The goal of this short note is to explain the relation between two "folklore" results on simple hypothesis testing, and, quite crucially, how they square with each other. Thanks to Hao-Chung Cheng for illuminating discussions.

For two *fixed* probability distributions $\mathbf{p}, \mathbf{q} \in \Delta(\Omega)$ over a known arbitrary domain Ω, we write $\Psi(\mathbf{p}, \mathbf{q}, \delta)$ for the sample complexity of deciding, with probability of error at most δ, which of these two distributions a sequence of i.i.d. samples from an (unknown) probability distribution $\mathbf{q} \in \{\mathbf{p}_0, \mathbf{p}_1\}$ originates from: specifically, given a uniform prior[1] on $(\mathbf{p}_0, \mathbf{p}_1)$, the error of a test $T\colon \Omega^n \to \{0,1\}$ taking n samples is

$$\delta := \frac{1}{2}\Pr_{\mathbf{p}_0}[\,T(X_1,\dots,X_n)=1\,] + \frac{1}{2}\Pr_{\mathbf{p}_1}[\,T(X_1,\dots,X_n)=0\,] \tag{1}$$

It is well-known (and described in one of these "short notes") that $\Psi(\mathbf{p}, \mathbf{q}, \delta)$ is characterized by the *squared Hellinger distance* between $\mathbf{p}$ and $\mathbf{q}$:

Fact 1 (Sample complexity of simple hypothesis testing)**.** *For any* $\mathbf{p}_0, \mathbf{p}_1$ *and* $\delta \in (0,1]$*,*

$$\Psi(\mathbf{p}_0, \mathbf{p}_1, \delta) = \Theta\left(\frac{\log(1/\delta)}{\mathrm{d_H}(\mathbf{p}_0, \mathbf{p}_1)^2}\right)$$

where $\mathrm{d_H}(\mathbf{p}_0, \mathbf{p}_1) = \frac{1}{\sqrt{2}}\|\sqrt{\mathbf{p}_0} - \sqrt{\mathbf{p}_1}\|_2$ *is the Hellinger distance.*

Flipping things around, one could ask, given n samples, what the best achievable probability of error δ^* (as in (1))) is as a function of $n, \mathbf{p}_0, \mathbf{p}_1$. Let's write $\mathcal{E}_n(\mathbf{p}_0, \mathbf{p}_1) := \frac{1}{n}\ln\frac{1}{\delta^*(n,\mathbf{p}_0,\mathbf{p}_1)}$ for this "finite-sample" *error exponent*, so that

$$\delta^* = e^{-n\mathcal{E}_n(\mathbf{p}_0,\mathbf{p}_1)} \tag{2}$$

Then, Fact 1 appears to state that

$$\mathcal{E}_n(\mathbf{p}_0, \mathbf{p}_1) = \Theta\left(\mathrm{d_H}(\mathbf{p}_0, \mathbf{p}_1)^2\right). \tag{3}$$

This is, however, quite annoying, as a classsical result in information theory and statistics, the Chernoff bound,[2] states that $\lim_{n\to\infty}\mathcal{E}_n(\mathbf{p}_0, \mathbf{p}_1) = C(\mathbf{p}_0, \mathbf{p}_1)$, where

$$C(\mathbf{p}_0, \mathbf{p}_1) := -\min_{\lambda\in[0,1]} \ln\sum_{x\in\Omega}\mathbf{p}_0(x)^\lambda \mathbf{p}_1(x)^{1-\lambda} \tag{4}$$

is the *Chernoff information* between $\mathbf{p}_0$ and $\mathbf{p}_1$ (with the straightforward generalization if Ω is not a discrete domain). *Annoying*, because $C(\mathbf{p}_0, \mathbf{p}_1)$ and $\mathrm{d_H}(\mathbf{p}_0, \mathbf{p}_1)^2$ are clearly not the same thing, and having two different (and seemingly *wildly* different) things characterize the same quantity is very confusing at best. So, erm, **how come?**

1 Hellinger distance: (3) is not wrong…

We first provide a self-contained proof of (3); actually, of a stronger version of it, with explicit constants. This is adapting and combining the contents from another of these short notes, "A short note on distinguishing discrete distributions" (2017), and [BY02, Theorem 4.7].

[1] One can generalize this to a non-uniform prior; the characterization of the error exponent as Chernoff information will remain the same, as the prior "disappears" asymptotically.

[2] No, not *that* Chernoff bound.

Lemma 2. *For any* $\mathbf{p}_0, \mathbf{p}_1$*, and* $n \geq 1$*, we have* $\frac{1}{2}e^{2n\ln(1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2)} \leq 2\delta^*(n,\mathbf{p}_0,\mathbf{p}_1) \leq e^{n\ln(1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2)}$, i.e.,

$$-\ln(1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2) - \frac{2\ln 2}{n} \leq \mathcal{E}_n(\mathbf{p}_0,\mathbf{p}_1) \leq -2\ln(1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2) - \frac{\ln 2}{n}\,, \tag{5}$$

which implies (3).

Proof. By the standard interpretation of total variation distance as characterization of the minimal sum of Type I and Type II errors, we have that

$$1 - 2\delta^*(n,\mathbf{p}_0,\mathbf{p}_1) = \mathrm{d_{TV}}\left(\mathbf{p}_0^{\otimes n},\mathbf{p}_1^{\otimes n}\right) \tag{6}$$

since $\delta^*(n,\mathbf{p}_0,\mathbf{p}_1)$ was defined in (1) as the optimal average error probability when distinguishing $\mathbf{p}_0$ and $\mathbf{p}_1$ from n i.i.d. samples. So our task boils down to establishing good enough upper and lower bounds on $\mathrm{d_{TV}}\left(\mathbf{p}_0^{\otimes n},\mathbf{p}_1^{\otimes n}\right)^2$.

To do so, we will rely on the following two relatively straightforward facts about Hellinger distance, with respect to total variation:

$$1 - \sqrt{1-\mathrm{d_{TV}}(\mathbf{p}_0,\mathbf{p}_1)^2} \leq \mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2 \leq \mathrm{d_{TV}}(\mathbf{p}_0,\mathbf{p}_1) \tag{7}$$

and products (tensoring):

$$\mathrm{d_H}\left(\mathbf{p}_0^{\otimes n},\mathbf{p}_1^{\otimes n}\right)^2 = 1 - \left(1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2\right)^n. \tag{8}$$

By (8), this implies $\mathrm{d_H}\left(\mathbf{p}_0^{\otimes n},\mathbf{p}_1^{\otimes n}\right)^2 = 1 - \left(1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2\right)^n = 1 - e^{n\ln(1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2)}$, and therefore , by (7),

$$\mathrm{d_{TV}}\left(\mathbf{p}_0^{\otimes n},\mathbf{p}_1^{\otimes n}\right) \geq 1 - e^{n\ln(1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2)} \tag{9}$$

Conversely, from the lower bound from (7) and using (8), we get

$$\mathrm{d_{TV}}\left(\mathbf{p}_0^{\otimes n},\mathbf{p}_1^{\otimes n}\right)^2 \leq 1 - \left(1-\mathrm{d_H}\left(\mathbf{p}_0^{\otimes n},\mathbf{p}_1^{\otimes n}\right)^2\right)^2 = 1 - \left(1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2\right)^{2n} = 1 - e^{2n\ln(1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2)} \tag{10}$$

and so, combining the two and recalling (6), we finally get

$$1 - \sqrt{1 - e^{2n\ln(1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2)}} \leq 2\delta^*(n,\mathbf{p}_0,\mathbf{p}_1) \leq e^{n\ln(1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2)} \tag{11}$$

and observing that $1-\sqrt{1-x} \geq x/2$ gives the claim. □

2 ... yet Chernoff is correct.

To square Lemma 2 with the Chernoff bound, which states that

$$\lim_{n\to\infty} \mathcal{E}_n(\mathbf{p}_0,\mathbf{p}_1) = C(\mathbf{p}_0,\mathbf{p}_1) \tag{12}$$

we need to argue that, while maybe not *equal*, $-\ln(1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2)$ and $C(\mathbf{p}_0,\mathbf{p}_1)$ are always within a factor 2 of each other. Basically, that constant factors *do*, after all, matter.

The first observation is to rewrite $1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2$ in an equivalent (and standard-ish) form involving the *Bhattacharyya coefficient*,

$$\mathrm{BC}(\mathbf{p}_0,\mathbf{p}_1) \coloneqq \sum_{x\in\Omega} \sqrt{\mathbf{p}_0(x)\mathbf{p}_1(x)}\,. \tag{13}$$

Namely, one can check that $1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2 = 1 - \mathrm{BC}(\mathbf{p}_0,\mathbf{p}_1)$. This is very convenient, as now we want to compare

$$-\ln(1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2) = -\ln\mathrm{BC}(\mathbf{p}_0,\mathbf{p}_1)$$

to

$$C(\mathbf{p}_0,\mathbf{p}_1) = -\min_{\lambda\in[0,1]} \ln \sum_{x\in\Omega} \mathbf{p}_0(x)^\lambda \mathbf{p}_1(x)^{1-\lambda} = -\ln \min_{\lambda\in[0,1]} \sum_{x\in\Omega} \mathbf{p}_0(x)^\lambda \mathbf{p}_1(x)^{1-\lambda}\,.$$

Getting rid of the logarithms, it would be enough to show that $\min_{\lambda\in[0,1]} \sum_{x\in\Omega} \mathbf{p}_0(x)^\lambda \mathbf{p}_1(x)^{1-\lambda}$ and $\sum_{x\in\Omega} \sqrt{\mathbf{p}_0(x)\mathbf{p}_1(x)}$ are within a quadratic factor of each other. Big if true! And, fortunately, true.

Lemma 3 (Skewed Bhattacharyya coefficients are quadratically related)**.** *For any* $\mathbf{p}_0,\mathbf{p}_1$ *and* $\lambda\in[0,1]$*, we have*

$$\left(\sum_{x\in\Omega} \sqrt{\mathbf{p}_0(x)\mathbf{p}_1(x)}\right)^2 \leq \sum_{x\in\Omega} \mathbf{p}_0(x)^\lambda \mathbf{p}_1(x)^{1-\lambda} \leq 1\,. \tag{14}$$

In particular, we have

$$\mathrm{BC}(\mathbf{p}_0,\mathbf{p}_1)^2 \leq \min_{\lambda\in[0,1]} \mathrm{BC}_\lambda(\mathbf{p}_0,\mathbf{p}_1) \leq \mathrm{BC}(\mathbf{p}_0,\mathbf{p}_1) \tag{15}$$

where $\mathrm{BC}_\lambda(\mathbf{p}_0,\mathbf{p}_1) = \sum_{x\in\Omega} \mathbf{p}_0(x)^\lambda \mathbf{p}_1(x)^{1-\lambda}$ *denotes the* λ-skewed Bhattacharyya coefficient.

Proof. Fix any $\lambda\in(0,1)$ (the cases $\lambda\in\{0,1\}$ being immediate). First, we have

$$\sum_{x\in\Omega} \mathbf{p}_0(x)^\lambda \mathbf{p}_1(x)^{1-\lambda} = \sum_{x\in\Omega} \mathbf{p}_1(x)\cdot\left(\frac{\mathbf{p}_0(x)}{\mathbf{p}_1(x)}\right)^\lambda \leq \left(\sum_{x\in\Omega} \mathbf{p}_1(x)\cdot\frac{\mathbf{p}_0(x)}{\mathbf{p}_1(x)}\right)^\lambda = \left(\sum_{x\in\Omega} \mathbf{p}_0(x)\right)^\lambda = 1$$

using Jensen's inequality for the concave function $x\mapsto x^\lambda$.

Second, let's use Hölder's (the generalized version, with 3 vectors) with exponents $(2/(1-\lambda), 2, 2/\lambda, 2)$, which satisfy $\frac{1-\lambda}{2}+\frac{1}{2}+\frac{\lambda}{2}=1$. We have

$$\begin{aligned}
\sum_{x\in\Omega} \mathbf{p}_0(x)^{1/2}\mathbf{p}_1(x)^{1/2} &= \sum_{x\in\Omega} \mathbf{p}_0(x)^{\frac{1-\lambda}{2}}\cdot \mathbf{p}_0(x)^{\frac{\lambda}{2}}\mathbf{p}_1(x)^{\frac{1-\lambda}{2}}\cdot\mathbf{p}_1(x)^{\frac{\lambda}{2}} \\
&\leq \left(\sum_{x\in\Omega} \mathbf{p}_0(x)\right)^{\frac{1-\lambda}{2}} \left(\sum_{x\in\Omega} \mathbf{p}_0(x)^\lambda \mathbf{p}_1(x)^{1-\lambda}\right)^{\frac{1}{2}} \left(\sum_{x\in\Omega} \mathbf{p}_1(x)\right)^{\frac{\lambda}{2}} && \text{(Hölder)} \\
&= \left(\sum_{x\in\Omega} \mathbf{p}_0(x)^\lambda \mathbf{p}_1(x)^{1-\lambda}\right)^{\frac{1}{2}},
\end{aligned}$$

concluding the proof. □

This readily implies that, for every $\mathbf{p}_0,\mathbf{p}_1$,

$$-\ln(1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2) \leq C(\mathbf{p}_0,\mathbf{p}_1) \leq -2\ln(1-\mathrm{d_H}(\mathbf{p}_0,\mathbf{p}_1)^2) \tag{16}$$

and sanity is restored.

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.