Learning discrete distributions with infinite support

Doron Cohen
Department of Computer Science
Ben-Gurion University of the Negev
Beer-Sheva, Israel
doronv@post.bgu.ac.il

Aryeh Kontorovich
Department of Computer Science
Ben-Gurion University of the Negev
Beer-Sheva, Israel
karyeh@cs.bgu.ac.il

Geoffrey Wolfer
Department of Computer Science
Ben-Gurion University of the Negev
Beer-Sheva, Israel
geoffrey@post.bgu.ac.il

Abstract

We present a novel approach to estimating discrete distributions with (potentially) infinite support in the total variation metric. In a departure from the established paradigm, we make no structural assumptions whatsoever on the sampling distribution. In such a setting, distribution-free risk bounds are impossible, and the best one could hope for is a fully empirical data-dependent bound. We derive precisely such bounds, and demonstrate that these are, in a well-defined sense, the best possible. Our main discovery is that the half-norm of the empirical distribution provides tight upper and lower estimates on the empirical risk. Furthermore, this quantity decays at a nearly optimal rate as a function of the true distribution. The optimality follows from a minimax result, of possible independent interest. Additional structural results are provided, including an exact Rademacher complexity calculation and apparently a first connection between the total variation risk and the missing mass.

1 Introduction

Estimating a discrete distribution in the total variation (TV) metric is a central problem in computer science and statistics (see, e.g., Han et al. [2015], Kamath et al. [2015], Orlitsky and Suresh [2015] and the references therein). The TV metric, which we use throughout the paper, is a natural and abundantly motivated choice [Devroye and Lugosi, 2001]. For support size \( d \), a sample of size \( O(d/\varepsilon^2) \) suffices for the maximum-likelihood estimator (MLE) to be \( \varepsilon \)-close (with constant probability) to the unknown target distribution. A matching lower bound is known [Anthony and Bartlett, 1999], and has been computed down to the exact constants [Kamath et al., 2015].

Classic VC theory — and, in particular, the aforementioned results — imply that for infinite support, no distribution-free sample complexity bound is possible. If \( \mu \) is the target distribution and \( \hat{\mu}_m \) is its empirical (i.e., MLE) estimate based on \( m \) iid samples, then Berend and Kontorovich [2013] showed that

\[
\frac{1}{4} \Lambda_m(\mu) - \frac{1}{4\sqrt{m}} \leq \mathbb{E}[\|\mu - \hat{\mu}_m\|_{TV}] \leq \Lambda_m(\mu), \quad m \geq 2,
\]

where

\[
\Lambda_m(\mu) = \sum_{j \in \mathbb{N}; \mu(j) < 1/m} \mu(j) + \frac{1}{2\sqrt{m}} \sum_{j \in \mathbb{N}; \mu(j) \geq 1/m} \sqrt{\mu(j)}.
\]

The quantity $\Lambda_m(\mu)$ has the advantage of always being finite and of decaying to 0 as $m \to \infty$. The bound in (1) suggests that $\Lambda_m(\mu)$, or a closely related measure, controls the sample complexity for learning discrete distributions in TV. Further supporting the foregoing intuition is the observation that for finite support size $d$ and $m \gg 1$, we have $\Lambda_m \lesssim \sqrt{d/m}$, recovering the known minimax rate. Additionally, a closely related measure turns out to control a minimax risk rate in a sense made precise in Theorem 2.5.

One shortcoming of (1) is that the lower bound only holds for the MLE, leaving the possibility that a different estimator could achieve significantly improved bounds. Another shortcoming of (1) and related estimates is that they are not empirical, in that they depend on the unknown quantity we are trying to estimate. A fully empirical bound, on the other hand, would give a high-probability estimate on $\|\mu - \hat{\mu}_m\|_{TV}$ solely in terms of observable quantities such as $\hat{\mu}_m$. Of course, such a bound should also be non-trivial, in the sense of improving with growing sample size and approaching 0 as $m \to \infty$.

A further desideratum might be something akin to instance optimality: We would like the rate at which the empirical bound decays to be “the best” possible for the given $\mu$, in an appropriate sense. Our analogue of instance optimality is inspired by, but distinct from, that of Valiant and Valiant [2016], as discussed in detail in Related work below.

**Our contributions.** We address the shortcomings of existing estimators detailed above by providing a fully empirical bound on $\|\mu - \hat{\mu}_m\|_{TV}$. Our main discovery is that the quantity $\Phi_m(\hat{\mu}_m) := \frac{1}{m} \sum_{j \in \mathbb{N}} \sqrt{\hat{\mu}_m(j)}$ satisfies all of the desiderata posed above for an empirical bound. As we show in Theorems 2.1 and 2.2, $\hat{\Phi}_m(\hat{\mu}_m)$ provides tight, high-probability upper and lower bounds on $\|\mu - \hat{\mu}_m\|_{TV}$. Further, Theorem 2.3 shows that $E[\Phi_m(\hat{\mu}_m)]$ behaves as $\Lambda_m(\mu)$ defined in (2). Finally, in the spirit of instance optimality, Theorem 2.4 shows that no other estimator-bound pair can improve upon $(\hat{\mu}_m, \Phi_m)$, other than by small constants. The latter follows from a minimax bound of independent interest, Theorem 2.5. Additional structural results are provided, including an exact Rademacher complexity calculation and a connection (apparently the first) between the total variation risk and the missing mass.

**Definitions, notation and setting.** As we are dealing with discrete distributions, there is no loss of generality in taking our sample space to be the natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$. For $k \in \mathbb{N}$, we write $[k] := \{i \in \mathbb{N} : i \leq k\}$. The set of all distributions on $\mathbb{N}$ will be denoted by $\Delta_\mathbb{N}$, which we enlarge to include the “deficient” distributions: $\Delta_\mathbb{N} \subset \Delta_\mathbb{N}^\circ := \{\mu \in [0, 1]^\mathbb{N} : \sum_{i \in \mathbb{N}} \mu(i) \leq 1\}$. For $d \in \mathbb{N}$, we write $\Delta_d \subset \Delta_\mathbb{N}$ to denote those $\mu$ whose support is contained in $[d]$.

For $\mu \in \Delta_\mathbb{N}^\circ$ and $I \subseteq \mathbb{N}$, we write $\mu(I) = \sum_{i \in I} \mu(i)$. We define the decreasing permutation of $\mu \in \Delta_\mathbb{N}^\circ$, denoted by $\mu^\downarrow$, to be the sequence $(\mu^\downarrow(i))_{i \in \mathbb{N}}$ sorted in non-increasing order, achieved by a permutation $\Pi_\mu : \mathbb{N} \to \mathbb{N}$; thus, $\mu^\downarrow(i) = \mu(\Pi_\mu(i))$. For $0 < \eta < 1$, define $T_\mu(\eta) \in \mathbb{N}$ as the least $t$ for which $\sum_{i \leq t} \mu^\downarrow(i) \leq \eta$. This induces a truncation of $\mu$, denoted by $\mu[\eta] \in \Delta_\mathbb{N}^\circ$ and defined by $\mu[\eta](i) = 1[\Pi^\downarrow(i) \leq T_\mu(\eta)] \mu(i)$.

For $\mu, \nu \in \Delta_\mathbb{N}^\circ$, we define the total variation distance in terms of the $\ell_1$ norm:

$$\|\mu - \nu\|_{TV} := \frac{1}{2} \|\mu - \nu\|_1 = \frac{1}{2} \sum_{i \in \mathbb{N}} |\mu(i) - \nu(i)|. \quad (3)$$

For $\mu \in \Delta_\mathbb{N}^\circ$, we also define the half-norm $\square$ as

$$\|\mu\|_{1/2} := \left( \sum_{i \in \mathbb{N}} \sqrt{\mu(i)} \right)^2; \quad (4)$$

note that while $\|\mu\|_{1/2}$ may be infinite, we have $\|\mu\|_{1/2} \leq \|\mu\|_0$, where the latter denotes the support size.

For $m \in \mathbb{N}$ and $\mu \in \Delta_\mathbb{N}^\circ$, we write $X = (X_1, \ldots, X_m) \sim \mu^m$ to mean that the components of the vector $X$ are drawn iid from $\mu$. We reserve $\hat{\mu}_m \in \Delta_\mathbb{N}$ for the empirical measure induced by the sample $X$, i.e. $\hat{\mu}_m(i) := \frac{1}{m} \sum_{t \in [m]} 1[X_t = i]$; the term MLE will be used interchangeably.

---

1While $\mu^\downarrow$ is uniquely defined, $\Pi_\mu^\downarrow$ is not. Uniqueness could be ensured by taking the lexicographically first permutation, but will not be needed for our results.

2The half-norm is not a proper vector-space norm, as it lacks sub-additivity.
For the class of boolean functions over the integers \( \{ f : \mathbb{N} \to \{0, 1\} \} \), which we denote by \( \{0, 1\}^\mathbb{N} \), recall the definition of the empirical Rademacher complexity \( \mathcal{R}_m(X) \) conditional on the sample \( X \):

\[
\hat{\mathcal{R}}_m(X) := \mathbb{E}_{\sigma} \left[ \sup_{f \in [0,1]^m} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(X_i) \right],
\]

(5)

where \( \sigma = (\sigma_1, \ldots, \sigma_m) \sim \text{Uniform}(\{-1, 1\}^m) \). The expectation of the above random quantity is the Rademacher complexity \( \mathcal{R}_m := \mathbb{E}_{X \sim \mu^m} \left[ \hat{\mathcal{R}}(X) \right] \).

(6)

Related work. Given the classical nature of the problem, a comprehensive literature survey is beyond our scope; the standard texts \cite{Devroye1985,Devroye2001} provide much of the requisite background. Chapter 6.5 of the latter makes a compelling case for the TV metric used in this paper, but see \cite{Waggoner2015} and the works cited therein for results on other \( \ell_p \) norms. Though surveying all of the relevant literature is a formidable task, a relatively streamlined narrative may be distilled. Conceptually, the simplest case is that of \( \|\mu\|_0 < \infty \) (i.e., finite support). Since learning a distribution over \([d]\) in TV is equivalent to agnostically learning the function class \( \{0, 1\}^d \), standard VC theory \cite{Anthony1999,Kontorovich2019} entails that the MLE achieves the minimax risk rate of \( \sqrt{d/m} \) over all \( \mu \in \Delta_N \) with \( \|\mu\|_0 \leq d \). An immediate consequence is that in order to obtain quantitative risk rates for the case of infinite support, one must assume some sort of structure \cite{Diakonikolas2016}. One can, for example, obtain minimax rates for \( \mu \) with bounded entropy \cite{Han2015}, or, say, bounded half-norm (as we do here). Alternatively, one can restrict one’s attention to a finite class \( Q \subseteq \Delta_N \); here too, optimal results are known \cite{Bousquet2019,Berend2013}. Our work departs from the paradigm of a-priori constraints on the unknown sampling distribution. Instead, our estimates hold for all \( \mu \in \Delta_N \). Of course, this must come at a price: no a-priori sample complexity bounds are possible in this setting. Absent any prior knowledge regarding \( \mu \), one can only hope for sample-dependent empirical bounds, and we indeed obtain these. Further, our empirical bounds are essentially the best possible, as formalized in Theorem 2.4. The latter result may be thought of as a learning-theoretic analogue of being instance-optimal, as introduced by \cite{Valiant2017} in the testing framework. Instance optimality is a very natural notion in the context of testing whether an unknown sampling distribution \( \mu \) is identical to or \( \varepsilon \)-far from a given reference one, \( \mu_0 \). For example, \cite{Valiant2017} discovered that a truncated \( 2/3 \)-norm of \( \mu_0 \) — i.e., a quantity closely related to \( \|\mu_0\|_{2/3} \) — controls the complexity of the testing problem in TV distance. Instance optimality is more difficult to formalize for distribution learning, since for any given \( \mu \in \Delta_N \), there is a trivial “learner” with \( \mu \) hard-coded inside. \cite{Valiant2016} defined this notion in terms of competing against an oracle who knows the distribution up to a permutation of the atoms, and did not provide empirical confidence intervals. We do derive fully empirical bounds, and further show that they are impossible to improve upon (i.e., any estimator) other than by constants. Our results suggest that the half-norm \( \|\mu\|_{1/2} \) plays a role in learning analogous to that of \( \|\mu\|_{2/3} \) in testing. As an intriguing aside, we note that the half-norm corresponds to the Tsallis \( q \)-entropy with \( q = 1/2 \), which was shown to be an optimal regularizer in some stochastic and adversarial bandit settings \cite{Zimmert2019}. We leave the question of investigating a deeper connection between the two results for future work.

2 Main results

In this section, we formally state our main results. Recall from the Definitions that the sample \( X = (X_1, \ldots, X_m) \sim \mu^m \) induces the empirical measure (MLE) \( \hat{\mu}_m \), and that a key quantity in our bounds is

\[
\Phi_m(\hat{\mu}_m) = \frac{1}{\sqrt{m}} \|\hat{\mu}_m\|_{1/2} = \frac{1}{\sqrt{m}} \sum_{j \in \mathbb{N}} \sqrt{\hat{\mu}_m(j)}.
\]

(7)

Our first result is a fully empirical, high-probability upper bound on \( \|\hat{\mu}_m - \mu\|_{TV} \) in terms of \( \Phi_m(\hat{\mu}_m) \).
Theorem 2.1. For all $m \in \mathbb{N}$, $\delta \in (0, 1)$, and $\mu \in \Delta_N$, we have that
\[
\|\hat{\mu}_m - \mu\|_{TV} \leq \Phi_m(\hat{\mu}_m) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}
\]
holds with probability at least $1 - \delta$. We also have
\[
\mathbb{E}[\|\hat{\mu}_m - \mu\|_{TV}] \leq \mathbb{E}[\Phi_m(\hat{\mu}_m)].
\]
Since $\|\hat{\mu}_m\|_{1/2} \leq \|\hat{\mu}_m\|_0 \leq \|\mu\|_0$, this recovers the minimax rate of $\sqrt{d/m}$ for $\mu \in \Delta_N$ with $\|\mu\|_0 \leq d$. We also provide a matching lower bound:

Theorem 2.2. For all $m \in \mathbb{N}$, $\delta \in (0, 1)$, and $\mu \in \Delta_N$, we have that
\[
\|\hat{\mu}_m - \mu\|_{TV} \geq \frac{1}{4\sqrt{2}} \Phi_m(\hat{\mu}_m) - 3\sqrt{\frac{\log \frac{2}{\delta}}{m}}
\]
holds with probability at least $1 - \delta$.

Our empirical measure $\Phi_m(\hat{\mu}_m)$ is never much worse than the non-empirical $\Lambda_m(\mu)$, defined in [2]:

Theorem 2.3. For all $m \in \mathbb{N}$ and $\mu \in \Delta_N$ we have
\[
\mathbb{E}[\Phi_m(\hat{\mu}_m)] \leq 2\Lambda_m(\mu)
\]
and, with probability at least $1 - \delta$,
\[
\Phi_m(\hat{\mu}_m) \leq 2\Lambda_m(\mu) + \sqrt{\log(1/\delta)/m}.
\]

Furthermore, no other estimator-bound pair $(\hat{\mu}_m, \Psi_m)$ can improve upon $(\hat{\mu}_m, \Phi_m)$, other than by a constant. This is the “instance optimality” result alluded to above:

Theorem 2.4. There exist universal constants $a, b > 0$ such that the following holds. For any estimator-bound pair $(\hat{\mu}_m, \Psi_m)$ and any continuous function $\theta : \mathbb{R}_+ \to \mathbb{R}_+$ such that
\[
\mathbb{E}[\|\hat{\mu}_m - \mu\|_{TV}] \leq \mathbb{E}[\Psi_m(\hat{\mu}_m)] \leq \theta\left(\mathbb{E}[\Phi_m(\hat{\mu}_m)]\right)
\]
holds for all $\mu \in \Delta_N$, $\theta$ necessarily verifies
\[
\inf_{0 < x < b} \frac{\theta(x)}{x} \geq \frac{1}{a}.
\]

The next result, framed in the high-probability setting, draws a direct parallel between our characterization of the learning sample complexity via the half-norm and Valiant and Valiant [2017]'s characterization of the testing sample complexity via the $2/3$-norm. The truncation is needed to ensure finiteness, since the $\|\mu\|_{1/2} = \infty$ for heavy-tailed distributions (e.g. $\mu(i) \propto 1/i^2$).

Theorem 2.5. There is a universal constant $C > 0$ such that for all $\Lambda \geq 2$ and $0 < \varepsilon, \delta < 1$, the MLE $\hat{\mu}_m$ verifies the following optimality property: For all $\mu \in \Delta_N$ with $\|\mu\|_{2\varepsilon/9}\|\mu\|_{1/2} \leq \Lambda$, we have
\[
m \geq C\varepsilon^{-2} \max\{\Lambda, \log(1/\delta)\} \implies \mathbb{P}(\|\hat{\mu}_m - \mu\|_{TV} < \varepsilon) \geq 1 - \delta.
\]

On the other hand, for any estimator $\hat{\mu}_m : \mathbb{N}^m \to \Delta_N$ there is a $\mu \in \Delta_N$ with $\max\{\|\mu\|_{1/2}, \|\mu\|_{2\varepsilon/9}\|\mu\|_{1/2}\} \leq \Lambda$ such that:
\[
m \leq C\varepsilon^{-2} \min\{\Lambda, \log(1/\delta)\} \implies \mathbb{P}(\|\hat{\mu}_m - \mu\|_{TV} \geq \varepsilon) \geq \min\{3/4, 1 - \delta\}.
\]

The above is a simplified statement chosen for brevity; a considerably refined version is stated and proved in Theorem 3.1.
3 Proofs

3.1 Proof of Theorem 2.1

The proof consists of two parts. The first is contained in Lemma 3.1, which provides a high-probability empirical upper bound, and an expectation bound, similar to Theorem 2.1 but in terms of $R_m(X)$ instead of $\Phi_m(X)$. The second part, contained in Lemma 3.2, provides an estimate of $R_m(X)$ in terms of $\Phi_m(X)$. 

Lemma 3.1. For all $m \in \mathbb{N}$, $\delta \in (0, 1)$, and $\mu \in \Delta_{\mathbb{N}}$, we have that

$$
|\hat{\mu}_m - \mu|_{TV} \leq 2R_m(X) + 3\sqrt{\log \frac{2}{\delta}}
$$

holds with probability at least $1 - \delta$. We also have,

$$
\mathbb{E}[|\hat{\mu}_m - \mu|_{TV}] \leq 2R_m.
$$

Proof. The high-probability bound from the observation,

$$
|\hat{\mu}_m - \mu|_{TV} := \sup_{A \subseteq \mathbb{N}} (\mu(A) - \hat{\mu}_m(A)) = \sup_{f \in F} \left( \mathbb{E}_{X \sim \mu} [f(X)] - \frac{1}{m} \sum_{i=1}^{m} f(X_i) \right)
$$

where $F := \{I_A | A \subseteq \mathbb{N}\} = \{0, 1\}^\mathbb{N}$, combined with [Mohri et al., 2012, Theorem 3.3], which states: Let $G$ be a family of functions from $\mathcal{Z}$ to $[0, 1]$ and let $\nu$ be a distribution supported on a subset of $\mathcal{Z}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over $Z = (Z_1, \ldots, Z_m) \sim \nu^m$, the following holds:

$$
\sup_{g \in G} \left( \mathbb{E}_{Z \sim \nu} [g(Z)] - \frac{1}{m} \sum_{i=1}^{m} g(Z_i) \right) \leq 2R_m(Z) + 3\sqrt{\log \frac{2}{\delta}}.
$$

Plugging in $F$ for $G$ and $\mu$ for $\nu$ in the above theorem completes the proof of the high-probability bound. The expectation bound (eq. (8)) follows from the observation at eq. (9) and a symmetrization argument [Mohri et al., 2012 eq. (3.8) to (3.13)].

In order to complete the proof, we apply

Lemma 3.2 (Empirical Rademacher estimates). Let $X = (X_1, \ldots, X_m)$ and let $\hat{\mu}_m$ be the empirical measure constructed from the sample $X$. Then,

$$
\frac{1}{2\sqrt{2}} \Phi_m(\hat{\mu}_m) \leq R_m(X) \leq \frac{1}{2} \Phi_m(\hat{\mu}_m).
$$

Proof. The proof is based on an argument that was also developed in [Scott and Nowak, 2006 Section 7.1, Appendix E.] in the context of histograms and dyadic decision trees, and that was credited to Gilles Blanchard.

Let $\hat{S} = \{X_i | i \in [m]\}$ be the empirical support according to the sample $X = (X_1, X_2, \ldots, X_m)$. Then,

$$
mR_m(X) = \mathbb{E}_{\sigma} \left[ \sup_{f \in \{\pm 1\}^\mathbb{N}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(X_i) \right] = \mathbb{E}_{\sigma} \left[ \sup_{A \subseteq \hat{S}} \sum_{i=1}^{m} \sigma_i I_A(X_i) \right]
$$

$$
= \sum_{x \in \hat{S}} \mathbb{E}_{\sigma} \left[ \sup_{A \subseteq \hat{S}} \sum_{i: X_i = x} \sigma_i I_A(X_i) \right] = \sum_{x \in \hat{S}} \mathbb{E}_{\sigma} \left[ \left( \sum_{i: X_i = x} \sigma_i \right)^+ \right] = \frac{1}{2} \mathbb{E}_{\sigma} \left[ \sum_{i=1}^{m} \sigma_i \right],
$$

where the last equality follows from counting $\{i : X_i = x\}$ and the symmetry of the random variable $\sum_{i=1}^{m} \sigma_i$ for all $n \in \mathbb{N}$. Now, by Khintchine’s inequality, for $0 < p < \infty$ and $x_1, x_2, \ldots, x_m \in \mathbb{C}$ we have

$$
A_p \left( \sum_{i=1}^{m} |x_i|^2 \right)^{1/2} \leq \left( \mathbb{E}_{\sigma} \left[ \left( \sum_{i=1}^{m} x_i \sigma_i \right)^p \right] \right)^{1/p} \leq B_p \left( \sum_{i=1}^{m} |x_i|^2 \right)^{1/2},
$$
where $A_p, B_p > 0$ are constants depending on $p$. Sharp values for $A_p, B_p$ were found by Haagerup [1981]. In particular, for $p = 1$ he found that $A_1 = \frac{1}{\sqrt{2}}$ and $B_1 = 1$. By using Khintchine’s inequality for each $\mathbb{E}_{\sigma}\left[\left|\sum_{i=1}^{m}\hat{\mu}_m(x)\sigma_i\right|\right]$ with these constants, we get

$$\frac{1}{\sqrt{2}}\sqrt{m\hat{\mu}_m(x)} \leq \mathbb{E}_{\sigma}\left[\sum_{i=1}^{m}\sigma_i\right] \leq \sqrt{m\hat{\mu}_m(x)},$$

and hence

$$\frac{1}{2\sqrt{2}}\sum_{x \in \mathcal{S}} \sqrt{m\hat{\mu}_m(x)} \leq m\hat{\mathcal{R}}_m(X) \leq \frac{1}{2} \sum_{x \in \mathcal{S}} \sqrt{m\hat{\mu}_m(x)}.$$

Dividing by $m$ completes the proof. \hfill \Box

Remark: We also give an exact expression for $\hat{\mathcal{R}}_m(X)$ in Lemma \ref{lem:exact_expression} and show in Corollary \ref{cor:exact_expression} with a more delicate analysis that

$$\frac{\|\hat{\mu}_m\|_{TV}^{1/2}}{\sqrt{2\pi m}} - \frac{3}{2}\sqrt{\frac{1}{2\pi} \frac{1}{m^{3/2}} \|\hat{\mu}_m\|^{-1/2}} \leq \mathcal{R}_m(X) \leq \frac{\|\hat{\mu}_m\|_{TV}^{1/2}}{\sqrt{2\pi m}} + \sqrt{\frac{1}{2\pi} \frac{1}{m^{3/2}} \|\hat{\mu}_m\|^{-1/2}}.$$

### 3.2 Proof of Theorem \ref{thm:main_result}

The proof follows from applying the lower bound of Lemma \ref{lem:lower_bound} to the following lemma:

**Lemma 3.3** (lower bound by empirical Rademacher). For all $m \in \mathbb{N}, \delta \in (0, 1)$, and $\mu \in \Delta_N$, we have that

$$\|\hat{\mu}_m - \mu\|_{TV} \geq \frac{1}{2}\mathcal{R}_m(X) - 3\sqrt{\frac{\log \frac{\delta}{\delta - 1}}{m}}$$

holds with probability at least $1 - \delta$.

**Proof.** The proof is closely based on [Wainwright 2019, Proposition 4.12], which states: Let $Y = (Y_1, \ldots, Y_m) \sim \nu^m$ for some distribution $\nu$ on $\mathcal{Z}$, let $\mathcal{G} \subseteq [-b, b]^m$ be a function class, and let $\sigma = (\sigma_1, \ldots, \sigma_m) \sim \text{Uniform}(\{-1, 1\}^m)$. Then

$$\sup_{g \in \mathcal{G}} \left|\mathbb{E}_{Y \sim \nu}[g(Y)] - \frac{1}{m}\sum_{i=1}^{m} g(Y_i)\right| \geq \frac{1}{2}\mathbb{E}_{\sigma, X}\left[\sup_{g \in \mathcal{G}} \left|\frac{1}{m}\sum_{i=1}^{m} \sigma_i g(Y_i)\right|\right] - \frac{\sup_{g \in \mathcal{G}} \mathbb{E}_{Y \sim \nu}[g(Y)]}{\sqrt{m}} - \delta \tag{10}$$

holds with probability at least $1 - e^{-\frac{m\delta^2}{2}}$. Plugging in $X$ for $Y$, $\mu$ for $\nu$, $\mathbb{N}$ for $\mathcal{Z}$, $1$ for $b$, and $\mathcal{F} := \{\mathbb{1}_A | A \subseteq \mathbb{N}\} = \{0, 1\}^\mathbb{N}$ for $\mathcal{G}$ in (10) together with observing that

$$\|\hat{\mu}_m - \mu\|_{TV} := \sup_{A \subseteq \mathbb{N}} \mathbb{E}_{\sigma, X}\left[\sup_{f \in \mathcal{F}} \left|\frac{1}{m}\sum_{i=1}^{m} \sigma_i f(X_i)\right|\right] \geq \mathcal{R}_m,$$

and

$$\mathbb{E}_{\sigma, X}\left[\sup_{f \in \mathcal{F}} \left|\frac{1}{m}\sum_{i=1}^{m} \sigma_i f(X_i)\right|\right] \geq \mathcal{R}_m,$$

followed by some algebraic manipulation we get

$$\|\hat{\mu}_m - \mu\|_{TV} \geq \frac{1}{2}\mathcal{R}_m - \frac{1}{2\sqrt{m}} - \sqrt{\frac{2\log \frac{\delta}{\delta - 1}}{m}}, \tag{11}$$

with probability at least $1 - \delta/2$. Applying McDiarmid’s inequality to the $1/m$-bounded-differences function $\mathcal{R}_m(X)$ (similar to [Mohri et al. 2012, Eq. (3.14)]) we get:

$$\frac{1}{2}\mathcal{R}_m \geq \frac{1}{2}\mathcal{R}_m(X) - \frac{1}{2}\sqrt{\frac{\log \frac{\delta}{\delta - 1}}{2m}} \tag{12}$$

6
with probability at least $1 - \delta/2$. To conclude the proof, combine (11) and (12) with the union bound to get:

$$\|\hat{\mu}_m - \mu\|_{TV} \geq \frac{1}{2} \mathcal{R}_m(X) - \frac{1}{2} \sqrt{\frac{\log \frac{2}{\delta}}{2m}} - \sqrt{\frac{2 \log \frac{2}{\delta}}{m}}$$

with probability at least $1 - \delta$, and use the fact $-\frac{1}{2\sqrt{m}} - \frac{1}{2} \sqrt{\frac{\log \frac{2}{\delta}}{2m}} - \sqrt{\frac{2 \log \frac{2}{\delta}}{m}} \geq -3 \sqrt{\frac{\log \frac{2}{\delta}}{m}}$ for all $m \in \mathbb{N}, \delta \in (0, 1)$.

**Remark 3.1.** We note that by using a more careful analysis, the constants of Theorem 2.2 can be improved to yield, under the same assumptions, $\|\hat{\mu}_m - \mu\|_{TV} \geq \frac{1}{2} \mathcal{R}_m(X) - \frac{3}{4\sqrt{m}} - \frac{3}{2} \sqrt{\frac{\log \frac{2}{\delta}}{m}}$ with probability at least $1 - \delta$.

### 3.3 Proof of Theorem 2.3

Invoking Fubini’s theorem, we write

$$\frac{1}{\sqrt{m}} \mathbb{E} \left[ \|\hat{\mu}_m\|_{1/2} \right] = \frac{1}{m} \sum_{i=1}^{\infty} \mathbb{E} \left[ \sqrt{X_{\sim \text{Bin}(m, \mu(i))}} \right].$$

Since $X \in \{0, 1, 2, \ldots\}$, we have $\sqrt{X} \leq X$ and hence $\mathbb{E} \left[ \sqrt{X} \right] \leq \mathbb{E} \left[ X \right]$. On the other hand, Jensen’s inequality implies $\mathbb{E} \left[ \sqrt{X} \right] \leq \sqrt{\mathbb{E} \left[ X \right]}$, whence

$$\frac{1}{\sqrt{m}} \mathbb{E} \left[ \|\hat{\mu}_m\|_{1/2} \right] \leq \frac{1}{m} \sum_{i=1}^{\infty} \min \{\sqrt{m\mu(i)}, m\mu(i)\} \leq \sum_{i: \mu(i) \leq 1/m} \mu(i) + \frac{1}{\sqrt{m}} \sum_{i: \mu(i) > 1/m} \sqrt{\mu(i)} \leq 2\Lambda_m(\mu). \quad (13)$$

The high-probability bound follows from applying McDiarmid’s inequality to the $2/m$-bounded-differences function: for all $\delta \in (0, 1)$, we have

$$\Phi_m(\hat{\mu}_m) \leq \mathbb{E} \left[ \Phi_m(\hat{\mu}_m) \right] + \sqrt{\log(1/\delta)/m}. \quad \Box$$

### 3.4 Statement and proof of the refined version of Theorem 2.5

**Theorem 3.1.** There is a universal constant $C > 0$ such that for all $\Lambda \geq 2$ and $0 < \varepsilon, \delta < 1$, the MLE verifies the following optimality property: For all $\mu \in \Delta_N$ with $\|\mu(2\varepsilon/9)\|_{1/2} \leq \Lambda$, if $(X_1, \ldots, X_m) \sim \mu^m$ and $m \geq C \max \{\Lambda, \ln \delta^{-1}\}$, then $\|\hat{\mu}_m - \mu\|_{TV} < \varepsilon$ holds with probability at least $1 - \delta$.

On the other hand, for all $\Lambda \geq 2$ and $0 < \varepsilon < 1$, $0 < \delta < 1$, for any estimator $\hat{\mu} : \mathbb{N}^m \to \Delta_N$ there is a $\mu \in \Delta_N$ with $\|\mu(\varepsilon/18)\|_{1/2} \leq \Lambda$ such that $\hat{\mu}$ must require at least $m \geq C \Lambda$ samples in order for $\|\hat{\mu} - \mu\|_{TV} < \varepsilon$ to hold with probability at least $3/4$, and for any estimator $\hat{\nu} : \mathbb{N}^m \to \Delta_N$ there is a $\nu \in \Delta_N$ with $\|\nu(2\varepsilon/9)\|_{1/2} \leq \Lambda$, such that $\hat{\nu}$ must require at least $m \geq C \frac{\Lambda}{\varepsilon^2} \ln \frac{1}{\delta}$ samples in order for $\|\hat{\nu} - \nu\|_{TV} < \varepsilon$ to hold with probability at least $1 - \delta$.

**Minimax risk.** For any $\Lambda \in [2, \infty), 0 < \varepsilon, \delta < 1$, we define the minimax risk

$$\mathcal{R}_m(\Lambda, \varepsilon, \delta) := \inf_{\mu, \nu : \mu, \nu \in \Delta_N} \sup_{\|\mu\|_{TV} \leq \Lambda} \mathbb{P}(\|\mu - \nu\|_{TV} > \varepsilon),$$

where the infimum is taken over all functions $\hat{\mu} : \mathbb{N}^m \to \Delta_N$, and the supremum is taken over the subset of distributions such that $\|\mu(2\varepsilon/9)\|_{1/2} < \Lambda$. 

7
Upper bound. Let $\Lambda \in [2, \infty), 0 < \epsilon, \delta < 1$, $\mu \in \Delta_N$, such that $\|\mu[2\epsilon\delta/9]\|_1 \leq \Lambda$, $m \in \mathbb{N}$, $(X_1, \ldots, X_m) \sim \mu$ and let $\hat{\mu}_m$ be the MLE. For $\eta > 0$, consider the two truncated distributions $\mu[\eta]$ and $\hat{\mu}'_m$, where we define the latter as

$$\hat{\mu}'_m(i) := \hat{\mu}_m(i)1[\mu[\eta](i) > 0], \quad i \in \mathbb{N}.$$  

By the triangle inequality, $\mathbb{P}\left(\|\hat{\mu}_m - \mu\|_{TV} > \epsilon\right) \leq \mathbb{P}\left(\mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 > \epsilon\right)$, where

$$\mathcal{E}_1 := \|\hat{\mu}_m - \hat{\mu}'_m\|_{TV}, \quad \mathcal{E}_2 := \|\hat{\mu}'_m - \mu[2\epsilon\delta/9]\|_{TV}, \quad \mathcal{E}_3 := \|\mu[2\epsilon\delta/9] - \mu\|_{TV}.$$  

By Markov’s inequality,

$$\mathbb{P}\left(\mathcal{E}_1 > \frac{\epsilon}{3}\right) \leq \frac{3}{\epsilon^2} \mathbb{E}\left[\|\hat{\mu}_m - \hat{\mu}'_m\|_{TV}\right] = \frac{3}{2\epsilon^2} \mathbb{E}\left[\sum_{i=1}^m |\hat{\mu}_m(i) - \hat{\mu}'_m(i)|\right]$$

$$= \frac{3}{2\epsilon^2} \mathbb{E}\left[\sum_{i \in \mathbb{N}} \mathbb{E}\left[\sum_{i=1}^m 1[X_i = i]\right]\right] = \frac{3}{2\epsilon^2} \mathbb{P}\left(\|\Pi_\mu(X_i) > T_\mu(\eta)\| \leq \frac{\delta}{3}\right).$$

Moreover, $\mathcal{E}_3 = \frac{1}{2} \sum_{\Delta_\eta(\mu)(i)} \mu^4(i) \leq \frac{\delta^2}{9} \leq \frac{\delta}{3}$. In order to apply the union bound, it remains to handle $\mathbb{P}\left(\mathcal{E}_2 > \epsilon/3\right)$. This is achieved in two standard steps. The first follows an argument similar to that of [Berend and Kontorovich, 2013, Lemma 5], that bounds from above the quantity in expectation using Jensen’s inequality, $\mathbb{E}\left[\mathcal{E}_2\right] \leq \frac{\|\mu[2\epsilon\delta/9]\|_{TV}^2}{\sqrt{m}} \leq \sqrt{\frac{\Lambda}{m}}$. An application of McDiarmid’s inequality controls the fluctuations around the expectation [Berend and Kontorovich, 2013, (7.5)] and concludes the proof.

\[\square\]

Sample complexity lower bound $m = \Omega\left(\frac{\log \frac{1}{\delta^2}}{\epsilon^2}\right)$. See Lemma B.1

Sample complexity lower bound $m = \Omega\left(\frac{\Lambda}{\epsilon}\right)$. Let $\epsilon \in (0, 1/16)$ and $\Lambda > 2$. First observe that $\Lambda/2 \leq 2[\Lambda/2] \leq \Lambda$, and $2[\Lambda/2] \in 2\mathbb{N}$. As a result,

$$\mathcal{R}_m(\Lambda, \epsilon, \delta) \geq \inf_{\mu} \sup_{\mu \in [2\Lambda/2], \mathbb{X} \sim \mu^m} \mathbb{P}\left(\|\hat{\mu} - \mu\|_{TV} > \epsilon\right)$$

$$\geq \inf_{\mu} \sup_{\mu \in [2\Lambda/2], \mathbb{X} \sim \mu^m} \mathbb{P}\left(\|\hat{\mu} - \mu\|_{TV} > \epsilon\right) \geq \frac{1}{2} \left(1 - \frac{mC\epsilon^2}{2\Lambda} \right) \geq \frac{1}{2} \left(1 - \frac{mC\epsilon^2}{\Lambda} \right)$$

where (i) and (ii) follow from taking the supremum over increasingly smaller sets, (iii) is Lemma B.2 invoked for $2[\Lambda/2] \in \mathbb{N}$, and $C > 0$ is a universal constant. To conclude, $m \leq \frac{\Lambda}{4\epsilon^2} \implies \mathcal{R}_m(\Lambda, \epsilon, \delta) \geq 1/4$, which yields the second lower bound.

Remark: The universal constant in the lower bound obtained by Tsybakov’s method at Lemma B.2 is suboptimal, and we give a short proof in the appendix for completeness. We refer the reader to the more involved methods of [Kamath et al., 2015] for obtaining tighter bounds.

3.5 Proof of Theorem 2.4

Let $d \in 2\mathbb{N}$ and $m \in \mathbb{N}$, and restrict the problem to $\mu \in \Delta_d$. Let $\epsilon \in (0, 1/16)$. By Lemma B.2

$$\mathcal{R}_m(d, \epsilon) := \inf_{\mu} \sup_{\mu \in \Delta_d} \mathbb{P}\left(\|\hat{\mu} - \mu\|_{TV} > \epsilon\right) \geq \frac{1}{2} \left(1 - \frac{Cm\epsilon^2}{d}\right)$$

for some $C > 0$, whence Markov’s inequality yields

$$\frac{1}{2} \left(1 - \frac{Cm\epsilon^2}{d}\right) \leq \frac{1}{\epsilon} \inf_{\mu} \sup_{\mu \in \Delta_d} \mathbb{E}\left[\|\hat{\mu} - \mu\|_{TV}\right].$$

Restrict $m \geq \frac{d}{\epsilon^2}$, with $b := \sqrt{3C/16}$ and set $\epsilon = \sqrt{\frac{d}{3Cm}}$, so that

$$\inf_{\mu} \sup_{\mu \in \Delta_d} \mathbb{E}\left[\|\hat{\mu} - \mu\|_{TV}\right] \geq \frac{1}{a} \sqrt{\frac{d}{m}}$$

where $a := \sqrt{27C}$ (15)

8
Suppose that \( \theta(\sqrt{d/m}) < \frac{1}{a} \sqrt{\frac{d}{m}} \), then by hypothesis,

\[
\inf_{\mu} \sup_{\mu \in \Delta_d} E \left[ \| \bar{\mu} - \mu \|_{TV} \right] \leq \sup_{\mu \in \Delta_d} E \left[ \Psi_m(\bar{\mu}_m) \right] \leq \sup_{\mu \in \Delta_d} \theta \left( E \left[ \Phi_m \right] \right).
\]

For \( \mu \in \Delta_d \), \( E \left[ \frac{\| \bar{\mu}_m \|_{TV}}{m^{1/2}} \right] \leq \frac{d}{m} \). It follows that

\[
\sup_{\mu \in \Delta_d} \theta \left( E \left[ \Phi_m \right] \right) \leq \theta \left( \frac{\sqrt{d}}{m} \right) < \frac{1}{a} \sqrt{\frac{d}{m}},
\]

which contradicts (15). We have therefore established, for

\[
r \in R := \left\{ \sqrt{d/m}: (m, d) \in \mathbb{N} \times \mathbb{N}, m \geq \frac{d}{b^2} \right\},
\]

the lower bound \( \theta(r) \geq r/a \). We extend the lower bound to the open interval \((0, b)\), by observing that \( R \) is dense in \((0, b)\) followed by a continuity argument. \(\square\)

**Broader Impact**

This work is of purely theoretical nature and does not present any foreseeable societal consequence.

**Acknowledgments and Disclosure of Funding**

We are thankful to Clayton Scott for the insightful conversations, and to the anonymous referees for their valuable comments. This research was partially supported by the Israel Science Foundation (grant No. 1602/19) and Google Research.

**References**


