

On Semi-Supervised Estimation of Distributions

H.S.Melihcan Erol, Erixhen Sula and Lizhong Zheng

Dept. EECS and RLE

Massachusetts Institute of Technology

Cambridge, MA 02139

{hsmerol, esula, lizhong}@mit.edu

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 $m = o(n)$.

I. INTRODUCTION

Estimating the probability mass function (pmf) is a crucial statistical task. A commonly used formulation for estimation is the minimax framework [1]. Early work resolved the minimax risk of pmf estimation under l_2^2 loss by identifying the minimax estimator [2]–[4]. Later studies determined the constant of the first order for the minimax risk under KL-divergence, l_1 and f -divergences [5]–[7].

Meanwhile, the last decades have witnessed a rapid expansion in the sizes of the available datasets for which the labeling efforts lag behind. This results in heterogeneous datasets where a significant portion of the samples lack some of the variables. However, the estimators considered in the works [2]–[7], can only operate in two modes in this setting: either the estimator should ignore the complete samples to estimate only the marginal pmf of the corresponding variable or the estimator should neglect the incomplete samples to form an estimate of the joint pmf. Adopting the naming conventions from machine learning literature [8], these modes of operation can be categorized as unsupervised and supervised estimation, respectively. However, both modes suffer from inefficiencies, prompting the need for estimators that can leverage both labeled and unlabeled samples, commonly referred to as semi-supervised learning. This paper focuses on investigating the fundamental limits of semi-supervised pmf estimators.

In particular, we study the case where there are two random variables X, Y jointly distributed with p_{XY} . We observe two datasets: m i.i.d. samples of (x_i, y_i) pairs drawn from p_{XY} and n samples of only x_j drawn from the marginal distribution p_X . Our goal is to find the minimax estimator of p_{XY} based on these observations. As a main result, we establish that the composition of minimax univariate estimators achieves the correct first-order term of the risk for the semi-supervised estimation problem in the regime $m = o(n)$.

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 [9] and [10], 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 [9], [10].

II. PRELIMINARIES & NOTATION

We employ Bachmann–Landau asymptotic notation and say $a_n = o(b_n)$ if $\limsup_n \frac{a_n}{b_n} = 0$, $a_n = O(b_n)$ if $\limsup_n \frac{a_n}{b_n} = K < \infty$, $\Theta(a_n) = b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. δ_x indicates a unit point mass at $x \in \mathcal{X}$. We write for $x, y \in \mathbb{R}$, $x \wedge y \triangleq \min(x, y)$, $x \vee y \triangleq \max(x, y)$. We denote the space of probability distributions over the finite set $\mathcal{X}$ by $\Delta_{\mathcal{X}}$. We reserve the symbols $k_x = |\mathcal{X}|$ and $k_y = |\mathcal{Y}|$. We use the upper case of a letter to denote a random variable and the lower case to indicate the realization of that random variable. Let $\mathcal{L} : \Delta_{\mathcal{X}} \times \Delta_{\mathcal{X}} \rightarrow \mathbb{R}$ be a loss function. For a set of joint samples $s = \{(x_i, y_i)\}_{i=1}^m$, we define their marginal sets as $s_X \triangleq \{x_i\}_{i=1}^m$ and $s_Y \triangleq \{y_i\}_{i=1}^m$ and its conditional subsets as $s_{Y|X=x} = \{(x_i, y_i) : 1 \leq i \leq m : x_i = x\}$. We denote by $T_x(s^m)$ the number of samples with value x in the set s^m . We denote by $\hat{p}_X(s^m)$ the maximum likelihood estimator for the pmf estimation problem, which coincides with the empirical counts in s^m , i.e. $\hat{p}_X(x; s^m) = \frac{T_x}{m}$.

In the minimax setting [1], we assume that nature adversarially chooses a distribution p_X ; n samples $u^n = \{x_i\}_{i=1}^n$ are drawn i.i.d. from this distribution; and our goal is to design an estimator $\hat{q} : \mathcal{X}^n \rightarrow \Delta_{\mathcal{X}}$ based on the samples u^n to minimize the expected loss. We denote the associated risk by $r_n^{\mathcal{L}}$, and the problem is formulated as:

$$r_n^{\mathcal{L}} \triangleq \min_{\hat{q}_X} \max_{p_X \in \Delta_{\mathcal{X}}} \mathbb{E}_{U^n} [\mathcal{L}(p_X, \hat{q}_X(U^n))] \quad (1)$$

The choices of l_2^2 , KL divergence, l_1 and general f -divergences for $\mathcal{L}$ has been considered in the prior work [2], [4]–[7]. In our case we chose $\mathcal{L}$ to be the general l_p^p norms for $p \geq 2$, namely for $p, q \in \Delta_{\mathcal{X}}$:

$$\mathcal{L}(p, q) = \|p - q\|_p^p \triangleq \sum_{x \in \mathcal{X}} \left(|p(x) - q(x)| \right)^p \quad (2)$$

and we denote the minimax estimation risk by r_n^p , then (1) becomes:

$$r_n^p \triangleq r_n^{l_p^p} = \min_{\hat{q}_X} \max_{p_X} \mathbb{E}_{U^n} [\|p_X - \hat{q}_X(U^n)\|_p^p] \quad (3)$$

The minimax risk scales as $r_n^p = \Theta(n^{-\frac{p}{2}})$. We defer the reader to the appendix of the full version of the paper for a proof. This rate is achieved by the maximum likelihood estimator. 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 the clarity of the presentation, we will group the parameters $\bar{C}_p, \underline{C}_p$ into C_p and adopt the notation $\simeq$ to denote that $g_n \simeq C_p f_n$ if $\limsup_n g_n/f_n \leq \bar{C}_p$ and $\liminf_n g_n/f_n \geq \underline{C}_p$. Similarly, we say $g_n \lesssim C_p f_n$ if $\liminf_n g_n/f_n \leq \underline{C}_p$ and $\limsup_n g_n/f_n \leq \bar{C}_p$.

This paper studies the extension of (3) in a multivariate semi-supervised setting. We assume that nature chooses a model p_{XY} of a pair of jointly distributed random variables X, Y . The estimator has access to two datasets: a collection of complete $l^m \triangleq \{(x'_i, y'_i)\}_{i=1}^m$ and a collection of incomplete samples $u^n \triangleq \{x_i\}_{i=1}^n$ generated i.i.d from p_{XY} and p_X . Our goal is to design an estimator with minimal expected risk. We formulate this new problem as:

$$R_{m,n}^p \triangleq \min_{\hat{q}_{XY}} \max_{p_{XY}} \mathbb{E}_{U^n, L^m} [\|p_{XY} - \hat{q}_{XY}(U^n, L^m)\|_p^p] \quad (4)$$

Throughout our study of (4), we will consider the following auxiliary problems:

$$R_m^p \triangleq \min_{\hat{q}_{Y|X}} \max_{p_{XY}} \mathbb{E}_{L^m} [\|p_{XY} - p_X \hat{q}_{Y|X}(L^m)\|_p^p]$$

$$\bar{R}_m^p \triangleq \min_{\hat{q}_{Y|X}} \max_{p_X} \mathbb{E}_{L_X^m} \left[\max_{p_{Y|X}} \mathbb{E}_{L_Y^m} [\|p_{XY} - p_X \hat{q}_{Y|X}(L^m)\|_p^p] \right]$$

The problem R_m^p corresponds to the limit of the problem $R_{m,n}^p$ as $n \rightarrow \infty$. 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 $\bar{R}_m^p$, nature has an additional advantage in forming the dataset l^m : it can first observe the realization of the X symbols, i.e. l_X^m , then choose $p_{Y|X}$ from which to generate the Y values and finish the construction of the joint samples.

For a given univariate estimator $\hat{q}_n^*$, we define the conditional estimator $\hat{q}_{Y|X}^{*,m}(l^m)$ as the *conditional composition based on $\hat{q}_n^*$* to be the concatenation of $\hat{q}_{n_x}^*(l_{Y|X=x})$, where n_x is number of samples in $l_{Y|X=x}$, which is equal to $T_x(l_X^m)$. We further define the estimator $\hat{q}_{XY}^{*,m,n}$ as the *joint composition based on $\hat{q}_n^*$* to be the estimator that estimates the conditional distribution $p_{Y|X}$ with the conditional composition estimator and the marginal p_X with the ML estimator, i.e. $\hat{q}_{XY}^{*,m,n}(u^n, l^m) = \hat{p}_X(u^n \cup l_X^m) \hat{q}_{Y|X}^{*,m}(l^m)$.

We say that an estimator $\{\hat{q}_n\}$ is *first order minimax optimal* for the problem r_n^p if $\max_{p_X} \mathbb{E}_{U^n} [\|p_X - \hat{q}_X\|_p^p] = r_n^p + o(r_n^p)$. The same definition carries over to the problems $R_{m,n}^p, R_m^p, \bar{R}_m^p$.

III. RESULTS

A. Main Theorems

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 $\bar{R}_m^p$:

$$\max_{p_X} \mathbb{E}_{L_X^m} \left[\max_{p_{Y|X}} \|p_{XY} - p_X \hat{q}_{Y|X}^{*,m}\|_p^p \right] = \bar{R}_m^p \quad (5)$$

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 :

$$\max_{p_{XY}} \mathbb{E}_{L^m} [\|p_{XY} - p_X \hat{q}_{Y|X}^{*,m}\|_p^p] = R_m^p + o(R_m^p) \quad (6)$$

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

$$|R_{m,n}^p - R_m^p| \leq O\left(m^{-\frac{p-1}{2}} (n)^{-1/2}\right) \quad (7)$$

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 $m = o(n)$.

B. Sketch of the proof

The main result of this paper is given in Theorem 4. To establish Theorem 4, we first show in Theorem 1 that the conditional composition of a minimax optimal estimator for r_n^p is a minimax optimal estimator for $\bar{R}_m^p$. In Theorem 2, we connect the problems $\bar{R}_m^p$ and R_m^p . In particular, we show that when $p \geq 2$ the adversarial distribution of $\bar{R}_m^p$ is δ_x , for which the problem $\bar{R}_m^p$ reduces to R_m^p . Finally in Theorem 3 we show that $R_{m,n}^p = R_m^p + o(m^{-\frac{p}{2}})$ by studying the regime $m = o(n)$.

IV. PROOFS FOR THEOREMS

A. Proof for Theorem 1

We define:

$$f(p_X, \hat{q}_{Y|X}) \triangleq \mathbb{E}_{L_X^m} \left[\max_{p_{Y|X}} \mathbb{E}_{L_Y^m} [\|p_{XY} - p_X \hat{q}_{Y|X}\|_p^p] \right] \quad (8)$$

Let $\hat{q}_{Y|X}^{**} : (\mathcal{X} \times \mathcal{Y})^m \rightarrow (\Delta_{\mathcal{Y}})^{|\mathcal{X}|}$ be an 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 :

$$\min_{\hat{q}_{Y|X}} f(p_X, \hat{q}_{Y|X}) = f(p_X, \hat{q}_{Y|X}^{**}) \quad (9)$$

since

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

where the first inequality is due to the substitution, the second inequality is the change of min max with max min, and the equality follows from (9). Now let us show (9) holds for the composition estimator $\hat{q}_{Y|X}^{*,m}$. Fix a p_X and for the compactness of notation we define $p_{i,x} \triangleq \mathbb{P}(T_x(L_X) = i)$. Then the left-hand side of (9) becomes:

$$\begin{aligned} &= \min_{\hat{q}_{Y|X}} \mathbb{E}_{L_X^m} \left[\max_{p_{Y|X}} \mathbb{E}_{L_Y^m} \left[\sum_{x \in \mathcal{X}} (p_X(x))^p \|p_{Y|X=x} - \hat{q}_{Y|X=x}\|_p^p \right] \right] \\ &= \sum_{x \in \mathcal{X}} p_X^p(x) \min_{\hat{q}_{Y|X}} \mathbb{E}_{L_X^m} \left[\max_{p_{Y|X=x}} \mathbb{E}_{L_Y^m} [\|p_{Y|X=x} - \hat{q}_{Y|X=x}\|_p^p] \right] \quad (10) \end{aligned}$$

$$= \sum_{x \in \mathcal{X}} (p_X(x))^p \sum_{i=0}^m p_{i,x} r_i^p \quad (11)$$

In (10) we note that the optimization variables are independent. Steps leading (10) to (11) are given below and these steps demonstrate that (9) holds for the estimator $\hat{q}_{Y|X}^{*,m}$:

$$\begin{aligned} & \min_{\hat{q}_{Y|X=x}} \mathbb{E}_{L_X^m} \left[\max_{p_{Y|X=x}} \mathbb{E}_{L_Y^m} [\|p_{Y|X=x} - \hat{q}_{Y|X=x}\|_p^p] \right] \\ &= \min_{\hat{q}_{Y|X=x}} \sum_{i=0}^m p_{i,x} \max_{p_{Y|X=x}} \mathbb{E}_{L_Y^i} [\|p_{Y|X=x} - \hat{q}_{Y|X=x}^i\|_p^p] \\ &= \sum_{i=0}^m p_{i,x} \min_{\hat{q}_{Y|X=x}^i} \max_{p_{Y|X=x}} \mathbb{E}_{L_Y^i} [\|p_{Y|X=x} - \hat{q}_{Y|X=x}^i\|_p^p] \quad (12) \\ &= \sum_{i=0}^m p_{i,x} \max_{p_{Y|X=x}} \mathbb{E}_{L_Y^i} [\|p_{Y|X=x} - \hat{q}_{Y|X=x}^{*,i}\|_p^p] \quad (13) \end{aligned}$$

To establish (13), we observe in (12) that the expression $\min_{\hat{q}_{Y|X=x}^i} \max_{p_{Y|X=x}} \mathbb{E}_{L_Y^i} [\|p_{Y|X=x} - \hat{q}_{Y|X=x}^i\|_p^p]$ is the problem r_i^p and hence is achieved by $\hat{q}_{Y|X}^{*,i}$.

B. Proof for Theorem 2

We start from (11) by recaliling the definition of $p_{i,x}$:

$$\bar{R}_m^p = \max_{p_X} \sum_{x \in \mathcal{X}} \sum_{i=0}^m \binom{m}{i} (p_X(x))^{i+p} (1 - p_X(x))^{m-i} r_i^p \quad (14)$$

$$\simeq \max_{p_X} \sum_{x \in \mathcal{X}} C_p \left(\frac{p_X(x)}{m} \right)^{\frac{p}{2}} + o(m^{-\frac{p}{2}}) \quad (15)$$

$$= \frac{C_p}{m^{\frac{p}{2}}} + o(m^{-\frac{p}{2}}) \quad (16)$$

(15) follows from Lemma 1. In (15), we note that for $p \geq 2$ the problem is convex and symmetric in variables $\{p_X(x)\}_{x \in \mathcal{X}}$ in the first order. Therefore the optimizer is a vertex of the probability simplex, which leads to (16). We obtain the matching lower bound by substituting $p_X = \delta_x$ for some $x \in \mathcal{X}$. We carry out the steps for completeness below:

$$\begin{aligned} R_m^p &\geq \min_{\hat{q}_{Y|X}} \max_{p_{Y|X}} \mathbb{E}_{L \sim \delta_x p_{Y|X}} [\|\delta_x p_{Y|X} - \hat{q}_{Y|X}\|_p^p] \\ &= \min_{\hat{q}_{Y|X}} \max_{p_{Y|X}} \mathbb{E}_{L_Y \sim p_{Y|X=x}} [\|p_{Y|X=x} - \hat{q}_{Y|X=x}\|_p^p] \\ &= \min_{\hat{q}_{Y|X=x}} \max_{p_{Y|X=x}} \mathbb{E}_{L_Y \sim p_{Y|X=x}} [\|p_{Y|X=x} - \hat{q}_{Y|X=x}\|_p^p] = r_m^p \end{aligned}$$

Therefore by $R_m^p \leq \bar{R}_m^p$ we obtain:

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

finally we use Theorem 1.

C. Proof for Theorem 3

By Lemma 3

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

We note that $r_{m+n}^p = \Theta(n+m)^{-\frac{p}{2}}$ and $R_m^p = \Theta(m^{-\frac{p}{2}})$, as proven in the full version of the paper. 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 . An alternative proof for the lower bound $R_m^p \leq R_{m,n}^p$ is also provided in the full version of the paper.

D. Proof for Theorem 4

By Theorem 2, the composition estimator $\hat{q}_{Y|X}^{*,m}$ is first order minimax optimal for R_m^p when $p \geq 2$. Finally by Theorem 3, $R_{m,n}^p$ and R_m^p has the same first order in the regime $m = o(n)$.

V. SUPPLEMENTARY RESULTS

For Lemma 1 and Lemma 2 we introduce:

$$H_p^n(x) \triangleq \sum_{i=0}^n \binom{n}{i} r_i^p x^{i+p} (1-x)^{n-i} \quad (17)$$

$$G_p^n(x) \triangleq \sum_{i=0}^n \binom{n}{i} r_i^p x^i \left(\frac{i}{n}\right)^p (1-x)^{n-i} \quad (18)$$

Lemma 1:

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

Proof: We fix the constant c given in Lemma 2. There are two cases:

In the first case $x \geq c \frac{\log^2(n)}{n}$:

$$H_p^n(x) \simeq \sum_{i=0}^n \binom{n}{i} \frac{C_p}{i^{\frac{p}{2}+1}} x^{i+p} (1-x)^{n-i} \quad (19)$$

$$= \frac{C_p}{n^{\frac{p}{2}}} \sum_{i=0}^n \binom{n}{i} \left(\frac{i}{n}\right)^{\frac{p}{2}} x^i (1-x)^{n-i} + O\left(\frac{H_p^n(x)}{\sqrt{\log n}}\right) \quad (20)$$

$$= \frac{C_p}{n^{\frac{p}{2}}} \left(x^{\frac{p}{2}} + O(n^{-1})\right) + O\left(\frac{H_p^n(x)}{\sqrt{\log n}}\right) \quad (21)$$

$$= C_p \left(\frac{x}{n}\right)^{\frac{p}{2}} + o\left(n^{-\frac{p}{2}}\right) \quad (22)$$

(19) holds since $r_n^p \simeq C_p n^{-\frac{p}{2}}$ whereas in (20) we use Lemma 2. To obtain (21), we utilize Theorem-3.1 in [11] as follows: we set $f(x) = x^{\frac{p}{2}}$ and bound the error of n th order Bernstein polynomial approximation B_n as:

$$\begin{aligned} |B_n(x; f) - f(x)| &= n^{-1} x(1-x) f''(x)/2 + o(n^{-1}) \\ &= n^{-1} p/4(p/2-1) x^{\frac{p}{2}-2} + o(n^{-1}) \\ &\leq n^{-1} p/4(p/2-1) + o(n^{-1}) \quad (23) \end{aligned}$$

where (23) follows since $p \geq 2$. Therefore we conclude that convergence is uniform with error $O(n^{-1})$ for all $x \in (0, 1)$. Finally, (21) implies that $H_p^n(x) = O(n^{-\frac{p}{2}})$ and in (22) we substitute this in the error term of (21).

For the second case we have $x \leq c \frac{\log^2(n)}{n}$:

$$H_p^n(x) = \sum_{i=0}^n \binom{n}{i} r_i^p x^{i+p} (1-x)^{n-i} \quad (24)$$

$$\leq C_p x^p \sum_{i=0}^n \binom{n}{i} \frac{1}{i^{\frac{p}{2}+1}} x^i (1-x)^{n-i} \quad (25)$$

$$\leq C_p x^p = O\left(\frac{\log^{2p}(n)}{n^p}\right) \quad (26)$$

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 $\left|C_p \left(\frac{x}{n}\right)^{\frac{p}{2}} - H_p^n(x)\right| = o(n^{-\frac{p}{2}})$. ■

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

$$|H_p^n(x) - G_p^n(x)| = O\left(\frac{H_p^n(x)}{\sqrt{\log n}}\right) \quad (27)$$

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:

$$|H_p^n(x) - G_p^n(x)| \leq \sum_{i=0}^n \binom{n}{i} x^i (1-x)^{n-i} r_i^p \Delta_p(x, i) \quad (28)$$

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 $x^p - \left(\frac{i}{n}\right)^p = (x - \frac{i}{n})\xi^{p-1}$, thus:

$$\Delta_p(x, i) = \left|x^p - \left(\frac{i}{n}\right)^p\right| \leq p \left|x - \frac{i}{n}\right| \left|x \vee \frac{i}{n}\right|^{p-1} \quad (29)$$

Now we analyze the sum in (28) over the ranges $i \leq nx$, $nx > i$ separately. For the case $i \leq nx$:

$$\sum_{i \leq nx} \binom{n}{i} x^i (1-x)^{n-i} r_i^p \Delta_p(x, i) \quad (30)$$

$$= \sum_{i \leq nx - \delta_1} \binom{n}{i} x^i (1-x)^{n-i} r_i^p \Delta_p(x, i) \quad (31)$$

$$+ \sum_{nx - \delta_1 < i \leq nx} \binom{n}{i} x^i (1-x)^{n-i} r_i^p \Delta_p(x, i) \quad (32)$$

$$\leq \sum_{i \leq nx - \delta_1} \binom{n}{i} x^i (1-x)^{n-i} 2 \quad (33)$$

$$+ \sum_{nx - \delta_1 \leq i \leq nx} \binom{n}{i} x^i (1-x)^{n-i} r_i^p p \left|x - \frac{i}{n}\right| x^{p-1} \quad (34)$$

$$\leq 2e^{-\frac{n}{x}\delta_1^2} + \sum_{nx - \delta_1 \leq i \leq nx} p \binom{n}{i} x^i (1-x)^{n-i} r_i^p \delta_1 x^{p-1} \quad (35)$$

$$= 2e^{-\frac{n}{x}\delta_1^2} + \frac{p}{x} \delta_1 H_p^n(x) \quad (36)$$

In (31) we use (29). In (32), we bound the lower tail of the binomial via lemma-2 in [12] 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 (32), $r_i^p \leq 2$ 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:

$$(34) \leq 2 \left(\frac{H_p^n(x)}{x}\right)^{c_1^2} + p c_1 \sqrt{-\frac{1}{nx} \log\left(\frac{1}{x} H_p^n(x)\right)} H_p^n(x) \quad (37)$$

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

$$\begin{aligned} \frac{1}{x} H_p^n(x) &\simeq C_p \sum_{i=0}^n \frac{x^{p-1}}{i^{\frac{p}{2}} + 1} \binom{n}{i} x^i (1-x)^{n-i} \\ &\geq C_p \frac{1}{n^{\frac{p}{2}} + 1} \left(\frac{c \log^2 n}{n}\right)^{p-1} \sum_{i=0}^n \binom{n}{i} x^i (1-x)^{n-i} \\ &= C_p \frac{1}{n^{\frac{p}{2}} + 1} \left(\frac{c \log^2 n}{n}\right)^{p-1} \geq \frac{k'}{n^{\frac{3}{2}p}} \end{aligned} \quad (38)$$

$$= C_p \frac{1}{n^{\frac{p}{2}} + 1} \left(\frac{c \log^2 n}{n}\right)^{p-1} \geq \frac{k'}{n^{\frac{3}{2}p}} \quad (39)$$

In (35), we notice $x \geq \frac{c \log^2 n}{n}$. On the right-hand side of (36) we collect the constants in k' . Therefore we establish that:

$$\sqrt{-\frac{1}{nx} \log\left(\frac{1}{x} H_p^n(x)\right)} \leq \sqrt{k'' \frac{1}{nx} \log n} \leq O\left(\frac{1}{\log(n)}\right) \quad (40)$$

where in the first inequality we use that $x \geq c \frac{\log^2 n}{n}$ 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:

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

Therefore by choosing c_1 large enough we ensure that $B\left(\frac{H_p^n(x)}{x}\right)^{c_1^2} = o(n^{-\frac{p}{2}})$. Whereas in the second case $i > nx$:

$$\begin{aligned} \sum_{nx < i \leq nx + \delta_2} \binom{n}{i} x^i (1-x)^{n-i} r_i^p \Delta_p(x, i) \\ + \sum_{i > nx + \delta_2} \binom{n}{i} x^i (1-x)^{n-i} r_i^p \Delta_p(x, i) \end{aligned} \quad (41)$$

$$\leq \sum_{nx < i \leq nx + \delta_2} \binom{n}{i} x^i (1-x)^{n-i} \left|x - \frac{i}{n}\right| \left(\frac{i}{n}\right)^{p-1} + 2 \sum_{i > nx + \delta_2} \binom{n}{i} x^i (1-x)^{n-i} \quad (42)$$

$$\leq \delta_2 \sum_{nx < i \leq nx + \delta_2} \binom{n}{i} x^i (1-x)^{n-i} (x + \delta_2)^{p-1} + 2e^{-\frac{n\delta_2^2}{2(x+\frac{\delta_2}{3})}} \quad (43)$$

$$\leq \delta_2 2^{p-1} \sum_{nx < i \leq nx + \delta_2} \binom{n}{i} x^i (1-x)^{n-i} (x \vee \delta_2)^{p-1} + e^{-\frac{n\delta_2^2}{\frac{2}{3}(x \vee \delta_2)}} \quad (44)$$

$$\leq \delta_2 2^{p-1} \sum_{nx < i \leq nx + \delta_2} \binom{n}{i} x^i (1-x)^{n-i} (x)^{p-1} + 2e^{-\frac{n\delta_2^2}{\frac{2}{3}x}} \quad (45)$$

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 lemma-2 of [12]. In (40), we see that the problem is identical to (34) 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 (36) 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 (40) is valid. ■

Lemma 3: For $p \geq 0$, there exists constants $\{c_i\}_{i=0}^p$ and c', c'' such that:

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

$$\text{with } \gamma_{m,n}^p = \sum_{i=0}^{\lfloor p \rfloor - 1} c_i (R_m^p)^{\frac{p-i}{p}} (r_{m+n}^p)^{\frac{i}{p}} + c' (R_m^p)^{\frac{p-\lfloor p \rfloor}{p}} (r_{m+n}^p)^{\frac{\lfloor p \rfloor}{p}} + c'' (R_m^p + r_{m+n}^p)^{\frac{p-\lfloor p \rfloor}{p}} (r_{m+n}^p)^{\frac{\lfloor p \rfloor}{p}}$$

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

$$\Gamma_{x,y}^{U,L}(u) \triangleq p_{XY}(x, y) - u \hat{q}_{Y|X}(y | x) \quad (46)$$

with derivatives:

$$\begin{aligned} \left| \frac{d^i}{du^i} |\Gamma_{x,y}^{U,L}(u)|^p \right| &= p^{(i)} (\hat{q}_{Y|X}(y|x))^i |\Gamma_{x,y}^{U,L}(u)|^{p-i} \\ &\leq p^{(i)} |\Gamma_{x,y}^{U,L}(u)|^{p-i} \end{aligned} \quad (42)$$

Continuing, we have:

$$\begin{aligned} &|\Gamma_{x,y}^{U,L}(\hat{q}_X(x))|^p \\ &\leq \sum_{i=0}^{\lfloor p \rfloor - 1} \frac{p^{(i)}}{i!} |\Gamma_{x,y}^{U,L}(p_X(x))|^{p-i} |\hat{q}_X(x) - p_X(x)|^i \\ &+ \left(|\Gamma_{x,y}^{U,L}(p_X(x))|^{p-\lfloor p \rfloor} \vee |\Gamma_{x,y}^{U,L}(\hat{q}_X(x))|^{p-\lfloor p \rfloor} \right) \end{aligned} \quad (43)$$

$$\begin{aligned} &|\hat{q}_X(x) - p_X(x)|^{\lfloor p \rfloor} \frac{p^{\lfloor p \rfloor}}{\lfloor p \rfloor!} \\ &\leq \sum_{i=0}^{\lfloor p \rfloor - 1} \frac{p^{(i)}}{i!} |\Gamma_{x,y}^{U,L}(p_X(x))|^{p-i} |\hat{q}_X(x) - p_X(x)|^i \\ &+ \frac{p^{\lfloor p \rfloor}}{\lfloor p \rfloor!} |\Gamma_{x,y}^{U,L}(p_X(x))|^{p-\lfloor p \rfloor} |\hat{q}_X(x) - p_X(x)|^{\lfloor p \rfloor} \\ &+ \frac{p^{\lfloor p \rfloor}}{\lfloor p \rfloor!} |\Gamma_{x,y}^{U,L}(\hat{q}_X(x))|^{p-\lfloor p \rfloor} |\hat{q}_X(x) - p_X(x)|^{\lfloor p \rfloor} \end{aligned} \quad (44)$$

$$\begin{aligned} \mathbb{E}_{U,L}[\|p_{XY} - \hat{q}_{XY}\|_p^p] &\leq \mathbb{E}_{U,L}[\|p_{XY} - p_X \hat{q}_{Y|X}\|_p^p] \\ &+ \sum_{i=1}^{\lfloor p \rfloor - 1} \frac{p^{(i)}}{i!} \mathbb{E}_{U,L} \left[\sum_{x,y} |\Gamma_{x,y}^{U,L}(p_X(x))|^{p-i} |\hat{q}_X(x) - p_X(x)|^i \right] \\ &+ \frac{p^{\lfloor p \rfloor}}{\lfloor p \rfloor!} \mathbb{E}_{U,L} \left[\sum_{x,y} |\Gamma_{x,y}^{U,L}(p_X(x))|^{p-\lfloor p \rfloor} |\hat{q}_X(x) - p_X(x)|^{p-\lfloor p \rfloor} \right] \\ &+ \frac{p^{\lfloor p \rfloor}}{\lfloor p \rfloor!} \mathbb{E}_{U,L} \left[\sum_{x,y} |\Gamma_{x,y}^{U,L}(\hat{q}_X(x))|^{p-\lfloor p \rfloor} |\hat{q}_X(x) - p_X(x)|^{p-\lfloor p \rfloor} \right] \end{aligned} \quad (45)$$

$$\begin{aligned} \mathbb{E}_{U,L}[\|p_{XY} - \hat{q}_{XY}\|_p^p] &\leq \mathbb{E}_{U,L}[\|p_{XY} - p_X \hat{q}_{Y|X}\|_p^p] \\ &+ \sum_{i=1}^{\lfloor p \rfloor - 1} \frac{p^{(i)} k_y^{\frac{i}{p}}}{i!} \mathbb{E}[\|p_{XY} - p_X \hat{q}_{Y|X}\|_p^p]^{\frac{p-i}{p}} \mathbb{E}[\|\hat{q}_X - p_X\|_p^p]^{\frac{i}{p}} \\ &+ \frac{p^{\lfloor p \rfloor} k_y^{\frac{\lfloor p \rfloor}{p}}}{\lfloor p \rfloor!} \mathbb{E}[\|p_{XY} - p_X \hat{q}_{Y|X}\|_p^p]^{\frac{p-\lfloor p \rfloor}{p}} \mathbb{E}[\|\hat{q}_X - p_X\|_p^p]^{\frac{\lfloor p \rfloor}{p}} \\ &+ \frac{p^{\lfloor p \rfloor} k_y^{\frac{\lfloor p \rfloor}{p}}}{\lfloor p \rfloor!} \mathbb{E}[\|p_{XY} - \hat{q}_X \hat{q}_{Y|X}\|_p^p]^{\frac{p-\lfloor p \rfloor}{p}} \mathbb{E}[\|\hat{q}_X - p_X\|_p^p]^{\frac{\lfloor p \rfloor}{p}} \end{aligned} \quad (46)$$

$$\triangleq \kappa_{m,n}(p_{XY}, \hat{q}_{XY})$$

In (43) we Taylor expand the $h(\hat{q}_X(x)) \triangleq |\Gamma_{x,y}^{U,L}(\hat{q}_X(x))|^p$ around the $p_X(x)$ and use the upper bound (42) for the derivatives. We bound the remainder of the Taylor expansion with the mean value theorem via the monotonicity of the derivatives as in (29), leading to (44). In (46) we sum over x, y and

take expectations with respect to $U, L \sim p_{XY}$ of both hand-sides (43), (45) by noting that $\mathbb{E}_{U,L}[\sum_{x,y} |\Gamma_{x,y}^{U,L}(\hat{q}(X))|^p] = \mathbb{E}_{U,L}[\|p_{XY} - \hat{q}_{XY}\|_p^p]$. To obtain (47), we apply Hölder's inequality to each summation inside the expectations in (46). 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 $\hat{q}_{XY}$:

$$\begin{aligned} R_{m,n}^p &= \min_{\hat{q}_{XY}} \max_{p_{XY}} \mathbb{E}_{U,L}[\|p_{XY} - \hat{q}_{XY}\|_p^p] \leq \max_{p_{XY}} \kappa_{m,n}(p_{XY}, \hat{q}_{XY}) \quad (48) \\ &\leq \max_{p_{XY}} \mathbb{E}_{U,L}[\|p_{XY} - p_X \hat{q}_{Y|X}\|_p^p] + \\ &\sum_{i=1}^{\lfloor p \rfloor - 1} c_i \max_{p_{XY}} \mathbb{E}[\|p_{XY} - p_X \hat{q}_{Y|X}\|_p^p]^{\frac{p-i}{p}} \max_{p_X} \mathbb{E}[\|\hat{q}_X - p_X\|_p^p]^{\frac{i}{p}} \\ &+ c' \max_{p_{XY}} \mathbb{E}[\|p_{XY} - p_X \hat{q}_{Y|X}\|_p^p]^{\frac{p-\lfloor p \rfloor}{p}} \max_{p_X} \mathbb{E}[\|\hat{q}_X - p_X\|_p^p]^{\frac{\lfloor p \rfloor}{p}} \\ &+ c''' \max_{p_{XY}} \mathbb{E}[\|p_{XY} - \hat{q}_X \hat{q}_{Y|X}\|_p^p]^{\frac{p-\lfloor p \rfloor}{p}} \max_{p_X} \mathbb{E}[\|\hat{q}_X - p_X\|_p^p]^{\frac{\lfloor p \rfloor}{p}} \end{aligned} \quad (49)$$

$$\begin{aligned} R_{m,n}^p &\leq \max_{p_{XY}} \mathbb{E}_{U,L}[\|p_{XY} - p_X \hat{q}_{Y|X}\|_p^p] + \\ &\sum_{i=1}^{\lfloor p \rfloor - 1} c_i \max_{p_{XY}} \mathbb{E}[\|p_{XY} - p_X \hat{q}_{Y|X}\|_p^p]^{\frac{p-i}{p}} \max_{p_X} \mathbb{E}[\|\hat{q}_X - p_X\|_p^p]^{\frac{i}{p}} \\ &+ c' \max_{p_{XY}} \mathbb{E}[\|p_{XY} - p_X \hat{q}_{Y|X}\|_p^p]^{\frac{p-\lfloor p \rfloor}{p}} \max_{p_X} \mathbb{E}[\|\hat{q}_X - p_X\|_p^p]^{\frac{\lfloor p \rfloor}{p}} \\ &+ c'' \max_{p_{XY}} \mathbb{E}[\|p_{XY} - p_X \hat{q}_{Y|X}\|_p^p]^{\frac{p-\lfloor p \rfloor}{p}} \max_{p_X} \mathbb{E}[\|\hat{q}_X - p_X\|_p^p]^{\frac{\lfloor p \rfloor}{p}} \\ &+ c'' \max_{p_X} \mathbb{E}[\|p_X - \hat{q}_X\|_p^p] \end{aligned} \quad (50)$$

In (49) we further upper bound (48) by taking the maximum of each summation separately. We define the constants are $c_i \triangleq p^{(i)} k_y^{\frac{i}{p}} / i!$, $c' = c''' \triangleq p^{(\lfloor p \rfloor)} k_y^{\frac{\lfloor p \rfloor}{p}} / \lfloor p \rfloor!$ based on (47). In (50) we note that:

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

which is a consequence of convexity of $|x|^p$ for $p \geq 1$. Finally, we choose $\hat{q}_X, \hat{q}_{Y|X}$ to be the minimax estimators of the r_m^p and R_m^p respectively to establish the Lemma 3. ■

VI. CONCLUSION

In this work, we considered the problem of minimax pmf estimation problem under l_p^p loss when there are m labeled and n unlabeled samples. In particular, we showed that for $p \geq 2$, the composition estimators of univariate minimax problems are optimal in the first order over the regime $m = o(n)$. Extending the results to $1 \leq p \leq 2$ and for f -divergences are possible future directions.

VII. ACKNOWLEDGEMENT

This research is supported by ONR Grant N00014-19-1-2621.

REFERENCES

- [1] A. Wald, "Statistical decision functions," *The Annals of Mathematical Statistics*, vol. 20, no. 2, pp. 165–205, 1949. [Online]. Available: <http://www.jstor.org/stable/2236853>
- [2] S. Trybula, "Some problems of simultaneous minimax estimation," *The Annals of Mathematical Statistics*, vol. 29, no. 1, pp. 245–253, 1958.
- [3] I. Olkin and M. Sobel, "Admissible and minimax estimation for the multinomial distribution and for k independent binomial distributions," *The Annals of Statistics*, vol. 7, no. 2, Mar. 1979. [Online]. Available: <https://doi.org/10.1214/aos/1176344613>
- [4] M. Wilczyński, "Minimax estimation for the multinomial and multivariate hypergeometric distributions," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 128–132, 1985.
- [5] D. Braess and T. Sauer, "Bernstein polynomials and learning theory," *Journal of Approximation Theory*, vol. 128, no. 2, pp. 187–206, 2004.
- [6] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of discrete distributions under l_1 loss," *CoRR*, vol. abs/1411.1467, 2014. [Online]. Available: <http://arxiv.org/abs/1411.1467>
- [7] S. Kamath, A. Orlitsky, D. Pichapati, and A. T. Suresh, "On learning distributions from their samples," in *Proceedings of The 28th Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, P. Grünwald, E. Hazan, and S. Kale, Eds., vol. 40. Paris, France: PMLR, 03–06 Jul 2015, pp. 1066–1100. [Online]. Available: <https://proceedings.mlr.press/v40/Kamath15.html>
- [8] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.
- [9] K. He, "An ancillarity paradox in the estimation of multinomial probabilities," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 824–828, Sep. 1990. [Online]. Available: <https://doi.org/10.1080/01621459.1990.10474946>
- [10] A. Kirichenko and P. Grünwald, "Minimax rates without the fixed sample size assumption," 2020. [Online]. Available: <https://arxiv.org/abs/2006.11170>
- [11] R. A. DeVore and G. G. Lorentz, *Constructive approximation*. Springer Science & Business Media, 1993, vol. 303.
- [12] F. Chung and L. Lu, "Connected components in random graphs with given expected degree sequences," *Annals of Combinatorics*, vol. 6, no. 2, pp. 125–145, Nov. 2002. [Online]. Available: <https://doi.org/10.1007/pl00012580>