Applications of Mathematics
Stochastic Modelling and Applied Probability
38
科学前沿丛书

Amir Dembo
Ofer Zeitouni

Large Deviations Techniques and Applications

大偏差技术和应用 第2版

Second Edition

Springer
世界图书出版公司
www.wpcbj.com.cn

Amir Dembo Ofer Zeitouni

Large Deviations Techniques and Applications

Second Edition

With 29 Figures

Springer

Amir Dembo
Department of Mathematics
Stanford University
Stanford, CA 94305, USA

Ofer Zeitouni
Department of Electrical Engineering
Technion–Israel Institute of Technology
Haifa 32000, Israel

Mathematics Subject Classification (1991): 60F10, 60E15, 60G57, 60H10, 60J15, 93E10

Library of Congress Cataloging-in-Publication Data
Dembo, Amir.
Large deviations techniques and applications / Amir Dembo, Ofer Zeitouni. — 2nd ed.
p. cm. — (Applications of mathematics ; 38)
Includes bibliographical references and index.
ISBN 0-387-98406-2 (hardcover : alk. paper)
1. Large deviations. I. Zeitouni, Ofer. II. Title. III. Series.
QA273.67.D46 1998
519.5e 34—dc21 97-45236

9 8 7 6 5 4 3 2

ISBN 0-387-98406-2 SPIN 10832459

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science+Business Media GmbH

Preface to the Second Edition

This edition does not involve any major reorganization of the basic plan of the book; however, there are still a substantial number of changes. The inaccuracies and typos that were pointed out, or detected by us, and that were previously posted on our web site, have been corrected. Here and there, clarifying remarks have been added. Some new exercises have been added, often to reflect a result we consider interesting that did not find its way into the main body of the text. Some exercises have been dropped, either because the new presentation covers them, or because they were too difficult or unclear. The general principles of Chapter 4 have been updated by the addition of Theorem 4.4.13 and Lemmas 4.1.23, 4.1.24, and 4.6.5.

More substantial changes have also been incorporated in the text.

1. A new section on concentration inequalities (Section 2.4) has been added. It overviews techniques, ranging from martingale methods to Talagrand's inequalities, to obtain upper bound on exponentially negligible events.

2. A new section dealing with a metric framework for large deviations (Section 4.7) has been added.

3. A new section explaining the basic ingredients of a weak convergence approach to large deviations (Section 6.6) has been added. This section largely follows the recent text of Dupuis and Ellis, and provides yet another approach to the proof of Sanov's theorem.

4. A new subsection with refinements of the Gibbs conditioning principle (Section 7.3.3) has been added.

5. Section 7.2 dealing with sampling without replacement has been completely rewritten. This is a much stronger version of the results, which

also provides an alternative proof of Mogulskii's theorem. This advance was possible by introducing an appropriate coupling.

The added material preserves the numbering of the first edition. In particular, theorems, lemmas and definitions in the first edition have retained the same numbers, although some exercises may now be labeled differently.

Another change concerns the bibliography: The historical notes have been rewritten with more than 100 entries added to the bibliography, both to rectify some omissions in the first edition and to reflect some advances that have been made since then. As in the first edition, no claim is being made for completeness.

The web site http://www-ee.technion.ac.il/~zeitouni/cor.ps will contain corrections, additions, etc. related to this edition. Readers are strongly encouraged to send us their corrections or suggestions.

We thank Tiefeng Jiang for a preprint of [Jia95], on which Section 4.7 is based. The help of Alex de Acosta, Peter Eichelsbacher, Ioannis Kontoyiannis, Stephen Turner, and Tim Zajic in suggesting improvements to this edition is gratefully acknowledged. We conclude this preface by thanking our editor, John Kimmel, and his staff at Springer for their help in producing this edition.

STANFORD, CALIFORNIA
HAIFA, ISRAEL

AMIR DEMBO
OFER ZEITOUNI
DECEMBER 1997

Preface to the First Edition

In recent years, there has been renewed interest in the (old) topic of large deviations, namely, the asymptotic computation of small probabilities on an exponential scale. (Although the term *large deviations* historically was also used for asymptotic expositions off the CLT regime, we always take large deviations to mean the evaluation of small probabilities on an exponential scale). The reasons for this interest are twofold. On the one hand, starting with Donsker and Varadhan, a general foundation was laid that allowed one to point out several "general" tricks that seem to work in diverse situations. On the other hand, large deviations estimates have proved to be the crucial tool required to handle many questions in statistics, engineering, statistical mechanics, and applied probability.

The field of large deviations is now developed enough to enable one to expose the basic ideas and representative applications in a systematic way. Indeed, such treatises exist; see, e.g., the books of Ellis and Deuschel–Stroock [Ell85, DeuS89b]. However, in view of the diversity of the applications, there is a wide range in the backgrounds of those who need to apply the theory. This book is an attempt to provide a rigorous exposition of the theory, which is geared towards such different audiences. We believe that a field as technical as ours calls for a rigorous presentation. Running the risk of satisfying nobody, we tried to expose large deviations in such a way that the principles are first discussed in a relatively simple, finite dimensional setting, and the abstraction that follows is motivated and based on it and on real applications that make use of the "simple" estimates. This is also the reason for our putting our emphasis on the projective limit approach, which is the natural tool to pass from simple finite dimensional statements to abstract ones.

With the recent explosion in the variety of problems in which large deviations estimates have been used, it is only natural that the collection of applications discussed in this book reflects our taste and interest, as well as applications in which we have been involved. Obviously, it does not represent the most important or the deepest possible ones.

The material in this book can serve as a basis for two types of courses: The first, geared mainly towards the finite dimensional application, could be centered around the material of Chapters 2 and 3 (excluding Section 2.1.3 and the proof of Lemma 2.3.12). A more extensive, semester-long course would cover the first four chapters (possibly excluding Section 4.5.3) and either Chapter 5 or Chapter 6, which are independent of each other. The mathematical sophistication required from the reader runs from a senior undergraduate level in mathematics/statistics/engineering (for Chapters 2 and 3) to advanced graduate level for the latter parts of the book.

Each section ends with exercises. While some of those are routine applications of the material described in the section, most of them provide new insight (in the form of related computations, counterexamples, or refinements of the core material) or new applications, and thus form an integral part of our exposition. Many "hinted" exercises are actually theorems with a sketch of the proof.

Each chapter ends with historical notes and references. While a complete bibliography of the large deviations literature would require a separate volume, we have tried to give due credit to authors whose results are related to our exposition. Although we were in no doubt that our efforts could not be completely successful, we believe that an incomplete historical overview of the field is better than no overview at all. We have not hesitated to ignore references that deal with large deviations problems other than those we deal with, and even for the latter, we provide an indication to the literature rather than an exhaustive list. We apologize in advance to those authors who are not given due credit.

Any reader of this book will recognize immediately the immense impact of the Deuschel–Stroock book [DeuS89b] on our exposition. We are grateful to Dan Stroock for teaching one of us (O.Z.) large deviations, for providing us with an early copy of [DeuS89b], and for his advice. O.Z. is also indebted to Sanjoy Mitter for his hospitality at the Laboratory for Information and Decision Systems at MIT, where this project was initiated. A course based on preliminary drafts of this book was taught at Stanford and at the Technion. The comments of people who attended these courses—in particular, the comments and suggestions of Andrew Nobel, Yuval Peres, and Tim Zajic—contributed much to correct mistakes and omissions. We wish to thank Sam Karlin for motivating us to derive the results of Sections 3.2 and 5.5 by suggesting their application in molecular biology. We thank Tom Cover and Joy Thomas for a preprint of [CT91], which influenced our treatment of Sections 2.1.1 and 3.4. The help of Wlodek Bryc, Marty

Day, Gerald Edgar, Alex Ioffe, Dima Ioffe, Sam Karlin, Eddy Mayer-Wolf, and Adam Shwartz in suggesting improvements, clarifying omissions, and correcting outright mistakes is gratefully acknowledged. We thank Alex de Acosta, Richard Ellis, Richard Olshen, Zeev Schuss and Sandy Zabell for helping us to put things in their correct historical perspective. Finally, we were fortunate to benefit from the superb typing and editing job of Lesley Price, who helped us with the intricacies of LaTeX and the English language.

STANFORD, CALIFORNIA
HAIFA, ISRAEL

AMIR DEMBO
OFER ZEITOUNI
AUGUST 1992

Contents

Chapter 1

Introduction

1.1 Rare Events and Large Deviations

This book is concerned with the study of the probabilities of very rare events. To understand why rare events are important at all, one only has to think of a lottery to be convinced that rare events (such as hitting the jackpot) can have an enormous impact.

If any mathematics is to be involved, it must be quantified what is meant by *rare*. Having done so, a *theory* of rare events should provide an analysis of the rarity of these events. It is the scope of the theory of large deviations to answer both these questions. Unfortunately, as Deuschel and Stroock pointed out in the introduction of [DeuS89b], there is no real "theory" of large deviations. Rather, besides the basic definitions that by now are standard, a variety of tools are available that allow analysis of small probability events. Often, the same answer may be reached by using different paths that seem completely unrelated. It is the goal of this book to explore some of these tools and show their strength in a variety of applications. The approach taken here emphasizes making ***probabilistic*** estimates in a finite dimensional setting and using ***analytical*** considerations whenever necessary to lift up these estimates to the particular situation of interest. In so doing, a particular device, namely, the projective limit approach of Dawson and Gärtner, will play an important role in our presentation. Although the reader is exposed to the beautiful convex analysis ideas that have been the driving power behind the development of the large deviations theory, it is the projective limit approach that often allows sharp results to be obtained in general situations. To emphasize this point, derivations for many of the large deviations theorems using this approach have been provided.

The uninitiated reader must wonder, at this point, what exactly is meant by *large deviations*. Although precise definitions and statements are postponed to the next section, a particular example is discussed here to provide both motivation and some insights as to what this book is about. Let us begin with the most classical topic of probability theory, namely, the behavior of the empirical mean of independent, identically distributed random variables. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent, standard Normal, real-valued random variables, and consider the empirical mean $\hat{S}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Since $\hat{S}_n$ is again a Normal random variable with zero mean and variance $1/n$, it follows that for any $\delta > 0$,

$$P(|\hat{S}_n| \geq \delta) \underset{n\to\infty}{\longrightarrow} 0, \tag{1.1.1}$$

and, for any interval A,

$$P(\sqrt{n}\hat{S}_n \in A) \underset{n\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}} \int_A e^{-x^2/2} dx\,. \tag{1.1.2}$$

Note now that

$$P(|\hat{S}_n| \geq \delta) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\delta\sqrt{n}}^{\delta\sqrt{n}} e^{-x^2/2} dx\,;$$

therefore,

$$\frac{1}{n} \log P(|\hat{S}_n| \geq \delta) \underset{n\to\infty}{\longrightarrow} -\frac{\delta^2}{2}\,. \tag{1.1.3}$$

Equation (1.1.3) is an example of a large deviations statement: The "typical" value of $\hat{S}_n$ is, by (1.1.2), of the order $1/\sqrt{n}$, but with small probability (of the order of $e^{-n\delta^2/2}$), $|\hat{S}_n|$ takes relatively large values.

Since both (1.1.1) and (1.1.2) remain valid as long as $\{X_i\}$ are independent, identically distributed (i.i.d.) random variables of zero mean and unit variance, it could be asked whether (1.1.3) also holds for non-Normal $\{X_i\}$. The answer is that while the limit of $n^{-1} \log P\ (|\hat{S}_n| \geq \delta)$ always exists, its value depends on the distribution of X_i. This is precisely the content of Cramér's theorem derived in Chapter 2.

The preceding analysis is not limited to the case of real-valued random variables. With a somewhat more elaborate proof, a similar result holds for d-dimensional, i.i.d. random vectors. Moreover, the independence assumption can be replaced by appropriate notions of *weak dependence*. For example, $\{X_i\}$ may be a realization of a Markov chain. This is discussed in Chapter 2 and more generally in Chapter 6. However, some restriction on the dependence must be made, for examples abound in which the rate of convergence in the law of large numbers is not exponential.

Once the asymptotic rate of convergence of the probabilities $P\left(|\frac{1}{n}\sum_{i=1}^n X_i| \geq \delta\right)$ is available for every distribution of X_i satisfying certain moment conditions, it may be computed in particular for $P\left(|\frac{1}{n}\sum_{i=1}^n f(X_i)| \geq \delta\right)$, where f is an arbitrary bounded measurable function. Similarly, from the corresponding results in $\mathbb{R}^d$, tight bounds may be obtained on the asymptotic decay rate of

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n f_1(X_i)\right| \geq \delta, \ldots, \left|\frac{1}{n}\sum_{i=1}^n f_d(X_i)\right| \geq \delta\right),$$

where $f_1, \ldots, f_d$ are arbitrary bounded and measurable functions. From here, it is only a relatively small logical step to ask about the rate of convergence of the empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, where δ_{X_i} denotes the (random) measure concentrated at X_i, to the distribution of X_1. This is the content of Sanov's impressive theorem and its several extensions discussed in Chapter 6. It should be noted here that Sanov's theorem provides a quite unexpected link between Large Deviations, Statistical Mechanics, and Information Theory.

Another class of large deviations questions involves the sample path of stochastic processes. Specifically, if $X^\epsilon(t)$ denotes a family of processes that converge, as $\epsilon \to 0$, to some deterministic limit, it may be asked what the rate of this convergence is. This question, treated first by Mogulskii and Schilder in the context, respectively, of a random walk and of the Brownian motion, is explored in Chapter 5, which culminates in the Freidlin–Wentzell theory for the analysis of dynamical systems. This theory has implications to the study of partial differential equations with small parameters.

It is appropriate at this point to return to the *applications* part of the title of this book, in the context of the simple example described before. As a first application, suppose that the mean of the Normal random variables X_i is unknown and, based on the observation $(X_1, X_2, \ldots, X_n)$, one tries to decide whether the mean is -1 or 1. A reasonable, and commonly used decision rule is as follows: Decide that the mean is 1 whenever $\hat{S}_n \geq 0$. The probability of error when using this rule is the probability that, when the mean is -1, the empirical mean is nonnegative. This is exactly the computation encountered in the context of Cramér's theorem. This application is addressed in Chapters 3 and 7 along with its generalization to more than two alternatives and to weakly dependent random variables.

Another important application concerns conditioning on rare events. The best known example of such a conditioning is related to Gibbs conditioning in statistical mechanics, which has found many applications in the seemingly unrelated areas of image processing, computer vision, VLSI design, and nonlinear programming. To illustrate this application, we return to the example where $\{X_i\}$ are i.i.d. standard Normal random variables, and assume that $\hat{S}_n \geq 1$. To find the conditional distribution of X_1 given this rare event, observe that it may be expressed as $P(X_1 | X_1 \geq Y)$, where $Y = n - \sum_{i=2}^{n} X_i$ is independent of X_1 and has a Normal distribution with mean n and variance $(n-1)$. By an asymptotic evaluation of the relevant integrals, it can be deduced that as $n \to \infty$, the conditional distribution converges to a Normal distribution of mean 1 and unit variance. When the marginal distribution of the X_i is not Normal, such a direct computation becomes difficult, and it is reassuring to learn that the limiting behavior of the conditional distribution may be found using large deviations bounds. These results are first obtained in Chapter 3 for X_i taking values in a finite set, whereas the general case is presented in Chapter 7.

A good deal of the preliminary material required to be able to follow the proofs in the book is provided in the Appendix section. These appendices are not intended to replace textbooks on analysis, topology, measure theory, or differential equations. Their inclusion is to allow readers needing a reminder of basic results to find them in this book instead of having to look elsewhere.

1.2 The Large Deviation Principle

The *large deviation principle* (LDP) characterizes the limiting behavior, as $\epsilon \to 0$, of a family of probability measures $\{\mu_\epsilon\}$ on $(\mathcal{X}, \mathcal{B})$ in terms of a *rate function*. This characterization is via asymptotic upper and lower exponential bounds on the values that μ_ϵ assigns to measurable subsets of $\mathcal{X}$. Throughout, $\mathcal{X}$ is a topological space so that open and closed subsets of $\mathcal{X}$ are well-defined, and the simplest situation is when elements of $\mathcal{B}_\mathcal{X}$, the Borel σ-field on $\mathcal{X}$, are of interest. To reduce possible measurability questions, all probability spaces in this book are assumed to have been completed, and, with some abuse of notations, $\mathcal{B}_\mathcal{X}$ always denotes the thus completed Borel σ-field.

Definitions *A rate function I is a lower semicontinuous mapping $I : \mathcal{X} \to [0, \infty]$ (such that for all $\alpha \in [0, \infty)$, the level set $\Psi_I(\alpha) \triangleq \{x : I(x) \leq \alpha\}$ is a closed subset of $\mathcal{X}$). A good rate function is a rate function for which all the level sets $\Psi_I(\alpha)$ are compact subsets of $\mathcal{X}$. The effective domain of I, denoted $\mathcal{D}_I$, is the set of points in $\mathcal{X}$ of finite rate, namely, $\mathcal{D}_I \triangleq \{x : I(x) < \infty\}$. When no confusion occurs, we refer to $\mathcal{D}_I$ as the domain of I.*

Note that if $\mathcal{X}$ is a metric space, the lower semicontinuity property may be checked on sequences, i.e., I is lower semicontinuous if and only if $\liminf_{x_n \to x} I(x_n) \geq I(x)$ for all $x \in \mathcal{X}$. A consequence of a rate function being good is that its infimum is achieved over closed sets.

The following standard notation is used throughout this book. For any set Γ, $\overline{\Gamma}$ denotes the closure of Γ, Γ^o the interior of Γ, and Γ^c the complement of Γ. The infimum of a function over an empty set is interpreted as ∞.

Definition $\{\mu_\epsilon\}$ *satisfies the large deviation principle with a rate function* I *if, for all* $\Gamma \in \mathcal{B}$,

$$-\inf_{x\in\Gamma^o} I(x) \le \liminf_{\epsilon\to 0} \epsilon \log \mu_\epsilon(\Gamma) \le \limsup_{\epsilon\to 0} \epsilon \log \mu_\epsilon(\Gamma) \le -\inf_{x\in\overline{\Gamma}} I(x) . \quad (1.2.4)$$

The right- and left-hand sides of (1.2.4) are referred to as the upper and lower bounds, respectively.

Remark: Note that in (1.2.4), $\mathcal{B}$ need not necessarily be the Borel σ-field. Thus, there can be a separation between the sets on which probability may be assigned and the values of the bounds. In particular, (1.2.4) makes sense even if some open sets are not measurable. Except for this section, we always assume that $\mathcal{B}_\mathcal{X} \subseteq \mathcal{B}$ unless explicitly stated otherwise.

The sentence "μ_ϵ satisfies the LDP" is used as shorthand for "$\{\mu_\epsilon\}$ satisfies the large deviation principle with rate function I." It is obvious that if μ_ϵ satisfies the LDP and $\Gamma \in \mathcal{B}$ is such that

$$\inf_{x\in\Gamma^o} I(x) = \inf_{x\in\overline{\Gamma}} I(x) \stackrel{\triangle}{=} I_\Gamma , \quad (1.2.5)$$

then

$$\lim_{\epsilon\to 0} \epsilon \log \mu_\epsilon(\Gamma) = -I_\Gamma . \quad (1.2.6)$$

A set Γ that satisfies (1.2.5) is called an I *continuity set*. In general, the LDP implies a precise limit in (1.2.6) only for I continuity sets. Finer results may well be derived on a case-by-case basis for specific families of measures $\{\mu_\epsilon\}$ and particular sets. While such results do not fall within our definition of the LDP, a few illustrative examples are included in this book. (See Sections 2.1 and 3.7.)

Some remarks on the definition now seem in order. Note first that in any situation involving non-atomic measures, $\mu_\epsilon(\{x\}) = 0$ for every x in $\mathcal{X}$. Thus, if the lower bound of (1.2.4) was to hold with the infimum over Γ instead of Γ^o, it would have to be concluded that $I(x) \equiv \infty$, contradicting the upper bound of (1.2.4) because $\mu_\epsilon(\mathcal{X}) = 1$ for all ϵ. Thus, some topological restrictions are necessary, and the definition of the LDP codifies a particularly convenient way of stating asymptotic results that, on the one hand, are accurate enough to be useful and, on the other hand, are loose enough to be correct.

Since $\mu_\epsilon(\mathcal{X}) = 1$ for all ϵ, it is necessary that $\inf_{x\in\mathcal{X}} I(x) = 0$ for the upper bound to hold. When I is a good rate function, this means that