The goal of this short note is to provide simple proofs or references for the "folklore facts" on the sample complexity of distinguishing between two fixed discrete probability distributions, with constant error probability. Thanks to Gautam Kamath for parts and possibly all of it.

For two *fixed* probability distributions $p, q \in \Delta(\Omega)$ over a known discrete domain Ω, we write $\Psi(p, q)$ for the sample complexity of deciding, with probability at least $2/3$, which of these two distributions a sequence of i.i.d. samples from an (unknown) probability distribution $r \in \{p, q\}$ originates from.

1 Total variation distance

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

Theorem 1. *$\Psi(p,q) = O\Big(1/\mathrm{d}_{\mathrm{TV}}(p,q)^2\Big)$ and $\Psi(p,q) = \Omega(1/\mathrm{d}_{\mathrm{TV}}(p,q))$.*

Proof. We start with the upper bound. Let $S_{p,q} \stackrel{\text{def}}{=} \operatorname{argsup}_{S\subseteq\Omega}(p(S) - q(S))$, and consider the obvious algorithm which, given n samples from $r \in \{p, q\}$, computes the fraction τ that falls in $S_{p,q}$ and returns ACCEPT if, and only if, $\tau \geq \frac{1}{2}(p(S_{p,q}) + q(S_{p,q}))$ (and REJECT otherwise). Since $p(S_{p,q}) = q(S_{p,q}) + \mathrm{d}_{\mathrm{TV}}(p,q)$ by definition, the test will be correct if we estimate $r(S_{p,q})$ to an additive $\frac{1}{2}\mathrm{d}_{\mathrm{TV}}(p,q)$; which by a standard Hoeffding bound can be done with probability $2/3$ with $n = O(1/\mathrm{d}_{\mathrm{TV}}(p,q)^2)$ samples, as claimed.

Now, for the lower bound it is enough to observe that any distinguisher A which takes n samples is intrinsically the indicator of a specific event $S_A \subseteq \Omega^n$. Therefore, to be successful – i.e., correct with probability at least 2/3–it must be the case that $|p^{\otimes n}(S_A) - q^{\otimes n}(S_A)| \geq \frac{1}{3}$. But since, again by definition and then by subadditivity of total variation,

$$\left|p^{\otimes n}(S_A) - q^{\otimes n}(S_A)\right| \leq \mathrm{d}_{\mathrm{TV}}\left(p^{\otimes n}, q^{\otimes n}\right) \leq n\mathrm{d}_{\mathrm{TV}}(p,q)$$

we need $n \geq \frac{1}{3\mathrm{d}_{\mathrm{TV}}(p,q)}$. □

2 Hellinger distance

Recall that $\mathrm{d}_{\mathrm{H}}(p,q) = \frac{1}{\sqrt{2}}\left(\sum_{x\in\Omega}\left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2\right)^{1/2} = \frac{1}{\sqrt{2}}\|\sqrt{p} - \sqrt{q}\|_2 \in [0,1]$ for any $p, q \in \Delta(\Omega)$.

Theorem 2. *$\Psi(p,q) = \Theta\Big(1/\mathrm{d}_{\mathrm{H}}(p,q)^2\Big)$.*

Proof. The lower bound can be found in [BY02, Theorem 4.7]; we only here establish the upper bound. We will rely on the following two relatively straightforward facts about Hellinger distance, with respect to total variation:

$$1 - \sqrt{1 - \mathrm{d}_{\mathrm{TV}}(p,q)^2} \leq \mathrm{d}_{\mathrm{H}}(p,q)^2 \leq \mathrm{d}_{\mathrm{TV}}(p,q) \tag{1}$$

and products (tensoring):

$$\mathrm{d}_{\mathrm{H}}\left(p^{\otimes n}, q^{\otimes n}\right)^2 = 1 - \left(1 - \mathrm{d}_{\mathrm{H}}(p,q)^2\right)^n . \tag{2}$$

For convenience, let $\varepsilon \stackrel{\text{def}}{=} \mathrm{d}_{\mathrm{H}}(p,q)$. By (2), this implies

$$\mathrm{d}_{\mathrm{H}}\left(p^{\otimes n}, q^{\otimes n}\right)^2 = 1 - \left(1 - \varepsilon^2\right)^n = 1 - e^{n\ln(1-\varepsilon^2)} \geq 1 - e^{-n\varepsilon^2}$$

and therefore $\mathrm{d}_{\mathrm{TV}}(p^{\otimes n}, q^{\otimes n}) \geq 1 - e^{-n\varepsilon^2}$ by (1). For $n \geq \frac{\ln 2}{\varepsilon^2}$, we therefore have $\mathrm{d}_{\mathrm{TV}}(p^{\otimes n}, q^{\otimes n}) \geq 1/2$, which by Theorem 1 implies that $O(1)$ samples from $r^{\otimes n}$ suffice to decide whether $r^{\otimes n} = p^{\otimes n}$ or $r^{\otimes n} = q^{\otimes n}$ with probability $2/3$. Equivalently, this means that $O(n) = O(1/\mathrm{d}_{\mathrm{H}}(p,q)^2)$ from r suffice to decide whether $r = p$ or $r = q$. □

Remark 3. In both Theorem 1 and Theorem 2, the upper bound can easily be, by standard repetition arguments, generalized to allow error probability δ instead of $1/3$, at the price of a factor $\log(1/\delta)$ in the sample complexity. Moreover, the lower bound of Theorem 2, as proven in [BY02, Theorem 4.7], also features this $\log(1/\delta)$ factor.

References

[BY02] Ziv Bar-Yossef. *The Complexity of Massive Data Set Computations*. PhD thesis, UC Berkeley, 2002. Adviser: Christos Papadimitriou. Available at http://webee.technion.ac.il/people/zivby/index_files/Page1489.html. 2, 3