A short note on learning discrete distributions

Clément L. Canonne*

February, 2020

Abstract

The goal of this short note is to provide simple proofs for the “folklore facts” on the sample complexity of learning a discrete probability distribution over a known domain of size $k$ to various distances $\varepsilon$, with error probability $\delta$.

Contents

1 Total variation distance
2 Hellinger distance
3 $\chi^2$ and Kullback—Leibler divergences
4 Briefly: Kolmogorov, $\ell_\infty$, and $\ell_2$ distances

Notation. For a given distance measure $d$, we write $\Phi(d, k, \varepsilon, \delta)$ for the sample complexity of learning discrete distributions over a known domain of size $k$, to accuracy $\varepsilon > 0$, with error probability $\delta \in (0, 1]$. As usual, asymptotics will be taken with regard to $k$ going to infinity, $\varepsilon$ going to 0, and $\delta$ going to 0, in that order. Without loss of generality, we hereafter assume the domain is the set $[k] := \{1, \ldots, k\}$.

1 Total variation distance

Recall that $d_{TV}(p, q) = \sup_{S \subseteq [k]} (p(S) - q(S)) = \frac{1}{2} \|p - q\|_1 \in [0, 1]$ for any $p, q \in \Delta([k])$.

Theorem 1. $\Phi(d_{TV}, k, \varepsilon, \delta) = \Theta \left( \frac{k + \log(1/\delta)}{\varepsilon^2} \right)$.

We focus here on the upper bound. The lower bound can be proven, e.g., via Assouad’s lemma (for the $k/\varepsilon^2$ term), and from the hardness of estimating the bias of a coin ($k = 2$) with high probability (for the $\log(1/\delta)/\varepsilon^2$ term).

First proof. Consider the empirical distribution $\tilde{p}$ obtained by drawing $n$ independent samples $s_1, \ldots, s_n$ from the underlying distribution $p \in \Delta([k])$:

$$\tilde{p}(i) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\{s_j = i\}}, \quad i \in [k]$$

(1)

*The latest version of this note can be found at github.com/ccanonne/probabilitydistributiontoolbox.
• First, we bound the expected total variation distance between \( \bar{p} \) and \( p \), by using \( \ell_2 \) distance as a proxy:

\[
\mathbb{E}[d_{TV}(\bar{p}, p)] = \frac{1}{2} \mathbb{E}[\|p - \bar{p}\|_1] = \frac{1}{2} \sum_{i=1}^{k} \mathbb{E}[|p(i) - \bar{p}(i)|] \leq \frac{1}{2} \sum_{i=1}^{k} \sqrt{\mathbb{E}[(p(i) - \bar{p}(i))^2]}
\]

the last inequality by Jensen. But since, for every \( i \in [k] \), \( n\bar{p}(i) \) follows a Bin\((n, p(i))\) distribution, we have \( \mathbb{E}[(p(i) - \bar{p}(i))^2] = \frac{1}{n^2} \text{Var}[n\bar{p}(i)] = \frac{1}{n^2}p(i)(1-p(i)) \), from which

\[
\mathbb{E}[d_{TV}(\bar{p}, p)] \leq \frac{1}{2\sqrt{n}} \sum_{i=1}^{k} \sqrt{p(i)} \leq \frac{1}{2} \sqrt{\frac{k}{n}}
\]

the last inequality this time by Cauchy–Schwarz. Therefore, for \( n \geq \frac{k}{\varepsilon} \) we have \( \mathbb{E}[d_{TV}(\bar{p}, p)] \leq \frac{\varepsilon}{2} \).

• Next, to convert this expected result to a high probability guarantee, we apply McDiarmid’s inequality to the random variable \( f(s_1, \ldots, s_n) := d_{TV}(\bar{p}, p) \), noting that changing any single sample cannot change its value by more than \( c := 1/n \):

\[
\mathbb{P}\left[ |f(s_1, \ldots, s_n) - \mathbb{E}[f(s_1, \ldots, s_n)]| \geq \frac{\varepsilon}{2} \right] \leq 2e^{-\frac{2\left(\frac{\varepsilon}{2}\right)^2}{2c^2}} = 2e^{-\frac{1}{4n}\varepsilon^2}
\]

and therefore as long as \( n \geq \frac{\varepsilon^2}{4c^2} \), we have \( |f(s_1, \ldots, s_n) - \mathbb{E}[f(s_1, \ldots, s_n)]| \leq \frac{\varepsilon}{2} \) with probability at least \( 1 - \delta \).

Putting it all together, we obtain that \( d_{TV}(\bar{p}, p) \leq \varepsilon \) with probability at least \( 1 - \delta \), as long as \( n \geq \max\left( \frac{k}{\varepsilon^2}, \frac{2}{\delta^2} \ln \frac{2}{\delta} \right) \). \( \square \)

**Second proof – the “fun” one.** Again, we will analyze the behavior of the empirical distribution \( \bar{p} \) over \( n \) i.i.d. samples from the unknown \( p \) (cf. (1)) – because it is simple, efficiently computable, and it works. Recalling the definition of total variation distance, note that \( d_{TV}(p, \bar{p}) > \varepsilon \) literally means there exists a subset \( S \subseteq [k] \) such that \( \bar{p}(S) > p(S) + \varepsilon \). There are \( 2^k \) such subsets, so… let us do a union bound.

Fix any \( S \subseteq [k] \). We have

\[
\bar{p}(S) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in S} 1_{\{s_j = i\}}
\]

and so, letting \( X_j := \sum_{i \in S} 1_{\{s_j = i\}} \) for \( j \in [n] \), we have \( \bar{p}(S) = \frac{1}{n} \sum_{j=1}^{n} X_j \) where the \( X_j \)'s are i.i.d. Bernoulli random variable with parameter \( p(S) \). Here comes the Chernoff bound (actually, Hoeffding, the other Chernoff):

\[
\mathbb{P}[\bar{p}(S) > p(S) + \varepsilon] = \mathbb{P}\left[ \frac{1}{n} \sum_{j=1}^{n} X_j > \mathbb{E}\left[ \frac{1}{n} \sum_{j=1}^{n} X_j \right] + \varepsilon \right] \leq e^{-2\varepsilon^2n}
\]

and therefore \( \mathbb{P}[\bar{p}(S) > p(S) + \varepsilon] \leq \frac{\delta}{2} \) for any \( n \geq \frac{k\ln 2 + \log(1/\delta)}{2\varepsilon^2} \). A union bound over these \( 2^k \) possible sets \( S \) concludes the proof:

\[
\mathbb{P}[\exists S \subseteq [k] \text{ s.t. } \bar{p}(S) > p(S) + \varepsilon] \leq 2^k \cdot \frac{\delta}{2^k} = \delta
\]

and we are done. **Badda bing badda boom**, as someone\(^1\) would say. \( \square \)

\(^1\)John Wright.
2 Hellinger distance

Recall that \( d_H(p, q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} \left( \sqrt{p(i)} - \sqrt{q(i)} \right)^2} = \frac{1}{\sqrt{2}} \| p - q \|_2 \) for any \( p, q \in \Delta([k]) \). The Hellinger distance has many nice properties: it is well-suited to manipulating product distributions, its square is subadditive, and is always within a quadratic factor of the total variation distance; see, e.g., [Can15, Appendix C.2].

**Theorem 2.** \( \Phi(d_H, k, \varepsilon, \delta) = O\left( \frac{k+\log(1/\delta)}{\varepsilon^2} \right) \).

This theorem is “highly non-trivial” to establish, however; for the sake of exposition, we will show increasingly stronger bounds, starting with the easiest to establish.

**Proposition 3 (Easy bound).** \( \Phi(d_H, k, \varepsilon, \delta) = O\left( \frac{k+\log(1/\delta)}{\varepsilon^2} \right) \), and \( \Phi(d_H, k, \varepsilon, \delta) = \Omega\left( \frac{k+\log(1/\delta)}{\varepsilon^2} \right) \).

**Proof.** This is immediate from **Theorem 9**, recalling that \( \frac{1}{2} \text{d}_{TV}^2 \leq d_H^2 \leq \text{d}_{TV} \). \( \square \)

**Proposition 4 (More involved bound).** \( \Phi(d_H, k, \varepsilon, \delta) = O\left( \frac{k+\log(1/\delta)}{\varepsilon^4} \right) \).

**Proof.** As for total variation distance, we consider the empirical distribution \( \hat{p} \) (cf. (1)) obtained by drawing \( n \) independent samples \( s_1, \ldots, s_n \) from \( p \in \Delta([k]) \).

- First, we bound the expected squared Hellinger distance between \( \hat{p} \) and \( p \); using the simple fact that
  \( d_H(p, q)^2 = 1 - \sum_{i=1}^{k} \sqrt{p(i)q(i)} \) for any \( p, q \in \Delta([k]) \),
  \[
  \mathbb{E}\left[ d_H(p, \hat{p})^2 \right] = 1 - \sum_{i=1}^{k} \sqrt{p(i) \cdot \mathbb{E}\left[ \sqrt{p(i)} \right]}
  \]

Now we would like to handle the square root inside the expectation, and of course Jensen’s inequality is in the wrong direction. However, for every nonnegative r.v. \( X \) with positive expectation, letting \( Y := X/\mathbb{E}[X] \), we have that
\[
\mathbb{E}\left[ \sqrt{X} \right] \geq \sqrt{\mathbb{E}[X]} \cdot \mathbb{E}\left[ \sqrt{Y} \right] = \sqrt{\mathbb{E}[X]} \cdot \mathbb{E}\left[ \sqrt{1 + (Y - \mathbb{E}[Y])} \right]
\]
\[
\geq \sqrt{\mathbb{E}[X]} \left( 1 + \frac{1}{2} \mathbb{E}[|Y - \mathbb{E}[Y]|] - \frac{1}{6} \mathbb{E}[\left( Y - \mathbb{E}[Y] \right)^2] \right) = \sqrt{\mathbb{E}[X]} \left( 1 - \frac{\text{Var} X}{6 \mathbb{E}[X]^2} \right)
\]
where we used the inequality \( \sqrt{1+x} \geq 1 + \frac{x}{2} - \frac{x^2}{8} \), which holds for \( x \geq 0 \). Since, for every \( i \in [k] \), \( np(i) \) follows a \( \text{Bin}(n, p(i)) \) distribution, we get
\[
\mathbb{E}\left[ d_H(p, \hat{p})^2 \right] \leq 1 - \frac{1}{\sqrt{n}} \sum_{i=1}^{k} \sqrt{p(i) \cdot np(i) \left( 1 - np(i) \frac{1 - np(i)}{6np(i)^2} \right)} \leq 1 - \sum_{i=1}^{k} \frac{p(i) \left( 1 - \frac{1}{6np(i)} \right)}{6n} = \frac{k}{6n}.
\]

Therefore, for \( n \geq \frac{k}{6} \), we have \( \mathbb{E}\left[ d_H(p, \hat{p})^2 \right] \leq \frac{k}{6n} \).

- Next, to convert this expected result to a high probability guarantee, we would like to apply McDiarmid’s inequality to the random variable \( f(s_1, \ldots, s_n) := d_H(p, \hat{p})^2 \) as in the (first) proof of **Theorem 9**; unfortunately, changing a sample can change the value by up to \( c \approx 1/\sqrt{n} \), and McDiarmid will yield only a vacuous bound.\(^3\) Instead, we will use a stronger, more involved concentration inequality:

\(^2\)And is inspired by the Taylor expansion \( \sqrt{1+x} = 1 + \frac{x}{2} - \frac{x^2}{8} + o(x^3) \): there is some intuition for it.

\(^3\)Try it: it’s a real bummer.
Theorem 5 ([BLM13, Theorem 8.6]). Let \( f : \mathcal{X}^n \to \mathbb{R} \) be a measurable function, and let \( X_1, \ldots, X_n \) be independent random variables taking values in \( \mathcal{X} \). Define \( Z := f(X_1, \ldots, X_n) \). Assume that there exist measurable functions \( c_i : \mathcal{X}^n \to [0, \infty) \) such that, for all \( x, y \in \mathcal{X}^n \),

\[
f(y) - f(x) \leq \sum_{i=1}^{n} c_i(x) \mathbb{I}_{\{x \neq y_i\}}.
\]

Then, setting \( v := \mathbb{E} \sum_{i=1}^{n} c_i(x)^2 \) and \( v_\infty := \sup_{x \in \mathcal{X}^n} \sum_{i=1}^{n} c_i(x)^2 \), we have, for all \( t > 0 \),

\[
\Pr[Z \geq \mathbb{E}[Z] + t] \leq e^{-\frac{t^2}{2v}} \quad \Pr[Z \leq \mathbb{E}[Z] - t] \leq e^{-\frac{t^2}{2v_\infty}}.
\]

For our \( f \) above, we have, for two any different \( x, y \in [k]^n \), that

\[
f(y) - f(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{k} \sqrt{p(i)} \left( \sqrt{\sum_{j=1}^{n} \mathbb{I}_{\{x_j = i\}}} - \sqrt{\sum_{j=1}^{n} \mathbb{I}_{\{y_j = i\}}} \right)
\]

\[
\leq \frac{1}{\sqrt{n}} \sum_{i=1}^{k} \sqrt{p(i)} \left( \sum_{j=1}^{n} \mathbb{I}_{\{x_j = i\}} - \sum_{j=1}^{n} \mathbb{I}_{\{y_j = i\}} \right)
\]

\[
\leq \frac{1}{\sqrt{n}} \sum_{i=1}^{k} \sqrt{p(i)} \sum_{j=1}^{n} \mathbb{I}_{\{x_j = i\}} \mathbb{I}_{\{y_j \neq i\}} = \sum_{j=1}^{n} \left( \frac{\sum_{j=1}^{n} \mathbb{I}_{\{x_j = i\}} \mathbb{I}_{\{y_j \neq i\}}}{c_j(x)} \right) \mathbb{I}_{\{x_j \neq y_j\}}.
\]

In view of Theorem 5, we then must evaluate

\[
v := \sum_{j=1}^{n} \mathbb{E}[c_j(X)^2] = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{k} p(i)^2 \cdot \mathbb{E} \left[ \frac{1}{1 + \sum_{\ell \neq j} \mathbb{I}_{\{X_i = i\}}} \right]
\]

where that last expectation is over \((x_\ell)_{\ell \neq j}\) drawn from \( p^{\otimes (n-1)} \). Since \( \sum_{\ell \neq j} \mathbb{I}_{\{X_i = i\}} \) is Binomially distributed with parameters \( n - 1 \) and \( p(i) \), we can use the simple fact that, for \( N \sim \text{Bin}(r, \rho) \),

\[
\mathbb{E} \left[ \frac{1}{N + 1} \right] = \frac{1 - (1 - \rho)^{r+1}}{\rho(r + 1)} \leq \frac{1}{\rho(r + 1)}
\]

to conclude that \( v \leq \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{k} p(i) = \frac{1}{n} \). By Theorem 5, we obtain

\[
\Pr \left[ \left| f(s_1, \ldots, s_n) - \mathbb{E}[f(s_1, \ldots, s_n)] \right| \geq \frac{\varepsilon^2}{2} \right] \leq e^{-\frac{1}{2}n \varepsilon^4}
\]

and therefore, as long as \( n \geq \frac{8}{\varepsilon^2} \ln \frac{1}{\delta} \), we have \( \left| f(s_1, \ldots, s_n) - \mathbb{E}[f(s_1, \ldots, s_n)] \right| \leq \frac{\varepsilon^2}{2} \) with probability at least \( 1 - \delta \).

Putting it all together, we obtain that \( d_H(p, \hat{p})^2 \leq \varepsilon^2 \) with probability at least \( 1 - \delta \), as long as \( n \geq \max \left( \frac{k}{\varepsilon^2}, \frac{8}{\varepsilon^2} \ln \frac{1}{\delta} \right) \).

We finally get to the final, optimal bound:

Proof of Theorem 2. We will rely on a recent – and quite involved – result due to Agrawal [Agr19], analyzing the concentration of the empirical distribution \( \hat{p} \) in terms of its Kullback–Leibler (KL) divergence with regard to the true \( p \),

\[
\text{KL}(\hat{p} \mid \mid p) = \sum_{i=1}^{k} \hat{p}(i) \ln \frac{\hat{p}(i)}{p(i)} \in [0, \infty].
\]

Observing that \( d_H(p, q)^2 \leq \frac{1}{2} \text{KL}(p \mid \mid q) \) for any distributions \( p, q \), the aforementioned result is actually stronger than what we need:
\textbf{Theorem 2} \cite{KOPS15, Agr19} \textit{Agr19}\textit{.} \textit{Suppose }n \geq \frac{k-1}{\alpha}. \textit{Then}

\[
\Pr[\text{KL}(\hat{p} \| p) \geq \alpha] \leq e^{-n\alpha \left(\frac{e\alpha n}{k-1}\right)^{k-1}}.
\]

In view of the above relation between Hellinger and KL, we will apply this convergence result with \(\alpha := 2\varepsilon^2\), obtaining

\[
\Pr[\text{d}_H(\hat{p}, p) \geq \varepsilon] \leq e^{-2n\varepsilon^2 + (k-1)\ln \frac{2n\varepsilon^2}{k}}.
\]

\textbf{Fact 7.} \textit{For }n \geq \frac{15k}{2k-2}, \textit{we have } (k-1)\ln \frac{2n\varepsilon^2}{k} \leq n\varepsilon^2.\]

\textit{Proof.} The conclusion is equivalent to \(2e \cdot \ln \frac{2n\varepsilon^2}{k} \leq 2n\varepsilon^2 + (k-1)\ln \frac{2n\varepsilon^2}{k}\), and thus follows from the fact that \(x \geq 2e \ln x\) for \(x \geq 15\). \(\square\)

This fact implies that, for \(n \geq \frac{15k}{2k-2}\), \(\Pr[\text{d}_H(\hat{p}, p) \geq \varepsilon] \leq e^{-n\varepsilon^2}\). Overall, we obtain that \(\text{d}_H(p, \hat{p}) \leq \varepsilon\) with probability at least \(1 - \delta\) as long as \(n \geq \max\left(\frac{15k}{2k-2}, \frac{1}{\varepsilon^2} \ln \frac{1}{\delta}\right)\), as desired. \(\square\)

\section{\(\chi^2\) and Kullback—Leibler divergences}

In view of the previous section, some remarks on Kullback--Leibler (KL) and chi-squared (\(\chi^2\)) divergences. Recall their definition, for \(p, q \in \Delta([k])\),

\[
\text{KL}(p \| q) = \sum_{i=1}^{k} p(i) \ln \frac{p(i)}{q(i)}, \quad \chi^2(p \| q) = \sum_{i=1}^{k} \frac{(p(i) - q(i))^2}{q(i)}
\]

both taking values in \([0, \infty]\); as well as the chain of (easily checked) inequalities

\[
2d_{TV}(p, q)^2 \leq \text{KL}(p \| q) \leq \chi^2(p \| q),
\]

where the first one is Pinsker’s. Importantly, KL and \(\chi^2\) divergences are unbounded and asymmetric, so the order of \(p\) and \(q\) matters \textit{a lot}: for instance, it is easy to show that, without strong assumptions on the unknown distribution \(p \in \Delta([k])\), the empirical estimator \(\hat{p}\) cannot achieve \(\text{KL}(p \| \hat{p}) < \infty\) (resp., \(\chi^2(p \| \hat{p}) < \infty\)) with any finite number of samples.\(^4\) So, that’s uplifting. (On the other hand, \textit{other} estimators than the empirical one, e.g., add-constant estimators, do provide good learning guarantees for those distance measures: see for instance \cite{KOPS15}).

We are going to focus here on getting \(\text{KL}(\hat{p} \| p)\) and \(\chi^2(\hat{p} \| p)\) down to \(\varepsilon\). Of course, in view of the inequalities above, the latter is at least as hard as the former, and a lower bound on both follows from that on \(d_{TV}\): \(\Omega\left((k + \log(1/\delta))/\varepsilon^2\right)\). And, behold! The result of Agrawal \cite{Agr19} used in the proof of \textbf{Theorem 2} does provide the optimal upper bound on learning in KL divergence – and it is achieved by the usual suspect, the empirical estimator:

\textbf{Theorem 8.} \(\Phi(\text{KL}, k, \varepsilon, \delta) \leq \Theta\left(\frac{k + \log(1/\delta)}{\varepsilon}\right), \text{where by KL we refer to minimizing } \text{KL}(\hat{p} \| p).\]

The optimal sample complexity of learning in \(\chi^2\) as a function of \(k, \varepsilon, \delta\), however, remains open.

\(^4\)You can verify this: intuitively, the issue boils down to having to non-trivially learn even the elements of the support of \(p\) that have arbitrarily small probability.
4 Briefly: Kolmogorov, $\ell_\infty$, and $\ell_2$ distances

To conclude, let us briefly discuss three other distance measures: Kolmogorov (a.k.a., “$\ell_\infty$ between cumulative distribution functions”), $\ell_\infty$, and $\ell_2$:

$$d_K(p, q) = \max_{i \in [k]} \left| \sum_{j=1}^{i} p(i) - \sum_{j=1}^{i} q(i) \right|$$

and

$$\ell_2(p, q) = \|p - q\|_2 = \sqrt{\sum_{i=1}^{k} (p(i) - q(i))^2}, \quad \ell_\infty(p, q) = \|p - q\|_\infty = \max_{i \in [k]} |p(i) - q(i)|.$$

A few remarks first. The Kolmogorov distance is actually defined for any distribution on $\mathbb{R}$, not necessarily discrete; one can equivalently define it as $d_K(p, q) = \sup_{f \in C} (\mathbb{E}_p[I_{f^{-1}(\infty, a)}] - \mathbb{E}_q[I_{f^{-1}(\infty, a)}])$. This has a nice interpretation: recalling the definition of TV distance, both are of the form $\sup_{f \in C} (\mathbb{E}_p[f] - \mathbb{E}_q[f])$ where $C$ is a class of measurable functions.\(^5\) For TV distance, $C$ is the class of indicators of all measurable subsets; for Kolmogorov, this is the (smaller) class of indicators of intervals of the form $(-\infty, a]$. (For Wasserstein/EMD distance, this will be the class of continuous, 1-Lipschitz functions.)

Second, because of the above, and also monotonicity of $\ell_p$ norms, Cauchy–Schwarz, the fact that $\ell_1(p, q) = 2d_{TV}(p, q)$, and elementary manipulations, we have

$$\ell_2(p, q) \leq 2d_{TV}(p, q) \leq \sqrt{k} \ell_2(p, q)$$

$$\ell_\infty(p, q) \leq \ell_2(p, q) \leq \sqrt{\ell_\infty(p, q)},$$

$$\frac{1}{2} \ell_\infty(p, q) \leq d_K(p, q) \leq d_{TV}(p, q).$$

That can be useful sometimes. Now, I will only briefly sketch the proof of the next theorem: the lower bounds follow from the simple case $k = 2$ (estimating the bias of a biased coin), the upper bounds are achieved by the empirical estimator (again). Importantly, the result for Kolmogorov distance still applies to continuous, arbitrary distributions.

**Theorem 9.** $\Phi(d_K, k, \varepsilon, \delta), \Phi(\ell_\infty, k, \varepsilon, \delta), \Phi(\ell_2, k, \varepsilon, \delta) = \Theta \left( \frac{\log(1/\delta)}{\varepsilon^2} \right)$, independent of $k$.

**Sketch.** The proof for Kolmogorov distance is the most involved, and follows from a very useful and non-elementary theorem due to Dvoretzky, Kiefer, and Wolfowitz from 1956 [DKW56] (with the optimal constant due to Massart, in 1990 [Mas90]):

**Theorem 10 (DKW Inequality).** Let $\hat{p}$ denote the empirical distribution on $n$ i.i.d. samples from $p$ (an arbitrary distribution on $\mathbb{R}$). Then, for every $\varepsilon > 0$,

$$\Pr\{d_K(\hat{p}, p) > \varepsilon\} \leq 2e^{-2n\varepsilon^2}.$$

Note, again, that this holds even if $p$ is a continuous (or arbitrary) distribution on an unbounded support.

The proof for $\ell_\infty$ just follows the Kolmogorov upper bound and the aforementioned inequality $\ell_\infty(p, q) \leq 2d_K(p, q)$ (which hinges on the fact that $p(i) = \sum_{j=1}^{i} p(i) - \sum_{j=1}^{i-1} p(i)$ and the triangle inequality). Finally, the proof for $\ell_2$ is a nice exercise involving analyzing the expectation of the $\ell_2$ distance achieved by the empirical estimator, and McDiarmid’s inequality.

**Acknowledgments.** I would like to thank Gautam Kamath and John Wright for suggesting “someone should write this up as a note,” and to Jiantao Jiao for discussions about the Hellinger case.

---

\(^5\)Such metrics on the space of probability distributions are called integral probability metrics.
References

[Agr19] Rohit Agrawal. Multinomial concentration in relative entropy at the ratio of alphabet and sample sizes. *CoRR*, abs/1904.02291, 2019. 6, 3

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013. 5

[Can15] Clément L. Canonne. A Survey on Distribution Testing: Your Data is Big. But is it Blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, 2015. 2

[DKW56] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27:642–669, 1956. 4

[KOPS15] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1066–1100. PMLR, 2015. 3

[Mas90] Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283, 1990. 4