Large Deviations and Applications
Timi Turpeinen – 100740833

1 Preface

In this paper, we address several exercises related to large deviations and their applications. Specifically, we will solve problems from the book "Large Deviations Techniques and Applications" by Amir Dembo and Ofer Zeitouni [1]. We adopt the same notation as the book and all theorems and lemmas we cite are from the given book, unless they are standard results from the probability theory. We assume the reader is familiar with these standard theorems.

In section 2, we first examine the Large Deviation Principle (LDP) for finite-dimensional spaces, and then extend it to the specific cases of $\mathbb{R}$ and $\mathbb{R}^d$. While the focus is primarily on i.i.d. samples, an extension to non-i.i.d. scenarios is explored in section 2.3. Finally, section 2.4 delves into concentration inequalities from an LDP perspective. In section 3 we study the applications of the LDP in finite-dimensional settings, e.g., for Markov processes.

2 LDP for finite Dimensional Spaces

2.1 Combinatorial Techniques for Finite Alphabets

Throughout this section exercises, all random variables assume values in a finite alphabet $\Sigma = \{a_1, \dots, a_n\}$ and $|\Sigma| = N$. In these exercises, $\mathbb{P}_\mu = \mu_n$ stands for family of probability measures generated by the probability measure $\mu \in M_1(\Sigma)$. Here, $M_1(\Sigma) \subset \mathbb{R}^{|\Sigma|}$ denotes the space of all probability measures on the finite alphabet Σ. In addition, let Σ_μ denote the support of the law μ, i.e., $\Sigma_\mu := \{a_i : \mu(a_i) > 0\}$, and some exercises (e.g. in section 2.1.2) it may be assumed without loss of generality $\Sigma = \Sigma_\mu$, when considering the single measure μ.

2.1.1 The method of Types of Sanov's theorem

In this section, let $Y_1, \dots, Y_n$ be a sequence of random variables that are independent and identically distributed, and let L_n^Y be the empirical measure associated with random sequence $Y = (Y_1, \dots, Y_n)$ which is a random element of the set $\mathcal{L}_n := \{\nu : \nu = L_n^y \text{ for some } y \in \Sigma\} \subset \mathbb{R}^{|\Sigma|}$. Also, let $H(\cdot \mid \cdot)$ denote the relative entropy defined as in definition 2.1.5 [1].

Ex. 2.1.16

Let us prove that for every open set Γ, we have

$$-\lim_{n\to\infty}\{\inf_{\nu\in\Gamma\cap\mathcal{L}_n} H(\nu \mid \mu)\} = \lim_{n\to\infty}\frac{1}{n}\log \mathbb{P}_\mu(L_n^Y \in \Gamma) = -\inf_{\nu\in\Gamma} H(\nu \mid \mu) = -I_\Gamma. \tag{1}$$

We shall proof (1) by showing the corresponding equalities one by one.

Proof.

First, we let us show that from (1), we have

$$\lim_{n\to\infty}\frac{1}{n}\log \mathbb{P}_\mu(L_n^Y \in \Gamma) = -\inf_{\nu\in\Gamma} H(\nu \mid \mu). \tag{2}$$

By using Sanov's theorem (theorem 2.1.10 [1]), we know that for every set Γ of probability vectors in $M_1(\Sigma)$ satisfies

$$-\inf_{\nu\in\Gamma^\circ} H(\nu \mid \mu) \leq \liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}_\mu(L_n^Y \in \Gamma) \leq \limsup_{n\to\infty} \frac{1}{n}\log \mathbb{P}_\mu(L_n^Y \in \Gamma) \leq -\inf_{\nu\in\Gamma} H(\nu \mid \mu).$$

Now, since we assumed that Γ is open set, we have that $\Gamma^\circ = \Gamma$, and since we know that $\liminf \leq \lim \leq \limsup$, we get immeaditly from the Sanov's theorem that equality (2) holds. Hence, it suffices only to show that equality

$$-\lim_{n\to\infty}\{\inf_{\nu\in\Gamma\cap\mathcal{L}_n} H(\nu \mid \mu)\} = -\inf_{\nu\in\Gamma} H(\nu \mid \mu) \tag{3}$$

from (1) is true as well.

As in proof of Sanov's theorem [1], we know that in this case also the equations (2.1.14) and (2.1.15) holds as well, that is,

$$\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}_\mu(L_n^Y \in \Gamma) = -\limsup_{n\to\infty}\{\inf_{\nu\in\Gamma\cap\mathcal{L}_n} H(\nu \mid \mu)\}$$

and

$$\limsup_{n\to\infty} \frac{1}{n}\log \mathbb{P}_\mu(L_n^Y \in \Gamma) = -\liminf_{n\to\infty}\{\inf_{\nu\in\Gamma\cap\mathcal{L}_n} H(\nu \mid \mu)\}$$

which together with Sanov's theorem yields inequality

$$-\inf_{\nu\in\Gamma^\circ} H(\nu \mid \mu) \leq -\limsup_{n\to\infty}\{\inf_{\nu\in\Gamma\cap\mathcal{L}_n} H(\nu \mid \mu)\} \leq -\liminf_{n\to\infty}\{\inf_{\nu\in\Gamma\cap\mathcal{L}_n} H(\nu \mid \mu)\} \leq -\inf_{\nu\in\Gamma} H(\nu \mid \mu).$$

This inequality immediately implies directly the equation (3) since we assumed that Γ is an open set, that is, $\Gamma^\circ = \Gamma$.

As a consequence, we have shown that (2) and (3) which together shows that the equality (1) holds. ■

Ex. 2.1.18

a) Let us extend the conclusion of exercise 2.1.16 to any subset Γ of $\{\nu \in M_1(\Sigma) : \Sigma_\nu \subseteq \Sigma_\mu\}$ that is contained in the closure of its interior.

Proof.

Now, let $\Gamma \subseteq \{\nu \in M_1(\Sigma) : \Sigma_\nu \subseteq \Sigma_\mu\}$ be such that we have $\Gamma^\circ \subseteq \Gamma \subseteq \overline{\Gamma^\circ}$.

Clearly, then by the monotonicity of infimum we have

$$\inf_{\nu\in\Gamma^\circ} H(\nu \mid \mu) \geq \inf_{\nu\in\Gamma} H(\nu \mid \mu) \geq \inf_{\nu\in\overline{\Gamma^\circ}} H(\nu \mid \mu) \Leftrightarrow -\inf_{\nu\in\Gamma^\circ} H(\nu \mid \mu) \leq -\inf_{\nu\in\Gamma} H(\nu \mid \mu) \leq -\inf_{\nu\in\overline{\Gamma^\circ}} H(\nu \mid \mu).$$

Hence, let us show the another direction, i.e., $\inf_{\nu\in\Gamma^\circ} H(\nu \mid \mu) \leq \inf_{\nu\in\Gamma} H(\nu \mid \mu)$.

Now, let $\nu \in \Gamma$. Then, since we have that $\Gamma \subseteq \overline{\Gamma^\circ}$, we know that there exists a sequence $\{\nu_n\}$ in Γ° such that $\nu_n \to \nu \in \Gamma$ as $n \to \infty$. By remark of definition 2.15 [1], we know that $H(\cdot \mid \mu)$ is a good rate function on the set $\{\nu \in M_1(\Sigma) : \Sigma_\nu \subseteq \Sigma_\mu\}$, so by lower semi continuity of $H(\cdot \mid \mu)$ and the observation made above yields that for every $\nu \in \Gamma$

$$H(\nu \mid \mu) \geq \liminf_{n\to\infty} H(\nu_n \mid \mu) \geq \inf_{\nu'\in\Gamma^\circ} H(\nu' \mid \mu).$$

This implies directly that $\inf_{\nu\in\Gamma} H(\nu \mid \mu) \geq \inf_{\nu\in\Gamma^\circ} H(\nu \mid \mu)$. Since we have shown both directions, we know that $\inf_{\nu\in\Gamma} H(\nu \mid \mu) = \inf_{\nu\in\Gamma^\circ} H(\nu \mid \mu)$. Combining this equality with Sanov's theorem [1], we have

$$-\inf_{\nu\in\Gamma} H(\nu \mid \mu) = -\inf_{\nu\in\Gamma^\circ} H(\nu \mid \mu) \leq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_\mu(L_n^Y \in \Gamma) \leq \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}_\mu(L_n^Y \in \Gamma) \leq -\inf_{\nu\in\Gamma} H(\nu \mid \mu).$$

Now, as in exercise 2.1.16, this implies that the equation (1) holds also for all sets $\Gamma \subseteq \{\nu \in M_1(\Sigma) : \Sigma_\nu \subseteq \Sigma_\mu\}$ be such that we have $\Gamma^\circ \subseteq \Gamma \subseteq \overline{\Gamma^\circ}$. ■

b) Let us prove that for any set mentioned in part a), $I_\Gamma := \inf_{\nu\in\Gamma} H(\nu \mid \mu) < \infty$ and $I_\Gamma = H(\nu' \mid \mu)$ for some $\nu' \in \overline{\Gamma^\circ}$.

Proof.

Clearly, since we assumed that $\Gamma \subseteq \overline{\Gamma^\circ}$, we know that every $\nu \in \Gamma$ belongs also to $\overline{\Gamma^\circ}$. Hence, if the infimum over Γ for the relative entropy $H(\nu \mid \mu)$ is attained at some point $\nu^* \in \Gamma$, we immeaditly have that

$$\exists \nu^* \in \overline{\Gamma^\circ} : \; H(\nu^* \mid \mu) = \inf_{\nu\in\Gamma} H(\nu \mid \mu) = I_\Gamma. \tag{4}$$

However, the infimum can be also attained outside of the set Γ if Γ is an open set. Hence, we need to cover also these cases as well.

By our assumption, we know that $\Gamma^\circ \subset \Gamma \subseteq K := \{\nu \in M_1(\Sigma) : \Sigma_\nu \subseteq \Sigma_\mu\}$, where K is a known to be a compact set, which is mentioned in the remark of definition 2.5 [1]. Also, the space of probability measures $M_1(\Sigma)$ is assumed to be Hausdorff space (start of the page 8 in the book [1]). Therefore, if we take closure from both sides we get from above observation that

$$\Gamma^\circ \subseteq K \Rightarrow \overline{\Gamma^\circ} \subseteq \overline{K} = K,$$

where the last equality holds since any compact set is also a closed set in Hausdorff space. Now, the fact that $\overline{\Gamma^\circ}$ is a closed subset of the compact set K in Hausdorff space, by Heine-Borel theorem we get that $\overline{\Gamma^\circ}$ is a compact set as well.

Now, suppose that ν^* is the point where $H(\nu^* \mid \mu) = I_\Gamma$. We aim to show that $\nu^* \in \overline{\Gamma^\circ}$. In this case, we assume that the infimum is not necessarily attained inside Γ, that is, $\nu^* \notin \Gamma$. However, we can still pick a sequence $\{\nu_n\} \in \Gamma$ which converges to ν^*. That is, we have $\nu_n \to \nu^*$ as $n \to \infty$.

Now, by our assumption $\Gamma \subset \overline{\Gamma^\circ}$, we know that the sequence $\{\nu_n\} \in \overline{\Gamma^\circ}$. Therefore, by (sequentially) compactness of the set $\overline{\Gamma^\circ}$, we know that the set $\{\nu_n\}$ has convergent subsequence in $\overline{\Gamma^\circ}$ which converges to same value as the sequence converges. That is, we know that for the sequence $\{\nu_n\}$ there exists convergent subsequence $\{\nu_{n_k}\} \to \nu^* \in \overline{\Gamma^\circ}$. This confirms that the statement (4) holds.

Lastly, let us show that I_Γ is finite. Since we assumed that $\Gamma \subseteq \{\nu \in M_1(\Sigma) : \Sigma_\nu \subseteq \Sigma_\mu\}$, we know that $\Sigma_\nu := \{a_i : \nu(a_i) > 0\} \subseteq \Sigma_\mu = \{a_i : \mu(a_i) > 0\}$. Hence, we get that by the definition of relative entropy

$$I_\Gamma = \inf_{\nu\in\Gamma} H(\nu \mid \mu) = \inf_{\nu\in\Gamma} \sum_{i=1}^{|\Sigma|} \nu(a_i) \log \frac{\nu(a_i)}{\mu(a_i)} \leq \sum_{i=1}^{|\Sigma_\mu|} \nu(a_i) \log \frac{\nu(a_i)}{\mu(a_i)} = \sum_{i=1}^{|\Sigma_\nu|} \overbrace{\nu(a_i)}^{>0} \cdot \log \overbrace{\frac{\nu(a_i)}{\mu(a_i)}}^{>0} < \infty.$$

Here, the first inequality follows from the fact that we do not take infimum anymore but we can still be in the range of Γ, that is, take only the values of $a_i \in \Sigma_\mu$ since from all other values $\nu(a_i) = \mu(a_i) = 0$. The second equality follows from the fact $\Sigma_\nu \subseteq \Sigma_\mu$, and lastly when taking the sum over Σ_ν, we know that for all i, we have $\mu(a_i), \nu(a_i) > 0$ which yields that the given terms are finite, and taking finite sum over finite elements is finite. ■

Ex. 2.1.21

Let us find a closed set Γ such that $\inf_{\nu\in\Gamma} H(\nu \mid \mu) < \infty$ and $\Gamma = \overline{\Gamma^\circ}$ while $\inf_{\nu\in\Gamma^\circ} H(\nu \mid \mu) = \infty$.

Solution:

For example, let us choose the closed set Γ such as $\Gamma = M_1(\Sigma)$, where the cardinality of Σ is $|\Sigma| = 2$. Suppose also, that $\Sigma \neq \Sigma_\mu = \{a_i : \mu(a_i) > 0\}$ such that $\mu(a_1) = 0$. That is, we must have that $\mu = \mathbf{1}_{a_2}$ in this case.

By our assumption, we know that Γ is the space of all probability space generated by two finite elements. Hence, Γ is a whole space, which means that the closure of Γ° must be the same as the whole space, that is, $\overline{\Gamma^\circ} = M_1(\Sigma) = \Gamma$.

Using this and the definition of relative entropy, we have

$$\inf_{\nu\in\Gamma} H(\nu \mid \mu) = \inf_{\nu\in\Gamma} \sum_{i=1}^{|\Sigma|} \nu(a_i) \log \frac{\nu(a_i)}{\mu(a_i)} = \inf_{\nu\in\Gamma} \Big(\nu(a_1) \cdot \log \frac{\nu(a_1)}{\mathbf{1}_{a_2}(a_1)} + \nu(a_2) \cdot \log \frac{\nu(a_2)}{\mathbf{1}_{a_2}(a_2)}\Big)$$
$$= \inf_{\nu\in\Gamma} \Big(\nu(a_1) \cdot \log \frac{\nu(a_1)}{\mathbf{1}_{a_2}(a_1)} + \nu(a_2) \cdot \log \nu(a_2)\Big) = 0 < \infty,$$

where the last equality follows from the fact that in order to first term be finite, we must have $\nu(a_1) = 0$ and therefore $\nu(a_2) = 1$. Otherwise, if $\nu(a_1) > 0$ the first term will diverge, and therefore we know that the infimum is attained when $\nu = \mathbf{1}_{a_2}$ as well.

Now, according of the construction of $\Gamma = M_1(\Sigma) = \{\nu \in \Sigma : \nu(a_i) \geq 0, \sum_{i=1}^{|\Sigma|} \nu(a_i) = 1\}$, its interior must be

$$\Gamma^\circ = \{\nu \in \Sigma : \nu(a_i) > 0, \sum_{i=1}^{|\Sigma|} \nu(a_i) = 1\}.$$

Since this construction yields that for all $\nu \in \Gamma^\circ$ we have $\nu(a_1) > 0$, by the relative entropy with respect to μ, the first term diverges for every $\nu \in \Gamma^\circ$. That is, we have

$$\inf_{\nu\in\Gamma^\circ} H(\nu \mid \mu) = \infty.$$

Hence, we have found a closed set Γ, which has the wanted properties.

2.1.2 Cramer's theorem for Finite Alphabets in $\mathbb{R}$

In this section, we note that the empirical mean is $\hat{S}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i = f(Y_i)$ with deterministic function $f : \Sigma \to \mathbb{R}$, and $Y_i \in \Sigma$ are i.i.d with law μ. Therefore, the random variables $X_1, \ldots, X_n$ are i.i.d as well since f can be assumed to be measurable function. Furthermore, without the loss of generality it is assumed that $\Sigma = \Sigma_\mu$ and that $f(a_1) < f(a_2) < f(a_{|\Sigma|})$. Hence, the random variables $\hat{S}_n$ assume values in the compact interval $K = [f(a_1), f(a_{|\Sigma|})]$.

Exercise 2.1.29

a) Let us prove that $I(x) = 0 \Leftrightarrow x = \mathbb{E}[X_1]$ with the terms made on the beginning of chapter 2.1.2.

Proof.

Let us show both implications. By Cramér's theorem (theorem 2.1.14 [1]) we know that the rate function $I(x)$ is continuous at $x \in K$, and satisfies there

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

where $\Lambda(\lambda) = \log \sum_{i=1}^{|\Sigma|} \mu(a_i) e^{\lambda f(a_i)} = \log \mathbb{E}[e^{\lambda X_i}]$ since $X_i = f(Y_i)$ and X_i's are i.i.d. Therefore, the given sum is the same as the moment generating function of X_i in finite alphabet in this case. In light of this, we get that

$$I(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \Lambda(\lambda)\}) = 0 \Leftrightarrow \frac{d}{d\lambda}(\lambda x - \log \mathbb{E}[e^{\lambda X_1}]) \mid_{\lambda=0} = 0 \Leftrightarrow (x - \frac{\mathbb{E}[X_1 e^{\lambda X_1}]}{\mathbb{E}[e^{\lambda X_1}]}) \mid_{\lambda=0} = 0$$
$$x - \frac{\mathbb{E}[X_1 \cdot e^{0 \cdot X_1}]}{\mathbb{E}[e^{0 \cdot X_1}]} = 0 \Leftrightarrow x = \mathbb{E}[X_1],$$

where the first step follows from the fact that since $\Lambda(\cdot)$ is strictly convex, we have that the supremum is attained when $\lambda = 0$ in this case. This proves the given equality. ■

This result was anticipated since according to the weak law of large numbers, we know that the empirical mean $\hat{S}_n$ approaches to $\mathbb{E}[X_1]$ in probability. That is, $\hat{S}_n \xrightarrow{\mathbb{P}} \mathbb{E}[X_1]$ as $n \to \infty$. This means that the empiricial mean $\hat{S}_n$ is concentrated around the mean of X_1, which suggests that the large deviations away from $\mathbb{E}[X_1]$ are increasingly unlikely. That is, in terms of the Large deviation principle (LDP) rate function, we should have that at $x = \mathbb{E}[X_1] \Leftrightarrow I(x) = 0$, and $I(x) > 0$ elsewhere which indicates exponentially small probabilities away from $\mathbb{E}[X_1]$.

b) Let us check that $H(\nu \mid \mu) = 0$ if and only if $\nu = \mu$.

Proof.

Let us show both implications. First, suppose $H(\nu \mid \mu) = 0$. Then by the definition of relative entropy, we have

$$H(\nu \mid \mu) = \sum_{i=1}^{|\Sigma|} \nu(a_i) \log \frac{\nu(a_i)}{\mu(a_i)} = 0 \Rightarrow \nu(a_i) \log \frac{\nu(a_i)}{\mu(a_i)} = 0 \ \forall i = 1, \ldots, |\Sigma|$$
$$\Rightarrow \nu(a_i) = \mu(a_i) \ \forall i = 1, \ldots, |\Sigma|.$$

Now, according to Dynkin's identification theorem this implies that $\nu = \mu$.

Secondly, we now suppose that $\nu = \mu$. Then directly by using the definition of relative entropy, we have

$$H(\nu \mid \mu) = H(\mu \mid \mu) = \sum_{i=1}^{|\Sigma|} \mu(a_i) \log \frac{\mu(a_i)}{\mu(a_i)} = \sum_{i=1}^{|\Sigma|} \mu(a_i) \overbrace{\log(1)}^{=0} = 0.$$

We conclude that $H(\nu \mid \mu) = 0 \Leftrightarrow \nu = \mu$. ■

In light of the result we discovered, we can interpreted the relative entropy $H(\nu \mid \mu)$ as a measure of difference ν is from μ, since $H(\cdot \mid \mu)$ is a nonnegative, and only zero if the measures equal. However, it should not be interpreted to be some kind of metric since it does not satisfy the triangle inequality. For example, if we consider $\Sigma = \{1, 2\}$. Then, for probability measures $\nu = (1, 0),\ \mu = (0, 1)$ and $\rho = (\frac{1}{2}, \frac{1}{2})$ the triangle inequality does not hold for the relative entropy.

c) Let us prove the strong law of large numbers by showing that, for all $\epsilon > 0$

$$\sum_{n=1}^{\infty} \mathbb{P}[|\hat{S}_n - \mathbb{E}[X_1]| > \epsilon] < \infty$$

Proof.

First, for the empirical mean $\hat{S}_n$, we shall estimate $\mathbb{P}[|\hat{S}_n - \mathbb{E}[X_1]| > \epsilon] = \mathbb{P}[\hat{S}_n \in A]$, where $A := (-\infty, \mathbb{E}[X_1] - \epsilon) \cup (\mathbb{E}[X_1] + \epsilon, \infty) \subset \mathbb{R}$.

Now, according to Cramer's theorem [1], we can get for this probability an upper bound

$$\frac{1}{n} \log \mathbb{P}[\hat{S}_n \in A] \leq - \inf_{x \in A} I(x) \Leftrightarrow \mathbb{P}[\hat{S}_n \in A] \leq e^{-n \cdot \inf_{x \in A} I(x)} = e^{-n \cdot \delta},$$

with the convention $\delta = \inf_{x \in A} I(x)$. By part b), we assured that $I(x) \geq 0$ and $I(x) = 0$ if and only if $x = \mathbb{E}[X_1]$. Hence, we know that $\delta > 0$ in this case, since $\mathbb{E}[X_1] \notin A$.

Using, this upper bound for the sum, we obtain that

$$\sum_{n=1}^{\infty} \mathbb{P}[|\hat{S}_n - \mathbb{E}[X_1]| > \epsilon] = \sum_{n=1}^{\infty} \mathbb{P}[\hat{S}_n \in A] \leq \sum_{n=1}^{\infty} e^{-n\delta} = \frac{e^{-\delta}}{1 - e^{-\delta}} < \infty,$$

where the last equality followed from the converging of the geometric series since $0 < e^{-\delta} < 1$. By Borell-Cantelli lemma of convergent part, this implies that the event $|\hat{S}_n - \mathbb{E}[X_1]| > \epsilon$ occurs only finitely often. In another words, we have $\mathbb{P}[\lim_{n \to \infty} \hat{S}_n = \mathbb{E}[X_1]] = 1$ which is the same as the strong law of large numbers. ■

Exercise 2.1.30

Let us guess the value of $\lim_{n \to \infty} \mathbb{P}_\mu[X_1 = f(a_i) \mid \hat{S}_n \geq q]$ for $q \in A = (\mathbb{E}[X_1], f(a_{|\Sigma|})$.

Solution:

First, we notice that the condition $\{\hat{S}_n \geq q\}$ is a rare event, since we have shown in exercise 2.1.29 that the empirical mean approaches to $\mathbb{E}[X_1]$ as $n \to \infty$. Thus, if we use the coniditonal probability formula, we get the giving formula of the limit into the formula

$$\mathbb{P}_\mu[X_1 = f(a_i) \mid \hat{S}_n \geq q] = \frac{\mathbb{P}_\mu[X_1 = f(a_i), \hat{S}_n \geq q]}{\mathbb{P}_\mu[\hat{S}_n \geq q]} = \frac{\mathbb{P}_\mu[X_1 = f(a_i), \frac{1}{n}(f(a_i) + \sum_{i=2}^{n} X_i) \geq q]}{\mathbb{P}_\mu[\hat{S}_n \geq q]}$$

$$\overset{\perp\!\!\!\perp}{=} \frac{\mathbb{P}_\mu[X_1 = f(a_i)] \cdot \mathbb{P}_\mu[\sum_{i=2}^n X_i \geq nq - f(a_i)]}{\mathbb{P}_\mu[\hat{S}_n \geq q]} = \frac{\mu(f(a_i)) \cdot \mathbb{P}_\mu[\frac{1}{n-1}\sum_{i=2}^n X_i \geq \frac{nq - f(a_i)}{n-1}]}{\mathbb{P}_\mu[\hat{S}_n \geq q]}$$
$$= \frac{\mu(f(a_i)) \cdot \mathbb{P}_\mu[\frac{1}{n-1}\sum_{i=1}^{n-1} X_i \geq \frac{nq - f(a_i)}{n-1}]}{\mathbb{P}_\mu[\hat{S}_n \geq q]} = \frac{\mu(f(a_i)) \cdot \mathbb{P}_\mu[\hat{S}_{n-1} \geq \frac{nq - f(a_i)}{n-1}]}{\mathbb{P}_\mu[\hat{S}_n \geq q]},$$

where we used the fact that $X_i = f(Y_i)$'s are i.i.d. since $Y_1, \ldots, Y_n$ are i.i.d with law μ. That is, in the first equality we used the conditional probability formula, third equality independence of X_1 to X_is when $i \geq 2$, when X_1 is set to be fixed. Second last equality, we use the shifting of the sum since $X_1, \ldots, X_n$ are identically distributed.

Now, since A is an open set, we have that the both events in denominator and numerator are open sets, which are the rare events. Hence, by using Cramer's theorem (theorem 2.1.24) the limit comes exact, and we get

$$\lim_{n\to\infty} \mathbb{P}_\mu[X_1 = f(a_i) \mid \hat{S}_n \geq q] = \lim_{n\to\infty} \frac{\mu(f(a_i) \cdot \mathbb{P}_\mu[\hat{S}_{n-1} \geq \frac{nq - f(a_i)}{n-1}]}{\mathbb{P}_\mu[\hat{S}_n \geq q]} = \lim_{n\to\infty} \frac{\mu(f(a_i)) \cdot e^{-(n-1)\cdot I(\frac{nq-f(a_i)}{n-1})}}{e^{-nI(q)}}.$$

Now, as $\frac{nq-f(a_i)}{n-1} \to q$ as $n \to \infty$. Hence, we can take first order Taylor expansion, to get $I(\frac{nq-f(a_i)}{n-1}) = I(q) + I'(q)(\frac{nq-f(a_i)}{n-1} - q) = I(q) + I'(q) \cdot \frac{q - f(a_i)}{n-1} + O(\frac{1}{n})$. Therefore, we get

$$\lim_{n\to\infty} \frac{\mu(f(a_i)) \cdot e^{-(n-1)\cdot I(\frac{nq-f(a_i)}{n-1})}}{e^{-nI(q)}} = \lim_{n\to\infty} \mu(f(a_i)) \cdot e^{-(n-1)\cdot(I(q)+I'(q)\cdot\frac{q-f(a_i)}{n-1}+O(\frac{1}{n}))+nI(q)}$$
$$= \mu(f(a_i))e^{I(q)-I'(q)(q-f(a_i))}.$$

Hence, the giving limit is

$$\lim_{n\to\infty} \mathbb{P}_\mu[X_1 = f(a_i) \mid \hat{S}_n \geq q] = \mu(f(a_i)) \cdot e^{I(q)-I'(q)(q-f(a_i))},$$

which also can be expressed in another form when we note, e.g., $\eta = I'(q)$. Then, $\Lambda(\eta) = \eta q - I(q)$ which yields

$$\lim_{n\to\infty} \mathbb{P}_\mu[X_1 = f(a_i) \mid \hat{S}_n \geq q] = \mu(f(a_i)) \cdot e^{I(q)-\eta(q-f(a_i))} = \mu(f(a_i)) \cdot e^{\eta f(a_i)-\Lambda(\eta)} = \tilde{\mu}(f(a_i)),$$

where $\tilde{\mu}$ is a new probability measure, known as tilted measure:

$$\frac{d\tilde{\mu}}{d\mu}(x) = e^{\eta x - \Lambda(\eta)}$$

as defined in the proof of theorem 2.2.3 [1].

2.1.3 Large Deviations for sampling Without replacement

Exercise 2.1.47

Let $I_\Gamma = \inf_{v\in\Gamma^\circ} I(\nu \mid \beta,\mu)$. Let us prove that when $\Gamma \subset \mathcal{D}_I := \{\nu : I(\nu \mid \beta,\mu) < \infty\}$ and $\Gamma \subseteq \overline{\Gamma^\circ}$, then

$$I_\Gamma = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}[L_n^Y \in \Gamma]. \tag{5}$$

Proof.

First, we fix $\mu \in M_1(\Sigma)$, and $\beta \in (0,1)$. Under the assumptions made on theorem 2.1.41, we know that

$$-\inf_{\nu\in\Gamma^\circ} I(\nu \mid \beta,\mu) \le \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}[L_n^Y \in \Gamma] \le \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}[L_n^Y \in \Gamma] \le -\inf_{\nu\in\overline{\Gamma}} I(\nu \mid \beta,\mu).$$

It now suffices to show that from this given inequality, the given results follows. In this case, we assumed also that $\Gamma^\circ \subseteq \Gamma \subseteq \overline{\Gamma^\circ}$ which implies when taking the closure $\overline{\Gamma^\circ} \subseteq \overline{\Gamma} \subseteq \overline{\overline{\Gamma^\circ}} = \overline{\Gamma^\circ} \Rightarrow \overline{\Gamma^\circ} = \overline{\Gamma}$. This result is the consequence of taking the closure of closed set which is the same as the set itself.

On the other hand, we know that $\Gamma \subset \mathcal{D}_I$, where $\mathcal{D}_I$ is a compact set. As a result of this and the continuity of $I(\cdot \mid \beta,\mu)$, the exercise 2.1.18 part b), tells us $\inf_{\nu\in\overline{\Gamma^\circ}} I(\nu \mid \beta,\mu) = \inf_{\nu\in\Gamma^\circ} I(\nu \mid \beta,\mu)$. Combining these result from the inequality provided from theorem 2.1.41, we get that

$$-I_\Gamma \le \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}[L_n^Y \in \Gamma] \le \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}[L_n^Y \in \Gamma] \le -\inf_{\nu\in\overline{\Gamma}} I(\nu \mid \beta,\mu) = -I_\Gamma,$$

from which directly we get the wanted result (5). ■

Exercise 2.1.48

Let us prove that the rate function $I(\cdot \mid \beta,\mu)$ is convex.

Proof.

First, let us note that the relative entropy is $H(\nu \mid \mu)$ is convex when $\mu,\nu \ge 0$ which they are as a probability measures. This can be checked easily by computing the Hessian matrix for $x \cdot \log(x/y)$, which is

$$A := \mathrm{Hes}\Big(x\log(\frac{x}{y})\Big) = \begin{pmatrix} \frac{\partial}{\partial x^2} x\log(\frac{x}{y}) & \frac{\partial}{\partial x}\frac{\partial}{\partial y} x\log(\frac{x}{y}) \\ \frac{\partial}{\partial y}\frac{\partial}{\partial x} x\log(\frac{x}{y}) & \frac{\partial}{\partial y^2} x\log(\frac{x}{y}) \end{pmatrix} = \begin{pmatrix} \frac{1}{x} & -\frac{1}{y} \\ -\frac{1}{y} & \frac{x}{y^2} \end{pmatrix}$$

Then, by using Sylvester's criterion, the matrix A is positive-semidefinitive since $1/x > 0$ for $x > 0$, and

$$\det(A) = \frac{1}{x}\cdot\frac{x}{y^2} - (-\frac{1}{y})(-\frac{1}{y}) = 0.$$

Therefore, the function $x \cdot \log(x/y)$ is convex (but not strictly convex). Since $H(\nu \mid \mu)$ is a sum of $\nu(a_i)\log(\frac{\nu(a_i)}{\mu(a_i)})$, and linear combination of convex functions is convex. Therefore, we know that $H(\nu \mid \mu)$ is a convex function.

Now, fix $\beta \in (0,1)$ and $\mu \in M_1(\Sigma)$. Then, if we recall the definition of $I(\cdot \mid \beta, \mu)$ from (2.1.32) in the book, we need only focus on the domain $D = \{\nu : \mu(a_i) \geq \beta\nu(a_i)\ \forall i\}$. Now, in this domain D, we get that the rate function is by definition

$$I(\nu \mid \beta, \mu) = H(\nu \mid \mu) + \frac{1-\beta}{\beta} \cdot H(\frac{\mu - \beta\nu}{1-\beta} \mid \mu).$$

Now, by above reasoaning we know that $\nu \to H(\nu \mid \mu)$ is convex. Also, since $\frac{\mu - \beta\nu}{1-\beta} \geq 0$ in the domain D, we get that also the second part is $\nu \to H(\frac{\mu-\beta\nu}{1-\beta} \mid \mu)$ is convex also. By multiplying convex function with positive scalar preserves convexity, and remembering that linear combination of convex functions is convex, we get that the rate function $I(\cdot \mid \beta, \nu)$ is convex. ■

2.2 Cramer's theorem

Throughout this section, we consider the empirical means $\hat{S}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$ for i.i.d. d-dimensional random vectors $X_1, \ldots, X_n, \ldots$, with X_1 distributed according tot the probability law $\mu \in M_1(\mathbb{R}^d)$. The logarithmic moment generating function associated with the law μ is defined as

$$\Lambda(\lambda) = \log M(\lambda) = \log \mathbb{E}[e^{\langle \lambda, X_1 \rangle}],$$

where $\langle \lambda, x \rangle = \sum_{j=1}^{d} \lambda_j x_j$ is the usual scalar product in $\mathbb{R}^d$ as stated beginning of section 2.2 [1]. Furthermore, we use definition of the Fenchel–Legendre transform of $\Lambda(\lambda)$ (definition 2.2.2 [1]) which is

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \{\langle \lambda, x \rangle - \Lambda(\lambda)\},$$

and use it here onward in another sections as well.

2.2.1 Cramer's theorem in $\mathbb{R}$

In this section, we first focus Cramer's theorem (theorem 2.2.3 and theorem 2.2.30 [1]) just in $\mathbb{R}$, and later study its extension into $\mathbb{R}^d$ in section 2.2.2. Also, we note two sets

$$\mathcal{D}_\Lambda := \{\lambda : \Lambda(\lambda) < \infty\} \quad \text{and} \quad \mathcal{D}_{\Lambda^*} := \{x : \Lambda^*(x) < \infty\}$$

which are useful in the derivation of Cramer's theorem, and we will use them in the following exercises.

Exercise 2.2.22

Let us prove by an application of Fatou's lemma that $\Lambda(\cdot)$ is lower semicontinous.

Proof.

For simplicity as in given section we will just focus on $\mathbb{R}$. Let us take sequence $\{\lambda_n\}$ such that $\lambda_n \to \lambda$ as $n \to \infty$. We aim to show that

$$\liminf_{n\to\infty} \Lambda(\lambda_n) \geq \Lambda(\lambda). \tag{6}$$

It is easy to prove that functions of the form $f_n(x) := e^{\langle \lambda_n, x \rangle} > 0$ are measurable functions since as it is composition of continuous functions (exponential and linear), and continuous functions are measurable functions. Hence, for measurable functions $f_n : \mathbb{R} \to \overline{\mathbb{R}}_+$ we can apply the Fatou's lemma (or Fatou's inequality) as an application to obtain

$$\liminf_{n\to\infty} \Lambda(\lambda_n) = \liminf_{n\to\infty} \log \mathbb{E}[e^{\langle \lambda_n, X_1 \rangle}] \geq \log \mathbb{E}[\liminf_{n\to\infty} e^{\langle \lambda_n, X_1 \rangle}] = \log \mathbb{E}[e^{\langle \lambda, X_1 \rangle}] = \Lambda(\lambda),$$

where the inequality follows in addition from the fact that $\log x$ in an increasing function so it preserves the inequality obtained from the Fatou's lemma. This implies directly that (6) holds. Therefore, $\Lambda(\cdot)$ is lower semicontinuous. ■

Exercise 2.2.25

a) Suppose A is a Borel measurable set such that $[y, z) \subset A \subset [y, \infty)$ for some $y < z$ and either $\mathcal{D}_\Lambda = \{0\}$ or $\overline{x} < z$. Let us prove that

$$\lim_{n\to\infty} \frac{1}{n} \log \mu_n(A) = -\inf_{x \in A} \Lambda^*(x). \tag{7}$$

Proof.

By our assumption we know that $[y,z) \subset A \subset [y,\infty)$ for some $y < z$, and since $[y,\infty)$ is known to be closed set in $\mathbb{R}$ induced by usual Euclidean norm, we can use the monotonicity of measures for $A \subset [y,\infty)$, and Cramer's theorem (theorem 2.2.3) for closed sets to get upper bound

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(A) \leq \limsup_{n\to\infty} \frac{1}{n} \log \mu_n([y,\infty)) \leq - \inf_{x\in[y,\infty)} \Lambda^*(x). \tag{8}$$

For lower bound, we use the observation that $(y+\delta, z) \subset [y+\delta, z) \subset [y,z) \subset A$ for small enough $\delta > 0$, and since $(y+\delta, z)$ is an open set we can apply Cramer's theorem for open set to get

$$\liminf_{n\to\infty} \log \mu_n(A) \geq \liminf_{n\to\infty} \log \mu_n((y+\delta, z)) = - \inf_{x\in(y+\delta,z)} \Lambda^*(x). \tag{9}$$

By using continuity of Λ^* at y, and taking $\delta \to 0$, we obtain that $\lim_{\delta\to 0} \inf_{x\in(y+\delta,z)} \Lambda^*(x) = \inf_{x\in(y,z)} \Lambda^*(x) = \inf_{x\in[y,z)} \Lambda^*(x)$. Combining this to (9), we obtain

$$\liminf_{n\to\infty} \log \mu_n(A) \geq - \inf_{x\in[y,z)} \Lambda^*(x). \tag{10}$$

Lastly, we assumed $\mathcal{D}_\Lambda = \{0\}$ or $\overline{x} < z$.

If $\mathcal{D}_\Lambda = \{0\}$, by Lemma 2.2.5 part b) [1], we know that $D_{\Lambda^*} = \{0\}$. This means that for any $y \in \mathbb{R}$ either $\inf_{x\in[y,z)} \Lambda^*(x) = \infty = \inf_{x\in[y,\infty)} \Lambda^*(x)$ or $\inf_{x\in[y,z)} \Lambda^*(x) = 0 = \inf_{x\in[y,\infty)} \Lambda^*(x)$.

If $\overline{x} < z$, then by Lemma 2.2.5 part b), for any $\lambda \in \mathbb{R}$, the infimum is achieved at $\overline{x}$. So depending of $y \in \mathbb{R}$, the infimum over set $[y,z)$ and $[y,\infty)$ is either achieved at $\overline{x} > y$ or $\overline{x} < y$. Hence, the infimum over the sets $[y,z)$ and $[y,\infty)$ coincide. That is

$$\inf_{[y,z)} \Lambda^*(x) = \inf_{[y,\infty)} \Lambda^*(x) = \inf_{A} \Lambda^*(x),$$

where the last equality follows from $[y,z) \subset A \subset [y,\infty)$. Therefore, combining this result for (8) and (10), we get

$$- \inf_{x\in A} \Lambda^*(x) \leq \liminf_{n\to\infty} \log \mu_n(A) \leq \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(A) \leq - \inf_{x\in A} \Lambda^*(x).$$

This inequality yields directly the wanted result (7). ■

b) Let us prove the conclusion of Corollary 2.2.19 [1] holds for $A = (y,\infty)$ when $y \in \mathcal{F}^\circ = \{\Lambda'(\lambda) : \lambda \in \mathcal{D}_\Lambda^\circ\}^\circ$ and $y > \overline{x}$.

Proof.

We aim to show that for $A = (y,\infty)$, we have

$$\lim_{n\to\infty} \frac{1}{n} \log \mu_n(A) = - \inf_{x\in A} \Lambda^*(x). \tag{11}$$

First, since $[y+\delta,\infty) \subset A \subset [y,\infty)$ for all $\delta > 0$, we know that $\lim_{\delta\to 0} [y+\delta,\infty) = (y,\infty) = A$. Using this observation and Corollary 2.2.19 for the set $[y+\delta,\infty)$, we obtain

$$\lim_{n\to\infty} \frac{1}{n} \mu_n(A) = \lim_{\delta\to 0} \lim_{n\to\infty} \frac{1}{n} \mu_n([y+\delta,\infty)) = - \lim_{\delta\to 0} \inf_{x\geq y+\delta} \Lambda^*(x).$$

We assumed that $y \in \mathcal{F}^\circ$ and $y > \overline{x}$ which with the help of exercise 2.2.24 means that Λ^* is strictly convex and C^∞ at neighborhood of y. Therefore, taking $\delta \to 0$ yields by continuity of Λ^* at neighborhood of y:

$$-\lim_{\delta \to 0} \inf_{x \geq y+\delta} \Lambda^*(x) = -\inf_{x>y} \Lambda^*(x) = -\inf_{x \in A} \Lambda^*(x).$$

Combining this result to above equality, we obtain the wanted result ■

2.2.2 Cramer's theorem in $\mathbb{R}^d$

In this section, we are considering Cramer's theorem in a multivariate case $\mathbb{R}^d$. As in the section 2.2.2 [1], we assume that $\mathcal{D}_\Lambda = \mathbb{R}^d$ unless otherwise mentioned.

Exercise 2.2.38

Let μ_n denote the law of $\hat{S}_n$, the empirical mean of the i.i.d. random vectors $X_i \in \mathbb{R}^d$, and $\Lambda(\cdot)$ denote the logarithmic moment generation function associated with the law of X_1. Also, let us not assume that $\mathcal{D}_{\Lambda^*} = \mathbb{R}^d$.

a) Let us use the Chebycheff's inequality to prove that for any measurable set $C \subset \mathbb{R}^d$, any n, and any $\lambda \in \mathbb{R}^d$,

$$\frac{1}{n} \log \mu_n(C) \leq -\inf_{y \in C} \langle \lambda, y \rangle + \Lambda(\lambda). \tag{12}$$

Proof.

Let $C \subset \mathbb{R}^n$. Then for the law of $\hat{S}_n$, we get for the event $\{\hat{S}_n \in C\} \subseteq \{\langle \lambda, \hat{S}_n \rangle \geq \inf_{y \in C} \langle \lambda, y \rangle\}$ when $\lambda > 0$ by using the Chebycheff's inequality (Markov's inequality in this case) that

$$\begin{aligned}
\mu_n(C) = \mathbb{P}[\hat{S}_n \in C] &\overset{\text{mon.}}{\leq} \mathbb{P}[n\langle \lambda, \hat{S}_n \rangle \geq n \inf_{y \in C} \langle \lambda, y \rangle] = \mathbb{P}[e^{n\langle \lambda, \hat{S}_n \rangle} \geq e^{n \inf_{y \in C} \langle \lambda, y \rangle}] \\
&\overset{\text{Cheb.}}{\leq} e^{-n \inf_{y \in C} \langle \lambda, y \rangle} \mathbb{E}[e^{n\langle \lambda, \hat{S}_n \rangle}] = e^{-n \inf_{y \in C} \langle \lambda, y \rangle} \mathbb{E}[e^{\sum_{i=1}^n \langle \lambda, X_i \rangle}] \overset{\text{i.i.d.}}{=} e^{-n \inf_{y \in C} \langle \lambda, y \rangle} \prod_{i=1}^n \mathbb{E}[e^{\langle \lambda, X_1 \rangle}] \\
&= e^{-n \inf_{y \in C} \langle \lambda, y \rangle} \mathbb{E}[e^{\langle \lambda, X_1 \rangle}]^n.
\end{aligned}$$

Taking logarithm from both sides then yields

$$\begin{aligned}
&\mu_n(C) = e^{-n \inf_{y \in C} \langle \lambda, y \rangle} \mathbb{E}[e^{\langle \lambda, X_1 \rangle}]^n \Leftrightarrow \\
&\log \mu_n(C) = \log \left(e^{-n \inf_{y \in C} \langle \lambda, y \rangle} \mathbb{E}[e^{\langle \lambda, X_1 \rangle}]^n \right) = -n \inf_{y \in C} \langle \lambda, y \rangle + n \log \mathbb{E}[e^{\langle \lambda, X_1 \rangle}] \\
&\frac{1}{n} \log \mu_n(C) = -\inf_{y \in C} \langle \lambda, y \rangle + \log \mathbb{E}[e^{\langle \lambda, X_1 \rangle}] = -\inf_{y \in C} \langle \lambda, y \rangle + \Lambda(\lambda),
\end{aligned}$$

which we sought to proof. ■

b) Let us apply the min-max theorem noted in exercise to justify the upper bound

$$\frac{1}{n} \log \mu_n(C) \leq -\sup_{\lambda \in \mathbb{R}^d} \inf_{y \in C} [\langle \lambda, y \rangle - \Lambda(\lambda)] = -\inf_{y \in C} \Lambda^*(y)$$

for every n, and every convex, compact set C.

Proof.

Let us denote $g(\lambda, y) = \langle \lambda, y \rangle - \Lambda(\lambda)$. Now, in exercise 2.22 we have shown that $\Lambda(\lambda)$ is lower semicontinuous by using Fatou's lemma, and convexity of Λ follows by Hölder's inequality as shown in Lemma 2.2.31 [1]. Therefore, we note that $-\Lambda(\lambda)$ must be concave. In addition, by bilinearity of standard inner product, it is easy to verify that $\langle \lambda, y \rangle$ is convex and concave function, and since it is linear function of λ and y, respectively. We note that $\langle \lambda, y \rangle$ is also continuous. Hence, it is upper and lower semicontinuous by definition.

Therefore, taking the linear combination $g(\lambda, y) = \langle \lambda, y \rangle - \Lambda(\lambda)$, we note that the function g is concave and upper semicontinuous in λ, and convex and lower semicontinous in y. Hence, we can apply the min-max theorem for the function g to conclude for part a) result that there holds inequality

$$\begin{aligned}\frac{1}{n}\log \mu_n(C) &\overset{a)}{\leq} -\sup_{\lambda\in\mathbb{R}^d}\inf_{y\in C}[\langle \lambda, y\rangle - \Lambda(\lambda)] = -\sup_{\lambda\in\mathbb{R}^d}\inf_{y\in C} g(\lambda, y)\\ &\overset{\text{min-max}}{=} -\inf_{y\in C}\sup_{\lambda\in\mathbb{R}^d}[g(\lambda, y)] = -\inf_{y\in C}[\sup_{\lambda\in\mathbb{R}^d}\{\langle \lambda, y\rangle - \Lambda(\lambda)\}] = -\inf_{y\in C}\Lambda^*(y),\end{aligned}$$

when $C \subset \mathbb{R}^d$ is convex and compact. ■

c) Let us show that the preceding upper bound holds for every n and every convex, closed set C by considering the convex, compact sets $C \cap [-\rho, \rho]^d$ with $\rho \to \infty$.

Proof.

Let us consider $C \subset \mathbb{R}^d$ to be convex, closet set, and let us consider $A_\rho = C \cap [-\rho, \rho]^d$.

Now, this setting implies that A_ρ is a convex, closed set as intersection of two closed and convex sets. In addition, since $[-\rho, \rho]^d \subset \mathbb{R}^d$ is bounded for every $\rho \geq 0$ which implies that A_ρ must be bounded as well since $A_\rho \subset [-\rho, \rho]^d$. Therefore, by Heine-Borel theorem, we note that A_ρ is a compact, convex set for every ρ. Hence, part b) result implies that

$$\frac{1}{n}\log \mu_n(A_\rho) \leq -\inf_{y\in A_\rho}\Lambda^*(y). \tag{13}$$

On the other hand, since $A_\rho \subseteq C$ for every $\rho \geq 0$, and for sufficiently large ρ we have $C \subset [-\rho, \rho]^d$. Hence, it follows that when taking $\rho \to \infty$, we get $\lim_{\rho\to\infty} A_\rho = \lim_{\rho\to\infty} C \cap [-\rho, \rho]^d = C$. Hence, the left and right hand-side of (13) becomes

$$\lim_{\rho\to\infty}\frac{1}{n}\log \mu_n(A_\rho) = \frac{1}{n}\log \mu_n(C) \quad \text{and} \quad \lim_{\rho\to\infty} -\inf_{y\in A_\rho}\Lambda^*(y) \leq -\inf_{y\in C}\Lambda^*(y),$$

respectively. Here, the right hand-side inequality follows from by definition of infimum and that $A_\rho \subset C$ for all $\rho \geq 0$. Combining these two sides in (13) then yields that for every convex, closed set C

$$\frac{1}{n}\log \mu_n(C) \leq -\inf_{y\in C}\Lambda^*(y).$$

Lastly, we note that if $C = \mathbb{R}^d$, then $\frac{1}{n}\log \mu_n(\mathbb{R}^d) = 0$, and due to fact Λ^* is convex and lower semi continuous we have $\inf_{y\in\mathbb{R}^d}\Lambda^*(y) \leq 0$. Hence, the inequality holds also for closed set $\mathbb{R}^d$. ■

d) Let us use this bound to show that the large deviations upper bound holds for all compact sets (with the rate function Λ^*).

Proof.

Let $D \subset \mathbb{R}^d$ be a compact set. We recall that by Heine-Borel theorem D is a compact set if and only if it is closed and bounded. Hence, the compactness of D implies that it is also closed on $\mathbb{R}^d$. Therefore, by part c), the large deviations upper bound holds for compact set D. That is,

$$\frac{1}{n}\log \mu_n(D) \leq -\inf_{y\in D} \Lambda^*(y)$$

for every n. Hence, by taking upper limit when $n \to \infty$, it follows that

$$\limsup_{n\to\infty} \frac{1}{n}\log \mu_n(D) \leq -\inf_{y\in D} \Lambda^*(y).$$

Since the set D was chosen arbitrary, we conclude that the large deviations upper bound holds for all compact sets (with the rate function Λ^*). ■

Exercise 2.2.40

Let $(w_{t_1}, \ldots, w_{t_d})$ be samples of a Brownian motion at the fixed times $t_1, \ldots, t_d$; so the increments $\{w_{t_{j+1}-w_{t_j}}\}$ are zero-mean, independent Normal random variables of variances $\{t_{j+1} - t_j\}$, respectively. Let us find the rate function for the empirical mean $\hat{S}_n$ of $X_i = (w^i_{t_1}, \ldots, w^i_{t_d})$, where $w^i_{t_j},\ i = 1, \ldots, n$ are samples of independent Brownian motions at time instances t_j.

Solution:

Without loss of generality, we may assume that $(w_{t_1}, \ldots, w_{t_d})$ be samples of a **standard** Brownian motion at the fixed times $t_1, \ldots, t_d$. Therefore, we get that for each $X_i = (w^i_{t_1}, \ldots, w^i_{t_d})$ is a Gaussian vector in $\mathbb{R}^d$, and since $w^i_{t_j},\ i = 1, \ldots, n$ are samples of independent Brownian motions at time instances t_j, this implies that the collection $\{X_i\}$ are i.i.d. Gaussian vectors. That is, $X_i \sim N(0, C)$ for all i, and where C is the covariance matrix with entries

$$C_{jk} = \mathbb{E}[w_j w_k] - \mathbb{E}[w_j]\overbrace{\mathbb{E}[w_k]}^{=0} = \mathbb{E}[w_j(w_k - w_j)] + \mathbb{E}[w_j^2] \overset{\perp\!\!\!\perp}{=} \mathbb{E}[w_j]\overbrace{\mathbb{E}[w_k - w_j]}^{=0} + \mathbb{E}[w_j^2] = \ldots = t_j \wedge t_k,$$

which follows from the definition of a Brownian motion.

Therefore, we get that the law of $\hat{S}_n$ is simply

$$\hat{S}_n = \frac{1}{n}\sum_{i=1}^n X_i \overset{\text{i.i.d.}}{\sim} \frac{1}{n} N(0, nC) = N(0, \frac{n}{n^2}C) = N(0, \frac{1}{n}C),$$

where we used the properties of covariance. This implies that $\hat{S}_n$ has the same law as for $\frac{1}{\sqrt{n}}(w_{t_1}, \ldots, w_{t_d}) \sim \frac{1}{\sqrt{n}} N(0, C) = N(0, \frac{1}{n}C)$.

By Cramer's theorem in $\mathbb{R}^d$ (Theorem 2.2.30) [1] then the empricial mean $\hat{S}_n$ satisfy the LDP with a good rate function $\Lambda^*(x) = \sup_{\lambda\in\mathbb{R}^d}\{\langle\lambda, x\rangle - \Lambda(\lambda)\}$, where $\Lambda(\lambda) = \log M(\lambda) = \log \mathbb{E}[e^{\langle\lambda, X_1\rangle}]$. Since $X_1 \sim N(0, C)$, we get that the moment generating function is in this case

$$M(\lambda) = \mathbb{E}[e^{\langle\lambda, X_1\rangle}] = \frac{1}{(2\pi)^{d/2}\det(C)^{1/2}} \int_{\mathbb{R}} \exp\{\langle\lambda, x\rangle\} \exp\{-\frac{1}{2}x^T C^{-1} x\} dx$$

$$= \frac{1}{(2\pi)^{d/2}\det(C)^{1/2}} \int_{\mathbb{R}} \exp\{\lambda^T x - \frac{1}{2}x^T C^{-1} x\} dx$$
$$= \frac{1}{(2\pi)^{d/2}\det(C)^{1/2}} \int_{\mathbb{R}} \exp\{-\frac{1}{2}(x - C\lambda)^T C^{-1}(x - C\lambda) + \frac{1}{2}\lambda^T C\lambda\} dx$$
$$= \exp\{\frac{1}{2}\lambda^T C\lambda\} \frac{1}{(2\pi)^{d/2}\det(C)^{1/2}} \int_{\mathbb{R}} \exp\{-\frac{1}{2}(x - C\lambda)^T C^{-1}(x - C\lambda)\} dx$$
$$= \exp\{\frac{1}{2}\lambda^T C\lambda\} \frac{1}{(2\pi)^{d/2}\det(C)^{1/2}} \int_{\mathbb{R}} \exp\{-\frac{1}{2}u^T C^{-1} u\} du = \exp\{\frac{1}{2}\lambda^T C\lambda\} \cdot 1 = \exp\{\frac{1}{2}\lambda^T C\lambda\}.$$

Hence, we get that the cumulative generating function is in this case $\Lambda(\lambda) = \log M(\lambda) = \frac{1}{2}\lambda^T C\lambda$. By Lemma 2.2.31 part a) [1], we know that Λ^* is a convex rate function which means that $\sup_{\lambda\in\mathbb{R}^d}\{\langle\lambda, x\rangle - \Lambda(\lambda)\}$ is achieved in the zero value of gradient of $\langle\lambda, x\rangle - \Lambda(\lambda)$. Hence, we get that

$$\frac{d}{d\lambda}(\langle\lambda, x\rangle - \Lambda(\lambda)) = \frac{d}{d\lambda}(\lambda^T x - \frac{1}{2}\lambda^T C\lambda) = x - C\lambda = 0 \Rightarrow \lambda = C^{-1}x.$$

Therefore, the rate function of $\hat{S}_n$ is

$$\Lambda^*(x) = \sup_{\lambda\in\mathbb{R}^d}\{\langle\lambda, x\rangle - \Lambda(\lambda)\} = (C^{-1}x)^T x - \frac{1}{2}(C^{-1}x)^T C(C^{-1}x) = x^T C^{-1}x - \frac{1}{2}x^T C^{-1}x = \frac{1}{2}x^T C^{-1}x.$$

2.3 The Gärtner-Ellis theorem

Last sections wer considered Cramer's theorem which is limited to i.i.d. case [1]. In this section, we will study the extension of the Cramer's theorem to non-i.i.d. case which is known as the Gärtner-Ellis theorem [1]. Throughout, in this section we have assumption 2.3.2 from the book [1]

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n}\Lambda_n(n\lambda),$$

where $\Lambda_n(\lambda) = \log \mathbb{E}[e^{\langle \lambda, Z_n\rangle}]$, and $Z_n \in \mathbb{R}^d$ is a sequence of random vectors which possesses the law μ_n. Using this assumption, it can be shown that the existence of a limit of properly scaled logarithmic moment generating functions indicates μ_n may satisfy the LDP.

Exercise 2.3.18

a) Let us prove that if $\Lambda(\cdot)$ is a steep logarithmic moment generating function, then $\exp(\Lambda(\cdot)) - 1$ is also a steep function.

Proof.

Let us recall that the function is steep function, namely when $\lim_{n\to\infty} |\nabla\Lambda(\lambda_n)| = \infty$ whenever a sequence $\{\lambda_n\}$ in $\mathcal{D}_\Lambda^\circ$ converging to a boundary point of $\partial\mathcal{D}_\Lambda^\circ$ by definition 2.3.5 [1].

Let us denote $\Lambda_Z(\lambda) = \exp(\Lambda(\lambda)) - 1$, and let a sequence $\{\lambda_n\} \in \mathcal{D}_\Lambda^\circ$ be such that $\lambda_n \to \lambda \in \partial\mathcal{D}_\Lambda^\circ$. Then, since $\Lambda(\cdot)$ is lower semicontinuous which implies that $\liminf_{n\to\infty} e^{\Lambda(\lambda_n)} \geq e^{\Lambda(\lambda)}$, and therefore $\Lambda_Z(\lambda)$ is lower semicontinuous also. This means that $\Lambda_Z(\lambda)$ is well-behaved on $\mathcal{D}_\Lambda^\circ$, and by taking the gradient, and using chain rule $\nabla\Lambda_Z(\cdot) = e^{\Lambda(\cdot)}\nabla\Lambda(\cdot)$. Therefore, we have that

$$\lim_{n\to\infty} |\nabla\Lambda_Z(\lambda_n)| = \lim_{n\to\infty} e^{\Lambda(\lambda_n)}|\nabla\Lambda(\lambda_n)| \geq e^{\Lambda(\lambda)} \cdot \infty = \infty.$$

Here, $e^{\Lambda(\lambda)} > 0$ follows from the fact that $\Lambda(\lambda_n) > -\infty$ by Lemma 2.3.9 [1]. Hence, by lower semicontinuity $\Lambda(\lambda) > -\infty$.

Thus, by definition of steepness, we get that the function Λ_Z is a steep function. ■

b) Let X_j be $\mathbb{R}^d$-valued i.i.d. random variables with a steep logarithmic moment generating function Λ such that $0 \in \mathcal{D}_\Lambda^\circ$. Let $N(t)$ be a Poisson process of unit rate that is independent of the X_j variables, and consider the random variables

$$\hat{S}_n = \frac{1}{n}\sum_{j=1}^{N(n)} X_j.$$

Let μ_n denote the law of $\hat{S}_n$ and let us prove that μ_n satisfies the LDP, with the rate function being the Fenchel-Legendre transform of $e^{\Lambda(\lambda)} - 1$.

Proof.

We know that the empirical mean $\hat{S}_n$ can be written as

$$\hat{S}_n = \frac{1}{n}\sum_{j=1}^{N(n)} X_j = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_i} X_j = \frac{1}{n}\sum_{i=1}^{n} Z_i, \tag{14}$$

where $N_j \sim \text{Poi}(1)$ are i.i.d. random variables. By this given formula, we can define the rate function of $\hat{S}_n$, when the cumulative generating function is

$$\Lambda_Z(\lambda) = \log \mathbb{E}[e^{\langle \lambda, Z_i \rangle}] = \log \mathbb{E}[e^{\sum_{j=1}^{N_i} \langle \lambda, X_j \rangle}] \overset{\text{i.i.d.}}{=} \log \mathbb{E}[e^{\sum_{j=1}^{N_i} \langle \lambda, X_1 \rangle}] = \log \mathbb{E}[\mathbb{E}[e^{\langle \lambda, X_1 \rangle}]^{N_1} \mid N_1]$$
$$\overset{\text{i.i.d.}}{=} \log \mathbb{E}[\mathbb{E}[e^{\langle \lambda, X_1 \rangle}]^{N_1}] = \log \mathbb{E}[M(\lambda)^{N_1}] = \log \big(\sum_{k=0}^{\infty} M(\lambda)^k \cdot \mathbb{P}[N_1 = k] \big) = \log \big(\sum_{k=0}^{\infty} M(\lambda)^k \cdot \frac{1^k}{k!} e^{-1} \big)$$
$$= \log \big(e^{-1} \sum_{k=0}^{\infty} \frac{M(\lambda)^k}{k!} \big) = \log \big(e^{-1} e^{M(\lambda)} \big) = \log \big(e^{M(\lambda) - 1} \big) = M(\lambda) - 1 = e^{\Lambda(\lambda)} - 1,$$

where $M(\lambda) = \mathbb{E}[e^{\langle \lambda, X_1 \rangle}]$ is the moment generating function of X_1, and $\Lambda(\lambda) = \log \mathbb{E}[e^{\langle \lambda, X_1 \rangle}]$ is the cumulative generating function.

By part a), we know that $\Lambda_Z(\lambda) = e^{\Lambda(\lambda)} - 1$ is a steep function, and lower semicontinuous. In addition, we assumed that $0 \in \mathcal{D}_\Lambda^\circ$, and so by exercise 2.3.16 part b) [1], when checking that Λ_Z is differentiable at $\mathcal{D}_\Lambda^\circ$, we get by Gärtner-Ellis theorem (theorem 2.3.6. [1]) part c) that μ_n satisfies the LDP with the rate function being the Fenchel-Legendre transform of $e^{\Lambda(\lambda)} - 1$. That is,

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \{ \langle \lambda, x \rangle - (e^{\Lambda(\lambda)} - 1) \} = \sup_{\lambda \in \mathbb{R}^d} \{ \langle \lambda, x \rangle - e^{\Lambda(\lambda)} + 1 \}.$$

■

Exercise 2.3.23

Let $X_1, \dots, X_n, \dots$ be real-valued zero mean, stationary Gaussian process with covariance sequence $R_i = \mathbb{E}[X_n X_{n+i}]$. Suppose the process has a finite power P defined via $P = \lim_{n \to \infty} \sum_{i=-n}^{n} R_i \Big(1 - \frac{|i|}{n} \Big)$. Let μ_n be the law of the empirical mean $\hat{S}_n$ of the first n samples of this process. Let us prove that $\{\mu_n\}$ satisfy the LDP with the good rate function $\Lambda^*(x) = \frac{x^2}{2P}$.

Proof.

We have that the empirical mean $\hat{S}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ is the sum of real-valued, zero-mean stationary Gaussian process (values) with covariance sequence R_i. Therefore, the empirical mean itself is Gaussian process with mean

$$\mathbb{E}[\hat{S}_n] = \mathbb{E}[\frac{1}{n} \sum_{i=1}^{n} X_i] \overset{\text{lin.}}{=} \frac{1}{n} \sum_{i=1}^{n} \overbrace{\mathbb{E}[X_i]}^{=0} = 0,$$

and (co)variance

$$\text{var}(\hat{S}_n) = \mathbb{E}[(\frac{1}{n} \sum_{i=1}^{n} X_i)^2] \overset{\text{lin.}}{=} \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}[X_i X_j] = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} R_{|i-j|}$$
$$\overset{k=|i-j|}{=} \frac{1}{n^2} \sum_{k=-n+1}^{n-1} R_k (n - |k|) = \frac{1}{n} \sum_{k=-n+1}^{n-1} R_k (1 - \frac{|k|}{n}) = \frac{1}{n} \sum_{k=-n}^{n} R_k (1 - \frac{|k|}{n}),$$

where in the second step we used the linearity of the expectation, and re-indexing the sums when $k = |i-j|$, which gives us the factor $n - |k|$, and last one is due to terms $k = n$ and $k = -n$ being zero in the sum. Taking $n \to \infty$, and recalling that the process has a finite power,

we get that the variance for $\hat{S}_n$ will be approximately $\text{var}(\hat{S}_n) = \frac{1}{n}\sum_{k=-n}^{n} R_k(1-\frac{|k|}{n}) \approx \frac{P}{n}$ for large enough n. Hence, we have that $\hat{S}_n \sim N(0, \frac{P}{n})$ when $n \to \infty$.

Therefore, we get that the moment generating function for $\hat{S}_n$ for large enough n is

$$M_n(\lambda) = \mathbb{E}[e^{\lambda \hat{S}_n}] = \exp(\frac{1}{2}\lambda^2 \frac{P}{n}) = \exp(\frac{P}{2n}\lambda^2),$$

and hence the logarithmic generating function (from assumption 2.3.2) will be

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n}\Lambda_n(n\lambda) = \lim_{n\to\infty} \frac{1}{n} \log M_n(n\lambda) = \lim_{n\to\infty} \frac{1}{n} \log \exp(\frac{P}{2n}(n\lambda)^2) = \lim_{n\to\infty} \frac{P}{2}\lambda^2 = \frac{P}{2}\lambda^2.$$

Now, since $\Lambda(\cdot)$ is quadratic function, it is clearly (lower semi-)continuous, convex function which is differentiable with respect to λ. In addition, we have $\mathcal{D}_\Lambda^\circ = \{\lambda \in \mathbb{R} : \Lambda(\lambda) < \infty\}^\circ = (-\infty, \infty)^\circ = (-\infty, \infty) \neq \varnothing$, and $\lim_{n\to\infty} |\nabla\Lambda(\lambda_n)| = \lim_{|\lambda|\to\infty} |\nabla\Lambda(\lambda)| = \lim_{|\lambda|\to\infty} P|\lambda| = \infty$. Therefore, by definition 2.3.5., it is smooth, in addition (lower semi)continuous function. Hence, by Gärtner-Ellis theorem (theorem 2.3.6.) part c), $\{\mu_n\}$ satisfies the LDP with the good rate function

$$\Lambda^*(x) = \sup_{\lambda\in\mathbb{R}}\{\lambda x - \frac{P}{2}\lambda^2\} \overset{\lambda = x/P}{=} \frac{x}{P}\cdot x - \frac{P}{2}\cdot\frac{x^2}{P^2} = \frac{x^2}{P} - \frac{x^2}{2P} = \frac{x^2}{2P}.$$

■

Exercise 2.3.25

Let us suppose that $0 \in \mathcal{D}_\Lambda^\circ$ for $\Lambda(\lambda) = \limsup_{n\to\infty} n^{-1}\Lambda_n(n\lambda)$.

a) Let us show first that the Lemma 2.3.9. and then part a) of the Gärtner-Ellis theorem hold for under this weaker form of assumption 2.3.2 [1].

Proof.

In this case, we assume that $0 \in \mathcal{D}_\Lambda^\circ$ for $\Lambda(\lambda) = \limsup_{n\to\infty} n^{-1}\Lambda_n(n\lambda)$, that is, the only difference from the stronger assumption is that we do not necessarily know that $\Lambda(\lambda) = \lim_{n\to\infty} n^{-1}\Lambda_n(n\lambda)$, and the given limit in the proof of Lemma 2.3.9 [1] is only used to show that Λ is a convex function in part a) of the proof.

Therefore, it suffices only to show that in this case $\Lambda(\lambda) = \limsup_{n\to\infty} n^{-1}\Lambda_n(n\lambda)$ is a convex function and the rest follows the same way as in Lemma 2.3.9. We know that Λ_n are convex functions, and therefore $\Lambda_n(n\cdot)/n$ is convex as well. Taking the upper limit preservers convexity since

$$\begin{aligned}&\limsup_{n\to\infty}\{n^{-1}\Lambda(n(t\lambda_1 - (1-t)\lambda_2)\} \leq \limsup_{n\to\infty}\{tn^{-1}\Lambda(n\lambda_1) - (1-t)n^{-1}\Lambda(n\lambda_2)\}\\ &= \limsup_{n\to\infty}\{tn^{-1}\Lambda(n\lambda_1)\} - \limsup_{n\to\infty}\{(1-t)n^{-1}\Lambda(n\lambda_2)\}\end{aligned}$$

for every $t \in [0,1]$ and $\lambda_1, \lambda_2 \in \mathbb{R}^d$. Therefore, we get that Λ is a convex function. The rest of the Lemma 2.3.9. part a) proof follows samely.

The part b) of the Lemma 2.3.9. follows by replacing the limits with upper limits, and the inequalities made on the proof holds still.

Now, since Lemma 2.3.9. hold for weaker assumption, we get that the part a) of the Gärtner-Ellis theorem hold also for this weaker form of assumption 2.3.2. same way as in the book [1]. ■

b) Let us show that if $z = \nabla\Lambda(0)$ then $\mathbb{P}[|Z_n - z| \geq \delta] \to 0$ exponentially in n for fixed $\delta > 0$.

Proof.

Suppose $z = \nabla\Lambda(0)$. Then by part b) of the lemma 2.3.9. (which we can use due to part a) of the exercise) we get that $\Lambda^*(z) = \langle 0, z\rangle - \Lambda(0) = 0$, since $0 \in \mathcal{D}_\Lambda^\circ$, and $\Lambda(0) = \log \mathbb{E}[e^{\langle 0, Z_n\rangle}] = \log(1) = 0$. Now, by lemma 2.3.9. we note that Λ^* is nonnegative, so this directly implies that $\Lambda^*(x) > \Lambda^*(z)$ for all $x \neq z$ by strict convexity of Λ^*.

Therefore, we can apply the Gärtner-Ellis theorem part a), for the probability in detail to get that

$$\limsup_{n\to\infty} \mathbb{P}[|Z_n - z| \geq \delta] \leq \limsup_{n\to\infty} \exp(-n \cdot \inf_{|x-z|\geq\delta} \Lambda(x)) = 0,$$

for $\delta > 0$ since $\inf_{|x-z|\geq\delta} \Lambda(x)) > 0$ by above reasoning. Hence, we conclude that $\mathbb{P}[|Z_n - z| \geq \delta] \to 0$ exponentially in n for fixed $\delta > 0$. ■

2.4 Concentration inequalities

In this section, we will focus on concentration inequalities from an LDP perspective. Specifically, we will first examine concentration inequalities for bounded martingale differences which are useful in many applications when studying random processes. Secondly, we briefly also study Talagrand's concentration inequalities.

Throughout this section, we assume that $\Sigma \subset \mathbb{R}$ is a Polish space (a complete separable metric space), in order to avoid measurability concerns.

2.4.1 Inequalities for bounded martingale differences

Exercise 2.4.21

Let $B(u) = 2u^{-2}[(1+u)\log(1+u) - u]$ for $u > 0$.

a) Let us show that for any $x, v > 0$,

$$H(\frac{x+v}{1+v} \mid \frac{v}{1+v}) \geq \frac{x^2}{2v} B(\frac{x}{v}), \tag{15}$$

and hence by (2.4.9) will the imply that for any $z > 0$,

$$\mathbb{P}[\hat{S}_n \geq z] \leq \exp(-\frac{z^2}{2nv} B(\frac{z}{v})). \tag{16}$$

Proof.

We aim to show that (15), which implies that (16). By using the definition of H, we get

$$\begin{aligned} H(\frac{x+v}{1+v} \mid \frac{v}{1+v}) &= \frac{x+v}{1+v}\log(\frac{x+v}{1+v}/\frac{v}{1+v}) + (1 - \frac{x+v}{1+v})\log((1-\frac{x+v}{1+v})/(1-\frac{v}{1+v})) \\ &= \frac{v+x}{1+v}\log(1+\frac{x}{v}) + \frac{1-x}{1+v}\log(1-x) = \frac{1}{1+v}[(v+x)\log(1+\frac{x}{v}) + (1-x)\log(1-x)] \end{aligned}$$

$$\geq \frac{1}{1+v}[(v+x)\log(1+\frac{x}{v}) - x] = \frac{1}{1+v}\cdot\frac{x^2}{2v}B(\frac{x}{v}),$$

where the inequality followed from the fact that $(1-x)\log(1-x) \geq -x$ when $x \in (0,1)$.

Here, the given inequality, $H(\frac{x+v}{1+v} \mid \frac{v}{1+v}) \geq \frac{x^2}{2v}B(\frac{x}{v})$ for any $x, v > 0$, is not achieved. The given inequality could perhaps be achieved by using convexity or some other algebraic manipulation, but I haven't been able to achieve the desired inequality after many tries.

Hence, we get that the inequality (2.4.9) [1] implies that for any $z > 0$

$$\mathbb{P}[\hat{S} \geq z] \leq \exp(-nH(\frac{z+v}{1+v} \mid \frac{v}{1+v})) \leq \exp(-\frac{1}{1+v}\cdot\frac{nz^2}{2v}B(\frac{z}{v}))$$

■

b) Suppose that $(S_n, \mathcal{F}_n)$ is a discrete time martingale such that $S_0 = 0$ and $Y_k = S_k - S_{k-1} \leq 1$ almost surely. Let $Q_n = \sum_{j=1}^{n}\mathbb{E}[Y_j^2 \mid \mathcal{F}_{j-1}]$ and show that for any $z, r > 0$,

$$\mathbb{P}[S_n \geq z, Q_n \leq r] \leq \exp(-\frac{z^2}{2r}B(\frac{z}{r})). \tag{17}$$

Proof.

Let us first check that $X_n = \exp(\lambda S_n - \theta Q_n)$ is super-martingale of filtration $\mathcal{F}_n$ when $\theta = e^\lambda - \lambda - 1 \geq \frac{\lambda^2}{2}$. Using the martingale property, we get that

$$\begin{aligned}
&\mathbb{E}[X_n \mid \mathcal{F}_{n-1}] = \mathbb{E}[\exp(\lambda S_n - \theta Q_n) \mid \mathcal{F}_{n-1}] \\
&= \mathbb{E}[\exp(\lambda S_{n-1} + \lambda Y_n - \theta\mathbb{E}[Y_n^2 \mid \mathcal{F}_{n-1}] - \theta Q_{n-1}) \mid \mathcal{F}_{n-1}] \\
&= \mathbb{E}[\exp((\lambda S_{n-1} - \theta Q_{n-1})\exp(\lambda Y_n - \theta\mathbb{E}[Y_n^2 \mid \mathcal{F}_{n-1}]) \mid \mathcal{F}_{n-1}] \\
&= \mathbb{E}[X_{n-1}\exp(\lambda Y_n - \theta\mathbb{E}[Y_n^2 \mid \mathcal{F}_{n-1}]) \mid \mathcal{F}_{n-1}] = X_{n-1}\mathbb{E}[\exp(\lambda Y_n - \theta\mathbb{E}[Y_n^2 \mid \mathcal{F}_{n-1}]) \mid \mathcal{F}_{n-1}] \\
&= X_{n-1}\exp(-\theta\mathbb{E}[Y_n^2 \mid \mathcal{F}_{n-1}])\mathbb{E}[\exp(\lambda Y_n) \mid \mathcal{F}_{n-1}],
\end{aligned}$$

where we used the fact that X_{n-1} by construction is $\mathcal{F}_{n-1}$ measurable so it can be taken out by properties of conditional expectation, and same followed for the term $\exp(-\theta\mathbb{E}[Y_n^2 \mid \mathcal{F}_{n-1}])$. Now, using Taylor expansion of $e^{\lambda x} = 1 + \lambda x + \sum_{k=1}^{\infty}\frac{\lambda^k}{k!}x^k$ for the conditional expectation yields

$$\mathbb{E}[\exp(\lambda Y_n) \mid \mathcal{F}_{n-1}] = 1 + \lambda\overbrace{\mathbb{E}[Y_n \mid \mathcal{F}_{n-1}]}^{=0} + \sum_{k=2}^{\infty}\frac{\lambda^k}{k!}\mathbb{E}[Y_n^k \mid \mathcal{F}_{n-1}] = 1 + \sum_{k=2}^{\infty}\frac{\lambda^k}{k!}\mathbb{E}[Y_n^k \mid \mathcal{F}_{n-1}],$$

where the $\mathbb{E}[Y_n \mid \mathcal{F}_{n-1}] = 0$ since S_n is a discrete time martingale which implies $\mathbb{E}[Y_n \mid \mathcal{F}_{n-1}] = \mathbb{E}[S_n - S_{n-1} \mid \mathcal{F}_{n-1}] \overset{\text{lin.}}{=} \mathbb{E}[S_n \mid \mathcal{F}_{n-1}] - S_{n-1} = S_{n-1} - S_{n-1} = 0$. Now, we also assumed that $Y_n \leq 1$ almost surely. This implies that $Y_n^k \leq Y_n^2$ for $k \geq 2$, and therefore we have

$$\begin{aligned}
&\mathbb{E}[\exp(\lambda Y_n) \mid \mathcal{F}_{n-1}] = 1 + \sum_{k=2}^{\infty}\frac{\lambda^k}{k!}\mathbb{E}[Y_n^k \mid \mathcal{F}_{n-1}] \leq 1 + \mathbb{E}[Y_n^2 \mid \mathcal{F}_{n-1}]\sum_{k=2}^{\infty}\frac{\lambda^k}{k!} \\
&= 1 + \mathbb{E}[Y_n^2 \mid \mathcal{F}_{n-1}]\cdot(e^\lambda - \lambda - 1) = 1 + \theta\mathbb{E}[Y_n^2 \mid \mathcal{F}_{n-1}] \leq \exp(\theta\mathbb{E}[Y_n^2 \mid \mathcal{F}_{n-1}]),
\end{aligned}$$

where we used the known inequality $1 + x \leq e^x$.

Combining this with first derivation, we get wanted property since

$$\mathbb{E}[X_n \mid \mathcal{F}_{n-1}] = X_{n-1}\exp(-\theta\mathbb{E}[Y_n^2 \mid \mathcal{F}_{n-1}])\mathbb{E}[\exp(\lambda Y_n) \mid \mathcal{F}_{n-1}]$$
$$\leq X_{n-1}\exp(-\theta\mathbb{E}[Y_n^2 \mid \mathcal{F}_{n-1}])\exp(\theta\mathbb{E}[Y_n^2 \mid \mathcal{F}_{n-1}]) = X_{n-1}$$

Therefore, we conclude that X_n is a supermartingale. The supermartingale property implies then $\mathbb{E}[X_n] \leq X_0 = \exp(\lambda S_0 + \theta Q_0) = \exp(\lambda \cdot 0 + \theta \cdot 0) = 1$. Using this observations, and the Markov property then yields

$$\mathbb{P}[S_n \geq z, Q_n \leq r] = \mathbb{P}[\lambda S_n \geq \lambda z, \theta Q_n \leq \theta r] \leq \mathbb{P}[\lambda S_n - \theta Q_n \geq \lambda z - \theta r]$$
$$= \mathbb{P}[\overbrace{\exp(\lambda S_n - \theta Q_n)}^{=X_n} \geq \exp(\lambda z - \theta r)] \overset{\text{Markov.}}{\leq} \exp(-\lambda z + \theta r)\mathbb{E}[X_n] \leq \exp(-\lambda z + \theta r),$$

where we used the monotonicity of probability measure for $\{\lambda S_n \geq \lambda z, \theta Q_n \leq \theta r\} \subset \{\lambda S_n - \theta Q_n \geq \lambda z - \theta r\}$ when $\lambda, \theta \geq \frac{\lambda^2}{2} > 0$. Hence, it suffices to check in which values of λ, we get the tightest bound. This can be done by differentiation

$$\frac{d}{d\lambda}(-\lambda z + (e^\lambda - \lambda - 1)r) = -z + re^\lambda - r = 0 \Rightarrow \lambda = \log(1 + \frac{z}{r}),$$

and therefore the tightest bound will be

$$\mathbb{P}[S_n \geq z, Q_n \leq r] \leq \exp(-\lambda z + \theta r) \leq \exp(-\log(1 + \frac{z}{r})z + [(1 + \frac{z}{r}) - \log(1 + \frac{z}{r}) - 1]r)$$
$$= \exp(-((r + z)\log(1 + \frac{z}{r}) - z)) = \exp(-r[(1 + \frac{z}{r})\log(1 + \frac{z}{r}) - \frac{z}{r}])$$
$$= \exp(-r\frac{z^2}{2r^2} \cdot 2(\frac{z}{r})^{-2}[(1 + \frac{z}{r})\log(1 + \frac{z}{r}) - \frac{z}{r}]) = \exp(-\frac{z^2}{2r}B(\frac{z}{r})),$$

where we used the definition of B. Therefore, we have shown the wanted result. ■

2.4.2 Talagrand's concentration inequalities

Exercise 2.4.43

Let us fix $a > 0$ and $A \in \mathcal{B}_{\Sigma^n}$. For any $x \in \Sigma^n$ let $V_A(x)$ denote the closed convex hull of $\{(\mathbf{1}_{y_1=x_1}, \ldots, \mathbf{1}_{y_n=x_n}) : y \in A\} \subset [0,1]^n$. Let us show that the measurable function $f_a(A, \cdot)$ of (2.4.25) [1] can be represented also as

$$f_a(A, x) = \inf_{s \in V_A(x)} \sum_{k=1}^n \phi_a(s_k). \tag{18}$$

Proof.

By equation (2.4.25) from the book [1], we get that the measurable function $f_a(A, \cdot)$ is defined as

$$f_a(A, x) = \inf_{\{\nu \in M_1(\Sigma^n) : \nu(A) = 1\}} \sum_{k=1}^n \phi_a(\nu(\{y : y_k = x_k\}))$$

and we aim to show that it can be defined also as (18).

Based on the above equation and (18), it only suffices to show that the infimum of $\{\nu \in M_1(\Sigma^n) : \nu(A) = 1\}$ is the same as for $V_A(x)$ when considering $s_k = \nu(\{y : y_k = x_k\})$. That is, in this case let us aim to show that $V_A(x) = S(A)$, where $S(A) := \{\nu \in M_1(\Sigma^n) :$

$\nu(A) = 1, s_k = \nu(\{y : y_k = x_k\})$ by showing both inclusions. First, note that here $V_A(x) = \text{Conv}(\{(\mathbf{1}_{y_1=x_1}, \ldots, \mathbf{1}_{y_n=x_n}) : y \in A\})$.

$\mathbf{S(A)} \subset \mathbf{V_A(x)}$:

Let $s \in S(A)$, that is, by definition there exists probability measure $\nu \in M_1(\Sigma^n)$ such that $\nu(A) = 1$, and the kth element of vector s is $s_k = \nu(\{y : y_k = x_k\}) = \mathbb{E}_\nu[\mathbf{1}_{x_k=y_k}]$, where $\mathbb{E}_\nu$ is the notation that we are taking expectation with respect to probability measure ν.

Since we have that $\nu(A) = 1$, ν is supported on A. Therefore, the given vector s can be written as expectation of a random vector $V(y) = (\mathbf{1}_{y_1=x_1}, \ldots, \mathbf{1}_{y_k=x_k}) \in \{0,1\}^n$, where $y \in A$ as

$$s = \mathbb{E}_\nu[V] = \int_{\Sigma^n} V(y)\nu(dy) = \int_A V(y)\nu(dy) \tag{19}$$

Clearly, we have that random variable $V(y)$ belongs to the convex hull of vectors $\{(\mathbf{1}_{y_1=x_1}, \ldots, \mathbf{1}_{y_n=x_n}) : y \in A\}$, that is, $V(y) \in V_A(x)$. Since s is taken as a convex combination of vectors of $V(y)$ where $y \in A$ by (19), we have that by definition of convex hull $s \in V_A(x)$.

$\mathbf{S(A)} \supset \mathbf{V_A(x)}$:

Let $s \in V_A(x)$, that is, we get that s can be represented as a convex combination of some points from the set $\{(\mathbf{1}_{y_1=x_1}, \ldots, \mathbf{1}_{y_n=x_n}) : y \in A\}$. That is, by definition of convexity of sets

$$s = \sum_{j=1}^n a_j v^{(j)},$$

where $a_j \geq 0$ for all j, $\sum_{j=1}^n a_j = 1$, and where $v^{(j)} \in \{(\mathbf{1}_{y_1^{(j)}=x_1}, \ldots, \mathbf{1}_{y_n^{(j)}=x_n}) : y \in A\}$. Then, if we contruct a measure ν as a discrete measure supported on the finite set $\{y^{(1)}, \ldots, y^{(n)}\} \subset A$ such that

$$\nu(x) = \sum_{j=1}^n a_j \delta_{y^{(j)}}(x) = \sum_{j=1}^n a_j \mathbf{1}_x(y^{(j)}),$$

we have that ν is probability measure. This is, since $\nu(\Sigma^n) = \sum_{j=1}^n a_j \mathbf{1}_{\Sigma^n}(y^{(j)}) = \sum_{j=1}^n a_j = 1$. Moreover, since $\{y^{(1)}, \ldots, y^{(n)}\} \subset A$, we have $\nu(A) = 1$, and the elements of s can be written as

$$s_k = \sum_{j=1}^n a_j v_k^{(j)} = \sum_{j=1}^n a_j \mathbf{1}_{y_k^{(j)}=x_k} = \sum_{j=1}^n a_j \mathbf{1}_{x_k}(y_k^{(j)}) = \nu(\{y : y_k = x_k\}).$$

Thus, by definition we have shown that $s \in S(A)$.

Since both inclusions hold, we conclude that $S(A) = V_A(x)$, which asserts that equation (2.4.25) [1] can be written as (18). ■

3 Applications - The finite dimensional case

In this chapter we will go through some application regarding of the theory represented in chapter 2. That is, we will study the applications of large deviation principles (mostly in finite state space) for example computing some interesting aspects for Markov chains.

3.1 Large Deviations for Finite State Markov Chains

The results of Section 2.1 are extended in this section to Markov chains $Y_!, \ldots, Y_n, \ldots$ taking values in a finite alphabet Σ, where without loss of generality Σ is identified with the set $\{1, \ldots, N\}$. Also, we note that $\Pi = \{\pi(i,j)\}_{i,j=1}^{|\Sigma|}$ be a stochastic matrix, and let $\mathbb{P}_\sigma^\pi$ denote the Markov probability measure associated with the transition probability Π and with the initial state σ. In short,

$$\mathbb{P}_\sigma^\pi[Y_1 = y_1, \ldots, Y_n = y_n] = \pi(\sigma, y_1) \prod_{i=1}^{n-1} \pi(y_i, y_{i+1}).$$

Furthermore, the expectations with respect to $\mathbb{P}_\sigma^\pi$ are denoted $\mathbb{E}_\sigma^\pi$.

3.1.1 LDP for Additive functionals of Markov Chains

In this subsection, the subject is to study the large deviations of the empirical means

$$Z_n = \frac{1}{n} \sum_{k=1}^{n} X_k,$$

where $X_k = f(Y_k)$ and $f : \Sigma \to \mathbb{R}^d$ is a given deterministic function but they can be extended to random functions as we will do in next exercise.

Exercise 3.1.4

Let us assume that $Y_1, \ldots, Y_n$ are distributed according to the joint law $\mathbb{P}_\sigma^\pi$ determined by the irreducible stochastic matrix Π. Let the conditional law of $\{X_k\}$ for each realization $\{Y_k = j_k\}_{k=1}^n$ be the product of the measures $\mu_j \in M_1(\mathbb{R}^d)$, i.e., the variables X_k are conditionally independent. Suppose that the logarithmic moment generating functions Λ_j associated with μ_j are finite everywhere (for all $j \in \Sigma$). Let us consider the empirical mean $Z_n = \frac{1}{n} \sum_{k=1}^n X_k$, and let us prove that Theorem 3.1.2 [1] holds for Borel measurable sets Γ with

$$\pi_\lambda(i,j) = \pi(i,j) e^{\Lambda_j(\lambda)}, \quad i,j \in \Sigma.$$

Proof.

We aim to show that Theorem 3.1.2 holds for Borel measurable sets Γ with $\pi_\lambda(i,j) = \pi(i,j)e^{\Lambda_j(\lambda)}, \quad i,j \in \Sigma$ with the given construction.

Let us define

$$\Lambda_n(\lambda) = \log \mathbb{E}[e^{\langle \lambda, Z_n \rangle}] = \log \mathbb{E}_\sigma^\pi[\mathbb{E}[e^{\langle \lambda, Z_n \rangle} \mid Y_1 = j_1, \ldots, Y_n = j_n]],$$

where we used the unbiasedness of conditional expectation since we know that $X_k = f(Y_k)$ where f is not necessarily deterministic function in this case, we take conditional expectation with respect to $Y_1, \ldots, Y_n$ which are distributed according to the joint law $\mathbb{P}_\sigma^\pi$ determined by the irreducible stochastic matrix Π.

Now, as in the proof of theorem 3.1.2, in view of the Gärtner-Ellis theorem (theorem 2.3.6 [1]), it is enough to check that limit

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n}\Lambda_n(n\lambda) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_\sigma^\pi[\mathbb{E}[e^{n\langle \lambda, Z_n\rangle} \mid Y_1 = j_1, \dots, Y_n = j_n]]$$

exists for every $\lambda \in \mathbb{R}^d$, that $\Lambda(\cdot)$ is finite and differentiable everywhere in $\mathbb{R}^d$, and that $\Lambda(\lambda) = \log \rho(\Pi_\lambda)$, where $\rho(\Pi_\lambda)$ is the Perron-Frobenius eigenvalue of the matrix Π_λ.

To begin with, note that

$$\Lambda_n(n\lambda) = \log \mathbb{E}_\sigma^\pi[\mathbb{E}[e^{n\langle \lambda, Z_n\rangle} \mid Y_1 = j_1, \dots, Y_n = j_n]] = \log \mathbb{E}_\sigma^\pi[\mathbb{E}[e^{\langle \lambda, \sum_{k=1}^n X_k\rangle} \mid Y_1 = j_1, \dots, Y_n = j_n]]$$

$$= \log \mathbb{E}_\sigma^\pi[\mathbb{E}[e^{\sum_{k=1}^n \langle \lambda, X_k\rangle} \mid Y_1 = j_1, \dots, Y_n = j_n]] \overset{\perp\!\!\!\perp}{=} \log \mathbb{E}_\sigma^\pi[\prod_{k=1}^n \mathbb{E}[e^{\langle \lambda, X_k\rangle} \mid Y_k = j_k]]$$

$$= \log \mathbb{E}_\sigma^\pi[\prod_{k=1}^n \int_{\mathbb{R}^d} e^{\langle \lambda, x\rangle} \mu_{j_k}(dx)] = \log \mathbb{E}_\sigma^\pi[\prod_{k=1}^n \exp\big(\Lambda_{j_k}(\lambda)\big)],$$

where we used fact that the conditional law of $\{X_k\}$ for each realization $\{Y_k = j_k\}_{k=1}^n$ is the product of the measures $\mu_j \in M_1(\mathbb{R}^d)$ (i.e., the conditional independence), and the definitions of logarithmic generating functions Λ_{j_k} associated with μ_{j_k}. By opening the expectation with respect to $\{Y_k\}$ a finite state Markov Chain with irreducible transition matrix Π yields

$$\begin{aligned}
\Lambda_n(n\lambda) &= \log \mathbb{E}_\sigma^\pi[\prod_{k=1}^n \exp\big(\Lambda_{j_k}(\lambda)\big)] = \log \sum_{j_1,\dots,j_k} \mathbb{P}_\sigma^\pi[Y_1 = j_1, \dots, Y_n = j_n] \prod_{k=1}^n \exp(\Lambda_{j_k}(\lambda)) \\
&= \log \sum_{j_1,\dots,j_k} \pi(\sigma, j_1) e^{\Lambda_{j_1}(\lambda)} \cdots \pi(j_{n-1}, j_n) e^{\Lambda_{j_n}(\lambda)} = \log \sum_{j_1,\dots,j_k} \pi_\lambda(\sigma, j_1) \cdots \pi_\lambda(j_{n-1}, j_n) \\
&= \log \sum_{j_n=1}^{|\Sigma|} (\Pi_\lambda)^n(\sigma, j_n),
\end{aligned}$$

where Π_λ is a nonnegative matrix whose elements are $\pi_\lambda(i,j) = \pi(i,j)e^{\Lambda_j(\lambda)}, \quad i,j \in \Sigma$. Since Π was assumed to be irreducible stochastic matrix, by the definition we get that Π_λ is irreducible stochastic matrix because we are multiplying all elements of Π of positive values $e^{\Lambda_j(\lambda)}$. Therefore, part e) of the Perron-Frobenius theorem (theorem 3.1.1) yields (with $\phi = (1, \dots, 1)$)

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n}\Lambda_n(n\lambda) = \log \rho(\Pi_\lambda).$$

Moreover, as in proof of theorem 3.1.2, since we know that $|\Sigma|$ is finite, $\rho(\Pi_\lambda)$, being an isolated root of the characteristic equation for the matrix Π_λ, is positive, finite and differentiable with respect to λ.

Hence, we conclude that Theorem 3.1.2 holds for Borel measurable sets Γ with $\pi_\lambda(i,j) = \pi(i,j)e^{\Lambda_j(\lambda)}, \quad i,j \in \Sigma$. ■

3.1.2 Sanov's Theorem for the Empiricial Measure of Markov chains

In this subsection, we will consider the LDP satisfied empirical measures $L_n^{\mathbf{Y}} = (L_n^{\mathbf{Y}}(1), \dots, L_n^{\mathbf{Y}}(|\Sigma|))$ of Markov chains. Here, $L_n^{\mathbf{Y}}$ will denote the vector of frequencies in which the Markov chain visits the different states, namely,

$$L_n^{\mathbf{Y}}(i) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_i(Y_k), \quad i = 1, \dots, |\Sigma|.$$

In addition to previous section, if we take $f(y) = (\mathbf{1}_1(y), \dots, \mathbf{1}_{|\Sigma|}(y))$, then by theorem 3.1.2 [1], the LDP holds for $\{L_n^{\mathbf{Y}}\}$ with rate function

$$I(q) = \sup_{\lambda \in \mathbb{R}^d} \{\langle \lambda, q\rangle - \log \rho(\Pi_\lambda)\},$$

where $\pi_\lambda(i,j) = \pi(i,j)e^{\lambda_j}$. Theorem 3.1.6 [1] gives alternative characterization for rate function which may be more useful. We will focus on that in the next exercises.

Exercise 3.1.8

Suppose that for every $i, j \in \Sigma, \pi(i,j) = \mu(j)$, where $\mu \in M_1(\Sigma)$. Let us show that $J(\cdot) = H(\cdot \mid \mu)$ (the relative entropy with respect to μ), and that $I(\cdot)$ is the Fenchel-Legendre transform of $\log[\sum_j e^{\lambda_j}\mu(j)]$. Thus theorem 3.1.6. is a natural extension of Exercise 2.2.36 to the Markov setup [1].

Proof.

We have that for every $i, j \in \Sigma, \pi(i,j) = \mu(j)$, where $\mu \in M_1(\Sigma)$, and by theorem 3.1.6

$$J(q) = \begin{cases} \sup_{u>>0} \sum_{j=1}^{|\Sigma|} q_j \log\left[\frac{u_j}{(u\Pi)_j}\right], & q \in M_1(\Sigma) \\ \infty, & q \notin M_1(\Sigma) \end{cases}.$$

We aim to show that $J(q) = H(q \mid \mu) = \sum_{i=1}^{|\Sigma|} q(j)\log(\frac{q(j)}{\mu(j)})$ for $\mu \in M_1(\Sigma)$ and arbitrarily $q \in \mathbb{R}^{|\Sigma|}$. Hence, first suppose that $q \in M_1(\Sigma)$. Since for every $i, j \in \Sigma, \pi(i,j) = \mu(j)$, from the definition $J(q)$ we note that

$$(u\Pi)_j = \sum_{i=1}^{|\Sigma|} u(i)\pi(i,j) = \sum_{i=1}^{|\Sigma|} u(i)\mu(j) = \mu(j)\sum_{i=1}^{|\Sigma|} u(i)$$

and therefore we have that

$$J(q) = \sup_{u>>0} \sum_{j=1}^{|\Sigma|} q(j)\log\left[\frac{u_j}{(u\Pi)_j}\right] = \sup_{u>>0} \sum_{j=1}^{|\Sigma|} q(j)[\log\left(\frac{u_j}{\mu(j)}\right) - \log\left(\sum_{i=1}^{|\Sigma|} u_i\right)]$$
$$= \sup_{u>>0} \sum_{j=1}^{|\Sigma|} q(j)\log\left(\frac{u_j}{\mu(j)}\right) - \sum_{j=1}^{|\Sigma|} q(j)\log\left(\sum_{i=1}^{|\Sigma|} u_i\right) = \sup_{u>>0} \sum_{j=1}^{|\Sigma|} q(j)\log\left(\frac{u_j}{\mu(j)}\right) - \log\left(\sum_{i=1}^{|\Sigma|} u_i\right),$$

where the last equality follows from the fact that $\sum_{j=1}^{|\Sigma|} q(j) = 1$. Now, we may find the supremum by computing the difference, since by construction the last term can be shown

to be strictly concave over the domain $u >> 0$ (due to logarithm functions). Therefore, we get that

$$\frac{d}{du_k}\Big[\sum_{j=1}^{|\Sigma|} q(j)\log\Big(\frac{u_j}{\mu(j)}\Big) - \log\Big(\sum_{i=1}^{|\Sigma|} u_i\Big)\Big] = \frac{q(k)}{u_k} - \frac{1}{\sum_{i=1}^{|\Sigma|} u_i} = 0 \Leftrightarrow u_k = q(k)\sum_{i=1}^{|\Sigma|} u_i \Rightarrow u_k = q(k),$$

and the term $J(q)$ yields then

$$\begin{aligned} J(q) &= \sup_{u>>0}\sum_{j=1}^{|\Sigma|} q(j)\log\Big(\frac{u_j}{\mu(j)}\Big) - \log\Big(\sum_{i=1}^{|\Sigma|} u_i\Big) = \sum_{j=1}^{|\Sigma|} q(j)\log\Big(\frac{q(j)}{\mu(j)}\Big) - \log\Big(\overbrace{\sum_{i=1}^{|\Sigma|} q(i)}^{=1}\Big) \\ &= \sum_{j=1}^{|\Sigma|} q(j)\log\Big(\frac{q(j)}{\mu(j)}\Big) = H(q \mid \mu). \end{aligned}$$

Now, if $q \notin M_1(\Sigma)$, by remark of the relative entropy definition (definition 2.1.5) we have that $H(q \mid \mu) = \infty = J(q)$ in this case. Therefore, we may conclude that $J(\cdot) = H(\cdot \mid \mu)$. By Cramer's theorem (theorem 2.1.24) we know that for $I(q) = H(q \mid \mu)$ the rate function can be defined as

$$I(q) = \sup_{\lambda\in\mathbb{R}^d}\{\langle\lambda, x\rangle - \Lambda(\lambda)\},$$

where $\Lambda(\lambda) = \log\sum_{j=1}^{|\Sigma|}\mu(j)e^{\lambda_j}$ in this case. Hence, theorem 3.1.6 is a natural extension of Exercise 2.2.36 to the Markov setup. ■

Exercise 3.1.9

a) Let us show that the relation in theorem 3.1.6 holds for any nonnegative irreducible matrix $B = \{b(i,j)\}$ (not necessarily stochastic).

Proof.

Let $\phi(i) = \sum_j b(i,j) > 0$ since for every $i, j \in \Sigma$ we have $b(i,j) > 0$. Then the matrix Π with entries $\pi(i,j) = \dfrac{b(i,j)}{\phi(i)}$ is stochastic since the row sums add up to one. Let I_B and J_B denote the rate functions I and J associated with the matrix B via (3.1.5) and (3.1.7) [1], respectively. Then by theorem 3.1.6, J_Π and J_B has a relation

$$\begin{aligned} J_\Pi(q) &= \sup_{u>>0}\sum_{j=1}^{|\Sigma|} q(j)\log\Big[\frac{u_j}{(u\Pi)_j}\Big] = \sup_{u>>0}\sum_{j=1}^{|\Sigma|} q(j)\log\Big[\frac{u_j}{(uB)_j/\phi(j)}\Big] \\ &= \sup_{u>>0}\sum_{j=1}^{|\Sigma|} q(j)\log\Big[\frac{u_j}{(uB)_j}\Big] + \sum_{j=1}^{|\Sigma|} q(j)\log(\phi(j)) = J_B(q) + \sum_{j=1}^{|\Sigma|} q(j)\log(\phi(j)), \end{aligned}$$

since $(u\Pi)_j = \sum_{i=1} u_i\pi(i,j) = \sum_i \frac{u_i}{\phi(i)}\pi(i,j) = (\dfrac{u}{\phi}\cdot B)_j = \dfrac{(u\cdot B)_j}{\phi(j)}$.

Similarly, for rate function I_Π we get

$$I_\Pi(q) = \sup_{\lambda\in\mathbb{R}^d}\{\langle\lambda, q\rangle - \log\rho(\Pi_\lambda)\} = \sup_{\lambda\in\mathbb{R}^d}\{\langle\lambda, q\rangle - \log\rho(B_\lambda) - \sum_{i=1}^{|\Sigma|} q(j)\log(\phi(j)^{-1})\}$$

$$= \sup_{\lambda\in\mathbb{R}^d}\{\langle\lambda,q\rangle - \log\rho(B_\lambda)\} + \sum_{i=1}^{|\Sigma|} q(j)\log(\phi(j)) = I_B(q) + \sum_{i=1}^{|\Sigma|} q(j)\log(\phi(j)),$$

since $\Pi_\lambda(i,j) = \pi(i,j)e^{\lambda_j} = \dfrac{b(i,j)}{\phi(i)}e^{\lambda_j} = \dfrac{1}{\phi(i)}b(i,j)e^{\lambda_j} = \dfrac{1}{\phi(i)}B_\lambda(i,j)$, the eigen vectors of Π_λ are scaled with the factors $\sum_{i=1}^{|\Sigma|} q(j)\log(\phi(j))$.

Using theorem 3.1.6 for Π, and above relations, we obtain that $J_B(q) = I_B(q)$ which we sought to proof. ■

b) Let us show that for any irreducible, nonnegative matrix B,

$$\log\rho(B) = \sup_{\nu\in M_1(\Sigma)}\{-J_B(\nu)\}.$$

Proof.

We know that rate function I_B is defined by the Fenchel-Legendre transform as

$$I_B(\nu) = \sup_{\lambda\in\mathbb{R}^d}\{\langle\lambda,\nu\rangle - \log\rho(B_\lambda)\} \Leftrightarrow \log\rho(B_\lambda) = \sup_{\nu\in M_1(z)}\{\langle\lambda,\nu\rangle - I_B(\nu)\}$$

where the equivalence comes from the duality. By part a), we know that $I_B = J_B$, and if we choose $\lambda = 0$, then $B_\lambda(i,j) = B(i,j)e^{\overbrace{\lambda_j}^{=0}} = B(i,j)$. That is, $B_\lambda = B$. Therefore, using part a) result with $\lambda = 0$ yields

$$\log\rho(B) = \sup_{\nu\in M_1(z)}\{\langle 0,\nu\rangle - I_B(\nu)\} = \sup_{\nu\in M_1(z)}\{-I_B(\nu)\} \overset{a)}{=} \sup_{\nu\in M_1(z)}\{-J_B(\nu)\}$$

which we sought to proof. ■

3.1.3 Sanov's Theorem for the Pair Empiricial Measure of Markov chains

According to Demob and Zeitouni, the rate function governing the LDP for the empirical measure of a Markov chain is still in the form of an optimization problem [1]. Hence, by obtaining appropriate LDP using Sanov's theorem for Markov chains, we have to consider the Pair Empricical measure for Markov chains. That is, in this section, consider $\Sigma^2 = \Sigma\times\Sigma$ which corresponds to consecutive pairs of elements from the sequence $\mathbf{Y}$. Then a Markov chain is recovered with a state space Σ^2 and transition matrix $\Pi^{(2)}$ via

$$\pi^{(2)}(k\times l, i\times j) = \mathbf{1}_l(i)\pi(i,j).$$

For simplicity, it is assumed in this section that Π is strictly positive which implies that $\Pi^{(2)}$ is an irreducible transition matrix. Thus, the results obtained in Section 3.1.2 may be applied to yield the rate function $I_2(q)$ with the large deviations of the pair empirical measure.

$$L_{n,2}^{\mathbf{Y}}(y) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_y(Y_{i-1}Y_i) \quad y\in\Sigma^2 \tag{20}$$

[1].

We will study the Sanov's theorem for Markov chains using the Pair Empirical Measure $L_{n,2}$ with good rate function $I_2(q)$ with the help of theorem 3.1.13 in next exercises [1].

Exercise 3.1.17

Let us prove that for any strictly positive stochastic matrix Π

$$J(\nu) = \inf_{\{q:q_2=\nu\}} I_2(q), \tag{21}$$

where $J(\cdot)$ is the rate function defined in (3.1.7), while $I_2(\cdot)$ is a specified in theorem 3.1.13 [1].

Proof.

Let $Y_0, Y_1, \dots$ be the Markov chain with the transition matrix Π started at $Y_0 = \sigma$, and let

$$L_n^Y(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_y(Y_k), y = 1, \dots, |\Sigma|$$

that is, L_n^Y is the empirical measure on Σ and $L_{n,2}^Y$ on Σ^2 as stated in (20). Now, the provided hint notes that for $Y_0 = \sigma$

$$L_n^Y \in A \Leftrightarrow L_{n,2}^Y \in \{q : q_2 \in A\},$$

that is, the empirical measure L_n^Y is determined by the projection q_2 of the empirical pair measure $L_{n,2}^Y$: For projection $P : M_1(\Sigma^2) \to M_1(\Sigma)$, we have $P(q) = q_2 \in M_1(\Sigma)$, where q_2 is the second marginal of measure q.

Now, since the projection of any measure $q \in M_1(\Sigma^2)$ to its marginal q_2 is continuous (because projection function are continuous), and by theorem 3.1.13 [1] function $I_2(\cdot)$ controls the LDP of $L_{n,2}^Y$ we have for any A

$$\limsup_{n\to\infty} \frac{1}{n}\log \mathbb{P}[L_n^Y \in A] \overset{\text{cont.}}{=} \limsup_{n\to\infty} \frac{1}{n}\log \mathbb{P}[L_{n,2}^Y \in P^{-1}(A)] \leq -\inf_{q\in P^{-1}(A)} I_2(q) = -\inf_{\nu\in A} \overbrace{\inf_{\{q:q_2=\nu\}} I_2(q)}^{I(\nu)}$$

if A is closed. If A is open

$$\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}[L_n^Y \in A] \overset{\text{cont.}}{=} \liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}[L_{n,2}^Y \in P^{-1}(A)] \geq -\inf_{q\in P^{-1}(A)} I_2(q) = -\inf_{\nu\in A} \overbrace{\inf_{\{q:q_2=\nu\}} I_2(q)}^{I(\nu)}.$$

That is, $I(\nu) = \inf_{\{q:q_2=\nu\}} I_2(q)$ is a rate function governing the LDP of L_n^Y. By theorem 3.1.6, we know that $I(\nu) = J(\nu)$ which proves the wanted itentity (21). ■

Exercise 3.1.21

a) Let us prove that for any sequence $\mathbf{y} = (y_1, \dots, y_n) \in \Sigma^n$ of nonzero $\mathbb{P}_\sigma^\pi$ probability,

$$\frac{1}{n}\log \mathbb{P}_\sigma^\pi[Y_1 = y_1, \dots, Y_n = y_n] = \sum_{i=1}^{|\Sigma|}\sum_{j=1}^{|\Sigma|} L_{n,2}^{\mathbf{y}}(i,j)\log \pi(i,j). \tag{22}$$

Proof.

Since $Y_1, \dots, Y_n, \dots$ are Markov chains taking values in a finite alphabet Σ, we get by doing straight calculation, and using the definition of $\mathbb{P}_\sigma^\pi$ when starting from the initial state $\sigma \in \Sigma$ that

$$\frac{1}{n}\log \mathbb{P}_\sigma^\pi[Y_1 = y_1, \dots, Y_n = y_n] = \frac{1}{n}\log\left(\pi(\sigma, y_1)\prod_{k=1}^{n-1}\pi(y_k, y_{k+1})\right) = \frac{1}{n}\sum_{k=1}^{n}\log(\pi(y_{k-1}, y_k))$$

$$= \sum_{i=1}^{|\Sigma|}\sum_{j=1}^{|\Sigma|}\frac{1}{n}\sum_{k=1}^{n}\mathbf{1}(Y_{k-1}=y_i, Y_k=y_j)\cdot\log(\pi(y_i,y_j)) = \sum_{i=1}^{|\Sigma|}\sum_{j=1}^{|\Sigma|}L_{n,2}^{\mathbf{y}}(i,j)\log\pi(i,j),$$

where we used the shifting of sum, when we notated that $y_0 = Y_0 = \sigma$, and the fact that $\mathbf{1}(Y_{k-1}=y_i, Y_k=y_j) = 1$ only when $y_i = y_{k-1}$, and $y_j = y_k$. ■

b) Let

$$\mathcal{L}_n = \{q : q = L_{n,2}^{\mathbf{y}}, \mathbb{P}_\sigma^\pi[Y_1 = y_1, \dots, Y_n = y_n] > 0 \text{ for some } \mathbf{y} \in \Sigma^n\}$$

be the set of possible types of pairs of the states of the Markov chain. Let us prove that $\mathcal{L}_n$ can be identified with a subset of $M_1(\Sigma_\Pi)$, where $\Sigma_\Pi = \{(i,j) : \pi(i,j) > 0\}$, and that $|\mathcal{L}_n| \le (n+1)^{|\Sigma|^2}$.

Proof.

First, let us show that $\mathcal{L}_n$ can be identified with a subset of $M_1(\Sigma_\Pi)$. We know that by part a) the given probability $\mathbb{P}_\sigma^\pi$ can be represented in terms of (22). Thus, suppose $\mathbb{P}_\sigma^\pi[Y_1 = y_1, \dots, Y_n = y_n] > 0$. Then equation (22) right hand-side yields that for every indices $i, j \in \{1, \dots, |\Sigma|\}$ where $L_{n,2}^{\mathbf{y}}(i,j) > 0$ we must have that $\pi(i,j) > 0$. Otherwise, if $L_{n,2}^{\mathbf{y}}(i,j) > 0$ and $\pi(i,j) = 0$, we get from (22) that

$$\mathbb{P}_\sigma^\pi[Y_1 = y_1, \dots, Y_n = y_n] = \sum_{i=1}^{|\Sigma|}\sum_{j=1}^{|\Sigma|}L_{n,2}^{\mathbf{y}}(i,j)\log\pi(i,j) = -\infty,$$

since $\log\pi(i,j) = -\infty$ with the specified indices. This is contradiction to our assumption. Therefore, we conclude $\mathcal{L}_n \subset M_1(\Sigma_\Pi)$, where $\Sigma_\Pi = \{(i,j) : \pi(i,j) > 0\}$.

Now, it only suffices to show that $|\mathcal{L}_n| \le (n+1)^{|\Sigma|^2}$. We note that every component of the vector $L_{n,2}^{\mathbf{y}}$ belongs to the set $\{\frac{0}{n}, \frac{1}{n}, \dots, \frac{n}{n}\}$ whose cardinality is $(n+1)$. This follows directly from the definition of $L_{n,2}^{\mathbf{y}}$. Thus, since the vector $L_{n,2}^{\mathbf{y}}$ is specified by at most $|\Sigma|^2$ such quantities because $y \in \Sigma^2$, the given bound follows directly from it. ■

c) Let $T_n(q)$ be the type class of $q \in \mathcal{L}_n$, namely, the set of sequences $\mathbf{y}$ of positive $\mathbb{P}_\sigma^\pi$ probability for which $L_{n,2}^{\mathbf{y}} = q$, and let $H(q) = -\sum_{i,j} q(i,j)\log q_f(j \mid i)$. Suppose that for any $q \in \mathcal{L}_n$,

$$(n+1)^{-(|\Sigma|^2+|\Sigma|)}e^{nH(q)} \le |T_n(q)| \le e^{nH(q)},$$

and moreover that for all $q \in M_1(\Sigma_\Pi)$,

$$\lim_{n\to\infty} d_V(q, \mathcal{L}_n) = 0 \Leftrightarrow q \text{ is shift variant.}$$

Let us prove by adapting the method of types Section 2.1.1 that $L_{n,2}^{\mathbf{Y}}$ satisfies the LDP with the rate function $I_2(\cdot)$ specified in (3.1.14).

Proof.

We aim to show that $L_{n,2}^{\mathbf{Y}}$ satisfies the LDP with the rate function $I_2(\cdot)$ specified in (3.1.14) by adapting the method of types Section 2.1.1 and above assumptions [1]. Hence, by using same methods as in theorem 2.1.10 proof [1], we can deduce that for every set of probability vectors Γ in $M_1(\Sigma_\Pi)$ has upper bound

$$\mathbb{P}_\sigma^\pi[L_{n,2}^{\mathbf{Y}} \in \Gamma] = \sum_{q\in\Gamma\cap\mathcal{L}_n}\mathbb{P}_\sigma^\pi[L_{n,2}^{\mathbf{Y}} = q] = \sum_{q\in\Gamma\cap\mathcal{L}_n}|T_n(q)|\mathbb{P}_\sigma^\pi[(Y_1,\dots,Y_n) = y, L_{n,2}^{\mathbf{y}} = q]$$

$$\overset{a)}{=} \sum_{q\in\Gamma\cap\mathcal{L}_n} |T_n(q)| \exp\Big(n\sum_{i,j} q(i,j)\log\pi(i,j)\Big) \le \sum_{q\in\Gamma\cap\mathcal{L}_n} e^{nH(q)} e^{n\sum_{i,j} q(i,j)\log\pi(i,j)}$$

$$\le |\Gamma\cap\mathcal{L}_n| e^{n\inf_{q\in\Gamma\cap\mathcal{L}_n}\left(H(q)+\sum_{i,j} q(i,j)\log\pi(i,j)\right)} \le |\mathcal{L}_n| e^{-n\inf_{q\in\Gamma\cap\mathcal{L}_n} I_2(q)} \overset{b)}{\le} (n+1)^{|\Sigma|^2} e^{-n\inf_{q\in\Gamma\cap\mathcal{L}_n} I_2(q)},$$

where we used part a) result (22), and the assumptions made for every $q \in \mathcal{L}_n \supset \Gamma\cap\mathcal{L}_n$, and part b) result for the upper bound to cardinality of $|\mathcal{L}_n|$. Moreover, in the last steps we obtained that

$$H(q) + \sum_{i,j} q(i,j)\log\pi(i,j) = -\sum_{i,j} q(i,j)\log q_f(j\mid i) + \sum_{i,j} q(i,j)\log\pi(i,j) = -\sum_{i,j} q(i,j)\log\frac{q_f(j\mid i)}{\pi(i,j)}$$

$$= -\sum_{i,j} q_1(i)\frac{\overbrace{q(i,j)}^{=q_f(j|i)}}{q_1(i)}\log\frac{q_f(j\mid i)}{\pi(i,j)} = -\sum_{i} q_1(i)\sum_{j} q_f(j\mid i)\log\frac{q_f(j\mid i)}{\pi(i,j)} \overset{(3.1.14)}{=} -I_2(q).$$

Samely, the accompanying lower bound is

$$\mathbb{P}_\sigma^\pi[L_{n,2}^{\mathbf{Y}}\in\Gamma] \ge \sum_{q\in\Gamma\cap\mathcal{L}_n} (n+1)^{-(|\Sigma|^2+|\Sigma|)} e^{nH(q)} e^{n\sum_{i,j} q(i,j)\log\pi(i,j)}$$

$$\ge |\Gamma\cap\mathcal{L}_n|(n+1)^{-(|\Sigma|^2+|\Sigma|)} e^{-n\inf_{q\in\Gamma\cap\mathcal{L}_n} I_2(q)} \ge (n+1)^{-(|\Sigma|^2+|\Sigma|)} e^{-n\inf_{q\in\Gamma\cap\mathcal{L}_n} I_2(q)}.$$

Since $\lim_{n\to\infty}\frac{1}{n}\log(n+1)^{-(|\Sigma|^2+|\Sigma|)} = -\lim_{n\to\infty}\frac{1}{n}\log(n+1)^{(|\Sigma|^2+|\Sigma|)} = 0$ and $\lim_{n\to\infty}\frac{1}{n}\log(n+1)^{|\Sigma|^2} = 0$, the normalized logarithmic limit of upper bound yields

$$\limsup_{n\to\infty}\frac{1}{n}\mathbb{P}_\sigma^\pi[L_n^{\mathbf{Y}}\in\Gamma] = -\liminf_{n\to\infty}\{\inf_{q\in\Gamma\cap\mathcal{L}_n} I_2(q)\} \le -\inf_{q\in\Gamma} I_2(q), \tag{23}$$

where the last step follows directly from $\Gamma\cap\mathcal{L}_n \subset \Gamma$ for all n. And the normalized logarithmic limit of lower bound yields

$$\liminf_{n\to\infty}\frac{1}{n}\mathbb{P}_\sigma^\pi[L_n^{\mathbf{Y}}\in\Gamma] = -\limsup_{n\to\infty}\{\inf_{q\in\Gamma\cap\mathcal{L}_n} I_2(q)\}. \tag{24}$$

In order to proof the lower bound completely, let us fix arbitrary point q in the interior of Γ such that $q\in\mathcal{L}_n$. Then, for some $\delta>0$ small enough $\{q' : d_V(q,q')<\delta\}\subset\Gamma$. Thus, by our second assumption, if q is shift variant, we know that $d_V(q,\mathcal{L}_n)=0$ which means that there exists sequence $q_n\in\Gamma\cap\mathcal{L}_n$ such that $q_n\to q$ as $n\to\infty$. Moreover, without loss of generality, it may be assumed that q_n are shift invariants, and hence

$$-\limsup_{n\to\infty}\{\inf_{q'\in\Gamma\cap\mathcal{L}_n} I_2(q')\} \ge -\lim_{n\to\infty} I_2(q_n) = I_2(q).$$

We recall that $I_2(q)=\infty$ whenever q is not shift invariant. Therefore, by preceding inequality,

$$-\limsup_{n\to\infty}\{\inf_{q'\in\Gamma\cap\mathcal{L}_n} I_2(q')\} \ge -\inf_{q\in\Gamma^\circ} I_2(q),$$

and the lower bound follows by combining above result to (24).

This proves that $L_{n,2}^{\mathbf{Y}}$ satisfies the LDP with the rate function $I_2(\cdot)$ specified in (3.1.14) [1]. ■

3.2 Long Rare Seqments in Random Walk

Let us consider now the random walk

$$S_0 = 0, \quad \text{and } S_k = \sum_{i=1}^{k} X_k, \quad k = 1, 2, \ldots,$$

where X_i are i.i.d. random variables taking values in $\mathbb{R}^d$. In this section we will study as an application how the stopping times of random walk T_r (or maximal length of the random walk up to time R_m) behaves asymptotically using Cramer's theorem [1].

Exercise 3.2.6

Let us prove that theorem 3.2.1 holds when $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are as in exercise 3.1.4. Specifically, Y_k are the states of a Markov chain on the finite set $\{1, 2, \ldots, |\Sigma|\}$ with an irreducible transition matrix Π, and the conditional law of X_k when $Y_k = j$ is $\mu_j \in M_1(\mathbb{R}^d)$; the random variables $\{X_k\}$ are independent given any realization of the Markov chain states, and the logarithmic moment generating functions Λ_j associated with μ_j are finite everywhere.

Proof.

Let us define $\hat{S}_n = \frac{1}{n} S_n = \frac{1}{n} \sum_{k=1}^{n} X_k$. According to exercise 3.1.4 [1] result, we know that for $\hat{S}_n$ Theorem 3.1.2 holds for any Borel meausrable set A with $\pi_\lambda(i,j) = \pi(i,j) e^{\Lambda_j(\lambda)}$, where $\pi(i,j)$ are the entries of stochastic matrix Π.

In another words, by theorem 3.1.2, we know that $\hat{S}_n$ satisfies the LDP for any Borel set $A \subset \mathbb{R}^d$ and initial state σ, and if we assume that $\inf_{x \in \overline{A}} I(x) = \inf_{x \in A^\circ} I(x)$ we know that the limit

$$I_A := \inf_A I(x) = -\lim_{n \to \infty} \log \frac{1}{n} \mathbb{P}_\sigma^\pi[\hat{S}_n \in A] = -\lim_{n \to \infty} \log \frac{1}{n} \mu_n(A)$$

exists, where μ_n is the law of $\hat{S}_n$. This is an assumption made on theorem 3.2.1.

We can not use directly theorem 3.2.1 since $\{X_k\}$ are conditionally independent which is not the same that stronger than unconditional independence assumption made before the theorem 3.2.1. Therefore, we have to proof explicitly theorem 3.2.1 when $\{X_k\}$ are conditionally independent.

By theorem 3.2.1 proof, the upper bound follows samely as in the given proof since assumption of independence is not used there. That is, we know that

$$\limsup_{m \to \infty} \frac{R_m}{\log m} = \limsup_{r \to \infty} \frac{r}{\log T_r} \leq \frac{1}{I_A} \quad \text{almost surely.}$$

Therefore, it suffices to show the lower bound (for $I_A < \infty$) by establishing that the right tail of T_r needs to be bounded. Hence, let

$$B_l := \{\frac{1}{r}(S_{lr} - S_{(l-1)r}) \in A\} = \{\frac{1}{r}(\sum_{k=1}^{lr} X_k - \sum_{k=1}^{(l-1)r} X_k) \in A\}$$

$$= \{\frac{1}{r} \sum_{k=(l-1)r+1}^{lr} X_k \in A\} = \{\frac{1}{r} \sum_{m=1}^{r} X_{m+(l-1)r} \in A\}.$$

By this definition, we note since B_l consist of disjoint segments of $\{X_k\}$s, it follows that

$$\mathbb{P}_\sigma^\pi[\bigcap_{l=1}^{\infty} B_l \mid Y_1 = j_1, \ldots, Y_n = j_n] = \mathbb{P}_\sigma^\pi[\bigcap_{l=1}^{\infty} \{\frac{1}{r}\sum_{m=1}^{r} X_{m+(l-1)r} \in A\} \mid Y_1 = j_1, \ldots, Y_n = j_n]$$
$$\overset{\perp\!\!\!\perp}{=} \prod_{l=1}^{\infty} \mathbb{P}_\sigma^\pi[\{\frac{1}{r}\sum_{m=1}^{r} X_{m+(l-1)r} \in A\} \mid Y_1 = j_1, \ldots, Y_n = j_n] = \prod_{l=1}^{\infty} \mathbb{P}_\sigma^\pi[B_l \mid Y_1 = j_1, \ldots, Y_n = j_n],$$

since random vectors $\{X_{m+(l+1)r}\}_{m=1}^{r}$ for each l are conditionally independent due to consisting disjoint segments of $\{X_k\}$s which are conditionally independent under Markov chain $\{Y_k\}$. Also, we note that

$$\mathbb{P}_\sigma^\pi[B_l] = \mathbb{P}_\sigma^\pi[\{\frac{1}{r}\sum_{m=1}^{r} X_{m+(l-1)r} \in A\}] = \mathbb{P}_\sigma^\pi[\hat{S}_r \in A] = \mu_r(A)$$

by definition. Therefore, using the inclusion

$$\bigcup_{l=1}^{\lfloor m/r \rfloor} B_l \subset \{T_r \leq m\}$$

yields the same lower bound samely as in theorem 3.2.1 proof under conditional independence. That is,

$$\liminf_{m\to\infty} \frac{R_m}{\log m} = \liminf_{r\to\infty} \frac{r}{\log T_r} \geq \frac{1}{I_A} \quad \text{almost surely},$$

which completes the proof for the given setting. ■

Exercise 3.2.7

Let us consider the i.i.d. random variables $\{\tilde{Y}_j\}, \{Y_j\}$ all distributed following $\tilde{\mu} \in M_1(\tilde{\Sigma})$ for $\tilde{\Sigma}$ a finite set. Let $\Sigma = \tilde{\Sigma}^2$ and $\mu = \tilde{\mu}^2 \in M_1(\Sigma)$. For any integers $s, r \geq 0$ let $L_k^{T^s\tilde{y}, T^r y}$ denote the empirical measure of the sequence $((\tilde{y}_{s+1}, y_{r+1}), \ldots, (\tilde{y}_{s+k}, y_{r+k}))$.

a) Using Lemma 2.1.9, let us show that for any $\nu \in \mathcal{L}_k, k \in \{1, \ldots, n\}$,

$$\mathbb{P}[\bigcup_{s,r\leq n-k} \{L_k^{T^s\tilde{Y}, T^r Y} = \nu\}] \leq n^2 e^{-kH(\nu|\mu)}.$$

Proof.

We first note that $\mathcal{L}_k := \{\nu : \nu = L_k^{T^s\tilde{y}, T^r y} \text{ for some } (\tilde{y}, y)\} \subset \mathbb{R}^{|\Sigma|}$. Now, using the inclusion $\bigcup_{s,r\leq n-k}\{L_k^{T^s\tilde{Y}, T^r Y} = \nu\} \subset \bigcup_{s,r\leq n}\{L_k^{T^s\tilde{Y}, T^r Y} = \nu\}$, and union bound, it follows that for any $\nu \in \mathcal{L}_k$

$$\mathbb{P}_\mu[\bigcup_{s,r\leq n-k} \{L_k^{T^s\tilde{Y}, T^r Y} = \nu\}] \leq \mathbb{P}_\mu[\bigcup_{s,r\leq n} \{L_k^{T^s\tilde{Y}, T^r Y} = \nu\}] \leq \sum_{s,r\leq n} \mathbb{P}_\mu[\{L_k^{T^s\tilde{Y}, T^r Y} = \nu\}]$$
$$\overset{\text{Lemma 2.1.9}}{\leq} \sum_{s,r\leq n} e^{-kH(\nu|\mu)} = n^2 e^{-kH(\nu|\mu)},$$

where we used the lemma 2.1.9 for $\mu = \tilde{\mu}^2 \in M_1(|\Sigma|)$. ■

b) Fix $f:\Sigma\to\mathbb{R}$ such that $\mathbb{E}[f(\tilde{Y}_1,Y_1)]<0$ and $\mathbb{P}[f(\tilde{Y}_1,Y_1)>0]>0$. Let us consider

$$M_n=\max_{k,0\le s,r\le n-k}\sum_{j=1}^{k}f(\tilde{Y}_{s+j},Y_{r+j}).$$

Let us prove that almost surely,

$$\limsup_{n\to\infty}\frac{M_n}{\log n}\le\sup_{\nu\in M_1(\Sigma)}\frac{2\int f d\nu}{H(\nu\mid\mu)}.$$

Proof.

First, let $X_j=f(\tilde{Y}_j,Y_j)$. Fix k, and consider for each pair $0\le s,r\le n-k$ let the event

$$B_{k,s,r}=\{\sum_{j=1}^{k}f(\tilde{Y}_{s+j},Y_{r+j})\ge 0\}.$$

Then using (2.2.12) [1] for X_j since we had that $\mathbb{E}[f(\tilde{Y}_1,Y_1)]<0$ and $\mathbb{P}[f(\tilde{Y}_1,Y_1)>0]>0$, we get

$$\mathbb{P}[M_n\ge 0]\le\mathbb{P}[\bigcup_{k\le n}\bigcup_{s,r\le n-k}B_{k,s,r}]\le\sum_{k\le n}\sum_{s,r\le n-k}\mathbb{P}[\sum_{j=1}^{k}f(\tilde{Y}_{s+j},Y_{r+j})\ge 0]\overset{(2.2.12)}{\le}\sum_{k\le n}\sum_{s,r\le n}e^{k\Lambda^*(0)}$$
$$=\sum_{k\le n}n^2e^{-k\Lambda^*(0)}\overset{k\ge(5/\Lambda^*(0))\log n}{\le}\sum_{k\le n}n^2e^{-(5/\Lambda^*(0))\log n\cdot\Lambda^*(0)}=\sum_{k\le n}n^2\cdot n^{-5}=n^3\cdot n^{-5}=n^{-2},$$

where we used the monotonicity for $\{\max_{k,0\le s,r\le n-k}\sum_{j=1}^{k}f(\tilde{Y}_{s+j},Y_{r+j})\ge 0\}\subset\bigcup_{k\le n}\bigcup_{s,r\le n-k}B_{k,s,r}$. This result implies that $k\ge(5/\Lambda^*(0))\log n$ is negligible when considering M_n since the over-harmonic series is finite, i.e., $\sum_{n=1}^{\infty}n^{-2}<\infty$, and $\mathbb{P}[M_n\ge t\log n]\le\mathbb{P}[M_n\ge 0]$ when $t>R=\sup_{\nu\in M_1(\Sigma)}\frac{2\int f d\nu}{H(\nu\mid\mu)}\ge\frac{2\int f d\nu}{H(\nu\mid\mu)}$. Therefore, it suffices to consider only cases when $k\le(5/\Lambda^*(0))\log n$. We can reform $\{M_n\ge t\log n\}$ into form of

$$\{M_n\ge t\log n\}=\{\max_{k,0\le s,r\le n-k}\sum_{j=1}^{k}f(\tilde{Y}_{s+j},Y_{r+j})\ge t\log n\}\subset\bigcup_{k\le n}\bigcup_{s,r\le n-k}\{\sum_{j=1}^{k}f(\tilde{Y}_{s+j},Y_{r+j})\ge t\log n\}$$
$$=\bigcup_{k\le n}\bigcup_{s,r\le n-k}\{k\int f d(L_k^{T^s\tilde{Y},T^rY})\ge t\log n\}=\bigcup_{k\le n}\bigcup_{\nu\in\mathcal{L}_k:k\int f d\nu\ge t\log n}\bigcup_{s,r\le n-k}\{L_k^{T^s\tilde{Y},T^rY}=\nu\}.$$

Using then union bound, and part a) result for $n_k=\lfloor e^k\rfloor$ we get

$$\mathbb{P}[M_{n_k}\ge t\log n_k]\le\sum_{k\le(5/\Lambda^*(0))\log n_k}\sum_{\nu\in\mathcal{L}_k:k\int f d\nu\ge t\log n_k}\mathbb{P}[\bigcup_{s,r\le n_k-k}\{L_k^{T^s\tilde{Y},T^rY}=\nu\}]$$
$$\overset{a)}{\le}\sum_{k\le(5/\Lambda^*(0))\log n_k}\sum_{\nu\in\mathcal{L}_k:k\int f d\nu\ge t\log n_k}n_k^2e^{-kH(\nu|\mu)}\overset{n_k=\lfloor e^k\rfloor}{\le}\sum_{k\le(5/\Lambda^*(0))\log n_k}\sum_{\nu\in\mathcal{L}_k:k\int f d\nu\ge t\log n_k}e^{2k}e^{-kH(\nu|\mu)}$$
$$=\sum_{k\le(5/\Lambda^*(0))\log n_k}\sum_{\nu\in\mathcal{L}_k:k\int f d\nu\ge t\log n_k}e^{-k(H(\nu|\mu)-2)}\le\sum_{k\le(5/\Lambda^*(0))\log n_k}\sum_{\nu\in\mathcal{L}_k:H(\nu|\mu)>2}e^{-k(H(\nu|\mu)-2)}$$
$$\le\sum_{k<\infty}|\mathcal{L}_k|e^{-k\cdot\delta}\overset{\text{Lemma 2.1.2}}{\le}\sum_{k<\infty}(k+1)^{|\Sigma|}e^{-k\cdot\delta}<\infty,$$

since $\delta = H(\nu \mid \mu) - 2 > 0$, and the finiteness follows from the ratio test. The result is due to the assumption that $t > \dfrac{2\int f d\nu}{H(\nu \mid \mu)}$, we have

$$k \int f d\nu \geq t \log n_k \Leftrightarrow k \int f d\nu > \frac{2\int f d\nu}{H(\nu \mid \mu)} \log e^k \Leftrightarrow H(\nu \mid \mu) > 2.$$

By monotonicity of $n \mapsto M_n$, we conclude with the help of first result that:

$$\sum_{n=1}^{\infty} \mathbb{P}[M_n \geq t \log n] < \infty,$$

which implies that by Borel-Cantelli lemma (of convergence part): $M_n \leq t \log n$ almost surely. In another words,

$$\limsup_{n \to \infty} \frac{M_n}{\log n} \leq \sup_{\nu \in M_1(\Sigma)} \frac{2\int f d\nu}{H(\nu \mid \mu)} \quad \text{almost surely.}$$

■

3.3 The Gibbs Conditioning principle for Finite Alphabets

In this section, let $Y_1, Y_2, \ldots, Y_n$ be a sequence of i.i.d. random variables with strictly positive law μ on the finite alphabet Σ, and $X_k = f(Y_k)$ for some deterministic $f : \Sigma \to \mathbb{R}$. Given the constaint type $\hat{S}_n = \frac{1}{n}\sum_{i=1}^n X_i \in A$, we will examine what is the conditional law of Y_1 when n is large. The Gibbs's principle (theorem 3.3.3) characterize this behavior which we will use in the next exercises [1].

Exercise 3.3.11

Let us prove the Gibbs conditioning principle for sampling without replacement, i.e., under the assumptions of Section 2.1.3.

a) Observe that Y_j again are identically distributed even under conditioning on their empirical measures. Let us conclude that (3.3.1) holds [1].

Proof.

We observe that Y_j again are identically distributed even under conditioning on the empirical measures, i.e., conditioning on the event $\{L_n^Y = \nu\}$. On this event the multiset of sampled types is fixed: for each $y \in \Sigma$ the number of sampled occurences of y is $n\nu(y)$. Given the multiset, all orderings of the n sampled items are equally likely (sampling without replacement but then conditioning on the counts removes ordering bias). Hence by symmetry each coordinate has the same conditional marginal. That is,

$$\mathbb{P}[Y_1 \in A \mid L_n^Y = \nu] = \frac{1}{n}\sum_{i=1}^n \mathbb{P}[Y_i \in A \mid L_n^Y = \nu] = \frac{1}{n} n\nu(A) = \nu(A).$$

Using this fact, we get that for every function $\phi : \Sigma \to \mathbb{R}$

$$\langle \phi, \mu_n^* \rangle = \mathbb{E}[\phi(Y_1) \mid \hat{S}_n \in A] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^n \phi(Y_i) \mid \hat{S}_n \in A] = \mathbb{E}[\langle \phi, L_n^Y \rangle \mid \langle f, L_n^Y \rangle \in A],$$

with the same conventions as in section 3.3 beginning. Then with the notation $\Gamma = \{\nu : \langle f, \nu \rangle \in A\}$, we conclude that the (3.3.1) holds in this case as well. That is

$$\mu_n^* = \mathbb{E}[L_n^Y \mid L_n^Y \in \Gamma].$$

■

b) Let us assume that Γ is such that

$$I_\Gamma = \inf_{\nu \in \Gamma^\circ} I(\nu \mid \beta, \mu) = \inf_{\nu \in \overline{\Gamma}} I(\nu \mid \beta, \mu) < \infty,$$

and define $\mathcal{M} = \{\nu \in \overline{\Gamma} : I(\nu \mid \beta, \mu) = I_\Gamma\}$. Let us prove both parts of theorem 3.3.3 hold [1].

Proof.

Since $|\Sigma| < \infty$, $\overline{\Gamma}$ is a compact set and thus $\mathcal{M}$ is non-empty. We will prove part a) first. Hence, let for every $U \subset M_1(\Sigma)$,

$$\mathbb{E}[L_n^Y \mid L_n^Y \in \Gamma] - \mathbb{E}[L_n^Y \mid L_n^Y \in U \cap \Gamma] = \mathbb{P}_\mu[L_n^Y \in U^c \mid L_n^Y \in \Gamma]\Big(\mathbb{E}[L_n^Y \mid L_n^Y \in U^c \cap \Gamma] - \mathbb{E}[L_n^Y \mid L_n^Y \in U \cap \Gamma]\Big).$$

By conditional expectation, we have that $\mathbb{E}[L_n^Y \mid L_n^Y \in U \cap \Gamma] \in \text{co}(U)$, while $\mu_n^* = \mathbb{E}[L_n^Y \mid L_n^Y \in \Gamma]$ as shown in part a). Thus, it follows that

$$d_V(\mu_n^*, \text{co}(U)) \leq \mathbb{P}_\mu[L_n^Y \in U^c \mid L_n^Y \in \Gamma] d_V(\mathbb{E}[L_n^Y \mid L_n^Y \in U^c \cap \Gamma], \mathbb{E}[L_n^Y \mid L_n^Y \in U \cap \Gamma]) \tag{25}$$

$$\leq \mathbb{P}_\mu[L_n^Y \in U^c \mid L_n^Y \in \Gamma] \tag{26}$$

where the last inequality is due to the bound $d_V(\cdot,\cdot) \leq 1$ since two probability measures distance can be at most one. Next, with $\mathcal{M}^\delta = \{\nu : d_V(\nu, \mathcal{M}) < \delta\}$, it is proved shortly that for every $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}_\mu[L_n^Y \in \mathcal{M}^\delta \mid L_n^Y \in \Gamma] = 1, \tag{27}$$

with an exponential (in n) rate of convergence. Consequently, when we use the result (27) and (26) for $U = \mathcal{M}^\delta$ results in $d_V(\mu_n^*, \text{co}(\mathcal{M}^\delta)) \to 0$. Since d_V is a convex function on $M_1(\Sigma) \times M_1(\Sigma)$ due to triangle inequality, each point in $\text{co}(\mathcal{M}^\delta)$ is within variational distance δ of some point in $\text{co}(\mathcal{M})$. With $\delta > 0$ being arbitrarily small, limit points of μ_n^* are necessarily in closure of $\text{co}(\mathcal{M})$.

To prove (27), we apply Theorem 2.1.41 [1] which is analog of Sanov's theorem, and the assumption to get

$$I_\Gamma = -\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_\mu[L_n^Y \in \Gamma], \tag{28}$$

and

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}_\mu[L_n^Y \in (\mathcal{M}^\delta)^c \cap \Gamma] \leq -\inf_{\nu \in (\mathcal{M}^\delta)^c \cap \Gamma} I(\nu \mid \beta, \mu) \leq -\inf_{\nu \in (\mathcal{M}^\delta)^c \cap \overline{\Gamma}} I(\nu \mid \beta, \mu). \tag{29}$$

Observe that $\mathcal{M}^\delta$ are open sets and, therefore $(\mathcal{M}^\delta)^c \cap \overline{\Gamma}$ are compact sets. Hence, for some $\tilde{\nu} \in (\mathcal{M}^\delta)^c \cap \overline{\Gamma}$,

$$\inf_{\nu \in (\mathcal{M}^\delta)^c \cap \overline{\Gamma}} I(\nu \mid \beta, \mu) = I(\tilde{\nu} \mid \beta, \mu) > I_\Gamma. \tag{30}$$

Now, we get that (27) follows from (28)-(30) because

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}_\mu[L_n^Y \in (\mathcal{M}^\delta)^c \mid L_n^Y \in \Gamma] = \limsup_{n\to\infty} \Big(\frac{1}{n} \log \mathbb{P}_\mu[L_n^Y \in (\mathcal{M}^\delta)^c \cap \Gamma] - \frac{1}{n} \log \mathbb{P}_\mu[L_n^Y \in \Gamma] \Big)$$

$$\leq -\inf_{\nu \in (\mathcal{M}^\delta)^c \cap \overline{\Gamma}} I(\nu \mid \beta, \mu) + I_\Gamma < -I_\Gamma + I_\Gamma = 0.$$

Therefore, we have that $d_V(\mu_n^*, \text{co}(\mathcal{M}^\delta)) \to 0$ so part a) of theorem 3.3.3 is proven.

Turning now to part b), we may use exercise 2.1.48 [1]. That is, suppose Γ is convex set of non-empty interior, and since by exercise 2.1.48 the rate $I(\cdot \mid \beta, \mu)$ is a (strictly) convex, we have by definition that $\mathcal{M} = \{\nu \in \overline{\Gamma} : I(\nu \mid \beta, \mu) = I_\Gamma\} = \{\nu^*\}$ where ν^* is the unique minimizer of $I(\cdot \mid \beta, \mu)$. By the compactness of $M_1(\Sigma)$, we get that $\mu_n^* \to \nu^*$ as $n \to \infty$. This proves the part b). ■

Exercise 3.3.12

a) Suppose that $\Sigma = (\Sigma')^k$ and $\mu = (\mu')^k$ are, respectively, a kth product alphabet and a kth product probability measure on it, and assume that μ' is strictly positive on Σ'. For any law $\nu \in M_1(\Sigma)$ and $j = \{1, \ldots, k\}$, let $\nu^{(j)} \in M_1(\Sigma')$, denote the jth marginal of ν on Σ'. Let us prove that

$$\frac{1}{k} H(\nu \mid \mu) \geq \frac{1}{k} \sum_{j=1}^k H(\nu^{(j)} \mid \mu') \geq H(\frac{1}{k} \sum_{j=1}^k \nu^{(j)} \mid \mu'), \tag{31}$$

with equality if and only if $\nu = (\nu')^k$ for some $\nu' \in M_1(\Sigma')$.

Proof.

Using the definition of relative entropy, we get that

$$\begin{aligned}
\frac{1}{k}H(\nu \mid \mu) &= \frac{1}{k}\sum_{i=1}^{|\Sigma|}\nu(a_i)\log\frac{\nu(a_i)}{\mu(a_i)} = \frac{1}{k}\sum_{i=1}^{|\Sigma|}\nu(a_i)\log\Big(\frac{\nu(a_i)}{\prod_{j=1}^{k}\nu^{(j)}(a_i^j)}\cdot\frac{\prod_{j=1}^{k}\nu^{(j)}(a_i^j)}{\mu(a_i)}\Big)\\
&= \frac{1}{k}\sum_{i=1}^{|\Sigma|}\nu(a_i)\log\Big(\frac{\nu(a_i)}{\prod_{j=1}^{k}\nu^{(j)}(a_i^j)}\Big) + \frac{1}{k}\sum_{i=1}^{|\Sigma|}\nu(a_i)\log\Big(\frac{\prod_{j=1}^{k}\nu^{(j)}(a_i^j)}{\mu(a_i)}\Big)\\
&= \frac{1}{k}\sum_{i=1}^{|\Sigma|}\nu(a_i)\log\Big(\frac{\nu(a_i)}{\prod_{j=1}^{k}\nu^{(j)}(a_i^j)}\Big) + \frac{1}{k}\sum_{i=1}^{|\Sigma|}\nu(a_i)\log\Big(\frac{\prod_{j=1}^{k}\nu^{(j)}(a_i^j)}{\prod_{j=1}^{k}\mu'(a_i^j)}\Big)\\
&= \frac{1}{k}\sum_{i=1}^{|\Sigma|}\nu(a_i)\log\Big(\frac{\nu(a_i)}{\otimes_{j=1}^{k}\nu^{(j)}(a_i)}\Big) + \frac{1}{k}\sum_{i=1}^{|\Sigma|}\nu(a_i)\sum_{j=1}^{k}\log\Big(\frac{\nu^{(j)}(a_i^j)}{\mu'(a_i^j)}\Big)\\
&= \frac{1}{k}\sum_{i=1}^{|\Sigma|}\nu(a_i)\log\Big(\frac{\nu(a_i)}{\otimes_{j=1}^{k}\nu^{(j)}(a_i)}\Big) + \frac{1}{k}\sum_{j=1}^{k}\sum_{i=1}^{|\Sigma'|}\nu^{(j)}(x_i)\log\Big(\frac{\nu^{(j)}(x_i)}{\mu'(x_i)}\Big)\\
&= \frac{1}{k}\sum_{j=1}^{k}H(\nu \mid \otimes_{j=1}^{k}\nu^{(j)}) + \frac{1}{k}\sum_{j=1}^{k}H(\nu^{(j)} \mid \mu') \geq \frac{1}{k}\sum_{j=1}^{k}H(\nu^{(j)} \mid \mu'),
\end{aligned}$$

where the inequality followed from the fact that the relative entropy H is non-negative, i.e., $\frac{1}{k}\sum_{j=1}^{k}H(\nu \mid \otimes_{j=1}^{k}\nu^{(j)}) \geq 0$ when Jensen's inequality is applied to the convex function $x\log x$. See the remark of definition 2.1.5 [1]. Here, we noted that $a_i^j \in \Sigma'$ is the jth element of $a_i \in \Sigma$, and $x_i = a_i^j$ as shorthand notation.

For the second inequality, we note that $H(\cdot \mid \mu)$ is a convex function. Therefore, it follows for every $\nu^j \in [0,1]$ that

$$\frac{1}{k}\sum_{j=1}^{k}H(\nu^{(j)} \mid \mu') \geq H(\frac{1}{k}\sum_{j=1}^{k}\nu^{(j)} \mid \mu').$$

Combining these inequalities yields the wanted result (31). Lastly, it is easy to see that if the marginal distributions coincide, i.e., $\nu = (\nu')^k$ for some $\nu' \in M_1(\Sigma)$, then equality follows in (31). ■

b) Let us assume that

$$\Gamma = \{\nu : \frac{1}{k}\sum_{j=1}^{k}v^{(j)} \in \Gamma'\} \tag{32}$$

for some $\Gamma' \subset M_1(\Sigma)$, which satisfies (3.3.2) with respect to μ' [1]. Let $\mathcal{M}' = \{\nu' \in \Gamma' : H(\nu' \mid \mu') = I_{\Gamma'}\}$ and let us prove that $\mathcal{M} = \{\nu : \nu = (\nu')^k, \nu' \in \mathcal{M}'\}$.

Proof.

Let $\nu \in \Gamma$. Then from part a) it follows that

$$H(\nu \mid \mu) \geq k \cdot H(\frac{1}{k}\sum_{j=1}^{k}\nu^{(j)} \mid \mu') \geq k \cdot \inf_{\nu' \in \Gamma'}H(\nu' \mid \mu') \overset{(3.3.2)}{=} k \cdot I_{\Gamma'}$$

where the last inequality is due to the fact that $\frac{1}{k}\sum_{j=1}^{k}\nu^{(j)} \in \Gamma'$. Since ν was chosen arbitrarily, we have that

$$I_\Gamma = \inf_{\nu\in\Gamma} H(\nu \mid \mu) \geq k \cdot I_{\Gamma'}.$$

Now, let $\nu' \in \mathcal{M}'$. Then, on the other hand we have by part a) that for $\nu = (\nu')^k \in \Gamma$ that,

$$H((\nu')^k \mid \mu) = k \cdot \frac{1}{k}\sum_{j=1}^{k} H(\nu' \mid \mu') = k \cdot H(\nu' \mid \mu') = k \cdot I_{\Gamma'}.$$

Combining these results, we get that

$$I_\Gamma = \inf_{\nu\in\Gamma} H(\nu \mid \mu) \geq H((\nu')^k \mid \mu) \Rightarrow I_\Gamma = H((\nu')^k \mid \mu),$$

that is, by definition we have that $\mathcal{M} = \{\nu : I_\Gamma = \inf_{\nu\in\Gamma} H(\nu \mid \mu)\} = \{\nu : \nu = (\nu')^k, \nu' \in \mathcal{M}'\}$ which we sought to proof. ■

c) Consider the kth joint conditional law

$$\mu_n^*(a'_{i_1}, \ldots, a'_{i_k}) = \mathbb{P}_{\mu'}[Y_1 = a'_{i_1}, \ldots, Y_k = a'_{i_k} \mid L_n^Y \in \Gamma'],$$

where Y_i are i.i.d. with marginal $\mu' \in M_1(\Sigma')$ and $\Gamma' \subset M_1(\Sigma')$ satisfies (3.3.2), with $\mathcal{M}'$ being a single point. Let $\mu = (\mu')^k$ be the law of $X_i = (Y_{k(i-1)+1}, \ldots, Y_{ki})$ on a new alphabet Σ. Let us prove that for any $n \in \mathbb{Z}_+$,

$$\mu_{nk}^*(a_i) = \mathbb{P}_\mu[X_1 = a_i \mid L_n^X \in \Gamma], \quad \forall a_i \in \Sigma, \tag{33}$$

where Γ is defined in (32). Furthermore, let us deduce the limit point μ_{nk}^* is a kth product of an element of $M_1(\Sigma')$. Hence, as $n \to \infty$ along the integer multiples of k, the random variables $Y_i, i = 1, \ldots, k$ are asymptotically conditionally i.i.d.

Proof.

Let us fix $k \in \mathbb{Z}_+$. Now, for we know that $L_{nk}^Y = \frac{1}{nk}\sum_{i=1}^{nk} \mathbf{1}\{Y_i\} \overset{\text{i.i.d.}}{=} \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i\} = \frac{1}{k}\sum_{j=1}^{k} L_n^X$. Therefore, we get that the events

$$\pi^{-1}(\{L_n^X \in \Gamma\}) = \{L_{nk}^Y \in \Gamma'\}$$

where $\pi : (\Sigma')^{nk} \to \Sigma^n$ is a measurable bijection $\pi(y_1, \ldots, y_{nk}) = ((y_1, \ldots, y_k), \ldots, (y_{(n-1)k+1}, \ldots, y_{nk}))$ and Γ is defined as (32). Now, we can also see that the cylinder events $a_i = (a'_{i_1}, \ldots, a'_{i_k}) \in \Sigma$ coincide too since,

$$\pi^{-1}(\{X_1 = a_i\}) = \{Y_{k(1-1)+1} = a'_{i_1}, \ldots, Y_{k\cdot 1} = a'_{i_k}\} = \{Y_1 = a'_{i_1}, \ldots, Y_k = a'_{i_k}\}.$$

Lastly, we note that since $\mu = (\mu')^k$, we have that $(\mu')^{nk} \circ \pi^{-1} = \mu^n$. Putting all things together then yields

$$\begin{aligned}
&\mu_{nk}^*(a_i) = \mu_{nk}^*(a'_{i_1}, \ldots, a'_{i_k}) = \mathbb{P}_{\mu'}[Y_1 = a'_{i_1}, \ldots, Y_k = a'_{i_k} \mid L_{nk}^Y \in \Gamma'] \\
&= \frac{\mathbb{P}_{\mu'}[\{Y_1 = a'_{i_1}, \ldots, Y_k = a'_{i_k}\} \cap \{L_{nk}^Y \in \Gamma'\}]}{\mathbb{P}_{\mu'}[\{L_{nk}^Y \in \Gamma'\}]} = \frac{\mathbb{P}_{\mu'}[\pi^{-1}(\{X_1 = a_i\} \cap \{L_n^X \in \Gamma\})]}{\mathbb{P}_{\mu'}[\pi^{-1}(\{L_n^X \in \Gamma\})]} \\
&= \frac{\mathbb{P}_\mu[\{X_1 = a_i\} \cap \{L_n^X \in \Gamma\}]}{\mathbb{P}_\mu[\{L_n^X \in \Gamma\}]} = \mathbb{P}_\mu[X_1 = a_i \mid L_n^X \in \Gamma] \quad \forall n \in \mathbb{Z}_+.
\end{aligned}$$

Therefore, the equation (33) holds.

Lastly, by part b), since $\mathcal{M}'$ consist single point, let us denote it $\nu' \in \mathcal{M}' \subset M_1(\Sigma')$, we know that $(\nu')^k \in \mathcal{M}$. By Gibbs's principle (theorem 3.3.3 part b)) [1], we note that $\mu^*_{nk} \to (\nu')^k$ by equation (33). That is, as $n \to \infty$ along the integers multiple of k, the random variables $Y_i, i = 1, \ldots, k$ are asymptotically conditionally i.i.d. ■

d) Let us prove that the preceding conclusion extends to n which need not to be integer multiples of k.

Proof.

We note that by part c)

$$\mu^*_{nk}(a_i) \to (\nu')^k \in \mathcal{M}.$$

Suppose n is not to be integer multiples of k, that is, it can be expressed as $n = q \cdot k + r$, for some $q \in \mathbb{Z}_+$ and $r \in [0, k]$. Then, we have that

$$\begin{aligned} L^Y_n &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i\} = \frac{1}{n}\Big(\sum_{i=1}^{qk} \mathbf{1}\{Y_i\} + \sum_{i=qk+1}^{qk+r} \mathbf{1}\{Y_i\}\Big) \\ &= \frac{1}{n}\sum_{i=1}^{qk} \mathbf{1}\{Y_i\} + \frac{1}{n}\sum_{i=qk+1}^{qk+r} \mathbf{1}\{Y_i\} = \frac{qk}{n}\frac{1}{qk}\sum_{i=1}^{qk} \mathbf{1}\{Y_i\} + \frac{r}{n}\frac{1}{r}\sum_{i=qk+1}^{qk+r} \mathbf{1}\{Y_i\} \\ &= \frac{qk}{n}L^Y_{qk} + \frac{r}{n}L^Y_r, \end{aligned}$$

where $L^Y_r = \sum_{i=qk+1}^{qk+r} \mathbf{1}\{Y_i\}$. We note that by total variation distance,

$$\begin{aligned} &d_{TV}(L^Y_n, L^Y_{qk}) = \frac{1}{2}\sum_{x\in\Sigma'} |L^Y_n(x) - L^Y_{qk}(x)| = \frac{|\Sigma'|}{2}|L^Y_n(x) - L^Y_{qk}(x)| \le \frac{|\Sigma'|}{2}\sup_{x\in\Sigma'}|L^Y_n(x) - L^Y_{qk}(x)| \\ &= \frac{|\Sigma'|}{2}\sup_{x\in\Sigma'}|\frac{qk}{n}L^Y_{qk}(x) + \frac{r}{n}L^Y_r(x) - L^Y_{qk}(x)| = \frac{|\Sigma'|}{2}\sup_{x\in\Sigma'}\Big|\frac{qk-n}{n}L^Y_{qk}(x) + \frac{r}{n}L^Y_r(x)\Big| \\ &= \frac{|\Sigma'|}{2}\sup_{x\in\Sigma'}\Big|\frac{-r}{n}L^Y_{qk}(x) + \frac{r}{n}L^Y_r(x)\Big| = \frac{|\Sigma'|}{2}\cdot\frac{r}{n}\sup_{x\in\Sigma'}\Big|L^Y_r(x) - L^Y_{qk}(x)\Big| \\ &\le \frac{|\Sigma'|}{2}\cdot\frac{r}{n}\cdot 1 = \frac{|\Sigma'|}{2}\cdot\frac{r}{n} \to 0 \end{aligned}$$

as $n \to \infty$. Here, we used in the first equality known results of total variation for with respect to probability mass functions (proposition 9.4 from Probability theory course), and the fact that the distance between two probability mass functions are at most 1.

Thus, this result implies that as $n \to \infty$ the law L^Y_n approaches the law of L^Y_{qk}, which is a empirical measure of integer multiples of k. By part c), we concluded that the conditional law approaches $\mu^*_{qk}(a_i) \to (\nu')^k \in \mathcal{M}$. Therefore, we note that the given conclusion extends to n which need not to be integer multiples of k. ■

3.6 Rate Distortion theory

In this section, we will deal with one of the basic problems in information theory, namely, namely, the problem of source coding. By looking at the source as a random source, one could hope to analyze this situation and find the fundamental performance limits and methods to achieve these limits [1]. Therefore, it is not surprising that the large deviation principle comes in hand in the source coding problem. We follow the notation done in the section 3.6 in the book [1] for source coding problem.

Exercise 3.6.10

a) Let us prove if Σ is a finite set and $R_1(D) > 0$, then there exists a probability measure Q on $\Sigma \times \Sigma$ for which $Q_X = \mathcal{P}_1,\ \rho_Q = D$, and $H(Q \mid Q_X \times Q_Y) = R_1(D)$.

Proof.

Suppose $|\Sigma| < \infty$ and $R_1(D) > 0$. We note that since Σ is finite, the space $M_1(\Sigma \times \Sigma)$ a compact and simplex. Since the set $\mathcal{C}_D = \{Q : \rho_Q \leq D, Q_X = \mathcal{P}_1\} \subset M_1(\Sigma \times \Sigma)$ is closed (and bounded), it is compact as well. Moreover, we know that relative entropy $H(Q \mid Q_x \times Q_Y)$ is lower semi-continuous on simplex and compact set $\mathcal{C}_D$. Therefore, we note that the infimum is attained by some probability measure $\hat{Q} \in \mathcal{C}_D$. That is,

$$\exists \hat{Q} \in \mathcal{C}_D : R_1(D) = \inf_{\mathcal{C}_D} H(Q \mid Q_X \times Q_Y) = H(\hat{Q} \mid \hat{Q}_X \times \hat{Q}_Y).$$

Hence, it now suffices to show that the given measure $\hat{Q}$ is in the boundary of $\mathcal{C}_D$, i.e., $\hat{Q} \in \partial\mathcal{C}_D = \{Q_X = \mathcal{P}_1, \rho_Q = D\}$. We aim to show this by proof by contradiction.

Suppose first that, $\rho_{\hat{Q}} < D$, and set $Q_t = (1-t)\hat{Q} + t(\mathcal{P}_1 \times \hat{Q}_Y)$, where $\mathcal{P}_1 \times \hat{Q}_Y$ is independent coupling. For this construction, we get that the marginals of Q_t are

$$(Q_t)_X = (1-t)\hat{Q}_X + t\mathcal{P}_1 = (1-t)\cdot\mathcal{P}_1 + t\cdot\mathcal{P}_1 = \mathcal{P}_1 \quad \text{and} \quad (Q_t)_Y = (1-t)\hat{Q}_Y + t\hat{Q}_Y = \hat{Q}_Y$$

for all $t \in [0,1]$. Therefore, since we know that the relative entropy $H(Q_t \mid \mathcal{P}_1 \times \hat{Q}_Y)$ is convex in t, and $H(\mathcal{P}_1 \times \hat{Q}_Y \mid \mathcal{P}_1 \times \hat{Q}_Y) = 0$, we have that

$$H(Q_t \mid \mathcal{P}_1 \times \hat{Q}_Y) \leq (1-t)H(\hat{Q} \mid \mathcal{P}_1 \times \hat{Q}_Y) + t \cdot H(\mathcal{P}_1 \times \hat{Q}_Y \mid \mathcal{P}_1 \times \hat{Q}_Y) = (1-t)H(\hat{Q} \mid \mathcal{P}_1 \times \hat{Q}_Y). \tag{34}$$

Moreover, by the definition of distortion and Q_t, we get that $\rho_{Q_t} = (1-t)\rho_{\hat{Q}} + t\rho_{\mathcal{P}_1 \times \hat{Q}_Y}$.

Now, suppose that $\rho_{\mathcal{P}_1 \times \hat{Q}_Y} > \rho_{\hat{Q}}$. Then, by continuity, we have that $\rho_{Q_t} = D$ for some $t \in (0,1]$. This implies that $Q_t \in \mathcal{C}_D$ for some $t \in (0,1]$, and due to (34), this contradicts the optimality of $\hat{Q}$.

If $\rho_{\mathcal{P}_1 \times \hat{Q}_Y} \leq \rho_{\hat{Q}} < D$, then $Q_t \in \mathcal{C}_D$ for every $t \in [0,1]$. Thus

$$0 \leq \lim_{t\to 1} H(Q_t \mid \mathcal{P}_1 \times \hat{Q}_Y) \leq \lim_{t\to 1}(1-t)H(\hat{Q} \mid \mathcal{P}_1 \times \hat{Q}_Y) = 0 \Rightarrow \lim_{t\to 1} H(Q_t \mid \mathcal{P}_1 \times \hat{Q}_Y) = 0,$$

and since the relative entropy is (lower semi-)continous, this implies that $H(Q_1 \mid \mathcal{P}_1 \times \hat{Q}_Y) = 0$. Thus, $R_1(D) = 0$ since $Q_1 \in \mathcal{C}_D$ which contradicts the assumption $R_1(D) > 0$. Therefore, we conclude that $\hat{Q} \in \partial\mathcal{C}_D$ which we aimed to show. ■

b) Let us prove that for measure in part a), we have also $\Lambda^*(\rho_Q) = H(Q \mid Q_X \times Q_Y)$.

Proof.

By part a), we know that there exists a measure $\hat{Q}$ such that $R_1(D) = H(\hat{Q} \mid \hat{Q}_X \times \hat{Q}_Y)$, $Q_X = \mathcal{P}_1$, and $\rho_{\hat{Q}} = D$. We note that the Fenchel-Legendre transform is

$$\Lambda^*(\theta) = \sup_{\lambda \in \mathbb{R}} \{\lambda \cdot \theta - \Lambda(\lambda)\},$$

where $\Lambda(\lambda)$ is the Lagrangian of constrained problem in this case, that is,

$$\Lambda(\lambda) = \inf_{\{Q \in M_1(\Sigma \times \Sigma), Q_X = \mathcal{P}_1\}} \{\lambda \cdot \rho_Q - H(Q \mid Q_X \times Q_Y)\}.$$

In the part a), we showed that the measure $\hat{Q}$ is the optimal measure for this optimization problem, i.e., the Lagrangian is minimized when $\hat{Q}$. Thus, the Lagrangian of constrained problem comes into form

$$\Lambda(\lambda) = \lambda \cdot \rho_{\hat{Q}} - H(\hat{Q} \mid \hat{Q}_X \times \hat{Q}_Y),$$

and by the Duality lemma (Lemma 4.5.8 [1]) this also ensures the supremum for Λ^*. That is, we have that for some λ^*, we have

$$\Lambda^*(\theta) = \lambda^* \cdot \theta - \Lambda(\lambda^*) = \lambda^* \cdot \theta - \lambda \cdot \rho_{\hat{Q}} + H(\hat{Q} \mid \hat{Q}_X \times \hat{Q}_Y).$$

Substituting $\theta = \rho_{\hat{Q}}$ then yields

$$\Lambda^*(\rho_{\hat{Q}}) = H(\hat{Q} \mid \hat{Q}_X \times \hat{Q}_Y).$$

■

Exercise 3.6.12

a) Let us show that for all integers m, n,

$$(m+n)R_{m+n}(D) \leq mR_m(D) + nR_n(D).$$

Proof.

First, we note that $R_J(D)$ is defined as $R_J(D) := \inf_{\{Q : \rho_Q^{(J)} \leq D, Q_X = \mathcal{P}_J\}} \frac{1}{J} H(Q \mid Q_X \times Q_Y)$, where $\mathcal{P}_J$ is the Jth marginal distribution of the stationary measure $\mathcal{P}$, and $\rho_Q^{(J)}$ is the J-symbol distortion of probability measure Q on $\Sigma^J \times \Sigma^J$ as in the book [1]. By these definition, we note that for small $\epsilon > 0$ we have

$$\frac{1}{m} H(Q^{(m)} \mid Q_X^{(m)} \times Q_Y^{(m)}) \leq R_m(D) + \epsilon \quad \text{and} \quad \frac{1}{n} H(Q^{(n)} \mid Q_X^{(n)} \times Q_Y^{(n)}) \leq R_n(D) + \epsilon. \tag{35}$$

Moreover, we note that

$$(m+n)R_{m+n}(D) \leq H(Q^{m+n} \mid Q_X^{m+n} \times Q_Y^{m+n}) \tag{36}$$

by the definition of $R_{m+n}(D)$. Here, Q^{m+n} is probability measure on Σ^{m+n}, and $Q_X^{m+n} = \mathcal{P}^{m+n}, Q_Y^{m+n}$ are the respective marginals. Using the definition of the relative entropy, we get that

$$H(Q^{(m+n)} \mid Q_X^{(m+n)} \times Q_Y^{(m+n)}) = \int_{\Sigma^{m+n} \times \Sigma^{m+n}} \log \Big(\frac{dQ^{(m+n)}}{d(Q_X^{(m+n)} \times Q_Y^{(m+n)})} \Big) dQ^{(m+n)}$$

$$
\begin{aligned}
&= \int_{\Sigma^{m+n}\times\Sigma^{m+n}} \log\Big(\frac{d(Q^{(m)}\otimes Q^{(n)})}{d\Big((Q_X^{(m)}\times Q_Y^{(m)})\otimes(Q_X^{(n)}\times Q_Y^{(n)})\Big)}\Big)dQ^{(m+n)} \\
&= \int_{\Sigma^{m+n}\times\Sigma^{m+n}} \log\Big(\frac{dQ^{(m)}\cdot dQ^{(n)}}{d(Q_X^{(m)}\times Q_Y^{(m)})\cdot d(Q_X^{(n)}\times Q_Y^{(n)})}\Big)dQ^{(m+n)} \\
&= \int_{\Sigma^{m+n}\times\Sigma^{m+n}} \log\Big(\frac{dQ^{(m)}}{d(Q_X^{(m)}\times Q_Y^{(m)})}\cdot\frac{dQ^{(n)}}{d(Q_X^{(n)}\times Q_Y^{(n)})}\Big)dQ^{(m+n)} \\
&= \int_{\Sigma^{m+n}\times\Sigma^{m+n}} \log\Big(\frac{dQ^{(m)}}{dQ_X^{(m)}\times Q_Y^{(m)}}\Big)dQ^{(m+n)} + \int_{\Sigma^{m+n}\times\Sigma^{m+n}} \log\Big(\frac{dQ^{(n)}}{dQ_X^{(n)}\times Q_Y^{(n)}}\Big)dQ^{(m+n)} \\
&= \int_{\Sigma^m\times\Sigma^m}\int_{\Sigma^n\times\Sigma^n} \log\Big(\frac{dQ^{(m)}}{dQ_X^{(m)}\times Q_Y^{(m)}}\Big)dQ^{(n)}dQ^{(m)} + \int_{\Sigma^n\times\Sigma^n}\int_{\Sigma^m\times\Sigma^m} \log\Big(\frac{dQ^{(n)}}{dQ_X^{(n)}\times Q_Y^{(n)}}\Big)dQ^{(m)}dQ^{(n)} \\
&= \int_{\Sigma^m\times\Sigma^m} \log\Big(\frac{dQ^{(m)}}{dQ_X^{(m)}\times Q_Y^{(m)}}\Big)dQ^{(m)} + \int_{\Sigma^m\times\Sigma^m} \log\Big(\frac{dQ^{(n)}}{dQ_X^{(n)}\times Q_Y^{(n)}}\Big)dQ^{(n)} \\
&= H(Q^{(m)} \mid Q_X^{(m)}\times Q_Y^{(m)}) + H(Q^{(n)} \mid Q_X^{(n)}\times Q_Y^{(n)}),
\end{aligned}
$$

where we used the fact that product measures factorize such that $Q^{(m)}\otimes Q^{(n)}(A\times B) = Q^{(m)}(A)\cdot Q^{(n)}(B)$ (proposition 6.6 from probability theory course), logarithmic rules, linearity of integrals, and Fubini's theorem in the fifth step since we may assume that $H(Q^{(m+n)} \mid Q_X^{(m+n)}\times Q_Y^{(m+n)}) < \infty$ given the rate distortion problem. In other terms, $\log\Big(\frac{dQ^{(m+n)}}{d(Q_X^{(m+n)}\times Q_Y^{(m+n)})}\Big)$ is integrable. In the second last step we used the fact that $Q^{(n)}$ and $Q^{(m)}$ are probability measures that are equal to one when we integrate over their entire probability spaces, respectively.

Combining this observation with (35) and (36) gives us

$$
\begin{aligned}
&(m+n)R_{m+n}(D) \le H(Q^{(m+n)} \mid Q_X^{(m+n)}\times Q_Y^{(m+n)}) \\
&= H(Q^{(m)} \mid Q_X^{(m)}\times Q_Y^{(m)}) + H(Q^{(n)} \mid Q_X^{(n)}\times Q_Y^{(n)}) \le mR_m(D) + nR_n(D) + (m+n)\epsilon,
\end{aligned}
$$

which holds for all $\epsilon > 0$. Hence, letting $\epsilon \to 0$ yields the desired result. ■

b) Let us conclude that

$$
\limsup_{J\to\infty} R_J(D) < \infty \Rightarrow R(D) = \lim_{J\to\infty} R_J(D).
$$

Proof.

Suppose that $\limsup_{J\to\infty} R_J(D) < \infty$. By part a), we also know that $J\cdot R_J(D)$ is a sub-additive. Thus, we get by using sub-additivity lemma (lemma 6.1.11 [1]) that

$$
\lim_{J\to\infty} R_J(D) = \lim_{J\to\infty} \frac{J\cdot R_J(D)}{J} = \inf_{J\ge 1} R_J(D) = R(D).
$$

■

3.7 Moderate Deviations and Exact Asymptotics in $\mathbb{R}^d$

In this section, we examine moderate deviations and exact asymptotics of them in $\mathbb{R}^d$. That is, we know that Cramer's theorem deals with the tails of the empirical mean $\hat{S}_n$ of i.i.d. random variables, and on finer scale, the random variables $\sqrt{n}\hat{S}_n$ possess a limiting Normal distribution by the central limit theorem. If we consider, $\beta \in (0, 1/2)$, it can be showed that the renormalized empirical mean $n^\beta \hat{S}_n$ satisfies an LDP but always with a quadratic (Normal-like) rate function [1]. Another refinement of Cramer's theorem involves a more accurate estimate of the law μ_n of $\hat{S}_n$. Specifically, for nice set A, one seeks an estimate J_n^{-1} of $\mu_n(A)$ in the sense that $\lim_{n\to\infty} J_n \mu_n(A) = 1$ [1].

Exercise 3.7.10.

a) Let $A = [q, q+a/n]$, where in the lattice case a/d is restricted to being an integer. Let us prove that for any $a \in (0, \infty)$, both (3.7.5) and (3.7.6) hold with $J_n = \eta\sqrt{\Lambda''(\eta)2\pi n}e^{n\Lambda^*(q)}/(1-e^{-\eta a})$ [1].

Proof.

We have exactly the same setting as in proof of theorem 3.7.4 except the set is $A = [q, q + a/n]$ where a/d is restricted to being an integer. Hence, we omit here some of the proof since they are identical to proof of theorem 3.7.4 [1]. We start noting that as in setting of the proof, since $\hat{S}_n = q + \sqrt{\Lambda''(\eta)/n}W_n$, it follows that

$$\mu_n(A) = \mu_n([q, q + a/n]) = \mu_n([q, \infty)) - \mu_n([q + \frac{a}{n}, \infty)).$$

For the set $[q, \infty)$ we may apply the Bahadur and Rao theorem (theorem 3.7.4 [1]). That is, $\lim_{n\to\infty} \hat{J}_n\mu([q, \infty)) = c(q)$ where $\hat{J}_n(q) = \eta\sqrt{\Lambda''(\eta)2\pi n}e^{n\Lambda^*(q)}$, $\eta = (\Lambda^*)'(q)$, and $c(q) = 1$ for non-lattice case, and $c(q) = \frac{\eta d}{1-e^{-\eta d}}$ for lattice case.

For the set $[q + \frac{a}{n}, \infty)$ we conclude that $[q + \frac{a}{n}, \infty) \to [q, \infty)$ as $n \to \infty$ when $a \in (0, \infty)$. Therefore, we may also apply theorem 3.7.4 for the set $[q + \frac{a}{n}, \infty)$, when

$$\hat{J}_n(q + \frac{a}{n}) = (\Lambda^*)'(q + \frac{a}{n})\sqrt{2\pi n/(\Lambda^*)''(q + \frac{a}{n})}e^{n\Lambda^*(q+a/n)}$$

by remarks of theorem 3.7.4. Since η and $\Lambda''(\eta)$ are smooth, and the fact that $(\Lambda^*)'(q) = \eta$, we get the following Taylor expansions

$$\begin{aligned}
\Lambda^*(q + \frac{a}{n}) &= \Lambda^*(q) + \frac{a}{n}(\Lambda^*)'(q) + O(n^{-2}) = \Lambda^*(q) + \frac{a}{n}\eta + O(n^{-2}),\\
(\Lambda^*)'(q + \frac{a}{n}) &= (\Lambda^*)'(q) + \frac{a}{n}(\Lambda^*)''(q) + O(n^{-2}) = \eta + O(n^{-1}), \text{ and}\\
(\Lambda^*)''(q + \frac{a}{n}) &= (\Lambda^*)''(q) + \frac{a}{n}(\Lambda^*)'''(q) + O(n^{-2}) = \Lambda''(\eta) + O(n^{-1}),
\end{aligned}$$

where we absorbed terms $\frac{a}{n}(\Lambda^*)''(q)$ and $\frac{a}{n}(\Lambda^*)'''(q)$ into term $O(n^{-1})$.

Using these Taylor expansions, we get the following ratio for $\hat{J}_n(q)$ and $\hat{J}_n(q + \frac{a}{n})$:

$$\begin{aligned}
\frac{\hat{J}_n(q)}{\hat{J}_n(q + \frac{a}{n})} &= \frac{\eta\sqrt{\Lambda''(\eta)2\pi n}e^{n\Lambda^*(q)}}{(\eta + O(n^{-1}))\sqrt{(\Lambda''(\eta) + O(n^{-1}))2\pi n}e^{n\Lambda^*(q+\frac{a}{n})}}\\
&= \frac{\eta}{\eta + O(n^{-1})} \cdot \frac{\sqrt{\Lambda''(\eta)}}{\sqrt{\Lambda''(\eta) + O(n^{-1})}} \cdot e^{n\Lambda^*(q)}e^{-n(\Lambda^*(q)+\frac{a}{n}\eta+O(n^{-2}))}
\end{aligned}$$

$$= \frac{\eta}{\eta}(1+O(n^{-1}))\cdot(1+O(n^{-1}))\cdot e^{-a\eta+O(n^{-1})} = (1+O(n^{-1}))e^{-a\eta-O(n^{-1})}$$
$$= (1+O(n^{-1}))e^{-a\eta}e^{-O(n^{-1})} = (1+O(n^{-1}))e^{-a\eta}(1-O(n^{-1})+O(n^{-2}))$$
$$= (1+O(n^{-1}))(1+O(n^{-1}))e^{-\eta a} = (1+O(n^{-1}))e^{-\eta a},$$

where we used the big-O notation computational rules, binomial approximation in computing the asymptotic of non-exponential terms, and the Taylor expansion of e^x. These steps are valid for large enough n since some terms comes negligible in the given case, that is, when A is a small interval of size of order $O(\frac{\log n}{n})$ as stated in remarks [1].

Using this ratio, we can show the wanted result as

$$\hat{J}_n(q)\mu_n(A) = \hat{J}_n(q)\mu_n([q,\infty)) - \hat{J}_n(q)\mu_n([q+\frac{a}{n},\infty))$$
$$= \hat{J}_n(q)\mu_n([q,\infty)) - \frac{\hat{J}_n(q)}{\hat{J}_n(q+\frac{a}{n})}\hat{J}_n(q+\frac{a}{n})\mu_n([q+\frac{a}{n},\infty))$$
$$= \hat{J}_n(q)\mu_n([q,\infty)) - e^{-\eta a}(1+O(n^{-1}))\hat{J}_n(q+\frac{a}{n})\mu_n([q+\frac{a}{n},\infty))$$
$$= c(q)+O(n^{-1}) - e^{-\eta a}(1+O(n^{-1}))(c(q)+O(n^{-1})) = c(q)+O(n^{-1}) - e^{-\eta a}(c(q)+O(n^{-1}))$$
$$= c(q)+O(n^{-1}) - e^{-\eta a} - e^{-\eta a}O(n^{-1}) = c(q)+O(n^{-1}) - e^{-\eta a} + O(n^{-1})$$
$$= c(q)(1-e^{-\eta a}) + O(n^{-1}),$$

where we used also the computational rules of big-O. In other terms, this results implies that

$$\lim_{n\to\infty}\frac{\hat{J}_n(q)}{1-e^{-\eta a}}\mu_n(A) = c(q) = \begin{cases} 1, & \text{for non-lattice case,} \\ \frac{\eta d}{1-e^{-\eta d}}, & \text{for lattice case.} \end{cases}$$

for the set $A = [q, q+\frac{a}{n}]$ when $a \in (0,\infty)$. That is, (3.7.5) and (3.7.6) holds when

$$J_n := J_n(q) = \frac{\hat{J}_n(q)}{1-e^{-\eta a}} = \frac{\eta\sqrt{\Lambda''(\eta)2\pi n}e^{n\Lambda^*(q)}}{1-e^{-\eta a}}.$$

■

b) As a consequence of part a), let us conclude that for any set $A = [q, q+b_n)$ both (3.7.5) and (3.7.6) hold for the J_n given in theorem 3.7.4 as long as $\lim_{n\to\infty} nb_n = \infty$.

Proof.

As a consequence of part a), we know that for the set $[q, q+\frac{a}{n})$ we have

$$\lim_{n\to\infty} J_n\mu_n([q,q+\frac{a}{n})) = (1-e^{-\eta a})c(q),$$

where $c(q) = 1$ for non-lattice case, and $c(q) = \frac{\eta d}{1-e^{-\eta d}}$ for lattice case as in part a). If we set $a = nb_n$, we have for the set $A = [q, q+b_n)$ that

$$\lim_{n\to\infty} J_n\mu_n(A) = \lim_{n\to\infty} J_n\mu_n([q,q+b_n)) = \lim_{n\to\infty} J_n\mu_n([q,q+\frac{nb_n}{n}))$$
$$= \lim_{n\to\infty}(1-e^{-\eta(nb_n)})c(q) = (1-0)c(q) = c(q)$$

where we assumed that $\lim_{n\to\infty} nb_n = \infty$. Thus, (3.7.5) and (3.7.6) hold for the J_n given in theorem 3.7.4 as long as $\lim_{n\to\infty} nb_n = \infty$.

■

Exercise 3.7.11.

Let $\eta > 0$ denote the minimizer of $\Lambda(\lambda)$ and suppose that $\Lambda(\lambda) < \infty$ in some open interval around η.

a) Based on the exercise 3.7.10, let us prove that when X_1 has a non-lattice distribution, the limiting distribution of $S_n = \sum_{i=1}^n X_i$ conditioned on $\{S_n \geq 0\}$ is Exponential(η).

Proof.

Let $\eta > 0$ denote the minimizer of $\Lambda(\lambda)$ and suppose that $\Lambda(\lambda) < \infty$ in some open interval around η. Fix $x > 0$. Then, we get that the limiting distribution of S_n conditioned on $\{S_n \geq 0\}$ can be expressed as

$$\mathbb{P}[S_n \geq x \mid S_n \geq 0] = \frac{\mathbb{P}[\{S_n \geq x\} \cap \{S_n \geq 0\}]}{\mathbb{P}[\{S_n \geq 0\}]} = \frac{\mathbb{P}[0 \leq S_n \leq x]}{\mathbb{P}[S_n \geq 0]} = \frac{\mathbb{P}[0 \leq \hat{S}_n \leq \frac{x}{n}]}{\mathbb{P}[\hat{S}_n \geq 0]}$$
$$= \frac{\mu_n([0, \frac{x}{n}])}{\mu_n((0,\infty)} = \frac{\mu_n([0, \frac{x}{n}))}{\mu_n((0,\infty)} \underset{\text{thm 3.7.4}}{\overset{\text{Ex. 3.7.10}}{\sim}} \frac{1 \cdot \frac{1 - e^{-\eta x}}{J_n(0)}}{1 \cdot J_n(0)} = 1 - e^{-\eta x},$$

for the J_n given in theorem 3.7.4 [1]. Here, we used the limiting estimate of the μ_n of $\hat{S}_n$ as $n \to \infty$ which was obtained in exercise 3.7.10, and theorem 3.7.4 when $q = 0$, and $a = x > 0$. Thus, the limiting distribution of $S_n = \sum_{i=1}^n X_i$ conditioned on $\{S_n \geq 0\}$ is Exponential(η) since the cumulative distribution function (cdf) of S_n conditioned on $\{S_n \geq 0\}$ is same as cdf of Exponential(η). ■

b) Suppose now that X_1 has a lattice distribution of span d and $1 > \mathbb{P}[X_1 = 0] > 0$. Let us prove that the limiting distribution of S_n/d conditioned on $\{S_n \geq 0\}$ is Geometric(p) with $p = 1 - e^{-\eta d}$ (i.e., $\mathbb{P}[S_n = kd \mid S_n \geq 0] \to p(1-p)^k$ for $k = 0, 1, 2, \ldots$).

Proof.

Let $\eta > 0$ denote the minimizer of $\Lambda(\lambda)$ and suppose that $\Lambda(\lambda) < \infty$ in some open interval around η. Moreover, we suppose X_1 has a lattice distribution of span d, and $1 > \mathbb{P}[X_1 = 0] > 0$. Fix $k \in \mathbb{N}_0$. Then we get that the limiting distribution of S_n/d conditioned on $\{S_n \geq 0\}$ is by application of exercise 3.7.10 and theorem 3.7.4,

$$\mathbb{P}[\frac{S_n}{d} = k \mid S_n \geq 0] = \mathbb{P}[S_n = kd \mid S_n \geq 0] = \frac{\mathbb{P}[\{S_n = kd\} \cap \{S_n \geq 0\}]}{\mathbb{P}[\{S_n \geq 0\}]} = \frac{\mathbb{P}[S_n = kd]}{\mathbb{P}[S_n \geq 0]}$$
$$= \frac{\mathbb{P}[\hat{S}_n = \frac{kd}{n}]}{\mathbb{P}[\hat{S}_n \geq 0]} \sim \frac{\mu_n([\frac{kd}{n}, \frac{kd}{n} + \frac{d}{n}))}{\mu_n((0,\infty)} \sim \frac{\frac{\eta d}{1 - e^{-\eta d}}(1 - e^{-\eta d})/J_n(\frac{kd}{n})}{\frac{\eta d}{1 - e^{-\eta d}}/J_n(0)}$$
$$= (1 - e^{-\eta d}) \cdot \frac{\eta\sqrt{\Lambda''(\eta)2\pi n}e^{n\Lambda^*(0)}}{\eta\sqrt{\Lambda''(\eta)2\pi n}e^{n\Lambda^*(kd/n)}} = (1 - e^{-\eta d})\frac{\eta\sqrt{\Lambda''(\eta)2\pi n}e^{n\Lambda^*(0)}}{\eta\sqrt{\Lambda''(\eta)2\pi n}e^{n(\Lambda^*(0) + \eta\frac{kd}{n})}}$$
$$= (1 - e^{-\eta d})e^{-\eta kd} = p(1-p)^k,$$

where $p = 1 - e^{-\eta d}$. Here, we used also the Taylor expansion

$$\Lambda^*(kd/n) = \Lambda^*(0) + \frac{kd}{n}(\Lambda^*)'(0) + O(n^{-2}) \approx \Lambda^*(0) + \frac{kd}{n}\eta$$

as in exercise 3.7.10 part a) proof when $q = 0$, and hence $\eta = (\Lambda^*)'(0)$. Thus, the limiting distribution of S_n/d conditioned on $\{S_n \geq 0\}$ is Geometric(p). ■

References

[1] Amir Dembo, Ofer Zeitouni, Large Deviations Techniques and Applications, 1998, Second edition.