Chapter VIII

Level-2 Large Deviations for I.I.D. Random Vectors

VIII.1. Introduction

Theorem II.4.3 stated the level-2 large deviation property for i.i.d. random vectors taking values in $\mathbb{R}^d$. This theorem follows from the results contained in Donsker and Varadhan (1975a, 1976a), which prove level-2 large deviation properties for Markov processes taking values in a complete separable metric space.[1] In Chapter VIII, we will give an elementary, self-contained proof of Theorem II.4.3 in the special case of i.i.d. random variables with a finite state space. This version of the theorem was applied in Chapter III to study the exponential convergence of velocity observables for the discrete ideal gas with respect to the microcanonical ensemble [Theorem III.4.4].

Theorems II.5.1 and II.5.2 stated contraction principles relating levels-1 and 2 for i.i.d. random vectors taking values in $\mathbb{R}^d$. In the chapters on statistical mechanics, the contraction principles were applied only in the case $d = 1$. Proofs for this case are given in Section VIII.3. We save for Section VIII.4 the more difficult proofs of the contraction principles for $d \geq 2$. Section VIII.4 can be skipped with no loss in continuity.

VIII.2. The Level-2 Large Deviation Theorem

We consider a sequence of i.i.d. random variables $X_1, X_2, \ldots$ with a finite state space Γ. Γ is topologized by the discrete topology. The Borel σ-field $\mathscr{B}(\Gamma)$ of Γ coincides with the set of all subsets of Γ. The empirical measure $L_n(\omega, \cdot) = n^{-1} \sum_{j=1}^{n} \delta_{X_j(\omega)}(\cdot)$, $n = 1, 2, \ldots$, takes values in the space $\mathscr{M}(\Gamma)$, which is the set of probability measures on $\mathscr{B}(\Gamma)$. The topology on $\mathscr{M}(\Gamma)$ is the topology of weak convergence. The following theorem is Theorem II.4.3 for the case of a finite state space.

Theorem VIII.2.1. *Let $X_1, X_2, \ldots$ be a sequence of* i.i.d. *random variables which take values in a set $\Gamma = \{x_1, x_2, \ldots, x_r\}$ with $x_1 < x_2 < \cdots < x_r$. Let $\rho \in \mathscr{M}(\Gamma)$ be the distribution of X_1; we assume that each $\rho_i = \rho\{x_i\} > 0$. If $\nu = \sum_{i=1}^{r} \nu_i \delta_{x_i}$ is a probability measure on $\mathscr{B}(\Gamma)$, then define the relative*

entropy of ν with respect to ρ by the formula

$$I_\rho^{(2)}(\nu) = \sum_{i=1}^{r} \nu_i \log \frac{\nu_i}{\rho_i}.$$

The following conclusions hold.

(a) $\{Q_n^{(2)}\}$, *the distributions on* $\mathcal{M}(\Gamma)$ *of the empirical measures* $\{L_n\}$, *have a large deviation property with* $a_n = n$ *and entropy function* $I_\rho^{(2)}$.

(b) $I_\rho^{(2)}(\nu)$ *is a convex function of* ν. $I_\rho^{(2)}(\nu)$ *measures the discrepancy between* ν *and* ρ *in the sense that* $I_\rho^{(2)}(\nu) \geq 0$ *with equality if and only if* $\nu = \rho$.

Part (b) of the theorem was proved in Proposition I.4.1. If A is a nonempty subset of $\mathcal{M}(\Gamma)$, then $I_\rho^{(2)}(A)$ denotes the infimum of $I_\rho^{(2)}$ over A. $I(\phi)$ equals ∞. In order to prove part (a) of the theorem, we must verify the following hypotheses.

(i) $I_\rho^{(2)}(\nu)$ is lower semicontinuous on $\mathcal{M}(\Gamma)$.
(ii) $I_\rho^{(2)}(\nu)$ has compact level sets in $\mathcal{M}(\Gamma)$.
(iii) $\limsup_{n\to\infty} n^{-1} \log Q_n^{(2)}\{K\} \leq -I_\rho^{(2)}(K)$ for each closed set K in $\mathcal{M}(\Gamma)$.
(iv) $\liminf_{n\to\infty} n^{-1} \log Q_n^{(2)}\{G\} \geq -I_\rho^{(2)}(G)$ for each open set G in $\mathcal{M}(\Gamma)$.

The set $\mathcal{M}(\Gamma)$ with the topology of weak convergence is homeomorphic to the compact convex subset of $\mathbb{R}^r$ consisting of all vectors $\nu = (\nu_1, \ldots, \nu_r)$ with $\nu_i \geq 0$ and $\sum_{i=1}^r \nu_i = 1$; that is, $\nu_n \Rightarrow \nu$ in $\mathcal{M}(\Gamma)$ if and only if the corresponding vectors converge in $\mathbb{R}^r$. Let $\mathcal{M}$ denote this compact convex subset of $\mathbb{R}^r$. Since $I_\rho^{(2)}(\nu)$ is continuous relative to $\mathcal{M}$, hypotheses (i) and (ii) follow. The vector in $\mathcal{M}$ corresponding to the empirical measure $L_n(\omega, \cdot)$ is the vector $L_n(\omega)$ whose ith component is $L_n(\omega, \{x_i\})$. The next theorem establishes a large deviation property for the distributions $\{\bar{Q}_n^{(2)}\}$ of $\{L_n(\omega)\}$ on $\mathbb{R}^r$. The entropy function $I(\nu)$ equals $I_\rho^{(2)}(\nu)$ for $\nu \in \mathcal{M}$ and equals ∞ for $\nu \in \mathbb{R}^r \setminus \mathcal{M}$. Hence

$$\limsup_{n\to\infty} \frac{1}{n} \log \bar{Q}_n^{(2)}\{K\} \leq -I(K) = -I_\rho^{(2)}(K \cap \mathcal{M}) \quad \text{for each closed set } K \text{ in } \mathbb{R}^r,$$

$$\liminf_{n\to\infty} \frac{1}{n} \log \bar{Q}_n^{(2)}\{G\} \geq -I(G) = -I_\rho^{(2)}(G \cap \mathcal{M}) \quad \text{for each open set } G \text{ in } \mathbb{R}^r.$$

Since the support of $\bar{Q}_n^{(2)}$ is contained in $\mathcal{M}$, $\bar{Q}_n^{(2)}\{A\}$ equals $\bar{Q}_n^{(2)}\{A \cap \mathcal{M}\}$ for any Borel subset A of $\mathbb{R}^r$. As the topology on $\mathcal{M}$ is its relative topology as a subset of $\mathbb{R}^r$, the last display implies that

$$\limsup_{n\to\infty} \frac{1}{n} \log \bar{Q}_n^{(2)}\{K\} \leq -I_\rho^{(2)}(K) \qquad \text{for each closed set } K \text{ in } \mathcal{M},$$

$$\liminf_{n\to\infty} \frac{1}{n} \log \bar{Q}_n^{(2)}\{G\} \geq -I_\rho^{(2)}(G) \qquad \text{for each open set } G \text{ in } \mathcal{M}.$$

These inequalities yield hypotheses (iii) and (iv) above.

Theorem VIII.2.2. *Let $\bar{Q}_n^{(2)}$ be the distribution of the random vector $L_n(\omega)$ on $\mathbb{R}^r$. Then the sequence $\{\bar{Q}_n^{(2)}; n = 1, 2, \dots\}$ has a large deviation property with $a_n = n$. The free energy function $c(t)$ and the entropy function $I(v)$ are given by*

$$c(t) = \log \sum_{i=1}^{r} e^{t_i}\rho_i \quad \text{and} \quad I(v) = \sup_{t \in \mathbb{R}^r} \{\langle t, v\rangle - c(t)\} = \begin{cases} I_\rho^{(2)}(v) & \text{for } v \in \mathcal{M}, \\ \infty & \text{for } v \notin \mathcal{M}. \end{cases}$$

Proof. We apply the large deviation theorem for random vectors, Theorem II.6.1. If t is a vector in $\mathbb{R}^r$, then define the function $f_t(x_i) = t_i$ for $x_i \in \Gamma$. We have

$$\langle t, L_n(\omega)\rangle = \frac{1}{n}\sum_{i=1}^{r} t_i \sum_{j=1}^{n} \delta_{X_j(\omega)}\{x_i\} = \frac{1}{n}\sum_{j=1}^{n} f_t(X_j(\omega)),$$

which is a sum of i.i.d. random variables. In the notation of Theorem II.6.1, W_n equals nL_n and a_n equals n. The free energy function of the sequence $\{nL_n; n = 1, 2, \dots\}$ is given by

$$c(t) = \lim_{n\to\infty} \frac{1}{n} \log E\{\exp(n\langle t, L_n\rangle)\} = \log E\{\exp f_t(X_1)\} = \log \sum_{i=1}^{r} e^{t_i}\rho_i.$$

The function $c(t)$ is differentiable for all $t \in \mathbb{R}^r$. Hence the theorem is proved once we identify the Legendre–Fenchel transform $I(v)$ of $c(t)$. The lower large deviation bound states that

$$\liminf_{n\to\infty} \frac{1}{n} \log \bar{Q}_n^{(2)}\{G\} \geq -I(G) \qquad \text{for each open set } G \text{ in } \mathbb{R}^r.$$

Since the support of $\bar{Q}_n^{(2)}$ is contained in $\mathcal{M}$, $I(G)$ equals ∞ whenever G is open and $G \cap \mathcal{M}$ is empty. Thus $I(v)$ equals ∞ for $v \notin \mathcal{M}$. The form of $I(v)$ for $v \in \mathcal{M}$ is given by part (c) of the next lemma.

Lemma VIII.2.3. *Let $\mathcal{N}$ denote the set of vectors $t \in \mathbb{R}^r$ of the form $t_i = \log(z_i/\rho_i)$ for some vector $z \in \operatorname{ri} \mathcal{M}$.* Then the following conclusions hold.*

(a) *$c(t) = \log \sum_{i=1}^{r} e^{t_i}\rho_i$ equals 0 if and only if t is in $\mathcal{N}$.*

(b) *For any vector t in $\mathbb{R}^r$, $h(t) = t - c(t)\mathbf{1}$ is in $\mathcal{N}$, where $\mathbf{1}$ is the constant vector $(1, \dots, 1)$.*

(c) *$I(v)$ equals $I_\rho^{(2)}(v) = \sum_{i=1}^{r} v_i \log(v_i/\rho_i)$ for any $v \in \mathcal{M}$.*

Proof (a) If t is in $\mathcal{N}$, then $\log \sum_{i=1}^{r} e^{t_i}\rho_i = \log \sum_{i=1}^{r} z_i = 0$. If $\log \sum_{i=1}^{r} e^{t_i}\rho_i = 0$, then $e^{t_i} = z_i/\rho_i$, where $z_i = \rho_i e^{t_i} > 0$. Hence t is in $\mathcal{N}$.

(b) $\log \sum_{i=1}^{r} e^{h(t)_i}\rho_i = \log \sum_{i=1}^{r} e^{t_i}\rho_i - c(t) = 0$. By part (a), $h(t)$ is in $\mathcal{N}$.

(c) For any $v \in \mathcal{M}$,

$$I(v) = \sup_{t \in \mathbb{R}^r} \{\langle t, v\rangle - c(t)\} \geq \sup_{t \in \mathcal{N}} \{\langle t, v\rangle - c(t)\} = \sup_{t \in \mathcal{N}} \langle t, v\rangle.$$

* ri $\mathcal{M}$ denotes the relative interior of $\mathcal{M}$, which is the set of $z = (z_1, \dots, z_r)$ satisfying $z_i > 0$ and $\sum_{i=1}^{r} z_i = 1$ [Rockafellar (1970, page 48)].

For any $v \in \mathscr{M}$, $\langle t, v\rangle - c(t) = \langle t - c(t)\mathbf{1}, v\rangle = \langle h(t), v\rangle$, and since $h(t)$ belongs to $\mathscr{N}$,

$$I(v) = \sup_{t \in \mathbb{R}^r} \langle h(t), v\rangle \leq \sup_{t^0 \in \mathscr{N}} \langle t^0, v\rangle.$$

It follows that for any $v \in \mathscr{M}$

$$I(v) = \sup_{t \in \mathscr{N}} \langle t, v\rangle = \sup_{z \in \mathrm{ri}\,\mathscr{M}} \sum_{i=1}^{r} v_i \log(z_i/\rho_i).$$

Since $I(v)$ is defined as a Legendre–Fenchel transform, $I(v)$ is a closed convex function on $\mathbb{R}^r$. Suppose we show that $I(v)$ equals $I_\rho^{(2)}(v)$ for all v in $\mathrm{ri}\,\mathscr{M}$. Since $I_\rho^{(2)}(v)$ is continuous relative to $\mathscr{M}$, the continuity property in Theorem VI.3.2 will imply that $I(v)$ equals $I_\rho^{(2)}(v)$ for all v in $\mathscr{M}$. For any v and z in $\mathrm{ri}\,\mathscr{M}$

$$\sum_{i=1}^{r} v_i \log \frac{z_i}{\rho_i} = \sum_{i=1}^{r} v_i \log \frac{v_i}{\rho_i} + \sum_{i=1}^{r} v_i \log \frac{z_i}{v_i} = I_\rho^{(2)}(v) - I_v^{(2)}(z).$$

Since $I_v^{(2)}(z) \geq 0$ with equality if and only if $z = v$, it follows that $I(v)$ equals $I_\rho^{(2)}(v)$ for all $v \in \mathrm{ri}\,\mathscr{M}$. □

Lemma VIII.2.3 completes the proof of Theorem VIII.2.2. By our remarks earlier in the section, Theorem VIII.2.1 follows.

VIII.3. The Contraction Principle Relating Levels-1 and 2 ($d = 1$)

In Theorem II.5.2, we stated a contraction principle relating levels-1 and 2 for Borel probability measures ρ on $\mathbb{R}$ whose support is a finite set. In the present section, a generalization is proved for any Borel probability measure ρ on $\mathbb{R}$ which is nondegenerate (not a unit point measure) and for which $c_\rho(t) = \log \int_{\mathbb{R}} \exp(tx)\rho(dx)$ is finite for all $t \in \mathbb{R}$. We also obtain additional properties of the level-1 entropy function $I_\rho^{(1)}$, including a relationship between the effective domain of this function and the support of ρ. These results will be proved in the next section for Borel probability measures ρ on $\mathbb{R}^d$, $d \geq 2$.

Let ρ be a Borel probability measure on $\mathbb{R}$. A point $x \in \mathbb{R}$ is said to be a *support point* for ρ if every neighborhood of x has positive ρ-measure. The *support* of ρ is defined as the set of all support points for ρ and is denoted by S_ρ. The support may be characterized as the smallest closed set having ρ-measure 1. The *convex hull* of S_ρ is defined as the intersection of all the convex sets containing S_ρ and is denoted by $\mathrm{conv}\, S_\rho$. Clearly, $\mathrm{conv}\, S_\rho$ is the smallest closed interval containing S_ρ; $\mathrm{conv}\, S_\rho$ has nonempty interior whenever ρ is nondegenerate. If $c_\rho(t)$ is finite for all $t \in \mathbb{R}$, then we define ρ_t to be the Borel probability measure on $\mathbb{R}$ which is absolutely continuous with respect to ρ and whose Radon–Nikodym derivative is given by

$$\frac{d\rho_t}{d\rho}(x) = \exp(tx) \cdot \frac{1}{\int_{\mathbb{R}} \exp(tx)\rho(dx)}. \tag{8.1}$$

According to Theorem VII.5.1(b), $c_\rho(t)$ is differentiable for all t and

$$c'_\rho(t) = \int_{\mathbb{R}} x \exp(tx)\rho(dx) \cdot \frac{1}{\int_{\mathbb{R}} \exp(tx)\rho(dx)} = \int_{\mathbb{R}} x\rho_t(dx).$$

If ν is a Borel probability measure on $\mathbb{R}$, then we define the relative entropy of ν with respect to ρ by the formula

$$I_\rho^{(2)}(\nu) = \begin{cases} \displaystyle\int_{\mathbb{R}} \log \frac{d\nu}{d\rho}(x)\nu(dx) & \text{if } \nu \ll \rho \text{ and } \displaystyle\int_{\mathbb{R}} \left|\log \frac{d\nu}{d\rho}\right| d\nu < \infty, \\ \infty & \text{otherwise.} \end{cases} \tag{8.2}$$

The next theorem states the contraction principle relating levels-1 and 2. The theorem shows that for each $z \in \mathbb{R}$

$$I_\rho^{(1)}(z) = \inf\left\{I_\rho^{(2)}(\nu) : \nu \in \mathcal{M}(\mathbb{R}), \int_{\mathbb{R}} x\nu(dx) = z\right\},$$

and it determines for $z \in \operatorname{conv} S_\rho$ where the infimum is attained. A related contraction principle was used in Chapter III to show the existence of the Maxwell–Boltzmann distribution for the discrete ideal gas [Lemma III.4.5].[2]

Theorem VIII.3.1. *Let ρ be a nondegenerate Borel probability measure on $\mathbb{R}$ such that $c_\rho(t)$ is finite for all $t \in \mathbb{R}$. Then the following conclusions hold.*

(a) *For each point $z \in \operatorname{int}(\operatorname{conv} S_\rho)$, there exists a unique point $t \in \mathbb{R}$ such that $\int_{\mathbb{R}} x\rho_t(dx) = z$. $I_\rho^{(2)}(\nu)$ attains its infimum over the set $\{\nu \in \mathcal{M}(\mathbb{R}) : \int_{\mathbb{R}} x\rho(dx) = z\}$ at the unique measure ρ_t and*

$$I_\rho^{(1)}(z) = I_\rho^{(2)}(\rho_t) = \inf\left\{I_\rho^{(2)}(\nu) : \nu \in \mathcal{M}(\mathbb{R}), \int_{\mathbb{R}} x\nu(dx) = z\right\} < \infty. \tag{8.3}$$

(b) *For $z \notin \operatorname{conv} S_\rho$,*

$$I_\rho^{(1)}(z) = \inf\left\{I_\rho^{(2)}(\nu) : \nu \in \mathcal{M}(\mathbb{R}), \int_{\mathbb{R}} x\nu(dx) = z\right\} = \infty. \tag{8.4}$$

(c) *Suppose that $\operatorname{conv} S_\rho$ has a finite endpoint α. If ρ has an atom at α ($\rho\{\alpha\} > 0$), then for $z = \alpha$ (8.3) is valid with ρ_t replaced by δ_α. If ρ does not have an atom at α, then (8.4) is valid for $z = \alpha$.*

Part (a) of the theorem states that for each point $z \in \operatorname{int}(\operatorname{conv} S_\rho)$, there exists a unique real number t such that $\int_{\mathbb{R}} x\rho_t(dx) = z$. Since $c'_\rho(t) = \int_{\mathbb{R}} x\rho_t(dx)$, we can prove the statement in part (a) by showing that the function c'_ρ defines a one-to-one mapping of $\mathbb{R}$ onto $\operatorname{int}(\operatorname{conv} S_\rho)$. Lemma VIII.3.2 and Theorem VIII.3.3 establish this fact together with other useful information. Theorem VIII.3.1 will be proved afterwards.

The following lemma is due to Lanford (1973, page 42).

Lemma VIII.3.2. *If the point* 0 *belongs to* $\operatorname{int}(\operatorname{conv} S_\rho)$, *then* $c'_\rho(t) = 0$ *for some* $t \in \mathbb{R}$.

Proof. A continuous function on $\mathbb{R}$ with compact level sets attains its infimum over $\mathbb{R}$ at some point t. If the function is differentiable, then the derivative vanishes at t. Hence it suffices to prove that the level sets $K_b = \{t \in \mathbb{R}: c_\rho(t) \le b\}$, b real, are compact. K_b is closed since c_ρ is a closed convex function. Since 0 belongs to $\operatorname{int}(\operatorname{conv} S_\rho)$, there exist numbers $\varepsilon > 0$ and $0 < \delta < 1$ such that

$$\rho\{x \in \mathbb{R}: x \ge \varepsilon\} \ge \delta > 0 \qquad \text{and} \qquad \rho\{x \in \mathbb{R}: x \le -\varepsilon\} \ge \delta > 0.$$

If t is non-negative, then

$$c_\rho(t) \ge \log \int_{\{x \ge \varepsilon\}} \exp(tx)\rho(dx) \ge t\varepsilon + \log \delta.$$

Similarly, if t is negative, then $c_\rho(t) \ge |t|\varepsilon + \log \delta$. Thus K_b is a subset of the interval $\{t \in \mathbb{R}: |t| \le (b - \log \delta)/\varepsilon\}$, and so K_b is bounded. □

Lemma VIII.3.2 will be used in the proof of part (a) of the next theorem.

Theorem VIII.3.3. *Let* ρ *be a nondegenerate Borel probability measure on* $\mathbb{R}$ *such that* $c_\rho(t)$ *is finite for all* $t \in \mathbb{R}$. *Denote the range of the function* $c'_\rho(t)$, $t \in \mathbb{R}$, *by* $\operatorname{ran} c'_\rho$. *Then the following conclusions hold.*

(a) c'_ρ *defines a one-to-one mapping of* $\mathbb{R}$ *onto* $\operatorname{int}(\operatorname{conv} S_\rho)$ *and*

$$\operatorname{int}(\operatorname{dom} I_\rho^{(1)}) = \operatorname{ran} c'_\rho = \operatorname{int}(\operatorname{conv} S_\rho).$$

(b) $\operatorname{dom} I_\rho^{(1)} \subseteq \operatorname{conv} S_\rho$; *i.e.*, $I_\rho^{(1)}(z) = \infty$ *if* $z \notin \operatorname{conv} S_\rho$.

(c) *Suppose that* $\operatorname{conv} S_\rho$ *has a finite endpoint* α. *Then* α *belongs to* $\operatorname{dom} I_\rho^{(1)}$ *if and only if* ρ *has an atom at* α; *in this case* $I_\rho^{(1)}(\alpha) = -\log \rho\{\alpha\}$. *In particular, if the support of* ρ *is a finite set, then* $\operatorname{dom} I_\rho^{(1)}$ *equals* $\operatorname{conv} S_\rho$ *and* $I_\rho^{(1)}$ *is continuous relative to* $\operatorname{conv} S_\rho$.

Proof. (a), (b). A short calculation shows that

$$c''_\rho(t) = \int_{\mathbb{R}} (x - \langle x \rangle_t)^2 \rho_t(dx), \tag{8.5}$$

where $\langle x \rangle_t = \int_{\mathbb{R}} x\rho_t(dx)$ and the measure ρ_t is defined in (8.1). Since ρ is nondegenerate, (8.5) implies that $c''_\rho(t)$ is positive. Thus $c_\rho(t)$ is strictly convex [Problem VI.7.2(b)] and the derivative $c'_\rho(t)$ is an increasing function on $\mathbb{R}$. In particular, if z is given, then the equation $c'_\rho(t) = z$ has a unique solution t whenever a solution exists. We now show that $\operatorname{ran} c'_\rho = \operatorname{int}(\operatorname{conv} S_\rho)$.

Step 1: $\operatorname{ran} c'_\rho \subseteq \operatorname{int}(\operatorname{conv} S_\rho)$. Since the support S_{ρ_t} of ρ_t equals the support of ρ, $c'_\rho(t) = \int_{S_{\rho_t}} x\rho_t(dx)$ must lie in $\operatorname{conv} S_\rho$ for each real t. If $c'_\rho(t)$ were

to lie in $\mathrm{bd}(\mathrm{conv}\, S_\rho)$, then ρ_t and ρ would have to be point measures. This would contradict the nondegeneracy of ρ.

Step 2: $\mathrm{int}(\mathrm{conv}\, S_\rho) \subseteq \mathrm{ran}\, c_\rho'$. Let z be a point in $\mathrm{int}(\mathrm{conv}\, S_\rho)$ and consider the translated measure $\rho(\cdot + z)$. The point 0 belongs to the interior of the set $\mathrm{conv}\, S_{\rho(\cdot+z)}$ and

$$c'_{\rho(\cdot+z)}(t) = \int_{\mathbb{R}} x\rho_t(dx) - z = c_\rho'(t) - z.$$

Thus finding a point t which satisfies $c_\rho'(t) = z$ is equivalent to finding a point t which satisfies $c'_{\rho(\cdot+z)}(t) = 0$. In other words, without loss of generality, it suffices to prove that if 0 belongs to $\mathrm{int}(\mathrm{conv}\, S_\rho)$, then there exists a point t which satisfies $c_\rho'(t) = 0$. This implication was proved in Lemma VIII.3.2.

We have shown that $\mathrm{ran}\, c_\rho' = \mathrm{int}(\mathrm{conv}\, S_\rho)$ and thus that c_ρ' defines a one-to-one mapping of $\mathbb{R}$ onto $\mathrm{int}(\mathrm{conv}\, S_\rho)$. We now describe the effective domain of the function

$$I_\rho^{(1)}(z) = \sup_{t\in\mathbb{R}} \{tz - c_\rho(t)\}, \qquad z \in \mathbb{R}.$$

Since $c_\rho(t)$ is convex, the supremum is attained at some point t if and only if $c_\rho'(t) = z$. In this case,

$$I_\rho^{(1)}(z) = t(z)\cdot z - c_\rho(t)) \qquad \text{where } c_\rho'(t(z)) = z.$$

Hence all points z in the range of c_ρ' are in the effective domain of $I_\rho^{(1)}$. Combining this with the previous result, we have

$$\mathrm{ran}\, c_\rho' = \mathrm{int}(\mathrm{conv}\, S_\rho) \subseteq \mathrm{dom}\, I_\rho^{(1)}.$$

We now show that $\mathrm{dom}\, I_\rho^{(1)} \subseteq \mathrm{conv}\, S_\rho$. This will yield part (b) of the theorem, and together with the last display it will show that

$$\mathrm{int}(\mathrm{dom}\, I_\rho^{(1)}) = \mathrm{ran}\, c_\rho' = \mathrm{int}(\mathrm{conv}\, S_\rho).$$

Let z be a point outside $\mathrm{conv}\, S_\rho$. If z lies to the right of $\mathrm{conv}\, S_\rho$, then there exist real numbers $\delta > 0$ and b such that

$$\sup\{x \in \mathbb{R} : x \in \mathrm{conv}\, S_\rho\} \le b - \delta < b + \delta \le z.$$

Since S_ρ is a subset of $\mathrm{conv}\, S_\rho$,

$$\begin{aligned} I_\rho^{(1)}(z) &\ge \sup_{t>0} \left\{ tz - \log \int_{S_\rho} \exp(tx)\rho(dx) \right\} \\ &\ge \sup_{t>0} \{t(b+\delta) - t(b-\delta)\} = \infty. \end{aligned}$$

A similar analysis shows that $I_\rho^{(1)}(z) = \infty$ if z lies to the left of $\mathrm{conv}\, S_\rho$. This proves that $\mathrm{dom}\, I_\rho^{(1)} \subseteq \mathrm{conv}\, S_\rho$.

(c) Suppose that α is a right-hand endpoint of $\mathrm{conv}\, S_\rho$. For fixed $x \in S_\rho$, $\exp[t(x-\alpha)]$ is a nonincreasing function of t which converges to $\chi_{\{\alpha\}}(x)$

as $t \to \infty$. Hence

$$I_\rho^{(1)}(\alpha) = -\inf_{t\in\mathbb{R}} \log \int_{S_\rho} \exp[t(x-\alpha)]\rho(dx) = -\log \int_{\mathbb{R}} \chi_{\{\alpha\}}(x)\rho(dx).$$

$I_\rho^{(1)}(\alpha)$ equals $-\log \rho\{\alpha\}$ or ∞ according to whether or not ρ has an atom at α. A similar proof works for a left-hand endpoint of $\operatorname{conv} S_\rho$. If the support of ρ is a finite set $\Gamma = \{x_1, x_2, \ldots, x_r\}$ with $x_1 < x_2 < \cdots < x_r$, then $\operatorname{conv} S_\rho = [x_1, x_r]$. Since ρ has atoms at x_1 and at x_r, $\operatorname{dom} I_\rho^{(1)}$ equals $\operatorname{conv} S_\rho$. $I_\rho^{(1)}$ is continuous relative to $\operatorname{conv} S_\rho$ by Theorems VI.3.1 and VI.3.2. □

Proof of the contraction principle, Theorem VIII.3.1. (a) Let z be a point in $\operatorname{int}(\operatorname{conv} S_\rho)$. Part (a) of Theorem VIII.3.3 implies that there exists a unique point $t \in \mathbb{R}$ such that $\int_{\mathbb{R}} x\rho_t(dx) = z$. We have

$$\begin{aligned} I_\rho^{(2)}(\rho_t) &= \int_{\mathbb{R}} \left[tx - \log \int_{\mathbb{R}} \exp(tx)\rho(dx) \right] \rho_t(dx) \\ &= tz - c_\rho(t) = I_\rho^{(1)}(z). \end{aligned}$$

Now let $\nu \neq \rho_t$ be any other Borel probability measure on $\mathbb{R}$ which has mean z and for which $I_\rho^{(2)}(\nu)$ is finite. Then ν is absolutely continuous with respect to ρ and to ρ_t and

$$\begin{aligned} I_\rho^{(2)}(\nu) &= \int_{\mathbb{R}} \log \frac{d\nu}{d\rho} d\nu = \int_{\mathbb{R}} \log \frac{d\nu}{d\rho_t} d\nu + \int_{\mathbb{R}} \log \frac{d\rho_t}{d\rho} d\nu \\ &= I_{\rho_t}^{(2)}(\nu) + \int_{\mathbb{R}} \left[tx - \log \int_{\mathbb{R}} \exp(tx)\rho(dx) \right] \nu(dx) \\ &= I_{\rho_t}^{(2)}(\nu) + tz - c_\rho(t) = I_{\rho_t}^{(2)}(\nu) + I_\rho^{(1)}(z). \end{aligned}$$

Since $\nu \neq \rho_t$, $I_{\rho_t}^{(2)}(\nu)$ is positive and so $I_\rho^{(2)}(\nu) > I_\rho^{(1)}(z)$. We conclude that for $\nu \in \mathcal{M}(\mathbb{R})$, $I_\rho^{(2)}(\nu) \geq I_\rho^{(1)}(z) = I_\rho^{(2)}(\rho_t)$ and that equality holds if and only if $\nu = \rho_t$.

(b) If z does not belong to $\operatorname{conv} S_\rho$, then any Borel probability measure ν on $\mathbb{R}$ which has mean z is not absolutely continuous with respect to ρ. Therefore

$$\inf\left\{ I_\rho^{(2)}(\nu) : \nu \in \mathcal{M}(\mathbb{R}), \int_{\mathbb{R}} x\nu(dx) = z \right\} = \infty.$$

$I_\rho^{(1)}(z)$ equals ∞ by Theorem VIII.3.3(b).

(c) If α is a finite endpoint of $\operatorname{conv} S_\rho$ and if ρ has an atom at α, then the measure δ_α is absolutely continuous with respect to ρ, δ_α has mean α, and δ_α is the only Borel probability measure on $\mathbb{R}$ with these properties. By Theorem VIII.3.3(c)

$$I_\rho^{(1)}(\alpha) = -\log \rho\{\alpha\} = I_\rho^{(2)}(\delta_\alpha) = \inf\left\{ I_\rho^{(2)}(\nu) : \nu \in \mathcal{M}(\mathbb{R}), \int_{\mathbb{R}} x\nu(dx) = \alpha \right\} < \infty.$$

If ρ does not have an atom at α, then

$$I_\rho^{(1)}(\alpha) = \inf\left\{I_\rho^{(2)}(v): v \in \mathscr{M}(\mathbb{R}), \int_{\mathbb{R}} xv(dx) = \alpha\right\} = \infty. \qquad \square$$

Our final result describes smoothness and mapping properties of the entropy function $I_\rho^{(1)}(z)$. We recall from Theorem VIII.3.3(a) that

$$\operatorname{int}(\operatorname{dom} I_\rho^{(1)}) = \operatorname{int}(\operatorname{conv} S_\rho).$$

Theorem VIII.3.4. *Let ρ be a nondegenerate Borel probability measure on $\mathbb{R}$ such that $c_\rho(t)$ is finite for all $t \in \mathbb{R}$. Then the following conclusions hold.*

(a) *$I_\rho^{(1)}(z)$ is essentially smooth; that is, $I_\rho^{(1)}(z)$ is differentiable for $z \in \operatorname{int}(\operatorname{dom} I_\rho^{(1)}) = \operatorname{int}(\operatorname{conv} S_\rho)$ and $|(I_\rho^{(1)})'(z)| \to \infty$ as $z \in \operatorname{int}(\operatorname{conv} S_\rho)$ converges to a boundary point of $\operatorname{int}(\operatorname{conv} S_\rho)$.*

(b) *The function $(I_\rho^{(1)})'$ defines a one-to-one mapping of $\operatorname{int}(\operatorname{conv} S_\rho)$ onto $\mathbb{R}$ with inverse c_ρ'.*

(c) *$I_\rho^{(1)}(z)$ is a real analytic function of $z \in \operatorname{int}(\operatorname{conv} S_\rho)$.*

Proof. (a) Since ρ is nondegenerate, $c_\rho(t)$ is strictly convex [see (8.5)] and so $I_\rho^{(1)}(z)$ is essentially smooth [Theorem VI.5.6].

(b) This follows from Theorem VIII.3.3(a) and the fact that for $z \in \operatorname{int}(\operatorname{conv} S_\rho)$, $t = (I_\rho^{(1)})'(z)$ if and only if $z = c_\rho'(t)$ [Theorem VII.5.5(d)].

(c) If z is a point in $\operatorname{int}(\operatorname{conv} S_\rho)$, then there exists a unique solution $t = t(z)$ of $c_\rho'(t) = z$ and

$$I_\rho^{(1)}(z) = t(z) \cdot z - c_\rho(t(z)).$$

It suffices to prove that the function $z \to t(z)$ is real analytic in some neighborhood of each fixed z. The function $c_\rho'(t)$ can be continued into the complex plane $\mathbb{C}$ to be an analytic function. Since $c_\rho''(t(z))$ is positive, this continuation defines a one-to-one analytic mapping of a complex open neighborhood of $t(z)$ onto a complex open neighborhood U of z. The inverse function is also analytic [Rudin (1974, page 231)]. Hence the restriction of the inverse function to points $z \in U \cap \mathbb{R}$ is real analytic. The restriction is the function $z \to t(z)$. $\square$

VIII.4. The Contraction Principle Relating Levels-1 and 2 ($d \geq 2$)

Let ρ be a nondegenerate Borel probability measure on $\mathbb{R}$ with support S_ρ and assume that $c_\rho(t)$ is finite for all $t \in \mathbb{R}$. In the previous section, we proved that the infimum in the contraction principle relating levels-1 and 2 is attained for all z in the interior of $\operatorname{conv} S_\rho$ (convex hull of S_ρ) and that the effective domain of $I_\rho^{(1)}$ is a subset of $\operatorname{conv} S_\rho$. These results will now be extended to

nondegenerate Borel probability measures ρ on $\mathbb{R}^d$, $d \geq 2$. The sets S_ρ and $\operatorname{conv} S_\rho$ are defined as on page 253.

The role played by the set $\operatorname{conv} S_\rho$ for $d = 1$ is played, for $d \geq 2$, by the closed convex hull of S_ρ. The latter is defined as the intersection of all the closed convex sets containing S_ρ and is denoted by $\operatorname{cc} S_\rho$.* We suppose that $c_\rho(t) = \log \int_{\mathbb{R}^d} \exp\langle t, x\rangle \rho(dx)$ is finite for all $t \in \mathbb{R}^d$, and we define ρ_t to be the Borel probability measure on $\mathbb{R}^d$ which is absolutely continuous with respect to ρ and whose Radon–Nikodym derivative is given by

$$\frac{d\rho_t}{d\rho}(x) = \exp\langle t, x\rangle \cdot \frac{1}{\int_{\mathbb{R}^d} \exp\langle t, x\rangle \rho(dx)}. \tag{8.6}$$

According to Theorem VII.5.1(b), $c_\rho(t)$ is differentiable for all t and

$$\frac{\partial c_\rho(t)}{\partial t_i} = \int_{\mathbb{R}^d} x_i \exp\langle t, x\rangle \rho(dx) \cdot \frac{1}{\int_{\mathbb{R}^d} \exp\langle t, x\rangle \rho(dx)}. \tag{8.7}$$

Thus $\nabla c_\rho(t) = \int_{\mathbb{R}^d} x \rho_t(dx)$.

Let ρ be a Borel probability measure on $\mathbb{R}^d$ such that $c_\rho(t)$ is finite for all $t \in \mathbb{R}^d$. The relative entropy $I_\rho^{(2)}(\nu)$ of $\nu \in \mathscr{M}(\mathbb{R}^d)$ with respect to ρ is defined as in (8.2). Donsker and Varadhan (1976a, page 425) prove that

$$I_\rho^{(1)}(z) = \inf\left\{ I_\rho^{(2)}(\nu) : \nu \in \mathscr{M}(\mathbb{R}^d), \int_{\mathbb{R}^d} x\nu(dx) = z \right\} \tag{8.8}$$

for each $z \in \mathbb{R}^d$.[3] We will prove (8.8) only for z in the relative interior of $\operatorname{cc} S_\rho$ and for $z \notin \operatorname{cc} S_\rho$. For z in the relative interior of $\operatorname{cc} S_\rho$, we will also determine where the infimum in (8.8) is attained.

Theorem VIII.4.1. *Let ρ be a nondegenerate Borel probability measure on $\mathbb{R}^d$, $d \geq 2$, such that $c_\rho(t)$ is finite for all $t \in \mathbb{R}^d$. Then the following conclusions hold.*

(a) *For $z \in \operatorname{ri}(\operatorname{cc} S_\rho)$, let A_z denote the set of points t for which $\int_{\mathbb{R}^d} x\rho_t(dx) = z$. Then A_z is nonempty,† the measures $\{\rho_t; t \in A_z\}$ are all equal, and $I_\rho^{(2)}(\nu)$ attains its infimum over the set $\{\nu \in \mathscr{M}(\mathbb{R}^d) : \int_{\mathbb{R}^d} x\nu(dx) = z\}$ at the unique measure ρ_t, $t \in A_z$. Furthermore*

$$I_\rho^{(1)}(z) = I_\rho^{(2)}(\rho_t) = \inf\left\{ I_\rho^{(2)}(\nu) : \nu \in \mathscr{M}(\mathbb{R}^d), \int_{\mathbb{R}^d} x\nu(dx) = z \right\} < \infty. \tag{8.9}$$

(b) *For $z \notin \operatorname{cc} S_\rho$,*

$$I_\rho^{(1)}(z) = \inf\left\{ I_\rho^{(2)}(\nu) : \nu \in \mathscr{M}(\mathbb{R}^d), \int_{\mathbb{R}^d} x\nu(dx) = z \right\} = \infty. \tag{8.10}$$

*Although S_ρ is closed, $\operatorname{conv} S_\rho$ need not be closed for $d \geq 2$; e.g., if $S_\rho = \{x \in \mathbb{R}^2 : x_1 \in \mathbb{R}, x_2 = 0 \text{ or } x_1 = 0, x_2 = 1\}$, then $\operatorname{conv} S_\rho = \{x \in \mathbb{R}^2 : x_1 \in \mathbb{R}, 0 \leq x_2 < 1 \text{ or } x_1 = 0, x_2 = 1\}$. In general, $\operatorname{ri}(\operatorname{cc} S_\rho) = \operatorname{ri}(\operatorname{conv} S_\rho)$.

†A_z consists of a unique point if and only if ρ is maximal [see Theorems VIII.4.3(a) and VIII.4.4(a)].

This theorem generalizes parts (a) and (b) of Theorem VIII.3.1 to $\mathbb{R}^d$, $d \geq 2$. The proof of the latter depended on the relation between the range of the function $c'_\rho(t)$ and the support of ρ expressed in Theorem VIII.3.3(a). A similar proof will yield Theorem VIII.4.1 once we establish an analogous relation between the range of the mapping $\nabla c_\rho(t)$, $t \in \mathbb{R}^d$, and the support of ρ.

It is convenient first to consider those measures ρ for which the convex function $c_\rho(t)$ is a strictly convex function on $\mathbb{R}^d$. The next proposition gives a simple criterion for this strict convexity. Let $\operatorname{aff} S_\rho$ be the affine hull of the support of ρ. A Borel probability measure ρ on $\mathbb{R}^d$ is said to be *maximal* if $\operatorname{aff} S_\rho$ equals all of $\mathbb{R}^d$. This condition is equivalent to S_ρ not being a subset of any hyperplane in $\mathbb{R}^d$. For example, if $d = 1$, then any nondegenerate measure is maximal. For $d \geq 1$, $\operatorname{cc} S_\rho$ has nonempty interior whenever ρ is maximal, and $\operatorname{int}(\operatorname{cc} S_\rho)$ and $\operatorname{int}(\operatorname{conv} S_\rho)$ coincide.

Proposition VIII.4.2. *Let ρ be a Borel probability measure on $\mathbb{R}^d$ such that $c_\rho(t)$ is finite for all $t \in \mathbb{R}^d$. Then $c_\rho(t)$ is a strictly convex function on $\mathbb{R}^d$ if and only if ρ is maximal.*

Proof. Hölder's inequality implies that for every t_1 and t_2 in $\mathbb{R}^d$ and $0 < \lambda < 1$

$$\int_{\mathbb{R}^d} \exp\langle \lambda t_1 + (1-\lambda)t_2, x\rangle \rho(dx) \leq \left\{ \int_{\mathbb{R}^d} \exp\langle t_1, x\rangle \rho(dx) \right\}^{\lambda} \cdot \left\{ \int_{\mathbb{R}^d} \exp\langle t_2, x\rangle \rho(dx) \right\}^{(1-\lambda)} \tag{8.11}$$

and that equality holds if and only if

$$\frac{\exp\langle t_1, x\rangle}{\int_{\mathbb{R}^d} \exp\langle t_1, x\rangle \rho(dx)} = \frac{\exp\langle t_2, x\rangle}{\int_{\mathbb{R}^d} \exp\langle t_2, x\rangle \rho(dx)} \qquad \rho\text{-a.s.} \tag{8.12}$$

If $c_\rho(t)$ is not strictly convex on $\mathbb{R}^d$, then there exist distinct points t_1 and t_2 and some $0 < \lambda < 1$ for which equality holds in (8.11). Hence by (8.12)

$$\langle t_1, x\rangle = \langle t_2, x\rangle + c_\rho(t_1) - c_\rho(t_2) \qquad \rho\text{-a.s.} \tag{8.13}$$

This equality implies that the support of ρ is a subset of the hyperplane

$$\{x \in \mathbb{R}^d : \langle x, t_1 - t_2\rangle = c_\rho(t_1) - c_\rho(t_2)\}.$$

Hence ρ is not maximal. Conversely, suppose that ρ is not maximal, but that its support is a subset of the hyperplane $H = \{x \in \mathbb{R}^d : \langle x, \gamma\rangle = b\}$, where γ is a unit vector. Let t_0 be a fixed element of H and let A be the set of vectors of the form

$$t = t_0 + \alpha\gamma, \qquad \alpha \text{ real} \quad (\alpha = \langle t - t_0, \gamma\rangle).$$

Since $\langle x, \gamma \rangle = b$ for all x in the support of ρ, we have for $t \in A$

$$c_\rho(t) = b\alpha + \log \int_{\mathbb{R}^d} \exp\langle t_0, x \rangle \rho(dx) = b\langle t - t_0, \gamma \rangle + c_\rho(t_0).$$

This shows that c_ρ is affine on A and thus is not strictly convex on $\mathbb{R}^d$. □

The next theorem generalizes Theorem VIII.3.3 to maximal probability measures on $\mathbb{R}^d$, $d \geq 2$. It is due to Barndorff-Nielsen (1970, 1978).[4] Measures which are not maximal will be considered later.

Theorem VIII.4.3. *Let ρ be a maximal Borel probability measure on $\mathbb{R}^d$, $d \geq 2$, such that $c_\rho(t)$ is finite for all $t \in \mathbb{R}^d$. Denote the range of the mapping $\nabla c_\rho(t)$, $t \in \mathbb{R}^d$, by $\operatorname{ran} \nabla c_\rho$. Then the following conclusions hold.*

(a) *∇c_ρ defines a one-to-one mapping of $\mathbb{R}^d$ onto $\operatorname{int}(\operatorname{cc} S_\rho)$ and*

$$\operatorname{int}(\operatorname{dom} I_\rho^{(1)}) = \operatorname{ran} \nabla c_\rho = \operatorname{int}(\operatorname{cc} S_\rho).$$

(b) $\operatorname{dom} I_\rho^{(1)} \subseteq \operatorname{cc} S_\rho$; *i.e.*, $I_\rho^{(1)}(z) = \infty$ *if* $z \notin \operatorname{cc} S_\rho$.

(c) *Let z be a boundary point of $\operatorname{cc} S_\rho$ and let $\mathscr{H}_z$ be the set of all supporting hyperplanes to $\operatorname{cc} S_\rho$ at z. If $\rho(H) = 0$ for some $H \in \mathscr{H}_z$, then z does not belong to $\operatorname{dom} I_\rho^{(1)}$.*

Proof of parts (a) and (b). It is convenient to divide the proof of parts (a) and (b) into four steps. The proof of part (c) is omitted [see Barndorff-Nielsen (1978, pages 140–143) and Problem VIII.6.8.]

Step 1: *∇c_ρ defines a one-to-one mapping.* Let t_1 and t_2 be any two distinct points in $\mathbb{R}^d$ and define the function

$$f(\lambda) = c_\rho(t_1 + \lambda(t_2 - t_1)), \qquad \lambda \in \mathbb{R}.$$

Since ρ is maximal, $c_\rho(t)$ is strictly convex on $\mathbb{R}^d$, and so $f(\lambda)$ is strictly convex on $\mathbb{R}$. Thus

$$f'(0) = \langle \nabla c_\rho(t_1), t_2 - t_1 \rangle < f'(1) = \langle \nabla c_\rho(t_2), t_2 - t_1 \rangle.$$

It follows that $\nabla c_\rho(t_1) \neq \nabla c_\rho(t_2)$.

Step 2: $\operatorname{ran} \nabla c_\rho \subseteq \operatorname{int}(\operatorname{cc} S_\rho)$. We first prove that $\operatorname{ran} \nabla c_\rho \subseteq \operatorname{cc} S_\rho$. Suppose that $z = \nabla c_\rho(t)$ lies in the complement of $\operatorname{cc} S_\rho$. By Lemma VI.5.4(b) there exists a hyperplane $H = \{x \in \mathbb{R}^d : \langle x, \gamma \rangle = b\}$ which separates $\operatorname{cc} S_\rho$ and $\{z\}$ strongly and

$$\sup\{\langle x, \gamma \rangle : x \in \operatorname{cc} S_\rho\} < b < \langle z, \gamma \rangle.$$

Since the support S_{ρ_t} of the measure ρ_t equals S_ρ and S_ρ is a subset of $\operatorname{cc} S_\rho$,

$$b < \langle z, \gamma \rangle = \langle \operatorname{grad} c_\rho(t), \gamma \rangle = \int_{S_{\rho_t}} \langle x, \gamma \rangle \rho_t(dx) < b.$$

This contradiction proves that $\operatorname{ran}\nabla c_\rho \subseteq \operatorname{cc} S_\rho$. Now suppose that $z = \nabla c_\rho(t)$ lies in $\operatorname{bd}(\operatorname{cc} S_\rho)$. By Lemma VI.5.4(a) there exists a hyperplane $H = \{x \in \mathbb{R}^d : \langle x, \gamma\rangle = b\}$ which separates $\operatorname{cc} S_\rho$ and $\{z\}$, and

$$\sup\{\langle x, \gamma\rangle : x \in \operatorname{cc} S_\rho\} \le b \le \langle z, \gamma\rangle.$$

As above, $b \le \langle z, \gamma\rangle = \int_{S_{\rho_t}} \langle x, \gamma\rangle \rho_t(dx) \le b$. This implies that $\langle x, \gamma\rangle$ equals b for all $x \in S_{\rho_t} = S_\rho$, or that S_ρ is a subset of the hyperplane H. As this contradicts the maximality of ρ, the proof of Step 2 is done.

Step 3: $\operatorname{int}(\operatorname{cc} S_\rho) \subseteq \operatorname{ran}\nabla c_\rho$. As in the proof of the analogous fact for $d = 1$ [page 256], it suffices to prove that if the point 0 belongs to $\operatorname{int}(\operatorname{cc} S_\rho)$, then there exists a point t which satisfies $\nabla c_\rho(t) = 0$. For $d = 1$, this was proved in Lemma VIII.3.2. The same proof applies to $d \ge 2$ once we show that the level sets $K_b = \{t \in \mathbb{R}^d : c_\rho(t) \le b\}$, b real, are compact.* K_b is closed since c_ρ is a closed convex function on $\mathbb{R}^d$. If $\operatorname{cc} S_\rho = \mathbb{R}^d$, then let $\varepsilon = \mathrm{R}$, where R is any fixed positive number. Otherwise, let $\varepsilon > 0$ be half the distance from 0 to the complement of $\operatorname{cc} S_\rho$. For any unit vector γ, define the half-space $A(\gamma) = \{x \in \mathbb{R}^d : \langle x, \gamma\rangle > \varepsilon\}$. We prove below that there exists a number $0 < \delta \le 1$ such that

$$\inf_{\|\gamma\|=1} \rho\{A(\gamma)\} \ge \delta. \tag{8.14}$$

From this it will follow that if t is nonzero, then

$$c_\rho(t) \ge \log \int_{A(t/\|t\|)} \exp\langle \|t\| \cdot t/\|t\|, x\rangle \rho(dx) \ge \varepsilon\|t\| + \log\delta.$$

The same inequality holds for $t = 0$ ($c_\rho(0) = 0 \ge \log\delta$). We conclude that K_b is a subset of the ball $\{t \in \mathbb{R}^d : \|t\| \le \varepsilon^{-1}(b - \log\delta)\}$ and thus that K_b is bounded.

If (8.14) were not true for some $\delta > 0$, then there would exist a sequence of unit vectors $\gamma_1, \gamma_2, \ldots$ such that $\rho\{A(\gamma_n)\} \to 0$ as $n \to \infty$. By the compactness of the unit sphere in $\mathbb{R}^d$, there would exist a subsequence $\{\gamma_{n'}\}$ and a unit vector γ such that $\gamma_{n'} \to \gamma$. Fatou's lemma would imply $\rho\{A(\gamma)\} = 0$. On the other hand, by the definiton of ε, the open half-space $A(\gamma)$ contains a point in $\operatorname{conv} S_\rho$, and therefore $A(\gamma)$ contains a point in S_ρ. It follows that $\rho\{A(\gamma)\}$ must be positive. This contradiction proves that (8.14) must hold for some $\delta > 0$ ($\delta \le 1$ since ρ is a probability measure).

Step 4: $\operatorname{int}(\operatorname{dom} I_\rho^{(1)}) = \operatorname{int}(\operatorname{cc} S_\rho) \subseteq \operatorname{dom} I_\rho^{(1)} \subseteq \operatorname{cc} S_\rho$. According to Theorem VI.5.3(c) $\operatorname{ran}\nabla c_\rho \subseteq \operatorname{dom} I_\rho^{(1)}$, and Steps 2 and 3 imply that $\operatorname{ran}\nabla c_\rho = \operatorname{int}(\operatorname{cc} S_\rho)$. Hence Step 4 is proved once we show that $\operatorname{dom} I_\rho^{(1)} \subseteq \operatorname{cc} S_\rho$; i.e., $I_\rho^{(1)}(z) = \infty$ for $z \notin \operatorname{cc} S_\rho$. If $z \notin \operatorname{cc} S_\rho$, then there exists a hyperplane $H = \{x \in \mathbb{R}^d : \langle x, \gamma\rangle = b\}$ which separates $\operatorname{cc} S_\rho$ and $\{z\}$ strongly, and

$$\sup\{\langle x, \gamma\rangle : x \in \operatorname{cc} S_\rho\} \le b - \delta < b + \delta \le \langle z, \gamma\rangle$$

for some $\delta > 0$. Since S_ρ is a subset of $\operatorname{cc} S_\rho$,

*The following proof is due to Lanford (1973, page 42).

$$I_\rho^{(1)}(z) = \sup_{t \in \mathbb{R}^d} \{\langle t, z\rangle - c_\rho(t)\} \geq \sup_{t=r\gamma, r>0} \left\{\langle t, z\rangle - \log \int_{S_\rho} \exp\langle t, x\rangle \rho(dx)\right\}$$
$$\geq \sup_{r>0} \{r(b+\delta) - r(b-\delta)\} = \infty.$$

This completes the proof of Step 4 and the proof of parts (a) and (b) of Theorem VIII.4.3. □

The previous theorem described the range of the mapping $\nabla c_\rho(t)$, $t \in \mathbb{R}^d$, and the effective domain of $I_\rho^{(1)}$ for maximal measures ρ. We now consider nondegenerate measures ρ which are not maximal. As a special case, assume that the affine hull of S_ρ equals the subspace $A = \{x \in \mathbb{R}^d : x_{m+1} = \cdots = x_d = 0\}$ of dimension $m < d$. Then $c_\rho(t) = \log \int_{\mathbb{R}^d} \exp(\sum_{i=1}^m t_i x_i) \rho(dx)$ is a function of $t \in A$. Since as a measure on A ρ is maximal, the previous theorem implies that $\operatorname{grad} c_\rho$ defines a one-to-one mapping of A onto $\operatorname{ri}(\operatorname{cc} S_\rho)$ (the interior of $\operatorname{cc} S_\rho$ relative to A) and that $\operatorname{ri}(\operatorname{dom} I_\rho^{(1)})$ equals $\operatorname{ri}(\operatorname{cc} S_\rho)$. In order to analyze an arbitrary nonmaximal measure, one reduces to this special case.

Theorem VIII.4.4. *Let ρ be a nondegenerate Borel probability measure on $\mathbb{R}^d$, $d \geq 2$, which is not maximal. Assume that $c_\rho(t)$ is finite for all $t \in \mathbb{R}^d$. Then the following conclusions hold.*

(a) *∇c_ρ defines a many-to-one mapping of $\mathbb{R}^d$ onto $\operatorname{ri}(\operatorname{cc} S_\rho)$ and*

$$\operatorname{ri}(\operatorname{dom} I_\rho^{(1)}) = \operatorname{ran} \nabla c_\rho = \operatorname{ri}(\operatorname{cc} S_\rho).$$

If $\nabla c_\rho(t_1) = \nabla c_\rho(t_2)$, then the measures ρ_{t_1} and ρ_{t_2} are equal.

(b) *∇c_ρ defines a one-to-one mapping of $\operatorname{aff} S_\rho$ onto $\operatorname{ri}(\operatorname{cc} S_\rho)$.*

(c) $\operatorname{dom} I_\rho^{(1)} \subseteq \operatorname{cc} S_\rho$.

We prove the assertion in part (a) that if $\nabla c_\rho(t_1) = \nabla c_\rho(t_2)$, then $\rho_{t_1} = \rho_{t_2}$. Suppose that $t_1 \neq t_2$ and define the convex function

$$f(\lambda) = c_\rho(\lambda t_1 + (1-\lambda)t_2), \qquad \lambda \in \mathbb{R}.$$

Since $\operatorname{grad} c_\rho(t_1) = \operatorname{grad} c_\rho(t_2)$, it follows that $f'(0) = f'(1)$ and thus that $f(\lambda)$ is affine on the interval $0 \leq \lambda \leq 1$. This implies that

$$c_\rho(\lambda t_1 + (1-\lambda)t_2) = \lambda c_\rho(t_1) + (1-\lambda)c_\rho(t_2) \qquad \text{for all } 0 \leq \lambda \leq 1.$$

The latter is equivalent to condition (8.12), which implies that $\rho_{t_1} = \rho_{t_2}$. The proof of the rest of the theorem is Problem VIII.6.9.

Proof of the contraction principle, Theorem VIII.4.1. (a) Since $\int_{\mathbb{R}^d} x \rho_t(dx) = \nabla c_\rho(t)$, the set A_z is the set of t for which $\nabla c_\rho(t) = z$. For $z \in \operatorname{ri}(\operatorname{cc} S_\rho)$, A_z is nonempty by Theorems VIII.4.3(a) and VIII.4.4(a). A_z consists of a unique point t if and only if ρ is maximal. We have just proved that if ρ is not maximal,

then the measures $\{\rho_t; t \in A_z\}$ are all equal. The rest of part (a) is proved exactly like Theorem VIII.3.1(a).

(b) If z does not belong to cc S_ρ, then any Borel probability measure ν on $\mathbb{R}^d$ which has mean z is not absolutely continuous with respect to ρ. Therefore

$$\inf\left\{I_\rho^{(2)}(\nu): \nu \in \mathscr{M}(\mathbb{R}^d), \int_{\mathbb{R}^d} x\nu(dx) = z\right\} = \infty.$$

$I_\rho^{(1)}(z)$ equals ∞ by Theorem VIII.4.3(b). □

For nondegenerate probability measures ρ on $\mathbb{R}$, Theorem VIII.3.4 described smoothness and mapping properties of the entropy function $I_\rho^{(1)}(z)$. This has a direct generalization to maximal probability measures ρ on $\mathbb{R}^d$, $d \geq 2$ [Problem VIII.6.10].

We have completed our discussion of level-2 large deviations. The level-3 problem is the topic of the next chapter.

VIII.5. Notes

1 (page 250). We describe the results in Donsker and Varadhan (1975a). Let $X_0, X_1, X_2, \ldots$ be a stationary Markov process with transition probabilities $\gamma(x, dy)$ taking values in a compact metric space $\mathscr{X}$. Assume that $X_0 = x$ and that

(a) $\gamma(x, dy)$ is a Feller transition probability;

(b) $\gamma(x, dy)$ has a density $\gamma(x, y)$ relative to some reference probability measure $\beta(dy)$;

(c) $\gamma(x, y)$ is uniformly bounded away from 0 and ∞.

Let $\mathscr{U}$ be the set of positive functions $u \in \mathscr{C}(\mathscr{X})$. Define for any Borel probability measure ν on $\mathscr{X}$

$$I_\gamma(\nu) = -\inf_{u \in \mathscr{U}} \int_{\mathscr{X}} \log\left(\frac{\gamma u}{u}\right)(x)\nu(dx),$$

where $\gamma u(x) = \int_{\mathscr{X}} u(y)\gamma(x, dy)$. It is then proved that the distributions on $\mathscr{M}(\mathscr{X})$ of the empirical measures $L_{n,x}(\omega, \cdot) = n^{-1}\sum_{j=0}^{n-1} \delta_{X_j(\omega)}(\cdot), n = 1, 2, \ldots,$ have a large deviation property with $a_n = n$ and entropy function $I_\gamma(\nu)$.

The paper also treats continuous parameter, stationary Markov processes $\{X_t; t \geq 0\}$ taking values in a compact metric space $\mathscr{X}$. Let $p(t, x, dy)$ be the transition probabilities. Assume that $X_0 = x$ and that for each $t > 0$ $\gamma(x, dy) = p(t, x, dy)$ satisfies hypotheses (a)–(c) above [$\beta(dy)$ fixed for all $t > 0$]. Let L be the infinitesimal generator of the semigroup associated with $p(t, x, dy)$ and $\mathscr{D}$ the domain of L. Define for any Borel probability measure ν on $\mathscr{X}$

$$I(\nu) = -\inf_{u > 0, u \in \mathscr{D}} \int_{\mathscr{X}} \left(\frac{Lu}{u}\right)(x)\nu(dx).$$

It is then proved that the distributions on $\mathscr{M}(\mathscr{X})$ of the empirical measures $L_t(\omega, \cdot) = t^{-1} \int_0^t \delta_{X_s(\omega)}(\cdot) ds$, $t > 0$, have a large deviation property with entropy function $I(\nu)$; that is, I is lower semicontinuous, I has compact level sets,

$$\limsup_{t\to\infty} \frac{1}{t} \log P\{L_t(\omega, \cdot) \in K\} \leq -\inf_{\nu \in K} I(\nu) \qquad \text{for each closed set } K \text{ in } \mathscr{X},$$

$$\liminf_{t\to\infty} \frac{1}{t} \log P\{L_t(\omega, \cdot) \in G\} \geq -\inf_{\nu \in G} I(\nu) \qquad \text{for each open set } G \text{ in } \mathscr{X}.$$

In Donsker and Varadhan (1976a), a large deviation property is proved for Markov processes taking values in a complete separable metric space $\mathscr{X}$. The transition probabilities must satisfy strong transitivity and recurrence properties. As a corollary, one obtains the level-2 large deviation property for i.i.d. random vectors taking values in $\mathscr{X}$. Theorem II.4.3 is a special case.

Donsker and Varadhan have applied these large deviation theorems to a number of interesting problems. See references at the end of the book. Large deviation results for empirical measures have been obtained by many people, including Sanov (1957), Sethuraman (1964), Gärtner (1977), Bahadur and Zabell (1979), Bretagnolle (1979), Groeneboom, Oosterhoff, and Ruymgaart (1979), Kac (1980), Chiang (1982), Jain (1982), Csiszár (1984), Stroock (1984), and Pinsky (1985). Luttinger (1982) presents a formal approach.

2 (page 254) Csiszár (1975) studies a large class of minimization problems involving relative entropy. Formula (8.3) is an example as is Problem VIII.6.2 below.

3 (page 259) Donsker and Varadhan (1976a, page 425) prove the contraction principle (8.8) for ρ a Borel probability measure on a Banach space $\mathscr{X}$ which satisfies $\int_{\mathscr{X}} e^{\sigma\|x\|} \rho(dx) < \infty$ for all $\sigma > 0$. If $\mathscr{X} = \mathbb{R}^d$, then this condition is equivalent to $c_\rho(t) < \infty$ for all $t \in \mathbb{R}^d$.

4 (page 261) Theorem VIII.4.3 is proved in Barndorff-Nielsen (1978, Section 9.1) for maximal measures ρ under less restrictive hypotheses on c_ρ.

VIII.6. Problems

VIII.6.1. Let ρ be a maximal Borel probability measure on $\mathbb{R}^d$ such that $c_\rho(t)$ is finite for all $t \in \mathbb{R}^d$.

(a) Show that the Hessian matrix of $c_\rho(t)$ is positive definite for all $t \in \mathbb{R}^d$ and thus that c_ρ is strictly convex on $\mathbb{R}^d$.

(b) Prove that $I_\rho^{(1)}$ is essentially strictly convex.

VIII.6.2 [Kagan, Linnik and Rao (1973, Section 13.2.2)]. We are given a Borel measure ρ on $\mathbb{R}$ (not necessarily a probability measure); bounded, measurable, real-valued functions $h_1, \ldots, h_r$ on $\mathbb{R}$; and real numbers $\alpha_1, \ldots, \alpha_r$. For ν a Borel probability measure on $\mathbb{R}$, define

$$I_\rho^{(2)}(\nu) = \begin{cases} \int_{\mathbb{R}} \log \frac{d\nu}{d\rho}(x)\nu(dx) & \text{if } \nu \ll \rho \text{ and } \int_{\mathbb{R}} \left|\log \frac{d\nu}{d\rho}\right| d\nu < \infty, \\ \infty & \text{otherwise.} \end{cases}$$

Let $\mathscr{P}$ denote the set of all Borel probability measures ν on $\mathbb{R}$ which are absolutely continuous with respect to ρ, for which $d\nu/d\rho$ has the form

$$\frac{d\nu}{d\rho}(x) = \exp(\beta_0 + \beta_1 h_1(x) + \cdots + \beta_r h_r(x)), \qquad \beta_0, \ldots, \beta_r \text{ real},$$

and which satisfy $\int_{\mathbb{R}} h_i(x)\nu(dx) = \alpha_i$, $i = 1, \ldots, r$. Prove that if $\mathscr{P}$ is not empty, then $\mathscr{P}$ equals the set of measures at which

$$\inf\left\{I_\rho^{(2)}(\nu) : \nu \in \mathscr{M}(\mathbb{R}), \int_{\mathbb{R}} h_i(x)\nu_i(dx) = \alpha_i, \qquad i = 1, \ldots, r\right\} \tag{8.15}$$

is attained.

VIII.6.3. Let ρ be Lebesgue measure on $\mathscr{B}(\mathbb{R})$. For each of the following probability densities, find functions $h_1, \ldots, h_r$ such that the probability measure on $\mathbb{R}$ with the given density solves the minimization problem (8.15) for some constants $\alpha_1, \ldots, \alpha_r$.

(i) $f(x) = x^{a-1}(1-x)^{b-1}/\beta(a,b)$ on $(0,1)$, $a > 0$, $b > 0$ (beta density).
(ii) $f(x) = ae^{-ax}$ on $(0, \infty)$, $a > 0$ (exponential density).
(iii) $f(x) = a^b x^{b-1} e^{-ax}/\Gamma(b)$ on $(0, \infty)$, $a > 0$, $b > 0$ (gamma density).
(iv) $f(x) = (2\pi\sigma^2)^{-1/2} \exp[-(x-m)^2/2\sigma^2]$ on $(-\infty, \infty)$, m real, $\sigma^2 > 0$ (Gaussian density).
(v) $f(x) = \frac{1}{2}ae^{-a|x|}$ on $(-\infty, \infty)$, $a > 0$ (Laplace density).

VIII.6.4. Let ρ be the probability measure $\frac{1}{3}(\delta_{(0,0)} + \delta_{(1,0)} + \delta_{(0,1)})$ on $\mathbb{R}^2$. Prove that for each $z \in \operatorname{cc} S_\rho$

$$I_\rho^{(1)}(z) = \inf\left\{I_\rho^{(2)}(\nu) : \nu \in \mathscr{M}(\mathbb{R}^2), \int_{\mathbb{R}^2} x\nu(dx) = z\right\} < \infty \tag{8.16}$$

and determine the measure ν at which the infimum is attained. Theorem VIII.4.1(a) states where the infimum in (8.16) is attained only for points $z \in \operatorname{int}(\operatorname{cc} S_\rho)$.

VIII.6.5. Let ρ be a Borel probability measure on $\mathbb{R}^d$. Prove that $I_\rho^{(2)}$ is a strictly convex function on $\mathscr{M}(\mathbb{R}^d)$; i.e., if μ and ν are distinct Borel probability measures on $\mathbb{R}^d$, then for all $0 < \lambda < 1$

$$I_\rho^{(2)}(\lambda\mu + (1-\lambda)\nu) < \lambda I_\rho^{(2)}(\mu) + (1-\lambda) I_\rho^{(2)}(\nu).$$

VIII.6.6. (a) Let ρ be a Borel probability measure on $\mathbb{R}^d$ and A a nonempty compact convex subset of $\mathscr{M}(\mathbb{R}^d)$. Prove that if $\inf_{\nu \in A} I_\rho^{(2)}(\nu)$ is finite, then $I_\rho^{(2)}(\nu)$ attains its infimum over A at a unique measure.

(b) Let ρ be a nondegenerate Borel probability measure on $\mathbb{R}^d$ such that $c_\rho(t)$ is finite for all $t \in \mathbb{R}^d$. Let z be a point in $\operatorname{rbd}(\operatorname{cc} S_\rho)$ such that

$$\inf\left\{ I_\rho^{(2)}(v) : v \in \mathscr{M}(\mathbb{R}^d), \int_{\mathbb{R}^d} x v(dx) = z \right\} < \infty. \tag{8.17}$$

Prove that the infimum is attained at a unique measure $\bar{\rho}$ and that $\bar{\rho}$ is the weak limit of measures $\{\rho_{t_n}; n = 1, 2, \dots\}$ for some sequence $\{t_n; n = 1, 2, \dots\}$ in $\mathbb{R}^d$ such that $\|t_n\| \to \infty$ [ρ_{t_n} is defined in (8.6)]. Theorem VIII.4.1(a) states where the infimum in (8.17) is attained only for points $z \in \operatorname{ri}(\operatorname{cc} S_\rho)$. [*Hint:* Use (8.8) and the following properties. $I_\rho^{(2)}(v)$ is strictly convex; $I_\rho^{(2)}(v)$ is lower semicontinuous; $I_\rho^{(2)}(v)$ has compact level sets; if $\sup_n I_\rho^{(2)}(v_n)$ is finite for some sequence $\{v_n\}$, then $\lim_{A \to \infty} \sup_n \int_{\|x\| \geq A} \|x\| v_n(dx) = 0$. The last three properties are proved in Donsker and Varadhan (1975a, 1976a).]

VIII.6.7 [Barndorff-Nielsen (1978, page 105)]. Let ρ be a Borel probability measure on $\mathbb{R}^d$ such that $c_\rho(t)$ is finite for all $t \in \mathbb{R}^d$. S_ρ denotes the support of ρ. Prove that for any nonzero $t \in \mathbb{R}^d$

$$\lim_{\lambda \to \infty} [-\lambda \sigma(t|S_\rho) + c_\rho(\lambda t)] = \log \rho\{x : \langle t, x \rangle = \sigma(t|S_\rho)\},$$

where $\sigma(t|S_\rho) = \sup_{x \in S_\rho} \langle t, x \rangle$ is the support function of S_ρ [see Problem VI.7.11].

VIII.6.8 [Barndorff-Nielsen (1978, pages 140–143)]. Let ρ be a maximal Borel probability measure on $\mathbb{R}^d$ such that $c_\rho(t)$ is finite for all $t \in \mathbb{R}^d$. Let z be a boundary point of $\operatorname{cc} S_\rho$ and $\mathscr{H}_z$ the set of all supporting hyperplanes to $\operatorname{cc} S_\rho$ at z. This problem shows that z does not belong to $\operatorname{dom} I_\rho^{(1)}$ if $\rho(H) = 0$ for some $H \in \mathscr{H}_z$. [Theorem VIII.4.3(c)].

(a) Any $H \in \mathscr{H}_z$ can be written as $\{x \in \mathbb{R}^d : \langle x - z, \gamma \rangle = 0\}$, where γ is a unit vector and $\langle x - z, \gamma \rangle \leq 0$ for all $x \in \operatorname{cc} S_\rho$. Prove this statement.

(b) Suppose that $H \in \mathscr{H}_z$ has unit normal vector γ. Prove that

$$I_\rho^{(1)}(z) \geq \sup_{\lambda \in \mathbb{R}} \{\lambda \langle \gamma, z \rangle - c_\rho(\lambda \gamma)\} = \lim_{\lambda \to \infty} \{\lambda \langle \gamma, z \rangle - c_\rho(\lambda \gamma)\}.$$

Conclude that if $\rho(H) = 0$, then z does not belong to $\operatorname{dom} I_\rho^{(1)}$. [*Hint:* Problem VIII.6.7.]

(c) Prove that if ρ is absolutely continuous with respect to Lebesgue measure, then $\operatorname{dom} I_\rho^{(1)} = \operatorname{int}(\operatorname{cc} S_\rho)$.

VIII.6.9. The purpose of this problem is to prove Theorem VIII.4.4. Assume the hypotheses of that theorem.

(a) Suppose that the dimension m of $\operatorname{aff} S_\rho$ is less than d. Prove that there exist an orthogonal matrix U and a vector γ such that $\operatorname{aff} S_\rho = UA + \gamma$, where $A = \{x \in \mathbb{R}^d : x_{m+1} = \cdots = x_d = 0\}$. For Borel subsets B of A, define $\bar{\rho}\{B\} = \rho\{UB + \gamma\}$. Prove that for suitable choice of γ

$$c_\rho(t) = \langle t, \gamma \rangle + c_{\bar{\rho}}(U^T(t - \gamma)), \qquad t \in \operatorname{aff} S_\rho,$$

where U^T is the transpose of U. Relate the range of $\nabla c_\rho(t)$, $t \in \operatorname{aff} S_\rho$, to the range of $\nabla c_\rho(s)$, $s \in A$, and prove parts (a) and (b) of Theorem VIII.4.4.

(b) Prove part (c) of Theorem VIII.4.4.

VIII.6.10. The purpose of this problem is to generalize Theorem VIII.3.4 to higher dimensions. Let ρ be a maximal Borel probability measure on $\mathbb{R}^d$, $d \geq 2$, such that $c_\rho(t)$ is finite for all $t \in \mathbb{R}^d$.

(a) Prove that $I_\rho^{(1)}(z)$ is essentially smooth.

(b) Prove that $\nabla I_\rho^{(1)}$ defines a one-to-one mapping of $\operatorname{int}(\operatorname{cc} S_\rho)$ onto $\mathbb{R}^d$ with inverse ∇c_ρ.

(c) Using the implicit function theorem for analytic functions [Bochner and Martin (1948, page 39)], prove that $I_\rho^{(1)}(z)$ is a real analytic function of $z \in \operatorname{int}(\operatorname{cc} S_\rho)$. [*Hint:* For any point $z \in \operatorname{int}(\operatorname{cc} S_\rho)$, there exists a unique point $t(z) \in \mathbb{R}^d$ such that $\nabla c_\rho(t(z)) = z$ [Theorem VIII.4.3(a)]. The Hessian matrix of c_ρ at $t(z)$ is positive-definite [Problem VIII.6.1(a)]. The mapping $\nabla c_\rho(t)$ can be continued into the complex space $\mathbb{C}^d$ to be an analytic mapping [Bochner and Martin (1948, page 34)].]