

---

# Learning Populations of Parameters

---

Kevin Tian, Weihao Kong, and Gregory Valiant

Department of Computer Science

Stanford University

Stanford, CA, 94305

(kjtian, whkong, valiant)@stanford.edu

## Abstract

Consider the following estimation problem: there are  $n$  entities, each with an unknown parameter  $p_i \in [0, 1]$ , and we observe  $n$  independent random variables,  $X_1, \dots, X_n$ , with  $X_i \sim \text{Binomial}(t, p_i)$ . How accurately can one recover the “histogram” (i.e. cumulative density function) of the  $p_i$ ’s? While the empirical estimates would recover the histogram to earth mover distance  $\Theta(\frac{1}{\sqrt{t}})$  (equivalently,  $\ell_1$  distance between the CDFs), we show that, provided  $n$  is sufficiently large, we can achieve error  $O(\frac{1}{t})$  which is information theoretically optimal. We also extend our results to the multi-dimensional parameter case, capturing settings where each member of the population has multiple associated parameters. Beyond the theoretical results, we demonstrate that the recovery algorithm performs well in practice on a variety of datasets, providing illuminating insights into several domains, including politics, sports analytics, and variation in the gender ratio of offspring.

## 1 Introduction

In many domains, from medical records, to the outcomes of political elections, performance in sports, and a number of biological studies, we have enormous datasets that reflect properties of a large number of entities/individuals. Nevertheless, for many of these datasets, the amount of information that we have about each entity is relatively modest—often too little to accurately infer properties about that entity. In this work, we consider the extent to which we can accurately recover an estimate of the *population* or *set* of property values of the entities, even in the regime in which there is insufficient data to resolve properties of each specific entity.

To give a concrete example, suppose we have a large dataset representing 1M people, that records whether each person had the flu in each of the past 5 years. Suppose each person has some underlying probability of contracting the flu in a given year, with  $p_i$  representing the probability that the  $i^{\text{th}}$  person contracts the flu each year (and assuming independence between years). With 5 years of data, the empirical estimates  $\hat{p}_i$  for each person are quite noisy (and the estimates will all be multiples of  $\frac{1}{5}$ ). Despite this, to what extent can we hope to accurately recover the population or set of  $p_i$ ’s? An accurate recovery of this population of parameters might be very useful—is it the case that most people have similar underlying probabilities of contracting the flu, or is there significant variation between people? Additionally, such an estimate of this population could be fruitfully leveraged as a prior in making concrete predictions about individuals’  $p_i$ ’s, as a type of *empirical Bayes* method.

The following example motivates the hope for significantly improving upon the empirical estimates:

**Example 1.** Consider a set of  $n$  biased coins, with the  $i^{\text{th}}$  coin having an unknown bias  $p_i$ . Suppose we flip each coin twice (independently), and observe that the number of coins where both flips landed heads is roughly  $\frac{n}{4}$ , and similarly for the number coins that landed HT, TH, and TT. We can safely conclude that almost all of the  $p_i$ ’s are almost exactly  $\frac{1}{2}$ . The reasoning proceeds in two

steps: first, since the average outcome is balanced between heads and tails, the average  $p_i$  must be very close to  $\frac{1}{2}$ . Given this, if there was any significant amount of variation in the  $p_i$ 's, one would expect to see significantly more HHs and TTs than the HT and TH outcomes, simply because  $\Pr[\text{Binomial}(2, p) = 1] = 2p(1 - p)$  attains a maximum for  $p = 1/2$ .

Furthermore, suppose we now consider the  $i^{\text{th}}$  coin, and see that it landed heads twice. The empirical estimate of  $p_i$  would be 1, but if we observe close to  $\frac{n}{4}$  coins with each pair of outcomes, using the above reasoning that argues that almost all of the  $p$ 's are likely close to  $\frac{1}{2}$ , we could safely conclude that  $p_i$  is likely close to  $\frac{1}{2}$ .

This ability to “denoise” the empirical estimate of a parameter based on the observations of a number of independent random variables (in this case, the outcomes of the tosses of the other coins), was first pointed out by Charles Stein in the setting of estimating the means of a set of Gaussians and is known as “Stein’s phenomenon” [14]. We discuss this further in Section 1.1. Example 1 was chosen to be an extreme illustration of the ability to leverage the large number of entities being studied,  $n$ , to partially compensate for the small amount of data reflecting each entity (the 2 tosses of each coin, in the above example).

Our main result, stated below, demonstrates that even for worst-case sets of  $p$ 's, significant “denoising” is possible. While we cannot hope to always accurately recover each  $p_i$ , we show that we can accurately recover the *set* or *histogram* of the  $p$ 's, as measured in the  $\ell_1$  distance between the cumulative distribution functions, or equivalently, the “earth mover’s distance” (also known as 1-Wasserstein distance) between the set of  $p$ 's regarded as a distribution  $P$  that places mass  $\frac{1}{n}$  at each  $p_i$ , and the distribution  $Q$  returned by our estimator. Equivalently, our returned distribution  $Q$  can also be represented as a set of  $n$  values  $q_1, \dots, q_n$ , in which case this earth mover’s distance is precisely  $1/n$  times the  $\ell_1$  distance between the vector of sorted  $p_i$ 's, and the vector of sorted  $q_i$ 's.

**Theorem 1.** *Consider a set of  $n$  probabilities,  $p_1, \dots, p_n$  with  $p_i \in [0, 1]$ , and suppose we observe the outcome of  $t$  independent flips of each coin, namely  $X_1, \dots, X_n$ , with  $X_i \sim \text{Binomial}(t, p_i)$ . There is an algorithm that produces a distribution  $Q$  supported on  $[0, 1]$ , such that with probability at least  $1 - \delta$  over the randomness of  $X_1, \dots, X_n$ ,*

$$\|P - Q\|_W \leq \frac{\pi}{t} + 3^t \sum_{i=1}^t \sqrt{\ln\left(\frac{2t}{\delta}\right) \frac{3}{n}} \leq \frac{\pi}{t} + O_\delta\left(\frac{3^t t \ln t}{\sqrt{n}}\right),$$

where  $P$  denotes the distribution that places mass  $\frac{1}{n}$  at value  $p_i$ , and  $\|\cdot\|_W$  denotes the Wasserstein distance.

The above theorem applies to the setting where we hope to recover a set of arbitrary  $p_i$ 's. In some practical settings, we might think of each  $p_i$  as being sampled independently from some underlying distribution  $P_{\text{pop}}$  over probabilities, and the goal is to recover this population distribution  $P_{\text{pop}}$ . Since the empirical distribution of  $n$  draws from a distribution  $P_{\text{pop}}$  over  $[0, 1]$  converges to  $P_{\text{pop}}$  in Wasserstein distance at a rate of  $O(1/\sqrt{n})$ , the above theorem immediately yields the analogous result in this setting:

**Corollary 1.** *Consider a distribution  $P_{\text{pop}}$  over  $[0, 1]$ , and suppose we observe  $X_1, \dots, X_n$  where  $X_i$  is obtained by first drawing  $p_i$  independently from  $P_{\text{pop}}$ , and then drawing  $X_i$  from  $\text{Binomial}(t, p_i)$ . There is an algorithm that will output a distribution  $Q$  such that with probability at least  $1 - \delta$ ,*

$$\|P_{\text{pop}} - Q\|_W \leq \frac{\pi}{t} + O_\delta\left(\frac{3^t t \ln t}{\sqrt{n}}\right).$$

The inverse linear dependence on  $t$  of Theorem 1 and Corollary 1 is information theoretically optimal, and is attained asymptotically for sufficiently large  $n$ :

**Proposition 1.** *Let  $P_{\text{pop}}$  denote a distribution over  $[0, 1]$ , and for positive integers  $t$  and  $n$ , let  $X_1, \dots, X_n$  denote random variables with  $X_i$  distributed as  $\text{Binomial}(t, p_i)$  where  $p_i$  is drawn independently according to  $P_{\text{pop}}$ . An estimator  $f$  maps  $X_1, \dots, X_n$  to a distribution  $f(X_1, \dots, X_n)$ . Then, for every fixed  $t$ , the following lower bound on the accuracy of any estimator holds for all  $n$ :*

$$\inf_f \sup_{P_{\text{pop}}} \mathbb{E} [\|f(X_1, \dots, X_n) - P_{\text{pop}}\|_W] > \frac{1}{4t}.$$

Our estimation algorithm, whose performance is characterized by Theorem 1, proceeds via the *method of moments*. Given  $X_1, \dots, X_n$  with  $X_i \sim \text{Binomial}(t, p_i)$ , and sufficiently large  $n$ , we can

obtain accurate estimates of the first  $t$  moments of the distribution/histogram  $P$  defined by the  $p_i$ 's. Accurate estimates of the first  $t$  moments can then be leveraged to recover an estimate of  $P$  that is accurate to error  $\frac{1}{t}$  plus a factor that depends (exponentially on  $t$ ) on the error in the recovered moments.

The intuition for the lower bound, Proposition 1, is that the realizations of  $\text{Binomial}(t, p_i)$  give *no* information beyond the first  $t$  moments. Additionally, there exist distributions  $P$  and  $Q$  whose first  $t$  moments agree exactly, but which differ in their  $t + 1^{st}$  moment, and have  $\|P - Q\|_W \geq \frac{1}{2t}$ . Putting these two pieces together establishes the lower bound.

We also extend our results to the practically relevant multi-parameter analog of the setting described above, where the  $i^{th}$  datapoint corresponds to a pair, or  $d$ -tuple of hidden parameters,  $p_{(i,1)}, \dots, p_{(i,d)}$ , and we observe independent random variables  $X_{(i,1)}, \dots, X_{(i,d)}$  with  $X_{(i,j)} \sim \text{Binomial}(t_{(i,j)}, p_{(i,j)})$ . In this setting, the goal is to recover the multivariate set of  $d$ -tuples  $\{p_{(i,1)}, \dots, p_{(i,d)}\}$ , again in an earth mover's sense. This setting corresponds to recovering an approximation of an underlying joint distribution over these  $d$ -tuples of parameters.

To give one concrete motivation for this problem, consider a hypothetical setting where we have  $n$  genotypes (sets of genetic features), with  $t_i$  people of the  $i$ th genotype. Let  $X_{(i,1)}$  denote the number of people with the  $i$ th genotype who exhibit disease 1, and  $X_{(i,2)}$  denote the number of people with genotype  $i$  who exhibit disease 2. The interpretation of the hidden parameters  $p_{i,1}$  and  $p_{i,2}$  are the respective probabilities of people with the  $i^{th}$  genotype of developing each of the two diseases. Our results imply that provided  $n$  is large, one can accurately recover an approximation to the underlying set or two-dimensional joint distribution of  $\{(p_{i,1}, p_{i,2})\}$  pairs, even in settings where there are too few people of each genotype to accurately determine which of the genotypes are responsible for elevated disease risk. Recovering this set of pairs would allow one to infer whether there are common genetic drivers of the two diseases—even in the regime where there is insufficient data to resolve *which* genotypes are the common drivers.

Our multivariate analog of Theorem 1 is also formulated in terms of multivariate analog of earth mover's distance (see Definition 1 for a formal definition):

**Theorem 2.** *Let  $\{p_{i,j}\}$  denote a set of  $n$   $d$ -tuples of hidden parameters in  $[0, 1]^d$ , with  $i \in \{1, \dots, n\}$  and  $j \in \{1, \dots, d\}$ , and suppose we observe random variables  $X_{i,j}$ , with  $X_{i,j} \sim \text{Binomial}(t, p_{i,j})$ . There is an algorithm that produces a distribution  $Q$  supported on  $[0, 1]^d$ , such that with probability at least  $1 - \delta$  over the randomness of the  $X_{i,j}$ s,*

$$\|P - Q\|_W \leq \frac{C_1}{t} + C_2 \sum_{|\alpha|=1}^t \frac{d(2t)^{d+1} 2^t}{3^{|\alpha|}} \sqrt{\ln\left(\frac{1}{\delta}\right) \frac{1}{n}} \leq \frac{C_1}{t} + O_{\delta,t,d}\left(\frac{1}{\sqrt{n}}\right),$$

for absolute constants  $C_1, C_2$ , where  $\alpha$  is a  $d$ -dimensional multi-index consisting of all  $d$ -tuples of nonnegative integers summing to at most  $t$ ,  $P$  denotes the distribution that places mass  $\frac{1}{n}$  at value  $p_i = (p_{i,1}, \dots, p_{i,d}) \in [0, 1]^d$ , and  $\|\cdot\|_W$  denotes the  $d$ -dimensional Wasserstein distance between  $P$  and  $Q$ .

## 1.1 Related Work

The seminal paper of Charles Stein [14] was one of the earliest papers to identify the surprising possibility of leveraging the availability of independent data reflecting a large number of parameters of interest, to partially compensate for having little information about each parameter. The specific setting examined considered the problem of estimating a list of unknown means,  $\mu_1, \dots, \mu_n$  given access to  $n$  independent Gaussian random variables,  $X_1, \dots, X_n$ , with  $X_i \sim \mathcal{N}(\mu_i, 1)$ . Stein showed that, perhaps surprisingly, that there is an estimator for the list of parameters  $\mu_1, \dots, \mu_n$  that has smaller expected squared error than the naive unbiased empirical estimates of  $\hat{\mu}_i = X_i$ . This improved estimator “shrinks” the empirical estimates towards the average of the  $X_i$ 's. In our setting, the process of recovering the set/histogram of unknown  $p_i$ 's and then leveraging this recovered set as a prior to correct the empirical estimates of each  $p_i$  can be viewed as an analog of Stein's “shrinkage”, and will have the property that the empirical estimates are shifted (in a non-linear fashion) towards the average of the  $p_i$ 's.

More closely related to the problem considered in this paper is the work on recovering an approximation to the unlabeled *set* of probabilities of domain elements, given independent draws from a

distribution of large discrete support (see e.g. [11, 2, 15, 16, 1]). Instead of learning the distribution, these works considered the alternate goal of simply returning an approximation to the multiset of probabilities with which the domain elements arise but without specifying which element occurs with which probability. Such a multiset can be used to estimate useful properties of the distribution that do not depend on the labels of the domain of the distribution, such as the entropy or support size of the distribution, or the number of elements likely to be observed in a new, larger sample [12, 17]. The benefit of pursuing this weaker goal of returning the unlabeled multiset is that it can be learned to significantly higher accuracy for a given sample size—essentially as accurate as the empirical distribution of a sample that is a logarithmic factor larger [15, 17].

Building on the above work, the recent work [18] considered the problem of recovering the “frequency spectrum” of rare genetic variants. This problem is similar to the problem we consider, but focuses on a rather different regime. Specifically, the model considered posits that each location  $i = 1, \dots, n$  in the genome has some probability  $p_i$  of being mutated in a given individual. Given the sequences of  $t$  individuals, the goal is to recover the set of  $p_i$ ’s. The work [18] focused on the regime in which many of the  $p_i$ ’s are significantly less than  $\frac{1}{nt}$ , and hence correspond to mutations that have never been observed; one conclusion of that work was that one can accurately estimate the number of such rare mutations that would be discovered in larger sequencing cohorts. Our work, in contrast, focuses on the regime where the  $p_i$ ’s are constant, and do not scale as a function of  $n$ , and the results are incomparable.

Also related to the current work are the works [9, 10] on *testing* whether certain properties of collections of distributions hold. The results of these works show that specific properties, such as whether most of the distributions are identical versus have significant variation, can be decided based on a sample size that is significantly sublinear in the number of distributions.

Finally, the papers [5, 6] consider the related by more difficult setting of learning “Poisson Binomials,” namely a sum of independent non-identical Bernoulli random variables, given access to samples. In contrast to our work, in the setting they consider, each “sample” consists of only the sum of these  $n$  random variables, rather than observing the outcome of each random variable.

## 1.2 Organization of paper

In Section 2 we describe the two components of our algorithm for recovering the population of Bernoulli parameters: obtaining accurate estimates of the low-order moments (Section 2.1), and leveraging those moments to recover the set of parameters (Section 2.3). The complete algorithm is presented in Section 2.2, and a discussion of the multi-dimensional extension to which Theorem 2 applies is described in Section 2.4. In Section 3 we validate the empirical performance of our approach on synthetic data, as well as illustrate its potential applications to several real-world settings.

## 2 Learning a population of binomial parameters

Our approach to recovering the underlying distribution or set of  $p_i$ ’s proceeds via the method of moments. In the following section we show that, given  $\geq t$  samples from each Bernoulli distribution, we can accurately estimate each of the first  $t$  moments. In Section 2.3 we explain how these first  $t$  moments can then be leveraged to recover the set of  $p_i$ ’s, to earth mover’s distance  $O(1/t)$ .

### 2.1 Moment estimation

Our method-of-moments approach proceeds by estimating the first  $t$  moments of  $P$ , namely  $\frac{1}{n} \sum_{i=1}^n p_i^k$ , for each integer  $k$  between 1 and  $t$ . The estimator we describe is unbiased, and also applies in the setting of Corollary 1 where each  $p_i$  is drawn i.i.d. from a distribution  $P_{pop}$ . In this case, we will obtain an unbiased estimator for  $\mathbb{E}_{p \leftarrow P_{pop}}[p^k]$ . We limit ourselves to estimating the first  $t$  moments because, as show in the proof of the lower bound, Proposition 1, the distribution of the  $X_i$ ’s are determined by the first  $t$  moments, and hence no additional information can be gleaned regarding the higher moments.

For  $1 \leq k \leq t$ , our estimate for the  $k^{th}$  moment is  $\beta_k = \frac{1}{n} \sum_{i=1}^n \frac{\binom{X_i}{k}}{\binom{t}{k}}$ . The motivation for this unbiased estimator is the following: Note that given any  $k$  i.i.d. samples of a variable distributed

according to  $\text{Bernoulli}(p_i)$ , an unbiased estimator for  $p_i^k$  is their product, namely the estimator which is 1 if all the tosses come up heads, and otherwise is 0. Thus, if we average over all  $\binom{t}{k}$  subsets of size  $k$ , and then average over the population, we still derive an unbiased estimator.

**Lemma 1.** *Given  $\{p_1, \dots, p_n\}$ , let  $X_i$  denote the random variable distributed according to  $\text{Binomial}(t, p_i)$ . For  $k \in \{1, \dots, t\}$ , let  $\alpha_k = \frac{1}{n} \sum_{i=1}^n p_i^k$  denote the  $k^{\text{th}}$  true moment, and  $\beta_k = \frac{1}{n} \sum_{i=1}^n \frac{\binom{X_i}{k}}{\binom{t}{k}}$  denote our estimate of the  $k^{\text{th}}$  moment. Then  $\mathbb{E}[\beta_k] = \alpha_k$ , and  $\Pr(|\beta_k - \alpha_k| \geq \epsilon) \leq 2e^{-\frac{1}{3}n\epsilon^2}$ .*

Given the above lemma, we obtain the fact that, with probability at least  $1 - \delta$ , the events  $|\alpha_k - \beta_k| \leq \sqrt{\ln(\frac{2t}{\delta}) \frac{3}{n}}$  simultaneously occur for all  $k \in \{1, \dots, t\}$ .

## 2.2 Distribution recovery from moment estimates

Given the estimates of the moments of the distribution  $P$ , as described above, our algorithm will recover a distribution,  $Q$ , whose moments are close to the estimated moments. We propose two algorithms, whose distribution recoveries are via the standard linear programming or quadratic programming approaches which will recover a distribution  $Q$  supported on some (sufficiently fine)  $\epsilon$ -net of  $[0, 1]$ : the variables of the linear (or quadratic) program correspond to the amount of probability mass that  $Q$  assigns to each element of the  $\epsilon$ -net, the constraints correspond to ensuring that the amount of mass at each element is nonnegative and that the total amount of mass is 1, and the objective function will correspond to the (possibly weighted) sum of the discrepancies between the estimated moments, and the moments of the distribution represented by  $Q$ .

To see why it suffices to solve this program over an  $\epsilon$ -net of the unit interval, note that any distribution over  $[0, 1]$  can be rounded so as to be supported on an  $\epsilon$ -net, while changing the distribution by at most  $\frac{\epsilon}{2}$  in Wasserstein distance. Additionally, such a rounding alters each moment by at most  $O(\epsilon)$ , because the rounding alters the individual contributions of point masses to the  $k^{\text{th}}$  moment by only  $O(\epsilon^k) < O(\epsilon)$ . As our goal is to recover a distribution with distance  $O(1/t)$ , it suffices to choose an  $\epsilon$ -net with  $\epsilon \ll 1/t$  so that the additional error due to this discretization is negligible. As this distribution recovery program has  $O(1/\epsilon)$  variables and  $O(t)$  constraints, both of which are independent of  $n$ , this program can be solved extremely efficiently both in theory and in practice.

We formally describe this algorithm below, which takes as input  $X_1, \dots, X_n$ , binomial parameter  $t$ , an integer  $m$  corresponding to the size of the  $\epsilon$ -net, and a weight vector  $w$ .

---

### Algorithms 1 and 2: Distribution Recovery with Linear / Quadratic Objectives

**Input:** Integers  $X_1, \dots, X_n$ , integers  $t$  and  $m$ , and weight vector  $w \in \mathbb{R}^t$ .

**Output:** Vector  $q = (q_0, \dots, q_m)$  of length  $m + 1$ , representing a distribution with probability mass  $q_i$  at value  $\frac{i}{m}$ .

- For each  $k \in \{1, \dots, t\}$ , compute  $\beta_k = \frac{1}{n} \sum \frac{\binom{X_i}{k}}{\binom{t}{k}}$ .

- (Algorithm 1) Solve the linear program over variables  $q_0, \dots, q_m$ :

$$\text{minimize: } \sum_{k=1}^t |\hat{\beta}_k - \beta_k| w_k, \text{ where } \hat{\beta}_k = \sum_{i=0}^m q_i \left(\frac{i}{m}\right)^k,$$

$$\text{subject to: } \sum q_i = 1, \text{ and for all } i, q_i \geq 0.$$

- (Algorithm 2) Solve the quadratic program over variables  $q_0, \dots, q_m$ :

$$\text{minimize: } \sum_{k=1}^t (\hat{\beta}_k - \beta_k)^2 w_k^2, \text{ where } \hat{\beta}_k = \sum_{i=0}^m q_i \left(\frac{i}{m}\right)^k,$$

$$\text{subject to: } \sum_i q_i = 1, \text{ and for all } i, q_i \geq 0.$$


---

### 2.2.1 Practical considerations

Our theoretical results, Theorem 1 and Corollary 1, apply to the setting where the weight vector,  $w$  in the above linear program objective function has  $w_k = 1$  for all  $k$ . It makes intuitive sense to penalize the discrepancy in the  $k$ th moment inversely proportionally to the empirically estimated standard deviation of the  $k^{th}$  moment estimate, and our empirical results are based on such a weighted objective.

Additionally, in some settings we observed an empirical improvement in the robustness and quality of the recovered distribution if one averages the results of running Algorithm 1 or 2 on several random subsamples of the data. In our empirical section, Section 3, we refer to this as a *bootstrapped* version of our algorithm.

### 2.3 Close moments imply close distributions

In this section we complete the high-level proof that Algorithm 1 accurately recovers  $P$ , the distribution corresponding to the set of  $p_i$ 's, establishing Theorem 1 and Corollary 1. The guarantees of Lemma 1 ensure that, with high probability, the estimated moments will be close to the true moments. Together with the observation that discretizing  $P$  to be supported on an  $\epsilon$ -net of  $[0, 1]$  alters the moments by  $O(\epsilon)$ , it follows that there is a solution to the linear program in the second step of Algorithm 1 corresponding to a distribution whose moments are close to the true moments of  $P$ , and hence with high probability Algorithm 1 will return such a distribution.

To conclude the proof, all that remains is to show that, provided the distribution  $Q$  returned by Algorithm 1 has similar first  $t$  moments to the true distribution,  $P$ , then  $P$  and  $Q$  will be close in Wasserstein (earth mover's) distance. We begin by formally defining the Wasserstein (earth mover's) distance between two distributions  $P$  and  $Q$ :

**Definition 1.** *The Wasserstein, or earth mover's, distance between distributions  $P, Q$ , is  $\|P - Q\|_W := \inf_{\gamma \in \Gamma(P, Q)} \int_{[0, 1]^{2d}} d(x, y) d\gamma(x, y)$ , where  $\Gamma(P, Q)$  is the set of all couplings on  $P$  and  $Q$ , namely a distribution whose marginals agree with the distributions. The equivalent dual definition is  $\|P - Q\|_W := \sup_{g \in \text{Lip}(1)} \int g(x) d(P - Q)(x)$  where the supremum is taken over Lipschitz functions,  $g$ .*

As its name implies, this distance metric can be thought of as the cost of the optimal scheme of "moving" the probability mass from  $P$  to create  $Q$ , where the cost per unit mass of moving from probability  $x$  and  $y$  is  $|x - y|$ . Distributions over  $\mathbb{R}$ , it is not hard to see that this distance is exactly the  $\ell_1$  distance between the associated cumulative distribution functions.

The following slightly stronger version of Proposition 1 in [7] bounds the Wasserstein distance between any pair of distributions in terms of the discrepancies in their low-order moments:

**Theorem 3.** *For two distributions  $P$  and  $Q$  supported on  $[0, 1]$  whose first  $t$  moments are  $\alpha$  and  $\beta$  respectively, the Wasserstein distance  $\|P - Q\|_W$  is bounded by  $\frac{\pi}{t} + 3^t \sum_{k=1}^t |\alpha_k - \beta_k|$ .*

The formal proof of this theorem is provided in the Appendix B, and we conclude this section with an intuitive sketch of this proof. For simplicity, first consider the setting where the two distributions  $P, Q$  have the *exact* same first  $t$  moments. This immediately implies that for any polynomial  $f$  of degree at most  $t$ , the expectation of  $f$  with respect to  $P$  is equal to the expectation of  $f$  with respect to  $Q$ . Namely,  $\int f(x)(P(x) - Q(x))dx = 0$ . Leveraging the definition of Wasserstein distance  $\|P - Q\|_W = \sum_{g \in \text{Lip}} \int g(x)(P(x) - Q(x))dx$ , the theorem now follows from the standard fact that, for any Lipschitz function  $g$ , there exists a degree  $t$  polynomial  $f_g$  that approximates it to within  $\ell_\infty$  distance  $O(1/t)$  on the interval  $[0, 1]$ .

If there is nonzero discrepancy between the first  $t$  moments of  $P$  and  $Q$ , the above proof continues to hold, with an additional error term of  $\sum_{k=1}^t c_k(\alpha_k - \beta_k)$ , where  $c_k$  is the coefficient of the degree  $k$  term in the polynomial approximation  $f_g$ . Leveraging the fact that any Lipschitz function  $g$  can be approximated to  $\ell_\infty$  distance  $O(1/t)$  on the unit interval using a polynomial with coefficients bounded by  $3^t$ , we obtain Theorem 3.

## 2.4 Extension: multivariate distribution estimation

We also consider the natural multivariate extension of the the problem of recovering a population of Bernoulli parameters. Suppose, for example, that every member  $i$  of a population of size  $n$  has two associated binomial parameters  $p_{(i,1)}, p_{(i,2)}$ , as in Theorem 2. One could estimate the marginal distribution of the  $p_{(i,1)}$  and  $p_{(i,2)}$  separately using Algorithm 1, but it is natural to also want to estimate the joint distribution up to small Wasserstein distance in the 2-d sense. Similarly, one can consider the analogous  $d$ -dimensional distribution recovery question.

The natural idea underlying our extension to this setting is to include estimates of the multivariate moments represented by multi-indices  $\alpha$  with  $|\alpha| \leq t$ . For example, in a 2-d setting, the moments for members  $i$  of the population would look like  $\mathbb{E}_{p_i \sim P}[p_{(i,1)}^a p_{(i,2)}^b]$ . Again, it remains to bound how close an interpolating polynomial can get to any  $d$ -dimensional Lipschitz function, and bound the size of the coefficients of such a polynomial. To this end, we use the following theorem from [3]:

**Lemma 2.** *Given any Lipschitz function  $f$  supported on  $[0, 1]^d$ , there is a degree  $s$  polynomial  $p(x)$  such that*

$$\sup_{x \in [0,1]^d} |p(x) - f(x)| \leq \frac{C_d}{t},$$

where  $C_d$  is a constant that depends on  $d$ .

In Appendix E, we prove the following bound on the magnitude of the coefficients of the interpolating polynomial:  $|c_\alpha| \leq \frac{(2t)^d 2^t}{3^{|\alpha|}}$ , where  $c_\alpha$  is the coefficient of the  $\alpha$  multinomial term. Together with the concentration bound of the  $\alpha^{th}$  moment of the distribution, we obtain Theorem 2, the multivariate analog of Theorem 1.

## 3 Empirical performance

### 3.1 Recovering distributions with known ground truth

We begin by demonstrating the effectiveness of our algorithm on several synthetic datasets. We considered three different choices for an underlying distribution  $P_{pop}$  over  $[0, 1]$ , then drew  $n$  independent samples  $p_1, \dots, p_n \leftarrow P_{pop}$ . For a parameter  $t$ , for each  $i \in \{1, \dots, n\}$ , we then drew  $X_i \leftarrow \text{Binomial}(t, p_i)$ , and ran our population estimation algorithm on the set  $X_1, \dots, X_n$ , and measured the extent to which we recovered the distribution  $P_{pop}$ . In all settings,  $n$  was sufficiently large that there was little difference between the histogram corresponding to the set  $\{p_1, \dots, p_n\}$  and the distribution  $P_{pop}$ . Figure 1 depicts the error of the recovered distribution as  $t$  takes on all even values from 2 to 14, for three choices of  $P_{pop}$ : the “3-spike” distribution with equal mass at the values  $1/4, 1/2$ , and  $3/4$ , a Normal distribution truncated to be supported on  $[0, 1]$ , and the uniform distribution over  $[0, 1]$ .

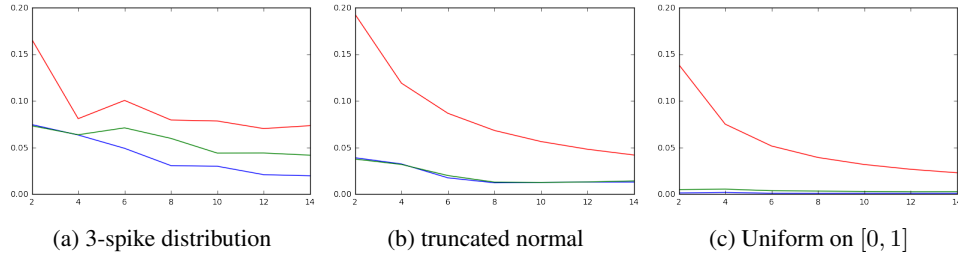


Figure 1: Earth mover’s distance (EMD) between the true underlying distribution  $P_{pop}$  and the distribution recovered by Algorithm 2 for three choices of  $P_{pop}$ : (a) the distribution consisting of equally weighted point masses at locations  $\frac{1}{4}, \frac{1}{2}, \frac{3}{4}$ ; (b) the normal distribution with mean 0.5 and standard deviation 0.15, truncated to be supported on  $[0, 1]$ ; and (c) the uniform distribution over  $[0, 1]$ . For each underlying distributions, we plot the EMD (median over 20 trials) between  $P_{pop}$  and the distribution recovered with Algorithm 2 as  $t$ , the number of samples from each of the  $n$  Bernoulli random variables, takes on all even values from 2 to 14. These results are given for  $n = 10,000$  (green) and  $n = 100,000$  (blue). For comparison, the distance between  $P_{pop}$  and the histogram of the empirical probabilities for  $n = 100,000$  is also shown (red).

Figure 2 shows representative plots of the CDFs of the recovered histograms and empirical histograms for each of the three choices of  $P_{pop}$  considered above.

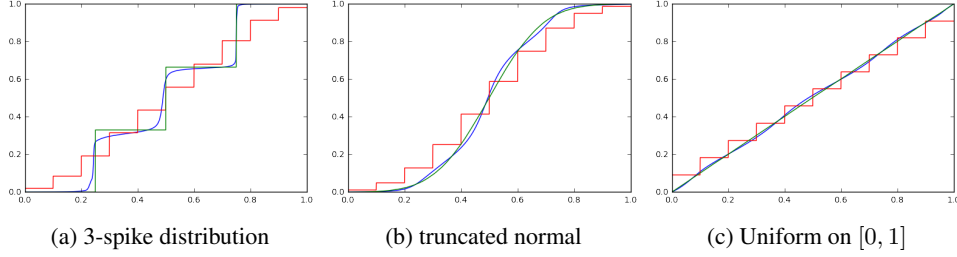


Figure 2: CDFs of the true distribution  $P$  (green), the histogram recovered by Algorithm 2 (blue) for  $P$ , and the empirical histogram (red) corresponding to  $t = 10$  samples and  $n = 100,000$ . Note that the empirical distribution is only supported on multiples of  $\frac{1}{10}$ .

We also considered recovering the distribution of probabilities that different flights are delayed (i.e. each flight—for example *Delta Airlines 123*—corresponds to a parameter  $p \in [0, 1]$  representing the probability that flight is delayed on a given day. Our algorithm was able to recover this non-parametric distribution of flight delay parameters extremely well based on few ( $\leq 10$ ) data points per flight. In this setting, we had access to a dataset with  $> 50$  datapoints per flight, and hence could compare the recovered distribution to a close approximation of the ground truth distribution. These results are included in the appendix.

### 3.2 Distribution of offspring sex ratios

One of the motivating questions for this work was the following naive sounding question: do all members of a given species have the same propensity of giving birth to a male vs female child, or is there significant variation in this probability across individuals? For a population of  $n$  individuals, letting  $p_i$  represent the probability that a future child of the  $i$ th individual is male, this question is precisely the question of characterizing the histogram or set of the  $p_i$ 's. This question of the uniformity of the  $p_i$ 's has been debated both by the popular science community (e.g. the recent BBC article “Why Billionaires Have More Sons”), and more seriously by the biology community.

Meiosis ensures that each male produces the same number of spermatozoa carrying the X chromosome as carrying the Y chromosome. Nevertheless, some studies have suggested that the difference in the amounts of genetic material in these chromosomes result in (slight) morphological differences between the corresponding spermatozoa, which in turn result in differences in their motility (speed of movement), etc. (see e.g. [4, 13]). Such studies have led to a chorus of speculation that the relative timing of ovulation and intercourse correlates with the sex of offspring.

While it is problematic to tackle this problem in humans (for a number of reasons, including sex-selective abortions), we instead consider this question for dogs. Letting  $p_i$  denote the probability that each puppy in the  $i$ th litter is male, we could hope to recover the distribution of the  $p_i$ 's. If this sex-ratio varies significantly according to the specific parents involved, or according to the relative timing of ovulation and intercourse, then such variation would be evident in the  $p_i$ 's. Conveniently, a typical dog litter consists of 4-8 puppies, allowing our approach to recover this distribution based on accurate estimates of these first moments.

Based on a dataset of  $n \approx 8,000$  litters, compiled by the Norwegian Kennel Club, we produced estimates of the first 10 moments of the distribution of  $p_i$ 's by considering only litters consisting of at least 10 puppies. Our algorithm suggests that the distribution of the  $p_i$ 's is indistinguishable from a spike at  $\frac{1}{2}$ , given the size of the dataset. Indeed, this conclusion is evident based even on the estimates of the first two moments:  $\frac{1}{n} \sum_i p_i \approx 0.497$  and  $\frac{1}{n} \sum_i p_i^2 \approx 0.249$ , since among distribution over  $[0, 1]$  with expectation  $1/2$ , the distribution consisting of a point mass at  $1/2$  has minimal variance, equal to  $0.25$ , and these two moments robustly characterize this distribution. (For example, any distribution supported on  $[0, 1]$  with mean  $1/2$  and for which  $> 10\%$  of the mass lies outside the range  $(0.45, 0.55)$ , must have second moment at least  $0.2505$ , though reliably resolving such small variation would require a slightly large dataset.)



### 3.3 Political tendencies on a county level

We performed a case study on the political leanings of counties. We assumed the following model: Each of the  $n = 3116$  counties in the US have an intrinsic “political-leaning” parameter  $p_i$  denoting their likelihood of voting Republican in a given election. We observe  $t = 8$  independent samples of each parameter, corresponding to whether each county went Democratic or Republican during the 8 presidential elections from 1976 to 2004.

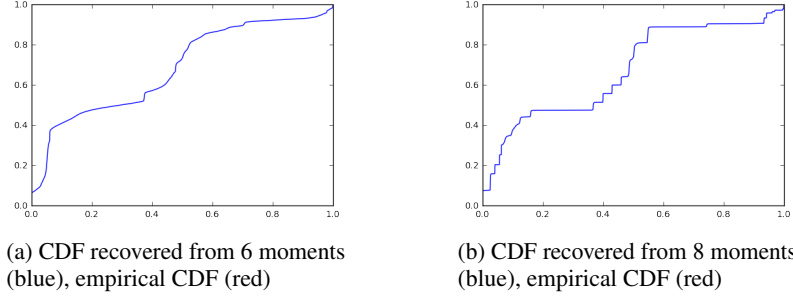


Figure 3: Output of bootstrapping Algorithm 2 on political data for  $n = 3,116$  counties over  $t = 8$  elections.

### 3.4 Game-to-game shooting of NBA players

We performed a case study on the scoring probabilities of two NBA players. One can think of this experiment as asking whether NBA players, game-to-game, have differences in their intrinsic ability to score field goals (in the sports analytics world, this is the idea of “hot / cold” shooting nights). The model for each player is as follows: for the  $i$ th basketball game there is some parameter  $p_i$  representing the player’s latent shooting percentage for that game, perhaps varying according to the opposing team’s defensive strategy. The empirical shooting percentage of a player varies significantly from game-to-game—recovering the underlying distribution or histogram of the  $p_i$ ’s allows one to directly estimate the consistency of a player. Additionally, such a distribution could be used as a prior for making decisions during games. For example, conditioned on the performance during the first half of a game, one could update the expected fraction of subsequent shots that are successful.

The dataset used was the per-game 3 point shooting percentage of players, with sufficient statistics of “3 pointers made” and “3 pointers attempted” for each game. To generate estimates of the  $k^{th}$  moment, we considered games where at least  $k$  3 pointers were attempted. The players chosen were Stephen Curry of the Golden State Warriors (who is considered a very consistent shooter) and Danny Green of the San Antonio Spurs (whose nickname “Icy Hot” gives a good idea of his suspected consistency).

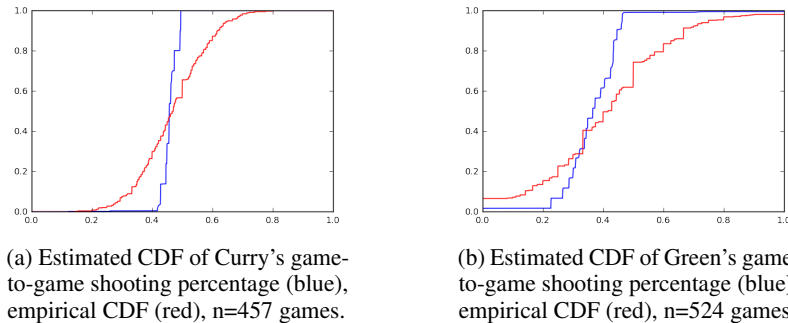


Figure 4: Estimates produced by bootstrapped version of Algorithm 2 on NBA dataset, 8 moments included

## Acknowledgments

We thank Kaja Borge and Ane Nødtvedt for sharing an anonymized dataset on sex composition of dog litters, based on data collected by the Norwegian Kennel Club. This research was supported by NSF CAREER Award CCF-1351108, ONR Award N00014-17-1-2562, NSF Graduate Fellowship DGE-1656518, and a Google Faculty Fellowship.

## References

- [1] Jayadev Acharya, Hirakendu Das, Alon Orlitsky, and Ananda Theertha Suresh. A unified maximum likelihood approach for optimal distribution property estimation. *arXiv preprint arXiv:1611.02960*, 2016.
- [2] Jayadev Acharya, Alon Orlitsky, and Shengjun Pan. Recent results on pattern maximum likelihood. In *Networking and Information Theory, 2009. ITW 2009. IEEE Information Theory Workshop on*, pages 251–255. IEEE, 2009.
- [3] Thomas Bagby, Len Bos, and Norman Levenberg. Multivariate simultaneous approximation. *Constructive Approximation*, 18(4), 2002.
- [4] P. Barlow and C.G. Vosa. The y chromosome in human spermatozoa. *Nature*, 226:961–962, 1970.
- [5] Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A Servedio. Learning poisson binomial distributions. *Algorithmica*, 72(1):316–357, 2015.
- [6] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Properly learning poisson binomial distributions in almost polynomial time. In *Conference on Learning Theory*, pages 850–878, 2016.
- [7] Weihao Kong and Gregory Valiant. Spectrum estimation from samples. *arXiv preprint arXiv:1602.00061*, 2016.
- [8] Nicolai Korneichuk and Nikolaj Pavlovic Korněichuk. *Exact constants in approximation theory*, volume 38. Cambridge University Press, 1991.
- [9] Reut Levi, Dana Ron, and Ronitt Rubinfeld. Testing properties of collections of distributions. *Theory of Computing*, 9(8):295–347, 2013.
- [10] Reut Levi, Dana Ron, and Ronitt Rubinfeld. Testing similar means. *Siam J. Discrete Math*, 28(4):1699–1724, 2014.
- [11] Alon Orlitsky, Narayana P Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. On modeling profiles instead of values. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 426–435. AUAI Press, 2004.
- [12] Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, page 201607774, 2016.
- [13] L.M. Penfold, C. Holt, W.V. Holt, G.R. Welch, D.G. Cran, and L.A. Johnson. Comparative motility of x and y chromosome-bearing bovine sperm separated on the basis of dna content by flow sorting. *Molecular Reproduction and Development*, 50(3):323–327, 1998.
- [14] Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 197–206, Berkeley, Calif., 1956. University of California Press.
- [15] Gregory Valiant and Paul Valiant. 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*, pages 685–694. ACM, 2011.
- [16] Gregory Valiant and Paul Valiant. Estimating the unseen: improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems*, pages 2157–2165, 2013.
- [17] Gregory Valiant and Paul Valiant. Instance optimal learning of discrete distributions. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 142–155. ACM, 2016.
- [18] James Zou, Gregory Valiant, Paul Valiant, Konrad Karczewski, Siu On Chan, Kaitlin Samocha, Monkol Lek, Shamil Sunyaev, Mark Daly, and Daniel G MacArthur. Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nature Communications*, 7, 2016.

## A Lateness in flights

We evaluated our algorithm on flight delays, based on the 2015 Flight Delays and Cancellations dataset. For each of  $n = 25,156$  different flights—where a “flight” is defined via the airline and flight number—we let the corresponding binomial parameter  $p$  correspond to the probability that flight departs at least 15 minutes late. Each flight considered had at least 50 records, and the empirical distribution of lateness parameters was our ground truth distribution  $P_{pop}$ . The estimates were very robust to repeated runs of the experiment, producing CDFs that matched the ground truth extremely closely for all settings of  $t$ .

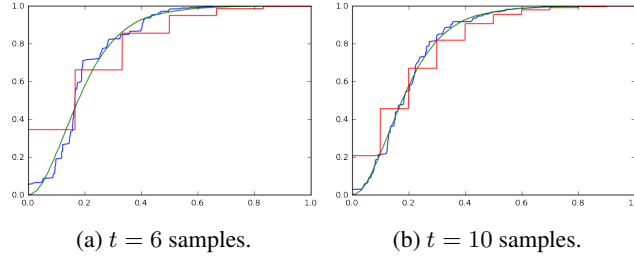


Figure 5: Recovering  $P_{pop}$ . Distributions recovered by bootstrapping Algorithm 2 on 6 and 10 samples of each flight (blue), ground truth distribution (green), and empirical distribution (red) shown.

## B Proof of Theorem 3, the Wasserstein distance bound

**Theorem 3** For two distributions  $P$  and  $Q$  supported on  $[0, 1]$  whose first  $t$  moments are  $\alpha$  and  $\beta$  respectively, the Wasserstein distance  $\|P - Q\|_W$  is bounded by  $\frac{\pi}{t} + 3^t \sum_{k=1}^t |\alpha_k - \beta_k|$ .

*Proof.* The natural approach to bounding the Wasserstein distance,

$$\sup_{f \in Lip_1} \int f(x) (P(x) - Q(x)) dx,$$

is to argue that for any Lipschitz function,  $f$ , there is a polynomial  $P_f$  of degree at most  $k$  that closely approximates  $f$ . To see this,

$$\begin{aligned} & \int_0^1 f(x)(P(x) - Q(x))dx \\ & \leq \int_0^1 |p_f(x) - f(x)|(P(x) - Q(x))dx + \int_0^1 p_f(x)(P(x) - Q(x))dx \\ & \leq 2\|f - p_f\|_\infty + \sum_{k=1}^t c_k(\alpha_k - \beta_k), \end{aligned}$$

where  $c_k$  be the coefficient of the degree- $k$  term of polynomial  $p_f$ . Hence all that remains is to argue that there is a good degree  $k$  polynomial approximation of any Lipschitz function  $f$ .

For convenience of the analysis, we generalize the domain of  $f$  from  $[0, 1]$  to  $[-1, 1]$  by letting  $f(-x) = f(x)$ . We further define function  $\phi(\theta) = f(\cos(\theta))$  which also has Lipschitz constant 1 since the cosine function has Lipschitz constant 1. Now we are ready to apply Theorem 4.2.1 of [8] to  $\phi(\theta)$ , which states that for any periodic- $2\pi$  function with Lipschitz constant 1 can be approximate by a degree  $t$  trigonometric polynomials with  $l_\infty$  approximation error  $\frac{K_1}{t} = \frac{\pi}{2t}$  where  $K_1$  is Favard constant which is equal to  $\frac{\pi}{2}$ . Let  $U_n(\theta)$  be the degree  $t$  trigonometric polynomials that achieves the stated approximation error. WLOG, by Proposition 2.1.6 of [8], we may assume  $U_n(\theta)$  is even. The algebraic polynomial to approximate  $f(x)$  can be defined as  $p_f(x) = U_t(\arccos(x))$  which again has degree  $t$ . Hence we have shown that  $\|f - p_f\|_\infty \leq \frac{\pi}{2t}$  and what remains is to bound the magnitude of  $c_k$ .

The plan is to first obtain sharp bound of the coefficients of the trigonometric polynomials  $U_t(\theta)$  explicitly, after which  $c_k$  can be bounded by being expressed in terms of these coefficients. Notice that the coefficient of term  $\cos(k\theta)$  in  $U_t(\theta)$ , denoted as  $u_k$ , is  $a_k \lambda_k^t$  by Formula 1.1 in Chapter 4 of [8] where  $a_k = \frac{1}{\pi} \int_0^{2\pi} \phi(\theta) \cos(k\theta) d\theta$  and  $\lambda_k^t = \frac{k\pi}{2(t+1)} \frac{1}{\tan(\frac{k\pi}{2(t+1)})}$  by Formula 1.42 in Chapter 4 of [8].

Given that  $\tan(x) \geq x$  for  $0 \leq x \leq \frac{\pi}{2}$  and  $\frac{k\pi}{2(t+1)} < \frac{\pi}{2}$ , we have  $\frac{1}{\tan(\frac{k\pi}{2(t+1)})} \leq \frac{2(t+1)}{k\pi}$  and hence  $0 \leq \lambda_k^t \leq 1$ . In order to bound  $a_k$ , notice that WLOG, we may assume  $\|f\|_\infty \leq 1/2$  and  $\|\phi\|_\infty \leq 1/2$  since  $f$  is Lipschitz-1. Hence  $|a_k| = \frac{1}{\pi} |\int_0^{2\pi} \phi(\theta) \cos(k\theta) d\theta| \leq \frac{1}{2\pi} \int_0^{2\pi} |\cos(k\theta)| d\theta \leq 1$ . We have shown that for all  $k$ ,  $u_k$  is at most 1.

The algebraic polynomial  $U_t(\arccos(x))$  can be expressed as  $\sum_{k=1}^t u_k T_k(x)$  where  $T_k(x)$  is Chebyshev polynomials of the first kind. Note the recurrence relation for Chebyshev polynomials given by  $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$ ,  $T_0(x) = 1$ ,  $T_1(x) = x$ , for the  $i$ th polynomial, we can loosely bound the magnitude of any of its coefficients by  $3^{i-1}$ . Since  $|u_i| < 1$  for all  $i$ , the magnitude of coefficient  $c_k$  can be upper bounded by  $\sum_{i=1}^t 3^{i-1} \leq 3^t$ . Thus, we have shown that:

$$\int_0^1 f(x)(P(x) - Q(x))dx \leq \frac{\pi}{t} + 3^t \sum_{k=1}^t |\alpha_k - \beta_k|.$$

□

## C Proof of Theorem 1

In this section, we prove the main theorem of our paper, Theorem 1, which establishes guarantees of the estimation accuracy of our algorithm. Before proving our main theorem, we first prove Lemma 1, the properties of our moment estimators:

**Lemma 1** *Given  $\{p_1, \dots, p_n\}$ , let  $X_i$  denote the random variable distributed according to  $\text{Binomial}(t, p_i)$ . For  $k \in \{1, \dots, t\}$ , let  $\alpha_k = \frac{1}{n} \sum_{i=1}^n p_i^k$  denote the  $k^{\text{th}}$  true moment, and*

*$\beta_k = \frac{1}{n} \sum_{i=1}^n \frac{\binom{X_i}{k}}{\binom{t}{k}}$  denote our estimate of the  $k$ th moment. Then*

$$\mathbb{E}[\beta_k] = \alpha_k, \text{ and } \Pr(|\beta_k - \alpha_k| \geq \epsilon) \leq 2e^{-\frac{1}{3}n\epsilon^2}.$$

*Proof.* First we show that for each  $i$  we have  $\mathbb{E}[\binom{X_i}{k}] = p_i^k \binom{t}{k}$ , then the claim  $\mathbb{E}[\beta_k] = \alpha_k$  holds trivially due to the additivity of expectation. Notice that the numerator counts the number of subsets of size  $k$  that are all 1, and the denominator is the number of subsets of size  $k$ . The probability that a certain subset of size  $k$  is all 1 is exactly  $p_i^k$ . Hence the claim about the expectation holds.

By Bernstein's Inequality, when  $\epsilon \leq 1$ ,  $\Pr(|\beta_k - \alpha_k| \geq \epsilon) \leq 2e^{-\frac{3}{8}n\epsilon^2} \leq 2e^{-\frac{1}{3}n\epsilon^2}$  holds. We have proved the claim about concentration. □

We are now ready to prove Theorem 1. For convenience, we restate the theorem:

**Theorem 1** *Consider a set of  $n$  probabilities,  $p_1, \dots, p_n$  with  $p_i \in [0, 1]$ , and suppose we observe the outcome of  $t$  independent flips of each coin, namely  $X_1, \dots, X_n$ , with  $X_i \sim \text{Binomial}(t, p_i)$ . There is an algorithm that produces a distribution  $Q$  supported on  $[0, 1]$ , such that with probability at least  $1 - \delta$  over the randomness of  $X_1, \dots, X_n$ ,*

$$\|P - Q\|_W \leq \frac{\pi}{t} + 3^t \sum_{i=1}^t \sqrt{\ln\left(\frac{2t}{\delta}\right) \frac{3}{n}} \leq \frac{\pi}{t} + O_\delta\left(\frac{3^t t \ln t}{\sqrt{n}}\right),$$

*where  $P$  denotes the distribution that places mass  $\frac{1}{n}$  at value  $p_i$ , and  $\|\cdot\|_W$  denotes the Wasserstein distance.*

*Proof.* Given Lemma 1, we obtain the fact that, with probability at least  $1 - \delta$ , the events  $|\alpha_k - \beta_k| \leq \sqrt{\ln\left(\frac{2t}{\delta}\right) \frac{3}{n}}$  simultaneously occur for all  $k \in \{1, \dots, t\}$ . Applying Theorem 3 yields the claimed accuracy guarantee. □

## D Proof of Proposition 1, the information-theoretic lower bound

In this section, we prove Proposition 1 establishing the tightness of the  $\Theta(1/t)$  dependence in our recovery guarantees. For convenience, we restate the proposition:

**Proposition 1** *Let  $P_{pop}$  denote a distribution over  $[0, 1]$ , and for positive integers  $t$  and  $n$ , let  $X_1, \dots, X_n$  denote independent random variables with  $X_i$  distributed as  $\text{Binomial}(t, p_i)$  where  $p_i$  is drawn independently according to  $P_{pop}$ . An estimator  $f$  maps  $X_1, \dots, X_n$  to a distribution  $f(X_1, \dots, X_n)$ . Then, for every fixed  $t$ , the following lower bound on the accuracy of any estimator holds for all  $n$ :*

$$\inf_f \sup_{P_{pop}} \mathbb{E} [\|f(X_1, \dots, X_n) - P_{pop}\|_W] > \frac{1}{4t}.$$

Our proof will leverage the following result from [7] which states that there exists a pair of distributions supported on  $[0, 1]$  whose first  $t$  moments agree, but have Wasserstein distance  $> 1/2t$ :

**Lemma 3.** *For any  $t$ , there exists a pair of distributions  $D_P, D_Q$  supported on  $[0, 1]$  that each consist of  $O(t)$  point masses, such that  $D_P$  and  $D_Q$  have identical first  $t$  moments, and  $\|D_P - D_Q\|_W > \frac{1}{2t}$*

*Proof of Proposition 1.* Consider the distributions  $D_P$  and  $D_Q$  whose existence is guaranteed by Lemma 3. Consider the distribution of  $X_i$ , where  $X_i$  is drawn by first drawing  $p_i$  according to  $D_P$ , and then drawing  $X_i \leftarrow \text{Binomial}(p_i, t)$ . Similarly, let  $Y_i$  denote the random variable defined by drawing  $q_i$  from  $D_Q$  and then drawing  $Y_i \leftarrow \text{Binomial}(q_i, t)$ .

We now claim that the distribution of  $X_i$  and  $Y_i$  are identical, and hence, for every  $n$ , the joint distribution of  $(X_1, \dots, X_n)$  is identical to that of  $(Y_1, \dots, Y_n)$ , and hence they cannot be distinguished.

Indeed, the distributions of  $X_i$  and  $Y_i$  are given by:

$$\begin{aligned} \mathbb{P}(X_i = k) &= \int_0^1 \binom{t}{k} x^k (1-x)^{t-k} D_P(x) dx \\ \mathbb{P}(Y_i = k) &= \int_0^1 \binom{t}{k} x^k (1-x)^{t-k} D_Q(x) dx \end{aligned}$$

Noting that the integrand is a degree- $t$  polynomial, and that  $D_P$  and  $D_Q$  have the same first  $t$  moments yields the conclusion that these two distributions are identical.

To conclude, note that if we are given  $(Z_1, \dots, Z_n)$  with the promise that, with probability  $1/2$ , they correspond to  $D_P$  and with probability  $1/2$  they correspond to  $D_Q$ , then no algorithm can correctly guess which of these distributions they were drawn from, with probability of success greater than  $1/2$ , and hence no estimator can achieve an expected error of recovery better than  $\frac{1}{2} \|D_P - D_Q\|_W > \frac{1}{4t}$ , as desired.  $\square$

## E Proof of Theorem 2, multivariate setting

The prove of Theorem 2 will be identical to Theorem 1, except that we will need the following slightly stronger version of Lemma 2:

**Lemma 2** *Given any Lipschitz function  $f$  supported on  $[0, 1]^d$ , there is a degree  $t$  polynomial  $p(x) = \sum_{|\alpha| \leq t} c_\alpha x^\alpha$  where  $\alpha$  is multi-index  $\{\alpha_1, \alpha_2, \dots, \alpha_d\}$  such that*

$$\sup_{x \in [0, 1]^d} |p(x) - f(x)| \leq \frac{C_d}{t}, \quad (1)$$

and  $c_\alpha \leq A_d \frac{(2t)^d 2^t}{3^{|\alpha|}}$ .

*Proof.* This polynomial approximation lemma is basically a restatement of Theorem 1 in [3]. What we need to do is only to give an explicit upper bound of the coefficients.

The high level idea is to first convolve  $f$  with a holomorphic bump function  $G$  which gives  $H = f * G$ , then the Maclaurin series of  $H$  is a good polynomial approximation of  $H$  and also  $f$ .

By the definition of Maclaurin series, the coefficient  $c_\alpha = \frac{\partial^\alpha H(0)}{|\alpha|!}$ . Suppose  $H$  is holomorphic on an open neighborhood of some polydisk  $E_S$  with radius  $S$ , assuming  $\sup_{z \in E_S} |H(z)| \leq M$ , by Cauchy's integral formula, we have  $|c_\alpha| = |\frac{1}{2\pi i} \oint_{|z|=S} \frac{H(z)}{z^{|\alpha|+1}}| \leq \frac{M}{S^{|\alpha|}}$ . By the definition of  $R$  in the proof of Theorem 1 in [3], we can set  $R = 1$  such that function  $f$  is supported on box  $B_R$ . Let  $S = 2R + 1 = 3$  and follow all the parameter settings, by Equation 14 in [3], we have  $|c_\alpha| \leq \frac{M}{S^{|\alpha|}} \leq A_d \frac{(2t)^d (t+1)2^t}{t3^{|\alpha|}} \leq A_d \frac{(2t)^d 2^t}{3^{|\alpha|}}$ , where  $A_d$  is a constant that depends on  $d$ .  $\square$