Journal of Machine Learning Research 26 (2025) 1-30 Submitted 7/24; Revised 1/25; Published 7/25

Distribution Estimation under the Infinity Norm

Aryeh Kontorovich KARYEH@CS.BGU.AC.IL
Department of Computer Science
Ben Gurion University of the Negev
Beer Sheva, Israel

Amichai Painsky AMICHAIP@TAUEX.TAU.AC.IL
School of Industrial Engineering
Tel Aviv University
Tel Aviv, Israel

Editor: Mladen Kolar

Abstract

We present novel bounds for estimating discrete probability distributions under the ℓ_∞ norm. These are nearly optimal in various precise senses, including a kind of instance-optimality. Our data-dependent convergence guarantees for the maximum likelihood estimator significantly improve upon the currently known results. A variety of techniques are utilized and innovated upon, including Chernoff-type inequalities and empirical Bernstein bounds. We illustrate our results in synthetic and real-world experiments. Finally, we apply our proposed framework to a basic selective inference problem, where we estimate the most frequent probabilities in a sample.

Keywords: Distribution Estimation, Multinomial Distribution, Count Data

1. Introduction

Consider a probability distribution p over $\mathbb{N} = \{1, 2, \ldots\}$. Let X^n be a sample of n independent observations from p. In this work we study the basic problem of estimating p from X^n. We focus our attention to the infinity norm, which is formally defined in (4). Also known as the *uniform* or *supremum* norm, this popular metric over distributions has a number of important applications, in addition to being a fundamental object of independent interest (Boucheron et al., 2003; Van Handel, 2014). Among the applications is a selective inference scheme for multinomial proportions, as discussed below. Our reference point is the following simple and classic bound, whose proof is an easy consequence of McDiarmid's inequality: for all $\delta \in (0, 1)$,

$$\sup_{i \in \mathbb{N}} |p_i - \hat{p}_i(X^n)| \lesssim \sqrt{\frac{\log(1/\delta)}{n}} \tag{1}$$

holds with probability at least $1 - \delta$, where $\hat{p}_i(X^n)$ is the maximum likelihood estimator (MLE) defined below and $\lesssim$ hides small absolute constants. This rate is known to be tight in the worst case (Lemma 23), but can certainly be improved upon for benign distributions. For example, when $p = (1-\theta, \theta)$ is the Bernoulli distribution, Bernstein's inequality (Boucheron

©2025 Aryeh Kontorovich and Amichai Painsky.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v26/24-1166.html.

et al., 2003, Corollary 2.11) yields

$$|\theta - \hat{\theta}| \lesssim \sqrt{\frac{\theta(1-\theta)}{n} \log \frac{1}{\delta}} + \frac{1}{n} \log \frac{1}{\delta}, \tag{2}$$

where $\hat{\theta}$ is the MLE. Furthermore, (2) has an *empirical Bernstein* version (Dasgupta and Hsu, 2008a, Lemma 5), in which the unknown quantity θ is replaced in the right-hand side by the empirically computable $\hat{\theta}$.

Drawing inspiration from (2) and its empirical version, we might expect something like

$$\sup_{i \in \mathbb{N}} |p_i - \hat{p}_i(X^n)| \stackrel{?}{\lesssim} \sqrt{\frac{v^* \log \frac{1}{\delta}}{n}} + \frac{1}{n} \log \frac{1}{\delta}, \tag{3}$$

where $v^* = \max_{i \in \mathbb{N}} p_i(1-p_i)$, or, even more ambitiously, some version of (3) with v^* replaced by its empirical version $\hat{v}^*$. It will turn out that (3) is too optimistic. Absent an oracle that tells us the index of the largest mass, some additional cost must be incurred for estimating many symbol probabilities simultaneously.

Our contributions. Our main result amounts to nearly achieving the ultimate goal. First, we derive a Chernoff-type upper bound in Theorem 1, which improves upon (1). Theorem 2 introduces its data-dependent counterpart, which demonstrates a significant improvement in small sample regimes. Next, we establish in Theorem 4 a version of (3) where v^* is replaced by $v^* \log \frac{1}{v^*}$ and provide even sharper bounds therein. These are matched by nearly optimal lower bounds, in distinct senses made precise below. Finally, we apply our results to the important problem of selective inference. Specifically, we study the basic problem of inferring the most frequent events in a sample and achieve a significant improvement over currently known schemes.

2. Definitions and Problem Statement

Consider a probability distribution p over $\mathbb{N}$, which induces the random variable $X \sim p$. The *support size* of X, $||p||_0$, is also the *alphabet size*, and unless stated otherwise, our results hold even when these are infinite. Let $X^n = (X_1, ..., X_n)$ be a sample consisting of n independent copies of X. Let $c_i(X^n)$ be the count (number of appearances) of the i^{th} symbol in the sample. Let $\hat{p}(X^n)$ be the maximum likelihood estimator (MLE) of p; namely, $\hat{p}_i(X^n) = c_i(X^n)/n$ for every $i \in \mathbb{N}$. In this work, we study empirical distribution estimation of p under the infinity norm. That is, given a prescribed $\delta > 0$, we seek a random variable $T_\delta(X^n)$ such that

$$||p - \hat{p}(X^n)||_\infty \triangleq \sup_{i \in \mathbb{N}} |p_i - \hat{p}_i(X^n)| \leq T_\delta(X^n) \tag{4}$$

with probability of at least $1-\delta$. In light of (1), we also require that, for fixed δ, $T_\delta(X^n) \to 0$ as $n \to \infty$, in some appropriate sense.

Notice that (4) may also be viewed as a multinomial inference scheme. Specifically, (4) implies that $p_i \in [\hat{p}_i \pm T(X^n)]$ simultaneously for all $i \in \mathbb{N}$. This *confidence region* (CR) defines a hypercube around the MLE, which covers p with a *confidence level* of $1 - \delta$.

3. Related Work

Discrete probability estimation is a fundamental problem in many fields. It is extensively studied under a variety of merits such as total variation (Jiao et al., 2017; Cohen et al., 2020), KL divergence (Orlitsky and Suresh, 2015), Hellinger distance (Hellinger, 1909) Wasserstein metric (Kantorovich, 1960), Kolmogorov-Smirnov distance (Smirnov, 1948) and others. The interested reader is referred to (Rice, 2006; Painsky and Wornell, 2019; Painsky, 2023b) for a comprehensive discussion. In this work we focus on the infinity norm. Here, the baseline is the bound implicit in (1), where, in the language of (4), $T_\delta(X^n) = \sqrt{1/n} + \sqrt{\log(1/\delta)/2n}$.

The infinity norm is difficult to analyze in the general case (4). In fact, it is later shown (Section 4) that (1) is only asymptotically tight and only for the worst-case distribution, and can be significantly improved in a limited sample regime. On the other hand, the binomial case, $||p||_0 = 2$, is fairly exhaustively understood as far as minimax optimal and fully empirical bounds.

In particular, if $Y \sim \text{Bin}(n, \theta)$ is a Binomial random variable and $\hat{\theta} = Y/n$ its MLE, then McAllester and Schapire (2000); Bousquet et al. (2003) and later Dasgupta and Hsu (2008b) showed that

$$|\theta - \hat{\theta}| \leq \sqrt{\frac{5\hat{\theta}(1-\hat{\theta})}{n} \log \frac{2}{\delta}} + \frac{5}{n} \log \frac{2}{\delta} \tag{5}$$

with probability of at least $1 - \delta$. A closely related line of work appears in the statistics literature. Let $F(y; n, \theta)$ be the cumulative distribution function of Y. For a given y and n, let θ_l and θ_u be the solutions (with respect to θ) of $F(y; n, \theta) = \delta/2$ and $F(y; n, \theta) = 1 - \delta/2$ respectively. Clopper and Pearson (1934) showed that

$$\mathbb{P}\left(\theta \in [\theta_l, \theta_u]\right) \geq 1 - \delta \tag{6}$$

for every $\theta \in [0, 1]$. The interval $[\theta_l, \theta_u]$ is widely known as the *exact* Clopper-Pearson (CP) confidence interval (CI). The exact notion refers to the fact that (6) holds for every n, as opposed to alternative approximations. In fact, CP is also known to be shortest possible CI, for most setups of interest. Specifically, let $\mathcal{T}$ be a collection of intervals $[t_l, t_u]$ that satisfy $\mathbb{P}\left(\theta \in [t_l, t_u]\right) \geq 1-\delta$, for every $[t_l, t_u] \in \mathcal{T}$. The shortest CI for θ is defined as the intersection of all intervals in $\mathcal{T}$. This notion also implies minimal expected length and minimal false coverage probability, uniformly. Wang (2006) showed that for $n \geq \log(\delta/2)/\log 0.5$, the CP CI is the shortest. Notice that for a nominal level of $\delta = 0.05$, this condition corresponds to $n \geq 6$. Hence, for practical setups of interest, the CP interval $[\theta_l, \theta_u]$ is the shortest possible CI for θ. Unfortunately, CP does not hold a closed-form expression. Yet, Thulin (2014) showed that for every $y \in \{1, ..., n-1\}$,

$$\theta_l = \hat{\theta} - z_{\frac{\delta}{2}} \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} + \frac{1}{3n}\left(\left(1 - 2\hat{\theta}\right) z_{\delta/2}^2 - 1 - \hat{\theta}\right)$$

$$\theta_u = \hat{\theta} + z_{\frac{\delta}{2}} \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} + \frac{1}{3n}\left(\left(1 - 2\hat{\theta}\right) z_{\delta/2}^2 + 2 - \hat{\theta}\right)$$

up to additive terms of order $n^{-3/2}$, where $z_\delta/2$ is the upper $\delta/2$ quantile of the standard normal distribution. This result implies that for every $y \in \{1, ..., n-1\}$, the shortest possible CI length for θ is

$$\theta_u - \theta_l = 2z_{\delta/2}\sqrt{\hat{\theta}(1-\hat{\theta})/n} + 1/n + O(n^{-3/2}). \tag{7}$$

Moreover, we have

$$|\theta - \hat{\theta}| \leq \max\{|\hat{\theta} - \theta_l|, |\hat{\theta} - \theta_u|\} \tag{8}$$

with probability of at least $1-\delta$. Importantly, it can be shown that $z_{\delta/2}$ behaves asymptotically like $\sqrt{2\log(2/\delta)}$. Comparing (1) to (8) (and (5)) , we observe that its sample complexity, $1/\sqrt{n}$, is tight. However, there may still be room for improvement by utilizing a data-dependent scheme.

The Clopper-Pearson interval (8) provides a tight solution for the binomial case $||p||_0 = 2$. Yet, the problem becomes more involved in the multinomial setting (4). Currently known methods focus on two basic regimes. The first considers an asymptotic setup, where n is much greater than the alphabet size (Quesenberry and Hurst, 1964; Goodman et al., 1964; Sison and Glaz, 1995). The second addresses the case where both n and $||p||_0$ are small (Chafai and Concordet, 2009; Malloy et al., 2020). Notice that while some of these methods provide rectangular CR (Quesenberry and Hurst, 1964; Goodman et al., 1964; Painsky, 2023a), others focus on hyper-cubes (Sison and Glaz, 1995). Yet, all of these methods assume a finite alphabet where performance guarantees are limited to relatively small $||p||_0$. To the best of our knowledge, no method considers the case where $||p||_0$ may be infinite.

4. Main Results

We begin our analysis by considering a data-indepedent bound under the infinity norm. Our proposed bound generalizes (1) by utilizing a Chernoff-like concentration bound.

Theorem 1 *Let $p = p_{i\in\mathbb{N}}$ be a distribution over $\mathbb{N}$. Let X^n be a sample of n independent observations from p. Let $\hat{p}(X^n)$ be the MLE of p. Then, with probability at least $1-\delta$,*

$$\|p - \hat{p}\|_\infty = \sup_{i\in\mathbb{N}} |p_i - \hat{p}_i(X^n)| \leq \tag{9}$$

$$\frac{1}{n}\left(\frac{1}{\delta^{1/m}}\right)\left(\sum_{k=1}^{m/2} k^{m-k} n^k \sum_i p_i^k (1-p_i)^k\right)^{1/m} \leq \frac{1}{\sqrt{n}}\frac{\sqrt{m/2}}{\delta^{1/m}}\exp\left(-\frac{1}{2}+\frac{1}{m}\right) + O\left(\frac{1}{n^{\frac{1}{2}+\frac{1}{m}}}\right)$$

for every even $m > 0$.

Theorem 1 relies on Markov's inequality for higher-order moments of the infinity norm. In addition, it applies higher-order properties of the MLE and the binomial distribution. The detailed proof appears Section 7.1. Next, similarly to Chernoff inequality, we minimize (9) with respect to m to obtain tighter convergence guarantees. Specifically, we minimize the leading term of (9) to obtain

$$\min_{m\in\mathbb{R}^+} \frac{\sqrt{m/2}}{\delta^{1/m}}\exp\left(-\frac{1}{2}+\frac{1}{m}\right) = \sqrt{1+\log\left(\frac{1}{\delta}\right)} \tag{10}$$

for the choice of $m^* = 2\log(1/\delta) + 2$. Hence the infimum of the bound is given by

$$\sqrt{\frac{1+\log(1/\delta)}{n}} + O\left(\frac{1}{n^{\frac{1}{2}+\frac{1}{m^*}}}\right). \tag{11}$$

Unfortunately, this is not a great improvement over the benchmark (1). However, it is shown in Section 7.1 that as m increases, the second inequality in (9) becomes tight for a worst-case case distribution $p = [1/2, 1/2, 0, \ldots, 0]$. This distribution is quite "unlikely" in a large alphabet regime. On the other hand, if we assume a "more likely" uniform distribution over a finite alphabet size $A := ||p||_0 < \infty$, we obtain

$$\|p - \hat{p}\|_\infty \leq \frac{1}{\sqrt{n}} \frac{\sqrt{m/2}}{\delta^{1/m}} A^{-\frac{1}{2}+\frac{1}{m}} + O\left(\frac{1}{n^{\frac{1}{2}+\frac{1}{m}}}\right)$$

for every even $m > 0$. Minimizing the leading term with respect to m yields

$$\min_{m\in\mathbb{R}^+} \frac{\sqrt{m/2}}{\delta^{1/m}} A^{-\frac{1}{2}+\frac{1}{m}} = \sqrt{\log\left(\frac{A}{\delta}\right)} \exp\left(\frac{1}{2} - \frac{1}{2}\log(A)\right), \tag{12}$$

for a choice of $m^* = 2\log(A/\delta)$. Notice the above vanishes with A. This result motivates our quest for a data-dependent bound, which considers an empirical estimate of p and does not assume a worst-case distribution as in (1) and (11). Theorem 2 below improves upon (9) and introduces a data-dependent bound which further accounts for $\hat{p}$.

Theorem 2 *Let $\delta_1 > 0$ and $\delta_2 > 0$. Let m be a positive even number. Then, with probability at least $1 - \delta_1 - \delta_2$,*

$$\|p - \hat{p}\|_\infty \leq \frac{1}{n}\left(\frac{1}{\delta_1}\left(\sum_i \sum_{k=1}^{m/2} k^{m-k}(n+1)^k(\hat{p}_i(1-\hat{p}_i))^k + \epsilon_{n+1}\right)\right)^{1/m} \tag{13}$$

for every even m, where

$$\epsilon_n = \sqrt{2n\log(1/\delta_2)} \cdot \sum_{k=1}^{m/2} k^{m-k} n^k \left(\frac{k}{n\cdot 4^{k-1}} + \frac{k(k-1)}{n^2\cdot 2^{2k-3}}\right).$$

To prove Theorem 2 we utilize the first inequality of (9) with $\delta = \delta_1$. Then, we apply McDiarmid's inequality to obtain a concentration bound for $\sum_{i\in\mathbb{N}} p_i^k(1-p_i)^k$ around its empirical counterpart, with probability $1 - \delta_2$. Finally, we apply the union bound to obtain (13). The detailed proof is provided in Section 7.2. To further clarify the proposed bound we introduce the following simplified corollary, whose proof is located in Section 7.3

Corollary 3 *Let $\delta_1 > 0$ and $\delta_2 > 0$. Let m be a positive even number. Then, with probability at least $1 - \delta_1 - \delta_2$,*

$$\|p - \hat{p}\|_\infty \leq$$
$$\frac{m}{\sqrt{2}\delta_1^{1/m}} \frac{1}{n}\left(\sum_{k=1}^{m/2}\sum_i (n\hat{p}_i(1-\hat{p}_i))^k\right)^{1/m} + a\frac{m}{\delta_1^{1/m}}\left(\log\left(\frac{1}{\delta_2}\right)\right)^{1/2m}\left(\frac{1}{n^{\frac{1}{2}\left(1+\frac{1}{m}\right)}} + \frac{1}{n^{\frac{1}{2}\left(1+\frac{3}{m}\right)}}\right)$$

for every even m, where $a = \sqrt{\exp(1/\mathrm{e})}$. Furthermore,

$$\inf_m \frac{m}{\delta_1^{1/m}} = \mathrm{e}\log(1/\delta_1),$$

where the infimum is obtained for a choice of $m^* = \log(1/\delta_1)$.

Let us compare Corollary 3 with the benchmark scheme (1) and our previous data independent bound (Theorem 1). First, we notice a similar sample complexity of order of $\sqrt{1/n}$ in all the three schemes. This is not entirely surprising, given (5). However, Corollary 3 demonstrates an improved dependency on the underlying distribution, which now depends on $\hat{p}$ and does not assume a worst-case distribution. Unfortunately, the dependency in $\hat{p}$ is somewhat involved, and does not hold the desired form of (3). Finally, we compare the dependency in the confidence level δ. Here, the data independent bounds introduce a squared root logarithmic dependency in $1/\delta$. On the other hand, Corollary 3 only attains a logarithmic dependency in $1/\delta_1$ (where $\delta_1 < \delta$) for the right choice of m. This difference is typically negligible compared to the other terms, especially in a fixed δ regime as later discussed.

To conclude, Theorem 2 and Corollary 3 introduce data-dependent upper bounds to the infinity norm. The proposed bounds generalize (1), as they both utilizes high-moments and, more importantly, do not consider a worst-case distribution. In that sense, the proposed results generalize and improve upon (1), which may be considered as their special case. Our experiments in Section 5 numerically demonstrate this improvement. Despite the above, we would still like to find more intuitive upper bounds, which are less complex and also introduce a dependency in p that is closer to the desired form (3). Hence, we present our second main result. First, we introduce some additional notation.

Notation. For any distribution $p_{i\in\mathbb{N}}$, define $v = v(p)$ by $v_i = p_i(1-p_i)$ and $v^* = \max_{i\in\mathbb{N}} v_i$ as above. Define the functional

$$V^*(p) = \sup_{i\in\mathbb{N}} v_i(p^{\downarrow})\log(i+1), \tag{14}$$

where $p^{\downarrow}$ is p sorted in non-increasing order. Define

$$\varphi(t) = t\log\frac{1}{t}, \qquad 0 \le t \le 1. \tag{15}$$

Theorem 4 *Let $p = p_{i\in\mathbb{N}}$ be a distribution over $\mathbb{N}$ and put $v^* = v^*(p)$, $V^* = V^*(p)$. For $n \ge 81$ and $\delta \in (0,1)$, we have that*

$$\|p-\hat{p}\|_\infty \le 2\sqrt{\frac{V^*}{n} + \frac{v^*}{n}\log\frac{2}{\delta}} + \frac{4}{3n}\log\frac{2(n+1)}{\delta} + \frac{\log n}{n} \tag{16}$$

$$\le 2\sqrt{\frac{\varphi(v^*)}{n} + \frac{v^*}{n}\log\frac{2}{\delta}} + \frac{4}{3n}\log\frac{2(n+1)}{\delta} + \frac{\log n}{n}; \tag{17}$$

$$\|p-\hat{p}\|_\infty \le 2\sqrt{\frac{v^*\log(n+1)}{n} + \frac{v^*}{n}\log\frac{2}{\delta}} + \frac{4}{3n}\log\frac{2(n+1)}{\delta} + \frac{\log n}{n} \tag{18}$$

holds with probability at least $1-\delta-81/n$.

Remark 5 *It seems that the* $\log(n)/n$ *term can be improved to* $\log(n)/(n \log\log\log n)$*; we shall explore this in a future work.*

It is instructive to compare Theorem 4 with our ambitious "dream" (3). The loosest bound therein, (18), features the desired dependency on v^*, but at the cost of a $\log n$ factor. The sharper bound (17) replaces the $\log n$ with $v^* \log \frac{1}{v^*}$. Finally, (16) gives the optimal (at least for the MLE, cf. Proposition 9) quantity V^*. The proof of Theorem 4 is located in Section 7.4. It relies on techniques from empirical process theory and large deviations.

Next, we provide the empirical counterpart of Theorem 4, which depends on $\hat{v}^* = \sup_{i\in\mathbb{N}} \hat{p}_i(1-\hat{p}_i)$.

Theorem 6 *Let* $p = p_{i\in\mathbb{N}}$ *be a distribution over* $\mathbb{N}$*. Let* $\hat{p}$ *be the MLE of* p*. Define*

$$a = \frac{4}{3n}\log\frac{2(n+1)}{\delta} + \frac{\log n}{n}, \qquad b = 2\sqrt{\frac{\log(n+1)}{n} + \frac{1}{n}\log\frac{2}{\delta}}.$$

Then, with probability at least $1 - \delta - 81/n$*,*

$$\|p - \hat{p}\|_\infty \le 3a/2 + \sqrt{a}b/2 + 5b^2/4 + 7b\sqrt{\hat{v}^*}/4. \tag{19}$$

Remark 7 *Note that the estimate in* (19) *is of order* $\sqrt{\frac{\hat{v}^* \log n}{n} + \frac{\hat{v}^*}{n}\log\frac{1}{\delta}} + \frac{1}{n}\log\frac{n}{\delta} + \frac{\log n}{n}$ *, matching, up to constants, the form of* (18)*.*

The proof of Theorem 6 follows the proof of Dasgupta and Hsu (2008a, Lemma 5) and is left for Section 7.5.

Open problem. The estimate in Theorem 6 gives an empirical analog of (18). We conjecture that some empirical analog of (16) should be possible as well: a bound of the general form $\sqrt{\frac{\hat{V}^*}{n} + \frac{\hat{v}^*}{n}\log\frac{1}{\delta}} + \frac{1}{n}\log\frac{n}{\delta} + \frac{\log n}{n}$.

Near-optimality. To argue the near-optimality[1] of the above bounds we introduce our lower bounds on $\sup_{i\in\mathbb{N}} |\hat{p}_i - p_i|$ for some fixed constant $\delta > 0$; this is equivalent to lower bounding $\mathbb{E}\sup_{i\in\mathbb{N}} |\hat{p}_i - p_i|$. Understanding the correct dependence on δ is left for future work.

For a fixed δ, the upper bound in (16) consists of two terms: one of order $\sqrt{V^*/n}$ and another one of order $\log(n)/n$. We shall argue below that the first is tight and the second nearly so, albeit in different senses.

The near-optimality of the $\log(n)/n$ term is proved in the following result, whose proof is provided in Section 7.6. Note that the lower bound obtained for this term is of a minimax type, meaning that it holds for *any* estimator, not just the MLE.

Proposition 8 *There is an absolute constant* $c > 0$ *such that the following holds for all sufficiently large* n*. For any estimator* $\tilde{p}(X^n)$*, there is a distribution* $p_{i\in\mathbb{N}}$ *on* $\mathbb{N}$ *such that*

$$\mathbb{E}\,\|p - \tilde{p}\|_\infty \;\ge\; \frac{c\log n}{n\log\log n} \tag{20}$$

for all sufficiently large n*.*

1. We use the term "instance-optimality" in the spirit of Theorems 2.3 and 2.4 of Cohen et al. (2020): fully empirical data-dependent bounds that cannot be significantly improved upon.

Our lower bound matching the $\sqrt{V^*/n}$ term will be limited to the MLE, but will have the advantage of holding pointwise for *any* given distribution, in constradistinction to the minimax bound in Proposition 8, which only holds for *some* adversarial distribution. The proof of Proposition 9 is provided in Section 7.7.

Proposition 9 *For any distribution $p_{i\in\mathbb{N}}$ and its corresponding MLE $\hat{p}$, we have*

$$\liminf_{n\to\infty} \sqrt{n}\mathbb{E}\left\|p-\hat{p}\right\|_\infty \geq c\sqrt{V^*(p)},$$

where $c>0$ is an absolute constant.

Remark 10 *We show in Section 7.7 that the lower bound is necessarily only asymptotic (rather than* finite-sample*, in the sense of holding for all n), as a consequence of previous results (Berend and Kontorovich, 2013)*

Finally, the following is a straightforward consequence of the Neyman-Pearson (Lemma 23):

Proposition 11 *For any estimator $\tilde{p}$ there exists a distribution $p_{i\in\mathbb{N}}$ such that*

$$\mathbb{E}\left\|p-\tilde{p}\right\|_\infty \geq c\sqrt{\frac{v^*(p)}{n}},$$

for all sufficiently large n, where $c>0$ is an absolute constant.

The proof of Proposition 11 is provided in Section 7.8. It is instructive to compare Propositions 9 and 11. The former holds for any fixed distribution p and the bound is stronger (since $V^* \geq v^*$), but only for the MLE estimate. The latter holds for all estimators, but the distribution can be aversarially chosen for each sample size n and the bound is weaker. We conjecture that the lower bound in Proposition 9 holds for all estimators and not just the MLE.

5. Experiments

Let us now demonstrate our proposed bounds. We focus on two benchmark distributions which represent two extreme cases. That is, we study the Zipf's law and the uniform distributions. The Zipf's law distribution is a typical benchmark in large alphabet probability estimation; it is a commonly used heavy-tailed distribution, mostly for modeling natural (real-world) quantities in physical and social sciences, linguistics, economics and others fields (Saichev et al., 2009). The Zipf's law distribution follows $p_i = i^{-s}/\sum_{r=1}^{A} r^{-s}$ where A is the alphabet size and s is a skewness parameter. We set $s=1.1$ throughout our experiments. In each experiment we draw n samples from a distribution over an alphabet size A to evaluate the proposed bounds for a given confidence level $1-\delta$. We repeat this process 10^4 times and report the average bound and coverage rate (that is, the number of times that the infinity norm is not greater than the bound).

In the first experiment we focus on $A=100$ and $\delta=0.05$. We examine three bounds. First, we consider the bound from Theorem 2 with $\delta_1 = 0.99\delta$, $\delta_2 = 0.01\delta$. We set m to minimize (13) over the worst-case distribution (see (10)). This results in $m* = \log(1/\delta_1) + 2 \approx 8$. Further, we examine Theorem 6 and the benchmark bound (1). To further assess

the tightness of our results we introduce an Oracle lower bound (OLB). The OLB knows the true distribution and evaluates the $1-\delta$ quantile of the desired infinity norm. Figure 1 summarizes the results we achieve. First, we observe that Theorem 2 outperforms both Theorem 6 and the benchmark. It is also relatively close to the OLB, especially as n increases. We emphasize that although Theorem 6 demonstrates a steep descent, it does not outperform Theorem 2, even for a relatively large $n = 10^5$. The reason for this phenomenon is the fixed δ regime, in which Theorem 2 is favorable. Importantly, all the examined bounds attain the prescribed coverage rate as desired.

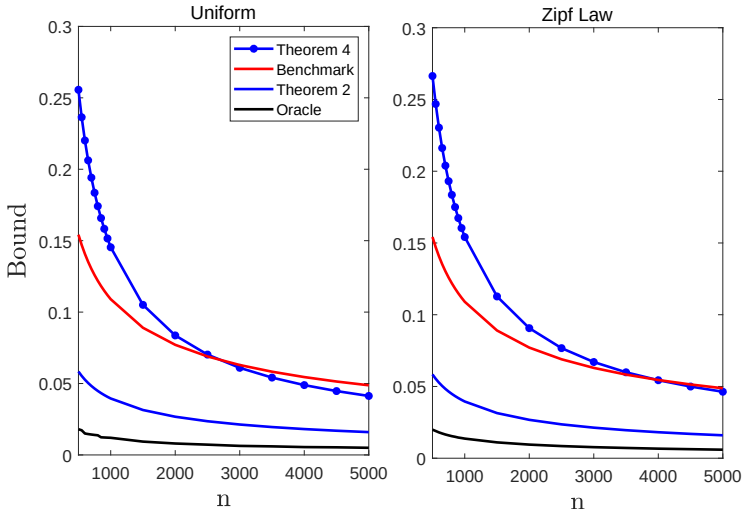

Figure 1: The proposed bounds compared to the benchmark and to an Oracle, as n grows and $\delta = 0.05$

Next, we examine the performance of our proposed schemes for a decaying confidence level, $\delta = 1/n^2$. As above, we set $A = 100$ and focus on the two benchmark distributions. Figure 2 demonstrates the results we achieve. Here, we see the advantage of Theorem 6, as it outperforms the alternatives for relatively large n. Once again, all bounds attain the prescribed confidence level.

To conclude, both Theorem 2 and Theorem 6 improve upon the benchmark (1) in the studied regimes. While Theorem 2 shows favorable performance in the fixed $\delta regime$, Theorem 6 demonstrates improved results in a decaying δ regime. It is also important to emphasize that while Theorem 6 is more intuitive and easier to apply, Theorem 2 is more complex and designed to attain the tightest results, even for smaller n. Notice that Theorems 1 and 4 are omitted from our experiments, as they require the unknown underlying distribution pp.

6. Application to Inference of Frequent Events

We now introduce an important application of our proposed scheme. Consider a survey asking individuals for their favorite food. We would like to report the k most popular foods along with their associated CIs. The common approach is to construct k marginal (binomial)

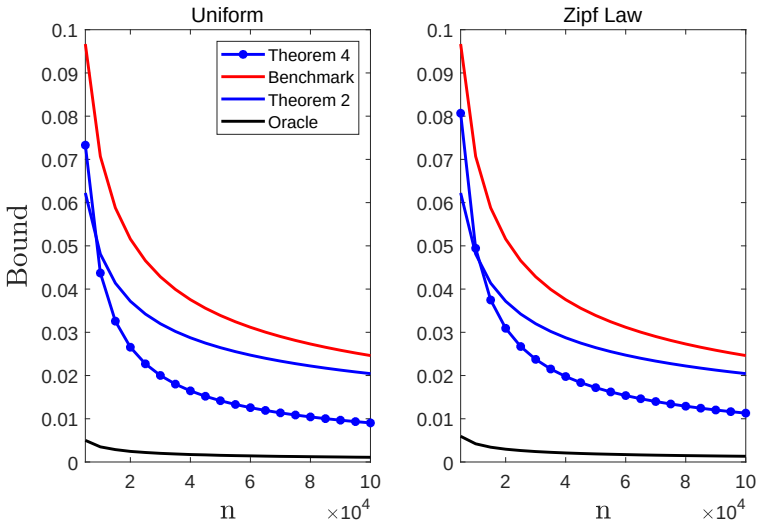

Figure 2: The proposed bounds compared to the benchmark and to an Oracle, as n grows and $\delta = 1/n^2$

intervals of confidence level $1-\delta$ each. This approach is genuinely wrong. For example, consider the case of $k = 1$, $n = 100$ and a uniform distribution over an alphabet size $A = ||p||_0$. By definition, the most popular food in the sample would attain at least a single vote. Therefore, its exact lower bound CI (of level $1-\delta = 0.95$) is at least $\theta_l = 2.5 \cdot 10^{-4}$. This means that for $A > 4000$, we attain zero coverage rate (!). This phenomenon is not all that surprising. Traditional (frequentist) inference assumes a fixed and unknown parameter θ. Here, the inferred parameter is data-dependent, as it corresponds to the most frequent symbols in the sample. That is, we may obtain different k most popular foods for different samples. This type of inference problem is known as *selective inference* (Ben-Hamou et al., 2017). Selective inference is a complicated task which is extensively studied in recent years (Tibshirani et al., 2016; Berk et al., 2013). One of the first major contributions to the problem is due to Benjamini and Yekutieli (2005). In their work, they showed that conditional coverage, following any selection rule for any set of (unknown) values for the parameters, is impossible to achieve. This means we cannot simply infer on the chosen parameters, given that they were selected.

Naturally, by controlling the infinity norm we implicitly control the k most frequent events. That is, assume we found $T_\delta(X^n)$ that satisfies (4). Then, we have $|p_i - \hat{p}_i(X^n)| \leq T_\delta(X^n)$ for every $i \in \mathbb{N}$ with probability $1-\delta$, including the k most frequent events. However, it is of a natural concern that such an approach is not tight enough, as it is oblivious to k. In the following we study this claim and discuss the tightness of the infinity norm with respect to the k most frequent events.

Theorem 12 *Let $p = p_{i\in\mathbb{N}}$ be a distribution over $\mathbb{N}$. Let $\hat{p}$ be the MLE of p. Let $j = \text{argmax}\ \hat{p}_i$ be the most frequent symbol in the sample so that $\hat{p}_j = \max_i \hat{p}_i$. Assume there*

exists $U_\delta(X^n)$ *such that*

$$\mathbb{P}\left(|p_j - \hat{p}_j| \geq U_\delta(X^n)\right) \leq \delta. \tag{21}$$

Then,

$$\mathbb{E}(U_\delta(X^n)) \geq z_{\delta/2}\sqrt{\frac{p_{[1]}(1-p_{[1]})}{n}} + O\left(\frac{1}{n}\right) \tag{22}$$

for sufficiently large n*, where* $p_{[1]} = max_i p_i$*.*

The proof of Theorem 12 is provided in Section 7.9. It utilizes the optimally of the CP CI and additional asymptotic properties. Let us now consider the k most frequent events. We would like to refrain from multiplicity corrections, so we seek an interval for $\sup_{i\in\mathcal{X}_k(X^n)} |p_i - \hat{p}_i(X^n)|$ where $\mathcal{X}_k(X^n)$ is the collection of the k most frequent events in the sample. This set naturally contains the single most frequent event, so a CI of average length (22) is inevitable.

Let us compare Theorem 12 with our proposed bounds. For this purpose we turn to real-world data sets. Notice that in the real-world settings, the true underlying probability is unknown. Hence, we treat the empirical distribution of the full data-set as the underlying distribution and sample from it accordingly. We begin with a census data; we consider the 2000 United States Census (Bureau, 2014), which lists the frequency of the top 1000 most common last names in the United States. We randomly sample n names (with replacement) and examine the studied bounds for $\delta = 0.05$. In addition, we present the Oracle CIs for the single most frequent symbol and the infinity norm. The left chart of Figure 3 demonstrates the results we achieve. As we can see, the Oracle CIs are very close to each other and the difference between them and Theorem 12 is also negligible. This shows that the infinity norm is a very good proxy to the k most frequent symbols in the alphabet. As we further examine our results, we see that for a typical experiment of $n = 10000$, the top $k = 5$ surnames are Smith, Johnson, Williams, Brown and Jones with $\hat{p}_i = [0.0213, 0.0165, 0.0137, 0.0134, 0.0130]$ respectively. Theorem 2 attains a bound of 0.0108 while the benchmark is about three times greater, 0.0345. Next, we consider a corpus linguistic experiment. The popular Broadway play *Hamilton* consists of 20,520 words, of which 3,578 are distinct. We randomly sample n words (with replacement), and evaluate the corresponding bounds. The right chart of Figure 3 demonstrates the results we achieve. Once again,it is quite evident that Theorem 2 outperforms its alternatives in this fixed $\delta = 0.05$ regime. Further, we observe that the the infinity norm is a tight proxy to the k most frequent symbols in the alphabet.

7. Proofs

7.1 Proof of Theorem 1

First, notice we have

$$\mathbb{E}\big(\sup_i |p_i - \hat{p}_i(X^n)|\big)^m \overset{\text{(i)}}{=} \mathbb{E}\left(\sup_i (p_i - \hat{p}_i(X^n))^m\right) \overset{\text{(ii)}}{\leq} \mathbb{E}\left(\sum_i (p_i - \hat{p}_i(X^n))^m\right) = \tag{23}$$

$$\frac{1}{n^m}\sum_i \mathbb{E}(n_i - np_i)^m \overset{\text{(iii)}}{\leq} \frac{1}{n^m}\sum_i \sum_{k=1}^{d} k^{m-k}(np_i(1-p_i))^k$$

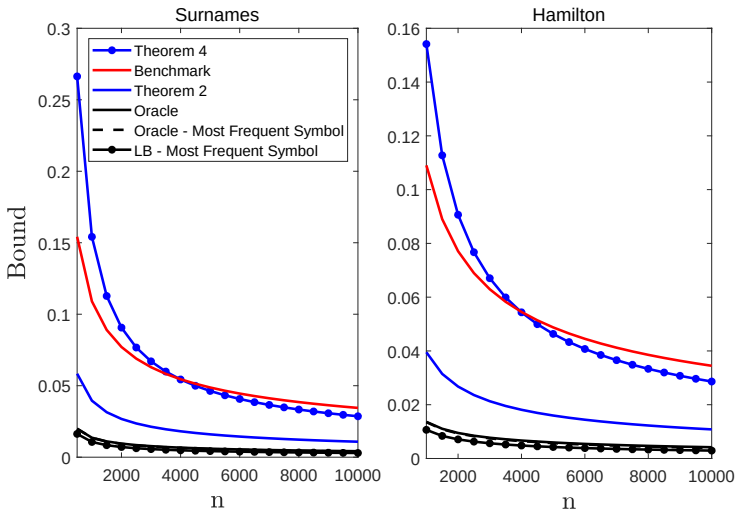

Figure 3: The proposed bounds compared to the benchmark and to an Oracle, as n grows. The lower bound for the most frequent symbol corresponds to Theorem 12

where $d = m/2$ and

(i) follows from the monotonicity of the power function.

(ii) The supremum of non-negative elements is bounded from above by their sum (Maddox, 1988).

(iii) follows from Theorem 4 of (Skorski, 2020)

Applying Markov's inequality we obtain

$$\mathbb{P}\Big(\sup_i |p_i - \hat{p}_i(X^n)| \geq a\Big) \leq \frac{1}{a^m}\mathbb{E}\left(\sup_i |p_i - \hat{p}_i(X^n)|\right)^m \leq \frac{1}{a^m}\frac{1}{n^m}\sum_i \sum_{k=1}^{d} k^{m-k}(np_i(1-p_i))^k. \tag{24}$$

Setting the right hand side to equal δ yields

$$a = \frac{1}{n}\left(\frac{1}{\delta_1}\sum_i \sum_{k=1}^{d} k^{m-k}(np_i(1-p_i))^k\right)^{1/m}.$$

Therefore, with probability $1-\delta$, we have

$$\sup_i |p_i - \hat{p}_i(X^n)| \leq \frac{1}{n}\left(\frac{1}{\delta}\sum_{k=1}^{d} k^{m-k}n^k \sum_i p_i^k(1-p_i)^k\right)^{1/m}. \tag{25}$$

Let us further bound from above the right hand side of (25). We have,

$$\sum_i p_i^k(1-p_i)^k \overset{\text{(i)}}{\leq} \max_{t\in[0,1]} t^{k-1}(1-t)^k \overset{\text{(ii)}}{\leq} \left(\frac{k-1}{2k-1}\right)^{k-1}\left(\frac{k}{2k-1}\right)^{k} \overset{\text{(iii)}}{\leq} \left(\frac{k}{2k-1}\right)^{2k-1} = \left(1+\frac{-k+1}{2k-1}\right)^{2k-1} \overset{\text{(iv)}}{\leq} \exp(-k+1) \tag{26}$$

(i) follows from $\sum_i p_i\psi(p_i) \leq \max_{t\in[0,1]}\psi(t)$ for $\psi(p_i) = p_i^{k-1}(1-p_i)^k$. That is, the mean of a random variable not greater than its maximum.

(ii) simple derivation shows that the maximum of $t^{k-1}(1-t)^k$ is attained for $t^* = \frac{k-1}{2k-1}$.

(iii) is due to $\left(\frac{k-1}{2k-1}\right)^{k-1} \leq \left(\frac{k}{2k-1}\right)^{k-1}$.

(iv) follows from Bernoulli inequality.

Importantly, notice that for a choice of $p = [1/2, 1/2, 0, \ldots, 0]$ we have

$$\sum_i p_i^k(1-p_i)^k = 2\left(\frac{1}{2}\right)^{2k} = \left(\frac{1}{2}\right)^{2k-1}, \tag{27}$$

which approaches the term on the right hand side of inequality (iii), as k increases. Finally, plugging (26) to (25) we obtain

$$\sup_i |p_i - \hat{p}_i(X^n)| \leq \frac{1}{n}\left(\frac{1}{\delta}\sum_{k=1}^{d} k^{m-k}n^k\exp(-k+1)\right)^{1/m} = \frac{1}{\sqrt{n}}\left(\frac{\sqrt{m/2}}{\delta^{1/m}}\right)\exp\left(-\frac{1}{2}+\frac{1}{m}\right) + O\left(\frac{1}{n^{\frac{1}{2}+\frac{1}{m}}}\right) \tag{28}$$

for every even $m > 0$.

7.2 Proof of Theorem 2

We begin with the following proposition.

Proposition 13 *Let $\delta_2 > 0$. Then, with probability $1-\delta_2$,*

$$\sum_i\sum_{k=1}^{m/2} k^{m-k}(n(p_i(1-p_i))^k \leq \left(\sum_i\sum_{k=1}^{m/2} k^{m-k}((n+1)\hat{p}_i(1-\hat{p}_i))^k + \epsilon_{n+1}\right) \tag{29}$$

for every even m, where

$$\epsilon_n = \sqrt{2n\log(1/\delta_2)}\sum_{k=1}^{d} k^{m-k}n^k\left(\frac{k}{n\cdot 4^{k-1}} + \frac{k(k-1)}{n^2\cdot 2^{2k-3}}\right) \tag{30}$$

.

Proof Define $\psi(n,d,\hat{p}) = \sum_i \sum_{k=1}^{d} k^{m-k}(n\hat{p}_i(1-\hat{p}_i))^k$. McDiarmid's inequality yields

$$\mathrm{P}\left(\psi(n,d,\hat{p}) - \mathbb{E}\left(\psi(n,d,\hat{p})\right) \leq -\epsilon_n\right) \leq \exp\left(\frac{-2\epsilon_n^2}{\sum_{j=1}^{n} c_j^2}\right)$$

where c_j is defined as

$$\sup_{x_j'} \left|\psi(n,d,\hat{p}) - \psi(n,d,\hat{p}')\right| \leq c_j. \tag{31}$$

Here, $\hat{p}'$ is the MLE over the same sample x^n, but with a different j^{th} observation, x_j'. First, let us find c_j. We have

$$\begin{aligned}
&\sup_{x_j'} \left|\psi(n,d,\hat{p}) - \psi(n,d,\hat{p}')\right| \leq \qquad (32)\\
&\sup_{p\in[0,1-1/n]} 2\Big|\sum_{k=1}^{d} k^{m-k}\left(np(1-p)\right)^k - \sum_{k=1}^{d} k^{m-k}\left(n(p+1/n)(1-(p+1/n))\right)^k)\Big| =\\
&\sup_{p\in[0,1-1/n]} 2\Big|\sum_{k=1}^{d} k^{m-k}n^k\left(p(1-p)\right)^k - \left((p+1/n)(1-(p+1/n))\right)^k)\Big|,
\end{aligned}$$

independently of j, where the inequality follows from the fact that changing a single observation effects only two symbols (for example, $\hat{p}_l$ and $\hat{p}_t$), where the change is $\pm 1/n$.

Now, we would like to bound from above (32). Denote $f_k(p) = p^k(1-p)^k$. Applying Taylor series to $f_k(p+1/n)$ around $f_k(p)$ yields

$$f_k\left(p+\frac{1}{n}\right) = f_k(p) + \frac{1}{n}f_k'(p) + r(p)$$

where $r(p) = \frac{1}{2!}\frac{1}{n^2}f''(c)$ is the residual and $c \in [p, p+1/n]$ (Stromberg, 2015). We have

$$\begin{aligned}
&f_k'(p) = k\left(p(1-p)\right)^{k-1}(1-2p) \leq k\left(p(1-p)\right)^{k-1} \qquad (33)\\
&f_k''(p) = k(k-1)(p(1-p))^{k-2}(1-2p)^2 - 2k(p(1-p))^{k-1} \leq k(k-1)(p(1-p))^{k-2}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sup_{p\in[0,1-1/n]} \left|\left(p(1-p)\right)^k - \left((p+1/n)(1-(p+1/n))\right)^k)\right| = \qquad (34)\\
&\sup_{p\in[0,1-1/n]} \left|-\frac{1}{n}f_k'(p) - \frac{1}{2!}\frac{1}{n^2}f''(c)\right| \leq \sup_{p\in[0,1-1/n]} \frac{1}{n}\left|f_k'(p)\right| + \frac{1}{2!}\frac{1}{n^2}\left|f''(c)\right| \overset{\text{(i)}}{\leq}\\
&\sup_{p\in[0,1-1/n]} \frac{k}{n}\left(p(1-p)\right)^{k-1} + \frac{k(k-1)}{2n^2}(p(1-p))^{k-2} \overset{\text{(ii)}}{\leq} \frac{k}{n\cdot 4^{k-1}} + \frac{k(k-1)}{n^2\cdot 2^{2k-3}}
\end{aligned}$$

where

(i) follows from (33).

(ii) follows from the concavity of $(p(1-p))^k$ for $k \geq 1$.

Therefore, we have

$$\sup_{x'_j} \left|\psi(n,d,\hat{p}) - \psi(n,d,\hat{p}')\right| \leq 2\sum_{k=1}^{d} k^{m-k} n^k \left(\frac{k}{n \cdot 4^{k-1}} + \frac{k(k-1)}{n^2 \cdot 2^{2k-3}}\right). \tag{35}$$

Next,

$$\begin{aligned}\mathbb{E}(\psi(n,d,\hat{p})) \geq & \sum_i \sum_{k=1}^{d} k^{m-k} n^k \left(\mathbb{E}(\hat{p}_i(1-\hat{p}_i))\right)^k = \\ & \sum_i \sum_{k=1}^{d} k^{m-k} n^k \left(\left(1 - \frac{1}{n}\right) p_i(1-p_i)\right)^k = \\ & \sum_i \sum_{k=1}^{d} k^{m-k} ((n-1)p_i(1-p_i))^k = \psi(n-1,d,p)\end{aligned} \tag{36}$$

where the first inequality follows from Jensen Inequality and the equality that follows is due to $\mathbb{E}(\hat{p}_i(1-\hat{p}_i)) = \mathbb{E}(\hat{p}_i) - \mathrm{Var}(\hat{p}_i) - \mathbb{E}^2(\hat{p}_i) = p(1-p)(1-1/n)$. Going back to McDiarmid's inequality, we have

$$\mathbb{P}\left(\mathbb{E}\psi(n,d,\hat{p}) \geq \psi(n,d,\hat{p}) + \epsilon_n\right) \leq \exp\left(\frac{-2\epsilon_n^2}{nc_j^2}\right) \tag{37}$$

In word, the probability that the random variable $Z = \psi(n,d,\hat{p})$ is smaller than a constant $C = \mathbb{E}(\psi(n,d,\hat{p})) - \epsilon_n$ is not greater that $\nu = \exp\left(-2\epsilon^2 / \sum_{j=1}^{n} c_j^2\right)$. Therefore, it necessarily means that the probability that Z is smaller than a constant smaller than C, is also not greater than ν. Hence, plugging (36) we obtain

$$\mathbb{P}\left(\psi(n-1,d,p) \geq \psi(n,d,\hat{p}) + \epsilon_n\right) \leq \exp\left(\frac{-2\epsilon_n^2}{\sum_j c_j^2}\right)$$

Setting the right hand side to equal δ_2 we get

$$\epsilon_n = \sqrt{2n\log(1/\delta_2)} \sum_{k=1}^{d} k^{m-k} n^k \left(\frac{k}{n \cdot 4^{k-1}} + \frac{k(k-1)}{n^2 \cdot 2^{2k-3}}\right) \tag{38}$$

and with probability $1 - \delta_2$,

$$\sum_i \sum_{k=1}^{d} k^{m-k} ((n-1)p_i(1-p_i))^k \leq \left(\sum_i \sum_{k=1}^{d} k^{m-k} (n\hat{p}_i(1-\hat{p}_i))^k + \epsilon_n\right). \tag{39}$$

Plugging $n+1$ to the above yields the desired result. ■

Finally, we apply the union bound to (25) with $\delta = \delta_1$ and Proposition 13 to obtain the stated Theorem 1.

7.3 A Proof for Corollary 3

We prove the Corollary with two propositions.

Proposition 14 *Let $\delta_1 > 0$. Then, with probability $1-\delta_1$,*

$$\sup_i |p_i - \hat{p}_i(X^n)| \leq \frac{m}{2n}\left(\frac{1}{\delta_1}\right)^{1/m}\left(\sum_i \sum_{k=1}^{m/2}(np_i(1-p_i))^k\right)^{1/m} \tag{40}$$

for every even $m > 0$.

Proof First, we have

$$\mathbb{E}\big(\sup_i |p_i - \hat{p}_i(X^n)|\big)^m \overset{\text{(i)}}{\leq} \frac{1}{n^m}\sum_i \sum_{k=1}^{d} k^{m-k}(np_i(1-p_i))^k \overset{\text{(ii)}}{\leq} \left(\frac{d}{n}\right)^m \sum_i \sum_{k=1}^{d}(np_i(1-p_i))^k$$

where $d = m/2$ and

(i) follows from (23).

(ii) follows from $k^{m-k} \leq d^m$ for every $k \in \{1, ..., d\}$.

Applying Markov's inequality we obtain

$$\begin{aligned}\mathbb{P}\left(\sup_i |p_i - \hat{p}_i(X^n)| \geq a\right) \leq & \frac{1}{a^m}\mathbb{E}\left(\sup_i |p_i - \hat{p}_i(X^n)|\right)^m \leq \\ & \frac{1}{a^m}\left(\frac{d}{n}\right)^m \sum_i \sum_{k=1}^{d}(np_i(1-p_i))^k.\end{aligned} \tag{41}$$

Setting the right hand side to equal δ_1 yields

$$a = \left(\frac{1}{\delta_1}\left(\frac{d}{n}\right)^m \sum_i \sum_{k=1}^{d}(np_i(1-p_i))^k\right)^{1/m} = \frac{m}{2n}\left(\frac{1}{\delta_1}\sum_i \sum_{k=1}^{m/2}(np_i(1-p_i))^k\right)^{1/m}.$$

■

Proposition 15 *Let $\delta_2 > 0$. Then, with probability $1-\delta_2$,*

$$\begin{aligned}&\sum_i \sum_{k=1}^{d}(np_i(1-p_i))^k \leq \\ &\left(\frac{n}{n-1}\right)^d \left(\sum_i \sum_{k=1}^{d}(n\hat{p}_i(1-\hat{p}_i))^k + d\sqrt{\frac{1}{2}\log(1/\delta_2)}\left(2n^{d-1/2} + 2n^{d-3/2}\right)\right)\end{aligned} \tag{42}$$

for every even m.

Proof McDiarmid's inequality suggests that

$$\mathbb{P}\left(\sum_i\sum_{k=1}^{d}(n\hat{p}_i(1-\hat{p}_i))^k - \mathbb{E}\left(\sum_i\sum_{k=1}^{d}(n\hat{p}_i(1-\hat{p}_i))^k\right) \le -\epsilon_n\right) \le \exp\left(\frac{-2\epsilon_n^2}{\sum_{j=1}^n c_j^2}\right)$$

where c_j follows

$$\sup_{x_j'}\Big|\sum_i\sum_{k=1}^{d}(n\hat{p}_i(1-\hat{p}_i))^k - \sum_i\sum_{k=1}^{d}\left(n\hat{p}_i'(1-\hat{p}_i')\right)^k)\Big| \le c_j. \tag{43}$$

First, let us find c_j. We have

$$\begin{aligned}
&\sup_{x_j'}\Big|\sum_i\sum_{k=1}^{d}(n\hat{p}_i(1-\hat{p}_i))^k - \sum_i\sum_{k=1}^{d}\left(n\hat{p}_i'(1-\hat{p}_i')\right)^k)\Big| \overset{\text{(i)}}{\le} \qquad (44)\\
&2\sup_{p\in[0,1-1/n]}\Big|\sum_{k=1}^{d}(np(1-p))^k - \sum_{k=1}^{d}(n(p+1/n)(1-(p+1/n)))^k)\Big| =\\
&2\sup_{p\in[0,1-1/n]}\Big|\sum_{k=1}^{d}n^k\left(p(1-p)\right)^k - ((p+1/n)(1-(p+1/n)))^k)\Big| \le\\
&2\sum_{k=1}^{d}n^k\sup_{p\in[0,1-1/n]}\Big|\left(p(1-p)\right)^k - ((p+1/n)(1-(p+1/n)))^k)\Big| \overset{\text{(ii)}}{=}\\
&2\sum_{k=1}^{d}n^k\left(\frac{k}{n\cdot 4^{k-1}} + \frac{k(k-1)}{n^2\cdot 2^{2k-3}}\right) \le\\
&n^{d-1}\sum_{k=1}^{d}\left(\frac{2k}{4^{k-1}} + \frac{k(k-1)}{n\cdot 4^{k-2}}\right) \overset{\text{(iii)}}{\le} 2dn^{d-1} + 2dn^{d-2}
\end{aligned}$$

where

(i) Changing a single observation effects only two symbols (for example, $\hat{p}_l$ and $\hat{p}_t$), where the change is $\pm 1/n$.

(ii) Follows from (34).

(iii) Follows from $\sum_{k=1}^{d}\frac{k}{4^{k-1}} = 4\sum_{k=1}^{d}\frac{k}{4^k} \le d$ and

$$\sum_{k=1}^{d}\frac{k(k-1)}{4^{k-2}} \le \sum_{k=1}^{d}\frac{k^2}{4^{k-2}} \le d\max_{k\in[1,d]}\frac{k^2}{4^{k-2}} \le 2d \tag{45}$$

where the maximum is obtain for $k^* = 2/\log(4)$.

Next, we have

$$\begin{aligned}
&\mathbb{E}\Big(\sum_i\sum_{k=1}^{d}(n\hat{p}_i(1-\hat{p}_i))^k\Big) \ge \sum_i\sum_{k=1}^{d}\left(\mathbb{E}(n\hat{p}_i(1-\hat{p}_i))\right)^k = \qquad (46)\\
&\qquad\sum_i\sum_{k=1}^{d}n^k\left(\left(1-\frac{1}{n}\right)p_i(1-p_i)\right)^k \ge \left(1-\frac{1}{n}\right)^d\sum_i\sum_{k=1}^{d}((n-1)p_i(1-p_i))^k
\end{aligned}$$

Going back to McDiarmid's inequality, we have

$$\mathbb{P}\left(\mathbb{E}\left(\sum_i\sum_{k=1}^{d}(n\hat{p}_i(1-\hat{p}_i))^k\right)\geq\sum_i\sum_{k=1}^{d}(n\hat{p}_i(1-\hat{p}_i))^k+\epsilon_n\right)\leq\exp\left(\frac{-2\epsilon_n^2}{\sum_{j=1}^{n}c_j^2}\right) \tag{47}$$

Plugging (46) we obtain

$$\mathbb{P}\left(\left(1-\frac{1}{n}\right)^d\sum_i\sum_{k=1}^{d}(np_i(1-p_i))^k\geq\sum_i\sum_{k=1}^{d}(n\hat{p}_i(1-\hat{p}_i))^k+\epsilon_n\right)\leq\exp\left(\frac{-2\epsilon_n^2}{\sum_j c_j^2}\right)$$

Setting the right hand side to equal δ_2 we get

$$\epsilon_n=\sqrt{\frac{n}{2}\log(1/\delta_2)}\left(2dn^{d-1}+2dn^{d-2}\right)$$

and with probability $1-\delta_2$,

$$\sum_i\sum_{k=1}^{d}(np_i(1-p_i))^k\leq \tag{48}$$
$$\left(\frac{n}{n-1}\right)^d\left(\sum_i\sum_{k=1}^{d}(n\hat{p}_i(1-\hat{p}_i))^k+d\sqrt{\frac{1}{2}\log(1/\delta_2)}\left(2dn^{d-1/2}+2dn^{d-3/2}\right)\right)$$

■

Finally, we apply the union bound to Propositions 14 and 15 to obtain

$$\sup_i|p_i-\hat{p}_i(X^n)|\leq$$
$$\frac{m}{2n}\left(\frac{1}{\delta_1}\left(\frac{n}{n-1}\right)^d\left(\sum_i\sum_{k=1}^{d}(n\hat{p}_i(1-\hat{p}_i))^k+d\sqrt{\frac{1}{2}\log(1/\delta_2)}\left(2n^{d-1/2}+2^{d-3/2}\right)\right)\right)^{1/m}\leq$$
$$\frac{m}{2\delta_1^{1/m}}\frac{1}{n}\left(\frac{n}{n-1}\right)^{1/2}\left(\sum_i\sum_{k=1}^{m/2}(n\hat{p}_i(1-\hat{p}_i))^k\right)^{1/m}+$$
$$\frac{m}{2\delta_1^{1/m}}\frac{1}{n}\left(\frac{n}{n-1}\right)^{1/2}(m/2)^{1/m}\left(\frac{1}{2}\log\left(\frac{1}{\delta_2}\right)\right)^{1/2m}\left(2n^{\frac{1}{2}-\frac{1}{2m}}+2n^{\frac{1}{2}-\frac{3}{2m}}\right)$$

with probability $1-\delta_1-\delta_2$. Notice that $\sqrt{n/(n-1)}\leq\sqrt{2}$. Define $g(m,\delta_1)=m/(\sqrt{2}\delta_1^{1/m})$. Further, it is immediate to show that $(m/2)^{1/m}\leq\sqrt{\exp(1/\exp(1))}$ and $(1/2)^{1/2m}\leq 1/\sqrt{2}$. Hence, with probability $1-\delta_1-\delta_2$,

$$\sup_i|p_i-\hat{p}_i(X^n)|\leq\frac{g(m,\delta_1)}{n}\left(\sum_i\sum_{k=1}^{m/2}(n\hat{p}_i(1-\hat{p}_i))^k\right)^{1/m}+$$
$$bg(m,\delta_1)(\log(1/\delta_2))^{1/2m}\left(n^{-\frac{1}{2}-\frac{1}{2m}}+n^{-\frac{1}{2}-\frac{3}{2m}}\right)$$

for every even m, where $b=\sqrt{2\exp(1/\exp(1))}$. Finally, we would like to choose m which minimizes $g(m,\delta_1)$. It is immediate to show that $\inf_m g(m,\delta_1)=(\exp(1)/2)\log(1/\delta_1)/\sqrt{2}$, where and the infimum is obtained for a choice of $m^*=\log(1/\delta_1)$.

7.4 A Proof of Theorem 4

Let us first introduce some auxiliary results and background

7.4.1 Auxiliary Results

Lemma 16 (contained in the proof of Lemma 10, Cohen and Kontorovich (2023)) *Let $Y_{i\in I\subseteq\mathbb{N}}$ be random variables such that, for each $i \in I$, there are $v_i > 0$ and $a_i \geq 0$ satisfying*

$$\mathbb{P}(Y_i \geq \varepsilon) \quad \leq \quad \exp\left(-\frac{\varepsilon^2}{2(v_i + a_i\varepsilon)}\right), \qquad \varepsilon \geq 0. \tag{49}$$

Put

$$v^* := \sup_{i\in I} v_i,\ V^* := \sup_{i\in I} v_i \log(i+1),\ a^* := \sup_{i\in I} a_i,\ A^* := \sup_{i\in I} a_i \log(i+1). \tag{50}$$

Then

$$\mathbb{P}\left(\sup_{i\in I} Y_i \geq 2\sqrt{V^* + v^* \log\frac{1}{\delta}} + 4A^* + 4a^* \log\frac{1}{\delta}\right) \quad \leq \quad \delta.$$

Remark 17 *When considering the random variable $Z = \sup_{i\in\mathbb{N}} |\hat{p}_i - p_i|$, there is no loss of generality in assuming that $p_i \leq 1/2$, $i \in \mathbb{N}$. Indeed, $|Y_i| = |\hat{p}_i - p_i|$ is distributed as $|n^{-1}\operatorname{Bin}(n, p_i) - p_i|$, and the latter distribution is invariant under the transformation $p_i \mapsto 1 - p_i$. This is easily verified via the calculation $\mathbb{P}(\operatorname{Bin}(n,p) = k) = \mathbb{P}(\operatorname{Bin}(n, 1-p) = n-k)$.*

Lemma 18 *For any distribution $p_{i\in\mathbb{N}}$,*

$$V^*(p) \quad \leq \quad \varphi(v^*(p)).$$

Proof (This elegant proof idea is due to Václav Voráček.) There is no loss of generality in assuming $p = p^{\downarrow}$. The claim then amounts to

$$\sup_{i\in\mathbb{N}} v_i \log(i+1) \leq v^* \log\frac{1}{v^*}.$$

The monotonicity of the p_i implies $p_i \leq (p_1 + \ldots + p_i)/i \leq 1/i$. Now $x \leq 1/i \implies x(1-x) \leq 1/(i+1)$ for $i \in \mathbb{N}$, and hence $v_i \leq 1/(i+1)$. Thus, $v_i \log(i+1) \leq v_i \log\frac{1}{v_i}$. Finally, since $x\log(1/x)$ is increasing on $[0, 1/4]$, which is the range of the v_i, we have $\sup_{i\in\mathbb{N}} v_i \log\frac{1}{v_i} \leq v^* \log\frac{1}{v^*}$. ■

Remark 19 *There is no reverse inequality of the form $\varphi(v^*(p)) \leq F(V^*(p))$, for any fixed $F : \mathbb{R}_+ \to \mathbb{R}_+$. This can be seen by considering p supported on $[k]$, with $p_1 = \log(k)/k$ and the remaining masses uniform. Then $V^*(p) \approx \log(k)/k$ while $\varphi(v^*(p)) \approx \log(k)\log(k/\log k)/k$.*

Proposition 20 *Let $n \geq 10$ and $\beta = \log(n)$. Then,*

$$f(n) := \frac{\beta^{-\beta} n^2 \left(\frac{n-\beta}{n}\right)^{\beta-n}}{2^\beta - 2} \leq \frac{81}{2}.$$

Proof To prove the above, we show that $f(n)$ is decreasing for $n > 200$. This means that the maximum of $f(n)$ may be numerically evaluated in the range $n \in \{10, ..., 200\}$. Finally, we verify that the maximum of $f(n)$ is attained for $n = 33$, and is bounded from above by $81/2$ as desired. It remains to verify that $f(n)$ is decreasing for $n > 200$. Since $f(n)$ is non-negative, it is enough to show that $g(n) = \log f(n)$ is decreasing. Denote

$$g(n) = -\beta \log \beta + 2 \log n + (n-\beta)\log(n-\beta) + (n-\beta)\log n - \log(2^\beta - 2). \tag{51}$$

Taking the derivative of $g(n)$ we have,

$$\begin{aligned} &g'(n) = \\ &-\frac{1}{n}(\log\beta + 1) + \frac{2}{n} + \left(1 - \frac{1}{n}\right)(-\log(n-\beta) - 1 + \log n) + \frac{n-\beta}{n} - \frac{1}{n}\frac{2^\beta \log 2}{2^\beta - 2} = \\ &\frac{1}{n}\left((n-1)\log\frac{n}{n-\beta} - \log\beta - \beta + 2 - \frac{2^\beta \log 2}{2^\beta - 2}\right) \leq \\ &\frac{1}{n}\left(n\log\frac{n}{n-\beta} - \log\beta - \beta + 2 - \log 2\right) \leq \frac{1}{n}\left(\frac{n\beta}{n-\beta} - \log\beta - \beta + 2 - \log 2\right) = \\ &\frac{1}{n}\left(\frac{\beta^2}{n-\beta} - \log\beta + 2 - \log 2\right), \end{aligned} \tag{52}$$

where the first inequality follows from $\log(n/(n-\beta)) \geq 1$ and $2^\beta/(2^\beta - 2) \geq 1$, while the second inequality is due to Bernoulli's inequality, $(n/(n-\beta))^n \leq \exp(n\beta/(n-\beta))$. Finally, it is easy to show that $\beta^2/(n-\beta)$ is decreasing for $n \geq 10$. This means that $\beta^2/(n-\beta) \leq (\log 10)^2/(10 - \log(10))$ and $g'(n) < 0$ for $n > 200$. ■

Lemma 21 (generalized Fano method (Yu, 1997), Lemma 3) *For $r \geq 2$, let $\mathcal{M}_r$ be a collection of r probability measures $\nu_1, \nu_2, ..., \nu_r$ with some parameter of interest $\theta(\nu)$ taking values in pseudo-metric space (Θ, ρ) such that for all $j \neq k$, we have*

$$\rho(\theta(\nu_j), \theta(\nu_k)) \geq \alpha$$

and

$$D(\nu_j \parallel \nu_k) \leq \beta.$$

Then

$$\inf_{\hat{\theta}} \max_{j \in [d]} \mathbb{E}_{Z \sim \mu_j} \rho(\hat{\theta}(Z), \theta(\nu_j)) \geq \frac{\alpha}{2}\left(1 - \left(\frac{\beta + \log 2}{\log r}\right)\right),$$

where the infimum is over all estimators $\hat{\theta} : Z \mapsto \Theta$.

Proposition 22 *Let p and q be two distributions with support size n. Define p by*

$$p_1 = \frac{\log n}{2n \log\log n}, \quad p_i = \frac{1 - p_1}{n-1}, \qquad i > 1,$$

and q by $q_2 = p_1$, and $q_i = p_2$ for $i \neq 2$. Then,

(i) $\|p-q\|_\infty \geq c\frac{\log n}{n\log\log n}$ *for some* $c>0$ *and all* n *sufficiently large.*

(ii) $\lim_{n\to\infty}\frac{n}{\log n}D(p||q)=\frac{1}{2}$

Proof For the first part, it is enough to show that

$$|p_1-p_2|\geq c\log(n)/n\log\log n$$

for some $c>0$ and sufficiently large n. First, we show that $p_1\geq p_2$ for $n\geq(\log n)^2$. That is,

$$p_1-\frac{1-p_1}{n-1}=\frac{np_1-1}{n-1}>0 \tag{53}$$

for $np_1>1$. Next, fix $0<c\leq 1/2$. We have,

$$\begin{aligned}|p_1-p_2|-\frac{c\log(n)}{n\log\log n}=&\frac{ap_1-1}{n-1}-\frac{c\log n}{n\log\log n}= \\ &\frac{1}{n-1}\left(\frac{\log n}{2\log\log n}-1-\frac{n-1}{n}\frac{c\log n}{\log\log n}\right)= \\ &\frac{1}{(n-1)2\log\log n}\left(\log n\left(1-\frac{n-1}{n}2c\right)-2\log\log n\right)>0\end{aligned} \tag{54}$$

where the last inequality holds for $c(n-1)/n<1/2$ and sufficiently large n, as desired. We now proceed to the second part of the proof.

$$\frac{n}{\log n}D(p||q)=\frac{n}{\log n}\left(p_1\log\frac{p_1}{q_1}+p_2\log\frac{p_2}{q_2}\right)=\frac{n}{\log n}(p_1-p_2)\log\frac{p_1}{p_2}. \tag{55}$$

First, we have

$$\begin{aligned}\frac{n}{\log n}(p_1-p_2)=&\frac{n}{\log n}\left(p_1-\frac{1-p_1}{n-1}\right)=\frac{n}{\log n}\left(\frac{np_1-1}{n-1}\right)= \\ &\frac{n}{\log n}\frac{\log n/2n\log\log n-1}{n-1}=\frac{n}{n-1}\left(\frac{1}{2\log\log n}-\frac{1}{\log n}\right).\end{aligned} \tag{56}$$

Next,

$$\begin{aligned}\log\frac{p_1}{p_2}=&\log(n-1)+\log\frac{p_1}{1-p_1}=\log(n-1)+\log\frac{\log n}{2n\log\log n-\log n}= \\ &\log(n-1)+\log\log n-2\log(2n\log\log n-\log n).\end{aligned} \tag{57}$$

Putting it all together we obtain

$$\begin{aligned}
&\frac{n}{\log n} D(p||q) = \qquad (58)\\
&\quad \frac{n}{n-1}\left(\frac{1}{2\log\log n} - \frac{1}{\log n}\right)(\log(n-1) + \log\log n - 2\log(2n\log\log n - \log n)) = \\
&\quad \frac{n}{n-1}\Bigg(\frac{\log(n-1)}{2\log\log n} - \frac{\log(n-1)}{\log n} + \frac{1}{2} - \frac{\log\log n}{\log n} - \\
&\qquad \frac{\log(2n\log\log n - \log n)}{2\log\log n} + \frac{\log(2n\log\log n - \log n)}{\log n}\Bigg) = \\
&\quad \frac{n}{n-1}\Bigg(\frac{1}{2} + \frac{\log(n-1) - \log(2n\log\log n - \log n)}{2\log\log n} + \\
&\qquad \frac{\log(2n\log\log n - \log n) - \log(n-1)}{\log n} - \frac{\log\log n}{\log n}\Bigg).
\end{aligned}$$

It is straightforward to show that the last three terms in the parenthesis above converge to zero for sufficiently large n, which leads to the stated result. ■

Lemma 23 (Peres (2017)) *When estimating a single Bernoulli parameter in the range $[0, p_0]$ for $p_0 \leq \frac{1}{2}$, $\Theta(p_0\varepsilon^{-2}\log(1/\delta))$ draws are both necessary and sufficient to achieve additive accuracy ε with probability at least $1-\delta$.*

7.4.2 Bernstein inequalities

Background: Let $Y \sim \mathrm{Bin}(n, \theta)$ be a Binomial random variable and let $\hat{\theta} = Y/n$ be the its MLE.

- Classic Bernstein (Boucheron et al., 2003):

$$\mathbb{P}\left(\hat{\theta} - \theta \geq \varepsilon\right) \quad \leq \quad \exp\left(-\frac{n\varepsilon^2}{2\theta(1-\theta) + \varepsilon/3}\right) \qquad (59)$$

  with an analogous bound for the left tail. This implies:

$$|\theta - \hat{\theta}| \quad \leq \quad \sqrt{\frac{2\theta(1-\theta)}{n}\log\frac{2}{\delta}} + \frac{2}{3n}\log\frac{2}{\delta}. \qquad (60)$$

- Empirical Bernstein (Dasgupta and Hsu, 2008a, Lemma 5):

$$|\theta - \hat{\theta}| \quad \leq \quad \sqrt{\frac{5\hat{\theta}(1-\hat{\theta})}{n}\log\frac{2}{\delta}} + \frac{5}{n}\log\frac{2}{\delta}. \qquad (61)$$

We are now ready to present the proof of Theorem 4. We assume without loss of generality that p is sorted in descending order: $p_1 \geq p_2 \geq \ldots$ and further, as per Remark 17, that $p_1 \leq 1/2$. The estimate $\hat{p}_i$ is just the MLE based on n iid draws.

Our strategy for analyzing $\sup_{i\in\mathbb{N}} |\hat{p}_i - p_i|$ will be to break up p into the "heavy" masses, where we apply a maximal Bernstein-type inequality, and the "light" masses, where we apply a multiplicative Chernoff-type bound.

We define the "heavy" masses as those with $p_i \geq 1/n$. Denote by $I \subset \mathbb{N}$ the set of corresponding indices and note that $|I| \leq n$. For $i \in I$, put $Y_i = \hat{p}_i - p_i$. Then (59) implies that each Y_i satisfies (49) with $v_i = p_i(1-p_i)/n$ and $a_i = 1/(3n)$; trivially, $\max_{i\in I} a_i \log(i+1) = \log(n+1)/(3n)$. Invoking Lemma 16 twice (once for Y_i and again for $-Y_i$) together with the union bound, we have, with probability $\geq 1-\delta$,

$$\max_{i\in I} |\hat{p}_i - p_i| \leq 2\sqrt{\frac{V^*}{n} + \frac{v^*}{n}\log\frac{2}{\delta}} + \frac{4\log(n+1)}{3n} + \frac{4}{3n}\log\frac{2}{\delta}. \tag{62}$$

Next, we analyze the light masses. Our first "segment" consisted of the $p_i \in [n^{-1}, 1]$; these were the heavy masses. We take the next segment to consist of $p_i \in [(2n)^{-1}, n^{-1}]$, of which there are at most $2n$ atoms. The segment after that will be in the range $[(4n)^{-1}, (2n)^{-1}]$, and, in general, the kth segment is in the range $[(2^k n)^{-1}, (2^{k-1}n)^{-1}]$, and will contain at most $2^k n$ atoms. To the kth segment, we apply the Chernoff bound $\mathbb{P}(\hat{p} \geq p+\varepsilon) \leq \exp(-nD(p+\varepsilon||p))$, where $p = (2^k n)^{-1}$ and $\varepsilon = \varepsilon_k = 2^k p\beta - p$, for some $\beta > 1$ to be specified below. [Note that $D(\alpha p||p)$ is monotonically increasing in p for fixed α, so we are justified in taking the left endpoint.] For this choice, in the kth segment we have

$$\begin{aligned} D(p+\varepsilon||p) &= D(2^k p\beta||p) = D\left(\frac{\beta}{n}\Big\|\frac{1}{2^k n}\right) \\ &= \frac{(n-\beta)\log\left(\frac{2^k(n-\beta)}{2^k n - 1}\right) + \beta\log\left(2^k\beta\right)}{n} \\ &\geq \frac{(n-\beta)\log\left(\frac{n-\beta}{n}\right) + \beta\log\left(2^k\beta\right)}{n}, \end{aligned}$$

since neglecting the $-1/2^k$ additive term in the denominator decreases the expression. Let E be the event that *any* of the p_is in any of the segments $k = 1, 2, \ldots$ has a corresponding $\hat{p}_i$ that exceeds β/n. Then

$$\mathbb{P}(E) \leq \sum_{k=1}^{\infty} 2^k n \exp\left(-(n-\beta)\log\left(\frac{n-\beta}{n}\right) - \beta\log\left(2^k\beta\right)\right) = \frac{2\beta^{-\beta} n\left(\frac{n-\beta}{n}\right)^{\beta-n}}{2^\beta - 2}.$$

For the choice $\beta = \log n$, we have

$$\mathbb{P}(E) \leq \frac{2\beta^{-\beta} n\left(\frac{n-\beta}{n}\right)^{\beta-n}}{2^\beta - 2} \leq \frac{81}{n}, \qquad n \geq 10, \tag{63}$$

which is proved in Proposition 20. Now E is the event that $\sup_{i:p_i<1/n}(\hat{p}_i - p_i) \geq \log(n)/n$. Since $p_i < 1/n$, there is no need to consider the left-tail deviation at this scale, as all of the probabilities will be zero. Combining (62) with (63) yields (16). Since Lemma 18 implies that $V^* \leq \varphi(v^*)$, (17) follows from (16). Finally, (18) follows from (16) via the obvious relation $V^* \leq \log(n+1)v^*$.

7.5 Proof of Theorem 6

We begin with an elementary observation: for $N \in \mathbb{N}$ and $a, b \in [0,1]^N$, we have

$$\left| \max_{i\in[N]} a_i(1-a_i) - \max_{i\in[N]} b_i(1-b_i) \right| \;\leq\; \max_{i\in[N]} |a_i - b_i| \,,$$

and this also carries over to $a, b \in [0,1]^{\mathbb{N}}$. Let us denote $v^* := \sup_{i\in\mathbb{N}} p_i(1-p_i)$ and $\hat{v}^* := \sup_{i\in\mathbb{N}} \hat{p}_i(1-\hat{p}_i)$.

Together with (18), this implies

$$|v^* - \hat{v}^*| \leq \|p - \hat{p}\|_\infty \leq a + b\sqrt{v^*}$$

where

$$\begin{aligned} a &= \frac{4}{3n}\log\frac{2(n+1)}{\delta} + \frac{\log n}{n}, \\ b &= 2\sqrt{\frac{\log(n+1)}{n} + \frac{1}{n}\log\frac{2}{\delta}}. \end{aligned}$$

Following the proof of Lemma 5 of Dasgupta and Hsu (2008a),

$$\begin{aligned} |v^* - \hat{v}^*| &\leq a + b\sqrt{v^*} \\ &\leq a + b\sqrt{\hat{v}^* + |v^* - \hat{v}^*|} \\ &\leq a + b\sqrt{\hat{v}^*} + b\sqrt{|v^* - \hat{v}^*|}, \end{aligned}$$

where we used $v^* \leq \hat{v}^* + |v^* - \hat{v}^*|$ and $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$.

Now we have an expression of the form

$$A \leq B\sqrt{A} + C,$$

where $A = |v^* - \hat{v}^*|$, $B = b$, $C = a + b\sqrt{\hat{v}^*}$, which implies $A \leq B^2 + B\sqrt{C} + C$, or

$$|v^* - \hat{v}^*| \;\leq\; b^2 + a + b\sqrt{\hat{v}^*} + b\sqrt{a + b\sqrt{\hat{v}^*}}.$$

Using $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ and $\sqrt{xy} \leq (x+y)/2$,

$$\begin{aligned} |v^* - \hat{v}^*| &\leq b^2 + a + b\sqrt{\hat{v}^*} + b\sqrt{a} + b\sqrt{b\sqrt{\hat{v}^*}} \\ &\leq b^2 + a + b\sqrt{\hat{v}^*} + b\sqrt{a} + b(b + \sqrt{\hat{v}^*})/2 \\ &= a + 3b^2/2 + b\sqrt{a} + 3b\sqrt{\hat{v}^*}/2. \end{aligned}$$

We still have

$$\begin{aligned} a + b\sqrt{v^*} &\leq a + b\sqrt{\hat{v}^*} + b\sqrt{|v^* - \hat{v}^*|} \\ &\leq a + b\sqrt{\hat{v}^*} + b^2/2 + |v^* - \hat{v}^*|/2, \end{aligned}$$

where $\sqrt{xy} \leq (x+y)/2$ was used. Hence, with probability $1-\delta$,

$$\begin{aligned} \|p - \hat{p}\|_\infty &\leq a + b\sqrt{v^*} \\ &\leq a + b\sqrt{\hat{v}^*} + b^2/2 + |v^* - \hat{v}^*|/2 \\ &\leq a + b\sqrt{\hat{v}^*} + b^2/2 + (a + 3b^2/2 + b\sqrt{a} + 3b\sqrt{\hat{v}^*}/2)/2 \\ &= 3a/2 + \sqrt{a}b/2 + 5b^2/4 + 7b\sqrt{\hat{v}^*}/4. \end{aligned}$$

7.6 A Proof of Proposition 8

The necessity of an additive term of order $\log(n)/(n\log\log n)$ can be intuited via a balls-in-bins analysis: If n balls are uniformly thrown into n bins, we expect a maximal load of about $\log(n)/(n\log\log n)$ balls in one of the bins (Raab and Steger, 1998). We proceed to formalize this intuition.

Let $\mu_1,\ldots,\mu_n$ be probability measures on $\mathbb{N}$ with support contained in $[n]$, defined as follows. For $a := \log(n)/(2n\log\log n)$ and $b := (1-a)/(n-1)$, we define, for $i \in [n]$, $\mu_i(i) = a$, for $j \neq i$, $\mu_i(j) = b$. For each $i \in [n]$, define the probability ν_i on $\mathbb{N}^n$ as the n-fold product measure μ_i^n.

It follows from Proposition 22 that

$$\frac{D(\nu_j||\nu_k)}{\log n} = \frac{nD(\mu_j||\mu_k)}{\log n} \underset{n\to\infty}{\longrightarrow} \frac{1}{2}$$

and $\|\mu_j - \mu_k\|_\infty \geq \alpha := c\log(n)/(n\log\log n)$ for $j \neq k$ and n sufficiently large. Invoking Lemma 21 with $r = n$, $\theta(\nu_j) = \mu_j$, $\rho = \|\cdot\|_\infty$, and $\beta = \frac{1}{2}\log n + o(\log n)$ completes the proof.

7.7 A Proof of Proposition 9

The analysis relies on a result of Cohen and Kontorovich (2023, Theorem 2). Let $X_i \sim \mathrm{Bin}(n,p_i)$, $i \in \mathbb{N}$ be a sequence of independent binomials with $1/2 \geq p_1 \geq p_2 \geq \ldots$ and define $Y_i := n^{-1}X_i - p_i$. Then Cohen and Kontorovich (2023) showed[2] that

$$c\sqrt{S} \leq \liminf_{n\to\infty} \sqrt{n}\mathbb{E}\sup_{i\in\mathbb{N}} Y_i, \tag{64}$$

where $S := \sup_{i\in\mathbb{N}} p_i \log(i+1)$ and $c > 0$ is an absolute constant.

Since $t \leq 2t(1-t)$ for $t \in [0,1/2]$ and $p_1 \leq 1/2$ (as per Remark 17), we have that $V^* \geq S/2$. However, the Cohen and Kontorovich (2023) lower bound is not immediately applicable to our case, because (64) requires the binomials to be independent. Fortunately, their dependence is of the *negative association* type (Dubhashi and Ranjan, 1998, Theorem 14), which futher implies negative right orthant dependence (Proposition 5, ibid.). Finally, (Kontorovich, 2023, Proposition 4) shows that

$$\mathbb{E}\sup_{i\in\mathbb{N}} Y_i \quad \geq \quad \frac{1}{2}\mathbb{E}\sup_{i\in\mathbb{N}} \tilde{Y}_i, \tag{65}$$

where the $\tilde{Y}_i$ are mutually independent and each one is distributed identically to its corresponding Y_i. This completes the proof.

Remark 24 *The lower bound is only asymptotic (rather than* finite-sample*, in the sense of holding for all n) — necessarily so. This is because even for a single binomial $Y \sim \mathrm{Bin}(n,p)$, the behavior of $\mathbb{E}|Y - np|$ is roughly $np(1-p)$ for $p \notin [1/n, 1-1/n]$ and $\approx \sqrt{np(1-p)}$ elsewhere (Berend and Kontorovich, 2013, Theorem 1). This precludes any finite-sample lower bound of the form $\mathbb{E}\|p - \hat{p}\|_\infty \geq c\sqrt{V(p)/n}$.*

2. The theorem therein claimed this for $\mathbb{E}\sup_{i\in\mathbb{N}}|Y_i|$ but in fact the proof shows this for $\mathbb{E}\sup_{i\in\mathbb{N}} Y_i$.

7.8 A Proof of Proposition 11

It follows from Lemma 23 that $\mathbb{E}|p-\hat{p}| \geq c\sqrt{p_0/n}$ for some universal $c > 0$. Since $\mathbb{E}\sup_{i\in\mathbb{N}}|p_i-\hat{p}_i| \geq \sup_{i\in\mathbb{N}}\mathbb{E}|p_i-\hat{p}_i|$, this completes the proof.

7.9 Proof Theorem 12

We begin with the following proposition.

Proposition 25 *Let $j = \arg\max_i \hat{p}_i$. Assume there exists $V_\delta(X^n)$ such that*

$$\mathbb{P}\left(|p_j - \hat{p}_j| \geq V_\delta(X^n) | p_j = p_{[1]}\right) \leq \delta, \tag{66}$$

where the conditioning $p_j = p_{[1]}$ implies that the probability of the most frequent event in the sample is the maximal probability, $p_{[1]}$. Then,

$$\mathbb{E}(V_\delta(X^n)) \geq z_{\delta/2}\sqrt{\frac{p_{[1]}(1-p_{[1]})}{n}} + O\left(\frac{1}{n}\right).$$

Proof Let $j = \text{argmax}_i \hat{p}_i$. Assume there exists $V_\delta(X^n)$ that satisfies (66) and

$$\mathbb{E}(V_\delta(X^n)) < z_{\delta/2}\sqrt{\frac{p_{[1]}(1-p_{[1]})}{n}} + O\left(\frac{1}{n}\right).$$

From (66), we have that

$$\mathbb{P}\left(|p_j - \hat{p}_j| \geq V_\delta(X^n) | p_j = p_{[1]}\right) = \mathbb{P}\left(|p_{[1]} - \hat{p}_j| \geq V_\delta(X^n) | p_j = p_{[1]}\right) \leq \delta. \tag{67}$$

Now, consider $Y \sim \text{Bin}(n, p_{[1]})$. Let Y^n be a sample of n independent observations. Notice we can always extend the Binomial case to a multinomial setup with parameters p, over any alphabet size $||p||_0$. That is, given a sample Y^n, we may replace every $Y = 0$ (or $Y = 1$) with a sample from a multinomial distribution over an alphabet size $||p||_0 - 1$. Further, we may focus on samples for which $p_{[1]}$ is the most likely event in the alphabet, and construct a CI for $p_{[1]}$ following (67). This means that we found a CI for $p_{[1]}$ with an expected length that is shorter than the CP CI, which contradicts its optimality. ■

Now, assume there exists $U_\delta(X^n)$ that satisfies

$$\mathbb{P}\left(|p_j - \hat{p}_j| \geq U_\delta(X^n)\right) \leq \delta. \tag{68}$$

and

$$\mathbb{E}(U_\delta(X^n)) < z_{\delta/2}\sqrt{\frac{p_{[1]}(1-p_{[1]})}{n}} + O\left(\frac{1}{n}\right). \tag{69}$$

For simplicity of notation, denote $v = \arg\max_i p_i$ as the symbol with the greatest probability in the alphabet. That is, $p_v = p_{[1]}$. We implicitly assume that v is unique, although the

proof holds in case of several maxima as well. We have that

$$\mathbb{P}\left(|p_j - \hat{p}_j| \geq U_\delta(X^n)\right) = \sum_u \mathbb{P}\left(|p_j - \hat{p}_j| \geq U_\delta(X^n) | j = u\right) \mathbb{P}(j = u) = \tag{70}$$
$$\mathbb{P}\left(|p_{[1]} - \hat{p}_j| \geq U_\delta(X^n) | j = v\right) \mathbb{P}(j = v) +$$
$$\sum_{u \neq v} \mathbb{P}\left(|p_j - \hat{p}_j| \geq U_\delta(X^n) | j = u\right) \mathbb{P}(j = u).$$

Proposition 25 together with assumption (69) suggest that

$$\mathbb{P}\left(|p_{[1]} - \hat{p}_j| \geq U_\delta(X^n) | j = v\right) > \delta.$$

On the other hand, it is well-known that $\hat{p}_{[1]} \to p_{[1]}$ for sufficiently large n (Gelfand et al., 1992; Shifeng and Guoying, 2005; Xiong and Li, 2009). This means that $\mathbb{P}(j = v) \to 1$ and (70) is bounded from below by δ, for sufficiently large n. This contradicts (67) as desired.

8. Discussion and Conclusions

In this work we study distribution estimation under the ℓ_∞ norm. We introduce two data-dependent upper bounds for the MLE, which significantly improve upon currently known results. Our first bound (Theorem 2) demonstrates favorable performance in small sample size and fixed δ regimes. However, its dependency in the data is somewhat involved, compared to our "dream" result (3). Our second bound (Theorem 6) improves the explicit dependency in the data, and demonstrates favorable performance in larger sample regimes, where δ decays with n. The above upper bounds are matched by nearly-optimal lower bounds, demonstrating the tightness of our analysis. Finally, we introduce an important application to our work in selective inference. We show that by utilizing ℓ_∞ results, we provide relatively tight confidence interval for the most frequent events in the sample.

Acknowledgments

A.K. is supported in part by the Israel Science Foundation (grant No. 1602/19), an Amazon Research Award and the Ben-Gurion University Data Science Research Center. A.P. is supported in part by the Israel Science Foundation (grant No. 963/21)

References

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

Yoav Benjamini and Daniel Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100 (469):71–81, 2005.

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

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, pages 802–837, 2013.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pages 208–240. Springer, 2003.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer school on machine learning*, pages 169–207. Springer, 2003.

US Census Bureau. Frequently occurring surnames from the census 2000. 2014.

Djalil Chafai and Didier Concordet. Confidence regions for the multinomial parameter with small sample size. *Journal of the American Statistical Association*, 104(487):1071–1079, 2009.

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Doron Cohen and Aryeh Kontorovich. Local glivenko-cantelli. In *The Thirty Sixth Annual Conference on Learning Theory, COLT*, volume 195, page 715, 2023.

Doron Cohen, Aryeh Kontorovich, and Geoffrey Wolfer. Learning discrete distributions with infinite support. In *Advances in Neural Information Processing Systems 33, 6-12*, 2020.

Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, pages 208–215, 2008a.

Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pages 208–215, 2008b.

Devdatt Dubhashi and Desh Ranjan. Balls and bins: a study in negative dependence. *Random Struct. Algorithms*, 13(2):99–124, September 1998. ISSN 1042-9832.

AE Gelfand, J Glaz, L Kuo, and T-M Lee. Inference for the maximum cell probability under multinomial sampling. *Naval Research Logistics (NRL)*, 39(1):97–114, 1992.

Leo A Goodman et al. Simultaneous confidence intervals for contrasts among multinomial populations. *The Annals of Mathematical Statistics*, 35(2):716–725, 1964.

Ernst Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136):210–271, 1909.

Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Maximum likelihood estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 63(10):6774–6798, 2017.

Leonid V Kantorovich. Mathematical methods of organizing and planning production. *Management science*, 6(4):366–422, 1960.

Aryeh Kontorovich. Decoupling maximal inequalities, 2023.

Ivor John Maddox. *Elements of functional analysis*. CUP Archive, 1988.

Matthew L Malloy, Ardhendu Tripathy, and Robert D Nowak. Optimal confidence regions for the multinomial parameter. *arXiv preprint arXiv:2002.01044*, 2020.

David A McAllester and Robert E Schapire. On the convergence rate of Good-Turing estimators. In *COLT*, pages 1–6, 2000.

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.

Amichai Painsky. Large alphabet inference. *Information and Inference: A Journal of the IMA*, 12(4), 2023a.

Amichai Painsky. Quality assessment and evaluation criteria in supervised learning. *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, pages 171–195, 2023b.

Amichai Painsky and Gregory W Wornell. Bregman divergence bounds and universality properties of the logarithmic loss. *IEEE Transactions on Information Theory*, 66(3): 1658–1673, 2019.

Yuval Peres. Learning a coin's bias (localized). Theoretical Computer Science Stack Exchange, 2017. https://cstheory.stackexchange.com/q/38931 (version: 2017-08-28).

Charles P Quesenberry and DC Hurst. Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics*, 6(2):191–195, 1964.

Martin Raab and Angelika Steger. "Balls into Bins" - A simple and tight analysis. In Michael Luby, José D. P. Rolim, and Maria J. Serna, editors, *Randomization and Approximation Techniques in Computer Science*, volume 1518, pages 159–170. Springer, 1998.

John A Rice. *Mathematical statistics and data analysis*. Cengage Learning, 2006.

Alexander I Saichev, Yannick Malevergne, and Didier Sornette. *Theory of Zipf's law and beyond*, volume 632. Springer Science & Business Media, 2009.

Xiong Shifeng and Li Guoying. Testing for the maximum cell probabilities in multinomial distributions. *Science in China Series A: Mathematics*, 48:972–985, 2005.

Cristina P Sison and Joseph Glaz. Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association*, 90(429):366–369, 1995.

Maciej Skorski. Handy formulas for binomial moments. *arXiv preprint arXiv:2012.06270*, 2020.

Nickolay Smirnov. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281, 1948.

Karl R Stromberg. *An introduction to classical real analysis*, volume 376. American Mathematical Soc., 2015.

Måns Thulin. The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics*, 8(1):817–840, 2014.

Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

Ramon Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2014.

Weizhen Wang. Smallest confidence intervals for one binomial proportion. *Journal of Statistical Planning and Inference*, 136(12):4293–4306, 2006.

ShiFeng Xiong and GuoYing Li. Inference for ordered parameters in multinomial distributions. *Science in China Series A: Mathematics*, 52(3):526–538, 2009.

Bin Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435, 1997.