6.883 Science of Deep Learning – Spring 2018 February 20, 2018

Lecture 10: Hypothesis Testing

Lecturer: Constantinos Daskalakis *Scribes: Yu Xia, Yi Sun, David Mayo*
(Revised by Andrew Ilyas and Manolis Zampetakis)

Disclaimer. The proofs presented in this note may require background knowledge in classical statistics and/or hypothesis testing (particularly for the uniform distribution case).

1 Introduction

Almost every machine learning problem involves making inferences about an unknown distribution based on random samples. In the real world, there are many large datasets where the data only represents a tiny fraction of an underlying distribution we hope to understand. For example, if we are given 100 samples from an unknown distribution supported on 1000 domain elements, then the empirical distribution given by the samples may be a poor approximation of the true distribution. Thus, it is important to understand sample complexities before making inferences on the data.

The idea of modeling unknown distributions based on a set of data samples has a long history in the field of statistics and is widely used in a number of fields of science.

For example, in psychology when measuring the prevalence of a particular trait in a population, a key question that must be answered is how large of a sample size is need to provide a certain level of confidence that the sampled population accurately measures the distribution of the entire population. Answering this question requires an understanding of distribution distance metrics and hypothesis testing, both of which will be explored in this lecture. We will also focus on the sample complexity needed to distinguish an unknown distribution from a known distribution which under the GAN paradigm is the key to training a generator function that accurately models an unknown distribution.

More specifically, we are trying to answer the following question:

Main question of this lecture: How many samples from an unknown probability distribution p do we need to distinguish p from an known distribution q with respect to some distance metric?

This fundamental question has received tremendous attention in statistics, and it's also essential for understanding deep learning. Recall from previous lectures the optimization goal of a GAN:

$$\inf_{\theta_g} \sup_{\theta_d} f(\theta_g, \theta_d),$$

where $f(\theta_g, \theta_d)$ quantifies how well a discriminator discriminates between a true distribution and a generator distribution. In particular, $\sup_{\theta_d} f(\theta_g, \theta_d)$ can be seen as a statistical estimation problem. An example of an algorithm that attempts to solve this estimation problem is the Wasserstein GAN where the discriminator problem is aimed at approximating the Wasserstein distance between the true distribution and the generator distribution (albeit for a restricted family of test fractions) [12].

In order to be able to use GANs in practice and understand how well they are learning a target distribution, it is important to have a method of measuring the generator and target distribution distance. Otherwise, its possible that a discriminator that was trained in tandem with a generator cannot distinguish between the true distribution and the generator distribution, but a better discriminator (e.g. human brain) could.

This lecture aims to understand the mathematical structure behind the problem: what is the sample size required to distinguish the generator distribution and the target distribution? We will start with some simple target distributions (Bernoulli, uniform) and then expand to more general distributions.

2 Notions of statistical distance

Before getting into some examples, lets get familiar with several notions of distances of probabilistic measures.

- We use $\boldsymbol{d_{TV}}$ to represent the **total variation (TV) distance**, which is given by

$$d_{TV}(p,q) = \frac{1}{2}\int_{x\in\mathcal{X}} |p(x) - q(x)| = \sup_{A\subseteq\mathcal{X}} |p(A) - q(A)|.$$

- **Kolmogorov distance** d_K is another distance metric similar to total variation, but uses the cumulative density function (CDF) rather than the density—it is only defined for probability measures on the real numbers.

$$\begin{aligned} d_K(p,q) &= \sup_{x\in\mathbb{R}} |p([-\infty,x]) - q([-\infty,x])| \\ &\leq d_{TV}(p,q). \end{aligned}$$

  The relationship with TV distance comes from the fact that the supremum in the definition of TV actually captures all sets of the form $[-\infty, x]$ (along with all other subsets of the space).

- **Kullback-Leibler (KL) divergence** is another popular way of measuring statistical distance. KL divergence is also known as relative entropy. We use $d_{KL}(p||q)$ to note the KL divergence from p to q.

$$d_{KL}(p||q) = \int_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}.$$

  Note that the KL divergence is not strictly a distance function, as it is not symmetric.

- $\boldsymbol{\chi^2}$ **distance** is derived from the well-known Pearson's χ^2 test statistic $\chi^2(p,q) = \sum_{x\in\mathcal{X}} \frac{(p(x)-q(x))^2}{p(x)}$, but is symmetric for p and q. χ^2 distance is often used in computer vision to estimate the distances between bag-of-visual-word representations of images. We use d_{χ^2} to denote χ^2 distance in this note.

$$d_{\chi^2}(p,q) = \frac{1}{2}\sum_{x\in\mathcal{X}} \frac{(p(x)-q(x))^2}{p(x)+q(x)}.$$

  χ^2 distance is also often used in constructing a kernel function out of histogram distances.

- **Hellinger distance** is yet another notion of statistical distance that we use in the proofs in this lecture. The *squared* Hellinger distance between two distributions P and Q is given by:

$$H(P,Q) = \int_{x\in\mathcal{X}} \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2.$$

  A crucial property of the Hellinger distance that we exploit in our proofs is that:

$$H(P,Q)^2 \leq d_{TV} \leq \sqrt{2}H(P,Q),$$

  a property which follows directly from the relationship between ℓ_1 and ℓ_2 norms, and the Cauchy-Schwarz inequality[1].

[1] For a step-by-step proof of this inequality, we direct the reader to http://www.tcs.tifr.res.in/~prahladh/teaching/2011-12/comm/lectures/l12.pdf

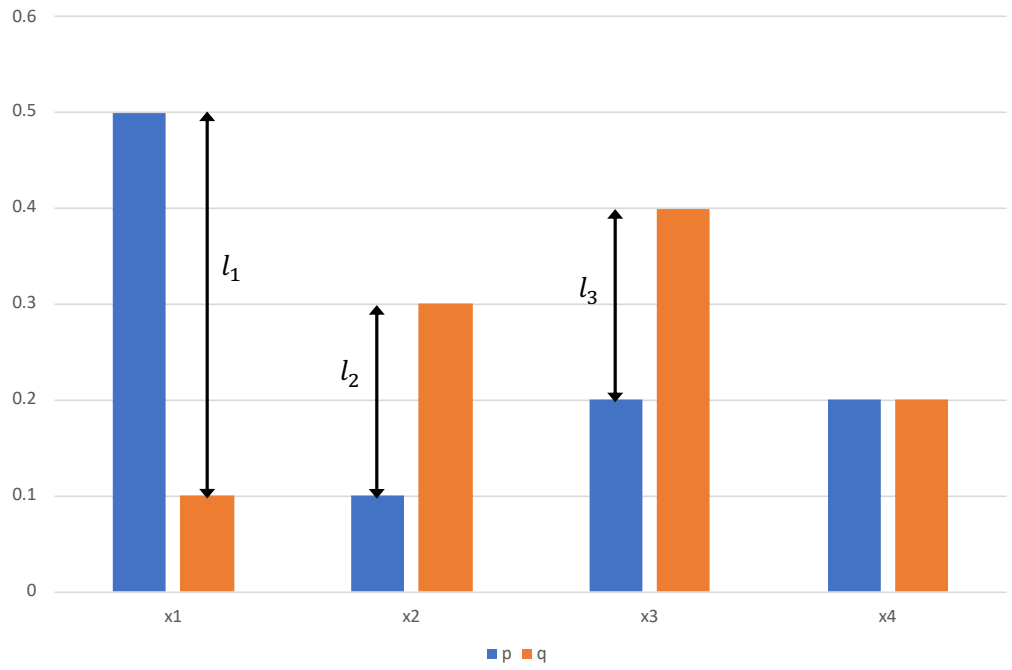

Figure 1: An example of total variation distance between two distributions.

It's worth noting that, d_{TV}, d_{KL}, and d_{χ^2} can be seen as special cases of general f-distances, which are defined by

$$d_f = \sum_{x \in \mathcal{X}} f\left(\frac{p(x)}{q(x)}\right),$$

where the corresponding function f for d_{TV}, d_{KL}, and d_{χ^2} are

$$\begin{aligned} f_{TV}(t) &= \frac{1}{2}|t-1| \\ f_{KL}(t) &= t \log t \\ f_{\chi^2}(t) &= \frac{(t-1)^2}{t+1}. \end{aligned}$$

Another useful result about statistical distances is Pinsker's inequality, which is formally stated as

Theorem 1 *If p and q are two probability distributions on a measurable space $\mathcal{X}$, then*

$$d_{TV}(p,q) \leq \sqrt{\frac{1}{2} d_{KL}(p,q)}.$$

Pinsker's inequality is known not to be tight. People have made improvements and generalizations of it [9]. For the proof of Pinsker's inequality, please refer to [8].

In this lecture we measure distances with respect to the total variation distance. The total variance distance represents the largest possible difference between the probabilities that two probability distributions can assign to the same event.

For example, Figure 1 shows two distribution over x_1, x_2, x_3, and x_4. The total variance d_{VT} measures half of the sum of differences of the probability iterating all possible events in either distributions, a.k.a., $\frac{1}{2}(l_1 + l_2 + l_3)$.

Since for every distribution d over a measurable set $\mathcal{X}$, we have $\sum_{x \in \mathcal{X}} d(x) = 1$, the effective difference are counted twice. Therefore we reduce the sum by half.

3 Goodness of Fit

One of the central problem of hypothesis testing is that of *goodness of fit*. The goodness of fit of a model describes how well a model fits a set of observations. Concretely, the goodness of fit problem can be formulated as follows:

We are given sample access to some unknown high-dimensional distribution $p \in \Delta(\mathcal{X})$, an explicit distribution $q \in \Delta(\mathcal{X})$, and some tolerance $\varepsilon > 0$. Our goal is to distinguish with probability of error $\leq \frac{1}{3}$, between $p = q$ v.s. $d(p,q) \geq \varepsilon$, for some notion of statistical distance d.

Goodness of fit is also considered in the form of the *tolerant goodness of fit* problem, in which we instead have two thresholds, ε_1 and ε_2. Our goal is to distinguish $d(p,q) \leq \varepsilon_1$ v.s. $d(p,q) \geq \varepsilon_2$.

Example 1: Testing fairness of a coin

To illustrate the goodness of fit problem, suppose that we are given a coin with an unknown weighting and we want to find out if the coin is a fair coin. Mathematically, you want to find out if the distribution of flip outcomes generated using that coin is:

$$\mathcal{X} = \{0,1\} \qquad q = \text{Bernouli}(1/2)$$

How many coin tosses do we need to test this hypothesis? It turns out that $\Theta\left(\frac{1}{\varepsilon^2}\right)$ samples are both necessary and sufficient for computing both goodness of fit, and tolerant goodness of fit (in terms of total variational distance).

Proof We prove this in two parts, first showing necessity and next, sufficiency.

Necessity. Suppose there exists a test using k samples that can distinguish with probability $\geq \frac{2}{3}$ a sample from $X = (x_1, x_2, \ldots, x_k) \sim P := \text{Bernoulli}\left(\frac{1}{2}\right)$ from a sample from $Y = (y_1, y_2, \ldots, y_k) \sim Q := \text{Bernoulli}\left(\frac{1}{2} + \varepsilon\right)$.

From our introduction of Hellinger distance above, we have the following upper bound on d_{TV}:

$$d_{TV}(X,Y) \leq 2H^2(X,Y) \tag{1}$$

$$= 2\left(\int_{x \in \{0,1\}^k} \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2\right) \tag{2}$$

$$\leq 2k\left(\int_{x \in \{0,1\}} \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2\right) \tag{3}$$

$$\in o\left(k\varepsilon^2\right) \tag{4}$$

Thus, in order for our test to be able to differentiate in TV distance between P and Q, we need $k \geq \Omega\left(\frac{1}{\varepsilon^2}\right)$.

Sufficiency. To show that $\Omega\left(\frac{1}{\varepsilon^2}\right)$ samples are sufficient, we use the binomial Chernoff bound (we omit the proof of this bound, which can be found in [10]).

In particular, we sample $n = \frac{1}{\varepsilon^2}$ results $x_1, x_2, \ldots, x_n$ and add them up. If the sum $X = \sum x_i$ is less than $\left(\frac{1}{2} + \frac{1}{2}\varepsilon\right) n = \frac{1}{2\varepsilon^2} + \frac{1}{2\varepsilon}$, then we will predict that the coin is fair, a.k.a., the underlying distribution is Bernouli$\left(\frac{1}{2}\right)$. Otherwise it is the biased distribution Bernouli$\left(\frac{1}{2} + \varepsilon\right)$.

Specifically, denote the sum of these $\frac{1}{\varepsilon^2}$ samples as X, if we are tossing the fair coin, we have the expected value of the sum as $\frac{1}{2\varepsilon^2}$. We can use a Chernoff bound to bound the probability that X deviates from its expected value:

$$\Pr\left[X > (1+\delta)\frac{1}{2\varepsilon^2}\right] \leq e^{-\frac{\delta^2}{2+\delta}\frac{1}{\varepsilon^2}} \tag{5}$$

$$\Rightarrow \Pr\left[X > (1+\varepsilon)\frac{1}{2\varepsilon^2}\right] \leq e^{-\frac{\varepsilon^2}{2+\varepsilon}\frac{1}{\varepsilon^2}} = e^{-\frac{1}{2+\varepsilon}} \sim e^{-\frac{1}{2}}, \tag{6}$$

We can repeat this sampling process, a.k.a., by using $K \cdot \frac{1}{\varepsilon^2}$ samples to exponentially decrease the constant on the right hand side of the inequality. ■

Example 2: Uniformity testing

Now we consider a more general problem. Instead of a Bernoulli distribution, now q is a uniform distribution over a discrete set $\mathcal{X}$ where $m \equiv |\mathcal{X}|$. Now we formally define the problem in the context of hypothesis testing.

Setting Consider the null hypothesis where p is a uniform distribution over support m

$$H_0 : p_i = \frac{1}{m}$$

and the non-parametric alternative:

$$H_A : d_{TV} = \sum_{i=1}^{m} |p(i) - \frac{1}{m}| > \varepsilon$$

Goal Given a unknown distribution p, how many samples do we need to have a consistent test distinguishing H_0 and H_A?

It's intuitive to think that the sample size needed for test I will increase as $|\mathcal{X}|$ increases. Indeed, we have the following results from Paninski [5].

Theorem 2 *In test I with $d = d_{TV}$, we need $\Theta\left(\frac{\sqrt{|\mathcal{X}|}}{\varepsilon^2}\right)$ samples from p to decide that p is ε-distant from uniform distribution q.*

Upper Bound

Our uniformity test will be based on "coincidences", that is, bins i for which more than one sample is observed. The basic idea, as in the birthday inequality [6], is that deviations from uniformity necessarily lead to an increase in the expected number of coincidences when we sample from the distribution. We define $m = |\mathcal{X}|$ as the total number of bins, K_1 as the the number of bins into which just one sample has fallen, and N as the sample size. For a fixed N, K_1 is negatively related to the number of coincidences. To see this, we can write the expectation of K_1 under a given distribution p as:

$$\mathbb{E}_p[K_1] = \sum_{i=1}^{m} \binom{N}{1} p_i (1 - p_i)^{N-1}.$$

where p_i is the probability a sample fall into bin i. In a uniform case, $p_i = \frac{1}{m}$ and

$$\mathbb{E}_u[K_1] = N\left(\frac{m-1}{m}\right)^{N-1}$$

We can compare the two expectations by computing the difference:

$$\mathbb{E}_u[K_1] - \mathbb{E}_p[K_1] = N\frac{m-1}{m}\sum_{i=1}^{m} p_i \left[1 - \left(\frac{m}{m-1}(1-p_i)\right)^{N-1}\right]$$

After some approximations and an application of Jensen's inequality, we have the following key lower bound on $\mathbb{E}[K_1]$ in terms of the distance from uniformity ε.

Lemma 3 $\mathbb{E}_u[K_1] - \mathbb{E}_p[K_1] \geq \frac{N^2\varepsilon^2}{m}\left[1 + O\left(\frac{N}{m}\right)\right], \forall p \in H_A$

Proof Let $f(p_i) = p_i\left[1 - \left(\frac{m}{m-1}(1-p_i)\right)^{N-1}\right]$. then we have:

$$\mathbb{E}_u[K_1] - \mathbb{E}_p[K_1] = N\left(\frac{m-1}{m}\right)\sum_{i=1}^{m} f(p_i)$$

Since $f(x)$ is not convex, we develop a convex lower bound on $f(x)$, valid for all $x \in [0,1]$ when $N \leq m$:

$$f(x) \geq g\left(|x - \frac{1}{m}|\right) + f'\left(\frac{1}{m}\right)\left(x - \frac{1}{m}\right)$$

where

$$g(z) = \begin{cases} f\left(z + \frac{1}{m}\right) - f'\left(\frac{1}{m}\right)z & z \in \left[0, \frac{1}{N} - \frac{1}{m}\right] \\ f\left(\frac{1}{N}\right) + \left(z + \frac{1}{m} - \frac{1}{N}\right) - f'\left(\frac{1}{m}\right)z & \text{o.w.} \end{cases}$$

The derivative can be computed to be:

$$f'\left(\frac{1}{m}\right) = \frac{N-1}{m-1}$$

Putting things together, we can bound $f(p_i)$ as:

$$\begin{aligned} \sum_i f(p_i) &\geq \sum_i \left[g\left(|x - \frac{1}{m}|\right) + f'\left(\frac{1}{m}\right)\left(x - \frac{1}{m}\right)\right] \\ &= \sum_i g\left(|x - \frac{1}{m}|\right) + \sum_i \frac{N-1}{m-1}\left(p_i - \frac{1}{m}\right) \\ &= \sum_i g\left(\left|x - \frac{1}{m}\right|\right) + \frac{N-1}{m-1}\left[\left(\sum_i p_i\right) - 1\right] \\ &= \sum_i g\left(\left|x - \frac{1}{m}\right|\right) \end{aligned}$$

Applying Jensen's inequalities, we can get:

$$\frac{1}{m}\sum_i g\left(\left|x - \frac{1}{m}\right|\right) \geq g\left(\frac{1}{m}\sum_i \left|p_i - \frac{1}{m}\right|\right) \geq g\left(\frac{\varepsilon}{m}\right)$$

where the last inequality is by the fact that g is increasing and $p \in H_A$. Near 0, we can approximate g as:

$$g(z) = \left(N + O\left(\frac{N^2}{m}\right)\right)z^2 + o\left(z^2\right)$$

We use this convex lower bound to bound the difference in expectation:

$$\begin{aligned} \mathbb{E}_u[K_1] - \mathbb{E}_p[K_1] &\geq Nm\left(\frac{m-1}{m}\right)^{N-1} g\left(\frac{\varepsilon}{m}\right) \\ &= \frac{N^2\varepsilon^2}{m}\left[1 + O\left(\frac{N}{m}\right)\right] \end{aligned}$$

which completes the proof. ■

On the other hand, we can also bound the variance of K_1 under p as follows:

Lemma 4

$$Var_p(K_1) \leq \mathbb{E}_u[K_1] - \mathbb{E}_p[K_1] + O\left(\frac{N^2}{m}\right)$$

Proof $Var_p(K_1)$ can be directly computed as:

$$Var_p(K_1) = \mathbb{E}_p[K_1] - \mathbb{E}_p[K_1]^2 + N(N-1)\sum_{i\neq j} p_i p_j (1-p_i-p_j)^{N-2}$$

However,this is hard to be bounded directly. Instead, we can use Efron-Stein inequality [11]:

$$Var(S) \leq \frac{1}{2}\mathbb{E}\left[\sum_{j=1}^{N}\left(S - S^{(i)}\right)^2\right]$$

where S is an arbitrary function of N independent random variable's x_i and

$$S^{(i)} = S(x_1, x_2, ..., x_i', ..., x_N)$$

denotes that S computed with x_i' substituted for x_i, where x_i' is an i.i.d. copy of x_i. We can apply this inequality to $S = K_1$, with x_i being the independent samples from p. We denote n_i as the number of samples observed to have fallen in bin i after N-1 samples have been drawn. Thus,

$$\begin{aligned}
\frac{1}{2}\mathbb{E}\left[\sum_{j=1}^{N}\left(S - S^{(i)}\right)^2\right] &= \frac{N}{2}\mathbb{E}_{\{x_i\}_{1\leq i\leq N-1}\sim p}\left[\sum_{i\leq j, j\leq m} p_i p_j \left(1(n_i = 0 \cap n_j > 0) + 1(n_j = 0 \cap n_i > 0)\right)\right] \\
&= N\sum_{i,j} p_i p_j \mathbb{P}_{\{x_i\}_{1\leq i\leq N-1}\sim p}(n_i = 0 \cap n_j > 0) \\
&= N\sum_{i,j} p_i p_j \left((1-p_i)^{N-1}\left(1 - \left(1 - \frac{p_j}{1-p_i}\right)^{N-1}\right)\right) \\
&= N\sum_{i,j} p_i p_j \left((1-p_i)^{N-1} - (1-p_i-p_j)^{N-1}\right) \\
&\leq N\sum_{j=1}^{m} p_j \left(1 - (1-p_j)^{N-1}\right) \\
&= \mathbb{E}_u[K_1] - \mathbb{E}_p[K_1] + N\left(1 - \left(\frac{m-1}{m}\right)^{N-1}\right) \\
&= \mathbb{E}_u[K_1] - \mathbb{E}_p[K_1] + O\left(\frac{N^2}{m}\right)
\end{aligned}$$

The inequality uses the fact that $(1-y)^n - (1-y-x)^n$ is a decreasing function of y for $n > 1, x \in [0,1]$, and $0 < y < 1 - x$. ■

Now we can construct our hypothesis test. Let $T \equiv \mathbb{E}_u[K_1] - K_1 = N\left(\frac{m-1}{m}\right)^{N-1} - K_1$. We reject null hypothesis H_0 if $T > T_\alpha$ for some threshold T_α.

Theorem 5 *The size of this test is*

$$\mathbb{P}_u(T \geq T_\alpha) = O\left(\frac{N^2}{mT_\alpha^2}\right)$$

. The power is greater than

$$\mathbb{P}_p(T \geq T_\alpha) = 1 - \frac{\mathbb{E}_u[K_1] - \mathbb{E}_p[K_1] + O\left(\frac{N^2}{m}\right)}{\left(\mathbb{E}_u[K_1] - \mathbb{E}_p[K_1] - T_\alpha\right)^2}$$

uniformly over all alternatives $p \in H_A$. If $\frac{N^2\varepsilon^4}{m} \longrightarrow \infty$, then the threshold T_α may be chosen so that the size tends to zero and the power to one, uniformly over all $p \in H_A$ (i.e., this condition is sufficient for the test to be uniformly consistent). A sufficient choice of T_α is

$$T_\alpha = \frac{N^2\varepsilon^2}{2m}$$

Proof Since we have $\mathbb{E}_u[T] = 0, Var_u[T] = O\left(\frac{N^2}{m}\right)$ from lemma 4, by Chebysheff the size is bounded by

$$\mathbb{P}_u(T \geq T_\alpha) = O\left(\frac{N^2}{mT_\alpha^2}\right)$$

For the power, again from lemma 4, we have that

$$\begin{aligned}\mathbb{P}_p(T < T_\alpha) &= \mathbb{P}_p(T - \mathbb{E}_p[T] < T_\alpha - \mathbb{E}_p[T]) \\ &\leq \frac{\mathbb{E}_p[T] + O\left(\frac{N^2}{m}\right)}{(\mathbb{E}_p[T] - T\alpha)^2}\end{aligned}$$

■

Thus we have showed that $\frac{N^2\varepsilon^4}{m} \longrightarrow \infty$ (that is, $N = O\left(\frac{\sqrt{m}}{\varepsilon^2}\right)$) is a sufficient condition for the existence of a uniformly consistent test of H_0 vs. H_A.

Lower Bound

Next we derive a lower bound on N to guarantee the consistency of any test.

Theorem 6 *If $\frac{N^2}{m}$ remains bounded, then no test reliably distinguishes H_0 from H_A.*

Proof W.l.o.g.,we assume m is even. The following bound is developed for one particular tractable mixing measure μ. We choose q randomly according to the following distribution $\mu(q)$: first we choose $\frac{2}{m}$ independent Bernoulli random variables $z_j \in \{-1, +1\}$ (i.e. z samples uniformly from the corners of the $\frac{m}{2}$-dimensional hypercube). Given $\{z_j\}$, set

$$q(i) = \begin{cases} \left(1 + \varepsilon z_{\frac{i}{2}}\right)/m & \text{i even} \\ \left(1 - \varepsilon z_{\frac{(i+1)}{2}}\right)/m & \text{i odd} \end{cases}$$

Let n_i denote the number of samples observed to have fallen into the i-th bin. We can write out the ratio of marginal likelihoods as:

$$\begin{aligned}\frac{L(\hat{n}|H_A)}{L(\hat{n}|H_0)} &= \mathbb{E}_{\hat{z}}\left[\prod_{i=2,4,..,m}\left(1 - z_{\frac{i}{2}}\varepsilon\right)^{n_{i-1}}\left(1 + z_{\frac{i}{2}}\varepsilon\right)^{n_i}\right] \\ &= \prod_{i=2,4,..,m}\mathbb{E}\left[\left(1 - z_{\frac{i}{2}}\varepsilon\right)^{n_{i-1}}\left(1 + z_{\frac{i}{2}}\varepsilon\right)^{n_i}\right] \\ &= \prod_{i=2,4,..,m}\frac{1}{2}\left[(1+\varepsilon)^{n_{i-1}}(1-\varepsilon)^{n_i} + (1+\varepsilon)^{n_i}(1-\varepsilon)^{n_{i-1}}\right] \\ &= \prod_{i=2,4,..,m}\left(1-\varepsilon^2\right)^{m_i}\left((1-\varepsilon)^{d_i} + (1+\varepsilon)^{d_i}\right)/2 \\ &= \prod_{i=2,4,..,m}\left(1-\varepsilon^2\right)^{m_i}\left(1 + \binom{d_i}{2}\varepsilon^2 + \binom{d_i}{4}\varepsilon^4 + ...\right)\end{aligned}$$

The second equality follows since z_j are i.i.d. In the fourth equality, we use abbreviations $m_i = \min(n_i, n_{i-1})$ and $d_i = |n_i - n_{i-1}|$. We interpret $\binom{d_i}{k}$ as 0 whenever $d_i < k$.

Note that the above multiplicands are greater than 1 only if $d_i \geq 2$, and less than 1 only if $m_i \geq 1$. Since the number of "two-bin coincidences" (pairs of bins into which two or more samples have fallen) is bounded in probability if $N = O(\sqrt{m})$, the likelihood ratio is bounded in probability as well,implying that the error probability of any test is bounded away from zero, and the proof is complete.

Finally, it is worth noting that the expected numbers of events $(m_i = 1, d_i = 0)$ and $(m_i = 0, d_i = 2)$ scale together, leading (after an expansion of the logarithm and a cancellation of the ε^2 terms) to exactly the $\frac{N^2\varepsilon^4}{m}$ scaling we observed previously. This implies that N must grow at least as quickly as $\frac{\sqrt{m}}{\varepsilon^2}$ to guarantee the consistency of any test. ■

Valiant and Valiant [3] construct a linear estimator for estimating distance to uniformity. They showed that given $\Theta\left(\frac{\sqrt{|\mathcal{X}|}}{\varepsilon^2}\right)$ independent samples from from a distribution of any support, their estimator will compute the TV distance to Uniform$(\mathcal{X})$ to within accuracy ε, with high probability. This is essential test II where $\varepsilon = \varepsilon_2 - \varepsilon_1$. Thus, we re-frame the results as the following theorem.

Theorem 7 *In test II with $d = d_{TV}$, we need $\Theta\left(\frac{|\mathcal{X}|}{\log|\mathcal{X}|\varepsilon^2}\right)$ samples from p to decide that p is ε-distant from uniform distribution q.*

Note that the same bounds apply for estimating symmetric properties of a distribution such as entropy and support size.

4 Goodness of Fit Testing

In a more general setting, can we test goodness of fit in terms of total variation distance if q is an unknown distribution over discrete support $\mathcal{X}$? Acharya-Dashalakis-Kamath '15 [2] develop a general testing framework which leads to the following results:

1. There exists an efficient test using $\mathcal{O}\left(\frac{\sqrt{|\mathcal{X}|}}{\varepsilon^2}\right)$ samples
2. This test requires $\Omega\left(\frac{\sqrt{|\mathcal{X}|}}{\varepsilon^2}\right)$ samples

Canonne and Diakonikolas [11] also showed similar results in their paper.

4.1 Upper Bound

Poisson Sampling In this proof, we use the standard Poissonization approach. Instead of drawing exactly m samples from a distribution p, we first draw $m' \sim Poisson(m)$, and then draw m' samples from p. As a result, the number of times different elements in the support of p occur in the sample become independent, giving much simpler analyses. In particular, the number of times we will observe domain element i will be distributed as $Poisson(mp_i)$, independently for each i. Since $Poisson(m)$ is tightly concentrated around m, this additional flexibility comes only at a sub-constant cost in the sample complexity with an inversely exponential in m, additive increase in the error probability.

The idea is to divide effective support into several intervals of roughly equal measure. It computes the statistic over each of these intervals, and we let our statistic Z be the sum of all but the largest t of these values. In the case when $p = q$, Z will only become smaller by performing this operation. We use Kolmogorov's maximal inequality to show that Z remains large when $d_{TV}(p, q) \geq \varepsilon$. The Algorithm is described in Figure 3, and the statistics Z is defined as:

$$Z = \sum_{i \in \mathcal{A}} \frac{(N_i - m\dot{q}_i)^2 - N_i}{m\dot{q}_i}$$

The terms $-N_i$ in the numerator is a correction for classical test statistics. It prevents Z from exploding when q_i is sufficiently small.

Claim 8 $\mathbb{E}[Z] = m\dot{\chi}^2(p, q)$ *where* $\chi^2(p, q) = \sum_i \frac{(p_i - q_i)^2}{q_i}$

Algorithm 1 Chi-squared testing algorithm

Input: ε; an explicit distribution q; (Poisson) m samples from a distribution p, where N_i denotes the number of occurrences of the ith domain element.
$\mathcal{A} \leftarrow \{i : q_i \geq \varepsilon/50n\}$
$Z \leftarrow \sum_{i \in \mathcal{A}} \frac{(N_i - mq_i)^2 - N_i}{mq_i}$
if $Z \leq m\varepsilon^2/10$ **then**
    return ACCEPT
else
    return REJECT
end if

Figure 2: chi-squared testing algorithm

Proof

$$\begin{aligned}
\mathbb{E}[Z] &= \sum_i \frac{\mathbb{E}[N_i]^2 - 2mq_i\mathbb{E}[N_i] + m^2q_i^2 - mp_i}{mq_i} \\
&= \sum_i \frac{m^2p_i^2 + mp_i - 2m^2q_ip_i + m^2q_i^2 - mp_i}{mq_i} \\
&= m\sum_i \frac{(p_i - q_i)^2}{q_i} \\
&= m\chi^2(p, q)
\end{aligned}$$

■

We demonstrate the separation in the means of the statistic Z in the two hypothesis of interest:

1. If $d_{TV}(p, q) = 0 \Rightarrow \chi^2(p, q) = 0 \Rightarrow \mathbb{E}[Z] = 0$ (1)

2. If $d_{TV}(p, q) \geq \epsilon \Rightarrow \chi^2(p, q) \geq 4\varepsilon^2 \Rightarrow \mathbb{E}[Z] \geq 4m\varepsilon^2 \approx \sqrt{|\mathcal{X}|}$ (2)

Claim 9

$$Var[Z] = \sum_i \left[2\frac{p_i^2}{q_i^2} + 4m\frac{p_i(p_i - q_i)^2}{q_i^2} \right]$$

Proof See Acharga-Dashalahis-Kamath NIPS'15 [1] for details. ■

We can bound the variance as:

$$Var[Z] \leq 4|\mathcal{X}| + 9\sqrt{|\mathcal{X}|}\mathbb{E}[Z] + \frac{2}{5}|\mathcal{X}|^{\frac{1}{4}}\mathbb{E}[Z]^{\frac{3}{2}}$$

We demonstrate the separation in the variance of the statistic Z in the two hypothesis of interest:

1. If $p = q \Rightarrow \mathbb{E}[Z] = 0 \Rightarrow Var[Z] \leq 4|\mathcal{X}|$ (3)

2. If $d_{TV}(p, q) \geq \epsilon \Rightarrow \mathbb{E}[Z] \geq \sqrt{|\mathcal{X}|} \Rightarrow Var[Z] \leq \left(\mathbb{E}[Z]^2\right)$ (4)

Together with conditions (1), (2), (3), and (4), we can distinguish between $p = q$ and $d_{TV}(p, q) \geq \varepsilon$ as in figure 3.

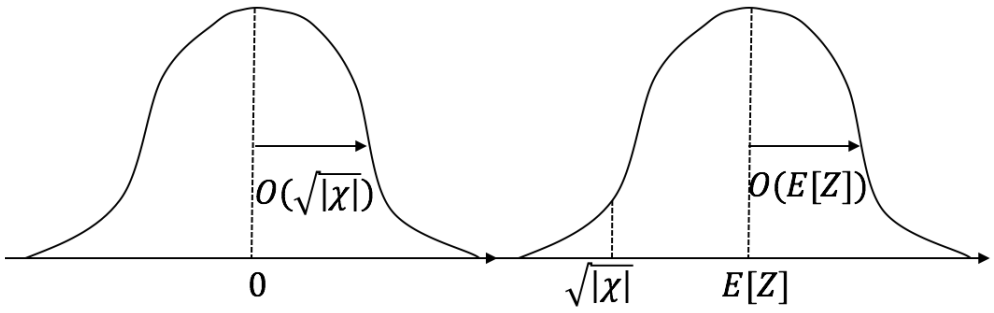

Figure 3: The left plot is the case where $p = q$. The right plot is the case where $d_{TV}(p, q) \geq \varepsilon$.

4.2 Lower Bound

The example studied by Paninski [5] to prove lower bounds on testing uniformity can be used to prove lower bounds for the classes we consider.This problem is not easier than distinguishing U_n from $\mu(q)$ as described in Theorem 9, where U_n is the uniform distribution over n. Therefore, by invoking Paninski's proof, we immediately get the lower bound of $\Omega\left(\frac{\sqrt{|\mathcal{X}|}}{\varepsilon^2}\right)$ samples.

5 Bad news for high-dimensional distributions

The above bounds imply exponential sample complexity lower bounds in high-dimensions.

Claim 10 *Suppose* $q = \textsc{Uniform}(\{0,1\}^n)$. *We are given sample access to* $p \in \Delta(\{0,1\}^n)$.

- $\Omega\left(\frac{2^{\frac{n}{2}}}{\varepsilon^2}\right)$ *samples are needed to distinguish whether* $p = q$ *vs* $d_{TV}(p, q) \geq \varepsilon$ *(or wasserstein* $(p, q) \geq \varepsilon$*)*
- $\Omega\left(\frac{2^n}{n\varepsilon^2}\right)$ *samples are needed to distinguish whether* $d_{TV}(p, q) \leq \frac{\varepsilon}{2}$ *vs* $d_{TV}(p, q) \geq \varepsilon$

Proof

- World 1: $p = \textsc{Uniform}(\{0,1\}^n)$
- World 2: fix a matching M of the vertices of a hyper-cube $\{0,1\}^n$. Thus, if $(u, v) \in M$, set $P_u = \frac{1+\varepsilon}{2^n}, p_v = \frac{1-\varepsilon}{2^n}$ with probability $\frac{1}{2}$. Otherwise, set $P_u = \frac{1-\varepsilon}{2^n}, p_v = \frac{1+\varepsilon}{2^n}$ with probability $\frac{1}{2}$.

By rewriting the previous construction, it can be seen that we cannot distinguish between world 1 and world 2 with less than $\left(\frac{2^{\frac{n}{2}}}{\varepsilon^2}\right)$ samples.

■

5.1 Back to GANs

The exponential sample complexity lower bounds are disappointing news for GANs. If uniformity cannot be tested from a practically feasible number of samples, are there other things that can be tested?

Recall that in world 2, since there are an exponential number of edges in the hyper-cube, it requires an exponential number of bits to index a distribution in the set. However, most real world distributions are not parametric. What if the distributions that are sampled and those we test against have low dimensional structures, like Markov Random Fields or Bayesian Networks, that can be exploited? Daskalakis,

Dikkala, & Kamath '18 [1] shows that under both Ising and Bayes net assumptions, goodness-of-fit testing requires only polynomial many samples in dimension, avoiding the curse of dimensionality.

The million dollar open research question that still remains is what is a reasonale structural assumption for real world distributions that combined with the structure of generators would allow us to rigorously test whether a real world distribution and the output of a generator are close.

References

[1] Daskalakis, C., Dikkala, N., & Kamath, G. (2018). Testing Ising Models. Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'18), 1989?2007.

[2] Acharya, J., Daskalakis, C., & Kamath, G. (2015). Optimal Testing for Properties of Distributions. Retrieved from http://arxiv.org/abs/1507.05952

[3] Valiant, P., & Valiant, G. (2013). Estimating the unseen: improved estimators for entropy and other properties. Advances in Neural Information Processing Systems, 2157-2165. https://doi.org/10.1145/3125643

[4] Valiant, G., & Valiant, P. (2011). The power of linear estimators. Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS, 403-412. https://doi.org/10.1109/FOCS.2011.81

[5] Paninski, L. (2008). A coincidence-based test for uniformity given very sparsely sampled discrete data. IEEE Transactions on Information Theory, 54(10), 4750âĂŞ4755. https://doi.org/10.1109/TIT.2008.928987

[6] Bloom, D. (1973). A birthday problem. The American Mathematical Monthly, 80:1141âĂŞ1142.

[7] Steele, J. (1986). An Efron-Stein inequality for nonsym- metric statistics. Annals of Statistics, 14:753âĂŞ758.

[8] M.S. Pinsker. (1964) Information and Information Stability of Random Variables and Processes. Holden-Day.

[9] Reid, M. D., & Williamson, R. C. (2009). Generalised pinsker inequalities. arXiv preprint arXiv:0906.1244.

[10] Hellman, M., & Raviv, J. (1970). Probability of error, equivocation, and the Chernoff bound. IEEE Transactions on Information Theory, 16(4), 368-372.

[11] ClÃľment L. Canonne, Ilias Diakonikolas, Themis Gouleakis, Ronitt Rubinfeld: Testing Shape Restrictions of Discrete Distributions. STACS 2016: 25:1-25:14

[12] Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. Retrieved from http://arxiv.org/abs/1701.07875