On Semi-supervised Estimation of Distributions

by

Hasan Sabri Melihcan Erol

B.S., Bilkent University (2020)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Authored by: Hasan Sabri Melihcan Erol
Department of Electrical Engineering and Computer Science
May 19, 2023

Certified by: Lizhong Zheng
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

On Semi-supervised Estimation of Distributions

by

Hasan Sabri Melihcan Erol

Submitted to the Department of Electrical Engineering and Computer Science
on May 19, 2023, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

We study the problem of estimating the joint probability mass function (pmf) over two random variables. In particular, the estimation is based on the observation of m samples containing both variables and n samples missing one fixed variable. We adopt the minimax framework with l_p^p loss functions, and we show that the composition of uni-variate minimax estimators achieves minimax risk with the optimal first-order constant for $p \geq 2$, in the regime $m = o(n)$.

Thesis Supervisor: Lizhong Zheng
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I would first and foremost like to express my sincere appreciation to my advisor, Prof. Lizhong Zheng, for his exceptional guidance and mentoring style. When I joined the group two years ago, I was initially apprehensive about his "hands-off" approach. As someone with limited experience in information theory research, it felt like being thrust into unfamiliar terrain without knowing how to ride a bike. However, Lizhong demonstrated tremendous patience during those early months. The visions he instilled in me became like training wheels on a bike, providing stability and direction as I pedaled forward. I am truly grateful to Prof. Zheng for encouraging me to pursue ambitious ideas. He gave me the freedom to take risks and follow my sometimes audacious instincts. Just as the best way to learn how to ride a bike is by actually hopping on and pedaling, I have learned that taking the plunge and embracing challenges is the most effective way to grow as a researcher.

I would like to thank my labmates Xiangxiang Xu, Jongha Ryu, Jiejun Jin, Eric Wang, Ibrahim Alhajri, and Alexander Morgan for their invaluable contributions and thought-provoking discussions. Their insights and stimulating conversations are deeply appreciated. Special mention goes to Erixhen Sula, whose mentorship and introduction to this line of research have been invaluable.

I owe a great deal of gratitude to my friends at MIT, and those scattered across the globe. This includes Nuri Vanli, Seyhmus Guler, Toros Arikan, Sarath Pattathil, Deniz Umut Yildirim, Tanay Wakhare, Chenghao Guo, Anzo Teh, Bai Lui, Echo Demirel, Ufuk Keskin, Bunyamin Kartal, Jason Luo, Kiril Bangachev, Enric Boix, Brice Huang, and beyond to Ridvan Yesiloglu, Hikmet Demir, Berk Tinaz, Deniz Mave, T. Kutay, Ege Aydin, Sami Mavi, Mahmut Yurt and Yee Yuen. Their friendship and unwavering support have played a huge part in my journey. A special acknowledgment for Tanay and Chenghoa, for carrying me.

The time spent at ADSC with Ertem Esiner and Utku Tefenk, solving interesting problems under the guidance of Prof. Zbigniew Kalbarczyk, was not only a pleasure but a privilege I hold dear.

Also, I cannot overlook the support of Safak Oguz and Mehmet Gokagac, who have stood by me since my high school years. Their assistance continues to be invaluable to this day.

The gratitude I have expressed to everyone thus far was possible due to a giant and *grpyrm*. However, my deepest appreciation and thanks will be reserved until I finally reach the stage of my dissertation.

I would like to express my deep gratitude to my dear parents, Nilüfer and Mehmet Seyfettin Erol, as well as my older sister, Asel Gulsen and younger sister Hatice. Your unwavering love and support have been the foundation of my journey. Simply put, I am who I am because of you. You are my strength, and it is to you that I wholeheartedly dedicate this thesis. Finally I would like to thank my grandmother and grandfather for their endless *readings*.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

pmf	probability mass function
a.s.	almost surely (with respect to appropriate probability measures
i.i.d.	independent and identically distributed

General Notation

$\mathbb{R}$	set of all real numbers
$\mathbb{N}$	set of all natural numbers $1, 2, 3, ...$
$\lfloor \ \rfloor$	floor function
$\lceil \ \rceil$	ceil function
$[\mathcal{U}]^*$	set of all sequences with elements from finite set $\mathcal{U}$
$\mathcal{X}, \mathcal{Y}$	finite sets with cardinality greater than 1
$[n]$	set of first $n \in \mathbb{N}$ non-negative integers

Information Theory

$D_{KL}(. \mid .)$	KL-divergence
$D_f(. \mid .)$	f-divergence

Bachmann-Landau Asymptotic Notation

Little-oh Notation: $a_n = o(b_n)$ if and only if:

$$\lim_{n\to\infty} \frac{a_n}{b_n} = 0$$

where we assume that $\forall n \in \mathbb{N}$, $a_n, b_n > 0$

Little-ω notation $a_n = \omega(b_n)$ if and only if:

$$\lim_{n\to\infty} \frac{b_n}{a_n} = 0$$

where we assume that $\forall n \in \mathbb{N}$, $a_n, b_n > 0$

Big-O notation $a_n = O(b_n)$ if and only if:

$$\lim_{n\to\infty} \frac{b_n}{a_n} = 0$$

where we assume that $\forall n \in \mathbb{N}$, $a_n > 0$

Big-Ω notation $a_n = \Omega(b_n)$ if and only if:

$$\limsup_{n\to\infty} \frac{a_n}{b_n} < \infty$$

where we assume that $\forall n \in \mathbb{N}$, $a_n > 0$

Big-Θ notation $a_n = \Theta(b_n)$ if and only if:

$$0 < \liminf_{n\to\infty} \frac{a_n}{b_n} \leq \limsup_{n\to\infty} \frac{a_n}{b_n} < \infty$$

where we assume that $\forall n \in \mathbb{N}$, $a_n, b_n > 0$

Probability Theory

$\mathbb{E}\left[.\right]$ expectation operator

$\mathbb{E}_{\mathrm{P}}[.]$ expectation operator with respect to distribution P

supp(.) support set of a measure

$\Delta_{\mathcal{X}}$ probability simplex of pmfs of a random variable X on the alphabet $\mathcal{X}$

$\mathbb{1}\{.\}$ indicator/characteristic function which equals 1 if its input proposition is true and 0 if its input proposition is false

Bernoulli(p) distribution of a Bernoulli random variable that equals 1 with probability $p \in [0,1]$ and 0 with probability $1p$

Binomial(p, n) distribution of a Binomial random variable with parameters n, o,i.e. sum of n i.i.d. Bernoulli(p) random variables

Other

$\mathbb{R}^n$ set of all n-dimensional real column vectors $n \in \mathbb{N}$

$||x||_p$ l^p norm of a column vector

$||x||_p^p$ pth power of p norm

Chapter 1

Introduction

1.1 Probability Mass Function (PMF) Estimation

1.1.1 Motivation

The vast realm of natural phenomena encompasses a multitude of phenomena that are commonly postulated to possess inherent probabilistic characteristics. These phenomena span a wide spectrum, including but not limited to written text, spoken language, stock prices, genomic composition, disease symptoms, physical characteristics, communication noise, traffic patterns, and various others. An underlying assumption in the scientific community is that these phenomena are generated according to an unknown distribution, which motivates the practical need to approximate such distributions from observed samples. The fundamental objective is to discover a distribution, denoted as q, that effectively approximates the true but elusive distribution p in a manner that aligns with the specific goals and requirements of the given context. Surprisingly, despite numerous years dedicated to statistical research and investigation, the understanding of this problem remains relatively limited, underscoring the intricacies involved and the complexity of the task at hand.

To formulate this problem in a rigorous manner, researchers often turn to the concept of minimax performance, which provides a solid foundation for analysis [13]. In this framework, an estimator is assessed based on its performance against the worst-case distribution, highlighting the importance of robustness and resilience in the face of challenging situations. In this regard, our primary focus is directed toward determining the least worst-case loss achieved by any estimator, often referred to as the minimax loss. This quantity captures the optimal performance attainable under the most adverse circumstances, providing valuable insights into the effectiveness and reliability of the estimator.

The investigation of the minimax loss, along with the identification of the optimal estimator that achieves it, carries significant practical implications. For instance, an estimator characterized by a small KL-loss can greatly enhance the realms of data compression and stock portfolio selection. Similarly, an estimator exhibiting a small l_1 loss can yield superior performance in classification tasks, fostering advancements in various domains.

In summary, the approximation of underlying distributions from observed samples in natural phenomena poses a profound and challenging problem. By embracing the minimax performance framework and judiciously selecting suitable loss functions, researchers can delve into the investigation of optimal estimators and the determination of the least worst-case loss. The pursuit of understanding and characterizing the minimax loss serves as a cornerstone of practical significance, offering valuable insights into the robustness and efficacy of estimators in various domains. An estimator with a minimized minimax loss can contribute to advancements in data compression techniques, stock portfolio selection, and classification tasks, fostering improved decision-making processes and enhancing overall system performance. By unraveling the complexities inherent in the approximation of underlying distributions, researchers can pave the way for advancements in statistical modeling, data analysis, and algorithmic design, enabling a deeper understanding of the probabilistic nature of diverse natural phenomena and facilitating the development of more accurate and reliable estimation techniques.

1.1.2 Related Work

Within the realm of this challenging problem, early contributions have successfully resolved the minimax risk associated with estimating probability mass functions (pmfs) under the l_2^2 loss by identifying the minimax estimator [11, 10, 14]. These seminal works shed light on the optimal strategies for pmf estimation under this loss function. Subsequent studies have extended these findings by determining the constant of the first order for the minimax risk under other prominent divergences, such as the Kullback-Leibler (KL) divergence, l_1 loss, and f-divergences [2, 5, 7]. These investigations have contributed to a more comprehensive understanding of the performance limits and optimal estimation approaches in different contexts.

1.2 Semi-Supervised PMf Estimation

1.2.1 Motivation

In recent decades, there has been a significant proliferation in the sizes of available datasets, surpassing the labeling efforts dedicated to them. Consequently, these datasets exhibit heterogeneity, characterized by a substantial number of samples lacking specific variables. This novel scenario necessitates a comprehensive theoretical analysis that effectively addresses the challenges it presents. The estimators discussed in the preceding section are confined to operating within two distinct modes within this framework: either disregarding complete samples in order to estimate solely the marginal probability mass function (pmf) of the corresponding variable, or excluding the incomplete samples to formulate an estimate of the joint pmf. Drawing from established naming conventions in the machine learning literature (Pattern Recognition), these modes are commonly classified as unsupervised and supervised estimation, respectively. Both modes exhibit inherent inefficiencies, thereby prompting the development of estimators capable of effectively utilizing both types of samples, commonly referred to as semi-supervised learning. The primary focus of this thesis is to delve into the fundamental limitations of semi-supervised pmf estimators.

Specifically, our study revolves around the case involving two random variables, denoted as X and Y, which are jointly distributed according to the probability distribution function p_{XY}. We are presented with two distinct datasets: m independent and identically distributed (i.i.d.) samples of (x_i, y_i) pairs drawn from the joint distribution p_{XY}, as well as n samples solely comprising of x_j, drawn from the marginal distribution p_X. Our ultimate objective is to determine the minimax estimator of the joint distribution p_{XY}, based on the observations obtained from these datasets. As a primary outcome, we establish that the amalgamation of minimax univariate estimators effectively accomplishes the correct first-order term of risk associated with the semi-supervised estimation problem, particularly in the regime where $m = o(n)$.

The exploration of semi-supervised pmf estimation carries important practical implications. By developing estimators that can effectively leverage both complete and incomplete samples, we can improve the accuracy and efficiency of various machine learning applications. However, to fully harness the potential of semi-supervised pmf estimation, it is crucial to prioritize a rigorous theoretical analysis. Such analysis is necessary to understand and address the challenges posed by the heterogeneity of datasets, and to develop robust estimators capable of effectively utilizing both complete and incomplete samples. By delving into the fundamental limitations of semi-supervised pmf estimators, we can make significant advancements in the field, leading to improved data analysis, modeling, and decision-making in practical applications.

1.2.2 Related Work

To the knowledge of the author, the minimax pmf estimation problem with labeled and unlabeled samples remains unexplored in the existing literature. In the multivariate case, the analysis is complicated by nature's control over the number of samples with a fixed marginal. Previous works, such as [6] and [8], have addressed related complications with slight variations. Unlike these works, where the number of samples is either generated from a fixed distribution or chosen adversarially, our study focuses on the case where the number of samples is generated from a distribution adversarially chosen by nature [6, 8].

1.3 Outline of the Thesis

In Chapter 2, we establish the necessary groundwork to prepare the reader for the subsequent unveiling of results in this thesis, which will be presented in Chapter 3. Chapter 4 encompasses the proofs for the theorems introduced in chapter 3, supplemented by relevant digressions. The appendix is dedicated to the proofs in Chapter 2 and the figures.

Chapter 2

Preliminaries

2.1 Probability Mass Function (PMF) Estimation

In order to formulate the problem of PMF estimation in a rigorous manner, researchers often turn to the concept of minimax performance, which provides a solid foundation for analysis. Within this framework, any distribution p over a discrete set $\mathcal{X}$ corresponds to an element residing within the simplex denoted as $\Delta_{\mathcal{X}}$:

$$\Delta_{\mathcal{X}} = \{p = (p_1, ..., p_{|\mathcal{X}|}) \in \mathbb{R}^{|\mathcal{X}|} : \sum_{i=1}^{|\mathcal{X}|} p_i = 1\} \tag{2.1}$$

By considering two distributions p and q, both of which belong to the simplex $\Delta_{\mathcal{X}}$. , the loss function $\mathcal{L}(p, q)$ serves as a metric to quantify the dissimilarity between the true distribution p and the estimated distribution q. It is important to emphasize that the selection of an appropriate loss function is highly dependent on the specific application domain being considered. For instance, in the context of compression and investment applications, the Kullback-Leibler (KL) divergence has emerged as a widely adopted loss measure. Conversely, classification tasks often leverage the l_1 loss as the pertinent measure of dissimilarity. Moreover, alternative loss functions such as l_2, Hellinger distance, and χ_2 distance find utility in diverse applications, catering to the unique characteristics and requirements of each respective field.

To further investigate the problem, we introduce the concept of a distribution estimator, which is a mapping denoted as $\hat{q} : [\mathcal{X}]^* \to \Delta_{\mathcal{X}}$, where $[\mathcal{X}]^*$ represents the set of finite sequences over the discrete set $[\mathcal{X}]$. The distribution estimator, $\hat{q}$, associates each observed sample, x^n, from the set $[\mathcal{X}]^*$ with a distribution $\hat{q}(x^n) = (\hat{q}_1(x^n), ..., \hat{q}_k(x^n))$ over the discrete set $\mathcal{X}$. We will denote by $\mathcal{E}_{\mathcal{X}}$ as the set of all valid pmf estimators. The average loss of the estimator $\hat{q}$, after observing n samples $X^n = X_1, ..., X_n$, which are generated independently and identically according to an unknown distribution $p \in \Delta_k$, serves as a good metric for the performance of the estimator:

$$\mathbb{E}_{X^n \sim p_X} \left[\mathcal{L}(p, \hat{q}(X^n))\right] \tag{2.2}$$

Of particular interest is the analysis of the worst-case scenario, where the performance of the estimator $\hat{q}$ is evaluated in relation to the most unfavorable distribution:

$$\max_{p \in \Delta_{\mathcal{X}}} \mathbb{E}_{X^n \sim p_X} \left[\mathcal{L}(p, \hat{q}(X^n))\right] \tag{2.3}$$

This entails measuring the loss of the estimator $\hat{q}$ under this worst-case distribution, highlighting the importance of robustness and resilience in the face of challenging situations. In this regard, our primary focus is directed towards determining the least worst-case loss achieved by any estimator, often referred to as the minimax risk:

$$r_n^{\mathcal{L}} \triangleq \min_{\hat{q} \in \mathcal{E}_{\mathcal{X}}} \max_{p \in \Delta_{\mathcal{X}}} \mathbb{E}_{X^n \sim p_X} \left[\mathcal{L}(p, \hat{q}(X^n))\right] \tag{2.4}$$

This quantity captures the optimal performance attainable under the most adverse circumstances, providing valuable insights into the effectiveness and reliability of the estimator.

2.1.1 l_2^2 Minimax PMF Estimation

Let us establish a concrete foundation for our discussion by focusing on the case of l_2^2 loss functions. By limiting our scope to l_2^2 loss functions, we can use it as a simplified example to illustrate the mathematical techniques that we will frequently encounter. It is worth mentioning that all the exercises presented below have long been included in textbooks [9, p. 349], with the earliest known reference found in [11].

Now, let's consider what could be a suitable estimator under the l_2^2 case. The initial inclination might be to consider the maximum likelihood estimator, as it is known to be asymptotically consistent and asymptotically efficient, thus leading to an estimator with asymptotically minimal square error. However, it turns out that the maximum likelihood estimator is not the optimal choice in-terms of worst case risk when dealing with a finite number of samples. Let us demonstrate this fact. It can be easily verified that for the problem of estimating the probability mass function (pmf), the maximum likelihood estimator coincides with the empirical distribution of the observed samples $X_1, ..., X_n$:

$$\hat{p}(x; \{X_i\}_{i=1}^n) \triangleq \frac{T_x}{n}$$

where we define $T_x \triangleq \sum_{i=1}^n \mathbb{1}\{X_i = x\}$. To ensure clarity in our presentation, let's set $k \triangleq |\mathcal{X}|$. Now, we will proceed to compute the worst case risk of the maximum likelihood estimator in terms of the l_2^2 loss. Specifically, we obtain the following expression:

$$\begin{aligned} \mathcal{R} &\triangleq \max_{p_X \in \Delta_\mathcal{X}} \mathbb{E}\left[\left\|p_X(x) - \frac{T_x}{n}\right\|_2^2\right] && (2.5)\\ &= \max_{p_X \in \Delta_\mathcal{X}} \sum_{x \in \mathcal{X}} \mathbb{E}\left[\left(p_X(x) - \frac{T_x}{n}\right)^2\right] && (2.6)\\ &= \max_{p_X \in \Delta_\mathcal{X}} \sum_{x \in \mathcal{X}} \operatorname{var}\left(\frac{T_x}{n}\right) && (2.7)\\ &= \max_{p_X \in \Delta_\mathcal{X}} \sum_{x \in \mathcal{X}} \frac{1}{n^2} \operatorname{var}\left(\sum_{i=1}^n X_i\right) && (2.8)\\ &= \frac{1}{n} \max_{p_X \in \Delta_\mathcal{X}} \sum_{x \in \mathcal{X}} p_X(x)(1 - p_X(x)) && (2.9) \end{aligned}$$

Upon observation, we note that the function $p_X(x)(1-p_X(x))$ is concave with respect to $p_X(x)$. Furthermore, the objective function exhibits symmetry in the variables $p_X(x) x \in \mathcal{X}$. Consequently, the maximum value is attained when all the variables

$p_X(x), x \in \mathcal{X}$ are selected to be equal, indicating that the adversarial distribution corresponds to the uniform distribution. With this understanding, we can proceed to compute the worst case risk as follows:

$$R = \frac{1}{n^2}\left(\sum_{x\in\mathcal{X}} \frac{1}{k}\left(1-\frac{1}{k}\right)\right) = \frac{1}{n}\left(1-\frac{1}{k}\right) \tag{2.10}$$

Indeed, it is natural to inquire whether there exists an estimator that outperforms the maximum likelihood estimator in terms of the minimax criterion. The answer is affirmative. In particular, let's examine the following estimator:

$$\hat{q}(x; \{X_i\}_{i=1}^n) \triangleq \frac{\sum_{i=1}^n \mathbb{1}\{X_i = x\} + \frac{\sqrt{n}}{k}}{n+\sqrt{n}} \tag{2.11}$$

Let us proceed to compute the worst case risk associated with this estimator, while considering a fixed value $x \in \mathcal{X}$:

$$\begin{aligned}
\mathbb{E}\left[\left(p_X(x) - \frac{T_x + \frac{\sqrt{n}}{k}}{n+\sqrt{n}}\right)^2\right] =& \frac{1}{(n+\sqrt{n})^2}\,\mathbb{E}\left[\left(\underbrace{p_X(x)(n+\sqrt{n}) - \left(T_x + \frac{\sqrt{n}}{k}\right)}_{\triangleq U}\right)^2\right] \\
=& \frac{1}{(n+\sqrt{n})^2}\left(\operatorname{var} U + \mathbb{E}\left[U\right]^2\right) \\
=& \frac{1}{(n+\sqrt{n})^2}\left(\operatorname{var}(T_x) + \left(\sqrt{n}p_X(x) - \frac{\sqrt{n}}{k}\right)^2\right) \\
=& \frac{1}{(n+\sqrt{n})^2}\left(np_X(x)(1-p_X(x)) + n\left(p_X(x) - \frac{1}{k}\right)^2\right) \\
=& \frac{1}{(1+\sqrt{n})^2}\left(p_X(x) - \frac{2}{k}p_X(x) + \frac{1}{k^2}\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\|p - \hat{q}(\{X_i\}_{i=1}^n)\|_2^2\right] =& \sum_{x\in\mathcal{X}} \mathbb{E}\left[p_X(x) - \hat{q}(x; \{X_i\}_{i=1}^n)^2\right] \\
=& \sum_{x\in\mathcal{X}} \frac{1}{(1+\sqrt{n})^2}\left(p_X(x) - \frac{2}{k}p_X(x) + \frac{1}{k^2}\right)
\end{aligned} \tag{2.12}$$

$$=\frac{1}{(1+\sqrt{n})^2}\left(1-\frac{1}{k}\right) \tag{2.13}$$

Upon observation, we notice that (2.13) remains constant for all $p_X \in \Delta_\mathcal{X}$, indicating that it is also the worst case risk for $\hat{q}$. By comparing (2.10) with (2.13), we can deduce that $\hat{q}$ exhibits slightly better performance than the maximum likelihood estimator, in the worst case scenario.

The natural question that arises is whether there exists another estimator that can surpass this result. Surprisingly, the answer is no. The intuition behind proving this stems from the idea that the maximum of a set can be lower bounded by the average over the set. In our specific problem, we will establish a lower bound on the maximum risk of any estimator using an average over the simplex, which is described by the averaging measure Π:

$$\max_{p_X \in \Delta_\mathcal{X}} \mathbb{E}_{X^n \sim p_x}\left[\|p_X - \hat{q}(X^n)\|_2^2\right] \geq \mathbb{E}_{p \sim \Pi}\left[\mathbb{E}_{X^n \sim p_X}\left[\|p_X - \hat{q}(X^n)\|_2^2\right]\right] \tag{2.14}$$

which implies:

$$\min_{\hat{q} \in \mathcal{E}_\mathcal{X}} \max_{p_X \in \Delta_\mathcal{X}} \mathbb{E}_{X^n \sim p_x}\left[\|p_X - \hat{q}(X^n)\|_2^2\right] \geq \min_{\hat{q} \in \mathcal{E}_\mathcal{X}} \mathbb{E}_{p \sim \Pi}\left[\mathbb{E}_{X^n \sim p_X}\left[\|p_X - \hat{q}(X^n)\|_2^2\right]\right] \tag{2.15}$$

When the lower bound matches the worst case risk of an estimator $\hat{q}$, we can consider the task essentially accomplished.

By employing the technique of lower bounding the maximum over the probability simplex with an average and seeking the optimal estimator as indicated in (2.15), we delve into the realm of Bayesian inference. This introduces additional terminology that is worth mentioning. Specifically, Π is now referred to as the *prior*, the right-hand side of (2.15) is known as the *Bayes risk of* Π, and finally, the estimator that achieves the Bayes risk is called the *Bayes estimator*. Continuing on (2.15):

$$\min_{\hat{q} \in \mathcal{E}_\mathcal{X}} \max_{p_X \in \Delta_\mathcal{X}} \mathbb{E}_{X^n \sim p_x}\left[\|p_X - \hat{q}(X^n)\|_2^2\right] \geq \min_{\hat{q} \in \mathcal{E}_\mathcal{X}} \mathbb{E}_{p \sim \Pi}\left[\mathbb{E}_{X^n \sim p_X}\left[\|p_X - \hat{q}(X^n)\|_2^2\right]\right]$$

$$\geq \min_{\{\hat{q}_x\}_{x\in\mathcal{X}}} \mathbb{E}_{p\sim\Pi}\left[\mathbb{E}_{X^n\sim p_X}\left[\|p_X - \hat{q}(X^n)\|_2^2\right]\right] \tag{2.16}$$

$$= \min_{\{\hat{q}_x\}_{x\in\mathcal{X}}} \sum_{x\in\mathcal{X}} \mathbb{E}_{p\sim\Pi}\left[\mathbb{E}_{X^n\sim p_X}\left[(p_X(x) - \hat{q}(x;X^n))_2^2\right]\right]$$

$$= \sum_{x\in\mathcal{X}} \min_{\{\hat{q}_x\}_{x\in\mathcal{X}}} \mathbb{E}_{p\sim\Pi}\left[\mathbb{E}_{X^n\sim p_X}\left[(p_X(x) - \hat{q}(x;X^n))_2^2\right]\right] \tag{2.17}$$

The expression (2.16) plays a crucial role in the analysis as it involves the minimization over a broader set of functions. However, it is important to note that in this context, the condition $\sum_{x\in\mathcal{X}} \hat{q}(x; X_{i=1}^n) = 1$ may no longer hold with high probability, and as a result, the resulting estimate may not satisfy the properties of a probability mass function.

Next, we will apply straightforward results from the Bayesian estimation literature. Consider random variables Y and Z following the distribution p_{YZ}. Let us construct an estimate $\hat{Y}(Z)$ based on the observation of Z. In this case, the minimum mean square error estimator should be $\mathbb{E}\left[Y \mid Z\right]$ a.s., which we illustrate in appendix B.3.

Now, by selecting Π as the Dirichlet distribution and by the discussion in appendix B.2, we derive the conditional Bayes estimator as follows:

$$\mathbb{E}\left[p_X(x) \mid X^n\right] = \frac{T_x(X^n) + \beta_n}{n + k\beta_n} \tag{2.18}$$

By selecting the $\beta_n = \frac{\sqrt{n}}{k}$ we obtain the estimator (2.11). As demonstrated in (2.13) the risk function of $\hat{q}_{+\frac{\sqrt{n}}{k}}$ is independent of p_X. Therefore, we infer that the average over $p_X \sim \Pi$ should also be equal to the same risk value. Consequently $\hat{q}_{+\frac{\sqrt{n}}{k}}$ is the minimax estimator. Another intriguing aspect of the estimator (2.18) is that although each $\hat{q}(x; \{X^n\})$ is seperately obtained in the step (2.16) the resulting estimator is self normalizing. Consequently, (2.16) actually is an equality. However, this is far from being a coincidence. Turns out that if the loss function is a Bregman divergence [1] then associated Bayes estimator will always be be the conditional expectation, thus giving the same estimator after decomposing in the step (2.16). However this is not true in general. For instance, if the loss function were l_1, the Bayes estimator for each component would be the median of observed samples observed in that component. This would not guarantee the normalization of the estimated components, leading to a loose inequality in the step (2.16).

2.1.2 Add-β Estimators

The estimators of the form shown in (2.19) are not exclusive to the analysis of the l_2^2 loss. In fact, one can argue that the entire literature on probability mass function (pmf) estimation is dedicated to finding the appropriate sequence β_n to achieve nearly minimax estimators. As a result, these estimators have earned a special designation and are commonly known as add-β_n estimators. When the value of β_n remains constant as the number of samples n increases, it is referred to as an add-constant rule. We will denote an add-β_n estimator as $\hat{p}{+}\beta_n$. Given a specific observation $\{X_i\}_{i=1}^n$, the add-β_n estimator produces the following output:

$$\hat{q}_{+\beta_n}(x;\{X_i\}_{i=1}^n) = \frac{\sum_{i=1}^n \mathbb{1}\{X_i = x\} + \beta}{n + \beta\,|\mathcal{X}|} \tag{2.19}$$

The operation of add-β estimator can be given an alternative interpretation, if the following decomposition of (2.19) is considered:

$$\begin{aligned}\hat{q}_{+\beta_n}(x;\{X_i\}_{i=1}^n) =& \frac{n}{n+\beta k}\frac{\sum_{i=1}^n \mathbb{1}\{X_i = x\}}{n} + \frac{\beta\,|\mathcal{X}|}{n+\beta\,|\mathcal{X}|}\frac{1}{k} \\ =& \left(1 - \frac{\beta k}{n+\beta k}\right)\underbrace{\frac{\sum_{i=1}^n \mathbb{1}\{X_i = x\}}{n}}_{I} + \frac{\beta\,|\mathcal{X}|}{n+\beta\,|\mathcal{X}|}\underbrace{\frac{1}{k}}_{II}\end{aligned}$$

where we recognize I as the MLE and II as the uniform distribution over the set $\mathcal{X}$. Furthermore, we observe their coefficients sum to one, and we see that as n grows, the coefficient of MLE becomes more dominant in the mixture. Therefore operationally and add-β_n rule performs MLE but *steps back* to the uniform distribution, which would have been the correct output if there were no samples observed, accordingly with the number of samples.

Perhaps the lag in the development of new results on pmf estimation can be explained by the non-existence of nice features of the l_2^2 loss highlighted in this section. In particular, the first result in the literature for a non l_2^2 loss is due to [2], which dates almost five decades later the work of Trybula [11]. In particular, [2] showed that a varying add-β rule achieves the minimax risk with the correct first-order constant for the KL divergence and established that for fixed $k \triangleq |\mathcal{X}|$:

$$r_{k,n}^{\text{KL}} = \frac{k-1}{2n} + o\left(\frac{1}{n}\right)$$

Loss Function	Regime	Minimax Estimator	Minimax Risk
l_2^2	-	$\hat{q}_{+\frac{\sqrt{n}}{k}}$	$\frac{1-\frac{1}{k}}{(\sqrt{n}+1)^2}$
χ_2	$k = \Theta(1)$	$\hat{q}_{+1}$	$\frac{k-1}{n} + O\left(\frac{\log n}{n^2}\right)$
l_1	$k = \Theta(1)$	$\hat{q}_{+0}$	$\sqrt{\frac{2(k-1)}{\pi n}} + O\left(\frac{1}{n^{\frac{3}{4}}}\right)$
f-divergences	$k = \Theta(1)$	$\hat{q}_{+\beta},\ \beta > 0$	$f''(1)\frac{k-1}{2n} + o\left(\frac{1}{n}\right)$

Table 2.1: First Order Constants of Minimax Risk determined in [7]

A decade later, [7] showed that add-constant estimators are optimal in the corresponding regimes for the losses l_1, χ_2, and 'smooth' f-divergences achieves the minimax rate with the correct first-order constant. We summarize the contributions of [7] in table 2.1 for the regimes and the loss functions where the first-order constant is determined: For general f-divergences, it is assumed that f is thrice differentiable and follows subexponential tails. Furthermore, it is assumed that p_X should be away from the boundaries of the probability simplex. Hence the risk depicted on the table 2.1 is for the problem:

$$r_n^{f,\delta} \triangleq \min_{\hat{q} \in \mathcal{E}_{\mathcal{X}}} \max_{p \in \Delta_{\mathcal{X}}} \mathbb{E}_{X^n \sim p_X}\left[\mathcal{L}(p, \hat{q}(X^n))\right]$$

where:

$$\Delta_\delta = \{p_x \in \Delta_{\mathcal{X}} : \forall x \in \mathcal{X}, \quad p_X(x) \geq \delta\}$$

2.1.3 l_p^p Minimax PMF Estimation

This work is mainly concerned with the l_p^p loss functions. The simple results given in this chapter seem missing in the literature. In particular for two distributions p, q over the alphabet $\mathcal{X}$, their l_p^p distance the l_p^p norm of their difference:

$$l_p^p(p,q) \triangleq \|p-q\|_p^p = \sum_{i=1}^{\mathcal{X}} |p_i - q_i|^p$$

and we denote the one variable minimax risk as:

$$r_n^p \triangleq r_n^{l_p^p} = \min_{\hat{q}} \max_{p \in \Delta_{\mathcal{X}}} \mathbb{E}_{X^n \sim p_X} \left[\|p - \hat{q}(X^n))\|_p^p \right]$$

The rate of decay of the r_n^p will be important for the proofs:

Lemma 2.1.1. *For $|\mathcal{X}| = O_n(1)$, $r_n^p = \Theta(n^{-\frac{p}{2}})$*

The proof for Lemma 2.1.1 is given in appendix B.4. We will be ultimately interested in the constants of r_n^p for the rate $n^{-\frac{p}{2}}$. We will denote the constants of this rate by $\bar{C}_p \triangleq \limsup_n n^{\frac{p}{2}} r_n^p$ and $\underline{C}_p = \liminf_n n^{\frac{p}{2}} r_n^p$. For convinience, we will group the constanst $\bar{C}_p, \underline{C}_p$ into C_p and use the operator $\simeq$ to denote that $g_n \simeq C_p f_n$ if $\limsup_n g_n/f_n \le \bar{C}_p$ and $\liminf_n g_n/f_n \ge \underline{C}_p$. Similarly, we say $g_n \lesssim C_p f_n$ if $\liminf_n g_n/f_n \le \underline{C}_p$ and $\limsup_n g_n/f_n \le \bar{C}_p$.

2.2 Semi-supervised PMF Estimation

We formalize the semi-supervised PMF estimation for the case when there are two random variables X, Y jointly distributed with $p_{XY} \in \Delta_{\mathcal{X}\times\mathcal{Y}}$. We observe two datasets: m i.i.d. samples of $\{(X_i, Y_i)\}_{i=1}^m$ pairs drawn from p_{XY} and n samples of only $\{X_j'\}_{j=1}^n$ drawn from the marginal distribution p_X. Now an estimator is a mapping $\check{q}_{XY} : [\mathcal{X}]^* \times [\mathcal{X} \times \mathcal{Y}]^* \to \Delta_{\mathcal{X}\times\mathcal{Y}}$.We can alternatively describe the set of estimators under consideration as the set $\mathcal{E}_{\mathcal{X}\times\mathcal{Y}} = \mathcal{E}_{\mathcal{X}} \times \mathcal{E}_{\mathcal{Y}}^{|\mathcal{X}|}$. Since there is no unpaired Y observations in this scenario, it is natural to decompose the estimator into marginal $\hat{q}_X$ and conditional part, $\hat{q}_{T|x}$ where labeled and unlabeled samples enhance the estimation of p_X where only labeled samples provide information about $p_{Y|X}$.

The worst case risk for the estimator $\hat{q}_{XY}$ is:

$$\max_{p_{XY}\in\Delta_{\mathcal{X}\times\mathcal{Y}}} \mathbb{E}_{\substack{\{X'_j\}_1^n\sim p_X \\ \{(X_i,Y_i)\}_{i=1}^n\sim p_{XY}}} \left[\mathcal{L}(p_{XY}, \hat{q}_{XY}(\{X'_j\}_{j=1}^n, \{(X_i,Y_i)\}_{i=1}^m))\right] \tag{2.20}$$

and the minimax risk is:

$$R^{\mathcal{L}}_{m,n} \triangleq \min_{\hat{q}_{XY}\in\mathcal{E}_{\mathcal{X}\times\mathcal{Y}}} \max_{p_{XY}\in\Delta_{\mathcal{X}\times\mathcal{Y}}} \mathbb{E}_{\substack{\{X'_j\}_1^n\sim p_X \\ \{(X_i,Y_i)\}_{i=1}^n\sim p_{XY}}} \left[\mathcal{L}(p_{XY}, \hat{q}_{XY}(\{X'_j\}_{j=1}^n, \{(X_i,Y_i)\}_{i=1}^m))\right] \tag{2.21}$$

We specialize it to the l_p^p loss:

$$R^{p}_{m,n} \triangleq \min_{\hat{q}_{XY}\in\mathcal{E}_{\mathcal{X}\times\mathcal{Y}}} \max_{p_{XY}\in\Delta_{\mathcal{X}\times\mathcal{Y}}} \mathbb{E}_{\substack{\{X'_j\}_1^n\sim p_X \\ \{(X_i,Y_i)\}_{i=1}^n\sim p_{XY}}} \left[\left\|p_{XY} - \hat{q}_{XY}(\{X'_j\}_{j=1}^n, \{(X_i,Y_i)\}_{i=1}^m)\right\|_p^p\right] \tag{2.22}$$

Attacking the problem $R^p_{m,n}$ seems formidable. Instead, we will study the following problems with increasing similarity to (2.21).

Auxilary Problem 1.

$$R^p_m \triangleq \min_{\hat{q}_{Y|X}\in\mathcal{E}_{\mathcal{Y}}^{|\mathcal{X}|}} \max_{p_{XY}\in\Delta_{\mathcal{X}\times\mathcal{Y}}} \mathbb{E}_{\{(X_i,Y_i)\}_{i=1}^m\sim p_{XY}} \left[\left\|p_{XY} - p_X\hat{q}_{Y|X}(\{(X_i,Y_i)\}_{i=1}^m)\right\|_p^p\right] \tag{2.23}$$

Auxilary Problem 2.

$$\bar{R}^p_m \triangleq \min_{\hat{q}_{Y|X}\in\mathcal{E}_{\mathcal{Y}}^{|\mathcal{X}|}} \max_{p_X\in\Delta_{\mathcal{X}}} \mathbb{E}_{\{X_i\}_{i=1}^m\sim p_X} \left[\max_{p_{Y|X}\in\Delta_{\mathcal{Y}}^{|\mathcal{X}|}} \mathbb{E}_{\{Y_i\}_{i=1}^m\sim p_{Y|X}} \left[\left\|p_{XY} - p_X\hat{q}_{Y|X}(\{(X_i,Y_i)\}_{i=1}^m)\right\|_p^p\right]\right] \tag{2.24}$$

The problem R^p_m corresponds to the limit of the problem $R^p_{m,n}$ as $n\to\infty$. Intuitively, in this case, there are sufficiently many incomplete samples to make the perfect estimation of p_X possible. The difference between the two problems is that for $\bar{R}^p_m$, nature has an additional advantage in forming the dataset $\{X_i\}_{i=1}^m$: it can first observe the realization of the $(X_1, ..., X_m)$ symbols then choose $p_{Y|X}$ from which to generate the $(Y_1, ..., Y_m)$ values accompanying the $(X_1, ..., X_m)$ to finish the construction of the joint samples. Mathematically this follows from the trivial comparison which holds for all $x_i \in \mathcal{X}^n, p_{XY} \in \Delta_{\mathcal{X}}$:

$$\begin{aligned} &\mathbb{E}_{\{Y_i\}_{i=1}^m \sim p_{Y|X}} \left[\left\| p_{XY} - p_X \hat{q}_{Y|X}(\{(x_i, Y_i)\}_{i=1}^m) \right\|_p^p \right] \\ \le &\max_{p_{Y|X} \in \Delta_{\mathcal{Y}}^{|\mathcal{X}|}} \mathbb{E}_{\{Y_i\}_{i=1}^m \sim p_{Y|X}} \left[\left\| p_{XY} - p_X \hat{q}_{Y|X}(\{(x_i, Y_i)\}_{i=1}^m) \right\|_p^p \right] \end{aligned}$$

and observing that all the following operations applying to both sides are monotonic and hence preserving the relationship:

$$R_m^p \triangleq \min_{\check{q}_{Y|X}} \max_{p_{XY}} \mathbb{E}\left[\| p_{XY} - \check{q}_{XY} \|_p^p \right] \le \min_{\check{q}_{Y|X}} \max_{p_X} \mathbb{E}\left[\max_{p_{Y|X}} \mathbb{E}\left[\| p_{XY} - p_X \check{q}_{Y|X} \|_p^p \right] \right] = \bar{R}_m \tag{2.25}$$

2.2.1 Composition Estimators

Our discussion focuses on composition estimators, which we aim to demonstrate their optimality. To aid in understanding, we can consider a given univariate estimator $\hat{q}$ as a set of estimators $\hat{q}i^{\infty}i = 0$, where each $\hat{q}i$ maps from $\mathcal{X}^i$ to $\Delta_{\mathcal{X}}$. This approach helps avoid any confusion that may arise due to changes in the adversarial distribution as the number of samples varies. We will utilize this notation for now to define the composition estimators.

Definition 2.2.1. *For a given univariate estimator $\hat{q}$, the conditional compostion estimator of $\hat{q}$ is:*

$$\check{q}_{Y|X}^{**,(m,n)}(\{(X_i, Y_i)\}_{i=1}^m) \triangleq \prod_{x \in \mathcal{X}} \check{q}^*_{m\hat{p}_X(x;\{(X_i,Y_i)\}_{i=1}^m)}(\{Y_j : 1 \le j \le m : X_j = x\})$$

Definition 2.2.2. *For a given univariate estimator $\hat{q}$, the joint compostion estimator of $\hat{q}$ is:*

$$\begin{aligned} \check{q}_{XY}^{**,(m,n)}(\{X_i\}_{i=1}^n, \{(X_i', Y_i')\}_{i=1}^m) \triangleq\ &\hat{p}_{n+m}(\{X_i\}_{i=1}^n \cup \{X_i'\}_{i=1}^m) \\ &\prod_{x \in \mathcal{X}} \check{q}^*_{m\hat{p}_X(x;\{(X_i',Y_i')\}_{i=1}^m)}(\{Y_j' : 1 \le j \le m : X_j' = x\}) \end{aligned}$$

2.2.2 First Order Optimality

Another fundamental concept in this study is the notion of first-order minimax optimality pertaining to the presented problems thus far.

Definition 2.2.3. *An estimator $\hat{q}$ is first order minimax optimal for the problem r_n^p if*

$$\max_{p_X} \mathbb{E}_{X^n \sim p_X} \left[\| p_X - \hat{q}_X \| \right] + o(r_n^p)$$

.

Definition 2.2.4. *An estimator $\hat{q}_{Y|X} \in \mathcal{E}_{\mathcal{Y}}^{|\mathcal{X}|}$ is first order minimax optimal for the problem R_n^p if*

$$\max_{p_{XY} \in \Delta_{\mathcal{X} \times \mathcal{Y}}} \mathbb{E}_{\{(X_i,Y_i)\}_{i=1}^m \sim p_{XY}} \left[\left\| p_{XY} - p_X \hat{q}_{Y|X}(\{(X_i, Y_i)\}_{i=1}^m) \right\|_p^p \right] + o(R_n^p)$$

Chapter 3

Main Results

Theorem 1. *Let $\hat{q}_n^*$ be a minimax optimal estimator for r_n^p. Then the conditional composition $\hat{q}_{Y|X}^{*,m}$ based on $\hat{q}_n^*$ is minimax optimal for $\bar{R}_m^p$:*

$$\max_{p_X \in \Delta_{\mathcal{X}}} \mathbb{E}_{\{X_i\}_{i=1}^m \sim p_X} \left[\max_{p_{Y|X} \in \Delta_{\mathcal{Y}}^{|\mathcal{X}|}} \mathbb{E}_{\{Y_i\}_{i=1}^m \sim p_{Y|X}} \left[\left\| p_{XY} - p_X \hat{q}_{Y|X}(\{(X_i, Y_i)\}_{i=1}^m) \right\|_p^p \right] \right] = \bar{R}_m^p \tag{3.1}$$

Theorem 2. *Let $p \geq 2$ and $\hat{q}_n^*$ be a first order minimax optimal estimator for r_n^p. Then the conditional composition $\hat{q}_{Y|X}^{*,m}$ based on $\hat{q}_n^*$ is first order minimax optimal for R_m^p:*

$$\max_{p_{XY}} \mathbb{E}_{\{(X_i,Y_i)\}_{i=1}^m} \left[\| p_{XY} - p_X \hat{q}_{Y|X}^{*,m}(\{(X_i, Y_i)\}_{i=1}^m) \|_p^p \right] = R_m^p + o\left(R_m^p\right) \tag{3.2}$$

Theorem 3. *Let $m = o(n)$:*

$$\left| R_{m,n}^p - R_m^p \right| \leq O\left(m^{-\frac{p-1}{2}} (n)^{-1/2} \right) \tag{3.3}$$

Theorem 4. *Let $p \geq 2$ and $\hat{q}_n^*$ be a first-order optimal estimator for r_n^p. Then the joint composition $\hat{q}_{XY}^{*,m,n}$ based on $\hat{q}_n^*$ is first order minimax optimal for $R_{m,n}^p$ in the regime $m = o(n)$.*

3.1 Sketch of the Proof

The primary conclusion of this paper, as delineated in Theorem 4, is established through a logical progression of Theorems 1-3. We commence by demonstrating in Theorem 1 that the conditional composition of a minimax optimal estimator for r_n^p acts as a minimax optimal estimator for $\bar{R}_m^p$. Subsequently, in Theorem 2, we draw a connection between the problems $\bar{R}_m^p$ and R_m^p. In detail, we ascertain that for $p \geq 2$, the adversarial distribution of $\bar{R}_m^p$ is δ_x, causing the problem $\bar{R}_m^p$ to simplify into R_m^p. Finally, we argue in Theorem 3 that $R_{m,n}^p = R_m^p + o(m^{-\frac{p}{2}})$, accomplished by scrutinizing the regime where $m = o(n)$

Chapter 4

Proofs for the Main Results

4.1 Proof for Theorem 1

Theorem 1. *Let $\hat{q}_n^*$ be a minimax optimal estimator for r_n^p. Then the conditional composition $\hat{q}_{Y|X}^{*,m}$ based on $\hat{q}_n^*$ is minimax optimal for $\bar{R}_m^p$:*

$$\max_{p_X \in \Delta_{\mathcal{X}}} \mathbb{E}_{\{X_i\}_{i=1}^m \sim p_X} \left[\max_{p_{Y|X} \in \Delta_{\mathcal{Y}}^{|\mathcal{X}|}} \mathbb{E}_{\{Y_i\}_{i=1}^m \sim p_{Y|X}} \left[\left\| p_{XY} - p_X \hat{q}_{Y|X}(\{(X_i, Y_i)\}_{i=1}^m) \right\|_p^p \right] \right] = \bar{R}_m^p \tag{3.1}$$

Proof. Now we define:

$$f(p_X, \hat{q}_{Y|X}) \triangleq \mathbb{E}_{X_1^m \sim p_X} \left[\max_{p_{Y|X}} \mathbb{E}_{Y_1^m \sim P_{Y|X}} \left[\|p_{XY} - p_X \hat{q}_{Y|X}\|_p^p \right] \right]$$

Let $\hat{q}_{Y|X} : (\mathcal{X} \times \mathcal{Y})^m \to (\Delta_{\mathcal{Y}})^{|\mathcal{X}|}$ be an arbitrary estimator for the conditional distribution $p_{Y|X}$. A sufficient condition for $\hat{q}_{Y|X}$ to achieve $\bar{R}_m^p$ is that for all p_X:

$$\min_{\hat{q}_{Y|X}} f(p_X, \hat{q}_{Y|X}) = f(p_X, \hat{q}_{Y|x}) \tag{4.1}$$

This follows from that:

$$\max_{p_X} f(p_X, \hat{q}_{Y|X}) \geq \bar{R}_m^p \triangleq \min_{\hat{q}} \max_{p_X} \; f(p_X, \hat{q}) \geq \max_{p_X} \min_{\hat{q}} f(p_X, \hat{q}) = \max_{p_X} f(p_X, \hat{q}_{Y|X})$$

where the first inequality is due to the substitution of the estimator $\hat{q}_{Y|X}$, and the second inequality is the change of $\min\max$ with $\max\min$. Next, we see the condition (4.1) implies that all terms are actually equal which establishes that that $\hat{q}_{Y|X}$ is minimax optimal for $\bar{R}^p_m$. Now let us show (4.1) holds for the conditional compositon estimator $\hat{q}^{**}_{Y|X}$ fix $p_X \in \Delta_{\mathcal{X}}$:

$$\begin{aligned}
\min_{\hat{q}_{Y|X}} f(p_X, \hat{q}_{Y|X}) &= \min_{\hat{q}_{Y|X}} \mathbb{E}_{X_1^m \sim p_X} \left[\max_{P_{Y|X}} \mathbb{E}_{Y_1^m \sim P_{Y|X}} \left[\|p_{XY} - p_X \hat{q}_{Y|X}\|_p^p \right] \right] \\
&= \min_{\hat{q}_{Y|X}} \sum_{x \in \mathcal{X}} (p_X(x))^p \, \mathbb{E}_{X_1^m \sim p_X} \left[\max_{p_{Y|X=x}} \mathbb{E}_{Y_1^m \sim p_{Y|X}} \left[\|p_{Y|X=x} - \hat{q}_{Y|X=x}\|_p^p \right] \right] \\
&= \sum_{x \in \mathcal{X}} (p_X(x))^p \min_{\hat{q}_{Y|X=x}} \mathbb{E}_{X_1^m \sim p_X} \left[\max_{p_{Y|X=x}} \mathbb{E}_{Y_1^m \sim p_{Y|X}} \left[\|p_{Y|X=x} - \hat{q}_{Y|X=x}\|_p^p \right] \right]
\end{aligned} \tag{4.2}$$

$$\begin{aligned}
&= \sum_{x \in \mathcal{X}} (p_X(x))^p \min_{\hat{q}_{Y|X=x}} \sum_{i=0}^{m} \binom{m}{i} (p_X(x))^i (1 - p_X(x))^{m-i} \\
&\qquad \mathbb{E}_{X_1^m \sim p_X} \left[\max_{p_{Y|X=x}} \mathbb{E}_{Y_1^m \sim p_{Y|X}} \left[\|p_{Y|X=x} - \hat{q}_{Y|X=x}\|_p^p \mid T_x(X_1^m) = i \right] \right] \\
&= \sum_{x \in \mathcal{X}} (p_X(x))^p \min_{\hat{q}_{Y|X=x}} \sum_{i=0}^{m} \binom{m}{i} (p_X(x))^i (1 - p_X(x))^{m-i} \\
&\qquad \mathbb{E}_{X_1^m \sim p_X} \left[\max_{p_{Y|X=x}} \mathbb{E}_{Y_1^m \sim p_{Y|X}} \left[\|p_{Y|X=x} - \hat{q}_{Y|X=x}\|_p^p \right] \mid T_x(X_1^m) = i \right] \\
&= \sum_{x \in \mathcal{X}} (p_X(x))^p \sum_{i=0}^{m} \binom{m}{i} (p_X(x))^i (1 - p_X(x))^{m-i} \\
&\qquad \underbrace{\min_{\hat{q}_{Y|X=x,i}} \mathbb{E}_{X_1^m \sim p_X} \left[\max_{p_{Y|X=x}} \mathbb{E}_{Y_1^m \sim p_{Y|X}} \left[\|p_{Y|X=x} - \hat{q}_{Y|X=x}\|_p^p \right] \mid T_x(X_1^m) = i \right]}_{\triangleq I_i}
\end{aligned} \tag{4.3}$$

$$= \sum_{x \in \mathcal{X}} (p_X(x))^p \sum_{i=0}^{m} \binom{m}{i} (p_X(x))^i (1 - p_X(x))^{m-i} r_i^p \tag{4.4}$$

$$= \sum_{x \in \mathcal{X}} \sum_{i=0}^{m} \binom{m}{i} (p_X(x))^{i+p} (1 - p_X(x))^{m-i} r_i^p \tag{4.5}$$

In (4.2), it is noted that the optimization variables are independent. In (4.3), each estimator is considered as a collection $\hat{q}_{Y|X=x} = \hat{q}_{Y|X=x,i}$, as previously described.

Under the conditioning $T_x(X_1^n) = i$, only the optimization variable $\hat{q}_{Y|X=x,i}$ is involved in the minimization problem. In (4.4), it is observed that problem I_i is the same problem as r_i^p. Hence by assumption $\hat{q}^{*,i}_{Y|X}$ and r_i^p achieves it. $\square$

4.2 Proof for Theorem 2

Theorem 2. *Let $p \geq 2$ and $\hat{q}_n^*$ be a first order minimax optimal estimator for r_n^p. Then the conditional composition $\hat{q}^{*,m}_{Y|X}$ based on $\hat{q}_n^*$ is first order minimax optimal for R_m^p:*

$$\max_{p_{XY}} \mathbb{E}_{\{(X_i,Y_i)\}_{i=1}^m} \left[\|p_{XY} - p_X \hat{q}^{*,m}_{Y|X}(\{(X_i,Y_i)\}_{i=1}^m)\|_p^p \right] = R_m^p + o\left(R_m^p\right) \tag{3.2}$$

Proof. We start with (4.5):

$$\begin{aligned} \bar{R}_m^p =& \max_{p_X} \sum_{x\in\mathcal{X}} \sum_{i=0}^{m} \binom{m}{i} \left(p_X(x)\right)^{i+p} \left(1 - p_X(x)\right)^{m-i} r_i^p && (4.6) \\ \simeq& \max_{p_X} \sum_{x\in\mathcal{X}} C_p \left(\frac{p_X(x)}{m} \right)^{\frac{p}{2}} + o(m^{-\frac{p}{2}}) && (4.7) \\ =& \frac{C_p}{m^{\frac{p}{2}}} + o(m^{-\frac{p}{2}}) && (4.8) \end{aligned}$$

The key step (4.7) follows from Lemma 4.2.1. Ignoring the lower order terms in (4.7), we note that for $p \geq 2$ the objective function is convex and symmetric in variables $\{p_X(x)\}_{x\in\mathcal{X}}$. Therefore the optimizer is a vertex of the probability simplex, which leads to (4.8).

In order to obtain the matching lower bound, we substitute $p_X = \delta_x$ for some arbitrary $x \in \mathcal{X}$. For completeness, we carry out the steps:

$$\begin{aligned} R_m^p &\geq \min_{\hat{q}_{Y|X}} \max_{p_{Y|X}} \mathbb{E}_{\{(X_i,Y_i)\}_{i=1}^m \sim \delta_x p_{Y|X}} \left[\|\delta_x p_{Y|X} - \delta_x \hat{q}_{Y|X}\|_p^p \right] && (4.9) \\ &= \min_{\hat{q}_{Y|X}} \max_{p_{Y|X}} \mathbb{E}_{\{Y_i\}_{i=1}^m \sim p_{Y|X=x}} \left[\|p_{Y|X=x} - \hat{q}_{Y|X=x}\|_p^p \right] && (4.10) \\ &= \min_{\hat{q}_{Y|X=x}} \max_{p_{Y|X=x}} \mathbb{E}_{\{Y_i\}_{i=1}^m \sim p_{Y|X=x}} \left[\|p_{Y|X=x} - \hat{q}_{Y|X=x}\|_p^p \right] = r_m^p && (4.11) \end{aligned}$$

Therefore by $R_m^p \leq \bar{R}_m^p$ (see the discussion in (2.2)) and (4.8) we obtain:

$$\frac{C_p}{m^{\frac{p}{2}}} \lesssim R_m^p \leq \bar{R}_m^p \simeq \frac{C_p}{m^{\frac{p}{2}}} + o\left(\frac{1}{m^{\frac{p}{2}}}\right)$$

□

4.2.1 Supplementary Results for Theorem 2

Before introducing the required lemmas, it is beneficial to provide some contextual information to facilitate comprehension of the topic at hand. Notably, one can assert that a pivotal advancement in the entire study lies in equation (4.7), where it states

$$\sum_{i=1}^{m} \binom{m}{i} (x)^{i+p} (1-x)^{m-i} \frac{1}{\left(\frac{i}{m}\right)^{\frac{p}{2}}} \approx x^{\frac{p}{2}} \tag{4.12}$$

To ensure clarity in the presentation, we have omitted the first term in the summation and cleared the denominator. This approximation method draws inspiration from the Bernstein polynomial basis in approximation theory. Specifically, the set of functions employed in this approach is the $n+1$-th order Bernstein basis polynomials, which can be defined as follows:

$$B_{i,n}(x) = \binom{n}{i} x^i (1-x)^{p-i}, i \in \{0, ..., n\}$$

where a function $f \in R$ is constructed from its Bernstein basis representation as:

$$B_n(f,x) = \sum_{i=0}^{n} f\left(\frac{i}{n}\right) x^i (1-x)^{n-i}$$

The Bernstein polynomial basis exhibits numerous intriguing and valuable properties. Notably, Bernstein polynomials are renowned for their ability to provide smooth approximations. They not only succeed in approximating a given differentiable function but also its higher-order derivatives, if they exist. This characteristic yields practical implications, such as the guarantee of convexity in the approximation over any domain where the approximated function is convex, among others. A brief illustration of this scenario is depicted in fig. A-1. However, this smooth approximation comes at the cost of a slower convergence rate of $\frac{1}{n}$, as elucidated by the following theorem:

Theorem 5 ([4],Theorem-3.1)**.** *If f is bounded on A, differentiable in some neighborhood of x, and has second derivative $f''(x)$ for some $x \in A$, then*

$$\lim_{n\to\infty} n\left[B_n(f,x) - f(x)\right] = \frac{x(1-x)}{2} f''(x).$$

It appears that this result cannot be directly employed to assert that:

$$\sum_{i=1}^{m} \binom{m}{i} (x)^i (1-x)^{m-i} \frac{1}{\frac{i}{m}^{\frac{p}{2}}} \approx \frac{1}{x^{\frac{p}{2}}}$$

In particular, as fig. A-2, we observe that convergence does not occur as anticipated. This observation aligns with the prediction made by Theorem 4.2.1, which states that:

$$\lim_{n\to\infty} n\left[B_n(f,x) - f(x)\right] = \frac{(1-x)}{x^{1+\frac{p}{2}}}$$

Indeed, the approximation error has unbounded growth as x approaches 0. On the other hand, (4.12) is an accurate approximation, as verified in fig. A-3. This can be explained by the fact that multiplying by the term x^p suppresses the approximation error near 0. In the following, we provide proof for this phenomenon, perhaps by examining the situation from a different perspective. Consider the following:

$$\begin{aligned}\sum_{i=0}^{m} \binom{m}{i} (x)^{i+p} (1-x)^{m-i} \frac{1}{\frac{i}{m}^{\frac{p}{2}}} =& \sum_{i=0}^{m} \binom{m}{i} (x)^i (1-x)^{m-i} \frac{(\frac{i}{m})^p}{\left(\frac{i}{m}\right)^{\frac{p}{2}}} &(4.13)\\ =& \sum_{i=0}^{m} \binom{m}{i} (x)^i (1-x)^{m-i} \left(\frac{i}{m}\right)^{\frac{p}{2}} &(4.14)\\ =& x^{\frac{p}{2}} + O\left(\frac{1}{m}\right) &(4.15)\end{aligned}$$

Let us assume that equation (4.13) holds with negligible error for now. Then equation (4.15) follows directly from equation (4.14) using Theorem 4.2.1, since $\frac{x}{(1-x)}\frac{d^2}{dx^2}(x^{\frac{p}{2}}) =$

$x^{\frac{p}{2}-1}$, which is finite for all $x \in [0,1]$ as long as $p \geq 2$. The only remaining task is to demonstrate the substitution x^p. Consider the identity:

$$x^p = \sum_{i=1}^{m} \binom{m}{i} x^i (1-x)^{m-i} \left(\frac{i}{m}\right)^p + O\left(\frac{1}{m}\right)$$

which follows again from (4.2.1). Then (4.13) is nothing but:

$$\sum_{i=1}^{m} \binom{m}{i} (x)^i (1-x)^{m-i} \frac{(\frac{i}{m})^p}{\left(\frac{i}{m}\right)^{\frac{p}{2}}} \approx \sum_{i=1}^{m} \binom{m}{i} (x)^i (1-x)^{m-i} \frac{\sum_{j=1}^{m} \binom{m}{j} (\frac{j}{m})^p x^j (1-x)^{m-j}}{\left(\frac{i}{m}\right)^{\frac{p}{2}}}$$

In other words, it must be the case that $\sum_{j=1}^{m} \binom{m}{j} (\frac{j}{m})^p x^j (1-x)^{m-j} = \left(\frac{i}{m}\right)$ whenever is needed. It turns out that this intuition is correct due to a double-sifting effect. Specifically, for any fixed value of x, only a relatively small number of terms $x^i(1-x)^{m-i}$ dominate the sum. These dominant terms correspond to indices i in the range $(mx - \epsilon_m, mx + \epsilon_m)$. Since this effect also holds for the sum $\sum_{j=1}^{m} \binom{m}{j} \left(\frac{j}{m}\right)^p x^j (1-x)^{m-j}$, it is effectively dominated by j in the range $(mx - \epsilon_m, mx + \epsilon_m)$, and its value is approximately $\left(\frac{j}{m}\right)^p \approx \left(\frac{i}{m}\right)^p$. To capture this sifting effect, we need to alleviate concentration and use sharp bounds on the tails of a binomial random variable. In this regard, we will utilize the following bound:

Theorem 6 ([3], lemma-2)**.** *Let* $X_1, \ldots, X_n$ *be independent* Bernoulli(p_i) *variables. We consider the sum* $X = \sum_{i=1}^{n} X_i$, *with expectation* $\mathrm{E}(X) = \sum_{i=1}^{n} p_i$. *Then, we have:*

$$\Pr(X \leq \mathrm{E}(X) - \lambda) \leq e^{-\lambda^2/2\mathrm{E}(X)}$$
$$\Pr(X \geq \mathrm{E}(X) + \lambda) \leq e^{-\frac{\lambda^2}{2(\mathbb{E}(X)+\lambda/3)}}$$

Now, let us proceed to formalize these intuitive arguments. First, let us introduce:

$$H_p^n(x) \triangleq \sum_{i=0}^{n} \binom{n}{i} r_i^p x^{i+p} (1-x)^{n-i} \tag{4.16}$$

$$G_p^n(x) \triangleq \sum_{i=0}^{n} \binom{n}{i} r_i^p x^i \left(\frac{i}{n}\right)^p (1-x)^{n-i} \tag{4.17}$$

and we proceed by presenting and proving the lemmas:

Lemma 4.2.1.

$$H_p^n(x) \simeq C_p \ \left(\frac{x}{n}\right)^{\frac{p}{2}} + o(n^{-\frac{p}{2}})$$

Proof. We fix the constant c given in Lemma 4.2.2. There are two cases: In the first case $x \geq c\frac{log^2(n)}{n}$:

$$\begin{aligned} H_p^n(x) &\simeq \sum_{i=0}^{n} \binom{n}{i} \frac{C_p}{i^{\frac{p}{2}}+1} x^{i+p}(1-x)^{n-i} && (4.18) \\ &= \frac{C_p}{n^{\frac{p}{2}}} \sum_{i=0}^{n} \binom{n}{i} \left(\frac{i}{n}\right)^{\frac{p}{2}} x^i(1-x)^{n-i} + O\left(\frac{H_p^n(x)}{\sqrt{\log n}}\right) && (4.19) \\ &= \frac{C_p}{n^{\frac{p}{2}}} \left(x^{\frac{p}{2}} + O\left(n^{-1}\right)\right) + O\left(\frac{H_p^n(x)}{\sqrt{\log n}}\right) && (4.20) \\ &= C_p \left(\frac{x}{n}\right)^{\frac{p}{2}} + o\left(n^{-\frac{p}{2}}\right) && (4.21) \end{aligned}$$

(4.18) holds since $r_n^p \simeq C_p n^{-\frac{p}{2}} = C_p n^{-\frac{p}{2}} + o(n^{-\frac{p}{2}})$ whereas in (4.19) we use Lemma 4.2.2. To obtain (4.20), we utilize as follows: we set $f(x) = x^{\frac{p}{2}}$ and bound the error of nth order Bernstein polynomial approximation B_n as:

$$\begin{aligned} |B_n(x;f) - f(x)| &= n^{-1}x(1-x)f''(x)/2 + o(n^{-1}) && (4.22) \\ &= n^{-1}p/4(p/2-1)x^{\frac{p}{2}-2} + o(n^{-1}) && (4.23) \\ &\leq n^{-1}p/4(p/2-1) + o(n^{-1}) && (4.24) \end{aligned}$$

where (4.24) follows since $p \geq 2$. Therefore we conclude that convergence is uniform with error $O(n^{-1})$ for all $x \in (0,1)$. Finally, (4.20) implies that $H_p^n(x) = O(n^{-\frac{p}{2}})$ and in (4.21) we substitute this in the error term of (4.20). For the second case we have $x \leq c\frac{log^2(n)}{n}$:

$$\begin{aligned} H_p^n(x) &= \sum_{i=0}^{n} \binom{n}{i} r_i^p x^{i+p}(1-x)^{n-i} && (4.25) \\ &\leq C_p \ x^p \sum_{i=0}^{n} \binom{n}{i} \frac{1}{i^{\frac{p}{2}}+1} x^i(1-x)^{n-i} && (4.26) \end{aligned}$$

$$\leq C_p\ x^p = O\left(\frac{\log^{2p}(n)}{n^p}\right) \tag{4.27}$$

Similarly $C_p \left(\frac{x}{n}\right)^{\frac{p}{2}} = O\left(\frac{\log^{2p}(n)}{n^p}\right)$ when $x \leq c\frac{\log^2 n}{n}$, therefore $\left|C_p \left(\frac{x}{n}\right)^{\frac{p}{2}} - H_p^n(x)\right| = o\left(n^{-\frac{p}{2}}\right)$. □

Lemma 4.2.2. *There exists a $c > 0$ such that for $x \geq c\frac{log^2 n}{n}$:*

$$\left|H_p^n(x) - G_p^n(x)\right| = O\left(\frac{H_p^n(x)}{\sqrt{\log n}}\right) \tag{4.28}$$

Proof. Fix $c > 0$, let $\delta_1, \delta_2 > 0$, whose values will be determined later, we define $\Delta_p(x,i) \triangleq \left|x^p - \left(\frac{i}{n}\right)^p\right|$. As a result of triangular inequality:

$$\left|H_p^n(x) - G_p^n(x)\right| \leq \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} r_i^p\ \Delta_p(x,i) \tag{4.29}$$

Before analyzing this sum, we note that by the mean value theorem, there exists $\xi \in (x \wedge \frac{i}{n}, x \vee \frac{i}{n})$ hence $x^p - \left(\frac{i}{n}\right)^p = \left(x - \frac{i}{n}\right)\xi^{p-1}$, thus:

$$\Delta_p(x,i) = \left|x^p - \left(\frac{i}{n}\right)^p\right| \leq p\left|x - \frac{i}{n}\right|\left|x \vee \frac{i}{n}\right|^{p-1} \tag{4.30}$$

To analyze the sum in (4.29), we inspect the intervals $i \leq nx$, $nx > i$ separately.

Case-1: $i \leq nx$**:**

$$\sum_{i \leq nx} \binom{n}{i} x^i (1-x)^{n-i} r_i^p\ \Delta_p(x,i) \tag{4.31}$$

$$= \sum_{i \leq nx - \delta_1} \binom{n}{i} x^i (1-x)^{n-i} r_i^p \Delta_p(x,i) + \sum_{nx-\delta_1 < i < nx} \binom{n}{i} x^i (1-x)^{n-i} r_i^p \Delta_p(x,i) \tag{4.32}$$

$$\leq \sum_{i \leq nx - \delta_1} \binom{n}{i} x^i (1-x)^{n-i} 2 + \sum_{nx-\delta_1 \leq i \leq nx} \binom{n}{i} x^i (1-x)^{n-i} r_i^p\ p\left|x - \frac{i}{n}\right| x^{p-1} \tag{4.33}$$

$$\leq 2e^{-\frac{n}{x}\delta_1^2} + \sum_{nx-\delta_1 \leq i \leq nx} p\binom{n}{i} x^i (1-x)^{n-i} r_i^p \delta_1 x^{p-1} \tag{4.34}$$

$$\leq 2e^{-\frac{n}{x}\delta_1^2} + \frac{p\ \delta_1}{x} H_p^n(x) \tag{4.35}$$

In (4.32) we use (4.30). In (4.33), we bound the lower tail of the binomial via Theorem 6 and observe that $\left|x - \frac{i}{n}\right| \leq \delta_1$ in the range $nx - \delta_1 \leq i \leq nx$. We also note that in (4.33), $r_i^p \leq 2$ and $\Delta_p(x, i) \leq 1$. Now we choose $\delta_1 = c_1 \sqrt{-\frac{x}{n} \log\left(\frac{1}{x} H_p^n(x)\right)}$ and obtain:

$$(4.35) \leq 2 \left(\frac{H_p^n(x)}{x}\right)^{c_1^2} + p\ c_1 \sqrt{-\frac{1}{nx} \log\left(\frac{1}{x} H_p^n(x)\right)} H_p^n(x) \tag{4.36}$$

Hence to establish the lemma, we first show that $\sqrt{-\frac{1}{nx} \log \frac{1}{x} H_p^n(x)} = O(\frac{1}{\sqrt{\log n}})$.

$$\begin{aligned}
\frac{1}{x} H_p^n(x) &\simeq C_p \sum_{i=0}^{n} \frac{x^{p-1}}{i^{\frac{p}{2}} + 1} \binom{n}{i} x^i (1-x)^{n-i} && (4.37)\\
&\geq C_p \frac{1}{n^{\frac{p}{2}} + 1} \left(\frac{c \log^2 n}{n}\right)^{p-1} \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} && (4.38)\\
&= C_p \frac{1}{n^{\frac{p}{2}} + 1} \left(\frac{c \log^2 n}{n}\right)^{p-1} && (4.39)\\
&\geq \frac{k'}{n^{\frac{3}{2}p}} && (4.40)
\end{aligned}$$

In (4.38), we notice $x \geq \frac{c \log^2 n}{n}$. On the right-hand side of (4.40) we collect the constants in k'. Therefore we establish that:

$$\sqrt{-\frac{1}{nx} \log \frac{1}{x} H_p^n(x)} \leq \sqrt{k'' \frac{1}{nx} \log n} \leq O\left(\frac{1}{\log(n)}\right) \tag{4.41}$$

where in the first inequality we use that $x \geq c \frac{\log n}{n}$ and collect the constants in k''. Secondly, we need to show that $\frac{H_p^n(x)}{x}$ decays sufficiently fast. To this end, we have:

$$\begin{aligned}
\sum_i \frac{C_p}{i^{\frac{p}{2}} + 1} \binom{n}{i} x^{i+p-1} (1-x)^{n-i} &\leq \sum_i \frac{c_3}{i+1} \binom{n}{i} x^i (1-x)^{n-i} && (4.42)\\
&\leq c_3 \frac{1 - (1-x)^{n+1}}{n+1} x^{p-2} && (4.43)\\
&\leq c_3 \frac{x^{p-2}}{n+1} && (4.44)
\end{aligned}$$

Therefore by choosing c_1 large enough we ensure that $B \left(\frac{H_p^n(x)}{x}\right)^{c_1^2} = o(n^{-\frac{p}{2}})$.

Case-2: $i > nx \quad$:

$$\sum_{i \geq nx} \binom{n}{i} x^i (1-x)^{n-i} r_i^p \ \Delta_p(x,i) \tag{4.45}$$

$$= \sum_{nx < i \leq nx+\delta_2} \binom{n}{i} x^i (1-x)^{n-i} r_i^p \Delta_p(x,i) + \sum_{i > nx+\delta_2} \binom{n}{i} x^i (1-x)^{n-i} r_i^p \Delta_p(x,i) \tag{4.46}$$

$$\leq \sum_{nx < i \leq nx+\delta_2} \binom{n}{i} x^i (1-x)^{n-i} \left| x - \frac{i}{n} \right| \left(\frac{i}{n} \right)^{p-1} \tag{4.47}$$

$$+ 2 \sum_{i > nx+\delta_2} \binom{n}{i} x^i (1-x)^{n-i} \tag{4.48}$$

$$\leq \delta_2 \sum_{nx < i \leq nx+\delta_2} \binom{n}{i} x^i (1-x)^{n-i} \left(x + \delta_2 \right)^{p-1} + 2e^{-\frac{n\delta_2^2}{2(x+\frac{\delta_2}{3})}} \tag{4.49}$$

$$\leq \delta_2 2^{p-1} \sum_{nx < i \leq nx+\delta_2} \binom{n}{i} x^i (1-x)^{n-i} \left(x \vee \delta_2 \right)^{p-1} + e^{-\frac{n\delta_2^2}{\frac{2}{3}(x \vee \delta_2)}} \tag{4.50}$$

$$\leq \delta_2 \ 2^{p-1} \sum_{nx < i \leq nx+\delta_2} \binom{n}{i} x^i (1-x)^{n-i} \left(x \right)^{p-1} + 2e^{-\frac{n\delta_2^2}{\frac{2}{3}x}} \tag{4.51}$$

Each step is justified in the corresponding step in the analysis for the range $i \leq nx$, except now we are using the upper tail in Theorem 6. In (4.51), we see that the problem is identical to (4.35) except for the constants. Hence we choose $\delta_2 = c_2 \sqrt{-\frac{x}{n} \log \left(\frac{1}{x} H_p^n(x) \right)}$ and by (4.40) we have $\delta_2 \leq c_3 \sqrt{x \frac{\log n}{n}}$. This ensures that $\delta_2 \leq x$ when $x \geq c \frac{\log^2 n}{n}$ and the step (4.51) is valid. □

4.3 Proof for Theorem 3

Theorem 3. *Let $m = o(n)$:*

$$\left| R_{m,n}^p - R_m^p \right| \leq O\left(m^{-\frac{p-1}{2}} (n)^{-1/2} \right) \tag{3.3}$$

Proof. We note that By Lemma 2.1.1 and Lemma 4.3.1 we have $r_{m+n}^p = \Theta(n+m)^{-\frac{p}{2}}$ and $R_m^p = \Theta(m^{-\frac{p}{2}})$. Therefore In the regime $m = o(n)$, $\gamma_{m,n}^p = O(m^{-\frac{p-1}{2}} (n)^{-1/2})$. Finally, we observe that $R_{m,n}^p$ monotonically decreases n, and in the limit it reduces

R_m^p. We forammly establish the lower bound $R_m^p \leq R_{m,n}^p$ in Lemma 4.3.4. □

4.3.1 Supplementary Results for Theorem-3

Lemma 4.3.1. *We have*

$$\frac{C_p}{m^{\frac{p}{2}}} \leq R_m^p \leq k_x \frac{C_p}{m^{\frac{p}{2}}} \tag{4.52}$$

Proof. The lower bound follows from choosing $p_X = \delta_x$ and is considered in the proof for Theorem 2. The upper bound follows from:

$$\|p_X p_{Y|X=x} - p_X \hat{q}_{Y|X}\|_p^p = \sum_{x \in \mathcal{X}} (p_X(x))^p \|p_{|X=x} - \hat{q}_{Y|X=x}\|_p^p$$

yielding:

$$\begin{aligned}\sup_{p_{XY}} \mathbb{E}_{L^m}\left[\|p_X p_{Y|X=x} - p_X \hat{q}_{Y|X}\|_p^p\right] &\leq \sum_{x \in \mathcal{X}} \sup_{p_X(x), p_{Y|X=x}} (p_X(x))^p \, \mathbb{E}_{L^m}\left[\|p_{Y|X=x} - \hat{q}_{Y|X=x}\|_p^p\right] \\ &\leq \sum_{x \in \mathcal{X}} \sup_{p_{Y|X=x}} \mathbb{E}_{L^m}\left[\|p_{Y|X=x} - \hat{q}_{Y|X=x}\|_p^p\right]\end{aligned}$$

Finally taking min over $\hat{q}_{Y|X}^m$ and noting that variables $\{\hat{q}_{Y|X=x}^*\}_{x \in \mathcal{X}}$ are independent:

$$\inf_{\hat{q}_{Y|X}} \sup_{p_{XY}} \mathbb{E}_{L^m}\left[\|p_X p_{Y|X=x} - p_X \hat{q}_{Y|X}\|_p^p\right] \leq \sum_{x \in \mathcal{X}} \inf_{\hat{q}_{Y|X=x}} \sup_{p_{Y|X=x}} \mathbb{E}_{L^m}\left[\|p_{Y|X=x} - \hat{q}_{Y|X=x}\|_p^p\right] = |\mathcal{X}| \, r_m^p \tag{4.53}$$

□

Lemma 4.3.2. *For $p \geq 0$, there exists constants $\{c_i\}_{i=0}^p$ and c', c'' such that:*

$$R_{m,n}^p \leq R_m^p + \gamma_{m,n}^p \tag{4.54}$$

for

$$\gamma^p_{m,n} = \sum_{i}^{\lfloor p\rfloor-1} c_i (R^p_m)^{\frac{p-i}{p}} (r^p_{m+n})^{\frac{i}{p}} + c'(R^p_m)^{\frac{p-\lfloor p\rfloor}{p}} (r^p_{m+n})^{\frac{\lfloor p\rfloor}{p}} + c''(R^p_m + r^p_{m+n})^{\frac{p-\lfloor p\rfloor}{p}} (r^p_{n+m})^{\frac{\lfloor p\rfloor}{p}} \tag{4.55}$$

Proof. First let us fix $\hat{q}_{XY}(U,L), U, L$ and let us define:

$$\Gamma^{U,L}_{x,y}(u) \triangleq p_{XY}(x,y) - u\hat{q}_{Y|X}(y \mid x) \tag{4.56}$$

with derivatives:

$$\begin{aligned} \left|\frac{d^i}{du^i}\left|\Gamma^{U,L}_{x,y}(u)\right|^p\right| &= p^{(i)}(\hat{q}_{Y|X}(y \mid x))^i \left|\Gamma^{U,L}_{x,y}(u)\right|^{p-i} && (4.57)\\ &\leq p^{(i)}\left|\Gamma^{U,L}_{x,y}(u)\right|^{p-i} && (4.58) \end{aligned}$$

Continuing, we have:

$$\begin{aligned} \left|\Gamma^{U,L}_{x,y}(\hat{q}_X(x))\right|^p &\leq \sum_{i=0}^{\lfloor p\rfloor-1} \frac{p^{(i)}}{i!}\left|\Gamma^{U,L}_{x,y}(p_X(x))\right|^{p-i} \left|\hat{q}_X(x) - p_X(x)\right|^i && (4.59)\\ +\Big(\left|\Gamma^{U,L}_{x,y}(p_X(x))\right|^{p-\lfloor p\rfloor} &\vee \left|\Gamma^{U,L}_{x,y}(\hat{q}_X(x))\right|^{p-\lfloor p\rfloor}\Big) \left|\hat{q}_X(x) - p_X(x)\right|^{\lfloor p\rfloor} \frac{p^{\lfloor p\rfloor}}{\lfloor p\rfloor!} && (4.60)\\ &\leq \sum_{i=0}^{\lfloor p\rfloor-1} \frac{p^{(i)}}{i!}\left|\Gamma^{U,L}_{x,y}(p_X(x))\right|^{p-i} \left|\hat{q}_X(x) - p_X(x)\right|^i \\ &+ \frac{p^{\lfloor p\rfloor}}{\lfloor p\rfloor!}\left|\Gamma^{U,L}_{x,y}(p_X(x))\right|^{p-\lfloor p\rfloor} \left|\hat{q}_X(x) - p_X(x)\right|^{\lfloor p\rfloor} \\ &+ \frac{p^{\lfloor p\rfloor}}{\lfloor p\rfloor!}\left|\Gamma^{U,L}_{x,y}(\hat{q}_X(x))\right|^{p-\lfloor p\rfloor} \left|\hat{q}_X(x) - p_X(x)\right|^{\lfloor p\rfloor} && (4.61) \end{aligned}$$

In the above display, we Taylor expand the $h(u) \triangleq \left|\Gamma^{U,L}_{x,y}(u)(\hat{q}_X)\right|^p$ around the $u = p_X(x)$ and use the upper bound (4.58) for the derivatives. We bound the remainder of the Taylor expansion with the mean value theorem via the monotonicity of the derivatives as in (4.30), leading to (4.60).

$$\mathbb{E}_{U,L}\left[\|p_{XY} - \hat{q}_{XY}\|^p_p\right] \leq \mathbb{E}_{U,L}\left[\|p_{XY} - p_X\hat{q}_{Y|X}\|^p_p\right]$$

$$
\begin{aligned}
&+ \sum_{i=1}^{\lfloor p\rfloor -1} \frac{p^{(i)}}{i!}\, \mathbb{E}_{U,L}\left[\sum_{x,y} \left|\Gamma^{U,L}_{x,y}(p_X(x))\right|^{p-i} \left|\hat{q}_X(x)-p_X(x)\right|^{i}\right] \\
&+ \frac{p^{\lfloor p\rfloor}}{\lfloor p\rfloor!}\, \mathbb{E}_{U,L}\left[\sum_{x,y} \left|\Gamma^{U,L}_{x,y}(p_X(x))\right|^{p-\lfloor p\rfloor} \left|\hat{q}_X(x)-p_X(x)\right|^{p-\lfloor p\rfloor}\right] \\
&+ \frac{p^{\lfloor p\rfloor}}{\lfloor p\rfloor}\, \mathbb{E}_{U,L}\left[\sum_{x,y} \left|\Gamma^{U,L}_{x,y}(\hat{q}_X(x))\right|^{p-\lfloor p\rfloor} \left|\hat{q}_X(x)-p_X(x)\right|^{p-\lfloor p\rfloor}\right] \qquad (4.62)
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}_{U,L}\left[\|p_{XY}-\hat{q}_{XY}\|_p^p\right] \le \mathbb{E}_{U,L}\left[\|p_{XY}-p_X\hat{q}_{Y|X}\|_p^p\right] \\
&+ \sum_{i=1}^{\lfloor p\rfloor -1} \frac{p^{(i)}k_y^{\frac{i}{p}}}{i!}\, \mathbb{E}\left[\|p_{XY}-p_X\hat{q}_{Y|X}\|_p^p\right]^{\frac{p-i}{p}} \mathbb{E}\left[\|\hat{q}_X-p_X\|_p^p\right]^{\frac{i}{p}} \\
&+ \frac{p^{\lfloor p\rfloor}k_y^{\frac{\lfloor p\rfloor}{p}}}{\lfloor p\rfloor!}\, \mathbb{E}\left[\|p_{XY}-p_X\hat{q}_{Y|X}\|_p^p\right]^{\frac{p-\lfloor p\rfloor}{p}} \mathbb{E}\left[\|\hat{q}_X-p_X\|_p^p\right]^{\frac{\lfloor p\rfloor}{p}} \\
&+ \frac{p^{\lfloor p\rfloor}k_y^{\frac{\lfloor p\rfloor}{p}}}{\lfloor p\rfloor!}\, \mathbb{E}\left[\|p_{XY}-\hat{q}_X\hat{q}_{Y|X}\|_p^p\right]^{\frac{p-\lfloor p\rfloor}{p}} \mathbb{E}\left[\|\hat{q}_X-p_X\|_p^p\right]^{\frac{\lfloor p\rfloor}{p}} \qquad (4.63) \\
&\triangleq \kappa_{m,n}(p_{XY},\hat{q}_{XY})
\end{aligned}
$$

In (4.62) we sum over x, y and take expectations with respect to $U, L \sim p_{XY}$ of both handsides (4.59) ,(4.61) by noting that $\mathbb{E}_{U,L}\left[\sum_{x,y}\left|\Gamma^{U,L}_{x,y}(\hat{q}(X))\right|^p\right] = \mathbb{E}_{U,L}\left[\|p_{XY}-\hat{q}_{XY}\|_p^p\right]$. To obtain (4.63), we apply Hölder's inequality to each summation inside the expectations in (4.62). Taking maximum of both sides over p_{XY} and taking the minimum of the left-hand side over $\hat{q}_{Y|X}$ we establish that for all $\hat{q}_{XY}$:

$$
R^p_{m,n} = \min_{\hat{q}_{XY}}\max_{p_{XY}} \mathbb{E}_{U,L}\left[\|p_{XY}-\hat{q}_{XY}\|_p^p\right] \le \max_{p_{XY}} \kappa_{m,n}(p_{XY},\hat{q}_{XY}) \qquad (4.64)
$$

Hence substituting $\kappa_{m,n}$ we obtain:

$$
\begin{aligned}
R^p_{m,n} \le &\max_{p_{XY}} \mathbb{E}_{U,L}\left[\|p_{XY}-p_X\hat{q}_{Y|X}\|_p^p\right] \; + \\
&\sum_{i=1}^{\lfloor p\rfloor -1} c_i \max_{p_{XY}} \mathbb{E}\left[\|p_{XY}-p_X\hat{q}_{Y|X}\|_p^p\right]^{\frac{p-i}{p}} \max_{p_X} \mathbb{E}\left[\|\hat{q}_X-p_X\|_p^p\right]^{\frac{i}{p}} \\
&+ c' \max_{p_{XY}} \mathbb{E}\left[\|p_{XY}-p_X\hat{q}_{Y|X}\|_p^p\right]^{\frac{p-\lfloor p\rfloor}{p}} \max_{p_X} \mathbb{E}\left[\|\hat{q}_X-p_X\|_p^p\right]^{\frac{\lfloor p\rfloor}{p}} \\
&+ c''' \max_{p_{XY}} \mathbb{E}\left[\|p_{XY}-\hat{q}_X\hat{q}_{Y|X}\|_p^p\right]^{\frac{p-\lfloor p\rfloor}{p}} \max_{p_X} \mathbb{E}\left[\|\hat{q}_X-p_X\|_p^p\right]^{\frac{\lfloor p\rfloor}{p}} \qquad (4.65)
\end{aligned}
$$

$$
\begin{aligned}
R^p_{m,n} \leq \max_{p_{XY}} \mathbb{E}_{U,L}\left[\|p_{XY} - p_X\hat{q}_{Y|X}\|_p^p\right] \; + \\
&\sum_{i=1}^{\lfloor p \rfloor - 1} c_i \max_{p_{XY}} \mathbb{E}\left[\|p_{XY} - p_X\hat{q}_{Y|X}\|_p^p\right]^{\frac{p-i}{p}} \max_{p_X} \mathbb{E}\left[\|\hat{q}_X - p_X\|_p^p\right]^{\frac{i}{p}} \\
&+ c' \max_{p_{XY}} \mathbb{E}\left[\|p_{XY} - p_X\hat{q}_{Y|X}\|_p^p\right]^{\frac{p-\lfloor p \rfloor}{p}} \max_{p_X} \mathbb{E}\left[\|\hat{q}_X - p_X\|_p^p\right]^{\frac{\lfloor p \rfloor}{p}} \\
&+ c'' \max_{p_{XY}} \mathbb{E}\left[\|p_{XY} - p_X\hat{q}_{Y|X}\|_p^p\right]^{\frac{p-\lfloor p \rfloor}{p}} \max_{p_X} \mathbb{E}\left[\|\hat{q}_X - p_X\|_p^p\right]^{\frac{\lfloor p \rfloor}{p}} \\
&+ c'' \max_{p_X} \mathbb{E}\left[\|p_X - \hat{q}_X\|_p^p\right] \qquad (4.66)
\end{aligned}
$$

In (4.65), we further upper bound (4.64) by taking the maximum of each summation separately. We define the constants are $c_i \triangleq p^{(i)} \frac{k_y^{\frac{i}{p}}}{i!}, c' = c''' \triangleq p^{(\lfloor p \rfloor)} \frac{k_y^{\frac{\lfloor p \rfloor}{p}}}{\lfloor p \rfloor !}$ based on (4.63). In (4.66) we note that:

$$
\max_{p_{XY}} \mathbb{E}\left[\|p_{XY} - \hat{q}_X\hat{q}_{Y|X}\|_p^p\right] \leq 2^{p-1} \max_{p_X} \mathbb{E}\left[\|p_{XY} - p_X\hat{q}_{Y|X}\|_p^p\right] + 2^{p-1} \max_{p_{XY}} \mathbb{E}\left[\|p_X - \hat{q}_X\|_p^p\right]
$$

which is a consequence of convexity of $|x|^p$ for $p \geq 1$. For completeness, we include a proof for this in Lemma 4.3.3. Finally, we choose $\hat{q}_X$, $\hat{q}_{Y|X}$ to be the minimax estimators of the r^p_m and R^p_m respectively to establish the Lemma 4.3.2. □

Lemma 4.3.3.

$$
\max_{p_{XY}} \mathbb{E}\left[\|p_{XY} - \hat{q}_X\hat{q}_{Y|X}\|_p^p\right] \leq 2^{p-1} \max_{p_X} \mathbb{E}\left[\|p_{XY} - p_X\hat{q}_{Y|X}\|_p^p\right] + 2^{p-1} \max_{p_{XY}} \mathbb{E}\left[\|p_X - \hat{q}_X\|_p^p\right]
$$

Proof. We have,

$$
\begin{aligned}
|p_{XY}(x,y) - \hat{q}_{XY}(x,y)|^p &= \left|p_{XY}(x,y) - p_X(x)\hat{q}_{Y|X}(y \mid x) + p_X(x)\hat{q}_{Y|X}(y \mid x) - \hat{q}_{XY}(x,y)\right|^p \\
&= 2^p\left|\frac{1}{2}(p_{XY}(x,y) - p_X(x)\hat{q}_{Y|X}(y \mid x)) + \frac{1}{2}(p_X(x)\hat{q}_{Y|X}(y \mid x) - \hat{q}_{XY}(x,y))\right|^p \\
&\leq 2^{p-1}\left|p_{XY}(x,y) - p_X(x)\hat{q}_{Y|X}(y \mid x)\right|^p + 2^{p-1}\left|p_X(x)\hat{q}_{Y|X}(y \mid x) - \hat{q}_{XY}(x,y)\right|^p \quad (4.67) \\
&\leq 2^{p-1}\left|p_{XY}(x,y) - p_X(x)\hat{q}_{Y|X}(y \mid x)\right|^p + 2^{p-1}\left|p_X(x) - \hat{q}_X(x)\right|^p \quad (4.68)
\end{aligned}
$$

where in (4.67) we use Jensen's inequality. Summing over $x, y \in \mathcal{X}, \mathcal{Y}$ and taking expectation over $\{X'_i\}_{i=1}^n, \{(X_j, Y_j)\}_{j=1}^m \sim p_{XY}$ of both sides (4.3.1) and (4.68) we obtain:

$$\mathbb{E}\left[\|p_{XY} - \hat{q}_X\hat{q}_{Y|X}\|_p^p\right] \leq 2^{p-1}\,\mathbb{E}\left[\|p_{XY} - p_X\hat{q}_{Y|X}\|_p^p\right] + 2^{p-1}\,\mathbb{E}\left[\|p_X - \hat{q}_X\|_p^p\right]$$

taking maximum over both sides over p_{XY} we obtain:

$$\begin{aligned}\max_{p_{XY}}\mathbb{E}\left[\|p_{XY} - \hat{q}_X\hat{q}_{Y|X}\|_p^p\right] &\leq \max_{p_{XY}}\left(2^{p-1}\,\mathbb{E}\left[\|p_{XY} - p_X\hat{q}_{Y|X}\|_p^p\right] + 2^{p-1}\,\mathbb{E}\left[\|p_X - \hat{q}_X\|_p^p\right]\right)\\ &\leq 2^{p-1}\max_{p_{XY}}\mathbb{E}\left[\|p_{XY} - p_X\hat{q}_{Y|X}\|_p^p\right] + 2^{p-1}\max_{p_X}\mathbb{E}\left[\|p_X - \hat{q}_X\|\right]\end{aligned}$$

□

Lemma 4.3.4. *For all m, n and $p \geq 1$:*

$$R_m^p \leq R_{m,n}^p$$

Proof. Let us denote the adversarial distribution of the R_m^p by p_{XY}^*. We choose a prior $\Pi_{p_{XY}}$ over $\Delta_{\mathcal{X}\times\mathcal{Y}}$ such that $\pi_{p_{XY}} = \delta_{p_X^*}\pi_{p_{Y|X}}$, we leave the choice of the $\pi_{p_{Y|X}}$ free as long as $\operatorname{supp}(\pi_{p_{Y|X=x}}) = \Delta_{\mathcal{Y}}$ for all $x \in \mathcal{X}$. Also let $U = \{X_i\}_{i=1}^m$ and $L = \{(X_i, Y_i)\}_{i=1}^n$ for notational convinence. Then,

$$\begin{aligned}R_{m,n}^p &= \min_{\hat{q}_{XY}}\max_{p_{XY}}\mathbb{E}\left[\|p_{XY} - \hat{q}_{XY}\|_p^p\right] &(4.69)\\ &\geq \min_{\hat{q}_{XY}}\mathbb{E}_{p_{XY}\sim\pi_{XY}}\left[\mathbb{E}_{U,L\sim p_{XY}}\left[\|\hat{q}_{XY} - p_{XY}\|_p^p\right]\right] &(4.70)\\ &= \mathbb{E}_{p_{XY}\sim\pi_{XY}}\left[\mathbb{E}_{U,L\sim p_{XY}}\left[\|\hat{q}_{XY}^{\pi_{XY}} - p_{XY}\|_p^p\right]\right] &(4.71)\\ &= \mathbb{E}_{p_{XY}\sim\pi_{XY}}\left[\mathbb{E}_{U,L\sim p_{XY}}\left[\|p_X^*\hat{q}_{Y|X}^{\pi_{XY}} - p_Xp_{Y|X}\|_p^p\right]\right] &(4.72)\\ &= \mathbb{E}_{p_{XY}\sim\pi_{XY}}\left[\mathbb{E}_{U,L\sim p_{XY}}\left[\|p_X^*\hat{q}_{Y|X}^{\pi_{XY}} - p_X^*p_{Y|X}\|_p^p\right]\right] &(4.73)\\ &= \min_{\hat{q}_{Y|X}}\max_{p_{Y|X}}\mathbb{E}\left[\|p_X^*\hat{q}_{Y|X} - p_X^*p_{Y|X}\|_p^p\right] &(4.74)\\ &= \min_{\hat{q}_{Y|X}}\max_{p_Xp_{Y|X}}\mathbb{E}\left[\|p_X\hat{q}_{Y|X} - p_Xp_{Y|X}\|_p^p\right] = R_m^p &(4.75)\end{aligned}$$

in (4.70) we lower bound the supremum with the average. In (4.71) we set $\hat{q}_{XY}^{\pi_{XY}}$ to be the Bayes estimator for the prior $\pi_{p_{XY}}$ which minimizes the posterior risk for an assignment U, L:

$$\hat{q}_{XY}^{\pi_{XY}}(U,L) \triangleq \operatorname*{arg\,min}_{q_{XY}\in\Delta_{\mathcal{X}\mathcal{Y}}} \mathbb{E}_{p_{XY}\sim\pi_{p_{XY}|U,L}}\left[\|p_{XY}-q_{XY}\|_p^p\right] \tag{4.76}$$

$$\triangleq \operatorname*{arg\,min}_{q_{XY}\in\Delta_{\mathcal{X}\mathcal{Y}}} F(q_{XY},U,L) \tag{4.77}$$

In (4.72), we note that the functional $F(q_{XY},U,L)$ is minimized by some element in the support of $\pi_{XY|U,L}$ by Lemma 4.3.5. Since for all U and L, $\operatorname{supp}(\pi_{XY|U,L}) \subseteq \{p_X^* q_{Y|X} : q_{Y|X} \in \Delta_{\mathcal{Y}}^{|\mathcal{X}|}\}$ we have $\hat{q}_X^{\pi_{XY}}(U,L) = p_X^*$. In (4.73) we again use that the $\pi_X = \delta_{p_X^*}$. In (4.74), we use that the prior $\pi_{Y|X}$ is essentially free, and since any minimax risk can be approximated arbitrarily by a sequence of priors by the minimax theorem [13]. Finally (4.75) follows since p^* is the adversarial distribution of R_m^p. □

We introduce the following trivial result for the sake of completeness:

Lemma 4.3.5. *Let $\mathcal{P} \subset \mathbb{R}^d$ be a bounded convex set. Let μ be a probability measure on $(\mathbb{R}^d, \mathbb{B}(\mathbb{R}^d))$ with support $\mathcal{P}$. Let be $x^* \in \mathbb{R}^d$ such that*

$$\inf_{x\in\mathbb{R}^d} \int \|x-y\|_p^p d\mu(y) = \int \|x^*-y\|_p^p d\mu(y)$$

then $x^* \in \mathcal{P}$.

Proof. Suppose for the contradiction that $x^* \notin \mathcal{P}$. Then we let,

$$x^{**} \triangleq \inf_{x\in\mathcal{P}} \|x^*-x\|_2^2$$

such $x^*{*}$ exists by the convexity of $\mathcal{P}$ and the convexity of l_2 norm. Then we note that $\forall y \in \mathcal{P}$:

$$\begin{aligned} \|x^{**}-y\|_p^p \leq \|x^*-y\|_p^p &\implies \|x^{**}-y\|_2^2 \leq \|x^*-y\|_2^2 \\ &\implies \int \|x^{**}-y\| d\mu(y) \leq \int \|x^*-y\| d\mu(y) \end{aligned}$$

□

4.4 Proof for Theorem 4

Theorem 4. *Let $p \geq 2$ and $\hat{q}_n^*$ be a first-order optimal estimator for r_n^p. Then the joint composition $\hat{q}_{XY}^{*,m,n}$ based on $\hat{q}_n^*$ is first order minimax optimal for $R_{m,n}^p$ in the regime $m = o(n)$.*

Proof. According to *Theorem* 2, the composition estimator $\hat{q}_{Y|X}^{*,m}$ achieves first-order minimax optimality for R_m^p when $p \geq 2$. Moreover, as stated in *Theorem* 3, in the regime where $m = o(n)$, both $R_{m,n}^p$ and R_m^p exhibit the same first-order behavior. □

Chapter 5

Conclusion

In this study, we embarked on tackling the challenging problem of minimax estimation of probability mass functions (pmfs). Our objective was to capture fundamental notions pertaining to this problem, shedding light on the optimal strategies for pmf estimation under different loss functions. Specifically, we focused on the minimax risk in the first-order constant for the l_p^p loss, considering scenarios where there are m labeled and n unlabeled samples.

Through our investigation, we were able to identify the optimal estimator that achieves the minimax risk in the first order constant for the l_p^p loss. Notably, we demonstrated that for $p \geq 2$, the composition estimators of univariate minimax problems emerged as the optimal choice in the first order over the regime where $m = o(n)$. This finding highlights the efficacy of composition estimators in achieving optimal performance within this specific context. However, it is important to note that the semisupervised pmf estimation problem remains an open and unresolved area of research, offering ample opportunities for future exploration and advancements.

Moving forward, there are several avenues for future research that can build upon the current findings. One potential direction involves extending the scope of the current results to encompass a broader range of loss functions, such as f-divergences. By considering alternative divergence measures, we can gain a more comprehensive understanding of the optimal estimation strategies under different contexts. Additionally, it would be valuable to investigate the case where $1 \leq p \leq 2$ within the framework of minimax pmf estimation. This extension would provide insights into the performance limits and optimal strategies in scenarios where the l_p^p loss function is applicable.

In conclusion, our work has made significant strides in addressing the problem of minimax pmf estimation, particularly under the l_p^p loss function. By identifying the optimal estimator in achieving the first-order constant for the minimax risk and considering the regime of labeled and unlabeled samples, we have contributed to the existing body of knowledge in this field. Nonetheless, there are still exciting research opportunities to explore, including the semisupervised pmf estimation problem and the extension of results to encompass f-divergences and for the l_p^p losses when $1 \leq p \leq 2$ regime. By pursuing these avenues, we can advance our understanding and capabilities in pmf estimation, enhancing decision-making processes and performance in various domains reliant on accurate probability estimation.

Appendix A

Figures

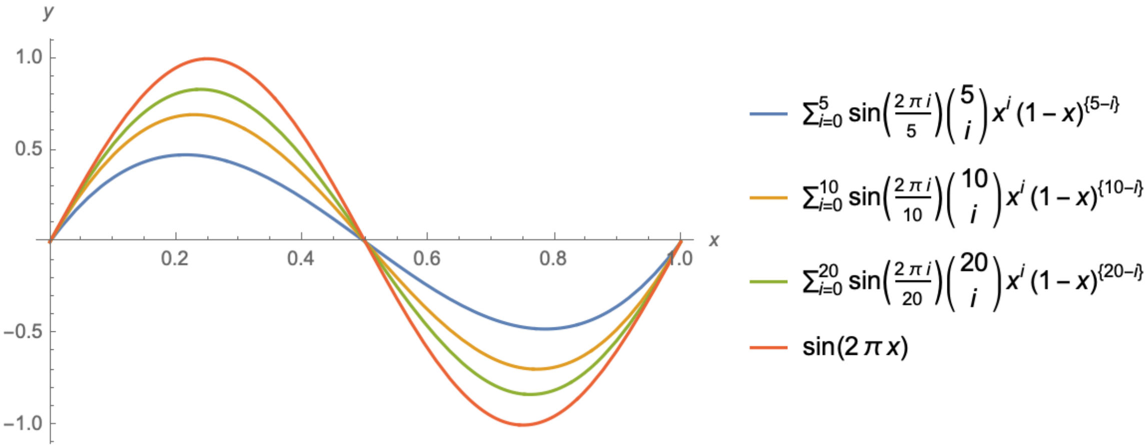

Figure A-1: Smooth Approximation of Polynomials.
We notice that all Bernstein approximations follow the convexity pattern of $\sin(2\pi x)$

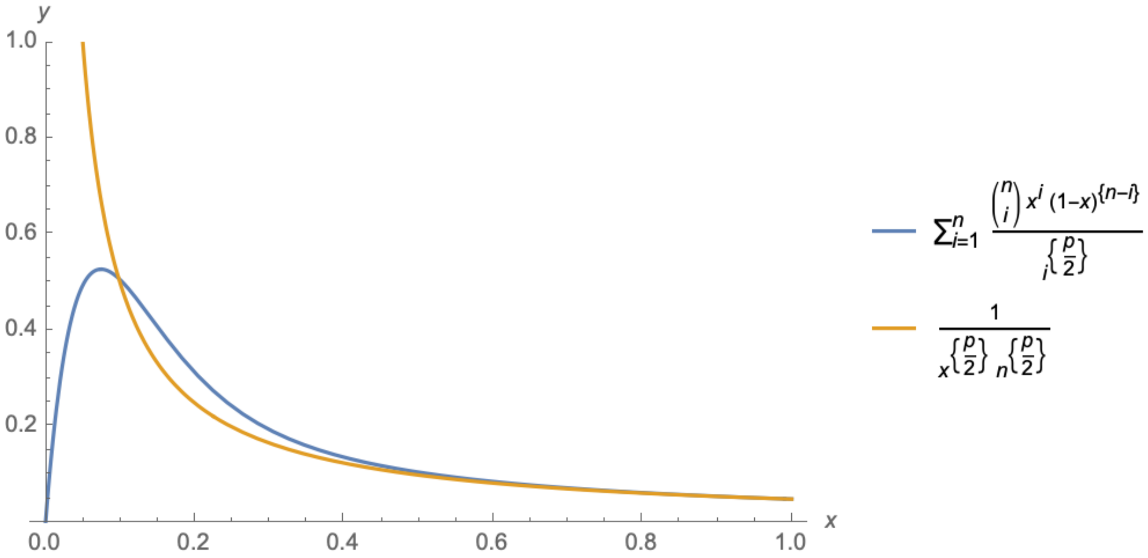

Figure A-2: Bernstein approximation for $(nx)^{\frac{-p}{2}}$ vs $(nx)^{\frac{-p}{2}}$.
In this plot $p = 2$ and $n = 20$, we notice the approximation error near 0.

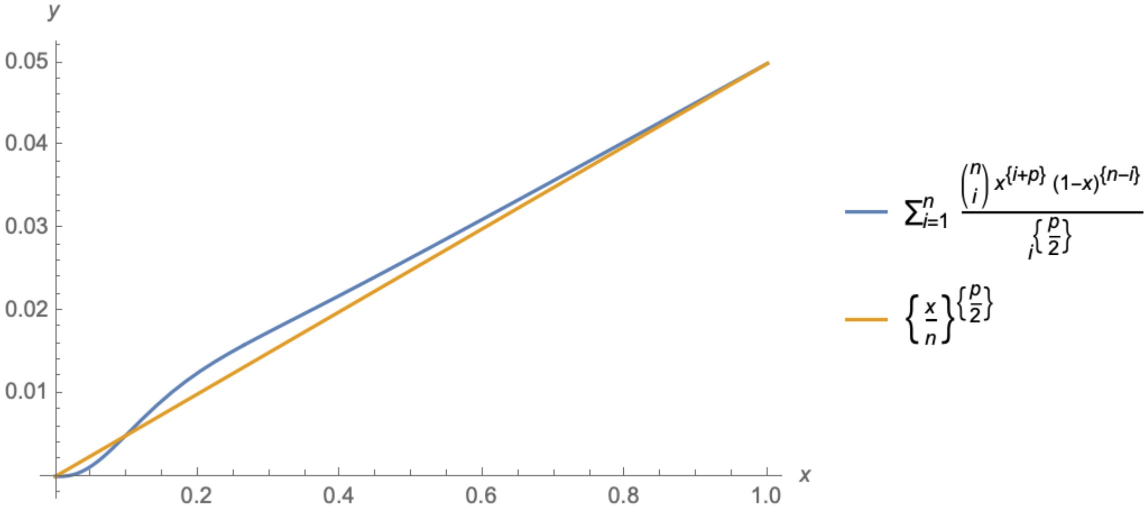

Figure A-3: Bernstein approximation for $\left(\frac{x}{n}\right)^{\frac{p}{2}}$ vs $\left(\frac{x}{n}\right)^{\frac{p}{2}}$.
In this plot $p = 2$ and $n = 20$.

Appendix B

Omitted Proofs for Chapter-2

B.1 Digression on Bayesian Methods

Let us consider the classical estimation setup, where given a model parametrized by θ, $p(x;\theta)$ we wish to estimate the θ^* after observing a set of samples $\{x_i\}_{i=1}^n \sim p(.;\theta^*)$. A belief encodes our initial thought about what parameter $\theta*$ may be through a probability measure Π whiich is a member of the set of all measures over the parameter space θ and an σ-algebra, which we denote by $\mathcal{M}(\theta,\sigma)$ over this parameter set. Intutively this . After the observation $\{x_i\}_{i=1}^n$ is made the belief Π is revisioned according to Bayes rule $\Gamma : \mathcal{M}(\theta,\sigma) \to \mathcal{M}(\theta,\sigma)$. Assuming that the likelihood $L(\{x_i\}_{i=1}^m;\theta)$ and Π admits a densities π, $l(\{x_i\}_{i=1}^m;\theta)$ relative to a given dominating measure, one can write the operation of the Bayes belief update using the Bayes' rule:

$$\Gamma(\theta;\Pi,\{x_i\}_{i=1}^n) = \frac{l(\{x_i\}_{i=1}^m;\theta)\ \pi(\theta)}{\int l(\{x_i\}_{i=1}^m;\theta)\ d\Pi(\theta)} \tag{B.1}$$

Under some mild assumptions, it can be shown that Bayes update yield a consistent estimator of θ^*, and Bernstein-von Mises Theorem [12, Theorem 10.1] contitutes one good example. This statement can be interpreted as that bayes methods work in a frequentist' world, and establishes a link between bayesian and frequentists points of views.

B.2 Conjugate Priors

A conjugate prior is an algebraic convenience yielding closed form solutions for the Bayesien belief updates ((B.1)), otherwise would be analytically intractable. In particular, one remains in the same parametrized family of distributions after performing successive belief updates, reducing the belief update procedure into a calculation of a new parameter pointing to a member in the family. We can derive for the multinomial likelihood function, as is the case in the PMF estimation problem, the corresponding conjugate family prior trivially. Let $\{x_i\}_{i=1}^n$ be a collection of observations from the p_X and we define for $x \in \mathcal{X}$, $T_x \triangleq \sum_{i=1}^n \mathbb{1}\{x_i = x\}$. Then the density of the likelihood becomes:

$$l(\{x_i\}_{i=1}^n; p_X) = \prod_{i=1}^n p_X(x_i) = \prod_{x \in \mathcal{X}} p_X(x)^{T_x}$$

We see that for a member of the conjugate prior family parametrized by $\beta \triangleq [\beta_1, ..., \beta_{|\mathcal{X}|}]$:

$$\Pi(p_X; \beta) \propto \prod_{x \in \mathcal{X}} (p_X(x))^{\beta_x} \tag{B.2}$$

the Bayesian updates Γ yields another member of the family:

$$\begin{aligned}\Gamma(p_X;\ \Pi(\beta), \{x_i\}_{i=1}^n) \propto &\Pi(p_X) l(\{x_i\}_{i=1}^n; p_X)\\ =&\prod_{x\in\mathcal{X}} (p_X(x))^{\beta} \prod_{x\in\mathcal{X}} p_X(x)^{T_x}\\ =&\prod_{x\mathcal{X}} (p_X(x))^{\beta+T_x}\\ \propto&\Pi(p_X; \beta + T)\end{aligned}$$

where $T \triangleq [T_1, ..., T_x]$, is the *occurences* of the samples $\{x_i\}_{i=1}^n$. The family of distributions (B.2) are called Dirichlet distribution and we will denote:

$$\mathrm{Dir}(p_X; \beta) = \frac{1}{\mathrm{B}(\beta)} \prod_{x\in\mathcal{X}} (p_X(x))^{\beta_x} \tag{B.3}$$

where the normalization constant is the Beta function defined in-terms of the Gamma function as:

$$\mathrm{B}(\beta) = \frac{\prod_{x\in\mathcal{X}} \Gamma(\beta_x)}{\Gamma(\sum_{x\in\mathcal{X}} \beta_x)} \tag{B.4}$$

Dirichlet distribution has nice analytical properties. Of particular interest is the mean:

$$\mathbb{E}_{p_X\sim\mathrm{Dir}(.;\beta)}\left[p_X(x)\right] = \frac{\beta_x}{\sum_{x\in\mathcal{X}} \beta_x}$$

Hence as per the discussion in appendix B.3, the bayes estimator for l_2^2 loss before the observation for $p_X(x)$ is $\frac{\beta_x}{\sum_{x\in\mathcal{X}} \beta_x}$. After observing the samples $\{x_i\}_{i=1}^n$ the bayes estimator becomes $\frac{\beta_x+T_x}{\sum_{x\in\mathcal{X}} \beta_x+T_x}$, which in a way *translates* according to the *occurences* of the observations $\{x_i\}_{i=1}^n$. Hence when all β_x is the same and β_n we obtain (2.18).

B.3 Minimum Mean Square Estimation

Theorem 7. *Let Y, Z be two random variables with finite second moments defined on the probability space $(\Sigma, \mathcal{F}, \mu)$. Let F denote all the measurable functions of Z w.r.t. Let $g \in \mathrm{F}$ such that,*

$$\inf_{f\in\mathrm{F}} \mathbb{E}\left[(Y - f(Z))^2\right]. = \mathbb{E}\left[(Y - g(Z))^2\right] \tag{B.5}$$

then

$$g(Z) = \mathbb{E}\left[Y \mid Z\right] \quad a.s.$$

Proof. This theorem can be derived directly from the definition of conditional expectation as a projection in the Hilbert space $L^2(\Sigma, \mathcal{F}, \mu)$ onto the subspace $L^2(\Sigma, \mathcal{F}_X, \mu)$, where $\mathcal{F}_X$ represents the sigma-algebra generated by the random variable X. However, we will present a more accessible and elementary proof, which is outlined below.

Let $h \in \mathrm{F}$. Then we gave:

$$\begin{aligned}
\mathbb{E}\left[(Y-h(Z))^2\right] = \mathbb{E}\left[(Y-g(Z)+g(Z)-h(Z))^2\right] & \\
=& \mathbb{E}\left[(Y-g(Z))^2\right]+\mathbb{E}\left[(g(Z)-h(Z))^2\right]+2\,\mathbb{E}\left[(Y-g(Z))(g(Z)-h(Z))\right] \\
=& \mathbb{E}\left[(Y-g(Z))^2\right]+\mathbb{E}\left[(g(Z)-h(Z))^2\right]+2\,\mathbb{E}\left[\mathbb{E}\left[(Y-g(Z))(g(Z)-h(Z)) \mid Z\right]\right] \\
=& \mathbb{E}\left[(Y-g(Z))^2\right]+\mathbb{E}\left[(g(Z)-h(Z))^2\right]+2\,\mathbb{E}\left[\underbrace{(\mathbb{E}\left[Y \mid Z\right]-g(Z))}_{=0}(g(Z)-h(Z))\right] \\
=& \mathbb{E}\left[(Y-g(Z))^2\right]+\mathbb{E}\left[(g(Z)-h(Z))^2\right]
\end{aligned}$$

In the last step, we see $\mathbb{E}\left[(Y-g(Z))^2\right]$ does not depend on the choice of h. Therefore if h can minimize the square error only if $\mathbb{E}\left[(g(Z)-h(Z))^2\right]$ which happens $g(Z)=h(Z)$ a.s. □

B.4 Proof for Lemma 2.1.1

Here we recall Lemma 2.1.1:

Lemma 2.1.1. *For* $|\mathcal{X}| = O_n(1)$, $r_n^p = \Theta(n^{-\frac{p}{2}})$

Proof. For the lower bound we utilize Lemma B.4.1 and obtain by letting $k \triangleq |\mathcal{X}|$:

$$\|p_X - \check{q}_X(X_1^n)\|_p \, k^{1-\frac{1}{p}} \geq \|p_X - \check{q}_X(X_1^n)\|_1 \tag{B.6}$$

$$\|p_X - \check{q}_X(X_1^n)\|_p \geq \|p_X - \check{q}_X(X_1^n)\|_1 \frac{1}{k^{1-\frac{1}{p}}} \tag{B.7}$$

$$\|p_X - \check{q}_X(X_1^n)\|_p^p \geq \left(\|p_X - \check{q}_X(X_1^n)\|_1 \frac{1}{k^{1-\frac{1}{p}}}\right)^p \tag{B.8}$$

$$\|p_X - \check{q}_X(X_1^n)\|_p^p \geq \left(\|p_X - \check{q}_X(X_1^n)\|_1\right)^p \frac{1}{k^{p-1}} \tag{B.9}$$

$$\|p_X - \check{q}_X(X_1^n)\|_p^p \geq \left(\|p_X - \check{q}_X(X_1^n)\|_1\right)^p k^{1-p} \tag{B.10}$$

□

Now taking the expectation of both sides of the inequality over $X_1^n \sim p_X$:

$$\mathbb{E}\left[\|p_X - \check{q}_X(X_1^n)\|_p^p\right] \geq \mathbb{E}\left[(\|p_X - \check{q}_X(X_1^n)\|)^p\right] k^{1-p} \tag{B.11}$$

$$\mathbb{E}\left[\|p_X - \check{q}_X(X_1^n)\|_p^p\right] \geq \left(\mathbb{E}\left[\|p_X - \check{q}_X(X_1^n)\|\right]\right)^p k^{1-p} \tag{B.12}$$

$$\min_{\check{q}_X} \max_{p_X} \mathbb{E}\left[\|p_X - \check{q}_X(X_1^n)\|_p^p\right] \geq \min_{\check{q}_X} \max_{p_X} \left(\mathbb{E}\left[\|p_X - \check{q}_X(X_1^n)\|\right]\right)^p k^{1-p} \tag{B.13}$$

$$\min_{\check{q}_X} \max_{p_X} \mathbb{E}\left[\|p_X - \check{q}_X(X_1^n)\|_p^p\right] \geq \left(\min_{\check{q}_X} \max_{p_X} \mathbb{E}\left[\|p_X - \check{q}_X(X_1^n)\|\right]\right)^p k^{1-p} \tag{B.14}$$

$$\min_{\check{q}_X} \max_{p_X} \mathbb{E}\left[\|p_X - \check{q}_X(X_1^n)\|_p^p\right] \geq \left(\sqrt{\frac{2(k-1)}{\pi n}} + O\left(\frac{1}{n^{3/4}}\right)\right)^p k^{1-p} \tag{B.15}$$

where in (B.12) we use Jensen's inequality and in (B.15) we use Corollary-9 from [7]. Hence,

$$r_n^p \geq \left(\sqrt{\frac{2(k-1)}{\pi n}} + O\left(\frac{1}{n^{3/4}}\right)\right)^p k^{1-p} \tag{B.16}$$

$$= \left(\left(\sqrt{\frac{2(k-1)}{\pi n}}\right)^p + O\left(\frac{1}{n^{3/4}}\right)\left(\sqrt{\frac{2(k-1)}{\pi n}}\right)^{p-1}\right) k^{1-p} \tag{B.17}$$

$$= \left(\sqrt{\frac{2(k-1)}{\pi n}}\right)^p k^{1-p} + O\left(\frac{1}{n^{\frac{2p+1}{4}}}\right) \tag{B.18}$$

For the upper bound, we note that the same strategy would not work because of the irreversibility of Jensen's step. However, we note that in the large sample regime, we should expect the random variable $\|p_X - \check{q}_X\|_p^p$ (B.12) to concentrate, allowing to reverse the Jensen inequality with some error margin. Although this is an interesting technique for the upper bound, we will plug in the MLE estimator and use its light-tails. In particular, we have by Hoeffding's inequality:

$$\mathbb{P}\left(|\hat{p}_X(x) - p_X(x)| \geq t\right) \leq 2e^{-2nt^2}$$

by Hoeffding's inequality, then for the MLE estimator $\hat{p}_X$:

$$\mathbb{E}\left[\|\hat{p}_X - p_X\|_p^p\right] = \sum_{x \in \mathcal{X}} \mathbb{E}\left[|\hat{p}_X(x; X_1^n) - p_X(x)|^p\right] \tag{B.19}$$

$$= \sum_{x \in \mathcal{X}} \mathbb{E}\left[|\hat{p}_X(x; X_1^n) - p_X(x)|^p\right] \tag{B.20}$$

$$= \sum_{x \in \mathcal{X}} \int_{t=0}^{\infty} \mathbb{P}\left(|\hat{p}_X(x; X_1^n) - p_X(x)|^p \geq t\right) dt \tag{B.21}$$

$$= \sum_{x \in \mathcal{X}} \int_{t=0}^{\infty} \mathbb{P}\left(|\hat{p}_X(x; X_1^n) - p_X(x)| \geq \sqrt[p]{t}\right) dt \tag{B.22}$$

$$= \sum_{x \in \mathcal{X}} \int_{u=0}^{\infty} \mathbb{P}\left(|\hat{p}_X(x; X_1^n) - p_X(x)| \geq u\right) p u^{p-1} du \tag{B.23}$$

$$= \sum_{x \in \mathcal{X}} \int_{u=0}^{\infty} 2e^{-2nu^2} p u^{p-1} du \tag{B.24}$$

We let $v = 2nu^2$:

$$= \sum_{x \in \mathcal{X}} \int_{u=0}^{\infty} 2e^{-2nu^2} p \left(u\right)^{p-2} u du \tag{B.25}$$

$$= \sum_{x \in \mathcal{X}} \int_{u=0}^{\infty} 2e^{-v} p \left(\frac{v}{2n}\right)^{\frac{p-2}{2}} \frac{1}{4n} dv \tag{B.26}$$

$$= \left(\frac{1}{2n}\right)^{\frac{p}{2}} p \sum_{x \in \mathcal{X}} \int_{u=0}^{\infty} e^{-v} \left(v\right)^{\frac{p-2}{2}} dv \tag{B.27}$$

$$= \left(\frac{1}{2n}\right)^{\frac{p}{2}} p \sum_{x \in \mathcal{X}} \Gamma\left(\frac{p}{2}\right) \tag{B.28}$$

$$= \left(\frac{1}{2n}\right)^{\frac{p}{2}} p k \Gamma\left(\frac{p}{2}\right) \tag{B.29}$$

$$\leq \left(\frac{1}{2n}\right)^{\frac{p}{2}} p k \left(\frac{p}{2}\right)^{\frac{p}{2}} \tag{B.30}$$

where there is a suboptimal k in front of the upper bound. For the purposes of proving the lemma, the gap between k^{1-p} and k is insignificant. Essentially, this leaves obtaining a general formula that reflects the interaction between k, n for $r_{k,n}^p$ open.

Lemma B.4.1 (Equivalence of norms on finite dimensional linear spaces)**.** *: Let $\frac{p}{q} \geq 1$ and $\|\|_p, \|\|_q$ be the l_p, l_q norms for $\mathbb{R}^n$. Then $\forall x \in \mathbb{R}^n$:*

$$\|x\|_q \leq n^{\frac{1}{q} - \frac{1}{p}} \|x\|_p \tag{B.31}$$

Proof. This is a classical application of Hölder's inequality:

$$\begin{aligned}
\|x\|_q^q =& \sum_{i=1}^{n} |x_i|^q \, 1 && \text{(B.32)}\\
\leq& \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{q}{p}} \left(\sum_{i=1}^{n} |1|^{\frac{q}{p-1}} \right)^{1-\frac{q}{p}} && \text{(B.33)}\\
=& \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{q}{p}} n^{1-\frac{q}{p}} && \text{(B.34)}
\end{aligned}$$

yielding:

$$\begin{aligned}
\|x\|_q = \left(\sum_{i=1}^{n} |x_i|^q \right)^{\frac{1}{q}} \leq& \left(\left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{q}{p}} n^{1-\frac{q}{p}} \right)^{\frac{1}{q}} && \text{(B.35)}\\
=& \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} n^{\frac{1}{q}-\frac{1}{p}} && \text{(B.36)}\\
=& \|x\|_p n^{\frac{1}{q}-\frac{1}{p}} && \text{(B.37)}
\end{aligned}$$

□

Bibliography

[1] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, dec 2005.

[2] Dietrich Braess and Thomas Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187–206, 2004.

[3] Fan Chung and Linyuan Lu. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6(2):125–145, November 2002.

[4] Ronald A DeVore and George G Lorentz. *Constructive approximation*, volume 303. Springer Science & Business Media, 1993.

[5] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions under l_1 loss. *CoRR*, abs/1411.1467, 2014.

[6] Kun He. An ancillarity paradox in the estimation of multinomial probabilities. *Journal of the American Statistical Association*, 85(411):824–828, September 1990.

[7] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1066–1100, Paris, France, 03–06 Jul 2015. PMLR.

[8] Alisa Kirichenko and Peter Grünwald. Minimax rates without the fixed sample size assumption. 2020.

[9] Erich L. Lehmann and George Casella. *Theory of Point Estimation.* Springer-Verlag, New York, NY, USA, second edition, 1998.

[10] Ingram Olkin and Milton Sobel. Admissible and minimax estimation for the multinomial distribution and for k independent binomial distributions. *The Annals of Statistics*, 7(2), March 1979.

[11] Stanislaw Trybula. Some problems of simultaneous minimax estimation. *The Annals of Mathematical Statistics*, 29(1):245–253, 1958.

[12] A.W. van der Vaart. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000.

[13] Abraham Wald. Statistical decision functions. *The Annals of Mathematical Statistics*, 20(2):165–205, 1949.

[14] Maciej Wilczyński. Minimax estimation for the multinomial and multivariate hypergeometric distributions. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 128–132, 1985.