Large deviations (Fall 2024)

Course instructor: B. Dagallier, dagallier@ceremade.dauphine.fr

Draft date: 27/11/24

These notes closely follow notes from previous versions of the course taught by F. Huveneers and S. Olla.

For accessible large deviations courses that go well beyond the scope of these notes, we refer to Varadhan's Saint Flour lecture notes. For further content one may look at the lecture notes by Jan Swart, available on his webpage, as well as the book “Large deviations" by A. Dembo and O. Zeitouni provides an extremely thorough course.

1 Large deviations for sums of independent, identically distributed random variables

1.1 Introduction

Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. random variables with law μ. Let $(S_n)_{n\geq 1}$ denote the sequence of their sum:

$$S_n := \sum_{k=1}^{n} X_k, \qquad n \geq 1. \tag{1.1}$$

If $\mathbb{E}[|X_1|] < \infty$, then the law of large number says:

$$\lim_{n\to\infty} \frac{S_n}{n} = m := \mathbb{E}[X_1] \qquad \mu - a.s. \tag{1.2}$$

At this level of generality we cannot say anything about how this limit is approached, meaning about how $S_n/n - m$ looks like when n is large. However, if we assume better tail bounds on X_1 in the form:

$$\mathbb{E}[X_1^2] < \infty, \tag{1.3}$$

then Chebychev inequality gives:

$$\mathbb{P}\Big(\Big|\frac{S_n}{n} - m\Big| \geq \varepsilon\Big) \leq \frac{\mathrm{Var}(S_n/n)}{\varepsilon^2} = \frac{\mathrm{Var}(X_1)}{\varepsilon^2 n}. \tag{1.4}$$

Without further assumptions on X_1 the $1/n$ decay rate cannot be improved to $1/n^{1+\varepsilon}$ for any $\varepsilon > 0$ (exercise: find a counterexample). However, if we assume $\mathbb{E}[X_1^{2k}] < \infty$ for $k > 1$ then the same proof gives:

$$\mathbb{P}\Big(\Big|\frac{S_n}{n} - m\Big| \geq \varepsilon\Big) \leq \frac{\mathbb{E}[|X_1 - m|^{2k}]}{\varepsilon^k n^{k-1}}. \tag{1.5}$$

Assume now that X_1 has moment generating function well-defined on the whole line:

$$M(\lambda) := \mathbb{E}[e^{\lambda X_1}] < \infty, \qquad \lambda \in \mathbb{R}. \tag{1.6}$$

Then:

$$\mathbb{P}\Big(\Big|\frac{S_n}{n} - m\Big| \geq \varepsilon\Big) \leq \mathbb{P}\Big(\frac{S_n}{n} - m \geq \varepsilon\Big) + \mathbb{P}\Big(\frac{S_n}{n} - m \leq -\varepsilon\Big), \tag{1.7}$$

and one can separately estimate each probability using the exponential Chebychev inequality: for each $\lambda > 0$,

$$\mathbb{P}\Big(\frac{S_n}{n} - m \geq \varepsilon\Big) \leq e^{-n\lambda(\varepsilon+m)}\mathbb{E}\big[e^{\lambda S_n}\big] = \exp\Big[n\big(-\lambda(\varepsilon+m) + \log M(\lambda)\big)\Big], \tag{1.8}$$

where we used the independence and identical distributions to get $\mathbb{E}[e^{\lambda S_n}] = M(\lambda)^n$. The left-hand side does not depend on λ, so we can optimise on λ to get:

$$\mathbb{P}\Big(\frac{S_n}{n} - m \geq \varepsilon\Big) \leq \exp\Big[-n\sup_{\lambda>0}\big\{\lambda(m+\varepsilon) - \log M(\lambda)\big\}\Big] \tag{1.9}$$

Since $M(0) = 1$, the supremum is always non-negative. We will in fact prove in the next section that it is strictly positive, and that:

$$\sup_{\lambda>0}\big\{\lambda(m+\varepsilon) - \log M(\lambda)\big\} = \sup_{\lambda\in\mathbb{R}}\big\{\lambda(m+\varepsilon) - \log M(\lambda)\big\} =: I(\varepsilon). \tag{1.10}$$

Similarly, taking $\lambda < 0$ and optimising:

$$\begin{aligned}\mathbb{P}\Big(\Big|\frac{S_n}{n} - m\Big| \geq \varepsilon\Big) &\leq 2\exp\Big[-n\sup_{\lambda<0}\big\{\lambda(-m+\varepsilon) - \log M(\lambda)\big\}\Big] \\ &= 2\exp\big[-nI(\varepsilon)\big]. \end{aligned}\tag{1.11}$$

The function I is the *Legendre transform* of $\log M$. It enjoys many useful analytical properties (convexity, lower semi-continuity...) which will be studied in Section 1.3.

Remark 1.1.

- *We will see with Cramér's theorem that the decay rate e^{-n} is optimal for i.i.d. variables, and a lower bound in terms of the same function I can be proven. Studying sequences of measures (here the laws of S_n, $n \in \mathbb{N}$) satisfying upper and lower bounds with a matching function I is the topic of this course. Such measures will be said to satisfy a large deviation principle with rate function I.*
- *Although the event $\{|S_n/n - m| > \varepsilon\}$ occurs with very small probability, the function I encoding such* rare *events also gives us properties about the* typical *behaviour of S_n through $\{I = 0\}$. In more complicated examples we will see that functions playing a role similar to I can give even more information on (S_n), such as how variables should arrange themselves in order to create a given rare event.*

We claim that $I(\varepsilon) > 0$ unless $\varepsilon = 0$, so the right-hand side in (1.11) is exponentially small in n unless $\varepsilon = 0$. The prefactor 2 is therefore irrelevant. More generally we will be interested in probabilities up to logarithmic equivalence in the following sense.

Definition 1.2. *Two sequences $(a_n), (b_n)$ of positive numbers are logarithmically equivalent, denoted $a_n \asymp b_n$, if:*

$$\lim_{n\to\infty} \frac{1}{n} \log\left(\frac{a_n}{b_n}\right) = 0. \tag{1.12}$$

We will also say that $a_n \asymp 0$ if:

$$\lim_{n\to\infty} \frac{1}{n} \log a_n = -\infty. \tag{1.13}$$

Thus $\mathbb{P}(|S_n/n - m| > \varepsilon)$ is logarithmically equivalent to $e^{-nI(\varepsilon)}$, and the same holds for $n^{10}\mathbb{P}(|S_n/n - m| > \varepsilon)$ or $e^{\sqrt{n}}\mathbb{P}(|S_n/n - m| > \varepsilon)$.

1.2 Two explicit examples

Example 1.3. *Let $\rho \in [0,1]$ and let $(X_k)_k$ be i.i.d random variables with values in $\{-1,1\}$, such that $\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = 1/2$. Then, for each $\varepsilon \geq 0$:*

$$\mathbb{P}\big(S_n > n\varepsilon\big) \asymp e^{-nI(\varepsilon)}, \tag{1.14}$$

where:

$$I(\varepsilon) := \begin{cases} \frac{1-\varepsilon}{2}\log\left(\frac{1-\varepsilon}{2}\right) + \frac{1+\varepsilon}{2}\log\left(\frac{1+\varepsilon}{2}\right) + \log(2) & \textit{if } \varepsilon \in [0,1], \\ +\infty \quad \textit{otherwise} . \end{cases} . \tag{1.15}$$

Proof. Since $|X_1| = 1$ is a fortiori bounded by 1, $S_n \leq n$. This implies the claim for $\varepsilon \geq 1$: $\mathbb{P}(S_n > n) = 0 \asymp 0 = e^{-\infty}$.

Take now $\varepsilon \in [0,1]$. The distribution of S_n is explicit. Indeed, it can take values $m \in \{-n, -n+2, ..., n-2, n\}$ provided exactly $(m+n)/2$ of the X_i are equal to $+1$ and $(m-n)/2$ to -1. Thus:

$$\mathbb{P}(S_n = m) = \binom{n}{(n+m)/2} 2^{-n}. \tag{1.16}$$

Let $\varepsilon \in [0,1]$. Then $\mathbb{P}(S_n > n\varepsilon) = \mathbb{P}(S_n \geq m_\varepsilon)$, with m_ε the largest integer in $\{-n, -n+2, ..., n-2, n\}$ such that $m_\varepsilon \leq n\varepsilon$; in particular $m_\varepsilon \geq 0$. It will in fact be enough to analyse $\mathbb{P}(S_n = m_\varepsilon)$, since on the one hand:

$$\mathbb{P}(S_n \geq m_\varepsilon) \geq \mathbb{P}(S_n = m_\varepsilon), \tag{1.17}$$

and on the other hand $m \geq 0 \mapsto \mathbb{P}(S_n = m)$ is decreasing from the explicit formula, therefore:

$$\mathbb{P}(S_n \geq m_\varepsilon) = \sum_{m \geq m_\varepsilon} \mathbb{P}(S_n = m) \leq n\mathbb{P}(S_n \geq m_\varepsilon). \tag{1.18}$$

Thus $\mathbb{P}(S_n = m_\varepsilon) \asymp \mathbb{P}(S_n \geq m_\varepsilon)$. Let us compute the former probability. Recall Stirling's formula:

$$\log p! = p\log p - p + o(p). \tag{1.19}$$

The fact that $|m_\varepsilon - n\varepsilon| < 2$ and Stirling's formula give:

$$\begin{aligned}\frac{1}{n}\log \mathbb{P}(S_n = m_\varepsilon) &= -\log 2 + \frac{1}{n}\log\binom{n}{(n+m_\varepsilon)/2}\\ &= -\log 2 - \frac{n+m_\varepsilon}{2n}\log\Big(\frac{n+m_\varepsilon}{2n}\Big) - \frac{n-m_\varepsilon}{2n}\log\Big(\frac{n-m_\varepsilon}{2n}\Big) + o_n(1)\\ &= -I(\varepsilon) + o_n(1). \qquad (1.20)\end{aligned}$$

□

Although large deviations for a sequence $(X_k)_k$ involve rare events, the *typical* value of certain functions of the X_k may be determined by a large deviation event.

Example 1.4 (Taken from Varadhan's Saint Flour lecture notes)**.** *Let $\alpha > 0$ and $(X_k)_k$ be an i.i.d. sequence of variables with values in $[1,2]$ and law μ. Let $P_n = \prod_{k=1}^n X_k$. Then:*

$$\lim_{n\to\infty}\frac{1}{n}\log P_n = \int \log(x)\mu(dx) \qquad a.s., \qquad (1.21)$$

while:

$$\frac{1}{n}\log \mathbb{E}[P_n] = \log\Big(\int x\mu(dx)\Big) > \int \log(x)\mu(dx), \qquad (1.22)$$

with the strict inequality given by Jensen inequality. Thus the typical *value of P_n is determined by* rare events *for the X_k. In fact,*

$$\mathbb{E}[P_n] = \mathbb{E}\Big[\exp\Big[\sum_{k=1}^n \log X_k\Big]\Big], \qquad (1.23)$$

with the argument of order e^n and probabilities that $\sum_{k=1}^n \log X_k$ deviate from their mean of order e^{-nI}. The mean value is obtained when the two balance each other out:

$$\mathbb{E}[P_n] \asymp \exp\Big[n\sup_{x\in\mathbb{R}}\{x - I(x)\}\Big] = \exp\Big[n\log\Big(\int x\mu(dx)\Big)\Big], \qquad (1.24)$$

with I the Legendre transform of $\phi = \log M_{\log X_1}$. The last equality comes from the fact that the supremum is reached at $x = \phi'(1)$ (see next section), and $I(\phi'(1)) = \phi'(1) - \phi(1)$, so that $\phi'(1) - I(\phi'(1)) = \phi(1) = \log\int x\mu(dx)$.

1.3 Properties of Legendre transform

Cramér's theorem, stated in the next section, generalises the argument of the introduction to prove large deviation for i.i.d real-valued random variables $(X_i)_{i\geq 1}$ with a function I given by the Legendre transform of the log-moment generating function of X_1. We first study properties of this Legendre transform.

In the following a random variable X_1 with log-moment generating function $\phi(\lambda) := \log\mathbb{E}[e^{\lambda X_1}]$ is fixed, and we set:

$$\mathcal{D}_\phi := \big\{\lambda\in\mathbb{R} : \phi(\lambda) < \infty\big\}. \qquad (1.25)$$

Throughout we assume that there is $\lambda_0 > 0$ such that $[-\lambda_0, \lambda_0] \subset \mathcal{D}_\phi$. The interior of $\mathcal{D}_\phi$ will be denoted by $\mathcal{D}_\phi^o$. In particular $0 \in \mathcal{D}_\phi^o$.

Proposition 1.5. *1. The function ϕ is convex and $\mathcal{D}_\phi$ is an interval.*

2. ϕ is C^∞ on $\mathcal{D}_\phi^o$, with e.g. for each $\lambda \in \mathcal{D}_\phi^o$:

$$\phi'(\lambda) = \frac{\mathbb{E}[X_1 e^{\lambda X_1}]}{M(\lambda)}, \quad \phi''(\lambda) = \frac{\mathbb{E}[X_1^2 e^{\lambda X_1}]}{M(\lambda)} - \frac{\mathbb{E}[X_1 e^{\lambda X_1}]^2}{M(\lambda)^2}. \tag{1.26}$$

3. If X_1 is not deterministic then ϕ is strictly convex on $\mathcal{D}_\phi^o$.

Proof. Let $\alpha \in [0,1]$ and $\lambda_1, \lambda_2 \in \mathbb{R}$. Hölder inequality with exponents $1/\alpha, 1/(1-\alpha)$ gives:

$$\begin{aligned} M(\alpha\lambda_1 + (1-\alpha)\lambda_2) &= \mathbb{E}\big[e^{\alpha\lambda_1 X_1 + (1-\alpha)\lambda_2 X_2}\big] \\ &\leq \mathbb{E}\big[e^{\lambda_1 X_1}\big]^\alpha \mathbb{E}\big[e^{\lambda_2 X_2}\big]^{1-\alpha} = M(\lambda_1)^\alpha M(\lambda_2)^{1-\alpha}. \end{aligned} \tag{1.27}$$

Thus $\phi = \log M$ is convex and $\mathcal{D}_\phi$ is an interval.

For the regularity, let us prove instead that M is C^∞ on $\mathcal{D}_\phi^o$. This is enough since $M(\lambda) \geq e^{\lambda \mathbb{E}[X_1]} > 0$ by Jensen inequality. E.g. for the first derivative, one has:

$$M'(\lambda) = \lim_{h\to 0} \frac{1}{h} \mathbb{E}\Big[e^{(\lambda+h)X_1} - e^{\lambda X_1}\Big]. \tag{1.28}$$

The difference in the expectation reads:

$$\frac{1}{h}\Big|e^{(\lambda+h)X_1} - e^{\lambda X_1}\Big| = \Big|X_1 \int_0^1 e^{(\lambda+th)X_1}\,dt\Big| \leq |X_1| e^{\lambda X_1} e^{\delta |X_1|}, \qquad |h| \leq \delta. \tag{1.29}$$

The right-hand side is integrable if δ is small enough as $\lambda \in \mathcal{D}_\phi^o$. The dominated convergence theorem then implies that M is differentiable, with derivative $M'(\lambda) = \mathbb{E}[X_1 e^{\lambda X_1}]$. A similar argument for higher derivatives yields the claim, and:

$$\frac{d^k}{d\lambda^k} M(\lambda) = \mathbb{E}[X_1^k e^{\lambda X_1}], \qquad k \geq 1. \tag{1.30}$$

For the last point, let $\lambda \in \mathcal{D}_\phi^o$ and notice that $\phi''(\lambda) = \mathrm{Var}_{\mu_\lambda}(x)$ where, if μ denotes the law of X_1, then μ_λ is the probability measure $\mu_\lambda(dx) = e^{\lambda x - \phi(\lambda)}\mu(dx)$. X_1 is deterministic if and only if μ is a Dirac measure if and only if $\mu_\lambda(dx) \ll \mu$ is a Dirac measure, which is a necessary and sufficient condition for $\mathrm{Var}_{\mu_\lambda}(x)$ to vanish. □

Define the Legendre transform I of ϕ (also called Fenchel-Legendre transform):

$$I(x) := \sup_{\lambda \in \mathbb{R}} \big\{\lambda x - \phi(\lambda)\big\} \in \mathbb{R} \cup \{+\infty\}, \tag{1.31}$$

and let $\mathcal{D}_I$ denote its domain:

$$\mathcal{D}_I := \big\{x \in \mathbb{R} : I(x) < \infty\big\}. \tag{1.32}$$

Proposition 1.6. *1. I is a convex, lower semi-continuous function on $\mathbb{R}$ with values in $\mathbb{R}_+ \cup \{+\infty\}$.*

2. *$\mathcal{D}_I^o$ is contained in $Im(\phi')$, $\mathcal{D}_I$ is an interval and I is C^∞ and strictly convex on $\mathcal{D}_I^o$.*

3. *I has compact sub-level sets and satisfies $\lim_{|x|\to\infty} I(x) = \infty$.*

4. *$I(x) = 0$ if and only if $x = \mathbb{E}[X_1]$, and I is increasing on $[\mathbb{E}[X_1], +\infty)$, decreasing on $(-\infty, \mathbb{E}[X_1]]$.*

5. *If $\mathbb{E}[X_1] \in \mathcal{D}_I^o$, then $I'(\mathbb{E}[X_1]) = 0$, $I''(\mathbb{E}[X_1]) = 1/\operatorname{Var}(X_1)$.*

Remark 1.7.

- *Lower semi-continuity and compact level sets will be important properties of more general large deviation statements. Indeed, if one has an upper bound of the form $\mathbb{P}(S_n/n \in [a,b]) \leq e^{-n\inf_{[a,b]} J}$ for a function J that either does not have compact sub-level sets or is not lower semi-continuous, then it could be that $\inf_A J = 0$ on some set A which does not contain $\{J = 0\}$.*

- *Item 5) allows for a heuristic connection between large deviations and central limit theorem. Indeed, assume:*

$$\mathbb{P}\Big(\frac{S_n}{n} - \mathbb{E}[X_1] \approx x\Big) = e^{nI(\mathbb{E}[X_1]+x)}, \qquad x \in \mathbb{R}. \tag{1.33}$$

  The central limit theorem scaling corresponds to $x = a/\sqrt{n}$, and gives convergence of the probability in the left hand side to the centred normal distribution with variance $\operatorname{Var}(X_1)$. On the other hand:

$$\begin{aligned} nI(x) &= nI\big(\mathbb{E}[X_1]\big) + \sqrt{n}aI'\big(\mathbb{E}[X_1]\big) + \frac{a^2}{2}I''\big(\mathbb{E}[X_1]\big) + o_n(1) \\ &= \frac{a^2}{2\operatorname{Var}(X_1)}, \end{aligned} \tag{1.34}$$

  Thus taking the CLT scaling, or taking the large deviation scaling and then expanding the result around 0 gives the same result.

Proof. 1) I is convex and lower semi-continuous as a supremum of linear, continuous functions. Taking $z = 0$ gives $I \geq 0$. In addition, as $\phi = +\infty$ outside of $\mathcal{D}_\phi$,

$$I(x) = \sup_{\lambda \in \mathcal{D}_\phi} \big\{\lambda x - \phi(\lambda)\big\}. \tag{1.35}$$

2) Note that if $x_1, x_2 \in \mathcal{D}_I$ then convexity of I implies $[x_1, x_2] \in \mathcal{D}_I$, thus $\mathcal{D}_I$ is an interval. If X_1 is deterministic then $\mathcal{D}_I$ is reduced to the point X_1 and $\phi' = X_1$. If X_1 is not deterministic, let us prove that $\mathcal{D}_I^o \subset \operatorname{Im}(\phi')$.

Let $x \in \mathcal{D}_\phi^o$. Convexity of ϕ on $\mathcal{D}_\phi$ and strict convexity on $\mathcal{D}_\phi^o$ implies that the supremum defining I may be reached at the boundaries of $\mathcal{D}_\phi$ (which may be $\pm\infty$) and at most one point in $\mathcal{D}_\phi^o$. As $x \in \mathcal{D}_I^o$, we claim that only the latter is possible. Indeed, let $\delta > 0$ be such that $x \pm \delta \in \mathcal{D}_I^o$. If the supremum defining I were reached at $+\infty$, say, then $\limsup_{\lambda\to\infty}[\lambda x - \phi(\lambda)] < \infty$ is incompatible with $I(x+\delta) > 0$ since:

$$I(x+\delta) \geq \limsup_{\lambda\to\infty} \Big\{\lambda\delta + (\lambda x - \phi(\lambda))\Big\} = +\infty. \tag{1.36}$$

There is thus $f(x) \in \mathcal{D}_\phi^o$ such that $I(x) = xf(x) - \phi(f(x))$. As $f(x)$ is a critical point for $\lambda \mapsto \lambda x - \phi(\lambda)$, it satisfies:

$$x = \phi'(f(x)). \tag{1.37}$$

As X_1 is not deterministic, then $\phi'' > 0$ by Proposition 1.5, item 2). This means that $f(x)$ is uniquely defined, and smooth by the implicit function theorem. Thus I is smooth on $\mathcal{D}_I^o$, and:

$$\begin{aligned} I'(x) &= f(x) + xf'(x) - f'(x)\phi'(f(x)) = f(x), \\ I''(x) &= f'(x) = \frac{1}{\phi''(f(x))} > 0. \end{aligned} \tag{1.38}$$

This implies strict convexity and 5). The identity $x = \phi'(f(x))$ also gives $x \in \mathrm{Im}(\phi')$, thus $\mathcal{D}_I^o \subset \mathcal{D}_\phi^o$.

3) If $x, \lambda > 0$, then $I(x) > \lambda x - \phi(\lambda)$. Thus $I(x)/x \geq \lambda - \phi(\lambda)/x$, and the right-hand side converges to λ when $x \to \infty$. If $x < 0$ the same reasoning applies for $\lambda < 0$. Thus $\{I \leq a\}$ is bounded for each $a \geq 0$, and since it is closed by lower semi-continuity it is in fact compact.

4) We have already seen $I(x) = 0$ if $x = \mathbb{E}[X_1]$. As I is convex on $\mathcal{D}_I$, it is constant on $[x, y]$ for any y with $I(y) = 0$. If such an y exists, then $(x, y) \subset \mathcal{D}_I^o$ on which set I is strictly convex, which is absurd. The convexity and the fact that $\mathbb{E}[X_1]$ is the only minimum of I implies that it is increasing on $[\mathbb{E}[X_1], \infty)$ and decreasing on $(-\infty, \mathbb{E}[X_1]]$. □

Exercise 1.8.

1. *Consider the Gaussian distribution with mean $m \in \mathbb{R}$ and variance σ^2. Show that $I(x) = (x-m)^2/2\sigma^2$*

2. *Let $p(dx) = \lambda e^{-\lambda x}\mathbf{1}_{x>0}$ be the exponential distribution of parameter λ. Show that $I(x) = \lambda x - 1 - \log(\lambda x)$ if $x > 0$ and $I(x) = +\infty$ if $x \leq 0$.*

3. *Let $I(x) = \sup_{\lambda \in \mathbb{R}}\{\lambda x - f(\lambda)\}$ be the Legendre transform of a convex function $f : \mathbb{R} \to \mathbb{R}$. Show that the Legendre transform of I is f:*

$$f(\lambda) = \sup_{x \in \mathbb{R}} \big\{ x\lambda - I(x) \big\}. \tag{1.39}$$

This is a special case of the more general Fenchel-Moreau theorem. What happens if f is not convex?

1.4 Cramér's theorem

We now state a general large deviation result for sequences $(X_k)_k$ of i.i.d. real random variables known as Cramér's theorem.

Theorem 1.9. *Let $(X_k)_k$ be a sequence of i.i.d random variables with values in $\mathbb{R}$. Assume that the moment generating function $M(\lambda) = \mathbb{E}[e^{\lambda X_1}]$ of X_1 is finite on $\mathbb{R}$:*

$$\forall \lambda \in \mathbb{R}, \qquad M(\lambda) < \infty. \tag{1.40}$$

Then, for any $a \leq b \in \mathbb{R} \cup \{-\infty, \infty\}$:

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \in [a,b]\Big) \leq -\inf_{[a,b]} I, \tag{1.41}$$

and:

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \in (a,b)\Big) \geq -\inf_{(a,b)} I. \tag{1.42}$$

where I is the Legendre transform *of $\phi = \log M$:*

$$I(x) := \sup_{\lambda\in\mathbb{R}} \big\{\lambda x - \phi(\lambda)\big\}, \qquad x \in \mathbb{R}. \tag{1.43}$$

Remark 1.10. *The assumptions of Theorem 1.9 can be relaxed considerably: it is not even needed that the X_1 have finite mean, much less that M be finite on the real line. See Section 2.2 in the book by Dembo and Zeitouni for generalisations.*

Proof. Consider the upper bound. If the mean $\mathbb{E}[X_1]$ is in $[a,b]$, then $\inf_{[a,b]} I = 0$ and the law of large numbers gives the claim. We henceforth assume $\mathbb{E}[X_1] \notin [a,b]$, say $a > m$. Then $\inf_{[a,b]} I = I(a)$ as I is increasing on $[a,+\infty)$, and the exponential Chebychev inequality gives as before:

$$\mathbb{P}\Big(\frac{S_n}{n} \in [a,b]\Big) \leq \mathbb{P}\Big(\frac{S_n}{n} \geq a\Big) \leq \exp\big[-nI(a)\big] = \exp\big[-n\inf_{[a,b]} I\big]. \tag{1.44}$$

The argument if $b < m$ is similar.

Consider now the lower bound. The proof involves one of the key principles in large deviation theory: to estimate the cost of a rare event, tilt the measure so that the rare event becomes typical under the tilted measure, and estimate the cost of the tilt.

We first prove a weaker lower bound as follows:

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \in (a,b)\Big) \geq -\inf_{(a,b)\cap\mathcal{D}_o^I} I. \tag{1.45}$$

The point of this reduction is that, for an element $c \in \mathcal{D}_I^o$, we know exactly how to tilt the measure so that S_n/n converges to c as we now explain.

Fix $c \in \mathcal{D}_I^o$ and $\delta > 0$ such that $[c-\delta, c+\delta] \in [a,b]$. Let $f(c)$ be the unique point such that $\phi'(f(c)) = c$, which exists by the proof of Proposition 1.6. If μ denotes the law of X_1, recall that $\phi'(f(c)) = c$ is the average value of X_1 under the tilted measure $\mu_{f(c)} = e^{f(c)X_1 - \phi(f(c))}\mu$. Thus, for each $\varepsilon \in (0, \delta]$

$$\begin{aligned}
\mathbb{P}\Big(\frac{S_n}{n} \in [a,b]\Big) \geq \mathbb{P}\Big(\frac{S_n}{n} \in [c-\varepsilon, c+\varepsilon]\Big) &= \int_{\mathbb{R}^n} \mathbf{1}_{S_n/n \in [c-\varepsilon, c+\varepsilon]} \prod_{i=1}^n \mu(dx_i) \\
&= \int_{\mathbb{R}^n} \exp\Big[-f(c)S_n + n\phi(f(c))\Big] \mathbf{1}_{S_n/n \in [c-\varepsilon, c+\varepsilon]} \prod_{i=1}^n \mu_{f(c)}(dx_i) \\
&\geq \exp\big[-n[I(c) - \varepsilon]\big] \int_{\mathbb{R}^n} \mathbf{1}_{S_n/n \in [c-\varepsilon, c+\varepsilon]} \prod_{i=1}^n \mu_{f(c)}(dx_i). \qquad (1.46)
\end{aligned}$$

The probability in the right-hand side converges to 1 when n is large, thus:

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \in [a,b]\Big) \geq -I(c) + \varepsilon. \qquad (1.47)$$

Since this is true for any $\varepsilon \in (0, \delta]$, any $c \in \mathcal{D}_I^o \cap [a,b]$ and the left-hand side is independent of c, ε, we have proven the lower bound restricted to $\mathcal{D}_I^o$.

We now improve the lower bound to $\inf_{(a,b)} I$. If $m \in (a,b)$ then the law of large numbers gives the claim. Otherwise assume e.g. $a > m$. Note first:

$$\inf_{(a,b)} I = \inf_{(a,b)\cap\mathcal{D}_I} I. \qquad (1.48)$$

Recall that $\mathcal{D}_I$ is an interval. If $\mathcal{D}_I$ is reduced to $\{m\}$, then $a > m$ implies $\inf_{(a,b)} I = +\infty = \inf_{\mathcal{D}_I^o \cap (a,b)} I$ using $\inf \emptyset = +\infty$. Otherwise, as $\mathcal{D}_I^o$ is an open set by definition, $\mathcal{D}_I^o \cap (a,b) = \mathcal{D}_I \cap (a,b)$ which concludes the proof. □

In more general situations than large deviations for sums of i.i.d. real random variables, the upper bound for closed sets/lower bound for open sets in Theorem 1.9 is the best one can hope for. In the sums of i.i.d. situation, however, the lower bound can actually be improved as stated next.

Proposition 1.11. *Let $a < b \in \mathbb{R} \cup \{\pm\infty\}$. Then:*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \in [a,b]\Big) \geq -\inf_{[a,b]} I. \qquad (1.49)$$

Thus $\mathbb{P}(S_n/n \in [a,b])$ and $e^{-n \inf_{[a,b]} I}$ are logarithmically equivalent.

Proof. If $m = \mathbb{E}[X_1] \in (a,b)$, there is nothing to prove. Assume otherwise that $a \geq m$, the argument for $n \leq m$ being identical. We would like to use the same tilt argument that in the proof of Theorem 1.9. There are two points to check:

- The existence of a tilt $f(a)$ so that the tilted measure has mean a.

- A lower bound on the probability $S_n \in [a, a+\varepsilon]$ for each small $\varepsilon > 0$ even for $a \notin \mathcal{D}_I^o$.

Case 1: $X_1 < a$ a.s. In this case $\mathbb{P}(S_n/n \in [a,b]) = 0$ and $I(a) = \sup_{\lambda\in\mathbb{R}}\{\lambda a - \phi(\lambda)\}$ is shown to diverge by taking $\lambda \to +\infty$.

Case 2: $X_1 \leq a$ a.s. **with** $\mathbb{P}(X_1 = a) = p > 0$. Then:

$$\mathbb{P}\Big(\frac{S_n}{n} \in [a,b]\Big) = \mathbb{P}\big(X_i = a, i \in \{1,...,n\}\big) = p^n. \tag{1.50}$$

On the other hand,

$$\lim_{\lambda\to\infty} \big\{a\lambda - \phi(\lambda)\big\} = -\log p. \tag{1.51}$$

Thus the claim also holds in that case.

Case 3: $\mathbb{P}(X_1 > a) > 0$. In this case we can check items (i) and (ii) above and conclude on the lower bound as in Theorem 1.9 as explained next. For item (i), we want to find λ such that $\phi'(\lambda) = a$. Let $M_a(\lambda) := e^{-\lambda a}M(\lambda)$. Notice:

$$\phi'(\lambda) - a = \frac{M_a'(\lambda)}{M_a(\lambda)}, \tag{1.52}$$

thus it is enough to prove that M_a' vanishes at some λ. By assumption on X_1, M_1 is C^∞ on the real line. We know $M_a'(0) = m - a < 0$ and:

$$M_a''(\lambda) = \mathbb{E}[(X_1 - a)^2 e^{\lambda(X_1 - a)}] \geq \mathbb{E}[(X_1 - a)^2 \mathbf{1}_{X_1 \geq a}] > 0, \qquad \lambda \in \mathbb{R}. \tag{1.53}$$

Thus M_a'' is uniformly convex and therefore M_a' vanishes at a value $f(a) > 0$.

We now explain item (ii). The tilted measure $\mu_{f(a)}$ admits moments of order 2 since ϕ is finite on the real line. In particular the central limit theorem applies and:

$$\begin{aligned}\liminf_{n\to\infty} \mu_{f(a)}^{\otimes n}\Big(\frac{1}{n}\sum_{i=1}^n (x_i - a) \in [0,\varepsilon]\Big) &= \liminf_{n\to\infty} \mu_{f(a)}^{\otimes n}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n (x_i - a) \in [0,\varepsilon\sqrt{n}]\Big)\\ &\geq \liminf_{n\to\infty} \mu_{f(a)}^{\otimes n}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n (x_i - a) \in [0,1]\Big) > 0. \end{aligned}\tag{1.54}$$

This concludes the proof. □

2 The Curie-Weiss model

2.1 Introduction

In this section we discuss constructions from statistical mechanics, the Curie-Weiss and Ising models, intended to describe magnetism in solids at a qualitative level. In statistical mechanics one considers a large number n of components, represented by a variable σ_i in a space S ($1 \leq i \leq n$) which can for instance be a position in $\mathbb{R}^d$, a colour, a charge, a combination of all that, etc.

The goal is then to describe the large n behaviour of the collection $(\sigma_i)_{1\leq i\leq n}$ assuming that any particular value of the $\sigma's$ occurs with a probability proportional to $e^{-\beta H(\sigma)}\mu^{\otimes n}(d\sigma)$, where μ is some base measure on S, $H : S^n \to \mathbb{R}$ is referred to as the energy and $\beta \in [0,\infty]$ represents the inverse of a temperature.

In the simplest case,

$$H(\sigma) = \sum_i h(\sigma_i), \tag{2.1}$$

which means that all components are independent and have the same distribution $\propto e^{-h(\sigma)}d\mu$. This corresponds to the i.i.d. situation treated in Section 1, and we now already know how to answer questions about the large n behaviour of the law of $\sum_i \sigma_i$ if they are real valued.

It is however much more interesting and realistic to consider interacting components, i.e. H cannot be written as a sum of functions of one component only. There are a myriad model and questions one can consider in this context, but we will focus on just one: phase transitions in the Ising model, a simple model for magnetism.

In the Ising model components are called spins and take values in $\{-1,1\} = \{\downarrow,\uparrow\}$. The space of all spins is $\{-1,1\}^\Lambda$, where Λ is a lattice (e.g. $\mathbb{Z}^d$ or a finite portion). The lattice is an ideal representation of the crystalline structure of a solid, with point of the lattice (“sites") representing atoms in the solid. Each spin is located at a site; can be thought of as an arrow, here only pointing up or down; and represents an elementary chunk of magnetism. If all spins/arrows point in the same direction, then the model is magnetic. If no direction is preferred, then the model is not magnetic. Spins interact, i.e. the spins are not i.i.d. The interaction induces a *phase transition* in the limt $|\Lambda| \to \infty$ as β is varied: the model will or will not be magnetic depending on β.

2.2 The Ising model

Let $d \geq 1$ and $L \in \mathbb{N}$. Write $\Lambda_L = \mathbb{Z}^d \cap [-L,L]^d$ and let $\beta \in\geq 0$. The Ising model in dimension d at inverse temperature β on $\{-1,1\}^{\Lambda_L}$ is the probability measure:

$$\mu_{\beta,L}(\sigma) = \frac{1}{Z_{\beta,L}} \exp\Big[\beta \sum_{\substack{i,j\in\Lambda_L\\|i-j|=1}} \sigma_i\sigma_j\Big], \qquad \sigma \in \{-1,1\}^{\Lambda_L}. \tag{2.2}$$

Above, $Z_{\beta,L}$ is a normalisation factor called the partition function. This is of the form described in the last section with $\mu = \delta_{-1} + \delta_1$ and βH the argument of the exponential. If $\beta = 0$ all spins are independent: $\mu_{0,L}$ is the product Bernoulli measure with parameter 1/2. If $\beta > 0$ spins are not independent and alignment with neighbouring spins is favoured. Note that $\mu_{\beta,L}(\sigma) = \mu_{\beta,L}(-\sigma)$ for each σ, so that the average magnetisation $|\Lambda_L|^{-1}\sum_i \sigma_i$ has mean 0:

$$\mathbb{E}_{\mu_{\beta,L}}\Big[\frac{1}{|\Lambda_L|}\sum_{i\in\Lambda_L}\sigma_i\Big] = 0. \tag{2.3}$$

One can however ask about how the mean magnetisation concentrates around 0. For instance, is it true that, whatever $\varepsilon > 0$:

$$\lim_{L\to\infty} \mu_{L,\beta}\Big(\Big|\frac{1}{|\Lambda_L|}\sum_{i\in\Lambda_L}\sigma_i\Big| > \varepsilon\Big) = 0? \tag{2.4}$$

It turns out that, when $d \geq 2$, the answer depends on the value of β; this is the phase transition alluded to above. More precisely, if $d \geq 2$, there is $\beta_c = \beta_c(d) \in (0,\infty)$ such that:

$$\begin{aligned} \forall \varepsilon > 0, \qquad & \lim_{L\to\infty} \mu_{L,\beta}\Big(\Big|\frac{1}{|\Lambda_L|}\sum_{i\in\Lambda_L}\sigma_i\Big| > \varepsilon\Big) = 0 \qquad \text{if } \beta < \beta_c \\ \exists \varepsilon > 0, \qquad & \liminf_{L\to\infty} \mu_{L,\beta}\Big(\Big|\frac{1}{|\Lambda_L|}\sum_{i\in\Lambda_L}\sigma_i\Big| > \varepsilon\Big) > 0 \qquad \text{if } \beta > \beta_c. \end{aligned} \tag{2.5}$$

Establishing this claim is beyond the scope of this lecture. We will however prove it for a simplified model known as the Curie-Weiss model.

2.3 The Curie-Weiss model

In the Curie-Weiss model, every spin has a weak interaction with every other spin (not just nearest neighbours). The lattice therefore plays no role any more and the state space is $\{-1,1\}^n$ $(n \geq 1)$. The model is defined as:

$$\mu^{\mathsf{CW}}_{\beta,n}(\sigma) = \frac{1}{Z^{\mathsf{CW}}_{\beta,n}} \exp\Big[\beta\sum_{i=1}^n \sigma_i\Big(\frac{1}{n}\sum_{j=1}^n\sigma_j\Big)\Big], \qquad \sigma \in \Sigma_n := \{-1,1\}^n \tag{2.6}$$

This model is called a mean-field approximation of the Ising model, as spins only interact with their empirical mean $m_n(\sigma) = \frac{1}{n}\sum_i \sigma_i$, and we can write:

$$\mu^{\mathsf{CW}}_{\beta,n}(\sigma) = \frac{1}{Z^{\mathsf{CW}}_{\beta,n}} \exp\big[n\beta m_n(\sigma)^2\big]. \tag{2.7}$$

In the Curie-Weiss model (2.5) can be established rigorously, by proving a large deviation principle for the magnetisation. To state it, define the energy density e, entropy density s and free energy density f_β as the following functions of $m \in [-1,1]$:

$$\begin{aligned} e(m) &:= -m^2, \\ s(m) &:= -\frac{1+m}{2}\log\Big(\frac{1+m}{2}\Big) - \frac{1-m}{2}\log\Big(\frac{1-m}{2}\Big), \\ f_\beta &:= \beta e - s. \end{aligned} \tag{2.8}$$

Theorem 2.1. *Let $\beta \geq 0$. For any interval $J \subset [-1,1]$ not reduced to a point,*

$$\mu^{\mathsf{CW}}_{n,\beta}(m_n \in J) \asymp e^{-n\inf_J I_\beta}, \qquad I_\beta := f_\beta - \inf_{[-1,1]} f_\beta \tag{2.9}$$

Before we prove the theorem, let us explain why it proves (2.5).

Proposition 2.2. *The rate function I_β is convex and has a unique minimiser at $m = 0$ when $\beta \leq 1/2$.*

For $\beta > 1/2$, there are two distinct minimisers at $\pm m_\beta$, with $m_\beta > 0$ converging to 1 when $\beta \to 0$. In particular I_β is not convex, in stark contrast with the i.i.d case.

Proof. The proof is an elementary computation. Let $m \in (-1,1)$ and notice:

$$f_\beta''(m) = -2\beta + \frac{1}{2(1+m)} + \frac{1}{2(1-m)} = \frac{1-2\beta(1-m)^2}{(1-m)^2}. \tag{2.10}$$

In particular $f_\beta'' > 0$ for $\beta < 1/2$, and for $\beta = 1/2$ if $m \neq 0$. Conversely, if $\beta > 1/2$, f_β is not strictly convex and vanishes at m solution of:

$$2\beta m - \frac{1}{2}\log\Big(\frac{1+m}{1-m}\Big) = 0 \Leftrightarrow m = \tanh(2\beta m). \tag{2.11}$$

$m = 0$ is always a solution, unstable as $f_\beta''(0) < 0$. As $m - \tanh(2\beta m)$ vanishes at 0, is strictly positive at 1 ($\tanh(1) \leq 0.77$) and its derivative changes sign exactly once on $(0,1)$, there is exactly one solution m_β on $(0,1)$, so exactly two solutions on $(-1,1)$ by symmetry. □

Proof of Theorem 2.1. Notice first that m_n takes values in $A_n := \{-1, -1+2/n, ..., 1-2/n, 1\}$. Then:

$$\mu_{\beta,n}^{\mathsf{CW}}(m_n \in J) = \frac{1}{Z_{\beta,n}^{\mathsf{CW}}} \sum_{m \in J \cap A_n} |\{\sigma : m_n(\sigma) = m\}| e^{\beta n m^2}. \tag{2.12}$$

Let $N_\pm$ denote the number of $\pm$ spins respectively. The averaged magnetisation is $m = (N_+ - N_-)/n$ and $N_+ + N_- = n$ always. Thus:

$$\begin{aligned} \mu_{\beta,n}^{\mathsf{CW}}(m_n \in J) &= \frac{1}{Z_{\beta,n}^{\mathsf{CW}}} \sum_{m \in J \cap A_n} \binom{n}{\frac{n(m+1)}{2}} e^{\beta n m^2} \\ &= \Big[\sum_{m \in A_n} \binom{n}{\frac{n(m+1)}{2}} e^{\beta n m^2}\Big]^{-1} \sum_{m \in J \cap A_n} \binom{n}{\frac{n(m+1)}{2}} e^{\beta n m^2}. \end{aligned} \tag{2.13}$$

Note that the sums contains at most n terms. Taking $\frac{1}{n}\log$, we therefore get:

$$\begin{aligned} \frac{1}{n}\log \mu_{\beta,n}^{\mathsf{CW}}(m_n \in J) = &\max_{m \in A_n \cap J} \Big\{\frac{1}{n}\log\binom{n}{\frac{n(m+1)}{2}} + \beta m^2\Big\} \\ &- \max_{m \in A_n} \Big\{\frac{1}{n}\log\binom{n}{\frac{n(m+1)}{2}} + \beta m^2\Big\} + O(\log n/n). \end{aligned} \tag{2.14}$$

The binomial coefficient was computed in Example 1.3:

$$\frac{1}{n}\log\binom{n}{\frac{n(m+1)}{2}} = s(m) + o_n(1). \tag{2.15}$$

Since $m^2 = -e(m)$, we get:

$$\begin{aligned} \frac{1}{n}\log \mu_{\beta,n}^{\mathsf{CW}}(m_n \in J) &= \max_{m \in A_n \cap J}(-f_\beta(m)) - \max_{m \in A_n}(-f_\beta(m)) \\ &= -\min_{m \in A_n \cap J}\big[f_\beta(m) - \min_{m \in A_n} f_\beta(m)\big]. \end{aligned} \tag{2.16}$$

It remains to check that A_n can be replaced by $[-1,1]$ in each minima, which follows from the uniform continuity of f_β. □

3 Sanov's theorem

3.1 Statement

So far we have looked, somewhat arbitrarily, at deviations of the sum $S_n = \sum_{i=1}^n X_i$ of i.i.d. random variables. Cramér's theorem and its proof then show that the best way to see a deviation $\{S_n \approx nm'\}$ for $m' \in \mathbb{R}^d$ is to tilt the law of X_1 in such a way that the resulting law has mean m'.

In this section we will refine the above pictures in two directions:

- by considering deviations of much more general functions of the (X_i) than their sum, for instance deviations of the typical number of X_i that take a prescribed value.
- By obtaining bounds on the probability of such deviations in terms of more general tilts of the law of X_1.

An object that captures the full distribution of $(X_i)_{1\leq i\leq n}$ is the empirical measure:

$$\pi_n(dx) = \pi_n^{(X_i)}(dx) := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}(dx). \tag{3.1}$$

This is a random probability measure on $\mathbb{R}^d$. The normalised sum S_n/n is related to the average value of a certain test function under this measure:

$$\frac{S_n}{n} = \langle \pi_n, \mathrm{id}\rangle := \int x\, d\pi_n(x). \tag{3.2}$$

We can also ask for much more detailed information on the law of the X_i. For instance, assuming that the X_i take discrete values $\{x_1, ..., x_p\}$, we can ask for the typical frequency at which $X_i = a$ for $a \in \mathbb{R}^d$:

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}_a(X_i) = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}(\{a\}) = \pi^n(\{a\}). \tag{3.3}$$

Notice that we know what π_n typically looks like in the large n limit. Indeed, the law of large number for the (X_i) and the dominated convergence theorem give that, for any continuous and bounded function $f : \mathbb{R}^d \to \mathbb{R}$

$$\lim_{n\to\infty}\langle \pi_n, f\rangle = \langle \alpha, f\rangle \qquad a.s., \qquad \alpha := \text{ law}(X_1). \tag{3.4}$$

In other words $(\pi_n)_n$ converges weakly to α for almost every realisation of the sequence $(X_n)_n$.

What about deviations? A theorem by Sanov, stated next, provides a counterpart to Cramér's theorem at the level of the empirical measure. To state it we need some more notations. For two probability measures μ, ν on $\mathbb{R}^d$, write $\mu \ll \nu$ if μ is absolutely continuous with respect to ν:

$$\mu \ll \nu \quad \Leftrightarrow \quad \text{there is } f \geq 0 \text{ and in } \mathbb{L}^1(\nu) \text{ such that } \mu = f\nu. \tag{3.5}$$

The relative entropy $H(\mu|\nu)$ (also: Kullback-Leibler divergence) reads:

$$H(\mu|\nu) := \begin{cases} \displaystyle\int f \log f \, d\nu & \text{if } \mu \ll \nu, \dfrac{d\mu}{d\nu} = f, \\ +\infty & \text{otherwise.} \end{cases} \tag{3.6}$$

Proposition 3.1 (Properties of relative entropy)**.** *The relative entropy on $E = \mathbb{R}^d$ admits the following characterisation:*

$$H(\mu|\pi) := \sup_{\substack{G : E \to \mathbb{R} \\ G \text{ bounded measurable}}} \Big\{ \mathbb{E}_\mu[G] - \log \mathbb{E}_\pi[e^G] \Big\}. \tag{3.7}$$

In addition $H(\cdot|\pi) \geq 0$, is convex, lower semi-continuous with respect to weak convergence and has compact sub-level sets.

Proof. For simplicity we will restrict to the case where the X_i take values in a finite set E. We leave the proof of the general claim as an exercise, see Lemmas 5.9-5.11 in Jan Swart's lecture notes on large deviations, available on his webpage.

Weak convergence of a measure $\mu = (\mu(i))_{i \in E}$ is then equivalent to convergence on $\mathbb{R}^p$, and the supremum (3.7) reads:

$$\tilde{H}(\mu|\pi) := \sup_{h \in \mathbb{R}^{|E|}} \Big\{ \sum_{i \in E} \mu(i) h(i) - \log \sum_{i \in E} \pi(i) e^{h(i)} \Big\}. \tag{3.8}$$

This is a lower semi-continuous, convex function of $\mu = (\mu(i))_{i \in E}$ as a supremum of linear functions, equal to 0 if $h = 0$. Moreover, if $\pi(i) = 0$ while $\mu(i) > 0$ for some i then the supremum blows up, thus $\tilde{H}(\mu|\pi) = +\infty$ unless $\mu \ll \pi$. Write for short S_ν for the support of a probability measure ν. When $\mu \ll \pi$, the supremum reads:

$$\begin{aligned} \tilde{H}(\mu|\pi) &= \sup_{h \in \mathbb{R}^{S_\pi}} \Big\{ \sum_{i \in S_\mu} \mu(i) h(i) - \log \sum_{i \in E} \pi(i) e^{h(i)} \Big\} \\ &= \sup_{g \in \mathbb{R}^{S_\mu}} \sup_{h \in \mathbb{R}^{S_\pi \setminus S_\mu}} \Big\{ \sum_{i \in S_\mu} \mu(i) h(i) - \log \Big(\sum_{i \in S_\mu} \pi(i) e^{h(i)} + \sum_{i \in S_\pi \setminus S_\mu} \pi(i) e^{g(i)} \Big) \Big\}. \end{aligned} \tag{3.9}$$

Fix $g \in \mathbb{R}^{S_\mu}$. As log is increasing, the supremum on h is reached when $g(i) = -\infty$ for each $i \in S_\pi \setminus S_\mu$. Thus:

$$\tilde{H}(\mu|\pi) = \sup_{h \in \mathbb{R}^{S_\mu}} \Big\{ \sum_{i \in S_\mu} \mu(i) h(i) - \log \sum_{i \in S_\mu} \pi(i) e^{h(i)} \Big\}. \tag{3.10}$$

The fonction in the supremum is strictly concave. If it admits a critical point then it must be unique and give the location of the global maximum. Critical points solve:

$$\mu(i) = \frac{e^{h(i)} \pi(i)}{\sum_{j \in E} \pi(j) e^{h(j)}}, \qquad i \in S_\mu. \tag{3.11}$$

This is equivalent to:

$$h(i) = \log\Big(\frac{\mu(i)}{\pi(i)} \Big), \qquad i \in S_\mu. \tag{3.12}$$

As a result $\tilde{H}(\mu|\pi)$ satisfies:

$$\tilde{H}(\mu|\pi) = \sum_{i\in\mathrm{supp}(\pi)} \mu(i)\log\Big(\frac{\mu(i)}{\pi(i)}\Big) = H(\mu|\pi). \tag{3.13}$$

□

We henceforth restrict to the case where the X_i take values in a finite set E. In this setting weak convergence of probability measures is equivalent to convergence in total variation distance, i.e. convergence for the metric:

$$d(\mu,\nu) = \frac{1}{2}\sum_{a\in E}|\mu(a)-\nu(a)|. \tag{3.14}$$

In addition, we can assume without loss of generality that $\alpha_k = \mathbb{P}(X_1 = k) > 0$ for each $k\in E$. Thus every probability measure on E is absolutely continuous with respect to α, so that the relative entropy reads:

$$H(\mu|\alpha) = \sum_{k\in E}\mu_k\log\Big(\frac{\mu_k}{\alpha_k}\Big), \qquad \mu \text{ probability measure on } E. \tag{3.15}$$

Theorem 3.2 (due to Sanov if the state space is $\mathbb{R}$)**.** *For a probability measure μ on E and $\varepsilon > 0$, let $B(\mu,\varepsilon)$ denote the open ball of radius ε around μ for the distance d and $\bar{B}(\mu,\varepsilon)$ denote its closure. Then:*

$$\mathbb{P}\big(\pi_n \in \bar{B}(\mu,\varepsilon)\big) \asymp \exp\Big[-\inf_{\nu\in\bar{B}(\mu,\varepsilon)} H(\nu|\alpha)\Big]. \tag{3.16}$$

Remark 3.3.

- *Contrary to the variable S_n/n which can only deviate from its typical value by a real number, there are many ways in which π_n can differ from α. Correspondingly, the rate function now takes as argument a measure rather than a scalar.*

  This is reflected in the characterisation of $H(\cdot|\alpha)$ as a generalised Legendre transform:

$$\sup_{\lambda\in\mathbb{R}} \to \sup_{\substack{G:\mathbb{R}^d\to\mathbb{R}\\ G \text{ bounded measurable}}}, \qquad \lambda x \to \int G\,d\mu, \qquad \phi(\lambda)\to\log\mathbb{E}_\alpha[e^G]. \tag{3.17}$$

- *The theorem is valid on general spaces, for instance any separable metric space, with the following modifications. 1) One should take balls with respect to weak convergence rather than than total variation distance. 2) The lower bound on the probability is given in terms of $\inf_{B(\mu,\varepsilon)} H(\cdot|\alpha)$, rather than $\inf_{\bar{B}(\mu,\varepsilon)} H(\cdot|\alpha)$. This difference between upper and lower bound is a general fact of large deviations to which we will come back later.*

Proof of Theorem 3.2. As we only prove the theorem for empirical measures of random variables taking values in a discrete space E, we can perform explicit computations. Write $E = \{1,...,p\}$ with $\min_E \alpha > 0$ without loss of generality, and note that probability measures on E are elements of the simplex S_p:

$$S_p := \Big\{(s_1,...,s_p)\in[0,1]^p : \sum_{i=1}^p s_i = 1\Big\}. \tag{3.18}$$

The empirical measure π_n is an element of a subset of the simplex:

$$\pi_n \in S_p^n := S_p \cap \frac{1}{n}\mathbb{N}^p = \Big\{(a_1, ..., a_p) \in \mathbb{N}^p : \sum_{i=1}^p a_i = n\Big\}. \tag{3.19}$$

Then, for any $a \in S_p^n$, one can compute explicitly:

$$\begin{aligned}\mathbb{P}\Big(\pi_n = \frac{a}{n}\Big) &= \mathbb{P}\Big(|\{i : X_i = 1\}| = a_1, \dots, |\{i : X_i = p\}| = a_p\Big) \\ &= \binom{n}{a_1}\mathbb{P}(X_1 = 1)^{a_1}\binom{n-a_1}{a_2}\mathbb{P}(X_1 = 2)^{a_2}...\binom{n-a_1-...-a_{p-1}}{a_p}\mathbb{P}(X_1 = p)^{a_p} \\ &= \prod_{k=1}^p \frac{\alpha_k^{a_k}(n-\sum_{\ell=1}^{k-1} a_\ell)!}{(n-\sum_{\ell=1}^{k} a_\ell)!a_k!} = n!\prod_{k=1}^p \frac{\alpha_k^{a_k}}{a_k!},\end{aligned} \tag{3.20}$$

where we recall that α is the law of X_1. One can easily check by recursion that $q^q e^{-q} \le q! \le eqq^q e^{-q}$. As a result,

$$\begin{aligned}\frac{1}{n}\log\mathbb{P}\Big(\pi_n = \frac{a}{n}\Big) &= \log(n/e) + \frac{1}{n}\sum_{k=1}^p a_k \log\Big(\frac{\alpha_k e}{a_k}\Big) + O(\log n/n) \\ &= \log(n) + \frac{1}{n}\sum_{k=1}^p a_k \log\Big(\frac{\alpha_k}{a_k}\Big) + O(\log n/n),\end{aligned} \tag{3.21}$$

where all error terms are uniform in a. Rewriting the right-hand side in terms of a/n makes the relative entropy appear:

$$\begin{aligned}\frac{1}{n}\log\mathbb{P}\Big(\pi_n = \frac{a}{n}\Big) &= -\sum_{k=1}^p \frac{a_k}{n}\log\Big(\frac{a_k/n}{\alpha_k}\Big) + O(\log n/n) \\ &= -H\Big(\frac{a}{n}\Big|\alpha\Big) + O(\log n/n).\end{aligned} \tag{3.22}$$

The fact that the number of elements in S_p^n is bounded by n^p gives us a first bound:

$$\frac{1}{n}\log\mathbb{P}\Big(\pi_n \in \bar{B}(\mu,\varepsilon)\Big) \le \max_{\frac{a}{n}\in\bar{B}(\mu,\varepsilon)}\Big[-H\Big(\frac{a}{n}\Big|\alpha\Big) + O(\log n/n)\Big]. \tag{3.23}$$

The lower bound is more direct:

$$\begin{aligned}\frac{1}{n}\log\mathbb{P}\Big(\pi_n \in \bar{B}(\mu,\varepsilon)\Big) &\ge \max_{\frac{a}{n}\in\bar{B}(\mu,\varepsilon)}\frac{1}{n}\log\mathbb{P}\Big(\pi_n = a/n\Big) \\ &= -\min_{\frac{a}{n}\in\bar{B}(\mu,\varepsilon)} H\Big(\frac{a}{n}\Big|\alpha\Big) + O(\log n/n).\end{aligned} \tag{3.24}$$

We have thus shown that $\mathbb{P}(\pi_n \in \bar{B}(\mu,\varepsilon)) \asymp \exp[-n\min_{\nu\in S_p^n\cap\bar{B}(\mu,\varepsilon)} H(\nu|\alpha)]$. It remains to replace this minimum with an infimum over all probability measures ν on E up to an $o_n(1)$ error, i.e. to prove:

$$\lim_{n\to\infty}\min_{S_p^n\cap\bar{B}(\mu,\varepsilon)} H(\cdot|\alpha) = \inf_{\bar{B}(\mu,\varepsilon)} H(\cdot|\alpha). \tag{3.25}$$

Suppose first that $\inf_{\bar{B}(\mu,\varepsilon)} H(\cdot|\alpha) = +\infty$. Then $\inf_{\bar{B}(\mu,\varepsilon)\cap S_p^n} H(\cdot|\alpha) = +\infty$ and the claim holds in this case. Suppose now that $\inf_{\bar{B}(\mu,\varepsilon)} H(\cdot|\alpha) < \infty$, and let $\delta > 0$ and ν be such that:

$$\inf_{\bar{B}(\mu,\varepsilon)} H(\cdot|\alpha) \geq H(\nu|\alpha) - \delta. \tag{3.26}$$

If we prove that one can find a sequence (a_n/n) with $a_n \in S_p^n$ such that $\lim_{n\to\infty} d(a_n/n, \nu) = 0$ and $\lim_{n\to\infty} H(a_n/n|\alpha) = H(\nu|\alpha)$, then for large enough n we will have:

$$\inf_{\bar{B}(\mu,\varepsilon)} H(\cdot|\alpha) \geq H(a_n/n|\alpha) - \delta/2 \geq \min_{S_p^n\cap\bar{B}(\mu,\varepsilon)} H(\cdot|\alpha) - \delta/2, \tag{3.27}$$

which is the claim since $\delta > 0$ is arbitrary. It is easy to find a sequence (a_n/n) approximating ν. Indeed, set $a_n(k) = \lfloor n\nu(k)\rfloor/Z_n$ for $k \in E$ with Z_n a normalisation:

$$Z_n = \sum_{k\in E} \lfloor n\nu(k)\rfloor. \tag{3.28}$$

We assume n is large enough to have $Z_n > 0$ so that a_n is a well defined element of S_p^n. By construction it satisfies $\lim_{n\to\infty} d(a_n/n, \nu) = 0$, and $a_n \ll \alpha$ is always true since α has full support. The relative entropy therefore reads:

$$H\Big(\frac{a_n}{n}\Big|\alpha\Big) = \sum_{k\in E} \frac{a_n(k)}{n} \log\Big(\frac{a_n(k)}{n}\frac{1}{\alpha_k}\Big). \tag{3.29}$$

The above expression converges to $H(\nu|\alpha)$, which concludes the proof. □

3.2 Around Sanov's theorem

We have proven Sanov's theorem by explicit computations. Let us see how to prove it using the same kind of idea as in Cramér's theorem: to observe a given deviation of π_n, one should tilt the law of π_n so that this deviation becomes typical and estimate the cost of tilting.

Proposition 3.4 (Sanov's theorem by tilting). *Let μ be a probability measure on E. Then:*

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\pi_n \in \bar{B}(\mu,\varepsilon)\big) = -H(\mu|\alpha). \tag{3.30}$$

Proof. Assume for the moment that $\min_{k\in E}\mu_k > 0$ and let $\varepsilon \in (0, \min_E \mu/2)$. Then:

$$\begin{aligned}\mathbb{P}\big(\pi_n \in \bar{B}(\mu,\varepsilon)\big) &= \int_{E^n} \mathbf{1}_{\pi_n\in\bar{B}(\mu,\varepsilon)}\alpha^{\otimes n}(dx) = \int_{E^n} \mathbf{1}_{\pi_n\in\bar{B}(\mu,\varepsilon)} \prod_{i=1}^n \frac{\alpha(x_i)}{\mu(x_i)}\mu^{\otimes n}(dx)\\ &= \int_{E^n} \mathbf{1}_{\pi_n\in\bar{B}(\mu,\varepsilon)} \prod_{k=1}^p \Big(\frac{\alpha_k}{\mu_k}\Big)^{n_k(x)}\mu^{\otimes n}(dx).\end{aligned} \tag{3.31}$$

where $n_k(x) = n\pi_n(k)$ is the number of variables equal to k. If $\pi_n \in \bar{B}(\mu,\varepsilon)$, one must have:

$$\forall k \in E, \qquad |\pi_n(k) - \mu(k)| \leq \varepsilon, \tag{3.32}$$

so that:

$$\mathbb{P}\big(\pi_n \in \bar{B}(\mu,\varepsilon)\big) \geq \exp\Big[n\sum_{k=1}^{p}\mu_k \log\Big(\frac{\alpha_k}{\mu_k}\Big) - n\varepsilon\sum_{k=1}^{p}\Big|\log\Big(\frac{\alpha_k}{\mu_k}\Big)\Big|\Big]\mathbb{P}_{\mu^{\otimes n}}\big(\pi_n \in \bar{B}(\mu,\varepsilon)\big)$$
$$\mathbb{P}\big(\pi_n \in \bar{B}(\mu,\varepsilon)\big) \leq \exp\Big[n\sum_{k=1}^{p}\mu_k \log\Big(\frac{\alpha_k}{\mu_k}\Big) + n\varepsilon\sum_{k=1}^{p}\Big|\log\Big(\frac{\alpha_k}{\mu_k}\Big)\Big|\Big]\mathbb{P}_{\mu^{\otimes n}}\big(\pi_n \in \bar{B}(\mu,\varepsilon)\big). \quad (3.33)$$

Note that the first sum in the exponential is precisely $-H(\mu|\alpha)$.

We know that the event $\pi_n \in \bar{B}(\mu,\varepsilon)$ is typical under $\mu^{\otimes n}$:

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}_{\mu^{\otimes n}}\big(\pi_n \in \bar{B}(\mu,\varepsilon)\big) = 0. \quad (3.34)$$

Taking $\frac{1}{n}\log$ and the large n limit, we thus get, for some $C > 0$:

$$-H(\mu|\alpha) - C\varepsilon \leq \liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}\big(\pi_n \in \bar{B}(\mu,\varepsilon)\big)$$
$$\leq \limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}\big(\pi_n \in \bar{B}(\mu,\varepsilon)\big) \leq -H(\mu|\alpha) + C\varepsilon. \quad (3.35)$$

Taking $\varepsilon \to 0$ thus gives the claim in the case where $\min_E \mu > 0$.

If $\min_E \mu = 0$, let $\mu^* = \min\{\mu(k) : \mu(k) \neq 0\}$ and, for $\varepsilon \in (0,1)$, approximate μ by $\mu_\varepsilon \in \bar{B}(\mu,\varepsilon)$ with $\mu_\varepsilon(k) = \mu^*\varepsilon/(p - |\mathrm{supp}(\mu)|)$ whenever $\mu_k = 0$, and $\mu_\varepsilon(k) = \mu(k) - \mu^*\varepsilon/|\mathrm{supp}(\mu)|$ otherwise (recall $p = |E|$). Then $\min_E \mu_\varepsilon > 0$, μ_ε is a probability measure on E and by continuity of the entropy:

$$\lim_{\varepsilon\to 0}\mu_\varepsilon = \mu, \qquad \lim_{\varepsilon\to 0}H(\mu_\varepsilon|\alpha) = H(\mu|\alpha). \quad (3.36)$$

This concludes the proof. □

As observed above,

$$\frac{S_n}{n} := \frac{1}{n}\sum_{i=1}^{n}X_i = \langle \pi_n, \mathrm{id}\rangle = \int x\, d\pi_n(x). \quad (3.37)$$

It therefore looks like Sanov's theorem is more general than Cramér's and should imply it. This is indeed the case, through what is known as a contraction principle. Contraction principles are a general way to obtain new large deviation principles from existing ones that we will later describe in an abstract way. Let us prove Sanov ⇒ Cramér when random variables only take a finite number of values; the proof is similar in the general case.

Proposition 3.5 (Contraction principle, first example)**.** *Let $\varepsilon > 0$. Then:*

$$\mathbb{P}\big(|S_n/n - \mathbb{E}[X_1]| \geq \varepsilon\big) \asymp \exp\big[-n \inf_{(\mathbb{E}[X_1]-\varepsilon,\mathbb{E}[X_1]+\varepsilon)^c} I\big], \quad (3.38)$$

with:

$$I : y \in \mathbb{R} \mapsto \inf_{\nu:\langle\nu,x\rangle = y} H(\nu|\alpha). \quad (3.39)$$

Proof. For a probability measure ν, $m_\nu := \int x\, d\nu(x)$ for its mean. In particular $\mathbb{E}[X_1] = m_\alpha$. Since $S_n/n = m_{\pi_n}$, we have:

$$\Big\{\Big|\frac{S_n}{n} - \mathbb{E}[X_1]\Big| \geq \varepsilon\Big\} = \big\{\pi_n \notin C(\varepsilon)\big\}, \qquad C(\varepsilon) := \{\nu : |m_\nu - m_\alpha| < \varepsilon\}. \tag{3.40}$$

Since $C(\varepsilon)^c$ is a closed set for the total variation distance, we can apply Sanov's theorem to find:

$$\mathbb{P}\big(|S_n/n - \mathbb{E}[X_1]| \geq \varepsilon\big) \asymp \exp\big[-n \inf_{C(\varepsilon)^c} H(\nu|\alpha)\big]. \tag{3.41}$$

As $\inf_{C(\varepsilon)^c} H(\nu|\alpha) = \inf_{(m_\alpha-\varepsilon, m_\alpha+\varepsilon)} I$ by definition, the lemma is proven. □

Exercise 3.6. *Check that the rate function I in Proposition 3.5 indeed corresponds to the one obtained in Cramér's theorem.*

Proof. By definition the relative entropy reads:

$$H(\nu|\alpha) = \sup_{h\in\mathbb{R}^E}\Big\{(\nu,h) - \log\sum_{e\in E}\alpha(e)e^{h_e}\Big\} \geq \sup_{\lambda\in\mathbb{R}}\Big\{\lambda\sum_{e\in E}\nu(e)e - \phi(\lambda)\Big\}, \tag{3.42}$$

where $\phi(\lambda) = \log\mathbb{E}[e^{\lambda X_1}] = \log\sum_{e\in E}\alpha(e)e^{\lambda e}$ and we chose specific vectors $h = (\lambda e)_{e\in\mathbb{E}}$ ($\lambda\in\mathbb{R}$). Thus:

$$\inf_{\nu: m_\nu = y} H(\nu|\alpha) \geq \sup_{\lambda\in\mathbb{R}}\Big\{\lambda y - \phi(\lambda)\Big\} = I_{\mathrm{Cramer}}(y). \tag{3.43}$$

It is therefore enough to prove that there is equality for a good choice of measure, or that both sides are equal to $-\infty$.

Write $E = \{e_1, ..., e_p\}$ with $e_1 \leq ... \leq e_p$. If $m_y \notin [e_1, e_p]$, then $\phi(\lambda) \in [\lambda e_1, \lambda e_p]$ (the boundaries are not necessarily ordered) implies $I_{\mathrm{Cramer}}(y) = +\infty$, therefore there is equality in this case.

Assume now $m_y \in [e_1, e_p]$. Then we claim that the supremum defining $I_{\mathrm{Cramer}}(y)$ is reached (possibly at $\pm\infty$) as we already saw in the Cramér case. Indeed, $M_y(\lambda) := e^{-\lambda y}\mathbb{E}[e^{\lambda X_1}]$ satisfies (recall α has full support):

$$M_y'(\lambda) = \mathbb{E}[(X_1 - y)e^{\lambda(X_1-y)}], \qquad M_y''(\lambda) = \mathbb{E}[(X_1-y)^2 e^{\lambda(X_1-y)}] > 0, \tag{3.44}$$

thus M_y'' is strictly convex and:

$$\lim_{\lambda\to\pm\infty} M_y'(\lambda) = \pm\infty \quad \text{ if } y \in (e_1, e_p). \tag{3.45}$$

This implies that for $y \in (e_1, e_p)$ M_y admits a unique global minimiser, therefore $\lambda \mapsto \lambda y - \phi(\lambda)$ has a unique global maximiser, call it $\lambda_y \in \mathbb{R}$, and:

$$I_{\mathrm{Cramer}}(y) = \lambda_y y - \phi(\lambda_y). \tag{3.46}$$

On the other hand if $y \in \{e_1, e_p\}$ then $-\log M_y(\lambda)$ is maximal at $\lambda = -\infty$ for e_1, $+\infty$ for e_2:

$$I_{\mathrm{Cramer}}(e_1) = \lim_{\lambda\to-\infty}\big\{\lambda y - \phi(\lambda)\big\}, \qquad I_{\mathrm{Cramer}}(e_p) = \lim_{\lambda\to\infty}\big\{\lambda y - \phi(\lambda)\big\}. \tag{3.47}$$

Write by extension $\lambda_{e_1} = -\infty$, $\lambda_{e_p} = +\infty$. By definition, as α has full support:

$$H(\nu|\alpha) = \sum_{i \in E} \nu(i) \log\left(\frac{\nu(i)}{\alpha(i)}\right) \in [0, \infty). \tag{3.48}$$

Let α^λ denote the tilted probability measure $\propto e^{\lambda_y e}\alpha(de)$ if $y \in (e_1, e_p)$. If $y \in \{e_1, e_p\}$ this is by extension the Dirac measure on e_1, e_p respectively. Then:

$$H(\alpha^{\lambda_y}|\alpha) = \begin{cases} \lambda_y y - \log \sum_{e \in E} \alpha(e) e^{\lambda_y e} = I_{\text{Cramer}}(y) & \text{if } y \in (e_1, e_p), \\ -\log(\alpha(e_1)) = I_{\text{Cramer}}(y) & \text{if } y = e_1, \\ -\log(\alpha(e_p)) = I_{\text{Cramer}}(y) & \text{if } y = e_p. \end{cases} \tag{3.49}$$

□

4 Abstract large deviation theory

In this section we give an abstract definition of large deviation principles and establish general properties. We work throughout with a polish space E, i.e. a complete and separable metric space, which we equip with its Borel σ-algebra $\mathcal{B}$. For a set $B \subset X$, its closure is denoted $\bar{B}$ and its interior B^o.

4.1 Large deviation principle

Definition 4.1 (Rate function). *A mapping $I : E \to [0, \infty]$ is said to be a* rate function *if I is lower semi-continuous, i.e. I has closed sub-level sets: $\{I \le \alpha\}$ is closed in E for each $\alpha \ge 0$.*

I is said to be a good *rate function if I has compact sub-level sets.*

Definition 4.2 (Large deviation principle). *Let $a_n > 0$ ($n \in \mathbb{N}$) be a diverging sequence. A sequence $(\mu_n)_{n \in \mathbb{N}}$ of probability measures on E is said to satisfy a large deviation principle with speed a_n and rate function I if I is a rate function in the sense of Definition 4.1 and the following holds. For every closed set C and every open set O,*

$$\begin{aligned} \limsup_{n \to \infty} \frac{1}{a_n} \log \mu_n(C) &\le -\inf_C I, \\ \liminf_{n \to \infty} \frac{1}{a_n} \log \mu_n(O) &\ge -\inf_O I. \end{aligned} \tag{4.1}$$

Remark 4.3.

- *One obtains an equivalent definition by asking that, for any Borel set B:*

$$-\inf_{B^o} I \le \liminf_{n \to \infty} \frac{1}{a_n} \log \mu_n(B) \le \limsup_{n \to \infty} \frac{1}{a_n} \log \mu_n(B) \le -\inf_{\bar{B}} I. \tag{4.2}$$

- *Since $\mu_n(E) = 1$ for each $n \in \mathbb{N}$, it must be that* $\inf I = 0$. *In particular, a rate function cannot be identically equal to $+\infty$. If I is a good rate function, then this implies that there is $x \in E$ with $I(x) = 0$. Without the goodness assumption this may not be true.*

- *For a measurable set B, in general:*

$$\lim_{n\to\infty} \frac{1}{n} \log \mu_n(B) \neq - \inf_B I. \tag{4.3}$$

  This is true for any set such that $\inf_{\bar{B}} I = \inf_{B^o} I$ (in which case both infima are equal to $\inf_B I$).

  In particular, if C is not open, one cannot have the lower bound with $\inf_C I$ instead of $\inf_{C^o} I$ in general. Any sequence of non-atomic measures gives a counterexample: as $\mu_n(\{x\}) = 0$ for each $x \in E$, one would have $I(x) = +\infty$ for each $x \in E$ which is impossible by the first item.

- *So far, all large deviation principles we have encountered had speed n, the number of variables under consideration. This is always the case in i.i.d settings (with suitable exponential moments assumption), but may fail for independent, non-identically distributed random variables or in the presence of interaction. For instance the magnetisation $\frac{1}{n^d}\sum_{i\in\{0,\dots,n\}^d} \sigma_i$ of the $d \geq 2$ Ising model at low temperature has deviations of order $e^{-cn^{d-1}}$ in a certain range of values.*

Lemma 4.4. *If (μ_n) satisfies two large deviation principles with the same speed a_n and rate functions I_1, I_2, then $I_1 = I_2$.*

Proof. Let $x \in E$, $\varepsilon > 0$ and consider the open ball $B(x,\varepsilon)$ around x. Then:

$$\begin{aligned} -I_1(x) \leq - \inf_{B(x,\varepsilon)} I_1 &\leq \liminf_{n\to\infty} \frac{1}{a_n} \log \mu_n(B(x,\varepsilon)) \\ &\leq \limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n\big(\overline{B(x,\varepsilon)}\big) \leq - \inf_{\overline{B(x,\varepsilon)}} I_2. \end{aligned} \tag{4.4}$$

The lower semi-continuity implies (exercise) that $\liminf_{\varepsilon\to 0} \inf_{\bar{B}(x,\varepsilon)} I_i \geq I_i(x)$ for $i \in \{1,2\}$ (in fact equality holds). Thus $I_1(x) \geq I_2(x)$ for all $x \in E$. Exchanging the roles of I_1, I_2 concludes the proof. □

4.2 Varadhan's lemma

Under suitable conditions, there is an equivalence between large deviation principles and control of exponential moments of a sufficiently large class of functions on E. Here, we will only state one direction, known as Varadhan's lemma (or the Laplace-Varadhan lemma).

Proposition 4.5. *Suppose μ_n satisfies a large deviation principle with speed a_n and rate function I (not necessarily good). Then, for every continuous function $F : X \to \mathbb{R}$ that is bounded from above,*

$$\lim_{n\to\infty} \frac{1}{a_n} \log \mathbb{E}_{\mu_n}\big[e^{a_n F}\big] = \sup_E \big\{F - I\big\}. \tag{4.5}$$

Remark 4.6.

- *If we imagine that μ_n is a measure on $\mathbb{R}$, say, with distribution given by $e^{-a_n I(\cdot)}$, then the above lemma turns into a generalisation of the Laplace principle:*

$$\lim_{n\to\infty} \frac{1}{a_n} \log \int_{\mathbb{R}} e^{a_n (F-I)(x)}\, dx = \sup_{\mathbb{R}} \{F - I\}. \tag{4.6}$$

- *To prove Cramér's theorem we deduced large deviations from a bound on all* $\mathbb{E}[e^{n\lambda(S_n/n)}]$, $\lambda \in \mathbb{R}$. *Writing* $S_n/n = \langle \pi_n, x\rangle$ *with* π_n *the empirical measure, this corresponds to* $F(\pi) = \langle \pi, x\rangle$, *which is not bounded above on the space of probability measures on* $\mathbb{R}$. *Varadhan's lemma can in fact be stated with much less stringent conditions on* F *than being bounded above, such as suitable exponential moment bounds. See the book by Dembo and Zeitouni, Theorem 4.3.1 for more on this topic.*

Proof. Consider first the lower bound. Let $x \in E$ and define $I_x(\varepsilon) := (F(x)-\varepsilon, F(x)+\varepsilon)$. The set $F^{-1}(I_x(\varepsilon))$ is open by continuity of F. The continuity of F implies that there is $\omega_x(\cdot) \geq 0$ with $\lim_{\varepsilon\to 0}\omega_x(\varepsilon) = 0$ such that, on $F^{-1}(I_x(\varepsilon))$, it holds that $F \geq F(x) - \omega_x(\varepsilon)$. The large deviation lower bound then gives:

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{a_n}\log \mathbb{E}_{\mu_n}\big[e^{a_n F}\big] &\geq \liminf_{n\to\infty}\frac{1}{a_n}\log\mathbb{E}_{\mu_n}\big[\mathbf{1}_{F^{-1}(I_x(\varepsilon))}e^{a_nF}\big] \\ &\geq F(x) - \omega_x(\varepsilon) - \inf_{F^{-1}(I_x(\varepsilon))} I \geq F(x) - \omega_x(\varepsilon) - I(x). \end{aligned} \tag{4.7}$$

As ε was arbitrary, we obtain the lower bound by taking ε to 0, then the supremum in x.

Consider next the upper bound. We decompose E into sets where F is approximately constant and use the large deviation upper bound on all these sets. Write $s = \sup_E F$ and $S = \sup_E\{F - I\}$. As F is bounded above and $I \geq 0$, we have $-\infty < S \leq s < \infty$, so we only need to care about the set $F^{-1}([S,s])$:

$$\limsup_{n\to\infty}\frac{1}{a_n}\log\mathbb{E}_{\mu_n}\big[e^{a_nF}\big] \leq \max\Big\{S, \limsup_{n\to\infty}\frac{1}{a_n}\log\mathbb{E}_{\mu_n}\big[\mathbf{1}_{F^{-1}([S,s])}e^{a_nF}\big]\Big\}. \tag{4.8}$$

Let $p \in \mathbb{N}\setminus\{0\}$ and define:

$$x_k := S + \frac{k(s-S)}{p}, \qquad k \in \{0,...,p\}. \tag{4.9}$$

Define $J_k = F^{-1}([x_k, x_{k+1}])$. Then $F^{-1}([S,s]) = \bigcup_{k=0}^{p-1} J_k$, and:

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{a_n}\log\mathbb{E}_{\mu_n}\big[\mathbf{1}_{F^{-1}([S,s])}e^{a_nF}\big] &\leq \max_{0\leq k\leq p-1}\limsup_{n\to\infty}\frac{1}{a_n}\log\mathbb{E}_{\mu_n}\big[\mathbf{1}_{J_k}e^{a_nF}\big] \\ &\leq \max_{0\leq k\leq p-1}\Big\{x_{k+1} + \limsup_{n\to\infty}\frac{1}{a_n}\log\mu_n(J_k)\Big\} \\ &\leq \max_{0\leq k\leq p-1}\Big\{x_{k+1} - \inf_{J_k} I\Big\}, \end{aligned} \tag{4.10}$$

since J_k is a closed set for each k. By definition of J_k, $x_{k+1} \leq \inf_{J_k} F + (s-S)/p$, thus:

$$\limsup_{n\to\infty}\frac{1}{a_n}\log\mathbb{E}_{\mu_n}\big[\mathbf{1}_{F^{-1}([S,s])}e^{a_nF}\big] \leq \max_{0\leq k\leq p-1}\Big\{\inf_{J_k}F - \inf_{J_k} I\Big\} + \frac{s-S}{p}. \tag{4.11}$$

It remains to notice that $\inf_{J_k} F - \inf_{J_k} I \leq \sup_{J_k}(F - I)$. Indeed, if $y_\varepsilon \in J_k$ satisfies $I(y_\varepsilon) \leq \inf_{J_k} I + \varepsilon$, then:

$$\inf_{J_k} F - \inf_{J_k} I \leq F(y_\varepsilon) - \inf_{J_k} I \leq F(y_\varepsilon) - I(y_\varepsilon) + \varepsilon \leq \sup_{J_k}\{F - I\} + \varepsilon, \tag{4.12}$$

and ε was arbitrary. At this point we have proven:

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{E}_{\mu_n}\big[e^{a_n F}\big] \leq \max\Big\{S, \max_{0\leq k\leq p-1} \sup_{J_k}\{I-F\} + \frac{s-S}{p}\Big\} \leq S + \frac{s-S}{p}. \tag{4.13}$$

Since p was arbitrary, this gives the desired upper bound and the claim. □

4.3 Generating new large deviation principles from known ones

4.3.1 Exponential tilts

Varadhan's lemma gives us a way to obtain new large deviation principles from existing ones as follows.

Proposition 4.7. *Suppose (μ_n) satisfies a large deviation principle with speed a_n and rate function I, and let F be a continuous function that is bounded from above. Define:*

$$\mu_n^F := \Big(\mathbb{E}_{\mu_n}\big[e^{a_n F}\big]\Big)^{-1} e^{a_n F}\mu_n. \tag{4.14}$$

Then (μ_n^F) satisfies a large deviation principle with speed a_n and rate function $I-F+\sup\{F-I\}$

Proof. The proof is directly adapted from the proof of Varadhan's lemma: one needs to compute $\int_B e^{a_n F} d\mu_n$ for suitable Borel sets B. It is left as an exercise. □

Proposition 4.7 gives access to large deviations of variables beyond the i.i.d setting. In particular one can use it to reprove Theorem 2.1 on the Curie-Weiss model, as well as the following generalisation to particles in $\mathbb{R}^d$ with mean-field interactions.

Exercise 4.8. *Let $V:\mathbb{R}^d\to\mathbb{R}$ be such that e^{-V} is Lebesgue-integrable, and let $W:(\mathbb{R}^d)^2\to\mathbb{R}$ be bounded. Consider the following probability measures on $(R^d)^n$:*

$$\mu_n(dx) \propto \exp\Big[-\frac{1}{2n}\sum_{i,j=1}^n W(x_i,x_j) - \sum_{i=1}^n V(x_i)\Big]\prod_{i=1}^n dx_i. \tag{4.15}$$

Prove that μ_n satisfies a large deviation principle with rate function:

$$I(\mu) = \mathcal{F}(\mu) - \inf_{\mathcal{M}_1(\mathbb{R}^d)} \mathcal{F}, \tag{4.16}$$

where $\mathcal{M}_1(\mathbb{R}^d)$ is the set of probability measures on $\mathbb{R}^d$ and the free energy $\mathcal{F}$ satisfies:

$$\mathcal{F}(\mu) := \int_{\mathbb{R}^d} f(x)\log f(x)\,dx + \int_{\mathbb{R}^d} V(x)\,\mu(dx) + \frac{1}{2}\int_{(\mathbb{R}^d)^2} W(x,y)\,\mu^{\otimes 2}(dx,dy),$$
$$\text{if } \mu \ll dx \text{ has density } f, \tag{4.17}$$

and $\mathcal{F}(\mu)=+\infty$ otherwise. You may admit that Sanov's theorem holds on $\mathbb{R}^d$.

4.3.2 Contraction principle

Another way of generating new large deviation principles from existing ones are contraction principles, which we have already encountered in Proposition 3.5.

Proposition 4.9. *Let (μ_n) satisfy a large deviation principle with speed a_n and good rate function I. Let E' be a Polish space equipped with its Borel σ-algebra, and let $T : E \to E'$ be a continuous mapping. Then the sequence of push-forwards $(\mu'_n) := (\mu_n \circ T^{-1})$ satisfies a large deviation principle with speed a_n and good rate function:*

$$I'(y) := \inf_{x \in E : T(x) = y} I(x), \qquad y \in E'. \tag{4.18}$$

Proof. Let C, O be closed, open sets in E' respectively. By definition,

$$\mu'_n(C) = \mu_n(T^{-1}(C)), \qquad \mu'_n(O) = \mu_n(T^{-1}(O)). \tag{4.19}$$

Since T is continuous, $T^{-1}(C)$ is closed, $T^{-1}(O)$ is open and both are measurable sets in E. The large deviation principle for μ_n then gives:

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{a_n} \log \mu'_n(C) &\leq -\inf_{T^{-1}(C)} I = -\big\{I(x) : x \in E, T(x) \in C\big\}\\ &= -\inf \bigcup_{y \in C} \big\{I(x) : x \in E, T(x) = y\big\}\\ &= -\inf_{y \in C} \inf \big\{I(x) : x \in E, T(x) = y\big\} = -\inf_C I'. \end{aligned} \tag{4.20}$$

The same argument gives the lower bound.

Let us now check that I' is a good rate function. As $I \geq 0$ the same is true for I'. It is therefore enough to check that I' has compact sub-level sets. Let $a \geq 0$ and $n \geq 1$. If $y \in \{I' \leq a\}$, then $\inf_{T^{-1}(y)} I \leq a < \infty$. Take then $x^y_n \in T^{-1}(\{y\})$ such that:

$$I(x^y_n) \leq \inf_{T^{-1}(y)} I + \frac{1}{n} \leq a + \frac{1}{n}. \tag{4.21}$$

As I is a good rate function x^y_n converges in E to some x^y, and $x^y \in T^{-1}(\{y\})$ because this set is closed. lower semi-continuity of I. The infimum of I on $T^{-1}(\{y\})$ is thus reached at x^y, and:

$$\begin{aligned}\big\{y \in E' : I'(y) \leq a\big\} &= \big\{y \in E' : \text{ there is } x_y \in E \text{ with } T(x_y) = y \text{ and } I(x_y) \leq a\big\}\\ &= \big\{y \in E' : \text{ there is } x \in \{I \leq a\} \text{ with } y = T(x)\big\}\\ &= T(\big\{I \leq a\}). \end{aligned} \tag{4.22}$$

Since $\{I \leq a\}$ is compact and T is continuous, $\{I' \leq a\}$ is also compact. □

Remark 4.10. *If I is a rate function but not a good rate function, then it may be that I' is not even a rate function. Consider e.g. $E = E' = \mathbb{R}$, $I = 0$ and $T(x) = e^x$. Then $I' = +\infty$ on $(-\infty, 0]$ and $I' = 0$ on $(0, \infty)$. In particular $\{I' = 0\}$ is not closed.*

5 Gärtner-Ellis theorem

We now consider large deviations of sequences $(S_n) \in (\mathbb{R}^d)^{\mathbb{N}}$ of variables without any structural assumption, i.e. we do not assume $S_n \propto \sum_{i=1}^n X_i$ with the X_i i.i.d. In that setting, as the Curie-Weiss example shows, properties of a rate function (such as convexity) depend on the particulars of the model, the existence of a large deviation principle is not guaranteed, and we briefly mentioned that the speed of a large deviation principle may not be given by the number of variables even if $S_n = \sum_{i=1}^n X_i$ when the X_i are not independent. It is however possible to give a general sufficient condition under which a large deviation principle with convex rate function must hold, through a construction mirroring what we did for Cramér's theorem.

Define:

$$\Phi_n(\lambda) := \log \mathbb{E}[e^{(\lambda, S_n)}], \qquad \lambda \in \mathbb{R}^d. \tag{5.1}$$

We make the following assumption throughout.

Assumption 5.1. *For some sequence $a_n > 0$ $(n \in \mathbb{N})$ with $\lim_{n\to\infty} a_n = \infty$,*

$$\Phi(\lambda) := \lim_{n\to\infty} \frac{1}{a_n} \log \Phi_n(a_n\lambda) \quad \textit{exists in } [-\infty, \infty]^d \ \textit{ for each } \lambda \in \mathbb{R}^d, \tag{5.2}$$

and the set $\mathcal{D}_\Phi := \{\lambda : \Phi(\lambda) < \infty\}$ contains 0 in its interior, i.e. $0 \in \mathcal{D}^o_\phi$.

Remark 5.2. *In the case $S_n = \frac{1}{n}\sum_{i=1}^n X_i$ with the X_i i.i.d and $a_n = n$, the Φ above is the moment generating function of X_1, called ϕ in Section 1.*

5.1 Properties of Φ and its Legendre transform

Lemma 5.3. *The function Φ is convex on $\mathbb{R}^d$ and $\mathcal{D}_\Phi$ is a convex set. Moreover, $\Phi > -\infty$ everywhere and Φ is continuous on $\mathcal{D}^o_\Phi$.*

Proof. As a pointwise limit of convex functions, Φ is convex on $\mathbb{R}^d$ which implies that $\mathcal{D}_\phi$ is a convex set. Convexity also implies continuity of Φ on the interior of $\{|\Phi| < \infty\}$. Let us conclude by showing that $\{|\Phi| < \infty\}$ coincides with the domain $\mathcal{D}_\Phi$ of Φ. As 0 is an interior point, there is $\delta > 0$ such that $B(0, \delta) \subset \mathcal{D}_\Phi$. Since $\Phi(0) = 0 > -\infty$, it must be that $\Phi(\lambda) > -\infty$ for any $\lambda \in \mathbb{R}^d \setminus \{0\}$. Indeed, if not convexity of Φ on the segment $[-\delta\lambda/|\lambda|, \lambda]$ gives a contradiction. Thus $\mathcal{D}_\Phi = \{|\Phi| < \infty\}$. □

Remark 5.4. *Contrary to the i.i.d. case, Φ need not be smooth on the interior of its domain.*

Define the Legendre transform I of Φ:

$$I(x) := \sup_{\lambda\in\mathbb{R}^d} \big\{(\lambda, x) - \Phi(\lambda)\big\}, \qquad x \in \mathbb{R}^d. \tag{5.3}$$

Lemma 5.5. *The function I is convex and a good rate function.*

Proof. From the definition $I \geq 0$, and I is convex and lower semi-continuous as a supremum of linear functions.

Let us prove that I has bounded sub-level sets. Recall that $0 \in \mathcal{D}_\Phi^0$ and let $\varepsilon > 0$ be such that $B(0, 2\varepsilon) \subset \mathcal{D}_\Phi^o$. Then Φ is continuous on $\bar{B}(0, \varepsilon)$, thus bounded. Moreover,

$$\begin{aligned} I(x) &\geq \sup_{u:|u|=1} \Big\{ \varepsilon(u, x) - \Phi(\varepsilon u) \Big\} \geq \sup_{u:|u|=1} \Big\{ \varepsilon(u, x) - \sup_{u:|u|=1} \Phi(\varepsilon u) \Big\} \\ &= \varepsilon |x| - \sup_{u:|u|=1} \Phi(\varepsilon u). \end{aligned} \tag{5.4}$$

This gives $\lim_{|x| \to \infty} I(x) = +\infty$ and the boundedness of sub-level sets. □

Definition 5.6 (Exposed point)**.** *We say that $x \in \mathbb{R}^d$ is an* exposed point *if, for some $\lambda \in \mathbb{R}^d$ and all $y \neq x$:*

$$I(y) > I(x) + (\lambda, y - x). \tag{5.5}$$

In other words I is above the affine hyperplane of normal λ containing $(x, I(x))$. This hyperplane is called the exposing hyperplane*. The set of exposed points with normal $\lambda \in \mathcal{D}_\Phi^o$ is denoted by $\mathcal{E}$.*

One can check that if x is an exposed point, then the rate function I satisfies:

$$(\lambda, x) - I(x) = \sup_{y \in \mathbb{R}^d} \big\{ (\lambda, y) - I(y) \big\} = \Phi(\lambda). \tag{5.6}$$

5.2 The theorem

Theorem 5.7. *Let $S_n \in \mathbb{R}^d$ ($n \in \mathbb{N}$) be a sequence of random variables for which Assumption 5.1 holds. Let μ_n denote the law of S_n and I denote the good convex rate function in* (5.3)*. Then, for any closed set C:*

$$\limsup_{n \to \infty} \frac{1}{a_n} \log \mu_n(C) \leq - \inf_C I, \tag{5.7}$$

and for any open set O:

$$\liminf_{n \to \infty} \frac{1}{a_n} \log \mu_n(O) \geq - \inf_{O \cap \mathcal{E}} I. \tag{5.8}$$

Remark 5.8. *Note that Theorem 5.7 is not a large deviation principle in the sense of Definition 4.2 since the lower bound only holds for exposed points. With additional conditions on (S_n) the lower bound can be strengthened, see Proposition 5.9.*

Proof. The proof resembles that of Cramér's theorem, with some additional ingredients due to being in $\mathbb{R}^d$ rather than $\mathbb{R}$ and the lack of a law of large numbers.

Upper bound. Let $x \in \mathbb{R}^d$ and $\delta > 0$. We first prove an upper bound for an open ball $B(x, \delta)$. By Chebychev exponential inequality, for each $\lambda \in \mathbb{R}^d \setminus \{0\}$:

$$\begin{aligned} \mu_n(B(x,\delta)) &= \mathbb{P}\big(|S_n - x| < \delta\big) \leq \mathbb{P}\big(|(S_n - x, \lambda)| < \delta|\lambda|\big) \\ &\leq \mathbb{P}\big((S_n - x, \lambda) > -\delta|\lambda|\big) \\ &\leq \mathbb{E}\Big[e^{a_n(S_n - x, \lambda)}\Big] e^{\delta\lambda a_n}. \end{aligned} \tag{5.9}$$

Taking $\frac{1}{a_n}\log$ and the large n limit yields:

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n(B(x,\delta)) \leq \delta|\lambda| - (x,\lambda) + \Phi(\lambda). \tag{5.10}$$

Note that the bound is also valid for $\lambda = 0$.

Now, by definition of the rate function I, for every $\varepsilon \in (0,1)$ one can find $\lambda_{x,\varepsilon} \in \mathbb{R}^d$ such that:

$$(x, \lambda_{x,\varepsilon}) - \Phi(\lambda_{x,\varepsilon}) \geq (I(x) - \varepsilon) \wedge \frac{1}{\varepsilon}, \tag{5.11}$$

where the $1/\varepsilon$ accounts for the possibility that $I(x) = +\infty$. Then:

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n(B(x,\delta)) \leq \delta|\lambda_{x,\varepsilon}| - (I(x) - \varepsilon) \wedge \frac{1}{\varepsilon}. \tag{5.12}$$

From this bound for balls we now deduce a bound for compact sets. For $x \in K$, define $\delta_{x,\varepsilon} := \varepsilon/\lambda_{x,\varepsilon} \wedge 1$. Then $\bigcup_{x\in K} B(x, \delta_{x,\varepsilon})$ is an open cover of K. Extract a finite subcover $B(x_i, \delta_{x_i,\varepsilon})$ $(1 \leq i \leq p_\varepsilon)$ and abbreviate $\delta_{x_i,\varepsilon}$ as δ_i and $\lambda_{x_i,\varepsilon}$ as λ_i. Then:

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n(K) &\leq \max_{1\leq i\leq p_\varepsilon} \limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n(B(x_i, \delta_i)) \\ &\leq \max_{1\leq i\leq p_\varepsilon} \Big\{\delta_i|\lambda_i| - (I(x_i) - \varepsilon) \wedge \frac{1}{\varepsilon}\Big\} \\ &\leq \varepsilon - \min_{1\leq i\leq p_\varepsilon} \Big\{(I(x_i) - \varepsilon) \wedge \frac{1}{\varepsilon}\Big\} \\ &= \varepsilon - \frac{1}{\varepsilon} \wedge \min_{1\leq i\leq p_\varepsilon} (I(x_i) - \varepsilon). \end{aligned} \tag{5.13}$$

Thus:

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n(K) \leq \varepsilon - \frac{1}{\varepsilon} \wedge (\inf_K I - \varepsilon). \tag{5.14}$$

Taking ε to 0 yields an upper bound for compact sets.

We now strengthen this bound to closed sets. Informally, this is done by showing that most of the mass of the measures μ_n is concentrated on compact sets at an exponential scale, a property known as exponential tightness. This is rigorously formulated as follows:

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n\big(B(0,M)^c\big) \underset{M\to\infty}{\longrightarrow} -\infty. \tag{5.15}$$

To prove (5.15), recall that $0 \in \mathcal{D}_I^o$ and let $\delta > 0$ be such that $B(0, 2\delta) \subset \mathcal{D}_I^o$. Then, writing $(e_i)_{1\leq i\leq d}$ for the canonical basis of $\mathbb{R}^d$ and using again the exponential Chebychev inequality:

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n\big(B(0,M)^c\big) &\leq \max_{1\leq i\leq d} \limsup_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}\Big(|(S_n, e_i)| \geq M/\sqrt{d}\Big)\\ &\leq \max_{\substack{1\leq i\leq d\\ s\in\{-,+\}}} \Big\{ -\frac{\delta M}{\sqrt{d}} + \Phi(\delta s e_i)\Big\}. \end{aligned} \tag{5.16}$$

The right-hand side goes to $-\infty$ when $M \to \infty$ as desired.

Using (5.15), the upper bound for closed sets is proven as follows. Let C be a closed set. Then, for each $M > 0$:

$$\begin{aligned}&\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n(C)\\ &\qquad \leq \max\Big\{ \limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n(C \cap \bar{B}(0,M)), \limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n(B(0,M)^c)\Big\}. \end{aligned} \tag{5.17}$$

The second term goes to $-\infty$ when $M \to \infty$ by (5.15). On the other hand, the set $C\cap\bar{B}(0,M)$ is compact, thus:

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n(C \cap \bar{B}(0,M)) \leq -\inf_{C\cap\bar{B}(0,M)} I. \tag{5.18}$$

As $C \cap \bar{B}(0,M)$ increases to C, the infimum decreases to $\inf_C I$ (check it). This concludes the proof of the upper bound for compact sets.

Lower bound. It is enough to prove that, for all x in the exposed set $\mathcal{E}$:

$$\liminf_{\delta\to 0} \liminf_{n\to\infty} \frac{1}{a_n} \log \mu_n(B(x,\delta)) \geq -I(x). \tag{5.19}$$

Indeed, if so then for any open set O and any $x \in \mathcal{E}, \delta > 0$ with $B(x,\delta) \subset O$,

$$\liminf_{n\to\infty} \frac{1}{a_n} \log \mu_n(O) \geq \liminf_{\delta\to 0} \liminf_{n\to\infty} \frac{1}{a_n} \log \mu_n(B(x,\delta)) \geq -I(x). \tag{5.20}$$

To prove (5.19), we again proceed as in the case of Cramér's theorem, tilting the measure so that $B(x,\delta)$ becomes typical. The difference is that we do not know the law of large numbers for S_n. We will instead use the large deviation upper bound.

Let $x \in \mathcal{E}$ and let $\lambda \in \mathcal{D}_\Phi^0$ be normal to an exposed hyperplane for x. Recall from (5.6) that I and $\Phi(\lambda)$ are then related as:

$$I(x) = (\lambda, x) - \Phi(\lambda). \tag{5.21}$$

Then:

$$\mu_n(B(x,\delta)) \geq e^{-a_n\delta|\lambda|}\mathbb{E}\Big[\mathbf{1}_{|S_n - x|<\delta} e^{a_n(\lambda, S_n - x)}\Big] = e^{-a_n\delta|\lambda|}\mathbb{E}\Big[e^{a_n(\lambda,S_n-x)}\Big]\mu_n^\lambda(B(x,\delta)), \tag{5.22}$$

with μ_n^λ the tilted probability measure proportional to $e^{a_n(\lambda,y-x)}\mu_n(dy)$. Taking $\frac{1}{a_n}\log$ and the large n limit,

$$\begin{aligned}\liminf_{n\to\infty}\frac{1}{a_n}\log\mu_n(B(x,\delta)) &\geq -\delta|\lambda| + \Phi(\lambda) - (x,\lambda) + \liminf_{n\to\infty}\frac{1}{a_n}\log\mu_n^\lambda(B(x,\delta))\\ &= -\delta|\lambda| - I(x) + \liminf_{n\to\infty}\frac{1}{a_n}\log\mu_n^\lambda(B(x,\delta)). \qquad (5.23)\end{aligned}$$

To estimate this last probability, notice that μ_n^λ satisfies Assumption 5.1. Indeed, for any $\tau\in\mathbb{R}^d$:

$$\lim_{n\to\infty}\frac{1}{a_n}\log\int e^{a_n(y,\tau)}\mu_n^\lambda(dy) = \Phi(\lambda+\tau)-\Phi(\lambda). \qquad (5.24)$$

In addition, 0 is an interior point of the domain of $\Phi(\lambda+\cdot)-\Phi(\lambda)$ since $\lambda\in\mathcal{D}_\Phi^o$ by definition of exposed points. In particular upper bound large deviations hold for μ_n^λ, and:

$$\limsup_{n\to\infty}\frac{1}{a_n}\log\mu_n^\lambda(B(x,\delta)^c) \leq -\inf_{\bar{B}(x,\delta)^c} I^\lambda, \qquad (5.25)$$

with the good rate function I^λ given by:

$$I^\lambda(y) = \sup_{\tau\in\mathbb{R}^d}\Big\{(\tau,y)-\Phi(\lambda+\tau)\Big\}+\Phi(\lambda), \qquad y\in\mathbb{R}^d. \qquad (5.26)$$

We claim that $\inf_{B(x,\delta)^c} I^\lambda > 0$. Indeed, I^λ is either identically equal to $+\infty$ on $B(x,\delta)^c$, or it achieves its infimum at some $y\in B(x,\delta)^c$ due to being a good rate function. In the latter case, the fact that x is an exposed point gives, using again (5.6):

$$\begin{aligned}I^\lambda(y_\delta) &= \sup_{\tau\in\mathbb{R}^d}\Big\{(\tau+\lambda,y_\delta)-\Phi(\lambda+\tau)\Big\}+\Phi(\lambda)-(\lambda,y) = I(y_\delta)+\Phi(\lambda)-(\lambda,y)\\ &= I(y_\delta)-(\lambda,y)-(I(x)-(\lambda,x))>0. \qquad (5.27)\end{aligned}$$

This implies, for some $C_{x,\delta}>0$ and all large enough n in the last inequality:

$$\mu_n(B(x,\delta)) = 1-\mu_n\big(B(x,\delta)^c\big)\geq 1-C_{\delta,x}e^{-a_nC_{\delta,x}}\geq\frac{1}{2}, \qquad (5.28)$$

which concludes the proof. □

The following proposition gives sufficient conditions on Φ to remove the set $\mathcal{E}$ in the lower bound of Theorem 5.7, turning its statement into a full large deviation principle.

Proposition 5.9. *Assume:*

1. *Φ is lower semi-continuous on $\mathbb{R}^d$.*
2. *Φ is differentiable in $\mathcal{D}_\Phi^0$.*
3. *$\mathcal{D}_\Phi=\mathbb{R}^d$, or $\lim_{\lambda\in\mathcal{D}_\Phi^o\to\partial\mathcal{D}_\Phi}|\nabla\Phi(\lambda)|=+\infty$.*

Then the lower bound in Theorem 5.7 holds without $\mathcal{E}$.

Remark 5.10. *If $d=1$, the condition on the slope of Φ is the analogue of what we needed in the proof of Cramér's theorem to define a tilt even at the boundary of a closed interval.*

6 A glance at some topics not seen in the course

6.1 Moderate deviations

Let (X_i) be i.i.d. real random variables with finite exponential moment in a neighbourhood of the origin and unit variance. The central limit theorem gives a bound of the form

$$\mathbb{P}\big(\frac{S_n}{\sqrt{n}} \in [a,b]\big) = \frac{1}{\sqrt{2\pi}} \int_{[a,b]} e^{-\frac{x^2}{2}}\, dx. \tag{6.1}$$

On the other hand Cramér's theorem gives:

$$\mathbb{P}\Big(\frac{S_n}{n} \in [a,b]\Big) \asymp e^{-n \inf_{[a,b]} I}. \tag{6.2}$$

What about S_n/n^α for $\alpha \in (1/2, 1)$? It turns out that the result is a mixture of both central limit theorem and large deviations:

$$\mathbb{P}\Big(\frac{S_n}{n^\alpha} \in [a,b]\Big) \asymp \exp\Big[-n^{2\alpha-1} \inf_{[a,b]} \frac{x^2}{2}\Big]. \tag{6.3}$$

In other words, deviations do hold but the rate function is always the Legendre transform of the moment generating function of a Gaussian. The proof is very similar to Cramér's theorem. The idea is that for intermediate scalings of the sums one can do a Taylor expansion of the moment generating function and terms higher than second order will be negligible. See Section 3.3.7 in Dembo-Zeitouni for more on this topic.

6.2 Heavy-tailed random variables

Consider i.i.d. X_i ($i \in \mathbb{N}$) but suppose that X_1 only has a finite number of moments, or worse, that $X_1 \geq 0$ and $\log X_1$ only has a finite number of moments. Can one then say anything about deviations of S_n? In such a regime deviations are dominated by the behaviour of single random variables, i.e. $S_n \approx nm$ if one variable is very large rather than all of them being close to m. If e.g. $\mathbb{P}(X_1 > t) \sim |t|^{-\alpha}$ as $|t| \to \infty$ for some $\alpha \in (0,\infty)$, one can still prove deviations of the form:

$$\lim_{n\to\infty} \frac{1}{\log n} \log \mathbb{P}(S_n > n^x) = 1 - \alpha x, \qquad x > \max\{1, 1/\alpha\}. \tag{6.4}$$

See e.g. Gantert, N. (2000). A note on logarithmic tail asymptotics and mixing. Stat. Probab. Lett. 49, 113–118.

6.3 Sample path large deviations

Consider a sequence of i.i.d real random variables $(X_i)_{i\geq 1}$. Assume:

$$\forall \lambda \in \mathbb{R}, \qquad \mathbb{E}[e^{\lambda X_1}] < \infty. \tag{6.5}$$

Consider the function of $t \in [0,1]$ given by:

$$t \in [0,1] \mapsto \frac{1}{n} \sum_{i=0}^{\lfloor nt \rfloor} X_i. \tag{6.6}$$

Let $Z_n(\cdot)$ denote the continuous process obtained by linear interpolation of the last object. The one can look at large deviations for $(Z_n(t))_{t\in[0,1]}$, viewed as a random element in the set $C_0([0,1],\mathbb{R})$ of continuous functions from $[0,1]$ to $\mathbb{R}$ vanishing at 0. To state a result, we need more notations. Let $\mathcal{C}_{\text{abs}}$ denote the subset of absolutely continuous functions:

$$\mathcal{C}_{\text{abs}} = \Big\{ f \in C_0([0,1],\mathbb{R}) \,\Big|\, \text{there is } \dot{f} \in \mathbb{L}^1((0,1)) \text{ such that } f(t) = \int_0^t \dot{f}(s)\, ds \Big\}. \tag{6.7}$$

Let also μ_n denote the law of Z_n for $n \geq 1$, and recall that I denotes the Legendre transform of the log-moment generating function:

$$I(x) := \sup_{\lambda \in \mathbb{R}} \Big\{ \lambda x - \phi(\lambda) \Big\}. \tag{6.8}$$

Theorem 6.1 (Mogulskii). *The measures $(\mu_n)_{n\geq 1}$ satisfy a large deviation principle on $C_0([0,1],\mathbb{R})$ with speed n and rate function:*

$$\mathcal{I}(f) = \begin{cases} \int_0^1 I(\dot{f}(s))\, ds & \textit{if } f \in \mathcal{C}_{\text{abs}}, \\ +\infty & \textit{otherwise.} \end{cases} \tag{6.9}$$

This theorem in particular provides sample-path large deviations for Brownian motion as we now explain. Let $(B_t)_{t\geq 0}$ denote a real Brownian motion and let B_n $(n \geq 1)$ be given by:

$$B_n(t) = \frac{1}{\sqrt{n}} B(t), \qquad t \in [0,1]. \tag{6.10}$$

Theorem 6.2 (Schilder). *The law ν_n of B_n $(n \geq 1)$ satisfies a large deviation principle on $C_0([0,1],\mathbb{R})$ with speed n and rate function:*

$$\mathcal{I}(f) = \begin{cases} \frac{1}{2} \int_0^1 [\dot{f}(s)]^2\, ds & \textit{if } f \in \mathcal{C}_{\text{abs}}, \\ +\infty & \textit{otherwise.} \end{cases} \tag{6.11}$$

Proof. We prove the theorem by comparing Brownian motion with a random walk and using Mogulskii's result. Notice that B_n reads:

$$B_n(t) = \frac{1}{\sqrt{n}} \sum_{k=1}^{\lfloor nt \rfloor} \Delta_k + \frac{1}{\sqrt{n}} \Big(B(t) - B_n(t) \Big), \tag{6.12}$$

where $\Delta_k = n^{-1/2} X_k$ and $X_k = \sqrt{n}[B(k/n) - B((k-1)/n)]$ is a standard normal random variables. The $(X_k)_{1\leq k \leq n}$ are independent. If $(Z_n(t))_{t\in[0,1]}$ denotes the continuous process built from the X_k as in (6.6), then in particular:

$$B_n(k/n) = Z_n(k/n), \qquad 0 \leq k \leq n, \tag{6.13}$$

and Mogulskii's theorem gives large deviations for Z_n, with the correct rate function (for standard normal random variables ones has $I(x) = x^2/2$). It is thus enough to prove that B_n and Z_n remain very close on $[0,1]$, i.e. to prove:

$$\forall \delta > 0, \qquad \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(\sup_{t\in[0,1]} \big|B_n(t) - Z_n(t)\big| \geq \delta\big) = -\infty. \tag{6.14}$$

A union bound gives:

$$\mathbb{P}\big(\sup_{t\in[0,1]} \big|B_n(t) - Z_n(t)\big| \geq \delta\big) \leq \sum_{k=0}^{n-1} \mathbb{P}\big(\sup_{k/n\leq t\leq (k+1)/n} \big|B_n(t) - Z_n(t)\big| \geq \delta\big). \tag{6.15}$$

Since $B_n(k/n) = Z_n(k/n)$, it will be enough to prove that both $B_n(t) - B_n(k/n)$ and $Z_n(t) - Z_n(k/n)$ satisfy the above bound. For Z_n, its linearity on $[k/n, (k+1)/n]$ gives:

$$\mathbb{P}\big(\sup_{k/n\leq t\leq (k+1)/n} \big|Z_n(t) - Z_n(k/n)\big| \geq \delta\big) = \mathbb{P}\big(n^{-1}|X_k| \geq \delta\big) \leq e^{-n^2\delta^2/2}. \tag{6.16}$$

The same bound is true for Brownian motion. Indeed, as $(e^{B(t)})_{t\geq 0}$ is a positive submartingale, Doob's inequality gives

$$\begin{aligned}\mathbb{P}\big(\sup_{k/n\leq t\leq (k+1)/n} \big|B_n(t) - B_n(k/n)\big| \geq \delta\big) &= \mathbb{P}\big(\sup_{t\leq 1/n} \big|B(t)\big| \geq \sqrt{n}\delta\big) = \mathbb{P}\big(\sup_{t\leq 1} \big|B(t)\big| \geq n\delta\big) \\ &\leq e^{-n^2\delta^2/2}. \end{aligned} \tag{6.17}$$

□