A **large deviation principle** (LDP) characterizes the limiting behavior as \( \epsilon \to 0 \) of a family of probability measures \( \mu_\epsilon \) on \((\mathcal{X},\mathcal{B})\) where \( \mathcal{X} \) is a topological space and \( \mathcal{B} \) is the Borel sigma algebra of the topology (smallest sigma algebra containing the open sets). We will suppose we are working in a metric space actually so that we can use sequences instead of nets.

**Definition 1.** A function \( f : \mathcal{X} \to \mathbb{R} \) is called **lower semicontinuous** if the level sets \( \Psi_f(\alpha) := \{ x : f(x) \leq \alpha \} \) are **closed** subsets of \( \mathcal{X} \).

**Proposition 2.** \( f \) is lower semicontinuous if and only if for all sequence \( x_n \to x_0 \) we have:

\[
\liminf_{n} f(x_n) \geq f(x_0)
\]

**Proof.** \((\Rightarrow)\) Suppose by contradiction that \( x_n \to x_0 \) but \( \liminf_n f(x_n) < f(x_0) \). Then \( \exists \epsilon_0 > 0 \) so that \( \liminf_n f(x_n) < f(x_0) - \epsilon_0 \).

For any \( \epsilon > 0 \) the set \( \{ x : f(x) > f(x_0) - \epsilon \} \) is an open set (by the def'n of lower semicontinuous) containing \( x_0 \). Hence since \( x_n \to x_0, x_n \in \{ x : f(x) > f(x_0) - \epsilon_0 \} \) eventually. But then \( \liminf_n f(x_n) \geq f(x_0) - \epsilon_0 \), a contradiction.

\((\Leftarrow)\) Fix an \( \alpha \) and suppose \( \{ x : f(x) \leq \alpha \} \) is non-empty. Given any convergent sequence \( y_n \in \{ x : f(x) \leq \alpha \}, y_n \to y \) we wish to show that \( y \in \{ x : f(x) \leq \alpha \} \). Suppose by contradiction \( y \notin \{ x : f(x) \leq \alpha \} \). Then there exists \( \epsilon_0 > 0 \) so that \( f(y) > \alpha + \epsilon_0 \). Now, \( y_n \to y \) and the given property gives that \( \liminf_n f(y_n) \geq f(y) > \alpha + \epsilon_0 \). But this is impossible as \( f(y_n) \leq \alpha \) for every \( n \) is given.

**Definition 3.** A **rate function** \( I \) is a lower semicontinuous mapping from \( I : \mathcal{X} \to [0,\infty] \). A **good** rate function is a rate function so that the level sets \( \{ x : f(x) \leq \alpha \} \) are **compact**. The effective domain of \( I \) is the set \( \mathcal{D}_I := \{ x : I(x) < \infty \} \). When there is no confusion, we call \( \mathcal{D}_I \) the domain of \( I \).

**Remark 4.** If a rate function \( I \) is good, the sets \( \{ x : f(x) \leq \alpha \} \) are compact. This means that \( \inf_{x} \{ x : f(x) \leq \alpha \} \) is achieved at some point \( x_0 \).

**Definition 5.** A collection of probability measure \( \mu_\epsilon \) are said to satisfy the large deviation principle with rate function \( I \) if for all \( A \in \mathcal{B} \) we have:

\[
- \inf_{x \in A^c} I(x) \leq \liminf_{\epsilon \to 0} \epsilon \log \mu_\epsilon(A) \leq \limsup_{\epsilon \to 0} \epsilon \log \mu_\epsilon(A) \leq - \inf_{x \in A} I(x)
\]

**Definition 6.** A set \( A \) that satisfies \( \inf_{x \in A^c} I(x) = \inf_{x \in \overline{A}} I(x) \) is called an \( I \)-continuity set.

**Remark 7.** What’s the deal with the \( A^c \) and \( \overline{A} \) in the above definition? Are they really needed?

Suppose we replaced \( A \) by \( \overline{A} \). Then take some non-atomic measures i.e. \( \mu_\epsilon(\{x\}) = 0 \forall x \in \mathcal{X} \). Then plugging in \( A = \{x\} \), the LDP would give \(-I(x_0) \leq \lim inf \epsilon \log(0) = -\infty \) so \( I(x_0) = \infty \). ... since this holds for any \( \{x_0\} \) we conclude \( I \) is a silly rate function!

The form of the LDP “codifies a particularly convenient way of stating asymptotic results that, on the one hand, are accurate enough to be useful and, on the other hand, are loose enough to be correct”.

**Remark 8.** Since \( \mu_\epsilon(\mathcal{X}) = 1 \), plugging in \( A = \mathcal{X} \) the whole space tells us that \( \inf_{x \in \mathcal{X}} I(x) = 0 \). When \( I \) is a good rate function, this means that there exists at least one point \( x \) for which \( I(x) = 0 \).

**Remark 9.** The following two conditions together are equivalent to the LDP (just unravel some definitions)

1. For every \( \alpha < \infty \) and every measurable set \( A \) with \( A \subset \{ x : I(x) > \alpha \} :=
   \limsup_{\epsilon \to 0} \epsilon \log \mu_\epsilon(A) \leq -\alpha
\]

2. For any \( x \in \mathcal{D}_I \) the effective domain of \( I \) and any measurable function \( A \) with \( x \in A^c \):
   \[
   \liminf_{\epsilon \to 0} \epsilon \log \mu_\epsilon(A) \geq -I(x)
   \]
1 LDP for Finite Dimensional Spaces

Let’s look at the simplest possible framework where we can have some large deviation results.

**Definition 10.** Table of definitions used in this section:

- \( \Sigma = \{a_1, \ldots, a_N\} \) : Underlying Alphabet with \( |\Sigma| = N \) elements
- \( M_1(\Sigma) \) : Probability measures on the alphabet \( \Sigma \)
  \[ M_1(\Sigma) = \{ \mu : \mathcal{P}(\Sigma) \to [0,1] : \mu \text{ is a probability measure} \} \]
- \( \Sigma_{\mu}; \mu \in M_1(\Sigma) \) : Elements of \( \Sigma \) that \( \mu \) takes with non-zero probability
  \[ \Sigma_{\mu} = \{ a_i : \mu(a_i) > 0 \} \subset \Sigma \]
- \( L_n^\Sigma; y = (y_1, \ldots, y_n) \in \Sigma^n \) : Empirical measure of a sequence \( y \), called the type of \( y \)
  \[ L_n^\Sigma(a_i) = \frac{1}{n} \sum_{j=1}^n 1_{a_i}(y_j) \]
- \( L_n^\Sigma(a_i) \) : Fraction of occurrences of \( a_i \) in \( y \)
- \( L_n^Y \) : Empirical measure of a random sequence \( Y \)
  \[ \text{with } Y = (Y_1, Y_2, \ldots, Y_n), Y_1 \overset{iid}{\sim} \mu \]
- \( \mathcal{L}_n \) : All possible types sequences of length \( n \)
  \[ \mathcal{L}_n = \{ \nu : \nu = L_n^Y \text{ for some } Y \} \]
- \( T_n(\nu); \nu \in \mathcal{L}_n \) : The type class of \( \nu \)
  \[ T_n(\nu) = \{ y \in \Sigma^n : L_n^Y = \nu \} \]

1.1 Basic Results and Sanov’s Theorem

**Lemma 11.** (Approximation lemma for sequence types)

\( a) |\mathcal{L}_n| \leq (n+1)^{|\Sigma|} \)

\( b)d_V(\nu, \mathcal{L}_n) := \inf_{\mu \in \mathcal{L}_n} d_V(\nu, \mu) \leq \frac{|\Sigma|}{2n}\)

Here \( d_V(\mu, \nu) \) is the total variation distance between two probability measures
\[ d_V(\mu, \nu) := \sup_{A \subset \Sigma} |\nu(A) - \mu(A)| \]

**Proof.** To prove a): For any \( \mu \in \mathcal{L}_n \), we know that \( \mu = L_n^Y \) for some \( Y \). Hence, each \( \mu(a_i) \in \{ 0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n} \} \), so there are at most \( n+1 \) choices for each “component” of \( \mu \). This means that there are at most \( (n+1)^{|\Sigma|} \) choices for such a measure \( \mu \) since there are \( |\Sigma| \)-components for the measure \( \mu \).

To prove b): Since \( \mathcal{L} \) contains all the probability vectors whose elements come from the set \( \{ 0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n} \} \), for any measure \( \nu \in M_1(\Sigma) \) we can always find a measure \( \mu \in \mathcal{L}_n \) so that \( |\mu(a_i) - \nu(a_i)| \leq \frac{1}{2n} \) for each \( a_i \). The result follows by the fact that:

\[ d_V(\mu, \nu) = \frac{1}{2} \sum_{i=1}^{|\Sigma|} |\nu(a_i) - \mu(a_i)| \]

(See the notes on Markov chains for details on this one...it’s actually pretty simple if you draw some pictures!)

\( \Box \)

**Remark 12.** The result in a) can be improved by using the fact that the components \( \mu(a_i) \) of \( \mu \) cannot all be chosen independently (since they must sum to 1). For example, the last component is determined once the first \( |\Sigma| - 1 \) components are fixed, so we could improve the inequality to \( |\mathcal{L}_n| \leq (n+1)^{|\Sigma|}-1 \).

**Remark 13.** From this lemma we know that the size of \( \mathcal{L}_n \) grows polynomial in \( n \) and that sets in \( \mathcal{L}_n \) approximate all the measures in \( M_1(\Sigma) \) uniformly and arbitrarily well as \( n \to \infty \). Both of these properties will fail to hold when \( |\Sigma| = \infty \).

**Definition 14.** The **entropy** of a probability vector \( \nu \in M_1(\Sigma) \) is:

\[ H(\nu) = -\sum_{i=1}^{|\Sigma|} \nu(a_i) \log(\nu(a_i)) \]

\[ = \sum_{i=1}^{|\Sigma|} \nu(a_i) \log \left( \frac{1}{\nu(a_i)} \right) \]
**Definition 15.** The **relative entropy** of a probability vector $\nu$ with respect to another probability vector $\mu$ is:

$$H(\nu|\mu) = \sum_{i=1}^{[\Sigma]} \nu(a_i) \log \left( \frac{\nu(a_i)}{\mu(a_i)} \right)$$

For the purposes of handling 0's in this formula, we will take the convention that $0 \log(0) = 0$ and $0 \log \left( \frac{0}{\mu} \right) = 0$.

**Remark 16.** By application of Jensen’s inequality to the convex function $x \log(x)$, it is possible to verify that $H(\nu|\mu) \geq 0$ with equality only when $\mu = \nu$. Also, notice that $H(\nu|\mu)$ is finite whenever $\Sigma_\nu \subset \Sigma_\mu$. Moreover, if we think of $H(\cdot|\mu) : M_1(\Sigma) \to \mathbb{R}$, we see that $H(\cdot|\mu)$ is a continuous function on the compact set $\{\nu : \Sigma_\nu \subset \Sigma_\mu\} \subset M_1(\Sigma)$ because $x \log(x)$ is continuous for $0 \leq x \leq 1$.

Also, $H(\cdot|\mu) = \infty$ for $\nu$ outside this set because we will have a $\log \left( \frac{\nu(a_i)}{\mu(a_i)} \right)$ when $\nu(a_i) \neq 0$ but $\mu(a_i) = 0$. These are some properties we expect of a good rate function!

We will now do some estimates relating the size of different sets (e.g. the number of sequences in a type class) to quantities involving these entropies.

**Lemma 17.** Choose a measure $\mu \in M_1(\Sigma)$ and consider a sequence $Y_1, \ldots, Y_N$ of i.i.d. $\mu$ random variables. (We will use $P_{\mu}$ when denoting probabilities involving these to remind us that the $Y$'s are $\mu$-distributed). Let $\nu \in \mathcal{L}_n$. If $y \in T_n(\nu)$ is in the type class of $\nu$, then:

$$P_{\mu}((Y_1, Y_2, \ldots, Y_n) = y) = \exp(-n[H(\nu) + H(\nu|\mu)])$$

**Proof.** If $\Sigma_\nu \not\subset \Sigma_\mu$, then there is an element $a_k$ with $\nu(a_k) > \frac{1}{2}$ that has $\mu(a_k) = 0$. In this case, there is at least one instance of $a_k$ on the sequence $y$ but there are NONE on the sequence $(Y_1, \ldots, Y_n)$ (almost surely). So the probability of the event we are interested in on the LHS is 0. The RHS is also 0 because $H(\nu|\mu) = \infty$ when $\Sigma_\nu \not\subset \Sigma_\mu$, so we have the desired equality.

Otherwise, we may assume that $\Sigma_\nu \subset \Sigma_\mu$ and that $H(\nu|\mu) < \infty$. Notice that:

$$P_{\mu}((Y_1, \ldots, Y_n) = y) = \prod_{k=1}^{n} P_{\mu}(Y_k = y_k) \text{ by independence}$$

$$= (P_{\mu}(Y = a_1))^{\#\{k: y_k = a_1\}} \cdot (P_{\mu}(Y = a_2))^{\#\{k: y_k = a_2\}} \ldots \text{ regroup terms}$$

$$= \prod_{i=1}^{[\Sigma]} \mu(a_i)^{\nu(a_i)} \text{ since } Y \sim \mu \text{ and } y \in T_n(\nu)$$

Hence:

$$\log (P_{\mu}(Y_1, \ldots, Y_n) = y) = n \sum_{i=1}^{[\Sigma]} \nu(a_i) \log(\mu(a_i))$$

$$= n \left( \sum_{i=1}^{[\Sigma]} \nu(a_i) \log(\nu(a_i)) - \nu(a_i) \log \left( \frac{\nu(a_i)}{\mu(a_i)} \right) \right)$$

$$= n (-H(\nu) - H(\nu|\mu))$$

$$= -n (H(\nu) + H(\nu|\mu))$$

Taking exp of both sides gives the desired result. \(\square\)

**Corollary 18.** $P_{\mu}((Y_1, \ldots, Y_n) = y) = \exp(-nH(\mu))$

**Proof.** Follows by the lemma since $H(\mu|\mu) = 0$. \(\square\)

**Lemma 19.** For every $\nu \in \mathcal{L}_n$:

$$(n + 1)^{-[\Sigma]} \exp(nH(\nu)) \leq |T_n(\nu)| \leq \exp(nH(\nu))$$

**Proof.** The upper bound is what you get when you take our last corollary and bound the probability by 1:

$$1 \geq P_{\nu}((Y_1, \ldots, Y_n) \in T_n(\nu))$$

$$= \sum_{y \in T_n(\nu)} P_{\nu}((Y_1, \ldots, Y_n) = y)$$

$$= |T_n(\nu)| \exp(-nH(\nu))$$

To prove the lower bound, we first aim to prove that for any measure $\mu \in \mathcal{L}_n$ that the empirical measure $L_n^Y$ of the sequence $(Y_1, \ldots, Y_n)$ has:

$$P_{\nu}(L_n^Y = \nu) \geq P_{\nu}(L_n^Y = \mu)$$
(Note: \(P_\nu(L_n^Y = \mu)\) is a bit like \(P_\nu((Y_1, \ldots, Y_n) = y)\) but not quite because the former does NOT care about the order of the \(Y_1, \ldots, Y_n\), while in the latter the order does matter. Hence there will be some factors counting the number of ways to rearrange the \(Y_i\)'s that will arise here. These factors are the additional difficulty we must overcome here.) When \(\Sigma_\mu \not\subseteq \Sigma_\nu\) the probability on the RHS is 0 and the result holds. Otherwise, consider:

\[
\frac{P_\nu(L_n^Y = \nu)}{P_\nu(L_n^Y = \mu)} = \frac{\sum_{y \in T_n(\nu)} P_\nu((Y_1, \ldots, Y_n) = y)}{\sum_{y \in T_n(\mu)} P_\nu((Y_1, \ldots, Y_n) = y)}
\]

\[
= \frac{|T_n(\nu)|}{|T_n(\mu)|} \prod_{i=1}^{[\Sigma]} \nu(a_i)^{n\nu(a_i)} \quad \text{by the previous lemma}
\]

\[
= \prod_{i=1}^{[\Sigma]} \left( \frac{m_i!}{(\nu(a_i))!} \right)^{n\nu(a_i)} \quad \text{by a counting argument}
\]

The last expression is a product of terms of the form \(\frac{m_i!}{(\nu(a_i))!} \nu(a_i)^{n\nu(a_i)}\) with \(\ell = n\nu(a_i)\) and \(m = n\mu(a_i)\). Considering the cases \(m \geq \ell\) and \(m < \ell\) separately, it is easily verified that \(\frac{m_i!}{(\nu(a_i))!} \nu(a_i)^{n\nu(a_i)} \geq \frac{\ell!}{(\mu(a_i))!} \mu(a_i)^{n\mu(a_i)}\) always holds. Hence \(\frac{m_i!}{(\nu(a_i))!} \nu(a_i)^{n\nu(a_i)} \geq \ell!^{n^{m-\ell}} \) so we have:

\[
\frac{P_\nu(L_n^Y = \nu)}{P_\nu(L_n^Y = \mu)} \geq \prod_{i=1}^{[\Sigma]} n^\mu(a_i)^{n\mu(a_i)} n^\nu(a_i)^{n\nu(a_i)} = n^{(\sum \mu(a_i) - \sum \nu(a_i))} = n^{n(1-1)} = 1
\]

Which proves the desired mini-result that \(P_\nu(L_n^Y = \nu) \geq P_\nu(L_n^Y = \mu)\). Finally, to get the lower bound on the inequality of the lemma, we have:

\[
1 = \sum_{\mu \in \mathcal{L}_n} P_\nu(L_n^Y = \mu)
\]

\[
\leq |\mathcal{L}_n| P_\nu(L_n^Y = \nu)
\]

\[
= |\mathcal{L}_n| |T_n(\nu)| \exp(-nH(\nu))
\]

So rearranging and using the bound \(|\mathcal{L}_n| \leq (n+1)^{[\Sigma]}\) gives the desired lower bound for \(|T_n(\nu)|\).

Lemma 20. For any \(\nu \in \mathcal{L}_n\) we have:

\((n+1)^{-[\Sigma]} \exp(-nH(\nu)|\mu)) \leq P_\mu(L_n^Y = \nu) \leq \exp(-nH(\nu)|\mu))\)

Proof. We have:

\[
P_\mu(L_n^Y = \nu) = \sum_{y \in T_n(\nu)} P ((Y_1, \ldots, Y_n) = y)
\]

\[
= |T_n(\nu)| \exp(-nH(\nu) + H(\nu)|\mu))
\]

So the bounds \((n+1)^{-[\Sigma]} \exp(nH(\nu)) \leq |T_n(\nu)| \leq \exp(nH(\nu))\) give exactly our desired result.

Theorem 21. (Sanov’s Theorem for Finite Alphabets) For every \(\Gamma \subseteq M_1(\Sigma)\) we have:

\[-\inf_{\nu \in \Gamma} H(\nu)|\mu) \leq \lim \inf_{n \to \infty} \frac{1}{n} \log P_\mu(L_n^Y \in \Sigma) \leq \lim \sup_{n \to \infty} \frac{1}{n} \log P_\mu(L_n^Y \in \Sigma) \leq -\inf_{\nu \in \Gamma} H(\nu)|\mu)\]

Proof. By the lemmas, we have the upper bound:

\[
P_\mu(L_n^Y \in \Gamma) = \sum_{\nu \in \Gamma \cap \mathcal{L}_n} P_\mu(L_n^Y = \nu)
\]

\[
\leq \sum_{\nu \in \Gamma \cap \mathcal{L}_n} \exp(-nH(\nu)|\mu))
\]

\[
\leq \sum_{\nu \in \Gamma \cap \mathcal{L}_n} \exp(-n \inf_{\nu \in \Gamma \cap \mathcal{L}_n} H(\nu)|\mu))
\]

\[
\leq |\Gamma \cap \mathcal{L}_n| \exp(-n \inf_{\nu \in \Gamma \cap \mathcal{L}_n} H(\nu)|\mu))
\]

\[
\leq (n+1)^{|\Sigma|} \exp(-n \inf_{\nu \in \Gamma \cap \mathcal{L}_n} H(\nu)|\mu))
\]
The accompanying lower bound is:

$$
P_\mu (L^n \in \Gamma) = \sum_{\nu \in \Gamma \cap L_n} P_\mu (L^n = \nu)$$

$$\geq \sum_{\nu \in \Gamma \cap L_n} (n + 1)^{-|\Omega|} \exp (-nH(\nu|\mu))$$

$$\geq (n + 1)^{-|\Omega|} \exp \left(-n \inf_{\nu \in \Gamma \cap L_n} H(\nu|\mu)\right) + \sum_{\nu \in \Gamma \cap L_n} 0$$

Now, since \(\lim_{n \to \infty} \frac{1}{n} \log (n + 1)^{|\Omega|} = 0\), this term has no contribution to the final logarithmic limit. So we have, by taking log and then limsup of these two inequalities (recall: a limsup(-x) = - \liminf(x)):

$$\limsup_{n \to \infty} \frac{1}{n} \log P_\mu (L^n \in \Gamma) \leq - \liminf_{n \to \infty} \left\{ \inf_{\nu \in \Gamma \cap L_n} H(\nu|\mu) \right\}$$

$$\limsup_{n \to \infty} \frac{1}{n} \log P_\mu (L^n \in \Gamma) \geq - \liminf_{n \to \infty} \left\{ \inf_{\nu \in \Gamma \cap L_n} H(\nu|\mu) \right\} + 0$$

Since these are equal, it must be that \(\limsup_{n \to \infty} \frac{1}{n} \log P_\mu (L^n \in \Gamma) = - \liminf_{n \to \infty} \left\{ \inf_{\nu \in \Gamma \cap L_n} H(\nu|\mu) \right\}\). Similarly, by taking \(\liminf\) of our two inequalities we get we get:

$$\liminf_{n \to \infty} \frac{1}{n} \log P_\mu (L^n \in \Gamma) = - \limsup_{n \to \infty} \left\{ \inf_{\nu \in \Gamma \cap L_n} H(\nu|\mu) \right\}$$

It remains only to argue that \(- \limsup_{n \to \infty} \left\{ \inf_{\nu \in \Gamma \cap L_n} H(\nu|\mu) \right\} \geq - \inf_{\nu \in \Gamma} H(\nu|\mu)\) and \(- \liminf_{n \to \infty} \left\{ \inf_{\nu \in \Gamma \cap L_n} H(\nu|\mu) \right\} \leq - \inf_{\nu \in \Gamma} H(\nu|\mu)\). The latter is trivial because \(\Gamma \cap L_n \subseteq \Gamma\) for every \(n\), so \(\inf_{\nu \in \Gamma \cap L_n} H(\nu|\mu) \geq \inf_{\nu \in \Gamma} H(\nu|\mu)\) always holds. The former holds because of the fact we saw earlier that \(L_n\) approximates any set \(\Gamma\) uniformly well, in the sense of \(d_V\). To be precise, take any \(\nu \in \Gamma^\circ\) with \(\Sigma_{\nu} \subset \Sigma_{\mu}\) and let \(\delta\) so small so that \(\{\nu' : d_V(\nu, \nu') < \delta\} \subset \Gamma\). Take \(N\) so large now so that \(|\Sigma|/2n < \delta\) for \(n > N\) and then we will be able to find \(\nu_n \in L_n\), so that \(\nu_n \in \{\nu' : d_V(\nu, \nu') < \delta\} \subset \Gamma\). Doing this for every \(n > N\), we can obtain a sequence \(\nu_n \to \nu\) as \(n \to \infty\) in the \(d_V\) sense. For this sequence, we have (keeping in mind that \(H(\cdot|\mu)\) is continuous here):

$$- \limsup_{n \to \infty} \left\{ \inf_{\nu \in \Gamma \cap L_n} H(\nu|\mu) \right\} \geq - \lim_{n \to \infty} H(\nu_n|\mu) = - H(\nu|\mu)$$

The inequality also holds trivially for points \(\nu \in \Gamma^\circ\) with \(\Sigma_{\nu} \not\subseteq \Sigma_{\mu}\) because in this case the RHS is \(\infty\). Since this holds for every point \(\nu \in \Gamma^\circ\), taking \(\lim\) over the LHS gives us the desired result.

**Exercise 22.** Prove that for every open set \(\Gamma\) that

$$- \lim_{n \to \infty} \left\{ \inf_{\nu \in \Gamma \cap L_n} H(\nu|\mu) \right\} = \lim_{n \to \infty} \frac{1}{n} \log (P_\mu (L^n \in \Gamma)) = - \inf_{\nu \in \Gamma} H(\nu|\mu)$$

If \(\Gamma\) is an open set, then \(\Gamma^\circ = \Gamma\). Hence, looking at the statement of the theorem, we notice that the LHS and RHS are equal, so we must have equality everywhere (its an equality sandwich!). Hence \(\lim_{n \to \infty} \frac{1}{n} \log (P_\mu (L^n \in \Gamma)) = - \inf_{\nu \in \Gamma} H(\nu|\mu)\).

(Rmk: For a general large deviation principle, one side of the inequality is \(\Gamma\) and the other side is \(\Gamma^\circ\) so the fact that the limit exists in this case is something special that doesn't happen in general.)

**Exercise 23.** Prove that if \(\Gamma\) is a subset of \(\{\nu \in M_1(\Sigma) : \Sigma_{\nu} \subset \Sigma_{\mu}\}\) and \(\Gamma \subset \overline{\Gamma^\circ}\), then the same conclusion as the previous exercise holds. Moreover, show that there is a unique \(\nu^* \in \overline{\Gamma^\circ}\) that achieves the minimum of \(H(\nu|\mu)\) in this case.

Suppose \(\Gamma \subset \{\nu \in M_1(\Sigma) : \Sigma_{\nu} \subset \Sigma_{\mu}\}\) and \(\Gamma \subset \overline{\Gamma^\circ}\). Since \(H(\cdot|\mu)\) is continuous on the set \(\{\nu \in M_1(\Sigma) : \Sigma_{\nu} \subset \Sigma_{\mu}\}\), we have that \(\inf_{\nu \in \Gamma} H(\nu|\mu) \geq \inf_{\nu \in \Gamma^\circ} H(\nu|\mu) = \inf_{\nu \in \Gamma^\circ} H(\nu|\mu)\) by continuity. This again completes the sandwich in this case, so we know \(\lim_{n \to \infty} \frac{1}{n} \log (P_\mu (L^n \in \Gamma)) = - \inf_{\nu \in \Gamma} H(\nu|\mu)\). Moreover in this case, since \(\Gamma^\circ\) is compact here, there must be a minimizing element \(\nu^* \in \Gamma^\circ\) that minimizes \(H(\nu^*|\mu) = \inf_{\nu \in \Gamma^\circ} H(\nu|\mu)\).

**Exercise 24.** Assume that \(\Sigma_{\mu} = \Sigma\) and that \(\Gamma\) is a convex subset of \(M_1(\Sigma)\) of non-empty interior. Prove that all of the conclusion of the previous exercise apply. Moreover, prove that the point \(\nu^*\) is unique.

First we claim that in this case that \(\Gamma \subset \overline{\Gamma^\circ}\) so we can apply the last exercise. If this is not the case then there is a \(\nu \in \Gamma\) with \(\nu \not\in \overline{\Gamma^\circ}\). Take any \(\nu_0 \in \Gamma^\circ\) and consider the line \(\nu_t = tv + (1-t)\nu_0\). It is sufficient to show that \(\nu_t \in \Gamma^\circ\) for \(t > 0\) for then we will have shown that \(\nu\) is a limit point of \(\nu\) and have our contradiction. To see that \(\nu_t \in \Gamma^\circ\), find a \(\epsilon_0\) so that for all \(\epsilon < \epsilon_0\) we have \(\nu_t \in \Gamma^\circ\) for every \(\nu_t \in B(\nu_0, \epsilon)\). For every such \(\nu_t\), the line segment \(\nu_t' = tv + (1-t)\nu_0\) is contained in \(\Gamma\) by convexity. Finally we notice that the collection of points \(\{\nu_t' : 0 < t < 1, 0 < \epsilon < \epsilon_0\}\) contains a small \(\epsilon_0\)’d of radius \(\epsilon_0\) around the point \(\nu_t\). (Draw a picture to see the geometry here). Hence each \(\nu_t \in \Gamma^\circ\) as desired.

To see that the limit point \(\nu^*\) is unique, it suffices to prove that \(H(\cdot|\mu)\) is strictly convex for if \(H(\cdot|\mu)\) is strictly convex, then taking a convex combination of different points \(\nu_1^*\) and \(\nu^*_2\) that minimize \(H(\cdot|\mu)\) would yield an even more extreme minimum: \(H(t\nu^*_1 + (1-t)\nu^*_2|\mu) < tH(\nu_1^*|\mu) + (1-t)H(\nu_2^*|\mu)\), which contradicts that \(\nu_1, \nu_2\) are minimizing.
Lemma 25. $H(\cdot|\mu)$ is strictly convex.

Proof. We will actually prove a slightly stronger statement, that $H(t\nu_1 + (1-t)\nu_2|t\mu_1 + (1-t)\mu_2) \leq tH(\nu_1|\mu_1) + (1-t)H(\nu_2|\mu_2)$ with equality only when $\nu_1\mu_2 = \nu_2\mu_1$. Setting $\mu_1 = \mu_2 = \mu$ will prove the strict convexity of $H(\cdot|\mu)$. The result is a consequence of the following inequality. Let $x_1,\ldots,x_n$ and $y_1,\ldots,y_n$ be non-negative numbers. Then the following inequality holds with equality only if $x_i/y_i$ is constant:

$$\sum x_i \log \left( \frac{x_i}{y_i} \right) \geq \left( \sum x_i \right) \log \left( \frac{\sum y_i}{\sum y_i} \right)$$

This holds by using the fact that $f(z) = z \log(z)$ is strictly convex, so by Jensen’s inequality:

$$\sum y_i f\left( \frac{x_i}{y_i} \right) \geq f\left( \frac{\sum y_i \left( \frac{x_i}{y_i} \right)}{\sum y_i} \right) = f\left( \frac{\sum x_i}{\sum y_i} \right)$$

To prove the inequality for $H(\cdot|\cdot)$ now, consider the above inequality with $n = 2$, and $x_1 = t\nu_1(a_i), x_2 = (1-t)\nu_2(a_i)$, $y_1 = t\mu_1(a_i)$, and $y_2 = (1-t)\mu_2(a_i)$. Have:

$$t\nu_1(a_i) \log\left( \frac{t\nu_1(a_i)}{t\mu_1(a_i)} \right) + (1-t)\nu_2(a_i) \log\left( \frac{(1-t)\nu_2(a_i)}{(1-t)\mu_2(a_i)} \right) \geq (t\nu_1(a_i) + (1-t)\nu_2(a_i)) \log\left( \frac{t\nu_1(a_i)}{t\mu_1(a_i)} + (1-t)\mu_2(a_i) \right)$$

Summing over $a_i$ now will give exactly $H(t\nu_1 + (1-t)\nu_2|t\mu_1 + (1-t)\mu_2) \leq tH(\nu_1|\mu_1) + (1-t)H(\nu_2|\mu_2)$ as desired. $\Box$

1.2 Cramer’s Theorem for Finite Alphabets in $\mathbb{R}$

Definition 26. Fix a function $f : \Sigma \rightarrow \mathbb{R}$ and, in the language of the last section, where $Y_i \overset{iid}{=} \mu$ let $X_i = f(Y_i)$. Without loss of generality assume further that $\Sigma = \Sigma_\mu$ and that $f(a_1) < \ldots < f(a_{|\Sigma|})$. Let $\hat{S}_n = \frac{1}{n} \sum_{j=1}^{n} X_j$ be the average of the first $n$ random variables. Cramér’s theorem deals with the LDP associated with the real-valued random variables $\hat{S}_n$.

Remark 27. In the case here, the random variable $X_i$’s take values in the compact interval $K = [f(a_1),\ldots,f(a_{|\Sigma|})]$. Consequently, so does the normalized average $\hat{S}_n$. Moreover, by our definitions:

$$\hat{S}_n = \sum_{i=1}^{\Sigma} f(a_i) L_n^Y(a_i) = \langle f, L_n^Y \rangle \text{ where } f := (f(a_1),\ldots,f(a_{|\Sigma|}))$$

Hence given any $A \subset \mathbb{R}$, we can associate a set $\Gamma_A \subset M_1(\mathbb{R})$ by $\Gamma_A = \{ \nu : \langle f, \nu \rangle \in A \}$. By the above and this definition, we have:

$$\hat{S}_n \in A \iff L_n^Y \in \Gamma_A$$

This naturally gives rise the LDP for the sum $\hat{S}_n$.

Theorem 28. (Cramer’s Theorem for finite subsets of $\mathbb{R}$) For any set $A \subset \mathbb{R}$ we have:

$$- \inf_{\nu \in \Gamma_A} H(\nu|\mu) = - \inf_{x \in A^*} I(x) \leq \liminf_{n \to \infty} \frac{1}{n} \log \left( P_\mu(\hat{S}_n \in A) \right) \leq \limsup_{n \to \infty} \frac{1}{n} \log \left( P_\mu(\hat{S}_n \in A) \right) \leq - \inf_{x \in A} I(x) = - \inf_{\nu \in \Gamma_A} H(\nu|\mu)$$

With $I(x) = \inf_{\langle f, \nu \rangle = x} H(\nu|\mu)$. One can verify that $I(x)$ is continuous at $x \in K$ and it satisfies there:

$$I(x) = \sup_{\lambda \in \mathbb{R}} \left\{ \lambda x - \Lambda(\lambda) \right\}$$

where

$$\Lambda(\lambda) = \log \left( \sum_{i=1}^{\Sigma} \mu(a_i) e^{\lambda f(a_i)} \right)$$
Proof. When the set $A$ is open, so is the set $\Gamma_A$, and the bounds directly follow. To get the fact about $\Lambda$, consider as follows.

By Jensen’s inequality, for any measure $\nu$, we have:

$$\Lambda(\lambda) = \log \left( \sum_{i=1}^{\Sigma} \mu(a_i) e^{\lambda f(a_i)} \right)$$

$$= \log \left( \sum_{i=1}^{\Sigma} \nu(a_i) \left( \frac{\mu(a_i) e^{\lambda f(a_i)}}{\nu(a_i)} \right) \right)$$

$$\geq \sum_{i=1}^{\Sigma} \nu(a_i) \log \left( \frac{\mu(a_i) e^{\lambda f(a_i)}}{\nu(a_i)} \right)$$

$$= \sum_{i=1}^{\Sigma} \nu(a_i) \left( \log \left( \frac{\mu(a_i)}{\nu(a_i)} \right) + \lambda f(a_i) \right)$$

$$= \lambda (f, \nu) - H(\nu|\mu)$$

Since $\log$ is strictly concave, equality holds when all the terms are equal. We will denote the special measure $\nu$ which has $\nu\lambda(a_i) = \mu(a_i) \exp(\lambda f(a_i) - \Lambda(\lambda))$. Rearranging the above inequality to put $H(\nu|\mu)$ and then taking inf gives:

$$I(x) = \inf_{\nu: (f, \nu) = x} H(\nu|\mu) \geq \lambda x - \Lambda(\lambda) \forall x$$

with equality at $x = (f, \nu)$. I.e. $I((f, \nu)) = \lambda (f, \nu) - \Lambda(\lambda)$ for every $\lambda$.

Now, the function $\Lambda(\lambda)$ is differentiable with $\Lambda'(\lambda) = (f, \nu)$. Fix an $x_0$ now in $\{\Lambda'(\lambda) : \lambda \in \mathbb{R}\}$, say $x_0 = \Lambda'(\lambda_0) = (f, \nu_{\lambda_0})$. Consider the optimization problem $\sup_{\lambda \in \mathbb{R}} \{\lambda x_0 - \Lambda(\lambda)\}$. By taking derivatives, it is clear the maximum occurs at $\lambda = \lambda_0$ and the optimal value here is $\lambda_0 (f, \nu_{\lambda_0}) - \Lambda(\lambda_0)$. This is equal to $I(x_0)$ by our earlier remark! This argument establishes that $I(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \Lambda(\lambda)\}$ for every $x \in \{\Lambda'(\lambda) : \lambda \in \mathbb{R}\}$.

We will now show that $\{\Lambda'(\lambda) : \lambda \in \mathbb{R}\} \subset K^c = (f(a_1), f(a_|\Sigma|)]$. First we notice that $\Lambda'(\cdot)$ is strictly increasing (since $\Lambda(\cdot)$ is strictly convex by the inequality $\Lambda(\lambda) \geq \lambda (f, \nu) - H(\nu|\mu) = \lambda \Lambda'(\lambda) + H(\nu|\mu)$). Notice moreover from $\Lambda'(\lambda) = (f, \nu_{\lambda})$ that $f(a_1) \leq \Lambda'(\lambda) \leq f(a_\Sigma)$. Moreover, since $\nu_{\lambda}(a_i) \propto \exp(\lambda f(a_i))$ by taking $\lambda \to -\infty$ or to $\infty$ we can achieve $\Lambda'(\lambda) = (f, \nu_{\lambda}) \to f(a_1)$ and $\to f(a_\Sigma)$. By continuity therefore, $\Lambda'$ can achieve every value in the open interval $K^o = (f(a_1), f(a_\Sigma)]$.

To include the endpoints, set $x = f(a_1)$ and let $\nu^*(a_i) = 1$ so that $(f, \nu^*) = f(a_1) = x$. Have $H(\nu^*|\mu) = -\log(\mu(a_1))$. Have then the string of inequalities:

$$-\log(\mu(a_1)) = H(\nu^*|\mu) \geq I(x) \geq \sup_{\lambda} \{\lambda x - \Lambda(\lambda)\}$$

$$\geq \lim_{\lambda \to -\infty} (\lambda x - \Lambda(\lambda)) = -\log(\mu(a_1))$$

So we get an equality sandwich and conclude that for $x = f(a_1)$ that $I(x) = \sup_{\lambda} \{\lambda x - \Lambda(\lambda)\}$. The same argument works to show that the point $x = f(a_\Sigma)$ works too.

Putting together all these arguments we indeed recover the desired result for $x$ anywhere in the whole interval $x \in K = (f(a_1), f(a_\Sigma)]$. \qed

1.3 Gibbs Conditioning Principle

We know that $\hat{S}_n \to (f, \mu)$ as $n \to \infty$, and Cramer’s theorem quantifies the rate for us. Let us the following question now, given some set $A \in \mathbb{R}$ what does the event $\{\hat{S}_n \in A\}$ look like? When $A$ does not contain the mean $(f, \mu)$ this is a rare event whose probability is very small, so when conditioning on it we will get some interesting behavior. To be precise, we will look at the conditional law of the constituent random variables $Y_1, \ldots, Y_n$ that make up $\hat{S}_n$ when we condition on $\hat{S}_n \in A$:

$$\mu^*(a_i) = P_\mu \left( Y_i = a_i | \hat{S}_n \in A \right), \ i = 1, \ldots, |\Sigma|$$

By symmetry between switching indices $i \leftrightarrow j$, we expect that $Y_i$ and $Y_j$ will be identically distributed (although not independent) when we condition like this. For this reason, we restrict our attention to $Y_1$ as above. Notice that for any function $\phi : \Sigma \to \mathbb{R}$ that:

$$\langle \phi, \mu^* \rangle = E \left[ \phi(Y_1) | \hat{S}_n \in A \right]$$

$$= E \left[ \phi(Y_j) | \hat{S}_n \in A \right] \text{ for any } j = 1, \ldots, |\Sigma|$$

$$= E \left[ \sum_{j=1}^{n} \phi(Y_j) | \hat{S}_n \in A \right]$$

$$= E \left[ \langle \phi, L_n^X \rangle | \langle f, L_n^X \rangle \in A \right] \text{ by writing } \hat{S}_n = \langle f, L_n^X \rangle$$

7
If we let $\Gamma = \{ \nu : \langle f, \nu \rangle \in A \}$, then this can be written simply as:

$$\mu^*_n = E [ L^Y_n | L^Y_n \in \Gamma ]$$

We will address the question of what are the possible limits as $n \to \infty$ of this conditional law $\mu^*_n$. This is called the Gibbs Conditioning Principle.

**Theorem 29.** (Gibbs’s Conditioning Principle) Suppose that we have a set $\Gamma$ for which $I_{\Gamma} := \inf_{\nu \in \Gamma} H(\nu | \mu) = \inf_{\nu \in \Gamma} H(\nu | \mu)$. Define the set of measures that minimize the relative entropy:

$$\mathcal{M} := \{ \nu \in \bar{\Gamma} : H(\nu | \mu) = I_{\Gamma} \}$$

Then:

a) All the limit points of $\{ \mu^*_n \}$ belong to $\bar{\mathcal{M}}(\mathcal{M})$-the closure of the convex hull of $\mathcal{M}$.

b) When $\Gamma$ is a convex set of non-empty interior, the set $\mathcal{M}$ consists of a single point to which $\mu^*_n$ converges as $n \to \infty$.

**Remark 30.** a) The condition that $I_{\Gamma}$ exists or the condition that $\Gamma$ is a convex set might seem strange. However, by earlier exercises we know these conditions where $I_{\Gamma}$ exists. Also by an earlier exercise, the condition that $\Gamma$ is convex also shows there is a unique minimizing element of $H(\nu | \mu)$, i.e. $\mathcal{M} = \{ \nu^* \}$.

b) The result is kind of intuitive in the following sense. We know by Sanov’s theorem that $P$

**Proof.** Firstly, we notice that part a) $\implies$ part b), because when $\Gamma$ is a convex set of non-empty interior. We know from an earlier exercise that $\mathcal{M}$ consists of a single point in this case. By compactness, every subsequence of $\mu^*_n$ must have a sub-subsequence converge to something, and the only candidate is the single point in $\mathcal{M}$ by part a). Hence $\mu^*_n$ converges to the single point in $\mathcal{M}$ too. To prove part a), we break up the main ideas into two claims:

**Claim 1:** For any $U \subseteq M_1(\Sigma)$, we have that $d_V(\mu^*_n; \text{co}(U)) \leq P_\mu (L^Y_n \in U | L^Y_n \in \Gamma)$

**Proof:** For any $U$, we have:

$$E [ L^Y_n | L^Y_n \in \Gamma ] - E [ L^Y_n | L^Y_n \in U \cap \Gamma ] = E [ L^Y_n | L^Y_n \in \Gamma \cap U ] P [ L^Y_n \in U | L^Y_n \in \Gamma ] + E [ L^Y_n | L^Y_n \in \Gamma \cap U^c ] P [ L^Y_n \in U^c | L^Y_n \in \Gamma ] - E [ L^Y_n | L^Y_n \in U \cap \Gamma ]$$

$$= E [ L^Y_n | L^Y_n \in \Gamma \cap U ] \left( \frac{P(L^Y_n \in U \cap \Gamma)}{P(L^Y_n \in \Gamma)} - 1 \right) + E [ L^Y_n | L^Y_n \in \Gamma \cap U^c ] \left( \frac{P(L^Y_n \in U^c \cap \Gamma)}{P(L^Y_n \in \Gamma)} - 1 \right)$$

$$= E [ L^Y_n | L^Y_n \in \Gamma \cap U ] \left( \frac{P(L^Y_n \in U \cap \Gamma)}{P(L^Y_n \in \Gamma)} - 1 \right) + E [ L^Y_n | L^Y_n \in \Gamma \cap U^c ] \left( \frac{P(L^Y_n \in U^c \cap \Gamma)}{P(L^Y_n \in \Gamma)} - 1 \right)$$

Hence, since the $d_V(\alpha, \beta) = \frac{1}{2} \sum_{i=1}^{\Sigma} |\alpha_i - \beta_i|$ depends only on the difference between the measures, we can factor out

$$P [ L^Y_n \in U^c | L^Y_n \in \Gamma ]$$

from any formulas we have. Now use the measure $E [ L^Y_n | L^Y_n \in \Gamma \cap U^c ]$ from the convex hull to escape the set $U$, might escape the set $U$ but you will still be in co$(U)$ since averaging is like a convex combination) and $\mu^*_n = E [ L^Y_n | L^Y_n \in \Gamma ]$ we have:

$$d_V(\mu^*_n; \text{co}(U)) \leq d_V( E [ L^Y_n | L^Y_n \in \Gamma ] , E [ L^Y_n | L^Y_n \in \Gamma \cap U ] )$$

$$\leq P [ L^Y_n \in U^c | L^Y_n \in \Gamma ] d_V( E [ L^Y_n | L^Y_n \in \Gamma \cap U^c ] , E [ L^Y_n | L^Y_n \in \Gamma \cap U ] )$$

$$\leq P [ L^Y_n \in U^c | L^Y_n \in \Gamma ] . 1$$

Which proves the result.

**Claim 2:** Let $\mathcal{M}^3 = \{ \nu : d_V(\nu, \mathcal{M}) < \delta \}$. Then for every $\delta > 0$:

$$\lim_{n \to \infty} P_\mu ( L^Y_n \in \mathcal{M}^3 | L^Y_n \in \Gamma ) = 1$$

**Proof:** By Sanov’s theorem in this case we know that $\inf_{\nu \in \Gamma} H(\nu | \mu) = I_{\Gamma} = - \lim_{n \to \infty} \frac{1}{n} \log \left( P_\mu ( L^Y_n \in \Gamma ) \right)$. We also have:

$$\lim_{n \to \infty} \inf_{\nu \in \mathcal{M}^3 \cap \Gamma} H(\nu | \mu) \leq H(\nu | \mu) \leq - \inf_{\nu \in \mathcal{M}^3 \cap \Gamma} H(\nu | \mu)$$
Since $\mathcal{M}^δ$ is an open set, we know that $(\mathcal{M}^δ)^c \cap \Gamma$ is a compact set, so this inf is achieved at some point $\tilde{\nu} \in (\mathcal{M}^δ)^c \cap \Gamma$. Since $\tilde{\nu} \notin \mathcal{M}$, we know that $H(\tilde{\nu}|\mu) > I_{\Gamma}$. So finally then, we put all this together to conclude that $P_\mu(L_n^Y \in (\mathcal{M}^δ)^c | L_n^Y \in \Gamma)$ goes to zero exponentially fast:

$$
\limsup_{n \to \infty} \frac{1}{n} \log P_\mu(L_n^Y \in (\mathcal{M}^δ)^c | L_n^Y \in \Gamma) = \limsup_{n \to \infty} \left\{ \frac{1}{n} \log \left( P_\mu(L_n^Y \in (\mathcal{M}^δ)^c \cap \Gamma) \right) - \frac{1}{n} \log \left( P_\mu(L_n^Y \in \Gamma) \right) \right\} \\
\leq -H(\tilde{\nu}|\mu) + I_{\Gamma} \\
< 0
$$

Which proves $\lim_{n \to \infty} P_\mu(L_n^Y \in (\mathcal{M}^δ)^c | L_n^Y \in \Gamma) = 0$. Taking complements gives the desired result.

Finally, applying the two claims together, we know that for any $\delta > 0$ that $d_V(\mu^*_n, \text{co}(\mathcal{M}^δ)) \leq P_\mu(L_n^Y \in (\mathcal{M}^δ)^c | L_n^Y \in \Gamma) \to 0$ as $n \to \infty$. Hence each limit point of $\mu^*_n$ must be in $\mathcal{M}^δ$. Since $\delta$ is arbitrary, it must be that all the limit points are in $\text{co}(\mathcal{M})$, as desired. $\square$