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 n to distance ε, with error probability δ, can be done with $O\left(\frac{n+\log(1/\delta)}{\varepsilon^2}\right)$. Thanks to Gautam Kamath and John Wright for suggesting "someone should write this up as a note."

For a given distance measure d, we write $\Phi(\mathrm{d}, n, \varepsilon, \delta)$ for the sample complexity of learning discrete distributions over a known domain of size n, to accuracy $\varepsilon > 0$, with error probability $\delta \in (0, 1]$. As usual, asymptotics will be taken with regard to n 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 $[n] \stackrel{\text{def}}{=} \{1, \dots, n\}$.

1 Total variation distance

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

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

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

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

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

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

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

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

  the last inequality this time by Cauchy–Schwarz. Therefore, for $m \ge \frac{n}{\varepsilon^2}$ 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 $f(s_1, \dots, s_m) \stackrel{\text{def}}{=} \mathrm{d}_{\mathrm{TV}}(p, \tilde{p})$, noting that changing any single sample cannot change its value by more than $c \stackrel{\text{def}}{=} 1/m$:

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

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

Putting it all together, we obtain that $\mathrm{d}_{\mathrm{TV}}(p, \tilde{p}) \le \varepsilon$ with probability at least $1 - \delta$, as long as $m \ge \max\left(\frac{n}{\varepsilon^2}, \frac{2}{\varepsilon^2}\ln\frac{2}{\delta}\right)$. $\square$

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

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

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

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

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

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

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

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

$\square$

[1] John Wright.