Extrapolating the profile of a finite population

Soham Jana SOHAM.JANA@YALE.EDU
Department of Statistics and Data Science, Yale University, New Haven, CT

Yury Polyanskiy YP@MIT.EDU
Department of EECS, MIT, Cambridge, MA

Yihong Wu YIHONG.WU@YALE.EDU
Department of Statistics and Data Science, Yale University, New Haven, CT

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We study a prototypical problem in empirical Bayes. Namely, consider a population consisting of k individuals each belonging to one of k types (some types can be empty). Without any structural restrictions, it is impossible to learn the composition of the full population having observed only a small (random) subsample of size $m = o(k)$. Nevertheless, we show that in the sublinear regime of $m = \omega(k/\log k)$, it is possible to consistently estimate in total variation the *profile* of the population, defined as the empirical distribution of the sizes of each type, which determines many symmetric properties of the population. We also prove that in the linear regime of $m = ck$ for any constant c the optimal rate is $\Theta(1/\log k)$. Our estimator is based on Wolfowitz's minimum distance method, which entails solving a linear program (LP) of size k. We show that there is a single infinite-dimensional LP whose value simultaneously characterizes the risk of the minimum distance estimator and certifies its minimax optimality. The sharp convergence rate is obtained by evaluating this LP using complex-analytic techniques.

Keywords: High-dimensional statistics, empirical Bayes, sublinear algorithms, minimax rate, H^∞-relaxation, Laguerre polynomials.

1. Introduction

Consider a finite population, say, an urn of at most k colored balls, with colors indexed by, without loss of generality, $[k] \triangleq \{1, \ldots, k\}$. Let θ_j denote the the number of balls of color $j \in [k]$ present in the urn. We observe a subsample, obtained by revealing each ball independently with probability p. This sampling scheme is referred to as the Bernoulli sampling model Bunge and Fitzpatrick (1993), a specific form of sampling without replacements. We will be interested in both the *linear* and the *sublinear* regime, in which the sampling probability p is a small constant or vanishing as k grows, respectively.

It is not hard to show (see Appendix B) that unless all but a vanishing fraction of the urn is observed, it is impossible to consistently estimate the empirical distribution of the colors, which aligns with the conventional wisdom that the sample size needs to exceed the number of parameters. Fortunately, many interesting properties about the population (such as entropy, number of distinct elements) are label-invariant and hence learnable through the *profile* of the population Orlitsky et al.

(2005), defined as the empirical distribution of $\theta = (\theta_1, \ldots, \theta_k)$:

$$\pi = \frac{1}{k}\sum_{j=1}^{k} \delta_{\theta_j}. \tag{1}$$

where δ_m denotes the Dirac measure (point mass) at m, Note that π is supported on $\{0, \ldots, k\}$ with mean at most one and probability mass function given by $\pi_m = \frac{1}{k}\sum_{j=1}^{k} \mathbf{1}_{\{\theta_j = m\}}$ for $m = 0, \ldots, k$. The profile provides information about the diversity of a population. For example, $\pi = (1-\frac{1}{k})\delta_0 + \frac{1}{k}\delta_k$ and $\pi = \delta_1$ correspond to the two extremes of all balls having the same color and different colors, respectively. Furthermore, π_0 encodes the total number c of distinct colors in urn, since $\pi_0 = 1 - c/k$.

Based on the subsampled population, our goal is to reconstruct the profile π of the full population. Since many symmetric properties can be expressed as its linear functionals, estimating π under the total variation (TV) distance allows simultaneous estimation of all such bounded properties. Our main result is that the profile can be estimated consistently even in the sublinear regime.

Let $X_j \sim \text{Binom}(\theta_j, p)$ be the number of observed balls of color j. The minimax TV risk of estimating π is defined as

$$R(k) = \inf \sup \mathbb{E}[\|\pi - \hat{\pi}\|_{\text{TV}}]. \tag{2}$$

where $\|\pi - \hat{\pi}\|_{\text{TV}} \triangleq \frac{1}{2}\sum_{m \geq 0} |\pi_m - \hat{\pi}_m|$, the supremum is over all urns of at most k balls, and the infimum is over all estimators $\hat{\pi}$ as a function of $X = (X_1, \ldots, X_k)$. Our main result is the following.

Theorem 1 *There exist absolute constants c, C, d_0, such that if $\log k \geq \frac{d_0}{\bar{p}}$, then*

$$\min\left\{\frac{\bar{p}}{p}, \sqrt{\log k}\right\}\frac{c}{\log k} \leq R(k) \leq \min\left\{\frac{C}{p\log k}, 1\right\}, \tag{3}$$

where $\bar{p} = 1 - p$. Furthermore, the upper bound in fact holds for all $p \in (0,1)$, achieved by a minimum-distance estimator computable in polynomial time.

In the linear regime, Theorem 1 shows that the optimal TV rate is $\Theta(\frac{1}{\log k})$ for any constant sampling probability p. This should be contrasted with the estimation of π_0, known as the distinct elements problem, which has been extensive studied in the literature Bunge and Fitzpatrick (1993); Charikar et al. (2000); Raskhodnikova et al. (2009); Valiant and Valiant (2011); Wu and Yang (2018). The precise behavior of the minimax risk of estimating π_0 was determined in Wu and Yang (2018). In particular, if $\frac{1}{\log k} \lesssim p \lesssim 1$, the optimal rate of π_0 is $k^{-\Theta(p)}$, much faster than estimating π itself. Our result refines this observation and reveals the following dichotomy: the polynomial rate $k^{-\Theta(p)}$ holds not just for estimating π_0 but for all π_m with $m = o(\log k)$; however, for $m = \Theta(\log k)$, π_m is much harder to estimate and the rate is no faster than $\Omega(\frac{1}{(\log k)^2})$. This explains the overall TV risk $\Omega(\frac{1}{\log k})$ for estimating the full distribution π.

In the sublinear regime, Theorem 1 shows that consistent estimation is possible if $p = \omega(\frac{1}{\log k})$. Although our current lower bound does not conclude its optimality, it is indeed the case based on existing impossibility results of the distinct element problem that shows π_0 cannot be estimated with vanishing error if $p = O(\frac{1}{\log k})$ Valiant (2012); Wu and Yang (2018).

For simplicity, we focus on the Bernoulli sampling model in this paper. The results can be extended to models such as iid sampling or Poisson sampling by the usual simulation or reduction argument (cf. (Wu and Yang, 2018, Appendix A)).

1.1. Related work

While the precise question we are considering here was not studied before, there is a long history of related work. First we observe that the goal of estimating functionals of $\theta = (\theta_1, \dots, \theta_k)$ is a "compound statistical decision problem", in the language of Robbins (1951). Instead of studying minimax risks of estimating θ or its functionals, Robbins (1951) proposed an alternative goal ("subminimaxity"), which in our case can be rephrased as follows: construct an estimator which has vanishing excess risk (regret) over that of the oracle estimator $\widehat{k_j}(X_j, \pi)$ having access to empirical distribution π of θ. The general recipe proposed in Robbins (1951) (and later promulgated by Robbins (1956) under the name of "empirical Bayes"), may roughly be described as a two-step procedure: first, one produces an estimate $\hat{\pi}$ of π, and then, second, substitutes it into the oracle estimator obtaining $\widehat{k_j}(X_j, \hat{\pi})$. Thus, Robbins (Robbins, 1951, p. 146) asked (his Problem I) how well can the first step be done? Our work addresses this question.

The main part of our theorem characterizes how well the "prior" π can be estimated. We mention that while empirical Bayes method is sometimes understood only as a way to derive estimates of a particular functional of the prior, as, for example, in the Good-Turing estimator for the number of unseen species, the idea of estimating the prior itself has also been proposed in Robbins (1956); Edelman (1988). Furthermore, the solution advocated therein, Wolfowitz's *minimum distance estimator* Wolfowitz (1957), is the one we employ in the proof of our result. In this regard, one of the main contributions of the paper is showing that performance of the minimum distance estimators is characterized by means of a certain function $\delta_{\mathrm{TV}}(t)$, defined as the value of an infinite-dimensional linear program, which simultaneously can also be used to produce a matching *lower bound*. This duality between the upper and the lower bound has previously been observed and operationalized in the context of estimating *a single linear functional* in Juditsky and Nemirovski (2009); Polyanskiy et al. (2017); Polyanskiy and Wu (2019). Here we extend this program to estimating the *full distribution*, and evaluate the relevant δ_{TV} function using complex-analytic techniques.

Arguably, the counterintuitive part of our result is the possibility of estimating the profile π consistently in TV, despite the absence of structural assumptions on the urn configuration and despite p possibly vanishing. In fact, this is a manifestation of the fascinating effect originally discovered by Orlitsky et al. (2005) and further developed in Valiant and Valiant (2013); Han et al. (2018), namely, although there exists no consistent estimator of the empirical color distribution, its sorted version can be estimated consistently. Nevertheless, the best upper bound that can be extracted (see Appendix A.1 for details) from existing results is $O(\frac{1}{\sqrt{\log k}})$ in the linear regime and there is no applicable lower bound. Theorem 1 shows that this rate is suboptimal by a square root factor, potentially due to the fact that these previous work did not exploit the finiteness of the population.

In terms of techniques, while the approach of Wu and Yang (2018) to the distinct elements problem relies on polynomial interpolation and approximation, both the scheme (minimum distance estimator) and the lower bound in the present paper involve linear programming (LP), which is more akin in spirit to the work of Valiant and Valiant (2011); Polyanskiy and Wu (2019). The technical novelty here is that we use tools from complex analysis to analyze the behavior of the LP.

Finally, we mention that a different line of research tracing back to Lord (1969) studies the "mirror image" of our problem: estimating the empirical distribution of parameters $p_1, \ldots, p_k$ from samples $X_j \sim \text{Binom}(\theta, p_j)$. The recent work of Tian et al. (2017) uses the method of moments to obtain the optimal rate for $\theta = o(\log k)$. This is further improved in Vinayak et al. (2019) by analyzing the nonparametric maximum likelihood. Alas, in this model, even for large population it is not possible to achieve consistent estimation without $\theta \to \infty$.

The rest of the paper is organized as follows. Section 2 introduces the minimum distance estimator and a general characterization of its risk by a linear program. Sections 3 and 4 are devoted to analyzing the behavior of this LP using complex-analytic techniques and Laguerre polynomials, completing the proof of Theorem 1. Appendix A contains a detailed discussion on related technical results and a list of open problems. Omitted proofs are contained in the rest of the appendices.

2. Minimum distance estimator and statistical guarantees

As mentioned in the last section, estimation of the profile revolves around the idea of minimum distance method, which fits a statistical model that is closest to the sample distribution with respect to some meaningful statistical distance. Examples of minimum distance estimators can be traced back to as early as Pearson (1900), which led to the discovery of the famous minimum chi-square method. In the 1950's, Wolfowitz studied minimum distance methods for the first time as a class, for obtaining strongly (almost surely) consistent estimators Wolfowitz (1957). The pioneering work of Beran (1977) demonstrates how minimum-Hellinger method can improve upon classical estimators such as the maximum likelihood in the presence of outliers. For a comprehensive account and more recent development we refer the readers to the monograph Basu et al. (2011).

To describe the paradigm of the minimum distance estimators we first introduce the general setting of Robbins' Problem I mentioned in Section 1.1. Consider a parametric family of distributions $\{P_\theta : \theta \in \Theta\}$ on some measurable space $\mathcal{X}$, viewed also as a Markov transition kernel P from Θ to $\mathcal{X}$. Let d be a distance on the space of priors $\mathcal{P}(\Theta)$. Select $\theta_1, \ldots, \theta_k$ from Θ such that $\frac{1}{k}\sum_{j=1}^k c(\theta_j) \le 1$, where $c : \Theta \to \mathbb{R}$ is some cost function (could be zero), resulting in the empirical distribution $\pi \triangleq \frac{1}{k}\sum_{j=1}^k \delta_{\theta_j}$. Given observations $X_j \overset{iid}{\sim} P_{\theta_j}$, an estimate $\hat{\pi}(X_1, \ldots, X_k)$ is produced with the goal of minimizing $\mathbb{E}[d(\hat{\pi}, \pi)]$. The minimax risk is defined as

$$R(k) = \inf_{\hat{\pi}} \sup_{\theta_1, \ldots, \theta_k} \mathbb{E}[d(\hat{\pi}, \pi)] .$$

Remark 2 *Note that Robbins also defined a related Problem II (Robbins, 1951, p. 147) in which* $\theta_j \overset{iid}{\sim} G$ *with* $\mathbb{E}_G[c(\theta)] \le 1$ *and the goal is to estimate the prior* G *instead of the (now random) empirical distribution* π*. The minimax risk* $R_2(k)$ *is similarly defined as the supremum over all such* G*. We argue that in many cases the difference between* $R(k)$ *and* $R_2(k)$ *is insignificant.*

Indeed, let $\tau_k = \sup_G \mathbb{E}[d(G, \pi)]$*, which due to concentration we assume is* $o(R(k))$*. The comparison* $R_2(k) \le R(k) + \tau_k$ *is by conditioning on* π*. In the opposite direction, if, for example,* $d(\cdot, \cdot) \le 1$*, then* $R(k) \le R_2(m) + \frac{m^2}{2k}$ *since by sampling* m *times from* $(X_1, \ldots, X_k)$ *with replacement we get* m *samples from Problem 2's setting with* $G = \pi$ *(except for a set of realizations of probability* $\frac{m^2}{2k}$ *on which we drew some* X_j *multiple times). Applying Problem 2's estimator for* m *samples we get the inequality. In interesting cases,* $R_2(k) \asymp R_2(k^\alpha) \ll k^{-\beta}$ *for any* $\alpha, \beta > 0$*, and thus we get* $R_1(k) \asymp R_2(k)$*.*

To solve this problem we proceed by choosing an auxiliary metric ρ on $\mathcal{P}(\Theta)$, the set of probability measures on Θ. Let $\hat{\nu} = \frac{1}{k}\sum_{j=1}^{k}\delta_{X_j}$ be the empirical distribution of the sample. Note that in expectation we have, for all $\theta_1, \ldots, \theta_k$,

$$\mathbb{E}[\hat{\nu}] = \pi P. \tag{4}$$

where $\pi P = \int P_\theta \pi(d\theta) = \frac{1}{k}\sum_{j=1}^{k} P_{\theta_j}$. This motivates the following minimum-distance estimator (putting existence of minimum aside):

$$\hat{\pi} = \operatorname*{argmin}_{\pi'} \left\{\rho(\hat{\nu}, \pi' P) : \mathbb{E}_{\pi'}[c(\theta)] \leq 1\right\} . \tag{5}$$

To analyze this estimator, suppose, in addition to (4), we have the high-probability guarantee:

$$\mathbb{P}[\rho(\pi P, \hat{\nu}) > t_k] \leq \epsilon_k$$

for some sequences $t_k, \epsilon_k \to 0$. By the triangle inequality we also have $\mathbb{P}[\rho(\hat{\pi} P, \pi P) > 2t_k] \leq \epsilon_k$. Finally, defining the following *deconvolution function*:

$$\delta(t) \triangleq \sup\{d(\pi, \pi') : \rho(\pi P, \pi' P) \leq t, \mathbb{E}_\pi[c(\theta)] \leq 1, \mathbb{E}_{\pi'}[c(\theta)] \leq 1\} ,$$

where the supremization is over all distributions $\pi, \pi' \in \mathcal{P}(\Theta)$. Then we immediately obtain the high-probability risk bound $\mathbb{P}[d(\hat{\pi}, \pi) > \delta(2t_k)] \leq \epsilon_k$. Using other properties of d and c, we can typically convert this into an upper bound for the average risk like $\mathbb{E}[d(\hat{\pi}, \pi)] \lesssim \delta(2t_k)$. Selecting different auxiliary metric ρ's results in different estimators. For example, the choice of ρ equal to the Kullback-Leibler divergence results in a the non-parametric maximum-likelihood estimator. As stated this is all well known. *Our key contribution is the following:* While ρ is left arbitrary so far, the choice of ρ being total variation (or Hellinger) distance is special since it comes with an essentially matching lower bound.

> *Meta-principle*. Suppose the loss function d is of seminorm-type, namely $d(\pi, \pi') = \sup_{T \in \mathcal{T}}\langle T, \pi - \pi'\rangle$ for some dual pairing $\langle \cdot, \cdot \rangle$ and a family of linear functionals $\mathcal{T}$ on $\mathcal{P}(\Theta)$. Take $\rho(\cdot, \cdot) = \| \cdot - \cdot \|_{\mathrm{TV}}$. Then under regularity conditions on $(\Theta, \mathcal{X}, c, P, \mathcal{T})$ we have
>
> $$\delta(1/k) \lesssim R(k) \lesssim \delta(t_k) .$$
>
> Thus, when $\delta(1/k) \asymp \delta(t_k)$ we get the sharp rate.

Working out general conditions for the applicability of this program is left for future work. Here we focus on the model discussed in the introduction. Recall $\pi = (\pi_0, \ldots, \pi_k)$ in (1) denotes the profile of the urn. In the Bernoulli sampling model, the observed numbers of balls with color j are independently distributed as

$$X_j \overset{\text{ind.}}{\sim} \text{Binom}(\theta_j, p), \quad j \in [k]. \tag{6}$$

Let $\hat{\nu} = \frac{1}{k}\sum_{j=1}^{k}\delta_{X_j}$ denote the empirical distribution of the X_j's. Then for each $m \geq 0$, we have $\hat{\nu}_m = \frac{Y_m}{k}$, where

$$Y_m = \sum_{j \in [k]} 1\{X_j = m\} \tag{7}$$

denotes the number of colors that are observed exactly m times.[1] Define the Markov kernel $P : \mathbb{Z}_+ \to \mathbb{Z}_+$ by $P(i, \cdot) = \text{Binom}(i, p)$, whose transition matrix $P = (P_{im})$ is given by

$$P_{im} = \binom{i}{m} p^m (1-p)^{i-m}, \quad i, m \geq 0. \tag{8}$$

Then as in (4), we have the unbiased relation $\mathbb{E}[\hat{\nu}] = \pi P$. Particularizing (5) with $\rho = \|\cdot\|_{\text{TV}}$ and $c(\theta) = \theta$, we obtain the following the minimum distance estimator:

$$\hat{\pi} = \operatorname*{argmin}_{\pi' \in \Pi_k} \|\pi' P - \hat{\nu}\|_{\text{TV}} \tag{9}$$

where

$$\Pi_k \triangleq \left\{ \pi' \in \mathcal{P}\{0, 1, \ldots, k\} : \sum_{m=0}^{k} m\pi'_m \leq 1 \right\}, \tag{10}$$

with $\mathcal{P}\{0, 1, \ldots, k\}$ being the set of all probability mass functions on $\{0, 1, \ldots, k\}$. As mentioned in Section 1, the true profile π belongs to Π_k. The estimator (9) is an LP with $k+1$ variables and can be solved in time that is polynomial in k. We will show that it attains the minimax upper bound in Theorem 1. As the first step, we relate the minimax risk $R(k)$ to the following LP of modulus of continuity type: for each $0 < t < 1$,

$$\delta_{\text{TV}}(t) \triangleq \sup\{\|\pi - \pi'\|_{\text{TV}} : \|\pi P - \pi' P\|_{\text{TV}} \leq t;\ \pi, \pi' \in \Pi\}, \tag{11}$$

where $\Pi \triangleq \Pi_\infty$ as in (10), that is, the set of all distributions on $\mathbb{Z}_+$ with mean at most one. The following result shows that the value of this LP characterizes the minimax risk.

Theorem 3 *There exist absolute constants C_1, C_2, d_0 such that for all $k \geq d_0$*

$$\frac{1}{72}\delta_{\text{TV}}\left(\frac{1}{6k}\right) - \frac{C_2}{\sqrt{k}} \leq R(k) \leq 2\delta_{\text{TV}}\left(\sqrt{\frac{C_1 \log k}{k}}\right), \tag{12}$$

where the upper bound is attained by the minimum distance estimator given in (9).

The proof of Theorem 3 is given in Appendix C. The main idea is as follows. By virtue of the minimum distance estimator $\hat{\pi}$ and the triangle inequality, we have:

$$\|\hat{\pi} P - \pi P\|_{\text{TV}} \leq \|\hat{\pi} P - \hat{\nu}\|_{\text{TV}} + \|\pi P - \hat{\nu}\|_{\text{TV}} \leq 2\|\pi P - \hat{\nu}\|_{\text{TV}},$$

which implies that $(\pi, \hat{\pi})$ is a feasible pair for $\delta_{\text{TV}}(t)$ with $t = 2\|\pi P - \hat{\nu}\|_{\text{TV}}$, and hence the following deterministic bound:

$$\|\hat{\pi} - \pi\|_{\text{TV}} \leq \delta_{\text{TV}}(2\|\pi P - \hat{\nu}\|_{\text{TV}}) \tag{13}$$

Recall from (4) that $\hat{\nu}$ is an unbiased estimator of πP. Furthermore, by concentration inequality one can show that with high probability that $\|\hat{\nu} - \pi P\|_{\text{TV}} = O(\sqrt{\frac{\log k}{k}})$, from which the upper bound quickly follows. The lower bound follows from that of estimating linear functionals developed in

[1] Technically, ν_0 is not directly observed from the sample. Nevertheless, one can compute it by $\hat{\nu}_0 \triangleq 1 - \sum_{m=1}^{k} \hat{\nu}_m$.

Polyanskiy and Wu (2019). Roughly speaking, we use the optimal solution (π, π') for $\delta_{\rm TV}(\Theta(1/k))$ to randomly generate two urns of size $\Theta(k)$ whose sampled version are statistically indistinguishable. With appropriate truncation argument, this can be turned into a valid minimax lower bound via Le Cam's method Tsybakov (2009).

Theorem 3 allows us to reduce the statistical problem (2) to studying the behavior of $\delta_{\rm TV}(t)$ for small t. This is characterized by the following lemma:

Lemma 4

(1) There exists absolute constant $C_3 > 0$ such that for all p, t we have

$$\delta_{\rm TV}(t) \le \min\left\{\frac{C_3}{p\log(1/t)}, 1\right\}. \tag{14}$$

(2) There exist absolute constants $C_4, t_0 > 0$ such that for any $p \in (0,1)$, $t \le t_0$,

$$\delta_{\rm TV}(t) \ge \min\left\{\frac{\bar{p}}{p}, \sqrt{\log(1/t)}\right\}\frac{C_4}{\log(1/t)}. \tag{15}$$

Combining Theorem 3 and Lemma 4 yields the main result in Theorem 1. The next two sections are devoted to the proof of Lemma 4.

Remark 5 (Reverse data processing) *Note that by the data processing inequality (DPI) of TV distance, we have $\|\pi P - \pi' P\|_{\rm TV} \le \|\pi - \pi'\|_{\rm TV}$ and hence $\delta_{\rm TV}(t) \ge t$. Therefore Lemma 4 can be understood as a* reverse DPI *for the binomial kernel P in (8). For example, if $p = \Theta(1)$, then (14) implies that (which is the best possible in view of (15)):*

$$\|\pi P - \pi' P\|_{\rm TV} \ge \exp\left\{-\Theta\left(\frac{1}{\|\pi - \pi'\|_{\rm TV}^2}\right)\right\}.$$

3. Upper bound on $\delta_{\rm TV}(t)$ by H^∞-relaxation

To bound $\delta_{\rm TV}(t)$ from above, we first relate it to the following LP

$$\delta_*(t) \triangleq \sup_{\Delta}\left\{\sum_{m=0}^{\infty}|\Delta_m| : \|\Delta P\|_1 \le t, \sum_{m=0}^{\infty} m|\Delta_m| \le 1\right\}. \tag{16}$$

The next lemma shows how the two LPs (11) and (16) are related. The proof is straightforward and deferred till Appendix D.

Lemma 6 *For all $t \in [0,1]$ we have $\frac{1}{2}(\delta_*(t) - t) \le \delta_{\rm TV}(t) \le \delta_*(t)$.*

Remark 7 *Note that our only goal is to substitute estimates on $\delta_{\rm TV}$ into (12). Therefore, due to the presence of the (unavoidable) second term in the LHS of (12), the slight difference between $\delta_*(t) - t$ and $\delta_*(t)$ in the lower bound in Lemma 6 is completely irrelevant and we can essentially think of $\delta_{\rm TV}$ and δ_* as universally within a factor of two of each other.*

Proof [Proof of upper bound in Lemma 4] We start with recalling a few facts from the complex analysis. Denote the sup-norm of a holomorphic function f over an open set $V \subset \mathbb{C}$ by $\|\cdot\|_{H_\infty(V)}$. Let $D = D_1$ be the open unit disk in $\mathbb{C}$ and denote the horodisks for $0 < p \le 1$ as

$$D_p \triangleq \bar{p} + pD = \{z \in \mathbb{C} : |z - \bar{p}| \le p\} .$$

In addition, we also define another norm for functions analytic in the neighborhood of the origin:

$$\|f\|_A \triangleq \sum_{j=0}^{\infty} |a_j|, \qquad f(z) \triangleq \sum_{j \ge 0} a_j z^j . \tag{17}$$

Since $f(re^{i\omega}) \le \sum_{n \ge 0} r^n |a_n| \le \|f\|_A$, we have

$$\|f\|_{H^\infty(D)} \le \|f\|_A . \tag{18}$$

In (Polyanskiy et al., 2017, (39)) by an application of Hadamard's three-lines theorem, it was shown that for any $q \in (0,1)$ and any holomorphic function f

$$\|f\|_{H^\infty(D_{1/2})} \le \|f\|_{H^\infty(D)}^{\frac{1-2q}{\bar{q}}} \|f\|_{H^\infty(D_q)}^{\frac{q}{\bar{q}}} . \tag{19}$$

Indeed, reparametrizing $f(z) = g(\frac{1+z}{1-z})$, we have

$$\|g\|_{H^\infty(\Re = r)} = \|f\|_{H^\infty(D_{1/(1+r)})} . \tag{20}$$

for $r \ge 0$. Then the Hadamard three-lines theorem applied to g shows that $r \mapsto \log \|f\|_{H^\infty(D_{1/(1+r)})}$ is convex, proving (19). A straightforward generalization (with a different choice of the middle line in the Hadamard theorem) shows that more generally for any $1 > q_1 > q > 0$ we have

$$\|f\|_{H^\infty(D_{q_1})} \le \|f\|_{H^\infty(D)}^{1 - \frac{q\bar{q}_1}{\bar{q}q_1}} \|f\|_{H^\infty(D_q)}^{\frac{q\bar{q}_1}{\bar{q}q_1}} . \tag{21}$$

Next, for any f holomorphic on λD for $\lambda > 0$ we have the following estimate

$$\frac{1}{\ell!} |f^{(\ell)}(0)| \le \lambda^{-\ell} \|f\|_{H^\infty(\lambda D)} . \tag{22}$$

which follows by a Cauchy integral formula: $\frac{f^{(\ell)}(0)}{\ell!} = \frac{1}{2\pi i} \oint_{|z|=\lambda} \frac{f(z)}{z^{\ell+1}} \, dz$.

With these preparations we move to the proof of (14). Consider any sequence Δ feasible for $\delta_*(t)$. For each absolutely summable sequence Δ, we consider its z-transform: $f_\Delta(z) \triangleq \sum_{m \ge 0} \Delta_m z^m$, which is a holomorphic function on the open unit disk D. Furthermore, using the definition of P in (8) and the binomial identity, it is straightforward to verify that $f_{\Delta P} = P f_\Delta$, where the Markov kernel P acts on f as a composition operator $(Pf)(z) \triangleq f(pz + \bar{p})$, where $\bar{p} \triangleq 1 - p$. Given this observation we see that the definition of $\delta_*(t)$ can also be restated as optimization over all holomorphic functions on the unit disk, cf. (17):

$$\delta_*(t) = \sup_f \left\{ \|f\|_A : \|Pf\|_A \le t, \|f'\|_A \le 1 \right\} . \tag{23}$$

For any feasible f in (23) we have that $\|f'\|_{H^\infty(D)} \le 1$ and $\|f\|_{H^\infty(D_p)} \le t$. Thus, integrating f' from some point in D_p we obtain that also $\|f\|_{H^\infty(D)} \le 1 + t \le 2$. Therefore, applying (21) to f we get $\|f\|_{H^\infty(D_{3/4})} \le 2t^{\min(\frac{p}{3\bar{p}},1)}$. Next, since $\frac{1}{2}D \subset D_{3/4}$ we have from (22)

$$|\Delta_\ell| = \frac{1}{\ell!}|f^{(\ell)}(0)| \le 2^\ell t^{\min(\frac{p}{3\bar{p}},1)} \le 2^\ell t^{p/3} . \tag{24}$$

Finally, since for any Δ feasible for $\delta_*(t)$ we have $\sum_m m|\Delta_m| \le 1$, Markov inequality implies $\sum_{m\ge J} |\Delta_m| \le \frac{1}{J}$ for any integer $J \ge 1$. Together with (24) we conclude that for any feasible Δ-sequence

$$\sum_m |\Delta_m| \le J2^J t^{\frac{p}{3}} + \frac{1}{J} \le \frac{1}{J}\left(1 + 6^J t^{p/3}\right) , \tag{25}$$

where in the last step we used $J^2 \le 3^J$. Hence, whenever $J \le \left\lfloor \frac{p \log \frac{1}{t}}{3 \log 6} \right\rfloor$, the right-hand side of (25) can be upper-bounded by $\frac{2}{J}$. This, in view of Lemma 6 completes the proof of (14) since by definition $\delta_{\mathrm{TV}} \le 1$. ■

Remark 8 *Note that functions that saturate (19) are $f(z) = e^{-m\frac{1+z}{1-z}}$ where $m \sim \log \frac{1}{t}$. Computing Taylor coefficients $[z^\ell]f(z)$ of $f(z)$ for $\ell = \Theta(m)$ can be done by applying the saddle-point method to the integral*

$$[z^\ell]f(z) = \frac{1}{2\pi i}\oint e^{-m\frac{1+z}{1-z} - (\ell+1)\log z} dz .$$

It turns out that these coefficients behave in the following way, when $\ell/m = \Theta(1)$:

$$[z^\ell]f(z) = \begin{cases} e^{-\Theta(m)}, & \ell/m < 1/2 \\ \Theta\left(\frac{1}{\sqrt{m}}\right), & \ell/m > 1/2 \end{cases}$$

This dichotomy corresponds to critical points of the function $\frac{1+z}{1-z} - \frac{\ell}{m}\log z$ leaving the unit circle when $\ell/m < 1/2$. This shows that the estimate in (25) is qualitatively tight. This effect of sudden jump in the magnitude of coefficients will be the basis of the lower bound in the next section.

4. Lower bound on $\delta_{\mathrm{TV}}(t)$

In view of Lemma 6 it suffices to consider $\delta_*(t)$ in (16). Given the equivalent definition (23), as a warm-up, let us naively replace all $\|\cdot\|_A$ norms with $\|\cdot\|_{H^\infty(D)}$. We then get the following optimization problem:

$$\delta_{H^\infty}(t) \triangleq \sup\{\|f\|_{H^\infty(D)} : \|f'\|_{H^\infty(D)} \le 1, \|f\|_{H^\infty(\bar{p}+pD)} \le t\} \tag{26}$$

Note that even though the objective function of (26) is smaller than that of $\delta_*(t)$, the feasible set is also a relaxation. Thus $\delta_{H^\infty}(t)$ does not constitute a valid lower bound to $\delta_*(t)$; nevertheless its solution, given in the following lemma, provides important insight on constructing a near-optimal solution for $\delta_*(t)$.

Lemma 9 $\delta_{H^\infty}(t) = \Theta_p\left(\frac{1}{\log(1/t)}\right)$.

Proof For the upper bound, as before we reparameterize $f(z) = g(w)$ with $w = \frac{1+z}{1-z}$. Then (20) with $r = 1/p - 1$ implies that $\|g\|_{H^\infty(\Re > \bar{p}/p)} = \|f\|_{H^\infty(\bar{p}+pD)} \le t$. By Cauchy's integral formula, we conclude that for some constant C_p (here and below possibly different on each line) we have $\|g'\|_{H^\infty(\Re > 2\bar{p}/p)} \le C_p t$.

Note that $g'(w) = \frac{2}{(1+w)^2} f'(\frac{w-1}{w+1})$. Applying (20) again with $r = 0$ yields $\|g'\|_{H^\infty(\Re>0)} \le 2$. Thus from Hadamard's three lines theorem we conclude for any $\epsilon \in (0, \bar{p}/p)$, $\|g'\|_{H^\infty(\Re=\epsilon)} \le C_p t^{\min\{\epsilon p/(2\bar{p}), 1\}}$.

Finally, for any $\omega \in \mathbb{R}$, integrating the derivative horizontally yields:

$$|g(i\omega) - g(i\omega + \bar{p}/p)| \le C_p \int_0^{\bar{p}/p} t^{\epsilon p/(2\bar{p})} d\epsilon \le C_p \frac{1}{\log \frac{1}{t}}$$

Since $|g(i\omega + \bar{p}/p)| \le \|g\|_{H^\infty(\Re = \bar{p}/p)} \le t$, we conclude that on $\{\Re = 0\}$ we have $\|g\|_{H^\infty(\Re=0)} = \|f\|_{H^\infty(D)} \le C_p \frac{1}{\log \frac{1}{t}}$, proving the upper bound part.

For the lower bound, consider the following function

$$f(z) = \frac{c_p}{\log(1/t)} (1-z)^2 t^{\frac{p}{\bar{p}} \frac{1+z}{1-z}} \tag{27}$$

for some constant $c_p > 0$. Then using (20) we have $\|f\|_{H^\infty(\bar{p}+pD)} \le \frac{4c_p}{\log(1/t)} \sup_{z \in \bar{p}+pD} |t^{\frac{p}{\bar{p}}\frac{1+z}{1-z}}| = \frac{4c_p t}{\log(1/t)}$, and

$$\begin{aligned}\|f'\|_{H^\infty(D)} &= c_p \left\| -\frac{2}{\log(1/t)}(1-z) t^{\frac{p}{\bar{p}}\frac{1+z}{1-z}} - \frac{2p}{\bar{p}} t^{\frac{p}{\bar{p}}\frac{1+z}{1-z}} \right\|_{H^\infty(D)} \\ &\le c_p \left(\frac{4}{\log(1/t)} + \frac{2p}{\bar{p}} \right) \left\| t^{\frac{p}{\bar{p}}\frac{1+z}{1-z}} \right\|_{H^\infty(D)} \overset{(20)}{=} c_p \left(\frac{4}{\log(1/t)} + \frac{2p}{\bar{p}} \right) \le \frac{2c_p(1+\bar{p})}{\bar{p}}\end{aligned}$$

where the last inequality follows from $\log(1/t) \ge 1$ for all small t. This shows f is feasible for $\delta_{H^\infty}(t)$ for small c_p. Finally noticing that $\|f\|_{H^\infty(D)} \ge |f(-1)| = \frac{c_p}{\log(1/t)}$ concludes the proof. ■

Next we modify (27) to produce a feasible solution for $\delta_*(t)$ leading to the following lower bound, which, in view of Lemma 6, provides the required bound in (15) on $\delta_{\mathrm{TV}}(t)$.

Lemma 10 *There exist absolute constants $C > 0$ and $\tilde{\beta}_0 > 0$ such that for all $t > 0$ and $p \in [0, 1)$,*

$$\delta_*(t) \ge \frac{C}{\tilde{\beta}}, \quad \tilde{\beta} \triangleq \max\left(\frac{p}{1-p} \log \frac{1}{t}, \sqrt{\frac{\log \frac{1}{t}}{1-p}} \right) \tag{28}$$

provided that $\tilde{\beta} \ge \tilde{\beta}_0$.

Proof Fix $p, t \in (0,1)$. Considering (23) our goal is to find a feasible function and bound its $\|\cdot\|_A$ norm from below. Our main tool for converting between the $\|\cdot\|_A$ norms in the definition (23) and the more convenient H^∞ norms is the following general result complementing (18): For any $r > 1$,

$$\|f\|_A \le \frac{1}{\sqrt{1 - r^{-2}}} \|f\|_{H^\infty(rD)} \,. \tag{29}$$

Indeed, let $f(z) = \sum_{n\geq 0} a_n z^n$ and let $\tilde{f}(z) = \sum_{n\geq 0} \tilde{a}_n z^n$ with $\tilde{a}_n = a_n r^n$ and thus $\tilde{f}(z) = f(rz)$. From the Plancherel identity we have

$$\sum_n |\tilde{a}_n|^2 = \frac{1}{2\pi}\int_0^{2\pi} |\tilde{f}(e^{i\omega})|^2 d\omega \leq \|\tilde{f}\|^2_{H^\infty(D)} = \|f\|^2_{H^\infty(rD)}\,.$$

Thus, (29) follows from an application of Cauchy-Schwarz inequality:

$$\sum_n |a_n| = \sum_n r^{-n}|\tilde{a}_n| \leq \sqrt{\sum_{n\geq 0} r^{-2n}}\|f\|_{H^\infty(rD)} = \frac{1}{\sqrt{1-r^{-2}}}\|f\|_{H^\infty(rD)}.$$

Next, fix some $\beta \geq \beta_0$ and $\tau \in (0,1)$, where $\beta_0 \geq 1$ is a numeric constant to be specified later, and let $\alpha = 1-\tau \in (0,1)$. Consider the function, a modified version of (27), given by

$$h(z) = \tilde{h}(\alpha z), \qquad \tilde{h}(z) = \exp\left(-\beta\frac{1+z}{1-z}\right). \tag{30}$$

Using (20), we can explicitly calculate that for any $0 < q \leq 1$:

$$\|\tilde{h}\|_{H^\infty(1-q+qD)} = e^{-\beta\frac{1-q}{q}}\,. \tag{31}$$

We will show below the following estimates (all positive numerical constants below, i.e. those that are independent of parameters p, t, β, are denoted by a common symbol C):

$$\|h\|_A \geq C\sqrt{\beta}(1-\tau)^{\frac{3\beta}{2}} \tag{32}$$

$$\|h(p\cdot+\bar{p})\|_A \leq \tau^{-\frac{1}{2}}e^{-\beta E}, \qquad E \triangleq \frac{\bar{\tau}\bar{p}}{p+\bar{p}\tau} \tag{33}$$

$$\|h'\|_A \leq 2\tau^{-\frac{3}{2}}\,. \tag{34}$$

Thus, taking $f(z) = \frac{1}{2}\tau^{\frac{3}{2}}h(z)$ in (23) proves that for all $\beta > \beta_0$ we have

$$\delta_*\left(\frac{\tau}{2}e^{-\beta E}\right) \geq C\sqrt{\beta\tau^3}(1-\tau)^{\frac{3\beta}{2}} \tag{35}$$

To show that (35) implies (28) we set $\tau = \frac{1}{\beta}$ and thus the last term in (35) can be lower bounded by $(1-1/\beta_0)^{3\beta_0/2}$ and be absorbed into C. Notice also that if $\beta \geq 2$ then $\bar{\tau} \geq 1/2$ and thus $E \geq \frac{\bar{p}}{2}\frac{1}{\frac{1}{\beta}+p}$. Since $\tau \leq 1$ and δ_* is monotone in its argument we can simplify

$$\delta_*\left(\exp\left\{-\frac{\beta}{\frac{1}{\beta}+p}\frac{\bar{p}}{2}\right\}\right) \geq \frac{C}{\beta} \tag{36}$$

Note next that for any $\mu, p > 0$, taking $x = \max(\mu p, \sqrt{\mu})$ implies $\frac{x}{\frac{1}{x}+p} \geq \frac{\mu}{2}$, which is verified by considering the two cases $\mu p \lesseqgtr \sqrt{\mu}$ separately. Then, defining $\mu \triangleq \frac{4}{\bar{p}}\log\frac{1}{t}$ and taking $\beta = \max(\mu p, \sqrt{\mu})$ ensures the argument of δ_* in (36) is at most t. In summary, we obtain the bound (28) for all $t \leq t_0$.

We proceed to proving (32)-(34). For (33) we set $r = \frac{1-\alpha\bar{p}}{\alpha p}$ in (29) and get

$$\|h(p\cdot+\bar{p})\|_A \le c\|h(p\cdot+\bar{p})\|_{H^\infty(rD)} = c\|h\|_{H^\infty(\bar{p}+prD)} = ce^{-\beta\frac{\alpha\bar{p}}{1-\alpha\bar{p}}}\,,$$

where we denoted $c = \sqrt{\frac{1}{1-r^{-2}}}$ and also applied (31) with $q = \alpha pr = 1-\alpha\bar{p}$. We next bound $c \le \left(1-r^{-1}\right)^{-1/2} = (1-\alpha\bar{p})^{1/2}\,(1-\alpha)^{-1/2} \le (1-\alpha)^{-1/2}$.

For (34) we first notice that for any function f holomorphic on r_2D we can estimate its derivative on r_1D, where $r_1 < r_2$ via Cauchy integral formula as $\|f'\|_{H^\infty(r_1D)} \le (r_2-r_1)^{-1}\,\|f\|_{H^\infty(r_2D)}$. Applying this with $f = h$, $r_1 = \frac{1+r_2}{2}$ and $r_2 = \frac{1}{\alpha}$ we get

$$\|h'\|_{H^\infty(r_2D)} \le \sqrt{2}\left(\alpha^{-1}-1\right)^{-1/2}\|h\|_{H^\infty(D/\alpha)} = \sqrt{2}\left(\alpha^{-1}-1\right)^{-1/2},$$

last step being again via (31) with $q = 1$. Applying now (29) with $r = r_2$ we obtain overall

$$\|h'\|_A \le \frac{2\alpha}{(1-\alpha)\sqrt{1-\alpha^2}} \le \frac{2}{(1-\alpha)^{3/2}}.$$

To show (32), we need to analyze the Taylor coefficients of h explicitly as the H^∞-norm bound is too weak. A natural and straightforward way is to apply the saddle-point method to study these coefficients. However, due to the special nature of h its coefficients have already been well understood. Indeed, in (Szegő, 1939, 5.1.9)) it shown that for each $x \in \mathbb{C}$ and $|v| < 1$

$$e^{-x\frac{v}{1-v}} = \sum_{n=0}^{\infty} v^n L_n^{(-1)}(x)\,, \tag{37}$$

where $L_n^{(-1)}(x)$ are *generalized Laguerre polynomial* of degree n. We will not need explicit formulae of these polynomials and only rely on their asymptotics (of Plancherel-Rotach type), cf. (Szegő, 1939, 8.22.9): For each $\epsilon > 0$ there exists a $C_\epsilon > 0$ such that for any $n \ge 0$, any $\epsilon \le \phi \le \pi/2 - \epsilon n^{-1}$, we have

$$L_n^{(-1)}(x) = e^{\frac{x}{2}}(-1)^n(\pi\sin\phi)^{-\frac{1}{2}}x^{\frac{1}{4}}n^{-\frac{3}{4}}\left\{\sin\left[n(\sin(2\phi)-2\phi)+\frac{3\pi}{4}\right]+(nx)^{-\frac{1}{2}}O_\epsilon(1)\right\} \tag{38}$$

where $x = 4n\cos^2\phi$ and the $O_\epsilon(1)$ is uniformly bounded by C_ϵ for all n and ϕ.

Comparing (37) with the definition of h we get $h(z) = e^{-\beta}\sum_{m\ge 0} L_m^{-1}(2\beta)z^m\alpha^m$. In other words, if we denote the m-th coefficient of $h(z)$ by Δ_m, then

$$\Delta_m = e^{-\beta}\alpha^m L_m^{-1}(2\beta)\,. \tag{39}$$

Due to the oscillatory nature of the Laguerre polynomial, it is not possible to bound $|\Delta_m|$ away from zero. Nevertheless, the following lemma shows that two consecutive terms cannot be simultaneously small:

Lemma 11 *For all $m \in (\beta, 3\beta/2)$ and for sufficiently large β,*

$$|\Delta_m| + |\Delta_{m+1}| \ge \alpha^{3\beta/2}\beta^{-1/2}\frac{\sqrt{2}}{6}\,. \tag{40}$$

From here (32) follows simply by $\|h\|_A \ge \sum_{\beta\le m\le 3\beta/2}|\Delta_m| \ge \alpha^{3\beta/2}\frac{\sqrt{2\beta}}{24}$. We note that the estimate (32) is tight. Indeed, applying (29) with $r = \frac{1}{\alpha}$ yields $\|h\|_A \le \frac{1}{\sqrt{1-\alpha^2}} \le 1/\sqrt{\tau}$, where we also used $\|h\|_{H^\infty(D/\alpha)} = \|\tilde{h}\|_{H^\infty(D)} = 1$ via (31) with $q = 1$. ■

References

J Acharya, H Das, A Jafarpour, A Orlitsky, and S Pan. Estimating multiple concurrent processes. In *2012 International Symposium on Information Theory Proceedings*, pages 1628–1632. IEEE, 2012.

Dragi Anevski, Richard D Gill, and Stefan Zohren. Estimating a probability mass function with unknown labels. *The Annals of Statistics*, 45(6):2708–2735, 2017.

A. Basu, H. Shioya, and C. Park. *Statistical inference: the minimum distance approach*. Chapman and Hall/CRC, Boca Raton, Florida, 2011.

R. Beran. Minimum Hellinger distance estimates for parametric models. *The annals of Statistics*, 5(3):445–463, 1977.

John Bunge and M Fitzpatrick. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.

Moses Charikar, Surajit Chaudhuri, Rajeev Motwani, and Vivek Narasayya. Towards estimation error guarantees for distinct values. In *Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pages 268–279. ACM, 2000.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory, 2nd Ed.* Wiley-Interscience, New York, NY, USA, 2006. ISBN 0471062596.

David Edelman. Estimation of the mixing distribution for a normal mean with applications to the compound decision problem. *The Annals of Statistics*, 16(4):1609–1622, 1988.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under wasserstein distance. In *Proc. 2018 Conference On Learning Theory (COLT)*, pages 3189–3221, 2018.

Yi Hao and Alon Orlitsky. The broad optimality of profile maximum likelihood. *Advances in Neural Information Processing Systems*, pages 10989–11001, 2019.

Anatoli B Juditsky and Arkadi S Nemirovski. Nonparametric estimation by convex programming. *The Annals of Statistics*, 37(5A):2278–2300, 2009.

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

A Orlitsky, NP Santhanam, K Viswanathan, and J Zhang. On estimating the probability multiset. 2008. URL http://alon.ucsd.edu/papers/pml1.pdf. draft.

Alon Orlitsky, NP Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. Convergence of profile based estimators. In *Proc. 2005 IEEE Int. Symp. Inf. Theory (ISIT)*, pages 1843–1847. IEEE, 2005.

Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

Y. Polyanskiy, A. T. Suresh, and Y. Wu. Sample complexity of population recovery. In *Proceedings of Conference on Learning Theory (COLT)*, Amsterdam, Netherland, Jul 2017. arXiv:1702.05574.

Yury Polyanskiy and Yihong Wu. Dualizing Le Cam's method, with applications to estimating the unseens. *arxiv preprint arxiv:1804.05436*, Feb 2019.

Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

Herbert Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The Regents of the University of California, 1951.

Herbert Robbins. An empirical bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1956.

Gabor Szegő. *Orthogonal polynomials*, volume 23. American Mathematical Society, 1939.

Kevin Tian, Weihao Kong, and Gregory Valiant. Learning populations of parameters. In *Advances in neural information processing systems*, pages 5778–5787, 2017.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Verlag, New York, NY, 2009.

G. Valiant. Private communication, May 2019.

Gregory Valiant. *Algorithmic Approaches to Statistical Questions*. PhD thesis, EECS Department, University of California, Berkeley, Sep 2012.

Gregory Valiant and Paul Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, pages 685–694, 2011.

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

Gregory Valiant and Paul Valiant. Instance optimal learning of discrete distributions. In *Proc. 48th Symp. on Th. of Comp. (STOC)*, pages 142–155, Cambridge, MA, USA, June 2016.

Ramya Korlakai Vinayak, Weihao Kong, Gregory Valiant, and Sham Kakade. Maximum likelihood estimation for learning populations of parameters. In *International Conference on Machine Learning*, pages 6448–6457, 2019.

Jacob Wolfowitz. The minimum distance method. *The Annals of Mathematical Statistics*, 28(1): 75–88, 1957.

Yihong Wu and Pengkun Yang. Sample complexity of the distinct element problem. *Mathematical Statistics and Learning*, 1(1):37–72, 2018.

Acknowledgments

The authors thank C. Daskalakis for pointing out Valiant and Valiant (2013) and G. Valiant for communicating Valiant (2019). S. Jana and Y. Wu were supported in part by the NSF Grant CCF-1900507, NSF CAREER award CCF-1651588, and an Alfred Sloan fellowship. Y. Polyanskiy's work was supported in part by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-09-39370, and the MIT-IBM Watson AI Lab.

Appendix A. Discussions

A.1. Comparison with previous results

In this section we review previous results on estimating sorted distribution or profile under different loss function and different sampling model. To this end, let us consider an urn with exactly k balls. Then its composition can be described by the distribution μ on $[k]$ with $\mu(x) = \theta_j/k$. When we go from μ to π we erase the "color labels" (i.e., if the balls in the urn are arranged as piles of distinct colors, going from μ to π is analogous to turning off the lights so that only the heights of each pile, but not their colors, are shown). This could have been done in a different way by sorting μ. Namely, let us define

$$\mu_i^{\downarrow} = i\text{-th largest atom of } \mu.$$

Note that π and $\mu^{\downarrow}$ can be expressed in terms of one another. In fact we have

$$\|\pi^1 - \pi^2\|_{\mathrm{TV}} \le 2\|\mu^{1\downarrow} - \mu^{2\downarrow}\|_{\mathrm{TV}} \le 2\|\mu^1 - \mu^2\|_{\mathrm{TV}} \tag{41}$$

Indeed, the second inequality follows from the fact that decreasing rearrangement minimizes the ℓ_1-distance. To prove the first inequality, note that

$$2\|\mu^{1\downarrow} - \mu^{2\downarrow}\|_{\mathrm{TV}} = \sum_j \left| \sum_{i \ge j} \pi_i^1 - \pi_i^2 \right| = W_1(\pi^1, \pi^2)\,. \tag{42}$$

where W_1 denotes the 1-Wasserstein distance between probability distributions and, in one dimension, coincides with the L_1-distance between the cumulative distribution functions (CDFs). Since π^1, π^2 are supported on $\mathbb{Z}$, the indicator function 1_E is 1-Lipschitz for any $E \subset \mathbb{Z}$ and thus $W_1(\pi^1, \pi^2) \ge \|\pi^1 - \pi^2\|_{\mathrm{TV}}$.

Can one estimate $\mu^{\downarrow}$ from the sample X? The answer is yes, in both ℓ_∞ and ℓ_1 (TV), as well as other metrics. However, to discuss these results let us move to the setting of Robbins Problem II. Namely, suppose we have $Z^M = (Z_1, \ldots, Z_M) \overset{iid}{\sim} \mu$ with μ some arbitrary distribution on $[k]$. The relevance to the Bernoulli sampling model comes from the following simple reduction: if μ is in fact the empirical distribution of colors, then given $\mathcal{N}$, which corresponds a sample of size $M' \sim \mathrm{Binom}(k, p)$ from μ without replacement, one can simulate an iid sample $Z_1, \ldots, Z_M$ with $M \approx (1 - e^{-\bar{p}})k$. Hence, any result regarding estimating $\mu^{\downarrow}$ from Z^M with $M = \Theta(k)$ implies a similar result about estimating $\mu^{\downarrow}$ from $\mathcal{N}$ with $p = \Theta(1)$.

We review several results regarding estimating $\mu^{\downarrow}$ from Z^M when μ is general. The pioneering result Orlitsky et al. (2005) only showed consistency, i.e. existence of estimator $\widehat{\mu^{\downarrow}}$ such that

$$\mathbb{E}\|\widehat{\mu^{\downarrow}} - \mu^{\downarrow}\|_{\mathrm{TV}} \to 0$$

without convergence rate. In a later draft Orlitsky et al. (2008) (see also (Anevski et al., 2017, Lemma 3) for a short proof) it was shown that simply estimating $\mu^\downarrow$ by a sorted empirical distribution achieves

$$\mathbb{E}[\|\widehat{\mu^\downarrow} - \mu^\downarrow\|_\infty] = O(k^{-\frac{1}{2}} \log k)\,.$$

A much more relevant result to us, however, is the one in Valiant and Valiant (2013). For any two π^1, π^2 they defined yet another distance:

$$D(\pi^1, \pi^2) = \inf_{\nu} \mathbb{E}[|\ln X_1 - \ln X_2|]\,, \tag{43}$$

where the infimum is over all couplings of X_1 and X_2 distributed on $\mathbb{Z}_+$ as $\mathbb{P}[X_i = j] = j\pi^i_j$ for $i \in \{1,2\}, j \in [k]$. They have shown that when $M = a\frac{k}{\log k}$ one can get

$$\mathbb{E}[D(\hat{\pi}, \pi)] \le O\left(\frac{1}{\sqrt{a}}\right),$$

which, per Valiant (2019), also holds for $a = \Theta(\log k)$. In addition (Valiant and Valiant, 2016, Appendix B) shows $W_1(\pi^1, \pi^2) \le 2D(\pi^1, \pi^2)$. Indeed, let $\boldsymbol{\nu}(\cdot,\cdot)$ be the optimal coupling in (43). Then define a coupling of π^1 to π^2 via

$$\tilde{\boldsymbol{\nu}}(j_1, j_2) = \begin{cases} \frac{1}{\max(j_1,j_2)}\boldsymbol{\nu}(j_1, j_2), & j_1 \neq 0, j_2 \neq 0 \\ \sum_{j \ge j_1} \left(\frac{1}{j_1} - \frac{1}{j}\right)\boldsymbol{\nu}(j_1, j), & j_2 = 0, j_1 > 0 \\ \sum_{j \ge j_2} \left(\frac{1}{j_2} - \frac{1}{j}\right)\boldsymbol{\nu}(j, j_2), & j_1 = 0, j_2 > 0 \end{cases}$$

and completing $j_1 = j_2 = 0$ as required. Letting $(X_1, X_2) \sim \boldsymbol{\nu}$ and $(\tilde{X}_1, \tilde{X}_2) \sim \tilde{\boldsymbol{\nu}}$ we have that

$$\mathbb{E}[|\tilde{X}_1 - \tilde{X}_2|] = 2\mathbb{E}[|\tilde{X}_1 - \tilde{X}_2|_+] = 2\mathbb{E}\left[\frac{|X_1 - X_2|}{\max(X_1, X_2)}\right] \le 2\mathbb{E}[|\ln X_1 - \ln X_2|] = 2D(\pi^1, \pi^2)\,.$$

In all, putting everything together we have that Valiant and Valiant showed that there exists an estimator of $\mu^\downarrow$ from $M = \Theta(k)$ samples such that

$$\mathbb{E}[\|\widehat{\mu^\downarrow} - \mu^\downarrow\|_{\mathrm{TV}}] = O\left(\frac{1}{\sqrt{\log k}}\right)\,. \tag{44}$$

In Han et al. (2018) it was shown that this rate is minimax optimal over all distributions supported on $[k]$. Note, however, that since the lower bound in Han et al. (2018) does not produce valid distributions on finite population (namely, μ with rational entries in $\frac{1}{k}\mathbb{Z}$), it does imply that the rate of estimating π in W_1 is $\frac{1}{\sqrt{\log k}}$, cf. (42), is sharp.

We also mention Acharya et al. (2012) and Hao and Orlitsky (2019) which discuss the use of profile maximum likelihood to estimate sorted distribution for certain sampling models. The later work deals with estimation of the sorted distribution under a truncated variant of ℓ_1 distance but it also could only achieve $\mathcal{O}\left(\frac{1}{\sqrt{\log k}}\right)$ risk bound for a sample size $\Theta(k)$.

In all, we see that following the trailblazing work Orlitsky et al. (2005) a number of works have established uniform convergence guarantees in various metrics. Relevant to us is that the best result available is $\|\hat{\pi} - \pi\|_{\mathrm{TV}} \le O\left(\frac{1}{\sqrt{\log k}}\right)$, which can obtained by first simulating samples drawn with replacement based on those without replacements, then combining (44) with (41). We show that this rate is suboptimal by a square root factor.

A.2. Open problems

For $1 \leq q \leq \infty$, let us define by $R_q(k)$ to be the minimax risk of estimating π in the ℓ_q-norm $(\sum_m |\pi_m - \hat{\pi}_m|^q)^{\frac{1}{q}}$. Then in the linear regime of $p = \Theta(1)$, Theorem 1 shows that

$$\left(\frac{1}{\log k}\right)^{2-\frac{1}{q}} \lesssim R_q(k) \lesssim \frac{1}{\log k}\,,$$

which is only tight for $q = 1$. Our complex-analytic methods seem to be especially well suited for studying the case of $q = 2$ and $q = \infty$, but we were not able to close the gap. The case of ℓ_∞ is of particularly interest as it concerns which individual profile is the hardest to estimate. Our result shows that for those colors that occur $m = \Theta(\log k)$ times, the corresponding π_m is particularly difficult and cannot be estimated better than $\Omega(\frac{1}{(\log k)^2})$. It is unclear if this is the hardest case.

Let us define by $R_{W_1}(k)$ to be the minimax risk of estimating π in the 1-Wasserstein distance $W_1(\pi, \hat{\pi})$. Given the equivalence (42), estimate (44) and lower bound $W_1(\pi, \hat{\pi}) \geq \|\pi - \hat{\pi}\|_{\mathrm{TV}}$ we get

$$\frac{1}{\log k} \lesssim R_{W_1}(k) \lesssim \frac{1}{\sqrt{\log k}}\,.$$

Due to W_1 being the L_1-distance between the CDFs, the minimax W_1 risk are also amenable to complex-analytic techniques, but so far resisted our attempts. An alternative approach is to generalize the W_1-lower bound construction of Han et al. (2018); however, as observed in previous work in the distinct elements problem Valiant (2012); Wu and Yang (2018) such moment-based construct is difficult to extend to finite population.

Appendix B. Impossibility of learning the empirical distribution

In this section we show that unless we observe all but a vanishing fraction of the urn, it is impossible to estimate the empirical distribution of the colors consistently. To this end, consider a k-ball urn and let μ denote the empirical distribution of the colors, with $\mu(j) = \frac{\theta_j}{k}, j \in [k]$. Compared to the profile π which is a distribution on $\mathbb{Z}_+$, here μ is a probability measure on the set of colors $[k]$. Similar to (2), we define the minimax TV risk for estimating μ:

$$\tilde{R}(k) = \inf_{\hat{\mu}} \sup_{\mu} \mathbb{E}[\|\mu - \hat{\mu}\|_{\mathrm{TV}}].$$

The following theorem shows that whenever the sampling ratio p is bounded away from one, it is impossible to estimate μ consistently. This observation agrees with the typical behavior in high-dimensional estimation that, absence any structural assumptions, the sample size need to exceed the number of parameters to achieve consistency.

Theorem 12

$$\tilde{R}(k) \geq \frac{k-1}{4k} h^{-1}\left(1 - p - \frac{\log_2(k+1)}{k-1}\right)$$

where $h : [0,1] \to [0,1]$ given by $h(x) = -x\log_2 x - (1-x)\log_2(1-x)$ is the binary entropy function, and h^{-1} is its inverse on $[0, \frac{1}{2}]$. Consequently, for any fixed $p < 1$, $\tilde{R}(k) = \Omega(1)$.

Proof The proof follows the mutual information method that compares the amount of information data provides and the minimum amount of information needed to reconstruct the parameters up to a certain accuracy. Consider the following Bayesian setting of a k-ball urn, where $\theta_j \overset{\text{i.i.d.}}{\sim} Ber(1/2)$ for $j = 1, \dots, k-1$ and $\theta_k = k - \sum_{j<k} \theta_j$. In other words, each of the first $k-1$ colors either is absent or appear exactly once with equal probability. Then for $j \in [k-1]$, the observed X_j is simply the erased version of θ_j with erasure probability $\bar{p}$. Thus the mutual information (in bits) between the parameters $\theta = (\theta_j : j \in [k-1])$ and the observations $X = (X_j : j \in [k])$ can be upper bounded as follows:

$$I(\theta; X) = \underbrace{I(\theta; X_1, \dots, X_{k-1})}_{=(k-1)p} + \underbrace{I(\theta; X_k | X_1, \dots, X_{k-1})}_{\le H(X_k) \le \log_2(1+k)}$$

where the inequality follows from the fact that X_k takes at most k values. On the other hand, suppose there exists $\hat{\mu} = \hat{\mu}(X)$, such that $\mathbb{E}[\|\mu - \hat{\mu}\|_{\text{TV}}] \le \epsilon$. Define $\hat{\theta}_j = \mathbf{1}_{\{\hat{\mu}_j > \frac{1}{2k}\}}$ for $j \in [k-1]$. Then $2\|\mu - \hat{\mu}\| \ge \sum_{i=1}^{k-1} \|\mu_j - \hat{\mu}_j\| \ge \frac{1}{2k} \sum_{i=1}^{k-1} \mathbf{1}_{\{\theta_j \neq \hat{\theta}_j\}}\|$. Thus $\hat{\theta}$ are close to θ in Hamming distance: $\sum_{i=1}^{k-1} \mathbb{P}[\theta_j \neq \hat{\theta}_j] \le 4\epsilon k$. By the rate-distortion function of Bernoulli distribution (Cover and Thomas, 2006, Chap. 10), their mutual information must be lower bounded by

$$I(\theta; \hat{\theta}) \ge (k-1)\left(1 - h\left(\frac{4\epsilon k}{k-1}\right)\right).$$

Combined with the data processing inequality $I(\theta; X) \ge I(\theta; \hat{\theta})$, the last two displays imply that $\epsilon \ge \frac{k-1}{4k} h^{-1}(\bar{p} - \frac{\log_2(k+1)}{k-1})$ which concludes the proof. ■

Appendix C. Proof of Theorem 3

Proof We first prove the upper bound by analyzing the minimum distance estimator (9). Let $\pi \in \Pi_k \subset \Pi$ denote the true profile. Denote the distribution $\nu \triangleq \pi P$. As outlined in Section 2 and in view of (13), the key step is to show that $\hat{\nu}$ is concentrated around ν in terms of total variation. To this end, observe that for $m \ge 1$, we have $\mathbb{E}[\hat{\nu}_m] = \nu_m$ from (4). Furthermore,

$$k \cdot \text{Var}[\hat{\nu}_m] = \frac{1}{k}\text{Var}[Y_m] = \frac{1}{k}\sum_{j \in \mathcal{X}} \text{Var}[1\{X_j = m\}] \le \frac{1}{k}\sum_{j \in \mathcal{X}} \mathbb{P}[X_j = m] = (\pi P)_m = \nu_m\,. \tag{45}$$

Thus $\mathbb{E}\left[|\hat{\nu}_m - \nu_m|\right] \le \sqrt{\nu_m/k}$. Summing over m we get

$$\begin{aligned} \mathbb{E}[\|\hat{\nu} - \nu\|_{\text{TV}}] \le \mathbb{E}\left[\sum_{m=1}^{k} |\hat{\nu}_m - \nu_m|\right] &\le \frac{1}{\sqrt{k}} \sum_{m=1}^{k} \sqrt{\nu_m} \\ &\overset{\text{(a)}}{\le} \frac{1}{\sqrt{k}} \left(\sum_{m=1}^{k} m\nu_m\right)^{1/2} \left(\sum_{m=1}^{k} \frac{1}{m}\right)^{1/2} \\ &\overset{\text{(b)}}{\le} O\left(\sqrt{\frac{\log k}{k}}\right), \end{aligned} \tag{46}$$

where (a) follows from Cauchy-Schwarz; (b) follows as follows: if we denote $U_1 \sim \pi$ and $U_2|U_1 \sim \text{Binom}(U_1, p)$, then $U_2 \sim \nu$ and hence $\mathbb{E}[U_2] = p\mathbb{E}[U_1] \leq p$ thanks to the mean constraint on $\pi \in \Pi$. Next we show that

$$\mathbb{P}\left[|\|\nu - \hat{\nu}\|_{\text{TV}} - \mathbb{E}\|\nu - \hat{\nu}\|_{\text{TV}}| \geq \epsilon\right] \leq \exp(-C_0 k \epsilon^2) \tag{47}$$

for some absolute constant C_0, all $\epsilon > 0$ and k large. For that we aim to show that $\|\nu - \hat{\nu}\|_{\text{TV}}$ satisfies the bounded difference property and then apply McDiarmid's inequality. Let $x_1, \ldots, x_{\tilde{k}}$ be the distinct colors present in the urn with $\tilde{k} \leq k$. Denote $\|\nu - \hat{\nu}\|_{\text{TV}} = d(N_{x_1}, \ldots, N_{x_{\tilde{k}}})$ for some function d. Then d satisfies the following: for any $i \in [\tilde{k}]$ and any $n_1, \ldots, n_{\tilde{k}}$ with $n_i' \neq n_i$, we have

$$\begin{aligned}
&\left|d(n_1, \ldots, n_{i-1}, n_i, n_{i+1}, \ldots, n_{\tilde{k}}) - d(n_1, \ldots, n_{i-1}, n_i', n_{i+1}, \ldots, n_{\tilde{k}})\right| \\
\leq& \frac{1}{2}\left| |\nu_{n_i} - \hat{\nu}_{n_i}| + |\nu_{n_i'} - \hat{\nu}_{n_i'}| - \left|\nu_{n_i} - \left(\hat{\nu}_{n_i} - \frac{1}{k}\right)\right| - \left|\nu_{n_i'} - \left(\hat{\nu}_{n_i'} + \frac{1}{k}\right)\right| \right| \\
\leq& \frac{1}{k}.
\end{aligned} \tag{48}$$

Furthermore, $(N_{x_1}, \ldots, N_{x_{\tilde{k}}})$ are independent. Then the desired exponential bound in (47) follows from McDiarmid's inequality.

Combining (46) and (47) we get

$$\mathbb{P}\left[\|\nu - \hat{\nu}\|_{\text{TV}} \geq \sqrt{\frac{C_1 \log k}{k}}\right] \leq k^{-1} \tag{49}$$

for some absolute constant C_1. Then taking expectations on both sides of (13), for sufficiently large k we get

$$\begin{aligned}
\mathbb{E}\|\hat{\pi} - \pi\|_{\text{TV}} &\leq \mathbb{E}[\delta_{\text{TV}}(2\|\pi P - \hat{\nu}\|_{\text{TV}})] \\
&\overset{\text{(a)}}{\leq} \delta_{\text{TV}}\left(\sqrt{\frac{C_1 \log k}{k}}\right) + k^{-1} \\
&\overset{\text{(b)}}{\leq} 2\delta_{\text{TV}}\left(\sqrt{\frac{C_1 \log k}{k}}\right),
\end{aligned}$$

where (a) follows from (49) and $\delta_{\text{TV}} \leq 1$, (b) follows from the universal fact that $\delta_{\text{TV}}(t) \geq t$ (Remark 5) and $\delta_{\text{TV}}(t)$ is increasing in t. This yields the desired upper bound on $R(k)$.

To show the lower bound, consider any bounded function $T : \mathbb{Z}_+ \to [-1, 1]$. Then for distribution π on $\mathbb{Z}_+$, define the linear functional T_π:

$$T_\pi = \sum_m \pi_m T(m).$$

Note that $2\|\hat{\pi} - \pi\|_{\text{TV}} = \sup_T |T_{\hat{\pi}} - T_\pi|$ for any estimator $\hat{\pi}$ of π. Hence the minimax TV risk of estimating π can be lower bounded by that of estimating T

$$R(k) \geq \frac{1}{2} R_T(k), \quad R_T(k) \triangleq \inf \sup \mathbb{E}\left[|\hat{T} - T_\pi|\right].$$

where the estimator $\hat{T}$ depends on $(X_j : j \in \mathcal{X})$ and the supremum is again over all k-ball urns. We are now in position to apply (Polyanskiy and Wu, 2019, Theorem 8) (with $\Theta = \mathbb{Z}_+$, $c(\theta) = \theta$, and $K_V = 1$) to obtain[2]

$$R_T(k) \geq \frac{1}{72}\delta_{\text{TV}}\left(\frac{1}{6k}\right) - \frac{C_2}{\sqrt{k}}$$

where

$$\delta_{\text{TV}}(t, T) = \sup\{|T_{\pi'} - T_\pi| : \text{TV}(\pi' P, \pi P) \leq t, \pi, \pi' \in \Pi\} \tag{50}$$

Finally optimizing over T observing that $\delta_{\text{TV}}(t) = \sup_T \delta_{\text{TV}}(t, T)$ for every $t > 0$ yields the result. ■

Appendix D. Proofs of technical lemmas

Proof [Lemma 6] We prove the lemma by showing how a feasible solution of one of the programs can be utilized to get a feasible solution of the other one, and vice-versa. Let us start with the second inequality. Given any pair (π,π') feasible for $\delta_{\text{TV}}(t)$, choose $\Delta = (\pi - \pi')/2$. We get

$$\sum_m m|\Delta_m| = \frac{1}{2}\sum_m m|\pi_m - \pi'_m| \leq \frac{1}{2}\sum_m m(\pi_m + \pi'_m) \leq 1.$$

The relation $\|\Delta P\|_1 \leq t$ follows directly from $\|\pi P - \pi' P\|_{\text{TV}} \leq t$. This shows Δ is feasible for $\delta_*(t)$ with $\|\Delta\|_1 = \|\pi - \pi'\|_{\text{TV}}$. This proves the second inequality in Lemma 6.

The first inequality is proven next. Take any non-zero feasible solution $\tilde{\Delta}$ to $\delta_*(t)$ (which exists because we can always choose $\tilde{\Delta} = 0$). Next, suppose that $\epsilon \triangleq \sum_m \tilde{\Delta}_m \neq 0$. Then, let us define $\Delta_j = \tilde{\Delta}_j$ for $j \geq 1$ and $\Delta_0 = \tilde{\Delta}_0 - \epsilon$. It is clear that

$$\sum_j \Delta_j = 0 \tag{51}$$

Furthermore, since $\langle \tilde{\Delta} P, \mathbf{1}\rangle = \langle \tilde{\Delta}, \mathbf{1}\rangle = \epsilon$ we conclude that $|\epsilon| \leq \|\Delta P\|_1 \leq t$. Therefore,

$$\sum_j |\Delta_j| \geq \sum_j |\tilde{\Delta}_j| - t\,. \tag{52}$$

Finally, because $\|rP\|_1 \leq \|r\|_1$ we also have from triangle inequality

$$\|\Delta P\|_1 \leq t + |\epsilon| \leq 2t\,. \tag{53}$$

Next we define $\Delta^+ = \max(\Delta, 0)$, $\Delta^- = \max(-\Delta, 0)$, where max is defined coordinate wise. We choose $\{\pi_m\}_{m=0}^\infty$ and $\{\pi'_m\}_{m=0}^\infty$ as

$$\pi_0 = 1 - \sum_{j=1}^\infty \Delta_j^+, \quad \pi'_0 = 1 - \sum_{j=1}^\infty \Delta_j^-,$$
$$\pi_m = \Delta_m^+, \quad \pi'_m = \Delta_m^-, \quad m \geq 1$$

[2] The result of (Polyanskiy and Wu, 2019, Theorem 8) is stated in terms of the χ^2-divergence. The TV version follows by applying (Polyanskiy and Wu, 2019, Proposition 1) to lower bound $\delta_{\chi^2}(t)$ via $\delta_{\text{TV}}(t)$.

Note that under constraints on Δ, we have $\pi, \pi' \in \Pi$. Indeed, $\sum_{m\geq 1} |\Delta_m| \leq \sum_m m|\Delta_m| = \sum_m m|\tilde{\Delta}_m| \leq 1$ and thus $\pi_0, \pi_0' \geq 0$. Furthermore, since $|\Delta_m| = \Delta_m^+ + \Delta_m^-$ we have $\sum_m m(\Delta_m^+ + \Delta_m^-) \leq 1$ which implies $\sum_m m(\pi_m + \pi_m') \leq 1$. This proves $\sum_m m\pi_m \leq 1$ and $\sum_m m\pi_m' \leq 1$. Next, observe that $\pi_0 - \pi_0' = \Delta_0$ due to (51) and thus $\pi - \pi' = \Delta$. From (53) we conclude that $\|(\pi - \pi')P\|_{\mathrm{TV}} \leq t$ and hence (π, π') is a feasible pair for $\delta_{\mathrm{TV}}(t)$. And thus via (52) we obtain

$$\delta_{\mathrm{TV}}(t) \geq \frac{1}{2}\left(\delta_*(t) - t\right) .$$

■

Proof [Lemma 11] In view of (39) and (38) the proof of (40) is straightforward but delicate. To simplify analysis we will assume $\beta \to \infty$ and denote by $o(1)$ the terms vanishing with β.

For $m \in \left(\beta, \frac{3\beta}{2}\right)$ we define $\phi_m = \arccos\sqrt{\beta/(2m)}$ and $\theta_m = F(\phi_m)$ where $F(\phi) = \sin(2\phi) - 2\phi$. Here $\phi_m \in (\arccos(1/2), \arccos(1/3))$ and hence is bounded away from both 0 and $\pi/2$ for all m in the above range. Then using (38) with $x = 2\beta$, we get that there exist absolute constants β_0, C_7 such that for all $\beta \geq \beta_0$,

$$\begin{aligned}
|L_m^{(-1)}(2\beta)| &\geq \frac{e^\beta}{\sqrt{\pi}\left(1-\frac{1}{3}\right)^{1/4}}(2\beta)^{1/4}m^{-3/4}\left\{\left|\sin\left(m\theta_m + \frac{3\pi}{4}\right)\right| - C_7\beta^{-1}\right\} \\
&\geq \frac{2e^\beta}{\sqrt{\pi}(2/3)^{1/4}3^{3/4}}\beta^{-1/2}\left\{\left|\sin\left(m\theta_m + \frac{3\pi}{4}\right)\right| - C_7\beta^{-1}\right\} \\
&\geq \frac{e^\beta\beta^{-1/2}}{2}\left\{\left|\sin\left(m\theta_m + \frac{3\pi}{4}\right)\right| - C_7\beta^{-1}\right\} . \qquad (54)
\end{aligned}$$

Now we consider any two consecutive integers m and $m+1$ in $\left(\beta, \frac{3\beta}{2}\right)$. Using (54) we get

$$\begin{aligned}
&|L_m^{(-1)}(2\beta)| + |L_{m+1}^{(-1)}(2\beta)| \\
\geq &\frac{e^\beta\beta^{-1/2}}{2}\left\{\left|\sin\left(m\theta_m + \frac{3\pi}{4}\right)\right| + \left|\sin\left((m+1)\,\theta_{m+1} + \frac{3\pi}{4}\right)\right| - 2C_7\beta^{-1}\right\} . \qquad (55)
\end{aligned}$$

The phase difference between the two sine terms comes out to be $m(\theta_m - \theta_{m+1}) - \theta_m$. Using the formula $\theta_m = F(\phi_m)$, we get

$$m(\theta_m - \theta_{m+1}) = m(\phi_m - \phi_{m+1})\frac{F(\phi_m) - F(\phi_{m+1})}{\phi_m - \phi_{m+1}}. \qquad (56)$$

We will show that the above is bounded away from 0 as m goes to infinity. We first consider the term $m(\phi_m - \phi_{m+1})$. Using $\frac{\mathrm{d}}{\mathrm{d}x} \arccos \sqrt{x} = -\frac{1}{2}\frac{1}{\sqrt{x(1-x)}}$ we deduce that

$$\begin{aligned}
m\left(\phi_m - \phi_{m+1}\right) &= m\left(\arccos\sqrt{\frac{\beta}{2m}} - \arccos\sqrt{\frac{\beta}{2m+2}}\right)\\
&= m\left(\arccos\sqrt{\beta/2m} - \arccos\sqrt{\beta/2m - \frac{\beta/2m}{m+1}}\right)\\
&= \frac{\beta}{2m}\cdot\frac{m}{m+1}\cdot\frac{\arccos\sqrt{\beta/2m} - \arccos\sqrt{\beta/2m - \frac{\beta/2m}{(m+1)}}}{\frac{\beta/2m}{(m+1)}}\\
&= -\frac{1}{2}\sqrt{\frac{\beta/2m}{1-\beta/2m}} + o(1)
\end{aligned}$$

where the $o(1)$ term goes to 0 as m, β tends to infinity with $\frac{\beta}{2m} \in \left(\frac{1}{3}, \frac{1}{2}\right)$. In view of (56) using $F'(\phi) = 2\cos(2\phi) - 2$ and $\cos^2(\phi_m) = \frac{\beta}{2m}$ we get

$$\begin{aligned}
m(\theta_m - \theta_{m+1}) &= -\frac{1}{2}\sqrt{\frac{\beta/2m}{1-\beta/2m}}F'(\phi_m) + o(1)\\
&= -2\sqrt{\frac{\beta/2m}{1-\beta/2m}}\left(\frac{\beta}{2m} - 1\right) + o(1)\\
&= 2\sqrt{\frac{\beta}{2m}\left(1 - \frac{\beta}{2m}\right)} + o(1) \qquad (57)
\end{aligned}$$

with the same last conditions on m, β. As $\frac{\beta}{2m} \in \left(\frac{1}{3}, \frac{1}{2}\right)$ the above quantity is bounded away from 0. Also (57) implies that θ_{m+1} can be approximated as $\theta_m + o(1)$. As we have

$$\begin{aligned}
\theta_m &= \sin(2\phi_m) - 2\phi_m\\
&= 2\sin\phi_m\cos\phi_m - 2\phi_m\\
&= 2\sqrt{\frac{\beta}{2m}\left(1-\frac{\beta}{2m}\right)} - 2\phi_m
\end{aligned}$$

continuing (55) and using (57) we get

$$\begin{aligned}
&|L_m^{(-1)}(2\beta)| + |L_{m+1}^{(-1)}(2\beta)|\\
\geq &\frac{e^\beta\beta^{-1/2}}{2}\left\{\left|\sin\left(m\theta_m + \frac{3\pi}{4}\right)\right| + \left|\sin\left(m\theta_m + \frac{3\pi}{4} + \theta_m - 2\sqrt{\frac{\beta}{2m}\left(1-\frac{\beta}{2m}\right)}\right)\right| + o(1)\right\}\\
= &\frac{e^\beta\beta^{-1/2}}{2}\left\{\left|\sin\left(m\theta_m + \frac{3\pi}{4}\right)\right| + \left|\sin\left(m\theta_m + \frac{3\pi}{4} - 2\phi_m\right)\right| + o(1)\right\}. \qquad (58)
\end{aligned}$$

Now we note that for any real number $a \in (0, \pi)$ the function $s(x) \triangleq |\sin(x)| + |\sin(x-a)|$ has period π and is piecewise concave on the intervals $(0, a)$ and (a, π). As $s(0) = s(a) = s(\pi) = \sin(a)$ we get

$$\inf_j \{|\sin(x)| + |\sin(x-a)|\} = \sin(a).$$

In view of the above, continuing (58) we get

$$\begin{aligned} |L_m^{(-1)}(2\beta)| + |L_{m+1}^{(-1)}(2\beta)| &\geq \frac{e^\beta \beta^{-1/2}}{2} \{\sin(2\phi_m) + o(1)\} \\ &= \frac{e^\beta \beta^{-1/2}}{2} \left\{ 2\sqrt{\frac{\beta}{2m}\left(1 - \frac{\beta}{2m}\right)} + o(1) \right\} \\ &\geq \frac{e^\beta \beta^{-1/2}}{2} \left(\frac{2\sqrt{2}}{3} + o(1) \right) \end{aligned}$$

for any $m \in \left(\beta, \frac{3\beta}{2}\right)$. In view of (39) this implies (40). ■