Central Limit Theorem and Large Deviations of the Maximum Likelihood Estimator

by

Clément Fortin

Department of Mathematics and Statistics

McGill University, Montréal

February, 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Master of Science

Abstract

We present asymptotic results for the maximum likelihood estimator of the dependence parameter arising naturally in the study of asymptotic efficiency. In particular, we demonstrate a Berry-Esseen-type estimate through a study of entropy functions. Next, we establish the large deviation principle under strict convexity assumptions and study the associated rate function. The results are shown for independent and identically distributed random variables, which we then generalize to finite state Markov chains.

Résumé

Nous présentons des résultats asymptotiques de l'estimateur du maximum de vraisemblance apparaissant naturellement dans le contexte de l'étude de l'efficacité asymptotique. Plus spécifiquement, nous démontrons un estimé de convergence à la Berry-Esseen à l'aide de la fonction d'entropie. Par la suite, nous établissons le principe de grandes déviations lorsque les fonctions d'entropie sont strictement convexes et nous étudions les propriétés de la fonction de taux. Ces résultats sont présentés pour des variables indépendantes et identiquement distribuées ainsi que pour des chaînes de Markov.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Prof. Vojkan Jakšić for his continued involvement in my mathematics training over the past two years. He has provided me with exceptional academic opportunities and has gone above and beyond in coordinating lectures and research activities at McGill and abroad. I thank him for everything and commend his commitment to training the next generation of mathematicians. Likewise, Prof. Armen Shirikyan and Dr. Renaud Raquépas have played a major role in my training, both in supervising my learnings and in making my time in the group a pleasant experience; I wish to thank them for their time and benevolence. Thanks are due to Gabriele Di Matteo, Jake Gerenraich and Anthony Jureidini for their companionship and their help during the preparation and writing of this thesis. Finally, I would like to thank my friends and family for their sympathy and limitless support.

Table of Contents

Chapter One

Introduction

Much of the field of statistical modelling is concerned with finding the probability distribution behind a given data-generating process, which would grant one who knows it the ability to calculate the probability of any future event that might occur. Although finding such probability distributions is considered practically impossible for most natural processes, one can instead follow the route of estimation theory, in which one considers the distribution to be dependent upon a set of parameters. To estimate the "true" parameter value of the process under consideration, one maximizes the likelihood that the model generated observed events. We call maximum likelihood estimator (MLE) the function that, given events, yields the argument which maximizes this likelihood. Whilst the MLE is rather unwieldy, its study is tightly linked to that of a more manageable object called the entropy function, which is minimized when evaluated at the MLE. Studying the maximum likelihood estimator thus amounts to examining the critical points of this entropy function. In this sense, the entropy function constitutes the protagonist of this work.

In 1922, Fisher published in [FR22] what would become a seminal work on the modern theory of the maximum likelihood estimation. Notably, he defined the notion of consistent estimators, which are those estimators that converge in probability to the parameter describing the experiment as the sample size increases to infinity. Under identifiability

conditions, the consistency of the maximum likelihood estimator can be derived from a uniform law of large numbers (LLN), which holds readily under moderate dependence assumptions. In evaluating this consistency, one inevitably comes across the so-called Fisher entropy (or Fisher information), defined as the variance of the entropy function's first derivative. Roughly, it gives a measure of the parameter dependence of a random variable and is thus central to the study of estimators.

Once the uniform consistency of the MLE is established, one might wonder whether the MLE also satisfies asymptotic assertions often considered in probability theory. For instance, does the central limit theorem (CLT) hold and can a convergence rate be determined? Is there a function that characterizes the exponential tail convergence of the MLE? These interrogations can both be answered positively using entropy functions. Namely, a specific convergence rate to the normal distribution can be obtained via the Berry-Esseen theorem, while exponential tail convergence of the MLE can be studied through the lens of large deviations theory, which allows for a rigorous treatment of questions akin to that aforementioned. Such a treatment is given herein.

Although our approach is comprehensive, much of what is presented below has been treated extensively by many authors. As early as 1922, Fisher mentions in [FR22] the asymptotic normality of maximum likelihood estimators for *iid* random variables. The result was later rigorously demonstrated in the influential book [Cra46] of Cramér, who followed a proof method presented by Dugué in [Dug37]. The latter also shows that the argument proving asymptotic normality of the MLE extends to Markov chains.

An optimal rate of convergence in distribution for asymptotically normal *iid* random variables can be traced back to Berry in [Ber41] and Esseen in [Ess42], who independently showed a rate of $N^{-1/2}$, for N the sample size. Via this result, Pfanzagl showed in [Pfa71] that asymptotic normality of minimum contrast estimates—a slight generalization of MLEs—is reached with rate $N^{-1/2}$ in the independent setting. This result was later generalized in [Pra73] to Markov processes satisfying Doeblin's condition.

The large deviation principle (LDP) for maximum likelihood estimators also has a rich literature, of which we by no means give an exhaustive account. Following the publication of [Bas56] by Basu, Bahadur instigated in [Bah60] the study of large deviations for consistent and asymptotically normal estimators of independent random variables. Moreover, the role of the relative entropy (or Kullback-Leibler divergence) was rapidly identified to be central to the study of exponential tail convergence of estimators, such as in [BZG80] and [Bah83]—the latter of which provides a finite state Markov chains framework. Large deviations were subsequently examined in more general settings, *e.g.*, [KK86] study the LDP for the MLE of exponential families over convex parameter spaces. More modern treatments of the MLE's large deviations, such as in [She01], have extended the analysis initiated by Basu and Bahadur to the possibly infinite-dimensional case.

Although our approach is similar to much of the arguments presented in the articles mentioned in the preceding paragraphs, the results of this thesis are in a very natural sense the continuation of the work found in the master's thesis [Mat23], which itself generalizes to the Markov case much of the *iid* results given in [Jak19, Chapter 7] using techniques of analytic perturbation theory found in [Sch12, Chapter 2].

In Chapter 2, we study asymptotics of the MLE for *iid* random variables on finite sets by building on the results established in [Jak19]. After presenting a few preliminary results in Section 2.1, we show in Section 2.2 that the MLE satisfies a central limit theorem and give a Berry-Esseen-type estimate for its convergence to the normal distribution. Although the uniform consistency result we offer through the uniform law of large numbers does not yield a sharp estimate, we manage to find a rate of $N^{-2/5}$ by optimizing our approach. In Section 2.3, we show that the maximum likelihood estimator satisfies a large deviation principle (LDP) when the entropy functions are assumed to be strictly convex functions of the parameter. We end our treatment of the *iid* MLE by studying a few properties of the LDP's associated rate function via the implicit function theorem.

In Chapter 3, we generalize our treatment to finite state Markov chains using various results proved in [Mat23]. In Section 3.1, we give a precise estimate for the MLE's consistency, which is again obtained through the uniform law of large numbers. In Section 3.2, we provide a central limit theorem for the MLE, and we prove a Berry-Esseen-type estimate. Our approach makes use of Mann's doctoral thesis [Man96, Theorem 1], in which the CLT with rate $N^{-1/2}$ is proved for countable state Markov chains via perturbations in the transition kernel. This time, we obtain an estimate of $N^{-1/4}$—worse than what we offer for *iid* measures. This is due to our use of the variance instead of the third moment in deriving a precise estimate for the uniform LLN. In Section 3.3, we study the logarithm of the spectral radius of a tilted stochastic matrix, which we refer to as the limiting cumulant-generating function (CGF). Assuming once more that the entropy functions admit strict convexity, we show that the limiting CGF is strictly convex. Further, we make use of its analyticity on the real line to compute its derivatives explicitly. After establishing the LDP for the Markovian MLE, we finish by studying the associated rate function.

The two chapters are written in a self-contained manner, in that the reader who wishes to do so can focus exclusively on handling the maximum likelihood estimator subject to the mild dependence of Markov chains. On the other hand, the reader curious to study the MLE in the more transparent setting of *iid* random variables can do so without having to deal with bivariate random variables and sequences of cumulant-generating functions.

Chapter Two

IID Measures

We start with independent and identically distributed (*iid*) random variables. This setting is the same as in [Jak19] and, as such, we base our analysis on a few results derived there. While we only use the results we require, we invite the reader to consult the reference for additional details.

2.1 Preliminaries

Let $[a, b] \subset \mathbb{R}$ be a fixed interval and let Ω be a finite set. To avoid trivialities, we assume $|\Omega| > 1$. Subsets of Ω are called events and elements $\theta \in [a, b]$ are called parameters. For $N \in \mathbb{N}$[1], let $\{P_{\theta N}\}_{\theta \in [a,b]}$ be a family of probability measures on Ω^N such that $P_{\theta N}$ is the N^{th} product measure of $P_{\theta 1} = P_\theta$. Further, we write $\mathbb{E}_{\theta N}(X)$ for the expectation value of a random variable $X \colon \Omega^N \to \mathbb{R}$ with respect to $P_{\theta N}$. Since we can always restrict the set Ω to $\overline{\Omega} = \operatorname{supp} P_\theta$, we assume without loss of generality that $P_\theta(\omega) > 0$ for all $\theta \in [a, b]$ and $\omega \in \Omega$. In addition, we assume throughout this section that $\theta \mapsto P_\theta$ is $C^3([a, b])$ and we set $\dot{P}_a \coloneqq \dot{P}_{a^+}$ and $\dot{P}_b \coloneqq \dot{P}_{b^-}$, ensuring that derivatives of functions of P_θ are defined on all of $[a, b]$. The second and third derivative are extended analogously.

[1] With the convention that $\mathbb{N} = 1, 2, \ldots$

We study the probability of events in a repeated probabilistic experiment described by an unknown parameter $\theta \in [a, b]$. Since the goal of parameter estimation is to estimate the probability measure that governs the experiment, we assume that probability measures are uniquely determined by the parameter. Explicitly,

$$\theta_1 \neq \theta_2 \implies P_{\theta_1} \neq P_{\theta_2}.$$

We refer to the latter as the identifiability property of $\{P_{\theta N}\}_{N\in\mathbb{N}}$ and assume it hereafter.

Definition 2.1.1. For $N \in \mathbb{N}$, we call maximum likelihood estimator (MLE) of order N the function $\hat{\theta}_N \colon \Omega^N \to [a, b]$ defined by

$$\hat{\theta}_N(\omega) \coloneqq \operatorname*{arg\,max}_{\theta\in[a,b]} P_{\theta N}(\omega) = \operatorname*{arg\,max}_{\theta\in[a,b]} \left(\prod_{i=1}^{N} P_\theta(\omega_i) \right).$$

Observe that the C^3 assumption on P_θ guarantees the existence of an argument of the maximum, whereas the identifiability property ensures that it is unique.

The key observation on which our work stands is that maximizing the map $\theta \mapsto P_{\theta N}(\omega)$ is equivalent to minimizing the following function.

Definition 2.1.2. Given $\omega \in \Omega$, we define the map $S_\bullet(\omega) \colon [a, b] \to \mathbb{R}$ by

$$S_\theta(\omega) \coloneqq -\log P_\theta(\omega)$$

and call it the entropy function. Moreover, we denote $S_{\theta N} = -\log P_{\theta N}$ for $N \in \mathbb{N}$.

The fact that P_θ is faithful for all $\theta \in [a, b]$ ensures that the entropy function is well-defined. Throughout, we will use the notation $\dot{f}(\theta) = \partial_\theta f(\theta)$ for derivatives of functions $f = f(\theta)$. Incidentally, we remark that the entropy function $\theta \mapsto S_\theta(\omega)$ enjoys a $C^3([a, b])$ regularity

with first and second derivative given by

$$\dot{S}_\theta = -\frac{\dot{P}_\theta}{P_\theta} \quad \text{and} \quad \ddot{S}_\theta = -\frac{\ddot{P}_\theta}{P_\theta} + \frac{\dot{P}_\theta^2}{P_\theta^2}.$$

Note that the expectation value of the first derivative of S_θ with respect to P_θ has

$$\mathbb{E}_\theta(\dot{S}_\theta) = \sum_{\omega\in\Omega} -\dot{P}_\theta(\omega) = -\partial_\theta\left(\sum_{\omega\in\Omega} P_\theta(\omega)\right) = -\partial_\theta(1) = 0$$

Moreover, $\mathbb{E}_\theta([\dot{S}_\theta]^2) = \mathrm{Var}_\theta(\dot{S}_\theta) = \mathbb{E}_\theta(\ddot{S}_\theta)$. In fact, this quantity has its own name and will play a central role in our analysis.

Definition 2.1.3. We call Fisher entropy the function $\mathcal{I}\colon [a,b] \to \mathbb{R}$ defined by

$$\mathcal{I}(\theta) \coloneqq \mathbb{E}_\theta\left([\dot{S}_\theta]^2\right) = \sum_{\omega\in\Omega} \frac{[\dot{P}_\theta(\omega)]^2}{P_\theta(\omega)}.$$

For an introductory account of the Fisher entropy in this setting, see [Jak19, Chapter 6]. Proceeding, we will assume that $\mathcal{I}(\theta)$ is nonvanishing for all $\theta \in [a,b]$. Before presenting our first result, we introduce an important function that will make sporadic appearances.

Definition 2.1.4. The relative entropy of P_θ with respect to $P_{\theta'}$ is defined by

$$S(P_\theta|P_{\theta'}) \coloneqq \sum_{\omega\in\Omega} P_\theta(\omega)\log\frac{P_\theta(\omega)}{P_{\theta'}(\omega)}.$$

Lemma 2.1.5. *For any $\theta,\theta' \in [a,b]$, we have $\mathbb{E}_\theta(S_{\theta'}) \geq \mathbb{E}_\theta(S_\theta)$ with equality if and only if $\theta = \theta'$.*

Proof. Notice that $\mathbb{E}_\theta(S_{\theta'}) - \mathbb{E}_\theta(S_\theta) = S(P_\theta|P_{\theta'})$ and by Jensen's inequality,

$$\sum_{\omega\in\Omega} P_\theta(\omega)\log\frac{P_\theta(\omega)}{P_{\theta'}(\omega)} = -\sum_{\omega\in\Omega} P_\theta(\omega)\log\frac{P_{\theta'}(\omega)}{P_\theta(\omega)} \geq -\log\left(\sum_{\omega\in\Omega} P_{\theta'}(\omega)\right) = 0$$

with equality if and only if $P_\theta = P_{\theta'}$. The result follows by the identifiability property. □

We invite the reader to consult [Jak19, Chapter 4] for additional details on the relative entropy. We now turn to an adaptation of the parametric law of large numbers discussed in [Jak19, Proposition 7.6].

Proposition 2.1.6. *Let $\theta \in [a,b]$ and $X_\theta \colon \Omega \to \mathbb{R}$ be random variables such that the maps $[a,b] \ni \theta \mapsto X_\theta(\omega)$ are continuously differentiable for all $\omega \in \Omega$. Set*

$$\mathcal{S}_{\theta N}(\omega = (\omega_1, \dots, \omega_N)) \coloneqq \sum_{k=1}^{N} X_\theta(\omega_k).$$

Then there exists a constant $\tilde{K} > 0$ such that for any $\epsilon > 0$,

$$\sup_{\theta \in [a,b]} P_{\theta N} \left\{ \omega \in \Omega^N \colon \sup_{\theta' \in [a,b]} \left| \frac{\mathcal{S}_{\theta' N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'}) \right| \geq \epsilon \right\} \leq \frac{\tilde{K}}{\epsilon^4 N^2}$$

for all $N \in \mathbb{N}$.

Proof. Let $\epsilon > 0$ and $N \in \mathbb{N}$. Further, let $\omega \in \Omega$ and $\theta, \theta' \in [a,b]$ be such that $\theta < \theta'$. By the mean-value theorem, there exists $\eta \in (\theta, \theta')$ such that

$$X_\theta(\omega) - X_{\theta'}(\omega) = \dot{X}_\eta(\omega)(\theta - \theta').$$

Since $|\Omega| < \infty$ and $[a,b] \ni \theta \mapsto X_\theta(\omega)$ is C^1, we can let $K < \infty$ be such that

$$K > \sup_{\substack{\theta \in [a,b] \\ \omega \in \Omega}} |\dot{X}_\theta(\omega)|$$

and we set $\Delta \coloneqq \frac{\epsilon}{4K} > 0$. Whenever $|\theta - \theta'| < \Delta$, we thus have $|X_\theta(\omega) - X_{\theta'}(\omega)| < \epsilon/4$ and

$$\sup_{\lambda \in [a,b]} |\mathbb{E}_\lambda(X_\theta) - \mathbb{E}_\lambda(X_{\theta'})| \leq \sup_{\lambda \in [a,b]} \sum_{\omega \in \Omega} |X_\theta(\omega) - X_{\theta'}(\omega)| \, P_\lambda(\omega) < \frac{\epsilon}{4},$$

Now, let $a = \theta'_0 < \theta'_1 < \cdots < \theta'_n = b$ be such that $\theta'_k - \theta'_{k-1} < \Delta$, for $n \coloneqq (b-a)/\Delta$, and suppose that $\omega \in \Omega^N$ satisfies $|\mathcal{S}_{\theta'_k N}(\omega)/N - \mathbb{E}_\theta(X_{\theta'_k})| < \epsilon/2$ for all $0 \leq k \leq n$. Then for any

$\theta' \in [a,b]$, there exists $0 \leq k \leq n$ such that $|\theta' - \theta'_k| < \Delta$. Thus, for all $\theta \in [a,b]$ we have

$$\begin{aligned}\left|\frac{\mathcal{S}_{\theta' N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'})\right| &\leq \left|\frac{\mathcal{S}_{\theta' N}(\omega)}{N} - \frac{\mathcal{S}_{\theta'_k N}(\omega)}{N}\right| + \left|\frac{\mathcal{S}_{\theta'_k N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'_k})\right| + \left|\mathbb{E}_\theta(X_{\theta'_k}) - \mathbb{E}_\theta(X_{\theta'})\right| \\ &< \frac{1}{N}\frac{N\epsilon}{4} + \frac{\epsilon}{2} + \frac{\epsilon}{4} = \epsilon.\end{aligned}$$

In other words,

$$\bigcap_{k=1}^{n}\left\{\omega \in \Omega^N \colon \left|\frac{\mathcal{S}_{\theta'_k N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'_k})\right| < \frac{\epsilon}{2}\right\} \subset \left\{\omega \in \Omega^N \colon \sup_{\theta' \in [a,b]}\left|\frac{\mathcal{S}_{\theta' N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'})\right| < \epsilon\right\}$$

and taking complements on both sides,

$$\left\{\omega \in \Omega^N \colon \sup_{\theta' \in [a,b]}\left|\frac{\mathcal{S}_{\theta' N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'})\right| \geq \epsilon\right\} \subset \bigcup_{k=1}^{n}\left\{\omega \in \Omega^N \colon \left|\frac{\mathcal{S}_{\theta'_k N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'_k})\right| \geq \frac{\epsilon}{2}\right\}.$$

For any $\theta \in [a,b]$, we hence have

$$\begin{aligned}&P_{\theta N}\left\{\omega \in \Omega^N \colon \sup_{\theta' \in [a,b]}\left|\frac{\mathcal{S}_{\theta' N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'})\right| \geq \epsilon\right\} \\ &\qquad \leq \sum_{k=1}^{n} P_{\theta N}\left\{\omega \in \Omega^N \colon \left|\frac{\mathcal{S}_{\theta'_k N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'_k})\right| \geq \frac{\epsilon}{2}\right\} \\ &\qquad = \sum_{k=1}^{n} P_{\theta N}\left\{\omega \in \Omega^N \colon \left|\frac{\mathcal{S}_{\theta'_k N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'_k})\right|^3 \geq \left(\frac{\epsilon}{2}\right)^3\right\} \\ &\qquad \leq \left(\frac{2}{\epsilon}\right)^3 \sum_{k=1}^{n} \mathbb{E}_{\theta N}\left(\left|\frac{\mathcal{S}_{\theta'_k N}}{N} - \mathbb{E}_\theta(X_{\theta'_k})\right|^3\right) \\ &\qquad \leq \left(\frac{2}{\epsilon}\right)^3 \frac{n}{N^3} \sup_{\theta, \theta' \in [a,b]} \sum_{\omega \in \Omega^N}\left(\sum_{j=1}^{N} |X_{\theta'}(\omega_j) - \mathbb{E}_\theta(X_{\theta'})|\right)^3 P_{\theta N}(\omega).\end{aligned}$$

Using the cumulative property of cumulants for independent random variables—which we demonstrate in Appendix A—we obtain

$$\sum_{\omega \in \Omega^N}\left(\sum_{j=1}^{N} |X_{\theta'}(\omega_j) - \mathbb{E}_\theta(X_{\theta'})|\right)^3 P_{\theta N}(\omega) = N \sum_{\omega \in \Omega} |X_{\theta'}(\omega) - \mathbb{E}_\theta(X_{\theta'})|^3 P_\theta(\omega).$$

Finally, let

$$C \coloneqq \sup_{\theta,\theta' \in [a,b]} \mathbb{E}_\theta \left(|X_{\theta'} - \mathbb{E}_\theta(X_{\theta'})|^3 \right)$$

and observe that $C < \infty$ by continuity of $\theta \mapsto X_\theta(\omega)$ and $\theta \mapsto P_\theta(\omega)$ for all $\omega \in \Omega$. Substituting $n = (b-a)/\Delta = 4K(b-a)/\epsilon$ yields

$$P_{\theta N} \left\{ \omega \in \Omega^N \colon \sup_{\theta' \in [a,b]} \left| \frac{\mathcal{S}_{\theta' N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'}) \right| \geq \epsilon \right\} \leq \frac{32KC(b-a)}{\epsilon^4 N^2}.$$

Taking the supremum over all $\theta \in [a,b]$ and setting $\tilde{K} = 32KC(b-a)$ gives the result. □

Remark 2.1.7. In Proposition 2.1.6, we can reduce the power of $\epsilon > 0$ in the estimate by one, provided we lose the uniformity condition on the parameter of the random variable. In particular, for $\theta' = \theta$ and $X_\theta = \ddot{S}_\theta$, we have

$$\begin{aligned} P_{\theta N} \left\{ \omega \in \Omega^N \colon \left| \frac{\ddot{S}_{\theta N}(\omega)}{N} - \mathcal{I}(\theta) \right| \geq \epsilon \right\} &= P_{\theta N} \left\{ \omega \in \Omega^N \colon \left| \frac{\ddot{S}_{\theta N}(\omega)}{N} - \mathcal{I}(\theta) \right|^3 \geq \epsilon^3 \right\} \\ &\leq \frac{1}{\epsilon^3} \mathbb{E}_{\theta N} \left(\left| \frac{\ddot{S}_{\theta N}(\omega)}{N} - \mathcal{I}(\theta) \right|^3 \right) \\ &= \frac{\mathbb{E}_\theta \left(|\ddot{S}_\theta - \mathcal{I}(\theta)|^3 \right)}{\epsilon^3 N^2}. \end{aligned}$$

Taking the supremum over all $\theta \in [a,b]$, we obtain

$$\sup_{\theta \in [a,b]} P_{\theta N} \left\{ \omega \in \Omega^N \colon \left| \frac{\ddot{S}_{\theta N}(\omega)}{N} - \mathcal{I}(\theta) \right| \geq \epsilon \right\} \leq \frac{\tilde{K}_1}{\epsilon^3 N^2} \tag{2.1}$$

for all $N \in \mathbb{N}$ and $\tilde{K}_1 = \mathbb{E}_\theta \left(|\ddot{S}_\theta - \mathcal{I}(\theta)|^3 \right)$.

This inequality will be used in the next section to estimate a rate of convergence for the uniform central limit theorem of the MLE. Another result which will be important to us is

the uniform consistency of the sequence of maximum likelihood estimators $(\hat{\theta}_N)_{N\in\mathbb{N}}$ obtained by making use of Proposition 2.1.6. The result is presented in [Jak19, Theorem 7.8] along with a proof containing a minor error, which we circumvent.

Theorem 2.1.8. *There exists a constant $\tilde{K}_2 > 0$ such that for any $\epsilon > 0$,*

$$\sup_{\theta\in[a,b]} P_{\theta N}\left\{\omega \in \Omega^N \colon \left|\hat{\theta}_N(\omega) - \theta\right| \geq \epsilon\right\} \leq \frac{\tilde{K}_2}{\epsilon^4 N^2}$$

for all $N \in \mathbb{N}$.

Proof. Let $\epsilon > 0$ and $I_\epsilon \coloneqq \{(\theta, \theta') \in [a,b]^2 \colon |\theta - \theta'| \geq \epsilon\}$. Since I_ϵ is compact, we can set

$$\delta_1 \coloneqq \min_{(u,v)\in I_\epsilon} \left(\mathbb{E}_u(S_v) - \mathbb{E}_u(S_u)\right).$$

By Lemma 2.1.5, we have that $\delta_1 > 0$. Let $(u^*, v^*) \in I_\epsilon$ be values at which the minimum is attained. By the mean value theorem, there exists η between u^* and v^* such that

$$\delta_1 = \mathbb{E}_{u^*}(S_{v^*}) - \mathbb{E}_{u^*}(S_{u^*}) = \left|\mathbb{E}_{u^*}(\dot{S}_\eta)\right||v^* - u^*|.$$

Letting $m \coloneqq \left|\mathbb{E}_{u^*}(\dot{S}_\eta)\right| > 0$, we have that $\delta_1 \geq m\epsilon$. Now, let

$$M = \sup_{\substack{\theta\in[a,b]\\ \omega\in\Omega}} |\dot{S}_\theta(\omega)|$$

and note that $M \geq m$. Choosing $\epsilon' \coloneqq \frac{m\epsilon}{2M}$ we have that $0 < \epsilon' < \epsilon/2$ and

$$\delta_2 \coloneqq \sup_{(u,v)\in[a,b]^2\setminus I_{\epsilon'}} \left(\mathbb{E}_u(S_v) - \mathbb{E}_u(S_u)\right) > 0$$

with $\delta_2 \leq m\epsilon/2$, by the mean value theorem. Thus, we have that $\delta_2 < \delta_1$. Furthermore, we let $\delta \coloneqq m\epsilon/4$ which satisfies $0 < \delta < m\epsilon/2 \leq \delta_1 - \delta_2$. Now, fix $\theta \in [a,b]$ and denote

$I_\epsilon(\theta) \coloneqq \{\theta' \in [a,b] \colon |\theta' - \theta| \geq \epsilon\}$. Lastly, define the sets

$$A \coloneqq \left\{ \omega \in \Omega^N \colon \sup_{\theta' \in I_\epsilon(\theta)} \left| \frac{S_{\theta' N}(\omega)}{N} - \mathbb{E}_\theta(S_{\theta'}) \right| < \frac{\delta}{2} \right\},$$
$$B \coloneqq \left\{ \omega \in \Omega^N \colon \sup_{\theta' \in [a,b] \setminus I_{\epsilon'}(\theta)} \left| \frac{S_{\theta' N}(\omega)}{N} - \mathbb{E}_\theta(S_{\theta'}) \right| < \frac{\delta}{2} \right\}.$$

For $\omega \in A$ and $\theta' \in I_\epsilon(\theta)$, we have

$$\frac{S_{\theta' N}(\omega)}{N} > \mathbb{E}_\theta(S_{\theta'}) - \frac{\delta}{2} \geq \mathbb{E}_\theta(S_\theta) + \delta_1 - \frac{\delta}{2}.$$

Similarly, for $\omega \in B$ and $\theta' \in [a,b] \setminus I_{\epsilon'}(\theta)$,

$$\frac{S_{\theta' N}(\omega)}{N} < \mathbb{E}_\theta(S_{\theta'}) + \frac{\delta}{2} \leq \mathbb{E}_\theta(S_\theta) + \delta_2 + \frac{\delta}{2}.$$

Recall that $\delta > 0$ was chosen so that $0 < \delta < \delta_1 - \delta_2$, hence $\delta_1 - \delta/2 > \delta_2 + \delta/2$. Therefore, for $\omega \in A \cap B$ with $\theta' \in I_\epsilon(\theta)$ and $\theta'' \in [a,b] \setminus I_{\epsilon'}(\theta)$, we have

$$\frac{S_{\theta' N}(\omega)}{N} > \mathbb{E}_\theta(S_\theta) + \delta_1 - \frac{\delta}{2} > \mathbb{E}_\theta(S_\theta) + \delta_2 + \frac{\delta}{2} > \frac{S_{\theta'' N}(\omega)}{N}.$$

Since $\hat{\theta}_N(\omega)$ minimizes $[a,b] \ni \theta \mapsto S_{\theta N}(\omega)$, we obtain that

$$\omega \in A \cap B \implies \left| \hat{\theta}_N(\omega) - \theta \right| < \epsilon.$$

Finally, since $\delta = m\epsilon/4$, we have

$$P_{\theta N} \left\{ \omega \in \Omega^N \colon \left| \hat{\theta}_N(\omega) - \theta \right| \geq \epsilon \right\} \leq P_{\theta N} \left\{ \omega \in \Omega^N \colon \sup_{\theta' \in [a,b]} \left| \frac{S_{\theta' N}(\omega)}{N} - \mathbb{E}_\theta(S_{\theta'}) \right| \geq \frac{m\epsilon}{8} \right\}$$

and applying Proposition 2.1.6 gives the result. □

We close this section by stating a central limit theorem for *iid* random variables.

Theorem 2.1.9. *Let $X\colon \Omega \to \mathbb{R}$ be a random variable with expectation value $\mathbb{E}(X) = \mu$ and variance $0 < \mathrm{Var}(X) = \sigma^2 < \infty$. Then for any $[A, B] \in \mathbb{R}$,*

$$\lim_{N\to\infty} P_N \left\{ \omega \in \Omega^N \colon \frac{1}{\sqrt{N\sigma^2}} \sum_{k=1}^{N} (X(\omega_k) - \mu) \in [A, B] \right\} = \int_A^B \mathrm{e}^{-x^2/2}\, \mathrm{d}x.$$

In addition, if $\mathbb{E}(|X - \mu|^3) < \infty$, then there is $K' > 0$ such that for all $[C, D] \subset \mathbb{R}$ and $N \geq 1$,

$$\left| P_N \left\{ \omega \in \Omega^N \colon \frac{1}{\sqrt{N\sigma^2}} \sum_{k=1}^{N} (X(\omega_k) - \mu) \in [C, D] \right\} - \frac{1}{\sqrt{2\pi}} \int_C^D \mathrm{e}^{-x^2/2}\, \mathrm{d}x \right| \leq \frac{K'}{\sqrt{N}}.$$

Both parts of the theorem can be found in [Fel71]. The second statement—which will be a key ingredient in providing convergence estimates for the maximum likelihood estimator—is due to [Ber41] and [Ess42] and is thus called the Berry-Esseen theorem.

2.2 Central Limit Theorem

Having laid the theoretical basis, we now turn to the central limit theorem of the MLE.

Theorem 2.2.1. *Suppose that $\theta \in (a, b)$. Then for any $[A, B] \subset \mathbb{R}$,*

$$\lim_{N\to\infty} P_{\theta N} \left\{ \omega \in \Omega^N \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B] \right\} = \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\, \mathrm{d}x.$$

Proof. Let $[A, B] \subset \mathbb{R}$, fix $\theta \in (a, b)$ and let $\epsilon > 0$ be such that $[\theta - \epsilon, \theta + \epsilon] \subset (a, b)$. Let

$$\Omega_\epsilon^N = \left\{ \omega \in \Omega^N \colon |\hat{\theta}_N(\omega) - \theta| < \epsilon \text{ and } \left| \frac{\ddot{S}_{\theta N}(\omega)}{N\mathcal{I}(\theta)} - 1 \right| < \epsilon \right\},$$

which has

$$\lim_{N\to\infty} P_{\theta N}(\Omega_\epsilon^N) = 1 \tag{2.2}$$

by Theorem 2.1.8 and (2.1). For $\omega \in \Omega_\epsilon^N$, we have $\hat{\theta}_N(\omega) \in (a,b)$ and also $\dot{S}_{\hat{\theta}_N(\omega)}(\omega) = 0$. By the mean value theorem, there is $\xi_N(\omega)$ between $\hat{\theta}_N(\omega)$ and θ such that

$$-\dot{S}_{\theta N}(\omega) = (\hat{\theta}_N(\omega) - \theta)\ddot{S}_{\xi_N(\omega)N}(\omega)$$

and hence

$$-\frac{\dot{S}_{\theta N}(\theta)}{\sqrt{N\mathcal{I}(\omega)}} = \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta)\frac{\ddot{S}_{\xi_N(\omega)N}(\omega)}{N\mathcal{I}(\theta)}. \tag{2.3}$$

Applying the mean value theorem again to $\theta \mapsto \ddot{S}_{\theta N}(\omega)$, we have

$$\frac{\ddot{S}_{\xi_N(\omega)N}(\omega) - \ddot{S}_{\theta N}(\omega)}{N\mathcal{I}(\theta)} = \frac{(\hat{\theta}_N(\omega) - \theta)\dddot{S}_{\zeta_N(\omega)N}(\omega)}{N\mathcal{I}(\theta)}$$

for some $\zeta_N(\omega)$ between θ and $\xi_N(\omega)$, thus

$$\begin{aligned}\left|\frac{\ddot{S}_{\xi_N(\omega)N}(\omega) - \ddot{S}_{\theta N}(\omega)}{N\mathcal{I}(\theta)}\right| &\le \epsilon \left(\sup_{\theta\in[a,b]} \frac{1}{\mathcal{I}(\theta)}\right)\left(\sup_{\theta\in[a,b]} \frac{\dddot{S}_{\theta N}(\omega)}{N}\right) \\ &= \epsilon \left(\sup_{\theta\in[a,b]} \frac{1}{\mathcal{I}(\theta)}\right)\left(\sup_{\substack{\theta\in[a,b]\\ \omega\in\Omega}} \left|\frac{\mathrm{d}^3}{\mathrm{d}\theta^3}\log P_\theta(\omega)\right|\right) \\ &=: \epsilon K.\end{aligned}$$

Since $\omega \in \Omega_\epsilon^N$, we have $1 - \epsilon < \frac{\ddot{S}_{\theta N}(\omega)}{N\mathcal{I}(\theta)} < 1 + \epsilon$ and summing it with

$$-K\epsilon \le \frac{\ddot{S}_{\xi_N(\omega)N}(\omega) - \ddot{S}_{\theta N}(\omega)}{N\mathcal{I}(\theta)} \le K\epsilon$$

we get

$$1 - \epsilon(K+1) < \frac{\ddot{S}_{\xi_N(\omega)N}(\omega)}{N\mathcal{I}(\theta)} < 1 + \epsilon(K+1). \tag{2.4}$$

Observe that there are three types of intervals $[A, B]$: either $0 \leq A < B$, $A < 0 \leq B$ or $A < B < 0$. In view of this, take $\epsilon > 0$ so small that

$$\epsilon(K+1)\big(|A| + |B|\big) < B - A \text{ and } 1 - \epsilon(K+1) \geq \gamma, \tag{2.5}$$

for some small $\gamma > 0$, and set

$$A_\epsilon \coloneqq \begin{cases} A(1+\epsilon(K+1)) & \text{if } 0 \leq A, \\ A(1-\epsilon(K+1)) & \text{if } A < 0, \end{cases} \qquad B_\epsilon \coloneqq \begin{cases} B(1-\epsilon(K+1)) & \text{if } 0 < B, \\ B(1+\epsilon(K+1)) & \text{if } B \leq 0. \end{cases} \tag{2.6}$$

Notice that the first condition of (2.5) ensures $A_\epsilon < B_\epsilon$ and whenever $\omega \in \Omega_\epsilon^N$ satisfies

$$A_\epsilon \leq \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta)\frac{\ddot{S}_{\xi_N(\omega)N}(\omega)}{N\mathcal{I}(\theta)} \leq B_\epsilon,$$

then $A \leq \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \leq B$ by (2.4). Making use of (2.3), we obtain

$$P_{\theta N}\left\{\omega \in \Omega_\epsilon^N \colon -\frac{\dot{S}_{\theta N}(\omega)}{\sqrt{N\mathcal{I}(\theta)}} \in [A_\epsilon, B_\epsilon]\right\} \leq P_{\theta N}\left\{\omega \in \Omega_\epsilon^N \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\}.$$

We can extend the sets to Ω^N on both sides to get

$$\begin{aligned} P_{\theta N}\left\{\omega \in \Omega^N \colon -\frac{\dot{S}_{\theta N}(\omega)}{\sqrt{N\mathcal{I}(\theta)}} \in [A_\epsilon, B_\epsilon]\right\} \\ \leq P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} + P_{\theta N}(\Omega^N \backslash \Omega_\epsilon^N) \end{aligned} \tag{2.7}$$

and taking the limit inferior, recalling (2.2) and applying Theorem 2.1.9 to the sequence $(-\dot{S}_{\theta N})_{N\in\mathbb{N}}$ with expectation $\mathbb{E}_\theta(-\dot{S}_\theta) = 0$ and variance $\mathrm{Var}_\theta(-\dot{S}_\theta) = \mathcal{I}(\theta)$,

$$\frac{1}{\sqrt{2\pi}}\int_{A_\epsilon}^{B_\epsilon} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \leq \liminf_{N\to\infty} P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\}.$$

Finally, taking $\epsilon \downarrow 0$ yields

$$\frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\,\mathrm{d}x \leq \liminf_{N\to\infty} P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\}. \tag{2.8}$$

We now show the reverse inequality holds with the limit superior. Let

$$A'_\epsilon \coloneqq \begin{cases} A(1-\epsilon(K+1)) & \text{if } 0 \leq A, \\ A(1+\epsilon(K+1)) & \text{if } A < 0, \end{cases} \qquad B'_\epsilon \coloneqq \begin{cases} B(1+\epsilon(K+1)) & \text{if } 0 < B, \\ B(1-\epsilon(K+1)) & \text{if } B \leq 0. \end{cases} \tag{2.9}$$

and note that $A'_\epsilon \leq A < B \leq B'_\epsilon$. In particular,

$$\begin{aligned} &\left\{\omega \in \Omega^N_\epsilon \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} \\ &\qquad \subset \left\{\omega \in \Omega^N_\epsilon \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta)\frac{\ddot{S}_{\xi_N(\omega)N}(\omega)}{N\mathcal{I}(\theta)} \in [A'_\epsilon, B'_\epsilon]\right\} \end{aligned}$$

by (2.4). Using (2.3) and extending both sides to Ω^N,

$$\begin{aligned} &P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} \\ &\qquad \leq P_{\theta N}\left\{\omega \in \Omega^N \colon -\frac{\dot{S}_{\theta N}(\omega)}{\sqrt{N\mathcal{I}(\theta)}} \in [A'_\epsilon, B'_\epsilon]\right\} + P_{\theta N}(\Omega^N \backslash \Omega^N_\epsilon). \end{aligned} \tag{2.10}$$

Taking the limit superior, recalling (2.2) and applying Theorem 2.1.9 to $(-\dot{S}_{\theta N})_{N\in\mathbb{N}}$,

$$\limsup_{N\to\infty} P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} \leq \frac{1}{\sqrt{2\pi}} \int_{A'_\epsilon}^{B'_\epsilon} \mathrm{e}^{-x^2/2}\,\mathrm{d}x.$$

Taking $\epsilon \downarrow 0$ and combining with (2.8) yields the desired result. □

We note that the assumption that $\theta \in (a, b)$ in the above theorem ensures that cases where the maximum likelihood estimator $\hat{\theta}_N(\omega)$ takes value a or b but $\dot{S}_{\hat{\theta}_N(\omega)N}(\omega) \neq 0$ are ruled out of our analysis, as required by our argument.

We now strengthen the above result to a uniform convergence with respect to the parameter by utilizing different results that were derived in the preliminaries section.

Theorem 2.2.2. *For any subinterval $[a', b'] \subset (a, b)$ and $N \in \mathbb{N}$ large enough, there is a constant $C > 0$ such that*

$$\sup_{\substack{\theta\in[a',b'] \\ [A,B]\subset\mathbb{R}}} \left| P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\,\mathrm{d}x \right| \leq \frac{C}{N^{2/5}}.$$

Proof. Let $[a', b'] \subset (a, b)$ and let $0 < \epsilon < \min\{\frac{a-a'}{2}, \frac{b'-b}{2}\}$ satisfy (2.5). Furthermore, let

$$\Omega_\epsilon^N(\theta) = \left\{\omega \in \Omega^N \colon |\hat{\theta}_N(\omega) - \theta| < \epsilon \text{ and } \left|\frac{\ddot{S}_{\theta N}(\omega)}{N\mathcal{I}(\theta)} - 1\right| < \epsilon\right\}.$$

By Theorem 2.1.8 and (2.1), there exist $\tilde{K}_1, \tilde{K}_2 > 0$ such that for all $N \in \mathbb{N}$,

$$\sup_{\theta\in[a,b]} P_{\theta N}(\Omega^N \backslash \Omega_\epsilon^N(\theta)) \leq \frac{\tilde{K}_2}{\epsilon^4 N^2} + \frac{\tilde{K}_1}{\epsilon^3 N^2} \tag{2.11}$$

By (2.7), we have

$$\begin{aligned} &P_{\theta N}\left\{\omega \in \Omega^N \colon -\frac{\dot{S}_{\theta N}(\omega)}{\sqrt{N\mathcal{I}(\theta)}} \in [A_\epsilon, B_\epsilon]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\mathrm{d}x \\ &\leq P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\mathrm{d}x + P_\theta(\Omega^N \backslash \Omega_\epsilon^N(\theta)), \end{aligned}$$

for $A_\epsilon < B_\epsilon$ as defined in (2.6). By Theorem 2.1.9, there exists some constant $K' > 0$ such that for all $\theta \in [a', b']$ and all intervals $[C, D] \subset \mathbb{R}$,

$$\left| P_{\theta N}\left\{\omega \in \Omega^N \colon -\frac{\dot{S}_{\theta N}(\omega)}{\sqrt{N\mathcal{I}(\theta)}} \in [C, D]\right\} - \frac{1}{\sqrt{2\pi}} \int_C^D \mathrm{e}^{-x^2/2}\mathrm{d}x \right| \leq \frac{K'}{\sqrt{N}}. \tag{2.12}$$

Thus, for all $\theta \in [a', b']$ and $[A, B] \subset \mathbb{R}$, we have

$$-\frac{K'}{\sqrt{N}} - \frac{1}{\sqrt{2\pi}} \int_{[A,B]\setminus[A_\epsilon,B_\epsilon]} \mathrm{e}^{-x^2/2} \mathrm{d}x - P_{\theta N}(\Omega^N \setminus \Omega_\epsilon^N(\theta))$$
$$\leq P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2} \, \mathrm{d}x.$$

On the other hand, by (2.10) we have

$$P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2} \, \mathrm{d}x$$
$$\leq P_{\theta N}\left\{\omega \in \Omega^N \colon -\frac{\dot{S}_{\theta N}(\omega)}{\sqrt{N\mathcal{I}(\theta)}} \in [A'_\epsilon, B'_\epsilon]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2} \, \mathrm{d}x + P_{\theta N}(\Omega^N \setminus \Omega_\epsilon^N(\theta))$$

for $A'_\epsilon < B'_\epsilon$ as defined in (2.9). Combining this with the Berry-Esseen bound of (2.12) applied to the interval $[A'_\epsilon, B'_\epsilon]$, we get

$$P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2} \, \mathrm{d}x$$
$$\leq \frac{K'}{\sqrt{N}} + \frac{1}{\sqrt{2\pi}} \int_{[A'_\epsilon,B'_\epsilon]\setminus[A,B]} \mathrm{e}^{-x^2/2} \, \mathrm{d}x + P_{\theta N}(\Omega^N \setminus \Omega_\epsilon^N(\theta))$$

and hence

$$\left|P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2} \, \mathrm{d}x\right|$$
$$\leq \frac{1}{\sqrt{2\pi}} \max\left\{\int_{[A'_\epsilon,B'_\epsilon]\setminus[A,B]} \mathrm{e}^{-x^2/2} \, \mathrm{d}x, \int_{[A,B]\setminus[A_\epsilon,B_\epsilon]} \mathrm{e}^{-x^2/2} \mathrm{d}x\right\} + \frac{K'}{\sqrt{N}} + P_{\theta N}(\Omega^N \setminus \Omega_\epsilon^N(\theta)).$$

Taking the supremum over $\theta \in [a', b']$ and all intervals $[A, B] \subset \mathbb{R}$ and making use of (2.11), we obtain

$$\sup_{\substack{\theta\in[a',b'] \\ [A,B]\subset\mathbb{R}}} \left|P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2} \, \mathrm{d}x\right|$$
$$\leq \sup_{[A,B]\subset\mathbb{R}} \frac{1}{\sqrt{2\pi}} \max\left\{\int_{[A'_\epsilon,B'_\epsilon]\setminus[A,B]} \mathrm{e}^{-x^2/2} \, \mathrm{d}x, \int_{[A,B]\setminus[A_\epsilon,B_\epsilon]} \mathrm{e}^{-x^2/2} \mathrm{d}x\right\} + \frac{K'}{\sqrt{N}} + \frac{\tilde{K}_2}{\epsilon^4 N^2} + \frac{\tilde{K}_1}{\epsilon^3 N^2}.$$

Lastly, we bound the maximum by

$$\sup_{[A,B]\subset\mathbb{R}} \frac{1}{\sqrt{2\pi}} \max\left\{ \int_{[A'_\epsilon,B'_\epsilon]\setminus[A,B]} \mathrm{e}^{-x^2/2}\,\mathrm{d}x, \int_{[A,B]\setminus[A_\epsilon,B_\epsilon]} \mathrm{e}^{-x^2/2}\mathrm{d}x \right\} \leq \sqrt{\frac{2}{e\pi}} \frac{\epsilon(K+1)}{(1-\epsilon(K+1))}$$

and refer the reader to Appendix C for a derivation of the estimate. Recall from one of our restrictions on $\epsilon > 0$ given in (2.5) that $1-\epsilon(K+1) \geq \gamma$ for some $\gamma > 0$, hence

$$\begin{aligned} \sup_{\substack{\theta\in[a',b'] \\ [A,B]\subset\mathbb{R}}} \left| P_{\theta N}\left\{ \omega\in\Omega^N \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega)-\theta)\in[A,B] \right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\,\mathrm{d}x \right| \\ \leq \epsilon\sqrt{\frac{2}{e\pi}}\frac{K+1}{\gamma} + \frac{K'}{\sqrt{N}} + \frac{\tilde{K}_2}{\epsilon^4 N^2} + \frac{\tilde{K}_1}{\epsilon^3 N^2}. \end{aligned} \tag{2.13}$$

Setting $\epsilon = N^{-z}$, it remains to find $z > 0$ that will yield the optimal rate of convergence. Since the first term is competing with the third and fourth term, and $N^{2-3z} \geq N^{2-4z}$ for $z > 0$, the optimal exponent will be one giving the same convergence rate to the first and third term. In other words, we want $z = 2-4z$ and hence $z = 2/5$. Thus, taking $N \in \mathbb{N}$ large enough that all of our restrictions on $\epsilon = N^{-2/5}$ hold, we obtain

$$\sup_{\substack{\theta\in[a',b'] \\ [A,B]\subset\mathbb{R}}} \left| P_{\theta N}\left\{ \omega\in\Omega^N \colon \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega)-\theta)\in[A,B] \right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\,\mathrm{d}x \right| \leq \frac{C}{N^{2/5}}.$$

for $C = \sqrt{\frac{2}{e\pi}}\frac{K+1}{\gamma} + K' + \tilde{K}_1 + \tilde{K}_2$. □

Remark 2.2.3. In the previous Berry-Esseen-type theorem, the factor preventing us from having a sharp decay estimate of $N^{-1/2}$, as in the second part of Theorem 2.1.9, comes from the third term of (2.13). Indeed, if one had $\sim 1/\epsilon^3 N^2$ as in the fourth term, taking $\epsilon = N^{-1/2}$ would give an optimal result. This extra factor of ϵ stems from the consistency of the MLE of Theorem 2.1.8, whose proof relies on Proposition 2.1.6. Since the latter involves a uniformity condition on the parameter of the random variable, an extra ϵ factor is introduced.

For the same reason that we took θ away from the boundary points in Theorem 2.2.1, the supremum in the uniform CLT must be taken over subintervals $[a', b'] \subset (a, b)$ instead of the whole $[a, b]$.

2.3 Large Deviations

In this section, we abandon the central limit theorem and concern ourselves with the large deviations of the MLE. Recall that a function $I : \mathbb{R} \to \mathbb{R}$ is called a rate function if it is nonnegative and lower semi-continuous on its domain, that is

$$\liminf_{x \to x_0} I(x) \geq I(x_0), \qquad \text{for all } x_0 \in \mathbb{R}.$$

Let $X : \Omega \to \mathbb{R}$ be a random variable, let $C(\alpha) \coloneqq \log \mathbb{E}(\mathrm{e}^{\alpha X})$ denote the corresponding cumulant-generating function, and observe that $C(\alpha)$ is strictly convex on $\mathbb{R}$. We shall study the Fenchel-Legendre transform of $C(\alpha)$, defined by

$$I(s) \coloneqq \sup_{\alpha \in \mathbb{R}} (\alpha s - C(\alpha)).$$

For now, we focus our attention on Cramér's theorem, given in [Jak19, Theorem 7.8].

Theorem 2.3.1. *Suppose $C(\alpha) = \log \mathbb{E}(\mathrm{e}^{\alpha X}) < \infty$ for all $\alpha \in \mathbb{R}$. Then the large deviation principle holds for $(\mathcal{S}_N/N)_{N \in \mathbb{N}}$ with rate function $I(s) = \sup_{\alpha \in \mathbb{R}} (\alpha s - C(\alpha))$. Explicitly,*

$$\lim_{N \to \infty} \frac{1}{N} \log P_N \left\{ \omega \in \Omega^N : \frac{\mathcal{S}_N(\omega)}{N} \in [A, B] \right\} = - \inf_{s \in [A,B]} I(s)$$

for any $[A, B] \subset \mathbb{R}$.

Since we will only be dealing with random variables that are defined on a finite space Ω and that are continuous with respect to the parameter $\theta \in [a, b]$, the condition on the cumulant-generating function will readily hold. Proceeding forward, we make the

additional assumption that the entropy functions

$$[a,b] \ni \theta \mapsto S_\theta(\omega) = -\log P_\theta(\omega)$$

are strictly convex for all $\omega \in \Omega$. Specifically, we will assume that

$$\ddot{S}_\theta(\omega) > 0, \text{ for all } \omega \in \Omega \text{ and } \theta \in [a,b].$$

In a sense, this assumption is the foundation of the work of this section, as it has many consequences that are essential in relating the entropy function to the MLE.

The first consequence is that the relative entropy function

$$[a,b] \ni \lambda \mapsto S(P_\theta|P_\lambda) = \sum_{\omega \in P_\theta} P_\theta(\omega) \log \frac{P_\theta(\omega)}{P_\lambda(\omega)}$$

is strictly convex. Second, for any $\lambda \in [a,b]$ there is $\omega \in \Omega$ such that $\dot{P}_\lambda(\omega) \neq 0$. Indeed, suppose this is not true for some $\lambda_0 \in [a,b]$, *i.e.*, for all $\omega \in \Omega$ we have $\dot{P}_{\lambda_0}(\omega) = 0$. Then

$$\partial_\lambda S(P_\theta|P_\lambda)|_{\lambda=\lambda_0} = -\sum_{\omega \in P_\theta} P_\theta(\omega) \frac{\dot{P}_{\lambda_0}(\omega)}{P_{\lambda_0}(\omega)} = 0$$

for any $\theta \in [a,b]$. By strict convexity, the local minima at $\lambda = \theta$ and $\lambda = \lambda_0$ are global, and we obtain a contradiction by uniqueness of the global minimum. In particular, since $\sum_{\omega \in \Omega} \dot{P}_\lambda(\omega) = 0$ for all $\lambda \in [a,b]$, then $\dot{P}_\lambda$ takes both positive and negative values on Ω. The next consequence of the convexity assumption is that for any $\lambda \in [a,b]$ and all $N \in \mathbb{N}$,

$$\left\{\omega \in \Omega^N \colon \hat{\theta}_N(\omega) \geq \lambda\right\} = \left\{\omega \in \Omega^N \colon \dot{S}_{\lambda N}(\omega) \leq 0\right\}, \tag{2.14}$$

$$\left\{\omega \in \Omega^N \colon \hat{\theta}_N(\omega) \leq \lambda\right\} = \left\{\omega \in \Omega^N \colon \dot{S}_{\lambda N}(\omega) \geq 0\right\}. \tag{2.15}$$

Let us denote by $J_\lambda^{(\theta)}$ the rate function for the sequence of random variables $(\dot{S}_{\lambda N}/N)_{N \in \mathbb{N}}$ with respect to $(P_{\theta N})_{N \in \mathbb{N}}$. In other words, denoting the cumulant-generating function by

$C_{\theta,\lambda}(\alpha) := \log \mathbb{E}_\theta\big(\mathrm{e}^{\alpha \dot{S}_\lambda}\big)$, we let

$$J_\lambda^{(\theta)}(s) := \sup_{\alpha \in \mathbb{R}} \big(\alpha s - C_{\theta,\lambda}(\alpha)\big).$$

By Theorem 2.3.1, for any interval $[A, B] \subset \mathbb{R}$,

$$\lim_{N\to\infty} \frac{1}{N} \log P_{\theta N} \left\{ \omega \in \Omega^N \colon \frac{\dot{S}_{\lambda N}(\omega)}{N} \in [A, B] \right\} = - \inf_{s \in [A,B]} J_\lambda^{(\theta)}(s).$$

In particular, making use of (2.14) and (2.15), we have

$$\begin{aligned}
\lim_{N\to\infty} \frac{1}{N} \log P_{\theta N} \left\{ \omega \in \Omega^N \colon \hat{\theta}_N(\omega) \geq \lambda \right\} &= - \inf_{s \leq 0} J_\lambda^{(\theta)}(s), \\
\lim_{N\to\infty} \frac{1}{N} \log P_{\theta N} \left\{ \omega \in \Omega^N \colon \hat{\theta}_N(\omega) \leq \lambda \right\} &= - \inf_{s \geq 0} J_\lambda^{(\theta)}(s).
\end{aligned}$$

Remark that since $J_\lambda^{(\theta)}$ is strictly convex on $\mathbb{R}$ with global minimum at $s = \mathbb{E}_\theta(\dot{S}_\lambda)$, then

$$\inf_{s \in [A,B]} J_\lambda^{(\theta)}(s) = \begin{cases} 0 & \text{if } \mathbb{E}_\theta(\dot{S}_\lambda) \in [A, B] \\ J_\lambda^{(\theta)}(A) & \text{if } A > \mathbb{E}_\theta(\dot{S}_\lambda) \\ J_\lambda^{(\theta)}(B) & \text{if } B < \mathbb{E}_\theta(\dot{S}_\lambda). \end{cases}$$

As $[a, b] \ni \lambda \mapsto S(P_\theta | P_\lambda)$ is strictly convex and $\partial_\lambda S(P_\theta | P_\lambda) = 0$ if and only if $\lambda = \theta$, it follows that for $\lambda \geq \theta$ we have $\mathbb{E}_\theta(\dot{S}_\lambda) = \partial_\lambda S(P_\theta | P_\lambda) \geq 0$, and hence

$$\lim_{N\to\infty} \frac{1}{N} \log P_{\theta N} \left\{ \omega \in \Omega^N \colon \hat{\theta}_N(\omega) \geq \lambda \right\} = -J_\lambda^{(\theta)}(0). \tag{2.16}$$

Similarly for $\lambda \leq \theta$,

$$\lim_{N\to\infty} \frac{1}{N} \log P_{\theta N} \left\{ \omega \in \Omega^N \colon \hat{\theta}_N(\omega) \leq \lambda \right\} = -J_\lambda^{(\theta)}(0). \tag{2.17}$$

Proposition 2.3.2. *The function* $[a,b] \ni \lambda \mapsto J_\lambda^{(\theta)}(0)$ *is finite, nonnegative, nonincreasing on* $[a,\theta]$*, nondecreasing on* $[\theta,b]$ *and vanishing at* $\lambda = \theta$*.*

Proof. Non-negativity is seen directly from

$$J_\lambda^{(\theta)}(0) = \sup_{\alpha\in\mathbb{R}}(-C_{\theta,\lambda}(\alpha)) \geq -C_{\theta,\lambda}(0) = 0.$$

To prove that $J_\lambda^{(\theta)}(0)$ is finite, recall that $C_{\theta,\lambda}(\alpha) \to \infty$ as $\alpha \to \pm\infty$, and that $\alpha \mapsto C_{\theta,\lambda}(\alpha)$ is real-analytic (in particular continuous) so $C_{\theta,\lambda}(\alpha) = -\infty$ is impossible. The fact that $\lambda \mapsto J_\lambda^{(\theta)}(0)$ is nonincreasing on $[a,\theta]$ and nondecreasing on $[\theta,b]$ follows from (2.16) and (2.17), respectively. Lastly, $J_\theta^{(\theta)}(0) = 0$ follows from (2.16) and the consistency of the maximum likelihood estimator. □

In fact, stronger results hold: $\lambda \mapsto J_\lambda^{(\theta)}(0)$ vanishes only at $\lambda = \theta$, is strictly decreasing on $[a,\theta]$ and strictly increasing on $[\theta,b]$. To see this, fix $\theta \in (a,b)$ and note that

$$\mathbb{R} \ni \alpha \mapsto \mathbb{E}_\theta\big(\mathrm{e}^{\alpha\dot{S}_\lambda}\big) = \sum_{\omega\in\Omega} \mathrm{e}^{-\alpha\dot{P}_\lambda(\omega)/P_\lambda(\omega)} P_\theta(\omega) \tag{2.18}$$

is strictly convex. Indeed, its second derivative has

$$\sum_{\omega\in\Omega} \left(\frac{\dot{P}_\lambda(\omega)}{P_\lambda(\omega)}\right)^2 \mathrm{e}^{-\alpha\dot{P}_\lambda(\omega)/P_\lambda(\omega)} P_\theta(\omega) > 0.$$

Since $\dot{P}_\lambda$ takes both positive and negative values on Ω, we have

$$\lim_{\alpha\to\pm\infty} \partial_\alpha \mathbb{E}_\theta\big(\mathrm{e}^{\alpha\dot{S}_\lambda}\big) = \lim_{\alpha\to\pm\infty} \sum_{\omega\in\Omega} \frac{\dot{P}_\lambda(\omega)}{P_\lambda(\omega)} \mathrm{e}^{-\alpha\dot{P}_\lambda(\omega)/P_\lambda(\omega)} P_\theta(\omega) = \mp\infty.$$

By the intermediate value theorem, there exists $\alpha_\lambda \in \mathbb{R}$ such that

$$\partial_\alpha \mathbb{E}_\theta\big(\mathrm{e}^{\alpha\dot{S}_\lambda}\big)\Big|_{\alpha=\alpha_\lambda} = \sum_{\omega\in\Omega} \frac{\dot{P}_\lambda(\omega)}{P_\lambda(\omega)} \mathrm{e}^{-\alpha_\lambda\dot{P}_\lambda(\omega)/P_\lambda(\omega)} P_\theta(\omega) = 0$$

and by strict convexity, $\alpha_\lambda \in \mathbb{R}$ is the unique minimum of (2.18). Therefore,

$$J_\lambda^{(\theta)}(0) = -\inf_{\alpha\in\mathbb{R}} C_{\theta,\lambda}(\alpha) = -\log \sum_{\omega\in\Omega} \mathrm{e}^{\alpha_\lambda \dot{S}_\lambda(\omega)} P_\theta(\omega).$$

Now, consider the function $F\colon [a,b]\times\mathbb{R}\to\mathbb{R}$ given by $F(\lambda,\alpha) \coloneqq \partial_\alpha C_{\theta,\lambda}(\alpha) = C'_{\theta,\lambda}(\alpha)$. Since $\theta \mapsto P_\theta(\omega)$ are C^3 on $[a,b]$ by assumption, then $F(\lambda,\alpha)$ is C^2 in $\lambda\in[a,b]$ and is infinitely differentiable in $\alpha\in\mathbb{R}$. It directly has $F(\lambda,\alpha_\lambda)=0$ with $\partial_\alpha F(\lambda,\alpha) = C''_{\theta,\lambda}(\alpha) > 0$ for any $(\lambda,\alpha)\in[a,b]\times\mathbb{R}$, by strict convexity of $C_{\theta,\lambda}$. By the implicit function theorem, there exists an open set on which $\lambda\mapsto\alpha_\lambda$ is C^2. Since $[a,b]\ni\lambda\mapsto\alpha_\lambda$ is unique, then it is actually C^2 on all of $[a,b]$, and so is $\lambda\mapsto J_\lambda^{(\theta)}(0)$. We can compute the derivatives of $J_\lambda^{(\theta)}(0)$ by implicit differentiation:

$$\begin{aligned}\partial_\lambda J_\lambda^{(\theta)}(0) &= -\frac{\mathbb{E}_\theta\left(\left(\dot{\alpha}_\lambda \dot{S}_\lambda + \alpha_\lambda \ddot{S}_\lambda\right)\mathrm{e}^{\alpha_\lambda \dot{S}_\lambda}\right)}{\mathbb{E}_\theta\left(\mathrm{e}^{\alpha_\lambda\dot{S}_\lambda}\right)} \\ &= -\alpha_\lambda \mathrm{e}^{J_\lambda^{(\theta)}(0)} \mathbb{E}_\theta\left(\ddot{S}_\lambda \mathrm{e}^{\alpha_\lambda \dot{S}_\lambda}\right), \end{aligned} \tag{2.19}$$

and the derivatives of $\lambda\mapsto\alpha_\lambda$ using the formula

$$\begin{aligned}\dot{\alpha}_\lambda &= -\left(\frac{\partial F(\lambda,\alpha)}{\partial\alpha}\right)^{-1}\left(\frac{\partial F(\lambda,\alpha)}{\partial\lambda}\right)\Bigg|_{\alpha=\alpha_\lambda} \\ &= \frac{\alpha_\lambda \mathbb{E}_\theta\left(\dot{S}_\lambda \mathrm{e}^{\alpha_\lambda\dot{S}_\lambda}\right)\mathbb{E}_\theta\left(\ddot{S}_\lambda \mathrm{e}^{\alpha_\lambda\dot{S}_\lambda}\right) - \mathbb{E}_\theta\left(\left(1+\alpha_\lambda\dot{S}_\lambda\right)\ddot{S}_\lambda \mathrm{e}^{\alpha_\lambda\dot{S}_\lambda}\right)\mathbb{E}_\theta\left(\mathrm{e}^{\alpha_\lambda\dot{S}_\lambda}\right)}{\mathbb{E}_\theta\left(\dot{S}_\lambda^2 \mathrm{e}^{\alpha_\lambda\dot{S}_\lambda}\right)\mathbb{E}_\theta\left(\mathrm{e}^{\alpha_\lambda\dot{S}_\lambda}\right) - \mathbb{E}_\theta\left(\dot{S}_\lambda \mathrm{e}^{\alpha_\lambda\dot{S}_\lambda}\right)^2}. \end{aligned} \tag{2.20}$$

From (2.19), we have $\alpha_\lambda = 0 \iff \partial_\lambda J_\lambda^{(\theta)}(0) = 0$, since $\ddot{S}_\lambda(\omega) > 0$ for all $\omega\in\Omega$, by assumption. Furthermore, note that

$$0 = C'_{\theta,\lambda}(\alpha_\lambda) \quad \text{and} \quad C'_{\theta,\lambda}(0) = \mathbb{E}_\theta(\dot{S}_\lambda) = -\partial_\lambda S(P_\theta|P_\lambda).$$

Therefore, $\alpha_\lambda = 0 \iff \partial_\lambda S(P_\theta|P_\lambda) = 0 \iff \theta = \lambda$ and hence $J_\lambda^{(\theta)}(0) = 0 \iff \lambda = \theta$. Moreover, $\partial_\lambda J_\lambda^{(\theta)}(0) = 0 \iff \lambda = \theta$. In particular, $\lambda\mapsto J_\lambda^{(\theta)}(0)$ is strictly decreasing on

$[a, \theta]$ and strictly increasing on $[\theta, b]$. Moreover, by (2.20), we have

$$\dot{\alpha}_\theta = -\frac{\mathbb{E}_\theta(\ddot{S}_\theta)}{\mathbb{E}_\theta(\dot{S}_\theta^2)} = -\frac{\mathcal{I}(\theta)}{\mathcal{I}(\theta)} = -1$$

and the second derivative of $\lambda \mapsto J_\lambda^{(\theta)}(0)$ yields

$$\begin{aligned}\partial_\lambda^2 J_\lambda^{(\theta)}(0) = -\dot{\alpha}_\lambda \mathrm{e}^{J_\lambda^{(\theta)}(0)} \mathbb{E}_\theta\big(\ddot{S}_\lambda \mathrm{e}^{\alpha_\lambda \dot{S}_\lambda}\big) + \alpha_\lambda^2 \mathrm{e}^{2J_\lambda^{(\theta)}(0)} \left(\mathbb{E}_\theta\big(\ddot{S}_\lambda \mathrm{e}^{\alpha_\lambda \dot{S}_\lambda}\big)\right)^2 \\ - \alpha_\lambda \mathrm{e}^{J_\lambda^{(\theta)}(0)} \mathbb{E}_\theta\left(\left(\dddot{S}_\lambda + \dot{\alpha}_\lambda \dot{S}_\lambda + \alpha_\lambda \ddot{S}_\lambda\right) \mathrm{e}^{\alpha_\lambda \dot{S}_\lambda}\right).\end{aligned}$$

Thus, we obtain $\partial_\lambda^2 J_\lambda^{(\theta)}(0)|_{\lambda=\theta} = \mathcal{I}(\theta) > 0$ and hence the map $\lambda \mapsto J_\lambda^{(\theta)}(0)$ is strictly convex around its minimum point $\lambda = \theta$, by continuity of the second derivative. Incidentally, we highlight the appearance of the Fisher entropy as the second derivative of the rate function at its minimum. In any case, extending the rate function to $J_\lambda^{(\theta)}(0) = \infty$ when $\lambda \notin [a, b]$ and utilizing (2.16) and (2.17), we derive the large deviation principle for the maximum likelihood estimator.

Theorem 2.3.3. *For any $[A, B] \subset \mathbb{R}$ and $\theta \in [a, b]$,*

$$\lim_{N\to\infty} \frac{1}{N} \log P_{\theta N}\left\{\omega \in \Omega^N : \hat{\theta}_N(\omega) \in [A, B]\right\} = -\inf_{\lambda\in[A,B]} J_\lambda^{(\theta)}(0).$$

Proof. Since $\hat{\theta}_N : \Omega \to [a, b]$ for all $N \in \mathbb{N}$, we assume without loss of generality that $[A, B] \subset [a, b]$. Let $\theta \leq A < B \leq b$. By (2.16), for any $\epsilon > 0$ there is $N_1 \in \mathbb{N}$ such that

$$P_{\theta N}\{\omega \in \Omega^N : \hat{\theta}_N(\omega) > B\} \leq P_{\theta N}\{\hat{\theta}_N(\omega) \geq B\} \leq \mathrm{e}^{-(J_B^{(\theta)}(0)-\epsilon)N}$$

for all $N \geq N_1$, and $N_2 \in \mathbb{N}$ such that

$$P_{\theta N}\{\omega \in \Omega^N : \hat{\theta}_N(\omega) \geq A\} \geq \mathrm{e}^{-(J_A^{(\theta)}(0)+\epsilon)N}$$

for all $N \geq N_2$. Thus for all $N \geq \max\{N_1, N_2\}$,

$$\frac{P_{\theta N}\{\omega \in \Omega^N : \hat{\theta}_N(\omega) > B\}}{P_{\theta N}\{\omega \in \Omega^N : \hat{\theta}_N(\omega) \geq A\}} \leq \mathrm{e}^{-(J_B^{(\theta)}(0) - J_A^{(\theta)}(0) - 2\epsilon)N}.$$

Pick $0 < 2\epsilon < J_B^{(\theta)}(0) - J_A^{(\theta)}(0)$ such that the above goes to 0 as $N \to \infty$, and hence

$$\begin{aligned} 0 &= \lim_{N\to\infty} \frac{1}{N} \log \left[1 - \frac{P_{\theta N}\{\omega \in \Omega^N : \hat{\theta}_N(\omega) > B\}}{P_{\theta N}\{\omega \in \Omega^N : \hat{\theta}_N(\omega) \geq A\}}\right] \\ &= \lim_{N\to\infty} \left(\frac{1}{N} \log P_{\theta N}\{\omega \in \Omega^N : \hat{\theta}_N(\omega) \in [A, B]\} - \frac{1}{N} \log P_{\theta N}\{\omega \in \Omega^N : \hat{\theta}_N(\omega) \geq A\}\right) \end{aligned}$$

and the result follows. The case $a \leq A < B \leq \theta$ is similar. Lastly suppose $\theta \in (A, B)$. By Theorem 2.1.8, for any $\epsilon > 0$, we have

$$\lim_{N\to\infty} \sup_{\theta\in[a,b]} P_{\theta N}\{\omega \in \Omega^N : |\hat{\theta}_N(\omega) - \theta| \geq \epsilon\} = 0.$$

Take $\epsilon = \min\{\theta - A, B - \theta\} > 0$ such that $A \leq \theta - \epsilon < \theta + \epsilon \leq B$, hence

$$1 \geq \lim_{N\to\infty} P_{\theta N}\left\{\omega \in \Omega^N : \hat{\theta}_N(\omega) \in [A, B]\right\} \geq \lim_{N\to\infty} P_{\theta N}\left\{\omega \in \Omega^N : \left|\hat{\theta}_N(\omega) - \theta\right| < \epsilon\right\} = 1.$$

Therefore,

$$\lim_{N\to\infty} \frac{1}{N} \log P_{\theta N}\left\{\omega \in \Omega^N : \hat{\theta}_N(\omega) \in [A, B]\right\} = 0 = -J_\theta^{(\theta)}(0) = -\inf_{\lambda\in[A,B]} J_\lambda^{(\theta)}(0). \qquad \square$$

We now study a special case which illustrates that the MLE rate function need not be convex everywhere, despite its corresponding random variables possessing strict convexity.

Example 2.3.4. Consider the probability measure given by the exponential families

$$P_\theta(\omega) = \mathrm{e}^{\theta H(\omega)}/Z(\theta), \quad \text{where} \quad Z(\theta) = \sum_{\omega\in\Omega} \mathrm{e}^{\theta H(\omega)}$$

and $H\colon \Omega \to \mathbb{R}$ is a non-constant function. The maps

$$\theta \mapsto S_\theta(\omega) = -\log P_\theta(\omega) = -\theta H(\omega) + \log Z(\theta)$$

are readily seen to be strictly convex:

$$\begin{aligned} \ddot{S}_\theta(\omega) &= \frac{\ddot{Z}(\theta)Z(\theta) - \dot{Z}(\theta)^2}{Z(\theta)^2} \\ &= \sum_{\omega' \in \Omega} \frac{H(\omega')^2 \mathrm{e}^{\theta H(\omega')}}{Z(\theta)} - \left(\sum_{\omega' \in \Omega} \frac{H(\omega') \mathrm{e}^{\theta H(\omega')}}{Z(\theta)} \right)^2 > 0 \end{aligned} \tag{2.21}$$

for all $\omega \in \Omega$, by Jensen's inequality. Given $\theta \in [a, b]$, let $\mathbb{R} \ni \lambda \mapsto J_\lambda^{(\theta)}(0)$ be the rate function from Theorem 2.3.3. Note that

$$\frac{\dot{P}_\lambda(\omega)}{P_\lambda(\omega)} = \frac{H(\omega)Z(\lambda) - \dot{Z}(\lambda)}{Z(\lambda)}.$$

Therefore, if $\mathcal{J}^{(\theta)}$ is the rate function associated to the sequence $(H(\omega_1) + \cdots + H(\omega_N))_{N \in \mathbb{N}}$ with respect to $P_{\theta N}$, then

$$\begin{aligned} J_\lambda^{(\theta)}(0) &= \sup_{\alpha \in \mathbb{R}} \left(-\log E_\theta\left(\mathrm{e}^{\alpha \dot{S}_\lambda}\right) \right) \\ &= \sup_{\alpha \in \mathbb{R}} \left(-\log \mathbb{E}_\theta\left(\mathrm{e}^{\alpha \dot{Z}(\lambda)/Z(\lambda)} \mathrm{e}^{-\alpha H}\right) \right) \\ &= \sup_{\alpha \in \mathbb{R}} \left(-\alpha \frac{\dot{Z}(\lambda)}{Z(\lambda)} - \log \mathbb{E}_\theta\left(\mathrm{e}^{-\alpha H}\right) \right) \\ &= \mathcal{J}^{(\theta)}\left(\dot{Z}(\lambda)/Z(\lambda)\right). \end{aligned} \tag{2.22}$$

In particular, for any interval $[A, B] \subset \mathbb{R}$, we have

$$\lim_{N \to \infty} \frac{1}{N} P_{\theta N} \left\{ \omega \in \Omega^N \colon \hat{\theta}_N(\omega) \in [A, B] \right\} = -\inf_{\lambda \in [A,B]} J_\lambda^{(\theta)}(0) = -\inf_{\lambda \in [A,B]} \mathcal{J}^{(\theta)} \left(\frac{\dot{Z}(\lambda)}{Z(\lambda)} \right).$$

We now turn to the rate function $\mathcal{J}^{(\theta)}$ and study its relation to $\lambda \mapsto J_\lambda^{(\theta)}(0)$. Let

$$m = \min_{\omega\in\Omega} H(\omega), \quad M = \max_{\omega\in\Omega} H(\omega),$$

and remark that the argument in (2.22) has

$$\frac{\dot{Z}(\lambda)}{Z(\lambda)} = \sum_{\omega\in\Omega} \frac{H(\omega)\mathrm{e}^{\lambda H(\omega)}}{Z(\lambda)} \in (m, M).$$

Further, the interval bounds are attained in the asymptotic limit. Indeed,

$$\begin{aligned}\frac{\dot{Z}(\lambda)}{Z(\lambda)} &= \frac{\sum_{\omega\in\Omega} H(\omega)\mathrm{e}^{\lambda H(\omega)}}{\sum_{\omega'\in\Omega} \mathrm{e}^{\lambda H(\omega')}} \\ &= \sum_{H(\omega)=M} \frac{M}{\sum_{\omega'\in\Omega} \mathrm{e}^{\lambda(H(\omega')-M)}} + \sum_{H(\omega)<M} \frac{H(\omega)}{\sum_{\omega'\in\Omega} \mathrm{e}^{\lambda(H(\omega')-H(\omega))}} \xrightarrow{\lambda\to\infty} M\end{aligned}$$

and

$$\begin{aligned}\frac{\dot{Z}(\lambda)}{Z(\lambda)} &= \frac{\sum_{\omega\in\Omega} H(\omega)\mathrm{e}^{\lambda H(\omega)}}{\sum_{\omega'\in\Omega} \mathrm{e}^{\lambda H(\omega')}} \\ &= \sum_{H(\omega)=m} \frac{m}{\sum_{\omega'\in\Omega} \mathrm{e}^{\lambda(H(\omega')-m)}} + \sum_{H(\omega)>m} \frac{H(\omega)}{\sum_{\omega'\in\Omega} \mathrm{e}^{\lambda(H(\omega')-H(\omega))}} \xrightarrow{\lambda\to-\infty} m.\end{aligned}$$

Applying [Jak19, Proposition 2.6], we observe that the function $\mathbb{R} \ni \lambda \mapsto J_\lambda^{(\theta)}$ is bounded with horizontal asymptotes given by

$$\begin{aligned}\lim_{\lambda\to\infty} J_\lambda^{(\theta)}(0) &= \mathcal{J}^{(\theta)}(M) = -\log P_\theta\{\omega \in \Omega\colon H(\omega) = M\}, \\ \lim_{\lambda\to-\infty} J_\lambda^{(\theta)}(0) &= \mathcal{J}^{(\theta)}(m) = -\log P_\theta\{\omega \in \Omega\colon H(\omega) = m\}.\end{aligned}$$

In particular, $\lambda \mapsto J_\lambda^{(\theta)}(0)$ cannot be convex on all of $\mathbb{R}$. Since our analysis shows that it is strictly convex around $\lambda = \theta$, the function $\lambda \mapsto J_\lambda^{(\theta)}(0)$ must have inflection points.

As a concrete example, consider the two-state system $\Omega = \{\pm 1\}$ with $H(\pm 1) = \pm 1$ and $\theta = 0$. Then $P_0(\pm 1) = \pm 1/2$ and for any λ, we have

$$\dot{Z}(\lambda)/Z(\lambda) = \tanh \lambda \quad \text{and} \quad \log \mathbb{E}_0(\mathrm{e}^{\alpha H}) = \log \cosh \alpha.$$

Therefore, the rate function has

$$J_\lambda^{(0)}(0) = \mathcal{J}^{(0)}(\tanh \lambda) = \sup_{\alpha \in \mathbb{R}} (\alpha \tanh \lambda - \log \cosh \alpha)$$

and the supremum is attained at $\alpha = \tanh^{-1} \lambda$. Thus,

$$J_\lambda^{(0)}(0) = \frac{1}{2}(1 + \tanh \lambda) \log(1 + \tanh \lambda) + \frac{1}{2}(1 - \tanh \lambda) \log(1 - \tanh \lambda) \tag{2.23}$$

with $\partial_\lambda^2 J_\lambda^{(0)}(0) = (1 - 2\lambda \tanh \lambda) \operatorname{sech}^2 \lambda$ and horizontal asymptote $-\log\{H(\pm 1)\} = \log 2$. The graph of function is shown in fig. 2.1.

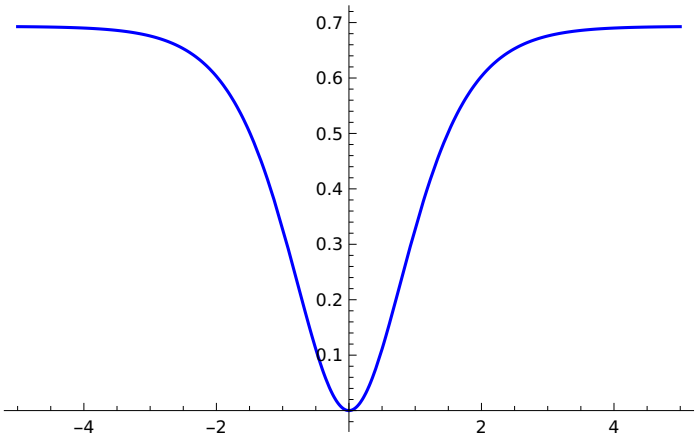

Figure 2.1: The rate function $\lambda \mapsto J_\lambda^{(0)}(0)$ given by (2.23) with horizontal asymptote of $\log 2$. The function is strictly convex around $\lambda = 0$ but has inflection points at $\lambda = \pm 0.7717$.

Chapter Three

Markov Measures

In this chapter, we adapt the results presented in Chapter 2 to the setting of finite state Markov measures by making use of results established in [Mat23].

3.1 Preliminaries

We start by introducing objects we shall work with for the rest of our analysis, as well as certain results that will prove useful to our purposes. Let $\Omega = \{1, \dots, L\}$ be a finite set with $L > 1$, and let $[a, b] \subset \mathbb{R}$ be an interval. We work with a family of irreducible and aperiodic (right) stochastic matrices $T(\theta) = [p_{ij}(\theta)] \in \mathbb{R}^{L \times L}$ whose entries are thrice-continuously differentiable maps

$$[a, b] \ni \theta \mapsto p_{ij}(\theta) \in (0, 1),$$

where $(\partial_\theta p_{ij})(a) = (\partial_\theta p_{ij})(a^+)$ and $(\partial_\theta p_{ij})(b) = (\partial_\theta p_{ij})(b^-)$. By the Perron-Frobenius theorem, there is a unique positive and invariant probability vector $\boldsymbol{p}(\theta) \in \mathbb{R}^L$ on Ω associated

to the transition matrix $T(\theta)$, in the sense that $p_i(\theta) > 0$ for any $i \in \Omega$,

$$\sum_{i\in\Omega} p_i(\theta) = 1 \quad \text{and} \quad \sum_{i\in\Omega} p_i(\theta)p_{ij}(\theta) = p_j(\theta) \quad \text{for all } j \in \Omega.$$

Moreover, the regularity of $\boldsymbol{p} = \boldsymbol{p}(\theta)$ is the same as that of $T = T(\theta)$, that is, given $\omega \in \Omega$, the maps $\theta \mapsto p_\omega(\theta)$ are $C^3([a,b])$, see [Mat23, Lemma 1.2.1]. For $N \in \mathbb{N}$, we construct the parameter-dependent Markov probability measure $P_{\theta N}$ on Ω^Nas follows. For any $N \in \mathbb{N}$[1] and any elementary event $\omega = (\omega_1, \ldots, \omega_N) \in \Omega^N$, we define the Markov measure by

$$P_{\theta N}(\omega) \coloneqq p_{\omega_1}(\theta) \prod_{k=2}^{N} p_{\omega_{k-1},\omega_k}(\theta) = p_{\omega_1}(\theta)p_{\omega_1,\omega_2}(\theta)\cdots p_{\omega_{N-1},\omega_N}(\theta) \tag{3.1}$$

with $P_\theta(\omega_1) \coloneqq P_{\theta 1}(\omega_1) = p_{\omega_1}(\theta)$ and write $\mathbb{E}_{\theta N}(X)$ for the expectation value of a random variable $X\colon \Omega^N \to \mathbb{R}$ with respect to $P_{\theta N}$. Hereafter, we make the assumption that

$$\theta_1 \neq \theta_2 \implies P_{\theta_1} \neq P_{\theta_2} \tag{3.2}$$

and refer to the latter as the identifiability property of the sequence $\{P_{\theta N}\}_{N\in\mathbb{N}}$.

The objective of this assumption is to allow for the obtainment of the "true" parameter value of the model. Specifically, it allows for a well-defined maximum argument function.

Definition 3.1.1. For $N \in \mathbb{N}$ with $N \geq 2$, we call maximum likelihood estimator (MLE) the function $\hat{\theta}_N\colon \Omega^N \to [a,b]$ defined by

$$\hat{\theta}_N(\omega) \coloneqq \arg\max_{\theta\in[a,b]} \underline{P_{\theta N}}(\omega), \tag{3.3}$$

where $\underline{P_{\theta N}}(\omega) \coloneqq p_{\omega_1,\omega_2}(\theta)\cdots p_{\omega_{N-1},\omega_N}(\theta)$.

We make the critical observation that maximizing (3.3) is equivalent to minimizing the following function.

[1] With the convention that $\mathbb{N} = 1, 2, \ldots$

Definition 3.1.2. Given $\omega_1, \omega_2 \in \Omega$, we define the function $S_\bullet(\omega_1, \omega_2)\colon [a,b] \to \mathbb{R}$ by

$$S_\theta(\omega_1, \omega_2) \coloneqq -\log \underline{P_{\theta 2}}(\omega_1, \omega_2) = -\log p_{\omega_1,\omega_2}(\theta)$$

and call it the entropy function. Moreover, for $\omega = (\omega_1, \dots, \omega_N) \in \Omega^N$ with $N \geq 2$, we write

$$S_{\theta N}(\omega) = -\log \underline{P_{\theta N}}(\omega) = -\sum_{k=1}^{N-1} \log p_{\omega_k,\omega_{k+1}}(\theta).$$

The entropy functions are well-defined since the transition matrix contains only strictly positive entries. Using the notation $\dot{f}(\theta) = \partial_\theta f(\theta)$ for derivatives of functions that depend on $\theta \in [a,b]$, we observe that the entropy functions are $C^3([a,b])$ and have

$$\begin{aligned}
\dot{S}_\theta(\omega_1, \omega_2) &= -\frac{\dot{\underline{P}}_{\theta 2}(\omega_1, \omega_2)}{\underline{P_{\theta 2}}(\omega_1, \omega_2)} = -\frac{\dot{p}_{\omega_1,\omega_2}(\theta)}{p_{\omega_1,\omega_2}(\theta)}, \\
\ddot{S}_\theta(\omega_1, \omega_2) &= -\frac{\ddot{\underline{P}}_{\theta 2}(\omega_1, \omega_2)}{\underline{P_{\theta 2}}(\omega_1, \omega_2)} + \left[\frac{\dot{\underline{P}}_{\theta 2}(\omega_1, \omega_2)}{\underline{P_{\theta 2}}(\omega_1, \omega_2)}\right]^2 = -\frac{\dot{p}_{\omega_1,\omega_2}(\theta)}{p_{\omega_1,\omega_2}(\theta)} + \left[\frac{\dot{p}_{\omega_1,\omega_2}(\theta)}{p_{\omega_1,\omega_2}(\theta)}\right]^2.
\end{aligned}$$

Although our analysis would be identical if we had defined the MLE to yield the value that, given $\omega \in \Omega^N$, maximized (3.1), it would make the notation more cumbersome.

Definition 3.1.3. We call Fisher entropy of the Markov measure generated by $(\boldsymbol{p}(\theta), T(\theta))$ the function $\mathcal{I}\colon [a,b] \to \mathbb{R}$ given by

$$\mathcal{I}(\theta) \coloneqq \mathbb{E}_{\theta 2}\left([\dot{S}_\theta]^2\right) = \sum_{\omega_1,\omega_2 \in \Omega} p_{\omega_1}(\theta) \frac{[\dot{p}_{\omega_1,\omega_2}(\theta)]^2}{p_{\omega_1,\omega_2}(\theta)}.$$

The Fisher entropy is intimately linked to estimation theory, in that, by the Cramér-Rao bound, it gives a lower bound for the variance of a parameter's estimator. For a derivation of this result for uniformly efficient consistent estimators in this setting, see [Mat23, Proposition 2.2.2]. Moreover, the derivation given in Appendix B highlights the central

role of $\mathcal{I}$ in the study of entropy functions; the special case $N = 2$ yields

$$\mathbb{E}_{\theta 2}([\dot{S}_\theta]^2) = \mathrm{Var}_{\theta 2}(\dot{S}_\theta) = \mathbb{E}_{\theta 2}(\ddot{S}_\theta) = \mathcal{I}(\theta).$$

We shall assume throughout that $\mathcal{I}$ is nonvanishing on $[a, b]$.

Before moving to our first result, we introduce another entropy function that will make important appearances, albeit infrequent.

Definition 3.1.4. The relative entropy of $P_{\theta N}$ with respect to $P_{\theta' N}$ is defined by

$$S(P_{\theta N}|P_{\theta' N}) := \sum_{\omega \in \Omega^N} P_{\theta N}(\omega) \log \frac{P_{\theta N}(\omega)}{P_{\theta' N}(\omega)}.$$

Lemma 3.1.5. *The relative entropy has $S(P_{\theta N}|P_{\theta' N}) \geq 0$, with equality if and only if $\theta = \theta'$.*

Proof. By Jensen's inequality,

$$S(P_{\theta N}|P_{\theta' N}) = -\sum_{\omega \in \Omega^N} P_{\theta N}(\omega) \log \frac{P_{\theta' N}(\omega)}{P_{\theta N}(\omega)} \geq -\log \left(\sum_{\omega \in \Omega^N} P_{\theta' N}(\omega) \right) = 0$$

with equality if and only if $P_{\theta N} = P_{\theta' N}$. The result follows by property (3.2). □

We infer that for fixed $\theta \in [a, b]$, the map $\theta' \mapsto S(P_{\theta N}|P_{\theta' N})$ attains a minimum at $\theta' = \theta$. We now turn to the uniform parametric law of large numbers given in [Mat23, Proposition 3.2.1]. Since we require a specific convergence rate different than that provided in the reference, we proceed a bit differently, starting with a generic law of large numbers.

Proposition 3.1.6. *Let $X : \Omega \to \mathbb{R}$ be a random variable and let $N \in \mathbb{N}$. Set*

$$\mathcal{S}_N(\omega = (\omega_1, \ldots, \omega_N)) := \sum_{k=1}^{N} X(\omega_k).$$

Then there exists a constant $C > 0$ such that for any $\epsilon > 0$,

$$P_N\left\{\omega \in \Omega^N \colon \left|\frac{\mathcal{S}_N(\omega)}{N} - \mathbb{E}(X)\right| \geq \epsilon\right\} \leq \frac{\mathrm{Var}(X) + C}{\epsilon^2 N}$$

for all $N \in \mathbb{N}$.

This result can be found in the proof of [Mat23, Proposition 3.1.1], where the estimate appears in the last line of the derivation, before the limit in $N \geq 1$ is taken. For random variables depending on a parameter, a similar uniform result holds.

Proposition 3.1.7. *Let $\theta \in [a, b]$ and $X_\theta \colon \Omega \to \mathbb{R}$ be a random variable such that the maps $[a, b] \ni \theta \mapsto X_\theta(\omega)$ are continuously differentiable for all $\omega \in \Omega$. Set*

$$\mathcal{S}_{\theta N}(\omega = (\omega_1, \ldots, \omega_N)) \coloneqq \sum_{k=1}^{N} X_\theta(\omega_k).$$

Then there exists a constant $\tilde{K} > 0$ such that for any $\epsilon > 0$,

$$\sup_{\theta \in [a,b]} P_{\theta N}\left\{\omega \in \Omega^N \colon \sup_{\theta' \in [a,b]} \left|\frac{\mathcal{S}_{\theta' N}(\omega)}{N} - E_\theta(X_{\theta'})\right| \geq \epsilon\right\} \leq \frac{\tilde{K}}{\epsilon^3 N}$$

for all $N \in \mathbb{N}$.

Proof. Let $\epsilon > 0$ and let $\omega \in \Omega$ and $\theta, \theta' \in [a, b]$ be such that $\theta < \theta'$. By the mean-value theorem, there exists $\eta \in (\theta, \theta')$ such that

$$X_\theta(\omega) - X_{\theta'}(\omega) = \dot{X}_\eta(\omega)(\theta - \theta').$$

Since $\theta \mapsto X_\theta$ is C^1 and the state space is finite, we can let $K < \infty$ be such that

$$K > \sup_{\substack{\theta \in [a,b] \\ \omega_1, \omega_2 \in \Omega}} |\dot{X}_\theta(\omega)|$$

and we set $\Delta = \frac{\epsilon}{4K} > 0$. Whenever $|\theta - \theta'| < \Delta$, we have $|X_\theta(\omega) - X_{\theta'}(\omega)| < \epsilon/4$ and

$$\sup_{\lambda \in [a,b]} |\mathbb{E}_\lambda(X_\theta) - \mathbb{E}_\lambda(X_{\theta'})| \leq \sup_{\lambda \in [a,b]} \sum_{\omega \in \Omega} |X_\theta(\omega) - X_{\theta'}(\omega)| P_\lambda(\omega) < \frac{\epsilon}{4}.$$

Now let $a = \theta'_0 < \theta'_1 < \cdots < \theta'_n = b$ be such that $\theta'_k - \theta'_{k-1} < \Delta$ for $n = (b-a)/\Delta$, and suppose that $\omega \in \Omega^N$ satisfies

$$\left| \frac{\mathcal{S}_{\theta'_k N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'_k}) \right| < \frac{\epsilon}{2}$$

for all $0 \leq k \leq n$. For any $\theta' \in [a,b]$, there exists $0 \leq k \leq n$ such that $|\theta' - \theta'_k| < \Delta$ and hence, for all $\theta \in [a,b]$,

$$\begin{aligned} \left| \frac{\mathcal{S}_{\theta' N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'}) \right| &\leq \left| \frac{\mathcal{S}_{\theta' N}(\omega)}{N} - \frac{\mathcal{S}_{\theta'_k N}(\omega)}{N} \right| + \left| \frac{\mathcal{S}_{\theta'_k N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'_k}) \right| + |\mathbb{E}_\theta(X_{\theta'_k}) - \mathbb{E}_\theta(X_\theta)| \\ &< \frac{1}{N}\frac{N\epsilon}{4} + \frac{\epsilon}{2} + \frac{\epsilon}{4} = \epsilon. \end{aligned}$$

Therefore,

$$\bigcap_{k=1}^{n} \left\{ \omega \in \Omega^N : \left| \frac{\mathcal{S}_{\theta'_k N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'_k}) \right| < \frac{\epsilon}{2} \right\} \subset \left\{ \omega \in \Omega^N : \sup_{\theta' \in [a,b]} \left| \frac{\mathcal{S}_{\theta' N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'}) \right| < \epsilon \right\}$$

and taking complements,

$$\left\{ \omega \in \Omega^N : \sup_{\theta' \in [a,b]} \left| \frac{\mathcal{S}_{\theta' N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'}) \right| \geq \epsilon \right\} \subset \bigcup_{k=1}^{n} \left\{ \omega \in \Omega^N : \left| \frac{\mathcal{S}_{\theta'_k N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'_k}) \right| \geq \frac{\epsilon}{2} \right\}.$$

Lastly, for any $\theta \in [a,b]$, we apply Proposition 3.1.6 to obtain

$$\begin{aligned} &P_{\theta N} \left\{ \omega \in \Omega^N : \sup_{\theta' \in [a,b]} \left| \frac{\mathcal{S}_{\theta' N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'}) \right| \geq \epsilon \right\} \\ &\qquad \leq \sum_{k=1}^{n} P_{\theta N} \left\{ \omega \in \Omega^N : \left| \frac{\mathcal{S}_{\theta'_k N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'_k}) \right| \geq \frac{\epsilon}{2} \right\} \leq \frac{4C_1 n}{\epsilon^2 N}, \end{aligned}$$

where $C_1 = \max_{\theta,\theta' \in [a,b]} \operatorname{Var}_\theta(X_{\theta'_k}) + C > 0$. Substituting $n = (b-a)/\Delta = 4K(b-a)/\epsilon$,

$$P_{\theta N}\left\{\omega \in \Omega^N : \sup_{\theta' \in [a,b]} \left|\frac{\mathcal{S}_{\theta' N}(\omega)}{N} - \mathbb{E}_\theta(X_{\theta'})\right| \geq \epsilon\right\} \leq \frac{16 C_1 K (b-a)}{\epsilon^3 N}.$$

Taking the supremum over all $\theta \in [a,b]$ and setting $\tilde{K} \coloneqq 16 C_1 K(b-a)$ yields the result. □

Remark 3.1.8. As stated, the previous two propositions only hold for univariate random variables, that is, $X \colon \Omega \to \mathbb{R}$. However, they can readily be generalized to bivariate random variables $X \colon \Omega^2 \to \mathbb{R}$ by considering an auxiliary Markov chain. To do this, we write $\mathcal{A} = \Omega^2$ and denote $\mathcal{S}_N(\omega) = \sum_{k=1}^{N-1} X(\omega_k, \omega_{k+1})$. We then consider the pair $(\overline{\boldsymbol{p}}, \overline{T})$ where $\overline{\boldsymbol{p}} = [p_a]_{a \in \mathcal{A}}$ and $\overline{T} = [p_{a,b}]_{(a,b) \in \mathcal{A}^2}$ have

$$p_{(\omega_1,\omega_2)} = p_{\omega_1} p_{\omega_1,\omega_2} \quad \text{and} \quad p_{(\omega_1,\omega_2),(\omega_3,\omega_4)} = \begin{cases} p_{\omega_3,\omega_4} & \text{if } \omega_2 = \omega_3, \\ 0 & \text{else.} \end{cases}$$

Thus, $\overline{\boldsymbol{p}} > 0$ and $\overline{\boldsymbol{p}}\overline{T} = \overline{T}$ with $\overline{T}$ an irreducible and aperiodic stochastic matrix. Denote $\overline{P}_N$ the Markov probability measure generated by the pair $(\overline{\boldsymbol{p}}, \overline{T})$. For $a \in \mathcal{A}^N$, we identify the bivariate random variable $X \colon \Omega^2 \to \mathbb{R}$ with the univariate random variable $\overline{X} \colon \mathcal{A} \to \mathbb{R}$ and we write $\overline{\mathcal{S}}_N(a) = \sum_{k=1}^{N} \overline{X}(a_k)$. Observe that

$$\overline{\mathbb{E}}_N(\mathrm{e}^{\alpha \overline{\mathcal{S}}_N}) = \mathbb{E}_{N+1}(\mathrm{e}^{\alpha \mathcal{S}_{N+1}})$$

for all $\alpha \in \mathbb{C}$ and $N \geq 1$. It is this correspondence that allows the extension to bivariate random variables, and we refer the reader to [Mat23, Section 3.2] for explicit computations. Letting $\theta \in [a,b]$, we can apply Proposition 3.1.6 to the sequence of second derivatives $(\ddot{S}_{\theta N})_{N \in \mathbb{N}}$ to obtain

$$P_{\theta N}\left\{\omega \in \Omega^N : \left|\frac{\ddot{S}_{\theta N}(\omega)}{N-1} - \mathcal{I}(\theta)\right| \geq \epsilon\right\} \leq \frac{\operatorname{Var}_{\theta 2}(\ddot{S}_\theta) + C}{\epsilon^2 (N-1)}.$$

Since $N/(N-1) \leq 2$, then for all $\epsilon > 0$ and $N \geq 2$,

$$\sup_{\theta \in [a,b]} P_{\theta N} \left\{ \omega \in \Omega^N : \left| \frac{\ddot{S}_{\theta N}(\omega)}{N-1} - \mathcal{I}(\theta) \right| \geq \epsilon \right\} \leq \frac{\tilde{K}_1}{\epsilon^2 N}, \tag{3.4}$$

where $\tilde{K}_1 \coloneqq 2 \sup_{\theta \in [a,b]} \mathrm{Var}_{\theta 2}(\ddot{S}_\theta) + C$.

Using the previous remark, we extend Proposition 3.1.7 to the bivariate case.

Proposition 3.1.9. *Let $\theta \in [a,b]$ and $X_\theta \colon \Omega^2 \to \mathbb{R}$ be a random variable such that the maps $[a,b] \ni \theta \mapsto X_\theta(\omega_1, \omega_2)$ are C^1 for all $\omega_1, \omega_2 \in \Omega$. Set*

$$\mathcal{S}_{\theta N}(\omega) = \sum_{k=1}^{N-1} X_\theta(\omega_k, \omega_{k+1}).$$

Then there exists a constant $\tilde{K} > 0$ such that for any $\epsilon > 0$,

$$\sup_{\theta \in [a,b]} P_{\theta N} \left\{ \omega \in \Omega^N : \sup_{\theta' \in [a,b]} \left| \frac{\mathcal{S}_{\theta' N}(\omega)}{N-1} - \mathbb{E}_{\theta 2}(X_{\theta'}) \right| \geq \epsilon \right\} \leq \frac{\tilde{K}}{\epsilon^3 N}$$

for all $N \geq 2$.

This allows us to provide a uniform consistency estimate for the Markovian MLE.

Theorem 3.1.10. *There exist a constant $\tilde{K}_2 > 0$ such that for any $\epsilon > 0$,*

$$\sup_{\theta \in [a,b]} P_{\theta N} \left\{ \omega \in \Omega^N : \left| \hat{\theta}_N(\omega) - \theta \right| \geq \epsilon \right\} \leq \frac{\tilde{K}_2}{\epsilon^3 N},$$

This result is an adaptation of [Mat23, Theorem 4.2.1] in that it gives an explicit bound in terms of $\epsilon > 0$ and $N \in \mathbb{N}$. Its proof follows the one given in the reference, except that we apply Proposition 3.1.9 in the last step to obtain our estimate.

Before moving on to the next section, we state the central limit theorem via a Berry-Esseen-type estimate for discrete Markov chains.

Theorem 3.1.11 ([Man96, Theorem 1]). *Let $X_N\colon \Omega^2 \to \mathbb{R}$ be a Markov chain with $N \geq 1$ an integer. Denote $\mathcal{S}_N = \sum_{k=1}^{N} X_k$ and suppose*

$$\sigma^2 = \lim_{N\to\infty} \frac{\mathrm{Var}_N(\mathcal{S}_N)}{N} < \infty.$$

Then there is a constant $K' > 0$ such that for all $[C, D] \subset \mathbb{R}$ and $N \geq 1$,

$$\left| P_N \left\{ \omega \in \Omega^N \colon \frac{1}{\sqrt{N\sigma^2}}(\mathcal{S}_N - \mathbb{E}(\mathcal{S}_N)) \in [C, D] \right\} - \frac{1}{\sqrt{2\pi}} \int_C^D \mathrm{e}^{-x^2/2}\,\mathrm{d}x \right| \leq \frac{K'}{\sqrt{N}}.$$

3.2 Central Limit Theorem

In this section we turn to the central limit theorem of the sequence of maximum likelihood estimators $(\hat{\theta}_N)_{N\in\mathbb{N}}$ and we recall that

$$\mathbb{E}_{\theta N}(-\dot{S}_{\theta N}) = 0 \quad \text{and} \quad \mathrm{Var}_{\theta N}(-\dot{S}_{\theta N}) = (N-1)\mathcal{I}(\theta),$$

for all $N \in \mathbb{N}$, as shown in Appendix B.

Theorem 3.2.1. *Let $\theta \in (a, b)$ and let $[A, B] \subset \mathbb{R}$, then*

$$\lim_{N\to\infty} P_{\theta N} \left\{ \omega \in \Omega^N \colon \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B] \right\} = \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\,\mathrm{d}x. \tag{3.5}$$

Proof. Let $[A, B] \subset \mathbb{R}$, fix $\theta \in (a, b)$ and let $\epsilon > 0$ be such that $[\theta - \epsilon, \theta + \epsilon] \subset (a, b)$. Furthermore, let $N \geq 2$ be an integer and denote

$$\Omega_\epsilon^N = \left\{ \omega \in \Omega^N \colon |\hat{\theta}_N(\omega) - \theta| < \epsilon \text{ and } \left| \frac{\ddot{S}_{\theta N}(\omega)}{(N-1)\mathcal{I}(\theta)} - 1 \right| < \epsilon \right\},$$

which has

$$\lim_{N\to\infty} P_{\theta N}(\Omega_\epsilon^N) = 1 \tag{3.6}$$

by Theorem 3.1.10 and (3.4). For $\omega \in \Omega_\epsilon^N$, we have $\hat{\theta}_N(\omega) \in (a,b)$ and also $\dot{S}_{\hat{\theta}_N(\omega)}(\omega) = 0$. By the mean value theorem, there is $\xi_N(\omega)$ between $\hat{\theta}_N(\omega)$ and θ such that

$$-\dot{S}_{\theta N}(\omega) = (\hat{\theta}_N(\omega) - \theta)\ddot{S}_{\xi_N(\omega)N}(\omega)$$

and hence

$$-\frac{\dot{S}_{\theta N}(\theta)}{\sqrt{(N-1)\mathcal{I}(\omega)}} = \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta)\frac{\ddot{S}_{\xi_N(\omega)N}(\omega)}{(N-1)\mathcal{I}(\theta)}. \tag{3.7}$$

Applying the mean value theorem again to $\theta \mapsto \ddot{S}_{\theta N}(\omega)$, we have

$$\frac{\ddot{S}_{\xi_N(\omega)N}(\omega) - \ddot{S}_{\theta N}(\omega)}{(N-1)\mathcal{I}(\theta)} = \frac{(\hat{\theta}_N(\omega) - \theta)\dddot{S}_{\zeta_N(\omega)N}(\omega)}{(N-1)\mathcal{I}(\theta)}$$

for some $\zeta_N(\omega)$ between θ and $\xi_N(\omega)$, thus

$$\begin{aligned}\left|\frac{\ddot{S}_{\xi_N(\omega)N}(\omega) - \ddot{S}_{\theta N}(\omega)}{(N-1)\mathcal{I}(\theta)}\right| &\le \epsilon \left(\sup_{\theta\in[a,b]} \frac{1}{\mathcal{I}(\theta)}\right)\left(\sup_{\theta\in[a,b]} \frac{|\dddot{S}_{\theta N}(\omega)|}{N-1}\right) \\ &= \epsilon \left(\sup_{\theta\in[a,b]} \frac{1}{\mathcal{I}(\theta)}\right)\left(\sup_{\substack{\theta\in[a,b]\\ \omega_1,\omega_2\in\Omega}} \left|\frac{\mathrm{d}^3}{\mathrm{d}\theta^3}\log p_{\omega_1,\omega_2}(\theta)\right|\right) \\ &=: \epsilon K.\end{aligned}$$

Since $\omega \in \Omega_\epsilon^N$, we have $1 - \epsilon < \frac{\ddot{S}_{\theta N}(\omega)}{(N-1)\mathcal{I}(\theta)} < 1 + \epsilon$ and summing it with

$$-K\epsilon \le \frac{\ddot{S}_{\xi_N(\omega)N}(\omega) - \ddot{S}_{\theta N}(\omega)}{(N-1)\mathcal{I}(\theta)} \le K\epsilon$$

we get

$$1 - \epsilon(K+1) < \frac{\ddot{S}_{\xi_N(\omega)N}(\omega)}{(N-1)\mathcal{I}(\theta)} < 1 + \epsilon(K+1). \tag{3.8}$$

Observe that there are three types of intervals $[A, B]$: either $0 \leq A < B$, $A < 0 \leq B$ or $A < B < 0$. In view of this, take $\epsilon > 0$ so small that

$$\epsilon(K+1)\big(|A| + |B|\big) < B - A \text{ and } 1 - \epsilon(K+1) \geq \gamma, \tag{3.9}$$

for some small $\gamma > 0$, and set

$$A_\epsilon \coloneqq \begin{cases} A(1+\epsilon(K+1)) & \text{if } 0 \leq A, \\ A(1-\epsilon(K+1)) & \text{if } A < 0, \end{cases} \qquad B_\epsilon \coloneqq \begin{cases} B(1-\epsilon(K+1)) & \text{if } 0 < B, \\ B(1+\epsilon(K+1)) & \text{if } B \leq 0. \end{cases} \tag{3.10}$$

Notice that the first condition of (3.9) ensures $A_\epsilon < B_\epsilon$ and whenever $\omega \in \Omega_\epsilon^N$ satisfies

$$A_\epsilon \leq \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta)\frac{\ddot{S}_{\xi_N(\omega)N}(\omega)}{(N-1)\mathcal{I}(\theta)} \leq B_\epsilon,$$

then $A \leq \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \leq B$ by (3.8). Making use of (3.7), we obtain

$$\begin{aligned} P_{\theta N}\left\{\omega \in \Omega_\epsilon^N \colon -\frac{\dot{S}_{\theta N}(\omega)}{\sqrt{(N-1)\mathcal{I}(\theta)}} \in [A_\epsilon, B_\epsilon]\right\} \\ \leq P_{\theta N}\left\{\omega \in \Omega_\epsilon^N \colon \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\}. \end{aligned}$$

We can extend the sets to Ω^N on both sides to get

$$\begin{aligned} P_{\theta N}\left\{\omega \in \Omega^N \colon -\frac{\dot{S}_{\theta N}(\omega)}{\sqrt{(N-1)\mathcal{I}(\theta)}} \in [A_\epsilon, B_\epsilon]\right\} \\ \leq P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} + P_{\theta N}(\Omega^N \backslash \Omega_\epsilon^N). \end{aligned} \tag{3.11}$$

Taking the limit inferior, recalling (3.6) and applying Theorem 3.1.11 to the sequence $(-\dot{S}_{\theta N})_{N\in\mathbb{N}}$ with expectation $\mathbb{E}_{\theta 2}(-\dot{S}_\theta) = 0$ and variance $\mathrm{Var}_{\theta 2}(-\dot{S}_\theta) = \mathcal{I}(\theta)$,

$$\frac{1}{\sqrt{2\pi}}\int_{A_\epsilon}^{B_\epsilon} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \leq \liminf_{N\to\infty} P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\}.$$

Finally, taking $\epsilon \downarrow 0$ yields

$$\frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\,\mathrm{d}x \leq \liminf_{N\to\infty} P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\}. \quad (3.12)$$

We now show the reverse inequality holds with the limit superior. Let

$$A'_\epsilon \coloneqq \begin{cases} A(1-\epsilon(K+1)) & \text{if } 0 \leq A, \\ A(1+\epsilon(K+1)) & \text{if } A < 0, \end{cases} \qquad B'_\epsilon \coloneqq \begin{cases} B(1+\epsilon(K+1)) & \text{if } 0 < B, \\ B(1-\epsilon(K+1)) & \text{if } B \leq 0. \end{cases} \quad (3.13)$$

and note that $A'_\epsilon \leq A < B \leq B'_\epsilon$. In particular,

$$\begin{aligned} &\left\{\omega \in \Omega^N_\epsilon \colon \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} \\ &\qquad \subset \left\{\omega \in \Omega^N_\epsilon \colon \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta)\frac{\ddot{S}_{\xi_N(\omega)N}(\omega)}{(N-1)\mathcal{I}(\theta)} \in [A'_\epsilon, B'_\epsilon]\right\} \end{aligned}$$

by (3.8). Using (3.7) and extending both sides to Ω^N,

$$\begin{aligned} &P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} \\ &\qquad \leq P_{\theta N}\left\{\omega \in \Omega^N \colon -\frac{\dot{S}_{\theta N}(\omega)}{\sqrt{(N-1)\mathcal{I}(\theta)}} \in [A'_\epsilon, B'_\epsilon]\right\} + P_{\theta N}(\Omega^N \backslash \Omega^N_\epsilon). \end{aligned} \quad (3.14)$$

Taking the limit superior, recalling (3.6) and applying Theorem 3.1.11 to $(-\dot{S}_{\theta N})_{N\in\mathbb{N}}$,

$$\limsup_{N\to\infty} P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} \leq \frac{1}{\sqrt{2\pi}} \int_{A'_\epsilon}^{B'_\epsilon} \mathrm{e}^{-x^2/2}\,\mathrm{d}x.$$

Taking $\epsilon \downarrow 0$ and combining with (3.12) yields the desired result. □

The assumption that $\theta \in (a, b)$ in Theorem 3.2.1 ensures that cases where the maximum likelihood estimator takes value a or b but $\dot{S}_{\hat{\theta}_N(\omega)N}(\omega) \neq 0$ are discarded from our analysis, as required by our argument.

We now strengthen the above result to a uniform convergence with respect to the parameter and we provide a rough estimate for the convergence rate by making use of different convergence estimates previously established.

Theorem 3.2.2. *For any $[a', b'] \subset (a, b)$ and $N \in \mathbb{N}$ large enough, there is $C > 0$ such that*

$$\sup_{\substack{\theta\in[a',b'] \\ [A,B]\subset\mathbb{R}}} \left| P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\,\mathrm{d}x \right| \leq \frac{C}{N^{1/4}}.$$

Proof. Let $[a', b'] \subset (a, b)$ and let $0 < \epsilon < \min\{\frac{a-a'}{2}, \frac{b'-b}{2}\}$ satisfy (3.9). Furthermore, let

$$\Omega_\epsilon^N(\theta) = \left\{\omega \in \Omega^N \colon |\hat{\theta}_N(\omega) - \theta| < \epsilon \text{ and } \left|\frac{\ddot{S}_{\theta N}(\omega)}{(N-1)\mathcal{I}(\theta)} - 1\right| < \epsilon\right\}.$$

By Theorem 3.1.10 and (3.4), there exist $\tilde{K}_1, \tilde{K}_2 > 0$ such that for all $N \in \mathbb{N}$,

$$\sup_{\theta\in[a,b]} P_{\theta N}(\Omega^N \backslash \Omega_\epsilon^N(\theta)) \leq \frac{\tilde{K}_2}{\epsilon^3 N} + \frac{\tilde{K}_1}{\epsilon^2 N} \tag{3.15}$$

By (3.11), we have

$$\begin{aligned}
&P_{\theta N}\left\{\omega \in \Omega^N \colon -\frac{\dot{S}_{\theta N}(\omega)}{\sqrt{(N-1)\mathcal{I}(\theta)}} \in [A_\epsilon, B_\epsilon]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\mathrm{d}x \\
&\leq P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\mathrm{d}x \\
&\quad + P_{\theta N}(\Omega^N \backslash \Omega_\epsilon^N(\theta)),
\end{aligned}$$

for $A_\epsilon < B_\epsilon$ as defined in (3.10). By Theorem 3.1.11, there exists some constant $K' > 0$ such that for all $\theta \in [a', b']$ and all intervals $[C, D] \subset \mathbb{R}$,

$$\left| P_{\theta N}\left\{\omega \in \Omega^N \colon -\frac{\dot{S}_{\theta N}(\omega)}{\sqrt{(N-1)\mathcal{I}(\theta)}} \in [C, D]\right\} - \frac{1}{\sqrt{2\pi}} \int_C^D \mathrm{e}^{-x^2/2}\mathrm{d}x \right| \leq \frac{K'}{\sqrt{N}}. \tag{3.16}$$

Thus, for all $\theta \in [a', b']$ and $[A, B] \subset \mathbb{R}$, we have

$$
\begin{aligned}
&-\frac{K'}{\sqrt{N}} - \frac{1}{\sqrt{2\pi}} \int_{[A,B]\setminus[A_\epsilon,B_\epsilon]} \mathrm{e}^{-x^2/2} \mathrm{d}x - P_{\theta N}(\Omega^N \setminus \Omega_\epsilon^N(\theta)) \\
&\qquad \le P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\, \mathrm{d}x.
\end{aligned}
$$

On the other hand, by (3.14) we have

$$
\begin{aligned}
&P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\, \mathrm{d}x \\
&\le P_{\theta N}\left\{\omega \in \Omega^N \colon -\frac{\dot{S}_{\theta N}(\omega)}{\sqrt{(N-1)\mathcal{I}(\theta)}} \in [A'_\epsilon, B'_\epsilon]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\, \mathrm{d}x + P_{\theta N}(\Omega^N \setminus \Omega_\epsilon^N(\theta))
\end{aligned}
$$

for $A'_\epsilon < B'_\epsilon$ as defined in (3.13). Combining this with the Berry-Esseen bound of (3.16) applied to the interval $[A'_\epsilon, B'_\epsilon]$, we get

$$
\begin{aligned}
&P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\, \mathrm{d}x \\
&\qquad \le \frac{K'}{\sqrt{N}} + \frac{1}{\sqrt{2\pi}} \int_{[A'_\epsilon,B'_\epsilon]\setminus[A,B]} \mathrm{e}^{-x^2/2}\, \mathrm{d}x + P_{\theta N}(\Omega^N \setminus \Omega_\epsilon^N(\theta))
\end{aligned}
$$

and hence

$$
\begin{aligned}
&\left|P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\, \mathrm{d}x\right| \\
&\le \frac{1}{\sqrt{2\pi}} \max\left\{\int_{[A'_\epsilon,B'_\epsilon]\setminus[A,B]} \mathrm{e}^{-x^2/2}\, \mathrm{d}x, \int_{[A,B]\setminus[A_\epsilon,B_\epsilon]} \mathrm{e}^{-x^2/2} \mathrm{d}x\right\} + \frac{K'}{\sqrt{N}} + P_{\theta N}(\Omega^N \setminus \Omega_\epsilon^N(\theta)).
\end{aligned}
$$

Taking the supremum over $\theta \in [a', b']$ and all intervals $[A, B] \subset \mathbb{R}$ and making use of (3.15), we obtain

$$
\begin{aligned}
&\sup_{\substack{\theta\in[a',b'] \\ [A,B]\subset\mathbb{R}}} \left|P_{\theta N}\left\{\omega \in \Omega^N \colon \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega) - \theta) \in [A, B]\right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\, \mathrm{d}x\right| \\
&\le \sup_{[A,B]\subset\mathbb{R}} \frac{1}{\sqrt{2\pi}} \max\left\{\int_{[A'_\epsilon,B'_\epsilon]\setminus[A,B]} \mathrm{e}^{-x^2/2}\, \mathrm{d}x, \int_{[A,B]\setminus[A_\epsilon,B_\epsilon]} \mathrm{e}^{-x^2/2} \mathrm{d}x\right\} + \frac{K'}{\sqrt{N}} + \frac{\tilde{K}_2}{\epsilon^3 N} + \frac{\tilde{K}_1}{\epsilon^2 N}.
\end{aligned}
$$

Lastly, we bound the maximum by

$$\sup_{[A,B]\subset\mathbb{R}} \frac{1}{\sqrt{2\pi}} \max\left\{ \int_{[A'_\epsilon,B'_\epsilon]\setminus[A,B]} \mathrm{e}^{-x^2/2}\,\mathrm{d}x, \int_{[A,B]\setminus[A_\epsilon,B_\epsilon]} \mathrm{e}^{-x^2/2}\mathrm{d}x \right\} \leq \sqrt{\frac{2}{e\pi}} \frac{\epsilon(K+1)}{(1-\epsilon(K+1))}$$

and refer the reader to Appendix C for a derivation of the estimate. Recall from one of our restrictions on $\epsilon > 0$ given in (3.9) that $1-\epsilon(K+1) \geq \gamma$ for some $\gamma > 0$, hence

$$\sup_{\substack{\theta\in[a',b']\\ [A,B]\subset\mathbb{R}}} \left| P_{\theta N}\left\{ \omega\in\Omega^N : \sqrt{(N-1)\mathcal{I}(\theta)}(\hat{\theta}_N(\omega)-\theta)\in[A,B] \right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\,\mathrm{d}x \right|$$
$$\leq \epsilon\sqrt{\frac{2}{e\pi}}\frac{K+1}{\gamma} + \frac{K'}{\sqrt{N}} + \frac{\tilde{K}_2}{\epsilon^3 N} + \frac{\tilde{K}_1}{\epsilon^2 N}.$$

Setting $\epsilon = N^{-z}$, it remains to find $z > 0$ that will yield the optimal rate of convergence. Since the first term is competing with the third and fourth term, and $N^{1-2z} \geq N^{1-3z}$ for $z > 0$, the optimal exponent will be one giving the same convergence rate to the first and third term. In other words, we want $z = 1-3z$ and hence $z = 1/4$. Thus, taking $N \in \mathbb{N}$ large enough that all of our restrictions on $\epsilon = N^{-1/4}$ hold, we obtain

$$\sup_{\substack{\theta\in[a',b']\\ [A,B]\subset\mathbb{R}}} \left| P_{\theta N}\left\{ \omega\in\Omega^N : \sqrt{N\mathcal{I}(\theta)}(\hat{\theta}_N(\omega)-\theta)\in[A,B] \right\} - \frac{1}{\sqrt{2\pi}} \int_A^B \mathrm{e}^{-x^2/2}\,\mathrm{d}x \right| \leq \frac{C}{N^{1/4}}.$$

for $C = \sqrt{\frac{2}{e\pi}}\frac{K+1}{\gamma} + K' + \tilde{K}_1 + \tilde{K}_2$. □

Remark 3.2.3. The last convergence estimate is worse than what we offered in the setting of Theorem 2.2.2 on *iid* random variables, since the proof of the law of large numbers in Proposition 2.1.6 made use of the third moment. In contrast, the variance was used in Proposition 3.1.6. However, we see no reason to believe that utilizing the third moment there is impossible. Nonetheless, in that case, one would only be able to obtain the same convergence rate provided in Section 2.2, *i.e.*, $N^{-2/5}$, still short of the sharp decay estimate of $N^{-1/2}$ given in Theorem 3.1.11. Again, this is due to the additional ϵ factor introduced by the uniformity condition on the random variable's parameter in Proposition 3.1.7.

3.3 Large Deviations

Although our analysis so far has been following closely that of Chapter 2, showing the LDP for the Markovian MLE requires establishing results that held trivially for *iid* random variables. This primarily stems from the loss of independence of the entropy functions

$$S_{\lambda N}(\omega) = -\sum_{k=1}^{N-1} \log p_{\omega_k,\omega_{k+1}}(\lambda), \quad \text{where } \omega \in \Omega^N, \lambda \in [a,b].$$

For $\theta, \lambda \in [a,b]$ and $\alpha \in \mathbb{R}$, we now have to consider the sequence $(C_{\theta,\lambda,N}(\alpha))_{N\geq 2}$ of cumulant-generating functions given by

$$C_{\theta,\lambda,N}(\alpha) \coloneqq \log \mathbb{E}_{\theta N}\left(\mathrm{e}^{\alpha \dot{S}_{\lambda N}}\right)$$

and worry about the limit of $C_{\theta,\lambda,N}(\alpha)/N$ as $N \to \infty$. This limit can be studied through the spectral radius of the tilted matrix $T_{\theta,\lambda}(\alpha) = [\mathrm{e}^{\alpha \dot{S}_\lambda(\omega_1,\omega_2)} p_{\omega_1,\omega_2}(\theta)]_{\omega_1,\omega_2}$. Indeed, we have

$$\lim_{N\to\infty} \frac{1}{N} C_{\theta,\lambda,N}(\alpha) = \lim_{N\to\infty} \log \left\| T_{\theta,\lambda}(\alpha)^N \right\|^{1/N}, \tag{3.17}$$

where the norm is unspecified as $|\Omega| < \infty$ and all norms are equivalent on finite spaces. By Gelfand's formula, we also have that the spectral radius $e_{\theta,\lambda}(\alpha)$ of $T_{\theta,\lambda}(\alpha)$ has

$$e_{\theta,\lambda}(\alpha) = \lim_{N\to\infty} \left\| T_{\theta,\lambda}(\alpha)^N \right\|^{1/N}.$$

Henceforth, we denote $\mathfrak{E}_{\theta,\lambda}(\alpha) \coloneqq \log e_{\theta,\lambda}(\alpha)$ and refer to it as the limiting cumulant-generating function. We now state a few results concerning the limit of (3.17). The next proposition follows from [Mat23, Proposition 3.1.2] and an adaption of the proof of [Mat23, Lemma 3.1.7] to the case where $[a,b] \ni \theta \mapsto p_{\omega_1,\omega_2}(\theta)$ are C^3 for all $\omega_1, \omega_2 \in \Omega$.

Proposition 3.3.1. *For any $\theta, \lambda \in [a,b]$, the function $\alpha \mapsto \mathfrak{E}_{\theta,\lambda}(\alpha)$ is real-analytic. Moreover, the function $\partial_\alpha \mathfrak{E}_{\theta,\lambda}(\alpha)$ is C^2 in $\lambda \in [a,b]$ and C^3 in $\theta \in [a,b]$.*

The sequence $(C_{\theta,\lambda,N}(\alpha)/N)_{N\geq 2}$ is readily seen to be real-analytic, giving us tremendous knowledge on the derivatives of the limiting CGF, as implied by the following result.

Theorem 3.3.2 ([Mat23, Lemma 3.1.6]). *For any $\alpha \in \mathbb{R}$, there exists an open set $U_\alpha \subset \mathbb{C}$ with $\alpha \in U_\alpha$ on which $C_{\theta,\lambda,N}(\alpha)$ is analytic for all $N \geq 2$ and the following limit is uniform for $\theta, \lambda \in [a,b]$ and α in compact subsets $K \subset U_\alpha$:*

$$\lim_{N\to\infty} \frac{1}{N} C_{\theta,\lambda,N}(\alpha) = \mathfrak{E}_{\theta,\lambda}(\alpha).$$

Corollary 3.3.3. *For any $k \in \mathbb{N}$ and $\alpha \in \mathbb{R}$, there is $U_\alpha \subset \mathbb{C}$ open with $U_\alpha \ni \alpha$ such that*

$$\lim_{N\to\infty} \frac{1}{N} C^{(k)}_{\theta,\lambda,N}(\alpha) = \mathfrak{E}^{(k)}_{\theta,\lambda}(\alpha)$$

is uniform for $\theta, \lambda \in [a,b]$ and α in compact subsets $K \subset U_\alpha$.

In particular, the first and second derivatives of the limiting CGF read

$$\begin{aligned}
\mathfrak{E}'_{\theta,\lambda}(\alpha) &= \lim_{N\to\infty} \frac{1}{N} \frac{\mathbb{E}_{\theta N}\left(\dot{S}_{\lambda N} \mathrm{e}^{\alpha \dot{S}_{\lambda N}}\right)}{\mathbb{E}_{\theta N}\left(\mathrm{e}^{\alpha \dot{S}_{\lambda N}}\right)}, \\
\mathfrak{E}''_{\theta,\lambda}(\alpha) &= \lim_{N\to\infty} \frac{1}{N} \left(\frac{\mathbb{E}_{\theta N}\left(\dot{S}^2_{\lambda N} \mathrm{e}^{\alpha \dot{S}_{\lambda N}}\right)}{\mathbb{E}_{\theta N}\left(\mathrm{e}^{\alpha \dot{S}_{\lambda N}}\right)} - \left[\frac{\mathbb{E}_{\theta N}\left(\dot{S}_{\lambda N} \mathrm{e}^{\alpha \dot{S}_{\lambda N}}\right)}{\mathbb{E}_{\theta N}\left(\mathrm{e}^{\alpha \dot{S}_{\lambda N}}\right)} \right]^2 \right)
\end{aligned}$$

We infer that, as the limit of strictly convex functions, $\mathfrak{E}_{\theta,\lambda}$ is convex on $\mathbb{R}$ for all $\theta, \lambda \in [a,b]$. As in Chapter 2, we make the assumption that for all $\theta \in [a,b]$, the entropy functions

$$[a,b] \ni \theta \mapsto S_\theta(\omega_1, \omega_2) = -\log p_{\omega_1,\omega_2}(\theta)$$

have $\ddot{S}_\theta(\omega_1, \omega_2) > 0$, and we refer to it as the strict convexity assumption of the entropy functions. This has many consequences that will allow us to drive our analysis further, the first of which is that $[a,b] \ni \theta \mapsto S\big(p_{\omega,\bullet}(\lambda) | p_{\omega,\bullet}(\theta)\big)$ is strictly convex for each $\omega \in \Omega$.

Lemma 3.3.4. *Let $\theta \in [a,b]$. There exists $\omega, \omega' \in \Omega$ such that $\dot{p}_{\omega,\omega'}(\theta) \neq 0$. In particular, for any $\omega \in \Omega$, there are $\omega_\pm \in \Omega$ such that*

$$\dot{p}_{\omega,\omega_+}(\theta) > 0 \quad \textit{and} \quad \dot{p}_{\omega,\omega_-}(\theta) < 0.$$

Proof. We start with the first assertion. Suppose by contradiction that for some $\theta_0 \in [a,b]$, we have $\dot{p}_{\omega,\omega'}(\theta) = 0$ for all $\omega, \omega' \in \Omega$. Then for any $\lambda \in [a,b]$ and $\omega \in \Omega$, we have

$$\partial_\theta S\big(p_{\omega,\bullet}(\lambda)|p_{\omega,\bullet}(\theta)\big)\big|_{\theta=\theta_0} = -\sum_{\omega' \in \Omega} p_{\omega,\omega'}(\lambda)\frac{\dot{p}_{\omega,\omega'}(\theta_0)}{p_{\omega,\omega'}(\theta_0)} = 0.$$

By strict convexity, both $\theta = \lambda$ and $\theta = \theta_0$ are global minima by Lemma 3.1.5, which is impossible. The second claim follows from the fact that for any $\omega \in \Omega$, we have $\sum_{\omega' \in \Omega} \dot{p}_{\omega,\omega'}(\theta) = 0$, by stochasticity. □

Before continuing with the strict convexity consequences, let us make a brief digression. Using Lemma 3.3.4 and the the following result on analytic functions, we will show that the limiting cumulant-generating possesses strict convexity.

Lemma 3.3.5 ([KP02, Corrolary 1.2.6])**.** *Let $U \subset \mathbb{R}$ be an open interval and f and g real-analytic functions on U. If $\{z \in U : f(z) = g(z)\}$ has an accumulation point, then $f \equiv g$ on U.*

Proposition 3.3.6. *For any $\theta, \lambda \in [a,b]$, the map $\mathbb{R} \ni \alpha \mapsto \mathfrak{E}_{\theta,\lambda}(\alpha)$ is strictly convex.*

Proof. As the pointwise limit of a sequence of strictly convex functions, $\mathfrak{E}_{\theta,\lambda}$ must also be convex. Now fix $\theta, \lambda \in [a,b]$ and recall that $\dot{S}_\theta(\omega_1,\omega_2) = -\dot{p}_{\omega_1,\omega_2}(\theta)/p_{\omega_1,\omega_2}(\theta)$. Set

$$Q_+(\theta,\lambda,\alpha) := \left\{ \mathrm{e}^{\alpha \dot{S}_\lambda(\omega_1,\omega_2)} p_{\omega_1,\omega_2}(\theta) : \omega_1,\omega_2 \in \Omega \text{ and } \dot{S}_\lambda(\omega_1,\omega_2) > 0 \right\},$$
$$Q_-(\theta,\lambda,\alpha) := \left\{ \mathrm{e}^{\alpha \dot{S}_\lambda(\omega_1,\omega_2)} p_{\omega_1,\omega_2}(\theta) : \omega_1,\omega_2 \in \Omega \text{ and } \dot{S}_\lambda(\omega_1,\omega_2) < 0 \right\},$$

and denote $f_\pm(\theta,\lambda,\alpha) \coloneqq \min Q_\pm(\theta,\lambda,\alpha) > 0$. It is clear that $\lim_{\alpha\to\pm\infty} f_\pm(\theta,\lambda,\alpha) = \infty$. Let $\alpha > 0$. Since $T_{\theta,\lambda}(\alpha)$ has nonnegative entries, taking the sup norm yields

$$\left\|T(\theta,\lambda,\alpha)^N\right\| = \max_{i\in\Omega} \sum_{\omega_1,\dots,\omega_{N-1},j\in\Omega} \mathrm{e}^{\alpha\dot{S}_\lambda(i,\omega_1)} p_{i,\omega_1}(\theta)\dots\mathrm{e}^{\alpha\dot{S}_\lambda(\omega_{N-1},j)} p_{\omega_{N-1},j}(\theta) \geq f_+(\theta,\lambda,\alpha)^N,$$

by Lemma 3.3.4. Therefore, by (3.17),

$$\mathfrak{E}_{\theta,\lambda}(\alpha) = \lim_{N\to\infty} \log \left\|T(\theta,\lambda,\alpha)^N\right\|^{1/N} \geq \log f_+(\theta,\lambda,\alpha).$$

Similarly, for any $\alpha < 0$ we have $\mathfrak{E}_{\theta,\lambda}(\alpha) \geq \log f_-(\theta,\lambda,\alpha)$ and thus

$$\lim_{\alpha\to\pm\infty} \mathfrak{E}_{\theta,\lambda}(\alpha) = \infty. \tag{3.18}$$

Now suppose that $\mathbb{R} \ni \alpha \mapsto \mathfrak{E}_{\theta,\lambda}(\alpha)$ is not strictly convex, in that there exists an interval $I \subset \mathbb{R}$ over which $\mathfrak{E}''_{\theta,\lambda}(\alpha) = 0$. Since $\mathbb{R} \ni \alpha \mapsto \mathfrak{E}_{\theta,\lambda}(\alpha)$ is real-analytic, it must be linear on all of $\mathbb{R}$ by Lemma 3.3.5, which is impossible by (3.18) and the fact that $\mathfrak{E}_{\theta,\lambda}(0) = 0$. □

From the proof, we observe that for each $\theta,\lambda \in [a,b]$, there exists a unique $\alpha_\lambda \in \mathbb{R}$ such that $\mathfrak{E}'_{\theta,\lambda}(\alpha_0) = 0$. Additionally, α_λ is a global minimum of the map $\alpha \mapsto \mathfrak{E}_{\theta,\lambda}(\alpha)$.

The next consequence of our convexity assumption is that for any $\lambda \in [a,b]$ and all N,

$$\left\{\omega \in \Omega^N \colon \hat{\theta}_N(\omega) \geq \lambda\right\} = \left\{\omega \in \Omega^N \colon \dot{S}_{\lambda N}(\omega) \leq 0\right\}, \tag{3.19}$$

$$\left\{\omega \in \Omega^N \colon \hat{\theta}_N(\omega) \leq \lambda\right\} = \left\{\omega \in \Omega^N \colon \dot{S}_{\lambda N}(\omega) \geq 0\right\}. \tag{3.20}$$

Having laid the groundwork, we turn to the LDP of the MLE and the study of its rate function. It is well-known that the LDP holds for sequences of finite state Markov chains with irreducible stochastic matrices. Making use of Remark 3.1.8, we state a version of [DZ10, Theorem 3.1.2] obtained via an application of the Gärtner-Ellis theorem—a generalization of Cramér's theorem to random variables that are not necessarily *iid*.

Theorem 3.3.7. *Let $X\colon \Omega^2 \to \mathbb{R}$ be a random variable defining a finite state Markov chain possessing an irreducible stochastic matrix $T = [p_{ij}]_{i,j\in\Omega}$. Consider the sequence*

$$\frac{\mathcal{S}_N(\omega)}{N} = \frac{1}{N}\sum_{k=1}^{N-1} X(\omega_k, \omega_{k+1}), \quad \textit{where } \omega \in \Omega^N.$$

Moreover, let $e(\alpha)$ be the spectral radius of tilted the matrix $T(\alpha) = [\mathrm{e}^{\alpha X(i,j)} p_{ij}]_{i,j\in\Omega}$ and set

$$I(s) = \sup_{\alpha\in\mathbb{R}}(\alpha s - \log e(\alpha)), \quad \textit{where } s \in \mathbb{R}.$$

Then the sequence $(\mathcal{S}_N/N)_{N\in\mathbb{N}}$ satisfies the LDP with convex, good rate function I.

We recall that a rate function is called good if its sublevel sets are compact, a fact we shall not use. Now, let $J_\lambda^{(\theta)}(s) \coloneqq \sup_{\alpha\in\mathbb{R}}(\alpha s - \mathfrak{E}_{\theta,\lambda}(\alpha))$, the Fenchel-Legendre transform of $\mathfrak{E}_{\theta,\lambda}$. Applying Theorem 3.3.7 to the sequence $(\dot{S}_{\lambda N}/N)_{N\geq 2}$, we obtain that for any set $E \subset \mathbb{R}$,

$$\begin{aligned} -\inf_{s\in\mathrm{int}(E)} J_\lambda^{(\theta)}(s) &\leq \liminf_{N\to\infty} \frac{1}{N}\log P_{\theta N}\left\{\omega\in\Omega^N\colon \frac{\dot{S}_{\lambda N}(\omega)}{N} \in \mathrm{int}(E)\right\} \\ &\leq \limsup_{N\to\infty} \frac{1}{N}\log P_{\theta N}\left\{\omega\in\Omega^N\colon \frac{\dot{S}_{\lambda N}(\omega)}{N} \in \mathrm{cl}(E)\right\} \leq -\inf_{s\in\mathrm{cl}(E)} J_\lambda^{(\theta)}(s). \end{aligned}$$

Since $J_\lambda^{(\theta)}(s) = 0 \iff s = \mathfrak{E}'_{\theta,\lambda}(0) = \mathbb{E}_{\theta 2}(\dot{S}_\lambda)$, then for any interval $[A, B] \subset \mathbb{R}$,

$$\inf_{s\in[A,B]} J_\lambda^{(\theta)}(s) = \begin{cases} 0 & \text{if } \mathbb{E}_{\theta 2}(\dot{S}_\lambda) \in [A, B], \\ J_\lambda^{(\theta)}(A) & \text{if } A > \mathbb{E}_{\theta 2}(\dot{S}_\lambda), \\ J_\lambda^{(\theta)}(B) & \text{if } B < \mathbb{E}_{\theta 2}(\dot{S}_\lambda). \end{cases}$$

Therefore, whenever $x \geq \mathbb{E}_{\theta 2}(\dot{S}_\lambda)$ we have

$$\lim_{N\to\infty} \frac{1}{N}\log P_{\theta N}\left\{\omega\in\Omega^N\colon \frac{\dot{S}_{\lambda N}(\omega)}{N} \geq x\right\} = -J_\lambda^{(\theta)}(x).$$

and whenever $x \leq \mathbb{E}_{\theta 2}(\dot{S}_\lambda)$,

$$\lim_{N\to\infty} \frac{1}{N} \log P_{\theta N} \left\{ \omega \in \Omega^N \colon \frac{\dot{S}_{\lambda N}(\omega)}{N} \leq x \right\} = -J_\lambda^{(\theta)}(x).$$

Extending $\lambda \mapsto J_\lambda^{(\theta)}(0)$ to $J_\lambda^{(\theta)}(0) = \infty$ when $\lambda \notin [a, b]$, we derive an LDP for the MLE.

Proposition 3.3.8. *For $\lambda \geq \theta$,*

$$\lim_{N\to\infty} \frac{1}{N} \log P_{\theta N} \left\{ \omega \in \Omega^N \colon \hat{\theta}_N(\omega) \geq \lambda \right\} = -J_\lambda^{(\theta)}(0). \tag{3.21}$$

For $\lambda \leq \theta$,

$$\lim_{N\to\infty} \frac{1}{N} \log P_{\theta N} \left\{ \omega \in \Omega^N \colon \hat{\theta}_N(\omega) \leq \lambda \right\} = -J_\lambda^{(\theta)}(0). \tag{3.22}$$

The following results establishes a few properties of the function $\lambda \mapsto J_\lambda^{(\theta)}(0)$.

Corollary 3.3.9. *The function $[a, b] \ni \lambda \mapsto J_\lambda^{(\theta)}(0)$ is finite, nonnegative, nonincreasing on $[a, \theta]$, nondecreasing on $[\theta, b]$ and vanishing at $\lambda = \theta$.*

Proof. Nonnegativity is seen directly from

$$J_\lambda^{(\theta)}(0) = \sup_{\alpha\in\mathbb{R}}(-\mathfrak{E}_{\theta,\lambda}(\alpha)) \geq -\mathfrak{E}_{\theta,\lambda}(0) = 0.$$

To prove that $J_\lambda^{(\theta)}(0)$ is finite, recall (3.18) and that $\alpha \mapsto \mathfrak{E}_{\theta,\lambda}(\alpha)$ is real-analytic (in particular continuous) so $\mathfrak{E}_{\theta,\lambda}(\alpha) = -\infty$ is impossible. The fact that $\lambda \mapsto J_\lambda^{(\theta)}(0)$ is nondecreasing on $[\theta, b]$ and nonincreasing on $[a, \theta]$ follows from (3.21) and (3.22) respectively. Lastly, $J_\theta^{(\theta)}(0) = 0$ follows from (3.21) and the uniform consistency of the maximum likelihood estimator given in Theorem 3.1.10. □

The function $[a,b] \ni \lambda \mapsto J_\lambda^{(\theta)}(0)$ actually enjoys stronger properties, and we shall now explore them. Fix $\theta \in (a,b)$ and given $\lambda \in [a,b]$, let $\alpha_\lambda \in \mathbb{R}$ be the global minimum of

$$\mathbb{R} \ni \alpha \mapsto \mathfrak{E}_{\theta,\lambda}(\alpha) = \lim_{N\to\infty} \frac{1}{N} \log \sum_{\omega\in\Omega^N} \mathrm{e}^{\alpha \dot{S}_{\lambda N}(\omega)} P_{\theta N}(\omega),$$

hence $J_\lambda^{(\theta)}(0) = -\mathfrak{E}_{\theta,\lambda}(\alpha_\lambda)$. Note that $0 = \mathfrak{E}'_{\theta,\lambda}(\alpha_\lambda)$ and recall that $\boldsymbol{p}(\theta) > 0$. Since

$$\mathfrak{E}'_{\theta,\lambda}(0) = \lim_{N\to\infty} \frac{\mathbb{E}_{\theta N}(\dot{S}_{\lambda N})}{N} = \mathbb{E}_{\theta 2}(\dot{S}_\lambda) = \sum_{\omega\in\Omega} p_\omega(\theta) \partial_\lambda S\big(p_{\omega,\bullet}(\theta) | p_{\omega,\bullet}(\lambda)\big),$$

then $J_\lambda^{(\theta)}(0) = 0 \iff \alpha_\lambda = 0 \iff \partial_\lambda S\big(p_{\omega,\bullet}(\lambda)|p_{\omega,\bullet}(\theta)\big) = 0$ for all $\omega \in \Omega \iff \lambda = \theta$.

We now consider the function $F\colon [a,b] \times \mathbb{R} \to \mathbb{R}$ given by

$$F(\lambda,\alpha) := \partial_\alpha \mathfrak{E}_{\theta,\lambda}(\alpha) = \lim_{N\to\infty} \frac{1}{N} \frac{\mathbb{E}_{\theta N}(\dot{S}_{\lambda N} \mathrm{e}^{\alpha \dot{S}_{\lambda N}})}{\mathbb{E}_{\theta N}(\mathrm{e}^{\alpha \dot{S}_{\lambda N}})}.$$

Note that the above map is real-analytic in α and C^2 in $\lambda \in [a,b]$ by Proposition 3.3.1, and observe that $F(\theta,\alpha_\theta) = 0$ by definition of α_θ. Furthermore, at $(\lambda,\alpha) = (\theta,\alpha_\theta)$, we have

$$\begin{aligned} \partial_\alpha F(\theta,\alpha)\big|_{\alpha=\alpha_\theta} &= \lim_{N\to\infty} \frac{1}{N} \left(\frac{\mathbb{E}_{\theta N}(\dot{S}_{\theta N}^2 \mathrm{e}^{\alpha_\theta \dot{S}_{\theta N}})}{\mathbb{E}_{\theta N}(\mathrm{e}^{\alpha_\theta \dot{S}_{\theta N}})} - \left[\frac{\mathbb{E}_{\theta N}(\dot{S}_{\theta N} \mathrm{e}^{\alpha_\theta \dot{S}_{\theta N}})}{\mathbb{E}_{\theta N}(\mathrm{e}^{\alpha_\theta \dot{S}_{\theta N}})} \right]^2 \right) \\ &= \lim_{N\to\infty} \frac{\mathbb{E}_{\theta N}(\dot{S}_{\theta N}^2)}{N} = \mathcal{I}(\theta). \end{aligned} \tag{3.23}$$

where the second equality follows from $\alpha_\theta = 0$ and $\mathbb{E}_{\theta N}(\dot{S}_{\theta N}) = 0$ for all $N \geq 2$, and the third equality follows from $\mathbb{E}_{\theta N}(\dot{S}_{\theta N}^2) = (N-1)\mathcal{I}(\theta)$, as derived in Appendix B. Since the Fisher entropy $\mathcal{I}$ is nonvanishing on $[a,b]$ by assumption, then $\partial_\alpha F(\theta,\alpha)\big|_{\alpha=\alpha_0} > 0$. By the implicit function theorem, there exists an open set $U_\theta \ni \theta$ such that the map $U_\theta \ni \lambda \mapsto \alpha_\lambda$ is twice-continuously differentiable and with derivative given by

$$\dot{\alpha}_\lambda = -\left(\frac{\partial F(\lambda,\alpha)}{\partial \alpha} \right)^{-1} \left(\frac{\partial F(\lambda,\alpha)}{\partial \lambda} \right) \Bigg|_{\alpha=\alpha_\lambda}.$$

Further, we infer that $U_\theta \ni \lambda \mapsto J_\lambda^{(\theta)}(0) = -\mathfrak{E}_{\theta,\lambda}(\alpha_\lambda)$ is C^2. Note that, at $\alpha = \alpha_\theta$, we have

$$F(\lambda, \alpha_\theta) = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{\theta N}(\dot{S}_{\lambda N}) = \mathbb{E}_{\theta 2}(\dot{S}_\lambda).$$

Therefore, $\partial_\lambda F(\lambda, \alpha_\theta)\big|_{\lambda=\theta} = \mathbb{E}_{\theta 2}(\ddot{S}_\theta) = \mathcal{I}(\theta)$ and combining this with (3.23), we obtain $\dot{\alpha}_\theta = -1$. Now, let $N \in \mathbb{N}$ and let $C_{\theta N} \colon [a, b] \to \mathbb{R}$ be the function defined by

$$C_{\theta N}(\lambda) := \log \mathbb{E}_{\theta N}\big(\mathrm{e}^{\alpha_\lambda \dot{S}_{\lambda N}}\big), \quad \text{for } \lambda \in U_\theta.$$

Note that this function is C^2 and has

$$C'_{\theta N}(\lambda) = \frac{\mathbb{E}_{\theta N}\big((\alpha_\lambda \ddot{S}_{\lambda N} + \dot{\alpha}_\lambda \dot{S}_{\lambda N})\mathrm{e}^{\alpha_\lambda \dot{S}_{\lambda N}}\big)}{\mathbb{E}_{\theta N}(\mathrm{e}^{\alpha_\lambda \dot{S}_{\lambda N}})} = \alpha_\lambda \frac{\mathbb{E}_{\theta N}(\ddot{S}_{\lambda N}\mathrm{e}^{\alpha_\lambda \dot{S}_{\lambda N}})}{\mathbb{E}_{\theta N}(\mathrm{e}^{\alpha_\lambda \dot{S}_{\lambda N}})}$$

by definition of $\alpha_\lambda \in \mathbb{R}$. Its second derivative has

$$\begin{aligned} C''_{\theta N}(\lambda) = \dot{\alpha}_\lambda \frac{\mathbb{E}_{\theta N}(\ddot{S}_{\lambda N}\mathrm{e}^{\alpha_\lambda \dot{S}_{\lambda N}})}{\mathbb{E}_{\theta N}(\mathrm{e}^{\alpha_\lambda \dot{S}_{\lambda N}})} + \alpha_\lambda \frac{\mathbb{E}_{\theta N}\big((\dddot{S}_{\lambda N} + \dot{\alpha}_\lambda \dot{S}_{\lambda N}\ddot{S}_{\lambda_N} + \alpha_\lambda \ddot{S}^2_{\lambda N})\mathrm{e}^{\alpha_\lambda \dot{S}_{\lambda N}}\big)}{\mathbb{E}_{\theta N}(\mathrm{e}^{\alpha_\lambda \dot{S}_{\lambda N}})} \\ - \alpha_\lambda^2 \left(\frac{\mathbb{E}_{\theta N}(\ddot{S}_{\lambda N}\mathrm{e}^{\alpha_\lambda \dot{S}_{\lambda N}})}{\mathbb{E}_{\theta N}(\mathrm{e}^{\alpha_\lambda \dot{S}_{\lambda N}})}\right)^2. \end{aligned}$$

Since $\dot{\alpha}_\theta = -1$ and $\alpha_\theta = 0$, then at $\lambda = \theta$ we have $-C''_{\theta N}(\theta) = (N-1)\mathcal{I}(\theta)$ for all $N \geq 2$. Thus, for each $N \geq 2$, there exists a neighborhood $\tilde{U}_{\theta N} \ni \theta$ on which $-C''_{\theta N}/N$ is convex. Proceeding, we make the assumption that there is a shared open set $\tilde{U}_\theta \ni \theta$ on which $-C_{\theta N}/N$ is strictly convex for any $N \geq 2$. We infer that $\tilde{U}_\theta \ni \lambda \mapsto J_\lambda^{(\theta)}(0)$ is convex and hence that the sequence of derivatives $(-C'_{\theta N}/N)_{N\geq 2}$ converges[2] to $\partial_\lambda J_\lambda^{(\theta)}(0)$:

$$\partial_\lambda J_\lambda^{(\theta)}(0) = -\alpha_\lambda \lim_{N\to\infty} \frac{1}{N} \frac{\mathbb{E}_{\theta N}(\ddot{S}_{\lambda N}\mathrm{e}^{\alpha_\lambda \dot{S}_{\lambda N}})}{\mathbb{E}_{\theta N}(\mathrm{e}^{\alpha_\lambda \dot{S}_{\lambda N}})}.$$

[2] A sequence of finite, convex and differentiable functions $(f_n)_n$ on an open convex set U converging pointwise to a finite, convex and differentiable function f on U has $f'_n \to f'$ uniformly on compact subsets of U. See [Roc70, Theorem 25.7]

Since $\ddot{S}_\lambda > 0$ on Ω^2 for all $\lambda \in [a,b]$, then for all $\lambda \in \tilde{U}_\theta$, $\partial_\lambda J_\lambda^{(\theta)}(0) = 0 \iff \alpha_\lambda = 0$ and thus $\lambda = \theta \iff \partial_\lambda J_\lambda^{(\theta)}(0)$. Moreover, we can compute the second derivative to obtain

$$\begin{aligned}\partial_\lambda^2 J_\lambda^{(\theta)}(0)\big|_{\lambda=\theta} &= -\dot{\alpha}_\theta \lim_{N\to\infty} \frac{1}{N} \frac{\mathbb{E}_{\theta N}(\ddot{S}_{\theta N} \mathrm{e}^{\alpha_\theta \dot{S}_{\theta N}})}{\mathbb{E}_{\theta N}(\mathrm{e}^{\alpha_\theta \dot{S}_{\theta N}})} - \alpha_\theta \left(\frac{\partial}{\partial_\lambda} \lim_{N\to\infty} \frac{1}{N} \frac{\mathbb{E}_{\theta N}(\ddot{S}_{\lambda N} \mathrm{e}^{\alpha_\lambda \dot{S}_{\lambda N}})}{\mathbb{E}_{\theta N}(\mathrm{e}^{\alpha_\lambda \dot{S}_{\lambda N}})} \right) \Bigg|_{\lambda=\theta} \\ &= \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{\theta N}(\ddot{S}_{\theta N}) = \mathcal{I}(\theta) > 0.\end{aligned}$$

As in the case of *iid* random variables, the Fisher entropy appears as the second derivative of the MLE's rate function at its minimum value.

Theorem 3.3.10. *For any fixed $\theta \in (a,b)$, there exists an open set $\tilde{U}_\theta \ni \theta$ with $\tilde{U}_\theta \subset [a,b]$ such that $\tilde{U}_\theta \ni \lambda \mapsto J_\lambda^{(\theta)}(0)$ is strictly convex around its minimum point $\lambda = \theta$.*

We now illustrate that, although the rate function can be strictly convex about its minimum point, it can still be bounded on $\mathbb{R}$.

Example 3.3.11. Consider the exponential families

$$p_{\omega_1,\omega_2}(\theta) = \mathrm{e}^{\theta H(\omega_1,\omega_2)}/Z(\theta), \qquad Z(\theta) = \frac{1}{|\Omega|} \sum_{\omega_1,\omega_2 \in \Omega} \mathrm{e}^{\theta H(\omega_1,\omega_2)}$$

where $H\colon \Omega \times \Omega \to \mathbb{R}$ is a non-constant function such that for all $\theta \in [a,b]$,

$$\sum_{\omega_2 \in \Omega} \mathrm{e}^{\theta H(\omega_1,\omega_2)} = \frac{1}{|\Omega|} \sum_{\omega_1,\omega_2} \mathrm{e}^{\theta H(\omega_1,\omega_2)} \quad \text{for any } \omega_1 \in \Omega.$$

In this case, the entropy functions $S_\theta(\omega_1,\omega_2) = -\log p_{\omega_1,\omega_2}(\theta) = -\theta H(\omega_1,\omega_2) + \log Z(\theta)$ are strictly convex since

$$\begin{aligned}\partial_\theta^2(-\log p_{\omega_1,\omega_2}(\theta)) &= \frac{\ddot{Z}(\theta)Z(\theta) - \dot{Z}(\theta)^2}{Z(\theta)^2} \\ &= \sum_{\omega_1,\omega_2 \in \Omega} \frac{H(\omega_1,\omega_2)^2 \mathrm{e}^{\theta H(\omega_1,\omega_2)}}{|\Omega| Z(\theta)} - \left(\sum_{\omega_1,\omega_2 \in \Omega} \frac{H(\omega_1,\omega_2) \mathrm{e}^{\theta H(\omega_1,\omega_2)}}{|\Omega| Z(\theta)} \right)^2 > 0,\end{aligned}$$

where the last inequality follows by Jensen's and our assumption on H. We have

$$\dot{S}_\lambda(\omega_1,\omega_2) = -\frac{\dot{p}_{\omega_1,\omega_2}(\lambda)}{p_{\omega_1,\omega_2}(\lambda)} = \frac{-Z(\lambda)}{\mathrm{e}^{\lambda H(\omega_1,\omega_2)}}\left(\frac{H(\omega_1,\omega_2)Z(\lambda)\mathrm{e}^{\lambda H(\omega_1,\omega_2)} + \dot{Z}(\lambda)\mathrm{e}^{\lambda H(\omega_1,\omega_2)}}{Z(\lambda)^2}\right)$$
$$= \frac{-H(\omega_1,\omega_2)Z(\lambda) + \dot{Z}(\lambda)}{Z(\lambda)}.$$

Denote by $\mathcal{J}^{(\theta)}$ the rate function associated to the sequence $H_N(\omega) = \sum_{k=1}^{N-1} H(\omega_k,\omega_{k+1})$ for $\omega \in \Omega^N$, and observe that

$$\begin{aligned}
J_\lambda^{(\theta)}(0) &= \sup_{\alpha\in\mathbb{R}}\left(-\lim_{N\to\infty}\frac{1}{N}\log\mathbb{E}_{\theta N}(\mathrm{e}^{\alpha\dot{S}_{\lambda N}})\right)\\
&= \sup_{\alpha\in\mathbb{R}}\left(-\lim_{N\to\infty}\frac{1}{N}\log\mathbb{E}_{\theta N}(\mathrm{e}^{N\alpha\dot{Z}(\lambda)/Z(\lambda)}\mathrm{e}^{-\alpha H_N})\right)\\
&= \sup_{\alpha\in\mathbb{R}}\left(-\alpha\frac{\dot{Z}(\lambda)}{Z(\lambda)} - \lim_{N\to\infty}\frac{1}{N}\log\mathbb{E}_{\theta N}(\mathrm{e}^{-\alpha H_N})\right)\\
&= \mathcal{J}^{(\theta)}(\dot{Z}(\lambda)/Z(\lambda)).
\end{aligned}$$

Now let

$$m = \min_{\omega_1,\omega_2\in\Omega} H(\omega_1,\omega_2) \quad \text{and} \quad M = \max_{\omega_1,\omega_2\in\Omega} H(\omega_1,\omega_2),$$

and note that

$$\frac{\dot{Z}(\lambda)}{Z(\lambda)} = \frac{1}{|\Omega|}\sum_{\omega_1,\omega_2\in\Omega}\frac{H(\omega_1,\omega_2)\mathrm{e}^{\lambda H(\omega_1,\omega_2)}}{Z(\lambda)} \in (m,M)$$

In particular, the bounds are saturated asymptotically:

$$\begin{aligned}
\frac{\dot{Z}(\lambda)}{Z(\lambda)} &= \frac{\sum_{\omega_1,\omega_2\in\Omega} H(\omega_1,\omega_2)\mathrm{e}^{\lambda H(\omega_1,\omega_2)}}{\sum_{\omega_1',\omega_2'\in\Omega}\mathrm{e}^{\lambda H(\omega_1',\omega_2')}}\\
&= \sum_{H(\omega_1,\omega_2)=M}\frac{M}{\sum_{\omega_1',\omega_2'\in\Omega}\mathrm{e}^{\lambda(H(\omega_1',\omega_2')-M)}} + \sum_{H(\omega_1,\omega_2)<M}\frac{H(\omega_1,\omega_2)}{\sum_{\omega_1',\omega_2'\in\Omega}\mathrm{e}^{\lambda(H(\omega_1',\omega_2')-H(\omega_1,\omega_2))}}\\
&\xrightarrow{\lambda\to\infty} M. \qquad (3.24)
\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{\dot{Z}(\lambda)}{Z(\lambda)} &= \frac{\sum_{\omega_1,\omega_2\in\Omega} H(\omega_1,\omega_2)\mathrm{e}^{\lambda H(\omega_1,\omega_2)}}{\sum_{\omega_1',\omega_2'\in\Omega}\mathrm{e}^{\lambda H(\omega_1',\omega_2')}} \\ &= \sum_{H(\omega_1,\omega_2)=m} \frac{m}{\sum_{\omega_1',\omega_2'\in\Omega}\mathrm{e}^{\lambda(H(\omega_1',\omega_2')-m)}} + \sum_{H(\omega_1,\omega_2)>m} \frac{H(\omega_1,\omega_2)}{\sum_{\omega_1',\omega_2'\in\Omega}\mathrm{e}^{\lambda(H(\omega_1',\omega_2')-H(\omega_1,\omega_2))}} \\ &\xrightarrow{\lambda\to-\infty} m. \end{aligned} \tag{3.25}$$

Denote by Λ_θ the limiting CGF of the sequence H_N, that is

$$\Lambda_\theta(\alpha) \coloneqq \lim_{N\to\infty}\frac{1}{N}\log\mathbb{E}_{\theta N}\left(\mathrm{e}^{\alpha H_N}\right).$$

Its first derivative is given by

$$\Lambda_\theta'(\alpha) = \lim_{N\to\infty}\frac{\mathbb{E}_{\theta N}\left(H_N\mathrm{e}^{\alpha H_N}\right)}{N\mathbb{E}_{\theta N}\left(\mathrm{e}^{\alpha H_N}\right)} \in (m,M).$$

By the intermediate value theorem, for any $s\in(m,M)$ there exists $\alpha_s\in\mathbb{R}$ such that $s=\Lambda_\theta'(\alpha_s)$ and $\alpha_s=(\Lambda_\theta')^{-1}(s)$ satisfies $\mathcal{J}^{(\theta)}(s)=\alpha_s\Lambda_\theta'(\alpha_s)-\Lambda_\theta(\alpha_s)$ with

$$\lim_{s\downarrow m}\alpha_s=-\infty \quad\text{and}\quad \lim_{s\uparrow M}\alpha_s=\infty.$$

The analysis becomes different than the one given in Example 2.3.4, since

$$P_{\theta N}\left\{H_N(\omega)=M(N-1)\right\} = (N-1)P_{\theta 2}\{H(\omega_1,\omega_2)=M\} > 0$$

for any $N\geq 2$, and thus,

$$\lim_{N\to\infty}\frac{1}{N}\log P_{\theta N}\left\{H_N(\omega)=M(N-1)\right\} = \lim_{N\to\infty}\frac{1}{N}\log(N-1)=0.$$

Evaluating the limits

$$\lim_{\lambda\to\infty} J_\lambda^{(\theta)}(0) = \lim_{\lambda\to\infty} \mathcal{J}^{(\theta)}(\dot{Z}(\lambda)/Z(\lambda)) = \lim_{s\uparrow M} \mathcal{J}^{(\theta)}(s) = \lim_{s\uparrow M} \big(\alpha_s\Lambda'_\theta(\alpha_s) - \Lambda_\theta(\alpha_s)\big)$$
$$\lim_{\lambda\to-\infty} J_\lambda^{(\theta)}(0) = \lim_{\lambda\to-\infty} \mathcal{J}^{(\theta)}(\dot{Z}(\lambda)/Z(\lambda)) = \lim_{s\downarrow m} \mathcal{J}^{(\theta)}(s) = \lim_{s\downarrow m} \big(\alpha_s\Lambda'_\theta(\alpha_s) - \Lambda_\theta(\alpha_s)\big)$$

is hence less clear. But considering a concrete example will support our analysis.

Consider the case $\Omega = \{1, -1\}$ with

$$H = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad Z(\theta) = \frac{1}{2}\sum_{\omega_1,\omega_2\in\Omega} \mathrm{e}^{\theta H(\omega_1,\omega_2)} = 1 + \mathrm{e}^{\theta}.$$

The associated stochastic matrix and its invariant probability vector are given by

$$[p_{\omega_1,\omega_2}(\theta)]_{\omega_1,\omega_2} = \frac{1}{1+\mathrm{e}^{\theta}} \begin{pmatrix} \mathrm{e}^{\theta} & 1 \\ 1 & \mathrm{e}^{\theta} \end{pmatrix}, \quad [p_{\omega_1}(\theta)]_{\omega_1} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

The first derivative of the entropy functions have

$$\left[\frac{-\dot{p}_{\omega_1,\omega_2}(\lambda)}{p_{\omega_1,\omega_2}(\lambda)}\right]_{\omega_1,\omega_2} = \frac{1}{1+\mathrm{e}^{\theta}} \begin{pmatrix} -1 & \mathrm{e}^{\theta} \\ \mathrm{e}^{\theta} & -1 \end{pmatrix}$$

and the tilted matrix yields

$$T(\theta,\lambda,\alpha) = [\mathrm{e}^{\alpha\dot{S}_\lambda(\omega_1,\omega_2)} p_{\omega_1,\omega_2}(\theta)]_{\omega_1,\omega_2} = \frac{1}{1+\mathrm{e}^{\theta}} \begin{pmatrix} \mathrm{e}^{-\alpha/(1+\mathrm{e}^{\lambda})+\theta} & \mathrm{e}^{\alpha\mathrm{e}^{\lambda}/(1+\mathrm{e}^{\lambda})} \\ \mathrm{e}^{\alpha\mathrm{e}^{\lambda}/(1+\mathrm{e}^{\lambda})} & \mathrm{e}^{-\alpha/(1+\mathrm{e}^{\lambda})+\theta} \end{pmatrix}.$$

The Perron root of this matrix is

$$e_{\theta,\lambda}(\alpha) = \mathrm{e}^{-\alpha/(1+\mathrm{e}^{\lambda})} \left(\frac{\mathrm{e}^{\alpha} + \mathrm{e}^{\theta}}{1+\mathrm{e}^{\theta}}\right)$$

and for $\theta = 0$, the function

$$\alpha \mapsto -\log e_{\theta,\lambda}(\alpha) = \frac{\alpha}{1+\mathrm{e}^{\lambda}} - \log\left(\frac{\mathrm{e}^{\alpha}+1}{2}\right)$$

is maximized at $\alpha = -\lambda$. Therefore, the rate function is given by

$$J_{\lambda}^{(0)}(0) = \sup_{\alpha\in\mathbb{R}}(-\log e_{0,\lambda}(\alpha)) = \frac{-\lambda}{1+\mathrm{e}^{\lambda}} - \log\left(\frac{\mathrm{e}^{-\lambda}+1}{2}\right).$$

Taking the limit in λ gives $\lim_{\lambda\to\pm\infty} J_{\lambda}^{(0)}(0) = \log 2$, and the second derivative yields

$$\partial_{\lambda}^{2} J_{\lambda}^{(\theta)}(0) = \frac{1-\lambda\tanh(\lambda/2)}{2(1+\cosh(\lambda))}.$$

The graph of the rate function is shown in fig. 3.1.

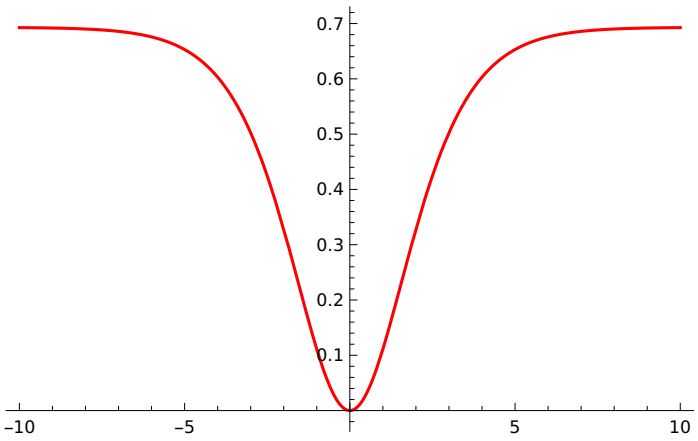

Figure 3.1: The rate function $\lambda \mapsto J_{\lambda}^{(0)}(0)$, which is not convex on $\mathbb{R}$ and has inflection points at $\lambda = \pm 1.5434$. The rate function possesses a horizontal asymptote at $\log 2$.

Chapter Four

Conclusion

Taking advantage of the elementary relationship that exists between the maximum likelihood estimator and the entropy function, the central limit theorems of the MLE is readily obtained by making use of a parameter-uniform law of large numbers. Further, we obtain explicit convergence estimates for both *iid* random variables and finite state Markov chains. Using our methods, we obtain an estimate of $N^{-2/5}$ in Chapter 2, short of the sharp bound of $N^{-1/2}$ that usually holds for Berry-Esseen-type results. The Markov estimate of $N^{-1/4}$ could likely be improved by utilizing the third moment instead of the second in proving the uniform LLN in Section 3.1.

The large deviations of the MLE are studied by exploiting a strict convexity assumption on the entropy functions. We observe that the rate function associated to the sequence of first derivatives of the entropy functions is tightly linked to an LDP statement for the MLE. In the *iid* setting, the MLE rate function is shown to be strictly convex on the entire parameter domain via the implicit function theorem. Additionally, it possesses one order of regularity less than that of the probability measures. In particular, in our setting, it is twice-continuously differentiable and—at its minimum—the second derivative matches the Fisher entropy. For Markov chains, the limiting cumulant-generating function makes the analysis more opaque, but its analyticity in neighborhoods of real values grants us

tremendous capabilities. Specifically, we compute its first and second derivatives and we show the limiting CGF admits strict convexity everywhere. We then present an LDP assertion for the Markovian MLE, and use the implicit function theorem to study the associated rate function about its minimum. Under a shared convexity assumption on the sequence of cumulant-generating functions, the MLE rate function is shown to be strictly convex around its minimum and, as in the *iid* setting, its second derivative matches the Fisher entropy at the minimum. We end Chapters 2 and 3 by providing an example which illustrates that the MLE rate function can admit inflection points despite the entropy functions' strict convexity.

Our study of the large deviations of the maximum likelihood estimator relies crucially on the strict convexity assumption imposed on the entropy functions. Naturally, one may wonder what can be said about the exponential tail convergence of the MLE when this assumption is dropped. Keeping an identifiability condition on the probability measures, we note that the entropy functions are forced to be monotone, by continuity. However, they are free to admit inflection points. A study of the LDP for the MLE in this case would thus require handling the convex and concave parts of the entropy functions simultaneously—a problem yet to be solved.

Appendix

A Cumulative Property of Cumulants

Let P be a probability measure on the finite set Ω and let $X, Y : \Omega \to \mathbb{R}$ be independent random variables. The cumulant-generating function of the sum $X + Y$ has

$$C_{X+Y}(\alpha) = \log \mathbb{E}(\mathrm{e}^{\alpha(X+Y)}) = \log \left[\mathbb{E}(\mathrm{e}^{\alpha X})\mathbb{E}(\mathrm{e}^{\alpha Y})\right] = C_X(\alpha) + C_Y(\alpha),$$

for $\alpha \in \mathbb{R}$. Taking the n^{th} derivative and evaluating at $\alpha = 0$, we obtain

$$C^{(n)}_{X+Y}(0) = C^{(n)}_X(0) + C^{(n)}_Y(0),$$

showing the cumulative property of cumulants for independent random variables.

Similarly, let $X : \Omega \to \mathbb{R}$ be a random variable. For $\omega = (\omega_1, \ldots, \omega_N) \in \Omega^N$, let $1 \leq j \leq N$ be an integer and define $X_j(\omega) \coloneqq X(\omega_j)$. The latter are readily seen to be an *iid* family of random variables. The cumulant-generating function of the sum $X_1 + \cdots + X_N$ thus has

$$C_{X_1+\cdots+X_N}(\alpha) = NC_X(\alpha) \implies C^{(n)}_{X_1+\cdots+X_N}(0) = NC^{(n)}_X(0).$$

B Expectations and Variance

We compute expectation values and the variance of derivatives of the entropy function in the Markov setting. Let $\theta \in [a, b]$, with $N \geq 2$ and $\omega \in \Omega^N$. Recall that

$$\begin{aligned}
S_{\theta N}(\omega) &= -\sum_{k=1}^{N-1} \log p_{\omega_k,\omega_{k+1}}(\theta), \\
\dot{S}_{\theta N}(\omega) &= -\sum_{k=1}^{N-1} \frac{\dot{p}_{\omega_k,\omega_{k+1}}(\theta)}{p_{\omega_k,\omega_{k+1}}(\theta)}, \\
\ddot{S}_{\theta N}(\omega) &= \sum_{k=1}^{N-1} \left(-\frac{\ddot{p}_{\omega_k,\omega_{k+1}}(\theta)}{p_{\omega_k,\omega_{k+1}}(\theta)} + \left[\frac{\dot{p}_{\omega_k,\omega_{k+1}}(\theta)}{p_{\omega_k,\omega_{k+1}}(\theta)}\right]^2 \right).
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}_{\theta N}(\dot{S}_{\theta N}) &= -\sum_{\omega\in\Omega^N} \sum_{k=1}^{N-1} \frac{\dot{p}_{\omega_k,\omega_{k+1}(\theta)}}{p_{\omega_k,\omega_{k+1}}(\theta)} p_{\omega_1}(\theta) p_{\omega_1,\omega_2}(\theta) \cdots p_{\omega_{N-1},\omega_N}(\theta) \\
&= -\sum_{k=1}^{N-1} \sum_{\omega\in\Omega^N} \dot{p}_{\omega_k,\omega_{k+1}(\theta)} p_{\omega_1}(\theta) \prod_{\substack{j=1 \\ j\neq k}}^{N-1} p_{\omega_j,\omega_{j+1}}(\theta) \\
&= -(N-1) \sum_{\omega_1\in\Omega} p_{\omega_1}(\theta) \frac{\mathrm{d}}{\mathrm{d}\theta} \left(\sum_{\omega_2\in\Omega} p_{\omega_1,\omega_2}(\theta) \right) \\
&= 0
\end{aligned}$$

and the expectation value of the second derivative has

$$\begin{aligned}
\mathbb{E}_{\theta N}(\ddot{S}_{\theta N}) &= \sum_{\omega\in\Omega^N} \sum_{k=1}^{N-1} \left(-\frac{\ddot{p}_{\omega_k,\omega_{k+1}}(\theta)}{p_{\omega_k,\omega_{k+1}}(\theta)} + \left[\frac{\dot{p}_{\omega_k,\omega_{k+1}}(\theta)}{p_{\omega_k,\omega_{k+1}}(\theta)}\right]^2 \right) p_{\omega_1}(\theta) p_{\omega_1,\omega_2}(\theta) \cdots p_{\omega_{N-1},\omega_N}(\theta) \\
&= -(N-1) \sum_{\omega_1\in\Omega} p_{\omega_1}(\theta) \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \left(\sum_{\omega_2\in\Omega} p_{\omega_1,\omega_2}(\theta) \right) + (N-1) \sum_{\omega_1,\omega_2\in\Omega} p_{\omega_1}(\theta) \frac{[\dot{p}_{\omega_1,\omega_2}(\theta)]^2}{p_{\omega_1,\omega_2}(\theta)} \\
&= (N-1)\mathcal{I}(\theta).
\end{aligned}$$

The variance of the first derivative has

$$
\begin{aligned}
\mathrm{Var}_{\theta N}(\dot{S}_{\theta N}) &= \mathbb{E}_{\theta N}(\dot{S}_{\theta N}^2) - \mathbb{E}(\dot{S}_{\theta N})^2 \\
&= \sum_{\omega \in \Omega^N} \sum_{j,k=1}^{N-1} \frac{\dot{p}_{\omega_j,\omega_{j+1}}(\theta)}{p_{\omega_j,\omega_{j+1}}(\theta)} \frac{\dot{p}_{\omega_k,\omega_{k+1}}(\theta)}{p_{\omega_k,\omega_{k+1}}(\theta)} p_{\omega_1}(\theta) p_{\omega_1,\omega_2}(\theta) \cdots p_{\omega_{N-1},\omega_N}(\theta) \\
&= \sum_{\omega \in \Omega^N} \left(\sum_{k=1}^{N-1} \frac{[\dot{p}_{\omega_k,\omega_{k+1}}(\theta)]^2}{p_{\omega_k,\omega_{k+1}}(\theta)} p_{\omega_1}(\theta) \prod_{j \neq k} p_{\omega_j,\omega_{j+1}}(\theta) \right. \\
&\quad \left. + 2 \sum_{k=2}^{N-1} \sum_{j=1}^{k-1} \dot{p}_{\omega_j,\omega_{j+1}}(\theta) \dot{p}_{\omega_k,\omega_{k+1}}(\theta) p_{\omega_1}(\theta) \prod_{i \neq j,k} p_{\omega_i,\omega_{i+1}}(\theta) \right) \\
&= (N-1) \sum_{\omega_1,\omega_2 \in \Omega} p_{\omega_1}(\theta) \frac{[\dot{p}_{\omega_1,\omega_2}(\theta)]^2}{p_{\omega_1,\omega_2}(\theta)} \\
&\quad + 2 \sum_{k=2}^{N-1} \sum_{j=1}^{k-1} \sum_{\omega_j,\ldots,\omega_{k+1} \in \Omega} \dot{p}_{\omega_j,\omega_{j+1}}(\theta) \dot{p}_{\omega_k,\omega_{k+1}}(\theta) p_{\omega_j}(\theta) p_{\omega_{j+1},\omega_{j+2}}(\theta) \cdots p_{\omega_{k-1},\omega_k}(\theta) \\
&= (N-1)\mathcal{I}(\theta).
\end{aligned}
$$

C Upper Bound of Gaussian Integrals

We want to find an upper bound in terms of $\epsilon > 0$ for the quantity

$$\sup_{[A,B]\subset\mathbb{R}} \frac{1}{\sqrt{2\pi}} \max\left\{ \int_{[A'_\epsilon,B'_\epsilon]\setminus[A,B]} \mathrm{e}^{-x^2/2}\,\mathrm{d}x, \int_{[A,B]\setminus[A_\epsilon,B_\epsilon]} \mathrm{e}^{-x^2/2}\mathrm{d}x \right\},$$

where $[A'_\epsilon, B'_\epsilon]$ and $[A_\epsilon, B_\epsilon]$ are defined as in (2.6) and (2.9). To this end, we evaluate the arguments of the maximum individually. Consider the first argument, which has

$$\begin{aligned}\int_{[A'_\epsilon,B'_\epsilon]\setminus[A,B]} \mathrm{e}^{-x^2/2}\,\mathrm{d}x &= \int_{A'_\epsilon}^{A} \mathrm{e}^{-x^2/2}\,\mathrm{d}x + \int_{B}^{B'_\epsilon} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \\ &\leq \max_{x\in[A'_\epsilon,A]} \mathrm{e}^{-x^2/2}(A - A'_\epsilon) + \max_{x\in[B,B'_\epsilon]} \mathrm{e}^{-x^2/2}(B'_\epsilon - B).\end{aligned}$$

No matter where the interval $[A, B]$ lies with respect to zero, we have

$$A - A'_\epsilon = |A|\epsilon(K+1) \quad \text{and} \quad B'_\epsilon - B = |B|\epsilon(K+1).$$

Since $\mathrm{e}^{-x^2/2}$ is greater when x^2 is closer to zero and $\epsilon > 0$ was chosen such that $\epsilon(K+1) < 1$ for $K > 0$, then

$$\int_{[A'_\epsilon,B'_\epsilon]\setminus[A,B]} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \leq \mathrm{e}^{-\frac{A^2}{2}(1-\epsilon(K+1))^2}|A|\epsilon(K+1) + \mathrm{e}^{-\frac{B^2}{2}(1-\epsilon(K+1))^2}|B|\epsilon(K+1).$$

Similarly for the second argument of the maximum, we have

$$A_\epsilon - A = |A|\epsilon(K+1) \quad \text{and} \quad B - B_\epsilon = |B|\epsilon(K+1),$$

and hence

$$\int_{[A,B]\setminus[A_\epsilon,B_\epsilon]} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \leq \mathrm{e}^{-\frac{A^2}{2}(1-\epsilon(K+1))^2}|A|\epsilon(K+1) + \mathrm{e}^{-\frac{B^2}{2}(1-\epsilon(K+1))^2}|B|\epsilon(K+1).$$

Therefore, we obtain the upper bound

$$\begin{aligned}
&\sup_{[A,B]\subset\mathbb{R}} \frac{1}{\sqrt{2\pi}} \max\left\{ \int_{[A'_\epsilon,B'_\epsilon]\setminus[A,B]} \mathrm{e}^{-x^2/2}\,\mathrm{d}x, \int_{[A,B]\setminus[A_\epsilon,B_\epsilon]} \mathrm{e}^{-x^2/2}\mathrm{d}x \right\} \\
&\qquad \leq \frac{1}{\sqrt{2\pi}} \sup_{[A,B]\subset\mathbb{R}} \left(\mathrm{e}^{-\frac{A^2}{2}(1-\epsilon(K+1))^2}|A|\epsilon(K+1) + \mathrm{e}^{-\frac{B^2}{2}(1-\epsilon(K+1))^2}|B|\epsilon(K+1) \right) \\
&\qquad \leq \frac{\epsilon(K+1)}{\sqrt{2\pi}} \left(\sup_{A\in\mathbb{R}} |A|\mathrm{e}^{-\frac{A^2}{2}(1-\epsilon(K+1))^2} + \sup_{B\in\mathbb{R}} |B|\mathrm{e}^{-\frac{B^2}{2}(1-\epsilon(K+1))^2} \right) \\
&\qquad = \frac{\epsilon\sqrt{2}(K+1)}{\sqrt{\pi}} \sup_{x\in\mathbb{R}} \left(|x|\mathrm{e}^{-\frac{x^2}{2}(1-\epsilon(K+1))^2} \right).
\end{aligned}$$

The maximum value of $f(x) = |x|\mathrm{e}^{-\frac{x^2}{2}(1-\epsilon(K+1))^2}$ is obviously not attained at $x = 0$, hence we can set $f'(x) = 0$ and solve for x assuming $x \neq 0$:

$$0 = f'(x) = -x|x|(1-\epsilon(K+1))^2\mathrm{e}^{-\frac{x^2}{2}(1-\epsilon(K+1))^2} + \frac{x}{|x|}\mathrm{e}^{-\frac{x^2}{2}(1-\epsilon(K+1))^2}$$

and thus $x = \pm(1-\epsilon(K+1))^{-1}$ are the two critical points. Since f continuous on $\mathbb{R}$, twice-differentiable on $\mathbb{R}\setminus\{0\}$, has $f(0) = 0$ with $f(x) \to 0$ as $x \to \pm\infty$ and has

$$f''(x = \pm(1-\epsilon(K+1))) = -\frac{2(1-\epsilon(K+1))}{\sqrt{e}} < 0,$$

then $x = \pm(1-\epsilon(K+1))^{-1}$ are points where f achieves its global maximum of

$$f(x = \pm(1-\epsilon(K+1))^{-1}) = \frac{1}{\sqrt{e}(1-\epsilon(K+1))}.$$

Therefore, we are left with the estimate

$$\sup_{[A,B]\subset\mathbb{R}} \frac{1}{\sqrt{2\pi}} \max\left\{ \int_{[A'_\epsilon,B'_\epsilon]\setminus[A,B]} \mathrm{e}^{-x^2/2}\,\mathrm{d}x, \int_{[A,B]\setminus[A_\epsilon,B_\epsilon]} \mathrm{e}^{-x^2/2}\mathrm{d}x \right\} \leq \sqrt{\frac{2}{e\pi}} \frac{\epsilon(K+1)}{(1-\epsilon(K+1))}.$$

References

[Bah60] R. R. Bahadur. "On the Asymptotic Efficiency of Tests and Estimates". In: *Sankhyā: The Indian Journal of Statistics (1933-1960)* 22.3/4 (1960), pp. 229–252.

[Bah83] R.R. Bahadur. "Large Deviations of the Maximum Likelihood Estimate in the Markov Chain Case". In: *Recent Advances in Statistics*. Ed. by M. Haseeb Rizvi, Jagdish S. Rustagi, and David Siegmund. Academic Press, 1983, pp. 273–286.

[Bas56] D. Basu. "The Concept of Asymptotic Efficiency". In: *Sankhyā: The Indian Journal of Statistics (1933-1960)* 17.2 (1956), pp. 193–196.

[Ber41] Andrew C. Berry. "The Accuracy of the Gaussian Approximation to the Sum of Independent Variates". In: *Transactions of the American Mathematical Society* 49.1 (1941), pp. 122–136.

[BZG80] R.L. Badahur, S.L. Zabell, and J.C. Gupta. "Large Deviations, Tests, and Estimates". In: *Asymptotic Theory of Statistical Tests and Estimation*. Ed. by I.M. Chakravarti. Academic Press, 1980, pp. 33–64.

[Cra46] Harald Cramér. *Mathematical Methods of Statistics (PMS-9)*. Princeton University Press, 1946.

[Dug37] Daniel Dugué. "Application des propriétés de la limite au sens du calcul des probabilités à l'étude de diverses questions d'estimation". Doctorat d'État. 1937.

[DZ10] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques And Applications*. 2nd ed. Springer, 2010.

[Ess42] Carl-Gustav Esseen. "On the Liapounoff limit of error in the theory of probability". In: *Ark. Mat., Astr. o. Fysik 28A* (1942), pp. 1–19.

[Fel71] William Feller. *An introduction to probability theory and its applications. Vol. II.* Second edition. New York: John Wiley & Sons Inc., 1971.

[FR22] R. A. Fisher and Edward John Russell. "On the mathematical foundations of theoretical statistics". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222.594-604 (1922), pp. 309–368.

[Jak19] Vojkan Jakšić. "Lectures on Entropy. I: Information-Theoretic Notions". In: *Bahns et al (Eds): Dynamical Methods in Open Quantum Systems, Tutorials, Schools and Workshops in the Mathematical Sciences*. Springer, 2019, pp. 141–268.

[KK86] A. D. M. Kester and W. C. M. Kallenberg. "Large Deviations of Estimators". In: *The Annals of Statistics* 14.2 (1986), pp. 648–664.

[KP02] Steven G. Krantz and Harold R. Parks. *A Primer of Real Analytic Functions*. Birkhäuser Advanced Texts Basler Lehrbücher. Birkhäuser Boston, MA, 2002.

[Man96] B.W. Mann. "Berry-Esseen Central Limit Theorems for Markov Chains". PhD thesis. Cambridge: Harvard University, 1996.

[Mat23] Gabriele Di Matteo. "Results in Parameter Estimation for Finite Markov Chains". MA thesis. Montréal: McGill University, 2023.

[Pfa71] J. Pfanzagl. "The Berry-Esseen bound for minimum contrast estimates". In: *Metrika* 17.1 (Dec. 1971), pp. 82–91.

[Pra73] B. L. S. Prakasa Rao. "On the rate of convergence of estimators for Markov processes". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 26.2 (June 1973), pp. 141–152.

[Roc70] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton: Princeton University Press, 1970.

[Sch12] Joel L. Schiff. *Perturbation Theory for Linear Operators*. 2nd ed. Classics in Mathematics. Springer Berlin, Heidelberg, 2012.

[She01] Xiaotong Shen. "On Bahadur Efficiency and Maximum Likelihood Estimation in General Parameter Spaces". In: *Statistica Sinica* 11.2 (2001), pp. 479–498.