Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under Wasserstein distance

Yanjun Han   YJHAN@STANFORD.EDU
Department of Electrical Engineering, Stanford University

Jiantao Jiao   JIANTAO@STANFORD.EDU
Department of Electrical Engineering, Stanford University

Tsachy Weissman   TSACHY@STANFORD.EDU
Department of Electrical Engineering, Stanford University

Abstract

We present *Local Moment Matching (LMM)*, a unified methodology for symmetric functional estimation and distribution estimation under Wasserstein distance. We construct an efficiently computable estimator that achieves the minimax rates in estimating the distribution up to permutation, and show that the plug-in approach of our unlabeled distribution estimator is "universal" in estimating symmetric functionals of discrete distributions. Instead of doing best polynomial approximation explicitly as in existing literature of functional estimation, the plug-in approach conducts polynomial approximation implicitly and attains the optimal sample complexity for the entropy, power sum and support size functionals.

Keywords: Distribution Estimation; Functional Estimation; Minimax Risk; Wasserstein Distance

1. Introduction and Main Results

Given n independent samples from a discrete distribution $P = (p_1, \cdots, p_S)$, we aim to estimate the distribution vector P up to permutation. In other words, let $P^< = (p_{(1)}, p_{(2)}, \cdots, p_{(S)})$ be the sorted version of P (i.e., $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(S)}$ are the order statistics of P), we would like to find an estimator $\hat{P}$ which comes close to minimizing the sorted ℓ_1 distance

$$\mathbb{E}_P \| \hat{P} - P^< \|_1$$

in the minimax sense.

Our study of estimating the sorted distribution $P^<$ is motivated by the following facts:

1. The sorted distribution $P^<$ can be interpreted as the distribution P up to permutation, or the multiset of probabilities in $P = (p_1, p_2, \ldots, p_S)$, or the "tail" of a distribution. In economics, the theory of *long tail* Anderson (2004) emphasizes the significance of products in the *tail*, and inferring the sorted distribution precisely shows the shape of the tail.

2. Estimating the sorted distribution $P^<$ turns out to require significantly less number of samples than that required to estimate the distribution P under the same ℓ_1 loss, as shown by Valiant and Valiant (2011a) using a different Wasserstein loss function.

3. The sorted distribution estimate proves to be useful in estimating symmetric functionals of the distribution, which are defined as functionals of P that can also be viewed as functionals of $P^<$. Indeed, Valiant and Valiant (2011a, 2013); Acharya et al. (2016) constructed estimators of the sorted distribution can be used to plug-in some symmetric functionals to achieve the information theoretic limits in certain parameter regimes, which performed significantly better than the approach of plugging-in the empirical distribution.

4. The whole distribution P can be decomposed into two parts: the sorted distribution $P^<$ and the permutation from $P^<$ to P. Being able to design computationally efficient schemes to achieve the information theoretic limit in estimating $P^<$ sheds light on the general question of inferring parameters up to group transformations, which is of fundamental significance in statistics and machine learning Kong and Valiant (2017); Tian et al. (2017).

The main idea to solve the traditional distribution estimation problem (i.e., estimating P) is to use the empirical frequency and/or its variants, which turn out to be minimax optimal for various loss functions including ℓ_2 Steinhaus (1957); Trybula (1958); Rutkowska (1977); Olkin and Sobel (1979), ℓ_1 Daskalakis et al. (2012); Diakonikolas (2014); Han et al. (2015); Kamath et al. (2015) and KL loss Kamath et al. (2015). To consistently estimate P, usually it is required to observe each symbol $i \in [S]$ sufficiently many times on average; for example, $n \gg S$ is a necessary and sufficient condition for the existence of an estimator which estimates P within a vanishing ℓ_1 error Han et al. (2015). However, recent studies suggested that estimating $P^<$ might be significantly easier than estimating P: compared with an oracle with the same observation $X_1, \cdots, X_n$ and perfect knowledge of $P^<$, there still exists some estimator which performs nearly as well as the oracle even if $S = \infty$ under the ℓ_1 loss Valiant and Valiant (2015) and KL loss Orlitsky and Suresh (2015). This observation shows that the "labeling" from $P^<$ to P is the difficult step in estimating P, and estimating the sorted distribution $P^<$ may only require sub-linear samples (i.e., $n \ll S$).

Two main approaches have been proposed in literature to estimate the sorted distribution $P^<$. One is the approach of *profile maximum likelihood (PML)* Orlitsky et al. (2004); Acharya et al. (2009), which aims at solving the sorted distribution that maximizes the likelihood of observing the *sorted* empirical distribution. It is not clear how to solve the corresponding optimization problem efficiently. Algorithms that approximately solve the PML have been proposed in the literature, including Orlitsky et al. (2004); Vontobel (2012); Pavlichin et al. (2017), without clear theoretical approximation guarantees. It was shown in Acharya et al. (2016) that plugging-in the profile maximum likelihood distribution into a variety of symmetric functionals (namely, entropy, support size, support coverage, and distance to uniformity) achieves the information theoretic limit when the number of samples is not "too" large.

A different approach, which initiated from Efron and Thisted (1976), proposed to use linear programming to find a sorted distribution that was consistent with the observed frequency counts. This approach was adapted and rigorously analyzed in Valiant and Valiant (2011a, 2013) under a Wasserstein distance loss function, where it was shown that plugging-in the inferred sorted distribution from the linear program into certain symmetric functionals (namely, entropy, support size, support coverage, and distance to uniformity) results in estimators that achieve the information theoretic limit in the constant error regime.

Various questions remain unsolved given existing literature. It is not clear how to efficiently provably solve the PML, and the proof of the optimality of PML in the plug-in machinery of symmetric functional estimation heavily relies on the fact that the observations can only take values in

a finite set. It does not apply to the Gaussian setting, where one observes a Gaussian random vector $X \sim \mathcal{N}(\mu, I_p)$, and would like to estimate the sorted version of the mean vector μ. For the linear programming approach, it was not shown to achieve the optimal dependence on ϵ in entropy estimation, and it was not clear whether it is near-optimal if we plug it in other functionals. Indeed, the general achievability proof is done through a Lipschitz continuity argument in Valiant and Valiant (2011a), and it was not clear whether they can match the lower bounds for individual functionals.

The main mathematical reason that motivated this paper is to develop relations between estimation of (nonsmooth) functionals of distributions, and estimation of the sorted distribution. In the first realm, Lepski et al. (1999) considered the problem of L_r norm estimation in Gaussian noise model and utilized Fourier approximation theory, while Cai and Low (2011) considered estimating the ℓ_1 norm of normal mean and applied best polynomial approximation. The work of Valiant and Valiant (2011b) developed "Chebyshev bump" based approximation and proposed linear estimators that achieve the optimal dependence on ϵ in estimating the entropy, distance to uniformity, and support size when the sample size n is not too "large". Minimax rates for estimation of a large variety of functionals were solved in the past few years, including entropy Wu and Yang (2016a); Jiao et al. (2015), Rényi entropy Acharya et al. (2015), support size Wu and Yang (2015), support coverage Orlitsky et al. (2016), distinct elements Wu and Yang (2016b), L_1 distance Jiao et al. (2016), Kullback–Leibler divergence Bu et al. (2016); Han et al. (2016), squared Hellinger divergence Han et al. (2016), χ^2 divergence Han et al. (2016), support coverage from multiple populations Raghunathan et al. (2017), L_r norm of a regression function in Gaussian white noise Han et al. (2017a), and differential entropy Han et al. (2017b). The latest batch of work have developed essentially a framework of proving minimax upper and lower bounds for functional estimation problems, which was called the *Approximation* approach in Jiao et al. (2015). The main idea is, we first use concentration inequalities to "zoom in" sets that are guaranteed to contain the true parameters with overwhelming probability, and then apply unbiased estimators of (best) approximation polynomials up to a certain degree in those sets. The minimax lower bounds are proved using the dual representation of best polynomial approximation over each individual sets that we may "zoom in". For a crisp illustration of the lower bound technique, we refer the readers to Jiao et al. (2017).

The *Approximation* approach requires to compute deterministic approximations for each individual functional separately, and is naturally a *non-plug-in* approach. It was shown in Jiao et al. (2017) that indeed any plug-in approach cannot hope to completely replace the *Approximation* approach: there exist certain functionals such that *any* plug-in approach fails to achieves the statistical limit.

This paper aims at bridging the *Approximation* approach illustrated above and the plug-in approach. We start with the following question.

Question 1 *Find an estimator of the sorted distribution that satisfies the following properties:*

1. *Plugging-in the estimator into a large variety of symmetric functionals achieves the information theoretic limit;*

2. *It has a clear mathematical correspondence with the* Approximation *approach (hence generalizable to Gaussian settings);*

3. *It achieves the minimax rates in estimating sorted distribution under ℓ_1 loss;*

4. *It is efficiently computable.*

We present *Local Moment Matching* (LMM), an approach that provably answers the question above.

The intuition behind LMM, on the highest level, is the following. It shares with *Approximation* the same first step, which is to use concentration inequalities to "zoom in" the smallest sets that are guaranteed to contain the true parameters with overwhelming probability. The difference appears in the second step, while *Approximation* tries to find a polynomial that closely approximates the functional over the specific set and then use unbiased estimators to estimate the polynomial, LMM aims to find a set of numbers, whose total number equals to the number of true parameters we believe are in the set, whose *moments* match the *unbiased* estimates of the moments of true parameters. In other words, *Approximation* conducts an *explicit* approximation of functional, and LMM conducts and *implicit* approximation, since LMM needs to achieve statistical optimality for a large variety of symmetric functionals.

The idea of combining moment matching and linear programming appeared before in Kong and Valiant (2017); Tian et al. (2017). However, the key contribution of our work is the feature of *local* rather than *global* moment matching in Kong and Valiant (2017); Tian et al. (2017). The advantage and necessity of *locality* can be seen from the following thought experiment. Suppose $p_1 = 0.1, p_2 = 0.1 + \epsilon, p_3 = 0.3$. Given sufficiently many number of samples, it is easy to infer that the probability corresponding to symbol 3 is larger than the probability corresponding to symbol 1, but it may be unclear whether $p_1 \geq p_2$ or $p_1 \leq p_2$ for ϵ small enough. The *global* moment matching approach tries to find a sorted distribution that matches the moments of (p_1, p_2, p_3), while *local* moment matching tries to only match the moments corresponding to (p_1, p_2). In other words, the *local* approach only uses linear program and moment matching when there is ambiguity about the relative magnitude of the true probabilities, while *global* moment matching discards the information we already know (such as $p_3 > p_1$) thus behaves sub-optimally unless the number of samples is very small. Indeed, it is the reason why Tian et al. (2017) is only statistically optimal for very small n ($n \leq \ln S$ in the notation of this paper). We also mention that the setting in Tian et al. (2017) is not entirely identical to ours, since in Tian et al. (2017) $\sum_{i=1}^{S} p_i$ is not necessarily one.

We present our main results below.

Theorem 1 *For $n \gtrsim \frac{S}{\ln S}$, we have[1]*

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P^<\|_1 \asymp \sqrt{\frac{S}{n \ln n}} + \tilde{\Theta}\left(\sqrt{\frac{S}{n}} \wedge n^{-\frac{1}{3}}\right).$$

Furthermore, the estimator $\hat{P}$ constructed in Section 2 does not require the knowledge of the support size S.

The following corollary is immediate.

Corollary 2 *There exists an estimator $\hat{P}$ for the unlabeled distribution $\hat{P}$ under sorted ℓ_1 loss if and only if $n \gg \frac{S}{\ln S}$.*

Corollary 2 shows that as opposed to the requirement $n = \omega(S)$ in consistently estimating P, estimating the sorted distribution $P^<$ only requires sub-linear samples $n = \omega(\frac{S}{\ln S})$. The following corollary shows that when the support size S is not too small, the empirical frequency P_n which is minimax optimal for estimating P is no longer optimal in estimating $P^<$.

---

1. Notation $a_n = \tilde{\Theta}(b_n)$ means that for any $\epsilon > 0$, we have $n^{-\epsilon} b_n \ll a_n \ll n^\epsilon b_n$.

Corollary 3 *For $n \gtrsim \frac{S}{\ln S}$, the minimax rate-optimal estimator $\hat{P}$ for $P^<$ outperforms the sorted empirical distribution $P_n^<$ if and only if $S \gg \tilde{\Theta}(n^{\frac{1}{3}})$, where P_n denotes the empirical distribution.*

Corollary 3 shows that the minimax rate-optimal estimator outperforms the baseline (i.e., the sorted empirical distribution) when $\frac{S}{\ln S} \lesssim n \ll \tilde{\Theta}(S^3)$. The constraint that the sample size n cannot be too large is indeed natural: for larger sample size, there is not enough ambiguity between the relative magnitude of probabilities of each symbol, and the problem of estimating sorted distribution is essentially reduced to that of estimating the original unsorted distribution. Specifically, we show that the error $\Theta(\sqrt{\frac{S}{n \ln n}})$ can be achieved if and only if $\frac{S}{\ln S} \lesssim n \lesssim \tilde{\Theta}(S^3)$; in contrast, Valiant and Valiant (2017) can only achieve it under a different Wasserstein distance when $\frac{S}{\ln S} \lesssim n \lesssim S$.

We then demonstrate the performance of plugging-in our sorted distribution estimate into certain symmetric functionals of the following form:

$$F(P) = \sum_{i=1}^{S} f(p_i), \qquad f \in C[0,1], f(0) = 0. \tag{1}$$

We will be mainly interested in the case where f is non-smooth and the estimation of $F(P)$ becomes challenging. Concretely, we consider entropy $H(P)$, the power sum function $F_\alpha(P)$ and the support size $S(P)$, which are given by

$$H(P) \triangleq \sum_{i=1}^{S} p_i \ln \frac{1}{p_i}, \qquad P \in \mathcal{M}_S,$$

$$F_\alpha(P) \triangleq \sum_{i=1}^{S} p_i^\alpha, \qquad P \in \mathcal{M}_S, 0 < \alpha < 1,$$

$$S(P) \triangleq \sum_{i=1}^{S} \mathbb{1}(p_i \neq 0), \qquad P \in \mathcal{D}_k \triangleq \left\{ P \in \mathcal{M}_S : p_i \geq \frac{1}{k}, i \in [S] \right\}.$$

The exact minimax rates for these functionals have been obtained in Wu and Yang (2016a); Jiao et al. (2015); Wu and Yang (2015), respectively. For these functionals, the plug-in approach $F(\hat{P})$ of the estimator $\hat{P}$ of the sorted distribution $P^<$ (as in Theorem 12) achieves the corresponding minimax risk when the sample size n is not too large.

Theorem 4 (Informal) *For $F(P) = H(P), F_\alpha(P)$ with $\alpha \in (0,1)$ and $S(P)$, the plug-in estimator $\hat{F} = F(\hat{P})$ achieves the corresponding minimax risk when n is not too large. In particular, $\hat{F}$ attains the optimal sample complexity $n \gg \frac{S}{\ln S}$ and $n \gg \frac{k}{\ln k}$ to achieve a vanishing error.*

We refer the explicit construction of the plug-in estimator $\hat{F} = F(\hat{P})$ and the precise statement of Theorem 4 to Section 3. In summary, Theorem 4 provides an affirmative answer to Question 1 and shows that our estimator $\hat{P}$ for the sorted distribution $P^<$ is "universal" in the sense that the plug-in approach yields a near-minimax estimator for various functionals.

The rest of this paper is organized as follows. Section 2 relates the problem of estimating sorted distribution $P^<$ to the distribution estimation problem under Wasserstein distance, where the idea of local moment matching is motivated and the final estimator $\hat{P}$ is constructed. Section 3 discusses the application of sorted distribution estimation to symmetric functional estimation in detail, and

proves Theorem 4. Theorem 1 is then proved via a combination of the achievability part in Section A and the converse part in Section B. Some auxiliary lemmas and their proofs are deferred in the appendices.

Notation: For a finite set A, let $|A|$ denote its cardinality; $[n] \triangleq \{1, \cdots, n\}$; lattice operations $\wedge, \vee$ are defined as $a \wedge b = \min\{a, b\}, a \vee b = \max\{a, b\}$; let $\mathcal{M}_S$ denote the probability simplex over S elements, and Poly_K be the space of all polynomials of degree at most K; for non-negative sequences $\{a_n\}$ and $\{b_n\}$, the notation $a_n \lesssim b_n$ (or $b_n \gtrsim a_n, a_n = O(b_n), b_n = \Omega(a_n)$) means $\limsup_{n\to\infty} \frac{a_n}{b_n} < \infty$, and $a_n \ll b_n$ ($b_n \gg a_n, a_n = o(b_n), b_n = \omega(a_n)$) means $\limsup_{n\to\infty} \frac{a_n}{b_n} = 0$, and $a_n \asymp b_n$ (or $a_n = \Theta(b_n)$) is equivalent to both $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

2. Estimator Construction

In this section, we make use of the duality in Wasserstein distance to relate sorted distribution estimation to the estimation of Lipschitz functionals, and introduce the duality between moment matching and polynomial approximation. Based on these insights, finally we construct the estimator via local moment matching.

2.1. Duality of Wasserstein Distance

We first introduce the Wasserstein distance.

Definition 5 (Wasserstein Distance) *Let (S, d) be a separable metric space, and P, Q be two Borel probability measures on S. The Wasserstein distance between P, Q is defined as*

$$W(P, Q) \triangleq \inf_{\mathcal{L}(X)=P, \mathcal{L}(Y)=Q} \mathbb{E}[d(X, Y)],$$

where the infimum is taken over all possible couplings between S-valued random variables X, Y with marginals P and Q, respectively.

The key reason why we introduce the Wasserstein distance lies on the following lemma.

Definition 6 *For any vector $P = (p_1, \cdots, p_S)$, we define μ_P to be the uniform probability measure on the multiset $\{p_1, \cdots, p_S\}$.*

Lemma 7 *For any two vectors P, Q, we have*

$$\|P^< - Q^<\|_1 = S \cdot W(\mu_P, \mu_Q)$$

with $d(x, y) = |x - y|$ being the usual Euclidean metric.

In other words, in order to estimate the sorted distribution $P^<$ in terms of the ℓ_1 distance, it is equivalent to finding some distribution $\hat{P}$ such that the Wasserstein distance between μ_P and $\mu_{\hat{P}}$ is small. However, $\mu_{\hat{P}}$ must be a discrete measure, which may complicate the estimator construction. Fortunately, the following randomization procedure and Lemma 9 show that it also suffices to find any (possibly non-atomic) probability measure $\hat{\mu}_P$ over the real line such that $W(\mu_P, \hat{\mu}_P)$ is small:

Definition 8 (Randomized Discretization) *Given a support size S and any probability measure μ over the real line, the following procedure outputs an S-dimensional vector $Q = (q_1, \cdots, q_S)$:*

1. *Let F be the CDF of μ, and $U_1, \cdots, U_S$ be S independent random variables uniformly distributed on $[0, 1]$;*

2. *For each $i = 1, \cdots, S$, define $q_i \triangleq F^{-1}\left(\frac{i-U_i}{S}\right)$, where the inverse $F^{-1}(\cdot)$ is defined as $F^{-1}(t) \triangleq \sup\{x : F(x) \leq t\}$.*

3. *Finally, form the vector $Q = (q_1, \cdots, q_S)$.*

Lemma 9 *Let P be an S-dimensional vector, and μ be any probability measure on $\mathbb{R}$. If Q is the returned vector of the previous randomization procedure, we have*

$$\mathbb{E}W(\mu_P, \mu_Q) = W(\mu_P, \mu)$$

where the expectation is taken with respect to the randomness in the randomized procedure.

To find a suitable probability measure $\hat{\mu}$ such that the Wasserstein distance $W(\mu_P, \hat{\mu})$ is small, it will be helpful to recall the well-known dual representation of the Wasserstein distance:

Lemma 10 *Kantorovich and Rubinstein (1958) For two Borel probability measures P, Q on a separable metric space (S, d), the following duality result holds:*

$$W(P, Q) = \sup_{f : \|f\|_{\mathrm{Lip}} \leq 1} \mathbb{E}_P f(X) - \mathbb{E}_Q f(X)$$

where X is a random variable taking value in S with distribution P or Q, and the Lipschitz norm is defined as $\|f\|_{\mathrm{Lip}} \triangleq \sup_{x \neq y \in S} \frac{|f(x) - f(y)|}{d(x,y)}$.

The previous lemma shows that we need to find some $\hat{\mu}$ such that $\mathbb{E}_{\hat{\mu}} f(X)$ is close to $\mathbb{E}_{\mu_P} f(X)$ for any real-valued function f with Lipschitz norm at most one. Moreover, by definition of μ_P, we have $\mathbb{E}_{\mu_P} f(X) = S^{-1} \sum_{i=1}^S f(p_i)$ for any function f. In other words, we need to tackle the problem of *functional estimation* of the form $\sum_{i=1}^S f(p_i)$ for all 1-Lipschitz functions *simultaneously*. There are two fundamental difficulties in this problem:

1. Estimation of functionals is hard in general;

2. The space of all 1-Lipschitz functions is infinite dimensional.

The next subsection will be devoted to overcoming these two difficulties.

2.2. Duality between Moment Matching and Approximation

In this subsection we present answers to the previous questions. The first step is to estimate the functional of the form $\sum_{i=1}^S f(p_i)$ for some fixed 1-Lipschitz function f. It may be tempted to use the plug-in approach $\sum_{i=1}^S f(\hat{p}_i)$, where $\hat{p}_i$ denotes the empirical probability of the symbol i. Note that this approach will return the empirical distribution as the distribution estimate in the end. However, it has been shown in previous works (e.g., Jiao et al. (2015)) that *bias* is the dominating error in the estimation of functionals, and the plug-in approach incurs too much bias. This observation motivates us to look for proper functions f such that there exists an unbiased estimator of

$\sum_{i=1}^{S} f(p_i)$, and the sufficient and necessary condition is that f must be a polynomial of degree at most n. Specifically, for $n\hat{p} \sim \mathsf{B}(n, p)$, we have

$$\mathbb{E}\left[\frac{n\hat{p}(n\hat{p}-1)\cdots(n\hat{p}-k+1)}{n(n-1)\cdots(n-k+1)}\right] = p^k, \qquad k = 0, 1, \cdots, n.$$

Hence, polynomials are *easy* functionals for estimation, and we may restrict the function f to be polynomials up to a certain degree.

The next step is to resolve the problem that the space of all 1-Lipschitz functions is infinite dimensional. However, if we could accurately estimate $\sum_{i=1}^{S} f(p_i)$ for monomials, i.e., $\sum_{i=1}^{S} \hat{p}_i^k \approx \sum_{i=1}^{S} p_i^k$ for $k = 0, 1, \cdots, K$, then for any 1-Lipschitz function f and any polynomial P of degree at most K, we have

$$\left|\sum_{i=1}^{S}(f(\hat{p}_i) - f(p_i))\right| \approx \left|\sum_{i=1}^{S}[(f(\hat{p}_i) - P(\hat{p}_i)) - (f(p_i) - P(p_i))]\right| \leq 2S \cdot \inf_{P \in \mathsf{Poly}_K} \|f - P\|_\infty$$

is small for large K. Hence, we can approximate the infinite-dimensional Lipschitz ball via a finite collection of functions, and then find an distribution to match these basis functions. When we choose monomials as the basis, we arrive at the *moment matching*; the reason why monomials are chosen as the basis will be detailed in the next subsection.

We emphasize that the final estimator requires that we conduct *moment matching* locally rather than globally, as shown in the next subsection.

2.3. Final Estimator

Before constructing the final estimator for $P^<$, we introduce the idea of Poissonization which has been widely used in related models. Specifically, in the Poissonized model, instead of drawing n i.i.d samples $X_1, \cdots, X_n$ from P, we draw N i.i.d samples $X_1, \cdots, X_N$ from P, where the number of samples $N \sim \mathsf{Poi}(n)$ is a random variable. The reason why we work on the Poissonized model is that, the empirical counts $n\hat{p}_j \sim \mathsf{Poi}(np_j)$ are independent under the Poissonized model. The following lemma relates the sorted ℓ_1 errors in these two models:

Lemma 11 *Let $R(n, S), R_P(n, S)$ be the minimax risk under sorted ℓ_1 loss in the Multinomial and Poissonized models, respectively. The following inequality holds:*

$$\frac{1}{2}R(2n, S) \leq R_P(n, S) \leq R(\frac{n}{2}, S) + 2\exp(-\frac{n}{8}).$$

By Lemma 11, it suffices to focus on the Poissonized model, where the estimator $\hat{P}$ for the sorted distribution $P^<$ is constructed as follows:

1. Split the samples into two parts, i.e., attach a random label uniformly distributed on $\{1, 2\}$ independently to each observation $X_1, \cdots, X_N$, and the observations are partitioned into two parts $X^{(1)}, X^{(2)}$ according to the label. By the property of Poisson distribution, each subset of the samples determines a Poissonized sampling model with rate replaced by $n/2$, and different subsets are independent. In the sequel we redefine $n/2$ as n for notational simplicity;

2. For each part of the samples and $i \in [S]$, conpute the empirical frequency $\hat{p}_{i,1}, \hat{p}_{i,2}$. The empirical frequency $\hat{p}_{i,1}$ in the first part will be used to determine the local domain where the true probability p_i lies, and $\hat{p}_{i,2}$ in the second part will be used for the estimation of $P^<$;

3. Fixing a universal constant $c_1 > 0$, partition the unit interval $[0, 1]$ into M sub-intervals $I_1, I_2, \cdots, I_M$, where $I_j \triangleq [\frac{c_1 \ln n}{n} \cdot (j-1)^2, \frac{c_1 \ln n}{n} \cdot j^2]$ for $j \in [M]$.[2] We define $x_j \triangleq \frac{c_1 j(j-1) \ln n}{n}$ as the "center" of the interval I_j. Similarly, we also define a slightly "enlarged" version of I_j: $\tilde{I}_j \triangleq [\frac{c_1 \ln n}{n} \cdot (j - \frac{3}{2})^2 \mathbb{1}(j \geq 2), \frac{c_1 \ln n}{n} \cdot (j+1)^2]$ for $j \in [M]$. Without loss of generality we assume that the number of sub-intervals $M = \sqrt{\frac{n}{c_1 \ln n}}$ is an integer;

4. Fixing universal constants $c_2, c_3 > 0$, in each sub-interval $\tilde{I}_j$ we solve the following convex optimization problems:

   - If $j \geq 2$, check whether there exists a measure μ_j on $\tilde{I}_j$ such that

   $$\mu_j(\tilde{I}_j) = S_j, \tag{2}$$

   $$\left| \int_{\tilde{I}_j} (x - x_j)^k \mu_j(dx) - \sum_{i=1}^{S} \mathbb{1}(\hat{p}_{i,1} \in I_j) g_{k,x_j}(\hat{p}_{i,2}) \right| \leq \sqrt{S_j \ln n} \cdot \left(\frac{c_3 j \ln n}{n} \right)^k \tag{3}$$

   hold simultaneously for $k = 1, 2, \cdots, K \triangleq c_2 \ln n$, where $S_j \triangleq \sum_{i=1}^{S} \mathbb{1}(\hat{p}_{i,1} \in I_j)$ is the number of symbols whose empirical probability lies in the interval I_j, and

   $$g_{k,x}(p) \triangleq \sum_{l=0}^{k} \binom{k}{l} (-x)^{k-l} \prod_{l'=0}^{l-1} \left(p - \frac{l'}{n} \right).$$

   If there exists a feasible solution, pick an arbitrary one; otherwise, report "failure" for the interval $\tilde{I}_j$;

   - If $j = 1$, solve the following minimization program over all measures μ_1 on $\tilde{I}_1$:

   $$\min \mu_1(\tilde{I}_1)$$
   $$\text{s.t. } \left| \int_{\tilde{I}_1} (x - x_1)^k \mu_1(dx) - \sum_{i=1}^{S} \mathbb{1}(\hat{p}_{i,1} \in I_1) g_{k,x_1}(\hat{p}_{i,2}) \right| \leq \sqrt{\mu_1(\tilde{I}_1) \ln n} \cdot \left(\frac{c_3 \ln n}{n} \right)^k$$
   $$\text{for all } k = 1, \cdots, K.$$
   $$\tag{4}$$

   Report the solution as μ_1; if this problem is infeasible, report "failure".

5. Construct a measure $\hat{\mu}$ on $[0, 1]$ as follows: if any previous step reports "failure", set $\hat{\mu}$ to be an arbitrary fixed distribution on $[0, 1]$; otherwise, set $\hat{\mu} = \sum_{j=1}^{M} \mu_j$;

6. Finally we need to output a vector. Let $S_0 \triangleq \lceil \hat{\mu}(\mathbb{R}) \rceil$, we add $S_0 - \hat{\mu}(\mathbb{R}) \geq 0$ to the point mass $\hat{\mu}(\{0\})$, and apply the randomized discretization in Definition 8 (with support size S_0) to transform the probability measure $\hat{\mu}/S_0$ into $\hat{P}$, which is our final estimator for $P^<$.

A few remarks are in order:

1. Choice of the partition $\{I_j\}_{j=1}^{M}$: the partition $\{I_j\}_{j=1}^{M}$ is chosen so that based on an observation $\hat{p}_{i,1} \in I_j$, the true probability mass p_i can be "localized" around I_j (i.e., p_i belongs to a

---

2. Boundary points can belong to either intervals, as long as $\{I_j\}_{j=1}^{M}$ constitutes a legitimate partition of $[0, 1]$.

slightly larger interval $\tilde{I}_j$) with high probability. Each interval I_j coincides with the definition of "confidence set" in Han et al. (2016), and the exact meaning of "localization" is referred to Lemma 17. As a result, probability masses p_i in the same partition are indistinguishable, while those in different partitions can be easily distinguished. Hence, at later stages it suffices to match moments locally since this is the range of indistinguishable probability masses.

2. Choice of $g_{k,x}(p)$: the key reason to choose $g_{k,x}(p)$ in the linear program is that in Poissonized model $n\hat{p} \sim \mathsf{Poi}(np)$, the statistic $g_{k,x}(\hat{p})$ is an unbiased estimator of $(p-x)^k$ (Withers, 1987, Example 2.8):

$$\mathbb{E}g_{k,x}(\hat{p}) = \sum_{l=0}^{k} \binom{k}{l}(-x)^{k-l}\mathbb{E}\prod_{l'=0}^{l-1}\left(\hat{p}-\frac{l'}{n}\right) = \sum_{l=0}^{k}\binom{k}{l}(-x)^{k-l}p^l = (p-x)^k. \tag{5}$$

We will see in Section 3 that estimating $P^<$ is similar to estimating symmetric functionals of P where bias is the dominating factor of the error, and this fact motivates us to apply an unbiased estimator of $(p-x)^k$ in (3), (4).

3. Convex optimization: the constraints in (2), (3) are linear in the measure μ_j, and thus they constitute an infinite-dimensional linear program. By squaring each sides of (4), the optimization problem for $j = 1$ becomes a conic quadratic programming and is thus convex. Due to its special structure, there is also a linear-programming-based way to solve (4): just do bisection search for $\mu_1(\tilde{I}_1)$, and solve a linear programming to check feasibility for each $\mu_1(\tilde{I}_1)$. To overcome the infinite dimensionality, in practice we can assume that μ_j is supported on a sufficiently fine grid to obtain a finite-dimensional problem. One can also transform the feasibility program in (2), (3) into a minimization problem, while the current form is sufficient for theoretical purposes. The idea of applying linear programming in related problems has appeared in several works, e.g., Valiant and Valiant (2011a, 2013, 2015); Kong and Valiant (2017); Tian et al. (2017).

4. Moment matching via convex optimization: the optimization problems (2), (3) and (4) are designed in such a way that the true measure $\mu_{P,j}$ (cf. (7), which requires the knowledge of the unknown P) is a feasible solution with high probability (cf. Lemma 18). Consequently, for any feasible solution μ_j, triangle inequality ensures that the local moments of our estimator will be close to the true moments (cf. (8)). The degree K of matched moments will be the main source of the *bias* of our estimator, and the RHS of (3) measures the fluctuation and will become the *variance*.

5. Choice of monomials $(x - x_j)^k$: there are two reasons to choose the monomials rather than other basis functions. Firstly, there exist unbiased estimators for monomials in the Poissonized model. Secondly, the subspace spanned by monomials is an optimal basis for approximating Lipschitz functions, i.e., it attains the Kolmogorov-n width of the Lipschitz ball Lorentz et al. (1996).

6. The knowledge of the support size S: we remark that our estimator construction is agnostic to the support size S. A key observation is that, although S appears in the definition of S_j, unseen symbols will not affect $\{S_j\}_{j\geq 2}$ since S_j only consists of symbols which have appeared in the first half samples for $j \geq 2$. The reason why we need a different program for

$j = 1$ is to deal with the unknown support size: if S was known, we could simply replace (4) by (2), (3) as well. The last step returns a vector $\hat{P}$ of length no longer than S with high probability (cf. Corollary 19), and we simply fill zeros to $\hat{P}$ when evaluating $\|\hat{P} - P^<\|_1$ with $S \geq S_0$. A key observation is that, filling m zeros to a vector $\hat{P}$ is equivalent to adding m units to $\hat{\mu}(\{0\})$ and then applying the randomized discretization, and thus Lemmas 7 and 9 still hold without knowing the support size S.

The performance of the estimator $\hat{P}$ is summarized in the following theorem.

Theorem 12 *Let $c_1 > 0$ be large enough as in Lemma 17, $c_1 > 2c_2, c_3 > 30c_1, c_2(6\ln 2 + \ln(5c_3/c_1)) < \epsilon$ and $c_2 \ln n \geq 1$. Then there exists a constant $C_0 > 0$ independent of n, S such that*

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P^<\|_1 \leq C_0 \left(\sqrt{\frac{S}{n \ln n}} + n^\epsilon \left(\sqrt{\frac{S}{n}} \wedge n^{-\frac{1}{3}} \right) \right).$$

3. Applications in Symmetric Functional Estimation

For functionals $F(\cdot)$ taking the form of (1), if we define the following estimator[3]

$$\hat{F} \triangleq \int_{\mathbb{R}} f(x)\hat{\mu}(dx) = S \cdot \int_{\mathbb{R}} f(x)\hat{\mu}^*(dx) \tag{6}$$

with $\hat{\mu}$ given by our estimator construction and $\hat{\mu}^* = S^{-1}\hat{\mu}$, it is straightforward to see that

$$|\hat{F} - F(P)| = S \cdot \left| \int_{\mathbb{R}} f(x)(\hat{\mu}^*(dx) - \mu_P(dx)) \right|.$$

If f is 1-Lipschitz, since Theorem 12 guarantees that the expected Wasserstein distance $\mathbb{E}_P W(\hat{\mu}^*, \mu_P)$ is small, it follows from the dual representation of Wasserstein distance (cf. Lemma 10) that $\mathbb{E}_P|\hat{F} - F(P)|$ is also small. For general non-smooth f, we have the following lemma:

Lemma 13 *Let the parameter configurations in Theorem 12 be fulfilled, $f(0) = 0$, $\mu_{P,j}$ and μ_j be defined in (7), (2), (3) and (4), respectively. Suppose that for each $j \in [M]$, there is a polynomial P_j of degree at most K with $\|f - P_j\|_{\infty, \tilde{I}_j} \leq M_j \triangleq \sup_{x \neq y \in \tilde{I}_j} |f(y) - f(x)|$, and $P_1(0) = 0$. For the estimator $\hat{F}$ defined in (6), with probability at least $1 - 3Sn^{-4}$, the following inequality holds conditioning on the first half samples:*

$$|\hat{F} - F(P)| \leq C_0 \sum_{j=1}^M \left(\int_{\mathbb{R}} |f(x) - P_j(x)|(\mu_j(dx) + \mu_{P,j}(dx)) + n^\epsilon M_j \sqrt{S_j} \right),$$

where $C_0 > 0$ is a constant independent of n, S, f and P_j. In particular,

$$|\hat{F} - F(P)| \leq C_0 \sum_{j=1}^M \left(S_j \cdot \inf_{P_j \in \mathsf{Poly}_K} \|f - P_j\|_{\infty, \tilde{I}_j} + n^\epsilon M_j \sqrt{S_j} \right).$$

---

3. Our construction of $\hat{\mu}$ does not depend on S, so is $\hat{F}$.

Remark 14 *The condition $\|f - P_j\|_{\infty, \tilde{I}_j} \leq M_j$ is mild since a reasonable approximating polynomial should approximate $f(\cdot)$ better than a constant function on $\tilde{I}_j$.*

Lemma 13 shows that the plug-in approach of $\hat{F}$ does polynomial approximation *implicitly* and *everywhere*. Specifically, the construction of $\hat{F}$ has nothing to do with polynomial approximation, while polynomial approximation emerges in the performance analysis of $\hat{F}$. Moreover, as opposed to the traditional approximation-based estimators where an explicit and functional-dependent polynomial is required, the plug-in estimator $\hat{F}$ can essentially approximate f using *any* polynomials. This property is desirable, since in general P_j may not be the best approximating polynomial and may be hard to design explicitly: we refer to Jiao et al. (2017) for such an example. Also, as opposed to the approximation-based estimators which split $\text{dom}(f)$ into "non-smooth" and "smooth" regimes, the plug-in estimator $\hat{F}$ does polynomial approximation everywhere. This property prevents $\hat{F}$ from achieving the optimal variance, but this is the price we need to pay to achieve a unified methodology without the dependence on f.

The following theorem characterizes the performance of $\hat{F}$ for $F(P)$ with $F = H, F_\alpha, S$:[4]

Theorem 15 *The plug-in estimator in (6) with $F = H, F_\alpha, S$ satisfies:*

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P |\hat{H} - H(P)| \leq C_0 \left(\frac{S}{n \ln n} + n^\epsilon \left(\sqrt{\frac{S}{n}} \wedge n^{-\frac{1}{3}} \right) \right),$$

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P |\hat{F}_\alpha - F_\alpha(P)| \leq C_0 \left(\frac{S}{(n \ln n)^\alpha} + n^\epsilon \left(\sqrt{\frac{S^{3-2\alpha}}{n}} \wedge n^{-\frac{\alpha}{3}} \right) \right),$$

$$\sup_{P \in \mathcal{D}_k} \mathbb{E}_P |\hat{S} - S(P)| \leq C_0 k \left(\exp \left(-\Theta \left(\sqrt{\frac{n \ln k}{k}} \right) \right) + \frac{n^\epsilon}{\sqrt{k}} \right), \qquad n \lesssim k \ln k,$$

where $C_0 > 0$ is a constant independent of n, S, k.

Remark 16 *An additional condition $n \lesssim k \ln k$ is required for the support size functional $S(P)$: if $n \gg k \ln k$, the minimax risk decays super-polynomially in n, which makes the $O(Sn^{-4})$ failure probability in Lemma 13 become non-negligible.*

Compared with the minimax rates of these functionals in Wu and Yang (2016a); Jiao et al. (2015); Wu and Yang (2015), the general plug-in approach in (6) achieves the optimal total bias term, which is the leading term when S or k is large. As a result, the plug-in approach attains the optimal sample complexity for all these functionals, establishing Theorem 4.

However, a comparison of Theorem 15 and the minimax rates shows that the variance term of $\hat{F}$ is not optimal, conforming to the aforementioned intuition that everywhere polynomial approximation may incur a too large variance. Hence, among the functionals considered in Theorem 15, the general plug-in approach in (6) attains the optimal bias and thus the optimal sample complexity, but need to pay a price on the variance.

---

4. For the support size functional $S(P)$, due to the additional constraint $p_i \geq \frac{1}{k}$ on the parameter set, an additional linear constraint $\mu_j((0, \frac{1}{k})) = 0$ should be imposed in addition to (2) and (3).

References

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.

Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. The complexity of estimating Rényi entropy. SODA, 2015.

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.

Chris Anderson. The long tail. *Wired magazine*, 12(10):170–177, 2004.

Daniel Berend and Aryeh Kontorovich. A sharp estimate of the binomial mean absolute deviation with applications. *Statistics & Probability Letters*, 83(4):1254–1259, 2013.

Yuheng Bu, Shaofeng Zou, Yingbin Liang, and Venugopal V Veeravalli. Estimation of KL divergence between large-alphabet distributions. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 1118–1122. IEEE, 2016.

T Tony Cai and Mark G Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A Servedio. Learning k-modal distributions via testing. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1371–1385, 2012.

Ronald A DeVore. Degree of approximation. *Approximation theory II*, 241(242):117–161, 1976.

Ilias Diakonikolas. Beyond histograms: Structure and distribution estimation. In *Workshop of the 46th ACM Symposium on Theory of Computing*, 2014.

Zeev Ditzian and Vilmos Totik. *Moduli of smoothness*. Springer, 1987.

John C Duchi and Martin J Wainwright. Distance-based and continuum fano inequalities with applications to statistical estimation. *arXiv preprint arXiv:1311.2669*, 2013.

Bradley Efron and Ronald Thisted. Estimating the number of unsen species: How many words did shakespeare know? *Biometrika*, 63(3):pp. 435–447, 1976. ISSN 00063444. URL http://www.jstor.org/stable/2335721.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions under ℓ_1 loss. *IEEE Transactions on Information Theory*, 61(11):6343–6354, 2015.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax rate-optimal estimation of divergences between discrete distributions. *arXiv preprint arXiv:1605.09124*, 2016.

Yanjun Han, Jiantao Jiao, Rajarshi Mukherjee, and Tsachy Weissman. On estimation of L_r-norms in Gaussian white noise models. *arXiv preprint arXiv:1710.03863*, 2017a.

Yanjun Han, Jiantao Jiao, Tsachy Weissman, and Yihong Wu. Optimal rates of entropy estimation over Lipschitz balls. *arXiv preprint arXiv:1711.02141*, 2017b.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *Information Theory, IEEE Transactions on*, 61(5):2835–2885, 2015.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the L_1 distance. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 750–754. IEEE, 2016.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the L_1 distance. *arXiv preprint arXiv:1705.00807*, 2017.

Sudeep Kamath, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Proceedings of The 28th Conference on Learning Theory*, pages 1066–1100, 2015.

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

Weihao Kong and Gregory Valiant. Spectrum estimation from samples. *The Annals of Statistics*, 45 (5):2218–2247, 2017.

Oleg Lepski, Arkady Nemirovski, and Vladimir Spokoiny. On estimation of the L_r norm of a regression function. *Probability theory and related fields*, 113(2):221–253, 1999.

George G Lorentz, Manfred von Golitschek, and Yuly Makovoz. *Constructive approximation: advanced problems*, volume 304. Springer Berlin, 1996.

Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

I. Olkin and M. Sobel. Admissible and minimax estimation for the multinomial distribution and for independent binomial distributions. *The Annals of Statistics*, 7:284–290, 1979.

Alon Orlitsky and Ananda Theertha Suresh. Competitive distribution estimation: Why is good-turing good. In *Advances in Neural Information Processing Systems*, pages 2143–2151, 2015.

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.

Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016.

Dmitri S Pavlichin, Jiantao Jiao, and Tsachy Weissman. Approximate profile maximum likelihood. *arXiv preprint arXiv:1712.07177*, 2017.

Giovanni Peccati and Murad S Taqqu. Some facts about charlier polynomials. In *Wiener Chaos: Moments, Cumulants and Diagrams*, pages 171–175. Springer, 2011.

Aditi Raghunathan, Greg Valiant, and James Zou. Estimating the unseen from multiple populations. *arXiv preprint arXiv:1707.03854*, 2017.

M. Rutkowska. Minimax estimation of the parameters of the multivariate hypergeometric and multinomial distributions. *Zastos. Mat.*, 16:9–21, 1977.

H. Steinhaus. The problem of estimation. *The Annals of Mathematical Statistics*, 28:633–648, 1957.

Kevin Tian, Weihao Kong, and Gregory Valiant. Learning populations of parameters. In *Advances in Neural Information Processing Systems*, pages 5780–5789, 2017.

S. Trybula. Some problems of simultaneous minimax estimation. *The Annals of Mathematical Statistics*, 29:245–253, 1958.

A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer-Verlag, 2008.

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 43rd annual ACM symposium on Theory of computing*, pages 685–694. ACM, 2011a.

Gregory Valiant and Paul Valiant. The power of linear estimators. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 403–412. IEEE, 2011b.

Gregory Valiant and Paul Valiant. Instance optimal learning. *arXiv preprint arXiv:1504.05321*, 2015.

Gregory Valiant and Paul Valiant. Estimating the unseen: Improved estimators for entropy and other properties. *Journal of the ACM (JACM)*, 64(6):37, 2017.

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

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

Abraham Wald. *Statistical decision functions*. Wiley, 1950.

Christopher Stroude Withers. Bias reduction by Taylor series. *Communications in Statistics-Theory and Methods*, 16(8):2369–2383, 1987.

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

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

Yihong Wu and Pengkun Yang. Sample complexity of the distinct elements problem. *arXiv preprint arXiv:1612.03375*, 2016b.

Appendix A. Estimator Analysis

A.1. Controlling "Bad Events"

There are several types of bad events in the construction of our estimator:

1. For some symbol $i = 1, \cdots, S$ and $j = 1, \cdots, M$, it may happen that $p_i \notin \tilde{I}_j$ but $\hat{p}_{i,1} \in I_j$;

2. For some $j = 1, \cdots, M$, it may happen that the linear programming in (2), (3), (4) does not have a solution;

3. In the last step, it may happen that $\hat{\mu}(\mathbb{R}) > S$.

In this subsection we show that the probability that any of these bad events occurs is negligible. The following lemma follows directly from the Poisson tail inequalities (cf. Lemma 29).

Lemma 17 *Let $c_1 > 0$ be large enough, then for any $i = 1, 2, \cdots, S$ and $j = 1, 2, \cdots, M$,*

$$\mathbb{P}(\hat{p}_{i,1} \in I_j | p_i \notin \tilde{I}_j) \leq n^{-5}.$$

Based on Lemma 17 and the union bound, we see that the first-type bad events occurs with a negligible probability. To upper bound the probability of other bad events, we need to come up with a solution μ_j which fulfills (2), (3), (4) with high probability. In the sequel we condition on a specific realization of the first half samples, and define the set of symbols falling in I_j as

$$A_j \triangleq \{i \in \{1, 2, \cdots, S\} : \hat{p}_{i,1} \in I_j\}.$$

Note that A_j is a random set depending only on the first half samples, and $|A_j| = S_j$. Moreover, $\{A_j, S_j\}_{j \geq 2}$ are uniquely determined by the first half samples, while $\{A_1, S_1\}$ may be unknown due to the unknown support size S and possibly unseen symbols. Now the key observation is that, the following measure

$$\mu_{P,j}(\cdot) \triangleq \sum_{i \in A_j} \mathbb{1}(p_i \in \cdot), \qquad j = 1, 2, \cdots, M \tag{7}$$

which requires the knowledge of the unknown P satisfies (2), (3), (4) with high probability. Obviously, if $p_i \in \tilde{I}_j$ for any $i \in A_j$ and $j \geq 2$, the measure $\mu_{P,j}$ will be supported on $\tilde{I}_j$ and thus (2) holds. The following lemma shows that, given the same assumption, the measure $\mu_{P,j}$ also satisfies (3) and (4) with high probability.

Lemma 18 *Let $c_1 > 0$ be large enough as in Lemma 17, and $c_1 > 2c_2, c_3 > 30c_1, c_2 \ln n \geq 1$. Further assume that $p_i \in \tilde{I}_j$ for any $i \in A_j$ and $j = 1, 2, \cdots, M$. Then conditioning on the first half samples, for $k = 1, 2, \cdots, K = c_2 \ln n$ we have*

$$\mathbb{P}\left(\left|\int_{\tilde{I}_j} (x - x_j)^k \mu_j(dx) - \sum_{i=1}^{S} \mathbb{1}(\hat{p}_{i,1} \in I_j) g_{k,x_j}(\hat{p}_{i,2})\right| > \sqrt{S_j \ln n} \cdot \left(\frac{c_3 j \ln n}{n}\right)^k\right) \leq 2n^{-4}.$$

Lemma 18 shows that $\mu_{P,j}$ is a feasible solution to (3) with high probability for $j \geq 2$. For $j = 1$, note that $\mu_{P,1}(\tilde{I}_1) = S_1$ conditioning on the events in Lemma 17, the measure $\hat{\mu}_{P,1}$ is also a feasible solution to (4). Moreover, in this case the returned solution μ_1 of (4) satisfies $\mu_1(\mathbb{R}) \leq S_1$, thus we must have $\hat{\mu}(\mathbb{R}) \leq S$ in the last step. Hence, based on Lemma 17 and 18, by the union bound we have the following corollary:

16

Corollary 19 *Let E be the event that any of the aforementioned bad events happens, then*

$$\mathbb{P}(E) \leq Sn^{-5} + 2Mn^{-4}.$$

By Corollary 19, the probability that any bad event happens is negligible, and it thus suffices to focus on the "good" events to analyze the performance of our estimator, which will be the focus of the next subsection.

A.2. Local Performance of Moment Matching

By Lemma 18, the measure $\mu_{P,j}$ using the unknown true knowledge of P is a feasible solution to (2), (3) and (4) with high probability, and our estimator returns a perfect answer if this solution is chosen among all feasible solutions. In this subsection, we show that any feasible solution μ_j is in fact close to the true measure $\mu_{P,j}$ in terms of the Wasserstein distance. By the dual representation of Wasserstein distance (cf. Lemma 10), we can fix any 1-Lipschitz function f on $\mathbb{R}$ and prove the following lemma:

Lemma 20 *For any $j = 1, \cdots, M$, let μ_j be any feasible solution to (2), (3) or (4), the true measure $\mu_{P,j}$ be given in (7), and c_2 be small enough such that $c_2(6\ln 2 + \ln(5c_3/c_1)) < \epsilon$ with $c_2 \ln n \geq 1$. Assuming all good events happen and conditioning on the first half samples, for any 1-Lipschitz function f on $\mathbb{R}$ with $f(0) = 0$,*

$$\left| \int_{\tilde{I}_j} f(x)\mu_j(dx) - \int_{\tilde{I}_j} f(x)\mu_{P,j}(dx) \right| \leq C_0 \left(\int_{\mathbb{R}} \sqrt{\frac{x}{n \ln n}} (\mu_j(dx) + \mu_{P,j}(dx)) + \frac{j\sqrt{S_j}}{n^{1-\epsilon}} \right)$$

where C_0 is a constant independent of n, S and f.

Remark 21 *The condition $f(0) = 0$ is important for $j = 1$: (4) cannot ensure that $\mu_1(\mathbb{R}) = \mu_{P,1}(\mathbb{R})$ since it does not contain the total mass constraint in (2).*

The remainder of this subsection is devoted to the proof of Lemma 20. For $j \geq 2$, by assumption both $\mu_{P,j}$ and μ_j are feasible solutions to (2) and (3), i.e., they are supported on $\tilde{I}_j$ with the same total mass S_j, and by triangle inequality we have

$$\left| \int_{\tilde{I}_j} (x - x_j)^k \mu_j(dx) - \int_{\tilde{I}_j} (x - x_j)^k \mu_{P,j}(dx) \right| \leq 2\sqrt{S_j \ln n} \cdot \left(\frac{c_3 j \ln n}{n} \right)^k \tag{8}$$

for any $k = 1, 2, \cdots, K = c_2 \ln n$. As a result, fixing any polynomial

$$P(x) = \sum_{k=0}^{K} a_k (x - x_j)^k$$

on $\tilde{I}_j$, triangle inequality together with (8) gives

$$
\left| \int_{\tilde{I}_j} f(x)(\mu_j(dx) - \mu_{P,j}(dx)) \right|
$$

$$
\leq \left| \int_{\tilde{I}_j} (f(x) - P(x))(\mu_j(dx) - \mu_{P,j}(dx)) \right| + \left| \int_{\tilde{I}_j} P(x)(\mu_j(dx) - \mu_{P,j}(dx)) \right|
$$

$$
\leq \underbrace{\int_{\tilde{I}_j} |f(x) - P(x)|(\mu_j(dx) + \mu_{P,j}(dx))}_{\triangleq B_1} + \underbrace{\sum_{k=1}^{K} |a_k| \cdot 2\sqrt{S_j \ln n} \left(\frac{c_3 j \ln n}{n} \right)^k}_{\triangleq B_2}. \qquad (9)
$$

It's straightforward to see that (8), (9) also hold for $j = 1$, while in (9) we need to add an additional assumption that the constant term of $P(x)$ is zero.

The inequality (9) holds for any polynomial $P(\cdot)$ of degree at most K, and both terms B_1 and B_2 depend on the choice of P. We shall choose P to be the best approximating polynomial of $f(x)$ on $\tilde{I}_j$ in the uniform norm, i.e.,

$$
P(x) \triangleq \arg \min_{Q \in \mathsf{Poly}_K} \max_{x \in \tilde{I}_j} |Q(x) - f(x)|.
$$

It is easy to see that this choice of P will result in a small value of B_1, while we need the following well-known Jackson's inequality in approximation theory to upper bound B_1 quantitatively:

Lemma 22 *DeVore (1976) Let $K > 0$ be any integer, and $[a, b] \subset \mathbb{R}$ be any bounded interval. For any 1-Lipschitz function f on $[a, b]$, there exists a universal constant C independent of K, f such that there exists a polynomial $P(\cdot)$ of degree at most K such that*

$$
|f(x) - P(x)| \leq \frac{C\sqrt{(b-a)(x-a)}}{K}, \qquad \forall x \in [a, b]. \qquad (10)
$$

In particular, the following norm bound holds:

$$
\sup_{x \in [a,b]} |f(x) - P(x)| \leq \frac{C(b-a)}{K}. \qquad (11)
$$

We use the pointwise bound (10) and the norm bound (11) to upper bound B_1 for the case $j = 1$ and $j \geq 2$, respectively. If $j = 1$, we have $\tilde{I}_j = [0, \frac{9c_1 \ln n}{4n}]$, then (10) with $x = 0$ and $f(0) = 0$ gives $P(0) = 0$, thus (9) holds for P. Moreover,

$$
B_1 \leq \frac{C}{K} \int_{\tilde{I}_1} \sqrt{\frac{9c_1 x \ln n}{4n}} (\mu_1(dx) + \mu_{P,1}(dx))
$$

$$
= \frac{C}{c_2} \sqrt{\frac{9c_1}{4}} \cdot \int_{\tilde{I}_1} \sqrt{\frac{x}{n \ln n}} (\mu_1(dx) + \mu_{P,1}(dx)). \qquad (12)
$$

18

If $j \geq 2$, recall that $\tilde{I}_j = [\frac{c_1(j-3/2)^2 \ln n}{n}, \frac{c_1(j+1/2)^2 \ln n}{n}]$, the norm bound (11) gives

$$
\begin{aligned}
B_1 &\leq \frac{C}{K} \left(\frac{c_1(j+1/2)^2 \ln n}{n} - \frac{c_1(j-3/2)^2 \ln n}{n} \right) \cdot \int_{\tilde{I}_j} (\mu_j(dx) + \mu_{P,j}(dx)) \\
&\leq \frac{C}{K} \left(\frac{c_1(j+1/2)^2 \ln n}{n} - \frac{c_1(j-3/2)^2 \ln n}{n} \right) \cdot \int_{\tilde{I}_j} \sqrt{\frac{x}{\frac{c_1(j-3/2)^2 \ln n}{n}}} (\mu_j(dx) + \mu_{P,j}(dx)) \\
&\leq \frac{12C\sqrt{c_1}}{c_2} \cdot \int_{\tilde{I}_j} \sqrt{\frac{x}{n \ln n}} (\mu_j(dx) + \mu_{P,j}(dx)).
\end{aligned}
\tag{13}
$$

A combination of (12) and (13) gives that for any $j \in [M]$,

$$
B_1 \leq \frac{12C\sqrt{c_1}}{c_2} \cdot \int_{\mathbb{R}} \sqrt{\frac{x}{n \ln n}} (\mu_j(dx) + \mu_{P,j}(dx)).
\tag{14}
$$

To upper bound B_2, we need to obtain upper bounds on the coefficients $|a_k|$ of the best approximating polynomial. We invoke Lemma 27 here: the polynomial $P_0(x) = P(x) - f(x_j)$ defined on $\tilde{I}_j$ satisfies

$$
\begin{aligned}
|P_0(x)| &\leq |P(x) - f(x)| + |f(x) - f(x_j)| \\
&\leq \frac{C}{K} \left(\frac{c_1(j+1)^2 \ln n}{n} - \frac{c_1(j-3/2)^2 \ln n}{n} \right) + \left(\frac{c_1(j+1)^2 \ln n}{n} - \frac{c_1 j(j-1) \ln n}{n} \right) \\
&\leq \left(1 + \frac{C}{K} \right) \frac{5c_1 j \ln n}{n}.
\end{aligned}
$$

Now applying Lemma 27 with $A = \left(1 + \frac{C}{K}\right) \frac{5c_1 j \ln n}{n}$, $[a,b] = [-\frac{c_1((2j-9/4)\vee 0) \ln n}{n}, \frac{c_1(3j+1) \ln n}{n}]$, for any $k = 1, 2, \cdots, K$ we have

$$
\begin{aligned}
|a_k| &\leq 2^{\frac{7K}{2}} \left(1 + \frac{C}{K}\right) \frac{5c_1 j \ln n}{n} \cdot \left(\frac{c_1 j \ln n}{5n} \right)^{-k} (5^K + 1) \\
&\leq 25 \left(1 + \frac{C}{K}\right) 2^{6K} \cdot \left(\frac{c_1 j \ln n}{5n} \right)^{1-k}.
\end{aligned}
$$

Hence, the quantity B_2 can be upper bounded as

$$
\begin{aligned}
|B_2| &\leq \sum_{k=1}^{K} 25 \left(1 + \frac{C}{K}\right) 2^{6K} \cdot \left(\frac{c_1 j \ln n}{5n} \right)^{1-k} \cdot 2\sqrt{S_j \ln n} \left(\frac{c_3 j \ln n}{n} \right)^k \\
&\leq 10c_1(1+C) n^{c_2(6 \ln 2 + \ln(5c_3/c_1))} (\ln n)^{\frac{3}{2}} \cdot \frac{j\sqrt{S_j}}{n}.
\end{aligned}
\tag{15}
$$

Now a combination of (14) and (15) completes the proof of Lemma 20.

A.3. Overall Performance

In this section we are about to establish Theorem 12. Note that finally the true measure μ_P and the measure $\hat{\mu}^*$ as the input of randomized discretization are given by (conditioning on no failures)

$$\mu_P = \frac{1}{S} \sum_{i=1}^{M} \mu_{P,j},$$

$$\hat{\mu}^* = \frac{1}{S} \sum_{i=1}^{M} \mu_j + \left(1 - \frac{\hat{\mu}(\mathbb{R})}{S}\right) \delta_0$$

where $\delta_0(\cdot)$ is the Dirac delta point mass at zero. By Lemma 7 and Lemma 9, the sorted ℓ_1 risk of $\hat{P}$ satisfies

$$\mathbb{E}_P \|\hat{P} - P^<\|_1 = S \cdot \mathbb{E}_P W(\mu_P, \hat{\mu}^*).$$

Using the dual representation of the Wasserstein distance (cf. Lemma 10), we further have

$$
\begin{aligned}
S \cdot \mathbb{E}_P W(\mu_P, \hat{\mu}) &= S \cdot \mathbb{E}_P \sup_{f:\|f\|_{\mathrm{Lip}}\leq 1} \int_{\mathbb{R}} f(x)(\mu_P(dx) - \hat{\mu}^*(dx)) \\
&\overset{(a)}{=} S \cdot \mathbb{E}_P \sup_{f:\|f\|_{\mathrm{Lip}}\leq 1, f(0)=0} \int_{\mathbb{R}} f(x)(\mu_P(dx) - \hat{\mu}^*(dx)) \\
&= \mathbb{E}_P \sup_{f:\|f\|_{\mathrm{Lip}}\leq 1, f(0)=0} \sum_{j=1}^{M} \int_{\mathbb{R}} f(x)(\mu_{P,j}(dx) - \mu_j(dx)) \quad (16)
\end{aligned}
$$

where (a) follows from $\mu_P(\mathbb{R}) = \hat{\mu}^*(\mathbb{R}) = 1$.

Suppose that the condition of Lemma 20 holds, then each summand admits the following "bias–variance" decomposition:

$$\int_{\mathbb{R}} f(x)(\mu_{P,j}(dx) - \mu_j(dx)) \leq C_0 \left(\int_{\mathbb{R}} \sqrt{\frac{x}{n \ln n}} (\mu_j(dx) + \mu_{P,j}(dx)) + \frac{j\sqrt{S_j}}{n^{1-\epsilon}} \right). \quad (17)$$

The first term in (17) corresponds to the "bias", which is the remaining error even after the first K moments are exactly matched. The second term in (17) corresponds to the "variance", which is caused by the imperfect moment matching. Since C_0 is independent of f, we have

$$\sup_{f:\|f\|_{\mathrm{Lip}}\leq 1, f(0)=0} \sum_{j=1}^{M} \int_{\mathbb{R}} f(x)(\mu_{P,j}(dx) - \mu_j(dx))$$

$$\leq C_0 \left(\underbrace{\int_{\mathbb{R}} \sqrt{\frac{x}{n \ln n}} (\hat{\mu}(dx) + S\mu_P(dx))}_{\triangleq \text{"total bias" } B} + \underbrace{\sum_{j=1}^{M} \frac{j\sqrt{S_j}}{n^{1-\epsilon}}}_{\triangleq \text{"total variance" } V} \right). \quad (18)$$

We first upper bound the total variance V. Using the fact that $\sum_{i=1}^{S} p_i = 1$, and $p_i \geq \frac{c_1 j^2 \ln n}{16n}$ for any $i \in A_j, j \geq 2$, we have

$$1 = \sum_{i=1}^{S} p_i \geq \sum_{j=2}^{M} \sum_{i \in A_j} p_i \geq \frac{c_1 \ln n}{16n} \sum_{j=2}^{M} \sum_{i \in A_j} j^2 = \frac{c_1 \ln n}{16n} \sum_{j=2}^{M} j^2 S_j. \tag{19}$$

Moreover, $S_1 \leq S$, thus by defining $J = \{j \in [M] : S_j \neq 0\}$, we have

$$\sum_{j=1}^{M} j\sqrt{S_j} \leq \sqrt{S} + \sum_{j=2}^{M} j\sqrt{S_j} \leq \sqrt{S} + \sum_{j \in J} j\sqrt{S_j} = \sqrt{S} + \left(|J| \cdot \sum_{j \in J} j^2 S_j \right)^{\frac{1}{2}} \leq \sqrt{S} + 4\sqrt{\frac{n|J|}{c_1 \ln n}}.$$

Now we obtain upper bounds for $|J|$, i.e., the number of sub-intervals in the partition which contains any symbol in the first half samples. Trivially, $|J| \leq S$, and (19) gives

$$1 \geq \frac{c_1 \ln n}{16n} \sum_{j=2}^{M} j^2 S_j \geq \frac{c_1 \ln n}{16n} \sum_{j \in J - \{1\}} j^2 \geq \frac{c_1 \ln n}{16n} \sum_{j=2}^{|J|} j^2 \geq \frac{c_1 \ln n}{48n} (|J| - 1)^3$$

implying that $|J| \leq \left(\frac{48n}{c_1 \ln n} \right)^{\frac{1}{3}} + 1$. As a result, we conclude that

$$\sum_{j=1}^{M} j\sqrt{S_j} \leq \sqrt{S} + 4\sqrt{\frac{n}{c_1 \ln n}} \left(\sqrt{S} \wedge \sqrt{\left(\frac{48n}{c_1 \ln n} \right)^{\frac{1}{3}} + 1} \right) \tag{20}$$

and consequently

$$V \lesssim n^{\epsilon} \left(\sqrt{\frac{S}{n}} \wedge n^{-\frac{1}{3}} \right). \tag{21}$$

Now we upper bound the total bias. By definition of μ_P, we know that

$$\int_{\mathbb{R}} x \cdot S\mu_P(dx) = \sum_{i=1}^{S} p_i = 1.$$

Moreover, summing over $j = 1, \cdots, M$ in (8) for $k = 1$ gives

$$\left| \int_{\mathbb{R}} x(S\mu_P(dx) - \hat{\mu}(dx)) \right| \leq \sum_{j=1}^{M} 2\sqrt{S_j \ln n} \cdot \frac{c_3 j \ln n}{n} = 2c_3 (\ln n)^{\frac{3}{2}} \cdot \sum_{j=1}^{M} \frac{j\sqrt{S_j}}{n}.$$

Hence, by (20) and the triangle inequality, we also have $\int_{\mathbb{R}} x\hat{\mu}(dx) = 1 + o(1)$. As a result, by Cauchy–Schwartz the total bias can be upper bounded as

$$B \leq \frac{1}{\sqrt{n \ln n}} \left(S \cdot \sqrt{\int_{\mathbb{R}} x\mu_P(dx)} + \sqrt{S \cdot \int_{\mathbb{R}} x\hat{\mu}(dx)} \right) \lesssim \sqrt{\frac{S}{n \ln n}}. \tag{22}$$

Finally, a combination of (16), (18), (21), (22), Corollary 19 and the fact that $\|\hat{P} - P^<\|_1 \leq 2$ gives that

$$
\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P^<\|_1 \lesssim \sqrt{\frac{S}{n \ln n}} + n^\epsilon \left(\sqrt{\frac{S}{n}} \wedge n^{-\frac{1}{3}} \right) + 2 \cdot (2Mn^{-4} + Sn^{-5})
$$

$$
\lesssim \sqrt{\frac{S}{n \ln n}} + n^\epsilon \cdot \left(\sqrt{\frac{S}{n}} \wedge n^{-\frac{1}{3}} \right)
$$

which completes the proof of Theorem 12.

Appendix B. Minimax Lower Bound

In this section we establish the following lower bound:

Theorem 23 *For $n \gtrsim \frac{S}{\ln S}$, there exists a constant $c_0 > 0$ independent of n, S such that*

$$
\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P^<\|_1 \geq c_0 \left(\sqrt{\frac{S}{n \ln n}} + \left(\sqrt{\frac{S}{n}} \wedge n^{-\frac{1}{3}} \right) \right).
$$

Notice that a combination of Theorem 12 and Theorem 23 completes the proof of Theorem 1. As in the proof of achievability, we call the first term in Theorem 23 as "bias" and the second term as "variance": the techniques used to lower bound these terms mimic those which have been widely used to lower bound the bias and the variance, respectively. The next two subsections are devoted to the proof of Theorem 23.

B.1. Lower Bound on the "Bias"

To prove the $\Omega(\sqrt{\frac{S}{n \ln n}})$ lower bound, we use the following "double duality" arguments:

1. Use the dual representation of Wasserstein distance (cf. Lemma 10) to transform into estimation of Lipschitz functionals;

2. Use the duality between moment matching and best polynomial approximation (cf. Lemma 25) to construct two measures used in the generalized Le Cam's method (cf. Lemma 24).

Note that both these dualities are also used in the proof of the achievability part (cf. Theorem 12), our arguments for the achievability and lower bound are in fact dual to each other.

We first make use of the first duality. Assume by contradiction that there exists an estimator $\hat{P}$ such that $\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P^<\|_1 \ll \sqrt{\frac{S}{n \ln n}}$, then for any 1-Lipschitz function $f(\cdot)$ on $\mathbb{R}$ and the symmetric functional $F(P)$ of the form

$$
F(P) \triangleq \sum_{i=1}^{S} f(p_i),
$$

a combination of Lemma 7 and 10 implies that for the estimator $F(\hat{P}) = \sum_{i=1}^{S} f(\hat{p}_i)$, we have

$$
\mathbb{E}_P |F(\hat{P}) - F(P)| \leq S \cdot \mathbb{E}_P W(\mu_P, \mu_{\hat{P}}) = \mathbb{E}_P \|\hat{P} - P^<\|_1 \ll \sqrt{\frac{S}{n \ln n}}.
$$

Hence, the existence of such an estimator $\hat{P}$ implies that, for any 1-Lipschitz function $f(\cdot)$ and the corresponding symmetric functional $F(\cdot)$, we have

$$\inf_{\hat{F}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P |\hat{F} - F(P)| \ll \sqrt{\frac{S}{n \ln n}}. \tag{23}$$

In other words, the estimation of Lipschitz functionals are *easier* than the estimation of the underlying distribution up to permutation. As a result, if we could prove that (23) breaks down for some Lipschitz functional $F(P)$, we would arrive at the desired contradiction.

Next we step into the second duality, which requires the following generalized Le Cam's method (also known as the method of two fuzzy hypotheses Tsybakov (2008)). The application of this method has appeared in several works in functional estimation Lepski et al. (1999); Cai and Low (2011); Jiao et al. (2015); Wu and Yang (2016a); Han et al. (2016); Jiao et al. (2017); Han et al. (2017a,b) to deal with the bias, which motivates us to call the first term in Theorem 23 as the "bias". Given a collection of distributions $\{P_\theta : \theta \in \Theta'\}$, suppose the observation $\mathbf{Z}$ is distributed as P_θ with $\theta \in \Theta \subset \Theta'$. Let $\hat{T} = \hat{T}(\mathbf{Z})$ be an arbitrary estimator of a function $T(\theta)$ based on $\mathbf{Z}$. Denote the total variation distance between two probability measures P, Q by

$$V(P, Q) \triangleq \sup_{A \in \mathcal{A}} |P(A) - Q(A)| = \frac{1}{2} \int |p - q| d\nu,$$

where $p = \frac{dP}{d\nu}, q = \frac{dQ}{d\nu}$, and ν is a dominating measure so that $P \ll \nu, Q \ll \nu$. The following general minimax lower bound follows from the same proof as (Tsybakov, 2008, Theorem 2.15):

Lemma 24 *Let σ_0 and σ_1 be two prior distributions on Θ'. Suppose there exist $\zeta \in \mathbb{R}, s > 0, 0 \le \beta_0, \beta_1 < 1$ such that*

$$\sigma_0(\theta \in \Theta : T(\theta) \le \zeta - s) \ge 1 - \beta_0,$$
$$\sigma_1(\theta \in \Theta : T(\theta) \ge \zeta + s) \ge 1 - \beta_1.$$

Then

$$\inf_{\hat{T}} \sup_{\theta \in \Theta} \mathbb{P}_\theta \left(|\hat{T} - T(\theta)| \ge s \right) \ge \frac{1 - V(F_1, F_0) - \beta_0 - \beta_1}{2},$$

where $F_i = \int P_\theta \sigma_i(d\theta)$ is the marginal distribution of $\mathbf{Z}$ under the prior σ_i for $i = 0, 1$, respectively.

In our application, we will set $\theta = P, T(\theta) = F(P)$, and $\sigma_i = \nu_i^{\otimes S}$ for $i = 0, 1$, where ν_0, ν_1 are priors on $[\frac{1}{S} - \sqrt{\frac{c \ln n}{nS}}, \frac{1}{S} + \sqrt{\frac{c \ln n}{nS}}]$ with some constant $c > 0$. The priors ν_0, ν_1 are chosen to be the solutions of the optimization program (24), whose optimal objective value is the best polynomial approximation error as shown in the following lemma.

Lemma 25 *Given a compact interval $I = [a, b]$ with $a > 0$, an integer $K > 0$ and a continuous function f on I, let*

$$E_K(f; I) \triangleq \inf_{\{a_i\}} \sup_{x \in I} \left| \sum_{i=0}^{K} a_i x^i - f(x) \right|$$

denote the best uniform approximation error of f by polynomials spanned by $\{1, x, \cdots, x^K\}$. Then

$$
\begin{aligned}
2E_K(f; I) = \max \int f(t)\nu_1(dt) - \int f(t)\nu_0(dt) \\
\text{s.t.} \int t^l \nu_1(dt) = \int t^l \nu_0(dt), \quad l = 0, \cdots, K
\end{aligned}
\tag{24}
$$

where the maximum is taken over pairs of probability measures ν_0 and ν_1 supported on I.

Lemma 25 establishes the duality between moment matching and best polynomial approximation, where moment matching helps to obtain a small total variation distance $V(F_1, F_0)$ in Lemma 24, and best polynomial approximation error gives the value s in Lemma 24. Moreover, here moment matching is also done locally, for $\operatorname{supp}(\nu_i) = [\frac{1}{S} - \sqrt{\frac{c \ln n}{nS}}, \frac{1}{S} + \sqrt{\frac{c \ln n}{nS}}], i = 0, 1$ takes the form of a local interval as $\{I_j\}_{j=1}^M$.

Next we specify the choice of the 1-Lipschitz function: $f(x) = |x - \frac{1}{S}| - \frac{1}{S}$, which was studied in Jiao et al. (2017). In particular, Jiao et al. (2017) shows that with these priors and $K \asymp \ln n$ in Lemma 25, we have $s \gtrsim \sqrt{\frac{S}{n \ln n}}$ and $\beta_0, \beta_1, V(F_1, F_0) \overset{n \to \infty}{\to} 0$ in Lemma 24 (with properly chosen Θ and ζ). Hence, by Markov's inequality,

$$
\inf_{\hat{F}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P |\hat{F} - F(P)| \geq s \cdot \inf_{\hat{F}} \sup_{P \in \mathcal{M}_S} \mathbb{P}_P \left(|\hat{F} - F(P)| \geq s \right) \geq \frac{s}{2} \gtrsim \sqrt{\frac{S}{n \ln n}}
$$

which is a desired contradiction to (23)!

B.2. Lower Bound on the "Variance"

To establish the second lower bound, we will essentially reduce the sorted distribution estimation problem to the traditional distribution estimation where labels are required. Specifically, we consider the scenario where we have known a priori that the probability vector is sorted, in which case an accurate estimator for $P^<$ can be easily transformed into an accurate estimator for P. Then in this scenario, we recover the traditional $\Omega(\sqrt{\frac{S}{n}})$ lower bound for estimating P. The reason why this lower bound does not hold for large S is that when S exceeds some threshold, the prior knowledge that the probability vector is sorted starts to make adjacent entries become informative on the inference of the entry in the middle.

For some constant $c > 0$ large enough, define

$$
S' \triangleq S \wedge \left(\frac{n}{c} \right)^{\frac{1}{3}}
$$

as the new support size, and without loss of generality we assume that $S' = 2T + 1$ for some integer T. Fixing some $\lambda > 0$ to be determined later, we associate a probability vector $P_\epsilon \in \mathcal{M}_S$ to any binary vector $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_T) \in \{\pm 1\}^T$ as follows:

$$
P_\epsilon \triangleq (0, 0, \cdots, 0, x_1 + \lambda \epsilon_1, \cdots, x_T + \lambda \epsilon_T, y_1 - \lambda \epsilon_1, \cdots, y_T - \lambda \epsilon_T, 1 - \sum_{i=1}^T (x_i + y_i))
$$

$$
\text{with} \quad x_i \triangleq \frac{c(t+i)^2}{100n}, \quad y_i \triangleq \frac{c(t+T+i)^2}{100n}, \quad \text{and} \quad t \triangleq \left(\frac{n}{cT} \right)^{\frac{1}{2}}.
$$

Hence, as long as

$$\lambda \in \left(0, \frac{ct}{100n}\right), \tag{25}$$

it is easy to check that P_ϵ is sorted in an ascending order for any $\epsilon \in \{\pm 1\}^T$. As a result, restricting to the subclass $P \in \mathcal{P} \triangleq \{P_\epsilon : \epsilon \in \{\pm 1\}^T\} \subset \mathcal{M}_S$, estimating $P^<$ is equivalent to estimating P. In other words,

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P^<\|_1 \geq \inf_{\hat{P}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \|\hat{P} - P\|_1. \tag{26}$$

Next we lower bound the RHS of (26). Consider a uniform prior on $\epsilon \in \{\pm 1\}^T$, the fact that the Bayes risk under any prior is always a lower bound for the minimax risk gives that

$$\inf_{\hat{P}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \|\hat{P} - P\|_1 \geq \inf_{\hat{P}} \mathbb{E}_\epsilon \mathbb{E}_{P_\epsilon} \|\hat{P} - P_\epsilon\|_1 \geq \frac{T\lambda}{10} \cdot \inf_{\hat{P}} \mathbb{P}\left(\|\hat{P} - P_\epsilon\|_1 \geq \frac{T\lambda}{10}\right).$$

Defining the test function $\hat{\epsilon} \triangleq \arg\min_{\epsilon \in \{\pm 1\}^T} \|\hat{P} - P_\epsilon\|_1$, by triangle inequality it is straightforward to see that the event $d_{\mathrm{H}}(\hat{\epsilon}, \epsilon) \geq \frac{T}{5}$ implies $\|\hat{P} - P_\epsilon\|_1 \geq \frac{T\lambda}{10}$, where $d_{\mathrm{H}}(\cdot, \cdot)$ is the Hamming metric $d_{\mathrm{H}}(x, y) = \sum_{i=1}^T \mathbb{1}(x_i \neq y_i)$. Consequently, we further have

$$\inf_{\hat{P}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \|\hat{P} - P\|_1 \geq \frac{T\lambda}{10} \cdot \inf_{\hat{\epsilon}} \mathbb{P}\left(d_{\mathrm{H}}(\hat{\epsilon}, \epsilon) \geq \frac{T}{5}\right). \tag{27}$$

To lower bound the RHS of (27), we introduce the distance-based Fano's inequality as follows:

Lemma 26 *([Duchi and Wainwright, 2013](#), Corollary 1) Let random variables V and $\hat{V}$ take value in $\mathcal{V}$, V be uniform on some finite $\mathcal{V}$, and $V - X - \hat{V}$ form a Markov chain. Let d be any metric on $\mathcal{V}$, and for $t > 0$, define*

$$N_{\max}(t) \triangleq \max_{v \in \mathcal{V}} |v' \in V : d(v, v') \leq t|, \qquad N_{\min}(t) \triangleq \min_{v \in \mathcal{V}} |v' \in V : d(v, v') \leq t|.$$

If $N_{\max}(t) + N_{\min}(t) < |\mathcal{V}|$, the following inequality holds:

$$\mathbb{P}(d(V, \hat{V}) > t) \geq 1 - \frac{I(V; X) + \ln 2}{\ln \frac{|\mathcal{V}|}{N_{\max}(t)}}$$

where $I(V; X) \triangleq \mathbb{E}_{P_{V,X}} \left[\ln \frac{dP_{V,X}}{dP_V \times dP_X}\right]$ denotes the mutual information between V and X.

Applying Lemma 26 to the Markov chain $\epsilon - X - \hat{\epsilon}$ with Hamming metric and $t = \frac{T}{5}$, by Lemma 29 we know that $\frac{N_{\max}(t)}{|\mathcal{V}|} \leq \exp(-\frac{T}{8})$. Moreover, in the Poissonized model we have

$$
\begin{aligned}
I(\epsilon; X) &\overset{(a)}{\leq} \mathbb{E}_\epsilon D(P_{X|\epsilon} \| P_0) \\
&= \sum_{i=1}^T \left(\mathbb{E}_\epsilon D(\mathsf{Poi}(n(x_i + \lambda\epsilon_i)) \| \mathsf{Poi}(nx_i)) + \mathbb{E}_\epsilon D(\mathsf{Poi}(n(y_i - \lambda\epsilon_i)) \| \mathsf{Poi}(ny_i))\right) \\
&\overset{(b)}{\leq} n\lambda^2 \cdot \sum_{i=1}^T \left(\frac{1}{x_i} + \frac{1}{y_i}\right) \\
&\overset{(c)}{\leq} 200T \cdot n\lambda^2
\end{aligned}
$$

where P_0 is the probability measure P_ϵ applied to $\epsilon = \mathbf{0}$, (a) follows from the variational representation of mutual information

$$I(X;Y) = \inf_{Q_Y} \mathbb{E}_{P_X} D(P_{Y|X} \| Q_Y),$$

(b) follows from $D(\mathsf{Poi}(\lambda_1) \| \mathsf{Poi}(\lambda_2)) = \lambda_1 \ln \frac{\lambda_1}{\lambda_2} - \lambda_1 + \lambda_2 \leq \frac{(\lambda_1 - \lambda_2)^2}{\lambda_2}$, and (c) follows from $x_i, y_i \geq (200T)^{-1}$. Consequently, a combination of (26), (27) and Lemma 26 yields

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \| \hat{P} - P^< \|_1 \geq \frac{T\lambda}{10} \left(1 - \frac{800nT^2\lambda^2 + \ln 2}{T/8} \right). \tag{28}$$

Choosing $\lambda = \frac{c'}{\sqrt{nT}}$, for c' small enough the condition (25) is fulfilled, and by (28) and the choice of T we conclude that

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \| \hat{P} - P^< \|_1 \gtrsim \sqrt{\frac{T}{n}} \gtrsim \sqrt{\frac{S}{n}} \wedge n^{-\frac{1}{3}},$$

establishing the second term of Theorem 23.

Appendix C. Auxiliary Lemmas

Lemma 27 *(Han et al., 2016, Lemma 28) Let $p_n(x) = \sum_{\nu=0}^{n} a_\nu x^\nu$ be a polynomial of degree at most n such that $|p_n(x)| \leq A$ for $x \in [a, b]$. Then*

1. *If $a + b \neq 0$, then*

$$|a_\nu| \leq 2^{7n/2} A \left| \frac{a+b}{2} \right|^{-\nu} \left(\left| \frac{b+a}{b-a} \right|^n + 1 \right), \qquad \nu = 0, \cdots, n.$$

2. *If $a + b = 0$, then*

$$|a_\nu| \leq A b^{-\nu} (\sqrt{2} + 1)^n, \qquad \nu = 0, \cdots, n.$$

Lemma 28 *Hoeffding (1963) For independent and identically distributed random variables $X_1, \cdots, X_n$ with $a \leq X_i \leq b$ for $1 \leq i \leq n$, denote $S_n = \sum_{i=1}^{n} X_i$, we have for any $t > 0$,*

$$\mathbb{P}\left\{ |S_n - \mathbb{E}[S_n]| \geq t \right\} \leq 2 \exp\left(-\frac{2t^2}{n(b-a)^2} \right).$$

Lemma 29 *(Mitzenmacher and Upfal, 2005, Theorem 5.4) For $X \sim \mathsf{Poi}(\lambda)$ or $X \sim \mathsf{B}(n, \frac{\lambda}{n})$ and any $\delta > 0$, we have*

$$\mathbb{P}(X \geq (1+\delta)\lambda) \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^\lambda \leq \exp\left(-\frac{(\delta^2 \wedge \delta)\lambda}{3}\right),$$

$$\mathbb{P}(X \leq (1-\delta)\lambda) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}} \right)^\lambda \leq \exp\left(-\frac{\delta^2\lambda}{2}\right).$$

Appendix D. Proof of Main Lemmas

D.1. Proof of Corollary 3

By Theorem 1, it suffices to prove that for the empirical distribution P_n, we have

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|P_n^< - P^<\|_1 \asymp \sqrt{\frac{S}{n}}, \qquad n \geq S.$$

By Han et al. (2015), the upper bound follows from

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|P_n^< - P^<\|_1 \leq \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|P_n - P\|_1 \leq \sqrt{\frac{S}{n}}.$$

For the lower bound, note that (Berend and Kontorovich, 2013, Theorem 1) shows that for $n \geq S$,

$$\mathbb{E} \left| \frac{1}{n} \mathsf{B}(n, \frac{1}{S}) - \frac{1}{S} \right| \geq \sqrt{\frac{S-1}{2nS^2}}.$$

Consider $P = (\frac{1}{S}, \cdots, \frac{1}{S})$ to be the uniform distribution, then there is no difference between estimating $P^<$ and estimating P. Hence,

$$\mathbb{E}_P \|P_n^< - P^<\|_1 = \mathbb{E}_P \|P_n - P\|_1 \geq S \cdot \mathbb{E} \left| \frac{1}{n} \mathsf{B}(n, \frac{1}{S}) - \frac{1}{S} \right| \geq \sqrt{\frac{S-1}{2n}}$$

as desired.

D.2. Proof of Lemma 7

Without loss of generality assume that P, Q are sorted in an ascending order. Consider the coupling on (X, Y) which is uniformly distributed on the multiset $\{(p_1, q_1), \cdots, (p_S, q_S)\}$, we immediately have

$$S \cdot W(\mu_P, \mu_Q) \leq \|P - Q\|_1.$$

For the opposite inequality, let $r_{ij} \triangleq \mathbb{P}(X = p_i, Y = q_j)$, it's straightforward to see that

$$W(\mu_P, \mu_Q) = \min \sum_{i=1}^{S} \sum_{j=1}^{S} r_{ij} |p_i - q_j|$$

$$\text{s.t.} \sum_{i=1}^{S} r_{ij} = \frac{1}{S}, \qquad j = 1, 2, \cdots, S$$

$$\sum_{j=1}^{S} r_{ij} = \frac{1}{S}, \qquad i = 1, 2, \cdots, S$$

$$r_{ij} \geq 0, \qquad i, j = 1, 2, \cdots, S.$$

For linear programming, there must be a vertex of the simplex which attains the minimum of the objective. In other words,

$$S \cdot W(\mu_P, \mu_Q) = \min_{\sigma} \sum_{i=1}^{S} |p_i - q_{\sigma(i)}|$$

where σ ranges over all permutations on $\{1, \cdots, S\}$. Finally, note that for $i < j, k < l$, we have (by symmetry we assume that $p_i \le q_k$)

$$|p_i - q_l| + |p_j - q_k| = |p_i - q_k| + |q_k - q_l| + |p_j - q_k| \ge |p_i - q_k| + |p_j - q_l|.$$

In other words, switching $\sigma(i)$ and $\sigma(j)$ whenever $i < j$ and $\sigma(i) > \sigma(j)$ can only make the value of the objective smaller. Hence, the minimum is attained at $\sigma = \mathrm{id}$, and $S \cdot W(\mu_P, \mu_Q) \ge \|P - Q\|_1$, as desired.

D.3. Proof of Lemma 9

We prove Lemma 9 via figure. In the following figure, all curves represent different CDFs. Note that μ_P is a discrete distribution supported on $S = 4$ elements, and μ is an arbitrary distribution. The area of the yellow region exactly represents the Wasserstein distance $W(\mu_P, \mu)$. The randomization procedure picks up one point uniformly at random from each small interval $[0, \frac{1}{4}], [\frac{1}{4}, \frac{1}{2}], [\frac{1}{2}, \frac{3}{4}], [\frac{3}{4}, 1]$ on the y-axis, and then returns the corresponding inverse on the x-axis. Now from a vertical viewpoint, it is straightforward to verify that $\mathbb{E}W(\mu_P, \mu_Q)$ is also the yellow area, as desired.
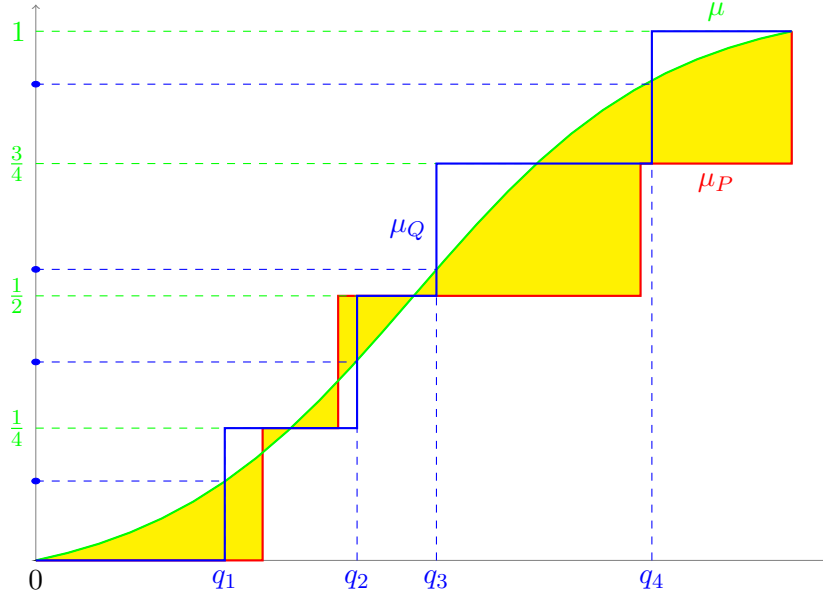
D.4. Proof of Lemma 11

Let $R(n, S, \pi), R_P(n, S, \pi)$ be the corresponding Bayes risks under prior π in the Multinomial and Poissonized models, respectively. By Jiao et al. (2015),

$$R_P(n, S, \pi) = \sum_{m=0}^{\infty} R(m, S, \pi) \cdot \mathbb{P}(\mathsf{Poi}(n) = m).$$

Note that $R(m, S, \pi)$ is non-increasing in m, by Lemma 29 and Markov's inequality we have

$$R_P(n, S, \pi) \leq 2 \cdot \mathbb{P}(\mathsf{Poi}(n) < \frac{n}{2}) + R(\frac{n}{2}, S, \pi) \leq 2\exp(-\frac{n}{8}) + R(\frac{n}{2}, S, \pi),$$

$$R_P(n, S, \pi) \geq R(2n, S, \pi) \cdot \mathbb{P}(\mathsf{Poi}(n) \leq 2n) \geq R(2n, S, \pi) \cdot \left(1 - \frac{\mathbb{E}\mathsf{Poi}(n)}{2n}\right) = \frac{1}{2}R(2n, S, \pi).$$

Now taking supremum over π and using the minimax theorem Wald (1950) complete the proof.

D.5. Proof of Lemma 18

By definition of $\mu_{P,j}$ and A_j, the claimed result is equivalent to

$$\mathbb{P}\left(\left|\sum_{i \in A_j}\left(g_{k,x_j}(\hat{p}_{i,2}) - (p_i - x_j)^k\right)\right| > \sqrt{S_j \ln n} \cdot \left(\frac{c_3 j \ln n}{n}\right)^k\right) \leq 2n^{-4}.$$

Let $Z_i \triangleq g_{k,x_j}(\hat{p}_{i,2}) - (p_i - x_j)^k$, and $M_i \triangleq \sup_{\hat{p}_{i,2}:|\hat{p}_{i,2}-x_j|\leq 3c_1 j \ln n/n}|Z_i|$. We first establish an upper bound for M_i with $i \in A_j$: firstly, the condition $i \in A_j$ implies that $p_i \in \tilde{I}_j$, and thus

$$|p_i - x_j|^k \leq \left|\frac{c_1(j+1)^2 \ln n}{n} - \frac{c_1 j(j-1) \ln n}{n}\right|^k \leq \left(\frac{4c_1 j \ln n}{n}\right)^k. \tag{29}$$

To upper bound $|g_{k,x_j}(\hat{p}_{i,2})|$, we introduce the following lemma:

Lemma 30 *Let $k \geq 1, np \in \mathbb{N}$ and $x \in [0,1]$. If $\max\{|x-p|, \sqrt{\frac{4pk}{n}}\} \leq \Delta$, we have*

$$|g_{k,x}(p)| = \left|\sum_{l=0}^{k}\binom{k}{l}(-x)^{k-l}\prod_{l'=0}^{l-1}\left(p - \frac{l'}{n}\right)\right| \leq (2\Delta)^k.$$

The proof of Lemma 30 is postponed to the end of this subsection. The conditions of Lemma 30 are fulfilled by $\Delta = \frac{3c_1 j \ln n}{n}$ as long as $c_1 > \frac{4}{3}c_2$, and thus

$$|g_{k,x_j}(\hat{p}_{i,2})| \leq \left(\frac{6c_1 j \ln n}{n}\right)^k. \tag{30}$$

As a result, a combination of (29) and (30) ensures that

$$M_i \leq \left(\frac{10c_1 j \ln n}{n}\right)^k, \qquad \forall i \in A_j. \tag{31}$$

Define a new random variable $\tilde{Z}_i \triangleq \max\{\min\{Z_i, M_i\}, -M_i\}$ as the truncated version of Z_i, the Hoeffding's inequality is about to be applied to the independent and bounded $\tilde{Z}_i$. We need to show that $\tilde{Z}_i$ and Z_i are indeed close in expectation. Clearly,

$$|\mathbb{E}(\tilde{Z}_i - Z_i)| \leq \mathbb{E}|g_{k,x_j}(\hat{p}_{i,2})|\mathbb{1}(|\hat{p}_{i,2} - x_j| > \Delta_j)$$

$$= \sum_{m:|m-nx_j|>n\Delta_j}|g_{k,x_j}(\frac{m}{n})| \cdot \mathbb{P}(\mathsf{Poi}(np_i) = m)$$

$$\leq \sum_{m:|m-nx_j|>n\Delta_j}2^k\left|\frac{m}{n} - x_j\right|^k \cdot \mathbb{P}(\mathsf{Poi}(np_i) = m)$$

where $\Delta_j \triangleq \frac{3c_1 j \ln n}{n}$ and we have used Lemma 30 in the last step. If $m + 1 > n(x_j + \Delta_j)$, we have

$$\frac{\mathbb{P}(\mathsf{Poi}(np_i) = m+1)}{\mathbb{P}(\mathsf{Poi}(np_i) = m)} = \frac{np_i}{m+1} \leq \frac{c_1(j+1/2)^2 \ln n}{c_1 j(j+2) \ln n} \leq 1 - \frac{1}{4j}.$$

Let $m_{\max}$ be the largest integer such that $m_{\max} \leq n(x_j + \Delta_j)$, by Lemma 29 and choosing $c_1 > 0$ large enough (as in Lemma 17) we have $\mathbb{P}(\mathsf{Poi}(np_i) = m_{\max}) \leq n^{-5}$. Hence,

$$\sum_{m:m-nx_j>n\Delta_j} \left| \frac{m}{n} - x_j \right|^k \cdot \mathbb{P}(\mathsf{Poi}(np_i) = m) \leq \sum_{l=0}^{\infty} (\Delta_j + \frac{l}{n})^k \cdot n^{-5} \left(1 - \frac{1}{4j} \right)^l$$

$$\leq n^{-5} \Delta_j^k \cdot \sum_{l=0}^{\infty} \exp\left(\frac{kl}{n\Delta_j} - \frac{l}{4j} \right)$$

$$\leq n^{-5} \Delta_j^k \left[1 - \exp\left(-\frac{1}{4j} + \frac{k}{n\Delta_j} \right) \right]^{-1}$$

$$\leq n^{-5} \Delta_j^k \cdot \left[1 - \exp\left(-\frac{1}{M} \left(\frac{1}{4} - \frac{c_2}{3c_1} \right) \right) \right]^{-1} \leq cMn^{-5}\Delta_j^k$$

where c is a constant depending only on c_1, c_2 when $3c_1 > 4c_2$.

The case where $m < n(x_j - \Delta_j)$ can be handled using similar arguments. As a result, as long as $c_1 > 2c_2$, we have

$$|\mathbb{E}(\tilde{Z}_i - Z_i)| \leq \frac{cM}{n^5} \cdot \left(\frac{6c_1 j \ln n}{n} \right)^k. \tag{32}$$

Note that $\mathbb{E}Z_i = 0$ by (5), and $|\tilde{Z}_i| \leq M_i \leq \left(\frac{9c_1 j \ln n}{n} \right)^k$ by (31), Hoeffding's inequality (cf. Lemma 28) with (32) yields

$$\mathbb{P}\left(\left| \sum_{i \in A_j} Z_i \right| > \sqrt{S_j \ln n} \cdot \left(\frac{c_3 j \ln n}{n} \right)^k \right)$$

$$\leq \mathbb{P}\left(\left| \sum_{i \in A_j} (\tilde{Z}_i - \mathbb{E}\tilde{Z}_i) \right| > \sqrt{S_j \ln n} \cdot \left(\frac{c_3 j \ln n}{n} \right)^k - \frac{24MS_j}{n^5} \cdot \left(\frac{6c_1 j \ln n}{n} \right)^k \right)$$

$$\leq 2 \exp\left(-\frac{\left(\sqrt{S_j \ln n} \cdot \left(\frac{c_3 j \ln n}{n} \right)^k - \frac{cMS_j}{n^5} \cdot \left(\frac{6c_1 j \ln n}{n} \right)^k \right)^2}{2S_j \left(\frac{10c_1 j \ln n}{n} \right)^{2k}} \right) \leq 2n^{-4}$$

as long as $c_3 > 30c_1$, as desired.

Proof [Proof of Lemma 30] The Charlier polynomial $c_k(u, a)$ for $u \in \mathbb{N}$ is defined as

$$c_k(u, a) \triangleq \sum_{l=0}^{k} (-1)^{k-l} \binom{k}{l} \frac{(u)_l}{a^l}$$

where $(u)_l \triangleq u(u-1)\cdots(u-l+1)$ is the falling factorial. For the Charlier polynomial, the following identities hold Peccati and Taqqu (2011), (Jiao et al., 2017, Eqn. (574)):

$$\sum_{k=0}^{\infty} \frac{c_k(u,a)}{k!} t^k = e^{-t}\left(1+\frac{t}{a}\right)^u,$$

$$\sum_{l=0}^{k} \binom{k}{l}(-b)^{k-l}a^l c_l(u,a) = (a+b)^k c_k(u,a+b).$$

The function $g_{k,x}(p)$ is related to the Charlier polynomial via the identity $g_{k,x}(p) = x^k c_k(np, nx)$, and thus the previous identities translate into the following:

$$\sum_{k=0}^{\infty} \frac{g_{k,p}(p)}{k!} t^k = e^{-pt}(1+\frac{t}{n})^{np}, \tag{33}$$

$$g_{k,x}(p) = \sum_{l=0}^{k} \binom{k}{l}(p-x)^{k-l}g_{l,p}(p). \tag{34}$$

We can rewrite (33) into the following form:

$$\sum_{k=0}^{\infty} \frac{g_{k,p}(p)}{k!} t^k = \left[e^{-\frac{t}{n}}\left(1+\frac{t}{n}\right)\right]^{np} = \left[\left(\sum_{l=0}^{\infty}\frac{1}{l!}(-\frac{t}{n})^l\right)\left(1+\frac{t}{n}\right)\right]^{np}$$

$$= \left(1 - \sum_{l=2}^{\infty}\frac{l-1}{l!}(-\frac{t}{n})^l\right)^{np}.$$

Comparing the coefficients of t^k at both sides yields

$$|g_{k,p}(p)| \leq \frac{k!}{n^k} \cdot \sum_{1\leq r\leq k/2}\binom{np}{r}\sum_{\sum_{i=1}^r l_i=k, l_i\geq 2}\prod_{i=1}^r \frac{l_i-1}{l_i!}$$

$$\leq \frac{k!}{n^k} \cdot \sum_{1\leq r\leq k/2}\binom{np}{r}\sum_{\sum_{i=1}^r l_i=k, l_i\geq 2} 1$$

$$= \frac{k!}{n^k} \cdot \sum_{1\leq r\leq k/2}\binom{np}{r}\binom{k-r-1}{r-1}$$

$$\leq \frac{k!}{n^k} \cdot \sum_{1\leq r\leq k/2}\binom{np}{r}2^{k-r-1}.$$

We distinguish into two cases: if $np \geq k$, we have $\binom{np}{r} \leq \binom{np}{k/2}$, and thus

$$|g_{k,p}(p)| \leq \frac{k!}{n^k} \cdot \binom{np}{k/2}2^k \leq \left(\frac{4pk}{n}\right)^{\frac{k}{2}}.$$

If $np < k$, we use the inequality $\binom{np}{r} \leq 2^{np} \leq 2^k$ to upper bound $|g_{k,p}(p)|$ as

$$|g_{k,p}(p)| \leq \frac{k!}{n^k} \cdot 2^{2k} = \left(\frac{4k}{n}\right)^k.$$

Combining these two cases, and using the assumption on Δ and (34), we conclude that

$$|g_{k,x}(p)| \leq \sum_{l=0}^{k} \binom{k}{l} \Delta^{k-l} \cdot \Delta^l = (2\Delta)^k$$

as desired. ∎

D.6. Proof of Lemma 13

By Corollary 19, the conditions of Lemma 20 are satisfied with probability at least $1 - Sn^{-5} - 2Mn^{-4} \geq 1 - 3Sn^{-4}$. Let $\mu_{P,j}$ and μ_j be given in (7) and (2), (3) respectively, we have

$$|\hat{F} - F(P)| \leq \sum_{j=1}^{M} \left| \int_{\mathbb{R}} f(x)(\mu_j(dx) - \mu_{P,j}(dx)) \right|.$$

By (9), for any degree-K polynomial $P_j(x) = \sum_{k=0}^{K} a_{k,j} x^k$ on $\tilde{I}_j$, we further have

$$|\hat{F} - F(P)| \leq \sum_{j=1}^{M} \left(\int_{\mathbb{R}} |f(x) - P_j(x)|(\mu_j(dx) + \mu_{P,j}(dx)) + \sum_{k=1}^{K} |a_{k,j}| \cdot 2\sqrt{S_j \ln n} \left(\frac{c_3 j \ln n}{n} \right)^k \right).$$

By assumption, $|P_j(x) - f(x_j)| \leq \|f - P\|_{\infty, \tilde{I}_j} + |f(x) - f(x_j)| \leq 2M_j$ for any j, by Lemma 27 we have

$$|a_{k,j}| \leq 2^{6K+3} M_j \left(\frac{c_1 j \ln n}{5n} \right)^{-k}, \qquad k \geq 1.$$

As a result, the second term can be upper bounded as

$$\sum_{k=1}^{K} |a_{k,j}| \cdot 2\sqrt{S_j \ln n} \left(\frac{c_3 \ln n}{n} \right)^k \leq \sum_{k=1}^{K} 2^{6K+3} M_j \left(\frac{c_1 j \ln n}{5n} \right)^{-k} \cdot 2\sqrt{S_j \ln n} \left(\frac{c_3 j \ln n}{n} \right)^k$$

$$= 16 c_2 n^{c_2(6 \ln 2 + \ln(5c_3/c_1))} (\ln n)^{\frac{3}{2}} \cdot M_j \sqrt{S_j} \lesssim n^\epsilon \cdot M_j \sqrt{S_j}$$

as long as $c_2(6 \ln 2 + \ln(5c_3/c_1)) < \epsilon$.

For the corollary, note that Lemma 18 ensures that $\mu_1(\mathbb{R}) \leq S_1$ with high probability. Moreover, the condition $f(0) = 0$ ensures that

$$\inf_{P_1 \in \mathsf{Poly}_K} \|f - P_1\|_{\infty, \tilde{I}_1} \leq 2 \cdot \inf_{P_1 \in \mathsf{Poly}_K, P_1(0)=0} \|f - P_1\|_{\infty, \tilde{I}_1}.$$

The proof is complete.

D.7. Proof of Theorem 15

We first consider the entropy functional $H(P)$ with $f(x) = -x \ln x$. By Jiao et al. (2015),

$$\inf_{P_1 \in \mathsf{Poly}_K} \|f - P_1\|_{\infty, \tilde{I}_1} \asymp \frac{1}{n \ln n}.$$

For $j \geq 2$, (Ditzian and Totik, 1987, Theorem 7.2.1) shows that

$$\inf_{P_j \in \mathsf{Poly}_K} \|f - P_j\|_{\infty, \tilde{I}_j} \leq \frac{|\tilde{I}_j|^2}{2K^2} \sup_{x \in \tilde{I}_j} |f''(x)| \lesssim \frac{j^2}{n} \sup_{x \in \tilde{I}_j} |f''(x)|.$$

For $f(x) = -x \ln x$, the previous inequality gives

$$\inf_{P_j \in \mathsf{Poly}_K} \|f - P_j\|_{\infty, \tilde{I}_j} \lesssim \frac{1}{n \ln n}, \qquad \forall j \geq 2.$$

As a result,

$$\sum_{j=1}^{M} S_j \cdot \inf_{P_j \in \mathsf{Poly}_K} \|f - P_j\|_{\infty, \tilde{I}_j} \lesssim \frac{1}{n \ln n} \sum_{j=1}^{M} S_j = \frac{S}{n \ln n}. \tag{35}$$

For the second term in Lemma 13, we have $M_j \lesssim \frac{j(\ln n)^2}{n}$. Hence, by (20),

$$\sum_{j=1}^{M} n^\epsilon M_j \sqrt{S_j} \lesssim \sum_{j=1}^{M} \frac{j\sqrt{S_j}}{n^{1-\epsilon}} \lesssim n^\epsilon \left(\sqrt{\frac{S}{n}} \wedge n^{-\frac{1}{3}} \right). \tag{36}$$

The desired result for $H(P)$ now follows from (35), (36) and Lemma 13. The results for $F_\alpha(P), 0 < \alpha < 1$ can also be obtained in a similar way.

Next we look at the support size functional $S(P)$ with $f(x) = \mathbb{1}(x \neq 0)$. Here to apply Lemma 13, it suffices to consider $j = 1$ with the corresponding interval $\tilde{I}'_1 = \{0\} \cup [\frac{1}{k}, \frac{c_1 \ln n}{n}]$. By Wu and Yang (2015),

$$\inf_{P_1 \in \mathsf{Poly}_K} \|f - P_1\|_{\infty, \tilde{I}'_1} \asymp \exp\left(-\Theta\left(\sqrt{\frac{n \ln k}{k}} \right) \right).$$

In addition, $M_1 = 1$ and $M_j = 0$ for any $j \geq 2$, by Lemma 13 we know that

$$\sup_{P \in \mathcal{D}_k} \mathbb{E}_P |\hat{S} - S(P)| \lesssim k \cdot n^{-3} + S_1 \cdot \exp\left(-\Theta\left(\sqrt{\frac{n \ln k}{k}} \right) \right) + n^\epsilon \sqrt{S_1}$$

$$\lesssim k \left[\exp\left(-\Theta\left(\sqrt{\frac{n \ln k}{k}} \right) \right) + \frac{n^\epsilon}{\sqrt{k}} \right].$$

The proof is complete.