The goal of this short document is to highlight a very useful result due to Valiant and Valiant [VV17], and slightly simplify a component of their proof (Lemma 2) by using the chi-squared distance instead of Hellinger. (We also discuss, in Section 2, the advantage of the former over the latter.) Throughout, we write $\mathbb{N}$ for the set of non-negative integers, and $a \wedge b$ (resp. $a \vee b$) to denote the minimum (resp. maximum) of $a$ and $b$.

**Theorem 1** ([VV17, Theorem 4.2])**.** *Given a distribution* $\mathbf{p}$ *over* $\mathbb{N}$*, and associated values* $\alpha_i$ *such that* $\alpha_i \in [0,1]$ *for all* $i \in \mathbb{N}$*, define the distribution over distributions* $Q$ *by the following process: independently for each* $i \in \mathbb{N}$*, uniformly choose* $z_i \in \{-1,1\}$*, set* $\tilde{\mathbf{q}}_i = (1+z_i\alpha_i)\mathbf{p}(i)$*,and then normalize* $\tilde{\mathbf{q}}$ *to obtain a distribution* $\mathbf{q}$*. Then there exists an absolute constant* $c > 0$ *such that it takes at least* $c(\sum_i \alpha_i^4 \mathbf{p}(i)^2)^{-1/2}$ *samples to distinguish* $\mathbf{p}$ *from* $Q$ *with success probability* $2/3$*. Further, with probability at least* $1/2$*, the total variation distance between a random distribution from* $Q$ *and* $\mathbf{p}$ *is at least* $\frac{1}{2}\min\left(\sum_{i\in\mathbb{N}} \alpha_i\mathbf{p}(i) - \max_i \alpha_i\mathbf{p}(i), \frac{1}{2}\sum_{i\in\mathbb{N}}\alpha_i\mathbf{p}(i)\right)$ *.*

*Proof.* The proof almost identically follows that of [VV17, Theorem 4.2]. We first argue the first part of the statement, about indistinguishability. Instead of considering directly $\mathbf{p}$ and the mixture $\mathbf{Q} \stackrel{\text{def}}{=} \mathbb{E}_{\mathbf{q}\sim Q}[\mathbf{q}]$, we will, fixing a target number of samples $n$, focus on the related Poisson processes $\Pi_{\mathbf{p}}, \Pi_{\mathbf{Q}}$ defined as follows:

- $\Pi_{\mathbf{p}}$ is the product distribution $\bigotimes_{i=1}^{\infty} \text{Poisson}(n\mathbf{p}(i))$ over $\mathbb{N}^{\mathbb{N}}$ (i.e., each coordinate $i$ is independent of all others, and is a $\text{Poisson}(n\mathbf{p}(i))$ r.v.);
- $\Pi_{\mathbf{Q}}$ is the mixture of product distributions $\mathbb{E}_{\mathbf{q}\sim Q}[\bigotimes_{i=1}^{\infty} \text{Poisson}(n\tilde{\mathbf{q}}(i))]$ over $\mathbb{N}^{\mathbb{N}}$ (i.e., each coordinate $i$ is independent of all others and is a $\text{Poisson}(n(1+z_i\alpha_i)\mathbf{p}(i))$ r.v., where $z_i$ is u.a.r. in $\{-1,1\}$).

We first argue that, given a draw from $\Pi_{\mathbf{p}}$ (resp. $\Pi_{\mathbf{Q}}$), one can generate, with probability at least $1/2$, $n$ i.i.d. samples from $\mathbf{p}$ (resp. $\mathbf{Q}$) as follows. Given a draw $\mathbf{x} \in \mathbb{N}^{\mathbb{N}}$ from either $\Pi_{\mathbf{p}}$ or $\Pi_{\mathbf{Q}}$:

- if $\sum_{i=1}^{\infty} \mathbf{x}_i < n$, then output fail.
- Otherwise, create a (finite)[1] multiset $T$ of length $\sum_{i=1}^{\infty} \mathbf{x}_i \geq n$ containing $\mathbf{x}_i$ occurrences of each $i \in \mathbb{N}$, and select u.a.r. a multiset $S$ of $T$ of size $n$. Output $S$.

It is not hard to see that, conditioned on not outputting fail, the above process outputs a set of $n$ i.i.d. samples distributed exactly according to $\mathbf{p}$ (resp. $\mathbf{Q}$). Thus, it suffices to show that the probability of outputting fail is at most $1/2$. This itself is a consequence of the fact that $\sum_{i=1}^{\infty} \mathbf{x}_i \sim \text{Poisson}(n)$, so that, $n$ being an integer, its median is exactly $n$.

Suppose that, for a given $n$, one cannot distinguish $\Pi_{\mathbf{p}}$ and $\Pi_{\mathbf{Q}}$ with advantage $1/12$; that is, given a sample drawn from either $\Pi_{\mathbf{p}}$ or $\Pi_{\mathbf{Q}}$ (each with probability $1/2$), the probability to guess correctly which one is less than $1/2 + 1/12 = 7/12$: then we claim that, given $n$ samples, one cannot distinguish $\mathbf{p}$ from $\mathbf{Q}$ with advantage $1/6$. Indeed, by contradiction, suppose we have an tester $\mathcal{T}$ for the latter task with sample complexity $n$ and advantage at least $1/6$:

$$\Pr_{\substack{b\sim\text{Bern}(1/2)\\ \mathbf{h}\sim b\mathbf{p}+(1-b)\mathbf{Q}}}[\,\mathcal{T},\ \text{given } n \text{ samples from } \mathbf{h},\ \text{outputs } b\,] \geq \frac{2}{3}\,. \tag{1}$$

Then we can get a distinguisher $\mathcal{T}'$ between $\Pi_{\mathbf{p}}$ and $\Pi_{\mathbf{Q}}$ with advantage $1/12$: given a sample a draw $\mathbf{x} \in \mathbb{N}^{\mathbb{N}}$ from either $\Pi_{\mathbf{p}}$ or $\Pi_{\mathbf{Q}}$, $\mathcal{T}'$ tries to generate $n$ i.i.d. samples as described above. If the output is fail, then it outputs a bit uniformly at random; otherwise, it runs $\mathcal{T}$ on the resulting $n$ samples and outputs what $\mathcal{T}$ returns. It follows that

$$\Pr_{\substack{b\sim\text{Bern}(1/2)\\ \mathbf{x}\sim b\Pi_{\mathbf{p}}+(1-b)\Pi_{\mathbf{Q}}}}[\,\mathcal{T}',\ \text{given } \mathbf{x},\ \text{outputs } b\,] \geq \frac{1}{2}\Pr[\,\mathsf{fail}\,] + \frac{2}{3}(1-\Pr[\,\mathsf{fail}\,]) \geq \frac{7}{12}\,, \tag{2}$$

a contradiction. Therefore, it suffices that one cannot distinguish $\Pi_{\mathbf{p}}$ and $\Pi_{\mathbf{Q}}$ with advantage $1/12$; for which it is enough to show that unless $n$ is large enough, the total variation distance $\mathrm{d}_{\mathrm{TV}}(\Pi_{\mathbf{p}}, \Pi_{\mathbf{Q}})$ is smaller than

[1] Note that $\sum_{i=1}^{\infty} \mathbf{x}_i$ is itself a $\text{Poisson}(n)$ random variable, so finite a.s.

some absolute constant $c > 0$. The crux is that, due to the definition of our two processes, both $\Pi_{\mathbf{p}}$ and $\Pi_{\mathbf{Q}}$ are product distributions. Therefore, using the subadditivity of Hellinger distance for product distributions, we can now bound

$$\mathrm{d_{TV}}(\Pi_{\mathbf{p}}, \Pi_{\mathbf{Q}})^2 \leq \mathrm{d_H}(\Pi_{\mathbf{p}}, \Pi_{\mathbf{Q}})^2 \leq \sum_{i\in\mathbb{N}} \mathrm{d_H}(\mathrm{Poisson}(n\mathbf{p}(i)), \Pi_{\mathbf{Q},i})^2 \tag{3}$$

where $\Pi_{\mathbf{Q},i} = \frac{1}{2}\left(\mathrm{Poisson}(n(1+\alpha_i)\mathbf{p}(i)) + \mathrm{Poisson}(n(1-\alpha_i)\mathbf{p}(i))\right)$. This is where we invoke Lemma 2, leading to

$$\mathrm{d_{TV}}(\Pi_{\mathbf{p}}, \Pi_{\mathbf{Q}})^2 \leq \sum_{i\in\mathbb{N}} \alpha_i^4 (n\mathbf{p}_i)^2 = n^2 \sum_{i\in\mathbb{N}} \alpha_i^4 \mathbf{p}_i^2 \,. \tag{4}$$

For the total variation to be at least $c$, we thus need $n \geq c^{1/2}\left(\sum_{i\in\mathbb{N}} \alpha_i^4 \mathbf{p}_i^2\right)^{-1/2}$, concluding the proof of indistinguishability.

We now turn to the second part of the claim about the distance. To ease notation, set $w_i \stackrel{\text{def}}{=} \alpha_i \mathbf{p}(i)$ for all $i \in \mathbb{N}$, and assume without loss of generality that the sequence $w$ is non-increasing; our goal is then to show

$$\Pr_{\mathbf{q}\sim Q}\left[ \mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) \geq \frac{1}{2}\min(\|w\|_1 - \|w\|_\infty, \frac{1}{2}\|w\|_1) \right] \geq \frac{1}{2}\,.$$

Observe that for any $\mathbf{q}$ (defined from the corresponding sequence $z \in \mathbb{N}^{\mathbb{N}}$), we have, since $\mathbf{q} = \tilde{\mathbf{q}}/\sum_i \tilde{\mathbf{q}}(i)$ and $\sum_i \tilde{\mathbf{q}}(i) = 1 + \sum_i z_i w_i$,

$$2\mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) = \sum_{i\in\mathbb{N}} |\mathbf{p}(i) - \mathbf{q}(i)| \geq \sum_{i\in\mathbb{N}} |\mathbf{p}(i) - \tilde{\mathbf{q}}(i)| - \sum_{i\in\mathbb{N}} |\tilde{\mathbf{q}}(i) - \mathbf{q}(i)| = \|w\|_1 - |\sum_{i\in\mathbb{N}} z_i w_i|\,. \tag{5}$$

Therefore, it suffices to show that $|\sum_{i\in\mathbb{N}} z_i w_i| \leq \|w\|_\infty \vee \frac{1}{2}\|w\|_1$ with probability at least $1/2$. We proceed by a distinction of cases: first, suppose $w_0 = \|w\|_\infty \geq \frac{1}{2}\|w\|_1$. Then

$$\Pr\left[ \Big|\sum_{i\geq 0} z_i w_i\Big| \leq \|w\|_\infty \right] = \Pr\left[ \Big| z_0\|w\|_\infty + \sum_{i\geq 1} z_i w_i \Big| \leq \|w\|_\infty \right] = \Pr\left[ \sum_{i\geq 1} z_i w_i \leq 0 \right] = \frac{1}{2}$$

by symmetry.

Otherwise, assume $\|w\|_\infty < \frac{1}{2}\|w\|_1$, and consider the index $t \geq 1$ such that $\sum_{i=0}^{t-1} w_i \leq \frac{1}{2}\|w\|_1 < \sum_{i=0}^{t} w_i$. Note that this implies $\sum_{i=0}^{t-1} w_i - \sum_{i=t}^{\infty} w_i \geq -\frac{1}{2}\|w\|_1$, as otherwise we could write

$$\|w\|_1 \geq w_0 + \sum_{i=t}^{\infty} w_i \geq w_t + \sum_{i=t}^{\infty} w_i > w_t + \sum_{i=0}^{t-1} w_i + \frac{1}{2}\|w\|_1 > \|w\|_1\,,$$

a contradiction. Then, since $\sum_{i=0}^{t-1} z_i w_i$ and $\sum_{i=t}^{\infty} z_i w_i$ have opposite signs with probability $1/2$,

$$\Pr\left[ \Big|\sum_{i\geq 0} z_i w_i\Big| \leq \frac{1}{2}\|w\|_1 \right] = \Pr\left[ \Big|\sum_{i=0}^{t-1} z_i w_i + \sum_{i=t}^{\infty} z_i w_i\Big| \leq \frac{1}{2}\|w\|_1 \right] \geq \frac{1}{2}$$

concluding the proof. □

# 1 Bounding distances between a Poisson and a mixture

**Lemma 2** (Hellinger bound)**.** *Let* $\lambda > 0$*, and* $\alpha \in [0,1]$*. Define* $\mathbf{Q} \stackrel{\text{def}}{=} \frac{1}{2}(\mathrm{Poisson}((1+\alpha)\lambda) + \mathrm{Poisson}((1-\alpha)\lambda))$*. Then* $\mathrm{d_H}(\mathrm{Poisson}(\lambda), \mathbf{Q}) \leq \alpha^2\lambda$*.*

(Note that proving the quadratically weaker bound $\mathrm{d_H}(\mathrm{Poisson}(\lambda)\,,\mathbf{Q})^2 \lesssim \alpha^2\lambda$ is straightforward, but insufficient to our purposes.) Lemma 2 will follow from the analogous (but easier to prove) claim for chi-squared distance, along with Fact 5. Note that as the chi-squared distance is not symmetric, the order of the distributions matters in Lemma 3.

**Lemma 3** ($\chi^2$ bound)**.** *Let $\lambda > 0$, and $\alpha \in [0,1]$. Define $\mathbf{Q} \stackrel{\text{def}}{=} \frac{1}{2}(\mathrm{Poisson}((1+\alpha)\lambda) + \mathrm{Poisson}((1-\alpha)\lambda))$. Then $1 \wedge \chi^2(\mathbf{Q} \mid\mid \mathrm{Poisson}(\lambda)) \leq \alpha^4\lambda^2$.*

*Proof.* We can assume in the rest of the proof that $\alpha^2\lambda \leq 1$, as otherwise there is nothing to prove (the LHS being at most 1 due to the minimum). For convenience, write $\mathbf{P} \stackrel{\text{def}}{=} \mathrm{Poisson}(\lambda)$. We can express the pmf of $\mathbf{Q}$ as

$$\mathbf{Q}(n) = \frac{1}{2}\left(e^{-\lambda(1+\alpha)}\frac{\lambda^n(1+\alpha)^n}{n!} + e^{-\lambda(1-\alpha)}\frac{\lambda^n(1-\alpha)^n}{n!}\right) = \mathbf{P}(n)\cdot\frac{e^{-\lambda\alpha}(1+\alpha)^n + e^{\lambda\alpha}(1-\alpha)^n}{2} \tag{6}$$

for $n \in \mathbb{N}$. It follows that

$$\chi^2(\mathbf{Q} \mid\mid \mathbf{P}) = -1 + \sum_{n\in\mathbb{N}}\frac{\mathbf{Q}(n)^2}{\mathbf{P}(n)} = -1 + e^{-\lambda}\sum_{n\in\mathbb{N}}\frac{\lambda^n}{n!}\left(\frac{e^{-\lambda\alpha}(1+\alpha)^n + e^{\lambda\alpha}(1-\alpha)^n}{2}\right)^2 \tag{7}$$

Focusing on the last sum, we expand the square and compute it explicitly:

$$\begin{aligned}\sum_{n\in\mathbb{N}}\frac{\lambda^n}{n!}\left(\frac{e^{-\lambda\alpha}(1+\alpha)^n + e^{\lambda\alpha}(1-\alpha)^n}{2}\right)^2 &= \sum_{n\in\mathbb{N}}\frac{\lambda^n}{n!}\frac{e^{-2\lambda\alpha}(1+\alpha)^{2n} + e^{2\lambda\alpha}(1-\alpha)^{2n} + 2(1-\alpha^2)^n}{4} \\ &= \frac{e^{-2\lambda\alpha}e^{\lambda(1+\alpha)^2} + e^{2\lambda\alpha}e^{\lambda(1-\alpha)^2} + 2e^{\lambda(1-\alpha^2)}}{4} \\ &= e^{\lambda}\cdot\frac{e^{\lambda\alpha^2} + e^{-\lambda\alpha^2}}{2}\,.\end{aligned}$$

Plugging this in (7), we get

$$\chi^2(\mathbf{Q} \mid\mid \mathbf{P}) = -1 + \frac{e^{\lambda\alpha^2} + e^{-\lambda\alpha^2}}{2} \leq \lambda^2\alpha^4 \tag{8}$$

where for the last inequality we used our bound $\lambda\alpha^2 \leq 1$, and the fact that $\cosh x \leq 1 + x^2$ for $|x| \leq 1$. This concludes the proof. ☐

## 2 Why $\chi^2$ instead of Hellinger

At first glance, our choice to use the chi-squared distance instead of Hellinger distance for the key technical lemma may seem peculiar. After all, the chi-squared distance (which is, at the end of the day, merely a first-order approximation to the Kullback–Leibler divergence[2]) is not bounded, and not even a distance (being asymmetric); while the Hellinger distance is bounded, behaves nicey with respect to product distributions (e.g., via subadditivity of its square), and overall looks clean and appealing.

However, a pervasice technique in proving sample complexity lower bounds involves reference distribution $\mathbf{p}$ and a family of *perturbations* of $\mathbf{p}$, $(\mathbf{p}_z)_{z\in\mathcal{Z}}$ (for some suitable parameter set $\mathcal{Z}$), such that $\mathbf{p}_z(i) =$

[2]Indeed, for the KL divergence in nats,

$$\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) = \sum_i \mathbf{p}(i)\ln\frac{\mathbf{p}(i)}{\mathbf{q}(i)} = \sum_i \mathbf{p}(i)\ln\left(1 + \frac{\mathbf{p}(i)-\mathbf{q}(i)}{\mathbf{q}(i)}\right) \approx \sum_i \frac{\mathbf{p}(i)^2}{\mathbf{q}(i)} - 1 = \chi^2(\mathbf{p} \mid\mid \mathbf{q})$$

since $\ln(1+x) \approx x$ and $\chi^2(\mathbf{p} \mid\mid \mathbf{q}) = \sum_i \frac{(\mathbf{p}(i)-\mathbf{q}(i))^2}{\mathbf{q}(i)} = \sum_i \frac{\mathbf{p}(i)^2}{\mathbf{q}(i)} - 1$.

$(1+\delta(i,z))\mathbf{p}(i)$ for all $i$. The key then is to upper bound the total variation distance between the reference distribution and the *mixture* of perturbations,

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbb{E}_Z[\mathbf{p}_Z])$$

(instead of the looser $\mathbb{E}_Z[\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{p}_Z)]$, which lacks a lot of useful cancellations and is typically much bigger). But using the Hellinger distance as a proxy will then involve a rather nasty square root: even assuming that $\mathbf{Q} \stackrel{\text{def}}{=} \mathbb{E}_Z[\mathbf{p}_Z]$

$$\mathrm{d}_{\mathrm{H}}(\mathbf{p}, \mathbb{E}_Z[\mathbf{p}_Z])^2 = \sum_i \left(\sqrt{\mathbf{p}(i)} - \sqrt{\mathbb{E}_Z[\mathbf{p}_Z(i)]}\right)^2 = \sum_i \mathbf{p}(i)\left(1 - \sqrt{1+\mathbb{E}_Z[\delta(i,Z)]}\right)^2$$

which is generally *not* a fun task. Yet, bounding the total variation distance by the (now bounded) quantity $1 \wedge \chi^2(\mathbb{E}_Z[\mathbf{p}_Z] \mid\mid \mathbf{p})$ leads to an expression of the form

$$\chi^2(\mathbb{E}_Z[\mathbf{p}_Z] \mid\mid \mathbf{p}) = \sum_i \frac{(\mathbb{E}_Z[\mathbf{p}_Z(i)] - \mathbf{p}(i))^2}{\mathbf{p}(i)} = \sum_i \mathbf{p}(i)\mathbb{E}_Z[\delta(i,Z)]^2$$

(the order of the distributions in the chi-squared distance will typically matter a lot: mixture first). Squares are in my experience nicer to handle than square roots.

Further, we have many other tools to deal with the chi-squared distance; for instance, the handy lemma below, due to [Pol03], which enables us to handle chi-square distances of mixtures with respect to a reference product distribution.

**Lemma 4** ([ACT19, Lemma 8])**.** *Consider a random variable $Z$ such that for each $Z = z$ the distribution $\mathbf{q}_z^n$ is defined as $\mathbf{q}_{1,z} \times \cdots \times \mathbf{q}_{n,z}$. Further, let $\mathbf{p}^n = \mathbf{p}_1 \times \cdots \times \mathbf{p}_n$ be a fixed product distribution. Then,*

$$\chi^2(\mathbb{E}_Z[\mathbf{q}_Z^n] \mid\mid \mathbf{p}^n) = \mathbb{E}_{ZZ'}[\prod_{j=1}^n (1+\Delta_j(Z,Z'))] - 1,$$

*where $Z'$ is an independent copy of $Z$, and with $\delta_j^z(x_j) = \frac{\mathbf{q}_{j,z}(x_j) - \mathbf{p}_j(x_j)}{\mathbf{p}_j(x_j)}$, $\Delta_j(z,z')$ is the* chi-squared correlation

$$\Delta_j(z,z') \stackrel{\text{def}}{=} \mathbb{E}\left[\delta_j^z(X_j)\delta_j^{z'}(X_j)\right],$$

*where the expectation is over $X_j$ distributed according to $\mathbf{p}_j$.*

# A Bounding Hellinger by $\chi^2$

For the sake of self-completeness, we give a simple proof of the fact that the chi-squared distance upper bounds the squared Hellinger one.

**Fact 5.** *For any two discrete distributions $\mathbf{p}_1, \mathbf{p}_2$, $\mathrm{d}_{\mathrm{H}}(\mathbf{p}_1, \mathbf{p}_2)^2 \leq 1 \wedge \frac{1}{2}\chi^2(\mathbf{p}_1 \mid\mid \mathbf{p}_2)$.*

*Proof.* This is easily shown from (i) $\mathrm{d}_{\mathrm{H}}(\mathbf{p}_1, \mathbf{p}_2) \leq 1$, and (ii) the identity $a - b = (\sqrt{w} - \sqrt{b})(\sqrt{w} + \sqrt{b})$, as

$$2\mathrm{d}_{\mathrm{H}}(\mathbf{p}_1, \mathbf{p}_2)^2 = \sum_i \left(\sqrt{\mathbf{p}_1(i)} - \sqrt{\mathbf{p}_2(i)}\right)^2 = \sum_i \frac{(\mathbf{p}_1(i) - \mathbf{p}_2(i))^2}{\left(\sqrt{\mathbf{p}_1(i)} + \sqrt{\mathbf{p}_2(i)}\right)^2} \leq \sum_i \frac{(\mathbf{p}_1(i) - \mathbf{p}_2(i))^2}{\mathbf{p}_2(i)},$$

which is exactly $\chi^2(\mathbf{p}_1 \mid\mid \mathbf{p}_2)$. See, e.g. [GS02] for the continuous case as well, or [DKW18, Proposition 1] for a generalization to subdistributions. ∎

# References


[ACT19] Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Inference under information constraints: Lower bounds from chi-square contraction. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 3–17, Phoenix, USA, 25–28 Jun 2019. PMLR.

[DKW18] Constantinos Daskalakis, Gautam Kamath, and John Wright. Which distribution distances are sublinearly testable? In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2747–2764. SIAM, Philadelphia, PA, 2018.

[GS02] Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[Pol03] David Pollard. Asymptopia, 2003. Manuscript.

[VV14] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *55th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2014*, pages 51–60. IEEE Computer Soc., Los Alamitos, CA, 2014.

[VV17] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM J. Comput.*, 46(1):429–455, 2017. Journal version of [VV14].