# Analysing Minimax prediction risk of Markov Chains

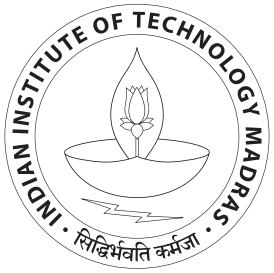
*A Project Report*

*submitted by*

**ADITYA PRADEEP**

*in partial fulfilment of the requirements*
*for the award of the degree of*

**BACHELOR OF TECHNOLOGY &**
**MASTER OF TECHNOLOGY**




**DEPARTMENT OF ELECTRICAL ENGINEERING**
**INDIAN INSTITUTE OF TECHNOLOGY, MADRAS.**
**May 2019**

# THESIS CERTIFICATE

This is to certify that the thesis entitled **Analysing Minimax prediction risk of Markov Chains**, submitted by **Aditya Pradeep** (**EE14B068**), to the Indian Institute of Technology, Madras, for the award of the degree of **Bachelors of Technology** and **Master of Technology**, is a bona fide record of the research work carried out by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

**Dr. Andrew Thangaraj**
Research Guide
Professor
Dept. of Electrical Engineering
IIT-Madras, 600 036

Place: Chennai

Date: 6th May 2019

# ACKNOWLEDGEMENTS

I would like to thank my advisor Prof. Andrew Thangaraj for his continuous support during my Master's project. His vast knowledge in this area allowed me to try out various ideas over the course of the last one year. His patience with me during the testing times were critical in motivating me to put more effort into this work.

I would also like to thank IIT Madras for providing me access to an outstanding education and great professors.

Lastly, I would like to thank my parents and family for being supportive of me during the course of this work.

# ABSTRACT

KEYWORDS: Markov chain, prediction risk, minimax risk

In the field of statistical learning, a significant problem is trying to estimate an unknown distribution from its samples. This problem has been studied very thoroughly with respect to iid distributions. However, the same problem in a Markov chain setting has not seen much research. Markov chains have a lot of practical significance as they increase the complexity of the unknown distribution. In many real life applications, like speech processing, words in sentences can be modelled as a markov chain - as the next word in a sentence depends on the previous words.

In this report, we first discuss the minimax squared error(SE) risk for the 2-state Markov chain and show that it is $O\left(\frac{1}{n}\right)$. We then extend this to the $k$-state Markov chain showing the lower bound $O\left(\frac{1}{kn}\right)$ and upper bound $O\left(\frac{k}{n}\right)$ respectively. Finally, we improve the upper bound for the KL-divergence risk to $O\left(\frac{k \log\log n}{n}\right)$ which also equals an existing lower bound.

# TABLE OF CONTENTS

# NOTATION

| | |
|---|---|
| $\alpha$ | Transition probability from state(0) to state(1) in 2-state Markov chain |
| $\beta$ | Transition probability from state(1) to state(0) in 2-state Markov chain |
| $\mathbb{M}_k$ | $k$-state Markov chain |
| SE | Square Error |
| KL | Kullback-Leibler divergence |
| $\rho_n$ | minimax risk |

# CHAPTER 1

# INTRODUCTION

## 1.1 Motivation

The prediction problem in statistical learning has been a fundamental problem for a long time. Essentially, the problem is to estimate an unknown finite distribution after observing a certain number of samples. This problem has been studied extensively for the iid distribution.

A similar problem on estimating Markov chains, has not been studied. Markov chains add a layer of complexity to the problem, while still remaining relatively simple. Markov chains can also be used to model lots of practical distributions like in speech processing - words going to appear in a sentence depends on the previous words; ecology, finance-stock markets etc.

In this report, we focus on the prediction problem for Markov chains. We quantify this by defining a quantity called risk and try to quantify the minimax risk ( essentially the "worst-case risk") by finding upper and lower bounds for it.

## 1.2 Problem Definition

A sequence of random variables $X^n = X_1, X_2, \ldots$ with $X_i \in [k] \triangleq \{0, 1, \ldots, k-1\}$ satisfying

$$\Pr(X^n = x^n) = \Pr(X_1 = x_1) \prod_{i=2}^{n} \Pr(X_i = x_i | X_{i-1} = x_{i-1}),\ n = 1, 2, \ldots,$$

is said to belong to a $k$-state, memory-1 stationary Markov chain if the state transition probability $P(v|u) \triangleq \Pr(X_i = v|X_{i-1} = u)$, $u, v \in [k]$, is independent of $i$, and the initial state distribution $\Pr(X_1)$ is the unique stationary distribution $\pi \triangleq \Pr(u)$, $u \in [k]$, for $P(v|u)$ satisfying $\pi(v) = \sum_u \pi(u)P(v|u)$, $u, v \in [k]$. The collection of such Markov chains is denoted $\mathbb{M}_k$ and each chain in the collection is parametrized by $P(v|u)$, and we will let $P$ denote the matrix of values $P(v|u)$, $u, v \in [k]$.

Given a sequence $X^n = [X_1, X_2, \ldots, X_n]$ from $\mathbb{M}_k$ with $P(v|u)$ unknown, we are interested in the prediction problem Falahatgar *et al.* [2016], which is the estimation of the random vector

$$\Theta(P, X_n) \triangleq [P(0|X_n)\ P(1|X_n)\ \cdots\ P(k{-}1|X_n)].$$

An estimator for $\Theta(P, X_n)$ using an observation $X^n = x^n$, denoted $\hat{\Theta}(x^n)$, is defined as

$$\hat{\Theta}(x^n) = [\hat{P}(0|x^n)\ \hat{P}(1|x^n)\ \cdots\ \hat{P}(k{-}1|x^n)],$$

where $\hat{P}(a|x^n) : [k] \times [k]^n \to [0, 1]$, $a \in [k]$.

We consider the minimax prediction squared-error risk, defined as

$$\begin{aligned}\rho_n^{SE}(\mathbb{M}_k) &= \min_{\hat{\Theta}} \max_{P \in \mathbb{M}_k} \mathop{\mathbb{E}}_{X^n \sim P} \left\| \Theta(P, X_n) - \hat{\Theta}(X^n) \right\|^2 \\ &= \min_{\hat{P}} \max_{P \in \mathbb{M}_k} \mathop{\mathbb{E}}_{X^n \sim P} \sum_{a=0}^{k-1} \left| P(a|X_n) - \hat{P}(a|X^n) \right|^2. \end{aligned} \tag{1.1}$$

## 1.3 Outline of Thesis

The rest of the thesis is organised as follows. Chapter 2 discussed a standard method called the Le Cam's method and a modification. Chapter 3 analyses the the squared error risk for a 2-state Markov chain. Chapter 4 extends similar ideas to the minimax risk of a $k$-state Markov chain.

# CHAPTER 2

# Useful Techniques for analysing minimax risk

The ideas from this chapter are discussed in Duchi [2016] and Kahlon [2018].

## 2.1 Le Cam's Method

Let $\mathcal{P}$ be a set of distributions and let $X_1, X_2, \ldots, X_n$ be a sample from some distribution $P \in \mathcal{P}$. Let $\theta = \theta(P)$ be some function of $P$. Let $\hat{theta} = \hat{theta}(X_1, X_2, \ldots, X_n)$ denote an estimator and $d$ be some distance metric satisfying triangle inequality and $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$ be a non-decreasing function with $\Phi(0) = 0$. Let the minimax risk be defined as

$$R_n^* = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} E_P \left[ \Phi(d(\hat{\theta}, \theta) \right] \tag{2.1}$$

Then, for any pair $P_0, P_1 \in \mathcal{P}$, let $\Delta = \dfrac{d(\theta(P_0), \theta(P_1))}{2}$, then

$$R_n^* \geq \frac{1}{2}\Phi(\Delta)\left[1 - ||P_0 - P_1||_{TV}\right] \tag{2.2}$$

For our lower bound calculations, our risk functions are not in the form of 2.1. Hence, we need to modify the Le Cam method.

## 2.2 Modified Le Cam's Method

Let $\mathcal{P}$ be a set of distributions and let $X_1, X_2, \ldots, X_n$ be a sample from some distribution $P \in \mathcal{P}$. Let $\theta_1 = \theta_1(P), \theta_2 = \theta_2(P), \gamma_1 = \gamma_1(P)$ and $\gamma_2 = \gamma_2(P)$ be some functions of $P$. Let $\hat{\theta_1} = \hat{\theta_1}(X_1, X_2, \ldots, X_n)$ and $\hat{\theta_2} = \hat{\theta_2}(X_1, X_2, \ldots, X_n)$ be the estimators of $\theta_1$ and $\theta_2$ respectively.$d$ is some metric distance satisfying triangle inequality and $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$ be a non-decreasing function with $\Phi(0) = 0$. Let the minimax risk be defined as

$$R_n^* = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} E_P \left[ \gamma_1 \Phi(d(\hat{\theta_1}, \theta_1)) + \gamma_2 \Phi(d(\hat{\theta_2}, \theta_2)) \right] \tag{2.3}$$

Then the lower bound on $R_n^*$ is given by

**Theorem 1.** *For any pair $P_0, P_1 \in \mathcal{P}$. Let $\Delta_1 = \frac{d(\theta_1(P_0), \theta_1(P_1))}{2}$ and $\Delta_2 = \frac{d(\theta_2(P_0), \theta_2(P_1))}{2}$. Then,*

$$\begin{aligned} R_n^* \geq & \frac{1}{2} \min(\gamma_1(P_0), \gamma_1(P_1)) \Phi(\Delta_1) \left[ 1 - \sqrt{\frac{1}{2} D_{KL}(P_0 || P_1)} \right] \\ & + \frac{1}{2} \min(\gamma_2(P_0), \gamma_2(P_1)) \Phi(\Delta_2) \left[ 1 - \sqrt{\frac{1}{2} D_{KL}(P_0 || P_1)} \right] \end{aligned} \tag{2.4}$$

*Proof.* An estimator $\hat{\theta}_1$ defines a test static $\psi_1$, namely,

$$\psi_1(X_1, X_2, \ldots, X_n) = \begin{cases} 1, & \text{if } d(\hat{\theta}, \theta_1(P_0)) \geq d(\hat{\theta}, \theta_1(P_1)) \\ 0, & \text{if } d(\hat{\theta}, \theta_1(P_0)) < d(\hat{\theta}, \theta_1(P_1)) \end{cases} \tag{2.5}$$

Similarly estimator $\hat{\theta}_2$ defines a test static $\psi_2$

$$\psi_2(X_1, X_2, \ldots, X_n) = \begin{cases} 1, & \text{if } d(\hat{\theta}, \theta_2(P_0)) \geq d(\hat{\theta}, \theta_2(P_1)) \\ 0, & \text{if } d(\hat{\theta}, \theta_2(P_0)) < d(\hat{\theta}, \theta_2(P_1)) \end{cases} \tag{2.6}$$

If $P = P_0$ and $\psi = 1$, then

$$2\Delta_1 = d(\theta_1(P_0), \theta_1(P_1)) \leq d(\theta_1(P_0), \hat{\theta}_1(P_1)) + d(\theta_1, \hat{\theta}) \leq 2d(\theta_1(P_0), \hat{\theta}) \tag{2.7}$$

$$\implies d(\theta_1(P_0), \hat{\theta}) \geq \Delta_1 \tag{2.8}$$

and so $\Phi(d(\theta_1(P_0), \hat{\theta})) \geq \Phi(\Delta)$. Hence,

$$\begin{aligned} E_{P_0}\Big[\gamma_1(P_0)\Phi(d(\hat{\theta}, \theta_1(P_0)))\Big] &\geq E_{P_0}\Big[\gamma_1(P_0)\Phi(d(\hat{\theta}, \theta_1(P_0)))I(\psi_1 = 1)\Big] \\ &\geq \gamma_1(P_0)\Phi(\Delta_1)E_{P_0}[I(\psi_1 = 1)] \\ &= \gamma_1(P_0)\Phi(\Delta_1)P_0(\psi_1 = 1) \end{aligned} \tag{2.9}$$

Similarly,

$$E_{P_1}\Big[\gamma_1(P_1)\Phi(d(\hat{\theta}, \theta_1(P_1)))\Big] \geq \gamma_1(P_1)\Phi(\Delta_1)P_1(\psi_1 = 0) \tag{2.10}$$

$$E_{P_0}\Big[\gamma_2(P_0)\Phi(d(\hat{\theta}, \theta_2(P_0)))\Big] \geq \gamma_2(P_0)\Phi(\Delta_2)P_0(\psi_2 = 1) \tag{2.11}$$

$$E_{P_1}\Big[\gamma_2(P_1)\Phi(d(\hat{\theta}, \theta_2(P_1)))\Big] \geq \gamma_2(P_1)\Phi(\Delta_2)P_1(\psi_2 = 0) \tag{2.12}$$

From the above, we can show that

$$\begin{aligned} R_{P_1} &= E_{P_0}\Big[\gamma_1(P_0)\Phi(d(\hat{\theta},\theta_1(P_0))) + \gamma_2(P_0)\Phi(d(\hat{\theta},\theta_2(P_0)))\Big] \\ &\geq \gamma_1(P_0)\Phi(\Delta_1)P_0(\psi_1=1) + \gamma_2(P_0)\Phi(\Delta_2)P_0(\psi_2=1) \end{aligned} \tag{2.13}$$

$$\begin{aligned} R_{P_2} &= E_{P_1}\Big[\gamma_1(P_1)\Phi(d(\hat{\theta},\theta_1(P_1))) + \gamma_2(P_1)\Phi(d(\hat{\theta},\theta_2(P_1)))\Big] \\ &\geq \gamma_1(P_1)\Phi(\Delta_1)P_1(\psi_1=0) + \gamma_2(P_1)\Phi(\Delta_2)P_1(\psi_2=0) \end{aligned} \tag{2.14}$$

We can thus write our risk as

$$\sup_{P\in\mathcal{P}} R_P \geq \max_{P\in P_0,P_1} R_P \geq \frac{R_{P_1}+R_{P_2}}{2} \tag{2.15}$$

$$\geq \frac{1}{2}\Big(\gamma_1(P_0)\Phi(\Delta_1)P_0(\psi_1=1) + \gamma_2(P_0)\Phi(\Delta_2)P_0(\psi_2=1)\Big) \tag{2.16}$$

$$+ \frac{1}{2}\Big(\gamma_1(P_1)\Phi(\Delta_1)P_1(\psi_1=0) + \gamma_2(P_1)\Phi(\Delta_2)P_1(\psi_2=0)\Big) \tag{2.17}$$

$$\geq \min(\gamma_1(P_0),\gamma_1(P_1))\Phi(\Delta_1)\inf_{\psi_1}\left[\frac{P_0(\psi_1=1)+P_1(\psi_1=0)}{2}\right] \tag{2.18}$$

$$+\min(\gamma_2(P_0),\gamma_2(P_1))\Phi(\Delta_2)\inf_{\psi_2}\left[\frac{P_0(\psi_2=1)+P_1(\psi_2=0)}{2}\right] \tag{2.19}$$

Using the result that

$$\inf_{\psi}\Big(P_0(\psi=1)+P_1(\psi=0)\Big) = 1-||P_0-P_1||_{TV} \tag{2.20}$$

we get,

$$R_n^* \geq \frac{1}{2}\min(\gamma_1(P_0), \gamma_1(P_1))\Phi(\Delta_1)[1 - ||P_0 - P_1||_{TV}] + \frac{1}{2}\min(\gamma_2(P_0), \gamma_2(P_1))\Phi(\Delta_2)[1 - ||P_0 - P_1||_{TV}] \tag{2.21}$$

We know,

$$||P_0 - P_1||_{TV}^2 \leq \frac{1}{2}D_{KL}(P_0||P_1) \tag{2.22}$$

and therefore

$$R_n^* \geq \frac{1}{2}\min(\gamma_1(P_0), \gamma_1(P_1))\Phi(\Delta_1)\left[1 - \sqrt{\frac{1}{2}D_{KL}(P_0||P_1)}\right] + \frac{1}{2}\min(\gamma_2(P_0), \gamma_2(P_1))\Phi(\Delta_2)\left[1 - \sqrt{\frac{1}{2}D_{KL}(P_0||P_1)}\right] \tag{2.23}$$

□

Similarly, if we extended to other $\gamma_i, \theta_i$ terms we have

$$R_n^* = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} E_P\left[\gamma_1\Phi(d(\hat{\theta_1}, \theta_1)) + \gamma_2\Phi(d(\hat{\theta_2}, \theta_2)) + \ldots \gamma_N\Phi(d(\hat{\theta_N}, \theta_N))\right] \tag{2.24}$$

Then the lower bound on $R_n^*$ is given by

**Theorem 2.** *For any pair $P_0, P_1 \in \mathcal{P}$. Let $\Delta_1 = \frac{d(\theta_1(P_0), \theta_1(P_1))}{2}$, $\Delta_2 = \frac{d(\theta_2(P_0), \theta_2(P_1))}{2}, \ldots$ and*

$\Delta_N = \frac{d(\theta_N(P_0),\theta_N(P_1))}{2}$ . *Then,*

$$\begin{aligned} R_n^* \geq & \frac{1}{2}\min(\gamma_1(P_0),\gamma_1(P_1))\Phi(\Delta_1)\left[1-\sqrt{\frac{1}{2}D_{KL}(P_0||P_1)}\right] \\ & + \frac{1}{2}\min(\gamma_2(P_0),\gamma_2(P_1))\Phi(\Delta_2)\left[1-\sqrt{\frac{1}{2}D_{KL}(P_0||P_1)}\right] \\ & + \dots \\ & + \frac{1}{2}\min(\gamma_N(P_0),\gamma_N(P_1))\Phi(\Delta_N)\left[1-\sqrt{\frac{1}{2}D_{KL}(P_0||P_1)}\right] \end{aligned} \tag{2.25}$$

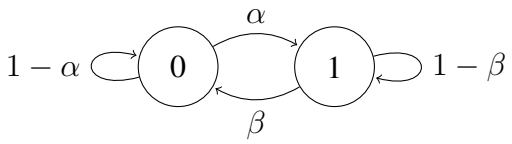
The proof follows similar to $Eqn$ 1.

# CHAPTER 3

# Analysis of squared error risk for 2-state Markov chain

## 3.1 Problem Definition and Results

The prediction problem defined in chapter 1 can be rewritten here specifically for the 2-state Markov chain.




The Markov chain represents the transition probabilities from the only two possible observable states 0 and 1. Given a sequence $X^n = [X_1, X_2, \ldots, X_n]$ from $\mathbb{M}_2$ with $\alpha, \beta$ unknown, we are interested in the prediction problem - that is what is the risk in estimating the next symbol $X_{n+1}$.

We show the following.

**Theorem 3.**

$$\begin{aligned}\rho_n^{SE}(\mathbb{M}_2) &= \min_{\hat{P}} \max_{P \in \mathbb{M}_2} \mathop{\mathbb{E}}_{X^n \sim P} \left( P(0|X_n) - \hat{P}(0|X^n) \right)^2 + \left( P(1|X_n) - \hat{P}(1|X^n) \right)^2 \\ &= 2 \min_{\hat{P}} \max_{P \in \mathbb{M}_2} \mathop{\mathbb{E}}_{X^n \sim P} \left( P(0|X_n) - \hat{P}(0|X^n) \right)^2 . \end{aligned} \tag{3.1}$$

$$\frac{0.9992}{27n} - o\left(\frac{1}{n}\right) \leq \rho_n^{SE}(\mathbb{M}_2) \leq O\left(\frac{1}{n}\right). \tag{3.2}$$

## 3.2 Lower Bound

The lower bound idea and proof in this subsection is got from Kahlon [2018].

Define $\rho$ as follows.

$$\begin{aligned}
\rho &= \mathop{\mathbb{E}}_{X^n \sim \mathbb{M}_2} \left[ |p(x_{n+1} = 0|x^n) - q(x_{n+1} = 0|x^n)|^2 \right] \\
&= p(x_n = 0) \mathop{\mathbb{E}}_{X^n \sim \mathbb{M}_2 | x_n = 0} \left[ |p(x_{n+1} = 0|x^n) - q(x_{n+1} = 0|x^n)|^2 \right] \\
&+ p(x_n = 1) \mathop{\mathbb{E}}_{X^n \sim \mathbb{M}_2 | x_n = 1} \left[ |p(x_{n+1} = 0|x^n) - q(x_{n+1} = 0|x^n)|^2 \right] \\
&= p(x_n = 0) \mathop{\mathbb{E}}_{X^n \sim \mathbb{M}_2 | x_n = 0} |\alpha - \hat{\alpha}|^2 + p(x_n = 1) \mathop{\mathbb{E}}_{X^n \sim \mathbb{M}_2 | x_n = 1} |\beta - \hat{\beta}|^2
\end{aligned} \tag{3.3}$$

Using the modified Le Cam method 1, we show that

$$\begin{aligned}
\rho_n^{SE}(\mathbb{M}^2) \geq \min \Big( p_0(x_n = 0), p_1(x_n = 0) \Big) \left( \frac{|\alpha_0 - \hat{\alpha_0}|^2}{2} \right) \left[ 1 - \sqrt{\frac{1}{2} D_{KL}(P_{0|X_n=0} || P_{1|X_n=0})} \right] \\
+ \min \Big( p_0(x_n = 1), p_1(x_n = 1) \Big) \left( \frac{|\beta_0 - \hat{\beta}_0|^2}{2} \right) \left[ 1 - \sqrt{\frac{1}{2} D_{KL}(P_{0|X_n=1} || P_{1|X_n=1})} \right]
\end{aligned} \tag{3.4}$$

where $P_0$ and $P_1$ are 2-state markov chains with parameters $(\alpha_0, \beta_0)$ and $(\alpha_1, \beta_1)$ respectively.$P_{0|X_n=0}$ is the conditional distribution of $P_0$ given that $X_n = 0$. Similarly the other terms are defined.

For finding the lower bound, the following distributions are considered.

$$P_0 : \alpha_0 = 1 - \beta_0 = \frac{1 + \delta}{2}$$

$$P_1 : \alpha_1 = 1 - \beta_1 = \frac{1 - \delta}{2}$$

Since $\alpha = 1 - \beta$, the samples $X^n = X_1, X_2, \ldots, X_n$ will become iid Bernoulli with $P(X = 0) = 1 - \alpha$ and $P(X = 1) = \alpha$. Now, since both $P_0$ and $P_1$ are iid Bernoulli, $P(X^n|X_n) = P(X^{n-1})$. Thus,

$$D_{KL}\bigg(P_0(X^n|X_n = 0)||P_1(X^n|X_n = 0)\bigg) = D_{KL}\bigg(P_0(X^{n-1})||P_1(X^{n-1})\bigg) \tag{3.5}$$

$$D_{KL}\bigg(P_0(X^n|X_n = 1)||P_1(X^n|X_n = 1)\bigg) = D_{KL}\bigg(P_0(X^{n-1})||P_1(X^{n-1})\bigg) \tag{3.6}$$

Since the distributions are iid,

$$D_{KL}\bigg(P_0(X^{n-1})||P_1(X^{n-1})\bigg) = (n-1)\delta \log\left(\frac{1+\delta}{1-\delta}\right) \tag{3.7}$$

Noting that $\delta \log\left(\frac{1+\delta}{1-\delta}\right) \leq 3\delta^2$ for $\delta \in [0, \frac{1}{2}]$, we obtain

$$D_{KL}\bigg(P_0(X^{n-1})||P_1(X^{n-1})\bigg) \leq 3(n-1)\delta^2 \tag{3.8}$$

Plugging the above in equation 3.4, we get

$$\begin{aligned}
\rho_n^{SE}(\mathbb{M}_2) \geq & \min\left(\frac{1+\delta}{2}, \frac{1-\delta}{2}\right)\left(\frac{\delta}{2}\right)^2\left[1 - \sqrt{\frac{3(n-1)\delta^2}{2}}\right] \\
& + \min\left(\frac{1+\delta}{2}, \frac{1-\delta}{2}\right)\left(\frac{\delta}{2}\right)^2\left[1 - \sqrt{\frac{3(n-1)\delta^2}{2}}\right] \\
\implies \rho_n^{SE}(\mathbb{M}_2) \geq & \frac{1}{4}(1-\delta)\delta^2\left(1 - \delta\sqrt{\frac{3(n-1)}{2}}\right)
\end{aligned} \tag{3.9}$$

For $\delta = \frac{2}{3}\sqrt{\frac{2}{3(n-1)}}$,

$$\rho_n^{SE}(\mathbb{M}_2) \geq \frac{2}{81n} - o\left(\frac{1}{n}\right) \tag{3.10}$$

## 3.3 Improvements to Lower Bound

First we consider these slightly modified distributions $P_0$ and $P_1$:

$$P_0 : \alpha_0 = 1 - \beta_0 = \frac{1 + a\delta}{2}$$

$$P_1 : \alpha_1 = 1 - \beta_1 = \frac{1 - a\delta}{2}$$

Since the distributions are iid, we have for some $\lambda$,

$$D_{KL}\left(P_0(X^{n-1})||P_1(X^{n-1})\right) = (n-1)a\delta \log\left(\frac{1+a\delta}{1-a\delta}\right) \leq (n-1)\lambda\delta^2 \tag{3.11}$$

We can then modify $Eqn$ 3.9 to get

$$\rho_n^{SE}(\mathbb{M}_2) \geq \frac{1}{4}(1-a\delta)a^2\delta^2\left(1 - \delta\sqrt{\frac{\lambda(n-1)}{2}}\right) \tag{3.12}$$

For $\delta = \frac{2}{3}\sqrt{\frac{2}{\lambda(n-1)}}$,

$$\rho_n^{SE}(\mathbb{M}_2) \geq \frac{2a^2}{27\lambda n} - o\left(\frac{1}{n}\right) \tag{3.13}$$

We find that for $a = 0.1$ and $\lambda = 0.020166$, $a^2/\lambda = 0.4996$. The author believes that the supremum of the set of values $a^2/\lambda$ is 0.5 while never equalling it.

Thus,

$$\rho_n^{SE}(\mathbb{M}_2) \geq \frac{0.9992}{27n} - o\left(\frac{1}{n}\right) \tag{3.14}$$

## 3.4 Upper Bound

In Falahatgar *et al.* [2016], they analyse the lower and upper bounds for the minimax KL divergence risk of the same problem. They show that

$$\rho_n^{KL}(\mathbb{M}_2) = \min_{\hat{P}} \max_{P \in \mathbb{M}_2} \rho_n^{KL}(P, \hat{P}) \leq \frac{2 \log \log n}{n} + O\left(\frac{1}{n}\right) \tag{3.15}$$

They prove this by taking 2 cases - single transition sequences and the remaining sequences. We modify these arguments to get results for the SE case.

### 3.4.1 Single transition sequences

Consider sequences $z_l = 0^{n-l}1^l$ of the form $00\ldots0011\ldots1111$, where the number of zeros are $l$. These sequences form the set $\overline{01}$. Similarly, we can define the set $\overline{10}$.

Define the add-$\beta$ estimator $\hat{p}(X_{n+1} = 0 | X^n = x^n) = \frac{N_{10}+\beta}{N_1+1}$ , where $N_{10}$ is the number of times a state $0$ follows state $1$ in $x^n$ ,$N_1$ is the number of occurrences of state $1$ in $x^n$ and any $\beta \in [0, 1]$. Here, for the single transition sequence $z_l$, $N_{10} = 0, N_1 = l$. Hence, $\hat{P} = \frac{\beta}{l+1}$. Then,

$$\begin{aligned}
\rho_n^{SE}(\overline{01}) &= \sum_{l=1}^{n-1} 2p(z_l)\left(p_0 - \frac{\beta}{(l+1)}\right)^2 \\
&= \sum_{l=1}^{n-1} \mu_0 p_1 (1-p_1)^{n-l-1}(1-p_0)^{l-1}\left[p_0^2 - \frac{2p_0\beta}{l+1} + \frac{\beta^2}{(l+1)^2}\right] \\
&\leq \sum_{l=1}^{n-1} \mu_0 p_1 (1-p_1)^{n-l-1}(1-p_0)^{l-1}\left[p_0^2 + \frac{\beta^2}{(l+1)^2}\right] \\
&\leq \frac{2\mu_0 p_0}{n} + \beta^2 \mu_0 \sum_{l=1}^{n-1} p_1 (1-p_1)^{n-l-1}(1-p_0)^{l-1}.\frac{1}{l^2}
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{2\mu_0 p_0}{n} + \beta^2 \mu_0 \sum_{l=1}^{n-1} p_1 (1-p_1)^{n-l-1} \frac{1}{l^2} \\
&\leq \frac{2\mu_0 p_0}{n} + \beta^2 \mu_0 \sum_{l=1}^{n-1} \frac{1}{n-l-1} \frac{1}{l^2} \\
&\leq \frac{2\mu_0 p_0}{n} + \beta^2 \mu_0 \sum_{l=1}^{n-1} \left( \frac{1}{(n-1)l^2} + \frac{1}{(n-1)^2 l} + \frac{1}{(n-1)^2 (n-l-1)} \right) \\
&\leq \frac{2\mu_0 p_0}{n} + \beta^2 \mu_0 \left( \frac{2}{(n-1)} + \frac{1}{(n-1)} + \frac{1}{(n-1)} \right) \\
&\leq O\left(\frac{1}{n}\right)
\end{aligned} \tag{3.16}
$$

We used the facts that $p_0, p_1 \leq 1$ and the lemma (from Falahatgar *et al.* [2016]) below

**Lemma 4.**

$$
\sum_{l=1}^{t} p_0 (1-p_0)^l p_1 (1-p_1)^{t-l} \leq \frac{1}{t+1}
$$

Note that his upper bound is true for any $\beta \in [0, 1]$.

Similarly, one can show for the $\overline{10}$ case and hence

$$
\rho_n^{SE}(\overline{01} \cup \overline{10}) \leq O\left(\frac{1}{n}\right) \tag{3.17}
$$

### 3.4.2 Remaining sequences

In Falahatgar *et al.* [2016] , they show that

$$
\rho_n^{KL}(X^n \setminus (\overline{01} \cup \overline{10})) \leq O\left(\frac{1}{n}\right) \tag{3.18}
$$

by assuming any add-$\beta$ estimator, and then analysing for the typical and atypical sets.

To show the similar result for the square error case, We first show the following result:

**Lemma 5.** *For* $0 \leq a, b \leq 1$*,*

$$(a-b)^2 \leq D(a||b) = a\log\left(\frac{a}{b}\right) + (1-a)\log\left(\frac{1-a}{1-b}\right)$$

*Proof.* From Pinkser's inequality we know that for any two probability distributions $P$ and $Q$,

$$d_{TV}(P,Q)^2 \leq \frac{1}{2}D_{KL}(P||Q)$$

where $d_{TV}$ is the total variation distance.

Now, define $P = a, 1-a$ and $Q = b, 1-b$, where $0 < a, b < 1$. As the set is finite, $d_{TV}(P,Q) = \frac{1}{2}||P-Q|||_1 = |a-b|$. Thus,

$$|a-b|^2 \leq \frac{1}{2}D_{KL}(a||b) \leq D_{KL}(a||b) = a\log\left(\frac{a}{b}\right) + (1-a)\log\left(\frac{1-a}{1-b}\right) \tag{3.19}$$

□

Using this lemma, we can show that

$$\rho_n^{SE}(X^n \setminus (\overline{01} \cup \overline{10})) \leq \rho_n^{KL}(X^n \setminus (\overline{01} \cup \overline{10})) \leq O\left(\frac{1}{n}\right) \tag{3.20}$$

By combining 3.17 and 3.20, we can see that

$$\rho_n^{SE}(\mathbb{M}_2) \leq O\left(\frac{1}{n}\right) \tag{3.21}$$

## 3.5 Understanding the Upper Bound

In our proof of the upper bound we used an already existing result. The proof of that result is very complicated and while it does give the correct order, the constant factor in the order is very large. Hence, we would like to see if we can find a simpler proof of the same.

Let the last symbol $X_n$ be 1. The probability of this occurring is $\frac{\beta}{\beta+\alpha}$. The square error risk given that the observed sequence ends in 1 is

$$\begin{aligned}
&\mathbb{E}_{X^n}(\alpha - \hat{\alpha})^2 \\
&= \sum_{k=1}^{n} \sum_{l} \binom{k-1}{l} \alpha^l (1-\alpha)^{k-l-1} \binom{n-k-1}{l-1} \beta^{l-1} (1-\beta)^{n-k-l} (\alpha - \hat{\alpha})^2 \\
&\leq \sum_{k=1}^{n} \sum_{l=1}^{\min(k-1,n-k)} \binom{k-1}{l} \alpha^l (1-\alpha)^{k-l-1} \binom{n-k-1}{l-1} \beta^{l-1} (1-\beta)^{n-k-l} \\
&\leq \sum_{k=1}^{n} \sum_{l=1}^{n/2} \binom{k-1}{l} \alpha^l (1-\alpha)^{k-l-1} \binom{n-k-1}{l-1} \beta^{l-1} (1-\beta)^{n-k-l}
\end{aligned} \tag{3.22}$$

The above steps follow as $\alpha, \hat{\alpha} \in (0,1)$ and $\max(\min(k-1, n-k)) = n/2$.

Now we split the summation into 2 parts and deal with each one. First we state a standard lemma which is used:

**Lemma 6.**

$$\sum_{l=0}^{\lambda n'} \binom{n'}{l} \beta^l (1-\beta)^{n'-l} \leq e^{-n' D(\lambda||\beta)} \tag{3.23}$$

### 3.5.1 For $0 \leq k \leq n/2$

Thus $n' = n - k \geq n/2$. Hence 3.22 is less than

$$\begin{aligned}
&\sum_{k=1}^{n'} \sum_{l=1}^{n/2} \binom{k-1}{l} \alpha^l (1-\alpha)^{k-l-1} \binom{n'-1}{l-1} \beta^{l-1} (1-\beta)^{n'-l} \\
&\leq \sum_{k=1}^{n/2} \sum_{l=0}^{n/2} \binom{n'}{l} \beta^l (1-\beta)^{n'-l} \\
&\leq \sum_{k=1}^{n/2} e^{-\frac{n}{2} D[\frac{1}{2}||p_1]} = \frac{n}{2} e^{-\frac{n}{2} D[\frac{1}{2}||p_1]}
\end{aligned} \tag{3.24}$$

### 3.5.2 For $n/2 \leq k \leq n$

Thus $n' = k - 1 \geq n/2$. Hence 3.22 is less than

$$\begin{aligned}
&\sum_{k=n/2}^{n} \sum_{l=0}^{n/2} \binom{k-1}{l} \alpha^l (1-\alpha)^{k-l-1} \\
&\leq \sum_{k=n/2}^{n} \sum_{l=0}^{n/2} \binom{n'}{l} \alpha^l (1-\alpha)^{n'-l} \\
&\leq \sum_{k=n/2}^{n} e^{-\frac{n}{2} D[\frac{1}{2}||p_0]} = \frac{n}{2} e^{-\frac{n}{2} D[\frac{1}{2}||p_0]}
\end{aligned} \tag{3.25}$$

Both 3.24 and 3.25 are of the form $f(x) = \frac{n}{2} e^{-\frac{n}{2} D[\frac{1}{2}||x]} = \frac{n}{2} e^{-\frac{n}{4} \log_e(\frac{1}{4x(1-x)})}$

Evaluating $f(x) \leq \frac{100}{n}$ ,we can see that this is true when $0 \leq x \leq 0.283$ and $0.717 \leq x \leq 1$.

One reason for not capturing the entire range is that we are dropping the $(\alpha - \hat{\alpha})^2$ term. This result is reassuring because it takes care of probabilities which are very low and very high. The other probability ranges centered around $0.5$ can be possibly dealt with by using concentration results.

# CHAPTER 4

# Analysis of squared error risk for $k$-state Markov chain

## 4.1 Lower bound

Define $\rho$ as follows.

$$\begin{aligned}\rho &= \mathop{\mathbb{E}}_{X^n \sim \mathbb{M}_k}\left[|p(x_{n+1}=0|x^n) - q(x_{n+1}=0|x^n)|^2\right] \\ &= \sum_{i=1}^{n} p(x_n=i) \mathop{\mathbb{E}}_{X^n \sim \mathbb{M}_2|x_n=i}\left[|p(x_{n+1}=i|x^n) - q(x_{n+1}=i|x^n)|^2\right]\end{aligned} \tag{4.1}$$

We first show the lower bound for a $k$-state markov chain with even $k$. The odd case follows similarly.

Let $P_0$ and $P_1$ be two distributions defined as

$$P_0(x_{n+1}=i|x_n=0) = P_0(x_{n+1}=i|x_n=1) = \cdots = P_0(x_{n+1}=i|x_n=k-1)$$

$$P_0(0|.) = P_0(1|.) = \cdots = P_0(k/2-1|.) = \frac{1+\delta}{k}$$

$$P_0(k/2|.) = P_0(k/2+1|.) = \cdots = P_0(k|.) = \frac{1-\delta}{k}$$

$$P_1(x_{n+1} = i|x_n = 0) = P_1(x_{n+1} = i|x_n = 1) = \cdots = P_1(x_{n+1} = i|x_n = k-1)$$
$$P_1(0|.) = P_1(1|.) = \cdots = P_1(k/2-1|.) = \frac{1-\delta}{k}$$
$$P_1(k/2|.) = P_1(k/2+1|.) = \cdots = P_1(k|.) = \frac{1+\delta}{k} \tag{4.2}$$

Now using extended Le cam Theorem $Eqn\ 2$,

$$\rho_n^{SE}(\mathbb{M}_k) \geq \frac{1}{2}\Big[\sum_{i=0}^{k} \min(P_0(i), P_1(i))\Big]\Big[\sum_{i=1}^{k} \frac{(P_0(0|i) - P_1(0|i))^2}{4}\Big]\left(1 - \sqrt{\frac{1}{2}D_{KL}(P_0|X_n = 0||P_1|X_n = 0)}\right) \tag{4.3}$$

From our construction of $P_0$ and $P_1$, for all $i$ we have :

- $\min(P_0(i), P_1(i)) = \dfrac{1-\delta}{k}$
- $\dfrac{(P_0(0|i) - P_1(0|i))^2}{4} = \dfrac{\delta^2}{4k^2}$
- $D_{KL}(P_0|X_n = 0||P_1|X_n = 0) = D_{KL}(P_0(X^{n-1}||P_1(X^{n-1})) = (n-1)D_{KL}(P_0||P_1) = \delta \log\left(\frac{1+\delta}{1-\delta}\right)$

Hence,

$$\rho_n^{SE}(\mathbb{M}_k) \geq \frac{1}{2}\left(1-\delta\right)\left(\frac{\delta^2}{4k}\right)\left(1 - \sqrt{\frac{n-1}{2}\delta \log\left(\frac{1+\delta}{1-\delta}\right)}\right) \tag{4.4}$$

As $\delta \log\left(\frac{1+\delta}{1-\delta}\right) \leq 3\delta^2$, with $\delta = \frac{2}{3}\sqrt{\frac{2}{3(n-1)}}$,

$$\rho_n^{SE}(\mathbb{M}_k) \geq \frac{4}{81kn} - o\left(\frac{1}{n}\right) \tag{4.5}$$

Similarly, when $k$ is odd we can show the same lower bound. We just have to make small changes to the construction of $P_0$ and $P_1$ which is described as:

$$P_0(x_{n+1} = i|x_n = 0) = P_0(x_{n+1} = i|x_n = 1) = \cdots = P_0(x_{n+1} = i|x_n = k-1)$$
$$P_0(0|.) = \frac{1+2\delta}{k}$$
$$P_0(1|.) = P_0(2|.) = \frac{1-\delta}{k}$$
$$P_0(3|.) = P_0(4|.) = \cdots = P_0\left(\frac{k+3}{2}|.\right) = \frac{1+\delta}{k}$$
$$P_0\left(\frac{k+5}{2}|.\right) = P_0\left(\frac{k+7}{2}|.\right) = \cdots = P_0(k|.) = \frac{1-\delta}{k}$$

$$P_1(x_{n+1} = i|x_n = 0) = P_1(x_{n+1} = i|x_n = 1) = \cdots = P_1(x_{n+1} = i|x_n = k-1)$$
$$P_1(0|.) = P_1(2|.) = \frac{1-\delta}{k}$$
$$P_1(1|.) = \frac{1+2\delta}{k}$$
$$P_0(3|.) = P_0(4|.) = \cdots = P_0\left(\frac{k+3}{2}|.\right) = \frac{1-\delta}{k}$$
$$P_0\left(\frac{k+5}{2}|.\right) = P_0\left(\frac{k+7}{2}|.\right) = \cdots = P_0(k|.) = \frac{1+\delta}{k}$$

## 4.2 Upper Bound

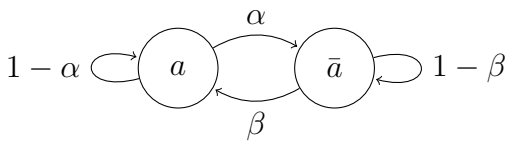
$$\begin{aligned}
\rho_n^{SE}(\mathbb{M}_k) &= \min_{\hat{P}} \max_{P \in \mathbb{M}_k} \rho_n^{SE}(\mathbb{M}_k, \hat{P}) \\
&= \min_{\hat{P}} \max_{P \in \mathbb{M}_k} \mathop{\mathbb{E}}_{X^n \sim P} \left[ \sum_{a=0}^{k-1} |p(x^{n+1} = a | X^n = x^n) - \hat{p}(x^{n+1} = a | X^n = x^n)|^2 \right] \\
&= \min_{\hat{P}} \max_{P \in \mathbb{M}_k} \sum_{a=0}^{k-1} \mathop{\mathbb{E}}_{X^n \sim P} \left[ |p(x^{n+1} = a | X^n = x^n) - \hat{p}(x^{n+1} = a | X^n = x^n)|^2 \right] \\
&\leq \min_{\hat{P}} \sum_{a=0}^{k-1} \max_{P \in \mathbb{M}_k} \mathop{\mathbb{E}}_{X^n \sim P} \left[ |p(x^{n+1} = a | X^n = x^n) - \hat{p}(x^{n+1} = a | X^n = x^n)|^2 \right] \quad (4.6)
\end{aligned}$$

Now, we consider the add-$\beta$ estimator $\hat{P}$, with $\beta = \frac{1}{k}$ as follows:

$$\hat{P}(x_{n+1} = a | X^n = x^n) = \frac{N_{x_n a} + \frac{1}{k}}{N_{x_n} + 1}, \ \forall \, 0 \leq a \leq k-1 \quad (4.7)$$

where $N_{x_n}$ is the number of occurrunces of state $x_n$ in $x^n$ and $N_{x_n a}$ is the number of times that state $a$ follows state $x_n$ in $x^n$.

For the $k$ state Markov chain and any state $a$, collapse the $k$-state Markov chain to a 2-state Markov chain with states $a$ and $\bar{a}$ , where $\bar{a} = [k] \setminus a$




From the 2-state Markov chain upper bound (setting $\beta = 1/k$), we can show that for any distribution $P$,

$$\mathop{\mathbb{E}}_{X^n\sim P}\Big[|p(x^{n+1}=a|X^n=x^n)-\hat{p}(x^{n+1}=a|X^n=x^n)|^2+|p(x^{n+1}=\overline{a}|X^n=x^n)-\hat{p}(x^{n+1}=\overline{a}|X^n=x^n)|^2\Big]\leq O\Big(\tfrac{1}{n}\Big)$$

$$\implies \mathop{\mathbb{E}}_{X^n\sim P}\Big[|p(x^{n+1}=a|X^n=x^n)-\hat{p}(x^{n+1}=a|X^n=x^n)|^2|\Big]\leq O\Big(\frac{1}{n}\Big) \tag{4.8}$$

Plugging 4.8 in 4.6 for our estimator $\hat{P}$.

$$\sum_{a=0}^{k-1}\max_{P\in\mathbb{M}_k}\mathop{\mathbb{E}}_{X^n\sim P}\Big[|p(x^{n+1}=a|X^n=x^n)-\hat{p}(x^{n+1}=a|X^n=x^n)|^2\Big]\leq O\Big(\frac{k}{n}\Big) \tag{4.9}$$

Thus, from our lower bound $Eqn$ 4.5 and upper bound $Eqn$ 4.9 we have

$$\frac{4}{81kn}-o\Big(\frac{1}{n}\Big)\leq\rho_n^{SE}(\mathbb{M}_k)\leq O\Big(\frac{k}{n}\Big) \tag{4.10}$$

## 4.3 Extending the Upper Bound for KL Divergence

In Falahatgar *et al.* [2016], they analyse the upper bound for the minimax KL divergence risk of the 2-state Markov chain. They show that

$$\rho_n^{KL}(\mathbb{M}_2)=\min_{\hat{P}}\max_{P\in\mathbb{M}_2}\rho_n^{KL}(P,\hat{P})\leq\frac{2\log\log n}{n}+O\Big(\frac{1}{n}\Big)\ \text{(or)}$$
$$\mathop{inf}_{\hat{p}}\mathop{sup}_{\alpha,\beta}\mathop{\mathbb{E}}_{X^n\sim\mathbb{M}_2}\Big[p(0|X_n)\log\frac{p(0|X_n)}{\hat{p}(0|X^n)}+p(1|X_n)\log\frac{p(1|X_n)}{\hat{p}(1|X^n)}\Big]\leq\frac{2\log\log n}{n}+O\Big(\frac{1}{n}\Big) \tag{4.11}$$

They show this by taking 2 cases

- **Single transition Sequence** : The consider sequences $z_l = 0^{n-l}1^l$ of the form $000\ldots0011\ldots11$, where the number of zeros are $l$.

  Using the estimator $\hat{P}(X_{n+1} = 0|X^n) = \frac{1}{l \log n}$, they show the upper bound here.

- **Remaining sequences** : They assume an add-$\beta$ estimator and show the upper bound for this case.

Now, consider the upper bound for the minimax KL divergence risk of the $k$-state Markov chain.

$$\begin{aligned}
\rho_n^{KL}(\mathbb{M}_k) &= \min_{\hat{P}} \max_{P \in \mathbb{M}_k} \rho_n^{KL}(\mathbb{M}_k, \hat{P}) \\
&= \min_{\hat{P}} \max_{P \in \mathbb{M}_k} \mathop{\mathbb{E}}_{X^n \sim P} \left[ \sum_{a=0}^{k-1} p(x^{n+1} = a|X^n) \log \left( \frac{p(x^{n+1} = a|X^n)}{\hat{p}(x^{n+1} = a|X^n)} \right) \right] \\
&= \min_{\hat{P}} \max_{P \in \mathbb{M}_k} \sum_{a=0}^{k-1} \mathop{\mathbb{E}}_{X^n \sim P} \left[ p(x^{n+1} = a|X^n) \log \left( \frac{p(x^{n+1} = a|X^n)}{\hat{p}(x^{n+1} = a|X^n)} \right) \right] \\
&\leq \min_{\hat{P}} \sum_{a=0}^{k-1} \max_{P \in \mathbb{M}_k} \mathop{\mathbb{E}}_{X^n \sim P} \left[ p(x^{n+1} = a|X^n) \log \left( \frac{p(x^{n+1} = a|X^n)}{\hat{p}(x^{n+1} = a|X^n)} \right) \right] \qquad (4.12)
\end{aligned}$$

Define the estimator $\hat{p}$ as follows:

$$\hat{P}(x_{n+1} = a|X^n) = \begin{cases} \frac{1}{l \log n} & x^n \text{ is a single transition sequence of form } a^l \bar{a}^{n-l} \\ 1 - \frac{1}{l \log n} & x^n \text{ is a single transition sequence of form } \bar{a}^l a^{n-l} \\ \dfrac{N_{x_n a} + \frac{1}{k}}{N_{x_n} + 1} & \text{else} \end{cases} \qquad (4.13)$$

From the 2-state Markov chain and for our estimator $\hat{p}$

$$\sup_{\alpha,\beta} \mathop{\mathbb{E}}_{X^n \sim \mathbb{M}_2} \left[ p(a|X_n) \log \frac{p(a|X_n)}{\hat{p}(a|X^n)} + p(\bar{a}|X_n) \log \frac{p(\bar{a}|X_n)}{\hat{p}(\bar{a}|X^n)} \right] \leq \frac{2 \log\log n}{n} + O\left(\frac{1}{n}\right)$$

$$\text{(or) } \sup_{\alpha,\beta} \mathop{\mathbb{E}}_{X^n \sim \mathbb{M}_2} \left[ p(a|X_n) \log \frac{p(a|X_n)}{\hat{p}(a|X^n)} \right] \leq \frac{2 \log\log n}{n} + O\left(\frac{1}{n}\right) \quad (4.14)$$

Plugging this in 4.12, we get

$$\rho_n^{KL}(\mathbb{M}_k) \leq \frac{2k \log\log n}{n} + O\left(\frac{1}{n}\right) \quad (4.15)$$

In Hao *et al.* [2018], they study the KL-divergence risk for a $k$-state Markov chain and show that

$$\frac{(k-1) \log\log n}{4en} \lesssim \rho_n^{KL}(\mathbb{M}_k) \lesssim \frac{2k^2 \log\log n}{n} \quad (4.16)$$

Thus, we improve the upper bound by a factor of $k$ to get the same order of $O\left(\frac{k \log\log n}{n}\right)$ in both the lower and upper bound. Thus,

$$\rho_n^{KL}(\mathbb{M}_k) = O\left(\frac{k \log\log n}{n}\right) \quad (4.17)$$

# CHAPTER 5

# Summary

In this report we have shown the following:

- For the 2-state Markov chain, we showed that

$$\frac{0.9992}{27n} - o\left(\frac{1}{n}\right) \leq \rho_n^{SE}(\mathbb{M}_2) \leq O\left(\frac{1}{n}\right). \tag{5.1}$$

- For the $k$-state Markov chain,

$$\frac{4}{81kn} - o\left(\frac{1}{n}\right) \leq \rho_n^{SE}(\mathbb{M}_k) \leq O\left(\frac{k}{n}\right) \tag{5.2}$$

- Lastly, for the KL-divergence risk of a $k$-state Markov chain

$$\rho_n^{KL}(\mathbb{M}_k) \leq \frac{2k \log\log n}{n} + O\left(\frac{1}{n}\right) \tag{5.3}$$

  and hence

$$\rho_n^{KL}(\mathbb{M}_k) = O\left(\frac{k \log\log n}{n}\right) \tag{5.4}$$

The gap between the lower and upper bound for the $k$-state Markov chain in the SE case can be improved.

# REFERENCES

**Duchi, J.** (2016). Lecture notes for statistics 311/electrical engineering 377. *URL: https://stanford. edu/class/stats311/Lectures/full_notes. pdf.*, **2**, 23.

**Falahatgar, M.**, **A. Orlitsky**, **V. Pichapati**, and **A. T. Suresh**, Learning markov distributions: Does estimation trump compression? *In IEEE International Symposium on Information Theory (ISIT)*. 2016.

**Hao, Y.**, **A. Orlitsky**, and **V. Pichapati** (2018). On learning markov chains. *arXiv preprint arXiv:1810.11754*.

**Kahlon, S. S.** (2018). Asymptotic properties of minimax risk for estimating the transition probabilities of markov chains.