Doubly-Competitive Distribution Estimation

Yi Hao [1] **Alon Orlitsky** [1]

Abstract

Distribution estimation is a statistical-learning cornerstone. Its classical *min-max* formulation minimizes the estimation error for the worst distribution, hence under-performs for practical distributions that, like power-law, are often rather simple. Modern research has therefore focused on two frameworks: *structural* estimation that improves learning accuracy by assuming a simple structure of the underlying distribution; and *competitive*, or *instance-optimal*, estimation that achieves the performance of a genie-aided estimator up to a small excess error that vanishes as the sample size grows, regardless of the distribution. This paper combines and strengthens the two frameworks. It designs a single estimator whose excess error vanishes both at a universal rate as the sample size grows, as well as when the (unknown) distribution gets simpler. We show that the resulting algorithm significantly improves the performance guarantees for numerous competitive- and structural-estimation results. The algorithm runs in near-linear time and is robust to model misspecification and domain-symbol permutations.

1. Introduction

Estimating large-alphabet distributions from their samples is a fundamental statistical-learning staple. Over the past few decades, distribution estimation has found numerous applications, ranging from language modeling (Chen & Goodman, 1999) to biological studies (Armaanzas et al., 2008), and has been extensively studied. In the following subsections, we formalize the discussion and present major research frameworks used in the field.

1.1. Distribution Estimation

Let Δ_k denote the collection of distributions over the discrete *alphabet* $[k] := \{1, \ldots, k\}$. Let $[k]^*$ be the set of finite-length sequences over $[k]$. An *estimator* is a mapping $\hat{p} : [k]^* \to \Delta_k$ that associates with every sequence x^n a distribution $\hat{p}(x^n) \in \Delta_k$. Let $X^n := X_1, \ldots, X_n$ be an i.i.d. sample sequence from an unknown p. Our objective is to find an estimator $\hat{p}$ such that $\hat{p}(X^n)$ approximates p well.

Specifically, for two distributions $p, q \in \Delta_k$, let $\ell(p, q)$ be the *loss* when approximating distribution p by estimate q. The loss of estimating p by $\hat{p}(X^n)$ is therefore $\ell(p, \hat{p}(X^n))$. We also consider the expected loss, known as *risk*,

$$r_n^\ell(p, \hat{p}) := \mathbb{E}_{X^n \sim p} L(p, \hat{p}(X^n)).$$

The two most important losses for distribution estimation are the KL-divergence $D(p \parallel q) := \sum_{x \in [k]} p_x \log \frac{p_x}{q_x}$, and the ℓ_1-distance $|p - q| := \sum_{x \in [k]} |p_x - q_x|$. *We study mainly the KL-loss, hence abbreviate* r^{KL} *as simply* r.

Next, we formalize the uncertainty about the distribution and the three common measures for the approximation quality: min-max, structural, and competitive estimation.

1.2. Previous Works

MIN-MAX

While the underlying distribution p is unknown, it often belongs to a known distribution collection $\mathcal{P}$. The *worst-case risk* of an estimator $\hat{p}$ over all distributions in $\mathcal{P}$ is

$$r_n^\ell(\mathcal{P}, \hat{p}) := \max_{p \in \mathcal{P}} r_n^\ell(p, \hat{p}),$$

and the minimal possible worst-case risk for $\mathcal{P}$, incurred by any estimator, is the *min-max risk*,

$$r_n^\ell(\mathcal{P}) := \min_{\hat{p}} r_n^\ell(\mathcal{P}, \hat{p}) = \min_{\hat{p}} \max_{p \in \mathcal{P}} r_n^\ell(p, \hat{p}).$$

The most classical and widely-studied class of distributions is simply the set Δ_k of all discrete distributions. The problem of determining $r_n^\ell(\Delta_k)$ up to the first order was introduced by (Cover, 1972) and studied in a sequence of papers (Krichevsky & Trofimov, 1981; Braess et al., 2002; Paninski, 2005). Among the many results on

[1] Department of Electrical and Computer Engineering, University of California, San Diego, USA. Correspondence to: Yi Hao <yih179@eng.ucsd.edu>, Alon Orlitsky <alon@ucsd.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

the topic, (Braess & Sauer, 2004) showed that for KL-divergence, as $n/k \to \infty$, the min-max KL-risk satisfies $r_n(\Delta_k) = (1 + o(1))\frac{k-1}{2n}$, achieved by a variant of the add-3/4 estimator. On the other hand, (Paninski, 2005) proved that as $k/n \to \infty$, the optimal KL-risk becomes $r_n(\Delta_k) = (1 + o(1)) \log \frac{k}{n}$, which is achieved by add-constant estimators. Similar results for other loss measures like ℓ_1-distance can be found in (Kamath et al., 2015).

Beyond min-max The success of add-constant estimators in achieving the classical min-max risks does not extend to practical applications. One possible explanation is that practical distributions, like power-law, or Poisson, are often rather simple and can be estimated more efficiently and accurately than the worst distribution targeted by the min-max paradigm. The desire to construct estimators that perform better on practical distributions has led to the following two frameworks.

STRUCTURAL

Instead of considering arbitrary underlying distributions, the structural approach focuses on learning distributions that posses a natural structure, such as monotonicity, log-concavity, and m-modality. In many cases, structural assumptions lead to more effective estimators that provably perform better on the corresponding distribution classes.

For example, (Kamath et al., 2015) showed that for fixed k, as n increases, the empirical estimator achieves the min-max ℓ_1-risk over Δ_k,

$$r_n^{\ell_1}(\Delta_k) = (1 + o(1))\sqrt{\frac{2(k-1)}{\pi n}}.$$

In many practical applications, the alphabet k is often large, hence several papers considered structured distributions (Acharya et al., 2017; Diakonikolas et al., 2016; Kamath, 2014; Chan et al., 2013; Daskalakis et al., 2012; Jankowski & Wellner, 2009; Feldman et al., 2008). For example, for the collection $\mathcal{M}_k^{t,m}$ of t-mixture m-modal distributions over $[k]$, more sophisticated estimators, e.g., (Acharya et al., 2017) attain

$$r_n^{\ell_1}(\mathcal{M}_k^{t,m}) = \Theta\left(\frac{tm \log k}{n}\right)^{1/3},$$

which for $k/\log k \gg n^{1/3}(tm)^{2/3}$, is lower than $r_n^{\ell_1}(\Delta_k)$.

Drawbacks The structural approach leverages the structure assumptions to design more efficient estimators, thus has the drawback of relying on the hypothetical models.

For example, to learn t-mixture m-modal distributions efficiently as above, one needs to ensure the correctness of the structure assumption and know both t and m up to constant factors. While it may seem possible to use hypothesis testing to find the best parameters, existing work on distribution property testing shows that even testing whether a distribution is m-modal requires a non-trivial number of samples (Canonne et al., 2018). Hence, when t and m are relatively large, finding the best parameters may require many samples.

In addition, many structures possessed by real-world distributions, for example, mixtures of log-concave and log-convex, have not been addressed before.

COMPETITIVE

Instead of relying on often-uncertain structural assumptions, the competitive distribution estimation framework takes a different view and aims to design universally near-optimal estimators. Any reasonable estimator for i.i.d. distributions would assign the same probability to all symbols appearing the same number of times in the sample, and we let $\mathcal{Q}_{\text{nat}}$ denote this collection of *natural estimators*.

Our objective is to design a distribution estimator $\hat{p}$ that estimates every distribution nearly as well as the best estimator designed with prior knowledge of the true distribution p, but is restricted to be natural. Specifically, for any distribution $p \in \Delta_k$, the lowest risk of a natural estimator knowing p is

$$\tilde{r}_n^{\ell}(p, \mathcal{Q}_{\text{nat}}) := \min_{\hat{p}' \in \mathcal{Q}_{\text{nat}}} r_n^{\ell}(p, \hat{p}'),$$

and the *excess risk* of an arbitrary estimator $\hat{p}$ is

$$\tilde{r}_n^{\ell}(p, \hat{p}) := r_n^{\ell}(p, \hat{p}) - \tilde{r}_n^{\ell}(p, \mathcal{Q}_{\text{nat}}).$$

Therefore, the worst-case excess risk, or *competitive risk*, of the estimator $\hat{p}$ over all distribution in Δ_k is

$$\tilde{r}_n^{\ell}(\hat{p}) := \max_{p \in \Delta_k} \tilde{r}_n^{\ell}(p, \hat{p}).$$

This formulation was introduced in (Orlitsky & Suresh, 2015) who showed that a simple variant of the Good-Turing estimator $\hat{p}_{\text{GT}}$ achieves a vanishing competitive KL-risk of $\tilde{r}_n(\hat{p}_{\text{GT}}) \le (3 + o(1))/n^{1/3}$, regardless of the alphabet size, and a more involved estimator $\hat{p}_{\text{MI}}$ achieves $\tilde{\Theta}(\min\{k/n, 1/\sqrt{n}\})$, For ℓ_1-distance, (Valiant & Valiant, 2016) designed a linear-programming-based estimator $\hat{p}_{\text{LP}}$ and proved $\tilde{r}_n^{\ell_1}(\hat{p}_{\text{LP}}) = \mathcal{O}(1/\text{polylog}(n))$.

Drawbacks The upper bounds provided by the competitive approach apply to all distributions, and similar to the min-max approach, track the excess error of the worst distribution. As we now show, they are too lax for many practical distributions. Consider the following generalization of the ubiquitous power-law distributions. For $c > 0$, $\alpha > 1$, and large alphabet-size k, define the enveloped distribution collection $\mathcal{P}_k^{\alpha,c} := \{p \in \Delta_k : p_x \le c \cdot x^{-\alpha}\}$. It can be shown that for $n \in [k^{0.1}, k^2]$ there is a constant $C_{\alpha,c}$ depending on

α and c, such that the min-max KL-risk of $\mathcal{P}_k^{\alpha,c}$ satisfies

$$r_n(\mathcal{P}_k^{\alpha,c}) = C_{\alpha,c} \cdot n^{-\frac{\alpha-1}{\alpha}+o(1)}.$$

By simple algebra, for $\alpha > 2$ and large n, this term is smaller than $\Theta(\min\{k/n, 1/\sqrt{n}\})$, the lowest competitive risk of any estimator (Orlitsky & Suresh, 2015). Hence the guarantees the competitive framework provides do not suffice to address relatively "simple" common distributions.

2. New Results

The foregoing section reviewed the merits and drawbacks of classical and modern approaches to distribution-estimation. It noted that the min-max approach is "pessimistic" and often performs sub-optimally in both theory and practice. Of the modern frameworks, the structural approach works well if the structural assumptions are both correct and accurate, but fails otherwise, hence this approach is "local" but not "global". The competitive approach constructs universally near-optimal estimators, but provides the same guarantees regardless of the distribution's structure, potentially resulting in sub-optimal estimators for practical distributions, hence this approach is "global" but not "local".

This raises the question of whether a single estimator can be both "global" and "local". Namely, without any assumptions on the distribution, provide universal excess-loss guarantees for general distributions, and stronger excess-loss guarantees for simple distributions. For example, an estimator $\hat{p}$ such that for any distribution p, $\tilde{r}_n(p,\hat{p}) \le n^{-1/2}$, and yet if the distribution p happens to be in the enveloped power-law class $\mathcal{P}_k^{3,c}$, then $\tilde{r}_n(p,\hat{p}) \le n^{-3/4}$.

We answer this question in the affirmative, and present the first competitive *and* structural distribution estimator.

2.1. Definitions

Instant competitive loss For consistency, let us instantiate the loss ℓ as the KL-divergence, i.e., for $p, q \in \Delta_k$,

$$\ell(p,q) := D(p \parallel q).$$

Let $p \in \Delta_k$ be an unknown discrete distribution, and let x^n be a realization of $X^n \sim p$. The best natural estimator, knowing both p and x^n, incurs the minimal possible loss

$$\tilde{\ell}_{x^n}(p, \mathcal{Q}_{\text{nat}}) := \min_{\hat{p}' \in \mathcal{Q}_{\text{nat}}} \ell(p, \hat{p}'(x^n)),$$

and for this particular pair (p, x^n), the excess loss of an arbitrary estimator $\hat{p}$ is

$$\tilde{\ell}_{x^n}(p,\hat{p}) := \ell(p,\hat{p}(x^n)) - \tilde{\ell}_{x^n}(p, \mathcal{Q}_{\text{nat}}).$$

Hence for sequence x^n, the worst-case excess loss of $\hat{p}$ over Δ_k, or simply the *instance competitive loss* of $\hat{p}$, is

$$\tilde{\ell}_{x^n}(\hat{p}) := \max_{p\in\Delta_k} \tilde{\ell}_{x^n}(p,\hat{p}).$$

Permutation class For any distribution $p \in \Delta_k$, we denote by $\langle p \rangle$ the collection of distributions in Δ_k that are equal to p up to some permutation over $[k]$. Knowing $\langle p \rangle$ is equivalent to knowing the multiset of p but not p itself.

General notation For $X^n \sim p \in \Delta_k$, the *multiplicity* of a symbol $x \in [k]$ is $N_x := \sum_{i=1}^n \mathbb{1}_{X_i = x}$, the number of times x appears in X^n. The *prevalence* of an integer μ is $\Phi_\mu := \sum_{x\in[k]} \mathbb{1}_{N_x=\mu}$, the number of symbols that appear μ times. Let $D := \sum_{\mu>0} \Phi_\mu$ be the number of distinct symbols in X^n, and let $D_\Phi := \sum_{\mu>0} \mathbb{1}_{\Phi_\mu>0}$ be the number of distinct positive multiplicities. Clearly, $D \ge D_\Phi$, and typically, $D \gg D_\Phi$. For example, if all symbols in the sequence X^n are distinct, then $D = n$, while D_Φ is just 1.

2.2. Main Results

We construct an explicit, near-linear-time computable distribution estimator $\hat{p}^*$ such that

Theorem 1. *For any distribution p, let $X^n \sim p$, then with probability at least $1 - n^{-8}$,*

$$\tilde{\ell}_{X^n}(\hat{p}^*) \le \widetilde{\mathcal{O}}\left(\frac{D_\Phi}{n}\right).$$

Note that the right-hand side is determined by just X^n, its computation requires no additional information about p.

The exact form of $\hat{p}^*$ can be found in Section 5, and the proof of Theorem 1 appears in the supplemental material.

Our main theorem implies the following new results and improvements on existing ones.

Global competitiveness In Section 3 we show that our estimator provides stronger estimation guarantees than many existing estimators: adaptive estimators (Corollary 3) such as the robust absolute discounting estimator (Ohannessian & Dahleh, 2012; Ben-Hamou et al., 2017); competitive estimators (Corollary 4) such as the modified Good-Turing estimator (Orlitsky & Suresh, 2015); and min-max estimators (Corollary 5).

Example: Section 3 shows that $D_\Phi \le \min\{\sqrt{2n}, k\}$. Corollary 4 then concludes that the excess loss $\tilde{\ell}_{X^n}(\hat{p}^*)$ is always at most $\widetilde{\mathcal{O}}\left(\min\{\sqrt{n}, k\}/n\right)$, providing a guarantee not only stronger than the $n^{-1/3}$ rate of the modified Good-Turing estimator, but also as strong as the more involved estimator in (Acharya et al., 2013; Orlitsky & Suresh, 2015).

Local competitiveness In Section 4, we use the theorem to establish eight new results on learning important structured distributions. We show that our estimator has strong excess-loss bounds for three important structured distribution families: T-value (Corollary 7 and 8), log-concave

(Corollary 9 and 10), and log-convex (Corollary 11, 12, and 13). Many common distributions are covered by these three classes. In particular, our results for power-law distributions (Corollary 12) are uniformly stronger than those in (Falahatgar et al., 2017) for all parameter regimes.

Example: Corollary 8 shows that for all uniform distributions, $\mathbb{E}[D_\Phi]$ is bounded above by $\tilde{\mathcal{O}}(n^{1/3})$, hence the algorithm's excess risk is at most $\tilde{\mathcal{O}}(n^{-2/3})$.

Robustness to model misspecification The structural approach often uses different estimators for different distribution classes. By contrast, our single estimator provides robust and adaptive guarantees for a variety of structural classes without any modification.

Example: Over uniform distributions, $\hat{p}^*$ achieves an excess risk of $\tilde{\mathcal{O}}(n^{-2/3})$ (Corollary 8), while for power law distributions with power parameter 1.5 , the same estimator achieves an excess risk of $\tilde{\mathcal{O}}(n^{-3/5})$ (Corollary 11).

Robustness to domain permutations The structural approach often assumes that we know how to order the symbols so that the underlying distribution would exhibit certain structure (such as power-law). As discussed in Section 4, this assumption may be impractical. By contrast, since the distribution of D_Φ is the same for all $p' \in \langle p \rangle$, the excess loss/risk guarantees of our algorithm are invariant under any permutation of the domain symbols.

Example: If under some unknown ordering of the domain symbols, the underlying distribution is a power-law with power parameter 1.5 , then Corollary 11 implies that our estimator achieves an excess risk of $\tilde{\mathcal{O}}(n^{-3/5})$.

Outline Besides Section 3 and 4 mentioned above, we present the exact form of our estimator in Section 5.

3. Global Competitiveness

In this section, we present several implications of Theorem 1 for the universal estimation guarantees of $\hat{p}^*$. In particular, we show that $\hat{p}^*$ is near-optimal under various classical and modern distribution learning frameworks, including min-max and competitive mentioned above.

Corollary 1. *For any distribution p,*

$$\tilde{r}_n(p, \hat{p}^*) \le \tilde{\mathcal{O}}\left(\frac{\mathbb{E}[D_\Phi]}{n}\right).$$

As in the proof of Theorem 1, $\tilde{\ell}_{X^n}(p, \hat{p}^*) \le \mathcal{O}(\log n)$ always. The corollary then follows from Theorem 1 itself.

Analogous to the previous definition of competitive distribution estimation, we can consider competing with an estimator that knows the probability multi-set. Specifically, for any distribution $p \in \Delta_k$, the lowest worst-case risk of a natural estimator knowing the multi-set of p is

$$\dot{r}^\ell_n(\langle p \rangle) := \min_{\hat{p}'} \max_{p' \in \langle p \rangle} r^\ell_n(p', \hat{p}'),$$

and an arbitrary estimator $\hat{p}$ has the *multi-set excess risk* of

$$\dot{r}^\ell_n(p, \hat{p}) := r^\ell_n(p, \hat{p}) - \dot{r}^\ell_n(\langle p \rangle).$$

For KL-divergence, the following lemma relates $\dot{r}^\ell_n$ to $\tilde{r}^\ell_n$.

Lemma 1. *(Orlitsky & Suresh, 2015) For any distribution $p \in \Delta_k$ and estimator $\hat{p}$,*

$$\max_{p' \in \langle p \rangle} \dot{r}_n(p', \hat{p}) \le \tilde{r}_n(p, \hat{p}).$$

Together with Corollary 2, the lemma yields,

Corollary 2. *For any distribution p,*

$$\max_{p' \in \langle p \rangle} \dot{r}_n(p, \hat{p}^*) \le \tilde{\mathcal{O}}\left(\frac{\mathbb{E}[D_\Phi]}{n}\right).$$

Adaptive optimality The min-max results (Krichevsky & Trofimov, 1981) imply that for any estimator, learning an arbitrary k-symbol distribution up to a certain KL-risk requires $\Omega(k)$ samples in the worst case. Since modern data science often considers applications over large alphabets, this is normally viewed as a negative result. However, as experience suggests, many practical distributions have small "effective alphabet sizes". For example, if we draw 10 samples from a geometric distribution with success probability 0.9, although the support size is infinite, with high probability, we shall observe at most 3 distinct symbols.

To formalize this intuition, for a given n, let the *effective alphabet size* of a distribution p be the expected number $\mathbb{E}[D]$ of distinct symbols that appear in $X^n \sim p$. As in (Falahatgar et al., 2017), given n, k, and d, let $\mathcal{P}_d$ be the collection of distributions in Δ_k satisfying $\mathbb{E}[D] \le d$. By Corollary 2, the performance of $\hat{p}^*$ over $\mathcal{P}_d$ is adaptive to d:

Corollary 3. *For all $d \ge 2$ and every distribution $p \in \mathcal{P}_d$,*

$$r_n(p, \hat{p}^*) \le \frac{d}{n} \log k + \tilde{\mathcal{O}}\left(\frac{d}{n}\right).$$

The following lemma shows the optimality of Corollary 3.

Lemma 2. *(Falahatgar et al., 2017) Let α be any constant greater than 1. There exist constants $c_0 > 0$ and n_0 such that for $d = n^{\frac{1}{\alpha}}$, any estimator $\hat{p}$, all $n > n_0$, and all $k > \max\{3n, 1.2^{\frac{1}{\alpha-1}} n^{\frac{1}{\alpha}}\}$,*

$$\max_{p \in \mathcal{P}_d} r_n(p, \hat{p}) \ge c_0 \frac{d}{n} \log k - \tilde{\mathcal{O}}\left(\frac{d}{n}\right).$$

Here we present two immediate implications. First, to learn a k-symbol distribution up to a certain KL-risk, the number of samples we need is at most $\tilde{\mathcal{O}}(\mathbb{E}[D]\log k)$, which is often much smaller than $\Omega(k)$. Second, in the extreme case when $k/n \to \infty$, the upper bound on $r_n(p, \hat{p}^*)$ is at most $(1+o(1))\log k$. Hence, our estimator achieves the min-max KL-risk over Δ_k to the right constant.

Competitive optimality Now we show that $\hat{p}^*$ is near-optimal under the competitive formulation described in Section 1.2. We begin by finding a simple upper bound for D_Φ, the number of distinct positive multiplicities. Since different multiplicities correspond to distinct symbols, D_Φ is at most the alphabet size k. On the other hand, since only distinct positive multiplicities count, $\sum_{\mu=1}^{D_\Phi} \mu \le n$. Hence, $D_\Phi \le \min\{k, \sqrt{2n}\}$, which together with Corollary 2 yields

Corollary 4. *For any distribution p,*

$$\tilde{r}_n(p, \hat{p}^*) \le \widetilde{\mathcal{O}}\left(\frac{\min\{k, \sqrt{n}\}}{n}\right).$$

The following lemma shows the optimality of Corollary 4.

Lemma 3. *(Orlitsky & Suresh, 2015) For any estimator $\hat{p}$,*

$$\max_{p\in\Delta_k} \tilde{r}_n(p, \hat{p}) \ge \widetilde{\Omega}\left(\frac{\min\{k, \sqrt{n}\}}{n}\right).$$

Min-max optimality The previous results show that $\hat{p}^*$ often achieves the min-max KL-risk $r_n(\Delta_k)$ to the right constant. Specifically,

Corollary 5. *Let α_0 be any constant greater than $1/2$. For any $\alpha > \alpha_0$ and $k > n^\alpha$,*

$$r_n(\Delta_k, \hat{p}^*) = (1 + o_n(1)) r_n(\Delta_k).$$

4. Local Competitiveness

We use Corollary 2 and 3 to establish eight new results on learning important structured distributions. We show that our estimator has strong excess-loss bounds for three important structured distribution families: T-value (Corollary 7 and 8), log-concave (Corollary 9 and 10), and log-convex (Corollary 11, 12, and 13). Many common distributions are covered by these three classes.

4.1. A Simple Bound on $\mathbb{E}[D_\Phi]$

By Corollary 2, the excess KL-risk $\tilde{r}_n(p, \hat{p}^*)$ of $\hat{p}^*$ in estimating p is upper bounded by $\widetilde{\mathcal{O}}\left(\mathbb{E}[D_\Phi]/n\right)$. Perhaps the most natural question to ask is: given n and p, how large is $\mathbb{E}[D_\Phi]$? To get a relatively simple closed-form expression for $\mathbb{E}[D_\Phi]$, we adopt the conventional "Poisson Sampling" technique where the sample size is an independent Poisson variable with mean n. By doing so, the multiplicities $N_x \sim \text{Poi}(np_x)$ independently of each other. Under Poisson sampling, the linearity of expectation implies

$$\mathbb{E}[D_\Phi] = n - \sum_{\mu>0} \prod_{x\in[k]} \left(1 - e^{-np_x}\frac{(np_x)^\mu}{\mu!}\right).$$

Expanding the right-hand side would give us an expression consisting of $n \cdot (2^k - 1)$ terms, which is hard to analyze. Hence, instead of evaluating $\mathbb{E}[D_\Phi]$ directly, we would like to work on its simple upper bounds. Given sampling parameter n, we partition the unit-length interval $(0, 1]$ into a sequence of sub-intervals,

$$I_j := \left((j-1)^2\frac{\log n}{n}, j^2\frac{\log n}{n}\right],\ 1 \le j \le \sqrt{\frac{n}{\log n}}.$$

For any distribution p, denote by p_{I_j} the number of probabilities p_x in I_j. Then,

Lemma 4. *For any distribution p,*

$$\mathbb{E}[D_\Phi] \le \mathcal{O}(\sum_{j\ge1} \min\{p_{I_j}, j\}) \cdot \log n.$$

In addition, since p is a distribution, for all j, $p_{I_j} \cdot \frac{j^2 \log n}{n} \le 1$, which in turn implies

$$\min\{p_{I_j}, j\} \le \min\left\{\frac{n}{j^2 \log n}, j\right\} < n^{\frac{1}{3}}.$$

More generally, let P_{I_j} denote the *sum* of probabilities p_x in I_j. Then,

$$\min\{p_{I_j}, j\} \le \min\left\{\frac{nP_{I_j}}{j^2 \log n}, j\right\} < (nP_{I_j})^{\frac{1}{3}}.$$

Combined, Corollary 2 and Lemma 4 yield

Corollary 6. *For any distribution p,*

$$\tilde{r}_n(p, \hat{p}^*) \le \widetilde{\mathcal{O}}\left(\frac{1}{n}\right)\sum_{j\ge1} \min\{p_{I_j}, j\}.$$

To illustrate the Corollary's significance, we present its implications for various distribution-learning problems.

4.2. T-Value Distributions

A uniform distribution can be described as a distribution whose positive probabilities take only a single value. As a generalization of this formulation, we call a distribution p a *T-value distribution* if its positive probabilities p_x can take T different values. Note that T-value distributions over $[k]$ can be viewed as mixtures of T uniform distributions over different subsets of $[k]$, and that these distributions generalize T-piecewise histogram distributions. Intuitively, for smaller values of T, we would expect the task of learning an unknown T-value distribution to be easier. The following corollary confirms this intuition.

Corollary 7. *For any T-value distribution p and $p' \in \langle p \rangle$,*

$$\tilde{r}_n(p', \hat{p}^*) \le \tilde{\mathcal{O}}\left(\frac{T^{\frac{2}{3}} \wedge n^{\frac{1}{6}}}{n^{\frac{2}{3}}}\right).$$

Note that $p \in \langle p \rangle$. To prove the corollary, observe that by our previous result, for all j,

$$\min\left\{p_{I_j}, j\right\} < (nP_{I_j})^{\frac{1}{3}}.$$

Note that for a T-value distribution, $p_{I_j} \neq 0$ for at most T different j values, say $j_1, \ldots, j_T$. By the above inequality and Corollary 6,

$$\tilde{r}_n(p, \hat{p}^*) \le \widetilde{\mathcal{O}}\left(\frac{1}{n}\right) \sum_{i=1}^{T} (nP_{I_{j_i}})^{\frac{1}{3}} \le \widetilde{\mathcal{O}}\left(\frac{T(n/T)^{\frac{1}{3}}}{n}\right).$$

combined with Corollary 4, this completes the proof.

Uniform Distributions Now we consider the collection $\mathcal{U}_k$ of 1-value distributions, i.e., uniform distributions over non-empty subsets of $[k]$. Our objective is to derive a result stronger than Corollary 7. Let S_p denote the support size of a distribution $p \in \mathcal{U}_k$. For all $x \in [k]$, p_x is either 0 or S_p^{-1}. Since $\{I_j, j \ge 1\}$ forms a partition of $(0, 1]$, there exists a unique j' such that $S_p^{-1} \in I_{j'}$, i.e.,

$$S_p^{-1} \in I_{j'} = \frac{\log n}{n}\left((j'-1)^2, j'^2\right],$$

which further implies $1 + \sqrt{n/(S_p \log n)} \ge j'$. Together with $D_\Phi \le D \le S_p$ and Corollary 6, this shows

Corollary 8. *Let p be an arbitrary distribution in $\mathcal{U}_k$, then*

$$\tilde{r}_n(p, \hat{p}^*) \le \tilde{\mathcal{O}}\left(\min\left\{\frac{1}{\sqrt{nS_p}}, \frac{S_p}{n}\right\}\right).$$

Note that the right-hand side is no more than $\tilde{\mathcal{O}}(n^{-2/3})$. Furthermore, in both the small alphabet regime where $S_p = \mathcal{O}(1)$ and the large alphabet regime where $S_p = \Omega(n)$, we have $\tilde{r}_n(p, \hat{p}^*) \le \tilde{\mathcal{O}}(n^{-1})$, which is fairly tight.

4.3. Log-Concave Distributions

The class of discrete log-concave distributions covers a variety of well-known distribution classes including binomial, Poisson, negative binomial, geometric, hypergeometric, hyperPoisson, Skellam, and Pólya-Eggenberger (Qu et al., 1990). We say a discrete distribution $p \in \Delta_k$ is *log-concave* if for all $x \in [k]$, $p_x^2 \ge p_{x-1} \cdot p_{x+1}$, and denote the collection of all such distributions by $\mathcal{L}_k$. Further, for all $\sigma > 0$, let $\mathcal{L}_k^{n,\sigma}$ denote the collection of $p \in \mathcal{L}_k$ whose standard deviation lies in $(\sigma \cdot \log^{-1} n, \sigma]$. Intuitively, one would expect the learning task over $\mathcal{L}_k^{n,\sigma}$ to be easier for smaller values of σ. The following corollary demonstrates the correctness of this intuition and shows the competitive performance of our estimator. Due to space considerations, we postpone its proofs to the supplemental material.

Corollary 9. *For any distribution $p \in \mathcal{L}_k^{n,\sigma}$ and $p' \in \langle p \rangle$,*

$$\tilde{r}_n(p', \hat{p}^*) \le \tilde{\mathcal{O}}\left((\sigma n)^{-\frac{1}{3}} \wedge \frac{\sigma}{n}\right).$$

For any $\sigma \gg 1$, the right-hand side is uniformly smaller than the bound $\tilde{\mathcal{O}}(\min\{k, \sqrt{n}\} \cdot n^{-1})$ in Corollary 4.

For mixtures of distributions in $\mathcal{L}_k^{n,\sigma}$, an analogous argument gives the following result.

Corollary 10. *Let p be a t-mixture of distributions in $\mathcal{L}_k^{n,\sigma}$ and p' be any distribution in $\langle p \rangle$,*

$$\tilde{r}_n(p', \hat{p}^*) \le \tilde{\mathcal{O}}\left((\sigma n)^{-\frac{1}{3}} \wedge \frac{t\sigma \wedge \sqrt{n}}{n}\right).$$

4.4. Log-Convex Distributions

While the T-value and log-concave families cover many common distributions, there are certainly more distribution classes to be explored. For example, a truncated power-law distribution is always log-convex. In this section, we consider two generic classes of log-convex distributions: power-law and HurwitzLerch Zeta distribution families.

Enveloped power-law distributions Consider the collection $\mathcal{P}_k^{\alpha,c} := \{p \in \Delta_k : p_x \le c \cdot x^{-\alpha}\}$ of enveloped (truncated) power-law distributions. Note that this definition generalizes power-law families, and that distributions in $\mathcal{P}_k^{\alpha,c}$ are not necessarily log-convex. We have the following result, whose proof appears in the supplemental material.

Corollary 11. *For any distribution $p \in \mathcal{P}_k^{\alpha,c}$ and $p' \in \langle p \rangle$,*

$$\tilde{r}_n(p', \hat{p}^*) \le \tilde{\mathcal{O}}_{c,\alpha}\left(n^{-\max\{\frac{\alpha}{\alpha+1}, \frac{1}{2}\}}\right).$$

The distribution collection $\mathcal{P}_k^{\alpha,c}$ has the interesting property that it is closed under mixtures. Hence, Corollary 11 also covers mixtures of enveloped power-law distributions.

Implications of Corollary 11 Let $p^\alpha \in \Delta_k$ be the truncated power-law distribution with power α that is truncated at k, i.e., $p_x^\alpha \propto x^{-\alpha}$, $\forall x \in [k]$. Clearly, we have $p^\alpha \in \mathcal{P}_k^{\alpha,c}$ for all $c \ge 1$. The recent work of (Falahatgar et al., 2017) shows that for $k > \{n, n^{\frac{1}{\alpha-1}}\}$ and any distribution $p' \in \langle p^\alpha \rangle$, the estimator $\hat{p}''$ proposed in (Ohannessian & Dahleh, 2012) satisfies

$$\dot{r}_n(p', \hat{p}'') \le \mathcal{O}_{c,\alpha}\left(n^{-\frac{2\alpha-1}{2\alpha+1}}\right).$$

A simple combination of Lemma 1 and Corollary 11 yields

Corollary 12. *For any distribution $p' \in \langle p^\alpha \rangle$,*

$$\dot{r}_n(p', \hat{p}^*) \le \tilde{r}_n(p, \hat{p}^*) \le \tilde{\mathcal{O}}_{c,\alpha}\left(n^{-\max\{\frac{\alpha}{\alpha+1}, \frac{1}{2}\}}\right).$$

Our approach has the following three advantages over the previous result in (Falahatgar et al., 2017). First, for all $\alpha > 0$, we have $-\alpha/(\alpha+1) < -(2\alpha-1)/(2\alpha+1)$, hence our guarantee is uniformly better than the previous one. Second, the previous result requires $k > \{n, n^{\frac{1}{\alpha-1}}\}$ to hold, which can be non-realistic for α close to 1. In comparison, our result does not require such conditions at all. Third, for small $\alpha < 1/2$, the previous result only implies a multi-set excess risk of $\mathcal{O}(n^{\Theta(1)})$, while Corollary 12 always yields $\tilde{\mathcal{O}}\left(n^{-1/2}\right)$ regardless of α.

Enveloped HurwitzLerch Zeta distributions For any distribution $p \in \Delta_k$, p is a (truncated) HurwitzLerch Zeta (HLZ) distribution (Gupta et al., 2008) if

$$p_x = \frac{1}{T(\theta, s, a, k)} \cdot \frac{\theta^x}{(a+x)^{s+1}},$$

for some parameter $s \ge 0$, $a \in [0, 1]$ and $\theta \in (0, 1]$, where the normalization factor $T(\theta, s, a, k) := \sum_{x \in [k]} \theta^x/(a+x)^{s+1}$. Analogously, consider the collection $\mathcal{H}_k^{\theta,s,a,c} := \{p \in \Delta_k : p_x \le c \cdot \theta^x/(a+x)^{s+1}\}$ of *enveloped* HLZ distributions. HLZ distributions include the well-known Riemann Zeta, Zipf-Mandelbrot, Lotka, Good, logarithmic-series, and Estoup distributions. These distributions have various applications in many fields. For example, the Good distribution (Zornig & Altmann, 1995) can be used to model species' frequencies and to estimate population parameters.

Note that $\mathcal{H}_k^{\theta,s,a,c} \subseteq \mathcal{P}_k^{s+1,c}$ for $\alpha \ge 1$. Hence, by Corollary 11, for any distribution $p \in \mathcal{H}_k^{\theta,s,a,c}$ and $p' \in \langle p \rangle$,

$$\tilde{r}_n(p', \hat{p}^*) \le \tilde{\mathcal{O}}_{c,s}\left(n^{-\frac{s+1}{s+2}}\right).$$

Let x_1 be the threshold parameter such that $c \cdot \theta^{x_1} = n^{-1}$. Direct computation gives $x_1 = \log(cn)/\log\frac{1}{\theta}$. The symbols $x \in [k]$ that are no larger than x_1 contribute at most x_1 to $\mathbb{E}[D_\Phi]$. Furthermore, the proof of Lemma 4 essentially shows that symbols with probability no larger than n^{-1} contributes at most $\mathcal{O}(\log n)$ to $\mathbb{E}[D_\Phi]$. Therefore, we conclude that $\mathbb{E}[D_\Phi] \le \mathcal{O}(\log(cn)/\log\frac{1}{\theta} + \log n)$.

Corollary 3 combines the above results and yields

Corollary 13. *For any $p \in \mathcal{H}_k^{\theta,s,a,c}$ and $p' \in \langle p \rangle$,*

$$\tilde{r}_n(p', \hat{p}^*) \le \tilde{\mathcal{O}}_{c,s}\left(\frac{n^{\frac{1}{s+2}}}{n} \wedge \frac{1 - \log^{-1}\theta}{n}\right).$$

Note that the right-hand side is the minimum of two quantities. For $\theta \in (1 - n^{-\frac{1}{s+2}}, 1]$, we can reduce the upper bound to $\tilde{\mathcal{O}}_{c,s}(n^{-\frac{s+1}{s+2}})$. On the other hand, for $\theta \in (0, 1 - n^{-\frac{1}{s+2}}]$, the upper bound becomes $\tilde{\mathcal{O}}_{c,s}((1 - \log^{-1}\theta) \cdot n^{-1})$.

4.5. Robustness to Domain Permutations

Our results on learning structured distribution families differ significantly from nearly all the existing ones. Prior work has mainly considered unknown distribution with a certain structure over a known and ordered domain. In our formulation, we assume that the underlying distribution has certain structure under some particular ordering of the domain elements, and this ordering is unknown to the estimator.

Below we illustrate this by a concrete example.

Let $\mathcal{F}$ be a finite discrete domain of size k. Consider learning an unknown log-concave distribution $P \in \Delta_{\mathcal{F}}$ from its sample sequence Y^n. Traditional formulations like (Chan et al., 2013) assume that we know an exact bijective mapping σ from $\mathcal{F}$ to $[k]$, such that reordering the probabilities of P according to σ yields a log-concave distribution $p \in \Delta_k$. Further applying σ to Y^n and denoting the resulting sequence by X^n transforms the problem into learning p from a sample sequence $X^n \sim p$. Here, the assumption that p is log-concave is equivalent to requiring $p_x^2 \ge p_{x-1} \cdot p_{x+1}$, for all $x \in [k] \setminus \{1, k\}$. We can see that such formulation may be non-practical. For example, in natural language processing, the observed samples are words and punctuation marks. Even we know these samples come from a log-concave distribution, we don't know how to order the alphabet, i.e., find the right mapping σ, so that the corresponding distribution $p \in \Delta_k$ would be log-concave.

5. The Estimator

Let p be an arbitrary distribution in Δ_k, and let X^n be a length-n sample sequence from p. For simplicity, abbreviate $\mathbb{1}_x^\mu := \mathbb{1}_{N_x=\mu}$. For any natural number μ, denote the total probability mass of the symbols that appear μ times by

$$M_\mu := \sum_{x \in [k]} p_x \mathbb{1}_x^\mu.$$

After observing X^n, an estimator $\hat{p}$ approximates M_μ by

$$\hat{M}_\mu := \sum_{x : N_x = \mu} \hat{p}_x(X^n).$$

Assume that $\hat{p}$ is a natural estimator. By (Orlitsky & Suresh, 2015), the excess loss of $\hat{p}$ over the best natural estimator that knows the underlying distribution p is

$$\tilde{\ell}_{X^n}(p, \hat{p}) = D(M \parallel \hat{M}) := \sum_{\mu \ge 0} M_\mu \log\frac{M_\mu}{\hat{M}_\mu}.$$

The above characterization of $\tilde{\ell}_{X^n}(p, \hat{p})$ converts the problem of finding good natural estimators for the underlying distribution to that of finding good estimators for

$$M := (M_0, \ldots, M_n).$$

Intuition We first motivate the estimator, whose form is similar to that in (Acharya et al., 2013), but with some modifications. Since the estimator is natural, it needs to approximate only $M := (M_0, \ldots, M_n)$. The construction is guided by analyzing the estimator bias and concentration properties for various multiplicities μ. To estimate M_0, we use the provably near-optimal (Rajaraman et al., 2017) Good-Turing estimator. For the remaining multiplicities, analysis shows that for moderate, yet frequent multiplicities, namely $\mu = \mathcal{O}(\log n)$ and $\Phi_\mu = \Omega(\log^2 n)$, the Good-Turing estimator performs nearly optimally. For infrequent multiplicities, the empirical estimator performs better. For the remaining multiplicities, both estimates are sub-optimal. Applying polynomial approximation techniques, we construct a more involved estimator that approximates the behavior of a genie that knows that expected M_μ values. The estimator is slightly simpler than that in (Acharya et al., 2013), yet achieves better performance.

Details Since our estimator $\hat{p}^*$ is natural, we simply specify $\hat{M}^*_\mu := \sum_{x:N_x=\mu} \hat{p}^*_x(X^n)$. To simplify the analysis, we adopt the standard "Poisson sampling" technique, and make the sample size a Poisson variable N with mean value n.

For $N < n \log n$, let c_1, c_2, and c_3 be properly chosen absolute constants. For any two natural numbers $\mu \geq \mu'$, denote $a^{\mu'}_\mu := \mu'!/\mu!$ and $E^{\mu'}_{x,\mu} := \mathbb{1}^{\mu'}_x a^{\mu'}_\mu (N_x)^{\underline{\mu-\mu'}}$, where $A^{\underline{B}}$ is the falling factorial of A of order B. Let

$$E_{x,\mu} = \frac{1}{c_1\sqrt{\mu/\log n}} \sum_{\mu'=\mu-c_1\sqrt{\mu/\log n}}^{\mu-1} E^{\mu'}_{x,\mu}.$$

We can show that $E_\mu := \sum_{x\in[k]} E_{x,\mu}$ is an unbiased estimator of $\mathbb{E}[\Phi_\mu]$. Empirical-frequency estimates M_μ by

$$\hat{\phi}_\mu := \Phi_\mu \frac{\mu}{n},$$

while Good-Turing estimates it by

$$\hat{G}_\mu := \Phi_{\mu+1} \frac{\mu+1}{n}.$$

To avoid zero probability estimates, slightly modify the Good-Turing estimator to $\hat{G}'_\mu := \max\{1/n, \hat{G}_\mu\}$ and let

$$\hat{O}_\mu := \Phi_\mu \frac{\mu+1}{n} \frac{E_{\mu+1}}{E_\mu},$$

and similarly set

$$\hat{O}'_\mu := \min\{\max\{1/n, \hat{O}_\mu\}, \log^2 n\}.$$

For $\mu < n \log n$, our estimator is

$$\hat{M}^*_\mu = \begin{cases} \hat{G}'_\mu & \text{if } \mu = 0, \\ \hat{\phi}_\mu & \text{if } \mu \geq 1 \text{ and } \Phi_\mu \leq c_2(\log^2 n), \\ \hat{O}'_\mu & \text{if } \mu > c_3 \log n \text{ and } \Phi_\mu > c_2(\log^2 n), \\ \hat{G}'_\mu & \text{if } c_3 \log n \geq \mu \geq 1 \text{ and } \Phi_\mu > c_2(\log^2 n). \end{cases}$$

As Poisson variables are concentrated around their mean, for $N \geq n \log n$, which rarely happens, and $\mu \in [0, N]$, we simply set $\hat{M}^*_\mu = 1/(N+1)$. If these probability estimates do not sum to 1, we normalize them by their sum.

Finally for each $x \in [k]$, our distribution estimator is

$$\hat{p}^*_x(X^n) = \frac{\hat{M}^*_{N_x}}{\Phi_{N_x}}.$$

6. Numerical Experiments

The estimator is easy to implement. In Section 1 of the supplemental material, we present experimental results on a variety of distributions, and show that the proposed estimator indeed outperforms the improved Good-Turing estimator in (Orlitsky & Suresh, 2015).

7. Future Directions

The results obtained in paper strengthen and extend the competitive approach to distribution estimation taken in (Orlitsky & Suresh, 2015). It would be of interest to obtain similar results for distribution estimation under ℓ_1 distance. (Kamath et al., 2015) showed that the simple empirical estimator achieves the min-max ℓ_1-risk $r^{\ell_1}_n(\Delta_k) = (1+o(1))\sqrt{2(k-1)/(\pi n)}$. Yet the excess risk of the estimator in the nice work of (Valiant & Valiant, 2016) is $\mathcal{O}(1/\text{polylog}(n))$. Hence, for $k \leq \tilde{\mathcal{O}}(n)$, this guarantee does not improve that of the empirical estimator, raising the possibility of strengthening the competitive results.

A similar approach can be applied to the related *property-estimation* task. A property, e.g., Shannon entropy, is simply a mapping $f : \Delta_k \to \mathbb{R}$. Most existing property-estimation results are worst-case (min-max) in nature. Yet practical and natural distributions are rarely the worst possible, and often possess a simple structure. To address this discrepancy, recent works (Hao et al., 2018; Hao & Orlitsky, 2019) took a competitive approach, constructing estimators whose performance is adaptive to the simplicity of the underlying distribution. Specifically, the widely-used empirical estimator estimates property values by evaluating the property at the empirical distribution. For every property in a broad class and *every* distribution in Δ_k, the expected error of the estimator in (Hao & Orlitsky, 2019) with sample size $n/\log n$ is at most that of the empirical estimator with sample size n, plus a distribution-free vanishing function of n.

These results cover several well-known properties such as entropy and support size, for which the $\log n$ factor is optimal up to constants, and also apply to any property in the form of $\sum_x f_x(p_x)$, such as the ℓ_1 distance to a given distribution, where f_x is 1-Lipschitz for all $x \in [k]$. It would be of interest to construct a doubly-competitive estimator for property estimation as well.

Acknowledgements

We thank Ananda Theertha Suresh for helpful comments and discussions, and are grateful to the National Science Foundation (NSF) for supporting this work through grants CIF-1564355 and CIF-1619448.

References

Acharya, J., Jafarpour, A., Orlitsky, A., and Suresh, A. T. Optimal probability estimation with applications to prediction and classification. In *Conference on Learning Theory*, pp. 764–796, 2013.

Acharya, J., Diakonikolas, I., Li, J., and Schmidt, L. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1278–1289. Society for Industrial and Applied Mathematics, 2017.

Armaanzas, R., Inza, I., Santana, R., Saeys, Y., Flores, J. L., Lozano, J. A., . . . , and Larraaga, P. A review of estimation of distribution algorithms in bioinformatics. *BioData Mining*, 1(6), 2008.

Ben-Hamou, A., Boucheron, S., and Ohannessian, M. I. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, 23(1):249–287, 2017.

Braess, D. and Sauer, T. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187–206, 2004.

Braess, D., Forster, J., Sauer, T., and Simon, H. U. How to achieve minimax expected kullback-leibler distance from an unknown finite distribution. In *International Conference on Algorithmic Learning Theory*, pp. 380–394, Berlin, Heidelberg, 2002. Springer.

Canonne, C. L., Diakonikolas, I., Gouleakis, T., and Rubinfeld, R. Testing shape restrictions of discrete distributions. *Theory of Computing Systems*, 62(1):4–62, 2018.

Chan, S. O., Diakonikolas, I., Servedio, R. A., and Sun, X. Learning mixtures of structured distributions over discrete domains. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM symposium on Discrete algorithms*, pp. 1380–1394. Society for Industrial and Applied Mathematics, 2013.

Chen, S. F. and Goodman, J. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394, 1999.

Cover, T. Admissibility properties or Gilbert's encoding for unknown source probabilities (corresp.). *IEEE Transactions on Information Theory*, 18(1):216–217, 1972.

Daskalakis, C., Diakonikolas, I., and Servedio, R. A. Learning k-modal distributions via testing. In *Proceedings of the Twenty-Third Annual ACM-SIAM symposium on Discrete Algorithms*, pp. 1371–1385. Society for Industrial and Applied Mathematics, 2012.

Diakonikolas, I., Kane, D. M., and Stewart, A. Optimal learning via the Fourier Transform for sums of independent integer random variables. In *Conference on Learning Theory*, pp. 831–849, 2016.

Falahatgar, M., Ohannessian, M. I., Orlitsky, A., and Pichapati, V. The power of absolute discounting: All-dimensional distribution estimation. In *Advances in Neural Information Processing Systems*, pp. 6660–6669, 2017.

Feldman, J., O'Donnell, R., and Servedio, R. A. Learning mixtures of product distributions over discrete domains. *SIAM Journal on Computing*, 37(5):1536–1564, 2008.

Gupta, P. L., Gupta, R. C., Ong, S. H., and Srivastava, H. M. A class of Hurwitz-Lerch zeta distributions and their applications in reliability. *Applied Mathematics and Computation*, 196(2):521–531, 2008.

Hao, Y. and Orlitsky, A. Data amplification: Instance-optimal property estimation. In *arXiv preprint arXiv:1903.01432.*, 2019.

Hao, Y., Orlitsky, A., Suresh, A. T., and Wu, Y. Data amplification: A unified and competitive approach to property estimation. In *Advances in Neural Information Processing Systems*, pp. 8848–8857, 2018.

Jankowski, H. K. and Wellner, J. A. Estimation of a discrete monotone distribution. *Electronic journal of statistics*, 3: 1567, 2009.

Kamath, G. G. C. On learning and covering structured distributions. *Doctoral dissertation, Massachusetts Institute of Technology*, 2014.

Kamath, S., Orlitsky, A., Pichapati, D., and Suresh, A. T. On learning distributions from their samples. In *Conference on Learning Theory*, pp. 1066–1100, July, 3-6, 2015.

Krichevsky, R. and Trofimov, V. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981.

Ohannessian, M. I. and Dahleh, M. A. Rare probability estimation under regularly varying heavy tails. In *Conference on Learning Theory*, pp. 21–1, 2012.

Orlitsky, A. and Suresh, A. T. Competitive distribution estimation: Why is Good-Turing good. *Advances in Neural Information Processing Systems*, pp. 2143–2151, 2015.

Paninski, L. Variational minimax estimation of discrete distributions under KL loss. In *Advances in Neural Information Processing Systems*, pp. 1033–1040, 2005.

Qu, Y., Beck, G. J., and Williams, G. W. Polya-Eggenberger distribution: Parameter estimation and hypothesis tests. *Biometrical journal*, 32(2):229–242, 1990.

Rajaraman, N., Thangaraj, A., and Suresh, A. T. Minimax risk for missing mass estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 3025–3029. IEEE, 2017.

Valiant, G. and Valiant, P. Instance optimal learning of discrete distributions. In *Proceedings of the forty-eighth Annual ACM symposium on Theory of Computing*, pp. 142–155. ACM, 2016.

Zornig, P. and Altmann, G. Unified representation of Zipf distributions. *Computational Statistics & Data Analysis*, 19(4):461–473, 1995.