

## 5 Large Deviations

### 5.1 Introduction

The theory of large deviations is concerned with the estimation of probabilities of rare events and expected values that are largely determined by rare events. It is also concerned with the following qualitative question: given that a rare event has occurred, how did it happen? As one might suspect, the theory has its roots in problems of insurance and risk. It now finds applications in many areas, including statistical mechanics, communication and information theories, and elsewhere. The theory has many parallels to the theory of weak convergence. In this section we will discuss the basic framework of large deviations, give a few examples, discuss the important change-of-measure technique that is often used to estimate lower bounds in large deviations, and also show how one can obtain large deviation results using weak convergence methods.

The definition of a Large Deviations Principle (LDP) in the Polish space setting is as follows.

**Definition 1.** *A sequence  $\{Y_n, n \in \mathbb{N}\}$  of  $S$ -valued random variables is said to satisfy the LDP with rate function  $I : S \rightarrow [0, \infty]$  if the following hold.*

1. *For any open set  $A \subset S$ ,*

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log P \{Y_n \in A\} \geq - \inf_{x \in A} I(x).$$

2. *For any closed set  $B \subset S$ ,*

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log P \{Y_n \in B\} \leq - \inf_{x \in B} I(x).$$

3. *For any  $K < \infty$ , the set  $\{x : I(x) \leq K\}$  is compact.*

You should observe the analogy with parts of the Portmanteau Theorem. Suppose that the set  $C \subset S$  satisfies

$$\inf_{x \in C^\circ} I(x) = \inf_{x \in \bar{C}} I(x).$$

Then in fact

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P \{Y_n \in C\} = - \inf_{x \in C} I(x),$$

which resembles yet another of the limits that appear in the Portmanteau Theorem. For such sets  $C$  we are told that, roughly speaking,

$$P\{Y_n \in C\} \approx e^{-n \inf_{x \in C} I(x)}.$$

Thus large deviations tells us that these probabilities decay exponentially, and also identifies the rate of decay. The condition  $\inf_{x \in C^\circ} I(x) = \inf_{x \in \bar{C}} I(x)$  does not mean that the event  $P\{Y_n \in \partial C\}$  is negligible from the point of view of large deviations, but rather that the boundary does not contribute enough of the total probability assigned to  $\bar{C}$  so as to distinguish its exponential rate of decay from that of  $C^\circ$ . Thus there is a close parallel, at least in form, to convergence in distribution. While it is possible that an LDP could involve a normalization sequence other than  $1/n$ , this normalization will hold for all the examples we consider.

Next suppose that the set  $C$  is also closed. In this case there is at least one infimizing point  $x^*$  in  $\inf_{x \in C} I(x)$ , and one can show that if  $G$  is the ensemble of all such infimizing points and  $\varepsilon > 0$ , then

$$P\{Y_n \in G^\varepsilon | Y_n \in C\} \rightarrow 1,$$

where  $G^\varepsilon \doteq \{x : d(x, G) \leq \varepsilon\}$ . Thus large deviations gives us not only quantitative information (how likely is the event  $\{Y_n \in C\}$ ), but also qualitative information (given that the unlikely event  $\{Y_n \in C\}$  occurred, it is most likely that in fact  $\{Y_n \in G^\varepsilon\}$ ).

In the setting of weak convergence, the convergence of expected values is taken as the definition, and the various convergences with respect to open and closed sets are then obtained as a consequence. In the setting of large deviations, the opposite situation has occurred. The definition given above in terms of lower and upper bounds for open and closed sets, respectively, was adopted as the definition of an LDP. However, one can also phrase the LDP in terms of limits of expected values against bounded continuous functions. In fact, in the setting of a Polish space, we have the following result.

**Theorem 1.** *A sequence of random variables  $\{Y_n\}$  that takes values in a Polish space  $S$  satisfies the LDP with rate function  $I$  if and only if it satisfies the following Laplace Principle with rate  $I$ :*

- for any  $K < \infty$ , the set  $\{x : I(x) \leq K\}$  is compact, and
- for any bounded and continuous function  $f : S \rightarrow \mathbb{R}$ ,

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log E e^{-nf(Y_n)} = \inf_{x \in S} [f(x) + I(x)].$$

The proof of this fact is very similar to one of the standard proofs of the Portmanteau Theorem, and involves showing how to properly approximate such an  $f$  by combinations of indicator functions of sets, and conversely. See, e.g., Theorems 1.2.1 and 1.2.3 of my book with Richard Ellis.

Another parallel with weak convergence theory is the following analogue of the Continuous Mapping Theorem.

**Theorem 2.** (CONTRACTION PRINCIPLE) *Let  $\{Y_n, n \in \mathbb{N}\}$  be  $S_1$ -valued random variables that satisfy the LDP with rate function  $I$ . Let  $G : S_1 \rightarrow S_2$ , where  $S_2$  is another Polish space and  $G$  is continuous. Then  $\{G(Y_n), n \in \mathbb{N}\}$  satisfy the LDP with rate function*

$$J(y) = \inf \{I(x) : G(x) = y\}.$$

**Proof:** There are only two items to prove. The first is that  $\{y \in S_2 : J(y) \leq K\}$  is compact. However, this is automatic, since this set is just the forward image of the compact set  $\{x \in S_1 : I(x) \leq K\}$  under the continuous function  $G$ . The second is that for any bounded and continuous function  $f : S_2 \rightarrow \mathbb{R}$ ,

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log E e^{-nf(G(Y_n))} = \inf_{y \in S_2} [f(y) + J(y)].$$

This is also easy, since

$$\begin{aligned} \lim_{n \rightarrow \infty} -\frac{1}{n} \log E e^{-nf(G(Y_n))} &= \inf_{x \in S_1} [f(G(x)) + I(x)] \\ &= \inf_{y \in S_2} \inf_{x \in S_1 : G(x)=y} [f(G(x)) + I(x)] \\ &= \inf_{y \in S_2} [f(y) + J(y)]. \end{aligned}$$

■

In the next section we prove what is probably the most well-known result in the theory, namely, Cramér's Theorem. This result identifies the large deviation properties of the empirical mean for a sequence of iid random variables. According to the LLN, the empirical mean converges to the mean of the distribution used to generate the random variables. Here we want to estimate the probability that the empirical mean falls in a set that does not contain the LLN limit. We present a somewhat classical proof that is based on Chebyshev's inequality for upper bounds and a clever change-of-measure argument for the lower bounds.

In the last section we use the theory of weak convergence to prove Sanov's Theorem, which is another of the fundamental results in large deviations

theory, and a cornerstone of information theory. We then derive Cramér’s Theorem (again) as a corollary by using the Contraction Principle. To keep the arguments in this section brief we use assumptions that are much stronger than necessary.

## 5.2 Cramér’s Theorem

The most basic large deviation result studies deviations of the empirical mean from the true mean for a sequence of iid random variables. It is due to H. Cramér, who happens to have been U. Grenander’s advisor. Here is the setup. Let  $\{X_n, n \in \mathbb{N}\}$  be a sequence of iid  $\mathbb{R}$ -valued random variables. Assume that the moment generating function

$$M(\alpha) \doteq Ee^{\alpha X_1}$$

is finite for all  $\alpha \in \mathbb{R}$ . Define  $H(\alpha) \doteq \log M(\alpha)$ . We will show in a moment that  $H$  is a convex function, and that  $H(0) = 0, H_\alpha(0) = \frac{d}{d\alpha} H(\alpha)|_{\alpha=0} = EX_1$ . Let  $L$  be the *Legendre transform* of  $H$ : for  $\beta \in \mathbb{R}$

$$L(\beta) = \sup_{\alpha \in \mathbb{R}} [\alpha\beta - H(\alpha)].$$

The geometry of the Legendre transform (as it applies in Cramér’s Theorem) is illustrated in a figure on the next page. The following are noteworthy.

- Consider first the line  $\beta_2\alpha$ , which is tangent to  $H(\alpha)$  at  $\alpha = 0$ . The slope of this line is of course  $H_\alpha(0)$ , which in the figure is negative (indicating that the mean of the distribution is negative, and hence the tendency of the process is to move to the left). Since the line is always *below*  $H(\alpha)$ , the supremum of  $\alpha\beta_2 - H(\alpha)$  is exactly zero, and it is achieved at  $\alpha = 0$ . Thus  $L(\beta_2) = 0$ .
- Next consider the line  $\beta_1\alpha$ . We are interested in the maximum vertical gap between this line and  $H(\alpha)$ . Clearly this will happen at a point where the slope of  $H(\alpha)$  and that of  $\beta_1\alpha$  coincide. In general, there need not be such a point, and in fact the maximum may be  $\infty$ . However, in the figure there is such a point, and it is denoted by  $\alpha^*$ . As indicated, this maximal vertical gap is  $L(\beta_1)$ .
- For *any* line  $\beta\alpha$  we can always consider the point  $\alpha = 0$ , and since  $H(0) = 0$  we always have  $L(\beta) \geq 0$ .

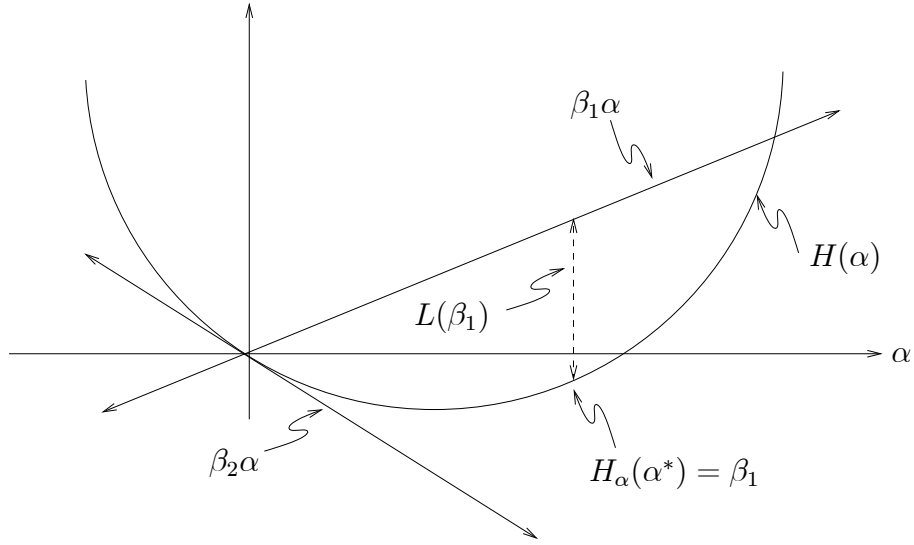


Figure 1. Geometry of the Legendre Transform.

**Lemma 1.** (PROPERTIES OF  $H$ ) Let  $H(\alpha) = \log E \exp[\alpha X_1]$  for  $\alpha \in \mathbb{R}$ , and assume that  $H(\alpha) < \infty$  for each  $\alpha \in \mathbb{R}$ . Then  $H$  has the following properties.

- $H(0) = 0$ .
- $H$  is convex.
- $H$  is continuously differentiable, and  $H_\alpha(0) = EX_1$ .

**Proof:** The first item follows directly from the definition of  $H$ . The convexity is a consequence of Holder's inequality. If  $p_1 + p_2 = 1$  and  $p_1 \geq 0, p_2 \geq 0$ , then

$$E e^{(p_1 \alpha_1 + p_2 \alpha_2) X_1} \leq [E e^{\alpha_1 X_1}]^{p_1} [E e^{\alpha_2 X_1}]^{p_2},$$

and so by taking logarithms

$$H(p_1 \alpha_1 + p_2 \alpha_2) \leq p_1 H(\alpha_1) + p_2 H(\alpha_2),$$

which is convexity.

The proof of the differentiability and its explicit form have a proof that is very similar to one of your homework assignments. One uses the LDCT and the dominating function

$$e^{|\alpha+1|X_1} \leq e^{|\alpha+1|X_1} + e^{-|\alpha+1|X_1},$$

which is integrable because  $H(|(\alpha + 1)|) < \infty$  and  $H(-|(\alpha + 1)|) < \infty$ . ■

**Lemma 2.** (PROPERTIES OF  $L$ ) *Let  $L$  be the Legendre transform of  $H$ , where  $H$  is as in the last lemma. Then  $L$  has the following properties.*

- $L : \mathbb{R} \rightarrow [0, \infty]$ ,
- $L$  is convex and lower semicontinuous,
- $L(\beta) = 0$  if  $\beta = EX_1$ , and  $L(\beta) > 0$  if  $\beta \neq EX_1$ .

**Proof:** As discussed previously, the choice  $\alpha = 0$  shows that

$$L(\beta) = \sup_{\alpha \in \mathbb{R}} [\alpha\beta - H(\alpha)] \geq -H(0) = 0.$$

The convexity is because  $L$  is itself the supremum of convex functions. Fix  $\beta \in \mathbb{R}$ . If  $p_1 + p_2 = 1$ ,  $p_1 \geq 0, p_2 \geq 0$  and  $\beta = p_1\beta_1 + p_2\beta_2$ , then for any  $\alpha \in \mathbb{R}$

$$\alpha\beta - H(\alpha) = p_1(\alpha\beta_1 - H(\alpha)) + p_2(\alpha\beta_2 - H(\alpha)) \leq p_1L(\beta_1) + p_2L(\beta_2).$$

Now supremize over  $\alpha$  to get  $L(\beta) \leq p_1L(\beta_1) + p_2L(\beta_2)$ . The proof of lower semicontinuity is similar. Let  $\beta_i \rightarrow \beta$  as  $i \rightarrow \infty$ . For any  $\alpha \in \mathbb{R}$

$$\liminf_{i \rightarrow \infty} L(\beta_i) \geq \liminf_{i \rightarrow \infty} [\alpha\beta_i - H(\alpha)] = \alpha\beta - H(\alpha).$$

Supremizing over  $\alpha$  shows that  $\liminf_{i \rightarrow \infty} L(\beta_i) \geq L(\beta)$ . The first part of the last item follows from calculus (see also the figure). To see the second part, observe that  $L(\beta) \leq 0$  implies  $\alpha\beta \leq H(\alpha)$  for all  $\alpha$ , and since these functions of  $\alpha$  agree at  $\alpha = 0$  they must have the same derivative there, i.e.,  $\beta = H_\alpha(0)$ . ■

Some examples of  $H - L$  pairs are as follows.

**Example 1.** *Suppose that  $X_1$  is Bernoulli with*

$$P\{X_1 = 0\} = 1 - p, \quad P\{X_1 = 1\} = p$$

for  $p \in (0, 1)$ . Then

$$H(\alpha) = \log((1 - p) + pe^\alpha)$$

and (with the understanding that  $0 \log 0 = 0$ )

$$L(\beta) = \begin{cases} \beta \log\left(\frac{\beta}{p}\right) + (1 - \beta) \log\left(\frac{1 - \beta}{1 - p}\right) & \beta \in [0, 1] \\ \infty & \beta \notin [0, 1] \end{cases}.$$

**Example 2.** Suppose that  $X_1$  is Poisson with parameter  $\lambda > 0$ , so that  $P\{X_1 = n\} = e^{-\lambda}\lambda^n/n!$  for  $n \in \mathbb{N}_0$ . Then

$$H(\alpha) = \lambda(e^\alpha - 1)$$

and

$$L(\beta) = \begin{cases} \beta \log\left(\frac{\beta}{\lambda}\right) - \beta + \lambda & \beta \geq 0 \\ \infty & \beta < 0 \end{cases}.$$

**Example 3.** Suppose that  $X_1$  is Gaussian  $N(b, \sigma^2)$ . Then

$$H(\alpha) = \frac{\alpha^2 \sigma^2}{2} + \alpha b$$

and

$$L(\beta) = \frac{1}{2}(\beta - b)^2.$$

**Example 4.** As with characteristic functions, we have nice scaling properties. Let  $H_x$  and  $L_x$  be associated with  $X_1$ , and let  $Y_1 = aX_1 + b$  for real numbers  $a \neq 0$  and  $b$ . If  $H_y$  and  $L_y$  be associated with  $Y_1$ , then

$$H_y(\alpha) = \log Ee^{\alpha a X + \alpha b} = H_x(\alpha a) + \alpha b,$$

and

$$\begin{aligned} L_y(\beta) &= \sup_{\alpha \in \mathbb{R}} [\alpha \beta - H_x(\alpha a) - \alpha b] \\ &= \sup_{\alpha \in \mathbb{R}} \left[ \alpha \frac{(\beta - b)}{a} - H_x(\alpha) \right] \\ &= L_x\left(\frac{(\beta - b)}{a}\right). \end{aligned}$$

Here is the statement of the theorem.

**Theorem 1.** (CRAMÉR'S THEOREM) Assume that  $H(\alpha) < \infty$  for  $\alpha \in \mathbb{R}$  and define  $S_n = \sum_{i=1}^n X_i$ . Then  $S_n/n$  satisfies the LDP with rate function  $L$ .

**A comparison of the CLT and LDP.** It is instructive to compare the information provided by a LDP with what one can get from the CLT. It turns out that these two limit theorems answer different questions, and provide complementary information. In the setting of sums of iid random variables, the CLT tells us something about probabilities for sets located

distance  $n^{1/2}$  from the mean, whereas the LDP considers sets that are much further into the tail (sets distance  $n$  from the mean). Suppose, for example, that  $EX_1 = 0$ , and that the moment generating function is finite for all  $\alpha$ . (Note that this implies a finite second moment, and hence the CLT is also valid.) To simplify the notation assume we have normalized so that  $EX_1^2 = 1$ . Consider a set  $A$  such that  $P\{\theta \in \partial A\} = 0$ , where  $\theta$  is  $N(0, 1)$ . Then by the CLT

$$P\left\{\frac{S_n}{n^{1/2}} \in A\right\} \rightarrow P\{\theta \in \partial A\},$$

and so

$$P\{S_n \in n^{1/2}A\} \approx P\{\theta \in \partial A\},$$

where  $\approx$  means that the ratio tends to 1 as  $n \rightarrow \infty$ . Note that the CLT can tell us nothing about an estimate of the form  $P\{S_n \in nA\}$ , because this would require a limit for

$$P\left\{\frac{S_n}{n^{1/2}} \in n^{1/2}A\right\},$$

which is not valid since the set  $n^{1/2}A$  depends on  $n$ . In contrast, the LDP tells us about  $P\{S_n \in nA\}$ . Suppose that  $L$  is the rate function, and that

$$C \doteq \inf_{\beta \in A^\circ} L(\beta) = \inf_{\beta \in \bar{A}} L(\beta).$$

Then

$$P\{S_n \in nA\} = P\left\{\frac{S_n}{n} \in A\right\} \approx e^{-nC},$$

in the sense that the ratio of the logarithms converge to 1. The LDP does not directly give us any CLT information, since the set  $A$  here is also fixed, and a CLT type statement would require

$$P\left\{\frac{S_n}{n} \in \frac{1}{n^{1/2}}A\right\}.$$

(It does turn out that there is some CLT information hidden in the rate function, in that the local expansion up to order 2 of the rate function around its minimum point is determined by just means and variances, and is the same as what one would expect from a formal CLT approximation. This is related to the fact that *moderate deviations* provide a bridge between large deviations and the CLT, but we will not pursue this point here.)

The following fact and some variations are used all the time in large deviations.

**Lemma 3.** Let sequences  $\{a_n\}, \{b_n\} \subset [0, 1]$  be given such that

$$\begin{aligned} -\frac{1}{n} \log a_n &\rightarrow u \in [0, \infty] \\ -\frac{1}{n} \log b_n &\rightarrow v \in [0, \infty]. \end{aligned}$$

Then

$$-\frac{1}{n} \log (a_n + b_n) \rightarrow u \wedge v.$$

This simply asserts that if two probabilities are decaying exponentially, then the one with the slower decay rate dominates the sum.

**Proof:** Suppose without loss that  $u \leq v$ . The bound

$$\limsup_{n \rightarrow \infty} -\frac{1}{n} \log (a_n + b_n) \leq \limsup_{n \rightarrow \infty} -\frac{1}{n} \log a_n \leq u$$

is trivial. For  $\varepsilon > 0$  choose  $N < \infty$  such that  $n \geq N$  implies  $-\frac{1}{n} \log b_n \geq v - \varepsilon$  and  $-\frac{1}{n} \log a_n \geq u - \varepsilon$ . In other words,  $b_n \leq e^{-n(v-\varepsilon)}$ ,  $a_n \leq e^{-n(u-\varepsilon)}$ . Then

$$(a_n + b_n) \leq e^{-n(u-\varepsilon)} + e^{-n(v-\varepsilon)} = e^{-n(u-\varepsilon)} \left(1 + e^{-n(v-u)}\right) \leq 2e^{-n(u-\varepsilon)}.$$

Thus

$$\liminf_{n \rightarrow \infty} -\frac{1}{n} \log (a_n + b_n) \geq u - \varepsilon,$$

and the result follows since  $\varepsilon > 0$  is arbitrary. ■

**Proof of Cramér's Theorem:** We follow Varadhan's proof.

UPPER BOUND. The large deviation upper bound follows from Chebyshev's inequality. Fix a closed set  $B \subset \mathbb{R}$ . If  $EX_1 \in B$  then  $\inf_{\beta \in B} L(\beta) = 0$ , and the upper bound is automatic since  $P\{S_n/n \in B\} \leq 1$  implies

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log P\{S_n/n \in B\} \leq 0.$$

We next show how to approximate  $B$  "from above" by a simpler closed set. Let

$$\begin{aligned} b_1 &= \inf \{\beta : \beta \in B, \beta > EX_1\} \\ b_2 &= \sup \{\beta : \beta \in B, \beta < EX_1\}. \end{aligned}$$

(If  $b_1 = \infty$  or  $b_2 = -\infty$  then the corresponding term can be ignored in proving the upper bound.) Then since  $L$  is convex with its minimum at  $\beta = EX_1$ ,

$$\begin{aligned}\inf_{\beta \in B, \beta > EX_1} L(\beta) &= L(b_1) \\ \inf_{\beta \in B, \beta < EX_1} L(\beta) &= L(b_2)\end{aligned}$$

and  $B \subset (-\infty, b_2] \cup [b_1, \infty)$ . We will prove

$$\begin{aligned}\limsup_{n \rightarrow \infty} \frac{1}{n} \log P \{S_n/n \in [b_1, \infty)\} &\leq -L(b_1) \\ \limsup_{n \rightarrow \infty} \frac{1}{n} \log P \{S_n/n \in (-\infty, b_2]\} &\leq -L(b_2),\end{aligned}$$

which using the lemma will imply the needed upper bound

$$\begin{aligned}\limsup_{n \rightarrow \infty} \frac{1}{n} \log P \{S_n/n \in B\} &\leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log (P \{S_n/n \in [b_1, \infty)\} + P \{S_n/n \in (-\infty, b_2]\}) \\ &\leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log (P \{S_n/n \in [b_1, \infty)\}) \vee \limsup_{n \rightarrow \infty} \frac{1}{n} \log (P \{S_n/n \in (-\infty, b_2]\}) \\ &\leq -(L(b_1) \wedge L(b_2)) \\ &= -\inf_{\beta \in B} L(\beta).\end{aligned}$$

The proof is given for just  $\limsup_{n \rightarrow \infty} \frac{1}{n} \log P \{S_n/n \in [b_1, \infty)\} \leq -L(b_1)$ , since the proof for the other half-interval is analogous. Let  $\alpha \geq 0$ . By Chebyshev's inequality and the independence

$$\begin{aligned}P \{S_n/n \geq b_1\} &= P \left\{ e^{\frac{1}{n}\alpha S_n} \geq e^{\alpha b_1} \right\} \\ &\leq e^{-\alpha b_1} E e^{\frac{1}{n}\alpha S_n} \\ &= e^{-\alpha b_1} e^{nH(\frac{1}{n}\alpha)},\end{aligned}$$

and so

$$\log P \{S_n/n \geq b_1\} \leq -\alpha b_1 + nH\left(\frac{1}{n}\alpha\right).$$

Now since  $b_1 \geq EX_1 = H_\alpha(0)$ , the supremum of  $[\alpha b_1 - H(\alpha)]$  over  $\alpha \in \mathbb{R}$  occurs at some  $\alpha \geq 0$  (see Figure 1). Therefore by letting  $\bar{\alpha} = \alpha/n$  in the

second equality,

$$\begin{aligned}
\log P \{S_n/n \geq b_1\} &\leq \inf_{\alpha \geq 0} - \left[ \alpha b_1 - nH \left( \frac{1}{n} \alpha \right) \right] \\
&= - \sup_{\alpha \geq 0} \left[ \alpha b_1 - nH \left( \frac{1}{n} \alpha \right) \right] \\
&= -n \sup_{\bar{\alpha} \geq 0} [\bar{\alpha} b_1 - H(\bar{\alpha})] \\
&= -n \sup_{\bar{\alpha} \in \mathbb{R}} [\bar{\alpha} b_1 - H(\bar{\alpha})] \\
&= -nL(b_1),
\end{aligned}$$

and so

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log P \{S_n/n \in [b_1, \infty)\} \leq -L(b_1).$$

This completes the proof of the upper bound.

**LOWER BOUND.** To prove the lower bound, we first assume that  $H_\alpha(\alpha) \rightarrow \infty$  as  $\alpha \rightarrow \infty$  and  $H_\alpha(\alpha) \rightarrow -\infty$  as  $\alpha \rightarrow -\infty$ . Let open  $A \subset \mathbb{R}$  be given. We first show how one can approximate  $A$  “from below” by a simpler open set. Given any  $\delta > 0$  there is  $\bar{\beta} \in A$  within  $\delta$  of the infimum:  $L(\bar{\beta}) \leq \inf_{\beta \in A} L(\beta) + \delta$ . Since  $A$  is open there is  $\varepsilon > 0$  such that  $(\bar{\beta} - \varepsilon, \bar{\beta} + \varepsilon) \subset A$ . We claim that to prove the lower bound it suffices prove

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log P \{S_n/n \in (\bar{\beta} - \varepsilon, \bar{\beta} + \varepsilon)\} \geq -L(\bar{\beta}).$$

This is true since

$$\begin{aligned}
&\liminf_{n \rightarrow \infty} \frac{1}{n} \log P \{S_n/n \in A\} \\
&\geq \liminf_{n \rightarrow \infty} \frac{1}{n} \log P \{S_n/n \in (\bar{\beta} - \varepsilon, \bar{\beta} + \varepsilon)\} \\
&\geq -L(\bar{\beta}) \\
&\geq - \inf_{\beta \in A} L(\beta) - \delta,
\end{aligned}$$

and since  $\delta > 0$  is arbitrary.

Owing to the temporary assumptions on  $H$ , there is  $\bar{\alpha}$  which achieves the supremum in  $[\alpha \bar{\beta} - H(\alpha)]$ . (It is given by  $\bar{\beta} = H_\alpha(\bar{\alpha})$ . Such a root exists by the monotonicity and continuity of  $H_\alpha(\alpha)$ , and the fact that its

range is all  $\mathbb{R}$ .) Letting  $\mu$  denote the distribution of  $X_1$  on  $\mathbb{R}$ , we define a new probability measure by

$$\frac{d\nu}{d\mu}(x) = e^{x\bar{\alpha} - H(\bar{\alpha})}.$$

Let  $\bar{X}_i$  be iid with distribution  $\nu$ , and to keep notations distinct assume they are defined on a probability space  $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{P})$ . Then for any bounded measurable function  $F$  and  $n$  we have

$$\begin{aligned} EF(X_1, \dots, X_n) &= \int_{\mathbb{R} \times \dots \times \mathbb{R}} F(x_1, \dots, x_n) \mu(dx_1) \cdots \mu(dx_n) \\ &= \int_{\mathbb{R} \times \dots \times \mathbb{R}} F(x_1, \dots, x_n) e^{-x_1 \bar{\alpha} + H(\bar{\alpha})} \cdots e^{-x_n \bar{\alpha} + H(\bar{\alpha})} \nu(dx_1) \cdots \nu(dx_n) \\ &= \bar{E}F(\bar{X}_1, \dots, \bar{X}_n) e^{-\bar{X}_1 \bar{\alpha} + H(\bar{\alpha})} \cdots e^{-\bar{X}_n \bar{\alpha} + H(\bar{\alpha})} \\ &= \bar{E}F(\bar{X}_1, \dots, \bar{X}_n) e^{-n[(\bar{S}_n/n)\bar{\alpha} - H(\bar{\alpha})]}.\end{aligned}$$

In particular, for  $\bar{\varepsilon} \in (0, \varepsilon)$

$$P\{S_n/n \in (\bar{\beta} - \bar{\varepsilon}, \bar{\beta} + \bar{\varepsilon})\} = \bar{E}1_{\{\bar{S}_n/n \in (\bar{\beta} - \bar{\varepsilon}, \bar{\beta} + \bar{\varepsilon})\}} e^{-n[(\bar{S}_n/n)\bar{\alpha} - H(\bar{\alpha})]}.$$

Now

$$\log \bar{E}e^{\theta \bar{X}_1} = \log \int_{\mathbb{R}} e^{\theta x} e^{x\bar{\alpha} - H(\bar{\alpha})} \mu(dx) = H(\theta + \bar{\alpha}) - H(\bar{\alpha}),$$

and so the choice of  $\bar{\alpha}$  implies

$$\bar{E}\bar{X}_1 = H_\alpha(\theta + \bar{\alpha})|_{\theta=0} = H_\alpha(\bar{\alpha}) = \bar{\beta}.$$

By the LLN

$$\bar{P}\{\bar{S}_n/n \in (\bar{\beta} - \bar{\varepsilon}, \bar{\beta} + \bar{\varepsilon})\} \rightarrow 1,$$

and so

$$\begin{aligned} &\liminf_{n \rightarrow \infty} \frac{1}{n} \log P\{S_n/n \in (\bar{\beta} - \varepsilon, \bar{\beta} + \varepsilon)\} \\ &\geq \liminf_{n \rightarrow \infty} \frac{1}{n} \log P\{S_n/n \in (\bar{\beta} - \bar{\varepsilon}, \bar{\beta} + \bar{\varepsilon})\} \\ &= \liminf_{n \rightarrow \infty} \frac{1}{n} \log \bar{E}1_{\{\bar{S}_n/n \in (\bar{\beta} - \bar{\varepsilon}, \bar{\beta} + \bar{\varepsilon})\}} e^{-n[(\bar{S}_n/n)\bar{\alpha} - H(\bar{\alpha})]} \\ &\geq -\sup_{|a| \leq \bar{\varepsilon}} [(\bar{\beta} + a)\bar{\alpha} - H(\bar{\alpha})] \\ &= -[\bar{\beta}\bar{\alpha} - H(\bar{\alpha})] - |\bar{\varepsilon}\bar{\alpha}| \\ &= -L(\bar{\beta}) - |\bar{\varepsilon}\bar{\alpha}|.\end{aligned}$$

The lower bound now follows by letting  $\bar{\varepsilon} \rightarrow 0$ .

Finally we remove the temporary assumption on  $H$ . The argument uses a clever mollification idea. Fix  $\beta \in \mathbb{R}$ . If  $L(\beta) = \infty$  there is nothing to prove, and so we can assume  $L(\beta) < \infty$ . Let  $\delta > 0$ , and let  $Y_i$  be iid Gaussian  $N(0, \delta)$  and independent of the  $X_i$ . Let  $Z_i = X_i + Y_i$ . A direct computation shows that if  $\bar{H}_\delta(\alpha) = \log E \exp \alpha Z_i$ , then  $\bar{H}_\delta(\alpha) = H(\alpha) + \frac{\delta}{2} |\alpha|^2$ . It is easy to check that  $\bar{H}_\delta(\alpha) \geq H(\alpha)$  for all  $\alpha \in \mathbb{R}$  implies  $\bar{L}_\delta(\beta) \leq L(\beta)$  for all  $\beta \in \mathbb{R}$ . Furthermore, since  $H(\alpha) \geq \alpha EX_1$  and  $H(\alpha)$  is convex,

$$\lim_{\alpha \rightarrow \infty} H_\alpha(\alpha) \geq EX_1, \quad \lim_{\alpha \rightarrow -\infty} H_\alpha(\alpha) \leq -EX_1$$

(recall that the derivative of a convex function is monotonic). Therefore

$$\lim_{\alpha \rightarrow \infty} (\bar{H}_\delta)_\alpha(\alpha) \geq \infty, \quad \lim_{\alpha \rightarrow -\infty} (\bar{H}_\delta)_\alpha(\alpha) \leq -\infty,$$

i.e.,  $\bar{H}_\delta$  satisfies the temporary assumptions. Let  $Q_n = Z_1 + \cdots + Z_n$  and  $R_n = Y_1 + \cdots + Y_n$ . Then by the previous lower bound

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log P \{Q_n/n \in (\beta - \varepsilon/2, \beta + \varepsilon/2)\} \geq -\bar{L}_\delta(\beta),$$

where  $\bar{L}_\delta$  is the Legendre transform of  $\bar{H}_\delta$ . However,  $Q_n/n = S_n/n + R_n/n$ , and by the large deviation upper bound for  $R_n/n$  (which has rate function  $\frac{1}{2\delta} |\beta|^2$ ),

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log P \{|R_n/n| \geq \varepsilon/2\} \leq -\frac{1}{8\delta} \varepsilon^2.$$

Since

$$P \{S_n/n \in (\beta - \varepsilon, \beta + \varepsilon)\} \geq P \{Q_n/n \in (\beta - \varepsilon/2, \beta + \varepsilon/2)\} - P \{|R_n/n| \geq \varepsilon/2\},$$

taking  $\delta > 0$  so that  $\varepsilon^2/8\delta \geq L(\beta) \geq \bar{L}_\delta(\beta)$  gives

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log P \{S_n/n \in (\beta - \varepsilon, \beta + \varepsilon)\} \geq -\bar{L}_\delta(\beta) \geq -L(\beta),$$

and therefore the proof is complete. ■

### 5.3 Sanov's Theorem

In this section we will use the theory of weak convergence to prove Sanov's Theorem, which is one of the fundamental results in large deviations theory. Cramér's Theorem will then be more-or-less obtained as a corollary. To keep the arguments brief we make assumptions that are stronger than necessary.

### 5.3.1 Relative Entropy

A key ingredient in the statement of Sanov's Theorem (and indeed, in all of large deviations theory) is the famous *relative entropy function*. Let  $S$  be a Polish space, and consider probability measures  $\mu$  and  $\nu$  on  $S$ . We define the relative entropy of  $\mu$  given  $\nu$  by

$$R(\mu \parallel \nu) = \begin{cases} \int_S \log \left( \frac{d\mu}{d\nu} \right) d\mu & \text{if } \mu \ll \nu \\ \infty & \text{else.} \end{cases}$$

In defining this integral the convention  $0 \log 0 = 0$  is used. Relative entropy plays a central role in information theory, statistical mechanics, and other disciplines, and is a well studied quantity. We will use the following properties of relative entropy. Item 2 was assigned as a homework problem, and you were allowed to assume the *Donsker-Varadhan formula for relative entropy*: for any pair  $\mu, \nu \in \mathcal{P}(S)$ ,

$$R(\mu \parallel \nu) = \sup_{g \in C_b(S)} \left[ \int_S g d\mu - \log \int_S e^g d\nu \right].$$

This latter result is the only hard one in the bunch, and you may find a proof in Section C.2 of my book with R.S. Ellis. Item 4 is a straightforward calculation and omitted. Proofs for items 1 and 3 are given after the statements.

**Lemma 1.** (PROPERTIES OF RELATIVE ENTROPY)

1.  $R(\mu \parallel \nu) \geq 0$ , and  $R(\mu \parallel \nu) = 0$  if and only if  $\mu = \nu$ .
2.  $R(\mu \parallel \nu)$  is a convex function and lower semicontinuous function of  $(\mu, \nu) \in \mathcal{P}(S)^2$ .
3. We have the following representation for exponential integrals. For any bounded and measurable  $f : S \rightarrow \mathbb{R}$ ,

$$-\log \int_S e^{-f} d\nu = \inf_{\mu \in \mathcal{P}(S)} \left[ \int_S f d\mu + R(\mu \parallel \nu) \right].$$

4. Suppose that  $S$  is of product form,  $S = S_1 \times S_2$ , where both  $S_1$  and  $S_2$  are Polish. Then we have the following chain rule for relative entropy. If  $(\mu, \nu) \in \mathcal{P}(S)^2$ , and if each distribution is factored into its marginal distribution on  $S_1$  times a conditional distribution on  $S_2$  given  $S_1$ :

$$\mu(dx_1 \times dx_2) = \mu_1(dx_1) \mu_{1,2}(dx_2 | x_1),$$

$$\nu(dx_1 \times dx_2) = \nu_1(dx_1)\nu_{1,2}(dx_2 | x_1),$$

then

$$R(\mu \| \nu) = R(\mu_1 \| \nu_1) + \int_{S_1} R(\mu_{1,2}(\cdot | x_1) \| \nu_{1,2}(\cdot | x_1)) \mu_1(dx_1).$$

**Proofs of Items 1 and 3.** In proving Item 1 we can assume  $R(\mu \| \nu) < \infty$ . In this case  $d\mu/d\nu$  is well defined. We use that  $s \log s \geq s - 1$  with equality if and only if  $s = 1$ . Thus

$$R(\mu \| \nu) = \int_S \frac{d\mu}{d\nu} \left( \log \frac{d\mu}{d\nu} \right) d\nu \geq \int_S \left( \frac{d\mu}{d\nu} - 1 \right) d\nu = 0,$$

and equality holds only when  $d\mu/d\nu = 1$ , which requires  $\mu = \nu$ .

With regard to item 3, it suffices to prove that

$$-\log \int_S e^{-f} d\nu = \inf \left[ \int_S f d\mu + R(\mu \| \nu) : R(\mu \| \nu) < \infty \right].$$

One can formally guess by a Lagrange multiplier argument that the minimizer should be given by

$$\frac{d\mu^*}{d\nu}(x) = e^{-f(x)} \cdot \frac{1}{\int_S e^{-f} d\nu}.$$

Since under  $R(\mu \| \nu) < \infty$   $\mu$  is absolutely continuous with respect to  $\nu$ , and since  $\nu$  is absolutely continuous with respect to  $\mu^*$ , it follows that  $\mu$  is absolutely continuous with respect to  $\mu^*$ . Using the definition of relative entropy twice, we write

$$\begin{aligned} \int_S f d\mu + R(\mu \| \nu) &= \int_S f d\mu + \int_S \log \left( \frac{d\mu}{d\nu} \right) d\mu \\ &= \int_S f d\mu + \int_S \log \left( \frac{d\mu}{d\mu^*} \right) d\mu + \int_S \log \left( \frac{d\mu^*}{d\nu} \right) d\mu \\ &= -\log \int_S e^{-f} d\nu + R(\mu \| \mu^*). \end{aligned}$$

Now use that  $R(\mu \| \mu^*) \geq 0$  with equality only when  $\mu = \mu^*$ . This not only proves the formula, but incidentally identifies the minimizer. ■

Let us give a random variable interpretation of items 3 and 4. Suppose that the random variables  $(X_1, X_2)$  have joint distribution  $\nu$  and  $(\bar{X}_1, \bar{X}_2)$

have distribution  $\mu$ , on some probability space  $(\Omega, \mathcal{F}, P)$  (one can certainly construct such a probability space, which will implicitly depend on  $\mu$ ). Further suppose that the  $\nu$  measure corresponds to *independent* random variables, so that  $\nu_{1,2}(\cdot|x_1) = \nu_2(\cdot)$  for some probability measure  $\nu_2$ . If  $f : S_1 \times S_2 \rightarrow \mathbb{R}$  is bounded and measurable, then

$$-\log Ee^{-f(X_1, X_2)} = \inf_{\mu \in \mathcal{P}(S)} E \left[ f(\bar{X}_1, \bar{X}_2) + \sum_{i=1}^2 R(\bar{\mu}_i \|\nu_i) \right],$$

where  $\bar{\mu}_1(\cdot) = \mu_1(\cdot)$  and  $\bar{\mu}_2(\cdot) = \mu_{1,2}(\cdot|\bar{X}_1)$ . Note that  $\bar{\mu}_2$  is a *random* measure.

By induction one can extend to any finite collection of independent random variables. Let  $\{X_i, i \in \mathbb{N}\}$  be iid  $S$ -valued random variables with distribution  $\nu$ . Let  $n \in \mathbb{N}$ . If  $f : S^n \rightarrow \mathbb{R}$  is bounded and measurable, then

$$-\log Ee^{-f(X_1, \dots, X_n)} = \inf E \left[ f(\bar{X}_1, \dots, \bar{X}_n) + \sum_{i=1}^n R(\bar{\mu}_i^n \|\nu) \right],$$

where the infimum is over all collections of random probability measures  $\{\bar{\mu}_i^n, i \in \{1, \dots, n\}\}$  that satisfy

1.  $\bar{\mu}_i^n$  is measurable with respect to the  $\sigma$ -algebra generated by  $\bar{X}_1, \dots, \bar{X}_{i-1}$ , and
2. the conditional distribution of  $\bar{X}_i$ , given  $\bar{X}_1, \dots, \bar{X}_{i-1}$ , is  $\bar{\mu}_i^n$ .

### 5.3.2 Sanov's Theorem

We can now state and give a weak convergence proof of (a special case of) Sanov's Theorem. To make the proof short, we assume that  $S$  is compact. This assumption is not needed, and the extra steps needed to prove tightness when this condition is absent are not particularly difficult.

**Theorem 1.** (SANOV'S THEOREM) *Let  $\{X_i, i \in \mathbb{N}\}$  be iid  $S$ -valued random variables with distribution  $\nu$ . Assume that  $S$  is compact, and for  $n \in \mathbb{N}$  define the empirical probability measure*

$$L^n(dx) \doteq \frac{1}{n} \sum_{i=1}^n \delta_{X_i}(dx).$$

*Then  $\{L^n, n \in \mathbb{N}\}$  satisfies the LDP on  $\mathcal{P}(S)$  with rate function  $I(\mu) = R(\mu \|\nu)$ .*

The proof turns on the following extension of the Glivenko-Cantelli Lemma.

**Theorem 2.** Consider a sequence of random probability measures  $\{\bar{\mu}_i^n, i \in \{1, \dots, n\}\}$  and random variables  $\{\bar{X}_i, i \in \{1, \dots, n\}\}$  such that

1.  $\bar{\mu}_i^n$  is measurable with respect to the  $\sigma$ -algebra generated by  $\bar{X}_1, \dots, \bar{X}_{i-1}$ , and
2. the conditional distribution of  $\bar{X}_i$ , given  $\bar{X}_1, \dots, \bar{X}_{i-1}$ , is  $\bar{\mu}_i^n$ .

Let  $\bar{L}^n(dx)$  denote the empirical measure of  $\bar{X}_1, \dots, \bar{X}_n$ , and let  $\theta^n(dx) \doteq \sum_{i=1}^n \bar{\mu}_i^n(dx)/n$ . Then the random measures  $\{(\bar{L}^n, \theta^n), n \in \mathbb{N}\}$  are tight, and if  $(\bar{L}, \theta)$  denotes the limit in distribution of any convergent subsequence, then  $\bar{L} = \theta$  a.s.

**Proof:** This is a nice application of martingales. Since  $\bar{L}^n$  and  $\theta^n$  are measures on a compact set, they are automatically tight. (By Prohorov's Theorem, if  $S$  is compact then so is  $\mathcal{P}(S)$ . Applying the argument again shows that  $\mathcal{P}(\mathcal{P}(S))$  is also compact.) Fix a subsequence along which there is convergence in distribution, and for convenience keep  $n$  as the label. Let  $(\bar{L}, \theta)$  denote the limit, and let  $g$  be any bounded and continuous function on  $S$ . Now for any  $i \in \{1, \dots, n\}$ ,

$$E \left[ g(\bar{X}_i) - \int_S g(x) \bar{\mu}_i^n(dx) \mid \bar{X}_1, \dots, \bar{X}_{i-1} \right] = 0.$$

Thus the sequence  $g(\bar{X}_i) - \int_S g(x) \bar{\mu}_i^n(dx)$  is a martingale difference with respect to  $\sigma(\bar{X}_1, \dots, \bar{X}_{i-1})$ . Hence

$$\sum_{i=1}^n \left[ g(\bar{X}_i) - \int_S g(x) \bar{\mu}_i^n(dx) \right]$$

is a martingale. A calculation we have done many times (see the proof of the WLLN) shows that the variance of this random variable is bounded above by  $4n \|g\|_\infty^2$ . Therefore the variance of

$$\int_S g d\bar{L}^n - \int_S g d\theta^n = \frac{1}{n} \sum_{i=1}^n \left[ g(\bar{X}_i) - \int_S g(x) \bar{\mu}_i^n(dx) \right]$$

is bounded above by  $4\|g\|_\infty^2/n$ . By the convergence in distribution version of Fatou's Lemma,

$$\begin{aligned} E \left| \int_S g d\bar{L} - \int_S g d\theta \right|^2 &\leq \liminf_{n \rightarrow \infty} E \left| \int_S g d\bar{L}^n - \int_S g d\theta^n \right|^2 \\ &\leq \liminf_{n \rightarrow \infty} 4\|g\|_\infty^2/n \\ &= 0. \end{aligned}$$

This shows that  $\int_S g d\bar{L} = \int_S g d\theta$  a.s. Does this imply that  $\bar{L} = \theta$  a.s.? This is a question we have raised several times: under what conditions do the integrals  $\int_S g d\mu$  for  $g$  in a given collection uniquely identify  $\mu$ ? Here we have a particular need: is there a *countable* collection of bounded continuous functions, such that  $\int_S g d\mu_1 = \int_S g d\mu_2$  for  $g$  in this collection implies  $\mu_1 = \mu_2$ ? If so, then the null sets will not pile up, and  $\int_S g d\bar{L} = \int_S g d\theta$  a.s. for all bounded continuous functions will indeed imply  $\bar{L} = \theta$  a.s. We already know (via the Inversion Formula for characteristic functions) that this is true when  $S = \mathbb{R}$ . Perhaps not surprisingly, such a countable collection in fact always exists when  $S$  is a Polish space. See, e.g., Lemma 3.1.4 in Stroock's book *Probability Theory, an Analytic View*. This observation completes the proof.  $\blacksquare$

**Proof of Sanov's Theorem:** Using the Laplace Principle formulation, it is enough to show that

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log E e^{-nf(L^n)} = \inf_{\mu \in \mathcal{P}(S)} [f(\mu) + R(\mu \|\nu)]$$

for any bounded and continuous function  $f$  on  $\mathcal{P}(S)$ . Using the representation

$$-\log E e^{-F(X_1, \dots, X_n)} = \inf E \left[ F(\bar{X}_1, \dots, \bar{X}_n) + \sum_{i=1}^n R(\bar{\mu}_i^n \|\nu) \right]$$

and the choice  $F(x_1, \dots, x_n) = nf(\sum_{i=1}^n \delta_{x_i}(dx)/n)$ , we get

$$-\frac{1}{n} \log E e^{-nf(L^n)} = \inf E \left[ f(\bar{L}^n) + \frac{1}{n} \sum_{i=1}^n R(\bar{\mu}_i^n \|\nu) \right].$$

As in the proof of Cramér's Theorem, we break the proof up into upper and lower bounds. For  $\varepsilon > 0$  let  $\mu^*$  satisfy

$$[f(\mu^*) + R(\mu^* \|\nu)] \leq \inf_{\mu \in \mathcal{P}(S)} [f(\mu) + R(\mu \|\nu)] + \varepsilon.$$

Then let  $\bar{\mu}_i^n = \mu^*$  for all  $n \in \mathbb{N}$  and  $i \in \{1, \dots, n\}$ , and apply the convergence theorem above. Consider any convergent subsequence, again labeled by  $n$ . It is automatic that  $\theta$  equals  $\mu^*$ . The fact that this particular choice of the  $\bar{\mu}_i^n$  is not necessarily infimizing gives the first inequality below, and the convergence in distribution version of LDCT gives the equality:

$$\begin{aligned} \limsup_{n \rightarrow \infty} -\frac{1}{n} \log E e^{-nf(L^n)} &\leq \limsup_{n \rightarrow \infty} E \left[ f(\bar{L}^n) + \frac{1}{n} \sum_{i=1}^n R(\bar{\mu}_i^n \|\nu) \right] \\ &= \left[ f(\mu^*) + \frac{1}{n} \sum_{i=1}^n R(\mu^* \|\nu) \right] \\ &\leq \inf_{\mu \in \mathcal{P}(S)} [f(\mu) + R(\mu \|\nu)] + \varepsilon. \end{aligned}$$

Since  $\varepsilon > 0$  and the subsequence are arbitrary, the upper bound follows.

Next, for each  $\varepsilon > 0$  let  $\{\bar{\mu}_i^n, i \in \{1, \dots, n\}\}$  and  $\{\bar{X}_i, i \in \{1, \dots, n\}\}$  satisfy

$$E \left[ f(\bar{L}^n) + \frac{1}{n} \sum_{i=1}^n R(\bar{\mu}_i^n \|\nu) \right] \leq -\frac{1}{n} \log E e^{-nf(L^n)} + \varepsilon.$$

Consider any convergent subsequence, again labeled by  $n$ . Using Jensen's inequality for the second inequality and the convergence in distribution,

$$\begin{aligned} \liminf_{n \rightarrow \infty} -\frac{1}{n} \log E e^{-nf(L^n)} + \varepsilon &\geq \liminf_{n \rightarrow \infty} E \left[ f(\bar{L}^n) + \frac{1}{n} \sum_{i=1}^n R(\bar{\mu}_i^n \|\nu) \right] \\ &\geq \liminf_{n \rightarrow \infty} E [f(\bar{L}^n) + R(\theta^n \|\nu)] \\ &= E [f(\bar{L}) + R(\bar{L} \|\nu)] \\ &\geq \inf_{\mu \in \mathcal{P}(S)} [f(\mu) + R(\mu \|\nu)], \end{aligned}$$

where the equality uses  $\bar{L} = \theta$  a.s. Since  $\varepsilon > 0$  and the subsequence are arbitrary, the lower bound follows. This completes the proof.  $\blacksquare$

We can now derive a version of Cramér's Theorem via the large deviation analogue of the Continuous Mapping Theorem. Although it is not as general in the sense that the compactness assumption on  $S$  requires the random variables  $X_i$  to be bounded, it is more general in that there is no need to restrict to one dimension. Since we assumed  $S$  was compact in Sanov's Theorem, we must assume here that the random variables  $\{X_n, n \in \mathbb{N}\}$  are

uniformly bounded. Suppose that  $|X_n| \leq B < \infty$  a.s. Note that the empirical average is just the integral of the empirical mean against the identity function. Now in general  $\mu \rightarrow \int_{\mathbb{R}} x\mu(dx)$  is not a continuous function (otherwise why would we need all these convergence theorems). However, when dealing with probability measures that are all supported on a compact set, the mapping  $\mu \rightarrow \int_{\mathbb{R}} x\mu(dx)$  is a continuous function with respect to weak convergence. Let  $L^n$  denote the empirical measure of  $\{X_i, i \in \{1, \dots, n\}\}$ . Then  $S_n/n = \int_{\mathbb{R}} xL^n(dx)$ , and the Contraction Principle implies the following.

**Theorem 3.** *Let  $\{X_n, n \in \mathbb{N}\}$  be a sequence of iid  $\mathbb{R}$ -valued random variables which satisfy  $|X_n| \leq B < \infty$  a.s. Let  $S_n = \sum_{i=1}^n X_i$  and let  $\nu$  denote the common distribution of the  $X_i$ . Then  $\{S_n/n, n \in \mathbb{N}\}$  satisfies the LDP with rate function*

$$I(\beta) = \inf \left\{ R(\mu \parallel \nu) : \int_{\mathbb{R}} x\mu(dx) = \beta \right\}.$$

This result identifies the rate function in terms of relative entropy, rather than as the Legendre transform of the log of the moment generating function. According to a homework problem two must be the same, and a direct proof of this fact is in my book with R.S. Ellis.