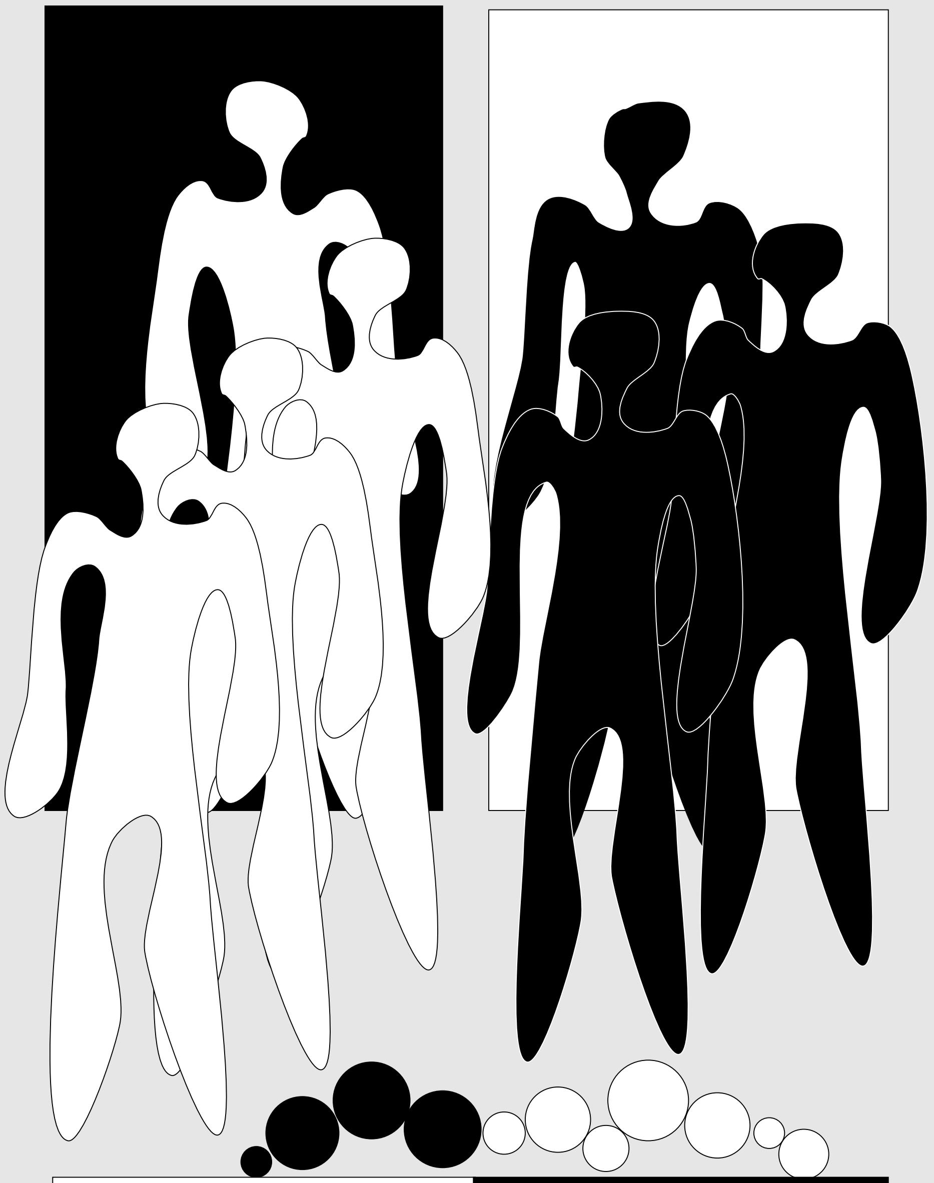

Adam Shwartz Alan Weiss

LARGE DEVIATIONS FOR PERFORMANCE ANALYSIS

This isn't the real cover. It's the one we wanted. Drawn by Freddy Bruckstein.

Table of contents

Chapter 0

What this Book Is, and What It Is Not

The field of communication and computer networks is bustling with activity. One of the active areas falls under the rubric "performance." Researchers and development engineers tackle systems that are huge, complex and fast; think of the telephone network in the United States. The resulting models are, for the most part, discrete-event, continuous time stochastic processes, technically known as jump Markov processes. The objective is to analyze the behavior of these systems, with the goal of designing systems that provide better service. "Better" may mean faster, less prone to error and breakdown, more efficient, or improved by many other criteria.

Until quite recently, the tools brought to bear on these problems were appropriate for small, simple systems. Some of these methods take into account only average behavior (or perhaps variances). But this is often not enough, as the performance of many systems is limited by events with a small probability of occurring, but with consequences that are severe. Clearly, new tools are needed. Computer simulation is one relatively new tool. But this method, for all its power, is limited in that it usually does not provide rules of thumb for design, may not give estimates on the sensitivity of results to various parameters, and can be extremely costly in terms of both computer time and programming (especially debugging) time. Analytic methods clearly retain some advantages. This book is about a fairly new analytic method called large deviations.

Large deviations is a mathematical theory that is very active at present. The theory deals with rare events, and is asymptotic in nature; it is thus a natural candidate for analyzing rare events in large systems. The theory of large deviations includes a set of techniques for turning hard probability problems that concern a class of rare events into analytic problems in the calculus of variations. It also provides a nice qualitative theory for understanding rare events. As an asymptotic technique, its effectiveness resides in the relative simplicity with which one may analyze systems whose size may be growing with the asymptotic parameter, or whose "conditioning" may be getting worse. The theory is often useful even when simulation or other numerical techniques become increasingly difficult as the parameter tends to its limit.

However, the theory is noted for being technically (mathematically) very demanding, and solving a problem in the calculus of variations is not typically an engineer's dream. Although the theory is being increasingly used for analyzing rare events in large systems, this is done by a relatively small number of researchers. We believe that the reason for this state of affairs is that the theory is not easily accessible to non-mathematicians, and the final results seem to require an additional translation to engineering lingo. Hence

Large deviations is useful.
Large deviations is formidably technical.
What's a student to do?

Herein is contained one point of view on what's to do. We develop the theory of large deviations from the beginning (independent, identically distributed (i.i.d.) random variables) through recent results on the theory for processes with boundaries, keeping to a very narrow path: continuous-time, discrete-state processes. By developing only what we need for the applications we present, we try to keep the theory to a manageable level, both in terms of length and in terms of difficulty. We make no particular claim to originality of the theory presented herein, except for the material concerning boundaries, which is the subject of Chapter 8. Most of the trailblazing work of Freidlin and Wentzell [FW], and of Donsker and Varadhan [DV1–DV4] goes further than we do. Also, others have subsequently treated the general theory much more thoroughly; e.g. Ellis [Ell], Wentzell [Wen], Deuschel and Stroock [DeS], Dembo and Zeitouni [DZ], and the recent work of Dupuis and Ellis [DE2]. We have, however, formulated a complete, self-contained set of theorems and proofs for jump Markov processes. Since our scope is limited to a class of relatively simple processes, the theory is much more accessible, and less demanding mathematically. To enhance the pedagogical value of this work, we have attempted to convey as much intuition as we could, and to keep the style friendly. In addition, we present for the first time a complete theory for processes with a flat boundary, and for some processes in a random environment. The level of the book is somewhat uneven, as indicated in the dependence chart Figure 0.1. This is purposeful—we believe that a neophyte would not want to read the difficult chapters, and that an expert doesn't want as much hand holding as a beginner.

We believe that our applications are important enough to require no apologies. As Mark Kac said, "Theorems come and go, but an application is forever." Our applications cover large areas of the theory of communication networks: circuit-switched transmission (Chapter 12), packet transmission (Chapter 13), multiple-access channels (Chapter 14), and the $M/M/1$ queue (Chapter 11). We cover aspects of parallel computation in a much more spotty fashion: basics of job allocation (Chapter 9), rollback-based simulation (Chapter 10), and assorted priority queuing models (Chapters 15 and 16) that may be used in performance models of various computer architectures.

The key word in the phrase "our applications" is "our." We present only our own results concerning the applications. We do not synthesize existing theory except in our narrow fashion for jump processes. We ignore possible improvements in order to remain within the realm of those large deviations bounds that we actually use. For example, Anantharam's beautiful results on the $G/G/1$ queue [An] are certainly relevant to the subjects we address, but his techniques are different. We do not obtain the best results known for jump Markov processes. It is certainly arguable whether this is a wise choice. However, we wanted to present a consistent, fully worked out point of view, avoiding digressions. Furthermore, once a student has learned the limited range of large deviations techniques we present, he or she should find it a much simpler matter to read both more abstract and complete works, and understand more wide-ranging applications. By limiting our range, we are able to give complete proofs for nearly all the results concerning our applications. We were also able to present a "bag of tricks" in the calculus of variations, which allows us to extract concrete information regarding these examples. We try to remedy some of the narrowness of our point of view in the end notes to the chapters and in the appendices.

On a less defensive note, we firmly believe that the large deviations of processes should be taught first for jump Markov processes. Diffusions are complicated objects, and the student does not need the extra burden of a subtle process to hinder the understanding of large deviations. Discrete time presents another unnecessarily difficult process, because the jumps are usually more general than those of the processes we consider. Furthermore, as we believe the book shows, there are many interesting applications of jump Markov processes. After all, we live in continuous time, and the events that occur in digital equipment are discrete.

As mentioned above, our book contains a new exposition of the theory of large deviations for jump Markov processes, but does not contain any new theory except for the results of §7.4 and Chapter 8. The applications contain many new results, though, and new derivations of previously known work. The new results include:

- A large deviations analysis of the $M/M/1$ queue that includes a surprising asymptotic formula for

  $$\mathbb{P}\left(x(0) = ny,\ x(nt) = nz\right)$$

  as n gets large, where $x(t)$ is the queue size at time t (§11.7).
- Fully proved large deviations principle for jump Markov processes with a flat boundary (Chapter 8).
- Analysis of a new class of Markov processes, "finite levels," for which both a fluid limit theorem and a large deviations principle are proved (Chapter 8).
- New analysis of an Aloha multiple-access protocol, using finite levels theory, gives the quasi-stability region for instant-feedback, continuous time Aloha (Chapter 14).
- New results for Erlang's model:
  - Transient analysis from any initial load (§12.5).
  - Transient analysis of a finite population model (§12.7.A).
  - Analysis of bulk service (large customers) (§12.7.B).
  - Transient analysis of trunk reservation (§12.8).
- New results for the AMS model:
  - Analysis of bit-dropping models (§13.7).
  - Calculation of buffer asymptotics for the multiple class case (§13.8).

- Analyses of a simple priority queue (§15.1), “serve the longest queue” (§ 15.6), and “join the shortest queue” (§15.10).
- Simple analysis of the Flatto-Hahn-Wright queueing system (Chapter 16).

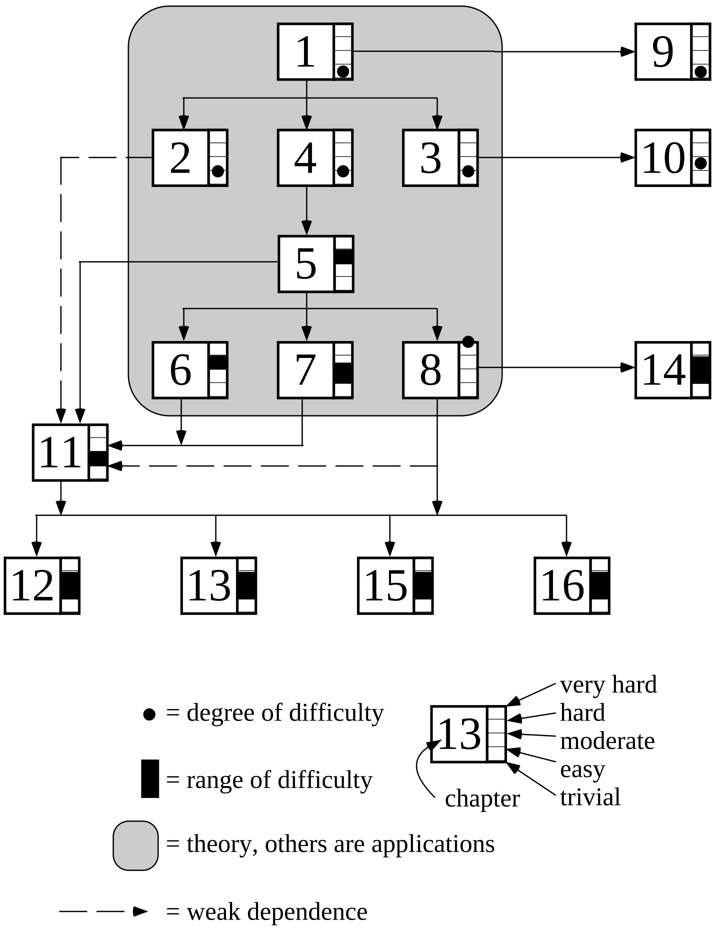

Figure 0.1. Dependence between the chapters.

0.1. What to Do with this Book

This book can be used as a basis for two types of one-semester courses. The first is an introduction to the theory of large deviations, through jump Markov processes. This course should cover most of Chapters 1, 2, 5, 6, Appendix D, and possibly the advanced material in Chapter 8. Such a course would prepare the student to read the more mathematical theory, and to fully appreciate the applications worked out in the rest of the book. It would be wise (in our opinion) to sprinkle such a theory-oriented course with some of the applications.

The second course is application-oriented. Such a course should probably start with Chapter 1 (at least §1.1–1.3), so that some flavor of the theory is provided. The results of §1.4, 2.1, and 2.3, and of Chapters 5–8 can then be stated without proof, with or without intuitive explanations. Some basic tools from the calculus of variations, at least to the extent summarized in Appendix C, should be covered. Then applications can be presented, according to the dependence chart shown in Figure 0.1.

Chapter 3 provides an easy application of the basic theory, and can thus be used to motivate the more general (and more technical) process-level theory. Chapter 4 summarizes some basic results concerning the Poisson process, and more generally jump Markov processes. There is nothing new in that chapter, but it is a strict prerequisite for the rest of the book. Finally, in the appendices we collect, for easy reference, some background material from analysis and probability theory.

In our judgment, the prerequisites for such courses (and for reading the book) are probability and analysis at a level of first-year graduate courses for engineering students, or senior-level courses for students of mathematics. The applications course can be done with much less background, provided the student is willing to believe the material as summarized in the appendices. However, some mathematical maturity (even affinity) is required.

0.2. About the Format of the Book

There are four types of exercises in the book. Some results that are easy to prove, important but not central to our development are presented as exercises. In some cases, extensions are relegated to an exercise when they are deemed not-too-hard but long; this is simply to save space. Examples and special cases are given as exercises, and are intended to help build intuition, or clarify a technical point. These exercises are an integral part of the text, and should at least be read, preferably solved. The last type of exercises are marked JFF (“just for fun”). The end of an Exercise is marked thus: ♠

There are two counters in the book: one for equations, one for all theorems, propositions, lemmas and corollaries, exercises, examples, figures, assumptions, and definitions. Equation numbers are written as (Chapter-Number.Equation-Number), and other numbers as Chapter-Number.Number. References appear in square brackets [ ], and we use either the first two letters of the author’s last name, or—in the case of multiple authors—all initials. Conflicts are resolved creatively.

The index identifies definitions by bold page numbers, and includes the frequently used symbols.

> We often wish to make a comment, or expand on a particular topic, in such a way that the reader may feel free to skip the comment, but will know that it is there. This is how we do it: in small type, in a narrow paragraph.

0.3. Acknowledgments

This project spanned many more years than we had ever anticipated. In the course of those years we have had help from many, many people. Preeminent among them are Armand Makowski, Debasis Mitra, and S.R.S. Varadhan. Debasis was steadfast in his support: moral, financial, and technical. He believed in us when we weren't sure we believed in ourselves. This project would never have been done without him. And we would never have gotten into the field (it is not certain that there would be much of a field to get into!) without Professor Varadhan. It was a tremendous comfort to know that there was no technical point, however difficult or subtle, that could not be answered almost instantly by a simple visit to NYU. Armand Makowski not only introduced us, and not only is he responsible for stating that the world would benefit from lecture notes on queueing applications of large deviations, but he can also be held accountable for doing something about it. With the support of John Baras, Armand arranged a sabbatical at the Systems Research Center where a first draft of these notes was hammered out by AW, and provided a sabbatical at the Systems Research Center where, somewhat unexpectedly, most of the time of AS was devoted to this project.

There are many more people who have helped over the years. Robert J. Vanderbei was, for a time, a coauthor of the book, and one appendix still bears his sole authorship. Several "field tests" of these ideas were graciously hosted by Armand Makowski at the University of Maryland, College Park, by Elja Arjas and the Finnish Summer School, and by Sid Browne and the Columbia Business School and Department of Mathematics. Within Bell Labs and the Technion, our home institutions, it seems that nearly everyone had something to contribute. Notable among those were co-large deviants Ofer Zeitouni and Amir Dembo. Also, Marty Reiman was a constant source of technical wisdom, moral support (what do you mean you aren't done yet?), and was an invaluable asset in transportation and living accommodations. Howard Trickey was our accessible TeX wizard, and justified his title hands down. Thanks also to Andrew Trevorrow for long-distance TeX help.

There were many students and colleagues who gave suggestions and feedback on everything from typos to approach, from early drafts to the first printing of this book. They include also those attending several courses given at the Technion, as well as lectures delivered at AT&T. We are particularly grateful for comments from Laurence Baxter, Henri Casanova, Hong Chen, Bill Cleveland, Ed Coffman, Amir Dembo, Amanda Galtman, Leo Flatto, Ben Fox, Predrag Jelenkovic, Ariel Landau, Armand Makowski, Colin Mallows, Bill Massey, Jim Mazo, Debasis

Mitra, Marty Reiman, Emre Telatar, Stephen R.E. Turner, Yashan Wang, Phil Whiting, Ward Whitt, Paul Wright, Aaron Wyner, Ofer Zeitouni, and Li Zhang.

The editor-in-chief of this series Laurence Baxter did yeoman's work. Our editor John Kimmel amazed us by answering "yes" to every one of our requests, and promptly, too!

Typists Sue Pope and Lesley Price helped turn scribbled handwriting into beautiful TEX, quickly, accurately, and cheerfully.

This book was produced using TEX, with AS serving as local TEXpert, and was set in Times Roman, with MathTimes and other math fonts from Y&Y. The figures were drawn by AW using Canvas©, and according to the egalitarian tendencies of the authors, was set on Macintosh©, UNIX©, and various PC computers and clones.

I (AS) am delighted for this opportunity to acknowledge Armand Makowski for his role as colleague, collaborator, and catalyst in my professional life and, above all, to express my appreciation for his friendship.

And I (AW) am eternally grateful for my two mentors, Debasis Mitra and Raghu Varadhan. These two fine men have unselfishly nurtured me throughout this and other projects. I hope that in some way they can find some recompense in this volume.

Finally, our families, particularly our wives Shuli Cohen Shwartz and Judy Weiss, deserve thanks for putting up with us during all these years of labor. While we'll never know, it was probably as hard on them as having children; it was certainly longer and with less reward at the end. We promise we'll never do it again.

Note: This printing incorporates all the corrections we accumulated during the first year the book was out. We thank our readers for reporting these mistakes, and our publisher for allowing us to make the changes. However, from the number of mistakes found so far, we know that more will be found. Please send comments to us at apdoo@research.bell-labs.com or adam@ee.technion.ac.il. You can obtain the latest errata sheet at any of the following locations:

http://cm.bell-labs.com/who/apdoo

http://www-ee.technion.ac.il/~adam

http://users.aol.com/apdoo

Chapter 1

Large Deviations of Random Variables

This chapter can be viewed as a guided tour through basic large deviations. Following a heuristic exposition, we derive large deviations estimates for random variables. We provide proofs when these provide insight, or are typical; otherwise, we provide references. The modern tools and approaches, especially those that have proved useful for the applications, are discussed in Chapter 2 and Appendix D.

The main results, Theorems 1.5, 1.10, and 1.22, as well as the computations of Examples 1.13–1.18 and Exercises 1.6, 1.17–1.25, will be used heavily throughout the book.

1.1. Heuristics and Motivation

Estimates of probabilities of rare events turn out to have an exponential form; i.e., these probabilities decrease exponentially fast as a function of the asymptotic parameter. To motivate the exponential form of the large deviations estimates, consider the following examples. Let $x_1, x_2, \ldots$ be a sequence of independent, identically distributed (i.i.d.) random variables with a common distribution function F and finite mean. Fix a number $a > \mathbb{E}x_1$. Now the probability that $x_1 + \cdots + x_n > na$ is clearly decreasing in n in a long-term sense, since by the (weak) law of large numbers

$$\mathbb{P}\left(\frac{x_1 + \cdots + x_n}{n} \geq a\right) \to 0 \text{ as } n \to \infty.$$

The next question would be: How fast does this probability decrease? Let us perform some quick calculations. First, if for some integer k,

$$x_{jk+1} + \ldots + x_{(j+1)k} \geq ak \quad \text{for all} \quad j = 0, \ldots, (n/k) - 1,$$

then clearly $x_1 + \cdots + x_n \geq na$. Therefore

$$\begin{aligned}\mathbb{P}&\left(\frac{x_1 + \cdots + x_n}{n} \geq a\right) \\ &\geq \mathbb{P}\left(x_{jk+1} + \ldots + x_{(j+1)k} \geq ak \quad \text{for all} \quad j = 0, \ldots, (n/k) - 1\right) \\ &= \left(\mathbb{P}\left(x_1 + \ldots + x_k \geq ak\right)\right)^{n/k}\end{aligned}$$

by independence. This immediately implies that the rate of convergence is at most exponential. On the other hand, for any positive θ, by Chebycheff's inequality

(Theorem A.113),

$$\begin{aligned}\mathbb{P}\left(x_1+\cdots+x_n\geq na\right) &= \mathbb{P}\left(e^{\theta(x_1+\cdots+x_n)}\geq e^{\theta na}\right)\\ &\leq e^{-\theta na}\,\mathbb{E}e^{\theta(x_1+\cdots+x_n)}\\ &= e^{-\theta na}\left(\mathbb{E}e^{\theta x_1}\right)^n\\ &= \left(e^{-\theta a}\,\mathbb{E}e^{\theta x_1}\right)^n\end{aligned}$$

by independence. For the right choice of θ, this exponential expression is decreasing:

Exercise 1.1. Show that if $a > \mathbb{E}x \geq 0$ and if $\mathbb{E}e^{\theta x} < \infty$ for all $|\theta|$ small, then $e^{-\theta a}\mathbb{E}e^{\theta x} < 1$ for some θ. Hint: compute $d\,\mathbb{E}e^{\theta x}/d\theta$ at $\theta = 0$. ♠

Thus, probabilities should decay exponentially in n. The questions are: Do the rates in the upper and lower bound agree, and if so, how do we compute the right exponent? In §1.2 we show that they are indeed the same, and give a formula. In §1.3 we compute several examples. Anticipating the shape of things to come, the arguments indicate that

$$\mathbb{P}\left(\sum_1^n x_i \geq na\right) = e^{-n\ell(a)+o(n)}, \tag{1.1}$$

where the function ℓ depends on the distribution F. For the meaning of $o(n)$ see Definition A.14.

Here is another view that some find quite intuitive. If, indeed, $x_1+\cdots+x_n \geq an$, then probably $x_1+\cdots+x_n \approx na$ (for an illustration see Exercise 1.2 below). Moreover, it is likely that this happens by nearly-equal splitting, i.e., $x_1+\ldots+x_{n/2} \approx an/2$ and $x_{(n/2+1)}+\ldots+x_n \approx an/2$, with an error of order $\sqrt{n}$. (This issue, of how improbable things happen, is explained in later chapters.)

Exercise 1.2. Show that in the case of fair coin flips, if x is the number of heads obtained in n flips and $0.8n$ is an integer,

$$\mathbb{E}\left(x-0.8n \;\middle|\; x\geq 0.8n\right) \to \frac{1}{3} \quad \text{as} \quad n\to\infty,$$

and does not grow with n ! Hint: $\binom{n}{0.8n+1} \approx \frac{1}{4}\cdot\binom{n}{0.8n}$. ♠

Exercise 1.3. Compare the chances of obtaining $(1/2+\alpha)\cdot n$ heads in n coin flips, with $0<\alpha<1/2$ in the following two ways: (i) by getting two series of $n/2$ flips, each with $\alpha\cdot n/2$ heads more than expected. (ii) by obtaining the additional heads in one series of length $n/2$ with the other series being "normal." Hint: use Stirling's formula. ♠

These heuristics imply that, for large n, we have the rough estimate

$$\begin{aligned}\mathbb{P}\left(\sum_1^n x_i \geq an\right) &\approx \mathbb{P}\left(\sum_1^{n/2} x_i \geq a\frac{n}{2}\,,\ \sum_{n/2+1}^{n} x_i \geq a\frac{n}{2}\right)\\ &\approx \left[\mathbb{P}\left(\sum_1^{n/2} x_i \geq a\frac{n}{2}\right)\right]^2 .\end{aligned}$$

Similarly, for any k much smaller than n,

$$\mathbb{P}\left(\sum_1^n x_i \geq an\right) \approx \left[\mathbb{P}\left(\sum_1^{n/k} x_i \geq a\frac{n}{k}\right)\right]^k ;$$

If we could choose k to be linear in n, we would see that this probability decreases exponentially fast. However, in general,

$$\mathbb{P}\left(\sum_1^n x_i \geq an\right) \not\approx \left[\mathbb{P}\left(x_1 \geq a\right)\right]^n$$

[for Bernoulli random variables with $a = 1/2$, the left-hand side is $1/2$ while the right is $(1/2)^n$!]. Thus the integer k above cannot quite grow linearly with n. This indicates that indeed (1.1) is to be expected, and that the "error term" $o(n)$ cannot be omitted.

Let us now illustrate some of the ideas from a different angle. To avoid technical difficulties, assume that the distribution function satisfies $F(1) = 0$, $F(2) = 1$. Let $\mu \stackrel{\triangle}{=} \mathbb{E}x_1$ and $\alpha \stackrel{\triangle}{=} \mathbb{E}\log x_1$. Then

$$\mathbb{E}\left(x_1 \cdot x_2 \cdots x_n\right) = \mu^n .$$

Let us estimate this expectation in a different way. Write

$$\begin{aligned}\mathbb{E}\left(x_1 \cdot x_2 \cdots x_n\right) &= \mathbb{E}\left(\exp\left(\log x_1 + \ldots + \log x_n\right)\right)\\ &= \mathbb{E}\left(\exp n\left(\frac{\log x_1 + \ldots + \log x_n}{n}\right)\right).\end{aligned} \tag{1.2}$$

By the strong law of large numbers,

$$\mathbb{P}\left(\frac{\log x_1 + \ldots + \log x_n}{n} \to \alpha\right) = 1$$

so that we expect $\mathbb{E}(x_1 \cdot x_2 \cdots x_n)$ to grow exponentially, roughly as $e^{n\alpha}$. However, by Jensen's inequality (A.11),

$$\mu \stackrel{\triangle}{=} \mathbb{E}x_1 = \mathbb{E}e^{\log x_1} > e^{\mathbb{E}\log x_1} = e^{\alpha}\ !$$

Clearly, the law of large numbers is not precise enough to estimate this expectation. Indeed, in this case we cannot expect convergence w.p.1 to imply convergence in expectation, since we are taking expectations of something that may grow quickly. Here is a refinement that will consolidate the two calculations. Suppose that, as in (1.1), we have an exponential estimate for the density of the sample averages of the sequence $\log x_1, \log x_2, \ldots$

$$\mathbb{P}\left(a \le \frac{\log x_1 + \cdots + \log x_n}{n} \le a + da\right) \approx e^{-n\ell(a)} da$$

for some non-negative function ℓ, and suppose that $\ell(a)/|a| \to \infty$ as $|a| \to \infty$. Then

$$\begin{aligned}\mathbb{E}\,(x_1 \cdot x_2 \cdots x_n) &= \mathbb{E}\exp\left(n \cdot \frac{\log x_1 + \cdots + \log x_n}{n}\right) \\ &\approx \int e^{na} e^{-n\ell(a)}\, da = \int e^{n(a-\ell(a))}\, da.\end{aligned}$$

Suppose the maximum $m \stackrel{\triangle}{=} \sup_a (a - \ell(a))$ is attained at some point and write

$$\int e^{n(a-\ell(a))}\, da = e^{nm} \int e^{n(a-\ell(a)-m)}\, da.$$

By the assumptions on ℓ, $a - \ell(a)$ diverges to $(-\infty)$ as $|a| \to \infty$, so that it is strictly negative outside a finite interval. Thus, the integrand in the last integral goes to zero (exponentially fast) as $n \to \infty$, except where the maximum is attained, so that

$$\mathbb{E}\,(x_1 \cdot x_2 \cdots x_n) \le e^{n(m+\varepsilon)}$$

for all $\varepsilon > 0$ and all n large. By looking at the points where $a - \ell(a) > m - \varepsilon$ we have

$$\mathbb{E}\,(x_1 \cdot x_2 \cdots x_n) \ge e^{n(m-\varepsilon)}$$

for every positive ε (this idea of estimating the rate of growth of an integral by considering the maximum of the integrand is called Laplace's method). We summarize these two inequalities using the notation

$$\mathbb{E}\,(x_1 \cdot x_2 \cdots x_n) \asymp \exp\left(n \cdot \sup_a (a - \ell(a))\right), \tag{1.3}$$

where the meaning of $\asymp$ is that the left-hand side grows exponentially fast, with rate $\sup_a (a - \ell(a))$. We will find in §2.2, as part of the derivation of large deviations estimates, that $\sup_a (a - \ell(a)) = \log \mu$, giving the correct exponential growth rate.

But this is just a formal calculation, and you are probably asking yourself now, "How can this be? I know that the mean is μ^n, and I've seen that the strong law of large numbers implies that the mean is almost surely near $e^{\alpha n}$, but how do I reconcile the two?" Let's consider what would happen if you would actually try to run an experiment to estimate $\mathbb{E}(x_1 \cdot x_2 \cdots x_n)$. You would collect n samples of the x_i, and then evaluate the product. You would undoubtedly (law of large numbers) come up with a number in the range of $e^{\alpha n}$. Repeat the experiment, and the results would be similar. However, after a great many experiments, you would come up with an unusually large observation—say something near μ^n. This observation is so large relative to the others that it completely dominates the mean you have been keeping, so that all of a sudden the mean looks like μ^n even though only one observation was of that order. Now what keeps an even more colossal observation from skewing further the observed mean? The answer is that it is too improbable for it to happen often [remember $\ell(a)$ grows quickly with a]. It will happen so rarely, that enough observations have been taken to completely dilute the effect of the "extra large" observation. This is the tradeoff we see between $\ell(a)$ ("the probability") and a ("the size"), and is the reason that $\sup_a(a - \ell(a))$ is the important quantity. It also serves to demonstrate that, sometimes, rare events are the most important to determine what's going on.

Stock and investment models

The last example has more than purely theoretical or pedagogical interest; it has monetary applications. Consider that investments usually pay an amount proportional to the investment. Suppose that an investment is risky; to be precise, an investment of one unit at the beginning of the i^{th} period yields x_i units at the end of the period [which is the beginning of the $(i+1)^{\text{th}}$ period]. Hence, after n periods, the value of a unit investment made at the beginning of the first period is $\prod_{i=1}^n x_i$.

How should we value an investment? This is a complicated question, but we have just seen that the return after a large number of periods is *most likely* to be near $\exp\left(n\mathbb{E}\log x_i\right)$, not near $\mathbb{E}\prod_{i=1}^n x_i$. Optimal investment strategies are based on this and related observations. See Kelly [KeJ], Algoet and Cover [AC], and references therein.

Beyond deviations from the mean.

Sanov's Theorem, introduced in §2.4, takes us one step up to "Level 2 Large Deviations." The question we ask there is: What do the random variables $x_1, x_2, \ldots$ look like when they do make a big excursion (such as making $\prod_{i=1}^n x_i \geq \mu^n$)? It turns out that *they all look like they are sampled from a "tilted distribution," one for which* $\mathbb{E}\log x = \log\mu$. In other words, the product becomes large because of conspiracies, **not** because of outliers. This conspiracy is a very rare occurrence, but when it occurs, its effect is huge. This is captured by the balance between the size of the effect e^{na}, and the rarity $e^{-n\ell(a)}$. Whereas in §1.1 we ask "How likely is it for the sample mean to deviate from the ensemble mean?," Sanov's Theorem addresses the question "How likely is it for the empirical distribution to deviate from the true distribution?" But let us establish first things first.

1.2. I.I.D. Random Variables

Chernoff's Theorem establishes (1.1) for i.i.d. random variables. The proof consists of an upper bound and a lower bound. The upper bound is just a parameterized version of Chebycheff's inequality (A.9) applied to the function $e^{\theta x}$. The lower bound uses a change of measure argument much as in importance sampling. These ideas generalize to random vectors and to processes, and will be used in all our large deviations proofs.

So, consider a sequence $x_1, x_2, \ldots$ of i.i.d. random variables with common distribution function F, and assume the mean $\mathbb{E}x_1$ exists. Define

$$M(\theta) = \mathbb{E}e^{\theta x_1} \tag{1.4a}$$

$$\ell(a) = -\log\left(\inf_\theta e^{-\theta a} M(\theta)\right) = \sup_\theta \left(\theta a - \log M(\theta)\right). \tag{1.4b}$$

$M(\theta)$ is the *moment generating function* of the random variable x_1. The function $\log M(\theta)$ is called the *logarithmic moment generating function* or *cumulant generating function* of x_1. Note that ℓ is non-negative [put $\theta = 0$ in (1.4 b)] and convex (by Theorem A.47, being the supremum of linear, hence convex functions); see Proposition 5.10, §5.2. The transformation applied to $\log M$ in (1.4b) is variously called the convex transform, Fenchel transform, Legendre transform, or Cramér transform.

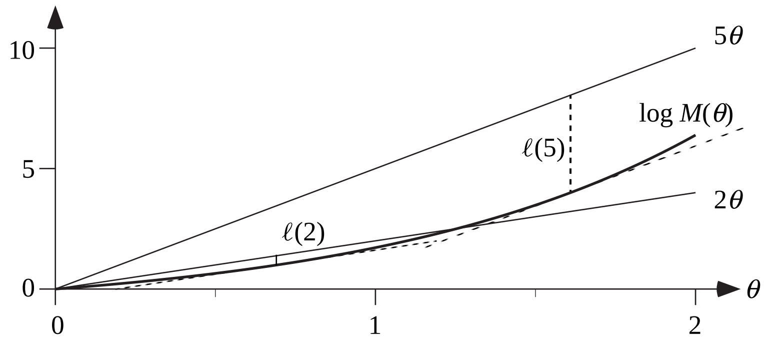

Figure 1.4. The ℓ-function: computing the Legendre transform.

By Exercise A.92, if the supremum in (1.4) is attained at a point θ^* in the interior of the interval where $M(\theta)$ is finite, then $M(\theta)$ is differentiable at θ^*, so that

$$\ell(a) = -\log \mathbb{E}e^{\theta^*(x_1 - a)}. \tag{1.5}$$

Theorem 1.5. *Consider the sequence $x_1, x_2, \ldots$ of i.i.d. random variables. For every $a > \mathbb{E}x_1$ and positive integer n,*

$$\mathbb{P}(x_1 + \cdots + x_n \geq na) \leq e^{-n\ell(a)}. \tag{1.6a}$$

Assume that $M(\theta) < \infty$ for θ in some neighborhood of 0 and that (1.5) holds for some θ^* in the interior of that neighborhood. Then for every $\varepsilon > 0$ there exists an integer n_0 such that whenever $n > n_0$,

$$\mathbb{P}(x_1 + \cdots + x_n \geq na) \geq e^{-n(\ell(a)+\varepsilon)}. \tag{1.6b}$$

Equations (1.6a)–(1.6b) imply that

$$\mathbb{P}(x_1 + \cdots + x_n \geq na) = e^{-n\ell(a)+o(n)}. \tag{1.6c}$$

Remark. This result holds, in fact, for i.i.d. random variables without any assumptions. The general case is proved using an approximation argument; see, e.g., Chernoff [Ch], Dembo and Zeitouni [DZ §2.2] and Theorem 1.10 below.

Proof. By virtue of Exercise 1.6 below, it suffices to establish the result when $\mathbb{E}x_1 = 0$. The upper bound is proved using an exponential estimate. First fix $\theta \geq 0$.

$$\begin{aligned}
\mathbb{P}(x_1 + \cdots + x_n \geq na) &\leq \mathbb{P}\left(e^{\theta(x_1+\cdots+x_n)} \geq e^{\theta na}\right) && e^{\theta x} \text{ is increasing} \\
&\leq e^{-\theta na}\mathbb{E}e^{\theta(x_1+\cdots+x_n)} && \text{Chebycheff} \\
&= \left(e^{-\theta a}\mathbb{E}e^{\theta x_1}\right)^n && \text{independence.}
\end{aligned}$$

Equality in the first relation holds whenever $\theta > 0$. Since $\theta \geq 0$ was arbitrary, taking the infimum would yield (1.6a) provided we show that we can ignore negative values of θ in (1.4b).

By Jensen's inequality,

$$M(\theta) = \mathbb{E}e^{\theta x_1} \geq e^{\mathbb{E}\theta x_1} = 1$$

for all θ. Thus, since $a > 0$,

$$e^{-\theta a}M(\theta) \geq 1 \text{ for } \theta \leq 0,$$

with equality for $\theta = 0$. Therefore in computing the middle term of (1.4), we can restrict the range of the infimum to $\theta \geq 0$, i.e.,

$$\inf_\theta e^{-\theta a}M(\theta) = \inf_{\theta\geq 0} e^{-\theta a}M(\theta).$$

This completes the proof of the upper bound (1.6a).

The lower bound is established using a change of measure (if you are unfamiliar with the idea of a change of measure, see §A.4). Let F be the distribution of x_1. Then G, defined by

$$G(x) = \left[M(\theta^*)\right]^{-1}\int_{-\infty}^{x} e^{\theta^* y}\,dF(y), \tag{1.7}$$

is a new distribution function (check!) (G is the tilted distribution referred to at the end of §1.1). For any real α, we clearly have

$$\begin{aligned}\mathbb{P}(x_1 \geq \alpha) &= \int_{-\infty}^{\infty} \mathbf{1}\left[y \geq \alpha\right]\, dF(y)\\ &= \int_{-\infty}^{\infty} \mathbf{1}\left[y \geq \alpha\right] e^{-\theta^* y} e^{\theta^* y}\, dF(y)\\ &= M(\theta^*) \int_{-\infty}^{\infty} \mathbf{1}\left[y \geq \alpha\right] e^{-\theta^* y}\, dG(y)\end{aligned}$$

by the definition of G. Applying this idea to the left-hand side of (1.6b),

$$\begin{aligned}\mathbb{P}(x_1 + \cdots + x_n \geq na) &= \int \cdots \int \mathbf{1}[y_1 + \cdots + y_n \geq na]\, dF(y_1) \cdots dF(y_n)\\ &= \int \cdots \int \mathbf{1}[y_1 + \cdots + y_n \geq na]\, e^{-\theta^*(y_1 + \cdots + y_n)}\\ &\qquad \times e^{\theta^* y_1}\, dF(y_1) \cdots e^{\theta^* y_n}\, dF(y_n).\end{aligned}$$

Changing to the measure G, we have, for any $\varepsilon' > 0$,

$$\begin{aligned}&\mathbb{P}(x_1 + \cdots + x_n \geq na) \qquad (1.8a)\\ &\quad = \left[M(\theta^*)\right]^n \int \cdots \int \mathbf{1}[y_1 + \cdots + y_n \geq na]\, e^{-\theta^*(y_1 + \cdots + y_n)}\\ &\qquad\quad dG(y_1) \cdots dG(y_n)\\ &\quad \geq \left[M(\theta^*)\right]^n \int \cdots \int \mathbf{1}[n(a+\varepsilon') \geq y_1 + \cdots + y_n \geq na]\, e^{-\theta^*(y_1 + \cdots + y_n)}\\ &\qquad\quad dG(y_1) \cdots dG(y_n)\\ &\quad \geq \left[M(\theta^*)\right]^n e^{-n\theta^*(a+\varepsilon')} \int \cdots \int \mathbf{1}[n(a+\varepsilon') \geq y_1 + \cdots + y_n \geq na]\\ &\qquad\quad dG(y_1) \cdots dG(y_n).\end{aligned}$$

Let $\tilde{x}_1, \ldots, \tilde{x}_n$ be i.i.d. random variables, distributed according to G. Then the probability in the first expression of (1.8a) is bounded below by

$$\left[M(\theta^*)\right]^n e^{-n\theta^*(a+\varepsilon')}\, \mathbb{P}\left(n(a+\varepsilon') \geq \tilde{x}_1 + \cdots + \tilde{x}_n \geq na\right). \qquad (1.8b)$$

We now provide a lower bound for the probability on the right of (1.8b). First, since $M(\theta)$ is finite in a neighborhood of θ^*, it is differentiable there by Exercise A.92. Therefore

$$\frac{d^n}{d\theta^n} M(\theta^*) = \mathbb{E} x_1^n e^{\theta^* x_1} < \infty, \qquad n = 1, 2, \ldots$$

and, in particular,

$$\mathbb{E}\tilde{x}_1^2 \triangleq \frac{\mathbb{E} x_1^2 e^{\theta^* x_1}}{M(\theta^*)} < \infty.$$

Since $\inf_\theta \mathbb{E}e^{\theta(x_1-a)} = \mathbb{E}e^{\theta^*(x_1-a)}$ and it is differentiable, the derivative vanishes at θ^*, so that

$$\mathbb{E}(x_1-a)e^{\theta^*(x_1-a)} = 0, \quad \text{or} \quad \mathbb{E}x_1 e^{\theta^* x_1} = aM(\theta^*).$$

This implies that the change of measure puts the mean of $\tilde{x}$ exactly at a since

$$\mathbb{E}\tilde{x}_1 = \int y dG(y) = \left[M(\theta^*)\right]^{-1} \int y e^{\theta^* y} dF(y) = a.$$

Consider the sum in (1.8b) of the i.i.d. random variables $\tilde{x}_1, \ldots, \tilde{x}_n$. Since these random variables have mean a and finite variance, the central limit theorem A.112 implies

$$\begin{aligned}\mathbb{P}\Big(n(a+\varepsilon') \geq \tilde{x}_1 + \cdots + \tilde{x}_n \geq na\Big) &= \mathbb{P}\left(\sqrt{n}\,\varepsilon' \geq \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\tilde{x}_i - a) \geq 0\right) \\ &\to \frac{1}{2} \quad \text{as} \quad n \to \infty.\end{aligned}$$

Let n_0 be such that the probability exceeds $1/4$ whenever $n \geq n_0$ (clearly n_0 depends on ε'). Then for $n \geq n_0$,

$$\begin{aligned}\mathbb{P}(x_1 + \cdots + x_n \geq na) &\geq \frac{1}{4}\left[e^{-\theta^* a} M(\theta^*)\right]^n e^{-n\theta^*\varepsilon'} \\ &= \frac{1}{4} e^{-n\ell(a)} e^{-n\theta^*\varepsilon'}.\end{aligned}$$

Now since $\theta^* > 0$ (why?) we can choose ε' so that $(1/4)e^{-n\theta^*\varepsilon'} \geq e^{-n\theta^*\varepsilon}$ whenever $n \geq n_0$. This proves (1.6b–1.6c). ■

Exercise 1.6. Let $y_i \stackrel{\triangle}{=} x_i + \bar{y}$ for each $i \geq 1$, where $\bar{y}$ is a fixed constant. Express the moment generating function M_y and the Cramér transform ℓ_y of its logarithm in terms of M_x and ℓ_x. Write Theorem 1.5 for $y_1, y_2, \ldots$ and conclude that the zero-mean assumption on $x_1, x_2, \ldots$ is without loss of generality. ♠

Exercise 1.7. Assume $\ell(a)$ is continuous, and re-derive the lower bound without invoking the central limit theorem. Hint: use the law of large numbers. ♠

Exercise 1.8. Let $y_1, y_2, \ldots$ be independent (but not necessarily identically distributed!) random variables so that for all i and M, $\mathbb{P}(|y_i| \geq M) \leq e^{-rM}$ for some $r > 0$. Then there exists a continuous function $f(\varepsilon)$, which depends only on r, so that for all $\varepsilon > 0$ we have $f(\varepsilon) > 0$ and

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}(y_i - \mathbb{E}y_i)\right| > n\varepsilon\right) \leq e^{-nf(\varepsilon)}.$$

Hint: compute separately for the case that the sum is larger than $n\varepsilon$ and smaller than $(-n\varepsilon)$, and start with the zero-mean case. Use Chebycheff's inequality as in the proof of Theorem 1.5. Prove $\mathbb{E}e^{|\theta y_i|}$ is finite for all θ small, uniformly in i. Use Exercise A.92 to conclude that the functions $f_i(\varepsilon, \theta) \stackrel{\triangle}{=} e^{-\theta\varepsilon}\mathbb{E}e^{\theta y_i}$ have continuous derivatives (of all orders!) near $(\varepsilon, \theta) = (0, 0)$. Now use a Taylor expansion in the two variables ε, θ to second order and set $\varepsilon = \theta$. Obtain a bound of the form $(1 - c\varepsilon^2)^n$ with $c > 0$ that holds for small ε. ♠

Remark. To compute $\mathbb{P}(x_1 + \cdots + x_n \leq na)$ for $a < \mathbb{E}(x_1)$, note from (1.4) that the ℓ-function for the sequence $-x_1, -x_2, \ldots$ is equal to the ℓ-function of $x_1, x_2, \ldots$ with the sign of its argument reversed, so that, for $a < 0$, by Chernoff's Theorem,

$$\mathbb{P}(x_1 + \cdots + x_n \leq na) = e^{-n\ell(a)+o(n)}.$$

A more detailed statement of a large deviations theorem in $\mathbb{R}^1$ and under weaker conditions is given in Theorem 1.10 below.

Theorem 1.5 gives us an estimate of the probability that the sample mean lies in the half-line above $a > \mathbb{E}x_1$, and the remark extends this to the half-line below $\mathbb{E}x_1$. This easily extends to more general sets. Define the real-valued function J on sets S in $\mathbb{R}$ by

$$J(S) \stackrel{\triangle}{=} \inf_{a \in S} \ell(a).$$

Corollary 1.9. *Assume $M(\theta) < \infty$ for θ in some neighborhood of zero. Then, for every open set S and positive integer n,*

$$\mathbb{P}\left(\frac{x_1 + \cdots + x_n}{n} \in S\right) = e^{-nJ(S)+o(n)}.$$

Note that such a result is not possible for closed sets: in particular, single points are closed sets, and if x_1 possesses a density, then the probability that the sample mean is in the set is zero.

For a proof of this corollary see Dembo and Zeitouni [DZ]. Here is a heuristic argument (when $\mathbb{E}x_1 = 0$). An application of Jensen's inequality (to the convex function $(-\log\alpha)$: use the definition of ℓ with θ fixed) shows that $\ell(0) \leq 0$, and since ℓ is non-negative, $\ell(0) = 0$. Thus the result is just the weak law of large numbers if $(0 =)\mathbb{E}x_1 \in S$. Now ℓ is non-negative and convex, so that it is increasing for $x > 0$ and decreasing for $x < 0$. But then there is a point, say a, in the closure of S so that $\ell(a) = J(S)$. Since S is open, there is an interval in S whose endpoint is a. The argument of the lower bound now applies, since (1.8) uses only a small interval, so that the same lower bound holds for all open sets for which a is a minimum point. For the upper bound, enclose S by the two smallest semi-infinite intervals $(-\infty, a^-]$ and $[a^+, \infty)$ and apply Theorem 1.5.

Actually, this discussion is generic in that lower bounds are usually proved locally, while upper bounds are established by increasing the sets.

The one-dimensional case is unique in that the upper bound holds for open sets. The typical large deviations statement consists of an upper bound for closed sets and a lower bound for open sets. Here is the best result for i.i.d. random variables, stated in generic large deviations form. For a proof, see [DZ §2.2].

Theorem 1.10. *Let $x_1, x_2, \ldots$ be i.i.d. random variables. Then the function ℓ defined in (1.4) is convex and lower semicontinuous. For any closed set F,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{x_1 + \cdots + x_n}{n} \in F\right) \leq - \inf_{a\in F} \ell(a)$$

and for any open set G,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{x_1 + \cdots + x_n}{n} \in G\right) \geq - \inf_{a\in G} \ell(a).$$

Note that no conditions, not even existence of the mean, are required.

1.3. Examples—I.I.D. Random Variables

In some cases, notably exponential families, the function ℓ of (1.4b) can be calculated explicitly (see, e.g., [MN]). We now present some simple calculations in order to develop a feeling for the scope of the large deviations estimates.

Example 1.11: Normal random variables. Let $x_1, x_2, \ldots$ be standard normal random variables. Then

$$M(\theta) = \frac{1}{\sqrt{2\pi}} \int e^{\theta y} e^{-\frac{1}{2}y^2} dy = e^{\frac{1}{2}\theta^2}$$

by completing the square in the exponent, so that $\ell(a) = \sup_\theta(\theta a - \frac{1}{2}\theta^2) = \frac{1}{2}a^2$. Thus Chernoff's Theorem states that, for any $a > 0$,

$$\mathbb{P}(x_1 + \cdots + x_n \geq na) \approx e^{-n\frac{1}{2}a^2}.$$

In this case, we can also perform a direct calculation: $x_1 + \cdots + x_n$ is a normal random variable distributed as $\sqrt{n}x_1$, so

$$\begin{aligned}\mathbb{P}(x_1 + \cdots + x_n \geq na) &= \mathbb{P}(x_1 \geq \sqrt{n}a) \\ &= \frac{1}{\sqrt{2\pi}} \int_{\sqrt{n}a}^{\infty} e^{-\frac{1}{2}t^2} dt.\end{aligned}$$

Using an estimate of this integral [Mc, p. 5],

$$\frac{1}{y + y^{-1}} e^{-\frac{1}{2}y^2} \leq \int_y^\infty e^{-\frac{1}{2}t^2} dt \leq \frac{1}{y} e^{-\frac{1}{2}y^2},$$

we obtain

$$\mathbb{P}(x_1 + \cdots + x_n \geq na) \approx \frac{1}{\sqrt{2\pi n}a} e^{-n\frac{1}{2}a^2},$$

which is in agreement with the exponential order of the large deviations estimates. The fact that $1/\sqrt{n}$ appears is also generic, as will be seen in the sequel.

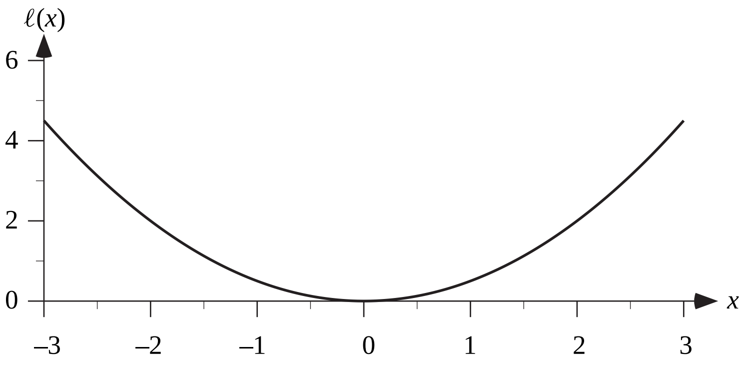

Figure 1.12. The rate function for Standard Normal random variables.

Example 1.13: Poisson random variables. Let x_i be Poisson with mean λ. Then $M(\theta) = e^{\lambda(e^{\theta}-1)}$, and for $a > 0$, $\theta^* = \log(a/\lambda)$. Thus

$$\ell(a) = a\left(\log\left(\frac{a}{\lambda}\right) - 1\right) + \lambda\ ,$$

and $\ell(0) = \lambda$, $\ell(a) = \infty$ for $a < 0$, with $|\theta^*| = \infty$ in the last two cases. Thus Chernoff's Theorem implies, for $a > \lambda$,

$$\mathbb{P}(x_1 + \cdots + x_n \geq na) = \left(\frac{a}{\lambda}\right)^{-na} e^{-n(\lambda-a)+o(n)}.$$

Let us compare this with a direct estimate. Since $x_1 + \cdots + x_n$ is a Poisson random variable with mean $n\lambda$,

$$\begin{aligned}
\mathbb{P}(x_1 + \cdots + x_n \geq na) &= \sum_{j=na}^{\infty} \frac{(n\lambda)^j}{j!} e^{-n\lambda} \approx \frac{(n\lambda)^{na}}{(na)!} e^{-n\lambda} \\
&\approx \frac{(n\lambda)^{na}}{\sqrt{2\pi na}(na)^{na}e^{-na}} e^{-n\lambda} \\
&= \frac{1}{\sqrt{2\pi na}} \left(\frac{a}{\lambda}\right)^{-na} e^{-n(\lambda-a)}
\end{aligned}$$

using Stirling's formula, and the factor $1/\sqrt{n}$ appears again.

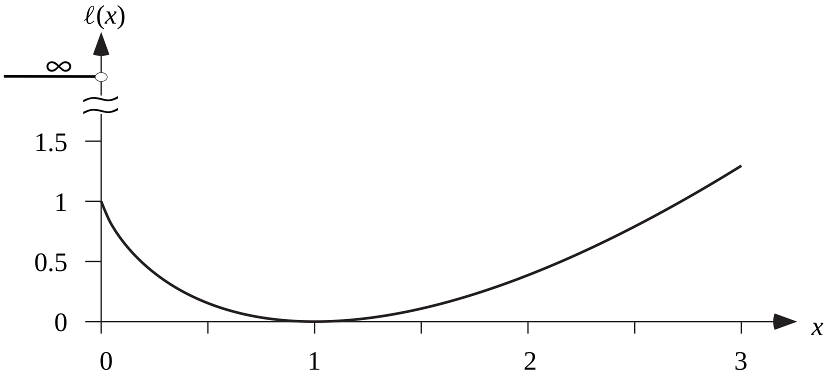

Figure 1.14. The rate function for Standard Poisson random variables.

Example 1.15: Bernoulli random variables. Let x_i take values zero and one with probability $1/2$. Then $M(\theta) = \frac{1}{2}\left(1 + e^{\theta}\right)$. When $\frac{1}{2} < a < 1$, straightforward calculus shows that $\theta^* = \log a - \log(1-a)$, so that in this range

$$\ell(a) = -\log\left(e^{-a\theta^*} M(\theta^*)\right) = a \log a + (1-a)\log(1-a) + \log 2. \quad (1.9)$$

Chernoff's Theorem thus implies that, for $1/2 < a < 1$,

$$\mathbb{P}(x_1 + \cdots + x_n \geq na) = e^{-n\ell(a)+o(n)} = 2^{-n} a^{-na} (1-a)^{-n(1-a)} e^{o(n)}.$$

We can obtain an estimate in a direct way, by approximating the binomial coefficient using Stirling's formula:

$$\begin{aligned}
&\mathbb{P}(x_1 + \cdots + x_n \geq na) \\
&\qquad = \sum_{j=na}^{n} \binom{n}{j} 2^{-n} \approx \frac{n!}{(na)!(n-na)!} 2^{-n} \\
&\qquad \approx 2^{-n}\sqrt{2\pi n}(n)^n e^{-n} \\
&\qquad\quad \times \left(\sqrt{2\pi na}(na)^{na} e^{-na} \sqrt{2\pi n(1-a)}(n(1-a))^{n(1-a)} e^{-n(1-a)}\right)^{-1} \\
&\qquad = \left(\sqrt{2\pi na(1-a)}\right)^{-1} 2^{-n} a^{-na} (1-a)^{-n(1-a)}.
\end{aligned}$$

The formula for M immediately implies that $\ell(a) = \infty$ whenever $a < 0$ or $a > 1$, and $\ell(0) = \ell(1) = \log 2$, with $|\theta^*| = \infty$ in all these cases. Chernoff's Theorem tells us that $\mathbb{P}(x_1 + \cdots + x_n \geq na) \leq e^{-n\infty} = 0$ whenever $a > 1$. For, $a = 1$ the theorem implies $\mathbb{P}(x_1 + \cdots + x_n \geq n) = \left(\frac{1}{2}\right)^n e^{o(n)}$, which is quite close to the exact probability $\left(\frac{1}{2}\right)^n$.

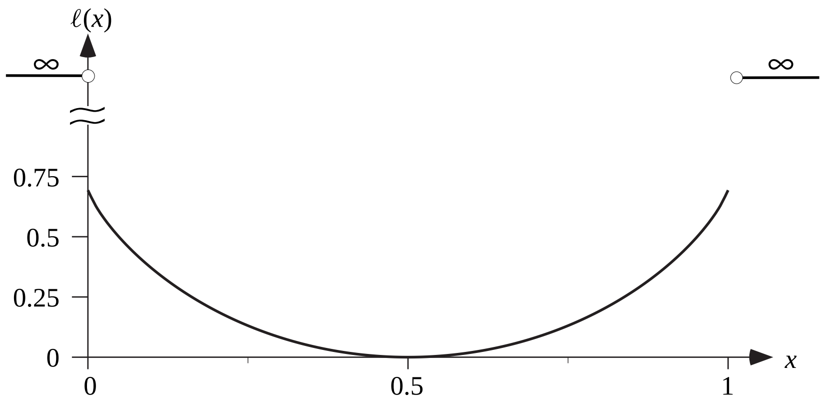

Figure 1.16. The rate function for Bernoulli-1/2 random variables.

Exercise 1.17. For Bernoulli random variables $b_1, b_2, \ldots$ with $\mathbb{P}(b_i = 1) = p$,

$$\ell(a) = a \log \frac{a}{p} + (1-a) \log \frac{1-a}{1-p}.$$ ♠

Example 1.18: Exponential random variables. Let x_i be exponential random variables with mean 1. Then $M(\theta) = 1/(1-\theta)$ for $\theta < 1$ and is infinite otherwise. Therefore $\theta^* = (a-1)/a$ whenever $a > 1$ and then

$$\ell(a) = a - 1 - \log a.$$

Chernoff's Theorem states that for $a > 1$,

$$\mathbb{P}(x_1 + \cdots + x_n \geq na) = a^n e^{-n(a-1)+o(n)}.$$

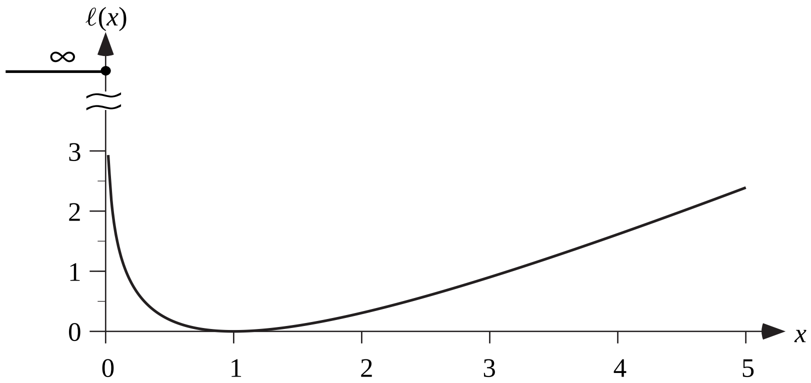

Figure 1.19. The rate function for Exponential random variables, rate one.

Exercise 1.20. For exponential random variables with parameter λ (mean $1/\lambda$),

$$\ell_\lambda(a) = \ell\,(a\lambda)\,.$$

Interpret this as a time change. ♠

From the examples, the following should be expected.

Exercise 1.21. Let c be the constant that is the greatest lower bound for a random variable x; that is, $\mathbb{P}(x < c) = 0$ and $\mathbb{P}(x \le c + \varepsilon) > 0$ for all $\varepsilon > 0$. Then $\ell(a) = \infty$ for $a < c$. Moreover, $\ell(c) < \infty$ if and only if $\mathbb{P}(x = c) > 0$. Hint: take $c = 0$, use dominated convergence, then extend by Exercise 1.6. ♠

1.4. I.I.D. Random Vectors

Large deviations in $\mathbb{R}^d$ are much more complex than in $\mathbb{R}^1$. The main reason for this is that open and closed sets are more complex. Fortunately, these results are not needed in the development of our theory. Sticking to our principle of proving just what we need, let us state a reasonable large deviations result, and provide rough intuition about a way a proof might go. A modern approach to this problem is discussed in §2.1.

Consider the $\mathbb{R}^d$-valued i.i.d. random vectors $\vec{x}_1, \vec{x}_2, \ldots$ with (vector) mean $\vec{0}$. Now let $\vec{\theta} \in \mathbb{R}^d$ and define (see Example A.8 for the notation)

$$M(\vec{\theta}) = \mathbb{E}e^{\langle\vec{\theta},\vec{x}_1\rangle} \tag{1.10a}$$

$$\ell(\vec{a}) = -\log\left(\inf_{\vec{\theta}} e^{-\langle\vec{\theta},\vec{a}\rangle} M(\vec{\theta})\right) = \sup_{\vec{\theta}}\left(\langle\vec{\theta},\vec{a}\rangle - \log M(\vec{\theta})\right). \tag{1.10b}$$

Define J as in Corollary 1.9, but for sets S in $\mathbb{R}^d$. That is,

$$J(S) \stackrel{\triangle}{=} \inf_{\vec{a}\in S} \ell(\vec{a}). \tag{1.11}$$

Theorem 1.22. *Assume $M(\vec{\theta}) < \infty$ for all $\vec{\theta}$. Then, for every closed set C and $\varepsilon > 0$, there exists an integer n_0 such that, whenever $n > n_0$,*

$$\mathbb{P}\left(\frac{\vec{x}_1 + \ldots + \vec{x}_n}{n} \in C\right) \le e^{-n(J(C)-\varepsilon)} \tag{1.12u}$$

and for every open set G and $\varepsilon > 0$, there exists an integer n_0 such that, whenever $n > n_0$,

$$\mathbb{P}\left(\frac{\vec{x}_1 + \ldots + \vec{x}_n}{n} \in G\right) \ge e^{-n(J(G)+\varepsilon)}. \tag{1.12l}$$

The proof of this theorem in $\mathbb{R}^d$ is much more involved than in $\mathbb{R}^1$. Furthermore, the "weakest assumptions" possible in $\mathbb{R}^d$ are much more restrictive than in $\mathbb{R}^1$. See, for example, §D.1, [DZ] and the remarks in §1.5 below. Although modern proofs rely on the technique of §D.1, we outline the extension of the one-dimensional arguments to $\mathbb{R}^d$.

The lower bound in $\mathbb{R}^1$ was based on estimating the probability of the sample mean being in a small interval around the point where ℓ is smallest. A similar argument works in the d-dimensional case: we need to consider small neighborhoods, or balls, around the minimizing point.

Exercise 1.23. Generalize Exercise 1.7 to $\mathbb{R}^d$. ♠

For the upper bound, consider half-spaces of the form $H_{s,a} \triangleq \{\vec{y} : \langle \vec{s}, \vec{y} \rangle \geq a\}$ for some $\vec{s} \in \mathbb{R}^d$ and $a > 0$. Then $\tilde{x}_i \triangleq \langle \vec{s}, \vec{x}_i \rangle$ are i.i.d. and

$$\frac{\vec{x}_1 + \ldots + \vec{x}_n}{n} \in H_{s,a} \iff \frac{\tilde{x}_1 + \ldots + \tilde{x}_n}{n} \geq a.$$

Exercise 1.24. Assume ℓ is continuous and finite. Prove the upper bound for convex sets in $\mathbb{R}^d$. Extend the proof to finite unions of convex sets. Hint: use Chernoff's Theorem for $\tilde{x}_1, \tilde{x}_2, \ldots$. Note that ℓ is convex as explained below (1.4) so that $C_\varepsilon \triangleq \{\vec{x} : \ell(\vec{x}) \leq J(C) - \varepsilon\}$ is convex, and has empty intersection with C. Therefore there is a half-space containing C that does not intersect C_ε. ♠

The following calculation will be used for our Poisson processes. It follows from (1.10)–(1.12).

Exercise 1.25. Suppose y_{ij} are i.i.d. with $y_{ij} \stackrel{\mathcal{L}}{=} Pois(\lambda_j)$. Define

$$\vec{x}_i \triangleq \sum_{j=1}^{J} y_{ij} \vec{e}_j .$$

Show that $\ell(\vec{a})$ defined in (1.10b) has the form

$$\ell(\vec{a}) \triangleq \sup_{\vec{\theta}} \left(\langle \vec{\theta}, \vec{a} \rangle - g(\vec{\theta}) \right), \tag{1.13}$$

where $g(\vec{\theta}) = \log M(\vec{\theta})$ is given by

$$g(\vec{\theta}) = \sum_{j=1}^{J} \lambda_j \left(e^{\langle \vec{\theta}, \vec{e}_j \rangle} - 1 \right). \tag{1.14}$$

Consequently, Exercise 1.24 implies that

$$\lim_{\varepsilon \downarrow 0} \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P} \left(\left| \frac{\vec{x}_1 + \cdots + \vec{x}_n}{n} - \vec{a} \right| < \varepsilon \right) = -\ell(\vec{a}).$$ ♠

1.5. End Notes

While the one-dimensional case is simple enough to make the ideas clear, it can (for the same reason) lull the reader into unwarranted carelessness. We conclude this chapter by identifying some potential pitfalls, and then discussing related works and possible extensions.

One-dimensional caveats

The reader should be aware of several delicate points. Some of these are discussed in Chapter 2 and Appendix D.

1. Properties of the rate function.
   a. Convexity. The calculations in Chapter 1 show that rate functions for random vectors are convex. This (unfortunately) does not extend to rate functions for processes, as will be seen in Chapter 5.
   b. Semicontinuity. The calculations in Chapter 1 show that rate functions are lower semicontinuous. Recall that this means that $\liminf_{x \to y} \ell(x) \geq \ell(y)$, so that ℓ can only jump down. When we formulate, in §2.1, the "axiomatic" large deviations principle, we shall *require* that rate functions be lower semicontinuous. There are several reasons for this restriction. Under this condition there is a convenient, equivalent formulation of the upper bound (see Lemma 2.11), and it guarantees uniqueness of the rate function (§2.1). This condition also makes the upper bound for closed sets particularly easy to prove. In addition, it implies that if a rate function is strictly positive at every point of a compact set, then the probability of that set decays exponentially fast (lower semicontinuous functions attain their minimum on compact sets).
   c. Compact level sets. The examples in Chapter 1 show that ℓ possesses compact level sets, i.e., the sets

$$\{x : \ell(x) \leq \alpha\}$$

   are closed and bounded for each α, if and only if the probability that the random variable actually takes its smallest possible value is zero: Exercise 1.21. In particular, these sets are closed, which implies lower semicontinuity—see Definition A.28 and Exercise A.29. The compactness condition is necessary in order to establish the important *contraction principle* (§2.3). This is illustrated further in §2.3. In Chapter 7 we provide an example of a non-negative birth-death process with constant drift (-1) for $x > 0$, but with the cost (in terms of the rate function) of going from 0 to ∞ being finite. It is easy to show that the process will explode (transition to infinity) if allowed to run for a long enough (finite) time.
2. Difficulties in higher dimensions. In one dimension, the use of Chernoff's Theorem for semi-infinite sets actually provides enough control to estimate the probability that

$$z_n \stackrel{\triangle}{=} (x_1 + \cdots + x_n)/n$$

is in a fairly arbitrary set. In higher dimensions, it is more complicated to estimate this probability, because the topology is more complicated. Thus we need stronger assumptions in $\mathbb{R}^d$ when $d \geq 2$.

3. Difficulties of processes, as opposed to finite dimension. Processes can be viewed, if you are so inclined, as random variables with values in some (infinite dimensional) space of functions. The topology that troubled us in $\mathbb{R}^d$ is simple compared to the topology in function spaces. In this book we restrict our attention to particularly simple processes: jump Markov processes, where the topology is well understood. This topology is discussed in §A.1.

Extensions and relations to other methodologies.

Extensions and generalizations of the results of Chapter 1 are discussed in Chapter 2 and Appendix D. Let us now mention briefly some extensions that will not be touched upon.

The only type of large deviations estimate we obtain in this book is on the order of

$$\mathbb{P}_n(S) \approx e^{-nI^*(S)}.$$

This is only the first term in an asymptotic expansion, though. Using formal methods such as WKB expansions, one finds [Ol] that the series usually continues as follows:

$$\mathbb{P}_n(S) \approx \frac{1}{n^{d/2}} e^{-nI^*(S)+C+O(1/n)} ,$$

where d is the number of dimensions of the process in question. There are a few cases where the full asymptotic expansion has been worked out, and there are many more cases where some *formal* terms have been calculated.

a. Formal expansions of singular equations (e.g., WKB methods). Several investigators, notably Knessl, Matkowski, Morrison, Schuss, and Tier [KMS, Mo1, Mo2] have calculated quite accurate and explicit asymptotic expressions for various large deviations problems using formal expansions. The main criticism of these techniques (there are several that are employed) is that there is no proof of their validity; in contrast, the student will note that in the present book, about 50% of the pages are devoted to proving the validity of the methods we employ. Martin Day provides rigorous proofs for the validity, in some cases, of formulae obtained by formal methods of the WKB type; see e.g., [D2]. Formal methods often give more terms in the asymptotic series than the "rough" methods we employ. They do not usually give sample path information, though, such as we obtain in Chapter 16.

b. Central limit expansions and moderate deviations. The quantities we estimate are generalizations of

$$\mathbb{P}\left(\frac{x_1 + x_2 + \ldots + x_n}{n} \geq a\right)$$

compared to the central limit quantity

$$\mathbb{P}\left(\frac{x_1 + x_2 + \ldots + x_n}{\sqrt{n}} \geq a\right).$$

Clearly there is room to investigate the quantities

$$\mathbb{P}\left(\frac{x_1 + x_2 + \ldots + x_n}{n^{b+1/2}} \geq a\right)$$

for $0 < b < 1/2$. Some of these questions have been approached by Ibragimov and Linnik [IL] and there has been a good deal of activity since then, for both random variables and processes. See, e.g., [DZ, § 3.7].

c. Spectral methods. Many of the applications we analyze may also be examined using spectral methods. For example, the AMS model (Chapter 13) has been investigated by A. Elwalid, D. Mitra, and T. Stern [EMS] among others. Some of the calculations we do are provably equivalent to calculations done on the spectrum; see, e.g., [C1].

d. Calculus of variations methods, optimal control. You can view our approach to probability problems as a method for turning them into problems in the calculus of variations; hence, anything you know about such problems is related to our methods. The type of variational problems arising here also appear in optimal control: of course optimal control and variational problems are themselves inextricably linked, e.g. [Yo, Ce]. In addition, there are several problems in recursive estimation theory, cf. [DK1, DK2], that can be solved via large deviations techniques.

e. Viscosity solutions. Variational problems can often be solved in terms of PDEs (partial differential equations). One of the technical problems that arises is smoothness of solutions. The so-called viscosity solutions turn out to be the correct object (in terms of the degree of smoothness) for many variational problems. Using this concept one can sometimes prove that formal calculations are correct, at least to first order. Barles, Evans, and Souganidis [BES] and Dupuis, Ishii, and Soner [DIS] have used viscosity techniques to prove large deviations principles for certain classes of systems. Reference [DIS] is notable because it proves the principle for the very important case of Jackson networks. In addition, viscosity solutions naturally lead to methods of solving PDEs (and hence variational problems) via successive approximations—a procedure that can facilitate the numerical solution of some large deviations problems.

f. Entropy. Entropy and large deviations are intimately related. We have deliberately avoided this relationship, but others have exploited it to good effect. Ellis [Ell] goes into great detail, proving results on the Ising model among others. All of information theory is based on Chernoff-type estimates; see, for example, Bucklew's book [Bu]. Donsker and Varadhan [DV3] showed how optimal change of measure can be calculated via entropy in a very general Markov process setting. Kullback-Leibler information can be viewed as a large deviations quantity. Again, the reasons we avoid this fruitful subject are lack of time, space, and our choice of applications.

g. Importance sampling. Importance sampling is, in essence, the use of change of measure to improve the accuracy of statistical estimates. It is increasingly important in the simulation of rare events. Our approach to the large deviations lower bound is equivalent to choosing an optimal importance sampling scheme among a class of changed measures. For more details, see [CFM, Bu].

h. Feynman path integrals. The Feynman-Kac formula can be viewed as showing that the exponential martingale we use is indeed a martingale. In other words, our method for proving upper bounds is based on the reasoning behind Feynman path integrals. For a more direct (but, so far as we know, unproven) connection, see Gunther [Gu]. See also Brydges and Maya [BrM].

i. Steepest descent methods. The first large deviations calculations were made by steepest descent methods. It is a natural method, since the transforms (Laplace or Fourier) of sums of independent random variables are simply powers, and steepest descent is then quite accurate for computing the inverse transform.