Chapter 31

Large Deviations for IID Sequences: The Return of Relative Entropy

Section 31.1 introduces the exponential version of the Markov inequality, which will be our major calculating device, and shows how it naturally leads to both the cumulant generating function and the Legendre transform, which we should suspect (correctly) of being the large deviations rate function. We also see the reappearance of relative entropy, as the Legendre transform of the cumulant generating *functional* of distributions.

Section 31.2 proves the large deviations principle for the empirical mean of IID sequences in finite-dimensional Euclidean spaces (Cramér's Theorem).

Section 31.3 proves the large deviations principle for the empirical distribution of IID sequences in Polish spaces (Sanov's Theorem), using Cramér's Theorem for a well-chosen collection of bounded continuous functions on the Polish space, and the tools of Section 30.2. Here the rate function is the relative entropy.

Section 31.4 proves that even the infinite-dimensional empirical process distribution of an IID sequence in a Polish space obeys the LDP, with the rate function given by the relative entropy rate.

The usual approach in large deviations theory is to establish an LDP for some comparatively tractable basic case through explicit calculations, and then use the machinery of Section 30.2 to extend it to LDPs for more complicated cases. This chapter applies this strategy to IID sequences.

31.1 Cumulant Generating Functions and Relative Entropy

Suppose the only inequality we knew in probability theory was Markov's inequality, $\mathbb{P}(X \geq a) \leq \mathbf{E}[X]/a$ when $X \geq 0$. How might we extract an exponential probability bound from it? Well, for any real-valued variable, e^{tX} is positive, so we can say that $\mathbb{P}(X \geq a) = \mathbb{P}\left(e^{tX} \geq e^{ta}\right) \leq \mathbf{E}\left[e^{tX}\right]/e^{ta}$. $\mathbf{E}\left[e^{tX}\right]$ is of course the moment generating function of X. It has the nice property that addition of independent random variables leads to multiplication of their moment generating functions, as $\mathbf{E}\left[e^{t(X_1+X_2)}\right] = \mathbf{E}\left[e^{tX_1}e^{tX_2}\right] = \mathbf{E}\left[e^{tX_1}\right]\mathbf{E}\left[e^{tX_2}\right]$ if $X_1 \perp\!\!\!\perp X_2$. If $X_1, X_2, \ldots$ are IID, then we can get a deviation bound for their sample mean $\overline{X}_n$ through the moment generating function:

$$\begin{aligned}
\mathbb{P}\left(\overline{X}_n \geq a\right) &= \mathbb{P}\left(\sum_{i=1}^{n} X_i \geq na\right) \\
\mathbb{P}\left(\overline{X}_n \geq a\right) &\leq e^{-nta}\left(\mathbf{E}\left[e^{tX_1}\right]\right)^n \\
\frac{1}{n}\log\mathbb{P}\left(\overline{X}_n \geq a\right) &\leq -ta + \log\mathbf{E}\left[e^{tX_1}\right] \\
&\leq \inf_t -ta + \log\mathbf{E}\left[e^{tX_1}\right] \\
&\leq -\sup_t ta - \log\mathbf{E}\left[e^{tX_1}\right]
\end{aligned}$$

This suggests that the functions $\log\mathbf{E}\left[e^{tX}\right]$ and $\sup ta - \log\mathbf{E}\left[e^{tX}\right]$ will be useful to us. Accordingly, we encapsulate them in a pair of definitions.

Definition 423 (Cumulant Generating Function) *The* cumulant generating function *of a random variable X in $\mathbb{R}^d$ is a function $\Lambda : \mathbb{R}^d \mapsto \mathbb{R}$,*

$$\Lambda(t) \equiv \log\mathbf{E}\left[e^{t\cdot X}\right] \tag{31.1}$$

Definition 424 (Legendre Transform) *The* Legendre transform *of a real-valued function f on $\mathbb{R}^d$ is another real-valued function on $\mathbb{R}^d$,*

$$f^*(x) \equiv \sup_{t\in\mathbb{R}^d} t\cdot x - f(t) \tag{31.2}$$

The definition of cumulant generating functions and their Legendre transforms can be extended to arbitrary spaces where some equivalent of the inner product (a real-valued form, bilinear in its two arguments) makes sense; f and f^* then must take arguments from the complementary spaces.

Legendre transforms are particularly important in convex analysis[1], since convexity is preserved by taking Legendre transforms. If f is not convex initially, then f^{**} is (in one dimension) something like the greatest convex lower bound on f; made precise, this statement even remains true in higher dimensions. I make these remarks because of the following fact:

[1] See Rockafellar (1970), or, more concisely, Ellis (1985, ch. VI).

Lemma 425 *The cumulant generating function $\Lambda(t)$ is convex.*

Proof: Simple calculation, using Hölder's inequality in one step:

$$\begin{aligned}
\Lambda(at+bu) &= \log \mathbf{E}\left[e^{(at+bu)X}\right] && (31.3)\\
&= \log \mathbf{E}\left[e^{atX}e^{buX}\right] && (31.4)\\
&= \log \mathbf{E}\left[\left(e^{tX}\right)^a\left(e^{uX}\right)^b\right] && (31.5)\\
&\leq \log \left(\mathbf{E}\left[e^{tX}\right]\right)^a\left(\mathbf{E}\left[e^{buX}\right]\right)^b && (31.6)\\
&= a\Lambda(t)+b\Lambda(u) && (31.7)
\end{aligned}$$

which proves convexity. □

Our previous result, then, is easily stated: if the X_i are IID in $\mathbb{R}$, then

$$\mathbb{P}\left(\overline{X}_n \geq a\right) \leq \Lambda^*(a) \quad (31.8)$$

where $\Lambda^*(a)$ is the Legendre transform of the cumulant generating function of the X_i. This elementary fact is, surprisingly enough, the foundation of the large deviations principle for empirical means.

The notion of cumulant generating functions can be extended to probability measures, and this will be useful when dealing with large deviations of empirical distributions. The definitions follow the pattern one would expect from the complementarity between probability measures and bounded continuous functions.

Definition 426 (Cumulant Generating Functional) *Let X be a random variable on a metric space Ξ, with distribution μ, and let $C_b(\Xi)$ be the class of all bounded, continuous, real-valued functions on Ξ. Then the* cumulant-generating functional $\Lambda : C_b(\Xi) \mapsto \mathbb{R}$ *is*

$$\Lambda(f) \equiv \log \mathbf{E}\left[e^{f(X)}\right] \quad (31.9)$$

Definition 427 *The* Legendre transform *of a real-valued functional F on $C_b(\Xi)$ is*

$$F^*(\nu) \equiv \sup_{f\in C_b(\Xi)} \mathbf{E}_\nu\left[f\right] - \Lambda(f) \quad (31.10)$$

where $\nu \in \mathcal{P}\left(\Xi\right)$, the set of all probability measures on Ξ.

Lemma 428 (Donsker and Varadhan) *The Legendre transform of the cumulant generating functional is the relative entropy:*

$$\Lambda^*(\nu) = D\left(\nu\|\mu\right) \quad (31.11)$$

PROOF: First of all, notice that the supremum in Eq. 31.10 can be taken over all bounded *measurable* functions, not just functions in C_b, since C_b is dense. This will let us use indicator functions and simple functions in the subsequent argument.

If $\nu \not\ll \mu$, then $D(\nu\|\mu) = \infty$. But then there is also a set, call it B, with $\mu(B) = 0$, $\nu(B) > 0$. Take $f_n = n\mathbf{1}_B$. Then $\mathbf{E}_\nu[f_n] - \Lambda(f_n) = n\nu(B) - 0$, which can be made arbitrarily large by taking n arbitrarily large, hence the supremum in Eq. 31.10 is ∞.

If $\nu \ll \mu$, then show that $D(\nu\|\mu) \leq \Lambda^*(\nu)$ and $D(\nu\|\mu) \geq \Lambda^*(\nu)$, so they must be equal. To get the first inequality, start with the observation then $\frac{d\nu}{d\mu}$ exists, so set $f = \log\frac{d\nu}{d\mu}$, which is measurable. Then $D(\nu\|\mu)$ is $\mathbf{E}_\nu[f] - \log\mathbf{E}_\mu\left[e^f\right]$. If f is bounded, this shows that $D(\nu\|\mu) \leq \Lambda^*(\nu)$. If f is not bounded, approximate it by a sequence of bounded, measurable functions f_n with $\mathbf{E}_\mu\left[e^{f_n}\right] \rightarrow 1$ and $\mathbf{E}_\nu[f_n] \rightarrow \mathbf{E}_\nu[f_n]$, again concluding that $D(\nu\|\mu) \leq \Lambda^*(\nu)$.

To go the other way, first consider the special case where $\mathcal{X}$ is finite, and so generated by a partition, with cells $B_1, \ldots B_n$. Then all measurable functions are simple functions, and $\mathbf{E}_\nu[f] - \Lambda(f)$ is

$$g(f) = \sum_{i=1}^{n} f_i\nu(B_i) - \log\sum_{i=1}^{n} e^{f_i}\mu(B_i) \tag{31.12}$$

Now, $g(f)$ is concave on all the f_i, and

$$\frac{\partial g(f)}{\partial f_i} = \nu(B_i) - \frac{1}{\sum_{i=1}^{n} e^{f_i}\mu(B_i)}\mu(B_i)e^{f_i} \tag{31.13}$$

Setting this equal to zero,

$$\begin{aligned} \frac{\nu(B_i)}{\mu(B_i)} &= \frac{1}{\sum_{i=1}^{n}\mu(B_i)e^{f_i}}e^{f_i} & (31.14)\\ \log\frac{\nu(B_i)}{\mu(B_i)} &= f_i & (31.15) \end{aligned}$$

gives the maximum value of $g(f)$. (Remember that $0\log 0 = 0$.) But then $g(f) = D(\nu\|\mu)$. So $\Lambda^*(\nu) \leq D(\nu\|\mu)$ when the σ-algebra is finite. In the general case, consider the case where f is a simple function. Then $\sigma(f)$ is finite, and $\mathbf{E}_\nu[f] - \log\mathbf{E}_\mu\left[e^f\right] \leq D(\nu\|\mu)$ follows by the finite case and smoothing. Finally, if f is not simple, but is bounded and measurable, there is a simple h such that $\mathbf{E}_\nu[f] - \log\mathbf{E}_\mu\left[e^f\right] \leq \mathbf{E}_\nu[h] - \log\mathbf{E}_\mu\left[e^h\right]$, so

$$\sup_{f\in C_b(\Xi)} \mathbf{E}_\nu[f] - \log\mathbf{E}_\mu\left[e^f\right] \leq D(\nu\|\mu) \tag{31.16}$$

which completes the proof. □

31.2 Large Deviations of the Empirical Mean in $\mathbb{R}^d$

Historically, the oldest and most important result in large deviations is that the empirical mean of an IID sequence of real-valued random variables obeys a large deviations principle with rate n; the oldest version of this proposition goes back to Harald Cramér in the 1930s, and so it is known as Cramér's theorem, even though the modern version, which is both more refined technically and works in arbitrary finite-dimensional Euclidean spaces, is due to Varadhan in the 1960s.

Theorem 429 (Cramér's Theorem) *If X_i are IID random variables in $\mathbb{R}^d$, and $\Lambda(t) < \infty$ for all $t \in \mathbb{R}^d$, then their empirical mean obeys an LDP with rate n and good rate function $\Lambda^*(x)$.*

Proof: The proof has three parts. First, the upper bound for closed sets; second, the lower bound for open sets, under an additional assumption on $\Lambda(t)$; third and finally, lifting of the assumption on Λ by means of a perturbation argument (related to Lemma 422).

To prove the upper bound for closed sets, we first prove the upper bound for sufficiently small balls around arbitrary points. Then, we take our favorite closed set, and divide it into a compact part close to the origin, which we can cover by a finite number of closed balls, and a remainder which is far from the origin and of low probability.

First the small balls of low probability. Because $\Lambda^*(x) = \sup_u u \cdot x - \Lambda(u)$, for any $\epsilon > 0$, we can find some u such that $u \cdot x - \Lambda(x) > \min 1/\epsilon, \Lambda^*(x) - \epsilon$. (Otherwise, $\Lambda^*(x)$ would not be the *least* upper bound.) Since $u \cdot x$ is continuous in x, it follows that there exists some open ball B of positive radius, centered on x, within which $u \cdot y - \Lambda(x) > \min 1/\epsilon, \Lambda^*(x) - \epsilon$, or $u \cdot y > \Lambda(x) + \min 1/\epsilon, \Lambda^*(x) - \epsilon$. Now use the exponential Markov inequality to get

$$\begin{aligned} \mathbb{P}\left(\overline{X}_n \in B\right) &\leq \mathbf{E}\left[e^{u \cdot n\overline{X}_n - n \inf_{y \in B} u \cdot y}\right] && (31.17) \\ &\leq e^{-n\left(\min \frac{1}{\epsilon}, \Lambda^*(x) - \epsilon\right)} && (31.18) \end{aligned}$$

which is small. To get the the compact set near the origin of high probability, use the exponential decay of the probability at large $\|x\|$. Since $\Lambda(t) < \infty$ for all t, $\Lambda^*(x) \to \infty$ as $\|x\| \to \infty$. So, using (once again) the exponential Markov inequality, for every $\epsilon > 0$, there must exist an $r > 0$ such that

$$\frac{1}{n} \log \mathbb{P}\left(\left\|\overline{X}_n\right\| > r\right) \leq -\frac{1}{\epsilon} \tag{31.19}$$

for all n.

Now pick your favorite closed measurable set $C \in \mathcal{B}^d$. Then $C \cap \{x : \|x\| \leq r\}$ is compact, and I can cover it by m balls $B_1, \ldots B_m$, with centers $x_1, \ldots x_m$, of the sort built in the previous paragraph. So I can apply a union bound to

$\mathbb{P}\left(\overline{X}_n \in C\right)$, as follows.

$$\begin{aligned}
& \mathbb{P}\left(\overline{X}_n \in C\right) & (31.20)\\
= \ & \mathbb{P}\left(\overline{X}_n \in C \cap \{x : \ \|x\| \leq r\}\right) + \mathbb{P}\left(\overline{X}_n \in C \cap \{x : \ \|x\| > r\}\right) & \\
\leq \ & \mathbb{P}\left(\overline{X}_n \in \bigcup_{i=1}^{m} B_i\right) + \mathbb{P}\left(\left\|\overline{X}_n\right\| > r\right) & (31.21)\\
\leq \ & \left(\sum_{i=1}^{m} \mathbb{P}\left(\overline{X}_n \in B_i\right)\right) + \mathbb{P}\left(\left\|\overline{X}_n\right\| > r\right) & (31.22)\\
\leq \ & \left(\sum_{i=1}^{m} e^{-n\left(\min \frac{1}{\epsilon}, \Lambda^*(x_i) - \epsilon\right)}\right) + e^{-n\frac{1}{\epsilon}} & (31.23)\\
\leq \ & (m+1)e^{-n\left(\min \frac{1}{\epsilon}, \Lambda^*(C) - \epsilon\right)} & (31.24)
\end{aligned}$$

with $\Lambda^*(C) = \inf_{x \in C} \Lambda^*(x)$, as usual. So if I take the log, normalize, and go to the limit, I have

$$\begin{aligned}
\limsup_{n} \frac{1}{n} \log \mathbb{P}\left(\overline{X}_n \in C\right) & \leq -\min \frac{1}{\epsilon}, \Lambda^*(C) - \epsilon & (31.25)\\
& \leq -\Lambda^*(C) & (31.26)
\end{aligned}$$

since ϵ was arbitrary to start with, and I've got the upper bound for closed sets.

To get the lower bound for open sets, pick your favorite open set $O \in \mathcal{B}^d$, and your favorite $x \in O$. Suppose, for the moment, that $\Lambda(t)/\|t\| \to \infty$ as $\|t\| \to \infty$. (This is the growth condition mentioned earlier, which we will left at the end of the proof.) Then, because $\Lambda(t)$ is smooth, there is some u such that $\nabla\Lambda(u) = x$. (You will find it instructive to draw the geometry here.) Now let Y_i be a sequence of IID random variables, whose probability law is given by

$$\mathbb{P}\left(Y_i \in B\right) = \frac{\mathbf{E}\left[e^{uX}\mathbf{1}_B(X)\right]}{\mathbf{E}\left[e^{uX}\right]} = e^{-\Lambda(u)}\mathbf{E}\left[e^{uX}\mathbf{1}_B(X)\right] \tag{31.27}$$

It is not hard to show, by manipulating the cumulant generating functions, that $\Lambda_Y(t) = \Lambda_X(t+u) - \Lambda_X(u)$, and consequently that $\mathbf{E}\left[Y_i\right] = x$. I construct these Y to allow me to pull the following trick, which works if $\epsilon > 0$ is sufficiently small that the first inequality holds (and I can always chose small enough ϵ):

$$\begin{aligned}
\mathbb{P}\left(\overline{X}_n \in O\right) & \geq \mathbb{P}\left(\left\|\overline{X}_n - x\right\| < \epsilon\right) & (31.28)\\
& = e^{n\Lambda(u)}\mathbf{E}\left[e^{-nu\overline{Y}_n}\mathbf{1}\{y : \ \|y - x\| < \epsilon\}(\overline{Y}_n)\right] & (31.29)\\
& \geq e^{n\Lambda(u) - nu\cdot x - n\epsilon\|u\|}\mathbb{P}\left(\left\|\overline{Y}_n - x\right\| < \epsilon\right) & (31.30)
\end{aligned}$$

By the strong law of large numbers, $\mathbb{P}\left(\left\|\overline{Y}_n - x\right\| < \epsilon\right) \rightarrow 1$ for all ϵ, so

$$\begin{aligned}
\liminf \frac{1}{n} \log \mathbb{P}\left(\overline{X}_n \in O\right) &\geq \Lambda(u) - u \cdot x - \epsilon\|u\| && (31.31)\\
&\geq -\Lambda^*(x) - \epsilon\|u\| && (31.32)\\
&\geq -\Lambda^*(x) && (31.33)\\
&\geq -\inf_{x \in O} \Lambda^*(x) = -\Lambda(O) && (31.34)
\end{aligned}$$

as required. This proves the LDP, as required, if $\Lambda(t)/\|t\| \rightarrow \infty$ as $\|t\| \rightarrow \infty$.

Finally, to lift the last-named restriction (which, remember, only affected the lower bound for open sets), introduce a sequence Z_i of IID standard Gaussian variables, i.e. $Z_i \sim \mathcal{N}(0, I)$, which are completely independent of the X_i. It is easily calculated that the cumulant generating function of the Z_i is $\|t\|^2/2$, so that $\overline{Z}_n$ satisfies the LDP. Another easy calculation shows that $X_i + \sigma Z_i$ has cumulant generating function $\Lambda_X(t) + \frac{\sigma^2}{2}\|t\|^2$, which again satisfies the previous condition. Since $\Lambda_{X+\sigma Z} \geq \Lambda_X$, $\Lambda^*_X \geq \Lambda^*_{X+\sigma Z}$. Now, once again pick any open set O, and any point $x \in O$, and an ϵ sufficiently small that all points within a distance 2ϵ of x are also in O. Since the LDP applies to $X + \sigma Z$,

$$\begin{aligned}
\mathbb{P}\left(\left\|\overline{X}_n + \sigma\overline{Z}_n - x\right\| \leq \epsilon\right) &\geq -\Lambda^*_{X+\sigma Z}(x) && (31.35)\\
&\geq -\Lambda^*_X(x) && (31.36)
\end{aligned}$$

On the other hand, basic probability manipulations give

$$\begin{aligned}
\mathbb{P}\left(\left\|\overline{X}_n + \sigma\overline{Z}_n - x\right\| \leq \epsilon\right) &\leq \mathbb{P}\left(\overline{X}_n \in O\right) + \mathbb{P}\left(\sigma\left\|\overline{Z}_n\right\| \geq \epsilon\right) && (31.37)\\
&\leq 2\max \mathbb{P}\left(\overline{X}_n \in O\right), \mathbb{P}\left(\sigma\left\|\overline{Z}_n\right\| \geq \epsilon\right) && (31.38)
\end{aligned}$$

Taking the liminf of the normalized log of both sides,

$$\begin{aligned}
&\liminf \frac{1}{n} \log \mathbb{P}\left(\left\|\overline{X}_n + \sigma\overline{Z}_n - x\right\| \leq \epsilon\right) && (31.39)\\
\leq\ & \liminf \frac{1}{n} \log\left(\max \mathbb{P}\left(\overline{X}_n \in O\right), \mathbb{P}\left(\sigma\left\|\overline{Z}_n\right\| \geq \epsilon\right)\right) &&\\
\leq\ & \liminf \frac{1}{n} \log \mathbb{P}\left(\overline{X}_n \in O\right) \vee \left(-\frac{\epsilon^2}{2\sigma^2}\right) && (31.40)\\
& && (31.41)
\end{aligned}$$

Since σ was arbitrary, we can let it go to zero, and obtain

$$\begin{aligned}
\liminf \frac{1}{n} \log \mathbb{P}\left(\overline{X}_n \in O\right) &\geq -\Lambda^*_X(x) && (31.42)\\
&\geq -\Lambda^*_X(O) && (31.43)
\end{aligned}$$

as required. □

31.3 Large Deviations of the Empirical Measure in Polish Spaces

The Polish space setting is, apparently, more general than $\mathbb{R}^d$, but we will represent distributions on the Polish space in terms of the expectation of a separating set of functions, and then appeal to the Euclidean result.

Proposition 430 *Any Polish space S can be represented as a Borel subset of a compact metric space, namely $[0,1]^{\mathbb{N}} \equiv M$.*

Proof: See, for instance, Appendix A of Kallenberg. □

Strictly speaking, there should be a function mapping points from S to points in M. However, since this is an embedding, I will silently omit it in what follows.

Proposition 431 *$C_b(M)$ has a countable dense separating set $\mathcal{F} = f_1, f_2, \ldots$.*

Proof: See Kallenberg again. □

Because $\mathcal{F}$ is separating, to specify a probability distribution on K is equivalent to specifying the expectation value of all the functions in $\mathcal{F}$. Write $f_1^d(X)$ to abbreviate the d-dimensional vector $(f_1(X), f_2(X), \ldots f_d(X))$, and $f_1^\infty(X)$ to abbreviate the corresponding infinite-dimensional vector.

Lemma 432 *Empirical means are expectations with respect to empirical measure. That is, let f be a real-valued measurable function and $Y_i = f(X_i)$. Then $\overline{Y}_n = \mathbf{E}_{\hat{P}_n}[f(X)]$.*

Proof: Direct calculation.

$$\overline{Y}_n \equiv \frac{1}{n}\sum_{i=1}^{n} f(X_i) \quad (31.44)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}_{\delta_{X_i}}[f(X)] \quad (31.45)$$

$$\equiv \mathbf{E}_{\hat{P}_n}[f(X)] \quad (31.46)$$

□

Lemma 433 *Let X_i be a sequence of IID random variables in a Polish space Ξ. For each d, the sequence of vectors $(\mathbf{E}_{\hat{P}_n}[f_1], \ldots \mathbf{E}_{\hat{P}_n}[f_d])$ obeys the LDP with rate n and good rate function J_d.*

Proof: For each d, the sequence of vectors $(f_1(X_i), \ldots f_d(X_i))$ are IID, so, by Cramér's Theorem (429), their empirical mean obeys the LDP with rate n and good rate function

$$J_d(x) = \sup_{t\in\mathbb{R}^d} t \cdot x - \log \mathbf{E}\left[e^{t\cdot f_1^d(X)}\right] \quad (31.47)$$

But, by Lemma 432, the empirical means are expectations over the empirical distributions, so the latter must also obey the LDP, with the same rate and rate function. □

Notice, incidentally, that the fact that the $f_i \in \mathcal{F}$ isn't relevant for the proof of the lemma; it will however be relevant for the proof of the theorem.

Theorem 434 (Sanov's Theorem) *Let X_i, $i \in \mathbb{N}$, be IID random variables in a Polish space Ξ, with common probability measure μ. Then the empirical distributions $\hat{P}_n$ obey an LDP with rate n and good rate function $J(\nu) = D(\nu\|\mu)$.*

PROOF: Combining Lemma 433 and Theorem 420, we see that $\mathbf{E}_{\hat{P}_n}\left[f_1^\infty(X)\right]$ obeys the LDP with rate n and good rate function

$$J(x) = \sup_d J_d(\pi_d x) \quad (31.48)$$

$$= \sup_d \sup_{t\in\mathbb{R}^d} t \cdot \pi_d x - \log \mathbf{E}\left[e^{t\cdot f_1^d(X)}\right] \quad (31.49)$$

Since $\mathcal{P}(M)$ is compact (so all random sequences in it are exponentially tight), and the mapping from $\nu \in \mathcal{P}(M)$ to $\mathbf{E}_\nu\left[f_1^\infty\right] \in \mathbb{R}^{\mathbb{N}}$ is continuous, apply the inverse contraction principle (Theorem 418) to get that $\hat{P}_n$ satisfies the LDP with good rate function

$$J(\nu) = J(\mathbf{E}_\nu\left[f_1^\infty\right]) \quad (31.50)$$

$$= \sup_d \sup_{t\in\mathbb{R}^d} t \cdot \mathbf{E}_\nu\left[f_1^d\right] - \log \mathbf{E}_\mu\left[e^{t\cdot f_1^d(X)}\right] \quad (31.51)$$

$$= \sup_{f\in \mathrm{span}\mathcal{F}} \mathbf{E}_\nu\left[f\right] - \Lambda(f) \quad (31.52)$$

$$= \sup_{f\in C_b(M)} \mathbf{E}_\nu\left[f\right] - \Lambda(f) \quad (31.53)$$

$$= D(\nu\|\mu) \quad (31.54)$$

Notice however that this is an LDP in the space $\mathcal{P}(M)$, not in $\mathcal{P}(\Xi)$. However, the embedding taking $\mathcal{P}(\Xi)$ to $\mathcal{P}(M)$ is continuous, and it is easily verified (see Lemma 27.17 in Kallenberg) that $\hat{P}_n$ is exponentially tight in $\mathcal{P}(\Xi)$, so another application of the inverse contraction principle says that $\hat{P}_n$ must obey the LDP in the restricted space $\mathcal{P}(\Xi)$, and with the same rate. □

31.4 Large Deviations of the Empirical Process in Polish Spaces

A fairly straightforward modification of the proof for Sanov's theorem establishes a large deviations principle for the finite-dimensional empirical distributions of an IID sequence.

Corollary 435 *Let X_i be an IID sequence in a Polish space Ξ, with common measure μ. Then, for every finite positive integer k, the k-dimensional empirical* *distribution $\hat{P}^k_n$, obeys an LDP with rate n and good rate function $J_k(\nu) = D\left(\nu \| \pi_{k-1}\nu \otimes \mu\right)$ if $\nu \in \mathcal{P}\left(\Xi^k\right)$ is shift invariant, and $J(\nu) = \infty$ otherwise.*

This leads to the following important generalization.

Theorem 436 *If X_i are IID in a Polish space, with a common measure μ, then the empirical process distribution $\hat{P}^\infty_n$ obeys an LDP with rate n and good rate function $J_\infty(\nu) = d(\nu\|\mu^\infty)$, the relative entropy rate, if ν is a shift-invariant probability measure, and $= \infty$ otherwise.*

Proof: By Corollary 435 and the projective limit theorem 420, $\hat{P}^\infty_n$ obeys an LDP with rate n and good rate function

$$J_\infty(\nu) = \sup_k J_k(\pi_k \nu) = \sup_k D\left(\pi_k\nu \| \pi_{k-1}\nu \otimes \mu\right) \tag{31.55}$$

But, applying the chain rule for relative entropy (Lemma 363),

$$\begin{aligned} D\left(\pi_n\nu\|\mu^n\right) &= D\left(\pi_n\nu\|\pi_{n-1}\nu\otimes\mu\right) + D\left(\pi_{n-1}\nu\|\mu^{n-1}\right) && (31.56)\\ &= \sum_{k=1}^{n} D\left(\pi_k\nu\|\pi_{k-1}\nu\otimes\mu\right) && (31.57)\\ \lim \frac{1}{n} D\left(\pi_n\nu\|\mu^n\right) &= \lim \frac{1}{n}\sum_{k=1}^{n} D\left(\pi_k\nu\|\pi_{k-1}\nu\otimes\mu\right) && (31.58)\\ &= \sup_k D\left(\pi_k\nu\|\pi_{k-1}\nu\otimes\mu\right) && (31.59)\end{aligned}$$

But $\lim n^{-1} D\left(\pi_n\nu\|\mu^n\right)$ is the relative entropy rate, $d(\nu\|\mu^\infty)$, and we've already identified the right-hand side as the rate function. □

The strength of Theorem 436 lies in the fact that, via the contraction principle (Theorem 410), it implies that the LDP holds for any continuous function of the empirical process distribution. This in particular includes the finite-dimensional distributions, the empirical mean, functions of finite-length trajectories, etc. Moreover, Theorem 410 also provides a means to calculate the rate function for all these quantities.