Chapter 23
Large Deviations

Except for the law of the iterated logarithm, so far we have encountered two types of limit theorems for partial sums $S_n = X_1 + \ldots + X_n$, $n \in \mathbb{N}$, of identically distributed, real random variables $(X_i)_{i\in\mathbb{N}}$ with distribution function F:

(1) (Weak) laws of large numbers state that (under suitable assumptions on the family $(X_i)_{i\in\mathbb{N}}$), for every $x > 0$,

$$\mathbf{P}\big[\big|S_n - n\mathbf{E}[X_1]\big| \geq xn\big] \overset{n\to\infty}{\longrightarrow} 0. \tag{23.1}$$

From this we get immediately that the empirical distribution functions

$$F_n : x \mapsto \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,x]}(X_i)$$

converge in probability; that is, $\|F_n - F\|_\infty \overset{n\to\infty}{\longrightarrow} 0$. In other words, for any distribution function $G \neq F$ and any $\varepsilon > 0$ with $\varepsilon < \|F - G\|_\infty$, we have

$$\mathbf{P}\big[\|F_n - G\|_\infty < \varepsilon\big] \overset{n\to\infty}{\longrightarrow} 0. \tag{23.2}$$

(2) Central limit theorems state that (under different assumptions on the family $(X_i)_{i\in\mathbb{N}}$) for every $x \in \mathbb{R}$

$$\mathbf{P}\big[S_n - n\mathbf{E}[X_1] \geq x\sqrt{n}\,\big] \overset{n\to\infty}{\longrightarrow} 1 - \Phi\left(\frac{x}{\sqrt{\mathbf{Var}[X_1]}}\right). \tag{23.3}$$

Here $\Phi : t \mapsto \mathcal{N}_{0,1}((-\infty, t])$ is the distribution function of the standard normal distribution.

In each case, the *typical value* of S_n is $n\mathbf{E}[X_1]$. Equation (23.3) makes a precise statement about the average size of the deviations (which are of order $\sqrt{n}$) from the typical value. A simple consequence is of course that the probability of *large deviations* (of order n) from the typical value goes to 0; that is, (23.1) holds.

A. Klenke, *Probability Theory*, Universitext, DOI 10.1007/978-1-4471-5361-0_23,

In this chapter, we compute the *speed of convergence* in (23.1) (Cramér's theorem) and in (23.2) (Sanov's theorem).

We follow in part the expositions in [31, 74].

23.1 Cramér's Theorem

Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbf{P}_{X_i} = \mathcal{N}_{0,1}$. Then, for every $x > 0$,

$$\mathbf{P}[S_n \geq xn] = \mathbf{P}[X_1 \geq x\sqrt{n}] = 1 - \Phi(x\sqrt{n}) = (1+\varepsilon_n)\frac{1}{x\sqrt{2\pi n}}e^{-nx^2/2},$$

where $\varepsilon_n \overset{n\to\infty}{\longrightarrow} 0$ (by Lemma 22.2). Taking logarithms, we get

$$\lim_{n\to\infty} \frac{1}{n}\log \mathbf{P}[S_n \geq xn] = -\frac{x^2}{2} \quad \text{for every } x > 0. \tag{23.4}$$

It might be tempting to believe that a central limit theorem could be used to show (23.4) for all centered i.i.d. sequences (X_i) with finite variance. However, in general, the limit might be infinite or might be a different function of x, as we will show below. The moral is that large deviations depend more subtly on the tails of the distribution of X_i than the average-sized fluctuations do (which are determined by the variance only). The following theorem shows this for Bernoulli random variables.

Theorem 23.1 *Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbf{P}[X_1 = -1] = \mathbf{P}[X_1 = 1] = \frac{1}{2}$. Then, for every $x \geq 0$,*

$$\lim_{n\to\infty} \frac{1}{n}\log \mathbf{P}[S_n \geq xn] = -I(x), \tag{23.5}$$

where the rate function *I is given by*

$$I(z) = \begin{cases} \frac{1+z}{2}\log(1+z) + \frac{1-z}{2}\log(1-z), & \textit{if } z \in [-1,1], \\ \infty, & \textit{if } |z| > 1. \end{cases} \tag{23.6}$$

Remark 23.2 Here we agree that $0\log 0 = 0$. This makes the restriction of I to $[-1,1]$ a continuous function with $I(-1) = I(1) = \log 2$. Note that I is strictly convex on $[-1,1]$ with $I(0) = 0$ and I is monotone increasing on $[0,1]$ and is monotone decreasing on $[-1,0]$. ◊

Proof of Theorem 23.1 For $x = 0$ and $x > 1$, the claim is trivial. For $x = 1$, we have $\mathbf{P}[S_n \geq n] = 2^{-n}$, and thus again (23.5) holds trivially. Hence, it is enough to consider $x \in (0,1)$. Since $\frac{S_n+n}{2} \sim b_{n,1/2}$ is binomially distributed, we have

$$\mathbf{P}[S_n \geq xn] = 2^{-n} \sum_{k\geq(1+x)n/2} \binom{n}{k}.$$

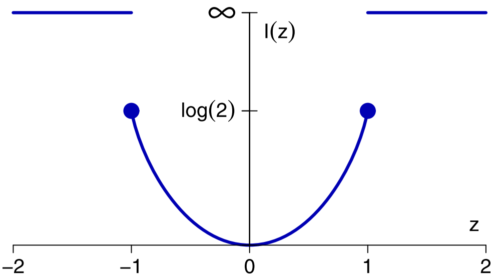

Fig. 23.1 Rate function $I(z)$ from (23.6)

Define $a_n(x) = \lceil n(1+x)/2 \rceil$ for $n \in \mathbb{N}$. Since $k \mapsto \binom{n}{k}$ is monotone decreasing for $k \geq \frac{n}{2}$, we get

$$Q_n(x) := \max\left\{ \binom{n}{k} : a_n(x) \leq k \leq n \right\} = \binom{n}{a_n(x)}. \tag{23.7}$$

We make the estimate

$$2^{-n} Q_n(x) \leq \mathbf{P}[S_n \geq xn] \leq (n+1) 2^{-n} Q_n(x). \tag{23.8}$$

By Stirling's formula

$$\lim_{n\to\infty} \frac{1}{n!} n^n e^{-n} \sqrt{2\pi n} = 1,$$

we obtain

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log Q_n(x) \\
&\quad = \lim_{n\to\infty} \frac{1}{n} \log \frac{n!}{a_n(x)! \cdot (n - a_n(x))!} \\
&\quad = \lim_{n\to\infty} \frac{1}{n} \log \frac{n^n}{a_n(x)^{a_n(x)} \cdot (n - a_n(x))^{n-a_n(x)}} \\
&\quad = \lim_{n\to\infty} \left[\log(n) - \frac{a_n(x)}{n} \log\big(a_n(x)\big) - \frac{n - a_n(x)}{n} \log\big(n - a_n(x)\big) \right] \\
&\quad = \lim_{n\to\infty} \left[\log(n) - \frac{1+x}{2} \left(\log\left(\frac{1+x}{2}\right) + \log(n) \right) \right. \\
&\qquad \left. - \frac{1-x}{2} \left(\log\left(\frac{1-x}{2}\right) + \log(n) \right) \right] \\
&\quad = -\frac{1+x}{2} \log\left(\frac{1+x}{2}\right) - \frac{1-x}{2} \log\left(\frac{1-x}{2}\right) = -I(x) + \log 2.
\end{aligned}$$

Together with (23.8), this implies (23.5). □

Under certain assumptions on the distribution of X_1, Cramér's theorem [29] provides a general principle to compute the rate function I.

Theorem 23.3 (Cramér (1938)) *Let $X_1, X_2, \dots$ be i.i.d. real random variables with finite logarithmic moment generating function*

$$\Lambda(t) := \log \mathbf{E}\big[e^{tX_1}\big] < \infty \quad \textit{for all } t \in \mathbb{R}. \tag{23.9}$$

Let

$$\Lambda^*(x) := \sup_{t\in\mathbb{R}} \big(tx - \Lambda(t)\big) \quad \textit{for } x \in \mathbb{R},$$

the Legendre transform of Λ. Then, for every $x > \mathbf{E}[X_1]$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_n \geq xn] = -I(x) := -\Lambda^*(x). \tag{23.10}$$

Proof By passing to $X_i - x$ if necessary, we may assume $\mathbf{E}[X_i] < 0$ and $x = 0$. (In fact, if $\tilde{X}_i := X_i - x$, and $\tilde{\Lambda}$ and $\tilde{\Lambda}^*$ are defined as Λ and Λ^* above but for $\tilde{X}_i$ instead of X_i, then $\tilde{\Lambda}(t) = \Lambda(t) - t \cdot x$ and thus $\tilde{\Lambda}^*(0) = \sup_{t\in\mathbb{R}}(-\tilde{\Lambda}(t)) = \Lambda^*(x)$.)

Define $\varphi(t) := e^{\Lambda(t)}$ and

$$\varrho := e^{-\Lambda^*(0)} = \inf_{t\in\mathbb{R}} \varphi(t).$$

By (23.9) and the differentiation lemma (Theorem 6.28), φ is differentiable infinitely often and the first two derivatives are

$$\varphi'(t) = \mathbf{E}\big[X_1 e^{tX_1}\big] \quad \text{and} \quad \varphi''(t) = \mathbf{E}\big[X_1^2 e^{tX_1}\big].$$

Hence φ is strictly convex and $\varphi'(0) = \mathbf{E}[X_1] < 0$.

First consider the case $\mathbf{P}[X_1 \leq 0] = 1$. Then $\varphi'(t) < 0$ for every $t \in \mathbb{R}$ and $\varrho = \lim_{t\to\infty} \varphi(t) = \mathbf{P}[X_1 = 0]$. Therefore,

$$\mathbf{P}[S_n \geq 0] = \mathbf{P}[X_1 = \ldots = X_n = 0] = \varrho^n$$

and thus the claim follows.

Now let $\mathbf{P}[X_1 < 0] > 0$ and $\mathbf{P}[X_1 > 0] > 0$. Then $\lim_{t\to\infty} \varphi(t) = \infty = \lim_{t\to-\infty} \varphi(t)$. As φ is strictly convex, there is a unique $\tau \in \mathbb{R}$ at which φ assumes its minimum; hence

$$\varphi(\tau) = \varrho \quad \text{and} \quad \varphi'(\tau) = 0.$$

Since $\varphi'(0) < 0$, we have $\tau > 0$. Using Markov's inequality (Theorem 5.11), we estimate

$$\mathbf{P}[S_n \geq 0] = \mathbf{P}\big[e^{\tau S_n} \geq 1\big] \leq \mathbf{E}\big[e^{\tau S_n}\big] = \varphi(\tau)^n = \varrho^n.$$

Thus we get the upper bound

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_n \geq 0] \leq \log \varrho = -\Lambda^*(0).$$

The remaining part of the proof is dedicated to verifying the reverse inequality:

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_n \geq 0] \geq \log \varrho. \tag{23.11}$$

We use the method of an exponential size-biasing of the distribution $\mu := \mathbf{P}_{X_1}$ of X_1, which turns the atypical values that are of interest here into typical values. That is, we define the *Cramér transform* $\hat{\mu} \in \mathcal{M}_1(\mathbb{R})$ of μ by

$$\hat{\mu}(dx) = \varrho^{-1} e^{\tau x} \mu(dx) \quad \text{for } x \in \mathbb{R}.$$

Let $\hat{X}_1, \hat{X}_2, \ldots$ be independent and identically distributed with $\mathbf{P}_{\hat{X}_i} = \hat{\mu}$. Then

$$\hat{\varphi}(t) := \mathbf{E}\big[e^{t\hat{X}_1}\big] = \frac{1}{\varrho} \int_{\mathbb{R}} e^{tx} e^{\tau x} \mu(dx) = \frac{1}{\varrho} \varphi(t+\tau).$$

Hence

$$\mathbf{E}[\hat{X}_1] = \hat{\varphi}'(0) = \frac{1}{\varrho} \varphi'(\tau) = 0,$$

$$\mathbf{Var}[\hat{X}_1] = \hat{\varphi}''(0) = \frac{1}{\varrho} \varphi''(\tau) \in (0, \infty).$$

Defining $\hat{S}_n = \hat{X}_1 + \ldots + \hat{X}_n$, we get

$$\begin{aligned}
\mathbf{P}[S_n \geq 0] &= \int_{\{x_1+\ldots+x_n \geq 0\}} \mu(dx_1) \ldots \mu(dx_n) \\
&= \int_{\{x_1+\ldots+x_n \geq 0\}} \big(\varrho e^{-\tau x_1}\big) \hat{\mu}(dx_1) \ldots \big(\varrho e^{-\tau x_n}\big) \hat{\mu}(dx_n) \\
&= \varrho^n \mathbf{E}\big[e^{-\tau \hat{S}_n} \mathbb{1}_{\{\hat{S}_n \geq 0\}}\big].
\end{aligned}$$

Thus, in order to show (23.11), it is enough to show

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{E}\big[e^{-\tau \hat{S}_n} \mathbb{1}_{\{\hat{S}_n \geq 0\}}\big] \geq 0. \tag{23.12}$$

However, by the central limit theorem (Theorem 15.37), for every $c > 0$,

$$\begin{aligned}
\frac{1}{n} \log \mathbf{E}\big[e^{-\tau \hat{S}_n} \mathbb{1}_{\{\hat{S}_n \geq 0\}}\big] &\geq \frac{1}{n} \log \mathbf{E}\big[e^{-\tau \hat{S}_n} \mathbb{1}_{\{0 \leq \hat{S}_n \leq c\sqrt{n}\}}\big] \\
&\geq \frac{1}{n} \log\left(e^{-\tau c \sqrt{n}} \mathbf{P}\left[\frac{\hat{S}_n}{\sqrt{n}} \in [0, c] \right] \right) \\
&\overset{n\to\infty}{\longrightarrow} \lim_{n\to\infty} \frac{-\tau c \sqrt{n}}{n} + \lim_{n\to\infty} \frac{1}{n} \log\big(\mathcal{N}_{0,\mathbf{Var}[\hat{X}_1]}\big([0,c]\big)\big) \\
&= 0.
\end{aligned}$$

□

Example 23.4 If $\mathbf{P}_{X_1} = \mathcal{N}_{0,1}$, then

$$\Lambda(t) = \log\big(\mathbf{E}\big[e^{tX_1}\big]\big) = \log\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx} e^{-x^2/2}\,dx\right) = \frac{t^2}{2}.$$

Furthermore,

$$\Lambda^*(z) = \sup_{t\in\mathbb{R}}\big(tz - \Lambda(t)\big) = \sup_{t\in\mathbb{R}}\left(tz - \frac{t^2}{2}\right) = \frac{z^2}{2}.$$

Hence the rate function coincides with that of (23.4). ◊

Example 23.5 If $\mathbf{P}_{X_1} = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, then $\Lambda(t) = \log\cosh(t)$. The maximizer $t^* = t^*(z)$ of the variational problem for Λ^* solves the equation $z = \Lambda'(t^*) = \tanh(t^*)$. Hence

$$\Lambda^*(z) = zt^* - \Lambda\big(t^*\big) = z\,\operatorname{arc\,tanh}(z) - \log\big(\cosh\big(\operatorname{arc\,tanh}(z)\big)\big).$$

Now $\operatorname{arc\,tanh}(z) = \frac{1}{2}\log\frac{1+z}{1-z}$ for $z \in (-1, 1)$ and

$$\cosh\big(\operatorname{arc\,tanh}(z)\big) = \frac{1}{\sqrt{1-z^2}} = \frac{1}{\sqrt{(1-z)(1+z)}}.$$

Therefore,

$$\begin{aligned}\Lambda^*(z) &= \frac{z}{2}\log(1+z) - \frac{z}{2}\log(1-z) + \frac{1}{2}\log(1-z) + \frac{1}{2}\log(1+z)\\ &= \frac{1+z}{2}\log(1+z) + \frac{1-z}{2}\log(1-z).\end{aligned}$$

However, this is the rate function from Theorem 23.1. ◊

Exercise 23.1.1 Let X be a real random variable with density

$$f(x) = c^{-1}\frac{e^{-|x|}}{1+|x|^3},$$

where $c = \int_{-\infty}^{\infty} \frac{e^{-|x|}}{1+|x|^3}\,dx$. Check if the logarithmic moment generating function Λ is continuous and sketch the graph of Λ.

23.2 Large Deviations Principle

The basic idea of Cramér's theorem is to quantify the probabilities of rare events by an exponential rate and a rate function. In this section, we develop a formal framework for the quantification of probabilities of rare events in which the complete theory of large deviations can be developed. For further reading, consult, e.g., [31, 32] or [74].

Let E be a Polish space with complete metric d. Recall that

$$B_\varepsilon(x) = \{y \in E : d(x, y) < \varepsilon\}$$

denotes the open ball of radius $\varepsilon > 0$ that is centered at $x \in E$.

A map $f : E \to \overline{\mathbb{R}} = [-\infty, \infty]$ is called *lower semicontinuous* if, for every $a \in \overline{\mathbb{R}}$, the *level set* $f^{-1}([-\infty, a]) \subset E$ is closed. (In particular, continuous maps are lower semicontinuous. On the other hand, $\mathbb{1}_{(0,1)} : \mathbb{R} \to \mathbb{R}$ is lower semicontinuous but not continuous.) An equivalent condition for lower semicontinuity is that

$$\liminf_{\varepsilon \downarrow 0} f\big(B_\varepsilon(x)\big) = f(x) \quad \text{for all } x \in E.$$

(Recall that $\inf f(A) = \inf\{f(x) : x \in A\}$.) If $K \subset E$ is compact and nonempty, then f assumes its infimum on K. Indeed, for the case where $f(x) = \infty$ for all $x \in K$, the statement is trivial. Now assume $\inf f(K) < \infty$. If $a_n \downarrow \inf f(K)$ is strictly monotone decreasing, then $K \cap f^{-1}([-\infty, a_n]) \neq \emptyset$ is compact for every $n \in \mathbb{N}$ and hence the infinite intersection also is nonempty:

$$f^{-1}\big(\inf f(K)\big) = K \cap \bigcap_{n=1}^{\infty} f^{-1}\big([-\infty, a_n]\big) \neq \emptyset.$$

Definition 23.6 (Rate function) A lower semicontinuous function $I : E \to [0, \infty]$ is called a *rate function*. If all level sets $I^{-1}([-\infty, a])$, $a \in [0, \infty)$, are compact, then I is called a *good rate function*.

Definition 23.7 (Large deviations principle) Let I be a rate function and $(\mu_\varepsilon)_{\varepsilon>0}$ be a family of probability measures on E. We say that $(\mu_\varepsilon)_{\varepsilon>0}$ satisfies a *large deviations principle* (LDP) with rate function I if

(LDP 1) $\liminf_{\varepsilon\to 0} \varepsilon \log(\mu_\varepsilon(U)) \geq -\inf I(U)$ for every open $U \subset E$,
(LDP 2) $\limsup_{\varepsilon\to 0} \varepsilon \log(\mu_\varepsilon(C)) \leq -\inf I(C)$ for every closed $C \subset E$.

We say that a family $(P_n)_{n\in\mathbb{N}}$ of probability measures on E satisfies an LDP with rate $r_n \uparrow \infty$ and rate function I if (LDP 1) and (LDP 2) hold with $\varepsilon_n = 1/r_n$ and $\mu_{1/r_n} = P_n$.

Often (LDP 1) and (LDP 2) are referred to as *lower bound* and *upper bound*. In many cases, the lower bound is a lot easier to show than the upper bound.

Before we show that Cramér's theorem is essentially an LDP, we make two technical statements.

Theorem 23.8 *The rate function in an LDP is unique.*

Proof Assume that $(\mu_\varepsilon)_{\varepsilon>0}$ satisfies an LDP with rate functions I and J. Then, for every $x \in E$ and $\delta > 0$,

$$\begin{aligned} I(x) &\geq \inf I\big(B_\delta(x)\big) \\ &\geq -\liminf_{\varepsilon\to 0} \varepsilon \log\big(\mu_\varepsilon\big(B_\delta(x)\big)\big) \\ &\geq -\limsup_{\varepsilon\to 0} \varepsilon \log\big(\mu_\varepsilon\big(\overline{B_\delta(x)}\big)\big) \\ &\geq \inf J\big(\overline{B_\delta(x)}\big) \overset{\delta\to 0}{\longrightarrow} J(x). \end{aligned}$$

Hence $I(x) \geq J(x)$. Similarly, we get $J(x) \geq I(x)$. □

Lemma 23.9 *Let $N \in \mathbb{N}$ and let a_ε^i, $i = 1, \ldots, N$, $\varepsilon > 0$, be nonnegative numbers. Then*

$$\limsup_{\varepsilon\to 0} \varepsilon \log \sum_{i=1}^N a_\varepsilon^i = \max_{i=1,\ldots,N} \limsup_{\varepsilon\to 0} \varepsilon \log\big(a_\varepsilon^i\big).$$

Proof The sum and maximum differ at most by a factor N:

$$\max_{i=1,\ldots,N} \varepsilon \log\big(a_\varepsilon^i\big) \leq \varepsilon \log \sum_{i=1}^N a_\varepsilon^i \leq \varepsilon \log(N) + \max_{i=1,\ldots,N} \varepsilon \log\big(a_\varepsilon^i\big).$$

The maximum and limit (superior) can be interchanged and hence

$$\begin{aligned} \max_{i=1,\ldots,N} \limsup_{\varepsilon\to 0} \varepsilon \log\big(a_\varepsilon^i\big) &= \limsup_{\varepsilon\to 0} \varepsilon \log\Big(\max_{i=1,\ldots,N} a_\varepsilon^i\Big) \\ &\leq \limsup_{\varepsilon\to 0} \varepsilon \log\left(\sum_{i=1}^N a_\varepsilon^i\right) \\ &\leq \limsup_{\varepsilon\to 0} \varepsilon \log(N) + \max_{i=1,\ldots,N} \limsup_{\varepsilon\to 0} \varepsilon \log\big(a_\varepsilon^i\big) \\ &= \max_{i=1,\ldots,N} \limsup_{\varepsilon\to 0} \varepsilon \log\big(a_\varepsilon^i\big). \end{aligned}$$

□

Example 23.10 Let $X_1, X_2, \ldots$ be i.i.d. real random variables that satisfy the condition of Cramér's theorem (Theorem 23.3); i.e., $\Lambda(t) = \log(\mathbf{E}[e^{tX_1}]) < \infty$ for every $t \in \mathbb{R}$. Furthermore, let $S_n = X_1 + \ldots + X_n$ for every n. We will show that Cramér's theorem implies that $P_n := \mathbf{P}_{S_n/n}$ satisfies an LDP with rate n and with good rate function $I(x) = \Lambda^*(x) := \sup_{t\in\mathbb{R}}(tx - \Lambda(t))$. Without loss of generality, we can assume that $\mathbf{E}[X_1] = 0$. The function I is lower semicontinuous, strictly convex (in the interval where it is finite) and has its unique minimum at $I(0) = 0$. By convexity, we have $I(y) > I(x)$ whenever $y > x \geq 0$ or $y < x \leq 0$.

Cramér's theorem says that $\lim_{n\to\infty}\frac{1}{n}\log(P_n([x,\infty)))=-I(x)$ for $x>0$ and (by symmetry) $\lim_{n\to\infty}\frac{1}{n}\log(P_n((-\infty,x]))=-I(x)$ for $x<0$. Clearly, for $x>0$,

$$\begin{aligned}-I(x)&\ge\liminf_{n\to\infty}\frac{1}{n}\log P_n\big((x,\infty)\big)\\&\ge\sup_{y>x}\liminf_{n\to\infty}\frac{1}{n}\log P_n\big([y,\infty)\big)=-\inf_{y>x}I(y).\end{aligned}$$

Similarly, $\liminf_{n\to\infty}\frac{1}{n}\log P_n((-\infty,x))\ge-\inf_{y<x}I(y)$ for $x<0$. Furthermore, by the law of large numbers, for any $x>0$, we have

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\log P_n\big((-x,\infty)\big)&=\lim_{n\to\infty}\frac{1}{n}\log P_n\big([-x,\infty)\big)\\&=\lim_{n\to\infty}\frac{1}{n}\log P_n\big((-\infty,x)\big)=\lim_{n\to\infty}\frac{1}{n}\log P_n\big((-\infty,x]\big)\\&=0=-I(0).\end{aligned}$$

The main work has been done by showing that the family $(P_n)_{n\in\mathbb{N}}$ satisfies conditions (LDP 1) and (LDP 2) at least for unbounded intervals. It remains to show by some standard arguments (LDP 1) and (LDP 2) for *arbitrary* open and closed sets, respectively.

First assume that $C\subset\mathbb{R}$ is closed. Define $x_+:=\inf(C\cap[0,\infty))$ as well as $x_-:=\sup(C\cap(-\infty,0])$. By monotonicity of I, on $(-\infty,0]$ and $[0,\infty)$, we get $\inf I(C)=I(x_-)\wedge I(x_+)$ (with the convention $I(-\infty)=I(\infty)=\infty$). If $x_-=0$ or $x_+=0$, then $\inf(I(C))=0$, and (LDP 2) holds trivially. Now let $x_-<0<x_+$.

Using Lemma 23.9, we get

$$\begin{aligned}&\limsup_{n\to\infty}\frac{1}{n}\log P_n(C)\\&\quad\le\limsup_{n\to\infty}\frac{1}{n}\log\big(P_n\big((-\infty,x_-]\big)+P_n\big([x_+,\infty)\big)\big)\\&\quad=\max\bigg\{\limsup_{n\to\infty}\frac{1}{n}\log P_n\big((-\infty,x_-]\big),\limsup_{n\to\infty}\frac{1}{n}\log P_n\big([x_+,\infty)\big)\bigg\}\\&\quad=\max\big\{-I(x_-),-I(x_+)\big\}=-\inf I(C).\end{aligned}$$

This shows (LDP 2).

Now let $U\subset\mathbb{R}$ be open. Let $x\in U\cap[0,\infty)$ with $I(x)<\infty$ (if such an x exists). Then there exists an $\varepsilon>0$ with $(x-\varepsilon,x+\varepsilon)\subset U$. Now

$$\begin{aligned}\liminf_{n\to\infty}\frac{1}{n}\log P_n\big((x-\varepsilon,\infty)\big)&\ge-I(x)>-I(x+\varepsilon)\\&=\lim_{n\to\infty}\frac{1}{n}\log P_n\big([x+\varepsilon,\infty)\big).\end{aligned}$$

Therefore,

$$\begin{aligned}\liminf_{n\to\infty}\frac{1}{n}\log P_n(U) &\geq \liminf_{n\to\infty}\frac{1}{n}\log P_n\big((x-\varepsilon,x+\varepsilon)\big)\\ &= \liminf_{n\to\infty}\frac{1}{n}\log\big(P_n\big((x-\varepsilon,\infty)\big)-P_n\big([x+\varepsilon,\infty)\big)\big)\\ &= \liminf_{n\to\infty}\frac{1}{n}\log\big(P_n\big((x-\varepsilon,\infty)\big)\big)\geq -I(x).\end{aligned}$$

Similarly, this also holds for $x\in U\cap(-\infty,0)$ with $I(x)<\infty$; hence

$$\liminf_{n\to\infty}\frac{1}{n}\log P_n(U)\geq -\inf I(U).$$

This shows the lower bound (LDP 1). ◊

In fact, the condition $\Lambda(t)<\infty$ for all $t\in\mathbb{R}$ can be dropped. Since $\Lambda(0)=0$, we have $\Lambda^*(x)\geq 0$ for every $x\in\mathbb{R}$. The map Λ^* is a convex rate function but is, in general, not a good rate function. We quote the following strengthening of Cramér's Theorem (see [31, Theorem 2.2.3]).

Theorem 23.11 (Cramér) *If $X_1,X_2,\ldots$ are i.i.d. real random variables, then $(\mathbf{P}_{S_n/n})_{n\in\mathbb{N}}$ satisfies an LDP with rate function Λ^*.*

Exercise 23.2.1 Let $E=\mathbb{R}$. Show that $\mu_\varepsilon:=\mathcal{N}_{0,\varepsilon}$ satisfies an LDP with good rate function $I(x)=x^2/2$. Further, show that strict inequality can hold in the *upper bound* (LDP 2).

Exercise 23.2.2 Let $E=\mathbb{R}$. Show that $\mu_\varepsilon:=\mathcal{N}_{0,\varepsilon^2}$ satisfies an LDP with good rate function $I(x)=\infty\cdot\mathbb{1}_{\mathbb{R}\setminus\{0\}}(x)$. Further, show that strict inequality can hold in the *lower bound* (LDP 1).

Exercise 23.2.3 Let $E=\mathbb{R}$. Show that $\mu_\varepsilon:=\frac{1}{2}\mathcal{N}_{-1,\varepsilon}+\frac{1}{2}\mathcal{N}_{1,\varepsilon}$ satisfies an LDP with good rate function $I(x)=\frac{1}{2}\min((x+1)^2,(x-1)^2)$.

Exercise 23.2.4 Compute Λ and Λ^* in the case $X_1\sim\exp_\theta$ for $\theta>0$. Interpret the statement of Theorem 23.11 in this case. Check that Λ^* has its unique zero at $\mathbf{E}[X_1]$. (Result: $\Lambda^*(x)=\theta x-\log(\theta x)-1$ if $x>0$ and $=\infty$ otherwise.)

Exercise 23.2.5 Compute Λ and Λ^* for the case where X_1 is Cauchy distributed and interpret the statement of Theorem 23.11.

Exercise 23.2.6 Let $X_\lambda\sim\text{Poi}_\lambda$ for every $\lambda>0$. Show that $\mu_\varepsilon:=\mathbf{P}_{\varepsilon X_{\lambda/\varepsilon}}$ satisfies an LDP with good rate function $I(x)=x\log(x/\lambda)+\lambda-x$ for $x\geq 0$ (and $=\infty$ otherwise).

Exercise 23.2.7 Let $(X_t)_{t\geq 0}$ be a random walk on $\mathbb{Z}$ in continuous time that makes a jump to the right with rate $\frac{1}{2}$ and a jump to the left also with rate $\frac{1}{2}$. Show that $(\mathbf{P}_{\varepsilon X_{1/\varepsilon}})_{\varepsilon>0}$ satisfies an LDP with convex good rate function

$$I(x) = 1 + x \operatorname{arc\,sinh}(x) - \sqrt{1+x^2}.$$

23.3 Sanov's Theorem

This section is close to the exposition in [31].

We present a large deviations principle that, unlike Cramér's theorem, is not based on a linear space. Rather, we consider empirical distributions of independent random variables with values in a finite set Σ, which often is called an *alphabet*.

Let μ be a probability measure on Σ with $\mu(\{x\}) > 0$ for any $x \in \Sigma$. Further, let $X_1, X_2, \ldots$ be i.i.d. random variables with values in Σ and with distribution $\mathbf{P}_{X_1} = \mu$. We will derive a large deviations principle for the empirical measures

$$\xi_n(X) := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

Note that by the law of large numbers, $\mathbf{P}$-almost surely $\xi_n(X) \overset{n\to\infty}{\longrightarrow} \mu$. Hence, as the state space we get $E = \mathcal{M}_1(\Sigma)$, equipped with the metric of total variation $d(\mu, \nu) = \|\mu - \nu\|_{TV}$. (As Σ is finite, in E vague convergence, weak convergence and convergence in total variation coincide.) Further, let

$$E_n := \big\{\mu \in \mathcal{M}_1(\Sigma) : n\mu\big(\{x\}\big) \in \mathbb{N}_0 \text{ for every } x \in \Sigma\big\}$$

be the range of the random variables $\xi_n(X)$.

Recall that the *entropy* of μ is defined by

$$H(\mu) := -\int \log\big(\mu\big(\{x\}\big)\big)\mu(dx).$$

If $\nu \in \mathcal{M}_1(\Sigma)$, then we define the *relative entropy* (or *Kullback–Leibler information*, see [104]) of ν given μ by

$$H(\nu \mid \mu) := \int \log\left(\frac{\nu(\{x\})}{\mu(\{x\})}\right)\nu(dx). \tag{23.13}$$

Since $\mu(\{x\}) > 0$ for all $x \in \Sigma$, the integrand ν-a.s. is finite and hence the integral also is finite. A simple application of Jensen's inequality yields $H(\mu) \geq 0$ and $H(\nu \mid \mu) \geq 0$ (see Lemma 5.26 and Exercise 5.3.3). Furthermore, $H(\nu \mid \mu) = 0$ if and only if $\nu = \mu$. In addition, clearly,

$$H(\nu \mid \mu) + H(\nu) = -\int \log\big(\mu\big(\{x\}\big)\big)\nu(dx). \tag{23.14}$$

Since the map $\nu \mapsto I_\mu(\nu) := H(\nu \mid \mu)$ is continuous, I_μ is a rate function.

Lemma 23.12 *For every $n \in \mathbb{N}$ and $\nu \in E_n$, we have*

$$(n+1)^{-\#\Sigma} e^{-nH(\nu|\mu)} \leq \mathbf{P}\big[\xi_n(X) = \nu\big] \leq e^{-nH(\nu|\mu)}. \tag{23.15}$$

Proof We consider the set of possible values for the n-tuple $(X_1, \ldots, X_n)$ such that $\xi_n(X) = \nu$:

$$A_n(\nu) := \left\{ k = (k_1, \ldots, k_n) \in \Sigma^n : \frac{1}{n} \sum_{i=1}^{n} \delta_{k_i} = \nu \right\}.$$

For every $k \in A_n(\nu)$, we have (compare (23.14))

$$\begin{aligned}
\mathbf{P}\big[\xi_n(X) = \nu\big] &= \#A_n(\nu)\mathbf{P}[X_1 = k_1, \ldots, X_n = k_n] \\
&= \#A_n(\nu) \prod_{x \in \Sigma} \mu\big(\{x\}\big)^{n\nu(\{x\})} \\
&= \#A_n(\nu) \exp\left(n \int \nu(dx) \log \mu\big(\{x\}\big) \right) \\
&= \#A_n(\nu) \exp\big(-n\big[H(\nu) + H(\nu \mid \mu)\big]\big).
\end{aligned}$$

Now let $Y_1, Y_2, \ldots$ be i.i.d. random variables with values in Σ and with distribution $\mathbf{P}_{Y_1} = \nu$. As in the calculation for X, we obtain (since $H(\nu \mid \nu) = 0$)

$$1 \geq \mathbf{P}\big[\xi_n(Y) = \nu\big] = \#A_n(\nu) e^{-nH(\nu)};$$

hence $\#A_n(\nu) \leq e^{nH(\nu)}$. This implies the second inequality in (23.15).

The random variable $n\xi_n(Y)$ has the multinomial distribution with parameters $(n\nu(\{x\}))_{x \in \Sigma}$. Hence the map $E_n \to [0,1]$, $\nu' \mapsto \mathbf{P}[\xi_n(Y) = \nu']$ is maximal at $\nu' = \nu$. Therefore,

$$\#A_n(\nu) = e^{nH(\nu)} \mathbf{P}\big[\xi_n(Y) = \nu\big] \geq \frac{e^{nH(\nu)}}{\#E_n} \geq (n+1)^{-\#\Sigma} e^{nH(\nu)}.$$

This implies the first inequality in (23.15). □

We come to the main theorem of this section, Sanov's theorem (see [149, 150]).

Theorem 23.13 (Sanov (1957)) *Let $X_1, X_2, \ldots$ be i.i.d. random variables with values in the finite set Σ and with distribution μ. Then the family $(\mathbf{P}_{\xi_n(X)})_{n \in \mathbb{N}}$ of distributions of empirical measures satisfies an LDP with rate n and rate function $I_\mu := H(\cdot|\mu)$.*

Proof By Lemma 23.12, for every $A \subset E$,

$$\begin{aligned}\mathbf{P}\big[\xi_n(X) \in A\big] &= \sum_{\nu \in A \cap E_n} \mathbf{P}\big[\xi_n(X) = \nu\big] \\ &\le \sum_{\nu \in A \cap E_n} e^{-nH(\nu|\mu)} \\ &\le \#(A \cap E_n) \exp\big(-n \inf I_\mu(A \cap E_n)\big) \\ &\le (n+1)^{\#\Sigma} \exp\big(-n \inf I_\mu(A)\big).\end{aligned}$$

Therefore,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}\big[\xi_n(X) \in A\big] \le -\inf I_\mu(A).$$

Hence the upper bound in the LDP holds (even for arbitrary A).

Similarly, we can use the first inequality in Lemma 23.12 to get

$$\mathbf{P}\big[\xi_n(X) \in A\big] \ge (n+1)^{-\#\Sigma} \exp\big(-n \inf I_\mu(A \cap E_n)\big)$$

and thus

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\big[\xi_n(X) \in A\big] \ge -\limsup_{n\to\infty} \inf I_\mu(A \cap E_n). \tag{23.16}$$

Note that, in this inequality, in the infimum we cannot simply replace $A \cap E_n$ by A. However, we show that, for open A this can be done at least asymptotically. Hence, let $A \subset E$ be open. For $\nu \in A$, there is an $\varepsilon > 0$ with $B_\varepsilon(\nu) \subset A$. For $n \ge (2\#\Sigma)/\varepsilon$, we have $E_n \cap B_\varepsilon(\nu) \ne \emptyset$ and hence there exists a sequence $\nu_n \overset{n\to\infty}{\longrightarrow} \nu$ with $\nu_n \in E_n \cap A$ for large $n \in \mathbb{N}$. As I_μ is continuous, we have

$$\limsup_{n\to\infty} \inf I_\mu(A \cap E_n) \le \lim_{n\to\infty} I_\mu(\nu_n) = I_\mu(\nu).$$

Since $\nu \in A$ is arbitrary, we get $\limsup_{n\to\infty} \inf I_\mu(A \cap E_n) = \inf I_\mu(A)$. □

Example 23.14 Let $\Sigma = \{-1, 1\}$ and let $\mu = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ be the uniform distribution on Σ. Define $m = m(\nu) := \nu(\{1\}) - \nu(\{-1\})$. Then the relative entropy of $\nu \in \mathcal{M}_1(\Sigma)$ is

$$H(\nu \mid \mu) = \frac{1+m}{2} \log(1+m) + \frac{1-m}{2} \log(1-m).$$

Note that this is the rate function from Theorem 23.1.

Next we describe formally the connection between the LDPs of Sanov and Cramér that was indicated in the previous example. To this end, we use Sanov's theorem to derive a version of Cramér's theorem for $\mathbb{R}^d$-valued random variables taking only finitely many different values. ◊

Example 23.15 Let $\Sigma \subset \mathbb{R}^d$ be finite and let μ be a probability measure on Σ. Further, let $X_1, X_2, \ldots$ be i.i.d. random variables with values in Σ and distribution $\mathbf{P}_{X_1} = \mu$. Define $S_n = X_1 + \ldots + X_n$ for every $n \in \mathbb{N}$. Let $\Lambda(t) = \log \mathbf{E}[e^{\langle t, X_1 \rangle}]$ for $t \in \mathbb{R}^d$ (which is finite since Σ is finite) and $\Lambda^*(x) = \sup_{t \in \mathbb{R}^d}(\langle t, x\rangle - \Lambda(t))$ for $x \in \mathbb{R}^d$.

We show that $(\mathbf{P}_{S_n/n})_{n \in \mathbb{N}}$ satisfies an LDP with rate n and rate function Λ^*.

Let $\xi_n(X)$ be the empirical measure of $X_1, \ldots, X_n$. Let $E := \mathcal{M}_1(\Sigma)$. Define the map

$$m : E \to \mathbb{R}^d, \qquad \nu \mapsto \int x \nu(dx) = \sum_{x \in \Sigma} x \nu\big(\{x\}\big).$$

That is, m maps ν to its first moment. Clearly, $\frac{1}{n} S_n = m(\xi_n(X))$. For $x \in \mathbb{R}^d$ and $A \subset \mathbb{R}^d$, define

$$E_x := m^{-1}\big(\{x\}\big) = \big\{\nu \in E : m(\nu) = x\big\}$$

and

$$E_A = m^{-1}(A) = \big\{\nu \in E : m(\nu) \in A\big\}.$$

The map $\nu \mapsto m(\nu)$ is continuous; hence E_A is open (respectively closed) if A is open (respectively closed). Let $\tilde{I}(x) := \inf I_\mu(E_x)$ (where $I_\mu(\nu) = H(\nu \mid \mu)$ is the relative entropy). Then, by Sanov's theorem for open $U \subset \mathbb{R}^d$,

$$\begin{aligned} \liminf_{n \to \infty} \frac{1}{n} \log \mathbf{P}_{S_n/n}(U) &= \liminf_{n \to \infty} \frac{1}{n} \log \mathbf{P}_{\xi_n(X)}\big(m^{-1}(U)\big) \\ &\geq -\inf I_\mu\big(m^{-1}(U)\big) = -\inf \tilde{I}(U). \end{aligned}$$

Similarly, for closed $C \subset \mathbb{R}^d$, we have

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbf{P}_{S_n/n}(C) \leq -\inf \tilde{I}(C).$$

In other words, $(\mathbf{P}_{S_n/n})_{n \in \mathbb{N}}$ satisfies an LDP with rate n and rate function $\tilde{I}$. Hence, it only remains to show that $\tilde{I} = \Lambda^*$.

Note that $t \mapsto \Lambda(t)$ is differentiable (with derivative Λ') and is strictly convex. Hence the variational problem for $\Lambda^*(x)$ admits a unique maximizer $t^*(x)$. More precisely,

$$\Lambda^*(x) = \big\langle t^*(x), x\big\rangle - \Lambda\big(t^*(x)\big),$$

$\Lambda^*(x) > \langle t, x\rangle - \Lambda(t)$ for all $t \neq t^*(x)$, and $\Lambda'(t^*(x)) = x$. By Jensen's inequality, for every $\nu \in \mathcal{M}_1(\Sigma)$,

$$\begin{aligned}\Lambda(t) &= \log \int e^{\langle t,y\rangle}\mu(dy)\\ &= \log \int \left(e^{\langle t,y\rangle}\frac{\mu(\{y\})}{\nu(\{y\})}\right)\nu(dy)\\ &\geq \int \log\left(e^{\langle t,y\rangle}\frac{\mu(\{y\})}{\nu(\{y\})}\right)\nu(dy)\\ &= \big\langle t, m(\nu)\big\rangle - H(\nu \,|\, \mu)\end{aligned}$$

with equality if and only if $\nu = \nu_t$, where $\nu_t(\{y\}) = \mu(\{y\})e^{\langle t,y\rangle - \Lambda(t)}$. Hence,

$$\langle t, x\rangle - \Lambda(t) \leq \inf_{\nu \in E_x} H(\nu \,|\, \mu)$$

with equality if $\nu_t \in E_x$. However, we now know that $m(\nu_t) = \Lambda'(t)$; hence we have $\nu_{t^*(x)} \in E_x$ and thus

$$\Lambda^*(x) = \big\langle t^*(x), x\big\rangle - \Lambda\big(t^*(x)\big) = \inf_{\nu \in E_x} H(\nu \,|\, \mu) = \tilde{I}(x). \qquad \Diamond$$

The method of the proof that we applied in the last example to derive the LDP with rate function $\tilde{I}$ is called a *contraction principle*. We formulate this principle as a theorem.

Theorem 23.16 (Contraction principle) *Assume the family* $(\mu_\varepsilon)_{\varepsilon>0}$ *of probability measures on* E *satisfies an LDP with rate function* I. *If* F *is a topological space and* $m : E \to F$ *is continuous, then the image measures* $(\mu_\varepsilon \circ m^{-1})_{\varepsilon>0}$ *satisfy an LDP with rate function* $\tilde{I}(x) = \inf I(m^{-1}(\{x\}))$.

23.4 Varadhan's Lemma and Free Energy

Assume that $(\mu_\varepsilon)_{\varepsilon>0}$ is a family of probability measures that satisfies an LDP with rate function I. In particular, we know that, for small $\varepsilon > 0$, the mass of μ_ε is concentrated around the zeros of I. In statistical physics, one is often interested in integrating with respect to μ_ε (where $1/\varepsilon$ is interpreted as "size of the system") functions that attain their maximal values away from the zeros of I. In addition, these functions are exponentially scaled with $1/\varepsilon$. Hence the aim is to study the asymptotics of $Z_\varepsilon^\phi := \int e^{\phi(x)/\varepsilon}\mu_\varepsilon(dx)$ as $\varepsilon \to 0$. Under some mild conditions on the continuity of ϕ, the main contribution to the integral comes from those points x that are not too unlikely (for μ_ε) and for which at the same time $\phi(x)$ is large. That is, those x for which $\phi(x) - I(x)$ is close to its maximum. These contributions are quantified in terms of the *tilted* probability measures $\mu_\varepsilon^\phi(dx) = (Z_\varepsilon^\phi)^{-1}e^{\phi(x)/\varepsilon}\mu_\varepsilon(dx)$,

$\varepsilon > 0$, for which we derive an LDP. As an application, we get the statistical physics principle of minimising the free energy. As an example, we analyze the Weiss ferromagnet.

We start with a lemma that is due to Varadhan [166].

Theorem 23.17 (Varadhan's Lemma (1966)) *Let I be a good rate function and let $(\mu_\varepsilon)_{\varepsilon>0}$ be a family of probability measures on E that satisfies an LDP with rate function I. Further, let $\phi : E \to \mathbb{R}$ be continuous and assume that*

$$\inf_{M>0} \limsup_{\varepsilon\to 0} \varepsilon \log \int e^{\phi(x)/\varepsilon} \mathbb{1}_{\{\phi(x)\geq M\}} \mu_\varepsilon(dx) = -\infty. \tag{23.17}$$

Then

$$\lim_{\varepsilon\to 0} \varepsilon \log \int e^{\phi(x)/\varepsilon} \mu_\varepsilon(dx) = \sup_{x\in E} \big(\phi(x) - I(x)\big). \tag{23.18}$$

Remark 23.18 (Moment condition) The tail condition (23.17) holds if there exists an $\alpha > 1$ such that

$$\limsup_{\varepsilon\to 0} \varepsilon \log \int e^{\alpha\phi/\varepsilon} d\mu_\varepsilon < \infty. \tag{23.19}$$

Indeed, for every $M \in \mathbb{R}$, we have

$$\begin{aligned}\varepsilon \log \int e^{\phi(x)/\varepsilon} \mathbb{1}_{\{\phi(x)\geq M\}} \mu_\varepsilon(dx) &= M + \varepsilon \log \int e^{(\phi(x)-M)/\varepsilon} \mathbb{1}_{\{\phi(x)\geq M\}} \mu_\varepsilon(dx)\\ &\leq M + \varepsilon \log \int e^{\alpha(\phi(x)-M)/\varepsilon} \mu_\varepsilon(dx)\\ &= -(\alpha-1)M + \varepsilon \log \int e^{\alpha\phi(x)/\varepsilon} \mu_\varepsilon(dx).\end{aligned}$$

Together with (23.19), this implies (23.17). ♢

Proof We use different arguments to show that the right-hand side of (23.18) is a lower and an upper bound for the left-hand side.

Lower bound. For any $x \in E$ and $r > 0$, we have

$$\begin{aligned}\liminf_{\varepsilon\to 0} \varepsilon \log \int e^{\phi/\varepsilon}\, d\mu_\varepsilon &\geq \liminf_{\varepsilon\to 0} \varepsilon \log \int_{B_r(x)} e^{\phi/\varepsilon}\, d\mu_\varepsilon\\ &\geq \inf \phi\big(B_r(x)\big) - I(x) \xrightarrow{r\to 0} \phi(x) - I(x).\end{aligned}$$

Upper bound. For $M > 0$ and $\varepsilon > 0$, define

$$F_M^\varepsilon := \int_{\{\phi\geq M\}} e^{\phi(x)/\varepsilon} \mu_\varepsilon(dx) \quad \text{and} \quad G_M^\varepsilon := \int_{\{\phi<M\}} e^{\phi(x)/\varepsilon} \mu_\varepsilon(dx).$$

Define

$$F_M := \limsup_{\varepsilon\to 0} \varepsilon \log F_M^\varepsilon \quad \text{and} \quad G_M := \limsup_{\varepsilon\to 0} \varepsilon \log G_M^\varepsilon .$$

By Lemma 23.9, for any $M > 0$,

$$\limsup_{\varepsilon\to 0} \varepsilon \log \int e^{\phi(x)/\varepsilon} \mu_\varepsilon(dx) = F_M \vee G_M .$$

As by assumption $\inf_{M>0} F_M = -\infty$, it is enough to show that

$$\sup_{M>0} G_M \le \sup_{x\in E} \big(\phi(x) - I(x)\big). \tag{23.20}$$

Let $\delta > 0$. For any $x \in E$ there is an $r(x) > 0$ with

$$\inf I\big(B_{2r(x)}(x)\big) \ge I(x) - \delta \quad \text{and} \quad \sup \phi\big(B_{2r(x)}(x)\big) \le \phi(x) + \delta .$$

Let $a \ge 0$. Since I is a *good* rate function, the level set $K := I^{-1}([0,a])$ is compact. Thus we can find finitely many $x_1, \ldots, x_N \in I^{-1}([0,a])$ such that $\bigcup_{i=1}^N B_{r(x_i)}(x_i) \supset K$. Therefore,

$$\begin{aligned} G_M^\varepsilon &\le \int_{\{\phi<M\}\cap K^c} e^{\phi(x)/\varepsilon} \mu_\varepsilon(dx) + \sum_{i=1}^N \int_{\{\phi<M\}\cap B_{r(x_i)}(x_i)} e^{\phi(x)/\varepsilon} \mu_\varepsilon(dx) \\ &\le e^{M/\varepsilon} \mu_\varepsilon\big(K^c\big) + \sum_{i=1}^N e^{(\phi(x_i)\wedge M + \delta)/\varepsilon} \mu_\varepsilon\big(B_{r(x_i)}(x_i)\big) \\ &= e^{(M+\varepsilon \log(\mu_\varepsilon(K^c)))/\varepsilon} + \sum_{i=1}^N e^{(\phi(x_i)\wedge M+\delta+\varepsilon\log(\mu_\varepsilon(B_{r(x_i)}(x_i))))/\varepsilon} . \end{aligned}$$

Using Lemma 23.9 and the LDP, we infer

$$\begin{aligned} G_M &\le (M-a) \vee \max_{i=1,\ldots,N} \big(\phi(x_i) - I(x_i) + 2\delta\big) \\ &\le (M-a) \vee \sup_{x\in E} \big(\phi(x) - I(x)\big) + 2\delta . \end{aligned}$$

By letting first $\delta \downarrow 0$ and then $a \uparrow \infty$, we obtain (23.20). □

Theorem 23.19 (Tilted LDP) *Assume that $(\mu_\varepsilon)_{\varepsilon>0}$ satisfies an LDP with good rate function I. Further, let $\phi : E \to \mathbb{R}$ be a continuous function that satisfies condition (23.17). Define $Z_\varepsilon^\phi := \int e^{\phi/\varepsilon}\, d\mu_\varepsilon$ and $\mu_\varepsilon^\phi \in \mathcal{M}_1(E)$ by*

$$\mu_\varepsilon^\phi(dx) = \big(Z_\varepsilon^\phi\big)^{-1} e^{\phi(x)/\varepsilon} \mu_\varepsilon(dx).$$

Further, define $I^\phi : E \to [0,\infty]$ by

$$I^\phi(x) = \sup_{z\in E}\big(\phi(z) - I(z)\big) - \big(\phi(x) - I(x)\big). \tag{23.21}$$

Then $(\mu_\varepsilon^\phi)_{\varepsilon>0}$ satisfies an LDP with rate function I^ϕ.

Proof This is left as an exercise. (Compare [32, Exercise 2.1.24], see also [43, Section II.7].) □

Varadhan's lemma has various applications in statistical physics. Consider a Polish space Σ that is interpreted as the space of possible states of a particle. Further, let $\lambda \in \mathcal{M}_1(\Sigma)$ be a distribution that is understood as the *a priori* distribution of this particle if the influence of energy could be neglected. If Σ is finite or is a bounded subset of an $\mathbb{R}^d$, then by symmetry, typically λ is the uniform distribution on Σ. If we place n indistinguishable particles independently according to λ on the random positions $z_1, \ldots, z_n \in \Sigma$, then the *state* of this ensemble can be described by $x := \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$. Denote by $\mu_n^0 \in \mathcal{M}_1(\mathcal{M}_1(\Sigma))$ the corresponding *a priori* distribution of x; that is, of the n-particle system.

Now we introduce the hypothesis that the energy $U_n(x)$ of a state has the form $U_n(x) = nU(x)$, where $U(x)$ is the average energy of one particle of the ensemble in state x.

Let $T > 0$ be the temperature of the system and let $\beta := 1/T$ be the so-called *inverse temperature*. In statistical physics, a key quantity is the so-called *partition function*

$$Z_n^\beta := \int e^{-\beta U_n}\, d\mu_n^0.$$

A postulate of statistical physics is that the distribution of the state x is the *Boltzmann distribution*:

$$\mu_n^\beta(dx) = \big(Z_n^\beta\big)^{-1} e^{-\beta U_n(x)} \mu_n^0(dx). \tag{23.22}$$

Varadhan's lemma (more precisely, the tilted LDP) and Sanov's theorem are the keys to building a connection to the variational principle for the free energy. For simplicity, assume that Σ is a finite set and $\lambda = \mathcal{U}_\Sigma$ is the uniform distribution on Σ. By Sanov's theorem, $(\mu_n^0)_{n\in\mathbb{N}}$ satisfies an LDP with rate n and rate function $I(x) = H(x|\lambda)$, where $H(x|\lambda)$ is the relative entropy of x with respect to λ. By (23.14), we have $H(x|\lambda) = \log(\#\Sigma) - H(x)$, where $H(x)$ is the entropy of x.

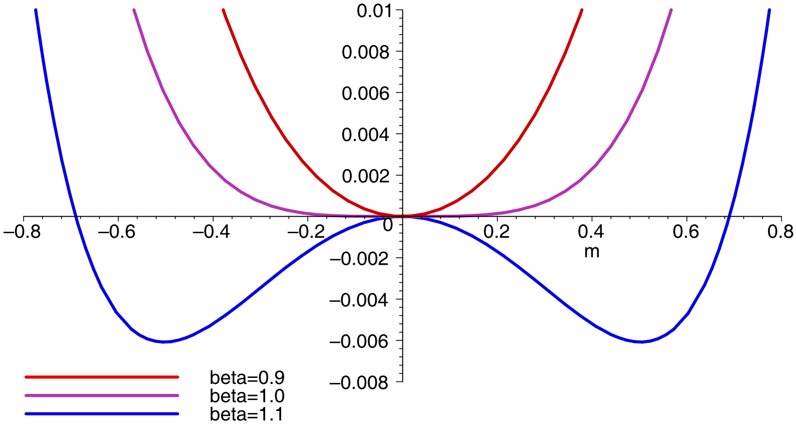

Fig. 23.2 The shifted free energy $F^\beta(m) - F^\beta(0)$ of the Weiss ferromagnet without exterior field $(h = 0)$

Define the *free energy* (or *Helmholtz potential*) per particle as

$$F^\beta(x) := U(x) - \beta^{-1} H(x).$$

The theorem on the tilted LDP yields that the sequence of Boltzmann distributions $(\mu_n^\beta)_{n\in\mathbb{N}}$ satisfies an LDP with rate n and rate function

$$I^\beta(x) = F^\beta(x) - \inf_{y\in\mathcal{M}_1(\Sigma)} F^\beta(y).$$

Thus, for large n, the Boltzmann distribution is concentrated on those x that minimize the free energy. For different temperatures (that is, for different values of β) these can be very different states. This is the reason for *phase transitions* at critical temperatures (e.g., melting ice).

Example 23.20 We consider the *Weiss ferromagnet*. This is a microscopic model for a magnet that assumes that each of n indistinguishable magnetic particles has one of two possible orientations $\sigma_i \in \Sigma = \{-1, +1\}$. The mean magnetization $m = \frac{1}{n}\sum_{i=1}^n \sigma_i$ describes the state of the system completely (as the particles are indistinguishable). Macroscopically, this is the quantity that can be measured. The basic idea is that it is energetically favorable for particles to be oriented in the same direction. We ignore the spatial structure and assume that any particle interacts with any other particle in the same way. This is often called the *mean field* assumption. In addition, we assume that there is an exterior magnetic field of strength h. Thus up to constants the average energy of a particle is

$$U(m) = -\frac{1}{2}m^2 - hm.$$

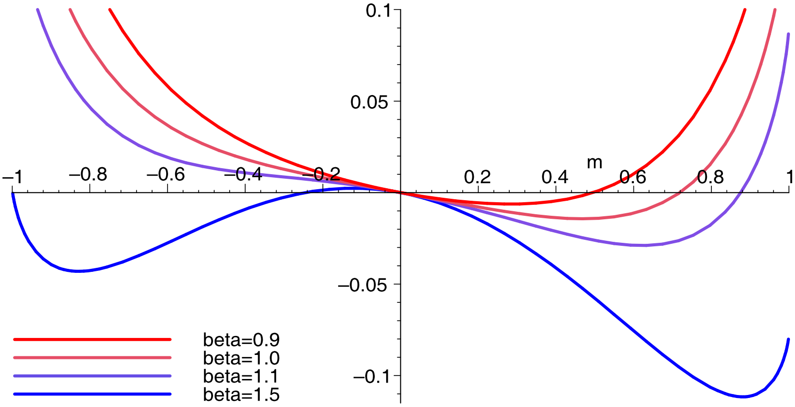

Fig. 23.3 Shifted free energy $F^\beta(m) - F^\beta(0)$ of the Weiss ferromagnet with exterior field $h = 0.04$

The entropy of the state m is

$$H(m) = -\frac{1+m}{2}\log\left(\frac{1+m}{2}\right) - \frac{1-m}{2}\log\left(\frac{1-m}{2}\right).$$

Hence the average free energy of a particle is

$$F^\beta(m) = -\frac{1}{2}m^2 - hm + \beta^{-1}\left[\frac{1+m}{2}\log\left(\frac{1+m}{2}\right) + \frac{1-m}{2}\log\left(\frac{1-m}{2}\right)\right].$$

In order to obtain the minima of F^β, we compute the derivative

$$0 \overset{!}{=} \tfrac{d}{dm}F^\beta(m) = -m - h + \beta^{-1}\operatorname{arc\,tanh}(m).$$

Hence, m solves the equation

$$m = \tanh\big(\beta(m+h)\big). \tag{23.23}$$

In the case $h = 0$, $m = 0$ is a solution of (23.23) for any β. If $\beta \le 1$, then this is the only solution and F^β attains its global minimum at $m = 0$. If $\beta > 1$, then (23.23) has two other solutions, $m_-^{\beta,0} \in (-1, 0)$ and $m_+^{\beta,0} = -m_-^{\beta,0}$, whose values can only be computed numerically.

In this case, F^β has a local maximum at 0 and has global minima $m_\pm^{\beta,0}$. For large n, only those values of m for which F^β is close to its minimal value can be attained and thus the distribution is concentrated around 0 if $\beta \le 1$ and around $m_\pm^{\beta,0}$ if $\beta > 1$. In the latter case, the absolute value of the mean magnetization is $|m_\pm^{\beta,0}| = m_+^{\beta,0} > 0$. Hence, there is a *phase transition* between the high temperature phase ($\beta \le 1$) without magnetization and the low temperature phase ($\beta > 1$) where so-called spontaneous magnetization occurs (that is, magnetization without an exterior field).

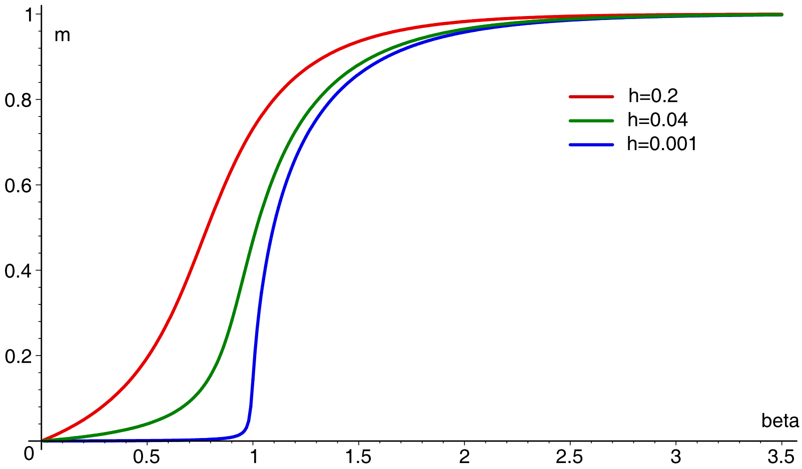

Fig. 23.4 Weiss ferromagnet: magnetization $m^{\beta,h}$ as a function of β

If $h \neq 0$, then F^β does not have a minimum at $m = 0$. Rather, F^β is asymmetric and has a global minimum $m^{\beta,h}$ with the same sign as h. Furthermore, for large β, there is another minimum with the opposite sign. Again, the exact values can only be computed numerically. However, for high temperatures (small β), we can approximate $m^{\beta,h}$ using the approximation $\tanh(\beta(m+h)) \approx \beta(m+h)$. Hence we get

$$m^{\beta,h} \approx \frac{h}{\beta^{-1}-1} = \frac{h}{T-T_c} \quad \text{for } T \to \infty, \tag{23.24}$$

where the *Curie temperature* $T_c = 1$ is the critical temperature for spontaneous magnetization. The relation (23.24) is called the *Curie–Weiss law*. ◊