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 total variation distance ε, with error probability δ. Thanks to 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.

For a given distance measure d, we write $\Phi(\mathrm{d}, k, \varepsilon, \delta)$ 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, ε going to 0, and δ going to 0, in that order. Without loss of generality, we hereafter assume the domain is the set $[k] \stackrel{\text{def}}{=} \{1, \dots, k\}$.

1 Total variation distance

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

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

We focus here on the upper bound. The lower bound can be proven, e.g., via Assouad's lemma (for the k/ε^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 $s_1, \dots, s_n$ from the underlying distribution $p \in \Delta([k])$:

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

- First, we bound the *expected* total variation distance between $\tilde{p}$ and p, by using ℓ_2 distance as a proxy:

$$\mathbb{E}[\mathrm{d}_{\mathrm{TV}}(p, \tilde{p})] = \frac{1}{2}\mathbb{E}[\|p - \tilde{p}\|_1] = \frac{1}{2}\sum_{i=1}^{k}\mathbb{E}[|p(i) - \tilde{p}(i)|] \le \frac{1}{2}\sum_{i=1}^{k}\sqrt{\mathbb{E}[(p(i) - \tilde{p}(i))^2]}$$

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

$$\mathbb{E}[\mathrm{d}_{\mathrm{TV}}(p, \tilde{p})] \le \frac{1}{2\sqrt{n}}\sum_{i=1}^{k}\sqrt{p(i)} \le \frac{1}{2}\sqrt{\frac{k}{n}}$$

  the last inequality this time by Cauchy–Schwarz. Therefore, for $n \ge \frac{k}{\varepsilon^2}$ we have $\mathbb{E}[\mathrm{d}_{\mathrm{TV}}(p, \tilde{p})] \le \frac{\varepsilon}{2}$.
- Next, to convert this expected result to a *high probability* guarantee, we apply McDiarmid's inequality to the random variable $f(s_1, \dots, s_n) \stackrel{\text{def}}{=} \mathrm{d}_{\mathrm{TV}}(p, \tilde{p})$, noting that changing any single sample cannot change its value by more than $c \stackrel{\text{def}}{=} 1/n$:

$$\Pr\Big[\, |f(s_1, \dots, s_n) - \mathbb{E}[f(s_1, \dots, s_n)]| \ge \frac{\varepsilon}{2} \,\Big] \le 2e^{-\frac{2\left(\frac{\varepsilon}{2}\right)^2}{nc^2}} = 2e^{-\frac{1}{2}n\varepsilon^2}$$

  and therefore as long as $n \ge \frac{2}{\varepsilon^2}\ln\frac{2}{\delta}$, we have $|f(s_1, \dots, s_n) - \mathbb{E}[f(s_1, \dots, s_n)]| \le \frac{\varepsilon}{2}$ with probability at least $1 - \delta$.

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

Second proof – the "fun" one. Again, we will analyze the behavior of the empirical distribution $\tilde{p}$ 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 $\mathrm{d_{TV}}(p,\tilde{p}) > \varepsilon$ literally means there exists a subset $S \subseteq [k]$ such that $\tilde{p}(S) > p(S) + \varepsilon$. There are 2^k such subsets, so... let us do a union bound.

Fix any $S \subseteq [k]$. We have

$$\tilde{p}(S) = \tilde{p}(i) \stackrel{(1)}{=} \frac{1}{n}\sum_{i\in S}\sum_{j=1}^{n} \mathbb{1}_{\{s_j=i\}}$$

and so, letting $X_j \stackrel{\text{def}}{=} \sum_{i\in S}\mathbb{1}_{\{s_j=i\}}$ for $j \in [n]$, we have $\tilde{p}(S) = \frac{1}{n}\sum_{j=1}^n X_j$ 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):

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

and therefore $\Pr[\,\tilde{p}(S) > p(S)+\varepsilon\,] \leq \frac{\delta}{2^k}$ 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:

$$\Pr[\,\exists S\subseteq[k] \text{ s.t. } \tilde{p}(S) > p(S)+\varepsilon\,] \leq 2^k\cdot\frac{\delta}{2^k} = \delta$$

and we are done. *Badda bing badda boom*, as someone[1] would say. ☐

2 Hellinger distance

Recall that $\mathrm{d_H}(p,q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^k(\sqrt{p(i)}-\sqrt{q(i)})^2} = \frac{1}{\sqrt{2}}\|p-q\|_2 \in [0,1]$ 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. $\Phi(\mathrm{d_H},k,\varepsilon,\delta) = \Theta\Big(\frac{k+\log(1/\delta)}{\varepsilon^2}\Big)$.

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)**.** $\Phi(\mathrm{d_H},k,\varepsilon,\delta) = O\Big(\frac{k+\log(1/\delta)}{\varepsilon^4}\Big)$*, and* $\Phi(\mathrm{d_H},k,\varepsilon,\delta) = \Omega\Big(\frac{k+\log(1/\delta)}{\varepsilon^2}\Big)$.

Proof. This is immediate from Theorem 1, recalling that $\frac{1}{2}\,\mathrm{d_{TV}^2} \leq \mathrm{d_H}^2 \leq \mathrm{d_{TV}}$. ☐

Proposition 4 (More involved bound)**.** $\Phi(\mathrm{d_H},k,\varepsilon,\delta) = O\Big(\frac{k}{\varepsilon^2}+\frac{\log(1/\delta)}{\varepsilon^4}\Big)$.

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

- First, we bound the *expected* squared Hellinger distance between $\widehat{p}$ and p: using the simple fact that $\mathrm{d_H}(p,q)^2 = 1-\sum_{i=1}^k\sqrt{p(i)q(i)}$ for any $p,q\in\Delta([k])$,

$$\mathbb{E}\Big[\mathrm{d_H}(p,\widehat{p})^2\Big] = 1-\sum_{i=1}^k\sqrt{p(i)}\cdot\mathbb{E}\Big[\sqrt{\widehat{p}(i)}\Big]\ .$$

[1] John Wright.

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 $Y \stackrel{\text{def}}{=} X/\mathbb{E}[X]$, we have that

$$\begin{aligned}\mathbb{E}\Big[\sqrt{X}\Big] &= \sqrt{\mathbb{E}[X]}\cdot\mathbb{E}\Big[\sqrt{Y}\Big] = \sqrt{\mathbb{E}[X]}\cdot\mathbb{E}\Big[\sqrt{1+(Y-\mathbb{E}[Y]))}\Big] \\ &\geq \sqrt{\mathbb{E}[X]}\left(1+\frac{1}{2}\mathbb{E}[Y-\mathbb{E}[Y]]-\frac{1}{6}\mathbb{E}\big[(Y-\mathbb{E}[Y])^2\big]\right) = \sqrt{\mathbb{E}[X]}\left(1-\frac{\operatorname{Var}X}{6\mathbb{E}[X]^2}\right)\end{aligned}$$

where we used the inequality $\sqrt{1+x}\geq 1+\frac{x}{2}-\frac{x^2}{6}$, which holds for $x\geq 0$.[2] Since, for every $i\in[k]$, $n\widehat{p}(i)$ follows a $\operatorname{Bin}(n,p(i))$ distribution, we get

$$\mathbb{E}\Big[\mathrm{d_H}(p,\widehat{p})^2\Big] \leq 1-\frac{1}{\sqrt{n}}\sum_{i=1}^k\sqrt{p(i)}\cdot\sqrt{np(i)}\left(1-\frac{np(i)(1-np(i))}{6n^2p(i)^2}\right) \leq 1-\sum_{i=1}^k p(i)\left(1-\frac{1}{6np(i)}\right) = \frac{k}{6n}\,.$$

Therefore, for $n\geq\frac{k}{3\varepsilon^2}$, we have $\mathbb{E}\Big[\mathrm{d_H}(p,\widehat{p})^2\Big]\leq\frac{\varepsilon^2}{2}$.

- Next, to convert this expected result to a high probability guarantee, we *would like* to apply McDiarmid's inequality to the random variable $f(s_1,\dots,s_n)\stackrel{\text{def}}{=}\mathrm{d_H}(p,\widehat{p})^2$ as in the (first) proof of Theorem 1; 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:

  Theorem 5 ([BLM13, Theorem 8.6])**.** *Let $f\colon\mathcal{X}^n\to\mathbb{R}$ be a measurable function, and let $X_1,\dots,X_n$ be independent random variables taking values in $\mathcal{X}$. Define $Z\stackrel{\text{def}}{=}f(X_1,\dots,X_n)$. Assume that there exist measurable functions $c_i\colon\mathcal{X}^n\to[0,\infty)$ such that, for all $x,y\in\mathcal{X}^n$,*

  $$f(y)-f(x)\leq\sum_{i=1}^n c_i(x)\mathbb{1}_{\{x_i\neq y_i\}}\,.$$

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

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

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

  $$\begin{aligned}f(y)-f(x) &= \frac{1}{\sqrt{n}}\sum_{i=1}^k\sqrt{p(i)}\left(\sqrt{\sum_{j=1}^n\mathbb{1}_{\{x_j=i\}}}-\sqrt{\sum_{j=1}^n\mathbb{1}_{\{y_j=i\}}}\right) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^k\sqrt{p(i)}\frac{\sum_{j=1}^n(\mathbb{1}_{\{x_j=i\}}-\mathbb{1}_{\{y_j=i\}})}{\sqrt{\sum_{j=1}^n\mathbb{1}_{\{x_j=i\}}}+\sqrt{\sum_{j=1}^n\mathbb{1}_{\{y_j=i\}}}} \\ &\leq \frac{1}{\sqrt{n}}\sum_{i=1}^k\sqrt{p(i)}\frac{\sum_{j=1}^n\mathbb{1}_{\{x_j=i\}}\mathbb{1}_{\{y_j\neq x_j\}}}{\sqrt{\sum_{j=1}^n\mathbb{1}_{\{x_j=i\}}}} = \sum_{j=1}^n\underbrace{\sqrt{\frac{p_{x_j}}{n\sum_{\ell=1}^n\mathbb{1}_{\{x_\ell=x_j\}}}}}_{c_j(x)}\cdot\mathbb{1}_{\{x_j\neq y_j\}}\,.\end{aligned}$$

  In view of Theorem 5, we then must evaluate

  $$v\stackrel{\text{def}}{=}\sum_{j=1}^n\mathbb{E}\big[c_j(X)^2\big]=\frac{1}{n}\sum_{j=1}^n\sum_{i=1}^k p(i)^2\cdot\mathbb{E}\left[\frac{1}{1+\sum_{\ell\neq j}\mathbb{1}_{\{X_\ell=i\}}}\right]$$

[2] And is inspired by the Tayor expansion $\sqrt{1+x}=1+\frac{x}{2}-\frac{x^2}{8}+o(x^2)$: there is *some* intuition for it.
[3] Try it: it's a real bummer.

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

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

to conclude that $v\leq\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{k}p(i)=\frac{1}{n}$. By Theorem 5, we obtain

$$\Pr\left[\,|f(s_1,\dots,s_n)-\mathbb{E}[f(s_1,\dots,s_n)]|\geq\frac{\varepsilon^2}{2}\,\right]\leq e^{-\frac{1}{8}n\varepsilon^4}$$

and therefore, as long as $n\geq\frac{8}{\varepsilon^4}\ln\frac{1}{\delta}$, we have $|f(s_1,\dots,s_n)-\mathbb{E}[f(s_1,\dots,s_n)]|\leq\frac{\varepsilon^2}{2}$ with probability at least $1-\delta$.

Putting it all together, we obtain that $\mathrm{d_H}(p,\widehat{p})^2\leq\varepsilon^2$ with probability at least $1-\delta$, as long as $n\geq\max\left(\frac{k}{3\varepsilon^2},\frac{8}{\varepsilon^4}\ln\frac{1}{\delta}\right)$. □

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 $\widehat{p}$ in terms of its Kullback–Leibler (KL) divergence with regard to the true p,

$$\mathrm{KL}(\widehat{p}\,\|\,p)=\sum_{i=1}^{k}\widehat{p}(i)\ln\frac{\widehat{p}(i)}{p(i)}\in[0,\infty]\,.$$

Observing that $\mathrm{d_H}(p,q)^2\leq\frac{1}{2}\mathrm{KL}(p\,\|\,q)$ for any distributions p,q, the aforementioned result is actually stronger than what we need:

Theorem 6 ([Agr19, Theorem 1.2])**.** *Suppose $n\geq\frac{k-1}{\alpha}$. Then*

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

In view of the above relation between Hellinger and KL, we will apply this convergence result with $\alpha\stackrel{\text{def}}{=}2\varepsilon^2$, obtaining

$$\Pr[\,\mathrm{d_H}(\widehat{p},p)\geq\varepsilon\,]\leq e^{-2n\varepsilon^2+(k-1)\ln\frac{2en\varepsilon^2}{k-1}}\,.$$

Fact 7. *For $n\geq\frac{15}{2e}\frac{k}{\varepsilon^2}$, we have $(k-1)\ln\frac{2en\varepsilon^2}{k-1}\leq n\varepsilon^2$.*

Proof. The conclusion is equivalent to $2e\cdot\ln\frac{2en\varepsilon^2}{k-1}\leq\frac{2en\varepsilon^2}{k-1}$, and thus follows from the fact that $x\geq 2e\ln x$ for $x\geq 15$. □

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

3 χ^2 and Kullback—Leibler divergences

To conclude, some remarks on Kullback–Leibler (KL) and chi-squared (χ^2) divergences. Recall their definition, for $p, q \in \Delta([k])$,

$$\mathrm{KL}(p \,\|\, q) = \sum_{i=1}^{k} p(i) \ln \frac{p(i)}{q(i)}\,, \qquad \chi^2(p \,||\, q) = \sum_{i=1}^{k} \frac{(p(i) - q(i))^2}{q(i)}$$

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

$$2\mathrm{d_{TV}}(p, q)^2 \leq \mathrm{KL}(p \,\|\, q) \leq \chi^2(p \,||\, q)\,,$$

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

We are going to focus here on getting $\mathrm{KL}(\widehat{p} \,\|\, p)$ and $\chi^2(\widehat{p} \,||\, p)$ down to ε. 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 $\mathrm{d_{TV}}$: $\Omega\big((k + \log(1/\delta))/\varepsilon^2\big)$. And, behold! The result of Agrawal [Agr19] used in the proof of Theorem 2 does provide the optimal upper bound on learning in KL divergence – and it is achieved by the usual suspect, the empirical estimator:

Theorem 8. $\Phi(\mathrm{KL}, k, \varepsilon, \delta) = \Theta\Big(\frac{k+\log(1/\delta)}{\varepsilon}\Big)$, *where by* KL *we refer to minimizing* $\mathrm{KL}(\widehat{p} \,\|\, p)$.

The optimal sample complexity of learning in χ^2 as a function of k, ε, δ, however, remains open.

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

[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

[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.