Large deviation theory and applications

Peter Mörters

November 10, 2008

Abstract

Large deviation theory deals with the decay of the probability of increasingly unlikely events. It is one of the key techniques of modern probability, a role which is emphasised by the recent award of the Abel prize to S.R.S. Varadhan, one of the pioneers of the subject. The subject is intimately related to combinatorial theory and the calculus of variations. Applications of large deviation theory arise, for example, in statistical mechanics, information theory and insurance.

1 Cramér's theorem and the moderate deviation principle

We start by looking at an example embedded in the most classical results of probability theory. Suppose that X and $X_1, X_2, \ldots$ are independent, identically distributed random variables with mean μ and variance $\sigma^2 < \infty$. We denote the partial sum by

$$S_n := \sum_{i=1}^{n} X_i.$$

The weak law of large numbers states that, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\Big\{ \big|\tfrac{1}{n} S_n - \mu\big| > \varepsilon \Big\} = 0,$$

and this means that the empirical means $\frac{1}{n}S_n$ converge *in probability* to μ.

The simplest way to prove this is by an application of Chebyshev's inequality. Indeed, we argue that

$$\begin{aligned}\mathbb{P}\Big\{\tfrac{1}{n}S_n > \mu+\varepsilon\Big\} &= \mathbb{P}\Big\{\Big(\sum_{i=1}^n X_i\Big) - n\mu > n\varepsilon\Big\}\\ &\le \frac{\mathbb{E}[((\sum_{i=1}^n X_i) - n\mu)^2]}{n^2\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0,\end{aligned}$$

and analogously for the opposite event $\frac{1}{n}S_n < \mu - \varepsilon$.

The more desirable result is however the *strong* law of large numbers, which states that

$$\mathbb{P}\Big\{\lim_{n\to\infty} \tfrac{1}{n}S_n = \mu\Big\} = 1.$$

Excursion: On the difference of weak and strong laws
Here is a typical example of a sequence of random variables converging weakly, but not strongly. Let $N_1, N_2, \ldots$ be an increasing sequence of random variables with values in $\mathbb{N}$ such that

$$\mathbb{P}\big\{n \in \{N_1, N_2, \ldots\}\big\} \to 0,$$

for example by choosing N_j uniformly from the set $\{2^j, \ldots, 2^{j+1}-1\}$. Pick $a_k \uparrow \infty$ very quickly, so that at least $a_k/a_{k-1} > 2$. Now choose $X_n = a_k$ if $N_k \le n < N_{k+1}$, then

$$\frac{X_{n+1}}{X_n} = \begin{cases} \frac{a_k}{a_{k-1}} & \text{if } n = N_k \text{ for some } k,\\ 1 & \text{otherwise.}\end{cases}$$

This sequence converges to one in probability but not almost surely. Moreover, $\limsup X_n/a_{\lfloor \log_2 n\rfloor - 2} \ge 1$ almost surely. □

To prove the *strong* law of large numbers, the previous argument is not good enough. Indeed, recall from the Borel-Cantelli lemma that

$$\sum_{n=1}^{\infty} \mathbb{P}\Big\{\tfrac{1}{n}S_n > \mu+\varepsilon\Big\} < \infty$$

would imply that, almost surely, $\frac{1}{n}S_n > \mu+\varepsilon$ for only finitely many n, and hence

$$\limsup_{n\to\infty} \tfrac{1}{n}S_n \le \mu,$$

and an analogous argument for the opposite event $\frac{1}{n}S_n < \mu - \varepsilon$ would imply the strong law. This line of reasoning fails, as our estimate for the probability of the *large deviation event* $\frac{1}{n}S_n > \mu + \varepsilon$ is of order $\frac{1}{n}$ and therefore not summable.

It is therefore desirable to find out *exactly how fast* the large deviation probabilities

$$\mathbb{P}\Big\{\tfrac{1}{n}S_n > \mu + \varepsilon\Big\}$$

decay. This depends on finer features of the random variable X than merely the finiteness of its variance. Our initial focus is on random variables satisfying

$$\varphi(\lambda) := \log \mathbb{E}e^{\lambda X} < \infty \text{ for all } \lambda \in \mathbb{R}. \tag{1}$$

In this case the large deviation probabilities *decay exponentially* and Cramér's theorem tells us exactly how fast.

Theorem 1.1 (Cramér's theorem). *Let $X_1, X_2, \ldots$ be independent identically distributed random variables with mean μ, satisfying* (1)*, and S_n their partial sums. Then, for any $x > \mu$ we have*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big\{\tfrac{1}{n}S_n \geq x\Big\} = -\varphi^*(x),$$

where φ^* given by

$$\varphi^*(x) := \sup_{\lambda\in\mathbb{R}} \big\{\lambda x - \varphi(\lambda)\big\}$$

is the Legendre transform *of φ.*

Remarks:

- This result holds without assumption (1) and, for $\mu < x <$ess sup X, the large deviation probability decreases exponentially as long as $\varphi(\varepsilon) < \infty$ for some $\varepsilon > 0$.

- In the example of a (possibly unfair) coin toss

$$X = \begin{cases} 0 & \text{with probability } 1-p, \\ 1 & \text{with probability } p, \end{cases}$$

  we obtain $\varphi(\lambda) = \log(pe^\lambda + (1-p))$ and therefore, for $0 < x < 1$,

$$\varphi^*(x) = x \log \frac{x}{p} + (1-x) \log \frac{1-x}{1-p},$$

  the *relative entropy* of $(x, 1-x)$ with respect to $(p, 1-p)$.

The proof allows a first glance at some typical techniques of large deviation theory. We shall give the argument for the lower and upper bound separately.

Proof of the upper bound. We use the Chebyshev inequality again, but in an optimised form. More precisely, for any nonnegative, increasing function ψ we get an upper bound of

$$\mathbb{P}\Big\{\tfrac{1}{n}S_n \geq x\Big\} \leq \mathbb{P}\Big\{\psi\big(S_n\big) \geq \psi(nx)\Big\} \leq \frac{1}{\psi(nx)}\,\mathbb{E}\psi\big(S_n\big).$$

In the proof of the weak law of large numbers, we have chosen $\psi(x) = x^2$, but now we choose $\psi(x) = e^{\lambda x}$ and optimise over $\lambda \geq 0$ later. This yields

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big\{\tfrac{1}{n}S_n \geq x\Big\} \\
&\qquad \leq \limsup_{n\to\infty} \frac{1}{n} \log e^{-\lambda n x} + \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}\Big[\exp(\lambda S_n)\Big] \\
&\qquad = -\lambda x + \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log \mathbb{E}\Big[\exp(\lambda X_i)\Big] \\
&\qquad = -\lambda x + \varphi(\lambda),
\end{aligned}$$

and therefore

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big\{\tfrac{1}{n}S_n \geq x\Big\} \leq -\sup_{\lambda\geq 0} \big\{\lambda x - \varphi(\lambda)\big\}.$$

As φ is a convex function with $\varphi'(0) = \mu < x$ the expression in the curly bracket on the right is negative for $\lambda < 0$, and vanishes for $\lambda = 0$. Hence the supremum may be taken over all $\lambda \in \mathbb{R}$, which completes the proof. □

Proof of the lower bound. We use a *change of measure* or *tilting* argument. The idea is to replace the law P of X by the law

$$dQ(X) = e^{-\varphi(\lambda)+\lambda X}\, dP(X).$$

If, for $\epsilon > 0$, the parameter $\lambda > 0$ can be chosen in such a way that for independent $X_1, \ldots, X_n$ with law

$$d\mathbb{Q} := dQ \otimes \overset{n}{\cdots} \otimes dQ = e^{-n\varphi(\lambda)+\lambda S_n}\, d\mathbb{P}(X_1 \ldots X_n),$$

the partial sums satisfy, for any $\epsilon > 0$,

$$\mathbb{Q}\Big\{x + \epsilon > \tfrac{1}{n} S_n \geq x\Big\} \to 1, \tag{2}$$

we infer that

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big\{\tfrac{1}{n} S_n \geq x\Big\} & \\
&\geq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big\{x + \epsilon > \tfrac{1}{n} S_n \geq x\Big\} \\
&= \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}_{\mathbb{Q}}\Big\{e^{n\varphi(\lambda) - \lambda S_n} \, 1\{x + \epsilon > \tfrac{1}{n} S_n \geq x\}\Big\} \\
&\geq \varphi(\lambda) - \lambda(x + \epsilon) + \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{Q}\big\{x + \epsilon > \tfrac{1}{n} S_n \geq x\big\} \\
&= \varphi(\lambda) - \lambda(x + \epsilon) \geq -\varphi^*(x + \epsilon),
\end{aligned}$$

and the result follows by letting $\epsilon \downarrow 0$.

It remains to show that $\lambda > 0$ can be chosen to satisfy (2). By the weak law of large numbers, it suffices to ensure that the expectation of X under Q equals $x + \frac{\epsilon}{2}$. Note that

$$\varphi'(\lambda) = e^{-\varphi(\lambda)} \mathbb{E}\big[X e^{\lambda X}\big] = \mathbb{E}_Q[X],$$

and hence $\varphi'(0) = \mathbb{E}[X] = \mu$ and $\varphi'(\infty) = \text{ess sup } X =: M$. If $\mu < x < M$, by the intermediate value theorem, we can find for every sufficiently small $\epsilon > 0$, some $\lambda > 0$ with $\varphi'(\lambda) = x + \frac{\epsilon}{2}$, as required.

To complete the argument note that, in the case $M < \infty$, for $x > M$ both sides of the statement in Theorem 1.1 are equal to $-\infty$, and if $x = M$ they are both equal to $\log \mathbb{P}\{X = M\}$. □

We have now seen that the probability of large deviation events of the type

$$\big\{S_n - \mu n \geq nx\big\} \qquad \text{for } x > 0,$$

i.e. the partial sum exceeds its average by more than nx, decay exponentially quickly. The *central limit theorem* tells us by how much the partial sum *normally* exceeds its average, namely by an order of $\sqrt{n}$. More precisely,

$$\mathbb{P}\big\{S_n - \mu n \geq \sqrt{n} x\big\} \to 1 - \Phi(x/\sigma) > 0,$$

where Φ is the distribution function of the standard normal law. This implies that for any sequence a_n with $\sqrt{n} \ll a_n \ll n$ we still have

$$\mathbb{P}\big\{S_n - \mu n \geq a_n\big\} \to 0,$$

and neither the central limit theorem nor Cramér's theorem tell us how fast this convergence is. This question is in the remit of the *moderate deviation principle* stated below.

Theorem 1.2 (Moderate deviation principle). *Under the same assumptions as in Theorem 1.1, if $\sqrt{n} \ll a_n \ll n$ we have, for all $x > 0$,*

$$\lim_{n\to\infty} \frac{n}{a_n^2} \log \mathbb{P}\Big\{S_n - \mu n \geq x a_n\Big\} = -\frac{x^2}{2\sigma^2}.$$

Remark: By contrast to Cramér's theorem, this is the result you would have got by replacing $\frac{S_n - \mu n}{\sigma\sqrt{n}}$ by a standard normal. The result, which can be extended significantly, is probably due to Feller.

Remark: In the next chapter we explain a large deviation framework that includes moderate deviation principles. The typical features of moderate deviation principles are:

- they explain the decay of deviations from the mean on a scale smaller than the mean itself and apply to a whole range of scales,
- they are associated to central limit type theorems and clarify when we can use the normal approximation to calculate tail probabilities,
- the decay rates are universal, i.e. independent of finer features of X.

Proof. To get the gist it suffices to prove the (easier) upper bound. Without loss of generality we may assume that $\mu = 0$ and $x = 1$. Because $\varphi(0) = \varphi'(0) = 0$ and $\varphi''(0) = \sigma^2$, a Taylor expansion of φ around zero gives

$$\varphi(\lambda) \sim \lambda^2 \, \tfrac{\sigma^2}{2} \qquad \text{as } \lambda \downarrow 0.$$

We use the Chebyshev inequality again to get, for any $\lambda > 0$,

$$\mathbb{P}\Big\{S_n \geq a_n\Big\} \leq \mathbb{P}\Big\{e^{S_n \frac{\lambda a_n}{n}} \geq e^{\frac{\lambda a_n^2}{n}}\Big\} \leq e^{-\frac{\lambda a_n^2}{n}} \, e^{n\varphi(\frac{\lambda a_n}{n})}.$$

Hence

$$\limsup_{n\to\infty} \frac{n}{a_n^2} \log \mathbb{P}\Big\{S_n \geq a_n\Big\} \leq -\lambda + \limsup_{n\to\infty} \frac{n^2}{a_n^2} \varphi\big(\tfrac{\lambda a_n}{n}\big) = -\lambda + \lambda^2 \, \tfrac{\sigma^2}{2}.$$

Maximising over $\lambda > 0$, i.e. choosing $\lambda = 1/\sigma^2$, yields the result. □

2 Framework of large deviation principles

Before starting to set up a framework for the theory we are going to develop, we formulate a simple but useful lemma, which is one of the cornerstones of large deviation theory. It states roughly that the rate of growth for a finite sum of sequences equals the maximal rate of growth of the summands.

Lemma 2.1 (Laplace principle). *Fix a sequence $a_n \to \infty$ and a finite number N of nonnegative sequences $(b_n^{(1)}), \dots, (b_n^{(N)})$. Then*

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \sum_{i=1}^{N} b_n^{(i)} = \max_{i=1}^{N} \limsup_{n\to\infty} \frac{1}{a_n} \log b_n^{(i)}.$$

Proof. Note that, for every fixed n,

$$0 \le \log \sum_{i=1}^{N} b_n^{(i)} - \max_{i=1}^{N} \log b_n^{(i)} \le \log N,$$

and dividing by a_n and taking limsup shows that

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \sum_{i=1}^{N} b_n^{(i)} = \limsup_{n\to\infty} \frac{1}{a_n} \max_{i=1}^{N} \log b_n^{(i)} = \max_{i=1}^{N} \limsup_{n\to\infty} \frac{1}{a_n} \log b_n^{(i)},$$

as required to complete the proof. □

From Cramér's theorem we can now infer the following result.

Corollary 2.2. *For every Borel set $A \subset \mathbb{R}$,*

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big\{ \tfrac{1}{n} S_n \in A \Big\} &\le - \inf_{x \in \operatorname{cl} A} \varphi^*(x), \\ \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big\{ \tfrac{1}{n} S_n \in A \Big\} &\ge - \inf_{x \in \operatorname{int} A} \varphi^*(x). \end{aligned} \tag{3}$$

Proof. We may assume that $A \subset [\mu, \infty)$, as otherwise we split $A = A_1 \cup A_2$ with $A_1 \subset [\mu, \infty)$ and $A_2 \subset (-\infty, \mu)$. Analogous arguments hold for the two

parts, and we obtain from the Laplace principle

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big\{\tfrac{1}{n}S_n\in A\Big\}\\
&\qquad=\limsup_{n\to\infty}\frac{1}{n}\log\Big[\mathbb{P}\Big\{\tfrac{1}{n}S_n\in A_1\Big\}+\mathbb{P}\Big\{\tfrac{1}{n}S_n\in A_2\Big\}\Big]\\
&\qquad=\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big\{\tfrac{1}{n}S_n\in A_1\Big\}\vee\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big\{\tfrac{1}{n}S_n\in A_2\Big\}\\
&\qquad\le\big(-\inf_{x\in\mathrm{cl}\,A_1}\varphi^*(x)\big)\vee\big(-\inf_{x\in\mathrm{cl}\,A_2}\varphi^*(x)\big)\\
&\qquad=-\inf_{x\in\mathrm{cl}\,A}\varphi^*(x),
\end{aligned}$$

and, more easily,

$$\begin{aligned}
&\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big\{\tfrac{1}{n}S_n\in A\Big\}\\
&\qquad\ge\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big\{\tfrac{1}{n}S_n\in A_1\Big\}\vee\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big\{\tfrac{1}{n}S_n\in A_2\Big\}\\
&\qquad\ge\big(-\inf_{x\in\mathrm{int}\,A_1}\varphi^*(x)\big)\vee\big(-\inf_{x\in\mathrm{int}\,A_2}\varphi^*(x)\big)\\
&\qquad=-\inf_{x\in\mathrm{int}\,A}\varphi^*(x).
\end{aligned}$$

Now let $a=\inf A\ge\mu$. As φ^* is increasing on $[\mu,\infty)$, the infimum of φ^* over cl A equals $\varphi^*(a)$. Then

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big\{\tfrac{1}{n}S_n\in A\Big\}\le\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big\{\tfrac{1}{n}S_n\ge a\Big\}=-\varphi^*(a).$$

For the lower bound we let $b\in\mathrm{int}\,A$ and $\varepsilon>0$ with $[b-\varepsilon,b+\varepsilon)\subset A$. We assume that $\varphi^*(b-\varepsilon)<\infty$ and $\varphi^*(b-\varepsilon)>0$. This implies that $\varphi^*(b-\varepsilon)<\varphi^*(b+\varepsilon)$ and hence

$$\begin{aligned}
&\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big\{\tfrac{1}{n}S_n\in A\Big\}\\
&\qquad\ge\lim_{n\to\infty}\frac{1}{n}\log\Big[\mathbb{P}\Big\{\tfrac{1}{n}S_n\ge b-\varepsilon\Big\}-\mathbb{P}\Big\{\tfrac{1}{n}S_n\ge b+\varepsilon\Big\}\Big]\\
&\qquad=-\varphi^*(b-\varepsilon).
\end{aligned}$$

Letting $b\downarrow\inf\mathrm{int}A$ and $\varepsilon\downarrow 0$ gives the result. □

We note from this that, for a general large deviation theory we may consider a sequence of random variable $X_1, X_2, \ldots$ in a general metric space M and consider events of the type $\{X_n \in A\}$ where $A \subset M$ is a Borel set. We start to set things up by looking at the functions that will in general replace φ^*.

Definition 2.3. *Fix a metric space M. A function $I\colon M \to [0,\infty]$ is called*

- *a* rate function *if it is lower semicontinuous, which means that the level sets $\{x \in M\colon I(x) \le a\}$ are closed for any $a \ge 0$;*
- *a* good rate function *if the level sets are compact for any $a \ge 0$.*

Now we can define the notion of a large deviation principle.

Definition 2.4. *A sequence of random variables $X_1, X_2, \ldots$ with values in a metric space is said to satisfy a* large deviation principle *with*

- *speed $a_n \to \infty$ and*
- *rate function I,*

if, for all Borel sets $A \subset M$,

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\Big\{X_n \in A\Big\} &\le -\inf_{x\in \operatorname{cl} A} I(x), \\ \liminf_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\Big\{X_n \in A\Big\} &\ge -\inf_{x\in \operatorname{int} A} I(x). \end{aligned} \tag{4}$$

Remark: Note that we *do not* require convexity of I. This is a special feature of the rate function φ^* in Cramér's theorem.
Remark: We do not mention the speed if it is clear from the context. The rate function of a large deviation principle is uniquely determined, due to its lower semicontinuity.

Note that for $S_n = \sum_{i=1}^n X_i$ as above

- $\frac{1}{n} S_n$ satisfies a large deviation principle with speed n and good rate function φ^*, by Cramér's theorem;

- for any sequence $\sqrt{n} \ll a_n \ll n$ the random variables

$$\frac{S_n - \mu n}{a_n}$$

satisfy a large deviation principle with speed $\frac{a_n^2}{n}$ and good rate function $I(x) = \frac{x^2}{2\sigma^2}$, by the moderate deviation principle.

Before returning to interesting examples we briefly discuss two important theoretical issues related to the general set-up. These are responses to the following questions:

(1) Can we get away with proving (4) not for all Borel sets A, but for a somewhat smaller family? Obviously it suffices to look at open sets for the lower bound and closed sets for the upper bound. But maybe it even suffices to check the lower bound for open balls, or the upper bound for compact sets?

(2) If we have a large deviation principle for X_n and $f\colon M \to M'$ is a continuous function, can we get a large deviation principle for $f(X_n)$?

We start with answers for the first question. The following simple lemma helps in the proof of the lower bounds.

Lemma 2.5. *If I is a rate function and A is a Borel set, such that for every $x \in A$ and $\varepsilon > 0$ with $B(x,\varepsilon) \subset A$,*

$$\liminf_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\big\{X_n \in B(x,\varepsilon)\big\} \geq -I(x),$$

then

$$\liminf_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\big\{X_n \in A\big\} \geq -\inf_{x\in \operatorname{int} A} I(x).$$

Proof. Choose $x_k \in \operatorname{int} A$ with $I(x_k) \to \inf_{x\in\operatorname{int} A} I(x)$ and for each x_k an $\varepsilon_k > 0$ such that $B(x_k,\varepsilon_k) \subset A$. Then

$$\liminf_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\big\{X_n \in A\big\} \geq \liminf_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\big\{X_n \in B(x_k,\varepsilon_k)\big\} \geq -I(x_k),$$

and the result follows as $k \uparrow \infty$. □

Replacing closed sets by compact sets in the upper bound requires a substantial condition, which is often hard to check.

Definition 2.6. *The sequence X_n of random variables is called* exponentially tight *if for every $N < \infty$ there exists a compact set $K \subset M$ such that*

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\big\{X_n \notin K\big\} < -N.$$

Lemma 2.7. *If the family X_n is exponentially tight and satisfies the large deviation upper bound upper bound for every compact set, then it holds for every closed set.*

Proof. Fix N and a compact set $K \subset M$ as in the definition of exponential tightness. If $A \subset M$ is closed, then $A \cap K$ is compact and therefore

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\Big\{X_n \in A\Big\} \\
&\qquad \le \limsup_{n\to\infty} \frac{1}{a_n} \log \Big[\mathbb{P}\Big\{X_n \in A \cap K\Big\} + \mathbb{P}\Big\{X_n \notin K\Big\}\Big] \\
&\qquad \le \big(-\inf_{x\in A\cap K} I(x)\big) \vee -N \\
&\qquad \le \big(-\inf_{x\in A} I(x)\big) \vee -N,
\end{aligned}$$

and the result follows by letting $N \to \infty$. □

Remark: Suppose the sequence X_n satisfies a large deviation principle with speed a_n and rate function I. If the sequence of random variables is exponentially tight, then I is a good rate function. Conversely, if M is separable and I is a good rate function, then the sequence is exponentially tight. See 1.2.18(b) and 4.1.10 in Dembo-Zeitouni.

Coming to the second question, the answer is almost as clean as one could wish for.

Lemma 2.8 (Contraction Principle). *If $X_1, X_2, \ldots$ satisfies a large deviation principle with speed a_n and good rate function I, and $f\colon M \to M'$ is a continuous mapping, then the sequence $f(X_1), f(X_2), \ldots$ satisfies a large deviation principle with speed a_n and good rate function J given by*

$$J(y) = \inf_{x\in f^{-1}(\{y\})} I(x).$$

Proof. Let $A \subset M'$ be a closed set. Then

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\Big\{ f(X_n) \in A \Big\} \\
&\qquad = \limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\Big\{ X_n \in f^{-1}(A) \Big\} \\
&\qquad \leq - \inf_{x \in f^{-1}(A)} I(x) = - \inf_{y \in A} \inf_{x \in f^{-1}(\{y\})} I(x) \\
&\qquad = - \inf_{y \in A} J(y),
\end{aligned}$$

using that $f^{-1}(A)$ is closed in M. An analogous argument applies to open sets. Note that we have not used the goodness of the rate function so far. This comes in when we check that J is indeed a rate function, i.e. is lower semicontinuous. The level sets of J are

$$\{y \in M \colon J(y) \leq a\} = \{f(x) \in M \colon I(x) \leq a\} = f\big(\{x \in M \colon I(x) \leq a\}\big).$$

Because the level sets of I are compact and f is continuous, the level sets of J are compact as well, so J is a good rate function. Observe that if I fails to be good, we could not ensure that J is a rate function. □

3 Sanov's theorem and the method of types

In this section we illustrate how combinatorial, or counting, arguments, can help providing large deviation principles. These techniques are often set up for a 'finite' context and then generalized using analytical machinery. Our interest at this point is in the former, so we now assume that $X_1, X_2, \dots$ are i.i.d. random variables taking values in a finite set $\mathcal{X}$. We are interested in the frequency of a symbol $x \in \mathcal{X}$ among the first n samples. While Cramér's theorem yields the precise rate of decay of probabilities of events of the form

$$\Big\{\frac{1}{n}\sum_{i=1}^{n} 1\{X_i = x\} \geq a\Big\}, \text{ for } 0 < a < 1, x \in \mathcal{X},$$

it does not help when we are interested in the frequency of more than one symbol, like

$$\Big\{\frac{1}{n}\sum_{i=1}^{n} 1\{X_i = x\} \geq a, \sum_{i=1}^{n} 1\{X_i = y\} \geq b\Big\}, \text{ for } a, b > 0, a + b < 1, x, y \in \mathcal{X}.$$

This type of problems is addressed in *Sanov's theorem*. It requires looking at the *empirical measure* L_n^X of a sample vector $X = (X_1, \dots, X_n)$ defined by

$$L_n^X(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i = x\}$$

and interpreted as a random element of the space $\mathcal{M}_1(\mathcal{X})$ of probability measures on $\mathcal{X}$ endowed with the metric inherited from the embedding into $\mathbb{R}^{|\mathcal{X}|}$ given by the mapping $\mu \mapsto (\mu(x) \colon x \in \mathcal{X})$.

Theorem 3.1 (Sanov's theorem). *Assume that $X_1, X_2, \dots$ are i.i.d. random variables taking values in a finite set $\mathcal{X}$ and denote by $\mu \in \mathcal{M}_1(\mathcal{X})$ their distribution. Then the empirical measures L_n^X satisfy a large deviation principle on the metric space $\mathcal{M}_1(\mathcal{X})$ with speed n and good rate function J given by the relative entropies*

$$J(\nu) = H(\nu\|\mu) := \sum_{x\in\mathcal{X}} \nu(x) \log \frac{\nu(x)}{\mu(x)}.$$

Remarks:
(a) The relative entropy $H(\nu\|\mu)$ is always nonnegative and zero only if $\nu = \mu$. This follows easily from Jensen's inequality for the strictly convex function $\phi(x) = x \log x$ on $[0, \infty)$,

$$\begin{aligned}\sum_{x\in\mathcal{X}} \nu(x) \log \frac{\nu(x)}{\mu(x)} &= \sum_{x\in\mathcal{X}} \mu(x) \phi\Big(\frac{\nu(x)}{\mu(x)}\Big) \\ &\geq \phi\Big(\sum_{x\in\mathcal{X}} \mu(x) \frac{\nu(x)}{\mu(x)}\Big) = \phi(1) = 0.\end{aligned}$$

(b) If our random variables take (finitely many) values in the real numbers we can derive Cramér's theorem from Sanov's theorem by contraction. Indeed, let

$$f\colon \mathcal{M}_1(\mathcal{X}) \to \mathbb{R}, \quad f(\nu) = \sum_{x\in\mathcal{X}} x\nu(x),$$

so that $f(L_n^X) = \frac{1}{n} S_n$. As f is obviously continuous, we obtain a large deviation principle for $\frac{1}{n} S_n$ with rate function

$$\begin{aligned}I(y) &= \inf_{\nu\in f^{-1}(\{y\})} J(\nu) \\ &= \inf\Big\{ \sum_{x\in\mathcal{X}} \nu(x) \log \frac{\nu(x)}{\mu(x)} \colon \nu \in \mathcal{M}_1(\mathcal{X}) \text{ with } \sum_{x\in\mathcal{X}} x\nu(x) = y \Big\} \\ &= \sup_{\lambda\in\mathbb{R}} \Big\{ \lambda y - \log\Big(\sum_{x\in\mathcal{X}} e^{\lambda x} \mu(x) \Big) \Big\}.\end{aligned}$$

The last identity is a typical variational identity arising in large deviation theory. Let us give a direct proof. The $\geq$ direction is based on Jensen's inequality and mimics the tilting argument of the proof of Cramér's theorem. Let ν be any permissible measure, λ arbitrary and abbreviate $Z = \sum_{x\in\mathcal{X}} e^{\lambda x} \mu(x)$. Then

$$\begin{aligned}\sum_{x\in\mathcal{X}} \nu(x) \log \frac{\nu(x)}{\mu(x)} &= \sum_{x\in\mathcal{X}} \nu(x) \log \frac{\nu(x)}{e^{\lambda x}\mu(x)} + \lambda y \\ &= \sum_{x\in\mathcal{X}} \phi\Big(\frac{\nu(x)}{\frac{1}{Z} e^{\lambda x}\mu(x)}\Big) \frac{1}{Z} e^{\lambda x} \mu(x) - \log Z + \lambda y \\ &\geq -\log Z + \lambda y.\end{aligned}$$

As in the proof of Cramér's theorem we can use the intermediate value theorem to find, for $\min \mathcal{X} < y < \max \mathcal{X}$ a $\lambda \in \mathbb{R}$ with

$$y = \frac{\sum x e^{\lambda x} \mu(x)}{\sum e^{\lambda x} \mu(x)},$$

and hence the choice $\nu(x) = \frac{1}{Z}\, \mu(x) e^{\lambda x}$ yields a permissible measure. Then

$$\sum_{x \in \mathcal{X}} \nu(x) \log \frac{\nu(x)}{\mu(x)} = \sum_{x \in \mathcal{X}} \nu(x) \log \big(\frac{1}{Z}\, e^{\lambda x} \big) = -\log Z + \lambda y,$$

proving the $\leq$ direction.

For the proof of Sanov's theorem, we use relatively simple combinatorics. We note that the probability of the events $\{(X_1, \ldots, X_n) = x\}$ for $x \in \mathcal{X}^n$ depends only on the *type* of x, which is the associated empirical measure L_n^x given by

$$L_n^x(y) = \frac{1}{n} \sum_{i=1}^{n} 1\{x_i = y\}, \qquad \text{for } y \in \mathcal{X}.$$

Denote by $T_n(\nu)$ the set of all vectors $x \in \mathcal{X}^n$ of type ν and define the *entropy* of a probability measure ν as

$$H(\nu) = -\sum_{x \in \mathcal{X}} \nu(x) \log \nu(x).$$

Lemma 3.2. *If $x \in T_n(\nu)$, then*

$$\mathbb{P}\big\{(X_1, \ldots, X_n) = x\big\} = e^{-n(H(\nu) + H(\nu \| \mu))}.$$

Proof. Note that

$$H(\nu) + H(\nu \| \mu) = -\sum_{x \in \mathcal{X}} \nu(x) \log \mu(x).$$

Then, using independence, for

$$\begin{aligned} \mathbb{P}\big\{(X_1, \ldots, X_n) = x\big\} &= \prod_{i=1}^{n} \mu(x_i) = \prod_{x \in \mathcal{X}} \mu(x)^{\nu(x)} \\ &= \exp\Big(\sum_{x \in \mathcal{X}} \nu(x)\, \log \mu(x) \Big), \end{aligned}$$

which completes the proof. □

Next, we count how many types and how many vectors of a given type there are. We denote by $\mathcal{L}_n$ the set of types of all possible vectors $x \in \mathcal{X}^n$.

Lemma 3.3.

(a) $\#\mathcal{L}_n \leq (n+1)^{\#\mathcal{X}}$.

(b) There exist polynomials p_1, p_2 with positive coefficients, such that for every $\nu \in \mathcal{L}_n$,

$$\frac{1}{p_1(n)}\, e^{nH(\nu)} \leq \#T_n(\nu) \leq p_2(n)\, e^{nH(\nu)}.$$

Proof. (*a*) For any $y \in \mathcal{X}$, the number $L_n^x(y)$ belongs to the set $\{\frac{0}{n}, \frac{1}{n}, \ldots, \frac{n}{n}\}$, whose cardinality is $n+1$. Hence the number of possible measures L_n^x is at most $(n+1)^{\#\mathcal{X}}$.

(*b*) $T_n(\nu)$ is in bijection to the number of ways one can arrange the objects from a collection containing object $x \in \mathcal{X}$ exactly $n\nu(x)$ times. Hence $\#T_n(\nu)$ is the multinomial

$$\#T_n(\nu) = \frac{n!}{\prod_{x\in\mathcal{X}}(n\nu(x))!}.$$

To find the exponential growth rate of the right hand side we use *Stirling's formula*, which states that (see, e.g., Feller I) for suitable $c_1, c_2 > 0$ and all positive integers $n \in \mathbb{N}$,

$$n \log \frac{n}{e} \leq \log n! \leq n \log \frac{n}{e} + c_1 \log n + c_2.$$

Now

$$\begin{aligned}
\log \#T_n(\nu) &\leq \log n! - \sum_{x\in\mathcal{X}} \log(n\nu(x))! \\
&\leq n \log \frac{n}{e} - \sum_{x\in\mathcal{X}} n\nu(x) \log \frac{n\nu(x)}{e} + c_1 \log n + c_2 \\
&= nH(\nu) + c_1 \log n + c_2,
\end{aligned}$$

which yields the upper bound with $p_1(n) = c_2 n^{c_1}$. The proof of the lower bound is analogous. □

Proof of Sanov's theorem. Take any Borel set $A \subset \mathcal{M}_1(\mathcal{X})$. Then, using the lemmas,

$$\begin{aligned}\mathbb{P}\big\{L_n^X \in A\big\} &= \sum_{\nu \in \mathcal{L}_n \cap A} \mathbb{P}\big\{L_n^X = \nu\big\} \\ &= \sum_{\nu \in \mathcal{L}_n \cap A} \sum_{x \in T_n(\nu)} \mathbb{P}\big\{X = x\big\} \\ &\leq \sum_{\nu \in \mathcal{L}_n \cap A} p_2(n)\, e^{nH(\nu)}\, e^{-n(H(\nu)+H(\nu\|\mu))} \\ &\leq p_2(n)\, \#\big(\mathcal{L}_n \cap A\big)\, e^{-n \inf_{\nu \in A} H(\nu\|\mu)}.\end{aligned}$$

Hence, as $\frac{1}{n} \log p_2(n) \to 0$ and $\frac{1}{n} \log \#\mathcal{L}_n \to 0$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big\{L_n^X \in A\big\} \leq - \inf_{\nu \in A} H(\nu\|\mu).$$

In particular, in this case there is no need for taking the closure of A in the upper bound.

For the lower bound we may assume $\mu(x) > 0$ for all $x \in \mathcal{X}$. Take an open set $A \subset \mathcal{M}_1(\mathcal{X})$ and assume, without loss of generality, $\inf_{\nu \in A} H(\nu\|\mu) < \infty$. Fix $\nu \in A$ such that $H(\nu\|\mu)$ differs from the infimum by no more than some fixed $\epsilon > 0$. For any sufficiently large n we can moreover find $\nu_n \in A \cap \mathcal{L}_n$ such that $H(\nu_n\|\mu)$ differs from $H(\nu\|\mu)$ by no more than ϵ. Hence,

$$\mathbb{P}\big\{L_n^X \in A\big\} \geq \mathbb{P}\big\{L_n^X = \nu_n\big\} = \sum_{x \in T_n(\nu_n)} \mathbb{P}\big\{X = x\big\} \geq \frac{1}{p_1(n)}\, e^{-nH(\nu_n\|\mu)}.$$

Hence, as $\frac{1}{n} \log 1/p_1(n) \to 0$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big\{L_n^X \in A\big\} \geq -H(\nu_n\|\mu) \geq - \inf_{\nu \in A} H(\nu\|\mu) - 2\epsilon.$$

The result follows as $\epsilon > 0$ was arbitrary. □

4 Large deviations for empirical pair measures

We now study a model example for the large deviation technique, hopefully raising all the major issues coming up in typical proofs.

We still look at i.i.d. random variables $X_1, \dots, X_n$ with values in a finite state space $\mathcal{X}$, and denote by $\mu \in \mathcal{M}_1(\mathcal{X})$ their distribution. Instead of the empirical measure, our interest is now focused on the *empirical pair measure*

$$M_n^X = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_{i-1}, X_i)} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{X})$$

where we put $X_0 := X_n$ for symmetry reasons. Note that this measure, other than L_n^X, reflects the linear indexing of our random variables. Interesting variations arise when the random variables are indexed by other graph structures.

Theorem 4.1. *The empirical pair measures $M_n^X \in \mathcal{M}_1(\mathcal{X} \times \mathcal{X})$ satisfy a large deviation principle with speed n and good rate function*

$$J(\nu) = \begin{cases} H(\nu \,\|\, \nu_1 \otimes \mu) & \text{if } \nu_1 = \nu_2, \\ \infty, & \text{otherwise,} \end{cases}$$

where $\nu_1, \nu_2 \in \mathcal{M}_1(\mathcal{X})$ are the two marginals of $\nu \in \mathcal{M}_1(\mathcal{X} \times \mathcal{X})$.

Remark: Sanov's theorem can be recovered from this result. Indeed, the contraction theorem applied to the mapping projecting ν onto its first marginal gives a large deviation principle for L_n^X with rate function

$$I(\omega) = \inf \big\{ H(\nu \,\|\, \omega \otimes \mu) \colon \nu_1 = \nu_2 = \omega \big\}.$$

Indeed, for any permissible choice of ν,

$$\begin{aligned} H(\nu \,\|\, \omega \otimes \mu) &= \sum_{x,y \in \mathcal{X}} \nu(x,y) \, \log \frac{\nu(x,y)}{\omega(x)\omega(y)} + \sum_{y \in \mathcal{X}} \omega(y) \log \frac{\omega(y)}{\mu(y)} \\ &\geq \sum_{y \in \mathcal{X}} \omega(y) \log \frac{\omega(y)}{\mu(y)} = H(\omega \,\|\, \mu), \end{aligned}$$

and the choice $\nu = \omega \otimes \omega$ gives equality.

How do we prove such a result?

I suggest to follow *two golden rules*:

(1) Make sure that the quantity you study is a type, i.e. we have (at least on an exponential scale) equal probability for all elementary events with the same empirical pair measure M_n^X. If this fails take a more informative quantity and reduce later, using the projection theorem.

(2) Make sure that the process class you study is big enough to contain a process for which the required large deviation behaviour is typical. If necessary look at a more general problem class.

In our case the *first* golden rule is satisfied, as we have observed in Lemma 3.2. But the *second* golden rule fails, as the typical empirical pair measure for any i.i.d sequence is of product structure $\omega \otimes \omega$. Hence we need to generalize the problem and ensure that we look at sequences $X_1, \dots, X_n$ no longer i.i.d. but with some control on the expected empirical pair measures. It is therefore easier to prove the result for *Markov chains*, which include i.i.d. sequences as a special case.

Let $X_1, X_2 \dots$ be a Markov chain with statespace $\mathcal{X}$, strictly positive transition matrix P and stationary distribution π. Assume that the initial distribution is π, so that the chain is stationary.

Theorem 4.2. *For a Markov chain as above, the empirical pair measures $M_n^X \in \mathcal{M}_1(\mathcal{X} \times \mathcal{X})$ satisfy a large deviation principle with speed n and good rate function*

$$J(\nu) = \begin{cases} H(\nu \,\|\, \nu_1 \otimes P) & \text{if } \nu_1 = \nu_2, \\ \infty, & \text{otherwise,} \end{cases}$$

where $\omega \otimes P$ for $\omega \in \mathcal{M}_1(\mathcal{X})$ is given as $\omega \otimes P(x,y) = \omega(x)P_{x,y}$.

Remark: Obviously, this is stronger than Theorem 4.1.

It can be checked that the first golden rule is still satisfied in this more general setup. This will be implicit in our proof. We give a proof of the upper bound based on Chebyshev's inequality, which is very versatile.

For any function $g\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ we define an auxiliary chain with transition matrix P^g given by

$$P^g_{x,y} = e^{g(x,y)-U_g(x)}\,P_{x,y}, \qquad \text{for } x,y\in\mathcal{X},$$

where

$$U_g(x) = \log\sum_{y\in\mathcal{X}} e^{g(x,y)}P_{x,y}.$$

Lemma 4.3. *For any closed set $A\subset\mathcal{M}_1(\mathcal{X}\times\mathcal{X})$ we have*

$$\limsup_{n\to\infty}\frac{1}{n}\,\log\mathbb{P}\big\{M_n^X\in A\big\} \\ \le -\inf_{\nu\in A}\sup_g\sum_{x,y\in\mathcal{X}}\big(g(x,y)-U_g(x)\big)\,\nu(x,y).$$

Proof. Note that (using the convention $x_0:=x_n$)

$$\begin{aligned}1 &\ge \sum_{x\in\mathcal{X}^n}\pi^g(x_1)\prod_{i=2}^n P^g_{x_{i-1},x_i}\\ &\ge \Big(\inf_{x,y\in\mathcal{X}}\tfrac{\pi^g(x)P_{x,y}}{\pi(x)P^g_{x,y}}\Big)\sum_{x\in\mathcal{X}^n}e^{\sum_{i=1}^n g(x_{i-1},x_i)-U_g(x_{i-1})}\,\pi(x_1)\prod_{i=2}^n P_{x_{i-1},x_i}\\ &= \Big(\inf_{x,y\in\mathcal{X}}\tfrac{\pi^g(x)P_{x,y}}{\pi(x)P^g_{x,y}}\Big)\,\mathbb{E}[e^{n\int(g-U_g)\,dM_n^X}].\end{aligned}$$

Given any $\nu\in A$ and $\epsilon>0$ we may fix g such that

$$\sum_{x,y\in\mathcal{X}}\big(g(x,y)-U_g(x)\big)\,\nu(x,y)\ge\sup_h\sum_{x,y\in\mathcal{X}}\big(h(x,y)-U_h(x)\big)\,\nu(x,y)-\epsilon.$$

Using continuity we also find $\delta>0$ such that, for all $\nu'\in B(\nu,\delta)$

$$\sum_{x,y\in\mathcal{X}}\big(g(x,y)-U_g(x)\big)\,\nu'(x,y)\ge\sum_{x,y\in\mathcal{X}}\big(g(x,y)-U_g(x)\big)\,\nu(x,y)-\epsilon.$$

Now, by Chebyshev's inequality,

$$\mathbb{P}\big\{M_n^X\in B(\nu,\delta)\big\}\le\mathbb{E}\big[e^{n\int(g-U_g)\,dM_n^X}\big]e^{-n[\int(g-U_g)\,d\nu-\epsilon]}.$$

Therefore,

$$\limsup_{n\to\infty}\frac{1}{n}\,\log\mathbb{P}\big\{M_n^X\in B(\nu,\delta)\big\}\le-\inf_{\nu'\in A}\sup_g\int(g-U_g)\,d\nu'-2\epsilon.$$

Covering the compact set A with finitely many balls of this type and using the Laplace principle we infer the result. □

Proof of the upper bound. To finish the proof of the upper bound, we fix ν with $\nu_1 = \nu_2$ and identify the variational problem in g as a relative entropy.

$$\begin{aligned}\sum_{x,y\in\mathcal{X}} \big(g(x,y) - U_g(x)\big)\,\nu(x,y) &= \sum_{x,y\in\mathcal{X}} \nu(x,y)\log\tfrac{P^g_{x,y}}{P_{x,y}}\\ &= H(\nu \parallel \nu_1\otimes P) - H(\nu\parallel \nu_1\otimes P^g)\\ &\le H(\nu\parallel \nu_1\otimes P),\end{aligned}$$

and equality holds if $\nu = \nu_1\otimes P^g$. This corresponds to the choice

$$g(x,y) = \log\frac{\nu(x,y)}{\nu_1(x)P_{x,y}}, \tag{5}$$

which makes $U_g = 0$. □

For the lower bound we *use the upper bound* to show that exactly this choice of g makes the event $M_n^X \approx \nu$ typical. We write $\mathbb{P}^g$, $\mathbb{E}^g$ for probabilities and expectations with respect to the stationary chain with transition matrix P^g.

Lemma 4.4. *Given $\nu\in\mathcal{M}_1(\mathcal{X}\times\mathcal{X})$ with $\nu_1=\nu_2$ and g as in (5), we have, for all $\epsilon>0$,*

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}^g\big\{M_n^X\notin B(\nu,\epsilon)\big\} < 0.$$

Proof. Applying the upper bound to the Markov chain with transition matrix P^g, we obtain

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}^g\big\{M_n^X\notin B(\nu,\epsilon)\big\} \le -\inf_{\nu'\notin B(\nu,\epsilon)} H(\nu'\,\|\,\nu'_1\otimes P^g).$$

By compactness and continuity, the right hand side is strictly negative if $\nu'\neq\nu'_1\otimes P^g$ for all $\nu'\notin B(\nu,\epsilon)$. This is the case, because $\nu'=\nu'_1\otimes P^g$ implies

$$\nu'(x,y) = \nu'_1(x)\frac{\nu(x,y)}{\nu_1(x)P_{x,y}}\,P_{x,y} = \nu'_1(x)\,\frac{\nu(x,y)}{\nu_1(x)}. \tag{6}$$

Summing over $x\in\mathcal{X}$ and using the equality of the marginals of ν' we infer that

$$\nu'_1(y) = \nu'_2(y) = \sum_{x\in\mathcal{X}}\nu'_1(x)\,\frac{\nu(x,y)}{\nu_1(x)}.$$

Hence ν_1' is the (unique) invariant distribution for the Markov chain with transition matrix $P'_{x,y} = \nu(x,y)/\nu_1(x)$. Obviously, ν_1 is also an invariant distribution for this Markov chain, so $\nu_1 = \nu_1'$. From (6) we can therefore infer that $\nu = \nu'$, as required. □

Proof of the lower bound. Having established a *law of large numbers* from the upper bound, we can complete the proof of the lower bound by change of measure. By Lemma 2.5 it suffices to show that, for every $\nu \in \mathcal{M}_1(\mathcal{X} \times \mathcal{X})$ with $\nu_1 = \nu_2$, and $\varepsilon > 0$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big\{ M_n^X \in B(\nu,\varepsilon) \big\} \geq -H(\nu \,\|\, \nu_1 \otimes P).$$

We write

$$\begin{aligned}
\mathbb{P}\big\{ M_n^X \in B(\nu,\varepsilon) \big\} &= \mathbb{E}^g\Big[\frac{\pi(X_1)}{\pi^g(X_1)} \prod_{i=2}^n \frac{P_{X_{i-1},X_i}}{P^g_{X_{i-1},X_i}} \, 1\{M_n^X \in B(\nu,\varepsilon)\} \Big] \\
&\geq \Big(\inf_{x,y\in\mathcal{X}} \tfrac{\pi(x)P^g_{y,x}}{\pi^g(x)P_{y,x}} \Big) \\
&\qquad \times \mathbb{E}^g\Big[\exp\Big(-\sum_{i=1}^n g(X_{i-1},X_i) - U_g(X_{i-1}) \Big) \, 1\{M_n^X \in B(\nu,\varepsilon)\} \Big].
\end{aligned}$$

The exponent is $-n\int (g - U_g)\, dM_n^X$ and, given $\delta > 0$, we may choose $0 < \tilde\varepsilon < \varepsilon$ small enough to ensure that, for every $\nu' \in B(\nu,\tilde\varepsilon)$,

$$\int (g - U_g)\, d\nu' \leq \int (g - U_g)\, d\nu + \delta.$$

Hence we can estimate the expectation from below by

$$\geq \exp\big(-n\big[\int (g - U_g)\, d\nu + \delta \big] \big) \, \mathbb{P}^g\big\{ M_n^X \in B(\nu,\tilde\varepsilon) \big\}.$$

Using Lemma 4.4, we conclude that

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big\{ M_n^X \in B(\nu,\varepsilon) \big\} &\geq -\int (g - U_g)\, d\nu - \delta \\
&= -\sum_{x\in\mathcal{X}} \nu(x,y) \log \frac{\nu(x,y)}{\nu_1(x) P_{x,y}} - \delta.
\end{aligned}$$

The result follows as $\delta > 0$ was arbitrary. □

A final remark about our two golden rules: Suppose we were interested in large deviations for the empirical measure L_n^X for a Markov chain. Then our *first* golden rule is violated, as L_n^X is not a type in this case. Our approach would be to move to a more inclusive statistic, namely the empirical pair measure, and then use the projection theorem to get a large deviation principle for L_n^X. The rate function remains in a variational form which cannot be solved explicitly.

5 The Dawson-Gärtner theorem and large deviation principles on the process level

In this section we give an example of a task which has to be carried out frequently in large deviation theory: lifting a large deviation principle from small to large spaces. As a reward we will see an example of the 'queen of large deviation principles', namely the principles on the process level.

We start by generalising Sanov's theorem for the empirical pair measures to empirical measures involving k-tuples. Let $X_1, X_2 \dots$ be a stationary Markov chain with statespace $\mathcal{X}$, strictly positive transition matrix P and stationary distribution π. Define the empirical k-measure

$$L_{n,k}^X = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, X_{i+1}, \dots, X_{i+k-1})},$$

which is a random probability measure on $\mathcal{M}_1(\mathcal{X} \times \overset{k}{\cdots} \times \mathcal{X})$.

Theorem 5.1. *The empirical k-measures $L_{n,k}^X$ satisfy a large deviation principle on $\mathcal{M}_1(\mathcal{X} \times \overset{k}{\cdots} \times \mathcal{X})$ with speed n and good rate function*

$$J_k(\nu) = \begin{cases} H(\nu \,\|\, \nu_{1,\dots,k-1} \otimes P) & \text{if } \nu_{1,\dots,k-1} = \nu_{2,\dots,k}, \\ \infty, & \text{otherwise,} \end{cases}$$

where $\nu_{i_1,\dots,i_j}$ is the marginal on the components with index $i_1, \dots, i_j$.

Proof. We look at the auxiliary Markov chain $Y_1, Y_2, \dots$ with statespace $\mathcal{X} \times \overset{k-1}{\cdots} \times \mathcal{X}$ given by

$$Y_j = (X_j, \dots, X_{j+k-2}).$$

This chain has the following properties

- its invariant distribution is $\pi^{(k)} = \pi \otimes P \otimes \overset{k-2}{\cdots} \otimes P$,
- its transition matrix is given by $P^{(k)}_{(x_1,\dots,x_{k-1}),(x_2,\dots,x_k)} = P_{x_{k-1},x_k}$,
- its empirical pair measure M_n^Y satisfies $L_{n,k}^X = M_n^Y \circ f^{-1}$ with

$$f\big((x_1, \dots, x_{k-1}), (x_2, \dots, x_k)\big) = (x_1, \dots, x_k).$$

From Theorem 4.2 we obtain a large deviation principle for M_n^Y with speed n and good rate function

$$J^Y(\nu) = \begin{cases} H(\nu \,\|\, \nu_1 \otimes P^{(k)}) & \text{if } \nu_1 = \nu_2, \\ \infty & \text{otherwise.} \end{cases}$$

The contraction principle hence gives a large deviation principle for $M_n^Y \circ f^{-1}$ with speed n and good rate function

$$J_k(\nu) = J^Y(\nu \circ f) = \begin{cases} H(\nu \,\|\, \nu_{1,\dots,k-1} \otimes P) & \text{if } \nu_{1,\dots,k-1} = \nu_{2,\dots,k}, \\ \infty & \text{otherwise,} \end{cases}$$

as required to complete the proof. □

Our aim is now to prove a large deviation principle for functionals of the Markov chain that may depend on an unbounded number of states, which means that we take a limit $k \to \infty$ in the preceding theorem. Formally, the Markov chain is a random element $(X_1, X_2, \dots)$ of the sequence space $\mathcal{X}^{\mathbb{N}}$ and we are interested in the empirical ∞-measure

$$L_{n,\infty}^X = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, X_{i+1}, \dots)}$$

which is in $\mathcal{M}(\mathcal{X}^{\mathbb{N}})$. The distribution of the Markov chain itself is an element of this space, therefore such a large deviation principle is said to be on the *process level*.

As $\mathcal{X}^{\mathbb{N}}$ is no longer finite, we need to discuss some topological subtleties now. The key point is that $\mathcal{M}(\mathcal{X}^{\mathbb{N}})$ is the *projective limit* of the spaces $\mathcal{M}(\mathcal{X}^k)$.

A *projective system* $(\mathcal{Y}_j, p_{ij})_{i \le j}$ is a family of metric spaces $\mathcal{Y}_j$ and continuous maps $p_{ij} \colon \mathcal{Y}_j \to \mathcal{Y}_i$ such that $p_{ik} = p_{ij} \circ p_{jk}$ whenever $i \le j \le k$, and p_{jj} are the identities on $\mathcal{Y}_j$. The *projective limit*, written as

$$\mathcal{Y}_\infty = \lim_{\leftarrow} \mathcal{Y}_j,$$

is the subspace of the product space $\mathcal{Y} = \prod_{j=1}^{\infty} \mathcal{Y}_j$ consisting of all sequences (y_j) for which $y_i = p_{ij}(y_j)$ whenever $i \le j$. There exist continuous projections

$$p_j \colon \mathcal{Y}_\infty \to \mathcal{Y}_j$$

given as the restrictions of the coordinate maps in the product space $\mathcal{Y}$.

Theorem 5.2 (Dawson-Gärtner). *Let (X_n) be a sequence of random variables on $\mathcal{Y}_\infty$ such that for any j the sequence of projections $(p_j(X_n))$ satisfy a large deviation principle on $\mathcal{Y}_j$ with good rate function J_j. Then (X_n) satisfies a large deviation principle on $\mathcal{Y}_\infty$ with good rate function*

$$J(x) = \sup_j J_j\big(p_j(x)\big), \qquad \text{for } x \in \mathcal{Y}_\infty.$$

Proof. See Dembo-Zeitouni Theorem 4.6.1. □

Let us apply this theorem to our situation. We take

$$\mathcal{Y}_j = \mathcal{M}(\mathcal{X}^j)$$

and, for $i \le j$, the marginal maps

$$p_{ij} \colon \mathcal{Y}_j \to \mathcal{Y}_i, \qquad \nu \mapsto \nu_{1,\dots,i}.$$

This defines a projective system. The associated projective limit consists formally of sequences $(\nu^{(n)})$ of probability measures on $\mathcal{X}^n$ such that

$$\nu^{(i)} = (\nu^{(j)})_{1,\dots,i} \quad \text{for } i \le j,$$

and the *Kolmogorov extension theorem* states that this implies that there exists a unique probability measure ν on $\mathcal{M}(\mathcal{X}^{\mathbb{N}})$ with $\nu^{(j)} = \nu_{1,\dots,j}$. Hence

$$\lim_{\leftarrow} \mathcal{M}(\mathcal{X}^j) = \mathcal{M}(\mathcal{X}^{\mathbb{N}})$$

with the topology given by

$$\nu^n \to \nu \qquad \text{iff} \qquad \nu^n_{1,\dots,j} \to \nu^n_{1,\dots,j} \text{ for all } j,$$

and projections given by

$$p_k(\nu) = \nu_{1,\dots,k}.$$

The given topology coincides with the weak topology on $\mathcal{M}(\mathcal{X}^{\mathbb{N}})$.

Note that, for every k,

$$p_k(L^X_{n,\infty}) = L^X_{n,k}.$$

The Dawson-Gärtner theorem therefore gives a large deviation principle for the empirical ∞-measure, although with an implicitly defined rate function.

To derive an explicit expression for the rate function we define the shifts

$$\theta_i \colon \mathcal{X}^{\mathbb{N}} \to \mathcal{X}^{\mathbb{N}}, \quad \theta_i(x_1, x_2, \ldots) = (x_{i+1}, x_{i+2}, \ldots).$$

A measure $\nu \in \mathcal{M}_1(\mathcal{X}^{\mathbb{N}})$ is *shift-invariant* if $\nu \circ \theta_1^{-1} = \nu$. Moreover, let

$$\mathbb{N}_i = \{\ldots, -1, 0, 1, \ldots, i\}$$

and define *backward measures* $\nu^*_{(i)} \in \mathcal{M}(\mathcal{X}^{\mathbb{N}_i})$ by

$$\nu^*_{(i)}\big(\{(\ldots, x_{i-1}, x_i) \colon (x_{i+1-k}, \ldots, x_i) \in A\}\big) = \nu_{1,\ldots,k}(A).$$

This is well-defined by the Kolmogorov extension theorem and shift-invariance. An analysis of the rate function yields the following result.

Theorem 5.3. *The empirical ∞-measures $L^X_{n,\infty}$ satisfy a large deviation principle on $\mathcal{M}_1(\mathcal{X}^{\mathbb{N}})$ with speed n and good rate function*

$$J_\infty(\nu) = \begin{cases} H(\nu^*_{(1)} \,\|\, \nu^*_{(0)} \otimes P) & \text{if } \nu \text{ shift-invariant,} \\ \infty, & \text{otherwise,} \end{cases}$$

where $\nu^*_{(i)}$ are the backwards measures and $\nu^*_0 \otimes P$ is given by

$$\begin{aligned} &\nu^*_{(0)} \otimes P\big(\{(\ldots, x_0, x_1) \colon (\ldots, x_0) \in A, x_1 = a\}\big) \\ &\qquad = \sum_{b \in \mathcal{X}} \nu^*_{(0)}\big(\{(\ldots, x_0) \colon (\ldots, x_{-1}, b) \in A\}\big) P_{b,a}. \end{aligned}$$

Proof. The rate function we get from the Dawson-Gärtner theorem is

$$J_\infty(\nu) = \begin{cases} \sup_k H(\nu_{1,\ldots,k} \,\|\, \nu_{1,\ldots,k-1} \otimes P) & \text{if } \nu_{1,\ldots,k-1} = \nu_{2,\ldots,k} \text{ for all } k, \\ \infty, & \text{otherwise.} \end{cases}$$

First note that ν is shift invariant if and only if

$$\nu_{1,\ldots,k-1} = (\nu \circ \theta_1^{-1})_{1,\ldots,k-1} = \nu_{2,\ldots,k} \quad \text{for all } k.$$

Next observe that $\nu^*_{(1),2-k,\ldots,1} = \nu_{1,\ldots,k}$ and $\nu^*_{(0),2-k,\ldots,0} = \nu_{1,\ldots,k-1}$, so that our result follows once we show that

$$H\big(\nu^*_{(1),-k,\ldots,1} \,\big\|\, \nu^*_{(0),-k,\ldots,0} \otimes P\big) \uparrow H(\nu^*_{(1)} \,\|\, \nu^*_{(0)} \otimes P)$$

for any shift-invariant ν. This follows directly from Pinsker's lemma, which is stated and proved below. □

Lemma 5.4 (Pinsker's lemma). *If $\nu^{(1)}, \nu^{(2)} \in \mathcal{M}(\mathcal{X}^{\mathbb{N}})$ then*

$$H\big(\nu^{(1)}_{1,\dots,k} \,\big\|\, \nu^{(2)}_{1,\dots,k}\big) \uparrow H(\nu^{(1)} \,\|\, \nu^{(2)}) := \int \big(\log \tfrac{d\nu^{(1)}}{d\nu^{(2)}}\big)\, d\nu^{(1)}.$$

Proof. We can represent entropy as

$$H(\nu^{(1)} \,\|\, \nu^{(2)}) = \sup_{\phi} \int \phi \, d\nu^{(1)} - \log \int e^{\phi} d\nu^{(2)},$$

where $\phi\colon \mathcal{X}^{\mathbb{N}} \to \mathbb{R}$ is continuous and bounded. We have shown a discrete variant of this on page 21. Permitting only functions ϕ depending on the first k, resp. $k-1$, coordinates we see that

$$H(\nu^{(1)} \,\|\, \nu^{(2)}) \geq H(\nu^{(1)}_{1,\dots,k} \,\|\, \nu^{(2)}_{1,\dots,k}) \geq H(\nu^{(1)}_{1,\dots,k-1} \,\|\, \nu^{(2)}_{1,\dots,k-1}).$$

Any bounded continuous function ϕ can be approximated, for sufficiently large k, by a function ϕ_k depending only on the first k coordinates, such that

$$\Big| \int \phi \, d\nu^{(1)} - \int \phi_k \, d\nu^{(1)} \Big| < \varepsilon,$$

and

$$\Big| \int e^{\phi} \, d\nu^{(2)} - \int e^{\phi_k} \, d\nu^{(2)} \Big| < \varepsilon.$$

This implies that $\lim_{k\to\infty} H(\nu^{(1)}_{1,\dots,k} \,\|\, \nu^{(2)}_{1,\dots,k}) = H(\nu^{(1)} \,\|\, \nu^{(2)})$. □

6 Varadhan's lemma and its inverse

Varadhan's lemma makes precise statements about the exponential growth of functionals of the form

$$\mathbb{E}\big[e^{nf(X_n)}\big], \qquad \text{as } n \to \infty.$$

Typical applications are the calculation of the free energy in statistical mechanics models, which are often of this form.

Theorem 6.1 (Varadhan's lemma). *If X_n is a family of random variables taking values in a metric space M satisfying a large deviation principle with speed n and rate function I, and $f\colon M \to \mathbb{R}$ is a continuous function, which is bounded from above, then*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\big[e^{nf(X_n)}\big] = \sup_{x\in M} \big\{f(x) - I(x)\big\}.$$

Remark: This results extends the Laplace principle in a natural way: Heuristically, we could write

$$\mathbb{E}\big[e^{nf(X_n)}\big] \approx \sum_x e^{nf(x)} \mathbb{P}\{X_n \approx x\} \approx \sum_x e^{n(f(x)-I(x))},$$

and then the biggest exponent in the sum determines the rate of the sum.

Proof. For the proof denote, for any $S \subset M$ Borel,

$$J_n(S) = \mathbb{E}\big[1\{X_n \in S\}\, e^{nf(X_n)}\big].$$

Denote by a the supremum of f, and by b the supremum of $f - I$, both are finite by our assumption on the boundedness of f.

Upper bound: We partition the space according o the values of f. Let $C = f^{-1}([b, a])$, define $c_j^N = b + \frac{j}{N}(a - b)$ and $C_j^N = f^{-1}[c_{j-1}^N, c_j^N]$, so that

$$C = \bigcup_{j=1}^{N} C_j^N.$$

All the cells C_j^N are closed and therefore

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\{X_n \in C_j^N\} \leq - \inf_{x\in C_j^N} I(x).$$

As $f(x) \leq c_j^N$ on C_j^N we obtain, from the Laplace principle,

$$\limsup_{n\to\infty} \frac{1}{n} \log J_n(C) \leq \max_{1\leq j\leq N} \big\{ c_j^N - \inf_{x\in C_j^N} I(x) \big\}.$$

Using now that

$$c_j^N \leq \inf_{x\in C_j^N} f(x) + \frac{1}{N}(a-b),$$

we get

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log J_n(C) &\leq \max_{1\leq j\leq N} \Big\{ \inf_{x\in C_j^N} f(x) - \inf_{x\in C_j^N} I(x) \Big\} + \frac{1}{N}(a-b) \\
&\leq \max_{1\leq j\leq N} \sup_{x\in C_j^N} \big\{ f(x) - I(x) \big\} + \frac{1}{N}(a-b) \\
&= \sup_{x\in C} \big\{ f(x) - I(x) \big\} + \frac{1}{N}(a-b) \\
&\leq b + \frac{1}{N}(a-b).
\end{aligned}$$

Letting $N \to \infty$ we get

$$\limsup_{n\to\infty} \frac{1}{n} \log J_n(C) \leq b.$$

As, trivially, $J_n(M \setminus C) \leq e^{nb}$ the Laplace principle implies

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}\big[e^{nf(X_n)}\big] \leq b = \sup_{x\in M} \big\{ f(x) - I(x) \big\}.$$

Lower bound: Pick $x \in M$ and $\varepsilon > 0$ arbitrary. Then

$$O(x,\varepsilon) = \{y \in M \colon f(y) > f(x) - \varepsilon\}$$

is an open neighbourhood of x, by continuity of f. Therefore

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\{X_n \in O(x,\varepsilon)\} \geq - \inf_{y\in O(x,\varepsilon)} I(y),$$

and

$$\liminf_{n\to\infty} \frac{1}{n} \log J_n\big(O(x,\varepsilon)\big) \geq f(x) - \varepsilon - I(x).$$

Now use that

$$\mathbb{E}\big[e^{nf(X_n)}\big] \geq J_n\big(O(x,\varepsilon)\big),$$

let $\varepsilon \downarrow 0$ and take the supremum over all $x \in M$ to find

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}\big[e^{nf(X_n)}\big] \geq \sup_{x\in M} \big\{f(x) - I(x)\big\}.$$

□

There is an inverse of Vardhan's lemma, due to Bryc.

Theorem 6.2 (Bryc's lemma). *If X_n is an exponentially tight family of random variables taking values in a metric space M such that*

$$\Lambda(f) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\big[e^{nf(X_n)}\big]$$

exists for all continuous and bounded functions $f\colon M \to \mathbb{R}$, then the random variables X_n satisfy a large deviation principle with speed n and good rate function

$$J(x) = \sup_{f\text{cts,bdd}} \big\{f(x) - \Lambda(f)\big\}.$$

Furthermore, we have the equation

$$\Lambda(f) = \sup_{x\in M} \big\{f(x) - J(x)\big\}.$$

The interest in Bryc's lemma consists in the fact that the existence of $\Lambda(f)$ need only be checked for bounded, continuous functions. However, for most practical purposes this class is too large, and one prefers to work in a more restrictive setting, in which the existence of $\Lambda(f)$ needs to be established for only a small class of functions. The most successful result in this direction is the Gärtner-Ellis theorem, which (under suitable assumptions) allows us to get away with testing only linear functions f.

Theorem 6.3 (Gärtner-Ellis theorem). *Suppose X_n is a family of random variables taking values in $\mathbb{R}^d$ such that*

$$\Lambda(x) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\big[e^{n\langle x, X_n\rangle}\big]$$

exists and is finite for all $x \in \mathbb{R}^d$. If Λ is additionally differentiable, then the random variables X_n satisfy a large deviation principle with speed n and good rate function

$$J(x) = \sup_{y\in\mathbb{R}^d} \big\{\langle x, y\rangle - \Lambda(y)\big\}.$$

Remark: The Gärtner-Ellis theorem for $d = 1$ is also a natural generalization of Cramér's theorem to dependent random variables. In this case J is given by a Legendre transform. As in our first result, we have not stated the Gärtner-Ellis theorem in its full strength, there are versions which allow Λ to be finite on some open set and steep at the boundary, see for example den Hollander V.2 for a precise statement and more details.

Proof. The proof is similar to that of Cramér's theorem, so we will only sketch it, pointing out some places where one needs to be more careful.

In the *upper bound* one can use Chebyshev's inequality, choosing for a given $x \in \mathbb{R}^d$ a $y \in \mathbb{R}^d$ which is a near maximizer in the definition of J and arguing that, given $\varepsilon > 0$ we find $\delta > 0$ with

$$\begin{aligned}\mathbb{P}\big\{X_n \in B_\delta(x)\big\} &\le \mathbb{P}\big\{\langle X_n - x, y\rangle \ge -\varepsilon\big\} \\ &\le e^{\varepsilon n}\mathbb{E}\big[e^{n\langle y, X_n\rangle}\big]e^{-n\langle x,y\rangle} \le e^{-n(J(x)-2\varepsilon)}.\end{aligned}$$

By covering with finitely many balls, this gives the result for compact sets. One then has to prove exponential tightness to pass to an upper bound for all closed sets. This can be proved using Chebyshev again, together with the fact that the $\Lambda(e_i) < \infty$ for all unit vectors e_i of $\mathbb{R}^d$. In fact, it suffices to have this for any small, positive multiple of the unit vectors.

For the *lower bound*, given $x \in \mathbb{R}^d$, we are using the changed measures $\hat{\mathbb{P}}_n$ given by

$$\frac{d\hat{\mathbb{P}}_n}{d\mathbb{P}_n}(X) = \frac{1}{\mathbb{E}[e^{n\langle y, X_n\rangle}]}\, e^{n\langle y, X\rangle},$$

where y is chosen as an *exposing hyperplane* for x, which means that

$$J(z) - J(x) > \langle z - x, y\rangle \qquad \text{for all } z \neq x.$$

Our conditions ensure the existence of such an exposing hyperplane, which is not a trivial fact. It is used in the proof of the weak law of large numbers, which we again derive from the upper bound. Let

$$\hat{\Lambda}_n(z) = \hat{\mathbb{E}}_n\big[e^{n\langle z,X\rangle}\big] = \frac{\mathbb{E}[e^{n\langle z+y,X_n\rangle}]}{\mathbb{E}[e^{n\langle y,X_n\rangle}]}.$$

Then

$$\lim_{n\to\infty}\frac{1}{n}\log\hat{\Lambda}_n(z) = \Lambda(y+z) - \Lambda(y),$$

so that the upper bound is applicable and yields

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}\big\{X_n \notin B_\varepsilon(x)\big\} \leq -\inf_{z\notin B_\varepsilon(x)}\big\{J(z) - \langle z,y\rangle + \Lambda(y)\big\}.$$

If z is a minimum in this variational problem, using the definition of J in the first and the exposing hyperplane property of y in the second step, gives

$$J(z) - \langle z,y\rangle + \Lambda(y) \geq \Big[J(z) - \langle z,y\rangle\Big] - \Big[J(x) - \langle x,y\rangle\Big] > 0,$$

establishing the weak law of large numbers. □

References

[DZ98] A. Dembo and O. Zeitouni. *Large deviations techniques and applications.* Springer, New York, (1998).

[dH00] F. den Hollander. *Large deviations.* AMS Fields Institute Monographs (2000).

[Ka01] O. Kallenberg. *Foundations of modern probability.* Springer, New York, (2001).