delta t^\alpha/2) + \cos(\delta t^\alpha)t^\alpha - \cos(\delta t^\alpha)o(t^\alpha).\end{aligned}$$

Using $\sin(\delta t^\alpha/2) \sim \delta t^\alpha/2$ as $t \downarrow 0$, we note that $\sin^2(\delta t^\alpha/2) = o(t^\alpha)$ as $t \downarrow 0$, which is enough to guarantee that (5.3) holds and thus complete the proof. □

5.4 The generalized central limit theorem

The characterization of stable distributions in the previous section gives us the context we need in order to move to the generalized central limit theorem. We already know that the limiting distributions of additive processes are stable (Theorem 5.3). Combining this with our characterization of stable distributions in Theorem 5.4 allows us to understand when different stable distributions emerge as the limiting distribution.

Theorem 5.8 (Generalized central limit theorem)**.** *Consider an infinite sequence of i.i.d. random variables*

$X_1, X_2, \ldots$ with distribution F. There exist deterministic sequences $\{a_n\}, \{b_n\}$ $(a_n > 0)$ such that

$$\frac{(X_1 + X_2 + \cdots + X_n) - b_n}{a_n} \xrightarrow{d} Z,$$

if and only if Z is α-stable for some $\alpha \in (0, 2]$. Further,

(i) Z is Gaussian, i.e., $\alpha = 2$, if and only if $\int_{-x}^{x} y^2 dF(y)$ is slowly varying as $x \to \infty$;

(ii) $|Z| \in \mathcal{RV}(\alpha)$, i.e., $\alpha \in (0, 2)$, if and only if

$$\bar{F}(x) = (p + o(1))x^{-\alpha}L(x), \quad F(-x) = (q + o(1))x^{-\alpha}L(x)$$

as $x \to \infty$, where $L(x)$ is slowly varying, and $p, q \geq 0$, $p + q > 0$.

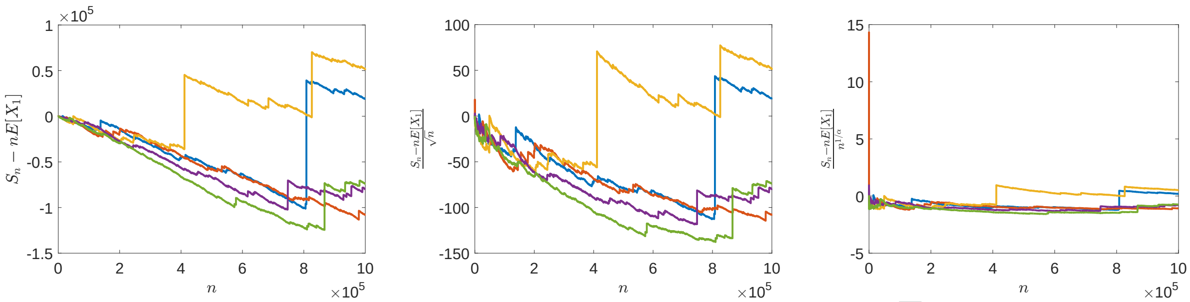
The most striking aspect of this result comes from contrasting it with the standard central limit theorem. Here, the emergence of the Gaussian distribution is just one of many possible options, and one that is, in some sense, a corner case since it corresponds only to $\alpha = 2$. Another interesting contrast to the standard central limit theorem is that the Gaussian distribution occurs as the limiting distribution even in some cases where the variance is infinite.[6] In these cases, the normalizing constant a_n must be chosen to be larger than $\sqrt{n}$, and includes a slowly varying function to counter the growth of $\int_{-x}^{x} y^2 dF(y)$. For a discussion of exactly how to determine a_n in such cases, see [70, Chapter 17].

The most novel part of the generalized central limit compared to the classical version central limit theorem is that it highlights that heavy-tailed stable distributions can emerge as the limit of additive processes. In particular, if one starts with finite variance distributions, the Gaussian emerges, but if one starts with regularly varying distributions with infinite variance, then heavy-tailed distributions can emerge. Further, the statement shows that the emerging distribution can have an infinite mean, and so the tail can be extremely heavy. Figure 5.3 illustrates the contrast with the behavior in the case when the X_is have finite variance. Note that there are big jumps in the process in the case of infinite variance that are not present in classical case illustrated in Figure 5.2(a) and, as a result, the limiting distribution looks dramatically different than the Gaussian. This is a visceral illustration of the catastrophe principle discussed in Chapter 3.

Not only does the generalized central limit theorem specifically highlight which limiting distributions may occur, it also characterizes exactly when specific stable distributions emerge in the limit, i.e., the "domain of attraction" for limiting distributions. Formally, the domain of attraction of a limiting distribution Z is the set of distributions for X_i that have Z as the limiting distribution. For example, the domain of attraction of the Gaussian distribution is the set of distributions where $\int_{-x}^{x} y^2 dF(y)$ is slowly varying as $x \to \infty$. The domain of attraction is also tightly coupled with the sequences of $\{a_n\}, \{b_n\}$ that are used for scaling the process. Though the statement of the generalized central limit theorem given above does not explicitly give $\{a_n\}, \{b_n\}$, they can be precisely specified. For example, when $\alpha \in (0, 2)$, the scaling constants a_n must be chosen to satisfy

$$\lim_{n\to\infty} \frac{nL(a_n)}{a_n^\alpha} = c \in (0, \infty)$$

[6] Indeed, note that the condition that $\int_{-x}^{x} y^2 dF(y)$ is slowly varying is satisfied by all finite variance distributions, as well as some distributions with infinite variance (see Exercise 4).



(a) (b) (c)

Figure 5.3: *Illustration of the error of the first-order approximation of S_n provided by the law of large numbers for Pareto distributed X_1 with $\alpha = 1.2$ and mean 1 (note that X_1 has an infinite variance). The plots illustrate 5 realizations of $S_n - n\mathbb{E}[X_1]$, each showing a different scaling: (a) shows the unscaled deviation around the mean; (b) shows the classical central limit theorem scaling, which does not lead to a limiting distribution; and (c) shows the generalized central limit theorem scaling, which leads to a non-Gaussian limiting distribution.*

and the translation constants may be chosen to satisfy

$$b_n = \begin{cases} 0 & \text{if } 0 < \alpha < 1 \\ na_n \int_{-\infty}^{\infty} sin(x/a_n)dF(x) & \text{if } \alpha = 1 \\ n\mathbb{E}[X_1] & \text{if } 1 < \alpha < 2 \end{cases}.$$

The case of $\alpha = 2$ is more involved, and so we refer the reader to [64, Section 2.2] Note that these sequences are both regularly varying, which again emphasizes the fundamental importance of the concept of regular variation.

Finally, it is important to discuss the proof of the generalized central limit theorem. Unfortunately, the full proof is quite technical and not appropriate for inclusion here. The interested reader can find the full details in [70, Chapter 17]. However, to give a sense for the generalized central limit theorem, we prove a restricted form of it here using an approach that mimics the way the standard central limit theorem is typically proven. This highlights that the generalized central limit theorem is no more mysterious than the standard central limit theorem, it simply requires a bit more technical machinery.

In particular, we prove a version of the generalized central limit theorem restricted to the case of symmetric random variables. The key reason that this restriction is simpler to prove is that symmetry means that the imaginary part of the characteristic function disappears, which allows the use of the Tauberian theorem in Theorem 5.6. Also, because of symmetry, the translation constants $\{b_n\} = 0$, and so they do not clutter the statement or the proof.

Theorem 5.9 (Generalized central limit theorem for symmetric random variables)**.** *Suppose that $\{X_i\}_{i\geq 1}$ are i.i.d. symmetric random variables with distribution F, where $Pr(|X| > x) \sim cx^{-\alpha}$, with $\alpha \in (0, 2)$. Then*

$$\frac{\sum_{i=1}^{n} X_i}{n^{1/\alpha}} \xrightarrow{d} G,$$

where G is a symmetric α-stable distribution.

Proof. Since the X_i are symmetric the imaginary part of their characteristic function equals zero. Thus, denoting the characteristic function of the i.i.d. X_i by ϕ_X, we have that $\phi_X = U_X(t)$. Moreover, since U_X is a symmetric function, i.e., $U_X(-t) = U_X(t)$, so is ϕ_X.

The proof of the standard central limit theorem begins by considering the Taylor expansion of the characteristic function of $Y_i = X_i - \mathbb{E}[X]$. Of course, since X_i are symmetric around zero here, we do not need to consider such a shift. But, a more fundamental difference is that, since the $\mathrm{Var}[X_i] = \infty$, we cannot write such a Taylor expansion of ϕ_X. Instead, we invoke the Tauberian theorem in Theorem 5.6 to obtain a similar representation of ϕ_X. In particular, we have

$$\phi_X(t) = 1 - b|t|^\alpha(1 + o(1)) \text{ as } t \downarrow 0,$$

where $b = \frac{c\pi}{2\Gamma(\alpha)\sin\left(\frac{\pi\alpha}{2}\right)}$. and $\alpha \in (0, 2)$. Recall that the corresponding representation in the proof of the standard central limit theorem has the form $1 - \frac{\sigma^2 t^2}{2} + o(t^2)$, which is parallel to the form above if one were to use $\alpha = 2$.

Given this representation of ϕ_X, the proof now mimics that for the standard central limit theorem. In particular, denote $Z_n = \frac{\sum_{i=1}^n X_i}{n^{1/\alpha}}$. Then the characteristic function of Z_n satisfies the following

$$\phi_{Z_n}(s) = \left(\phi_X(s/n^{1/\alpha})\right)^n = \left(1 - b(1 + o(1))\frac{|t|^\alpha}{n}\right)^n \to e^{-b|t|^\alpha} \text{ as } n \to \infty.$$

The limit is simply a scaled version of the canonical characteristic function of the symmetric ($\beta = 0$) α-stable distribution, as desired. Again, applying the Lévy continuity theorem then completes the proof. □

5.5 A variation: The emergence of heavy tails in random walks

To this point in the chapter we have considered a general, but abstract, version of an additive process: $S_n = X_1 + \ldots + X_n$. Within this context, we focused entirely on understanding the behavior of S_n as $n \to \infty$, and we showed that if the X_is have finite variance then the Gaussian distribution emerges, but that if the X_is have infinite variance then heavy-tailed distributions can emerge.

Of course, additive processes come in many guises and, depending on the setting, one may ask a variety of other questions about S_n besides simply how it behaves as $n \to \infty$. In this section, we highlight that, when different aspects of additive processes are studied, heavy-tailed may emerge even in contexts where the random process is made up of entirely light-tailed components.

In particular, in this chapter we consider a classical example of an additive process, *random walks*. Random walks are an example of a very simple process that has found application in a surprising number of disciplines. In fact, few mathematical models have found applications in as diverse a range of areas, including finance, computer science, physics, biology, and more.

The most simple example of a random walk is one in which a walker is equally likely take a single step in one direction or the other at each time step. Formally, the walker that starts at 0 and takes a sequence of independent steps $X_1, \ldots, X_n$ where

$$X_i = \begin{cases} 1, & \text{with probability } 1/2; \\ -1, & \text{with probability } 1/2. \end{cases}$$

Thus, the position of the walker at time n is simply the additive process

$$S_n = X_1 + X_2 + \ldots + X_n.$$

This is referred to as a simple symmetric random walk in one-dimension, where simple refers to the fact that the step size is one, symmetric refers to the fact that steps up and down are equally likely, and one-dimensional refers to the fact that the random walk is on a line. You have almost certainly come across this version of a random walk in your introductory probability course. It is often described under the guise of a drunken persona leaving a bar and wandering aimlessly up and down the street trying to get home.

Of course, there are many more complicated versions of random walks too. In general, the random walk may be asymmetric or biased, e.g., the probability of taking a positive step may be $p \neq 1/2$; the random walk may be in more than one dimension or over a general graph; or the random walk might allow step sizes other than 1, e.g., the step size could be random. But, for this section, we stick to the simple, symmetric, one-dimensional case because this case is already sufficient to highlight the emergence of heavy-tailed distributions. We study other forms of random walks in Chapters 3 and 7.

One of the most natural questions to ask about a random walk is where the walker is likely to be after a certain (large) number of steps, i.e., what is the behavior of S_n as n grows large. Of course, this is the same question that we have addressed throughout this chapter for more general additive processes. In particular, the law of large numbers and the central limit theorem are enough to give us an answer. In fact, in the case of the simple one-dimensional random walk we are studying, the X_is are light-tailed and so the classical version of the central limit theorem applies, which means that $S_n/\sqrt{n}$ converges to a Gaussian distribution. However, it is not even necessary to apply the central limit theorem since the distribution of the position of the random walk can be easily seen to be a Binomial distribution. Note that if the random walk is no longer "simple" and has random step sizes, this question becomes more complex. We discuss that case in Section 3.4, where we highlight the impact of heavy-tailed step sizes on the behavior of the walker.

Though the position of the random walk at time n can be understood quite easily, this is only one of many questions one may ask about a random walk. For example, two other important questions that are often asked are "when will the walker first return to its starting point?" and "what is the maximum position of the walker after n steps?" It is the first of these that we focus on here. We discuss the second question in Section 7.4.

More specifically, in this section we study the "return time" of a random walk, which is denoted by T and defined as the first time when the walker returns to its starting point. The return time of a random walk is often of crucial importance.

The return time is of particular interest in the context of this chapter because it is an example where a heavy-tailed distribution emerges from an additive process. Further, it is a jarring example of the emergence of heavy-tailed distributions because a heavy-tailed distribution emerges even though the underlying process is defined entirely using bounded (and thus extremely light-tailed) distributions. Specifically, the tail of the distribution of the return time of a simple, symmetric, one-dimensional random walk can be characterized as follows.

Theorem 5.10. *Consider a simple, symmetric, one-dimensional random walk. The distribution of the return time T satisfies Pr* $(T > x) \sim \frac{\sqrt{2/\pi}}{\sqrt{x}}$.

This result highlights that, not only is the return time heavy-tailed, it is extremely heavy-tailed. In particular, it is regularly varying with index $1/2$, which highlights that both the mean and variance are infinite.

There are many methods for proving Theorem 5.10. Below, we use the proof as an opportunity to illustrate another application of Tauberian theorems. In this case, we make use of Karamata's Tauberian theorem introduced in Chapter 2, i.e., Theorem 2.9. Thus, the bulk of the proof is simply to derive the form of the Laplace-Stieltjes transform of the return time and then we apply Karamata's Tauberian theorem (Theorem 2.8) in order to infer the tail behavior of the return time from the behavior of the Laplace-Stieltjes transform around zero.

Proof. For $n = 0, 1, 2, \cdots$, let $u(n)$ denote the probability that the random walk hits zero at time n, i.e., $u(n) = \Pr(S_n = 0)$.[7] Clearly, $u(n) = 0$ for all odd n, since the random walk can only return to zero in an even number of steps. Also, since the random walk starts at zero, $u(0) = 1$.

We analyze the distribution of T by relating it to the sequence $u(\cdot)$, which, as we will see, can be computed explicitly. Let $f(n) = \Pr(T = n)$. As before, note that $f(n) = 0$ for all odd n. Also, note that since $T > 0$, $f(0) = 0$.

We first relate $u(\cdot)$ and $f(\cdot)$ as follows. For $n \geq 1$, if the random walk hits zero at time $2n$, then the time T of *first* return to zero is necessarily $\leq 2n$. We can therefore represent the probability $u(2n)$ as

$$u(2n) = \Pr(S_{2n} = 0) = \sum_{j=1}^{n} \Pr(T = 2j) \Pr(S_{2n} = 0 \mid T = 2j).$$

Moreover, note that $\Pr(S_{2n} = 0 \mid T = 2j)$ is simply the probability that the random walk hits zero after starting at zero after $2n - 2j$ steps, i.e., $\Pr(S_{2n} = 0 \mid T = 2j) = u(2n - 2j)$. Therefore, we obtain the following recursive relation for $n \geq 1$.

$$u(2n) = \sum_{j=1}^{n} f(2j) u(2n - 2j) \tag{5.4}$$

We now use the above recursion to relate the Laplace transform of the random variable T to that of the function $u(\cdot)$. Specifically, define

$$\psi_T(s) := \mathbb{E}\left[e^{-sT}\right] = \sum_{n=1}^{\infty} f(2n) e^{-2ns},$$

$$\psi_u(s) := \sum_{m=0}^{\infty} u(m) e^{-ms} = \sum_{n=0}^{\infty} u(2n) e^{-2ns}.$$

[7] It is important to note that when we consider the event $S_n = 0$, the random walk could have hit zero several times before time n. $u(n)$ captures the probability of all such paths.

Now, using (5.4), we may express ψ_u as follows.

$$\begin{aligned}\psi_u(s) &= \sum_{n=0}^{\infty} u(2n)e^{-2ns} \\ &= 1 + \sum_{n=1}^{\infty}\sum_{j=1}^{n} f(2j)u(2n-2j)e^{-2ns}\end{aligned}$$

Interchanging the order of the summations above, we obtain

$$\begin{aligned}\psi_u(s) &= 1 + \sum_{j=1}^{\infty} e^{-2js} f(2j) \sum_{n=j}^{\infty} e^{-s(2n-2j)} u(2n-2j) \\ &= 1 + \psi_T(s)\psi_u(s),\end{aligned}$$

which gives us

$$\psi_T(s) = 1 - \frac{1}{\psi_u(s)}. \tag{5.5}$$

Note that this interchange can be justified via Fubini's theorem.

Now that we have related the Laplace transform of T to that of the sequence $u(\cdot)$, we move on to computing $u(\cdot)$ and its transform explicitly.

For $n \geq 1$, let us consider the event $S_{2n} = 0$. This event simply means that, after starting at zero, the random walk made n positive steps, and n negative steps. Therefore, the total number of paths of the random walk that correspond to the event $S_{2n} = 0$ equals $\binom{2n}{n}$. Since the total number of possible paths after $2n$ steps equals 2^{2n}, and each of these is equally likely, we conclude that

$$u(2n) = \frac{\binom{2n}{n}}{2^{2n}}.$$

Therefore, the transform ψ_u is given by

$$\psi_u(s) = 1 + \sum_{n=1}^{\infty} \frac{\binom{2n}{n}}{2^{2n}} e^{-2ns}.$$

Using the binomial expansion

$$\frac{1}{\sqrt{1-x}} = 1 + \sum_{n=1}^{\infty} \frac{\binom{2n}{n}}{2^{2n}} x^n,$$

we may express ψ_u as

$$\psi_u(s) = \frac{1}{\sqrt{1-e^{-2s}}}.$$

Therefore, using the relation (5.5), we obtain

$$\psi_T(s) = 1 - \sqrt{1-e^{-2s}}.$$

With an explicit expression for the Laplace transform in hand, we can now apply the Theorem 2.9 to deduce the tail behavior of T from the behavior of $\psi_T(s)$ near the origin. Specifically, as $s \downarrow 0$, since $e^{-2s} = 1 - 2s(1 + o(1))$,

$$\begin{aligned}\psi_T(s) &= 1 - \sqrt{1 - (1 - 2s(1 + o(1)))} \\ &= 1 - \sqrt{2s(1 + o(1))} \\ &= 1 - \sqrt{2s}(1 + o(1))\end{aligned}$$

Therefore, from Theorem 2.9,

$$\Pr(T > x) \sim \frac{\sqrt{2}}{\Gamma(1/2)} \frac{1}{\sqrt{x}}.$$

Noting that $\Gamma(1/2) = \sqrt{\pi}$, we have our desired tail characterization. □

5.6 Additional notes

The central limit theorems presented in this chapter are part of a rich literature with a long history that originated in the study of coin tossing-type problems by De Moivre, Gauss and Laplace in the 18th and 19th centuries. Within this literature, the (partial) universality of the Gaussian distribution emerged around 1900 and was the result of contributions from a number of renowned mathematicians including Markov, Lyapunov, Lindeberg and Kolmogorov culminating in the classical central limit theorem. A detailed history can be found in [72].

Our treatment of the central limit theorem focuses on its role as a second-order approximation of additive processes. For more discussion of this viewpoint, see [190]. Given this viewpoint it is natural to consider higher order expansions. It is indeed possible to derive such higher order expansions, and an overview of these can be found in [26, 155, 156].

The process of moving beyond the classical central limit theorem toward the generalized central limit theorem was initiated by Lévy, who was the first to show that stable distributions beyond the Gaussian can can occur when the summands of the random walk have infinite variance [124]. This led to a number of variations of "generalized" central limit theorems for stable distributions such as the one presented in this chapter. Detailed mathematical treatments of these results can be found in [13, 28, 70, 87, 104, 127, 201]; see [64, Section 2.2] for a discussion. A comprehensive history of the development of generalized central limit theorems can be found in [72].

Motivated by the importance of stable distributions for generalized versions of the central limit theorem, stable distributions have become a topic worthy of study in their own right. Lévy was among the first to develop properties of stable distributions, and now there is a rich literature focused on characterizing stable distributions and stable processes, which generalize the Gaussian processes. A comprehensive overview can be found in [169].

We ended the chapter by studying the return times of random walks in Section 5.5. This is a classical topic and a rich theory exists. In particular, we have illustrated the heavy-tailed nature of return times using a simple random walk, but the phenomenon is more general. Power laws persist when the step size distribution is generally distributed, as long as the mean is zero; in this case one typically considers the event of crossing, rather than hitting the zero boundary. More complicated boundary crossing probabilities (particularly time-dependent boundaries) have been considered as well. In all of these cases, heavy-tailed behavior persists. An excellent survey of classical results in this space is [58].

5.7 Exercises

1. To ground the ideas in this chapter it is important to get a feel of the classical and generalized central limit theorems using data.

   (a) *Classical central limit theorem:* Simulate a sequence of i.i.d. Weibull random variables $\{X_i\}$. For some (large) fixed n, generate a large number of samples of $\frac{\sum_{i=1}^{n} X_i - n\mathbb{E}[X_1]}{\sqrt{n\mathrm{Var}(X_1)}}$. Plot a histogram of the data you have generated. Does your data 'look' Gaussian?

   (b) *Generalized central limit theorem:* Simulate a sequence of i.i.d. Pareto random variables $\{X_i\}$ with shape parameter $\alpha \in (1, 2)$. For some (large) fixed n, generate a large number of samples of $\frac{\sum_{i=1}^{n} X_i - b_n}{a_n}$, where the scaling constants $\{a_n\}$ and translation constants $\{b_n\}$ are as obtained in Exercise 6.

      i. Plot a histogram of the data you have generated. Does it look qualitatively different from the histogram from Exercise 1? What are the issues with visualizing your data using a histogram?

      ii. Next, plot the empirical c.c.d.f. of the data, i.e., plot the fraction of data points exceeding x as a function of x. Does the visualization get better?

      iii. Finally, re-plot the above empirical c.c.d.f. using a logarithmic scale for both axes. (You will have to restrict yourself to positive x for this.) Explain the picture you see using the generalized central limit theorem.

2. In this chapter, we gave a number of examples of stable distributions, but did not verify them formally. In this problem you will verify that they are indeed stable. Specifically, prove that the following distributions are stable by checking that they satisfy the condition in Definition 5.1. See Chapter 1 for definitions of these distributions.

   (a) The standard Gaussian

   (b) The Cauchy distribution

   (c) The Lévy distribution

3. The goal of this exercise is to come up with an alternative, equivalent definition of stable distributions that can sometimes be easier to work with. Specifically, prove that a distribution F is stable if and only if, for i.i.d. random variables X_1, X_2 with distribution F, and any constants $a_1, a_2 > 0$, there exist constants $a > 0$ and $b \in \mathbb{R}$ such that

$$a_1 X_1 + a_2 X_2 \stackrel{d}{=} a X_1 + b.$$

4. As mentioned in the chapter, the Gaussian distribution is the limiting distribution for additive processes beyond the setting of the classical central limit theorem. Your task in this problem is to give

an example of this phenomenon. Specifically, construct a distribution having infinite variance that belongs to the domain of attraction of the Gaussian distribution.

5. To get a better understanding of the limiting distributions for the generalized central limit theorem, in this problem your task is to construct specific distributions that lead to α-stable limiting distributions. In particular, for each $\alpha \in (0, 2)$, construct an explicit distribution over the non-negative reals that lies in the domain of attraction of an $\alpha-$stable distribution.

6. To practice applying the generalized central limit theorem, let us consider the case of Pareto random variables. Consider the running sum $S_n = \sum_{i=1}^{n} X_i$ of i.i.d. Pareto random variables $\{X_i\}_{i \geq 1}$ with shape parameter $\alpha \in (0, 2)$. Identify the scaling and translation constants for required for the generalized central limit theorem to apply to the sequence $\{S_n\}_{n \geq 1}$.

7. Suppose that $X_1, X_2, \cdots, X_n$ are i.i.d., standard Cauchy random variables. Prove that $\frac{1}{n} \sum_{i=1}^{n} X_i$ is a standard Cauchy random variable.

   Hint: You might want to exploit the characteristic function of the standard Cauchy.

Chapter 6

Multiplicative processes

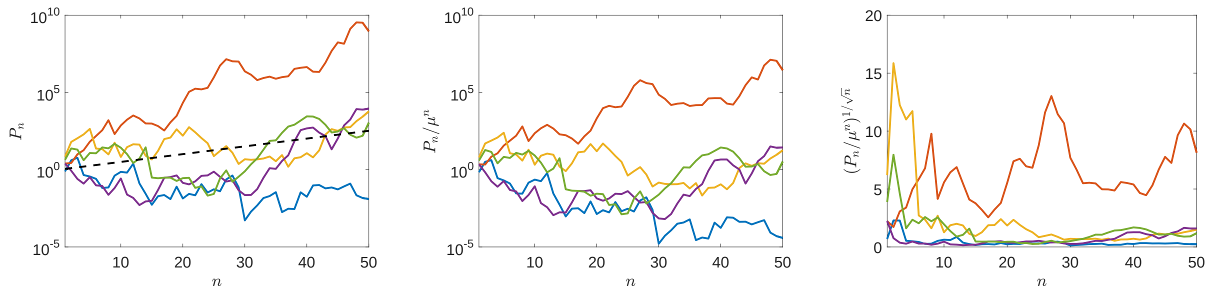
The prevailing view that heavy-tailed phenomena are a surprising curiosity is due, in large part, to the view that the emergence of the Gaussian distribution is almost a "law of nature". As we have discussed, this view stems from the prominence of the central limit theorem, and more generally the view that additive processes frequently govern the growth of processes in the world around us.

Though it is certainly true that additive processes are prominent in the world around us, they are far from the only way for things to evolve. In fact, they are probably not even the most common way for things to evolve. One could easily argue that it is as common, or even more common, for things to evolve according to some form of *multiplicative process*, where growth happens proportionally to the current size. For example, investments and incomes tend to grow multiplicatively since interest rates and raises tend to be in terms of percentages, e.g., a 5% interest rate or a 3% raise, which may vary over time. Another example is the growth of populations: the number of kids that families have on average, which varies over generation to generation, provides a multiplicative growth rate for a population, e.g., if families tend to have 2.5 kids then the population will grow multiplicatively at a rate of 2.5/2=1.25 per generation.

Multiplicative processes are relevant not only for growth processes like the simple examples above, they are also relevant for studying situations involving fragmentation. For example, consider a situation where a stick or rock breaks into two pieces, and then the resulting pieces break into two pieces, and so on. In each case, the resulting pieces represent some fraction of the original item, and so the fragmentation follows a multiplicative process. Such processes have been suggested as models for the evolution of the sizes of meteors (and consequently craters), which tend to break up as they collide with each other, for example.

These simple examples highlight that multiplicative processes tend to be more representative models than additive processes for many parts of our physical and economic world. The same is true for the digital world. For example, when someone watches a video online and likes it, she is likely to share it with friends, and thus the number of views will grow proportionally with the number of people who have watched the video already. Similarly, if someone has a lot of twitter followers, this may lead to a lot of retweets, which then leads to more followers, and so on.

The prevailing theme of these examples is a sense that the "rich get richer", i.e., once something becomes large or popular it tends to grow more quickly, often in proportion to its increased size. If a song has has been downloaded 100,000 times, then it is likely that it will quickly get many more downloads. If one considers the number of links pointing to a website, then as the website is linked to more often, more people are likely to find out about it, and thus the number of links pointing to it will grow even more quickly. This concept of



(a) (b) (c)

Figure 6.1: *Illustration of the multiplicative process P_n. The plots illustrate 5 realizations of P_n with $Y_i \sim \mathrm{Exp}(1/2)$. Figure (a) shows the (unscaled) process itself. Note that the scale on the y-axis is logarithmic, indicating an exponential growth in the process P_n. The dotted line is the first-order approximation μ^n, where $\mu = e^{\mathbb{E}[\log(Y_1)]}$. Figure (b) shows the same process normalized with respect to the first order approximation. Note that the y-axis is still on a logarithmic scale. Finally, Figure (c) shows the process scaled as per the multiplicative central limit theorem. Note that the y-axis is now linear. Under this scaling, Theorem 6.1 shows that the process converges in distribution to the LogNormal distribution.*

"rich get richer" is fundamentally tied to multiplicative processes and heavy tails.

In fact, the examples of multiplicative processes that we have discussed above should sound very familiar, as we have used them often in this book as examples of heavy-tailed distributions. This is not an accident or coincidence. Multiplicative processes are fundamentally tied to the emergence of heavy-tailed distributions. We illustrate and explain this link in this chapter by studying a number of variations of multiplicative processes, including multiplicative processes with barriers, with noise, and the celebrated preferential attachment, a.k.a., Yule process, model. Our study of these variations highlight that different limiting distributions may emerge depending on the specific form of the multiplicative process, but that the emerging distribution from a multiplicative process is nearly always heavy-tailed. This quite different than the behavior of additive processes. Though heavy-tailed distributions can emerge from additive processes, they emerge only when the additive process is made up of infinite variance heavy-tailed distributions. Thus, additive processes do not *create* heavy-tailed distributions from light-tailed distributions, they only *maintain* heavy-tails. In contrast, multiplicative processes can *create* heavy-tailed distributions from light-tailed distributions, and so they provide a much more satisfying explanation for the prevalence of heavy-tailed phenomena.

6.1 The multiplicative central limit theorem

Let us begin our discussion of multiplicative processes with a simple, generic example of a process that grows via the product of random events. In particular, we consider the following multiplicative process:

$$P_n = Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n = P_{n-1} \cdot Y_n,$$

where $P_0 = 1$ and Y_i are i.i.d. and strictly positive. See Figure 6.1(a) for an illustration of such a process.

As in the case of additive processes, our goal is to study the behavior of the multiplicative process P_n as $n \to \infty$. However, to derive intuition for why multiplicative processes are tied to heavy-tailed distributions it is useful to start with some simple examples of small, finite n. In particular, let us first look at the case where the Y_i are exponentially distributed with rate λ. In this case, though the Y_i are light-tailed, P_2 already turns out to be heavy-tailed. To see this, we can use the following simple bound on P_2:

$$\Pr\left(P_2 > t\right) \geq \Pr\left(Y_1 > \sqrt{t}\right)\Pr\left(Y_2 > \sqrt{t}\right) = e^{-2\lambda\sqrt{t}}.$$

The above is already enough to show that P_2 is heavy-tailed.

More generally, the same phenomena happens with distributions that have lighter than exponential tails, but as the tail gets lighter an increasing number of terms are necessary in order to generate a heavy-tail. For example, consider the case where Y_i follows a Weibull distribution with scale parameter β and shape parameter $\alpha > 1$. Then for all $k > \alpha$,

$$\Pr\left(P_k > t\right) \geq \left(\Pr\left(Y_i > t^{1/k}\right)\right)^k = e^{-k\beta t^{\alpha/k}},$$

which guarantees that P_k is heavy-tailed.

The two examples we have looked at so far already show that multiplicative processes behave quite differently from additive processes, and are much more strongly tied to heavy-tailed distributions. In fact, given the two examples above, it is quite natural to expect P_n as $n \to \infty$ to yield a heavy-tailed distribution. This is indeed true and, in fact, it turns out that it is possible to prove this fact via a connection between multiplicative processes and additive processes. In particular, a simple translation of P_n allows its behavior to be described using the law of large numbers and the central limit theorem. The key observation that provides this connection is that the logarithm of a multiplicative process yields an additive process. Specifically, defining $X_i = \log Y_i$ yields

$$\begin{aligned}\log P_n &= \log Y_1 + \log Y_2 + \ldots + \log Y_n \\ &= X_1 + X_2 + \ldots + X_n\end{aligned}$$

A consequence of this translation is that the emergence of heavy-tailed distributions under multiplicative processes is quite natural. It is essentially a consequence of the central limit theorem: the logarithm of a multiplicative process is an additive process and so, instead of the Gaussian distribution, the LogNormal distribution emerges. To show this more concretely, let us walk through the application of the law of large numbers and the central limit theorem. Assuming that the X_i have finite mean $\mathbb{E}\left[X_i\right]$ and finite variance σ^2. Starting with the law of large numbers, we have

$$\log P_n \overset{a.s.}{=} \mathbb{E}\left[X_i\right] n + o(n),$$

which yields

$$P_n \overset{a.s.}{=} e^{\mathbb{E}[X_i]n + o(n)} \overset{a.s.}{=} \left(e^{\mathbb{E}[\log Y_i]}\right)^n e^{o(n)}. \tag{6.1}$$

Just as we saw in the case of additive processes, the law of large numbers provides a first order approximation of the process P_n. However, there are interesting contrasts in the structure of the first order approximation between the two cases. First, for the multiplicative process P_n, the first order approximation is an exponential function $\left(e^{\mathbb{E}[\log Y_i]}\right)^n$ as opposed to the linear approximation of additive processes provided by the law of large numbers. Second, note that the error of this first order approximation is much larger for multiplicative processes than for additive processes. Expressed multiplicatively, the error is $e^{o(n)}$ in the case of multiplicative processes; whereas there is only an additive $o(n)$ error in the case of an additive processes. Figure 6.1(b) provides an illustration of the error of the first order approximation of the multiplicative process P_n that emphasizes these contrasts.

That the first order approximation of a multiplicative process is an exponential should not in itself be surprising. However, what might be surprising is that the base of this exponential approximation is $e^{\mathbb{E}[\log Y_i]}$. Indeed, one might guess from the fact that we are applying the law of large numbers that the dominant growth of P_n should be $\mathbb{E}[Y_i]^n$, since the Y_i are independent and thus $\mathbb{E}[P_n] = \prod_{i=1}^n \mathbb{E}[Y_i]$. However, instead, the dominant growth is $\left(e^{\mathbb{E}[\log Y_i]}\right)^n$, which can be seen to be smaller then $\mathbb{E}[P_n]$ as a consequence of Jensen's inequality.[1] Thus, the mean is not actually a good predictor of the behavior of P_n.

A simple example makes this point clear. Consider Y_i that takes on either $1/2$ or $3/2$ with equal probability. Thus, $\mathbb{E}[Y_i] = 1$ and so $\mathbb{E}[P_n] = 1$ too. However, the median can be far from the mean. To see this, note that the number of Y_i that are equal to $1/2$ follows a Binomial distribution, and thus the median of P_n has $n/2$ samples as 1/2 and 3/2. So, the median of P_n is

$$\left(\frac{1}{2}\right)^{n/2}\left(\frac{3}{2}\right)^{n/2} = \left(\frac{3}{4}\right)^{n/2} < 1 = \mathbb{E}[P_n].$$

In fact, not only is the median of P_n smaller than the mean, the gap between them grows exponentially with n. Thus, the mean is not a good predictor of the behavior of the distribution, which is typical for heavy-tailed distributions.

Of course, we can improve the characterization of P_n provided by the law of large numbers by additionally applying the central limit theorem. In particular, this gives the following approximation for $\log P_n$.

$$\log P_n \stackrel{d}{=} \mathbb{E}[X]\, n + Z\sqrt{n} + o(\sqrt{n}), \text{ where } Z \sim\text{Gaussian}(0, \sigma^2).$$

Again, we can then recover P_n by exponentiating:

$$P_n \stackrel{d}{=} e^{\mathbb{E}[X_i]n + Z\sqrt{n} + o(\sqrt{n})} \stackrel{d}{=} \left(e^{\mathbb{E}[X_i]}\right)^n \left(e^Z\right)^{\sqrt{n}} e^{o(\sqrt{n})}, \text{ where } Z \sim\text{Gaussian}(0, \sigma^2)$$

The above shows that, similarly to the case of additive processes, the combination of the law of large numbers and the central limit theorem gives a second order approximation for the growth of a multiplicative process. But, the form of this approximation illustrates that it is exponentially less accurate for multiplicative processes than the parallel approximation is for additive processes, i.e., the error term is of the order $e^{o(\sqrt{n})}$ instead of $o(\sqrt{n})$. However, this increased error does not affect the use of the approximation to study the limiting behavior of P_n as $n \to \infty$. In particular, by looking at the centered and scaled version of P_n we

[1] Jensen's inequality (see [31]) states that given a random variable X and a function f that is concave over the support of X, $\mathbb{E}[f(X)] \le f(\mathbb{E}[X])$, provided the expectations exist. Taking $f(x) = \log(x)$, this implies that $\mathbb{E}[\log(Y_i)] \le \log(\mathbb{E}[Y_i])$, which implies that $e^{\mathbb{E}[\log(Y_i)]} \le \mathbb{E}[Y_i]$.

can obtain the following characterization of the limiting distribution of multiplicative processes:

$$\left(\frac{P_n}{e^{\mathbb{E}[X_i]n}}\right)^{1/\sqrt{n}} \stackrel{d}{=} e^{Z+o(\sqrt{n})/\sqrt{n}} \xrightarrow{d} e^{Z}, \text{ as } n\to\infty, \text{ where } Z \sim\text{Gaussian}(0,\sigma^2).$$

The above highlights that, when properly normalized, P_n converges to a LogNormal distribution in the limit, and thus the behavior of multiplicative processes is tightly tied to heavy-tailed distributions. See Figure 6.1(c) for an illustration of P_n under the above normalization.

Though the above sketch is not completely rigorous, it can easily be made so, which gives the following statement of the multiplicative central limit theorem. This is sometimes referred to as Gibrat's law, named after Robert Gibrat, who studied multiplicative growth in the context of the growth of firms in the 1930s [85].

Theorem 6.1 (The multiplicative central limit theorem)**.** *Suppose $\{Y_i\}_{i\geq 1}$ is an i.i.d. sequence of strictly positive random variables satisfying Var $[\log Y_i] = \sigma^2 < \infty$ and define $\mu = e^{\mathbb{E}[\log Y_i]}$. Then*

$$\left(\frac{Y_1\cdots Y_n}{\mu^n}\right)^{\frac{1}{\sqrt{n}}} \xrightarrow{d} H, \textit{where } H \sim \textit{LogNormal}(0,\sigma^2).$$

The form of the multiplicative central limit theorem essentially parallels that of the central limit theorem except that the centering and scaling of the process happens in a different manner due to the fact that the processes grows multiplicatively instead of additively. Note that we have stated a version of the multiplicative central limit theorem that uses only the standard central limit theorem, not the generalized central limit theorem. However, because the condition of finite variance applies only to $\log Y_i$ rather than to Y_i itself, this case already includes all common distributions, even many distributions with infinite mean. For example, recall that if Y_i is Pareto distributed than $\log Y_i$ is Exponentially distributed, and thus all of its moments are finite. More generally, a sufficient condition to ensure that Var $[\log Y_i] = \sigma^2 < \infty$ is $\bar{F}_Y(x) = o\left(\frac{1}{\log^2(x)}\right)$, which shows that all regularly varying distributions with non-zero index satisfy the condition – and thus only extremely heavy-tailed distributions are excluded. Of course, if one desires to state a "generalized" multiplicative central limit theorem for that case, the derivation above can be adjusted accordingly.

The contrast between the multiplicative central limit theorem with the central limit theorem for additive processes is striking. While the emergence of heavy-tailed limiting distributions under the central limit theorem requires starting with heavy-tailed distributions with infinite variance, the emergence of heavy-tailed limiting distributions under the multiplicative process is *guaranteed* – even when the Y_i are extremely light-tailed. So, one truly should *expect* heavy-tailed distributions whenever multiplicative growth governs a process.

6.2 Variations on multiplicative processes

The simple form of the multiplicative central limit theorem that we have described so far already highlights the strong connection between multiplicative processes and heavy-tailed distributions, and specifically LogNormal distributions. More broadly, the connection between heavy-tailed distributions and multiplicative processes is extremely robust, but one should not always expect the LogNormal distribution to emerge.

To illustrate the robustness of the connection between multiplicative processes and heavy-tailed distributions, as well as the brittleness of the emergence of the LogNormal distribution, we study two variants of

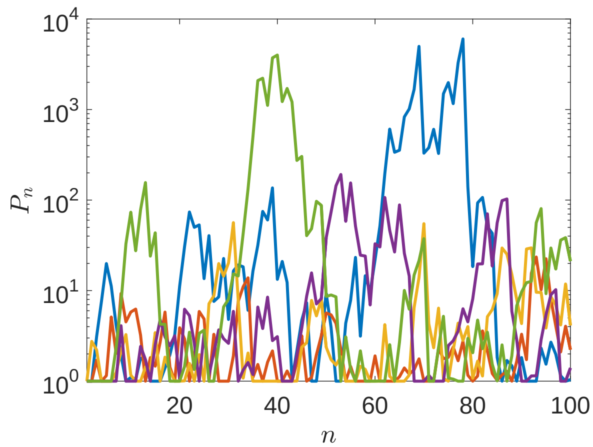


Figure 6.2: *Illustration of a multiplicative process with a lower barrier. The plots show 5 realizations of P_n with $Y_i \sim \mathrm{Exp}(3/4)$ and $l = 1$. Note that the process gets occasionally 'reset' to the lower barrier, but can take large values between successive resets.*

multiplicative processes in this section. The first variation is a multiplicative process that has a strictly positive lower barrier which cannot be crossed, and the second variation is a multiplicative process that includes additive noise at each step. Each of these variations leads to the emergence of heavy-tailed distributions; however in both cases, though the variations are small, the specific heavy-tailed distributions that emerge are no longer LogNormal distributions. Instead, distributions that are approximately power-law emerge. This illustrates that, even though the emergence of the LogNormal distribution under multiplicative processes is brittle, the connection between multiplicative processes and heavy-tailed distributions is robust.

6.2.1 A multiplicative processes with a lower barrier

An important variation of the classical multiplicative process that we have considered so far is to add a lower barrier on the process that ensures that the process never drops below a certain level, $l > 0$. Specifically,

$$P_n = \max(P_{n-1}Y_n, l) = \max(\dots \max(\max(Y_1, l) \cdot Y_2, l) \cdots Y_n, l),$$

where the Y_n are strictly positive i.i.d. random variables and $P_0 = 1$ is the starting point of the process. See Figure 6.2 for a visualization of this process.

This variation of multiplicative processes captures the idea that it is often unreasonable to have the process take on arbitrarily small values. A frequently cited example where this is the case is incomes: if there is a minimum wage, then there is a point beneath which the income can never drop. Similarly, minimum levels also are natural if one considers the population of a city, the number of links to a website, and many other common examples of multiplicative processes.

Initially, it is natural to think that the addition of a lower bound should have little impact on the behavior of the multiplicative process, especially since the unbounded process is already guaranteed to be strictly positive, and so has a lower bound already. Indeed, in many cases it is true that the lower bound plays no role. For example, if the process grows quickly then, no matter what the lower bound, the process quickly grows away from it and is then not influenced by it for large n. Thus, for the lower bound to play a role, it is necessary to impose that the process drifts downward, i.e., that $\mu := e^{\mathbb{E}[\log(Y_i)]} < 1$. Alternatively, there are also some situations where the lower bound plays too strong of a role. For example, if the multiplicative process never grows, then the process simply converges to the lower bound. To avoid this trivial case it is enough to ensure that the process grows with some probability, i.e., that Pr $(Y_i > 1) > 0$.

It turns out that the two conditions we have just described are enough to ensure that the lower bound plays a role, but does not lead to trivial limiting behavior. In particular, they ensure that the multiplicative process P_n neither drifts upward to infinity nor downward to the lower bound. As a consequence, the process no longer needs to be centered or scaled as in the multiplicative central limit theorem. This leads to a slightly simpler form for the multiplicative central limit theorem in this case.

Theorem 6.2. *Consider the multiplicative process with a lower barrier* $P_n = \max(P_{n-1}Y_n, l)$, *where* $\{Y_n\}_{n\geq 1}$ *is an i.i.d. sequence of strictly positive random variables,* $l > 0$ *is a lower barrier enforced on the process, and* $P_0 = 1$. *Suppose that the distribution of* Y_1 *satisfies the following conditions:*

(a) $\mu := e^{\mathbb{E}[\log(Y_1)]} < 1$.

(b) Pr $(Y_1 > 1) > 0$.

(c) $\mathbb{E}[Y_1^s] < \infty$ *for some* $s > 0$.

Then, $P_n \xrightarrow{d} H$, *where* H *is heavy-tailed and has a distribution* F *that satisfies* $\lim_{x\to\infty} -\frac{\log \bar{F}(x)}{\log(x)} = s^*$, *with* $s^* = \sup\{s \geq 0 \mid \mathbb{E}[Y_1^s] \leq 1\}$.

Interestingly, the above theorem highlights that, though it is in some sense a mild variation of the classical multiplicative process, the addition of a lower barrier leads to a qualitatively different limiting behavior. In particular, the LogNormal distribution is no longer the emergent distribution; instead a distribution that is approximately power-law, and thus has a much heavier tail, emerges. The limiting distribution is asymptotically linear on a log-log scale, i.e., $\log \bar{F}(x)/\log x$ converges to a constant. This condition is not satisfied by the LogNormal distribution, but is satisfied by regularly varying distributions, e.g., the Pareto distribution. More specifically, this property nearly corresponds to the class of regularly varying distributions; however it is more general. Recall that Lemma 2.3 states that a regularly varying distribution with index $-\alpha$, has $\lim x \to \infty \log \bar{F}(x)/\log(x) = -\alpha$, but the converse does not hold.

The conditions imposed by Theorem 6.2 appear technical at first glance, but are all quite natural. We have already explained the motivation for conditions (a) and (b) above. Condition (c) is also natural since it is parallel to the condition that Var $[\log Y_i] < \infty$ in the multiplicative central limit theorem. That is, condition (c) ensures the distribution of Y_i is not "too heavy-tailed" for standard analytic techniques to be used, i.e., that the tail of Y_1 is bounded above by a power law, Pr $(Y_1 > x) = o(x^{-\alpha})$ for some $\alpha > 0$. Of course, condition (c) is not particularly restrictive since it already includes distributions that are quite heavy-tailed, e.g., regularly varying distributions with an infinite mean. In particular, the condition corresponds to imposing that $\log Y_i$ is light-tailed.

Proof of Theorem 6.2. As in the case of the multiplicative central limit theorem, the key step in the proof of Theorem 6.2 is to apply a logarithmic transformation on our multiplicative process to convert it into an additive process. This allows us to deduce the limiting behavior of our multiplicative process from analysis of the behavior of random walks.

For simplicity, we prove the result only in the case of $l = 1$. However, the case of general l can be reduced to this case without too much additional effort.

Assuming that $l = 1$, we may inductively express P_n as follows.

$$\begin{aligned} P_n &= \max(\max(P_{n-1}Y_{n-1}, 1)Y_n, 1) \\ &= \max(P_{n-2}Y_{n-1}Y_n, Y_n, 1) \\ &= \max(\max(P_{n-3}Y_{n-2}, 1)Y_{n-1}Y_n, Y_n, 1) \\ &= \max(P_{n-3}Y_{n-2}Y_{n-1}Y_n, Y_{n-2}Y_{n-1}Y_n, Y_{n-1}Y_n, Y_n, 1) \end{aligned}$$

Proceeding in this manner, and noting that $P_0 = 1$, it is easy to see that

$$P_n = \max(Y_1Y_2, Y_3 \cdots Y_{n-1}Y_n,\ Y_2, Y_3 \cdots Y_{n-1}Y_n,\ \cdots,\ Y_{n-1}Y_n,\ Y_n,\ 1).$$

Now, since Y_i are i.i.d., we may write

$$P_n \stackrel{d}{=} \max(1,\ Y_1,\ Y_1Y_2,\ Y_1Y_2Y_3,\ \cdots,\ Y_1Y_2 \cdots Y_{n-1},\ Y_1Y_2 \cdots Y_{n-1}Y_n).$$

At this point, we make the logarithmic transformation $X_i = \log(Y_i)$ for $i \geq 1$, and $Z_n = \log(P_n)$ to obtain

$$Z_n \stackrel{d}{=} \max(S_0, S_1, \cdots, S_n), \tag{6.2}$$

where $S_0 = 0$, and $S_i = \sum_{j=1}^{i} X_j$ for $j \geq 1$. We have thus expressed Z_n as the maximum over n time steps of a random walk starting at zero, with i.i.d. increments X_i. Moreover, our assumptions on the distribution of Y_1 imply that

(a) $\mathbb{E}[X_1] < 0$, i.e., the random walk has a negative drift,

(b) $\Pr(X_1 > 0) > 0$, i.e., the random walk makes positive increments with some probability,

(c) $\mathbb{E}\left[e^{sX_1}\right] < \infty$ for some $s > 0$, i.e., the increments are light-tailed to the right.

Now, it is clear that as $n \to \infty$, $\max(S_0, S_1, \cdots, S_n) \to S_{max} = \max_{j \geq 0}\{S_j\}$, the all time maximum of the random walk, whenever S_{max} is finite. We study S_{max} in detail in Section 7.4, and we can now apply one of the results we prove there. In particular, under the above assumptions on X_1, we show that S_{max} is finite with probability 1, and

$$\lim_{x \to \infty} -\frac{\log \Pr(S_{max} > x)}{x} = s^*, \tag{6.3}$$

where $s^* = \sup\{s \geq 0 \mid \mathbb{E}\left[e^{sX_1}\right] \leq 1\}$ (see Theorem 7.6).

From (6.2), it now follows that as $n \to \infty$,

$$Z_n \xrightarrow{d} S_{max},$$

implying that

$$P_n = e^{Z_n} \xrightarrow{d} e^{S_{max}}.$$

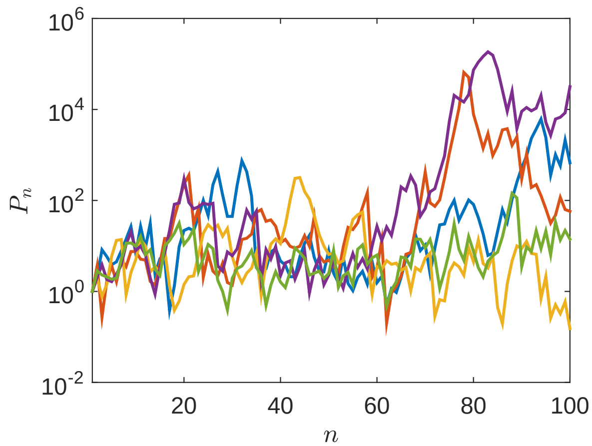


Figure 6.3: *Illustration of a multiplicative process with additive noise. The plots show 5 realizations of P_n with $Y_i \sim \mathrm{Exp}(0.75)$, Q_i being distributed as the absolute value of a standard Gaussian.*

Finally, we note that the distribution F of $e^{S_{max}}$ satisfies

$$\begin{aligned}\lim_{x\to\infty} -\frac{\log \bar{F}(x)}{\log(x)} &= \lim_{x\to\infty} -\frac{\log \Pr\left(e^{S_{max}} > x\right)}{\log(x)} \\ &= \lim_{x\to\infty} -\frac{\log \Pr\left(S_{max} > \log(x)\right)}{\log(x)} = s^*,\end{aligned}$$

where $s^* = \sup\{s \geq 0 \mid \mathbb{E}\left[e^{sX_1}\right] \leq 1\} = \sup\{s \geq 0 \mid \mathbb{E}\left[Y_1^s\right] \leq 1\}$. □

6.2.2 A noisy multiplicative process

We have just seen the phenomenon of multiplicative processes leading to heavy-tailed distributions is robust to a lower barrier. Here, we highlight that it is also robust to additive noisy in the process. In particular, in this section we discuss a "noisy" multiplicative process where there is a small amount of additive noise between the steps of multiplicative growth. Specifically,

$$P_n = P_{n-1}Y_n + Q_n \tag{6.4}$$

where the Y_n are i.i.d. and strictly positive, the Q_n provide additive noise and are i.i.d. and non-negative, and $P_0 = 1$ is the starting point of the process.[2] See Figure 6.3 for a visualization of the above process.

This noisy variation of a multiplicative process captures the idea that there is often an opportunity for small additive changes to occur between multiplicative growth stages. For example, in the case of incomes, the effect of a bonus or a promotion is an additive change that happens between yearly raises; and, in the case of populations, the effect of immigration may be viewed as an additive change that combines with the

[2] We adopt the terminology of "noise" to describe Q_n following Newman [146]. However, because Q_n is non-negative it may be more descriptive to think of Q_n as a "repulsive force" that keeps the process from shrinking to zero, which is the phrasing used in Sornette [178] and Gabaix [80].

multiplicative effect of birth rate. Alternatively, the noise could be a result of error in observations of the growth of some physical process.

Interestingly, the addition of additive noise to a multiplicative process causes a qualitative change to the limiting distribution that emerges. In fact, the impact is the same as in the previous case where we studied the addition of a lower barrier: the limiting distribution is no longer always LogNormal, instead it is approximately power law.

Theorem 6.3. *Consider the noisy multiplicative process $P_n = P_{n-1} \cdot Y_n + Q_n$, where $P_0 = 1$ and $\{Y_n, Q_n\}_{n \geq 1}$ is an i.i.d. sequence of random variable pairs, where Y_n is strictly positive and Q_n is non-negative. Suppose that the distribution of Y_1 satisfies the following conditions.*

(a) $Pr\,(Y_1 > 1) > 0$.

(b) $\mu := e^{\mathbb{E}[\log(Y_1)]} < 1$.

(c) $\mathbb{E}\,[Y_1^s] < \infty$ *for some* $s > 0$

Additionally, suppose that Q_1 is light-tailed and $Pr\,(Q_1 > 0) > 0$. Then, $P_n \xrightarrow{d} F$, where F is heavy-tailed and satisfies $\lim_{x\to\infty} -\frac{\log \bar{F}(x)}{\log(x)} = s^*$, with $s^* = \sup\{s \geq 0 \mid \mathbb{E}\,[Y_1^s] \leq 1\}$.

Note that Theorem 6.3 essentially parallels Theorem 6.2 for the case of a multiplicative process with a lower barrier. The same conditions are sufficient for a limiting distribution to emerge, and the limiting distribution that emerges satisfies the same logarithmic asymptotics. Further, it is again the case that a LogNormal distribution does not emerge, and instead the emerging distribution is approximately power law, i.e., asymptotically linear on a log-log scale.

Though it may not be apparent initially why Theorem 6.3 and Theorem 6.2 parallel each other, there is some intuition for the similarity. Roughly, the additive noise provides a "soft" lower bound on the process. That is, because the noise is i.i.d., the multiplicative process cannot spend much time below the median of Q_n – 1/2 of the noise samples will be at least that large. Thus, though the noise does not give a strict lower barrier, it does provide a significant boost to the process when it is near zero.

However, the intuitive connection between noisy multiplicative processes and multiplicative processes with lower bounds does not translate directly to a proof. In fact, the addition of noise to the process means that a structurally different proof technique is required to prove Theorem 6.3. The consequence is that the full proof is more technical, and we do not include it here. The interested reader can find a proof in [80, 109, 178]. Here, we give a proof of only the special case where $Q_n = 1$, which can be done cleanly and highlights the connection between the case of additive noise and the case of a lower barrier.

Proof of Theorem 6.3 for $Q_n = 1$. We begin by obtaining a clearer characterization of our process for the special case $Q_n = 1$. It is not hard to show from (6.4) that

$$P_n = 1 + Y_n + Y_nY_{n-1} + Y_nY_{n-1}Y_{n-2} + \cdots + Y_nY_{n-1}\cdots Y_1.$$

Thus,

$$P_n \stackrel{d}{=} 1 + Y_1 + Y_1Y_2 + Y_1Y_2Y_3 + \cdots + Y_1Y_2\cdots Y_n.$$

As in the proof of Theorem 6.2, we now make the logarithmic transformation $X_i = \log(Y_i)$ for $i \geq 1$, yielding

$$P_n \stackrel{d}{=} \sum_{k=0}^{n} e^{S_k},$$

where $S_0 = 0$, and $S_i = \sum_{j=1}^{i} X_j$ for $j \geq 1$. It now follows that

$$P_n \stackrel{d}{\to} P = \sum_{k=0}^{\infty} e^{S_k}. \tag{6.5}$$

Having identified the limiting distribution of P_n, it now remains to the characterize the tail behavior of this limiting distribution. Our approach will be to derive asymptotically matching lower and upper bounds on the tail. In both cases there is strong connection to the case of a multiplicative process with a lower barrier. We start with the lower bound.

Lower bound: The lower bound is obtained by noting that (6.5) implies that

$$P \geq P' := \sup_{k\geq 0} e^{S_k}$$

almost surely. This lower bound has a simple interpretation: note that P' is simply the limit of the multiplicative process with a lower barrier defined by $P'_n = \max\{P'_{n-1}Y_n, 1\}$, $P'_0 = 1$. Indeed, it is not hard to see directly that the $P_n \geq P'_n$ for all n. The above bound allows us to invoke Theorem 6.2 for multiplicative processes with a lower barrier to conclude that

$$\liminf_{x\to\infty} \frac{\log \mathbf{Pr}\left(P > x\right)}{\log x} \geq \lim_{x\to\infty} \frac{\log \mathbf{Pr}\left(P' > x\right)}{\log x} = -s^*.$$

Upper bound: The proof of the upper bound is also highly related to the proof of Theorem 6.2. It follows from (6.5) that for $\epsilon > 0$,

$$P \leq e^{\sup_{k\geq 0}[S_k + \epsilon k]} \sum_{k=0}^{\infty} e^{-\epsilon k} = \frac{1}{1 - e^{-\epsilon}}\, e^{\sup_{k\geq 0}[S_k+\epsilon k]} =: c_\epsilon P'_\epsilon.$$

We now note that $\sup_{k\geq 0}[S_k + \epsilon k]$ is simply the supremum of the random walk with increment process $X_n + \epsilon$. Taking ϵ small enough so that $\mathbb{E}\left[X_n + \epsilon\right] < \infty$, we can conclude from Theorem 6.2 that

$$\lim_{x\to\infty} \frac{\log \mathbf{Pr}\left(P'_\epsilon > x\right)}{\log x} = -s^*_\epsilon,$$

where

$$s^*_\epsilon = \max\{s : \mathbb{E}\left[e^{s(X_1+\epsilon)}\right] \leq 1\} = \max\{s : \mathbb{E}\left[Y_1^s\right] \leq e^{-\epsilon s}\}.$$

We thus obtain the upper bound

$$\limsup_{x\to\infty} \frac{\log \mathbf{Pr}\left(P'_\epsilon > x\right)}{\log x} \leq -s^*_\epsilon,$$

To see that the upper bound matches the lower bound, observe that s_ϵ^* is decreasing in ϵ. Furthermore, as $E[Y_1^s]$ is continuous on $(0, s^*)$, it follows that $s_\epsilon^* \uparrow s^*$ as $\epsilon \downarrow 0$, removing the gap between the lower and upper bound. □

6.3 An example: Preferential attachment and Yule processes

So far in this chapter we have focused on generic multiplicative processes without delving into any particular example in depth. To end the chapter, we focus in some depth on one specific, celebrated example of a multiplicative process that comes from the area of *network science*.

Network science has provided some of the most popular, and controversial, examples of heavy-tailed phenomena. There was an enormous amount of excitement surrounding the discovery of heavy-tailed phenomena in complex networks during the late 1990s and early 2000s as large data sets about social, biological, and communication networks began to emerge. This period was particularly exciting due to the diversity of settings where large data sets were becoming available and the apparent "universality" of the features observed in these networks. One of the particularly striking features that was observed repeatedly in complex networks from a variety settings was that the degree distribution tended to be heavy-tailed, often power-law, and thus scale-invariant. Such observations of scale-invariant behavior emerged in the context of social networks, citation networks, biological networks, communication networks, power networks, and beyond.[3]

The emergence of power law degree distributions in diverse complex networks, a.k.a., the emergence of scale-free networks, motivated a deep scientific question that was (and still is) particularly seductive for the network science research community:

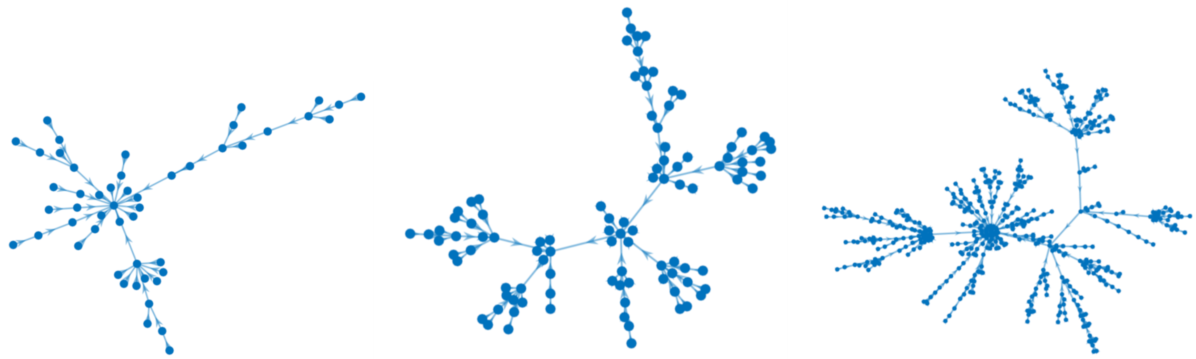
Is there some sort of unifying explanation, or fundamental law, governing the emergence of heavy-tailed, scale-invariant behavior in complex networks?

The seductive nature of this question has led to an enormous literature that seeks to develop and analyze models for the evolution of networks and, at this point, there is an enormous array of models that lead to heavy-tailed degree distributions in networks. For example, [7, 22, 39, 119, 131, 151]; excellent surveys of this space can be found in [24, 146, 184]. These networks use a variety of growth mechanisms; however many of them rely on some form of multiplicative process at their core. To highlight this fact, in this section we describe a particularly celebrated example of such a mechanism: *preferential attachment*.

The preferential attachment mechanism

The preferential attachment mechanism has a long history, and has been reinvented under many guises over the years. In particular, it is essentially equivalent to the growth of the so-called Yule processes [198], which are named after the statistician who first studied them in 1925, Udny Yule. Initially, Yule processes were introduced as a simplified model for studying the sizes of biological taxa; however it did not take long for them to be adopted in other disciplines as well. For example, Herbert Simon [174] applied Yule processes to study the distribution of wealth and, in the process, gave them their name. It did not take long for Yule

[3] As we discuss in Chapter 8, in some cases the initial claims of power-law distributions have since been refuted in favor of either other heavy-tailed distributions (e.g., LogNormal distributions or Weibull distributions) or, in some cases, light-tailed distributions. Examples of such refutations include the internet graph [45, 196] and protein-protein interaction networks [182]. Debate about which networks exhibit power-law degree distributions continues to this day, e.g. [34, 186].



(a) 50 nodes (b) 100 nodes (c) 500 nodes

Figure 6.4: *Illustrations of graphs generated by the preferential attachment model. For clarity, self-edges are not shown. Note that there are a few 'core' high-degree nodes, surrounded by a periphery of many low-degree nodes.*

processes to find application in complex networks: In 1968 Derek Price [160] adopted the Yule process to study the evolution of citation networks, renaming the mechanism "cumulative advantage". More recently, during the excitement surrounding the discoveries of heavy-tailed phenomena in complex networks in the 1990s, Albert-László Barabási and Réka Albert [22] rediscovered Price's mechanism giving it the name that is commonly used today: preferential attachment. It is this name and context that we adopt in the following; however we would like to emphasize that there is nothing inherent in preferential attachment that restricts its application to complex networks, and in fact it has many applications outside of this context.

The preferential attachment mechanism essentially formalizes the "rich get richer" phenomena in the context of a network. In particular, under preferential attachment the network grows in a manner where new edges are "preferentially attached" to vertices that already have high degree, i.e., the "rich" vertices.

More specifically, consider a discrete time process on $t = 0, 1, \ldots, n$ for creating a directed graph. At each time t, one vertex arrives, with a single link to itself, and chooses a pre-existing vertex to connect to via a directed edge from itself to the selected vertex. The key to the preferential attachment mechanism is that the vertex to connect with is chosen with probability proportional to its in-degree.

Formally, we study a model where, with probability $1 - \alpha > 0$, the newly arriving vertex use a preferential attachment mechanism and chooses a pre-existing vertex to connect to proportionally to its in-degree; and, with probability α, the newly arriving vertex chooses a pre-existing vertex uniformly at random. So, $1 - \alpha$ is the probability of preferential attachment and α is the probability of random choice. Thus, the probability that the vertex arriving at time t chooses a certain vertex with in-degree k to connect to equals

$$\frac{\alpha}{t} + \frac{(1-\alpha)k}{2t}.$$

Note that there are $2t$ pre-existing edges in the network when the new vertex arrives at time t.[4]

[4] To make this model consistent at $t = 0$, let us assume that the first node creates 2 links to itself, i.e., it starts with an in-degree of 2. Note that each subsequent node starts with in-degree 1.

Some examples of networks that emerge from the preferential attachment mechanism are shown in Figure 6.4. These examples highlight some of the properties that are commonly associated with preferential attachment. In particular, it is easy to see that there is extreme diversity in the in-degrees of the vertices.[5] A consequence is that high-degree nodes tend to be very "central" to the network and, in particular, the network exhibits a so-called "core-periphery" structure where older, higher degree nodes form a tightly connected core surrounded by a periphery of lower degree vertices. Similarly to heavy-tailed degree distributions, this core-periphery structure is another aspect of complex networks that is, in some sense, "universal".

The degree distribution under preferential attachment

The preferential attachment mechanism is closely tied to the multiplicative processes that we have studied in this chapter. At each time, the probability that a new vertex chooses to connect to a given vertex is proportional to the in-degree of the vertex, and thus the expected increase in the in-degree is proportional to the current degree. As a result, we should *expect* the emergent degree distribution under the preferential attachment mechanism to be heavy-tailed. Indeed this is the case. However, a formal proof does not follow directly from the generic results about multiplicative processes we have discussed to this point in the chapter. It requires a more detailed analysis, which we perform in the following.

More specifically, the following theorem states that the degree distribution that emerges under the preferential attachment mechanism is not just heavy-tailed, but also power-law and scale invariant. To state the theorem, let F_t denote the distribution of the in-degree when the total number of nodes equals t, i.e., $F_t(k)$ equals the probability that a randomly selected node has in-degree $\leq k$, just before the arrival of the node at time t.

Theorem 6.4. *Under the preferential attachment model, $F_t \xrightarrow{d} F$, as $t \to \infty$, where $\bar{F}(x) \sim \beta x^{-\frac{2}{1-\alpha}}$ for some positive constant β.*

Interestingly, this theorem not only shows that the degree distribution under preferential attachment is a power-law, it also gives the dependence of the power-law on the probability of preferential attachment $(1-\alpha)$. In particular, as the probability of preferential attachment increases, i.e., as α decreases, the tail of the limiting degree distribution becomes heavier. Moreover, by varying the parameter α over $[0, 1)$, the above model generates power-law degree distributions with the exponent varying over $[2, \infty)$. However, it is important to note that the specific range on the exponent of the power-law is not a fundamental consequence of preferential attachment itself, it is simply a consequence of the particular way we have chosen to model it. There are a wide variety of variations on the preferential attachment model that lead to different expressions for the power-law exponent (see, for example, [22, 59, 116, 160, 174]). Examples of such model variations include allowing the incoming node to attach to more than one pre-existing node, allowing the pre-existing nodes to add edges to one another in a preferential manner, and so on.

It is also important to point out that the preferential attachment mechanism is only one of many mechanisms that can lead to heavy-tailed phenomena in networks. In particular, though the mechanism provides a simple explanation for the emergence of a heavy-tailed degree distribution, one should be careful to conclude that it is an accurate model for complex networks more broadly. In particular, the simplicity of the preferential attachment mechanism means that many other properties that are associated with complex networks are not captured. And, in fact, if one simply compares the examples of preferential attachment networks

[5] The degree of a vertex is strongly correlated with its arrival time, i.e., older vertices tend to have larger in-degree.

in real-world examples of complex networks it is clear that there are significant differences. As a result, since the introduction of the preferential attachment model, many other examples of generative mechanisms for networks have been proposed, e.g., the configuration model [184], hyperbolic graphs [152], Kronecker graphs [118], the forest-fire model [120], dot-product graphs [197], and many more.

Finally, let us move on to the proof of Theorem 6.4. The proof may appear technical at first glance, but the structure of the proof is actually quite simple.

Proof of Theorem 6.4. The idea of this proof is to first understand the evolution of the average number of nodes with degree k at time t and then use that understanding to study the limiting degree distribution as $t \to \infty$. While the second piece of the proof requires some technical calculation, the first piece is simple and already provides clear intuition as to why preferential attachment is related to multiplicative processes and, thus, why a heavy-tailed degree distribution emerges.

To start, let $m_t(k)$ denote the average number of pre-existing nodes with degree k at time t. Let $p_t(k) = \frac{m_t(k)}{t}$. Note that $p_t(k)$ is the probability that a randomly selected (pre-existing) node at time t has degree k, i.e., $p_t(\cdot)$ is the probability mass function corresponding to the distribution F_t. Our objective is to analyze the behavior of $p_t(\cdot)$ as $t \to \infty$.

Our first step toward this goal is to develop a recursive relation for $m_t(k)$. Consider the newly arriving node at time t. As we have noted, the probability that this node connects to a particular pre-existing node with in-degree k equals $\frac{\alpha}{t} + \frac{(1-\alpha)k}{2t}$. Since there are, on average, $m_t(k)$ nodes with in-degree k, the probability that the newly arriving node connects with a node with in-degree k equals $m_t(k)\left(\frac{\alpha}{t} + \frac{(1-\alpha)k}{2t}\right)$. We can therefore express the evolution of the average number of nodes with in-degree $k > 1$ as follows.

$$\begin{aligned} m_{t+1}(k) &= m_t(k) + m_t(k-1)\left(\frac{\alpha}{t} + \frac{(1-\alpha)(k-1)}{2t}\right) - m_t(k)\left(\frac{\alpha}{t} + \frac{(1-\alpha)k}{2t}\right) \\ &= m_t(k)\left(1 - \frac{2\alpha + (1-\alpha)k}{2t}\right) + m_t(k-1)\left(\frac{2\alpha + (1-\alpha)(k-1)}{2t}\right) \end{aligned} \tag{6.6}$$

Note that the above evolution of the average number of nodes with in-degree k accounts for an increase coming from the possibility that a node with in-degree $k-1$ at the previous time step gained an incoming edge, and a decrease coming from the possibility that a node in-degree k at the previous time step gained an incoming edge. The evolution for the case $k = 1$ is different. Since the increase in the number of nodes with degree 1 comes only from the newly arrived node,

$$\begin{aligned} m_{t+1}(1) &= m_t(1) + 1 - m_t(k)\left(\frac{\alpha}{t} + \frac{(1-\alpha)}{2t}\right) \\ &= m_t(1)\left(1 - \frac{1+\alpha}{2t}\right) + 1. \end{aligned} \tag{6.7}$$

Already, the form of the evolution of $m_t(k)$ already suggests that a heavy-tailed degree distribution is likely. Note that the existence of terms $m_t(k-1)\cdot(k-1)$ highlights the multiplicative nature of the growth process and, given the results in this chapter, that already indicates that a heavy-tailed degree distribution is likely.

Proceeding with the proof, we now use the above recursions for $m_t(k)$ to analyze the asymptotic behavior of $p_t(\cdot)$ as $t \to \infty$. We denote, for $k \geq 1$,

$$p(k) := \lim_{t\to\infty} p_t(k) = \lim_{t\to\infty} \frac{m_t(k)}{t}.$$

We prove that the limit exists later in the proof. Note that $p(k)$ may be interpreted as long-term probability of picking a node with degree k. In other words, $p(\cdot)$ is the probability mass function corresponding to our limiting distribution F.

Our analysis of the behavior of $p_t(\cdot)$ as $t \to \infty$ is based on the technical result, the proof of which is left to the reader as Exercise 7. Consider the recursion

$$a_{t+1} = a_t \left(1 - \frac{b_t}{t}\right) + c_t.$$

If $\lim_{t\to\infty} b_t = b \in (0, \infty)$, and $\lim_{t\to\infty} c_t = c \in \mathbb{R}$, then $\frac{a_t}{t}$ converges as $t \to \infty$, and

$$\lim_{t\to\infty} \frac{a_t}{t} = \frac{c}{1+b}.$$

Applying this result to the recursion (6.7) (taking $a_t = m_t(1)$, $b_t = \frac{1+\alpha}{2}$, and $c_t = 1$) we conclude that $\lim_{t\to\infty} \frac{m_t(1)}{t}$ exists, and

$$p(1) = \lim_{t\to\infty} \frac{m_t(1)}{t} = \frac{1}{1+\frac{1+\alpha}{2}} = \frac{2}{3+\alpha}.$$

Next, we study the asymptotic behavior of $p(k)$ for $k \geq 2$ by applying the same result to the recursion (6.6). To show that $\frac{m_t(k)}{t}$ converges as $t \to \infty$, we can proceed inductively. Suppose that $\frac{m_t(k-1)}{t}$ converges as $t \to \infty$. We can now apply the above result to recursion (6.6), setting $a_t = m_t(k)$, $b_t = \frac{2\alpha+(1-\alpha)k}{2}$, and $c_t = m_t(k-1)\left(\frac{2\alpha+(1-\alpha)(k-1)}{2t}\right)$. This gives that $\frac{m_t(k)}{t}$ converges as $t \to \infty$, and

$$p(k) = \lim_{t\to\infty} \frac{m_t(k)}{t} = \frac{c}{1+\frac{2\alpha+(1-\alpha)k}{2}},$$

where

$$c = \lim_{t\to\infty} m_t(k-1)\left(\frac{2\alpha+(1-\alpha)(k-1)}{2t}\right) = p(k-1)\left(\frac{2\alpha+(1-\alpha)(k-1)}{2}\right).$$

Therefore, we have

$$\begin{aligned} p(k) &= p(k-1)\left[\frac{2\alpha+(1-\alpha)(k-1)}{2\left(1+\frac{2\alpha+(1-\alpha)k}{2}\right)}\right] \\ &= p(k-1)\left[\frac{\frac{2\alpha}{1-\alpha}+k-1}{\frac{2+2\alpha}{1-\alpha}+k}\right]. \end{aligned}$$

The above relation can be iterated to obtain

$$p(k) = p(1)\left[\frac{\left(\frac{2\alpha}{1-\alpha}+1\right)\left(\frac{2\alpha}{1-\alpha}+2\right)\cdots\left(\frac{2\alpha}{1-\alpha}+k-1\right)}{\left(\frac{2+2\alpha}{1-\alpha}+2\right)\left(\frac{2+2\alpha}{1-\alpha}+3\right)\cdots\left(\frac{2+2\alpha}{1-\alpha}+k\right)}\right].$$

The remainder of the argument consists of manipulating the above expression into a more intuitive form. To this end, we can represent $p(k)$ in terms of the gamma function[6] using the property $\Gamma(a) = (a-1)\Gamma(a-1)$ as follows:

$$p(k) = p(1)\left[\frac{\Gamma\left(\frac{2\alpha}{1-\alpha}+k\right)}{\Gamma\left(\frac{2\alpha}{1-\alpha}+1\right)}\,\frac{\Gamma\left(\frac{2+2\alpha}{1-\alpha}+2\right)}{\Gamma\left(\frac{2+2\alpha}{1-\alpha}+k+1\right)}\right].$$

Collecting the factors that do not depend on k into a single constant β', we obtain

$$p(k) = \beta'\frac{\Gamma\left(\frac{2\alpha}{1-\alpha}+k\right)}{\Gamma\left(\frac{2+2\alpha}{1-\alpha}+k+1\right)}.$$

We can now deduce the asymptotic behavior of $p(k)$ using the following property of the gamma function: for fixed y,

$$\frac{\Gamma(x)}{\Gamma(x+y)} \sim x^{-y}.$$[7]

Therefore, we conclude that

$$p(k) \sim \beta'\left(k+\frac{2\alpha}{1-\alpha}\right)^{-\left(1+\frac{2}{1-\alpha}\right)} \sim \beta' k^{-\left(1+\frac{2}{1-\alpha}\right)}.$$

This form shows that the probability mass function of F decays as a power-law. It now follows easily (see Exercise 8) that

$$\bar{F}(k) = \sum_{j>k} p(j) \sim \frac{\beta'(1-\alpha)}{2}k^{-\frac{2}{1-\alpha}},$$

which completes the proof. □

6.4 Additional notes

Multiplicative processes are tightly tied to many generative models for heavy-tailed phenomena. The observation that the simple multiplicative process we consider has a LogNormal limit is quite classical; the book by Aitchison and Brown [6] is an early reference. This idea has been used to claim the emergence of the LogNormal distribution in several domains, including finance (see, for example, [47, 103]), basic

[6] The gamma function, for non-negative z, is defined as $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$. It is easy to see that $\Gamma(0) = 1$, and $\Gamma(z+1) = z\Gamma(z)$. It thus follows that for $n \in \mathbb{N}$, $\Gamma(n+1) = n!$. Thus, the gamma function may be interpreted as an extension of the factorial function over the real line.

[7] See Exercise 6.

sciences like ecology, geology, and atmospheric science (see the edited volume [50]), and more recently, communication networks (see, for example, [5, 102]).

The observation that multiplicative mechanisms with a lower barrier result in a power-law limiting distribution was first made by Champernowne [41]. The additive noise variant was first studied by Kesten [109], who analysed a multidimensional version of the processes we introduce in Section 6.2.2. Kesten's result has been extended and simplified during the past several decades, e.g., see [36, 113, 137]. A recent book devoted to this subject is [37]. These mechanisms have been used to argue the emergence of power laws in distributions of incomes [41], city sizes [80], and stock market returns [123], to name a few.

More generally, all of the multiplicative processes we consider in this chapter are special cases of a general recursion of the form

$$P_{n+1} = \Phi_n(P_n), n \geq 0, \tag{6.8}$$

with $\Phi_n, n \geq 0$ a sequence of i.i.d. random functions satisfying a linear growth condition, i.e., $\Phi_n(x) = A_n x + o(x)$ as $x \to \infty$. Under certain conditions (especially on the distribution of the slope A_n), it has been shown that power laws appear for processes of this form. A seminal paper on this topic is due to Goldie, [89]. These results have further been extended to settings with branching mechanisms in [108].

To illustrate the application of multiplicative processes in a specific application setting, we discussed the preferential attachment generative model for complex networks. This model has a long history and has been reinvented under many guises. The model is usually credited to Yule, who used the model to study biological taxa in 1925 [198]. However, Pólya actually used the same model a few years earlier in the study of his celebrated urn model [63]. Later, Simon applied them to the modeling of the populations of cities and the distribution of wealth in the 1950s [174]. In the 60s, Price applied the model to citation networks [160] and, most recently, Barabási and Albert suggested that the model could be used more generally to study complex networks such as the internet graph and the world wide web [22].

Our focus on preferential attachment should not suggest that it is the only (or even the most realistic) generative model for complex networks that exhibits a heavy-tailed degree distribution. Heavy-tailed degrees are only one of many salient features that have been observed in complex networks and the preferential attachment mechanism does not generate networks that have other important properties, such as (i) searchability, the ability for agents to find short paths quickly [111]; (ii) densification, the fact that complex networks often become more dense as they grow [121]; and (iii) shrinking diameters, the fact that the diameter of complex networks often shrinks as the network grows [121].

There are a wide variety of other probabilistic generative models that have been presented, e.g., the configuration model [184], hyperbolic graphs [152], Kronecker graphs [118], the forest-fire model [120], dot-product graphs [197], and many more. Nearly all of these can be tied to some form of multiplicative process, as was the case for the preferential attachment model.

There are also many other classes of generative models that are used to provide explanations for heavy-tailed phenomena in complex networks. One particularly important class of models is based on optimization rather than randomness. This direction was pioneered by Mandelbrot, who argued that power-laws form a sort of information-theoretic optimal for language [131]. In particular, Mandelbrot argued that power-law frequency of word lengths optimizes the average amount of information per letter in a language such as English. This idea was later adopted and extended to a variety of other settings, e.g., Carson and Doyle suggest models for file sizes and forest fires [39] and Fabrikant et al. suggest an application to modeling the internet graph [67].

Interestingly, Mandelbrot's optimization-based model and Simon's preferential attachment model were introduced around the same time and there was a heated debate between the two about the contrasting assumptions and explanations of heavy-tailed behavior; see [132–134, 175–177]. This was mirrored decades later as there was again a heated debated between the preferential attachment model and optimization based approaches in the early 2000s; see [125, 196].

6.5 Exercises

1. The goal of this exercise is to get a feel for the multiplicative mechanisms described in this chapter using data. Let $\{Y_i\}_{i\geq 1}$ be a sequence of i.i.d. exponential random variables with mean $1/\lambda$. Figure out how to set λ to make $\mu = e^{\mathbb{E}[\log(Y_1)]} < 1$. (Jensen's inequality should come in handy.) Also figure out how to set λ to make $\mu = e^{\mathbb{E}[\log(Y_1)]} < 1$.

   (a) *Simple multiplicative process:* Set λ such that $\mu = e^{\mathbb{E}[\log(Y_1)]} > 1$. For some (large) fixed N, generate a large number of samples of $\left(\frac{P_N}{\mu^N}\right)^{1/\sqrt{N}}$, where $P_0 = 1$, $P_n = P_{n-1} \cdot Y_n$ for $n \geq 1$. Does your data look LogNormal, as predicted by the multiplicative central limit theorem? You might want to make a logarithmic transformation on your data to see this.

   Repeat the above exercise by setting λ such that $\mu < 1$.

   (b) *Multiplicative process process with a lower barrier:* Set λ such that $\mu = e^{\mathbb{E}[\log(Y_1)]} < 1$. For some (large) fixed N, generate a large number of samples of P_N, where $P_0 = 1$, $P_n = \max(P_{n-1} \cdot Y_n, 1)$ for $n \geq 1$.

   Plot the empirical c.c.d.f. of the data (i.e., the fraction of data points exceeding x as a function of x) using a logarithmic scale on both axes. Is your visualization consistent with the conclusion of Theorem 6.2?

   (c) *Multiplicative process process with additive noise:* Set λ such that $\mu = e^{\mathbb{E}[\log(Y_1)]} < 1$. For some (large) fixed N, generate a large number of samples of P_N, where $P_0 = 1$, $P_n = P_{n-1} \cdot Y_n + Q_n$ for $n \geq 1$, where $\{Q_n\}_{n\geq 1}$ is a sequence of i.i.d. uniform random variables taking values in $[0, 1]$. Once again, plot the empirical c.c.d.f. of the data using a logarithmic scale on both axes. Is your visualization consistent with the conclusion of Theorem 6.3?

2. Show that the LogNormal distribution is *stable under multiplication*. Specifically, for $n \geq 2$, if $Y_1, Y_2, \cdots Y_n$ are i.i.d. LogNormal random variables, show that there exist positive constants $c, d > 0$ such that

$$\prod_{i=1}^{n} Y_i \overset{d}{=} c Y_1^d.$$

3. Consider the simple multiplicative process defined in Section 6.1. Prove that

   (a) If $\mu > 1$, then $P_n \to \infty$ almost surely as $n \to \infty$

   (b) If $\mu < 1$, then $P_n \to 0$ almost surely as $n \to \infty$

   Now, construct an example where as $n \to \infty$, $P_n \to 0$ almost surely, but $\mathbb{E}[P_n] \to \infty$. How can you explain this phenomenon?

4. The goal of this exercise is gain some experience in computing the power-law tail index s^* that emerges in the limit of a multiplication process with a lower barrier (Theorem 6.2) or with additive noise (Theorem 6.3). Specifically, compute $s^* := \{s \geq 0 \mid \mathbb{E}[Y^s] \leq 1\}$ for the following examples (you should also verify that these examples are consistent with assumptions of Theorems 6.2 and 6.3).

   - For $p > 1/2$,
   $$Y = \begin{cases} 1/2 & \text{w.p. } p \\ 2 & \text{w.p. } (1-p) \end{cases}$$
   - $Y \sim \mathrm{LogNormal}(\mu, \sigma^2)$, where $\mu < 0$

5. The goal of this exercise is to get a feel for the preferential attachment model. For $\alpha \in [0, 1]$, write a program for generating a random graph with n nodes using the preferential attachment model presented in Section 6.3.

   (a) Plot the graphs you obtain for $n = 10, 100, 1000$. Do you see the "rich get richer" phenomenon?

   (b) For large enough n, plot the empirical c.c.d.f. of the in-degree distribution on a log-log scale. Is your visualization consistent with the statement of Theorem 6.4?

6. For $y \in \mathbb{N}$, show that as $x \to \infty$,
$$\frac{\Gamma(x)}{\Gamma(x+y)} \sim x^{-y}.$$
Note: The above statement also holds for all positive real y, though this is harder to prove.

7. The goal of this exercise is to prove the following result, which was used in the proof of Theorem 6.4.

   Consider a sequence $\{a_t\}$ that satisfies the recursion
$$a_{t+1} = a_t \left(1 - \frac{b_t}{t}\right) + c_t,$$
where $\lim_{t\to\infty} b_t = b \in (0, \infty)$ and $\lim_{t\to\infty} c_t = c \in \mathbb{R}$. Prove that
$$\lim_{t\to\infty} \frac{a_t}{t} = \frac{c}{1+b}.$$

8. The goal of this exercise is to prove the following result which was used in the proof of Theorem 6.4. It shows that if a probability mass function is asymptotically power law, then the corresponding c.c.d.f. is also asymptotically power law. Note that this may be viewed as a highly simplified Karamata theorem.

   Let p_X denote the probability mass function of a random variable X taking integer values. As $k \to \infty$, if $p_X(k) \sim ck^{-(1+\gamma)}$, where $c, \gamma > 0$, show that $\bar{F}_X(k) \sim \frac{c}{\gamma} k^{-\gamma}$.

   Hint: You can bound $\sum_{m>k} m^{-(1+\gamma)}$ as follows.
$$\int_{k+1}^{\infty} x^{-(1+\gamma)} dx \leq \sum_{m>k} m^{-(1+\gamma)} \leq \int_{k}^{\infty} x^{-(1+\gamma)} dx.$$

Chapter 7

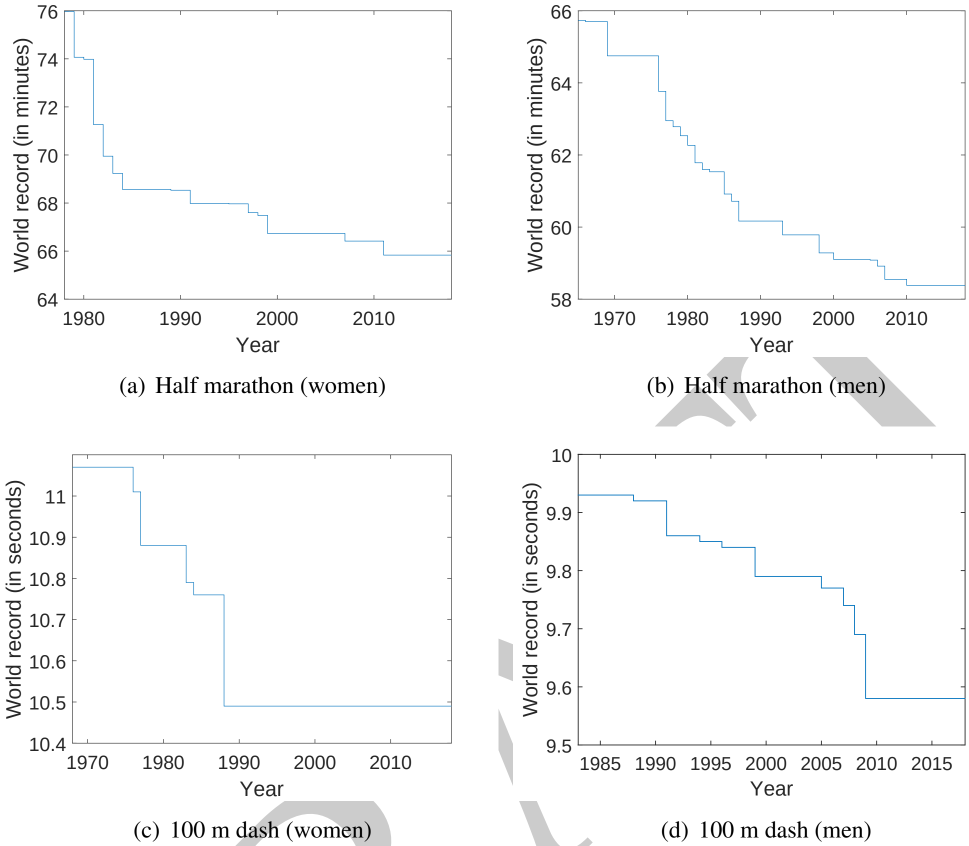
Extremal processes

Extreme events, the good and the bad, shape our lives dramatically. Catastrophic events such as large earthquakes, hurricanes, floods, and stock market crashes have huge societal tolls, and exhilarating events such as world record breaking times at the Olympics leave a sense of amazement. Though such events are almost always unexpected, as a society we must plan for them carefully – dams and dikes must be built in anticipation of possible flooding, buildings must be built to withstand the largest earthquakes, and insurance companies must be prepared for any number of catastrophic events. These situations motivate the importance of understanding the behavior of *extremal processes*, which grow as the maximum (or minimum) of a sequence of events. Further, as we illustrate in Part III of this book, extremal processes are crucial tools for statistical analysis of heavy-tailed distributions. In particular, they are key to the analysis of procedures for fitting the tail of heavy-tailed distributions.

Though extremal processes are more exotic than additive and multiplicative processes, we all have strong intuition about their behavior based on our life experiences. In particular, most typically we think of extremal processes as being analogous to the evolution of world record times. For example, consider the evolution of the world record for the half marathon (Figures 7.1(a) and 7.1(b)). Initially, the world record dropped quickly, sometimes by over a minute every few years, but recently the world record has tended to drop only a minute per *decade*. A similar slowing of the progression of world record times is evident for other events too, e.g. the 100m dash (Figures 7.1(c) and 7.1(d)). We see related phenomena in many other aspects of our lives as well. Consider the progression of the size of the largest snowfall you have experienced or the height of the tallest mountain you have climbed. Thus, it is natural to assume that extremal processes have a decreasing rate of change and may even eventually asymptote to some limiting value – thus they seem to be light-tailed.

This intuition leads to the view that heavy-tailed distributions are surprising curiosities in the context of extremal processes. Of course, at this point in the book we have already seen how heavy-tailed distributions can emerge in the context of both additive and multiplicative processes, and so it should not be surprising that they are more than just curiosities here too. In particular, there are many places where extremes yield heavy-tailed distributions, e.g., the progression of size of the largest stock market crashes, earthquakes, and floods. As we show in Figure 7.2, these examples can exhibit very different behavior than the progression of world records.

The goal of this chapter is to highlight why the emergence of heavy-tailed phenomena in the context of extremal processes should not be surprising. Interestingly, the behavior of extremal processes closely



(a) Half marathon (women) (b) Half marathon (men)

(c) 100 m dash (women) (d) 100 m dash (men)

Figure 7.1: *Progression of world record as ratified by the International Association of Athletics Federations (IAAF).*

parallels the behavior of additive processes, which we study in Chapter 5. Specifically, as we saw for additive processes, extremal processes can yield either light-tailed limiting distributions or heavy-tailed limiting distributions. In the following we introduce the "extremal central limit theorem", which characterizes when heavy-tailed distributions can emerge from extremal processes similarly to the characterization of additive processes provided by the generalized central limit theorem. Further, analogously to the class of stable distributions that we discussed in the context of additive processes, we highlight a class of max-stable distributions that precisely characterizes which distributions can emerge as the limit of an extremal process.

Interestingly, though there are many parallels between additive and extremal processes, one important distinction between the two relates to the emergence of heavy-tailed distributions. Heavy-tailed distributions can emerge from additive processes, but only when one starts with infinite-variance heavy-tailed distributions. In contrast, under extremal processes, heavy-tailed distributions can emerge when starting from distributions with finite variance. In fact, heavy-tailed distributions can emerge even when starting from bounded, light-tailed distributions. Thus, like multiplicative processes, extremal processes can *create* heavy-tailed distributions. However, extremal processes are not as closely connected with heavy-tailed distributions as

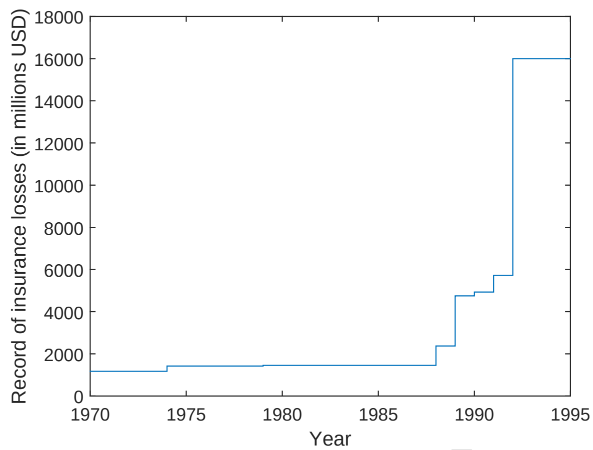


Figure 7.2: *Record of insurance loss due to catastrophies worldwide in millions of US Dollars (at 1992 prices) between 1970–1995 [162].*

multiplicative processes and there are many situations where (extremely) light-tailed distributions can also emerge from extremal processes.

7.1 A limit theorem for maxima

The study of extremal processes is typically referred to as "extreme value theory", and has a long history. Our focus in this chapter is on only a small piece of this theory. In particular, we focus on a generic one-dimensional extremal process of the following form:

$$M_n = \max(X_1, X_2, \cdots, X_n), \text{ where } X_i \text{ are i.i.d. with distribution } F.$$

This simple process corresponds to the evolution of the sample maxima; however it is straightforward to translate our discussion into results for the evolution of the sample minima via the relation

$$\min(X_1, X_2, \cdots, X_n) = -\max(-X_1, -X_2, \cdots, -X_n).$$

Our goal in this chapter is to characterize the behavior of M_n as $n \to \infty$, and to use this simple process to develop intuition about the connection between heavy tails and extremal processes. In some sense, studying extremal processes is easier than studying additive or multiplicative processes. Unlike the cases of additive and multiplicative processes, it is easy to explicitly write down the distribution of M_n in a compact, closed form:

$$\Pr(M_n \le t) = \Pr(X_1 \le t, \ldots, X_n \le t) = F(t)^n.$$

This highlights a simple, but important observation about the limiting distribution of M_n: extremes happen only at the upper end of the distribution. In particular, define x_F as the upper end of the support of

F, i.e., $x_F = \sup\{x : F(x) < 1\}$ and consider a fixed t. Then

$$\lim_{n\to\infty} \Pr(M_n \le t) = \lim_{n\to\infty} F(t)^n = \begin{cases} 0, & \text{if } t < x_F; \\ 1, & \text{if } t = x_F. \end{cases}$$

This calculation illustrates that, if the distribution has a finite upper bound, e.g., the Uniform distribution, then M_n converges with probability 1 to this upper bound. Similarly, if the distribution has no finite bound, the above illustrates that for any finite value t, M_n is eventually larger than t with probability 1.

Of course, one does not learn much about the behavior of the maxima from statements like the above. In fact, those statements are no more informative than simply stating that the additive process $S_n = X_1 + \ldots + X_n$ approaches $+/-\infty$ as $n \to \infty$ depending on whether X_i has positive/negative mean. Just as the law of large numbers and the central limit theorem tell us *how S_n scales* to $+/-\infty$, here we seek a similar understanding of the behavior of M_n. Specifically, the goal is to obtain a more fine grained understanding of the behavior of the M_n, e.g., by understanding the rate at which M_n grows with n and the typical deviations of M_n around this growth rate. Thus, as we did for additive and multiplicative processes, we are interested in studying the centered and normalized maxima, deriving convergence results of the form

$$\frac{M_n - b_n}{a_n} \xrightarrow{d} Z, \text{ where } a_n > 0, b_n \in \mathbb{R}. \tag{7.1}$$

Clearly, this form parallels the form of the cental limit theorems for additive and multiplicative processes we have discussed in the previous two chapters. Like in those cases, one can think of this form as providing a second order approximation of M_n as

$$M_n \stackrel{d}{=} b_n + a_n Z + o(a_n).$$

However, unlike in those cases, it is not easy to anticipate the proper choices of a_n and b_n nor the limiting distribution of Z.

To obtain intuition for the behavior of the centered and normalized version of M_n it is useful to consider

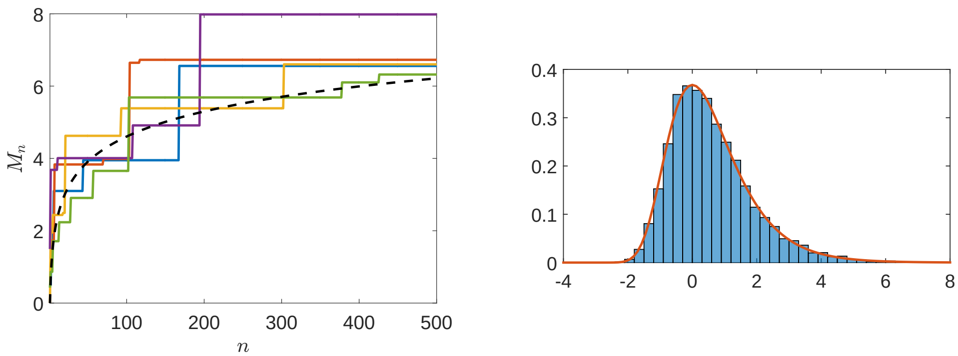


Figure 7.3: *Illustration of an extremal process. (a) shows five realizations of M_n with $X_i \sim \text{Exp}(1)$. The dotted line is the first order approximation $\log(n)$. (b) shows a histogram of the error $M_n - \log(n)$ of the first first order approximation for $n = 500$ over 5000 realizations. The orange curve is the probability distribution function of the Gumbel distribution.*

a simple example. For this, let us consider the Exponential distribution.

Example: Exponential distribution
For the case of the Exponential distribution we have a sequence X_i with $F(t) = 1 - e^{-t}$, as illustrated in Figure 7.3. In this case we can immediately write down the limit we're interested in. This gives

$$\text{Pr}\left((M_n - b_n)/a_n \leq t\right) = F(a_n t + b_n)^n = (1 - e^{-a_n t - b_n})^n.$$

It is not immediate to see how to choose a_n and b_n to obtain a meaningful limit, but with some foresight one can see that $a_n = 1$ and $b_n = \log n$ is a natural choice to take advantage of the fact that $(1-1/n)^{-n} \to 1/e$. In particular, with these choices, we obtain

$$\text{Pr}\left(M_n - B_n \leq t\right) = F(t + \log n)^n = (1 - e^{-t-\log n})^n = \left(1 - \frac{e^{-t}}{n}\right)^n \to e^{-e^{-t}} \text{ as } n \to \infty.$$

This highlights that a limiting distribution for M_n does indeed emerge as $n \to \infty$ and, though you may not recognize this immediately, the limiting distribution is the Gumbel distribution Γ (see Figure 7.3). The c.d.f. of the Gumbel distribution, typically denoted by Λ, is defined by

$$\Lambda(x) = e^{-e^{-x}} \quad (x \in \mathbb{R}).$$

Note that to obtain this limiting distribution we chose $b_n = \log n$, and so $M_n = \log(n) + Z + o(1)$. This form nicely aligns with the intuition we described at the start of the chapter, which suggests that extremes tend to change quickly initially before eventually settling down. The connection with this intuition is strengthened further by the fact that the Gumbel distribution is extremely light-tailed. In fact, the left tail of the Gumbel decays as a double exponential, and the right tail decays exponentially: $\bar{\Lambda}(x) = e^{-x} + o(e^{-x})$ as $x \to \infty$.

The emergence of the Gumbel distribution in the example above provides a candidate for the limiting distribution of extremal processes more broadly than just the case of the Exponential distribution. To investigate this further, let us now repeat the calculation above starting with the Gumbel distribution instead of the Exponential distribution.

$$\Pr\left(M_n - \log n \le t\right) = F(t + \log n)^n = \left(e^{-e^{-t-\log n}}\right)^n = e^{-ne^{-t-\log n}} = e^{-e^{-t}}$$

Note that we again obtain the Gumbel distribution, and this time the equality holds for all finite n, not just in the limit. This highlights that the Gumbel distribution is "stable" with respect to maxima, similarly to the way that the Gaussian distribution is "stable" with respect to sums. Thus, it is natural to define a class of "max-stable" distributions in a parallel manner to the class of stable distributions that we introduced for additive processes in Chapter 5.

Definition 7.1. *A distribution F is said to be max-stable if, for any $n \ge 2$ i.i.d. random variables $X_1, X_2, \cdots, X_n$ with distribution F, there exist constants $c_n > 0$ and $d_n \in \mathbb{R}$ such that*

$$\max(X_1, X_2, \cdots, X_n) \stackrel{d}{=} c_n X_1 + d_n.$$

A random variable is said to be max-stable if its distribution function is max-stable.

Like the class of stable distributions, the class of max-stable distributions is difficult to understand from the definition alone. We have seen that the Gumbel distribution is one example, but it is not immediately clear which other distributions satisfy the definition. However, it is clear that the class of max-stable distributions is tightly coupled to the form of the limit theorems for extremal processes that we are looking for, i.e., to (7.1). In particular, it is easy to see that a max-stable distribution can occur as the limiting distribution in (7.1), and it turns out that max-stable distributions are the *only* non-degenerate distributions that can appear in such a limit. More specifically, just like the class of stable distributions characterizes precisely those distributions that can serve as the limiting distribution of additive processes; max-stable distributions characterizes precisely those distributions that can serve as the limiting distribution of extremal processes.

Theorem 7.1. *A random variable Z is max-stable if and only if there exists an infinite sequence of i.i.d. random variables $X_1, X_2, \cdots$, and deterministic sequences $\{a_n\}$, $\{b_n\}$ ($a_n > 0$), such that*

$$\frac{\max(X_1, X_2, \cdots, X_n) - b_n}{a_n} \xrightarrow{d} Z.$$

In a sense, Theorem 7.1 gives a central limit theorem for extremal processes since it characterizes exactly which limiting distributions can emerge. However, alone, it is quite vague and unsatisfying since it does not provide insight into the properties of the limiting distributions or when different limiting distributions may emerge. In this way, Theorem 7.1 for max-stable distributions exactly parallels Theorem 5.3 for stable distributions. In fact, the parallels between the two theorems is more than just superficial – the proofs of the two results parallel one another as well.

Proof sketch of Theorem 7.1. To prove Theorem 7.1, we first show that if F is a max-stable distribution, then it is the limit, in distribution, of a centered, normalized extremal process. To do this, let $\{X_i\}_{i\ge 1}$

denote an i.i.d. sequence of random variables with distribution F. By Definition 7.1, for any $n \geq 2$, there exist constants $c_n > 0$, $d_n \in \mathbb{R}$ such that

$$\max(X_1, X_2, \cdots, X_n) \stackrel{d}{=} c_n X_1 + d_n.$$

In other words, for any $n \geq 2$,

$$\frac{\max(X_1, X_2, \cdots, X_n) - d_n}{c_n} \stackrel{d}{=} X_1.$$

It therefore follows trivially that as $n \to \infty$,

$$\frac{\max(X_1, X_2, \cdots, X_n) - d_n}{c_n} \xrightarrow{d} F.$$

Next, we show that if the distribution F is the limit in distribution of a centered, normalized extremal process, then F is max-stable. Accordingly, suppose that

$$\frac{\max(X_1, X_2, \cdots, X_n) - b_n}{a_n} \xrightarrow{d} F,$$

where $\{X_i\}_{i\geq 1}$ is an i.i.d. sequence of random variable, and $\{a_n\}$, $\{b_n\}$ are deterministic sequences satisfying $a_n > 0$. Now, fix integer $k \geq 2$, and define, for $m = jk$, $j \in \mathbb{N}$,

$$Y_m = \frac{\max(X_1, X_2, \cdots, X_m) - b_m}{a_m},$$
$$Z_m = \frac{\max(X_1, X_2, \cdots, X_m) - b_j}{a_j}.$$

Consider the limit as $m \to \infty$ by taking $j \to \infty$. Clearly, $Y_m \xrightarrow{d} F$. On the other hand, note that Z_m is the maximum of k i.i.d. random variables, each distributed as $\frac{\max(X_1,X_2,\cdots,X_j)-b_j}{a_j}$. Therefore, $Z_m \xrightarrow{d} F^k$. Moreover, since Y_m and Z_m differ only via translation and scaling parameters, it can be shown that their limiting distributions also only differ via translation and scaling parameters. In other words, F^k and F differ only via translation and scaling parameters. Since this is true for all $k \geq 2$, it then follows from Definition 7.1 that F is max-stable.

To formalize the above argument, one has to show that F is neccesarily a continuous function, and invoke a technical result relating the limits of sequences of random variables that are themselves related via translation and scaling (see [64, Appendix A1.5]). The interested reader is referred to [64, Chapter 2] for the details. $\square$

While we have so far seen parallels between limits of additive and extremal processes, there there are also some important differences. For additive processes, the generalized central limit theorem (Theorem 5.8) states that a non-degenerate limiting distribution exists for most commonly encountered distributions of the summands X_i (with suitable normalization coefficients). In contrast, there are several commonly encountered distributions of X_i for which no non-degenerate limiting distribution is possible for the extremal process $\{M_n\}$, for any choice of normalization coefficients. For example, if X_i has an atom at its right endpoint, or is geometric, or Poisson; see Exercises 3–5. In this sense, limits of extremal process are more fragile than those of additive processes.

7.2 Understanding max-stable distributions

To this point, we have characterized the limiting distributions for extremal processes via the class of max-stable distributions. But, beyond knowing that the Gumbel distribution is an example of a max-stable distribution, we do not yet understand much about max-stable distributions or about when different max-stable distributions may emerge from extremal processes. So, in order to move toward a general statement of the extremal central limit theorem, we need to first develop a more detailed understanding of the class of max-stable distributions.[1]

In Chapter 5 we were able to characterize the class of stable distributions explicitly via their characteristic functions using the representation in Theorem 5.4. The fact that the representation was in terms of the characteristic function was natural given that stable distributions emerge from an additive process and the sums of independent random variables can be cleanly represented as the product of their characteristic functions. In the case of extremal processes, the maximum of independent random variables has a clean representation in terms of the product of their distribution functions. Thus, it is natural for a representation theorem to be in terms of the distribution function. Of course, this is appealing since it is much more direct to infer distributional properties from the distribution function than from the characteristic function. In particular, the class of max-stable distributions has a very concrete representation, given by the following theorem.

Theorem 7.2. *A non-degenerate random variable X is max-stable if and only if $X \stackrel{d}{=} aZ + b$ where $a > 0$, $b \in \mathbb{R}$, and the distribution function of Z is one of the following.*

(i) Fréchet: $\Phi_\alpha(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \exp\{-x^{-\alpha}\} & \text{for } x > 0 \end{cases} \quad (\alpha > 0)$

(ii) Weibull: $\Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^{\alpha}\} & \text{for } x \leq 0 \\ 1 & \text{for } x > 0 \end{cases} \quad (\alpha > 0)$

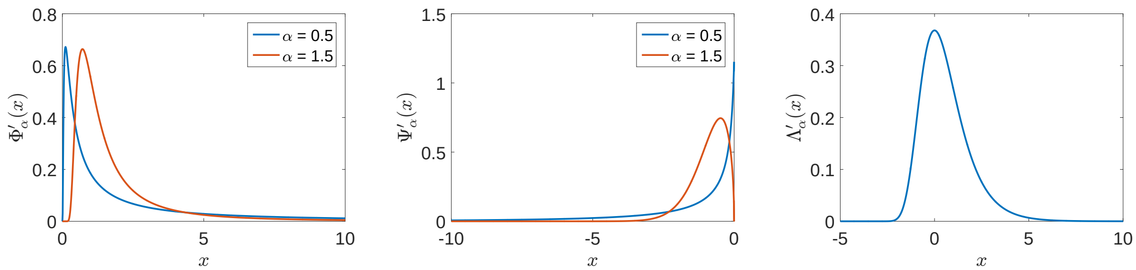
(iii) Gumbel: $\Lambda(x) = \exp\{-e^{-x}\}$.

This representation of max-stable distributions leads the Fréchet, Weibull, and Gumbel distributions to be termed the "extreme value distributions." Interestingly, these three distributions have quite different properties, as is illustrated in Figure 7.4, which shows the p.d.f. of each. For example, while the Fréchet is a heavy-tailed distribution, the Gumbel is a light-tailed distribution, and the Weibull can be either heavy-tailed or light-tailed. Similarly, the supports of the three distributions also differ.

More specifically, the Fréchet distribution, denoted by Φ_α, is parameterized by $\alpha > 0$, has support $[0, \infty)$, and has a tail that satisfies

$$\bar{\Phi}_\alpha(x) = 1 - (1 - x^{-\alpha} + o(x^{-\alpha})) = x^{-\alpha} + o(x^{-\alpha}).$$

[1]Degenerate distributions are trivially max-stable, but they are not particularly interesting to consider and so we focus on non-degenerate distributions throughout this chapter.



(a) Fréchet (b) Weibull (c) Gumbel

Figure 7.4: *Probability density functions corresponding to the standard max-stable distributions*

So, $\bar{\Phi}_\alpha(x) \sim x^{-\alpha}$ as $x \to \infty$, and thus is a regularly varying distribution with index $-\alpha$.

In contrast, the Weibull distribution, denoted by Ψ_α is parameterized by $\alpha > 0$, and has support $(-\infty, 0]$. Note that Ψ_α is actually the mirror image of the Weibull that we have typically discussed in this book. That is, we have so far defined the Weibull distribution as having support $[0, \infty)$, and defined by $F_\alpha(x) = 1 - e^{-x^\alpha}$ for $x \geq 0$. However, all the properties that we associate with the Weibull still hold true. Most importantly, Ψ_α is heavy-tailed (to the left) when $0 < \alpha < 1$, and light-tailed (to the left) when $\alpha \geq 1$.

Finally, the Gumbel distribution, $\Lambda(x)$ has support over the full real line and, as we have already discussed, is light-tailed both to the right and to the left. Interestingly, the left tail decays doubly-exponentially fast, i.e., $\Lambda(-x) = e^{-e^x}$, while the right tail decays exponentially, i.e., $\bar{\Lambda}(x) = 1 - (1 - e^{-x} + o(e^{-x})) = e^{-x} + o(e^{-x})$.

Though the three extreme value distributions that make up the max-stable class behave very differently, it turns out that they are actually quite connected. In particular, the following lemma highlights that the Weibull and Gumbel distributions can be represented as simple functions of the Fréchet distribution. These relationships play a crucial role in the proof of the extremal central limit theorem, which we sketch in the next section.

Lemma 7.1. *Consider a random variable X. The following statements are equivalent.*

1. *X has the Fréchet distribution Φ_α*
2. *$-1/X$ has the Weibull distribution Ψ_α*
3. *$\log(X^\alpha)$ has the Gumbel distribution Λ*

We leave the proof of Lemma 7.1 as an exercise for the reader, see Exercise 6.

Another connection between the three extreme value distributions can be developed by included all three extreme value distributions in one parametric form via

$$H(x) = \exp\left\{ -\left(1 + \xi \frac{x - \mu}{\psi}\right)^{-1/\xi} \right\}. \tag{7.2}$$

This is often referred to as the *generalized extreme value distribution*. If $\xi > 0$ this corresponds to the Fréchet distribution (i.e., the regularly varying case, with $\xi = 1/\alpha$), while $\xi < 0$ corresponds to the Weibull distribution. Using the fact that $(1+\xi y)^{-1/\xi} \to e^{-y}$ as $\xi \to 0$, we see that $\xi = 0$ corresponds to the Gumbel distribution. This representation is used heavily in Chapter 9 when applying extremal techniques to estimate the tail behavior of a distribution from data.

7.3 The extremal central limit theorem

At this point we have seen that the class of max-stable distributions precisely corresponds to the limiting distributions of extremal processes, and we have seen that the class of max-stable distributions is made up of three families of distributions: the Fréchet, the Weibull, and the Gumbel. However, we have not yet understood when each of these limiting distributions can emerge from extremal processes. Formally, just like the generalized central limit theorem specifies the domain of attraction for stable distributions in the context of additive processes, we would like to specify the *maximum domain of attraction* (MDA) for each of the max-stable distributions. These results are provided by the following extremal central limit theorem.

Theorem 7.3 (Extremal central limit theorem)**.** *Consider an infinite sequence of i.i.d. random variables $X_1, X_2, \dots$ with distribution F. Let $x_F := \sup\{x : \bar{F}(x) > 0\}$ denote the right endpoint of F. There exist deterministic sequences $\{a_n\}, \{b_n\}$ $(a_n > 0)$ such that*

$$\frac{\max(X_1, X_2, \cdots, X_n) - b_n}{a_n} \xrightarrow{d} G$$

if and only if G is max-stable. Further,

(i) G follows the Fréchet distribution Φ_α if and only if $\bar{F}(x) = x^{-\alpha} L(x)$, where L is slowly varying;

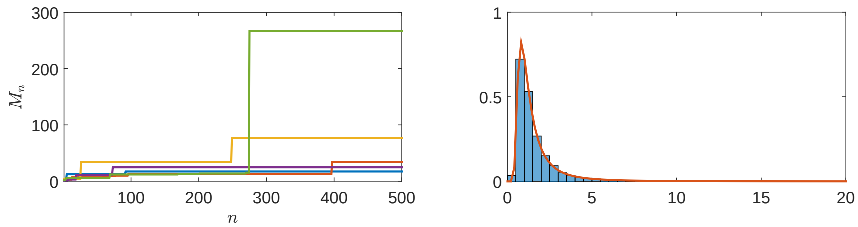
(ii) G follows the Weibull distribution Ψ_α if and only if $x_F < \infty$, and $\bar{F}(x_F - 1/x) = x^{-\alpha} L(x)$, where L is slowly varying;

(iii) G follows the Gumbel distribution Λ if and only if there exists $z < x_F$, such that F satisfies $\bar{F}(x) = c(x)\exp\left\{-\int_z^x \frac{\beta(t)}{g(t)} dt\right\}$ for $x \in (z, x_F)$, where $\lim_{x \uparrow x_F} c(x) = c \in (0, \infty)$, $\lim_{x \uparrow x_F} \beta(x) = 1$, and g is a positive, absolutely continuous function satisfying $\lim_{x \uparrow x_F} g'(x) = 0$.

See Figures 7.5, 7.6, and 7.7 for an illustration of the extremal central limit theorem under the three regimes described in Theorem 7.3.

To some extent, the extremal central limit theorem takes on an expected form: for the heavy-tailed Fréchet distribution to emerge, it must be that the X_i are already heavy-tailed, specifically regularly varying. This is similar to the statement of the generalized central limit theorem, except that for additive processes heavy-tailed distributions can emerge only from distributions that already have infinite variance, while whether the variance is finite or infinite is irrelevant for extremal processes.

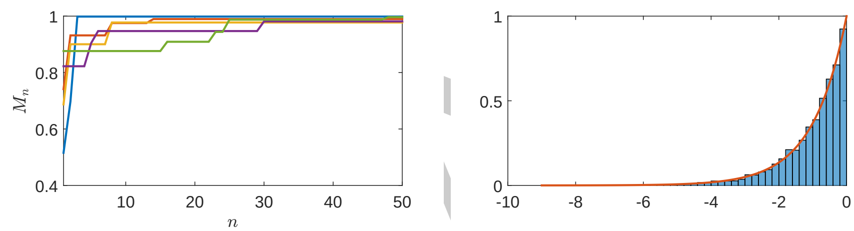
Another difference is that heavy-tailed distributions can emerge from extremal processes even in some situations where the X_i are light-tailed. In particular, the MDA for the Weibull distribution highlights that a heavy-tailed distribution (Ψ_α with $\alpha < 1$) can emerge even when the X_i have bounded support. Thus, heavy-tailed distributions are, in some sense, more likely to emerge under extremal processes than under additive processes.



(a) Illustrations of five sample paths of extremal process

(b) Histogram of $\frac{M_n}{\sqrt{n}}$ for $n = 500$ over 5000 realizations. The orange line is the density function of Φ_2.

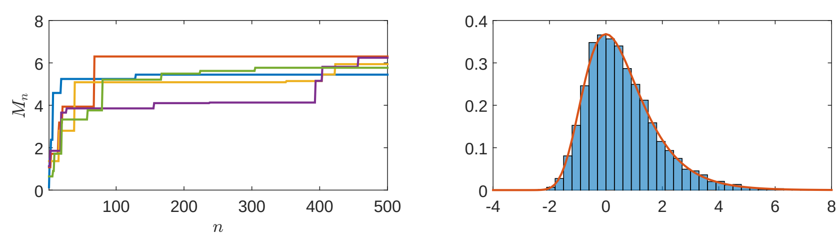
Figure 7.5: $X_i \sim \text{Pareto}$ *($x_m = 1$, $\alpha = 2$), which lies in the MDA of the Fréchet distribution*



(a) Illustrations of five sample paths of extremal process

(b) Histogram of $n(M_n - 1)$ for $n = 500$ over 5000 realizations. The orange line is the density function of Ψ_1

Figure 7.6: $X_i \sim \text{Uniform}([0, 1])$, *which lies in the MDA of the Weibull distribution*



(a) Illustrations of five sample paths of extremal process

(b) Histogram of $M_n - \log(n)$ for $n = 500$ over 5000 realizations. The orange line is the density function of Λ

Figure 7.7: $X_i \sim \text{Exp}(1)$, *which lies in the MDA of the Gumbel distribution*

However, light-tailed distribution can also emerge under extremal processes. In fact, the MDA of the Gumbel distribution highlights that a light-tailed distribution can emerge even when the X_i are heavy-tailed, as long as the tail decays faster than regularly varying tails. To see this, let us contrast the representation given in case (iii) for $x_F = \infty$ with the representation theorem for regularly varying functions (Theorem 2.6). The two have the same form except that in the representation for $\mathcal{RV}(\rho)$, $\beta(t) \to \rho$ and $g(t) = t$. But, since $g(t) = o(t)$ for distributions in the the MDA of the Gumbel distribution, it follows that these distributions have much lighter tails than regularly varying distributions; see Exercise 7.

To keep the statement of the extremal central limit theorem from becoming bloated, we have not described explicitly how to define the scaling and centering sequences, $\{a_n\}$ and $\{b_n\}$, for each case. However, these can also be defined explicitly. To state them, we need to use the generalized inverse, $F^{\leftarrow}$, which is defined as $F^{\leftarrow}(t) = \inf\{x \in \mathbb{R} \mid F(x) \geq t\}$ for $t \in (0, 1)$. Given this definition, we have that if F belongs to the MDA of the Fréchet distribution, Φ_α, then

$$\frac{M_n}{F^{\leftarrow}(1 - 1/n)} \xrightarrow{d} \Phi_\alpha.$$

In contrast, if F belongs to the MDA of the Weibull distribution, Ψ_α, then

$$\frac{M_n - x_F}{x_F - F^{\leftarrow}(1 - 1/n)} \xrightarrow{d} \Psi_\alpha.$$

Finally, if F belongs to the MDA of the Gumbel distribution, Λ, then

$$\frac{M_n - F^{\leftarrow}(1 - 1/n)}{g(F^{\leftarrow}(1 - 1/n))} \xrightarrow{d} \Lambda,$$

where g is defined as in Theorem 7.3, case (iii).

While the MDA of the Fréchet distribution is easy to understand, it is instructive to further explore MDA of the Weibull and the Gumbel distribution.

The MDA of the Weibull distribution is intimately tied to that of the Fréchet distribution. Indeed, the distribution F with right endpoint $x_F < \infty$ belongs to the MDA of the Ψ_α if and only if the distribution G defined by $\bar{G}(x) = \bar{F}(x_F - 1/x)$ belongs to the the MDA of Φ_α. This connection is consistent with the relationship between Ψ_α and Φ_α given in Lemma 7.1; see Exercise 8. The simplest example of a distribution that belongs to the MDA of the Weibull distribution is the uniform distribution. If $F \sim \mathrm{Uniform}([a, b])$, then $\bar{F}(x) = \frac{b-x}{b-a}$ for $x \in [a, b]$, which implies that for large enough x,

$$\bar{F}(b - 1/x) = \frac{1}{x(b - a)}.$$

It follows that F lies in the MDA of Ψ_1. Another common distribution in the MDA of the Weibull is the beta distribution (see Exercise 9).

The MDA of the Gumbel distribution includes distributions with a finite right endpoint, as well as distributions with $x_F = \infty$ that have lighter tails than regularly varying distributions. While we have already contrasted the representation in case (iii) with the representation theorem for regularly varying distributions, it is also instructive to contrast it with the the representation of the complementary c.c.d.f. in terms of the hazard rate (see Equation (4.6) in Chapter 4). Setting $c(\cdot) \equiv c$, $\beta(\cdot) \equiv 1$, we see that $a(x)$ becomes the reciprocal of the hazard rate function $q(x)$. It follows that a sufficient condition for F to be in the MDA of the Gumbel distribution is

$$\lim_{x \uparrow x_F} \frac{d}{dx}\left(\frac{1}{q(x)}\right) = 0. \tag{7.3}$$

For example, consider the (non-negative) Weibull distribution, characterized by the c.c.d.f. $\bar{F}(x) = e^{-\beta t^{\alpha}}$, for $\beta, \alpha > 0$. In this case, it is not hard to see that $q(x) = \alpha\beta t^{\alpha-1}$ satisfies (7.3), which means that the class of (non-negative) Weibull distributions belongs to the MDA of the Gumbel distribution. The condition (7.3) is actually quite powerful — it can be used to show that a large class of distributions, including the Erlang, the hyperexponential, the Gaussian, and the LogNormal, belong to the MDA of the Gumbel distribution (see Exercises 10–13). Note that this MDA includes heavy-tailed as well as light-tailed distributions. For an example of a distribution with a finite right endpoint that lies in the MDA of the Gumbel, consider the distribution F defined as:

$$\bar{F}(x) = e^{-\frac{\alpha}{x_F - x}} \qquad x \in (-\infty, x_F),$$

where $\alpha > 0$. An elementary application of the condition (7.3) confirms that F lies in the MDA of the Gumbel. For an in-depth understanding of the relationship between (7.3) and the MDA of the Gumbel, we refer the reader to [64, Chapter 3].

It is important to not get confused by the appearance of heavy-tailed distributions (like the LogNormal) in the MDA of the Gumbel distribution, which is itself light-tailed. For example, taking X_i in our definition of an extremal process to be LogNormal, it follows from Lemma 4.4 that for any n, $M_n = max(X_1, X_2, \cdots, X_n)$ is long-tailed. On the other hand, we see that the limiting distribution of M_n (with suitable centering and scaling) is light-tailed (specifically, Gumbel)! However, note that the former statement holds for any fixed n, and characterizes the asymptotic behavior of $\Pr(M_n > t)$ as $t \to \infty$. On the other hand, the latter statement involves the limit as $n \to \infty$ for fixed t. In particular, for finite (but large) n, the Gumbel approximation would apply to the *body* of M_n, whereas Lemma 4.4 provides an approximation for its *tail*.

Finally, let us move to the proof of the extremal central limit theorem. As in the case of the generalized central limit theorem, the full proof is too technical for inclusion here. However, a proof of a restricted version of the theorem is possible using elementary techniques.

Sketch of the proof of Theorem 7.3. In what follows, we show that distributions that satisfy the conditions of case (i) of Theorem 7.3 belong to the MDA of Φ_α. Using similar aguments, it can be shown that distributions that satisfy the conditions of case (ii) of Theorem 7.3 belong to the MDA of Ψ_α (see Exercise 8). It is harder to prove the converse implications of the above, and to tackle the MDA of the Gumbel distribution. The interested reader is referred to [64, 165] for the details.

Suppose that F satisfies the condition $\bar{F}(x) = x^{-\alpha}L(x)$, with L slowly varying. To prove that F belongs to the maximum domain of attraction of the Fréchet distribution Φ_α, it suffices to prove that

$$\frac{M_n}{a_n} \xrightarrow{d} \Phi_\alpha, \tag{7.4}$$

where $a_n := F^{\leftarrow}(1 - 1/n)$. Note that $a_n = \inf\{x \in \mathbb{R} \mid \bar{F}(x) \leq n^{-1}\}$, which implies that $a_n \to \infty$ as $n \to \infty$. Moreover, it can be shown that $\bar{F}(a_n) \sim n^{-1}$ (this is trivial if F is continuous; proving this statement for general F is the goal of Exercise 14).

For $x > 0$,

$$\Pr\left(\frac{M_n}{a_n} \leq x\right) = F(a_n x)^n = (1 - \bar{F}(a_n x))^n.$$

Since F is regularly varying with index $-\alpha$, $\bar{F}(a_n x) \sim x^{-\alpha}\bar{F}(a_n) \sim x^{-\alpha}n^{-1}$. It then follows that

$$\lim_{n\to\infty} \Pr\left(\frac{M_n}{a_n} \leq x\right) = e^{-x^{-\alpha}}.$$

For $x \leq 0$,

$$\Pr\left(\frac{M_n}{a_n} \leq x\right) = F(a_n x)^n \leq F(0)^n.$$

Since $F(0) < 1$, it follows that

$$\lim_{n\to\infty} \Pr\left(\frac{M_n}{a_n} \leq x\right) = 0,$$

which completes the proof of (7.4).

□

7.4 An example: Extremes of random walks

Until this point in the chapter we have considered a generic extremal process and the results we have presented highlight that the behavior of extremal processes parallels the behavior of additive processes in the sense that both light-tailed and heavy-tailed limiting distributions may emerge depending on the distribution of the X_i. However, unlike the case of an additive process, the results illustrate that heavy-tailed distributions may emerge from extremal processes even when the X_i are bounded, light-tailed distributions.

To end the chapter, we move from a generic extremal process to a specific example of an extremal process that is important to a wide variety of disciplines. In particular, we return to the setting of random walks and study the behavior of *extremes* of random walks.

Random walks are one of our favorite examples, and we have discussed them already in Section 3.4, where we analyzed the probability that a random walk exhibits a large deviation from its expected value, and in Section 5.5, where we focused on the return time of a symmetric random walk where the walker was equally likely to take a unit step up or down at each time.

The random walks we have considered have been simple so far, but random walks come in a variety of forms and, in general, random walks can be quite complex, e.g., they can be asymmetric; they can allow steps to be of different sizes; and they can happen in more than one-dimension. In this section, we consider a random walk that is asymmetric and allows arbitrary step sizes, but is still one-dimensional. In particular, we consider a random walk in which a walker starts at 0 and takes a sequence of i.i.d. steps $X_1, \ldots, X_n$. The position of the walker at time n is thus simply the additive process

$$S_n = X_1 + X_2 + \ldots + X_n, \text{ for } n > 0 \text{ and } S_0 = 0.$$

The first natural question to ask about such a random walk is "where will the walker be after n steps?", i.e., "what is the behavior of S_n?" Since S_n is simply an additive process, our discussion in Chapter 5 provides the tools to address this question. However, this is only one of many questions one may ask about a random walk. Two other important questions that are often asked are "when will the walker first return to its starting point?" and "what is the maximum position the walker visited?". Since the focus of this chapter is on extremal processes, it is the second of these that we focus on here. Note that we have discussed the first of these already in Section 5.5.

It is important to begin by formalizing the question we are asking. To do this, denote the all time maximum of the random walk by

$$S_{max} = \max_{n \geq 0}\{S_n\}.$$

Our goal will be to characterize the distribution of S_{max}. Note that while our earlier analyses in this chapter focused on the maximum of i.i.d. random variables, S_{max} is the maximum of the positions S_n of the random walk, which are neither independent, nor identically distributed. Because of this, a precise characterization of S_{max} is beyond our reach, and we focus on understanding the behavior of the tail of S_{max}, i.e., $\Pr(S_{max} > x)$ as $x \to \infty$.

As a first step toward characterizing the maximum position of a random walk, consider the examples shown in Figure 7.8. These highlight that understanding the maximum is simple if the random walk has positive drift, i.e., if $\mathbb{E}[X_i] > 0$. In such cases, the random walk grows unboundedly with probability 1, and so there is no finite maximum as we let $n \to \infty$. Thus, it makes sense to focus on random walks with negative drift, i.e., with $\mathbb{E}[X_i] < 0$. In this case, the random walk approaches $-\infty$ as $n \to \infty$ with probability 1, which means that S_{max} is well defined and finite with probability 1.

Another observation from the examples in Figure 7.8 is that there seems to be a distinction between the behavior of S_{max} under light-tailed X_i and under heavy-tailed X_i. This is not too surprising given the results we have described already in the chapter. We have already seen that light-tailed or heavy-tailed limiting distributions can emerge from extremal processes depending on the distributions of X_i. Indeed a similar phenomena happens here, and the behavior of S_{max} differs depending on the weight of the tail of X_i. So, in the following we treat these two cases separately.

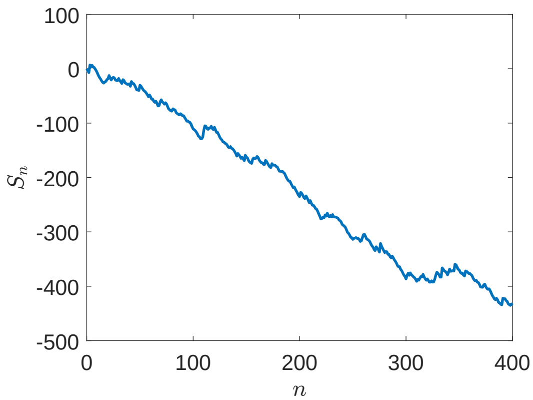
Heavy-tailed step sizes

We start with the case where X_i is heavy-tailed, specifically regularly varying. In this case, based on the extremal central limit theorem, it is natural to expect the maximum of the random walk to also be heavy-tailed. So, the more interesting component of the result that follows is the characterization of the tail itself. In particular, it is interesting to see how the tail of S_{max} relates to the tail of X_i.

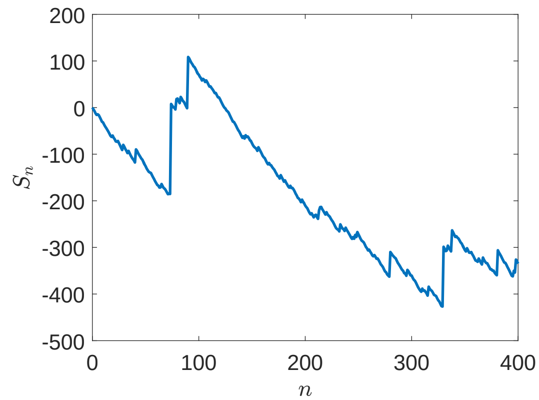
Theorem 7.4. *Consider a one-dimensional random walk S_n with heavy-tailed i.i.d. step sizes X_i such that $\mathbb{E}[X_i] = -a < 0$ and $X_i \in \mathcal{RV}(-\alpha)$, where $\alpha > 1$. Then, the all time maximum $S_{max} = \max_{n\geq 0}\{S_n\}$ satisfies*

$$Pr(S_{max} > x) \sim \frac{1}{a(\alpha-1)} x Pr(X_i > x).$$

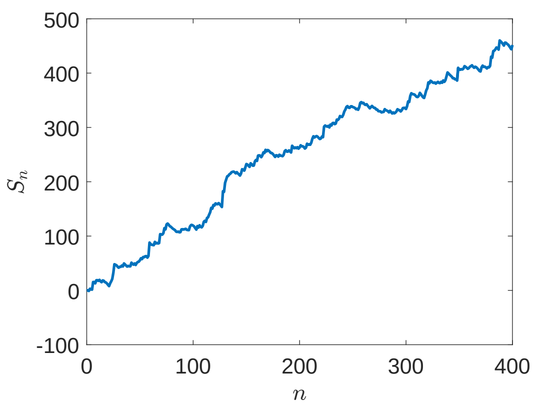
The key point made by the result above is that the tail of S_{max} is one degree heavier than the tail of X_i. Thus, for example, if X_i has an infinite variance and finite mean, then S_{max} has an infinite mean. This highlights that the maximum of the random walk is significantly more variable than the step size. It is interesting to observe that this result is in contrast to the extremal central limit theorem. Specifically, in the case where $X_i \in \mathcal{RV}(-\alpha)$, the extremal limit theorem gives that the scaled M_n converges to a Fréchet



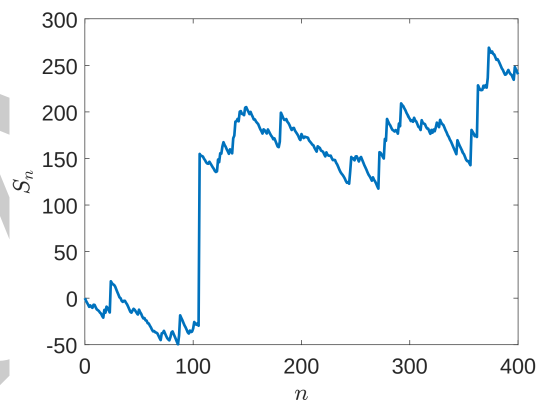
(a) Random walk with negative draft, light-tailed increments; $X_i \sim \text{Exp}(1/4) - 5$ so that $\mathbb{E}[X_i] = -1$



(b) Random walk with negative draft, heavy-tailed increments; $X_i \sim Y_i - 7$ where Y_i is Pareto with $\alpha = 1.5$, $x_m = 2$, so that $\mathbb{E}[X_i] = -1$



(c) Random walk with positive draft, light-tailed increments; $X_i \sim \text{Exp}(1/4) - 3$ so that $\mathbb{E}[X_i] = 1$



(d) Random walk with positive draft, heavy-tailed increments; $X_i \sim Y_i - 5$ where Y_i is Pareto with $\alpha = 1.5$, $x_m = 2$, so that $\mathbb{E}[X_i] = 1$

Figure 7.8: *Illustrations contrasting the behavior of random walks with negative/positive drifts and light-tailed/heavy-tailed increments.*

distribution, Φ_α, which is regularly varying with the same index $-\alpha$. Thus, the dependency inherent in S_n is leading S_{max} to have a heavier tail than predicted by the extremal central limit theorem.

Theorem 7.4 is proved by establishing asymptotically matching upper and lower bounds on $\Pr(S_{max} > x)$. Below, we provide a proof of the lower bound, which confirms the insight that a large all time maximum is most likely caused by a single big jump in the random walk. The proof of the matching upper bound is more involved. Here, we only prove a weaker upper bound that is of the same order as the lower bound. In other words, our upper bound is asymptotically within a constant factor of the lower bound. For a proof of the precise upper bound, we refer the reader to [199].

Our upper bound is based on the following maximal inequality (see [156]).

Theorem 7.5. *Consider an independent sequence of random variables $\{Y_i\}_{i\geq 1}$, and the associated random walk $\{W_n\}_{n\geq 1}$, where $W_n := \sum_{i=1}^n Y_i$. For a given $m \in \mathbb{N}$, if there exist constants $C, q > 0$ such that*

$$Pr\left(W_m - W_k > -C\right) \geq q \quad \forall\, k \in \{1, 2, \cdots, m-1\},$$

then

$$Pr\left(\max_{1\leq k\leq m} W_k > x\right) \leq \frac{1}{q} Pr\left(W_m > x - C\right) \quad \forall\, x.$$

Sketch of the proof of Theorem 7.4. We break the proof into two parts, first showing the lower bound and then moving to the upper bound.

Lower Bound: The idea behind the lower bound is to make use of Karamata's Theorem (Theorem 2.4) to bound $\Pr\left(S_{max} > x\right)$. A key step in accomplishing this is the following result, which is a consequence of the weak law of large numbers: given $\epsilon, \delta > 0$, there exists $L > 0$ such that

$$\Pr\left(S_n > -L - n(a+\epsilon)\right) \geq 1 - \delta \quad \forall\, n \geq 0.$$

We leave the proof of this statement as an exercise for the reader (see Exercise 15).

Using this result, we can now bound $\Pr\left(S_{max} > x\right)$ in terms of the running maximum of the random walk, $S_{max}^{(n)} := \max_{0\leq k\leq n} S_n$, as follows.

$$\begin{aligned}
\Pr\left(S_{max} > x\right) &= \sum_{n=0}^{\infty} \Pr\left(S_{max}^{(n)} \leq x,\ S_{n+1} > x\right) \\
&\geq \sum_{n=0}^{\infty} \Pr\left(S_{max}^{(n)} \leq x,\ S_n > -L - n(a+\epsilon),\ X_{n+1} > x + L + n(a+\epsilon)\right) \\
&= \sum_{n=0}^{\infty} \Pr\left(S_{max}^{(n)} \leq x,\ S_n > -L - n(a+\epsilon)\right) \bar{F}(x + L + n(a+\epsilon)) \\
&\geq \sum_{n=0}^{\infty} \left(\Pr\left(S_{max}^{(n)} \leq x\right) - \delta\right) \bar{F}(x + L + n(a+\epsilon)) \\
&\geq \left(\Pr\left(S_{max} \leq x\right) - \delta\right) \sum_{n=0}^{\infty} \bar{F}(x + L + n(a+\epsilon)) \\
&\geq \frac{\Pr\left(S_{max} \leq x\right) - \delta}{a+\epsilon} \int_{x+L}^{\infty} \bar{F}(s)ds
\end{aligned}$$

Noting that Karamata's theorem implies that $\int_x^\infty \bar{F}(s) \sim x\bar{F}(x)$, we obtain the lower bound:

$$\liminf_{x\to\infty} \frac{\Pr\left(S_{max} > x\right)}{x\bar{F}(x)} \geq \frac{1-\delta}{a+\epsilon}.$$

The desired bound now follows letting $\epsilon, \delta \downarrow 0$.

Order-matching upper bound: The proof of a tight upper bound is quite technical, and so we prove a weaker order-matching bound here. Additionally, we assume the increments X_i have finite second moment in order to keep the argument simple and informative.

The key idea of the proof is to use the union bound to upper bound the probability that S_{max} exceeds a threshold x by a sum of probabilities of the random walk exceeding x within suitably chosen finite intervals. We then invoke the maximal inequality of Theorem 7.5 to further upper bound the probability that the random walk exceeds x over finite intervals.

We start with the following union bound.

$$\begin{aligned}
\Pr\left(S_{max} > x\right) &\leq \sum_{k\geq 0} \Pr\left(\max_{2^k\leq n\leq 2^{k+1}} S_n > x\right) \\
&= \sum_{k\geq 0} \Pr\left(\max_{2^k\leq n\leq 2^{k+1}} (S_n + na) > x + na\right) \\
&\leq \sum_{k\geq 0} \Pr\left(\max_{2^k\leq n\leq 2^{k+1}} (S_n + na) > x + a2^k\right) \\
&\leq \sum_{k\geq 0} \Pr\left(\max_{1\leq n\leq 2^{k+1}} (S_n + na) > x + a2^k\right)
\end{aligned}$$

Next, we rewrite the above using a change of variables. Define $Y_i = X_i + a$, $W_n = \sum_{i=1}^{n} Y_i = S_n + na$. Thus,

$$\Pr\left(S_{max} > x\right) \leq \sum_{k\geq 0} \Pr\left(\max_{1\leq n\leq 2^{k+1}} W_n > x + a2^k\right)$$

Now, since Y_i are zero mean and i.i.d., it follows from the central limit theorem that

$$\lim_{n\to\infty} \Pr\left(\sum_{i=1}^{n} Y_i > -\sigma\sqrt{n}\right) \geq \frac{1}{2}.$$

It now follows that there exists large enough $D > 0$ such that

$$\Pr\left(\sum_{i=1}^{n} Y_i > -D\sqrt{n}\right) \geq \frac{1}{2} \quad \forall n.$$

Thus, we see that the conditions of Theorem 7.5 are satisfied, with $C = D\sqrt{m}$, and $q = 1/2$. It therefore follows that

$$\Pr\left(\max_{1\leq k\leq m} W_k > y\right) \leq 2\Pr\left(W_m > y - D\sqrt{m}\right).$$

Taking $m = 2^{k+1}$ and $y = x + a2^k$ we obtain

$$\Pr\left(\max_{1\leq n\leq 2^{k+1}} W_n > x + a2^k\right) \leq 2\Pr\left(W_{2^{k+1}} > x + a2^k - D2^{(k+1)/2}\right).$$

For x large enough, one can find $\epsilon \in (0,1)$ such that for all k,

$$\Pr\left(W_{2^{k+1}} > x + a2^k - D2^{(k+1)/2}\right) \leq \Pr\left(W_{2^{k+1}} > (x + a2^k)(1-\epsilon)\right).$$

Consequently,

$$\begin{aligned}\Pr\left(S_{max} > x\right) &\leq \sum_{k\geq 0} \Pr\left(\max_{1\leq n\leq 2^{k+1}} W_n > x + a2^k\right) \\ &\leq 2\sum_{k\geq 0} \Pr\left(W_{2^{k+1}} > (x+a2^k)(1-\epsilon)\right).\end{aligned}$$

We now invoke (3.24) to further bound the sum in the above inequality. Specifically, it follows from (3.24) that, given $\delta > 0$, there exists $n_0 \in \mathbb{N}$ such that for $n > n_0$,

$$\sup_{t:\ t > C\sqrt{n\log(n)}} \left|\frac{\Pr\left(W_n > t\right)}{n\Pr\left(Y_1 > t\right)} - 1\right| < \delta.$$

Since, for large enough k,

$$a2^k(1-\epsilon) > C\sqrt{2^{k+1}\log(2^{k+1})},$$

it follows that there exists $k_0 \in \mathbb{N}$ such that for $k > k_0$,

$$\Pr\left(W_{2^{k+1}} > (x+a2^k)(1-\epsilon)\right) \leq (1+\delta)2^{k+1}\Pr\left(Y_1 > (x+a2^k)(1-\epsilon)\right).$$

For $k \leq k_0$, it follows from Exercise 2.7 that for large enough x,

$$\Pr\left(W_{2^{k+1}} > (x+a2^k)(1-\epsilon)\right) \leq (1+\delta)2^{k+1}\Pr\left(Y_1 > (x+a2^k)(1-\epsilon)\right).$$

Finally, combining the bounds above, for large enough x, we have

$$\begin{aligned}\Pr\left(S_{max} > x\right) &\leq 2(1+\delta)\sum_{k\geq 0} 2^{k+1}\Pr\left(Y_1 > (x+a2^k)(1-\epsilon)\right). \\ &\leq \gamma\int_{x(1-\epsilon)}^{\infty} \Pr\left(Y_i > t\right) dt \\ &\overset{(a)}{\sim} \hat{\gamma}\, x\Pr\left(Y_1 > x\right) \\ &\overset{(b)}{\sim} \hat{\gamma}\, x\Pr\left(X_1 > x\right),\end{aligned}$$

for positive constants γ, $\hat{\gamma}$. Note that the step (a) above follows from Karamata's theorem (Theorem 2.4), while (b) is a consequence of the long-tail property. $\square$

Light-tailed step sizes

We now move from the case where X_i is heavy-tailed to the case where X_i is light-tailed. In this setting it is natural to expect the maximum position of the random walk to also be light-tailed, and the following result highlights that this is indeed the case. Further, the result provides a precise description of the logarithmic asymptotics of the tail of the maximum of the random walk.

Theorem 7.6. *Consider a one-dimensional random walk S_n with light-tailed i.i.d. step sizes X_i such that $\mathbb{E}[X_i] < 0$. Then, the all time maximum $S_{max} = \max_{n\geq 0}\{S_n\}$ satisfies*

$$\lim_{x\to\infty} -\frac{\log Pr\left(S_{max} > x\right)}{x} = s^*,$$

where $s^* = \sup\{s \geq 0 \mid \mathbb{E}\left[e^{sX_i}\right] \leq 1\}$.

The characterization of the tail of S_{max} in the above result is less precise than the characterization in Theorem 7.4. In this case, only the logarithmic asymptotics of the tail are characterized, i.e., the asymptotics of $\log \Pr\left(S_{max} > x\right)$. However, this is already enough to provide some interesting information. In particular, we see that the tail decays approximately exponentially. This is similar to what the extremal central limit theorem gives in the case of i.i.d. processes. In particular, the Gumbel distribution would be the emergent distribution in this setting if the steps were independent, and the right-tail of the Gumbel distribution decays exponentially. However, the specific decay rate of S_{max} is more mysterious, though s^* may appear familiar. This familiarity is because we already applied this result in Chapter 6 in order to characterize the behavior of multiplicative processes with a lower bound, and so the same s^* appeared in Theorem 6.2.

Proof. For simplicity, we prove the result under the assumption that there exists $\bar{s} > 0$ such that $\mathbb{E}\left[e^{\bar{s}X_i}\right] > 1$. (For a proof of the result without this assumption, we refer the reader to [149].) Under this assumption, $\mathbb{E}\left[e^{sX_i}\right] < \infty$ for all $s \in (0, \bar{s})$ and moreover, there exists a unique $s^* \in (0, \bar{s})$ such that $\mathbb{E}\left[e^{s^*X_i}\right] = 1$, $\mathbb{E}\left[e^{sX_i}\right] < 1$ for $s \in (0, s^*)$, and $\mathbb{E}\left[e^{sX_i}\right] > 1$ for $s \in (s^*, \bar{s})$.

The key idea behind the proof is that for any $m \in \mathbb{N}$,

$$\Pr\left(S_m > x\right) \leq \Pr\left(S_{max} > x\right) \leq \sum_{n\geq 0} \Pr\left(S_n > x\right) \tag{7.5}$$

While the second inequality in (7.5) yields the desired upper bound on the tail of $\log \Pr\left(S_{max} > x\right)$, the matching lower bound is obtained by utilizing the first inequality of (7.5) with a carefully chosen value of m. Indeed, our proof highlights that the all time maximum of the random walk is caused due to a 'conspiracy' between a large number of 'larger than usual' increments. Moreover, the proof reveals the most likely time when the maximum occurs.

We first utilize the second inequality in (7.5) to obtain the asymptotic upper bound on $\log \Pr\left(S_{max} > x\right)$.

Let $s^\delta = s^* - \delta$ and write

$$\begin{aligned}
\Pr\left(\max_{n\geq 0}\{S_n\} > x\right) &\leq \sum_{m\geq 0} \Pr\left(S_m > x\right) \\
&= \sum_{m\geq 0} \Pr\left(e^{s^\delta S_m} > e^{s^\delta x}\right) \\
&\leq e^{-s^\delta x} \sum_{m\geq 0} \mathbb{E}\left[e^{s^\delta X_1}\right]^m \\
&= e^{-s^\delta x} \frac{1}{1-\mathbb{E}\left[e^{s^\delta X_1}\right]}.
\end{aligned}$$

which is finite for any $\delta > 0$. Consequently,

$$\liminf_{x\to\infty} \frac{-\log \Pr\left(\max_{n\geq 0}\{S_n\} > x\right)}{x} \geq s_\delta. \tag{7.6}$$

Letting $\delta \downarrow 0$, we obtain

$$\liminf_{x\to\infty} \frac{-\log \Pr\left(\max_{n\geq 0}\{S_n\} > x\right)}{x} \geq s_0. \tag{7.7}$$

Next, we utilize the first inequality in (7.5) to obtain an asymptotically matching lower bound on $\log \Pr\left(S_{max} > x\right)$. Let $a = \mathbb{E}\left[X_1 e^{s^* X_1}\right]$ and set $m = x/a$. Invoking Cramér's theorem (see Theorem 3.4), we obtain

$$\begin{aligned}
\lim_{x\to\infty} \frac{-\log \Pr\left(S_{x/a} > x\right)}{x} &= \lim_{y\to\infty} \frac{-\log \Pr\left(S_y > ya\right)}{ya} \\
&= \frac{1}{a} \sup_{s>0}[as - \log \mathbb{E}\left[e^{sX_1}\right]] \\
&= s^*,
\end{aligned}$$

since s^* solves this optimization problem. Consequently,

$$\limsup_{x\to\infty} \frac{-\log \Pr\left(\max_{n\geq 0}\{S_n\} > x\right)}{x} \leq \lim_{x\to\infty} \frac{-\log \Pr\left(S_{x/a} > x\right)}{x} = s^*.$$

□

7.5 A variation: The time between record breaking events

We started this chapter with a discussion of the progression of world records in the half marathon and the 100m dash (Figure 7.1). Given that these are such canonical examples of extremal processes, it is natural to end by coming full circle back to these examples. As we already highlighted, at first glance it these world record progressions shows a leveling off, and thus suggests a light-tailed limiting distribution. But, on a second look, you may start to see that the time between improvements looks like it could be heavy-tailed. Basically, there are lots of periods where the record changes frequently, but there are also others where the records go unchanged for long periods. In fact, this has been observed and tested empirically in a variety of contexts, e.g., records for rainfall, earthquakes, and other extreme events (see, for example, [136, 140]). Very commonly, the times between "records" seem to exhibit heavy-tailed behavior.

With this observation in mind, it is natural to ask "Why?" and, unfortunately, the results we have discussed so far in the chapter do not provide an explanation. So, providing an explanation the goal of this section.

To formalize the setting, let us consider the following. Suppose we observe a sequence $\{X_i\}_{i\geq 1}$ of i.i.d. random variables, with distribution F. As before, define $M_n := \max(X_1, X_2, \cdots, X_n)$. Let L_k denote the instant of the kth record, i.e., $L_1 = 1$ and $L_{k+1} = \min\{i > L_k \mid X_i > M_{i-1}\}$ for $k \geq 1$. For $k \geq 1$, let $T_k := L_{k+1} - L_k$ denote the time between the $k+1$st record and the kth record.

Then, we can prove the following theorem, which shows that the time between records is indeed heavy-tailed, specifically regularly varying.

Theorem 7.7. *Suppose that F is continuous. Then for any $k \geq 1$, T_k is heavy-tailed, with*

$$Pr\left(T_k > n\right) \sim \frac{2^{k-1}}{n}.$$

Note that this is a somewhat delicate situation that we are studying since T_k is not stationary with respect to k. Indeed, you expect that, as the record gets bigger, the time it takes to break it gets larger. Also, a interesting observation about the theorem is that it is not required that F have infinite support for T_k to be heavy-tailed. So, looking at records can create heavy tails from things that are extremely light-tailed.

Proof. To begin the proof, we start by showing that we may assume that F is exponentially distributed with no loss of generality, since the distribution of T_k does not depend on the distribution F. To do this, we focus on the function $Q(x) = -\log \bar{F}(x)$, which you should recall is the cumulative hazard function corresponding to the distribution F (see Chapter 4). The key step in this first part of the proof is to show that the random variable $Q(X_1)$ is exponentially distributed with mean 1. This follows easily from the fact that $U := \bar{F}(X_1)$ is a uniform random variable over $[0, 1]$:

$$\Pr\left(Q(X_1) > x\right) = \Pr\left(-\log(U) > x\right) = \Pr\left(U < e^{-x}\right) = e^{-x}.$$

Let $E_i = Q(X_i)$. Clearly, $\{E_i\}$ is an i.i.d. sequence of exponential random variables. Moreover, since Q is non-decreasing, records of the sequence $\{X_i\}$ coincide with those of the sequence $\{E_i\}$. Thus, for the purpose of studying the time between records, we may assume without loss of generality that F is an exponential distribution with mean 1. We make this assumption in the remainder of this proof.

Next, we study the distribution of the records. Specifically, we show that the kth record $M_{(k)} := X_{L_k}$ has an Erlang distribution with shape parameter k and rate parameter 1 (i.e., $M_{(k)}$ is distributed as the sum of k i.i.d. exponential random variables, each having mean 1). We proceed inductively as follows. Clearly, the above claim is true for $k = 1$, since $M_{(1)} = X_1$. Assume that the claim is true for some $k \in \mathbb{N}$. Note that for $x > 0$,

$$\Pr\left(M_{(k+1)} > M_{(k)} + x\right) = \Pr\left(E > M_{(k)} + x \mid E > M_{(k)}\right),$$

where E is exponentially distributed with mean 1, and independent of $M_{(k)}$. From the memoryless property of the exponential distribution, it now follows that $\Pr\left(M_{(k+1)} > M_{(k)} + x\right) = e^{-x}$, which implies that $M_{(k+1)} \stackrel{d}{=} M_{(k)} + E$. This proves our claim that $M_{(k+1)}$ has an Erlang distribution.

We are now ready to analyze the tail of T_k. Once again, we proceed inductively, and first consider the case $k = 1$. Note that conditioned on the value of X_1, T_1 is a geometrically distributed with

$$\Pr(T_1 > n \mid X_1 = x) = (1 - e^{-x})^n.$$

Therefore, unconditioning with respect to X_1,

$$\Pr(T_1 > n) = \int_0^\infty (1 - e^{-x})^n e^{-x} dx.$$

Making the substitution $y = e^{-x}$, we get

$$\Pr(T_1 > n) = \int_0^1 (1 - y)^n dy = \frac{1}{n+1}.$$

It follows that $\Pr(T_1 > n) \sim \frac{1}{n}$.

Next, we assume that, for some $k \in \mathbb{N}$, $\Pr(T_k > n) \sim \frac{2^{k-1}}{n}$ and analyze the tail of T_{k+1}. Recall that $M_{(k+1)} \overset{d}{=} M_{(k)} + E$, where E is exponentially distributed with mean 1, and independent of $M_{(k)}$. Therefore, we can think of T_{k+1} as the time until a new sample exceeds $M_{(k)} + E$. Note that the time until a new sample exceeds $M_{(k)}$ is distributed as T_k. Moreover, conditioned on a new sample X_i exceeding $M_{(k)}$, the probability that it exceeds $M_{(k)} + E$ equals

$$\Pr\left(X_i > M_{(k)} + E \mid X_i > M_{(k)}\right) = \Pr(X_i > E) = 1/2.$$

The above calculation exploits the memoryless property of the exponential distribution, and the fact that X_i and E are i.i.d. Thus, when a new sample exceeds $M_{(k)}$, it also exceeds $M_{(k)} + E$ (and thus sets a new record) with probability 1/2. Therefore, T_{k+1} is simply distributed as a geometric random sum of i.i.d. random variables, each distributed as T_k, i.e.,

$$T_{k+1} \overset{d}{=} \sum_{i=1}^{N} Y_k(i),$$

where $\{Y_k(i)\}$ is an i.i.d. sequence of random variables with the same distribution as T_k, and N is a geometric random variable independent of $\{Y_k(i)\}$ with success probability 1/2.

Finally, since T_k is assumed to be regularly varying (and therefore subexponential), we may invoke Theorem 3.1 to obtain the tail behavior of T_{k+1}. We therefore have

$$\Pr(T_{k+1} > n) \sim \mathbb{E}[N] \Pr(T_k > n) = 2 \Pr(T_k > n),$$

which proves our desired induction step. □

7.6 Additional notes

The focus in this chapter has been on a few examples of extremal processes and there is much more material that the interested reader can consult in continuing their study of this important area. In particular, our discussion of max-stable distributions and the extremal central limit theorem are only a brief introduction into the area of extreme value theory, which is one of the main subdisciplines within probability theory and statistics, and is still a very active area of research. There exist many papers and books with which the interested reader can continue their study. Fisher and Tippett [73] is a classic in the field, as are the works of Gnedenko [86], Gumbel [93] and the PhD thesis of De Haan [51]. Excellent textbooks include [8], focusing on the phenomenology of large exceedances of dependent variables appearing in clumps, [117], covering maxima of dependent Gaussian sequences, and books by Resnick [164, 165] exhibiting the deep connections with regular variation and point processes. Embrechts covers many links with financial and insurance models, and also the theory of order statistics in [64] .

An important note about the presentation here is that all of the processes we have considered have a discrete time parameter. In the mathematical literature, it is common for a continuous-time process $M(t), t \geq 0$, to be called an extremal process, if, for any non-decreasing sequence $t_1, ..., t_n$, the following holds: there exist independent random variables $U_1, ..., U_n$ such that $(M(t_1), ..., M(t_n))$ has the same joint distribution as $\{U_1, U_1 + U_2, \cdots, U_1 + \cdots + U_n\}$ and

$$\Pr(U_i \leq u) = \Pr(M(t_i - t_{i-1}) \leq u).$$

Beyond the simple discrete extremal process that we consider in the chapter, we also focuses on examples related to all-time maxima random walks. This material is classical and can be found in many textbooks, such as [14] and [15]. An important distinctive feature of the heavy-tailed and light-tailed examples this chapter is perhaps not so much where the ruin probability has an exponential tail, or a power tail, but the behavior of *the time until ruin, given ruin occurs*. In particular, set $\tau(x) = \inf\{n : S_n > x\}$. We see that $\sup_{n\geq 0} S_n > x$ if and only if $\tau(x) < \infty$. We are interested in the time until ruin, given that ruin occurs, i.e. in the behavior of $\tau(x)$, given $\tau(x) < \infty$. It now turns out that, if the claim sizes are light-tailed, $\tau(x)/x$, conditional on $\tau(x) < \infty$ converges to a deterministic constant. This is in sharp contrast with the case of power-law claim sizes, where $\tau(x)/x$, conditional on $\tau(x) < \infty$, converges to a random variable which is heavy-tailed in itself. This provides an interesting explanation of the unpredictability of so-called "black swans" [181] and a rigorous proof can be found in [16].

A final note about this chapter is that we have not touched upon connections with statistical applications at all. This will be the subject of the Part III of this book, where several of the results appearing in this chapter will find application.

7.7 Exercises

For all the exercises in this chapter we will use the following notation: $\{X_n\}_{n\geq 1}$ is an i.i.d. sequence with distribution F, and $M_n = \max(X_1, X_2, \cdots, X_n)$ for $n \geq 1$. Also, we use x_F to denote the right endpoint of the distribution F, i.e., $x_F = \sup\{x : \bar{F}(x) > 0\}$.

1. Prove that for a real sequence $\{x_n\}_{n\geq 1}$ and $\tau \in [0, \infty]$,

$$\lim_{n\to\infty} n\bar{F}(x_n) = \tau \quad \iff \quad \lim_{n\to\infty} \Pr\left(M_n \leq x_n\right) = e^{-\tau}.$$

2. For deterministic sequences $\{a_n\}$ and $\{b_n\}$, where $a_n > 0$, and a max-stable distribution G, prove that

$$\frac{M_n - b_n}{a_n} \xrightarrow{d} G \iff \lim_{n\to\infty} n\bar{F}(a_n x + b_n) = -\log G(x) \,\forall\, x \in \mathbb{R}.$$

Here $-\log G(x)$ should be interpreted as ∞ when $G(x) = 0$.

Hint: Use the result of Exercise 1.

3. The goal of this exercise is to show that if F has a jump at its right endpoint, then the corresponding extremal process cannot have a non-degenerate limit under for any choice of translation and scaling parameters.

Specifically, show that if $x_F < \infty$ and $\Pr\left(X_1 = x_F\right) > 0$ (or equivalently, F has a jump discontinuity at x_F), then $\lim_{n\to\infty} \Pr\left(M_n \leq x_n\right) = \theta$ implies that θ is either 0 or 1.

Hint: Use the result of Exercise 1.

4. In this exercise you will show that if F is geometrically distributed, then the corresponding extremal process cannot have a non-degenerate limit under for any choice of translation and scaling parameters.

Specifically, if F is geometric with parameter $p \in (0, 1)$, show that $\lim_{n\to\infty} \Pr\left(M_n \leq x_n\right) = \theta$ implies that θ is either 0 or 1.

Hint: You have to use the result of Exercise 1. Prove that if $n\bar{F}(x_n) \to \tau \in (0, \infty)$, then that would imply that $x_n \to \infty$ and ${\bar{F}(x_n)}/{\bar{F}(x_{n-1})} \to 1$, which is not possible given that F is geometric.

5. Like the previous exercises, in this problem you will show that if F is Poisson, then the corresponding extremal process cannot have a non-degenerate limit under for any choice of translation and scaling parameters. Specifically, if F is Poisson with parameter $\lambda > 0$, show that $\lim_{n\to\infty} \Pr\left(M_n \leq x_n\right) = \theta$ implies that θ is either 0 or 1.

6. Prove Lemma 7.1.

7. The goal of this exercise is to show that distributions in the maximum domain of attraction of the Gumbel distribution have lighter tails than regularly varying distributions.

Consider a distribution F with $x_F = \infty$ in the maximum domain of attraction of the Gumbel distribution. Prove that if the distribution G is regularly varying with index $-\rho$, where $\rho > 0$, then

$$\lim_{x\to\infty} \frac{\bar{F}(x)}{\bar{G}(x)} = 0.$$

Hint: Compare the representation of F given by Theorem 7.3 with that for G by the Karamata representation theorem (Theorem 2.6. Exploit the fact that the function g in the former representation satisfies $\frac{g(t)}{t} \to 0$ as $t \to \infty$.

8. Suppose that the distribution F satisfies the conditions of the case (ii) of Theorem 7.3. Prove that

$$\frac{M_n - x_F}{x_F - F^{\leftarrow}(1-1/n)} \xrightarrow{d} \Psi_\alpha.$$

Hint: From the result of Exercise 2, it suffices to show that $n\bar{F}(a_n x + b_n) \to -\log \Psi_\alpha(x)$ for $x < 0$, where a_n and b_n are the given normalization constants. Define G such that $\bar{G}(y) = \bar{F}(x_F - 1/y)$ and $\tilde{a}_n := G^{\leftarrow}(1-1/n)$. Prove that $\tilde{a}_n = 1/a_n$. Finally, show that for $x < 0$,

$$n\bar{F}(a_n x + b_n) = n\bar{G}\left(a_n/x\right) \to \Phi_\alpha\left(1/x\right) = \Psi_\alpha(x).$$

9. The beta distribution has support $[0,1]$, and is characterized by the density function

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \quad (x \in [0,1]),$$

where $a, b > 0$ and $\Gamma(\cdot)$ denotes the gamma function. Prove that the beta distribution lies in the MDA of Ψ_b.

Hint: Use Karamata's theorem (Theorem 2.4) to show that $\bar{F}(b - 1/x)$ is regularly varying with index $-b$.

10. Recall that the Erlang distribution with parameters (k, μ), where $k \in \mathbb{N}$ and $\mu > 0$, is associated with the c.c.d.f.

$$\bar{F}(x) = \begin{cases} e^{-\mu x} \displaystyle\sum_{i=0}^{k-1} \frac{(\mu x)^i}{i!} & \text{for } x \geq 0 \\ 1 & \text{for } x < 0 \end{cases}.$$

Prove that the Erlang distribution belongs to the MDA of the Gumbel distribution.

11. Recall that the hyperexponential distribution is the mixture of independent exponentials. Specifically, consider the hyperexponential distribution F defined by the c.c.d.f.

$$\bar{F}(x) = \sum_{i=1}^{n} p_i e^{-\mu_i x} \qquad (x \geq 0).$$

Here, $\mu_i > 0$ and $p_i > 0$ for $1 \leq i \leq n$, with $\sum_{i=1}^n p_i = 1$. Prove that the hyperexponential distribution is belongs to the MDA of the Gumbel distribution.

12. Prove that the standard Gaussian belongs to the MDA of the Gumbel distribution.

You may want use the fact that $q_N(x) \sim x$ as $x \to \infty$, where q_N denotes the hazard rate of the standard Gaussian. This can be proved easily using L'Hospital's rule.

13. Prove that the LogNormal distribution belongs to the MDA of the Gumbel distribution.

Hint: Use the result of Exercise 12.

14. Suppose that the distribution F is regularly varying. Define

$$a_n := \inf\{x \in \mathbb{R} \mid \bar{F}(x) \leq n^{-1}\}.$$

Prove that $\bar{F}(a_n) \sim n^{-1}$.

Hint: Use Karamata's representation theorem (Theorem 2.6) to show that there exists a continuous distribution G such that $\bar{F}(x) \sim \bar{G}(x)$ as $x \to \infty$.

15. Suppose that $\mathbb{E}[X_i] = \mu$. Let $S_0 = 0$, and $S_n = \sum_{i=1}^{n} X_i$ for $n \geq 1$. Prove that given $\epsilon, \delta > 0$, there exists $L > 0$ such that

$$\Pr(S_n > n(\mu - \epsilon) - L) \geq 1 - \delta \quad \forall\, n \geq 0.$$

Part III

Estimation

Given the mystique and excitement that surrounds the discovery of heavy-tailed phenomena, the detection and estimation of heavy-tails in data is a task that is often (over)zealously pursued. Discovery of heavy-tailed phenomena is often viewed with surprise and curiosity; however, one must be careful in such pursuits since the estimation of heavy-tailed distributions is fraught with pitfalls and booby traps. Many simple and intuitive approaches that are commonly used for detection and estimation of heavy-tailed distributions can be misleading. In fact, it is not uncommon for the misuse of statistical tools to result in mistaken detections of heavy-tailed phenomena, which in turn leads to the controversy surrounding heavy-tails that still exists today.

In Part III of this book we focus on providing an introduction to the statistical tools used for the estimation of heavy-tailed phenomena. Unfortunately, there is no perfect recipe for how to "properly" detect and estimate heavy-tailed distributions in data. Thus, our treatment seeks to highlight a handful of important approaches, and to provide insight into when each approach is appropriate and when each may be misleading. In particular, in Chapter 8 we focus on classical parametric approaches for estimating power-law distributions, such as linear regression and maximum likelihood estimation. These approaches assume the data come from a precise power-law distribution and can thus use data from the body of the distribution to estimate the tail. However, often it is not the case that exact power-law *distributions* are present in data; instead only power-law *tails* are present. When only a power-law tail is present, classical parametric approaches can be misleading. Thus, Chapter 9 focuses on semi-parametric statistical tools that are appropriate for detecting power-law tails. These techniques provide a dramatic contrast to the classical approaches; they tend to throw away large amounts of data about the body of the distribution, keeping only data about the tail, i.e., the outliers. Finally, Chapter 10 focuses on the additional challenges presented by estimating heavy-tailed phenomena in multivariate settings. Again, in multivariate settings it is unusual to have precise heavy-tailed *distributions* present in data; it is typically only the *tails* that are power-law. However, in the multivariate world, the body is even less informative about the tail due to the fact that correlations between variables are not parametric. This means that classical approaches can be even more problematic than in the univariate case. However, by reparameterizing the data using polar coordinates it is possible to estimate the tail directly using techniques that parallel what we present in Chapter 9; thus obtaining reliable estimation tools.

Combined, these three chapters highlight a crucial point: *one must proceed carefully when seeking to estimate heavy-tailed phenomena in real-world data*. In particular, it is typically naive to seek to estimate exact heavy-tailed distributions in data. Instead, the focus should be on estimating the tail of heavy-tailed phenomena. However, even in doing this, one should not rely on a single method for estimation. Instead, it is a necessity to build confidence through the use of multiple, complementary estimation approaches. These are important lessons to take to heart since there are many examples of top-notch researchers enthusiastically reaching flawed conclusions about heavy-tailed phenomena as a result of methodological mistakes.

Chapter 8

Estimating power-law distributions: Listen to the body

Chapter 9

Estimating power-law tails: Let the tail do the talking

Chapter 10

Estimating multivariate power-law tails: Cautionary tales of tails

Bibliography

[1] National Centers for Environment Information. https://www.ngdc.noaa.gov/. Accessed: 2019-08-06.

[2] United States Census Bureau. https://www.census.gov/data/tables/time-series/demo/popest/2010s-total-cities-and-towns.html. Accessed: 2019-08-06.

[3] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. *Journal of the ACM (JACM)*, 56(4):21, 2009.

[4] J. Aczél. *On applications and theory of functional equations*. Academic Press, 2014.

[5] L. A. Adamic and B. A. Huberman. The nature of markets in the world wide web. *Quarterly Journal of Electronic Commerce*, 2000.

[6] J. Aitchison and J. Brown. The lognormal distribution, 1957.

[7] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[8] D. Aldous. *Probability approximations via the Poisson clumping heuristic*. 1989), publisher=Springer.

[9] A. P. Allen, B.-L. Li, and E. L. Charnov. Population fluctuations, power laws and mixtures of lognormal distributions. *Ecology Letters*, 4(1):1–3, 2001.

[10] L. J. Allen. *An introduction to stochastic processes with applications to biology*. Chapman and Hall/CRC, 2010.

[11] S. Amari, R. S. Lewis, and E. Anders. Interstellar grains in meteorites: I. isolation of sic, graphite and diamond; size distributions of sic and graphite. *Geochimica et Cosmochimica Acta*, 58(1):459–470, 1994.

[12] C. Anderson. *The long tail: Why the future of business is selling less of more*. Hachette Books, 2006.

[13] A. Araujo and E. Giné. *The central limit theorem for real and Banach valued random variables*, volume 431. Wiley New York, 1980.

[14] S. Asmussen. *Applied probability and queues*, volume 51. Springer Science & Business Media, 2008.

[15] S. Asmussen and H. Albrecher. *Ruin probabilities*. World Scientific Publishing, 2010.

[16] S. Asmussen and C. Klüppelberg. Large deviations results for subexponential tails, with applications to insurance risk. *Stochastic processes and their applications*, 64(1):103–125, 1996.

[17] S. Asmussen, C. Klüppelberg, and K. Sigman. Sampling at subexponential times, with queueing applications. *Stochastic processes and their applications*, 79(2):265–286, 1999.

[18] K. B. Athreya and P. Jagers. *Classical and modern branching processes*, volume 84. Springer Science & Business Media, 2012.

[19] K. B. Athreya and P. E. Ney. *Branching processes*. Courier Corporation, 2004.

[20] F. Baccelli and S. Foss. Moments and tails in monotone-separable stochastic networks. *The Annals of Applied Probability*, 14(2):612–650, 2004.

[21] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[22] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[23] A.-L. Barabási and E. Bonabeau. Scale-free networks. *Scientific American*, 288(5):60–69, 2003.

[24] A.-L. Barabási et al. *Network science*. Cambridge university press, 2016.

[25] A.-L. Barabási et al. *Network science*. Cambridge university press, 2016.

[26] R. N. Bhattacharya and R. R. Rao. *Normal approximation and asymptotic expansions*. SIAM, 2010.

[27] P. Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[28] N. Bingham, C. Goldie, and J. Teugels. *Regular variation*. Cambridge University Press, 1989.

[29] O. Boxma and B. Zwart. Tails in scheduling. *Performance Evaluation Review*, 34(4):13–20, 2007.

[30] O. J. Boxma and V. Dumas. Fluid queues with long-tailed activity period distributions. *Computer Communications*, 21(17):1509–1529, 1998.

[31] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[32] T. Britton. Stochastic epidemic models: a survey. *Mathematical biosciences*, 225(1):24–35, 2010.

[33] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer networks*, 33(1-6):309–320, 2000.

[34] A. D. Broido and A. Clauset. Scale-free networks are rare. *Nature communications*, 10(1):1017, 2019.

[35] J. A. Bucklew. *Large deviation techniques in decision, simulation, and estimation*. Wiley New York, 1990.

[36] D. Buraczewski, J. F. Collamore, E. Damek, and J. Zienkiewicz. Large deviation estimates for exceedance times of perpetuity sequences and their dual processes. *The Annals of Probability*, 44(6):3688–3739, 2016.

[37] D. Buraczewski, E. Damek, and T. Mikosch. *Stochastic Models with Power-Law Tails*. Springer, 2016.

[38] G. Caldarelli. *Scale-free networks: complex webs in nature and technology*. Oxford University Press, 2007.

[39] J. M. Carlson and J. Doyle. Highly optimized tolerance: A mechanism for power laws in designed systems. *Physical Review E*, 60(2):1412, 1999.

[40] D. G. Champernowne. A model of income distribution. *The Economic Journal*, 63(250):318–351, 1953.

[41] H. Chen and D. D. Yao. *Fundamentals of queueing networks: Performance, asymptotics, and optimization*, volume 46. Springer Science & Business Media, 2013.

[42] V. Chistyakov. A theorem on sums of independent positive random variables and its applications to branching random processes. *Theory of Probability and its Applications*, 9:640–648, 1964.

[43] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[44] R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. 2001.

[45] J. C. Cox, S. A. Ross, and M. Rubinstein. Option pricing: A simplified approach. *Journal of Financial Economics*, 7(3):229–263, 1979.

[46] H. Cramér. Sur un nouveau théoreme-limite de la théorie des probabilités. *Actualités scientifiques et industrielles*, 736(5-23):115, 1938.

[47] M. E. Crovella, M. S. Taqqu, and A. Bestavros. Heavy-tailed probability distributions in the world wide web. *A practical guide to heavy tails*, 1:3–26, 1998.

[48] E. L. Crow and K. Shimizu. *Lognormal distributions*. Marcel Dekker, 1987.

[49] L. F. M. deHaan. *On regular variation and its application to the weak convergence of sample extremes*. PhD thesis, University of Amsterdam, 1970.

[50] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*. Springer-Verlag, Berlin, 2010.

[51] F. Den Hollander. *Large deviations*. American Mathematical Society, 2008.

[52] D. Denisov, A. B. Dieker, and V. Shneer. Large deviations for random walks under subexponentiality: The big-jump domain. *The Annals of Probability*, pages 1946–1991, 2008.

[53] D. Denisov and B. Zwart. On a theorem of breiman and a class of random difference equations. *Journal of Applied Probability*, 44(4):1031–1046, 2007.

[54] J.-D. Deuschel and D. W. Stroock. *Large deviations*, volume 137. Academic press Boston, 1989.

[55] L. Devroye. Branching processes and their applications in the analysis of tree structures and tree algorithms. In *Probabilistic methods for algorithmic discrete mathematics*, pages 249–314. 1998.

[56] R. Doney. Local behaviour of first passage probabilities. *Probability Theory and Related Fields*, 152(3):559–588, 2012.

[57] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking. *Physical review letters*, 85(21):4633, 2000.

[58] A. B. Downey. The structural cause of file size distributions. In *MASCOTS 2001, Proceedings Ninth International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, pages 361–370. IEEE, 2001.

[59] P. Dupuis and R. S. Ellis. *A weak convergence approach to the theory of large deviations*, volume 902. John Wiley & Sons, 2011.

[60] R. Durrett. *Probability: Theory and examples*. Cambridge university press, 2010.

[61] F. Eggenberger and G. Pólya. Über die statistik verketteter vorgänge. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 3(4):279–289, 1923.

[62] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events: for insurance and finance*, volume 33. Springer Verlag, 1997.

[63] C. Estan and G. Varghese. New directions in traffic measurement and accounting: Focusing on the elephants, ignoring the mice. *ACM Transactions on Computer Systems (TOCS)*, 21(3):270–313, 2003.

[64] J.-B. Estoup. *Gammes sténographiques: méthode et exercices pour l'acquisition de la vitesse*. Institut sténographique, 1916.

[65] A. Fabrikant, E. Koutsoupias, and C. Papadimitriou. Heuristically optimized trade-offs: A new paradigm for power laws in the internet. *Automata, languages and programming*, pages 781–781, 2002.

[66] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, volume 29, pages 251–262. ACM, 1999.

[67] W. Feller. *An introduction to probability theory and its applications: Volume I*. John Wiley & Sons, 1968.

[68] W. Feller. *An introduction to probability theory and its applications: Volume II*. John Wiley & Sons, 1971.

[69] J. Feng and T. G. Kurtz. *Large deviations for stochastic processes*. Number 131. American Mathematical Society, 2006.

[70] H. Fischer. *A history of the central limit theorem: From classical to modern probability theory*. Springer Science & Business Media, 2010.

[71] R. A. Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190, 1928.

[72] S. Foss, T. Konstantopoulos, and S. Zachary. Discrete and continuous time modulated random walks with heavy-tailed increments. *Journal of Theoretical Probability*, 20(3):581–612, 2007.

[73] S. Foss and D. Korshunov. Heavy tails in multi-server queue. *Queueing Systems*, 52(1):31–48, 2006.

[74] S. Foss, D. Korshunov, and S. Zachary. *An introduction to heavy-tailed and subexponential distributions*. Springer, 2011.

[75] S. Foss, Z. Palmowski, S. Zachary, et al. The probability of exceeding a high boundary on a random time interval for a heavy-tailed random walk. *The Annals of Applied Probability*, 15(3):1936–1957, 2005.

[76] M. Fréchet. Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Math. Polon.*, 6:93–116, 1927.

[77] X. Gabaix. Zipf's law and the growth of cities. *American Economic Review*, 89(2):129–132, 1999.

[78] X. Gabaix. Zipf's law for cities: an explanation. *The Quarterly journal of economics*, 114(3):739–767, 1999.

[79] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. E. Stanley. A theory of power-law distributions in financial market fluctuations. *Nature*, 423(6937):267, 2003.

[80] F. Galton. Problem 4001. *Educational Times*, 1, 1873.

[81] A. Ganesh, N. O'Connell, and D. Wischik. *Big Queues*. Springer, 2004.

[82] R. Gibrat. Les inégalités economiques (sirey, paris). *A. MAJUMDER AND SR CHAKRAVARTY*, 1931.

[83] B. Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, pages 423–453, 1943.

[84] B. Gnedenko and A. Kolmogorov. *Limit distributions for sums of independent random variables*. Addison-Wesley, 1954.

[85] C. Goldie and C. Klüppelberg. Subexponential distributions. In *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*. Birkhäuser, Boston, 1998.

[86] C. M. Goldie. Implicit renewal theory and tails of solutions of random equations. *The Annals of Applied Probability*, pages 126–166, 1991.

[87] A. Goswami and B. V. Rao. *A course in applied stochastic processes*, volume 40. Springer, 2006.

[88] G. Grimmett, G. R. Grimmett, and D. Stirzaker. *Probability and random processes*. Oxford university press, 2001.

[89] F. Guess and F. Proschan. 12 mean residual life: Theory and applications. *Handbook of statistics*, 7:215–224, 1988.

[90] E. J. Gumbel. *Statistics of extremes*. Columbia Univiversity Press, 1958.

[91] L. Guo and I. Matta. The war between mice and elephants. In *Proceedings of the IEEE Conference on Network Protocols*, pages 180–188, 2001.

[92] B. Gutenberg and C. F. Richter. Frequency of earthquakes in california. *Bulletin of the Seismological Society of America*, 34(4):185–188, 1944.

[93] A. P. Hackett. *70 Years of Best Sellers: 1895-1965*. New York: RR Bowker Company, 1967.

[94] B. Hajek. *Random processes for engineers*. Cambridge University Press, 2015.

[95] M. Harchol-Balter and A. B. Downey. Exploiting process lifetime distributions for dynamic load balancing. *ACM Transactions on Computer Systems*, 15(3):253–285, 1997.

[96] D. Heath, S. Resnick, and G. Samorodnitsky. Patterns of buffer overflow in a class of queues with long memory in the input stream. *The Annals of Applied Probability*, pages 1021–1057, 1997.

[97] P. Hines, K. Balasubramaniam, and E. C. Sanchez. Cascading failures in power grids. *IEEE Potentials*, 28(5):24–30, Sep. 2009.

[98] B. A. Huberman and L. A. Adamic. Internet: growth dynamics of the world-wide web. *Nature*, 401(6749):131, 1999.

[99] B. A. Huberman and L. A. Adamic. Internet: growth dynamics of the world-wide web. *Nature*, 401(6749):131, 1999.

[100] J. Hull. *Introduction to futures and options markets*. Prentice Hall, 1991.

[101] I. Ibragimov and Y. Linnik. *Independent and stationary sequences of random variables*. Walters—Noordoff, 1971.

[102] R. Jain and S. Ramakumar. Stochastic dynamics modeling of the protein sequence length distribution in genomes: implications for microbial evolution. *Physica A: Statistical Mechanics and its Applications*, 273(3-4):476–485, 1999.

[103] S. Janson. Stable distributions. *arXiv preprint arXiv:1112.0220*, 2011.

[104] P. Jelenković and P. Momčilović. Large deviation analysis of subexponential waiting times in a processor-sharing queue. *Mathematics of Operations Research*, 28(3):587–608, 2003.

[105] P. R. Jelenković and M. Olvera-Cravioto. Implicit renewal theorem for trees with general weights. *Stochastic Processes and their Applications*, 122(9):3209–3238, 2012.

[106] H. Kesten. Random difference equations and renewal theory for products of random matrices. *Acta Mathematica*, 131(1):207–248, 1973.

[107] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170, 2000.

[108] C. Klüppelberg. Subexponential distributions and integrated tails. *Journal of Applied Probability*, 25(1):132–141, 1988.

[109] C. Klüppelberg, S. Pergamenchtchikov, et al. The tail of the stationary distribution of a random coefficient ar (q) model. *The Annals of Applied Probability*, 14(2):971–1005, 2004.

[110] L. Knopoff and D. Sornette. Eartkquake death tolls. *Journal de Physique I*, 5(12):1681–1668, 1995.

[111] E. V. Koonin, Y. I. Wolf, and G. P. Karev. *Power laws, scale-free networks and genome biology*. Springer, 2006.

[112] P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Physical review letters*, 85(21), 2000.

[113] M. R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and related properties of random sequences and processes*. Springer Science & Business Media, 2012.

[114] J. Leskovec, D. Chakrabarti, J. Kleinberg, and C. Faloutsos. Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In *European conference on principles of data mining and knowledge discovery*, pages 133–145. Springer, 2005.

[115] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *Journal of Machine Learning Research*, 11(Feb):985–1042, 2010.

[116] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM, 2005.

[117] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1):2, 2007.

[118] J. Leslie. On the non-closure under convolution of the subexponential family. *Journal of Applied Probability*, 26(1):58–66, 1989.

[119] M. Levy, S. Solomon, and G. Ram. Dynamical explanation for the emergence of power law in a stock market model. *International Journal of Modern Physics C*, 7(01):65–72, 1996.

[120] P. Lévy. Théorie des erreurs: La Loi de Gauss et Les Lois Exceptionelles. *Bull. Soc. Math. France*, 52:49–85, 1924.

[121] L. Li, D. Alderson, J. C. Doyle, and W. Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, 2(4):431–523, 2005.

[122] F. Lindskog, S. I. Resnick, J. Roy, et al. Regularly varying measures on metric spaces: Hidden regular variation and hidden jumps. *Probability Surveys*, 11:270–314, 2014.

[123] M. Loeve. *Probability Theory II*. Springer, 1978.

[124] E. T. Lu and R. J. Hamilton. Avalanches and the distribution of solar flares. *The astrophysical journal*, 380:L89–L92, 1991.

[125] N. M. Luscombe, J. Qian, Z. Zhang, T. Johnson, and M. Gerstein. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome biology*, 3(8):research0040–1, 2002.

[126] A. Mahanti, N. Carlsson, A. Mahanti, M. Arlitt, and C. Williamson. A tale of the tails: Power-laws in internet measurements. *IEEE Network*, 27(1):59–64, 2013.

[127] B. Mandelbrot. An informational theory of the statistical structure of language. *Communication theory*, 84:486–502, 1953.

[128] B. Mandelbrot. A note on a class of skew distribution functions: Analysis and critique of a paper by ha simon. *Information and Control*, 2(1):90–99, 1959.

[129] B. Mandelbrot. Final note on a class of skew distribution functions: analysis and critique of a model due to ha simon. *Information and Control*, 4(2-3):198–216, 1961.

[130] B. Mandelbrot. Post scriptum to "final note". *Information and Control*, 4(2-3):300–304, 1961.

[131] B. B. Mandelbrot. *The fractal geometry of nature*. WH Freeman New York, 1982.

[132] A. Mazzarella and N. Diodato. The alluvial events in the last two centuries at sarno, southern Italy: Their classification and power-law time-occurrence. *Theoretical and applied climatology*, 72(1-2):75–84, 2002.

[133] T. Mikosch. *Extreme Values in Finance, Telecommunications, and the Environment*, chapter Modeling dependence and tails of financial time series. Chapman and Hall, 2003.

[134] T. Mikosch and A. Nagaev. Large deviations of heavy-tailed sums with applications in insurance. *Extremes*, 1(1):81–110, 1998.

[135] A. Müller and D. Stoyan. Comparison methods for stochastic models and risks. *John Wiley & Sons*, 2002.

[136] R. Musson, T. Tsapanos, and C. Nakas. A power-law function for earthquake interarrival time and magnitude. *Bulletin of the Seismological Society of America*, 92(5):1783–1794, 2002.

[137] A. V. Nagaev. Integral limit theorems taking large deviations into account when cramér's condition does not hold. i. *Theory of Probability & Its Applications*, 14(1):51–64, 1969.

[138] S. V. Nagaev. Large deviations of sums of independent random variables. *The Annals of Probability*, pages 745–789, 1979.

[139] T. Nakajima and A. Higurashi. A use of two-channel radiances for an aerosol characterization from space. *Geophysical Research Letters*, 25(20):3815–3818, 1998.

[140] G. Neukum and B. Ivanov. Crater size distributions and impact probabilities on earth from lunar, terrestrial-planet, and asteroid cratering data. *Hazards due to Comets and Asteroids*, 1:359–416, 1994.

[141] M. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[142] M. A. Nowak and R. M. May. Mathematical biology of HIV infections: antigenic variation and diversity threshold. *Mathematical Biosciences*, 106(1):1–21, 1991.

[143] R. Núñez-Queija. Queues with equally heavy sojourn time and service requirement distributions. *Annals of Operations Research*, 113(1-4):101–117, 2002.

[144] M. Nuyens and B. Zwart. A large-deviations analysis of the gi/gi/1 srpt queue. *Queueing Systems*, 54(2):85–97, 2006.

[145] A. G. Pakes. On the tails of waiting-time distributions. *Journal of Applied Probability*, 12(3):555¡V564, 1975.

[146] F. Papadopoulos, D. Krioukov, M. Boguñá, and A. Vahdat. Greedy forwarding in dynamic scale-free networks embedded in hyperbolic metric spaces. In *Proceedings of IEEE INFOCOM*, 2010.

[147] F. Papadopoulos, D. Krioukov, M. Boguñá, and A. Vahdat. Greedy forwarding in dynamic scale-free networks embedded in hyperbolic metric spaces. In *2010 Proceedings IEEE INFOCOM*, pages 1–9. IEEE, 2010.

[148] K. Papagiannaki, N. Taft, S. Bhattacharyya, P. Thiran, K. Salamatian, and C. Diot. A pragmatic definition of elephants in internet backbone traffic. In *Proceedings of ACM SIGCOMM Workshop on Internet measurment*, pages 175–176, 2002.

[149] V. Petrov. *Sums of independent random variables*. Springer, 1975.

[150] V. V. Petrov. *Limit theorems of probability theory: Sequences of independent random variables*. Oxford University Press, 1995.

[151] V. Pisarenko and M. Rodkin. *Heavy-tailed distributions in disaster analysis*, volume 30. Springer Science & Business Media, 2010.

[152] E. Pitman. On the behaviour of the characteristic function of a probability distribution in the neighbourhood of the origin. *Journal of the Australian Mathematical Society*, 8(03):423–443, 1968.

[153] E. Pitman. Subexponential distribution functions. *Journal of the Australian Mathematical Society (Series A)*, 29(03):337–347, 1980.

[154] D. d. S. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the Association for Information Science and Technology*, 27(5):292–306, 1976.

[155] Y. V. Prokhorov. An extremal problem in probability theory. *Theory of Probability & Its Applications*, 4(2):201–203, 1959.

[156] S. Re. Natural catastrophes and major losses in 1995. *Sigma*, (2), 1996.

[157] S. Resnick. *Heavy-tail phenomena: Probabilistic and statistical modeling*. Springer, 2006.

[158] S. I. Resnick. *Extremal processes*. John Wiley & Sons, 1982.

[159] S. I. Resnick. *Extreme values, regular variation and point processes*. Springer, 2013.

[160] C.-H. Rhee, J. Blanchet, and B. Zwart. Sample path large deviations for heavy-tailed l\'evy processes and random walks. *arXiv preprint arXiv:1606.02795*, 2016.

[161] P. Rooney. Microsoft's ceo: 80-20 rule applies to bugs, not just features. *CRN*, 2002.

[162] P. ROSIN and E. RAMMLER. The laws governing the fineness of powdered coal.-j. institute of fuel, 7: 29-36. evaluation of the folk and ward graphic measures. *J. Sedimen. Res*, 48:863–878, 1933.

[163] G. Samorodnitsky and M. S. Taqqu. *Stable non-Gaussian random processes: Stochastic models with infinite variance*. CRC press, 1994.

[164] A. Scheller-Wolf. Necessary and Sufficient Conditions for Delay Moments in FIFO Multiserver Queues with an Application Comparing s Slow Servers with One Fast One. *OPERATIONS RESEARCH*, 51(5):748–758, 2003.

[165] M. Shaked and J. G. Shanthikumar. *Stochastic orders*. Springer Science & Business Media, 2007.

[166] C. Shirky. Pareto principle. https://www.edge.org/responses/what-scientific-concept-would-improve-everybodys-cognitive-toolkit, 2011. Accessed: 2019-08-06.

[167] A. Shwartz and A. Weiss. *Large deviations for performance analysis: queues, communication and computing*. CRC Press, 1995.

[168] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, 1955.

[169] H. A. Simon. Some further notes on a class of skew distribution functions. *Information and Control*, 3(1):80–88, 1960.

[170] H. A. Simon. Reply to dr. mandelbrot's post scriptum. *Information and Control*, 4(2-3):305–308, 1961.

[171] H. A. Simon. Reply to "final note" by benoit mandelbrot. *Information and Control*, 4(2-3):217–223, 1961.

[172] D. Sornette and R. Cont. Convergent multiplicative processes repelled from zero: Power laws and truncated power laws. *Journal de Physique I*, 7(3):431–444, 1997.

[173] D. Stoyan. *Comparison methods for queues and other stochastic models*. John Wiley & Sons, 1983.

[174] J. Sutton. Gibrat's legacy. 1996.

[175] N. N. Taleb. *The black swan: The impact of the highly improbable*, volume 2. Random house, 2007.

[176] R. Tanaka, T.-M. Yi, and J. Doyle. Some protein interaction data do not exhibit power law statistics. *FEBS letters*, 579(23):5140–5144, 2005.

[177] J. L. Teugels. The class of subexponential distributions. *The Annals of Probability*, pages 1000–1011, 1975.

[178] R. Van Der Hofstad. *Random graphs and complex networks*, volume 1. Cambridge university press, 2016.

[179] N. Veraverbeke. Asymptotic behaviour of wiener-hopf factors of a random walk. *Stochastic Processes and their Applications*, 5(1):27–37, 1977.

[180] I. Voitalov, P. van der Hoorn, R. van der Hofstad, and D. Krioukov. Scale-free networks well done. *arXiv preprint arXiv:1811.02071*, 2018.

[181] I. Voitalov, P. van der Hoorn, R. van der Hofstad, and D. Krioukov. Scale-free networks well done. *arXiv preprint arXiv:1811.02071*, 2018.

[182] H. W. Watson and F. Galton. On the probability of the extinction of families. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 4:138–144, 1875.

[183] M. S. Wheatland and P. Sturrock. Avalanche models of solar flares and the distribution of active regions. *The Astrophysical Journal*, 471(2):1044, 1996.

[184] W. Whitt. *Stochastic-process limits: An introduction to stochastic-process limits and their application to queues*. Springer Science & Business Media, 2002.

[185] Wikipedia. List of earthquakes in 2010 — wikipedia, the free encyclopedia, 2017. [Online; accessed 7-August-2017].

[186] Wikipedia. List of earthquakes in 2011 — wikipedia, the free encyclopedia, 2017. [Online; accessed 7-August-2017].

[187] Wikipedia contributors. Scale invariance — Wikipedia, the free encyclopedia, 2018. [Online; accessed 7-September-2018].

[188] E. Willekens. On the supremum of an infinitely divisible process. *Stochastic processes and their applications*, 26:173–175, 1987.

[189] W. Willinger, D. Alderson, and J. C. Doyle. Mathematics and the internet: A source of enormous confusion and great potential. *Notices of the American Mathematical Society*, 56(5):586–599, 2009.

[190] W. Willinger, D. Alderson, and J. C. Doyle. Mathematics and the internet: A source of enormous confusion and great potential. *Notices of the American Mathematical Society*, 56(5):586–599, 2009.

[191] S. J. Young and E. R. Scheinerman. Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 138–149. Springer, 2007.

[192] G. U. Yule. A mathematical theory of evolution, based on the conclusions of dr. jc willis, f.r.s. *Philosophical transactions of the Royal Society of London. Series B, containing papers of a biological character*, 213:21–87, 1925.

[193] S. Zachary. A note on veraverbeke's theorem. *Queueing Systems*, 46(1):9–14, 2004.

[194] G. K. Zipf. Human behavior and the principle of least effort. 1949.

[195] V. M. Zolotarev. *One-dimensional stable distributions*, volume 65. American Mathematical Society, 1986.

[196] A. P. Zwart. Tail asymptotics for the busy period in the gi/g/1 queue. *Mathematics of Operations Research*, 26(3):485–493, 2001.

[197] B. Zwart. Rare events and heavy tails in stochastic systems, 2009.

[198] B. Zwart, S. Borst, and K. Debicki. Subexponential asymptotics of hybrid fluid and ruin models. *The Annals of Applied Probability*, 15(1A):500–517, 2005.

[199] B. Zwart, S. Borst, M. Mandjes, et al. Exact asymptotics for fluid queues fed by multiple heavy-tailed on–off flows. *The Annals of Applied Probability*, 14(2):903–957, 2004.