# Maximum Likelihood Estimation for Learning Populations of Parameters

**Ramya Korlakai Vinayak** [1] **Weihao Kong** [2] **Gregory Valiant** [2] **Sham Kakade** [1]

## Abstract

Consider a setting with $N$ independent individuals, each with an unknown parameter, $p_i \in [0, 1]$ drawn from some unknown distribution $P^\star$. After observing the outcomes of $t$ independent Bernoulli trials, i.e., $X_i \sim \text{Binomial}(t, p_i)$ per individual, our objective is to accurately estimate $P^\star$. This problem arises in numerous domains, including the social sciences, psychology, healthcare, and biology, where the size of the population under study is usually large while the number of observations per individual is often limited. Our main result shows that, in the regime where $t \ll N$, the maximum likelihood estimator (MLE) is both statistically minimax optimal and efficiently computable. Precisely, for sufficiently large $N$, the MLE achieves the information theoretic optimal error bound of $\mathcal{O}(\frac{1}{t})$ for $t < c \log N$, with regards to the earth mover's distance (between the estimated and true distributions). More generally, in an exponentially large interval of $t$ beyond $c \log N$, the MLE achieves the minimax error bound of $\mathcal{O}(\frac{1}{\sqrt{t \log N}})$. In contrast, regardless of how large $N$ is, the naive "plug-in" estimator for this problem only achieves the sub-optimal error of $\Theta(\frac{1}{\sqrt{t}})$.

## 1. Introduction

The problem of learning a distribution of parameters over a population arises in several domains such as social sciences, psychology, medicine, and biology (Lord, 1965; Lord & Cressie, 1975; Millar, 1986; Palmer & Dixon, 1990; Colwell & Coddington, 1994; Bell et al., 2000). While the number of individuals in the population can be very large, the number of observations available per individual is often very limited, which prohibits accurate estimation of the parameter of interest per individual. In such sparse observation scenarios, *how accurately can we estimate the distribution of parameters over the population?*

In the 1960's F. M. Lord studied the problem of estimating the distribution of parameters over a population in the context of psychological testing (Lord, 1965; 1969). Consider a study involving a large number of independent individuals. Each individual has an unknown probability $p_i$ of answering a question correctly. Given the scores of these individuals on a test with small set of questions, the goal is to estimate the underlying distribution of $p_i$'s. Such an estimated distribution can be used in downstream tasks, like testing if the distribution of scores is uniform or multimodal, or comparing two tests of the same psychological trait.

We use the lens of sparse regime analysis for this problem of learning a population of parameters. Our analysis is inspired by the recent advances in a related problem of estimating discrete distributions and their properties such as, entropy and support size, when the number of observations is much smaller than the support size of the distribution (Valiant & Valiant, 2013; Jiao et al., 2015; Wu & Yang, 2015; 2016; Orlitsky et al., 2016; Acharya et al., 2017; Jiao et al., 2018; Han et al., 2018). However, we note that our setting is not the same as estimating a discrete distribution. For instance, the probabilities sum to 1 for a discrete distribution, where as, the true parameters in our setting need not sum to 1.

There have been several classical works on non-parametric mixture models in general (Turnbull, 1976; Simar, 1976; Laird, 1978; Lindsay, 1983a;b; Böhning, 1989; Lesperance & Kalbfleisch, 1992) and binomial mixture models in particular (Cressie, 1979; Wood, 1999) which have studied the geometry of the maximum likelihood estimator (MLE), the optimality conditions, identifiability, and uniqueness of the MLE solution, and algorithms for computing the optimal solution to the MLE. However, the statistical analysis of how accurately the MLE recovers the underlying distribution has not been addressed. In this paper, we fill this gap, and show that MLE achieves the optimal error bound with regards to the earth mover's distance (or Wasserstein-1 distance, Definition 3.1) between the estimated and true distributions (equivalently, the $l_1$-distance between the CDF's).

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle [2]Department of Computer Science, Stanford University, Stanford. Correspondence to: Ramya Korlakai Vinayak <ramya@cs.washington.edu>.

*Proceedings of the $36^{th}$ International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. 

### 1.1. Problem set-up and summary of results

The setting considered in (Lord, 1969) can be modeled as follows. Consider a set of $N$ independent coins, each with its own unknown bias $p_i \in [0,1]$ drawn independently from some unknown distribution $P^\star$ over $[0,1]$. That is, the probability of seeing a head when coin $i$ is tossed is $p_i$. For each coin $i$, we get to observe the outcome of $t$ independent tosses, denoted by, $X_i \sim \text{Binomial}(t, p_i)$. Our goal is to estimate the unknown distribution $P^\star$ from $\{X_i\}_{i=1}^N$.

The MLE for this problem is as follows:

$$\arg\max_{Q \in \mathcal{D}} \sum_{i=1}^{N} \log \int_0^1 \binom{t}{X_i} y^{X_i}(1-y)^{t-X_i} dQ(y),$$

where $\mathcal{D}$ is set of all distributions on $[0,1]$.

**Our Contribution:** We bound the earth mover's distance (or the Wasserstein-1 distance) between the true distribution $P^\star$ and the MLE solution $\hat{P}_{\text{mle}}$, and show that:

- The MLE achieves an error bound of
$$W_1(P^\star, \hat{P}_{\text{mle}}) = \mathcal{O}_\delta\left(\frac{1}{t}\right),$$ [1]
when $t = \mathcal{O}(\log N)$. For sufficiently large $N$, the bound of $\Theta\left(\frac{1}{t}\right)$ is information theoretically optimal.
- The MLE achieves an error bound of
$$W_1(P^\star, \hat{P}_{\text{mle}}) = \mathcal{O}_\delta\left(\frac{1}{\sqrt{t \log N}}\right),$$
when $t \in \left[\Omega(\log N), \mathcal{O}\left(N^{2/9-\epsilon}\right)\right]$, and this bound is information theoretically optimal in this regime.

Table 1 summarizes our results in comparison to other estimators. While the moment matching estimator (Tian et al., 2017) achieves the same error bound as the MLE when $t = \mathcal{O}(\log N)$, if fails when $t = \Omega(\log N)$ due to high variance in the larger moments. While the local moment matching (Han et al., 2018) could potentially avoid this weakness, it involves hyperparameter tuning which makes it difficult to work with in practice (Remark 3.3 in Section 3). In contrast, the MLE naturally adapts itself and achieves the optimal rates in different regimes without the need for any parameter tuning. We demonstrate that the MLE works well in practice on both synthetic as well as real datasets. Furthermore, our analysis involves bounding the coefficients of Bernstein polynomials approximating Lipschitz-1 functions (Proposition 4.1). We believe that this question is of independent interest with implications to general polynomial approximation theory as well as applications in computer graphics.

## 2. Related Works

Starting from (Lord, 1969), there has been a great deal of interest in the problem of estimating the distribution of true scores of a population of independent entities. Maximum likelihood estimation for non-parametric mixture models has been studied extensively (Lord & Cressie, 1975; Cressie, 1979; Laird, 1978; Turnbull, 1976; Lesperance & Kalbfleisch, 1992). (Lindsay, 1983a) and (Lindsay, 1983b) delineate the geometry of the MLE landscape for non-parametric mixture models in general, and specifically for exponential family respectively. (Wood, 1999) further discusses the issue of uniqueness of the solution for mixture of binomials and the relationship with the moment space. As mentioned in the introduction, the accuracy of the MLE solution for this formulation has not been studied in the literature. Our work fills in this gap by showing that the MLE solution is minimax optimal when $t \ll N$.

*Table 1.* Comparison of results

| Estimators | Bound on EMD |
|---|---|
| Empirical | $\Theta\left(\frac{1}{\sqrt{t}}\right) + \Theta\left(\frac{1}{\sqrt{N}}\right)$ in all regimes |
| Moment Matching (Tian et al., 2017) | • $\Theta\left(\frac{1}{t}\right)$ when $t = \mathcal{O}(\log N)$ • Fails when $t = \Omega(\log N)$ |
| MLE (this paper) | • $\Theta\left(\frac{1}{t}\right)$ when $t = \mathcal{O}(\log N)$ • $\Theta\left(\frac{1}{\sqrt{t \log N}}\right)$, when $t \in \left[\Omega(\log N), \mathcal{O}\left(N^{2/9-\epsilon}\right)\right]$ |

In a recent work (Tian et al., 2017), the authors proposed a *moment matching estimator* to estimate the unknown distribution of the biases in the regime where the number of tosses per coin $t \leq \mathcal{O}(\log N)$. This estimator finds a distribution on $[0,1]$ that closely matches the first $t$ empirical moments of the unknown distribution that can be estimated using the observations. This moment matching estimator incurs $\mathcal{O}\left(\frac{1}{t}\right) + \mathcal{O}_\delta\left(2^t t \sqrt{\frac{\log t}{N}}\right)$. Furthermore, (Tian et al., 2017) also showed that $\Omega\left(\frac{1}{t}\right)$ is a lower bound in this setting. The main weakness of this method of moments approach is that it fails to obtain the optimal rate when $t > c \log N$.

A tangentially related problem is that of estimating a discrete distribution and its symmetric properties[2] like entropy, and support size, when the number of observations is much smaller than the support size of the distribution. This is a well-studied classical problem in statistics (Fisher et al., 1943; Good & Toulmin, 1956; Efron & Thisted, 1976). It has received a lot of interest in the past decade and con-

[1] $\mathcal{O}_\delta(.)$ hides $\log(1/\delta)$ in the bound for it to hold with probability at least $1 - 2\delta$.

[2] A function over a discrete distribution is said to be a *symmetric function* if it remains invariant to the relabeling of the domain symbols.

tinues to be a very active area of research (Paninski, 2003; Orlitsky et al., 2004; Valiant & Valiant, 2011; 2013; Jiao et al., 2015; Wu & Yang, 2015; 2016; Orlitsky et al., 2016). Recent work (Han et al., 2018) used local moment matching to provide bounds on estimating symmetric properties of discrete distributions under the Wasserstein-1 distance. This technique of local moment matching can be used in our setting to improve the bounds obtained in (Tian et al., 2017) in the regime where $t > c \log N$. We discuss this more in Section 3.2. In a similar spirit to our work, a series of works (Acharya et al., 2009; 2010; 2017) examined the *profile* or *pattern* maximum likelihood as a unifying framework for estimating symmetric properties of a discrete distribution. Unlike in our setting, it is computationally challenging to computing the exact maximum likelihood estimator, and the question becomes how to efficiently approximate it (see e.g. (Vontobel, 2012; Charikar et al., 2019)).

## 3. Main Results

Before formally stating our results, we introduce some notation, discuss the MLE objective and define the Wasserstein-1 metric used to measure the accuracy of estimation.

**Notation:** Recall that $N$ is the number of independent coins and $t$ is the number of tosses per coin. The biases of the coins are denoted by $\{p_i\}_{i=1}^N$, where each $p_i \in [0,1]$ is drawn from some unknown distribution $P^\star$ on $[0,1]$. The set of observations is $\{X_i\}_{i=1}^N$, where $X_i \sim \text{Binomial}(t, p_i)$. For $s \in \{0,1,...,t\}$, let $n_s$ denote the number of coins that show $s$ heads out of $t$ tosses. Let $h_s^{\text{obs}}$ denote the fraction of coins that show $s$ heads.

$$n_s := \sum_{i=1}^{N} \mathbf{1}_{\{X_i = s\}}, \; h_s^{\text{obs}} := \frac{n_s}{N}, \tag{1}$$

where $\mathbf{1}_{\mathcal{A}}$ is indicator function for set $\mathcal{A}$. $\mathbf{h}^{\text{obs}} := \{h_0^{\text{obs}}, h_1^{\text{obs}}, ..., h_t^{\text{obs}}\}$ is the observed *fingerprint*. Since the identity of the coins is not important to estimate the distribution of the biases, the observed fingerprint is a sufficient statistics for the estimation problem.

**MLE Objective:** The MLE estimate of the distribution of biases given the observations $\{X_i\}_{i=1}^N$ is,

$$\begin{aligned}\hat{P}_{\text{mle}} &\in \arg\max_{Q \in \mathcal{D}} \sum_{i=1}^{N} \log \int_0^1 \binom{t}{X_i} y^{X_i}(1-y)^{t-X_i} dQ(y) \\ &= \arg\max_{Q \in \mathcal{D}} \sum_{s=0}^{t} n_s \log \underbrace{\int_0^1 \binom{t}{s} y^s (1-y)^{t-s} q(y) dy}_{=: E_Q[h_s]},\end{aligned}$$

where $\mathcal{D}$ is the set of all distributions on $[0,1]$, $n_s$ is the number of coins that that see $s$ heads out of $t$ tosses, and $E_Q[h_s]$ is the expected fraction of the population that sees $s$ heads out of $t$ tosses under the distribution $Q$. Equivalently, the MLE can be written in terms of the fingerprint as follows,

$$\hat{P}_{\text{mle}} \in \arg\max_{Q \in \mathcal{D}} \sum_{s=0}^{t} h_s^{\text{obs}} \log E_Q[h_s], \tag{2}$$

$$= \arg\min_{Q \in \mathcal{D}} \text{KL}\left(\mathbf{h}^{\text{obs}}, E_Q[\mathbf{h}]\right), \tag{3}$$

where $\text{KL}(A\ , B)$ is the Kullback-Leibler divergence[3] between distributions $A$ and $B$, $\mathbf{h}^{\text{obs}}$ is the observed fingerprint vector and $E_Q[\mathbf{h}]$ denotes the expected fingerprint vector when the biases are drawn from distribution $Q$.

**Remark 3.1.** The set $\mathcal{D}$ of all distributions over $[0,1]$ is convex. Furthermore, the objective function of the MLE (Equation 3) is convex in $Q$ and strictly convex in the valid fingerprints, $\{E_Q[h_s]\}_{s=0}^t$. While there is a unique $E_{\hat{P}_{\text{mle}}}[\mathbf{h}]$ that minimizes the objective (3), there can be many distributions $Q^\star \in \mathcal{D}$ that can give rise to the optimal expected fingerprint. Moreover, while the fingerprint vector $\mathbf{h}$ lives in $\Delta^t$, the $t$-dimensional simplex in $\mathbb{R}^{t+1}$, not all vectors in $\Delta^t$ can be valid fingerprints. The set of all valid fingerprints is a small convex subset of $\Delta^t$. Very often $\mathbf{h}^{\text{obs}}$ falls outside the set of valid fingerprints and the solution to the MLE is the closest projection under the KL divergence onto the valid fingerprint set. Furthermore, the fingerprints are related to moments via a linear transform. The geometry of the set of valid fingerprints therefore can also be described using moments. For more details on this geometric description we refer the reader to (Wood, 1999).

**Wasserstein-1 Distance:** We measure the accuracy of our estimator using the Wasserstein-1 distance or the earth mover's distance (EMD) between two probability distributions over the interval $[0,1]$ which is defined as:

**Definition 3.1** (Wasserstein-1 or earth mover's distance)**.**

$$W_1(P,Q) := \inf_{\gamma \in \Gamma(P,Q)} \int_{x=0}^{1} \int_{y=0}^{1} |x-y| \, d\gamma(x,y), \tag{4}$$

where $\Gamma(P,Q)$ is a collection of all the joint distributions on $[0,1]^2$ with marginals P and Q. A dual definition due to Kantarovich and Rubinstein (Kantorovich & Rubinstein, 1958) of this metric is as follows:

$$W_1(P,Q) := \sup_{f \in \text{Lip}(1)} \int_0^1 f(x)(p(x) - q(x))dx \tag{5}$$

$$= \sup_{f \in \text{Lip}(1)} \left(\text{E}_P[f] - \text{E}_Q[f]\right), \tag{6}$$

where $p$ and $q$ are the probability density functions of the distributions P and Q respectively, and $\text{Lip}(1)$ denotes the set of Lipschitz-1 functions.

Wassertein-1 distance is a natural choice to measure the accuracy of estimator in our setting. E.g., suppose the true

[3] KL divergence between two discrete distributions A and B supported on $\mathcal{X}$ is defined as $\text{KL}(A,B) = \sum_{x \in \mathcal{X}} A(x) \log \frac{A(x)}{B(x)}$.

distribution $P^\star$ is $\delta(0.5) = 1$. Let $P_1$ with $\delta(0.45) = 1$ and $P_2$ with $\delta(0) = \delta(1) = \frac{1}{2}$ be the output of two estimators. The Wassertein-1 distance, $W_1(P^\star, P_1) = 0.05$ and $W_1(P^\star, P_2) = 0.5$, clearly distinguishes the first estimate to be much better than the second. In contrast, the total variation distance between both $P_1$ and $P_2$ to the truth is 1 and the KL divergence to the truth in both cases is infinite.

### 3.1. Small sample regime

We first focus on the regime where the number of observations per coin, $t = \mathcal{O}(\log N)$. Consider the problem setup in Section 1.1. The following theorem gives a bound on the Wasserstein-1 distance between the MLE (Equation 3) and the true underlying distribution.

**Theorem 3.1** (Small Sample Regime)**.** When $t = \mathcal{O}(\log N)$, the Wassertein-1 distance between an optimal solution to the MLE, denoted by $\hat{P}_{\text{mle}}$ and the true underlying distribution $P^\star$ can be bounded with probability at least $1 - 2\delta$ as follows,

$$W_1(P^\star, \hat{P}_{\text{mle}}) \leq \mathcal{O}_\delta\left(\frac{1}{t}\right). \tag{7}$$

For constant $\delta$, this $O(1/t)$ rate is information theoretically optimal due to the following result (Proposition 1 in (Tian et al., 2017)):

**Proposition 3.1** (Lower Bound (Tian et al., 2017))**.** Let $P$ denote a distribution over $[0, 1]$. Let $\mathbf{X} := \{X_i\}_{i=1}^N$ be random variables with $X_i \sim \text{Binomial}(t, p_i)$ where $p_i$ is drawn independently from $P$. Let $f$ be an estimator that maps $\mathbf{X}$ to a distribution $f(\mathbf{X})$. For every fixed $t$, the following lower bound holds for all $N$:

$$\inf_f \sup_P \mathrm{E}\left[W_1(P, f(\mathbf{X}))\right] > \frac{1}{4t}. \tag{8}$$

### 3.2. Medium sample regime

In this section we consider the regime where the number of observations per coin $t$ is larger than $\Omega(\log N)$. For the same setting as before (Section 1.1), the following theorem provides a bound on the Wasserstein-1 distance between the MLE solution and the true distribution.

**Theorem 3.2** (Medium Sample Regime)**.** There exists $\epsilon > 0$, such that, for $t \in \left[\Omega(\log N), \mathcal{O}\left(N^{2/9-\epsilon}\right)\right]$, with probability at least $1 - 2\delta$,

$$W_1(P^\star, \hat{P}_{\text{mle}}) \leq \mathcal{O}_\delta\left(\frac{1}{\sqrt{t \log N}}\right). \tag{9}$$

**Remark 3.2.** We conjecture that the interval for this bound should be $t \in \left[\Omega(\log N), \mathcal{O}\left(N^{2/3-\epsilon}\right)\right]$. Details on why the sub-optimal bound of $\mathcal{O}\left(N^{2/9-\epsilon}\right)$ arises in our analysis is described in Remark 4.1 and discussion after Lemma 4.2.

We prove a $\Theta(\frac{1}{\sqrt{t \log N}})$ lower bound of the minimax rate for estimating the population of parameters under Wasserstein-1 distance. This lower bound, combining with the $\Theta(\frac{1}{t})$ lower bound shown in (Tian et al., 2017), implies that the MLE is minimax optimal up to a constant factor in both the regimes. The lower bound is formalized in Theorem 3.3.

**Theorem 3.3.** Let $P$ be a distribution over $[0, 1]$. Let $\mathbf{X} := \{X_i\}_{i=1}^N$ be random variables with $X_i \sim \text{Binomial}(t, p_i)$ where $p_i$ is drawn independently from $P$. Let $f$ be an estimator that maps $\mathbf{X}$ to a distribution $f(\mathbf{X})$. For every $t, N$ s.t. $t \leq \frac{N^{2(e^4-1)}}{36}$, the following lower bound holds:

$$\inf_f \sup_P \mathrm{E}\left[W_1(P, f(\mathbf{X}))\right] > \frac{1}{3e^4\sqrt{t \log N}}. \tag{10}$$

**Remark 3.3. Local Moment Matching:** The moment matching estimator in (Tian et al., 2017) fails when $t$ is larger than $\Omega(\log N)$ because the $t$-th order moments cannot be estimated accurately in that regime. This causes the second term in the error bound $\mathcal{O}\left(\frac{1}{t}\right) + \mathcal{O}_\delta\left(2^t t\sqrt{\frac{\log t}{N}}\right)$ to become large. Naturally, one might consider matching only the first $\log N$ moments which can be reliably estimated. In addition, the parameter interval $[0, 1]$ can be split into blocks, and the moment matching can be done in each block locally by utilizing the fact that for large $t$, $X_i/t$ tightly concentrates around $p_i$. The local moment matching was first introduced in a recent work by (Han et al., 2018) in the setting of learning discrete distributions. Potentially, one may apply the local moment matching approach to our setting of learning populations of parameters which will likely yield an algorithm that achieves Wasserstein-1 distance error $\mathcal{O}(\max(\frac{1}{\sqrt{t \log N}}, \frac{1}{t}))$ in the $t \ll N$ regime. The algorithm will degenerate to the one developed in (Tian et al., 2017) in the $t = \mathcal{O}(\log N)$ regime. However, from a practical perspective, the local moment matching algorithm is quite unwieldy. It involves significant parameter tuning and special treatment for the edge cases. Some techniques used in local moment matching, e.g. using a fixed blocks partition of $[0, 1]$ and matching the first $\log N$ for all the blocks, are quite crude and likely lose large constant factors both in theory and in practice. Therefore, we expect the local moment matching to have inferior performance than the MLE approach in practice. We include a brief sketch of how one may apply the local moment matching approach to our setting in the supplementary material.

**Remark 3.4. Empirical Estimator:** The naive "plug-in" estimator for the underlying distribution is the sorted estimates of the biases of the coins. This incurs an error of $\mathcal{O}\left(\frac{1}{\sqrt{t}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ in the earth movers distance (or $l_1-$ distance between the estimated and the true CDFs), where the first term is due to the error in estimating the biases of the coins from $t$ outcomes, and the second term is due to estimating the error in the estimated CDF using $N$

coins. If the number of tosses per coin is very large, that is, $t \gg N$, then we can estimate individual biases pretty well, and obtain empirical CDF that can estimate $P^\star$ incurring overall error rate of $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$. However, in the regime of interest, the number of observation per coin is small, i.e., $t \ll N$ (*sparse regime*). The empirical estimates of the biases in this regime are very crude. Thus, when $t$ is small, even with a very large population (large $N$), the empirical estimator does not perform better on the task of estimating the underlying distribution than on estimating the biases itself which incurs a $\Theta\left(\frac{1}{\sqrt{t}}\right)$.

## 4. Proof Sketches

### 4.1. Bound on Wasserstein-1 distance

Proofs of Theorems 3.1 and 3.2 involve bounding the Wasserstein-1 distance between the true distribution $P^\star$ and the MLE estimate $\hat{P}_{\text{mle}}$. Recall the dual definition of Wasserstein-1 distance or the earth movers distance between two distributions $P$ and $Q$ supported on $[0, 1]$,

$$W_1(P, Q) = \sup_{f \in \text{Lip}(1)} \int_{x=0}^{1} f(x)(p(x) - q(x))dx,$$

where $p$ and $q$ are the probability density functions of the distributions P and Q respectively, and $\text{Lip}(1)$ denotes the set of Lipschitz-1 functions. Any Lipschitz-1 function $f$ on $[0, 1]$ can be approximated using Bernstein polynomials as, $\hat{f}(x) := \sum_{j=0}^{t} b_j \binom{t}{j} x^j (1-x)^{t-j}$. Using this approximation we have,

$$\begin{aligned}
&\int_0^1 f(x)(p(x) - q(x))dx \\
&= \left\{ \int_0^1 \left(f(x) - \hat{f}(x)\right)(p(x) - q(x))dx \right. \\
&\quad \left. + \int_0^1 \hat{f}(x)(p(x) - q(x))dx \right\},
\end{aligned}$$

which can be bounded by,

$$\begin{aligned}
&2||f - \hat{f}||_\infty \\
&+ \int_0^1 \sum_{j=0}^{t} b_j \binom{t}{j} x^j (1-x)^{t-j} (p(x) - q(x))dx \\
&= 2||f - \hat{f}||_\infty + \sum_{j=0}^{t} b_j \left(\text{E}_P[h_j] - \text{E}_Q[h_j]\right), \qquad (11)
\end{aligned}$$

where $||f - \hat{f}||_\infty := \max_{x \in [0,1]} |f(x) - \hat{f}(x)|$ is the approximation error. Therefore, the Wasserstein-1 distance (Definition 5) between the true distribution $P^\star$ and MLE estimate $\hat{P}_{\text{mle}}$ can be bounded as follows,

$$\begin{aligned}
W_1(P, \hat{P}_{\text{mle}}) \leq \sup_{f \in \text{Lip}(1)} \Bigg\{ & \underbrace{2||f - \hat{f}||_\infty}_{(a)} \\
& + \underbrace{\sum_{j=0}^{t} b_j \left(\text{E}_{P^*}[h_j] - h_j^{\text{obs}}\right)}_{(b)} \\
& + \underbrace{\sum_{j=0}^{t} b_j \left(h_j^{\text{obs}} - \text{E}_{\hat{P}_{\text{mle}}}[h_j]\right)}_{(c)} \Bigg\} \qquad (12)
\end{aligned}$$

The first term $(a)$ in the above bound (Equation 12) is the approximation error for using Bernstein polynomials to approximate Lipschitz-1 functions. The second term $(b)$ is the error due to sampling. The third term $(c)$ is the estimation error in matching the fingerprints. We bound the second and third term using the following lemmas.

**Lemma 4.1.** With probability at least $1 - \delta$,

$$\left| \sum_{j=0}^{t} b_j \left(h_j - \text{E}(h_j)\right) \right| \leq \mathcal{O}\left( \max_j |b_j| \sqrt{\frac{\log 1/\delta}{N}} \right) \qquad (13)$$

**Lemma 4.2.** For $3 \leq t \leq \sqrt{C_0 N} + 2$, w. p. $1 - \delta$,

$$\begin{aligned}
&\left| \sum_{j=0}^{t} b_j \left(h_j - \text{E}_{P_{\text{mle}}}(h_j)\right) \right| \\
&\underset{(1)}{\leq} \max_j |b_j| \sum_{j=0}^{t} \left| \left(h_j - \text{E}_{P_{\text{mle}}}(h_j)\right) \right| \\
&\leq \max_j |b_j| \sqrt{2 \ln 2} \sqrt{\frac{t}{2N} \log \frac{4N}{t} + \frac{1}{N} \log \frac{3e}{\delta}}. \qquad (14)
\end{aligned}$$

We prove Lemma 4.1 using McDiarmid's inequality for concentration of fingerprints and Lemma 4.2 using optimality of MLE, Pinsker's inequality and recent results bounding the KL divergence between empirical observations and the true distribution for discrete distributions (Mardia et al., 2018).[4] We believe that the $\sqrt{t}$ dependence in the bound in Lemma 4.2 is spurious due to our analysis via the first inequality.

### 4.2. Bounding the polynomial approximation error

In this section we bound term $\max_j |b_j|$. Let $f$ be any Lipschitz-1 function on $[0, 1]$. Let $\hat{f}_t$ be degree $t$ polynomial

[4] The details are available in the supplementary material and arXiv version (Vinayak et al., 2019).

approximation of $f$ using Bernstein polynomials:

$$\hat{f}_t(x) = \sum_{j=0}^{t} b_j \binom{t}{j} x^j (1-x)^{t-j} := \sum_{j=0}^{t} b_j B_j^t(x), \quad (15)$$

where, $B_j^t(x) := \binom{t}{j} x^j (1-x)^{t-j}$, is $j-$th Bernstein polynomial of degree $t$, for $j = 0, 1, ..., t$. Our goal is to bound the uniform approximation error, $||f - \hat{f}||_\infty := \max_{x\in[0,1]} |f(x) - \hat{f}(x)|$ while controlling the magnitude $|b_j|$, of the coefficients. We note that $\max_j |b_j|$ appears in the bounds of error terms $(b)$ and $(c)$ in Equation (12), and hence it is important to control it to obtain tight bounds on the Wasserstein-1 metric in different regimes of $t$ and $N$. Bernstein used $t+1$ uniform samples of the function $f$ on $[0,1]$, $f(\frac{j}{t})$, as the coefficients in Equation (15) and showed that the uniform approximation error of such approximation is $||f - \hat{f}||_\infty \leq \frac{C}{\sqrt{t}}$, where $C$ is a constant. This approximation is not sufficient to show the bounds in Theorems 3.1 and 3.2. *Can we obtain better uniform approximation error using Bernstein polynomials with other bounded coefficients?* The following proposition answers this question.

**Proposition 4.1.** Any Lipschitz-1 funciton on $[0,1]$ can be approximated using Bernstein polynomials (Equation 15) of degree $t$, with an uniform approximation error of

- $\mathcal{O}(\frac{1}{t})$ with $\max_j |b_j| \leq \sqrt{t} 2^t$.
- $\mathcal{O}(\frac{1}{k})$ with $\max_j |b_j| \leq \sqrt{k}(t+1)e^{\frac{k^2}{t}}$, for $k < t$.

For $t$ above $\Omega(\log N)$, we set $k = \sqrt{t \log N^c}$, for appropriate choice of $c > 0$, obtaining a bound of $\max_j |b_j| \leq t^{1/4}(t+1)\left(\log N^c\right)^{1/4} N^c$. Combining these bounds with Lemmas 4.1 and 4.2 gives the results in Theorems 3.1 and 3.2. In the reminder of this section, we sketch out the proof of Proposition 4.1. Key idea is to approximate $f$ using Chebyshev polynomials of lower degree, $k \leq t$, and transform to Bernstein polynomials of degree $t$ to obtain appropriate bounds on the coefficients $|b_j|$.

Let $\tilde{T}_m$ denote Chebyshev polynomial of degree $m$ shifted to $[0,1]$ which satisfy the following recursive relation:

$$\tilde{T}_m(x) = (4x-2)\tilde{T}_{m-1} - \tilde{T}_{m-2}(x), \; m = 2, 3, ....,$$

and $\tilde{T}_0(x) = 1$, $\tilde{T}_1(x) = 2x - 1$. We use the following lemma regarding Chebyshev polynomial approximation [5].

**Lemma 4.3.** Given any Lipschitz-1 function $f(x)$ on $[0,1]$, there exists a degree $k$ polynomial in the form of $f_k(x) = \sum_{m=0}^{k} a_m \tilde{T}_m(x)$ that approximates $f(x)$ with error $\max_{x\in[0,1]} |f(x) - f_k(x)| = O(\frac{1}{k})$, where $\tilde{T}_m(x)$ denotes Chebyshev polynomial of degree $m$ shifted to $[0,1]$. Further, the coefficients $(a_1, a_2, \ldots, a_k)$ satisfies $\|a\|_2 \leq 1$.

[5] The proof is available in the supplementary material and in the arXiv version (Vinayak et al., 2019).

Chebyshev polynomial $\tilde{T}_m$, can be written in terms of Bernstein-Bezier polynomials of degree $m$ as follows (Rababah, 2003):

$$\tilde{T}_m(x) = \sum_{i=0}^{m} (-1)^{m-i} \frac{\binom{2m}{2i}}{\binom{m}{j}} B_i^m(x). \quad (16)$$

Note that the coefficients of $B_i^m$ can be as large as $2^m$. For $m = t$, this gives an upper bound of $2^t$ on the coefficients of Bernstein polynomial. This along with Equation 25 gives the first part of Proposition 4.1. To show the second part, we need to bound the coefficents when $m < t$.

**Degree raising:** Bernstein polynomials of degree $m < t$ can be raised to degree $t$ as:

$$B_i^m(x) = \sum_{j=i}^{i+t-m} \frac{\binom{m}{i}\binom{t-m}{j-i}}{\binom{t}{j}} B_j^t(x). \quad (17)$$

Using degree raising of Bernstein polynomials, we can write shifted Chebyshev polynomials of degree $m < t$ in terms of Bernstein polynomials of degree $t$ as,

$$\begin{aligned}\tilde{T}_m(x) &= \sum_{i=0}^{m} (-1)^{m-i} \frac{\binom{2m}{2i}}{\binom{m}{i}} \sum_{j=i}^{i+t-m} \frac{\binom{m}{i}\binom{t-m}{j-i}}{\binom{t}{j}} B_j^t(x), \\ &=: \sum_{j=0}^{t} C(t,m,j) B_j^t(x), \end{aligned} \quad (18)$$

where the coefficient of $j$-th Bernstien polynomial of degree $t$ is given by[6],

$$C(t,m,j) := \sum_{l=0}^{j} (-1)^{m-l} \frac{\binom{2m}{2l}\binom{t-m}{j-l}}{\binom{t}{j}}. \quad (19)$$

Following is a generating function for the coefficients,

$$\begin{aligned}&(1+z)^{t-m} \frac{(1+i\sqrt{z})^{2m} + (1-i\sqrt{z})^{2m}}{2} \\ &= \sum_{j=0}^{t} C(t,m,j) \binom{t}{j} z^{t-j}. \end{aligned} \quad (20)$$

Using Beta function, the binomial terms in the denominator can be written as, $\binom{t}{j}^{-1} = (t+1)\int_0^1 (1-u)^j u^{t-j} du$. We bound the generating function of the coefficients on the unit circle and use Parseval's theorem to prove the following lemma (details are available in the supplementary material and in arXiv version (Vinayak et al., 2019)).

**Lemma 4.4.** The $l_2$-norm of the coefficents of $B_j^t$ can be bounded as follows,

$$\sqrt{\sum_{j=0}^{t} |C(t,m,j)|^2} \leq (t+1)e^{\frac{m^2}{t}}. \quad (21)$$

[6] For positive integers $a, b > 0$, $\binom{a}{b} = 0$ when $a < b$.

From Lemma 4.4 (Equation 21), we can obtain the following bound on the coefficients:

$$|C(t,m,j)| \leq (t+1)e^{\frac{m^2}{t}}. \tag{22}$$

**Remark 4.1.** Bounding $|C(t,m,j)|$ by the $l_2$-norm of the coefficients gives a weak bound. We conjecture that the right bound on the coefficients of $B_j^t$ for every $m \leq t$ to be,

$$|C(t,m,j)| \leq e^{\frac{m^2}{t}},\ j = 1,2,...,t. \tag{23}$$

In fact, for a fixed $m$, the coefficients $C(t,m,j)$ should converge to points sampled uniformly from $T_m(x)$ as $t \to \infty$ by Bernstein's approximation. So the bound on the coefficients should converge to 1 as $t \to \infty$.

Let $f$ be a Lipschitz-1 function on $[0,1]$. We first let $f_k$ be the polynomial approximation using Chebyshev polynomials upto degree $k = \sqrt{t \log N^c}$ obtained from Lemma 4.3. Then we re-write each $\tilde{T}_m$ use Bernstein polynomials of degree $k$ followed by degree raising to $t$.

$$\begin{aligned} f_k(x) &= \sum_{m=0}^{k} a_m \tilde{T}_m(x) = \sum_{m=0}^{k} a_m \left( \sum_{j=0}^{t} C(t,m,j) B_j^t(x) \right) \\ &= \sum_{j=0}^{t} \left( \sum_{m=0}^{k} a_m\, C(t,m,j) \right) B_j^t(x) \\ &=: \sum_{j=0}^{t} b_j B_j^t(x). \end{aligned} \tag{24}$$

Since $||a||_2 \leq 1$,and from Equation 22, we have the following bound on the coefficients, for $j = 1,2,....,t$,

$$\begin{aligned} |b_j| &= \left| \sum_{m=0}^{k} a_m\, C(t,m,j) \right| \leq \sum_{m=0}^{k} |a_m|\ |C(t,m,j)|\,, \\ &\leq \sqrt{k}\ \max_j |C(t,m,j)| \leq \sqrt{k}(t+1)e^{\frac{k^2}{t}}. \end{aligned} \tag{25}$$

### 4.3. Lower bound for medium $t$ regime

The basic idea of the proof of Theorem 3.3 is to construct a pair of distributions $P, Q$ such that $W_1(P,Q) = \Theta(\frac{1}{\sqrt{t \log N}})$. With $N$ coins sampled from theses distributions each with $t$ flips, we argue that it is information theoretically hard to distinguish the two distributions. We use the following two propositions[7]:

**Proposition 4.2.** Given two distributions $P, Q$, supported on $[1/2 - \sqrt{\frac{\log N}{t}}, 1/2 + \sqrt{\frac{\log N}{t}}]$, whose first $L := e^4 \log N$ moments match, let $p \sim P$, $X \sim \text{Binomial}(t,p)$, $q \sim Q$ and $Y \sim \text{Binomial}(t,q)$. The total variation distance between $X$ and $Y$ satisfies $TV(X,Y) \leq \frac{2\sqrt{t}}{N^{e^4}}$.

[7]The proofs of these propositions are provided in supplementary material and arXiv version (Vinayak et al., 2019)

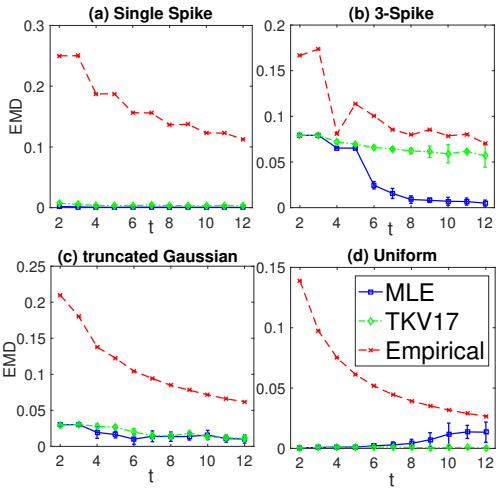


*Figure 1.* EMD between estimated and true distribution for small $t$ (number of coins $N = 1e6$) for various distributions using MLE (blue), moment matching (TKV17) (Tian et al., 2017) (magenta) and empirical estimate (red). Results are averaged over 10 runs.

**Proposition 4.3.** For any $t$, there exists a pair of distributions $P, Q$ supported on $[a,b]$ where $0 < a < b$ such that $P$ and $Q$ have identical first $t$ moments, and $W_1(P - D_Q) \geq \frac{(b-a)}{2t}$

*Proof of Theorem 3.3.* First we apply Proposition 4.3 to construct a pair of distributions $P, Q$ supported on $[1/2 - \sqrt{\frac{\log N}{t}}, 1/2 + \sqrt{\frac{\log N}{t}}]$ such that $P$ and $Q$ have identical first $L := e^4 \log N$ moments, and $W_1(P - Q) \geq \frac{1}{e^4} \frac{1}{\sqrt{t \log N}}$. Let $\mathbf{X} := \{X_i\}_{i=1}^N$ be random variables with $X_i \sim \text{Binomial}(t, p_i)$ where $p_i$ is drawn independently from $P$. Let $\mathbf{Y} := \{Y_i\}_{i=1}^N$ be random variables with $Y_i \sim \text{Binomial}(t, q_i)$ where $q_i$ is drawn independently from $Q$. Denote $P_N$ as the joint distribution of $\mathbf{X}$ and $Q_N$ as the joint distribution of $\mathbf{Y}$. It follows from Proposition 4.2 that $TV(X_i, Y_i) \leq \frac{2\sqrt{t}}{N^{e^4}}$. By the property of the product distribution and $t \leq \frac{N^{2(e^4-1)}}{36}$, $TV(P_N, Q_N) \leq \frac{2\sqrt{t}}{N^{e^4-1}} \leq 1/3$, which implies the minimax error is at least $\frac{1}{3e^4\sqrt{t \log N}}$. □

## 5. Numerical Experiments

Recall that the MLE (Equation 2) is a convex optimization problem,

$$\hat{P}_{\text{mle}} \in \arg\max_{Q \in \mathcal{D}} \sum_{s=0}^{t} h_s^{\text{obs}} \log E_Q[h_s],$$

where $\mathcal{D}$ is the set of all distributions on $[0,1]$. We discretize the interval $[0,1]$ into a uniform grid of width $\frac{1}{m}$. Note that as long as the error due to discretization $\mathcal{O}(\frac{1}{m})$ is smaller than the expected error in earth mover's distance (EMD), we will not be losing much numerically. Unless otherwise specified, we use grid length of $m = 1000$. The discretized set $\hat{\mathcal{D}}$ can then be written as $\hat{\mathcal{D}}_m := \{ q \in \mathbb{R}^{m+1} : q \geq 0, \mathbf{1}^\top q = 1 \}$. We then solve the

MLE which is convex on this discrete convex set using cvx (Grant & Boyd, 2014; 2008) for Matlab®.

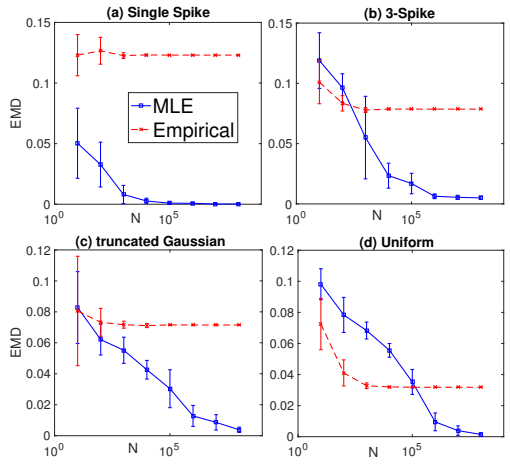


*Figure 2.* Comparing EMD between estimated distribution and truth for varying number of coins (with $t = 10$) for various distributions using MLE (blue) and empirical estimate (red).

### 5.1. Simulations on synthetic data:

We demonstrate the performance of the MLE on synthetic datasets where we know the ground truth. We consider 4 distributions, (1) single spike at $0.5$, (2) mixture of 3 spikes of equal mass at $0.25, 0.5$ and $0.75$, (3) truncated Gaussian on $[0, 1]$ with mean $0.5$ and variance $0.1$ and (4) uniform distribution on $[0, 1]$.

**Varying $t$:** In the first set of experiments we consider the vary $t$ regime where $t = \mathcal{O} \log N$. With the population size $N = 1e6$, we vary $t$ from 2 to 12. We run MLE, moment matching estimator and empirical estimator. Figure 1 shows the earth mover's distance (EMD) of the estimates from the true distribution as a function of $t$.

**Varying $N$:** For $t = 10$, we vary the population size $N$ from 10 to $10^8$ in multiples of 10. Figure 2 shows the comparison of performance of the MLE and empirical estimator in EMD. As $N$ increases, the second term in EMD which depends on $N$ decreases. We note that the error in EMD for MLE is much lower than that for empirical distribution when $t \ll N$ as predicted by our analysis.

**Varying $t$:** For $N = 1e6$, we vary the number of tosses $t$ from 2 to 10 in steps of two and then $t = [50, 100, 500, 1000]$ to illustrate the performance of the MLE as $t$ varies widely. Figure 3 shows the comparison of performance of the MLE and empirical estimator in EMD.

### 5.2. Experiments on real datasets

We ran the MLE on two real datasets used in (Tian et al., 2017): (1) A dataset on *political leanings* of counties in the US with data on whether a county leaned Democratic or Republican for $N = 3116$ counties in $t = 8$ presidential elections from 1976 to 2004. Here, each county $i$ is assumed to have a true probablity $p_i$ with which the county leans Republican in a given election (assuming the independence

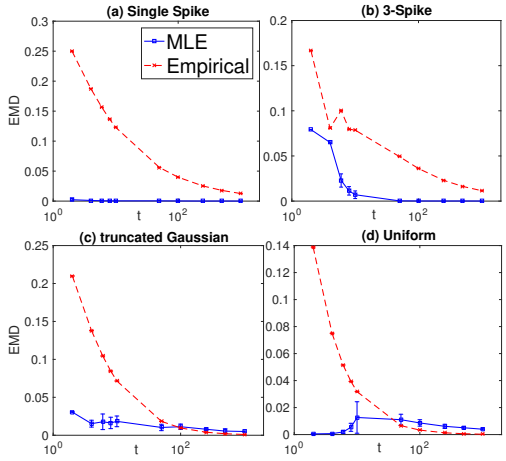


*Figure 3.* Comparing EMD between estimated distribution and truth for varying number of tosses (with $N = 1e6$) for various distributions using MLE (blue) and empirical estimate (red).

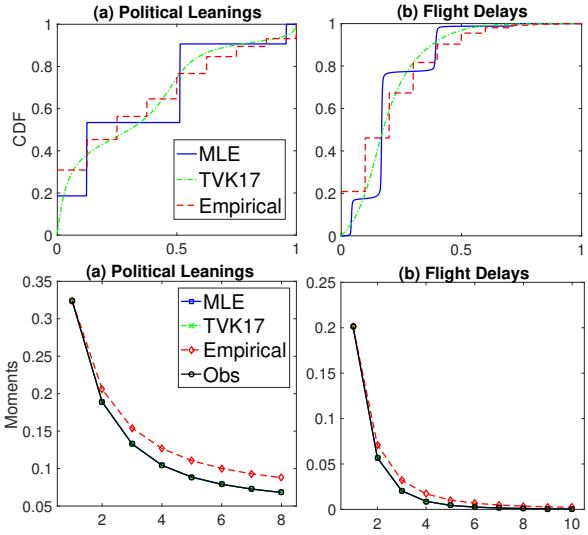


*Figure 4.* CDF of estimated distribution (top row), and first $t$ moments (bottom row) for (a) political leaning ($t = 8$, $N = 3116$) and (b) flight delay datasets ($t = 10$, $N = 25,156$) using MLE (blue), moment matching (TKV17) (Tian et al., 2017) (green) and empirical estimate (red). Observed moments is in black.

of counties and elections). (2) A dataset of *delays of flights* with $N = 25,156$ flights. Each flight is assumed to have an intrinsic probability $p_i$ of being delayed more than 15 minutes. Figure 4 shows the CDF output by MLE, moment matching (Tian et al., 2017) and empirical estimators on these datasets. We note that the observed fingerprints often lie outside the set of valid fingerprints (see Remark 3.1), and hence the solution to the MLE is the projection on the surface. Thus, the MLE tends to give sparser solutions. While the CDF output by the MLE and moment matching qualitatively look very different, their first $t$ moments match and fit the observed first $t$ moments extremely well (bottom row of Figure 4). Without any side information to enforce constraints like smoothness, it is not possible to pick one over the other with sparse observations.

## Acknowledgements

Sham Kakade acknowledges funding from the Washington Research Foundation for Innovation in Data-intensive Discovery, the National Science Foundation Grant under award CCF-1637360 (Algorithms in the Field) and award CCF-1703574, and the Office of Naval Research (Minerva Initiative) under award N00014-17-1-2313. Gregory Valiant and Weihao Kong were supported by National Science Foundation award CCF-1704417 and Office of Naval Research award N00014-18-1-2295.

## References

Acharya, J., Orlitsky, A., and Pan, S. Recent results on pattern maximum likelihood. In *Networking and Information Theory, 2009. ITW 2009. IEEE Information Theory Workshop on*, pp. 251–255. IEEE, 2009.

Acharya, J., Das, H., Mohimani, H., Orlitsky, A., and Pan, S. Exact calculation of pattern probabilities. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pp. 1498–1502. IEEE, 2010.

Acharya, J., Das, H., Orlitsky, A., and Suresh, A. T. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *International Conference on Machine Learning*, pp. 11–21, 2017.

Bell, G., Lechowicz, M. J., and Waterway, M. J. Environmental heterogeneity and species diversity of forest sedges. *Journal of Ecology*, 88(1):67–87, 2000.

Böhning, D. Likelihood inference for mixtures: geometrical and other constructions of monotone step-length algorithms. *Biometrika*, 76(2):375–383, 1989.

Charikar, M., Shiragur, K., and Sidford, A. Efficient profile maximum likelihood for universal symmetric property estimation. In *Proceedings of the 51st Annual Symposium on Theory of Computing*. ACM, 2019.

Colwell, R. K. and Coddington, J. A. Estimating terrestrial biodiversity through extrapolation. *Phil. Trans. R. Soc. Lond. B*, 345(1311):101–118, 1994.

Cressie, N. A quick and easy empirical bayes estimate of true scores. *Sankhyā: The Indian Journal of Statistics, Series B*, pp. 101–108, 1979.

Efron, B. and Thisted, R. Estimating the number of unseen species: How many words did shakespeare know? *Biometrika*, 63(3):435–447, 1976.

Fisher, R. A., Corbet, A. S., and Williams, C. B. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pp. 42–58, 1943.

Good, I. and Toulmin, G. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

Grant, M. and Boyd, S. Graph implementations for nonsmooth convex programs. In Blondel, V., Boyd, S., and Kimura, H. (eds.), *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pp. 95–110. Springer-Verlag Limited, 2008. http://stanford.edu/~boyd/graph_dcp.html.

Grant, M. and Boyd, S. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.

Han, Y., Jiao, J., and Weissman, T. Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under wasserstein distance. *arXiv preprint arXiv:1802.08405*, 2018.

Jiao, J., Venkat, K., Han, Y., and Weissman, T. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

Jiao, J., Han, Y., and Weissman, T. Minimax estimation of the l1 distance. *IEEE Transactions on Information Theory*, 2018.

Kantorovich, L. V. and Rubinstein, G. S. On a space of completely additive functions. *Vestnik Leningrad. Univ, 13(7):52–59*, 1958.

Laird, N. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

Lesperance, M. L. and Kalbfleisch, J. D. An algorithm for computing the nonparametric mle of a mixing distribution. *Journal of the American Statistical Association*, 87(417): 120–126, 1992.

Lindsay, B. G. The geometry of mixture likelihoods: a general theory. *The annals of statistics*, pp. 86–94, 1983a.

Lindsay, B. G. The geometry of mixture likelihoods, part ii: the exponential family. *The Annals of Statistics*, 11(3): 783–792, 1983b.

Lord, F. M. A strong true-score theory, with applications. *Psychometrika*, 30(3):239–270, 1965.

Lord, F. M. Estimating true-score distributions in psychological testing (an empirical bayes estimation problem). *Psychometrika*, 34(3):259–299, 1969.

Lord, F. M. and Cressie, N. An empirical bayes procedure for finding an interval estimate. *Sankhyā: The Indian Journal of Statistics, Series B*, pp. 1–9, 1975.

Mardia, J., Jiao, J., Tánczos, E., Nowak, R. D., and Weissman, T. Concentration inequalities for the empirical distribution. *arXiv preprint arXiv:1809.06522*, 2018.

Millar, W. J. Distribution of body weight and height: comparison of estimates based on self-reported and observed measures. *Journal of Epidemiology & Community Health*, 40(4):319–323, 1986.

Orlitsky, A., Santhanam, N. P., Viswanathan, K., and Zhang, J. On modeling profiles instead of values. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 426–435. AUAI Press, 2004.

Orlitsky, A., Suresh, A. T., and Wu, Y. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016.

Palmer, M. W. and Dixon, P. M. Small-scale environmental heterogeneity and the analysis of species distributions along gradients. *Journal of Vegetation Science*, 1(1):57–65, 1990.

Paninski, L. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

Rababah, A. Transformation of chebyshev–bernstein polynomial basis. *Computational Methods in Applied Mathematics Comput. Methods Appl. Math.*, 3(4):608–622, 2003.

Simar, L. Maximum likelihood estimation of a compound poisson process. *The Annals of Statistics*, pp. 1200–1209, 1976.

Tian, K., Kong, W., and Valiant, G. Optimally learning populations of parameters. *arXiv preprint arXiv:1709.02707*, 2017.

Turnbull, B. W. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 290–295, 1976.

Valiant, G. and Valiant, P. Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new clts. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 685–694. ACM, 2011.

Valiant, P. and Valiant, G. Estimating the unseen: improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems*, pp. 2157–2165, 2013.

Vinayak, R. K., Kong, W., Valiant, G., and Kakade, S. M. Maximum likelihood estimation for learning populations of parameters. *arXiv preprint arXiv:1902.04553*, 2019.

Vontobel, P. O. The bethe approximation of the pattern maximum likelihood distribution. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012.

Wood, G. R. Binomial mixtures: geometric estimation of the mixing distribution. *The Annals of Statistics*, 27(5): 1706–1721, 1999.

Wu, Y. and Yang, P. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *arXiv preprint arXiv:1504.01227*, 2015.

Wu, Y. and Yang, P. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.