Topics and Techniques in Distribution Testing

A Biased but Representative Sample

Suggested Citation: Clément L. Canonne (2018), "Topics and Techniques in Distribution Testing", : Vol. xx, No. xx, pp 1–18. DOI: 10.1561/XXXXXXXXX.

Clément L. Canonne
University of Sydney
clement.canonne@sydney.edu.au

now
the essence of knowledge
Boston — Delft

Contents

Topics and Techniques in Distribution Testing

Clément L. Canonne[1]

[1] *University of Sydney; clement.canonne@sydney.edu.au*

ABSTRACT

We focus on some specific problems in distribution testing, taking goodness-of-fit as a running example. In particular, we do not aim to provide a comprehensive summary of all the topics in the area; but will provide self-contained proofs and derivations of the main results, trying to highlight the unifying techniques.

Clément L. Canonne (2018), "Topics and Techniques in Distribution Testing", : Vol. xx, No. xx, pp 1–18. DOI: 10.1561/XXXXXXXXX.

Nomenclature

$a_n \lesssim b_n$ There exists a constant $C > 0$ such that $a_n \leq C \cdot b_n$ for every n. Similar to $a_n = O(b_n)$, but not necessarily asymptotic, and easier to write.

$a_n \asymp b_n$ Both $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Similar to $a_n = \Theta(b_n)$, but not necessarily asymptotic, and easier to write.

$a_n = O(b_n)$ There exist a constant $C > 0$ and a value n_0 such that $a_n \leq C \cdot b_n$ for every $n \geq n_0$.

$a_n = \Omega(b_n)$ There exist a constant $C > 0$ and a value n_0 such that $a_n \geq C \cdot b_n$ for every $n \geq n_0$. This is equivalent to $b_n = O(a_n)$.

$a_n = \Theta(b_n)$ Both $a_n = \Omega(b_n)$ and $a_n = O(b_n)$.

$a_n \underset{n\to\infty}{\sim} b_n$ Asymptotic equivalence: $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$ (stronger than $a_n = \Theta(b_n)$, as the "hidden constant" is 1).

$a_n \gg b_n$ (Informal) a_n is much larger than, or "sufficiently" large compared to, b_n.

log, ln Throughout, log denotes the logarithm in base 2 and ln the natural logarithm.

1

What is distribution testing?

This survey serves as an introduction and detailed overview of some topics in (probability) distribution testing, an area of theoretical computer science which falls under the general umbrella of *property testing*, and sits at the intersection of computational learning, statistical learning and hypothesis testing, information theory, and (depending whom one asks) the theory of machine learning. Broadly speaking, distribution testing is concerned with the following type of questions:

> Given a **small** number of independent data points from some blackbox random source, can we **efficiently** decide whether the distribution of the data follows some purported model ("property"), or is statistically far from doing so?

Of course, there are many details to be made precise here. What type of assumptions on the data do we make – is it discrete, continuous, univariate, high-dimensional? What do we mean by "efficiently" – the number of data points (data efficiency), the running time of our algorithms (time efficiency), both? What do we mean by "far" – what notion of distance are we considering? And what type of error do we

allow – false positives (Type I), false negative (Type II)?

Some of these are left flexible, as we will see below when formally introducing the setting of distribution testing. However, the general idea is to focus on *finite sample guarantees* (no qualitative limiting statements as data size grows to infinity), for a *fixed error probability target* δ controlling both Type I and Type II, and making *as few assumptions as possible* under the (composite) alternative hypothesis. That is, we will answer questions of the form "either the distribution of the data satisfies the property, or it is *pretty much anything* far from that."

Adopting a Computer Science viewpoint, we will also assume that the "size" of the object considered – typically, the domain size for discrete data – is large, which allows us to focus on the first-order dependence on this quantity. This also implies we typically consider a *worst-case* (minimax) setting with respect to this quantity, making statements about the worst-case data size, or time, required to achieve our goal. This does not means the algorithms and ideas obtained do not lead to "practical" algorithms: rather, that people working in distribution testing are quite pessimistic and paranoid in nature, and want the guarantee that things are *never* too slow before the promise that they *often* are quite fast. (Moreover, as we will see later, the worst-case instances for most of our testing tasks are actually quite natural, and likely to arise in practice! Paranoia, for once, may be warranted.)

A note. For simplicity, throughout this survey we will sweep under the rug many measure-theoretic subtleties, and assume probability distributions, probability density functions (pdf), and probability mass functions (pmf) exist whenever required, and are suitably well-behaved. We will also typically identify a probability distribution with its pdf or pmf, and by a slight abuse of notation use $\mathbf{p}$ indifferently for the distribution itself and its pdf. Most, if not all, of those subtleties can be handled by inserting the words "Radon–Nikodym," "measurable," and "counting measure" in suitable places and order.

1.1 Formulation, and relation to Hypothesis Testing

In what follows, $k \in \mathbb{N}$ will be used to parametrize the domain of the probability distributions: namely, Δ_k will denote the set of probability distributions over a (known) domain $\mathcal{X}_k$.

We begin with the notion of distance we will be concerned about, the total variation distance (also known as *statistical distance*):

Definition 1.1 (Total variation distance)**.** The *total variation distance* between two probability distributions $\mathbf{p}, \mathbf{q} \in \Delta_k$ is given by

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = \sup_{S \subseteq \mathcal{X}_k} (\mathbf{p}(S) - \mathbf{q}(S)) \,.$$

Given a subset $\mathcal{C} \subseteq \Delta_k$ of distributions, we further define the distance from $\mathbf{p} \in \Delta_k$ to $\mathcal{C}$ as $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{C}) := \inf_{\mathbf{q} \in \mathcal{C}} \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})$, and will say that $\mathbf{p}$ is ε-*far from* $\mathcal{C}$ if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{C}) > \varepsilon$.

One can check that d_{TV} defines a metric on Δ_k, and takes values in $[0, 1]$. Moreover, the total variation distance exhibits several important properties, some of which will be detailed at length in Appendix B; we recall a crucial one below.

E: Check it!

Fact 1.1 (Data Processing Inequality)**.** Suppose X and Y are independent random variables with distributions $\mathbf{p}$ and $\mathbf{q}$, and let f be any (possibly randomized) function independent of X, Y. Then the probability distributions $\mathbf{p}_f$ and $\mathbf{q}_f$ of $f(X)$ and $f(Y)$ satisfy

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_f, \mathbf{q}_f) \leq \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \,.$$

That is, *postprocessing cannot increase the total variation distance.*

Assuming that $\mathbf{p}, \mathbf{q}$ are absolutely continuous with respect to some dominating measure μ,

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \int \left| \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}\mu} - \frac{\mathrm{d}\mathbf{q}}{\mathrm{d}\mu} \right| \mathrm{d}\mu \tag{1.1}$$

In the discrete case where $\mathbf{p}, \mathbf{q}$ are both over $\mathbb{N}$ or a finite domain, this leads to

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \|\mathbf{p} - \mathbf{q}\|_1 \tag{1.2}$$

that is, "total variation is half the ℓ_1 distance between pmfs." This turns out to be a very useful connection, since ℓ_p norms are quite well-studied beasts: we get to use our arsenal of geometric inequalities — Hölder, Cauchy–Schwarz, and monotonicity of ℓ_p norms, to name a few.

One last piece of terminology: a *property* of distributions is a predicate we are interested in (*e.g.*, "is the probability distribution unimodal?"). By identifying the predicate with the set of objects which satisfy it, we can equivalently view a property of distributions as a *subset* of probability distributions (typically, with some interesting structure). Which is what we will do: throughout, a property is just an arbitrary subset of distributions we are interested in. With this in hand, we are ready to provide a formal definition of what a "testing algorithm" is.

Definition 1.2 (Testing algorithm)**.** Let $\mathcal{P} = \bigcup_{k=1}^{\infty} \mathcal{P}_k$ and $\mathcal{C} = \bigcup_{k=1}^{\infty} \mathcal{C}_k$ be two properties of probability distributions, where $\mathcal{P}_k, \mathcal{C}_k \subseteq \Delta_k$ for all k; and $n \colon \mathbb{N} \times (0,1] \times (0,1] \to \mathbb{N}$, $t \colon \mathbb{N} \times (0,1] \times (0,1] \to \mathbb{N}$ be two functions. A *testing algorithm for $\mathcal{P}$ under $\mathcal{C}$ with sample complexity n and time complexity t* is a (possibly randomized) algorithm $\mathcal{A}$ which, on input $k \in \mathbb{N}, \varepsilon \in (0,1], \delta \in (0,1]$, and a multiset S of $n(k,\varepsilon,\delta)$ elements of $\mathcal{X}_k$, runs in time at most $t(k,\varepsilon,\delta)$ and outputs $\mathbf{b} \in \{0,1\}$ such that the following holds.

- If S is i.i.d. from some $\mathbf{p} \in \mathcal{P}_k$, then $\Pr_{S,\mathcal{A}}[\,\mathbf{b} = 1\,] \geq 1 - \delta$;
- If S is i.i.d. from some $\mathbf{p} \in \mathcal{C}_k$ such that $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}_k) > \varepsilon$, then $\Pr_{S,\mathcal{A}}[\,\mathbf{b} = 0\,] \geq 1 - \delta$,

where in both cases the probability is over the draw of the i.i.d. sample S from the (unknown) $\mathbf{p}$, and the internal randomness of $\mathcal{A}$.

The *sample complexity of testing $\mathcal{P}$ under $\mathcal{C}$* is then the minimum sample complexity $n(k,\varepsilon,\delta)$ achievable by a testing algorithm.

A few remarks are in order. First, in most of our applications we will take $\mathcal{C}_k = \Delta_k$, so that the unknown distribution $\mathbf{p}$ is *a priori* arbitrary, and the goal is to check whether it belongs to the subset (property) of interest $\mathcal{P}_k$. However, this need not always be the case, and we may want to choose $\mathcal{C}_k$ differently to perform hypothesis testing

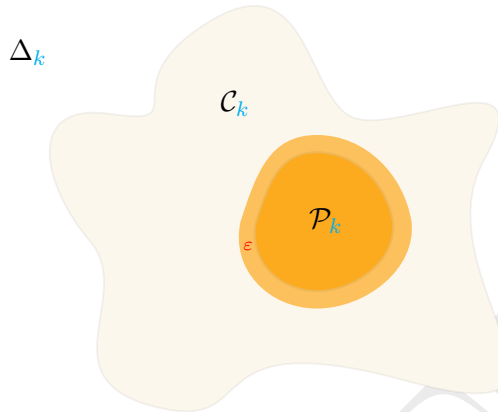

Figure 1.1: An example of property to test. Here, $\mathcal{P}_k \subseteq \mathcal{C}_k \subseteq \Delta_k$, where the property $\mathcal{P}_k$ is depicted as the inner orange area ("yolk"), and the "egg white" is the area of rejection, *i.e.*, the subset of $\mathcal{C}_k$ at total variation distance at least ε from $\mathcal{P}_k$.[1]

under structural assumptions: for instance, to test whether an unknown unimodal distribution is actually Binomial (in this case, $\mathcal{P}_k \subsetneq \mathcal{C}_k \subsetneq \Delta_k$), or if say a log-concave distribution is monotone (in which case there is no inclusion relation between $\mathcal{P}_k$ and $\mathcal{C}_k$, and both are strict subsets of Δ_k).

Another important point is that, while our main focus will be on *discrete* distributions, Definition 1.2 allows for continuous distributions as well. Finally, the above definition is quite flexible, and can be seen to allow for testing *multiple* distributions: for instance, taking $\mathcal{X}_k = [k] \times [k]$, $\mathcal{C}_k := \{ \mathbf{p} \in \Delta_k : \mathbf{p} = \mathbf{p}_1 \otimes \mathbf{p}_2 \}$ (product distributions), and $\mathcal{P}_k := \{ \mathbf{p}_1 \otimes \mathbf{p}_2 \in \mathcal{C}_k : \mathbf{p}_1 = \mathbf{p}_2 \}$, we obtain the question of two-sample testing (a.k.a. closeness testing), which asks to test whether two unknown distributions over $[k]$ are equal, or far from each other.

Dependence on the error probability δ. Our definition of testing algorithm leaves the error probability δ as a free parameter; however, it

[1] TikZ code for Fig. 1.1 adapted from https://tex.stackexchange.com/a/598086/31516.

is quite common in the distribution testing literature to set it as some arbitrary constant smaller than 1/2 (usually 1/3). Indeed, by a standard argument, an error probability 1/3 can be driven down to arbitrary δ at the price of a $O(\log(1/\delta))$ factor in the sample complexity.

Lemma 1.1 (Error probability amplification)**.** Fix $\mathcal{P}$ and $\mathcal{C}$, and suppose there exists a testing algorithm $\mathcal{A}$ for $\mathcal{P}$ under $\mathcal{C}$ with sample complexity $n(k, \varepsilon, 1/3)$ and time complexity $t(k, \varepsilon, 1/3)$. Then there is a testing algorithm $\mathcal{A}'$ for $\mathcal{P}$ under $\mathcal{C}$ with sample and time complexities $n'(k, \varepsilon, \delta) := n(k, \varepsilon, 1/3)\lceil 18 \ln(1/\delta)\rceil$ and $t'(k, \varepsilon, \delta) := O(t(k, \varepsilon, 1/3) \log(1/\delta))$.

Proof sketch. Fix $\mathcal{P}$, $\mathcal{C}$, $\mathcal{A}$ as in the statement. Given k, ε, and $\delta \in (0, 1]$, let $\mathcal{A}'$ be the algorithm which takes as input a multiset of $n'(k, \varepsilon, \delta)$ elements, partitions it (arbitrarily) in $m := \lceil 18 \ln(1/\delta)\rceil$ disjoint multisets $S_1, \dots, S_m$, runs $\mathcal{A}$ independently on those m multisets with error probability 1/3 to get $\mathbf{b}_1, \dots, \mathbf{b}_m$, and finally outputs the majority answer $\mathbf{b} := \mathbb{1}\{\sum_{i=1}^m \mathbf{b}_i \geq m/2\}$. The running time is dominated by the m executions, giving the claimed $O(m \cdot t(k, \varepsilon, 1/3))$ bound. Thus, it suffices to check that the output is correct with probability at least $1 - \delta$; this in turn follows from a Hoeffding bound (Theorem A.3). Indeed, by assumption, each $\mathbf{b}_i$ is independently correct with some probability $p \geq 2/3$. Letting $X_i \sim \mathrm{Bern}(p)$ be the indicator of the event "$\mathbf{b}_i$ is the correct output," we have

$$\Pr[\,\mathbf{b} \text{ incorrect}\,] = \Pr\left[\frac{1}{m}\sum_{i=1}^{m} X_i < \frac{1}{2}\right] \leq e^{-2(p-1/2)^2 m} \leq e^{-m/18} \leq \delta\,,$$

where we used our setting of m in the last inequality. □

Importantly, this logarithmic dependence is not always the right one: as we will see in Chapter 2, there exist natural problems for which the right dependence on the error probability only scales as $\sqrt{\log(1/\delta)}$.

The learning baseline. Before setting out to design specific algorithms for various testing tasks and analyze their performance, it is good to have some sort of baseline to compare the result to. The most natural one is the *testing-by-learning* approach, which can essentially be summarized as follows: the sample complexity of testing $\mathcal{P} = \bigcup_{k=1}^{\infty} \mathcal{P}_k$

under $\mathcal{C} = \bigcup_{k=1}^{\infty} \mathcal{C}_k$ is at most the sample complexity of, given k, *learning* an arbitrary distribution from $\mathcal{P}_k \cup \mathcal{C}_k$. More specifically, we have the following:

Lemma 1.2 (Testing by Learning)**.** Fix any $\mathcal{P} = \bigcup_{k=1}^{\infty} \mathcal{P}_k$ and $\mathcal{C} = \bigcup_{k=1}^{\infty} \mathcal{C}_k$, and let $n_{\mathcal{L}}(k, \varepsilon, \delta)$ denote the sample complexity of learning an arbitrary probability distribution from $\mathcal{P}_k \cup \mathcal{C}_k \subseteq \Delta_k$ to total variation ε with error probability at most δ. Then, the sample complexity n of testing $\mathcal{P}$ under $\mathcal{C}$ satisfies

$$n(k, \varepsilon, \delta) \leq n_{\mathcal{L}}(k, \tfrac{\varepsilon}{2}, \delta)\,.$$

This is not necessarily achieved by a computationally efficient tester.

Proof. Fix a learning algorithm $\mathcal{A}$ for $\mathcal{P}_k \cup \mathcal{C}_k$ with sample complexity $n := n_{\mathcal{L}}(k, \frac{\varepsilon}{2}, \delta)$. By running it on n i.i.d. samples from $\mathbf{p}$ (which we are promised either belongs to $\mathcal{P}_k$ or $\mathcal{C}_k$), we obtain a distribution $\hat{\mathbf{p}}$ such that $\mathrm{d}_{\mathrm{TV}}(\hat{\mathbf{p}}, \mathbf{p}) \leq \varepsilon/2$ with probability at least $1 - \delta$. Assuming this is the case, then (i) if $\mathbf{p} \in \mathcal{P}_k$, then of course $\mathrm{d}_{\mathrm{TV}}(\hat{\mathbf{p}}, \mathcal{P}) \leq \varepsilon/2$; while (ii) if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}) > \varepsilon$, by the triangle inequality (since total variation distance is a metric) we must have $\mathrm{d}_{\mathrm{TV}}(\hat{\mathbf{p}}, \mathcal{P}) > \varepsilon/2$.

But we have an explicit description of $\hat{\mathbf{p}}$ in our hands, so we can check which of the two cases holds – this may not be computationally efficient, but does not require any additional sample from $\mathbf{p}$. Thus, we have a *bona fide* testing algorithm for $\mathcal{P}$ under $\mathcal{C}$. □

Importantly, this baseline is with respect to the sample complexity of learning distributions from $\mathcal{P}_k \cup \mathcal{C}_k$, *not* just $\mathcal{P}_k$: the latter is in general much larger! For instance, if $\mathcal{P}_k$ is a singleton but $\mathcal{C}_k = \Delta_k$ (*e.g.*, as in identity testing, which we shall see in Chapter 2) then learning $\mathcal{P}_k$ has sample complexity 0, but learning $\mathcal{P}_k \cup \mathcal{C}_k = \Delta_k$ has sample complexity $\Omega(k)$. This leads us to our baseline: since $\mathcal{P}_k \cup \mathcal{C}_k \subseteq \Delta_k$, the sample complexity of *any* distribution testing problem is at most the sample complexity of learning an arbitrary distribution over a known domain of the same size, which we record below:

Theorem 1.3 (Learning baseline)**.** The sample complexity of learning an arbitrary probability distribution from Δ_k to total variation ε with

error probability at most δ is

$$n_{\mathcal{L}}(k, \varepsilon, \delta) = \Theta\left(\frac{k + \log(1/\delta)}{\varepsilon^2}\right),$$

giving an upper bound on the sample complexity of any testing problem.

The proof can be found in various places; *e.g.*, Kamath *et al.* (2015) and Canonne (2020a). This testing-by-learning baseline, which is linear in the domain size k, motivates the name commonly given to testing algorithms which achieve significantly better sample complexity: *sublinear algorithms*.

Worst-case distance parameter ε. As defined, a testing algorithm must reject all distributions which are at distance greater than ε from the property, where ε is provided as an input parameter. In particular, the requirement is oblivious to the *true* distance $\varepsilon(\mathbf{p}) := \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}_k) > \varepsilon$ of the unknown distribution $\mathbf{p}$ to the property, and the sample complexity is just expressed as a function of the "worst-case" ε. Instead of this, one may want an *adaptive* algorithm which only takes the number of samples "needed" to reject, as a function of $\varepsilon(\mathbf{p})$: after all, in cases where $\varepsilon(\mathbf{p}) \gg \varepsilon$, one may reject after taking much fewer samples.

As it turns out, our focus on "worst-case ε" readily implies this adaptive setting, via the use of a *doubling search*. The idea is quite simple: given a testing algorithm $\mathcal{A}$, we create an adaptive testing algorithm $\mathcal{A}'$ by repeatedly trying to guess the true distance $\varepsilon(\mathbf{p})$, starting at $\varepsilon_0 = 1$ and halving our current guess ε_j at every stage until we reach $\varepsilon_L = \varepsilon$, and calling $\mathcal{A}$ for every guess, with parameters k, ε_j, and a suitable probability of failure δ_j at stage j. If any of these (at most $L := \lceil \log(1/\varepsilon) \rceil$) calls leads to a rejection, $\mathcal{A}'$ outputs 0; otherwise, it outputs 1. The key is to choose δ_j suitably so that (1) by a union bound all invocations of $\mathcal{A}$ are correct with probability at least $1 - \delta$, but (2) the union bound does not cost too much in terms of sample complexity. A standard way to do so is to set $\delta_j := \frac{\delta}{2(j+1)^2}$ (though many other choices of convergent series would do), so that $\sum_{j=0}^{\infty} \delta_j \leq \delta$.

The resulting sample complexity will then be, in the case $\varepsilon(\mathbf{p}) > \varepsilon$,

$$\sum_{j=0}^{\lceil \log(1/\varepsilon(\mathbf{p})) \rceil} n(k, \varepsilon_j, \delta_j) = \sum_{j=0}^{\lceil \log(1/\varepsilon(\mathbf{p})) \rceil} n\left(k, 2^{-j}, \frac{\delta}{2(j+1)^2}\right),$$

where $n(\cdot,\cdot,\cdot)$ denotes the sample complexity of $\mathcal{A}$. Under very mild conditions on n, this will be of the order $n\Big(k, \varepsilon(\mathbf{p}), \frac{\delta}{\log(1/\varepsilon(\mathbf{p}))}\Big)$, and recalling that the dependence on the error probability is at worst logarithmic, this means that adapting to the true value of $\varepsilon(\mathbf{p})$ incurs a cost at most doubly logarithmic in $\varepsilon(\mathbf{p})$. Of course, when $\mathbf{p} \in \mathcal{P}_k$, our adaptive algorithm $\mathcal{A}'$ should run for all of the $L := \lceil \log(1/\varepsilon) \rceil$ iterations (until ε_L) in order to output 1, in which case the sample complexity will be bounded as

$$\sum_{j=0}^{\lceil \log(1/\varepsilon) \rceil} n\left(k, 2^{-j}, \frac{\delta}{2(j+1)^2}\right).$$

We will see a concrete example of this technique in Exercise 2.11.

1.2 Why total variation distance?

The standard formulation of distribution testing, as stated in Definition 1.2, is tied to a specific metric between probability distribution: the total variation distance (Definition 1.1). It is natural to wonder of that choice is arbitrary, and, if not, what motivates it.

- Total variation distance provides a *very strong guarantee*, and for instance is the most stringent of all ℓ_p norms. This has practical consequences: if a source of data passes the test, then it will by nearly "as good as if it had the desired property" as far as *any* algorithm is concerned.

- It is *well-behaved*: total variation distance defines a proper metric, and thus satisfies for instance the triangle inequality (which cannot be said about, for instance, Kullback–Leibler divergence). It is also nicely bounded, and will not take infinite values due to pathological reasons.

- It satisfies the *data processing inequality* (Fact 1.1), which means it is robust to preprocessing. If data comes from two sources close in total variation distance, then post-processing this data cannot make their distribution statistically further apart. This is not the case for, among others, the ℓ_2 metric.

- Its relation to hypothesis testing: total variation distance has a natural and precise interpretation in terms of *distinguishability*. This is formalized in Lemma 1.4, and makes total variation distance the "right" notion of distance in applications such as cryptography, and when arguing about indistinguishability of data sources.

- Its connection to other distance measures. Total variation distance enjoys various inequalities relating it to other distance measures such as Kullback–Leibler divergence, ℓ_p metrics, Hellinger distance, Kolmogorov distance, and Wasserstein (Earthmover) metric. Some of those are elaborated upon in Appendix B.

Of course, total variation distance also has its drawbacks: it is sometimes too stringent, especially when considering distributions over continuous domains: in that case, absent further assumptions on the unknown continuous density, the testing problem becomes trivially impossible (Le Cam, 1973). It also does not "tensorize" well (as opposed to, say, Hellinger distance or Kullback–Leibler divergence), meaning that the total variation distance between product measures does not have a nice form with respect to the total variation distances between individual marginals. And indeed, there are pros and cons to each choice; although in this case the above should convince you that the pros vastly outnumber the cons.

Relation to hypothesis testing. As aforementioned, there is a natural connection between total variation distance and hypothesis testing, which we recall below.

Lemma 1.4 (Pearson–Neyman)**.** Any (possibly randomized) statistical test which distinguishes between $\mathbf{p}_0$ and $\mathbf{p}_1$ from a single sample must have Type I (false positive) and Type-II (false negative) errors satisfying

$$\text{Type I} + \text{Type II} \geq 1 - \mathrm{d}_{\mathrm{TV}}(\mathbf{p}_0, \mathbf{p}_1)$$

Moreover, this is achieved by the test which outputs 1 if, and only if, the sample belongs to the set $S^* := \{\, x : \mathbf{p}_1(x) > \mathbf{p}_0(x) \,\}$.

Proof. Fix any test $\mathcal{A}$ distinguishing between two distributions $\mathbf{p}_0$ and $\mathbf{p}_1$, given a single observation. Letting α and β denote the Type I and Type-II errors of $\mathcal{A}$, we have

$$\begin{aligned}\alpha + \beta &= \Pr_{\mathbf{p}_0,R}[\,\mathcal{A}(X,R) = 1\,] + \Pr_{\mathbf{p}_1,R}[\,\mathcal{A}(X,R) = 0\,] \\ &= \mathbb{E}_R[\Pr_{\mathbf{p}_0}[\,\mathcal{A}(X,R) = 1\,]] + \mathbb{E}_{\mathcal{A}}[\Pr_{\mathbf{p}_1}[\,\mathcal{A}(X,R) = 0\,]] \\ &= \mathbb{E}_R[\Pr_{\mathbf{p}_0}[\,\mathcal{A}(X,R) = 1\,] + \Pr_{\mathbf{p}_1}[\,\mathcal{A}(X,R) = 0\,]]\end{aligned}$$

where we denote by R the internal randomness of $\mathcal{A}$. Since, for any fixed realization r of this randomness R, the resulting test $\mathcal{A}(\cdot, r)$ is deterministic, we can define for any r the *acceptance region* $S_{\mathcal{A},r} := \{\, x : \mathcal{A}(x,r) = 1 \,\}$, and write

$$\begin{aligned}\alpha + \beta &= \mathbb{E}_R[\Pr_{\mathbf{p}_0}[\,X \in S_{\mathcal{A},R}\,] + \Pr_{\mathbf{p}_1}[\,X \notin S_{\mathcal{A},R}\,]] \\ &= 1 + \mathbb{E}_R[\mathbf{p}_0(S_{\mathcal{A},R}) - \mathbf{p}_1(S_{\mathcal{A},R})] \\ &\geq 1 + \inf_S(\mathbf{p}_0(S) - \mathbf{p}_1(S)) \\ &= 1 - \sup_S(\mathbf{p}_1(S) - \mathbf{p}_0(S)) \\ &= 1 - \mathrm{d}_{\mathrm{TV}}(\mathbf{p}_0, \mathbf{p}_1)\,,\end{aligned}$$

as claimed. Finally, it is immediate from the definition of total variation distance that the proposed test satisfies Type I+Type II $= 1 + \mathbf{p}_0(S^*) - \mathbf{p}_1(S^*) = 1 - \mathrm{d}_{\mathrm{TV}}(\mathbf{p}_0, \mathbf{p}_1)$. □

1.3 The road not taken: tolerant testing

In Definition 1.2 and throughout this survey, we focus on the standard formulation version of testing, where the unknown distribution $\mathbf{p}$ either *belongs* to the property $\mathcal{P}_k$ or is far from it. A natural generalization of this, allowing for some "tolerance" to noise or misspecification in the former case, would be to ask to distinguish $\mathbf{p}$ *close* to $\mathcal{P}_k$ from $\mathbf{p}$ far from it. This is called *tolerant testing* Parnas *et al.*, 2006, and is formalized by introducing two parameters $0 \leq \varepsilon' < \varepsilon \leq 1$, and relaxing the first item of Definition 1.2 to

> If S is i.i.d. from some $\mathbf{p} \in \Delta_k$ such that $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}_k) \leq \varepsilon'$, then $\Pr_{S,\mathcal{A}}[\,\mathbf{b} = 1\,] \geq 1 - \delta$;

(Note then that our regular, "non-tolerant" testing corresponds to taking $\varepsilon' = 0$.) The tolerant testing task, sometimes called in Statistics testing with an *imprecise null*, typically requires a much higher sample complexity than the non-tolerant one (Valiant and Valiant, 2011), and both algorithms and lower bounds are obtained via significantly different techniques. For this reason, we will not here discuss tolerant testing in much, or indeed any detail: the interested reader is referred to, *e.g.*, Wu and Yang (2020) for a primer on some of those techniques, and to Canonne *et al.* (2021) and references within for an overview of results on tolerant goodness-of-fit testing.

1.4 Historical notes

Hypothesis testing has a long and rich history in Statistics, starting with the work of Pearson (1900) introducing the χ^2 test, and leading to substantial advances over the next century. While it is difficult and slightly dangerous to reduce twelve decades of intense research and study in a few sentences,[2] standard approaches in Statistics share a few common features. First, they are *asymptotic* in nature (as opposite to focusing on finite-sample guarantees), establishing and studying the limiting distribution of a given test as the sample size goes to infinity. This enables one to compute confidence intervals, and obtain a swath of information from the limiting distribution; but by itself provides little insight regarding the intermediate, finite-sample regime. Second, they tend to focus on the so-called Type I error (significance of the test), *i.e.*, the probability to mistakenly reject the null hypothesis $\mathcal{H}_0$, and only after fixing this significance level set out to minimize the Type II error (that is, maximize the *power* of the test), which is the probability to mistakenly accept the alternative hypothesis $\mathcal{H}_1$. This is, again, an oversimplification; the reader should refer to, *e.g.*, Balakrishnan and Wasserman (2018) for a complementary and more detailed view. Nonetheless, these features are two of the most salient points of contrast with the very recent and related take on hypothesis testing from the theoretical computer science community, *distribution testing*, which

[2] Which is exactly what the following paragraph will set out to do regardless.

perhaps shares most similarity with the work of Ingster (Ingster, 1986; Ingster, 1997).

Distribution testing was first introduced in an influential paper by Goldreich *et al.* (1998), which formally defined the broader field of property testing; Goldreich and Ron (2000) then specifically considered the question of testing uniformity of an unknown probability distribution (in an ℓ_2 sense), using the collision-based tester to test whether a random walk had (approximately) reached its stationary distribution.

This was, however, only implicitly using uniformity testing as a subroutine in the context of testing some property (expansion) of bounded-degree graphs. The work of Batu *et al.* (2000) first considers distribution testing for its own sake, studying the question of *closeness testing* (*i.e.*, two-sample testing), where one seeks to decide from samples if two unknown distributions are equal. This initiated a long line of work on testing many properties – including uniformity, identity, closeness, monotonicity, independence (being a product distribution), to name only a few.

While the early papers focused on the dependence on the domain size k, treating the distance parameter ε as a small constant or a second-order concern, later works, beginning with Chan *et al.* (2014), started looking for the tight dependence on ε as well. Even more recently, the "right" dependence on δ, the error probability, has come into focus as well (Diakonikolas *et al.*, 2018; Diakonikolas *et al.*, 2021). This, in some sense, brings the theoretical computer science closer to the information theory literature, where the focus on the *error exponent* (that is, the rate at which the error probability decays exponentially, as a function of the other parameters) is the standard view.

Another recent trend in distribution testing has been to consider different "accesses" to the data, rather than i.i.d. samples: for instance, access to so-called conditional samples (Canonne *et al.*, 2015; Chakraborty *et al.*, 2013), or the ability to ask for, or observe, the probability of individual elements of the domain (Rubinfeld and Servedio, 2009; Canonne and Rubinfeld, 2014; Onak and Sun, 2018). These types of access allow for much more efficient testing algorithms, but require significantly different algorithmic tools and proof techniques, and we will not discuss them here. For more on this, we defer the interested reader to another

survey, Canonne (2020b).

Finally, over the past few years distribution testing has ventured into adjacent areas of computer science and information theory, by incorporating various constraints and resources into its formulation. Examples include data privacy (and, more specifically, *differential privacy* (Dwork *et al.*, 2006) and its variants) – see, *e.g.*, (Kamath and Ullman, 2020), memory constraints, and bandwidth constraints (Tsitsiklis, 1993); of which we will cover a fraction in Chapter 4. This has been done by borrowing, extending, and (re)discovering ideas and techniques from these areas and Statistics; somewhat satisfyingly, leading distribution testing back to some of its roots.

This survey aims to describe some of these connections, and provide an overview of these ideas and techniques which took years for the author to learn about.

2

Testing goodness-of-fit of univariate distributions

In this chapter, we take an in-depth look at one of the most natural and "basic" properties to test, *uniformity*:

> Is the unknown distribution *the* uniform distribution $\mathbf{u}_k$ over the domain $[k]$, or does it significantly deviate from being uniform?

Formally, recalling Definition 1.2, we define uniformity testing by letting $\mathcal{P}_k := \{\mathbf{u}_k\}$ (a singleton), and $\mathcal{C}_k = \Delta_k$. Of course, while the uniform distribution is a rather fundamental one, one might instead have a different reference distribution in mind, *e.g.*, a particular Binomial distribution, or a Zipf one, or an arbitrary probability distribution provided as a vector of k probabilities. This generalization is known (in the computer science community) as *identity testing*, and also goes by the name of *one-sample goodness-of-fit*. While identity testing might seem to be a strictly more general problem than uniformity testing, the two are closely related. Clearly, identity testing is at least as hard as uniformity testing; but it turns out that a converse holds, and that uniformity is also the "hardest" case of identity testing, in a very strong sense. Namely, as we will see in Section 2.2, any testing algorithm for

uniformity implies one for identity testing, with essentially the same sample complexity![1]

Before delving into the details of this chapter, two more comments. First, "uniformity testing" refers to testing whether the unknown distribution $\mathbf{p}$ is uniform *on the whole, known domain*: the different question of testing whether $\mathbf{p}$ is uniform on its (unknown) *support*, *i.e.*,, testing $\mathcal{P}_k := \{ \mathbf{u}_S : S \subseteq [k] \}$ is known as *generalized* uniformity testing (Batu and Canonne, 2017), and has a strictly higher sample complexity.

Second, the formulation of identity testing crucially assumes that the reference distribution is fully known to the algorithm. One can also consider the task of testing, given samples from two distributions $\mathbf{p}, \mathbf{q}$ (both unknown), whether $\mathbf{p} = \mathbf{q}$ or $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$. This question, *closeness testing* (or *two-sample goodness-of-fit*), which we briefly alluded to in Chapter 1), can be seen to be at least as hard as identity testing: any algorithm for closeness testing can be used to perform identity testing by generating i.i.d. samples from the known reference distribution $\mathbf{q}$.[2] But closeness testing, again, is *strictly* harder than identity testing: while, as we will see, the latter has sample complexity $\Theta(\sqrt{k}/\varepsilon^2)$, the former requires $\Theta(k^{2/3}/\varepsilon^{4/3} + \sqrt{k}/\varepsilon^2)$ samples (Chan *et al.*, 2014), which is much larger as long as $\varepsilon \gg 1/k^{1/4}$.

In the rest of the chapter, we will go through an in-depth overview of uniformity and identity testing, with a full derivation of many results. This is the most detailed section of this monograph. We will start in Section 2.1 by describing and analyzing seven different uniformity testing algorithms, before turning in Section 2.2 to identity testing. There, we will describe the aforementioned reduction between uniformity and identity testing, which lets us use any of the algorithms from Section 2.1 to solve the more general problem; as well as a testing algorithm to perform identity testing directly, without reducing it to uniformity. The goal of this chapter is not to provide a list of various testing algorithms to the reader, then left to ponder what to do with it; but rather to highlight some general techniques and ideas in the course

[1] There is a small print, of course, and some constant factors are lost.

[2] This adds a computational overhead, but as far as sample complexity is concerned this is not a problem.

of their analysis, applicable beyond these specific examples.

2.1 Uniformity testing

The sample complexity of uniformity testing with distance parameter $\varepsilon \in (0, 1]$ is known to be $\Theta(\sqrt{k}/\varepsilon^2)$ (Paninski, 2008); this is, for a large range of parameters, significantly smaller than the domain size k, and in particular *much* better than the learning baseline of Theorem 1.3. This also means that we can reliably infer interesting properties of the distribution after observing only a negligible fraction of the domain elements! That is quite surprising, and nice. Now, *how do we perform uniformity testing?* And what should we keep in mind while doing so?

Although the sample complexity is of course a key aspect, there are several things to consider in a testing algorithm. To name a few:

Data efficiency: does the algorithm achieve the optimal sample complexity $\Theta(\sqrt{k}/\varepsilon^2)$?

Time efficiency: how fast is the algorithm to run (as a function of k, ε, and the number of samples n)?

Memory efficiency: how much memory does the algorithm require (as a function of k, ε, and n)?

Simplicity: is the algorithm simple to describe and implement?

"Simplicity": is the algorithm simple to *analyze*? This is not just a one-time thing: adapting and building upon a given algorithm will be much easier if the analysis does not require a long succession of technical, very specific, and complex steps.

Robustness: how *tolerant* is the algorithm to breaches of the promise? That is, does it accept distributions which are not *exactly* uniform as well, or is it very brittle?

Elegance: This is somewhat subjective, and each of us might have a different view of what it means. If you have strong feelings about this, however, know that you are not alone.

Generalizable: Does the algorithm have other features that might be desirable in other settings? For instance, if the algorithm has low sensitivity (in the Lipschitz sense), then it will be more resilient to adversarial perturbations of some of the samples – this is great for, *e.g.*, differential privacy applications.

In this chapter, we will have a detailed look at 7 different uniformity testing algorithms, each with its pros and cons. For each of them, we will provide a full proof of their performance, which will help us illustrate several techniques and ideas underlying the analysis of distribution testing algorithms. Table 2.1 summarizes the names, and some of the specificities, of those seven algorithms: we will discuss them in more detail in Section 2.1.9.

2.1.1 The ℓ_2 distance, and why

A key insight, which underlies a lot of the algorithms we will cover, is that here ℓ_2 *distance is a good proxy for total variation distance*:

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) = \frac{1}{2}\|\mathbf{p} - \mathbf{u}_k\|_1 \leq \frac{\sqrt{k}}{2}\|\mathbf{p} - \mathbf{u}_k\|_2 \tag{2.1}$$

the inequality being Cauchy–Schwarz. So if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) > \varepsilon$, then $\|\mathbf{p} - \mathbf{u}_k\|_2^2 > 4\varepsilon^2/k$ (and, well, if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) = 0$ then $\|\mathbf{p} - \mathbf{u}_k\|_2^2 = 0$ too, of course). Moreover, we have the very convenient fact, specific to the distance to uniform: for any distribution $\mathbf{p}$ over $[k]$,

$$\|\mathbf{p} - \mathbf{u}_k\|_2^2 = \sum_{i=1}^{k} (\mathbf{p}(i) - 1/k)^2 = \sum_{i=1}^{k} \mathbf{p}(i)^2 - 1/k = \|\mathbf{p}\|_2^2 - 1/k\,, \tag{2.2}$$

so combining the two we get that $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) > \varepsilon$ implies $\|\mathbf{p}\|_2^2 > (1 + 4\varepsilon^2)/k$.

Remark 2.1. The quantity $\|\mathbf{p}\|_2^2$ is commonly known as the *collision probability* of $\mathbf{p}$, due to the following easy fact: if X, Y are i.i.d. random variables distributed according to $\mathbf{p}$, then

$$\Pr[\,X = Y\,] = \sum_{i \in \mathcal{X}} \Pr[\,X = i, Y = i\,] = \sum_{i \in \mathcal{X}} \mathbf{p}(i)^2 = \|\mathbf{p}\|_2^2 \tag{2.3}$$

	Sample complexity	Notes	References
Collision-based	$\frac{k^{1/2}}{\varepsilon^2}$	"Natural"	Goldreich and Ron, 2000; Diakonikolas *et al.*, 2019b
Unique elements	$\frac{k^{1/2}}{\varepsilon^2}$	Low sensitivity $\varepsilon \gg 1/k^{1/4}$	Paninski, 2008
Modified χ^2	$\frac{k^{1/2}}{\varepsilon^2}$	(None)	Valiant and Valiant, 2017; Acharya *et al.*, 2015; Diakonikolas *et al.*, 2015
Empirical distance to uniform	$\frac{k^{1/2}}{\varepsilon^2}$	Low sensitivity	Diakonikolas *et al.*, 2018
Random binary hashing	$\frac{k}{\varepsilon^2}$	Suboptimal, but fast	Acharya *et al.*, 2020d
Bipartite collisions	$\frac{k^{1/2}}{\varepsilon^2}$	Tradeoff possible	Diakonikolas *et al.*, 2019a
Empirical subset weighting	$\frac{k^{1/2}}{\varepsilon^2}$	Tradeoff possible $\varepsilon \gg 1/k^{1/4}$	Acharya *et al.*, 2022

Table 2.1: The current landscape of uniformity testing, based on the algorithms covered in this survey. For ease of reading, we omit the $O(\cdot)$, $\Theta(\cdot)$, and $\Omega(\cdot)$'s from the table: all results should be read as asymptotic with regard to the parameters, up to absolute constants.

(this generalizes to higher-order collisions, with $\|\mathbf{p}\|_j^j$). It is easy to see, from Eq. (2.2), that among all probability distributions over a given support size k the collision probability is minimized for the uniform distribution: indeed, $\|\mathbf{p}\|_2^2 = \frac{1}{k} + \|\mathbf{p} - \mathbf{u}_k\|_2^2 \geq \frac{1}{k}$.

Remark 2.2. Eq. (2.1) provides an upper bound on the total variation distance in terms of the ℓ_2 distance. Recalling further that ℓ_p norms are monotone (non-increasing) in p, we further get that $\|x\|_1 \geq \|x\|_2$ for any $x \in \mathbb{R}^d$, and therefore

E: Prove it: Exercise 2.1.

$$\frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_2 \leq \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \leq \frac{\sqrt{k}}{2}\|\mathbf{p} - \mathbf{q}\|_2 \tag{2.4}$$

for any two $\mathbf{p}, \mathbf{q} \in \Delta_k$. Although loose by a factor $\sqrt{k}$ (where k is the domain size), this relation turns out to be surprisingly useful in many occasions.

2.1.2 Collision-based

In view of Eq. (2.2), a very natural idea is to estimate $\|\mathbf{p}\|_2^2$, in order to distinguish between (i) $\|\mathbf{p}\|_2^2 = 1/k$ (uniform) and (ii) $\|\mathbf{p}\|_2^2 > (1+4\varepsilon^2)/k$ (ε-far from uniform). How to do that? Upon recalling Remark 2.1, the probability that two independent samples from $\mathbf{p}$ take the same value (a "collision") is exactly $\|\mathbf{p}\|_2^2$. Thus, an obvious approach is to take n samples $X_1, \ldots, X_n$, count the number of pairs that show a collision, and use that as an unbiased estimator Z_1 for $\|\mathbf{p}\|_2^2$:

$$Z_1 = \frac{1}{\binom{n}{2}} \sum_{1 \le s < t \le n} \mathbb{1}\{X_s = X_t\}\,. \tag{2.5}$$

By the above, $\mathbb{E}[Z_1] = \|\mathbf{p}\|_2^2$. If we threshold Z_1 somewhere between (i) and (ii), at say $\tau := (1 + 2\varepsilon^2)/k$, we should be able to distinguish between our two cases and get a valid tester. But how large must n be for this to work?

Algorithm 1 Collision-Based Tester

Require: Multiset of n samples $x_1, \ldots, x_n \in \mathcal{X}$, parameters $\varepsilon \in (0, 1]$ and $k = |\mathcal{X}|$

1: Set $\tau \leftarrow \frac{1+2\varepsilon^2}{k}$

2: Compute ▷ Can be done in $O(n)$ time if $\mathcal{X}$ is known, $O(n \log n)$ if only k is.

$$Z_1 = \frac{1}{\binom{n}{2}} \sum_{1 \le s < t \le n} \mathbb{1}\{x_s = x_t\} = \frac{1}{\binom{n}{2}} \sum_{j \in \mathcal{X}} \binom{N_j}{2}$$

where $N_j \leftarrow \sum_{t=1}^{n} \mathbb{1}\{x_t = j\}$.

3: **if** $Z_1 \ge \tau$ **then return** 0 ▷ Not uniform

4: **else return** 1 ▷ Uniform

Intuitively, we expect the test to work as long as the standard deviation of Z_1 (the "noise") is smaller than the gap between the

expectations in our two cases (the "signal"); that is,

$$\sqrt{\mathrm{Var}[Z_1]} \ll \Delta\mathbb{E}[Z_1] = \frac{4\varepsilon^2}{k} \tag{2.6}$$

as this is the condition for the random fluctuations of our statistic Z_1 not to "cross" our threshold too often and lead to a wrong answer.

To make this quantitative, we can use Chebyshev's inequality, which requires us to bound $\mathrm{Var}[Z_1]$. This is where things get tricky, since Z_1 is the sum of $\binom{n}{2}$ random variables which are *not* pairwise independent.[3]

We first show how to derive a (suboptimal) bound $n = O\left(\sqrt{k}/\varepsilon^4\right)$:

$$\begin{aligned}\mathrm{Var}[Z_1] &= \mathbb{E}\left[Z_1^2\right] - \mathbb{E}[Z_1]^2 \\ &= \frac{1}{\binom{n}{2}^2} \sum_{1\le s<t\le n} \sum_{1\le s'<t'\le n} \mathbb{E}[\mathbb{1}\{X_s = X_t\}\mathbb{1}\{X_{s'} = X_{t'}\}] - \|\mathbf{p}\|_2^4\end{aligned}$$

To handle this last sum despite the lack of independence of the summands, we will break it in 3 groups depending on the cardinality of $\{s, t, s', t'\}$, which can be either 4 (all indices are distinct), 3 (one index is common to the two pairs), or 2 (both pairs of indices are the same).

- In the first case, we have independence of the two indicator random variables, and

  $$\mathbb{E}[\mathbb{1}\{X_s = X_t\}\mathbb{1}\{X_{s'} = X_{t'}\}] = \mathbb{E}[\mathbb{1}\{X_s = X_t\}]\mathbb{E}[\mathbb{1}\{X_{s'} = X_{t'}\}] = \|\mathbf{p}\|_2^4.$$

- In the third case, the two indicator random variables are the same, and since $\mathbb{1}\{\}^2 = \mathbb{1}\{\}$ we get

  $$\mathbb{E}[\mathbb{1}\{X_s = X_t\}\mathbb{1}\{X_{s'} = X_{t'}\}] = \mathbb{E}[\mathbb{1}\{X_s = X_t\}] = \|\mathbf{p}\|_2^2.$$

- The second case is the messiest one; still, one can verify that in this case $\mathbb{1}\{X_s = X_t\}\mathbb{1}\{X_{s'} = X_{t'}\}$ is 1 if, and only if, the three distinct samples corresponding to the 3 distinct indices among s, t, s', t' take the same value, from which

  $$\mathbb{E}[\mathbb{1}\{X_s = X_t\}\mathbb{1}\{X_{s'} = X_{t'}\}] = \|\mathbf{p}\|_3^3.$$

[3] Namely, the summands $\mathbb{1}\{X_s = X_t\}$ in the definition of Z_1 are *positively correlated*: $\mathrm{Cov}(\mathbb{1}\{X_s = X_t\}, \mathbb{1}\{X_{s'} = X_{t'}\}) \ge 0$, and are only independent if s, s', t, t' are all distinct.

It remains to count how many summands of each type we have. Clearly, we have exactly $\binom{n}{2}$ summands of the third type; it is also not too hard to see that we have $\binom{n}{2}\binom{n-2}{2} = 6\binom{n}{4}$ summands of the first, and $6\binom{n}{3}$ of the second. (As a sanity check, $6\binom{n}{4} + 6\binom{n}{3} + \binom{n}{2} = \binom{n}{2}^2$, so all our summands are accounted for.)

Getting back to our variance computation, this yields

$$\begin{aligned}
\mathrm{Var}[Z_1] &= \frac{1}{\binom{n}{2}^2}\left(6\binom{n}{4}\|\mathbf{p}\|_2^4 + 6\binom{n}{3}\|\mathbf{p}\|_3^3 + \binom{n}{2}\|\mathbf{p}\|_2^2\right) - \|\mathbf{p}\|_2^4 \\
&= \frac{1}{\binom{n}{2}^2}\left(\left(6\binom{n}{4} - \binom{n}{2}^2\right)\|\mathbf{p}\|_2^4 + 6\binom{n}{3}\|\mathbf{p}\|_3^3 + \binom{n}{2}\|\mathbf{p}\|_2^2\right) \qquad (2.7) \\
&\le \frac{4}{n}\|\mathbf{p}\|_3^3 + \frac{4}{n^2}\|\mathbf{p}\|_2^2 \\
&\le \frac{4}{n}\mathbb{E}[Z_1]^{3/2} + \frac{4}{n^2}\mathbb{E}[Z_1]
\end{aligned}$$

first using that $6\binom{n}{4} < \binom{n}{2}^2$ to discard a negative term, then that $n \ge 2$ to get a simpler-looking upper bound on binomial coefficients, and finally writing $\|\mathbf{p}\|_3 \le \|\mathbf{p}\|_2$ by monotonicity of ℓ_p norms.

In the uniform case (often called the *completeness* case for historical reasons), we seek to control the probability that Z_1 crosses our threshold $\tau := \frac{1+2\varepsilon^2}{k}$, that is

$$\Pr[Z_1 \ge \tau] = \Pr\Big[Z_1 \ge (1+2\varepsilon^2)\mathbb{E}[Z_1]\Big] \le \Pr\Big[Z_1 \ge (1+\varepsilon^2)\mathbb{E}[Z_1]\Big]$$

while in the "far" case (often called the *soundness* case), we want to control

$$\Pr[Z_1 < \tau] \le \Pr\left[Z_1 < \frac{(1-\varepsilon^2)(1+4\varepsilon^2)}{k}\right] \le \Pr\Big[Z_1 < (1-\varepsilon^2)\mathbb{E}[Z_1]\Big]$$

using first that $(1-\varepsilon^2)(1+4\varepsilon^2) \ge 1+2\varepsilon^2$ (for $\varepsilon \le 1/2$), and then the fact that in the "far" case $\mathbb{E}[Z_1] > \frac{1+4\varepsilon^2}{k}$.

To control our probability of error in both cases, it is thus sufficient to upper bound $\Pr[|Z_1 - \mathbb{E}[Z_1]| \ge \varepsilon^2\mathbb{E}[Z_1]]$; by Chebyshev's inequality

(Theorem A.2), this is at most

$$\begin{aligned}\Pr\Big[|Z_1 - \mathbb{E}[Z_1]| \geq \varepsilon^2 \mathbb{E}[Z_1]\Big] &\leq \frac{\mathrm{Var}[Z_1]}{\varepsilon^4 \mathbb{E}[Z_1]^2} \\ &\leq \frac{4}{\varepsilon^4 n \mathbb{E}[Z_1]^{1/2}} + \frac{4}{\varepsilon^4 n^2 \mathbb{E}[Z_1]} \\ &\leq \frac{4\sqrt{k}}{\varepsilon^4 n} + \frac{4k}{\varepsilon^4 n^2}\end{aligned}$$

which is at most $1/3$, as desired, for $n \geq \frac{13\sqrt{k}}{\varepsilon^4}$. (For the third inequality, we relied on the fact that $\mathbb{E}[Z_1] = \|\mathbf{p}\|_2^2 \geq 1/k$ (*cf.* Remark 2.1).) □

The above argument establishes that the collision-based tester has sample complexity $O(\sqrt{k}/\varepsilon^4)$. This is not the best one can get; in fact, by being (much) more careful one can show that this tester achieves the optimal sample complexity (as a function of k and ε)!

Theorem 2.1. The collision-based tester (Algorithm 1) is a testing algorithm for uniformity with sample complexity $n(k, \varepsilon, 1/3) = O(\sqrt{k}/\varepsilon^2)$ and time complexity $O(n)$.

Proof. Clearly, if we are to prove this tighter bound, we have to be less heavy-handed in one of the steps of the previous analysis, specifically in bounding the variance. An obvious candidate is the first step featuring an inequality instead of an equality, just after Eq. (2.7): there, we discarded a term since its coefficient $6\binom{n}{4} - \binom{n}{2}^2$ was negative.

Maybe we should not have. Starting again at Eq. (2.7) and recalling the relation $\binom{n}{2}^2 = 6\binom{n}{4} + 6\binom{n}{3} + \binom{n}{2}$ we saw earlier, we have

$$\begin{aligned}\binom{n}{2}^2 \mathrm{Var}[Z_1] &= \left(6\binom{n}{4} - \binom{n}{2}^2\right)\|\mathbf{p}\|_2^4 + 6\binom{n}{3}\|\mathbf{p}\|_3^3 + \binom{n}{2}\|\mathbf{p}\|_2^2 \\ &= \binom{n}{2}\|\mathbf{p}\|_2^2(1 - \|\mathbf{p}\|_2^2) + 6\binom{n}{3}(\|\mathbf{p}\|_3^3 - \|\mathbf{p}\|_2^4) \qquad (2.8)\end{aligned}$$

A detour: this quite interesting! The first term is exactly the variance we would get had we had independent summands (this is exactly the variance of a Binomial with parameters $\binom{n}{2}$ and $\|\mathbf{p}\|_2^2$), while the second

is the "positive correlation" term. Indeed, one can see that the second term is always nonegative, as

$$\|\mathbf{p}\|_2^4 = \left(\sum_i \mathbf{p}(i)^{3/2}\mathbf{p}(i)^{1/2}\right)^2 \le \sum_i \mathbf{p}(i)^3 \sum_i \mathbf{p}(i) = \|\mathbf{p}\|_3^3$$

by Cauchy–Schwarz (with equality if, and only if, $\mathbf{p}$ is uniform on its support). Going back to our variance analysis, we will simplify a little the RHS above for the sake of conciseness, leading to

$$\mathrm{Var}[Z_1] \le \frac{4}{n^2}\|\mathbf{p}\|_2^2 + \frac{4}{n}(\|\mathbf{p}\|_3^3 - \|\mathbf{p}\|_2^4) \tag{2.9}$$

at the cost of some small constant factors (and assuming without loss of generality that $n \ge 2$).

- In the uniform case, we have $\|\mathbf{u}_k\|_3^3 = \|\mathbf{u}_k\|_2^4$ and Eq. (2.9) gives $\mathrm{Var}[Z_1] \le \frac{4}{n^2 k}$, and similarly as before

  $$\Pr[Z_1 \ge \tau] = \Pr\Big[Z_1 \ge (1+2\varepsilon^2)\mathbb{E}[Z_1]\Big] \le \frac{\mathrm{Var}[Z_1]}{4\varepsilon^4 \mathbb{E}[Z_1]^2} \le \frac{k}{\varepsilon^4 n^2}$$

  by Chebyshev's inequality, and using $\mathbb{E}[Z_1] = 1/k$. This in turn is less than $1/3$ as long as $n \ge \sqrt{3k}/\varepsilon^2$.

- In the "far" case, we will be a little more careful. Let $\alpha^2 := k\|\mathbf{p} - \mathbf{u}_k\|_2^2 \ge 4\varepsilon^2$, so that $\mathbb{E}[Z_1] = \frac{1+\alpha^2}{k}$. The probability that our tester errs in this case is

  $$\begin{aligned}
  \Pr[Z_1 < \tau] &= \Pr\left[Z_1 < \frac{1+2\varepsilon^2}{1+\alpha^2}\mathbb{E}[Z_1]\right] \\
  &= \Pr\left[Z_1 < \left(1 - \frac{\alpha^2 - 2\varepsilon^2}{1+\alpha^2}\right)\mathbb{E}[Z_1]\right] \\
  &\le \Pr\left[Z_1 < \left(1 - \frac{\alpha^2}{2(1+\alpha^2)}\right)\mathbb{E}[Z_1]\right] \\
  &\le \frac{4(1+\alpha^2)^2}{\alpha^4} \cdot \frac{\mathrm{Var}[Z_1]}{\mathbb{E}[Z_1]^2} \\
  &\le \frac{16(1+\alpha^2)^2}{\alpha^4 n^2 \|\mathbf{p}\|_2^2} + \frac{16(1+\alpha^2)^2}{\alpha^4 n} \cdot \frac{\|\mathbf{p}\|_3^3 - \|\mathbf{p}\|_2^4}{\|\mathbf{p}\|_2^4}
  \end{aligned}$$

the first inequality using $\alpha^2 \geq 4\varepsilon^2$, the second being Chebyshev's, and the third being Eq. (2.9). The first term is relatively familiar: since $\|\mathbf{p}\|_2^2 = (1+\alpha^2)/k$ by definition of α, we have

$$\frac{16(1+\alpha^2)^2}{\alpha^4 n^2 \|\mathbf{p}\|_2^2} = \frac{16(1+\alpha^2)k}{\alpha^4 n^2} \leq \frac{5k}{\varepsilon^4 n^2} \tag{2.10}$$

the inequality since $x > 0 \mapsto \frac{1+x}{x^2}$ is decreasing and $\alpha^2 \geq 4\varepsilon^2$ (and $\varepsilon \leq 1$ to write $1 + 4\varepsilon^2 \leq 5$).

To handle the second, we write $\mathbf{p} = (\mathbf{p} - \mathbf{u}_k) + \mathbf{u}_k$ and expand:

$$\begin{aligned}
\|\mathbf{p}\|_3^3 - \|\mathbf{p}\|_2^4 &\leq \|\mathbf{p} - \mathbf{u}_k + \mathbf{u}_k\|_3^3 - \frac{1}{k^2} \\
&= \|\mathbf{p} - \mathbf{u}\|_3^3 + \frac{3}{k}\|\mathbf{p} - \mathbf{u}\|_2^2 \qquad (2.11) \\
&\leq \|\mathbf{p} - \mathbf{u}\|_2^3 + \frac{3}{k}\|\mathbf{p} - \mathbf{u}\|_2^2 \\
&= \frac{\alpha^3}{k^{3/2}} + \frac{3\alpha^2}{k^2} \qquad (2.12)
\end{aligned}$$

where the first inequality uses $\|\mathbf{p}\|_2^2 \geq 1/k$, the equality follows from expanding the cubes in $\sum_i ((\mathbf{p}(i) - \mathbf{u}_k(i)) + \mathbf{u}_k(i))^3$ and observing the cancellations, the second inequality is monotonicity of ℓ_p norms, and the last equality stems from the definition of α. This gives

$$\begin{aligned}
\frac{16(1+\alpha^2)^2}{\alpha^4 n} \cdot \frac{\|\mathbf{p}\|_3^3 - \|\mathbf{p}\|_2^4}{\|\mathbf{p}\|_2^4} &= \frac{16k^2}{\alpha^4 n} \cdot (\|\mathbf{p}\|_3^3 - \|\mathbf{p}\|_2^4) \\
&\leq \frac{16\sqrt{k}}{\alpha n} + \frac{48}{\alpha^2 n} \\
&\leq \frac{8\sqrt{k}}{\varepsilon n} + \frac{12}{\varepsilon^2 n}. \qquad (2.13)
\end{aligned}$$

Combining Eqs. (2.10) and (2.13), we get

$$\Pr[Z_1 < \tau] \leq \frac{5k}{\varepsilon^4 n^2} + \frac{8\sqrt{k}}{\varepsilon n} + \frac{12}{\varepsilon^2 n} \leq \frac{5k}{\varepsilon^4 n^2} + \frac{20\sqrt{k}}{\varepsilon^2 n}$$

which is less than $1/3$ for $n \geq 64\sqrt{k}/\varepsilon^2$, as can be seen by solving the inequality $5x^2 + 20x \leq 1/3$ (where $x = \frac{\sqrt{k}}{\varepsilon^2 n}$).

Combining the uniform and far cases proves the theorem, showing that $n = O(\sqrt{k}/\varepsilon^2)$ samples suffice for the collision-based tester to be correct with probability at least $2/3$ in both cases. □

2.1.3 Unique elements

Another idea: count the number of elements that appear exactly *once* among the n samples taken. Why is that a sensible thing to do? We have seen that the uniform distribution will have the fewer collisions, so, equivalently, will have the maximum number of unique elements. In this case, the estimator Z_2 (the number of unique elements) is defined as

$$Z_2 = \frac{1}{n} \sum_{j \in \mathcal{X}} \mathbb{1}\{N_j = 1\}\,, \tag{2.14}$$

again with $N_j := \sum_{t=1}^{n} \mathbb{1}\{X_t = j\}$. It is a simple matter to verify that this statistic has expectation

E: Check it: Exercise 2.2.

$$\mathbb{E}[Z_2] = \sum_{i \in \mathcal{X}} \mathbf{p}(i)(1 - \mathbf{p}(i))^{n-1} \tag{2.15}$$

which is… a thing? Note that under the uniform distribution $\mathbf{u}_k$, this is exactly $(1-1/k)^{n-1} \approx 1 - \frac{n}{k}$, and under arbitrary $\mathbf{p}$ this is (making a few approximations not always valid) $\approx \sum_{i=1}^{k} \mathbf{p}(i)(1 - n\mathbf{p}(i)) = 1 - n\|\mathbf{p}\|_2^2$. So the gap in expectation between the two cases "should" be around $4\varepsilon^2 n/k$, and, if the variance analysis goes well and the stars align, we will be able to use Chebyshev's inequality and argue that we can distinguish the two for n large enough.

Now, before we actually delve into this analysis, it is worth mentioning a limitation of this tester, which is that we only expect it to work for $n \ll k$. Indeed, we count the number of *distinct* elements, and there will never ever be more than k of them if the domain size is k.[4] That explains, intuitively, the condition for the test to work: we need n (the number of samples taken) to be smaller than k (the maximum number of distinct elements one can ever hope to see), which gives, since we will eventually get $n = \Theta(\sqrt{k}/\varepsilon^2)$, the condition $\varepsilon \gg 1/k^{1/4}$.

[4] More quantitatively, for $n \to \infty$ the approximations made in the above discussion completely break down; and we will instead get $\mathbb{E}[Z_2] \to 0$ in both the uniform and the far cases.

Algorithm 2 UNIQUE-ELEMENTS TESTER

Require: Multiset of n samples $x_1, \dots, x_n \in \mathcal{X}$, parameters $\varepsilon \in (0,1]$ and $k = |\mathcal{X}|$ such that $\varepsilon \geq \frac{15}{k^{1/4}}$

1: Set $\tau \leftarrow (1 - \frac{1}{k})^{n-1} - \frac{n\varepsilon^2}{8k}$ ▷ This is $\mathbb{E}_{\mathbf{u}_k}[Z_2] - \frac{n\varepsilon^2}{8k}$

2: Compute ▷ Can be done in $O(n)$ time if $\mathcal{X}$ is known, $O(n \log n)$ if only k is.

$$Z_2 = \frac{1}{n} \sum_{j \in \mathcal{X}} \mathbb{1}\{N_j = 1\}$$

where $N_j \leftarrow \sum_{t=1}^{n} \mathbb{1}\{x_t = j\}$.

3: **if** $Z_2 \leq \tau$ **then return** 0 ▷ Not uniform

4: **else return** 1 ▷ Uniform

Our first step is to make our back-of-the-envelope computation above rigorous, and lower bound the gap $\Delta(\mathbf{p}) := \mathbb{E}_{\mathbf{u}_k}[Z_2] - \mathbb{E}_{\mathbf{p}}[Z_2]$ between the far and uniform cases.

Lemma 2.2. If $n \leq k$, we have

$$\mathbb{E}_{\mathbf{u}_k}[Z_2] - \mathbb{E}_{\mathbf{p}}[Z_2] \geq \frac{n}{16k} \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k)^2 .$$

Proof. Denote this gap by $\Delta(\mathbf{p})$. From (2.15), we can explicitly write

$$\begin{aligned}
\Delta(\mathbf{p}) &= (1 - 1/k)^{n-1} - \sum_{i \in \mathcal{X}} \mathbf{p}(i)(1 - \mathbf{p}(i))^{n-1} \\
&= \sum_{i \in \mathcal{X}} \mathbf{p}(i) \Big((1 - 1/k)^{n-1} - (1 - \mathbf{p}(i))^{n-1} \Big) \\
&= (1 - 1/k)^{n-1} \sum_{i \in \mathcal{X}} \mathbf{p}(i) \left(1 - \left(\frac{1 - \mathbf{p}(i)}{1 - 1/k} \right)^{n-1} \right)
\end{aligned}$$

where in the second line we used $\sum_i \mathbf{p}(i) = 1$ to "hide 1." Defining $f\colon [0,1] \to \mathbb{R}$ by

$$f(x) = x \left(1 - \left(\frac{1 - x}{1 - 1/k} \right)^{n-1} \right)$$

and using the fact that $n \leq k$ to write $(1 - 1/k)^{n-1} \geq (1 - 1/k)^k \geq 1/4$ (as $k \geq 2$), we are left with $\Delta(\mathbf{p}) \geq \frac{1}{4} \sum_{i \in \mathcal{X}} f(\mathbf{p}(i))$. At this point, we would like to rely on tools from our arsenal (convexity or concavity for

Jensen's inequality, monotonicity, etc.); unfortunately, the function f is not very well-behaved, and is neither concave, convex, or monotone. It does satisfy $f(0) = f(1/k) = 0$, $f(1) = 1$, but that is not quite enough. Instead, we will find a "good" lower bound g on f, which will allow us to reason more easily: specifically, we will set, for $x \in [0, 1]$,

$$g(x) = \frac{n-1}{k-1} \cdot (x - 1/k) + \frac{n-1}{2(1-1/k)}(x - 1/k)^2 \mathbb{1}\{x \le 1/k\}\,.$$

Let us demystify this choice a little. The first coefficient is simply $f'(1/k)$, while the second has been chosen as the largest possible value such that $g'(0) \ge 0$.[5] Moreover, when summing $\sum_i g(\mathbf{p}(i))$, the linear term will just cancel out and we will be left with a quadratic term of the form $\sum_i g(\mathbf{p}(i)) \asymp \sum_i (\mathbf{p}(i) - 1/k)^2 \mathbb{1}\{\mathbf{p}(i) \le 1/k\}$, which we can hope to relate to $\|\mathbf{p} - \mathbf{u}_k\|_2^2$ (the same thing, without the indicator). An illustration of f and g is given in Fig. 2.1.

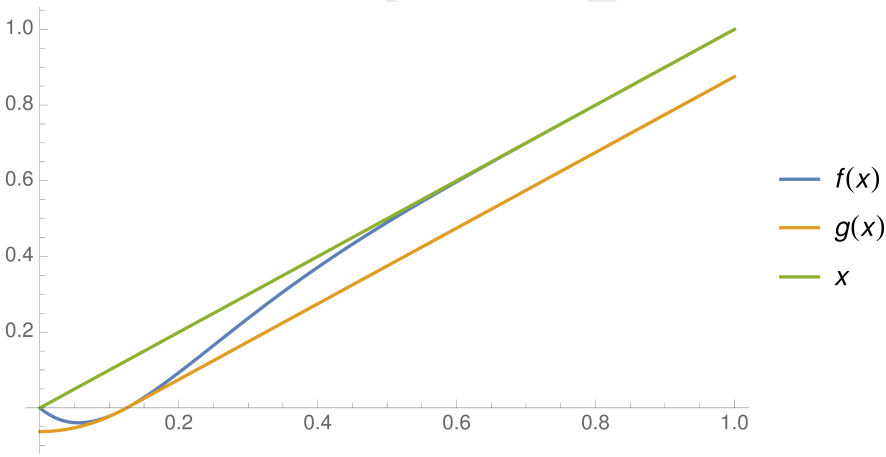

Figure 2.1: Our choice of g, here depicted for $k = 8$, $n = 7$.

E: You are the reader.

To continue our analysis, we will rely on the following technical claim, whose proof is calculus and left to the reader.

Claim 2.1. Fix $\alpha \in (0, 1)$ and $\beta \ge 1$ such that $\frac{\alpha\beta}{1-\alpha} \le 1$; and define

[5] While we will not require it, this ensures g is nondecreasing, which is nice.

$f_{\alpha,\beta}, g_{\alpha,\beta}\colon [0,1] \to \mathbb{R}$ by $f_{\alpha,\beta}(x) = x\left(1 - \left(\frac{1-x}{1-\alpha}\right)^{\beta}\right)$ and

$$g_{\alpha,\beta}(x) = \frac{\alpha\beta}{1-\alpha}(x-\alpha) + \frac{\beta}{2(1-\alpha)}(x-\alpha)^2 \mathbb{1}\{x \le \alpha\}\,.$$

Then $g_{\alpha,\beta}$ is nondecreasing, and $f_{\alpha,\beta} \ge g_{\alpha,\beta}$.

Applying this to $\alpha := 1/k$ and $\beta := n-1$ (which, since $n \le k$, satisfy the assumptions) leads to

$$\begin{aligned}
\Delta(\mathbf{p}) &\ge \frac{1}{4}\sum_{i\in\mathcal{X}} g(\mathbf{p}(i)) \\
&= \frac{n-1}{8(1-1/k)}\sum_{i\in\mathcal{X}} (\mathbf{p}(i) - 1/k)^2 \mathbb{1}\{\mathbf{p}(i) \le 1/k\} \\
&\ge \frac{n-1}{8(k-1)}\left(\sum_{i\in\mathcal{X}} |\mathbf{p}(i) - 1/k| \mathbb{1}\{\mathbf{p}(i) \le 1/k\}\right)^2 \\
&= \frac{n-1}{8(k-1)} \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k)^2\,,
\end{aligned}$$

where the first line uses $f \ge g$, the second from cancellation of the linear term of g, the third is Cauchy–Schwarz, and the last follows from the definition of total variation distance. Using $n \ge 2$ to write $\frac{n-1}{k-1} \ge \frac{n}{2k}$ concludes the proof. □

At this point, we have half of the puzzle: it remains to get a handle on the variance of Z_2 in both the far and uniform cases. We start by providing an exact (albeit unwieldy) expression for it.

$$\begin{aligned}
\mathrm{Var}[Z_2] &= \frac{1}{n^2}\sum_{i,j\in\mathcal{X}} \mathbb{E}[\mathbb{1}\{N_i = 1\}\mathbb{1}\{N_j = 1\}] - \mathbb{E}[Z_2]^2 \\
&= \frac{1}{n^2}\sum_{i\in\mathcal{X}} \mathbb{E}[\mathbb{1}\{N_i = 1\}] + \frac{1}{n^2}\sum_{i\ne j} \mathbb{E}[\mathbb{1}\{N_i = 1\}\mathbb{1}\{N_j = 1\}] - \mathbb{E}[Z_2]^2 \\
&= \frac{1}{n}\mathbb{E}[Z_2] - \mathbb{E}[Z_2]^2 + \frac{n-1}{n}\sum_{i\ne j} \mathbf{p}(i)\mathbf{p}(j)(1 - (\mathbf{p}(i) + \mathbf{p}(j)))^{n-2}\,,
\end{aligned}$$

where the last line relied on Eq. (2.14) to recognize $\mathbb{E}[Z_2]$ in the first term, and on the fact that $\mathbb{E}[\mathbb{1}\{N_i = 1\}\mathbb{1}\{N_j = 1\}] = n(n-1)\mathbf{p}(i)\mathbf{p}(j)(1 - (\mathbf{p}(i) + \mathbf{p}(j)))^{n-2}$ for $i \ne j$. (Indeed, $\mathbb{1}\{N_i = 1\}\mathbb{1}\{N_j = 1\}$ is equal to

one if, and only if, out of the n samples two fall on i and j, respectively, and the $n-2$ others hit $\mathcal{X} \setminus \{i,j\}$.)

This looks quite unwieldy; however, we can rearrange the term $\mathbb{E}[Z_2]^2$ to obtain the nicer expression

$$\begin{aligned}\mathrm{Var}[Z_2] &= \frac{\mathbb{E}[Z_2](1-\mathbb{E}[Z_2])}{n} + \frac{n-1}{n}\left(\sum_{i\neq j}\mathbf{p}(i)\mathbf{p}(j)(1-\mathbf{p}(i)-\mathbf{p}(j))^{n-2} - \mathbb{E}[Z_2]^2\right)\\ &\leq \frac{1-\mathbb{E}[Z_2]}{n} + \sum_{i\neq j}\mathbf{p}(i)\mathbf{p}(j)(1-\mathbf{p}(i)-\mathbf{p}(j))^{n-2} - \mathbb{E}[Z_2]^2\end{aligned} \tag{2.16}$$

For the uniform case, Eq. (2.16) simplifies quite a bit, and we get

$$\begin{aligned}\mathrm{Var}_{\mathbf{u}_k}[Z_2] &\leq \frac{1}{n}\left(1-\left(1-\frac{1}{k}\right)^{n-1}\right) + \left(1-\frac{2}{k}\right)^{n-2} - \left(1-\frac{1}{k}\right)^{2(n-1)}\\ &\leq \frac{n-1}{nk} + \left(1-\frac{1}{k}\right)^{2(n-2)} - \left(1-\frac{1}{k}\right)^{2(n-1)}\\ &\leq \frac{1}{k} + \left(1-\frac{1}{k}\right)^{2(n-2)}\left(1-\left(1-\frac{1}{k}\right)^{2}\right)\\ &= \frac{1}{k} + \frac{2}{k}\left(1-\frac{1}{k}\right)^{2(n-2)}\left(1-\frac{1}{2k}\right) \leq \frac{3}{k}\end{aligned} \tag{2.17}$$

where the second inequality follows from $1-2x \leq (1-x)^2$, and $(1-x)^m \geq 1-mx$ for $m \geq 1$ and $x \leq 1$. (As a side note, we proved along the way that $\frac{1}{n}(1-\mathbb{E}_{\mathbf{u}_k}[Z_2]) \leq \frac{1}{k}$, which will come in handy.)

This looks great! We just proved that, at least in the uniform case, $\mathrm{Var}_{\mathbf{u}_k}[Z_2] \leq 3/k$. By the same rule of thumb as in the previous argument (Eq. (2.6)), we expect our test to work as long as the standard deviation (the "noise") of our statistic is smaller than the gap in expectations (the "signal"), which by Lemma 2.2 gives the condition

$$\mathrm{Var}_{\mathbf{u}_k}[Z_2] \leq \frac{3}{k} \ll \frac{\varepsilon^4 n^2}{16k^2} \leq \Delta(\mathbf{p})^2\,.$$

Reorganizing, this yields the condition $n \gg \sqrt{k}/\varepsilon^2$, which is exactly what we want to prove. The problem, of course, is that we so far only have bounded the variance in one of the two cases; to conclude, we need the last quarter of the puzzle.

To do so, we will invoke the following:

Lemma 2.3. Fix $m \geq 1$ and $k \in \mathbb{N}$. For any $x_1, \dots, x_k \geq 0$ such that $\sum_{i=1}^k x_i = 1$, we have

$$\frac{m \sum_{1 \leq i < j \leq k} x_i x_j \big((1 - x_i - x_j)^{m-1} - (1 - x_i)^m (1 - x_j)^m\big)}{\sum_{i=1}^k x_i (1 - (1 - x_i)^m)} \leq 1$$

Unfortunately, there is not much intuition we can provide about *why* this inequality holds; and we defer its proof to the end of the chapter (Section 2.5), focusing for now on how it will provide us with the last piece of said puzzle. In view of resuming from Eq. (2.16), we expand $\mathbb{E}[Z_2]^2$ to write

$$\begin{aligned}
&\sum_{i \neq j} \mathbf{p}(i)\mathbf{p}(j)(1 - \mathbf{p}(i) - \mathbf{p}(j))^{n-2} - \mathbb{E}[Z_2]^2 \\
&\quad = \sum_{i \neq j} \mathbf{p}(i)\mathbf{p}(j)(1 - \mathbf{p}(i) - \mathbf{p}(j))^{n-2} - \sum_{i,j} \mathbf{p}(i)\mathbf{p}(j)(1 - \mathbf{p}(i))^{n-1}(1 - \mathbf{p}(j))^{n-1} \\
&\quad \leq \sum_{i \neq j} \mathbf{p}(i)\mathbf{p}(j)\Big((1 - \mathbf{p}(i) - \mathbf{p}(j))^{n-2} - (1 - \mathbf{p}(i))^{n-1}(1 - \mathbf{p}(j))^{n-1}\Big) \\
&\quad \leq \frac{1}{n-1} \sum_{i=1}^k \mathbf{p}(i)(1 - (1 - \mathbf{p}(i))^{n-1}) = \frac{1 - \mathbb{E}[Z_2]}{n-1}
\end{aligned}$$

where the last inequality is Lemma 2.3 applied with $m = n - 1$ and $x_i = \mathbf{p}(i)$. Then, using this in Eq. (2.16) (and bounding $1/(n-1) \leq 2/n$) leads to

$$\begin{aligned}
\mathrm{Var}[Z_2] &\leq 3 \cdot \frac{1 - \mathbb{E}_{\mathbf{p}}[Z_2]}{n} = 3\left(\frac{1 - \mathbb{E}_{\mathbf{u}_k}[Z_2]}{n} + \frac{\mathbb{E}_{\mathbf{u}_k}[Z_2] - \mathbb{E}_{\mathbf{p}}[Z_2]}{n}\right) \\
&\leq 3\left(\frac{1}{k} + \frac{\Delta(\mathbf{p})}{n}\right). \qquad (2.18)
\end{aligned}$$

Before invoking Chebyshev's inequality, let us see why this is wonderful news. The first term of the bound, $3/k$, is the same as in the uniform case, and we have discussed before how it would by itself lead to the desired sufficient condition $n \gg \sqrt{k}/\varepsilon^2$. The second is new; by the same rule of thumb, it will lead to the condition

$$\frac{3\Delta(\mathbf{p})}{n} \ll \Delta(\mathbf{p})^2$$

where one $\Delta(\mathbf{p})$ crucially cancels out, leaving us with $\Delta(\mathbf{p}) \gg 1/n$. Since $\Delta(\mathbf{p}) \geq \frac{\varepsilon^2 n}{16k}$, a sufficient condition then becomes $\frac{\varepsilon^2 n}{k} \gg \frac{1}{n}$, which then will be satisfied as soon as $n \gg \sqrt{k}/\varepsilon$.

Now that we have all the pieces of the puzzle, let us establish the main result of this subsection:

Theorem 2.4. The unique-elements tester (Algorithm 2) is a testing algorithm for uniformity with sample complexity $n(k, \varepsilon, 1/3) = O(\sqrt{k}/\varepsilon^2)$ and time complexity $O(n)$, provided that $\varepsilon \geq 15/k^{1/4}$.

Proof. For any $\mathbf{p} \in \Delta_k$, let as before $\Delta(\mathbf{p}) := \mathbb{E}_{\mathbf{u}_k}[Z_2] - \mathbb{E}_{\mathbf{p}}[Z_2]$. Of course, if $\mathbf{p} = \mathbf{u}_k$ then $\Delta(\mathbf{p}) = 0$, and we know from Lemma 2.2 that $\Delta(\mathbf{p}) \geq \Delta := \frac{n\varepsilon^2}{16k}$ whenever $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) \geq \varepsilon$.[6] We also obtained earlier (in Eqs. (2.17) and (2.18)) a bound in the variance of Z_2 in both the uniform and "far" cases, so we have all the ingredients we need. Define our threshold

$$\tau := \mathbb{E}_{\mathbf{u}_k}[Z_2] - \frac{\Delta}{2}$$

as in Algorithm 2.

- In the uniform case, the probability to output 0 (and thus make a mistake) is bounded as

$$\Pr[Z_2 \leq \tau] = \Pr\left[Z_2 \leq \mathbb{E}_{\mathbf{u}_k}[Z_2] - \frac{\Delta}{2}\right] \leq \frac{4 \operatorname{Var}_{\mathbf{u}_k}[Z_2]}{\Delta^2} \leq \frac{3072k}{\varepsilon^4 n^2}$$

  by Chebyshev's inequality, using $\operatorname{Var}_{\mathbf{u}_k}[Z_2] \leq 3/k$ and the definition of Δ. This in turn is less than $1/3$ as long as $n \geq 96\sqrt{k}/\varepsilon^2$.

- In the "far" case, since $\mathbb{E}_{\mathbf{p}}[Z_2] = \mathbb{E}_{\mathbf{u}_k}[Z_2] - \Delta(\mathbf{p})$ and $\Delta(\mathbf{p}) \geq \Delta$ the probability to err by outputting 1 is

$$\begin{aligned}
\Pr[Z_2 > \tau] &= \Pr\left[Z_2 > \mathbb{E}_{\mathbf{p}}[Z_2] + \frac{\Delta(\mathbf{p})}{2} + \frac{\Delta(\mathbf{p}) - \Delta}{2}\right] \\
&\leq \Pr\left[Z_2 > \mathbb{E}_{\mathbf{p}}[Z_2] + \frac{\Delta(\mathbf{p})}{2}\right] \\
&\leq \frac{4 \operatorname{Var}_{\mathbf{p}}[Z_2]}{\Delta(\mathbf{p})^2} \leq \frac{12}{k\Delta(\mathbf{p})^2} + \frac{12}{n\Delta(\mathbf{p})} \\
&\leq \frac{3072k}{\varepsilon^4 n^2} + \frac{192k}{\varepsilon^2 n^2} \leq \frac{3264k}{\varepsilon^4 n^2}
\end{aligned}$$

[6] We assume throughout $n \leq k$, and will enforce this at the end.

using $\mathrm{Var}_{\mathbf{p}}[Z_2] \le 3/k + 3\Delta(\mathbf{p})/n$. The resulting bound is then less than $1/3$ for $n \ge 99\sqrt{k}/\varepsilon^2$.

The above analysis shows that both errors are less than $1/3$ for, say, $n = \lceil 99\sqrt{k}/\varepsilon^2 \rceil$. However, we did rely on Lemma 2.2, which requires $n \le k$; given our choice of n, this in turns imposes a condition on ε (for instance, one can check that $\varepsilon \ge 15/k^{1/4}$ suffices). □

2.1.4 Modified χ^2

If you are a statistician, or just took a Statistics class, or even got lost on Wikipedia at some point and ended up on the wrong page at the wrong time, you may know of Pearson's χ^2 test for goodness-of-fit: for every element i of the domain, count how many times it appeared in the samples, N_i. Compute $\sum_i \frac{(N_i - n/k)^2}{n/k}$. Compare the result to a predetermined threshold. This very natural idea, maybe not surprisingly, works well! In particular, since $N_i \sim \mathrm{Bin}(n, \mathbf{p}(i))$, one can derive

$$\mathbb{E}\left[\sum_{i=1}^{k} \frac{(N_i - n/k)^2}{n/k}\right] = \frac{k}{n}\sum_{i=1}^{k} \mathbb{E}\left[\left(N_i - \frac{n}{k}\right)^2\right] = k(n-1)\|\mathbf{p} - \mathbf{u}_k\|_2^2 + k - 1\,,$$

E: Check it!

using moments of a Binomial random variable and Eq. (2.2). This does look like a reasonable way to estimate the distance to uniformity *via* the ℓ_2 distance again... unfortunately, the variance will be quite annoying, due to the correlations between terms of the sums.

To make the task easier, it is helpful to think of taking $\mathrm{Poisson}(n)$ samples instead of exactly n, which will greatly simplifies the analysis. Then, under this (slightly different) sampling model the N_i's become independent, with $N_i \sim \mathrm{Poisson}(n\mathbf{p}(i))$: this is called *Poissonization*, and can be done more or less without loss of generality since a $\mathrm{Poisson}(n)$ random variable will be between $0.99n$ and $1.01n$ with overwhelming probability. (See Appendix C for more on Poissonization, and why we can use it "without loss of generality").

The bad news is that it does not actually lead to the optimal sample complexity: Poissonization introduces a bit more variance (as we introduce extra randomness ourselves by taking a random number of samples), and so the variance of this χ^2 test can be too big due to

Algorithm 3 CHI-SQUARE TESTER

Require: Multiset of n samples $x_1, \ldots, x_n \in \mathcal{X}$, parameters $\varepsilon \in (0, 1]$ and $k = |\mathcal{X}|$ ▷ Assumes Poissonization

1: Set $\tau \leftarrow 2n\varepsilon^2$

2: Compute ▷ Can be done in $O(n)$ time if $\mathcal{X}$ is known, $O(n \log n)$ if only k is.

$$Z_3 = \sum_{j \in \mathcal{X}} \frac{(N_j - n/k)^2 - N_j}{n/k}$$

where $N_j \leftarrow \sum_{t=1}^{n} \mathbb{1}\{x_t = j\}$.

3: **if** $Z_3 \geq \tau$ **then return** 0 ▷ Not uniform

4: **else return** 1 ▷ Uniform

the elements we only expect to see zero or once (so, most of them). The *good* news is that a simple correction of that test, of the form

$$Z_3 = \sum_{i=1}^{k} \frac{(N_i - n/k)^2 - N_i}{n/k} \tag{2.19}$$

does have a much smaller variance, and a threshold test of the form "$Z_3 > \tau$?" will yield the right sample complexity.[7] Recalling that $N_i \sim \text{Poisson}(n\mathbf{p}(i))$ for all i, the expectation of Z_3 will then just be

$$\mathbb{E}[Z_3] = nk\|\mathbf{p} - \mathbf{u}_k\|_2^2 \tag{2.20}$$

which is perfect. Analyzing this test boils down, again, to bounding the variance of Z_3 and invoking Chebyshev's inequality... Before doing so, we will make a change which looks innocuous, but will come in quite handy in Section 2.2 when generalizing beyond the uniform distribution: let us rewrite

$$Z_3 = \sum_{i=1}^{k} \frac{(N_i - n\mathbf{u}_k(i))^2 - N_i}{n\mathbf{u}_k(i)}$$

where, of course, $\mathbf{u}_k(i) = 1/k$ for all $i \in [k]$. Recalling that, by Poissonization, all the N_i's are independent Poisson random variables, we

[7] In the multinomial ("non-Poissonized") case, substracting N_i was not necessary, since that would correspond to removing overall $\frac{k}{n}\sum_{i=1}^{k} N_i = k$, a constant term. In the Poissonized case, however, $\sum_{i=1}^{k} N_i \sim \text{Poisson}(n)$, and $\frac{k}{n}\sum_{i=1}^{k} N_i$ is not a constant – it is a random variable with expectation k and (large) variance k^2/n.

will invoke the technical claim below, which follows from (somewhat tedious, but straightforward) computations involving the moments of Poisson random variables:

E: Verify it: Exercise 2.3.

Claim 2.2. Let $\mu, \lambda \geq 0$. If $X \sim \text{Poisson}(\lambda)$, then $\mathbb{E}[(X-\mu)^2 - X] = (\lambda-\mu)^2$ and $\mathbb{E}[((X-\mu)^2 - X)^2] = (\lambda-\mu)^4 + 2\lambda^2 + 4\lambda(\lambda-\mu)^2$.

Given this, by linearity of expectation we immediately get[8]

$$\begin{aligned}\mathbb{E}[Z_3] &= \sum_{i=1}^{k} \frac{\mathbb{E}[(N_i - n\mathbf{u}_k(i))^2 - N_i]}{n\mathbf{u}_k(i)} = n\sum_{i=1}^{k} \frac{(\mathbf{p}(i)-\mathbf{u}_k(i))^2}{\mathbf{u}_k(i)} \\ &= n \cdot \chi^2(\mathbf{p} \,||\, \mathbf{u}_k) \end{aligned} \tag{2.21}$$

which here can further be simplified by $\chi^2(\mathbf{p} \,||\, \mathbf{u}_k) = k\|\mathbf{p}-\mathbf{u}_k\|_2^2$, as the denominator is constant and equal to $1/k$; this establishes Eq. (2.20).

Turning to the variance, we want to relate $\text{Var}[Z_3]$ to known quantities, and in particular $\mathbb{E}[Z_3]$. To do so, we use independence of the N_i's followed by the second part of the above claim to get

$$\begin{aligned}\text{Var}[Z_3] &= \sum_{i=1}^{k} \frac{\text{Var}[(N_i - n\mathbf{u}_k(i))^2 - N_i]}{n^2\mathbf{u}_k(i)^2} \\ &= \sum_{i=1}^{k} \frac{2n^2\mathbf{p}(i)^2 + 4n^3\mathbf{p}(i)(\mathbf{p}(i)-\mathbf{u}_k(i))^2}{n^2\mathbf{u}_k(i)^2} \\ &= 2\sum_{i=1}^{k} \frac{\mathbf{p}(i)^2}{\mathbf{u}_k(i)^2} + 4n\sum_{i=1}^{k} \frac{\mathbf{p}(i)(\mathbf{p}(i)-\mathbf{u}_k(i))^2}{\mathbf{u}_k(i)^2} \\ &\leq 2\sum_{i=1}^{k} \frac{\mathbf{p}(i)^2}{\mathbf{u}_k(i)^2} + 4n\sqrt{\sum_{i=1}^{k} \frac{\mathbf{p}(i)^2}{\mathbf{u}_k(i)^2}} \cdot \sqrt{\sum_{i=1}^{k} \frac{(\mathbf{p}(i)-\mathbf{u}_k(i))^4}{\mathbf{u}_k(i)^2}} \\ &\leq 2\sum_{i=1}^{k} \frac{\mathbf{p}(i)^2}{\mathbf{u}_k(i)^2} + 4n\sqrt{\sum_{i=1}^{k} \frac{\mathbf{p}(i)^2}{\mathbf{u}_k(i)^2}} \cdot \sum_{i=1}^{k} \frac{(\mathbf{p}(i)-\mathbf{u}_k(i))^2}{\mathbf{u}_k(i)} \\ &= 2\sum_{i=1}^{k} \frac{\mathbf{p}(i)^2}{\mathbf{u}_k(i)^2} + 4\sqrt{\sum_{i=1}^{k} \frac{\mathbf{p}(i)^2}{\mathbf{u}_k(i)^2}}\,\mathbb{E}[Z_3]\,, \end{aligned} \tag{2.22}$$

[8] Here, $\chi^2(\mathbf{p} \,||\, \mathbf{q}) = \sum_{i=1}^{k} \frac{(\mathbf{p}(i)-\mathbf{q}(i))^2}{\mathbf{q}(i)}$ denotes the *chi-square divergence* between distributions $\mathbf{p}$ and $\mathbf{q}$; for a refresher on this notion of distance, the reader is referred to Appendix B.

where the first inequality is Cauchy–Schwarz, and the second is monotonicity of ℓ_p norms: namely, $\ell_2 \le \ell_1$. In order to proceed further, we need to bound the quantity $\sum_{i=1}^{k} \frac{\mathbf{p}(i)^2}{\mathbf{u}_k(i)^2}$. The trick here will be to write, using $(a+b)^2 \le 2a^2 + 2b^2$,

$$\mathbf{p}(i)^2 = ((\mathbf{p}(i) - \mathbf{u}_k(i)) + \mathbf{u}_k(i))^2 \le 2(\mathbf{p}(i) - \mathbf{u}_k(i))^2 + 2\mathbf{u}_k(i)^2 \,,$$

since then we have

$$\sum_{i=1}^{k} \frac{\mathbf{p}(i)^2}{\mathbf{u}_k(i)^2} \le 2k + 2\sum_{i=1}^{k} \frac{(\mathbf{p}(i) - \mathbf{u}_k(i))^2}{\mathbf{u}_k(i)^2} = 2k + 2k \sum_{i=1}^{k} \frac{(\mathbf{p}(i) - \mathbf{u}_k(i))^2}{\mathbf{u}_k(i)}$$
$$= 2k\left(1 + \frac{\mathbb{E}[Z_3]}{n}\right).$$

Putting this back in Eq. (2.22), we get

$$\mathrm{Var}[Z_3] \le 4k\left(1 + \frac{\mathbb{E}[Z_3]}{n}\right) + 4\sqrt{2}k^{1/2}\mathbb{E}[Z_3] + 4\sqrt{2}\frac{k^{1/2}}{n^{1/2}}\mathbb{E}[Z_3]^{3/2} \quad (2.23)$$

In particular, in the uniform case $\mathbb{E}_{\mathbf{u}_k}[Z_3] = 0$, and so $\mathrm{Var}_{\mathbf{u}_k}[Z_3] \le 4k$. Before formally analyzing the resulting sample complexity via (once more) Chebyshev's inequality, let us do the usual check and compare standard deviation (noise) to expectation gap (signal), and see if things look promising. The gap in expectation will be, given Eq. (2.21), at least $\Delta := nk \cdot \frac{4\varepsilon^2}{k} = 4n\varepsilon^2$; so, in the uniform case, we need $\mathrm{Var}_{\mathbf{u}_k}[Z_3] \ll \Delta^2$ which, given the above, is satisfied as long as $n \gg \sqrt{k}/\varepsilon^2$, since then

$$\mathrm{Var}_{\mathbf{u}_k}[Z_3] \le 4k \ll 16n^2\varepsilon^4 \le \Delta^2 \,.$$

In the "far" case, for $\mathbf{p}$ at total variation distance at least ε from uniform, the condition is $\mathrm{Var}_{\mathbf{p}}[Z_3] \ll \Delta(\mathbf{p})^2 = \mathbb{E}_{\mathbf{p}}[Z_3]^2$, which by Eq. (2.23) will require

$$\max\left(k, \frac{k}{n}\mathbb{E}_{\mathbf{p}}[Z_3], k^{1/2}\mathbb{E}_{\mathbf{p}}[Z_3], \frac{k^{1/2}}{n^{1/2}}\mathbb{E}_{\mathbf{p}}[Z_3]^{3/2}\right) \ll \mathbb{E}_{\mathbf{p}}[Z_3]^2 \,.$$

Considering each term separately, simplifying, and recalling that $\mathbb{E}_{\mathbf{p}}[Z_3] \ge 4n\varepsilon^2$, we see that this will also hold as long as $n \gg \sqrt{k}/\varepsilon^2$.

We will make this formal, and show the following:

Theorem 2.5. The χ^2-based tester (Algorithm 3) is a testing algorithm for uniformity with sample complexity $n(k, \varepsilon, 1/3) = O(\sqrt{k}/\varepsilon^2)$ and time complexity $O(n)$ in the Poissonized setting.

Proof. For any $\mathbf{p} \in \Delta_k$, let as before $\Delta(\mathbf{p}) := \mathbb{E}_{\mathbf{p}}[Z_3]$. By Eq. (2.20), we know that if $\mathbf{p} = \mathbf{u}_k$ then $\Delta(\mathbf{p}) = 0$, and that $\Delta(\mathbf{p}) \geq \Delta := 4n\varepsilon^2$ whenever $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) \geq \varepsilon$ (recalling Eq. (2.1)). We also have our variance bound from Eq. (2.23). Define our threshold

$$\tau := \frac{\Delta}{2}$$

as in Algorithm 3.

- In the uniform case, where $\mathbb{E}_{\mathbf{u}_k}[Z_3] = 0$, the probability to output 0 (and thus make a mistake) is bounded as

$$\Pr[Z_3 \geq \tau] \leq \frac{4 \operatorname{Var}_{\mathbf{u}_k}[Z_3]}{\Delta^2} \leq \frac{k}{\varepsilon^4 n^2}$$

  by Chebyshev's inequality, using $\operatorname{Var}_{\mathbf{u}_k}[Z_3] \leq 4k$ and the definition of Δ. This in turn is less than $1/3$ as long as $n \geq \sqrt{3k}/\varepsilon^2$.

- In the "far" case, since $\mathbb{E}_{\mathbf{p}}[Z_3] \geq \Delta$ the probability to err by outputting 1 is

$$\begin{aligned}
\Pr[Z_3 < \tau] &\leq \Pr\left[|Z_3 - \mathbb{E}_{\mathbf{p}}[Z_3]| > \frac{\Delta(\mathbf{p})}{2}\right] \\
&\leq \frac{4 \operatorname{Var}_{\mathbf{p}}[Z_3]}{\Delta(\mathbf{p})^2} \\
&\leq \frac{16k\left(1 + \frac{\mathbb{E}[Z_3]}{n}\right) + 16\sqrt{2}k^{1/2}\mathbb{E}[Z_3] + 16\sqrt{2}\frac{k^{1/2}}{n^{1/2}}\mathbb{E}[Z_3]^{3/2}}{\mathbb{E}[Z_3]^2} \\
&\leq \frac{k}{n^2\varepsilon^4} + \frac{4k}{n^2\varepsilon^2} + \frac{4\sqrt{2}k^{1/2}}{n\varepsilon^2} + \frac{8\sqrt{2}k^{1/2}}{n\varepsilon} \\
&\leq \frac{5k}{n^2\varepsilon^4} + \frac{12\sqrt{2}k^{1/2}}{n\varepsilon^2}
\end{aligned}$$

  first using Eq. (2.23), simplifying, then $\mathbb{E}[Z_3] \geq n\varepsilon^2$. By solving the inequality $5x^2 + 12\sqrt{2}x \leq 1/3$, we see that the result is at most $1/3$ for $n \geq 52\sqrt{k}/\varepsilon^2$.

The above analysis shows that both errors are less than $1/3$ for, say, $n = \lceil 52\sqrt{k}/\varepsilon^2 \rceil$; this concludes the proof. □

2.1.5 Empirical distance to uniform

Let us take a break from ℓ_2 and consider another, very natural thing to try: the *plugin estimator*. Since we have n samples from $\mathbf{p}$, we can compute the empirical estimator of the distribution, $\hat{\mathbf{p}}_n$, based on these n samples. Now, we want to test $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) = 0$ vs. $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) > \varepsilon$? Why not consider

$$Z_4 := \mathrm{d}_{\mathrm{TV}}(\hat{\mathbf{p}}, \mathbf{u}_k) \tag{2.24}$$

the empirical distance to uniform? A reason might be: *this sounds like a terrible idea.* Unless $n = \Omega(k)$ (which is much more than what we want), we will not have observed most of the domain elements even once, and the empirical distribution $\hat{\mathbf{p}}_n$ will be at distance $1 - o(1)$ from uniform, *even* if $\mathbf{p}$ is actually uniform.

That's the thing, though: the devil is in the $o(1)$ details. Sure, $\mathbb{E}[Z_4]$ will be *almost* 1 whether $\mathbf{p}$ is uniform or far from it unless $n = \Omega(k)$. But this "almost" will be different in the two cases! Carefully analyzing this tiny gap in expectation, and showing that Z_4 concentrates well enough around its expectation to preserve this tiny gap, amazingly leads to a tester with optimal sample complexity $n = \Theta(\sqrt{k}/\varepsilon^2)$. Let us see how.

For simplicity of exposition, we focus here on the case $n \leq k$, though the analysis of the test can be performed for extended to all parameter regimes – see Diakonikolas *et al.* (2018) for the full general case. The argument proceeds in two steps: first, computing and bounding the expectation of Z_4 under the uniform and the far cases separately is not going to really work, so instead we will bound the expectation gap $\Delta(\mathbf{p}) := \mathbb{E}_{\mathbf{p}}[Z_4] - \mathbb{E}_{\mathbf{u}_k}[Z_4]$ directly (a little like in the case of the unique elements-based tester). Then, once the gap in expectation is established, we will once again argue concentration of Z_4 as usual by a variance-based (Chebyshev's inequality) argument – but this time diversifying our toolkit and using a different tool, the Efron–Stein lemma, to bound the variance.

Algorithm 4 EMPIRICAL-DISTANCE TESTER

Require: Multiset of n samples $x_1, \dots, x_n \in \mathcal{X}$, parameters $\varepsilon \in (0, 1]$ and $k = |\mathcal{X}|$

▷ Set the threshold τ, depending on the parameter regime.

if $n \le k$ **then**

  $\tau \leftarrow \frac{n^2\varepsilon^2}{32ek^2}$

else if $k < n \le k/\varepsilon^2$ **then**

  $\tau \leftarrow c_1 \cdot \varepsilon^2\sqrt{n/k}$ ▷ $c_1 > 0$ is some absolute constant.

else

  $\tau \leftarrow c_2 \cdot \varepsilon$ ▷ $c_2 > 0$ is some absolute constant.

Compute ▷ Can be done in $O(n)$ time if $\mathcal{X}$ is known, $O(n \log n)$ if only k is.

$$Z_4 = \mathrm{d}_{\mathrm{TV}}(\hat{\mathbf{p}}, \mathbf{u}_k) = \frac{1}{2}\sum_{j=1}^{k}\left|\frac{N_j}{n} - \frac{1}{k}\right|$$

where $N_j \leftarrow \sum_{t=1}^{n} \mathbb{1}\{x_t = j\}$.

if $Z_4 \ge \mathbb{E}_{\mathbf{u}_k}[Z_4] + \tau$ **then return** 0 ▷ Not uniform

else return 1 ▷ Uniform

The expectation gap $\boldsymbol{\Delta(\mathbf{p})}$. From the definition of total variation distance, we can rewrite

$$Z_4 = \sum_{i=1}^{k}\left(\hat{\mathbf{p}}_n(i) - \frac{1}{k}\right)\mathbb{1}\{\hat{\mathbf{p}}_n(i) > 1/k\} = \frac{1}{n}\sum_{i=1}^{k}\left(N_i - \frac{n}{k}\right)_+$$

where as previously N_i denotes the number of samples falling on element i, and $x_+ := \max(x, 0)$. For the sake of the analysis, we will introduce the multivariate function

$$S(x_1, \dots, x_n) := \frac{1}{n}\sum_{i=1}^{k}\left(N_i(x_1, \dots, x_n) - \frac{n}{k}\right)_+ \tag{2.25}$$

and the function $\mu\colon \Delta_k \to \mathbb{R}$ given by

$$\mu(\mathbf{p}) := \mathbb{E}_{\mathbf{p}}[S(X_1, \dots, X_n)]\,.$$

(Note that $\mathbb{E}_{\mathbf{p}}[Z_4] = \mu(\mathbf{p})$.) Then, since $N_i \sim \mathrm{Bin}(n, \mathbf{p}(i))$ under $\mathbf{p}$, we have the exact expression

$$\begin{aligned}\mu(\mathbf{p}) &= \frac{1}{n}\sum_{i=1}^{k}\sum_{\ell=0}^{n}\binom{n}{\ell}\mathbf{p}(i)^{\ell}(1-\mathbf{p}(i))^{n-\ell}\left(\ell - \frac{n}{k}\right)_+ \\ &= \frac{1}{n}\sum_{i=1}^{k}\sum_{\ell=1}^{n}\binom{n}{\ell}\mathbf{p}(i)^{\ell}(1-\mathbf{p}(i))^{n-\ell}\left(\ell - \frac{n}{k}\right)\end{aligned}$$

since $(\ell - n/k)_+ > 0$ for $\ell \geq \lceil n/k \rceil = 1$ (recall that we assume $n \leq k$). Using the facts that $\sum_{\ell=0}^{n}\binom{n}{\ell}\mathbf{p}(i)^{\ell}(1-\mathbf{p}(i))^{n-\ell} = 1$ and $\sum_{\ell=0}^{n}\ell\binom{n}{\ell}\mathbf{p}(i)^{\ell}(1-\mathbf{p}(i))^{n-\ell} = n\mathbf{p}(i)$, the inner sum considerably simplifies and we get

$$\mu(\mathbf{p}) = \frac{1}{k}\sum_{i=1}^{k}(1-\mathbf{p}(i))^{n} \tag{2.26}$$

Since our goal is to lower bound $\Delta(\mathbf{p}) = \mu(\mathbf{p}) - \mu(\mathbf{u}_k)$, the view of μ as a multivariate function suggests a Taylor expansion around $\mu(\mathbf{u}_k)$. Namely, for any $\mathbf{p} \in \Delta_k$, we can write by Taylor's theorem

$$\mu(\mathbf{p}) = \mu(\mathbf{u}_k) + \nabla\mu(\mathbf{u}_k)^{\top}(\mathbf{p}-\mathbf{u}_k) + \frac{1}{2}(\mathbf{p}-\mathbf{u}_k)^{\top}\mathbf{H}(\mathbf{q})(\mathbf{p}-\mathbf{u}_k)$$

where $\mathbf{q} = (1-\theta)\mathbf{u}_k + \theta\mathbf{p}$ for some $\theta \in [0,1]$, and $\mathbf{H}$ is the Hessian of μ: $\mathbf{H}_{i,j}(\mathbf{q}) = \frac{\partial^2\mu}{\partial x_i \partial x_j}(\mathbf{q})$. Given Eq. (2.26), we can compute explicitly both the gradient and the Hessian: first, denoting by $\mathbf{1}_k$ the all-one vector,

$$\nabla\mu(\mathbf{u}_k) = \left(\frac{\partial\mu}{\partial x_1}(\mathbf{u}_k), \ldots, \frac{\partial\mu}{\partial x_k}(\mathbf{u}_k)\right) = -\frac{n}{k}\left(1-\frac{1}{k}\right)^{n-1}\mathbf{1}_k$$

so $\nabla\mu(\mathbf{u}_k)^{\top}(\mathbf{p}-\mathbf{u}_k) = -\frac{n}{k}(1-\frac{1}{k})^{n-1}\sum_{i=1}^{k}(\mathbf{p}(i)-\frac{1}{k}) = 0$. Then, as μ is separable, $\mathbf{H}$ will be diagonal: $\mathbf{H}_{i,j}(\mathbf{q}) = 0$ for $i \neq j$, while

$$\mathbf{H}_{i,i}(\mathbf{q}) = \frac{n(n-1)}{k}(1-\mathbf{q}(i))^{n-2}$$

for all $i \in [k]$. Recalling our Taylor expansion, this means that

$$\begin{aligned}\Delta(\mathbf{p}) &= \frac{1}{2}(\mathbf{p}-\mathbf{u}_k)^{\top}\mathbf{H}(\mathbf{q})(\mathbf{p}-\mathbf{u}_k) \\ &= \frac{n(n-1)}{2k}\sum_{i=1}^{k}(1-\mathbf{q}(i))^{n-2}\left(\mathbf{p}(i)-\frac{1}{k}\right)^2\end{aligned} \tag{2.27}$$

where again $\mathbf{q}$ is a probability distribution such that $\mathbf{q} = (1-\theta)\mathbf{u}_k + \theta\mathbf{p}$ for some $\theta \in [0,1]$. To proceed, a natural idea is to restrict the sum to indices for which we can non-trivially bound $(1-\mathbf{q}(i))^{n-2}$; in particular, focusing on the indices i for which $\mathbf{p}(i) < 1/k$ will do, since for those we must have $\mathbf{q}(i) \le 1/k$ as well. This leads to writing

$$\begin{aligned}
\Delta(\mathbf{p}) &\ge \frac{n(n-1)}{2k}\sum_{i=1}^{k}(1-\mathbf{q}(i))^{n-2}\left(\mathbf{p}(i)-\frac{1}{k}\right)^2\mathbb{1}\{\mathbf{p}(i)<1/k\} \\
&\ge \frac{n(n-1)}{2k}\left(1-\frac{1}{k}\right)^{n-2}\sum_{i=1}^{k}\left(\mathbf{p}(i)-\frac{1}{k}\right)^2\mathbb{1}\{\mathbf{p}(i)<1/k\} \\
&\ge \frac{n(n-1)}{2k}\left(1-\frac{1}{k}\right)^{n-2}\frac{1}{k}\left(\sum_{i=1}^{k}\left(\frac{1}{k}-\mathbf{p}(i)\right)\mathbb{1}\{\mathbf{p}(i)<1/k\}\right)^2 \\
&\ge \frac{n(n-1)}{2k^2}\left(1-\frac{1}{k}\right)^{k-2}\varepsilon^2 \ge \frac{n^2\varepsilon^2}{4ek^2}
\end{aligned} \tag{2.28}$$

where the third inequality is Cauchy–Schwarz, the fourth is the definition of total variation distance along with $\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{u}_k) \ge \varepsilon$ (and, in the exponent, $n \le k$), and the last is $(1-1/k)^{k-2} \ge 1/e$ for $k \ge 2$.

Before turning to proving concentration of Z_4 (that is, showing that the expectation gap, our signal, is not drowned by the random fluctations of Z_4 around its expectation), we will make an observation which will help in the analysis. Since we restricted ourselves to the regime $n/k \le 1$, the i-th summand of S in Eq. (2.25) is non-zero only when $N_i \ge 1$, and thus we can rewrite

$$\begin{aligned}
Z_4 &= \frac{1}{n}\sum_{i=1}^{k}\left(\left(N_i-\frac{n}{k}\right)+\frac{n}{k}\mathbb{1}\{N_i=0\}\right) \\
&= \frac{1}{n}\sum_{i=1}^{k}\left(N_i-\frac{n}{k}\right)+\frac{1}{k}\sum_{i=1}^{k}\mathbb{1}\{N_i=0\} \\
&= \frac{1}{k}\sum_{i=1}^{k}\mathbb{1}\{N_i=0\}
\end{aligned} \tag{2.29}$$

where the last equality follows from $\sum_{i=1}^{k} N_i = n$. That is, in this regime, Z_4 is exactly the fraction of *unseen elements* of the domain; this readily allows one to retrieve Eq. (2.26), and also lets us relate Z_4

to the "unique elements" statistic Z_2 from Section 2.1.3.[9]

Concentration of Z_4. As you may have guessed, we will show that Z_4 concentrates around its expectation via Chebyshev's inequality, which requires us to bound $\mathrm{Var}[Z_4]$ in both the uniform and "far" cases. In order to diversify our toolkit, we will do so by invoking the *Efron–Stein inequality*, which allows us to bound the variance of a function of n independent random variables by the expected quadratic change of this function when only one sample is re-randomized: for a function f of n variables,

$$\mathrm{Var}[f(X)] \leq \frac{1}{2}\sum_{t=1}^{n}\mathbb{E}\left[\left(f(X) - f(X^{(t)})\right)^2\right] \tag{2.30}$$

where $X = (X_1 \ldots, X_n)$ and $X^{(t)} = (X_1, \ldots, X'_t, \ldots, X_n)$, with X'_t being an independent copy of X_t.

We apply this to Z_4, *i.e.*, the function S defined in Eq. (2.25); given that S is symmetric and that all X_t's are i.i.d. and distributed according to $\mathbf{p}$, we get

$$\begin{aligned}\mathrm{Var}[Z_4] &\leq \frac{n}{2}\mathbb{E}_{\mathbf{p}}\left[\left(S(X_1, \ldots, X_{n-1}, X_n) - S(X_1, \ldots, X_{n-1}, X'_n)\right)^2\right] \\ &= \frac{n}{2k^2}\mathbb{E}_{\mathbf{p}}\left[\left(\sum_{i=1}^{k}\left(\mathbb{1}\{N_i = 0\} - \mathbb{1}\{N'_i = 0\}\right)\right)^2\right]\end{aligned}$$

where we wrote $N_i := N_i(X_1, \ldots, X_{n-1}, X_n)$ and $N'_i := N_i(X_1, \ldots, X_{n-1}, X'_n)$, and used the expression of S from Eq. (2.29). To handle this quantity, observe that since only the n-th sample changes, $\sum_{i=1}^{k}|\mathbb{1}\{N_i = 0\} - \mathbb{1}\{N'_i = 0\}|$ is at most 1: and this happens when X_n falls on an element i not yet observed in the first $n-1$ samples while X'_n falls on an element j already

[9] Again, this relation only holds in the specific regime $n \leq k$; while we do nto cover the case $n > k$ here, the guarantees of Z_4 extend to this other regime; the identity Eq. (2.29), however, does not.

observed, or vice-versa. That is, we can write

$$\begin{aligned}\mathrm{Var}[Z_4] &\le \frac{n}{k^2}\sum_{i\neq j}\mathbf{p}(i)\mathbf{p}(j)\Pr[N_i(X_1,\dots,X_{n-1})=0, N_j(X_1,\dots,X_{n-1})\ge 1]\\ &= \frac{n}{k^2}\sum_{i\neq j}\mathbf{p}(i)\mathbf{p}(j)\Big((1-\mathbf{p}(i))^{n-1}-(1-\mathbf{p}(i)-\mathbf{p}(j))^{n-1}\Big)\\ &\le \frac{n}{k^2}\sum_{i,j}\mathbf{p}(i)\mathbf{p}(j)\Big(1-(1-\mathbf{p}(j))^{n-1}\Big)\\ &= \frac{n}{k^2}\Big(1-\sum_{j=1}^{k}\mathbf{p}(j)(1-\mathbf{p}(j))^{n-1}\Big) \qquad (2.31)\end{aligned}$$

where the first inequality follows the above discussion, taking the expectation over X_n, X'_n and using symmetry between the two cases mentioned; the second inequality uses monotonicity of the function $y \mapsto (1-x)^m - (1-x-y)^m$ on $[0,x]$, and adds back the diagonal terms $i=j$ to the sum afterwards; and the last equality uses $\sum_i \mathbf{p}(i) = 1$.[10]

Very conveniently, in the uniform case Eq. (2.31) leads to the bound

$$\mathrm{Var}_{\mathbf{u}_k}[Z_4] \le \frac{n}{k^2}\left(1-\left(1-\frac{1}{k}\right)^{n-1}\right) \le \frac{n(n-1)}{k^3} \qquad (2.32)$$

which combined with our bound Eq. (2.28) on the expectation gap leads to the condition

$$\frac{n^2}{k^3} \ll \frac{n^4\varepsilon^4}{k^4}$$

satisfied for $n \gg \sqrt{k}/\varepsilon^2$. In the far case, a similar bound is easy to obtain for any $\mathbf{p}$ with $\|\mathbf{p}\|_\infty \lesssim 1/k$, since

$$\mathrm{Var}_{\mathbf{p}}[Z_4] \le \frac{n}{k^2}\Big(1-\sum_{j=1}^{k}\mathbf{p}(j)(1-\|\mathbf{p}\|_\infty)^{n-1}\Big) \le \frac{n(n-1)}{k^2}\cdot\|\mathbf{p}\|_\infty\,, \qquad (2.33)$$

so it would be quite nice if we could argue this was, in some sense, the *only* case we needed to worry about. Which brings us to the second new tool to add to our toolkit: an argument based on *stochastic dominance*, which will let us do exactly that. First, let us recall the key concept:

[10] Interestingly, the bound we just obtained is exactly $\mathrm{Var}[Z_4] \le \frac{n}{k^2}(1-\mathbb{E}[Z_2])$, where Z_2 is the "unique elements tester" from Section 2.1.3.

Definition 2.1. Let A, B be two (real-valued) random variables. We say that A *stochastically dominates* B if $\Pr[A \geq t] \geq \Pr[B \geq t]$ for all $t \in \mathbb{R}$; or, in terms of cumulative distribution functions, $F_A(t) \leq F_B(t)$ for all t.

In our case, we have a random variable Z_4, and in the "far" case we want to prove $\Pr[Z_4 \geq \tau]$ is large (where τ is our threshold, based on the expectation gap). So if, for each $\mathbf{p}$ in the 'far" case, we can argue there is a $\mathbf{p}'$ with $\|\mathbf{p}'\|_\infty \lesssim 1/k$ (i.e., which we know how to analyze) such that $Z_4(\mathbf{p})$ stochastically dominates $Z_4(\mathbf{p}')$, then we are good.

To do so, we need two more steps: first, the notion of *majorization*, and how this relates to stochastic dominance.

Definition 2.2. For a vector $x \in \mathbb{R}^k$, define $x^\downarrow \in \mathbb{R}^k$ as the vector with the same components, but sorted in non-increasing order. Then, given two vectors $x, y \in \mathbb{R}^k$, we say that y *majorizes* x (denoted $y \succeq x$) if $\sum_{i=1}^\ell y_i^\downarrow \geq \sum_{i=1}^\ell x_i^\downarrow$ for all $1 \leq \ell \leq k$, and $\sum_{i=1}^k y_i^\downarrow = \sum_{i=1}^k x_i^\downarrow$.

The key relation between vector majorization and stochastic dominance is given in the following theorem:

Theorem 2.6. Let $f\colon \mathbb{R}^k \to \mathbb{R}$ by a symmetric convex function. For a probability distribution $\mathbf{p} \in \Delta_k$ and $n \in \mathbb{N}$, define the random variable $X_n(\mathbf{p})$ as follows: let $N_1, \ldots, N_k$ be the counts of each domain element among n i.i.d. samples from $\mathbf{p}$, and set $X_n(\mathbf{p}) = f(N_1, \ldots, N_k)$. Then, for any $\mathbf{p}, \mathbf{q}$ such that $\mathbf{p} \succeq \mathbf{q}$, $X_n(\mathbf{p})$ stochastically dominates $X_n(\mathbf{q})$.

We will not prove this theorem, but instead use it as follows. First, we show that for every $\mathbf{p}$ which is far from uniform there exists some $\bar{\mathbf{p}}$ which majorizes $\mathbf{p}$, is still far from uniform, and importantly has small ℓ_∞ norm:

Lemma 2.7. For any $\mathbf{p} \in \Delta_k$, let $\bar{\mathbf{p}} \in \Delta_k$ denote the probability distribution obtained by averaging $\mathbf{p}$ over its $K := \lceil k/2 \rceil$ heaviest elements. Then (1) $\mathbf{p} \succeq \bar{\mathbf{p}}$, (2) $\|\bar{\mathbf{p}}\|_\infty \leq 2/k$, and (3) $\mathrm{d}_{\mathrm{TV}}(\bar{\mathbf{p}}, \mathbf{u}_k) \geq \frac{1}{2}\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k)$.

Proof. Clearly, all properties are invariant by permutation of the domain, so we can assume for simplicity that $\mathbf{p}$ is non-increasing: $\mathbf{p}(1) \geq \mathbf{p}(2) \geq$

$\cdots \geq \mathbf{p}(k)$, and $\bar{\mathbf{p}}(1) = \cdots = \bar{\mathbf{p}}(K) = \frac{1}{K}\sum_{i=1}^{K}\mathbf{p}(i)$, $\bar{\mathbf{p}}(i) = \mathbf{p}(i)$ for $i \geq K$. In particular, this immediately implies (1), as well as (2) (since $\|\bar{\mathbf{p}}\|_\infty = \bar{\mathbf{p}}(1) \leq 2/k$).

Let us prove (3), which states that this averaging does not decrease the distance too much. Consider the minimal set $S \subseteq [k]$ such that $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{u}_k) = \mathbf{p}(S) - \mathbf{u}_k(S)$: we know that $S = \{\, x : \mathbf{p}(x) > 1/k \,\}$. By motonicity of $\mathbf{p}$, this implies that $S = \{1, \ldots, \ell\}$ for some $\ell \geq 1$.

- If $\ell \geq K$, then $\bar{\mathbf{p}}(S) = \mathbf{p}(S)$, and so $\mathrm{d_{TV}}(\bar{\mathbf{p}}, \mathbf{u}_k) = \mathrm{d_{TV}}(\mathbf{p}, \mathbf{u}_k)$;
- otherwise, $\ell < K$, in which case we look at $T = [k] \setminus S$, which satisfies $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{u}_k) = \mathbf{u}_k(T) - \mathbf{p}(T)$. Let $T' := \{K+1, \ldots, k\} \subsetneq T$; by monotonicity, $\frac{\mathbf{p}(T \setminus T')}{|T \setminus T'|} \geq \frac{\mathbf{p}(T')}{|T'|}$, implying $\mathbf{p}(T') \leq \frac{|T'|}{|T|}\mathbf{p}(T)$. But then, since $\mathbf{p} = \bar{\mathbf{p}}$ on T',

$$\begin{aligned}\mathrm{d_{TV}}(\bar{\mathbf{p}}, \mathbf{u}_k) &= \sup_{R \subseteq [k]} (\mathbf{u}_k(R) - \bar{\mathbf{p}}(R)) \geq \mathbf{u}_k(T') - \bar{\mathbf{p}}(T') \\ &= \frac{|T'|}{|T|}\mathbf{u}_k(T) - \mathbf{p}(T') \geq \frac{|T'|}{|T|}(\mathbf{u}_k(T) - \mathbf{p}(T)) \\ &= \frac{|T'|}{|T|}\mathrm{d_{TV}}(\mathbf{p}, \mathbf{u}_k) \geq \frac{1}{2}\mathrm{d_{TV}}(\mathbf{p}, \mathbf{u}_k)\,,\end{aligned}$$

the last inequality following from $\frac{|T'|}{|T|} = \frac{k-K}{k-\ell} \geq \frac{k-\lceil k/2 \rceil}{k-1} \geq \frac{1}{2}$.

This concludes the proof of the lemma. □

Note that by the data processing inequality Fact 1.1, we have $\mathrm{d_{TV}}(\bar{\mathbf{p}}, \mathbf{u}_k) \leq \mathrm{d_{TV}}(\mathbf{p}, \mathbf{u}_k)$, so the averaging can change the distance to uniformity by a factor at most 2.[11]

Combining Theorem 2.6 and Lemma 2.7 (conveniently, the function S from Eq. (2.25) is indeed symmetric and convex), we get that for every $\mathbf{p}$ such that $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{u}_k) \geq \varepsilon$, there exists some "nicer" $\bar{\mathbf{p}}$ such that $\mathrm{d_{TV}}(\bar{\mathbf{p}}, \mathbf{u}_k) \geq \varepsilon/2$ with $\|\bar{\mathbf{p}}\|_\infty \leq 2/k$ and, for any choice of threshold τ,

$$\Pr_{\mathbf{p}}[\, Z_4 \geq \tau \,] \geq \Pr_{\bar{\mathbf{p}}}[\, Z_4 \geq \tau \,]\,. \tag{2.34}$$

[11] Moreover, this factor 2 cannot be improved upon, as one can check with, e.g., $\mathbf{p} = \frac{k}{k-1}\varepsilon\delta_1 + \left(1 - \frac{k}{k-1}\varepsilon\right)\mathbf{u}_k$, for which $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{u}_k) = \varepsilon$, but $\mathrm{d_{TV}}(\bar{\mathbf{p}}, \mathbf{u}_k) = \frac{1+o(1)}{2}\varepsilon$.

Thus, it suffices to choose a suitable τ for which the RHS is at least $2/3$ (along with $\Pr_{\mathbf{u}_k}[\,Z_4 \geq \tau\,] \leq 1/3$, of course) to conclude. By Eq. (2.28) (but for $\varepsilon/2$, not ε), we know that the expectation gap for such a $\bar{\mathbf{p}}$ is at least $\Delta(\bar{\mathbf{p}}) \geq \frac{n^2\varepsilon^2}{16ek^2}$. Moreover, now we can use Eq. (2.33) to get the variance bound

$$\mathrm{Var}_{\bar{\mathbf{p}}}[Z_4] \leq \frac{2n(n-1)}{k^3}\,, \tag{2.35}$$

which lets us establish the following theorem:

Theorem 2.8. The empirical-distance tester (Algorithm 4) is a testing algorithm for uniformity with sample complexity $n(k,\varepsilon,1/3) = O(\sqrt{k}/\varepsilon^2)$ and time complexity $O(n)$.

Proof. As discussed above, we only prove here the case $n \leq k$, *i.e.*, $\varepsilon = \Omega\big(1/k^{1/4}\big)$; however, the statement holds for all regimes. For any $\mathbf{p} \in \Delta_k$, let $\Delta(\mathbf{p}) := \mathbb{E}_{\mathbf{p}}[Z_4] - \mathbb{E}_{\mathbf{u}_k}[Z_4]$. Clearly, if $\mathbf{p} = \mathbf{u}_k$ then $\Delta(\mathbf{p}) = 0$; if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{u}_k) \geq \varepsilon$, by stochastic dominance it suffices to consider the corresponding $\bar{\mathbf{p}}$, which satisfies $\Delta(\bar{\mathbf{p}}) \geq \frac{n^2\varepsilon^2}{16ek^2} := \Delta$. We also have our variance bounds from Eqs. (2.32) and (2.33). Define our threshold

$$\tau := \frac{\Delta}{2} = \frac{n^2\varepsilon^2}{32ek^2}$$

as in Algorithm 4 (for this regime of parameters).

- In the uniform case, the probability to output 0 (and thus make a mistake) is bounded as

$$\Pr[Z_4 \geq \mathbb{E}_{\mathbf{u}_k}[Z_4] + \tau] \leq \frac{\mathrm{Var}_{\mathbf{u}_k}[Z_4]}{\tau^2} \leq \frac{n(n-1)}{k^3} \cdot \frac{(32e)^2k^4}{n^4\varepsilon^4}$$

  which is less than $1/3$ as long as $n \geq 32e\sqrt{3k}/\varepsilon^2$.

- In the "far" case, the probability to err by outputting 1 is

$$\begin{aligned}
\Pr_{\mathbf{p}}[Z_4 < \mathbb{E}_{\mathbf{u}_k}[Z_4] + \tau] &\leq \Pr_{\bar{\mathbf{p}}}[Z_4 < \mathbb{E}_{\mathbf{u}_k}[Z_4] + \tau] \\
&\leq \Pr_{\bar{\mathbf{p}}}[Z_4 < \mathbb{E}_{\bar{\mathbf{p}}}[Z_4] - \tau] \\
&\leq \frac{\mathrm{Var}_{\mathbf{p}}[Z_4]}{\tau^2} \\
&\leq \frac{2n(n-1)}{k^3} \cdot \frac{(32e)^2k^4}{n^4\varepsilon^4}
\end{aligned}$$

using Eq. (2.33), simplifying, then $\mathbb{E}[Z_3] \geq n\varepsilon^2$. This is less than $1/3$ as long as $n \geq 32e\sqrt{6k}/\varepsilon^2$.

The above analysis shows that both errors are less than $1/3$ for, say, $n = \lceil 32e\sqrt{6k}/\varepsilon^2 \rceil$; this concludes the proof. □

Remark 2.3. It is unclear whether one could avoid using the first (very convenient) "stochastic dominance hammer" (Theorem 2.6), and instead relate directly the bound on $\mathrm{Var}_{\mathbf{p}}[Z_4]$ from Eq. (2.31) to $\Delta(\mathbf{p})$ for arbitrary "far" distribution $\mathbf{p}$ in order to obtain the right sample complexity.

2.1.6 Random binary hashing

Now we turn to a tester that is *not* sample-optimal – but has other advantages, and whose analysis contains a couple insighful aspects. The main idea is that, while large domains are complicated, if there is one thing we know how to do optimally it is estimating the bias of a coin. That is, we know how to handle the case $k = 2$:

Fact 2.1 (Bias of a coin). Given i.i.d. samples from a Bernoulli with unknown parameter $\alpha \in [0, 1]$, estimating α to an additive η with probability $1 - \delta$ can be done with (and requires) $\Theta\left(\frac{\log(1/\delta)}{\eta^2}\right)$ samples.

E: Prove this! Exercise 2.4.

Of course, we do not have a probability distribution over $\{0, 1\}$ here, we have a much more problematic $(k-1)$-dimensional object. However, what prevents us from *making* our n samples into i.i.d. samples from a Bernoulli? Let us uniformly randomly partition the domain $[k]$ in two sets S and $[k] \setminus S$, and convert each sample into a $\{0, 1\}$ value accordingly: $X'_i := \mathbb{1}\{X_i \in S\}$, for $1 \leq i \leq n$.

This gives us n i.i.d. samples from a Bernoulli random variable with parameter $\alpha_S(\mathbf{p}) := \mathbf{p}(S)$: let us estimate it! Since we know S, we know exactly what this should be under the uniform distribution: $\alpha_S(\mathbf{u}_k) = \mathbf{u}_k(S) = |S|/k$. If only we could argue that $\alpha_S(\mathbf{p})$ noticeably differs from $\mathbf{u}_k(S)$ (with high probability over the random choice of S) whenever $\mathbf{p}$ is ε-far from uniform, we would have a tester: just estimate the bias $\alpha_S(\mathbf{p})$ to high enough accuracy. Luckily for us, this is, indeed, the case:

Algorithm 5 BINARY HASHING TESTER

Require: Multiset of n samples $x_1, \dots, x_n \in \mathcal{X}$, parameters $\varepsilon \in (0, 1]$ and $k = |\mathcal{X}|$

Set $\tau \leftarrow \frac{\varepsilon}{2\sqrt{2k}}$

Pick a random subset $S \subseteq [k]$ by including each $i \in [k]$ independently with probability $1/2$. ▷ 4-wise independence suffices.

Compute ▷ Can be done in $O(n)$ time.

$$Z_5 = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i \in S\}$$

if $\left| Z_5 - \frac{|S|}{k} \right| \geq \tau$ **then return** 0 ▷ Not uniform

else return 1 ▷ Uniform

Algorithm 6 AMPLIFYING THE BINARY HASHING TESTER

Require: Multiset of n samples $x_1, \dots, x_n \in \mathcal{X}$, parameters $\varepsilon \in (0, 1]$ and $k = |\mathcal{X}|$

▷ $T \in \mathbb{N}$ is an absolute constant, whose value is related to the hidden constant in Fact 2.1.

Partition the n samples into multisets $M_1 \dots, M_T$ of size $\lfloor n/T \rfloor$

for all $1 \leq t \leq T$ **do** ▷ Repeat the tester independently

  Run Algorithm 5 on the samples from M_t and parameters ε, k

  Let $\mathbf{b}_i \in \{0, 1\}$ be the output ▷ $\mathbf{b}_1, \dots, \mathbf{b}_T$ are i.i.d.

Use Fact 2.1 on $\mathbf{b}_1, \dots, \mathbf{b}_T$ with parameters $\eta \leftarrow \frac{1}{200}$ and $\delta \leftarrow \frac{1}{3}$ to get $\widehat{\mathbf{b}} \in [0, 1]$: estimate of the probability Algorithm 5 returns 0

if $\widehat{\mathbf{b}} \geq \frac{1}{100} + \eta$ **then return** 0 ▷ Not uniform

else return 1 ▷ Uniform

Lemma 2.9 (Random Binary Hashing). Let $\mathbf{p}, \mathbf{q} \in \Delta_k$. Then

$$\Pr_S \left[|\mathbf{p}(S) - \mathbf{q}(S)| \geq \frac{1}{2\sqrt{2}} \|\mathbf{p} - \mathbf{q}\|_2 \right] \geq \frac{1}{48},$$

where $S \subseteq [k]$ is a uniformly random subset of $[k]$.

We defer the proof of this lemma to the end of the subsection, and show how to use it. The first observation is that, regardless of the random choice of S, if $\mathbf{p} = \mathbf{u}_k$ then $\mathbf{p}(S) = \mathbf{u}_k(S) = |S|/k$ with probability

one, and so estimating the bias $\alpha_S(\mathbf{p})$ will (with high probability) not lead to any rejection. However, if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) \geq \varepsilon$, then by Eq. (2.1) $\|\mathbf{p} - \mathbf{u}_k\|_2 \geq 2\varepsilon/\sqrt{k}$, and so $|\mathbf{p}(S) - |S|/k| \geq \varepsilon/\sqrt{2k}$ with constant probability (over the choice of S). Whenever this happens, estimating the bias $\alpha_S(\mathbf{p})$ to $\pm\varepsilon/(2\sqrt{2k})$ will allow us to detect that $\mathbf{p}$ was far from uniform, and this can be done with

$$n \asymp \frac{1}{(\varepsilon/\sqrt{k})^2} = \frac{k}{\varepsilon^2}$$

samples by Fact 2.1. Of course, there is a catch: we will only detect it when there *is* some bias to detect, *i.e.*, when the random set S we choose is "good;" which we only proved happens with probability at least $1/48$. But a constant probability to distinguish between uniform and far from uniform is enough: repeating independently the test several times will let us amplify the success probability to $2/3$.

Theorem 2.10. The binary hashing tester (Algorithm 6) is a testing algorithm for uniformity with sample complexity $n(k, \varepsilon, 1/3) = O(k/\varepsilon^2)$ and time complexity $O(n)$.

Proof. Set $\tau \leftarrow \frac{\varepsilon}{2\sqrt{2k}}$, as in Algorithm 5, and for $\mathbf{p} \in \Delta_k$ let $\alpha_S(\mathbf{p}) := \mathbf{p}(S)$ denote the bias of the resulting coin; note that this is a random variable, over the choice of $S \subseteq [k]$, and that $\mathbb{E}[Z_5 \mid S] = \alpha_S(\mathbf{p})$. We will use Fact 2.1 to get, for $n = O(1/\tau^2)$, a probability at least $99/100$ to correctly estimate the bias up to an additive τ.

- In the uniform case, $\alpha_S(\mathbf{u}_k) = |S|/k$ always, and the probability to output 0 (and thus make a mistake) is therefore bounded for every S by

$$\Pr[\,|Z_5 - \alpha_S(\mathbf{u}_k)| \geq \tau \mid S\,] \leq 1/100\,,$$

  where the inequality holds by Fact 2.1.

- In the "far" case, denote by $\mathcal{E}$ the event that S is "good," that is

$$\mathcal{E} := \{|\alpha_S(\mathbf{p})) - \alpha_S(\mathbf{u}_k)| \geq \varepsilon/\sqrt{2k} = 2\tau\}$$

  which by Lemma 2.9 we know satisfies $\Pr_S[\mathcal{E}] \geq 1/48$.

The probability to err by outputting 1 is then at most

$$\begin{aligned}\Pr_{\mathbf{p},S}[|Z_5 - \alpha_S(\mathbf{u}_k)| < \tau] &\le \Pr_{\mathbf{p},S}[|Z_5 - \alpha_S(\mathbf{u}_k)| < \tau \mid \mathcal{E}] \Pr_S[\mathcal{E}] + \Pr_S\left[\overline{\mathcal{E}}\right] \\ &\le \frac{1}{48} \Pr_{\mathbf{p},S}[|Z_5 - \alpha_S(\mathbf{p})| > \tau \mid \mathcal{E}] + \frac{47}{48} \\ &\le \frac{1}{48} \cdot \frac{1}{100} + \frac{47}{48} < \frac{98}{100}\end{aligned}$$

where we again used Fact 2.1.

The above analysis shows that Algorithm 5 will output 0 (reject) with probability less than 1/100 under the uniform distribution, but with probability at least 2/100 under any "far" distribution. This is enough to be able to distinguish between the two cases: by repeating the test independently $O(1)$ times to estimate the rejection probability (which can be seen as a Bernoulli random variable with bias either more than 2/100 or less than 1/100), Fact 2.1 (again!) guarantees that we can decide which of the two cases holds, and be correct with probability at least 2/3: this is what Algorithm 6 does. This concludes the proof. □

It only remains to provide the proof of the binary hashing lemma:

Proof of Lemma 2.9. If $\mathbf{p} = \mathbf{q}$, then the statement trivially holds as $|\mathbf{p}(S) - \mathbf{q}(S)| = 0$ with probability one; we thus assume $\mathbf{p} \neq \mathbf{q}$. Write $\delta := \mathbf{p} - \mathbf{q} \in \mathbb{R}^k$, so that $\mathbf{p}(S) - \mathbf{q}(S) = \sum_{i=1}^k \delta_i S_i$, where $S_1, \dots, S_k$ are independent Bern(1/2) (where S_i indicates whether $i \in S$). Equivalently, since $\sum_{i=1}^k \delta_i = 0$ we can write $\mathbf{p}(S) - \mathbf{q}(S) = \frac{1}{2}Z$, where $Z := \sum_{i=1}^k \delta_i \xi_i$ for $\xi_1, \dots, \xi_k$ i.i.d. Rademacher (uniform on $\{\pm 1\}$). By linearity of expectation, it is immediate to check that $\mathbb{E}[Z] = 0$ (although we will not use this), and that by independence

$$\mathbb{E}\left[Z^2\right] = \sum_{1 \le i,j \le k} \delta_i \delta_j \mathbb{E}[\xi_i \xi_j] = \sum_{i=1}^k \delta_i^2 = \|\delta\|_2^2 ,$$

so what we want to prove can be rewritten $\Pr\left[Z^2 \ge \frac{1}{2}\mathbb{E}\left[Z^2\right]\right] \ge 1/48$. That is, we want an *anticoncentration* result,[12] which suggests one of the most versatile tools for this, the Paley–Zygmund inequality:

[12] Indeed, we aim to show that, with constant probability, Z stays away from its expectation $\mathbb{E}[Z] = 0$ – *i.e.*, that it does not concentrate too much around $\mathbb{E}[Z]$.

Theorem 2.11 (Paley–Zygmund inequality)**.** Let U be a non-negative random variable with finite variance. Then, for every $\theta \in [0,1]$,

$$\Pr[U \geq \theta \mathbb{E}[U]] \geq (1-\theta)^2 \frac{\mathbb{E}[U]^2}{\mathbb{E}[U^2]}\,.$$

We will apply this to Z^2, which is indeed non-negative. Unfortunately, this means we need to compute $\mathbb{E}\left[Z^4\right]$, in order to compare it to $\mathbb{E}\left[Z^2\right]^2$. We could do so directly, by expanding $\left(\sum_{i=1}^k \delta_i \xi_i\right)^4$, keeping track of the various products appearing and using linearity of expectation. This is quite fastidious; instead, we will take a shortcut and bound the moment-generating function (MGF) of Z. For any $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda Z}\right] = \prod_{i=1}^k \mathbb{E}\left[e^{\lambda \delta_i \xi_i}\right] \leq \prod_{i=1}^k e^{\frac{\lambda^2}{2}\delta_i^2} = e^{\frac{\lambda^2}{2}\|\delta\|_2^2}$$

relying on, *e.g.*, Hoeffding's Lemma for the inequality. Using the Taylor expansion of e^x along with the fact that Z is symmetric (so all its odd moments cancel out), we have $\mathbb{E}\left[e^{\lambda Z}\right] = \sum_{\ell=0}^\infty \frac{\lambda^{2\ell}}{(2\ell)!}\mathbb{E}\left[Z^{2\ell}\right] \geq \frac{\lambda^4}{4!}\mathbb{E}\left[Z^4\right]$, and so, for any $\lambda \neq 0$,

$$\mathbb{E}\left[Z^4\right] \leq \frac{24}{\lambda^4} e^{\frac{\lambda^2}{2}\|\delta\|_2^2} = \frac{3}{2}\|\delta\|_2^4 \cdot \left(\frac{4}{\|\delta\|_2^2 \lambda^2} \cdot e^{\frac{\lambda^2}{4}\|\delta\|_2^2}\right)^2 .$$

In particular, by studying the function $x > 0 \mapsto e^x/x$ we see that the RHS is minimized for $\lambda = 2/\|\delta\|_2$, showing that $\mathbb{E}\left[Z^4\right] \leq \frac{3e^2}{2}\|\delta\|_2^4$. We can finally apply Theorem 2.11, getting that, for every $\theta \in [0,1]$,

$$\Pr\left[Z^2 \geq \theta \mathbb{E}\left[Z^2\right]\right] \geq (1-\theta)^2 \frac{\mathbb{E}\left[Z^2\right]^2}{\mathbb{E}\left[Z^4\right]} \geq (1-\theta)^2 \frac{2}{3e^2} \geq \frac{(1-\theta)^2}{12}\,.$$

Choosing $\theta = 1/2$ then concludes the proof. □

Remark 2.4. The proof of this lemma established a little more than stated: first, we actually showed a tradeoff, where for every $\alpha \in [0, 1/2]$

$$\Pr_S[\,|\mathbf{p}(S) - \mathbf{q}(S)| \geq \alpha \|\mathbf{p} - \mathbf{q}\|_2\,] \geq \frac{1}{12}(1-4\alpha^2)^2\,.$$

Second, although our proof via the MGF used full independence of the ξ_i's (and thus a truly uniformly random set S), this was only to obtain

bounds on the first 4 moments of Z, which is all that the Paley–Zygmund lemma eventually requires. Any distribution for S with the same first 4 moments will then lead to the same guarantees! In particular, instead of a truly uniformly random S, one can instead use a 4-wise independent hash function, which requires significantly less randomness and can be much easier.

Finally, we conclude this section by mentioning a generalization of the binary hashing lemma (Lemma 2.9) to an arbitrary number of parts:

Theorem 2.12 (Domain Compression Lemma)**.** There exist absolute constants $c_1, c_2 > 0$ such that the following holds. For any $2 \le \ell \le k$ and any $\mathbf{p}, \mathbf{q} \in \Delta_k$,

$$\Pr_{\Pi}\left[\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_\Pi, \mathbf{q}_\Pi) \ge c_1\sqrt{\frac{\ell}{k}}\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})\right] \ge c_2\,,$$

where $\Pi = (\Pi_1, \dots \Pi_\ell)$ is a uniformly random partition of $[k]$ in ℓ subsets, and $\mathbf{p}_\Pi \in \Delta_\ell$ denotes the probability distribution on $[\ell]$ induced by $\mathbf{p}$ and Π via $\mathbf{p}_\Pi(i) = \mathbf{p}(\Pi_i)$.

This theorem, which we will not prove, can be seen as some type of "one-sided" dimensionality reduction for probability distributions, and lets us trade domain size for distances. In some settings, this can lead to better sample complexities, *e.g.*, by starting with some testing algorithm and optimizing its sample complexity $n(\ell, \varepsilon\sqrt{\ell/k}, 1/3)$ as a function of ℓ: we will get back to this in Section 4.3.

2.1.7 Bipartite collisions

Recall that in the collision-based tester from Section 2.1.2, we took a multiset S of n samples from $\mathbf{p}$ and defined our statistic Z_1 as the (normalized) number of "collisions" in S. That is fine, but requires to keep in memory all the samples observed so far. One related idea would be to instead take *two* multisets S_1, S_2 of size n_1 and n_2, and only count "bipartite collisions" – i.e., collisions between a sample of S_1 and one of

S_2:

$$Z_6 = \frac{1}{n_1 n_2} \sum_{(x,y)\in S_1\times S_2} \mathbb{1}\{x = y\} \tag{2.36}$$

One can check that $\mathbb{E}[Z_6] = \|\mathbf{p}\|_2^2$: we are back to using ℓ_2 as a proxy! Compared to the "vanilla" collision-based test, this is more flexible (S_1, S_2 need not be of the same size), and thus lends itself to some settings where a tradeoff between n_1 and n_2 is desirable: we will see that one needs $n_1 n_2 \gtrsim k/\varepsilon^4$ and $\min(n_1, n_2) \gtrsim 1/\varepsilon^2$, and the resulting sample complexity is $n = n_1 + n_2$. For the case $n_1 = n_2$, this retrieves the optimal $n \asymp \sqrt{k}/\varepsilon^2$.

Algorithm 7 Bipartite Collision-Based Tester

Require: Multisets S_1, S_2 of n_1 and n_2 samples $x_1, \dots, x_{n_1} \in \mathcal{X}$, $y_1, \dots, y_{n_2} \in \mathcal{X}$, parameters $\varepsilon \in (0, 1]$ and $k = |\mathcal{X}|$

1: Set $\tau \leftarrow \frac{1+\frac{1}{2}\varepsilon^2}{k}$

2: Compute ▷ Can be done in $O(n)$ time if $\mathcal{X}$ is known, $O(n \log n)$ if only k is.

$$Z_6 = \frac{1}{n_1 n_2} \sum_{s=1}^{n_1} \sum_{t=1}^{n_2} \mathbb{1}\{x_s = y_t\}\,.$$

3: **if** $Z_6 \geq \tau$ **then return** 0 ▷ Not uniform

4: **else return** 1 ▷ Uniform

To bound the variance of Z_6, we start by expanding the square Z_6^2 and breaking the resulting double sum in 4 cases to get

$$\begin{aligned}
{n_1}^2 {n_2}^2 \mathbb{E}\left[Z_6^2\right] &= \sum_{(x,y)\in S_1\times S_2} \sum_{(x',y')\in S_1\times S_2} \mathbb{E}\left[\mathbb{1}\{x = y\}\mathbb{1}\{x' = y'\}\right] \\
&= \sum_{x\in S_1} \sum_{y\in S_2} \mathbb{E}\left[\mathbb{1}\{x = y\}\right] \\
&\quad + \sum_{x\in S_1} \sum_{y\neq y'\in S_2} \mathbb{E}\left[\mathbb{1}\{x = y = y'\}\right] \\
&\quad + \sum_{x\neq x'\in S_1} \sum_{y\in S_2} \mathbb{E}\left[\mathbb{1}\{x = x' = y\}\right] \\
&\quad + \sum_{x\neq x'\in S_1} \sum_{y\neq y'\in S_2} \mathbb{E}\left[\mathbb{1}\{x = y\}\right]\mathbb{E}\left[\mathbb{1}\{x' = y'\}\right] \\
&= n_1 n_2 \|\mathbf{p}\|_2^2 + n_1 n_2 (n_1 + n_2 - 2)\|\mathbf{p}\|_3^3 + n_1 n_2 (n_1 - 1)(n_2 - 1)\|\mathbf{p}\|_2^4\,.
\end{aligned}$$

From $\mathrm{Var}[Z_6] = \mathbb{E}[Z_6^2] - \mathbb{E}[Z_6]^2$, we obtain

$$\begin{aligned}\mathrm{Var}[Z_6] &= \frac{1}{n_1 n_2}\|\mathbf{p}\|_2^2 + \frac{n_1+n_2-2}{n_1 n_2}\|\mathbf{p}\|_3^3 - \frac{n_1+n_2-1}{n_1 n_2}\|\mathbf{p}\|_2^4 \\ &\leq \frac{1}{n_1 n_2}\|\mathbf{p}\|_2^2 + \frac{n_1+n_2}{n_1 n_2}(\|\mathbf{p}\|_3^3 - \|\mathbf{p}\|_2^4)\,, \end{aligned} \tag{2.37}$$

using as in the proof of Theorem 2.1 the fact that $\|\mathbf{p}\|_3^3 \geq \|\mathbf{p}\|_2^4$ to obtain the slightly nicer-looking upper bound of Eq. (2.37). Note that when $n_1 = n_2 = \frac{n}{2}$, we retrieve the bound on $\mathrm{Var}[Z_1]$ from Eq. (2.9)! Specifically, we have exactly the same expression as in Eq. (2.9), but with the factor $4/n$ replaced by $\frac{1}{n_1 n_2}$ and $4/n^2$ replaced by $\frac{n_1+n_2}{n_1 n_2}$. This is good – we could follows the exact same analysis as in the proof of Theorem 2.1 to obtain, in the uniform case,

E: Exercise 2.6.

$$\Pr_{\mathbf{u}_k}\left[Z_6 \geq \frac{1+2\varepsilon^2}{k} \right] \leq \frac{k}{4\varepsilon^4 n_1 n_2}\,,$$

and, in the "far" case,

$$\Pr_{\mathbf{p}}\left[Z_6 < \frac{1+2\varepsilon^2}{k} \right] \leq \frac{5k}{4\varepsilon^4 n_1 n_2} + \frac{n_1+n_2}{n_1 n_2}\left(\frac{2\sqrt{k}}{\varepsilon} + \frac{3}{\varepsilon^2} \right). \tag{2.38}$$

At first glance, this looks perfect: our tester will be correct in both cases as long as $n_1 n_2 \gtrsim k/\varepsilon^4$, which is what we want, and $\frac{n_1 n_2}{n_1+n_2} \gtrsim \max\left(\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2}\right)$. However, that last condition is an issue: if we wanted to set $n_1 \gg n_2$, for instance, then we would still need

$$n_2 \gtrsim \max\left(\frac{\sqrt{k}}{\varepsilon}, \frac{1}{\varepsilon^2} \right),$$

which, given that the "vanilla" collision-based tester corresponded to $n_1 = n_2 \asymp \frac{\sqrt{k}}{\varepsilon^2}$, is not much of a tradeoff between n_1 and n_2 at all...

So, where did we go wrong? Recall that in order to handle the "far" case,[13] in the proof of Theorem 2.1 we had set $\alpha^2 := k\|\mathbf{p} - \mathbf{u}_k\|_2^2 \geq 4\varepsilon^2$, and bounded

$$\|\mathbf{p}\|_3^3 - \|\mathbf{p}\|_2^4 \leq \frac{\alpha^3}{k^{3/2}} + \frac{3\alpha^2}{k^2}$$

[13] We only have to worry about the far case, as the uniform case is already good – the issue arises in the second term of the variance, which is zero under the uniform distribution.

While this was enough for the "vanilla" collision-based tester, for the bipartite one this is too loose: specifically, we are losing too much after Eq. (2.11), when bounding $\|\mathbf{p}-\mathbf{u}\|_3^3$ by $\|\mathbf{p}-\mathbf{u}\|_2^3$. We could, instead of monotonicity of ℓ_p norms, write $\|\mathbf{p}-\mathbf{u}\|_3^3 \le \|\mathbf{p}-\mathbf{u}\|_\infty \|\mathbf{p}-\mathbf{u}\|_2^2$ to get

$$\|\mathbf{p}\|_3^3 - \|\mathbf{p}\|_2^4 \le \|\mathbf{p}-\mathbf{u}\|_\infty \|\mathbf{p}-\mathbf{u}\|_2^2 + \frac{3}{k}\|\mathbf{p}-\mathbf{u}\|_2^2 \le 2\|\mathbf{p}\|_\infty \frac{\alpha^2}{k} + \frac{3\alpha^2}{k^2}$$

where the second inequality uses the triangle inequality and $\|\mathbf{p}\|_\infty \ge 1/k$ to bound $\|\mathbf{p}-\mathbf{u}\|_\infty$. Is that better? This is not immediately clear, since $\|\mathbf{p}\|_\infty$ could be much larger than $\alpha/\sqrt{k}$. Yet, *if* we were lucky enough to have $\|\mathbf{p}\|_\infty \lesssim 1/k$, then this bound would be perfect! For instance, if we had $\|\mathbf{p}\|_\infty \le 2/k$, then we *could* write

$$\mathrm{Var}_{\mathbf{p}}[Z_6] \overset{?}{\le} \frac{1}{n_1 n_2} \cdot \frac{1+\alpha^2}{k} + \frac{n_1+n_2}{n_1 n_2} \cdot \frac{7\alpha^2}{k^2}\,. \tag{2.39}$$

If only...

As mentioned above, we clearly do *not* have that for every $\mathbf{p}$ such that $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) \ge \varepsilon$. Still, we saw in Section 2.1.5 an argument, based on stochastic dominance, which effectively allowed to assume this was the case, losing only a factor 2 in the distance parameter ε. This was done by considering, for any given $\mathbf{p}$, the "averaging" $\bar{\mathbf{p}}$ defined in Lemma 2.7.

Can we do the same here? Unfortunately, *no*. This stochastic dominance argument relied on our statistic Z_4 being a symmetric and convex function of the sample counts $N_1, \dots, N_k$. This is clearly not the case here: letting $N_1, \dots, N_k$ and $M_1, \dots, M_k$ be the counts of each domain element in the samples from S and T, respectively, we can rewrite

$$Z_6 = \frac{1}{n_1 n_2} \sum_{i=1}^{k} N_i M_i \tag{2.40}$$

which is neither symmetric nor convex in the N_i, M_i's, and so we cannot our stochastic dominance hammer (Theorem 2.6). Sure, but maybe we could still use some stochastic mallet? After all, we have $Z_6 = S(N_1 M_1, N_2 M_2, \dots, N_k M_k)$ for a function $S\colon \mathbb{R}^k \to \mathbb{R}$ which is symmetric and linear (and so *a fortiori* convex) in its arguments! Unfortunately... *still no*. As we will verify in Exercise 2.7, stochastic dominance here fails in a spectacular way. So we are left with the variance

bound which follows from what we had established, but features an $\|\mathbf{p}\|_\infty$ where we would like $O(1/k)$.

$$\mathrm{Var}_{\mathbf{p}}[Z_6] \le \frac{1}{n_1 n_2} \cdot \frac{1+\alpha^2}{k} + \frac{n_1+n_2}{n_1 n_2} \cdot \frac{5\|\mathbf{p}\|_\infty \alpha^2}{k}. \tag{2.41}$$

Nonetheless we can enforce *some* bound on $\|\mathbf{p}\|_\infty$, by using an extra number of samples n_3 to detect if something looks amiss. This is exactly what Algorithm 8 will allow us to do, ensuring that $\|\mathbf{p}\|_\infty \lesssim 1/n_3$. For technical reasons, this will require $n_3 \le k^{2/3}$; as we will see in Exercise 2.9 we could improve this mild restriction to any $n_3 \le k^{(s-1)/s}$, for any constant $s \ge 3$, at the cost of worse constants.

Algorithm 8 ℓ_∞ Tester via 3-way collision

Require: Multiset S of n_3 samples $x_1, \dots, x_{n_3} \in \mathcal{X}$ and $k = |\mathcal{X}|$

1: Check if any value $i \in [k]$ appears at least 3 times in S
2: **if** this happens and $n_3 \le k^{2/3}$ **then return** 0 ▷ Not uniform
3: **else return** 1 ▷ Uniform

Lemma 2.13. Given n_3 i.i.d. samples from a distribution $\mathbf{p} \in \Delta_k$, Algorithm 8 distinguishes with probability at least 5/6 between (i) $\mathbf{p} = \mathbf{u}_k$ (*i.e.*, $\|\mathbf{p}\|_\infty = 1/k$) and (ii) $\|\mathbf{p}\|_\infty > 10/n_3$, provided that $n_3 \le k^{2/3}$.

Proof. Suppose that $\mathbf{p} = \mathbf{u}_k$. For any $s \ge 2$, the probability $p(n_3, k, s)$ to observe an s-way collision among n_3 i.i.d. samples from the uniform distribution on k elements is bounded as

$$p(n_3, k, s) \le \frac{1}{k^{s-1}} \binom{n_3}{s} \tag{2.42}$$

(see, *e.g.*, Suzuki *et al.* (2006, Theorem 2)), which for $s = 3$ gives the bound $p(n_3, k, 3) \le \frac{{n_3}^3}{6k^2}$, which is at most 1/6 for $n_3 \le k^{2/3}$.

Now, assume $\|\mathbf{p}\|_\infty > 10/n_3$, and consider any element $i \in [k]$ such that $\mathbf{p}(i) = \|\mathbf{p}\|_\infty$. The number of times N_i this element i appears among the n_3 samples follows a Binomial distribution with parameters n_3 and $\|\mathbf{p}\|_\infty$ (and thus mean $n_3\|\mathbf{p}\|_\infty > 10$), and so by a Chernoff bound (Theorem A.6) we have $\Pr[N_i < 3] < 1/6$. □

We will prove a more general theorem first; before showing how to instantiate it to take advantage of Lemma 2.13:

Theorem 2.14. The bipartite collision-based tester (Algorithm 7) is a testing algorithm for uniformity with sample complexity $n(k, \varepsilon, 1/6) = n_1 + n_2$ and time complexity $O(n)$, provided that $n_1 n_2 \geq 96k/\varepsilon^4$ and $\min(n_1, n_2) \geq 480 \min(k\|\mathbf{p}\|_\infty, \sqrt{2k})/\varepsilon^2$.

Proof. Fix any $\mathbf{p} \in \Delta_k$. We also have our variance bounds from Eqs. (2.37) and (2.41). Define our threshold

$$\tau := \frac{1 + \frac{1}{2}\varepsilon^2}{k}$$

as in Algorithm 7.

- In the uniform case, $\mathbb{E}_{\mathbf{u}_k}[Z_6] = 1/k$, and the probability to output 0 (and thus make a mistake) is bounded as

$$\Pr[Z_6 \geq \tau] = \Pr\left[Z_6 \geq (1 + \tfrac{\varepsilon^2}{2})\mathbb{E}_{\mathbf{u}_k}[Z_6]\right] \leq \frac{4 \operatorname{Var}_{\mathbf{u}_k}[Z_6]}{\varepsilon^4 \mathbb{E}_{\mathbf{u}_k}[Z_6]^2} \leq \frac{4k}{n_1 n_2 \varepsilon^4}$$

  using Eq. (2.37); this is less than 1/6 as long as $n_1 n_2 \geq 24k/\varepsilon^4$.

- In the "far" case, define α by $\mathbb{E}_{\tilde{\mathbf{p}}}[Z_6] = \frac{1+\alpha^2}{k}$, so that $\alpha^2 \geq \varepsilon^2$ but also $\|\mathbf{p}\|_\infty \leq \|\mathbf{p}\|_2 = \sqrt{(1+\alpha^2)/k}$. The probability to err by outputting 1 is

$$\begin{aligned}
\Pr_{\mathbf{p}}[Z_6 < \tau] &\leq \Pr_{\tilde{\mathbf{p}}}[Z_6 < \tau] = \Pr_{\tilde{\mathbf{p}}}\left[Z_6 < \frac{1 + \frac{1}{2}\varepsilon^2}{1+\alpha^2}\mathbb{E}_{\tilde{\mathbf{p}}}[Z_6]\right] \\
&\leq \Pr_{\tilde{\mathbf{p}}}\left[Z_6 < \left(1 - \frac{\alpha^2}{2(1+\alpha^2)}\right)\mathbb{E}_{\tilde{\mathbf{p}}}[Z_6]\right] \\
&\leq \frac{4(1+\alpha^2)^2}{\alpha^4} \cdot \frac{\operatorname{Var}_{\mathbf{p}}[Z_6]}{\mathbb{E}_{\tilde{\mathbf{p}}}[Z_6]^2} \\
&\leq \frac{k}{n_1 n_2} \cdot \frac{4(1+\alpha^2)}{\alpha^4} + \frac{n_1 + n_2}{n_1 n_2} \cdot \frac{20k\|\mathbf{p}\|_\infty}{\alpha^2} \\
&\leq \frac{8k}{n_1 n_2 \varepsilon^4} + \frac{40}{\min(n_1, n_2)} \cdot \min\left(\frac{k\|\mathbf{p}\|_\infty}{\alpha^2}, \frac{\sqrt{k}\sqrt{1+\alpha^2}}{\alpha^2}\right) \\
&\leq \frac{8k}{n_1 n_2 \varepsilon^4} + \frac{40}{\min(n_1, n_2)\varepsilon^2} \min\left(k\|\mathbf{p}\|_\infty, \sqrt{2k}\right)
\end{aligned}$$

using Eq. (2.41), then $n_1 + n_2 \leq 2\max(n_1, n_2)$, then $\alpha^2 \geq \varepsilon^2$ and that the function $x \mapsto \frac{1}{x^2}\sqrt{1+x^2}$ is decreasing (and $\varepsilon \leq 1$). This is less than $1/12 + 1/12 = 1/6$ as long as (1) $n_1 n_2 \geq 96k/\varepsilon^4$ and (2) $\min(n_1, n_2) \geq 480 \min\left(k\|\mathbf{p}\|_\infty, \sqrt{2k}\right)/\varepsilon^2$.

The above analysis shows that both errors are less than $1/6$ for any $n = n_1 + n_2$ satisfying the two conditions from the statement; this concludes the proof. □

By choosing $n_1 = n_2 = \frac{n}{2}$, we first obtain:

Corollary 2.15. The bipartite collision-based tester (Algorithm 7) is a testing algorithm for uniformity with sample complexity $n(k, \varepsilon, 1/6) = O\left(\sqrt{k}/\varepsilon^2\right)$ and time complexity $O(n)$.

Perhaps more interesting is combining Corollary 2.16 and Lemma 2.13, which leads to the next result:

Corollary 2.16. The bipartite collision-based tester (Algorithm 7) combined with the ℓ_∞ testing algorithm (Algorithm 8) form a testing algorithm for uniformity with sample complexity $n(k, \varepsilon, 1/3) = n_1 + n_2$ and time complexity $O(n)$, provided that $n_1 n_2 \geq 4800k/\varepsilon^4$ and $\max(n_1, n_2) \leq k^{2/3}$.

Proof. Without loss of generality, suppose that $n_2 \leq n_1 \leq k^{2/3}$. The overall algorithm runs Algorithm 8 on S_1, and records the output as $b_1 \in \{0, 1\}$; and in parallel runs Algorithm 7 on the multisets S_1 and S_2, and records the output as $b_2 \in \{0, 1\}$. The output of the algorithm is then $b_1 b_2$, *i.e.*, the AND of the two decisions: it declares "uniform" only if both sub-algorithms declared it.

- In the uniform case, by a union bound both algorithms return 1 with probability at least $1 - (1/6 + 1/6) = 2/3$, as desired.
- In the "far" case, we have two possibilities. If $\|\mathbf{p}\|_\infty > 10/n_3$, then Algorithm 8 will return $b_1 = 0$ with probability at least $5/6$, and then regardless of the output of Algorithm 7 we have $b_1 b_2 = 0$ and will return 0. If $\|\mathbf{p}\|_\infty \leq 10/n_3$, however, then we have $n_1 n_2 \geq 96k/\varepsilon^4$ and $\min(n_1, n_2) \geq 480k\|\mathbf{p}\|_\infty/\varepsilon^2$, and

so Algorithm 7 will return $b_2 = 0$ with probability at least $5/6$. Either way, we return 0 with probability at least $5/6 > 2/3$.

This concludes the proof. □

2.1.8 Empirical subset weighting

To conclude, we will look at a somewhat different type of algorithms, in that this one can inherently be seen as *adaptive*: it works in two stages, where the second depends on what the outcome of the first stage was. As in the previous section, it also allows for some tradeoff, dividing the n total samples into two sets of size n_1 and n_2.

Fix an integer $1 \leq n_1 \leq n$. Take n_1 i.i.d. samples from $\mathbf{p}$, and consider the *set* $S \subseteq \mathcal{X}$ (not multiset) induced by those n_1 samples. The quantity of interest will be $\mathbf{p}(S)$, the (unknown) probability weight of that random set S:

$$\mathbf{p}(S) = \sum_{j=1}^{k} \mathbf{p}(j) \mathbb{1}\{N_j \geq 1\} \tag{2.43}$$

where, as before, N_j is the number of occurrences of j among the n_1 samples. One can check the expectation of this random variable is

$$\mathbb{E}[\mathbf{p}(S)] = \sum_{i=1}^{k} \mathbf{p}(i)(1 - (1 - \mathbf{p}(i))^{n_1}) \tag{2.44}$$

which should be roughly (making a few not necessarily warranted approximations) $\mathbb{E}[\mathbf{p}(S)] \approx n_1 \|\mathbf{p}\|_2^2$. Under the uniform distribution, this is exactly $(1 - (1 - 1/k)^{n_1}) \approx n_1/k$, where the approximation is valid for $n_1 \ll k$.

Great: we have a new estimator for (more or less) the ℓ_2 norm. Now, assuming things went well, at the end of this first stage we have a set S such that $\mathbf{p}(S)$ is approximately either n_1/k or $n_1\|\mathbf{p}\|_2^2 \geq n_1(1+\Omega(\varepsilon^2))/k$ (we just argued that this is what happens *in expectation*).[14] So, let's do a second stage! Take the next $n_2 := n - n_1$ samples, and count the

[14] The tricky part, of course, will be to argue that $\mathbf{p}(S)$ does concentrate enough around its expectation for this to also happen with constant probability.

number of them which fall in S: this is allows you to estimate $\mathbf{p}(S)$ up to an additive $n_1\varepsilon^2/k$, as long as

$$n_2 \gg \frac{k}{n_1\varepsilon^4}$$

(see Fact 2.2 below). This lets us retrieve the same condition $n_1 n_2 \gg k/\varepsilon^4$ as in the previous section; and, similarly, for $n_1 = n_2 = n/2$ this leads to the optimal $n \asymp \sqrt{k}/\varepsilon^2$! Only drawback: we need $n_1 \ll k$ for our approximations to be valid – indeed, after that, $\mathbb{E}[\mathbf{p}(S)]$ cannot be approximately $n_1\|\mathbf{p}\|_2^2$ anymore: the former is at most one, the latter at least n_1/k. This is, at its core, the same issue as with the "unique elements" algorithm from Section 2.1.3, and this imposes the condition $\varepsilon \gg 1/k^{1/4}$.

Algorithm 9 Empirical Subset Weighting Tester

Require: Multisets S_1, S_2 of n_1 and n_2 samples $x_1, \dots, x_{n_1} \in \mathcal{X}$, $y_1, \dots, y_{n_2} \in \mathcal{X}$, parameters $\varepsilon \in (0,1]$ and $k = |\mathcal{X}|$

1: Set $\tau \leftarrow \frac{n_1\varepsilon^2}{64k}$

2: Compute S, the *set* of samples from S_1. ▷ Can be done in $O(n_1 \log n_1)$ time, and $O(n_1)$ expected (*e.g.*,, via cuckoo hashing).

3: Compute ▷ Can be done in $O(n_2)$ time.

$$Z_7 = \frac{1}{n_2}\sum_{t=1}^{n_2} \mathbb{1}\{y_t \in S\}\,.$$

4: **if** $Z_7 \geq 1 - (1-\frac{1}{k})^{n_1} + 2\tau$ **then return** 0 ▷ Not uniform

5: **else return** 1 ▷ Uniform

The key part of the analysis is in analyzing the random variable $\mathbf{p}(S)$ from the first stage of the algorithm; the second stage, whose goal is to *estimate* $\mathbf{p}(S)$, is much simpler to handle. To do so, as usual by now, we begin by bounding the expectation gap

$$\Delta(\mathbf{p}) := \mathbb{E}_{\mathbf{p}}[\mathbf{p}(S)] - \mathbb{E}_{\mathbf{u}_k}[\mathbf{u}_k(S)] \tag{2.45}$$

when $\mathbf{p}$ is far from uniform. From Eq. (2.44), we can explicitly write,

for every $\mathbf{p} \in \Delta_k$,

$$\Delta(\mathbf{p}) = \sum_{i=1}^{k} \mathbf{p}(i)(1-(1-\mathbf{p}(i))^{n_1}) - (1-(1-1/k)^{n_1})$$
$$= \sum_{i=1}^{k} \mathbf{p}(i)((1-1/k)^{n_1} - (1-\mathbf{p}(i))^{n_1}),$$

"hiding one" by writing $(1-(1-1/k)^{n_1}) = \sum_{i=1}^{k} \mathbf{p}(i)(1-(1-1/k)^{n_1})$. We are in luck: the resulting expression turns out to be exactly the same as in the expectation for Z_2 (Lemma 2.2), with n_1 instead of $n-1$! We can therefore reuse the analysis we had then, and immediately get:

Lemma 2.17. If $n \le k$, we have

$$\Delta(\mathbf{p}) \ge \frac{n_1}{16k} \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k)^2 .$$

Here, we will introduce (and use) another concept we have not seen yet, that of *negative association* between random variables. This is a stronger notion than negative correlation, and is very useful when dealing with random variables which are dependent "but in a way which helps us" (to prove concentration):

Definition 2.3 (Negative Association)**.** The random variables $X_1, \dots, X_n$ are said to be *negatively associated* if, for all disjoint subsets $I, J \subseteq [n]$ and functions $f\colon \mathbb{R}^{|I|} \to \mathbb{R}$, $g\colon \mathbb{R}^{|J|} \to \mathbb{R}$, we have

$$\mathbb{E}[f((X_i)_{i\in I})g((X_j)_{j\in J})] \le \mathbb{E}[f((X_i)_{i\in I})]\mathbb{E}[g((X_j)_{j\in J})]$$

whenever f, g are both non-increasing or both non-decreasing.

As it turns out, the sample counts $N_1, \dots, N_k$ are negatively associated (see, *e.g.*,, (Dubhashi and Ranjan, 1998, Section 2.2)), which we will use below.[15] We will rely on this to bound the variance explicitly,

[15] What we will use does not require the full power of negative associativity and could be obtained directly, but it is a good concept to know, so – why not?

starting with the expected square:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{p}(S)^2\right] &= \sum_{i=1}^{k}\sum_{j=1}^{k}\mathbf{p}(i)\mathbf{p}(j)\mathbb{E}[\mathbb{1}\{N_i \geq 1\}\mathbb{1}\{N_j \geq 1\}] \\
&= \sum_{i=1}^{k}\mathbf{p}(i)^2\mathbb{E}[\mathbb{1}\{N_i \geq 1\}] + 2\sum_{i<j}\mathbf{p}(i)\mathbf{p}(j)\mathbb{E}[\mathbb{1}\{N_i \geq 1\}\mathbb{1}\{N_j \geq 1\}] \\
&\leq \sum_{i=1}^{k}\mathbf{p}(i)^2\mathbb{E}[\mathbb{1}\{N_i \geq 1\}] + 2\sum_{i<j}\mathbf{p}(i)\mathbf{p}(j)\mathbb{E}[\mathbb{1}\{N_i \geq 1\}]\mathbb{E}[\mathbb{1}\{N_j \geq 1\}] \\
&= \sum_{i=1}^{k}\mathbf{p}(i)^2\Pr[N_i \geq 1] + \left(\sum_{i=1}^{k}\mathbf{p}(i)\Pr[N_i \geq 1]\right)^2 - \sum_{i=1}^{k}\mathbf{p}(i)^2\Pr[N_i \geq 1]^2 \\
&= \sum_{i=1}^{k}\mathbf{p}(i)^2\Pr[N_i \geq 1](1-\Pr[N_i \geq 1]) + \mathbb{E}[\mathbf{p}(S)]^2,
\end{aligned}$$

where the inequality follows from negative associativity, and we got the third equality by completing the sum $2\sum_{i<j}x_{i,j} = \sum_{i,j}x_{i,j} - \sum_i x_{i,i}$. That is, we just proved

$$\mathrm{Var}[\mathbf{p}(S)] \leq \sum_{i=1}^{k}\mathbf{p}(i)^2\Pr[N_i \geq 1]\Pr[N_i = 0]\,. \tag{2.46}$$

By upper bounding the last factor by 1, we then get

$$\mathrm{Var}[\mathbf{p}(S)] \leq \|\mathbf{p}\|_\infty\mathbb{E}[\mathbf{p}(S)] \leq n_1\|\mathbf{p}\|_\infty^2 \tag{2.47}$$

where for the last step we used that

$$\mathbb{E}_{\mathbf{p}}[\mathbf{p}(S)] = \sum_{i=1}^{k}\mathbf{p}(i)(1-(1-\mathbf{p}(i))^{n_1}) \leq 1-(1-\|\mathbf{p}\|_\infty)^{n_1} \leq n_1\|\mathbf{p}\|_\infty\,.$$

Of course, Eq. (2.47) requires a bound on $\|\mathbf{p}\|_\infty$, which we do *not* have for the "far" case.[16] Yet, by observing that the function $x \in [0,1] \mapsto x(1-x)^m$ is maximized at $\frac{1}{m+1}$ (where it is at most $\frac{1}{2(m+1)}$) we can get

$$\mathrm{Var}[\mathbf{p}(S)] \leq \sum_{i=1}^{k}\mathbf{p}(i)\Pr[N_i \geq 1]\cdot\mathbf{p}(i)(1-\mathbf{p}(i))^{n_1} \leq \frac{\mathbb{E}[\mathbf{p}(S)]}{2n_1}\,. \tag{2.48}$$

[16] Unfortunately, we cannot in this case use our stochastic dominance hammer, or even gavel, as it is not true in general that $\mathbf{p}(S)$ stochastically dominates $\mathbf{q}(S)$ whenever $\mathbf{p} \succeq \mathbf{q}$.

This allows us to prove that the first stage of the algorithm will (with high probability) succeed, in that the random variable $\mathbf{p}(S)$ will significantly differ under the uniform and "far" cases.

Lemma 2.18. For $\mathbf{p} \in \Delta_k$ such that $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{u}_k) \geq \varepsilon$, we have

$$\Pr_{\mathbf{u}_k}[\mathbf{u}_k(S) \geq \mathbb{E}_{\mathbf{u}_k}[\mathbf{u}_k(S)] + \tau] \leq \frac{1}{6}, \quad \Pr_{\mathbf{p}}[\mathbf{p}(S) \leq \mathbb{E}_{\mathbf{u}_k}[\mathbf{u}_k(S)] + 3\tau] \leq \frac{1}{6},$$

where $\tau := \frac{n_1\varepsilon^2}{64k}$; provided that $k \geq n_1 \geq 115\sqrt{k}/\varepsilon^2$ and $\varepsilon \geq 15/k^{1/4}$.

Proof. Let $\mathbf{p} \in \Delta_k$. As usual by now, we will invoke Chebyshev's inequality to prove the desired bounds.

- In the uniform case,

$$\Pr_{\mathbf{u}_k}\left[\mathbf{u}_k(S) \geq \mathbb{E}_{\mathbf{u}_k}[\mathbf{u}_k(S)] + \frac{n_1\varepsilon^2}{64k}\right] \leq \frac{4096k^2 \operatorname{Var}_{\mathbf{u}_k}[\mathbf{u}_k(S)]}{{n_1}^2\varepsilon^4} \leq \frac{4096}{n_1\varepsilon^4}$$

  using Eq. (2.47), which implies $\operatorname{Var}_{\mathbf{u}_k}[\mathbf{u}_k(S)] \leq \frac{n_1}{k^2}$; this is less than $1/6$ as long as[17] $n_1 \geq 24576/\varepsilon^4$.

- In the "far" case, we have by Lemma 2.17 we have $\Delta(\mathbf{p}) \geq \frac{n_1\varepsilon^2}{16k} \geq \frac{\varepsilon^2}{16}\mathbb{E}_{\mathbf{u}_k}[\mathbf{u}_k(S)]$,

$$\begin{aligned}
\Pr\left[\mathbf{p}(S) < \mathbb{E}_{\mathbf{u}_k}[\mathbf{u}_k(S)] + \frac{3n_1\varepsilon^2}{64k}\right] &\leq \Pr\left[\mathbf{p}(S) < \mathbb{E}_{\mathbf{u}_k}[\mathbf{u}_k(S)] + \frac{3}{4}\Delta(\mathbf{p})\right] \\
&= \Pr\left[\mathbf{p}(S) < \mathbb{E}_{\mathbf{p}}[\mathbf{p}(S)] - \frac{1}{4}\Delta(\mathbf{p})\right] \\
&\leq \frac{16 \operatorname{Var}[\mathbf{p}(S)]}{\Delta(\mathbf{p})^2} \\
&\leq \frac{128k}{{n_1}^2\varepsilon^2} \cdot \frac{\mathbb{E}_{\mathbf{p}}[\mathbf{p}(S)]}{\Delta(\mathbf{p})} \\
&\leq \frac{2176k}{{n_1}^2\varepsilon^4}
\end{aligned}$$

  using Eq. (2.47), then $\frac{\mathbb{E}_{\mathbf{p}}[\mathbf{p}(S)]}{\Delta(\mathbf{p})} = 1 + \frac{\mathbb{E}_{\mathbf{u}_k}[\mathbf{p}(S)]}{\Delta(\mathbf{p})} \leq 1 + \frac{16}{\varepsilon^2} \leq \frac{17}{\varepsilon^2}$. This is less than $1/6$ whenever $n_1 \geq 115\sqrt{k}/\varepsilon^2$.

[17] It is worth pointing out, in case this was not apparent, that this survey does *not* focus on optimizing the constants.

Observing that our condition $\varepsilon \geq 15/k^{1/4}$ implies $115\sqrt{k}/\varepsilon^2 \geq 24576/\varepsilon^4$ concludes the proof of the lemma. □

The second stage then boils down to using n_2 samples to distinguish between the two cases: (1) $\mathbf{p}(S) \leq \mathbb{E}_{\mathbf{u}_k}[\mathbf{u}_k(S)] + \tau$ and (2) $\mathbf{p}(S) \geq \mathbb{E}_{\mathbf{u}_k}[\mathbf{u}_k(S)] + 3\tau$, assuming one of the two holds (which will if the first stage is successful). To do so, we will require the following result, which can be seen as a "localized" variant of Fact 2.1.[18]

E: Prove this! Exercise 2.5.

Fact 2.2 (Bias of a coin, distinguishing)**.** Given i.i.d. samples from a Bernoulli with unknown parameter $\alpha \in [0,1]$, and parameters $\beta, \eta \in (0,1]$, distinguishing with probability $1-\delta$ between $\alpha \leq \beta$ and $\alpha \geq \beta(1+\eta)$ can be done with (and requires) $\Theta\left(\frac{\log(1/\delta)}{\beta\eta^2}\right)$ samples. This is achieved by the empirical estimator.

We are now ready to establish the guarantees of our algorithm.

Theorem 2.19. The empirical subset weighting tester (Algorithm 9) is a testing algorithm for uniformity with sample complexity $n(k,\varepsilon,1/3) = n_1 + n_2$ and time complexity $O(n_1 \log n_1 + n_2)$ (expected $O(n_1+n_2)$), provided that $n_1 n_2 \geq Ck/\varepsilon^4$, $115\sqrt{k}/\varepsilon^2 \leq n_1 \leq k$, and $\varepsilon \geq 15/k^{1/4}$.

Proof. Fix any $\mathbf{p} \in \Delta_k$, and define

$$\tau := \frac{n_1\varepsilon^2}{64k}$$

as in Algorithm 9 and Lemma 2.18. We assume that $n_2 \geq C \cdot \frac{k}{n_1\varepsilon^4}$, where $C > 0$ is the constant hidden in the $O(\cdot)$ of Fact 2.2, so that we can apply this lemma with $\beta := \mathbb{E}_{\mathbf{u}_k}[\mathbf{u}_k(S)] + \tau$, $\eta := \frac{2\tau}{\beta}$, and $\delta := 1/6$. Here, we implicitly used that $\mathbb{E}_{\mathbf{u}_k}[\mathbf{u}_k(S)] \leq \frac{n_1}{k}$, to get

$$\frac{1}{\beta\eta^2} = \frac{\beta}{4\tau^2} \lesssim \frac{k}{n_1\varepsilon^4}\,.$$

- In the uniform case, by Lemma 2.18 and Fact 2.2 we have that, with probability at least $(1-\frac{1}{6})^2 \geq 2/3$, $\mathbf{u}_k(S) \leq \mathbb{E}_{\mathbf{u}_k}[\mathbf{u}_k(S)] + \tau = \beta$ and our estimate Z_7 of $\mathbf{u}_k(S)$ detects it, so that we output 1.

[18] Localized, as it depends on the "location" β we focus on, while Fact 2.1 holds for all possible values – and is thus a worst-case result (over the unknown value α).

- In the "far" case, similarly by Lemma 2.18 and Fact 2.2, with probability at least $(1-\frac{1}{6})^2$, $\mathbf{p}(S) \geq \mathbb{E}_{\mathbf{u}_k}[\mathbf{u}_k(S)] + 3\tau = \beta(1+\eta)$ and our estimate Z_7 of $\mathbf{p}(S)$ detects it, so that we output 0.

The above analysis shows that both errors are less than $1/3$ for any $n = n_1 + n_2$ such that $n_1 n_2 \geq Ck/\varepsilon^4$ and $k \geq n_1 \geq \left\lceil 115\sqrt{k}/\varepsilon^2 \right\rceil$; this concludes the proof. □

Setting $n_1 = n_2$ in the above theorem, we get as an immediate corollary:

Corollary 2.20. The empirical subset weighting tester (Algorithm 9) is a testing algorithm for uniformity with sample complexity $n(k, \varepsilon, 1/3) = O(\sqrt{k}/\varepsilon^2)$ and time complexity $O(n \log n)$ (expected $O(n)$), provided that $\varepsilon \geq 15/k^{1/4}$.

2.1.9 Discussion

In this section, we have covered and analyzed seven different algorithms for uniformity testing;[19] some, such as the collision-based tester and the bipartite collision tester, or the unique element tester and the empirical-distance tester, were related, and many (but not all) relied on the use of ℓ_2 distance as a proxy for total variation distance. Before concluding this section, let us have a look at some of their differences, and specific advantages of each of them.

- The collision-based tester, besides being a very natural idea (which turns out to yield the optimal sample complexity!), is as a byproduct an estimator of the ℓ_2 norm of the distribution, *i.e.*, its collision probability. This could be useful by itself, if one is interested in this quantity for its own sake. Perhaps more importantly, the use of ℓ_2 as proxy (via Cauchy–Schwarz), combined with monotonicity of ℓ_p norms ($\ell_2 \leq \ell_1$), implies that this tester actually provides some amount of *tolerance* (robustness to model misspecification): instead of just between $\mathbf{p} = \mathbf{u}_k$ and $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) \geq \varepsilon$, it enables one

[19] There may (and surely are) other possible algorithms one could consider, of course; yet, we hope that this chapter provided a good and representative overview of the ideas, techniques, and tools one could use to analyze such algorithms.

to distinguish between $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{u}_k) \le \varepsilon/(2\sqrt{k})$ and $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{u}_k) \ge \varepsilon$. This seemingly small tolerance happens to be quite helpful in many settings. Of course, the same comment applies to the bipartite collision tester, but also to the random binary hashing tester (which also relies on, and estimates, the squared ℓ_2 distance).

- Both the unique elements tester and the empirical-distance tester have *low sensitivity*, meaning that changing the value of a single one of the n samples cannot change the value of the resulting statistic (Z_2 or Z_4) by much: roughly, $\asymp 1/n$ or $\asymp 1/k$. While the different renormalization makes it a little tricky to compare, the collision-based tester, for instance, has much worse sensitivity, as changing one sample can have a much larger effect on the statistic (since Z_1 is a quadratic function of the sample counts). Having low sensitivity, which is essentially an ℓ_1 Lipschitzness guarantee, in turn is very desirable to obtain robust algorithms (*i.e.*, robust to data corruption, adversarial or not), or, importantly, to obtain *differentially private* algorithms (Dwork *et al.*, 2006), since the amount of random noise added to ensure data privacy is directly related to this sensitivity. In particular, the unique elements tester and the empirical-distance tester each have been used to obtain optimal, or near-sample-optimal, differentially private uniformity testing algorithms (Acharya *et al.*, 2018; Aliakbarpour *et al.*, 2018).

- This low sensitivity is one of the aspects which allows the empirical-distance tester to achieve the optimal "high-probability bound," that is, the optimal dependence on the error probability δ. Indeed, while our bounds for $\delta = 1/3$, combined with the standard amplification from Lemma 1.1, yield the upper bound

$$n(k, \varepsilon, \delta) = O\left(\frac{\sqrt{k}}{\varepsilon^2} \log \frac{1}{\delta}\right) \tag{2.49}$$

on the sample complexity of uniformity testing for general $\delta \in (0, 1]$, the "right" bound is

$$n(k, \varepsilon, \delta) = \Theta\left(\frac{1}{\varepsilon^2}\left(\sqrt{k \log \frac{1}{\delta}} + \log \frac{1}{\delta}\right)\right) \tag{2.50}$$

which can be much smaller for vanishing δ, *e.g.*, $\delta = 1/2^{\Omega(k)}$. The bound Eq. (2.50) is achieved by the empirical-distance tester, but with a different analysis to prove concentration around the mean: namely, instead of a Chebyshev-based (variance) bound, one can use instead the so-called *bounded differences inequality* (McDiarmid's inequality) and its "Bernstein-type variant" (bounded variances inequality) (Dubhashi and Panconesi, 2009, Chapters 5.4 and 8) to show that the statistic Z_4 concentrates very tightly. The low sensitivity of Z_4, as it happens, plays a crucial role when applying those bounds.

- The random binary hashing testing, albeit not sample-optimal, only requires to store a *single* bit of information per sample. (Moreover, this can be generalized to ℓ bits by hashing to 2^ℓ elements instead of two, leveraging Theorem 2.12.) This can be very valuable in memory-limited or communication-limited settings (we will get back to this in Chapter 4), or in the case of *local* differential privacy (Kasiviswanathan *et al.*, 2011), where this can be used to obtain a sample-optimal locally private uniformity testing algorithm (Acharya *et al.*, 2021a) (as optimally privatizing *one* bit is much easier than a full $\log k$-bit sample).

- Finally, the χ^2 tester... just works. It also nicely generalizes to other problems (identity testing, as we will see in Section 2.2; or even, with appropriate modifications, to "closeness testing" (two-sample goodness-of-fit) or other related testing tasks. Among other enjoyable properties, its asymptotic distribution (as the number of samples $n \to \infty$) can be obtained, letting us (asymptotically) obtain confidence intervals. It also does provide, as the collision-based tester, some amount of tolerance, in the so-called χ^2 divergence; while beyond the scope of this survey, this χ^2-divergence tolerance can be used as a blackbox to obtain testing algorithms for other properties than uniformity (Acharya *et al.*, 2015).

2.2 Identity testing

Having developed techniques, insights, and mathematical muscle memory for the task of uniformity testing, we now turn to its natural generalization, *identity testing*, where we now seek to test whether the unknown distribution $\mathbf{p}$ is equal to some specific, fixed reference distribution $\mathbf{q}$ – our "model" – of which we have the full, explicit description. Fortunately, many of the ideas from the previous chapter can be extended or reused for this more general problem; as we will see in the last subsection, there even exists a way to reuse the exact same algorithms as for uniformity. But first, let us get reacquainted with a familiar algorithm, the χ^2 tester.

2.2.1 The return of χ^2

We will only make a small change to our statistic Z_3 from Section 2.1.4, to replace the uniform distribution by our reference distribution $\mathbf{q}$; the analysis itself, besides a simple "trick," will also be almost identical. We will again work here in the Poissonized sampling model, and let

$$Z = \sum_{i=1}^{k} \frac{(N_i - n\mathbf{q}(i))^2 - N_i}{n\mathbf{q}(i)}\,; \tag{2.51}$$

that is, we now have $\mathbf{q}(i)$ instead of $\mathbf{u}_k(i) = 1/k$ in the numerator and denominator, and as before N_i denotes the number of occurrences of $i \in [k]$ among our Poisson(n) samples.

This is where we face a technical hurdle, due to the fact that $\mathbf{q}(i)$ in the denominator, while known, could be arbitrarily close to 0, which would make the variance of Z blow up for no particularly good reason. Fortunately, there is an easy fix: we can just replace $\mathbf{q}$ by the mixture

$$\mathbf{q}' := \frac{1}{2}\mathbf{q} + \frac{1}{2}\mathbf{u}_k \tag{2.52}$$

which will guarantee that each element has probability at least $1/(2k)$ under $\mathbf{q}'$. Moreover, given samples from an unknown distribution $\mathbf{p}$ it is easy to generate the same number of samples from the (also unknown) distribution $\mathbf{p}' := \frac{1}{2}\mathbf{p} + \frac{1}{2}\mathbf{u}_k$. If $\mathbf{p} = \mathbf{q}$, then of course $\mathbf{p}' = \mathbf{q}'$, and if

E: We will check this in Exercise 2.12.

$\mathrm{d_{TV}}(\mathbf{p},\mathbf{q}) > \varepsilon$ then $\mathrm{d_{TV}}(\mathbf{p}',\mathbf{q}') > \varepsilon/2$, since

$$2\mathrm{d_{TV}}(\mathbf{p}',\mathbf{q}') = \sum_{i=1}^{k} |\mathbf{p}'(i) - \mathbf{q}'(i)| = \sum_{i=1}^{k} \left|\frac{1}{2}\mathbf{p}(i) - \frac{1}{2}\mathbf{q}(i)\right| = \mathrm{d_{TV}}(\mathbf{p},\mathbf{q})\,.$$

That is, this transformation only costs us a constant factor in the distance parameter. For simplicity, in what follows we will write directly $\mathbf{p},\mathbf{q}$ instead of $\mathbf{p}',\mathbf{q}'$, and assume throughout that $\min_{i\in[k]} \mathbf{q}(i) \geq 1/(2k)$.

Algorithm 10 Chi-Square Tester (for Identity)

Require: Multiset of n samples $x_1,\ldots,x_n \in \mathcal{X}$, parameters $\varepsilon \in (0,1]$ and explicit description of $\mathbf{q} \in \Delta_k$. ▷ Assumes Poissonization

1: Independently replace, with probability $1/2$, each x_i by a uniformly random element of $\mathcal{X}$. ▷ Generate samples from $\mathbf{p}' = \frac{1}{2}\mathbf{p} + \frac{1}{2}\mathbf{u}_k$ in time $O(n \log k)$.

2: Set $\tau \leftarrow \frac{1}{2}n\varepsilon^2$ ▷ This is $2n(\varepsilon/2)^2$, since the "mixture trick" above leads to a factor-2 loss in the distance parameter ε.

3: Compute

$$Z = \sum_{j\in\mathcal{X}} \frac{(N_j - n\mathbf{q}'(i))^2 - N_j}{n\mathbf{q}'(i)}$$

where $N_j \leftarrow \sum_{t=1}^{n} \mathbb{1}\{x_t = j\}$ and $\mathbf{q}'(i) := \frac{\mathbf{q}(i)+1/k}{2}$.

4: **if** $Z \geq \tau$ **then return** 0 ▷ Not $\mathbf{q}$

5: **else return** 1 ▷ Equal to $\mathbf{q}$

Since, by Poissonization, the N_i's are independent Poisson random variables with $N_i \sim \mathrm{Poisson}(n\mathbf{p}(i))$, we can once more use Claim 2.2, leading to

$$\mathbb{E}[Z] = n\sum_{i=1}^{k} \frac{(\mathbf{p}(i) - \mathbf{q}(i))^2}{\mathbf{q}(i)} = n \cdot \chi^2(\mathbf{p} \,||\, \mathbf{q}) \tag{2.53}$$

From this, we immediately get that if $\mathbf{p} = \mathbf{q}$ then $\mathbb{E}[Z] = 0$. Moreover, if $\mathrm{d_{TV}}(\mathbf{p},\mathbf{q}) > \varepsilon$, then the relation between χ^2 divergence and total variation distance (Lemma B.2) implies that $\mathbb{E}[Z] > 4n\varepsilon^2$.

To compute the variance, we use independence of the N_i's and the

second part of the same Claim 2.2 to get

$$\begin{aligned}\mathrm{Var}[Z] &= \sum_{i=1}^{k} \frac{\mathrm{Var}[(N_i - n\mathbf{q}(i))^2 - N_i]}{n^2\mathbf{q}(i)^2} \\ &\le 2\sum_{i=1}^{k} \frac{\mathbf{p}(i)^2}{\mathbf{q}(i)^2} + 4\sqrt{\sum_{i=1}^{k} \frac{\mathbf{p}(i)^2}{\mathbf{q}(i)^2}\mathbb{E}[Z]}\,, \end{aligned} \tag{2.54}$$

following the exact same steps as for Eq. (2.22) in the uniformity testing case. As before, to continue we need to bound the quantity $\sum_{i=1}^{k} \frac{\mathbf{p}(i)^2}{\mathbf{q}(i)^2}$; this is where the fact that $\min_{i\in[k]} \mathbf{q}(i) \ge 1/(2k)$ will be crucial. Indeed, the first inequality is as in the uniformity testing case, the second now relies on this fact:

$$\begin{aligned}\sum_{i=1}^{k} \frac{\mathbf{p}(i)^2}{\mathbf{q}(i)^2} &\le 2k + 2\sum_{i=1}^{k} \frac{(\mathbf{p}(i) - \mathbf{q}(i))^2}{\mathbf{q}(i)^2} \\ &\le 2k + 4k\sum_{i=1}^{k} \frac{(\mathbf{p}(i) - \mathbf{q}(i))^2}{\mathbf{q}(i)} \\ &= 2k\left(1 + 2\frac{\mathbb{E}[Z]}{n}\right). \end{aligned}$$

Compared to variance analysis in the uniformity testing case, we only lose a factor 2 in the second term. Plugging this bound in Eq. (2.54), we get

$$\mathrm{Var}[Z] \le 4k\left(1 + 2\frac{\mathbb{E}[Z]}{n}\right) + 4\sqrt{2}k^{1/2}\mathbb{E}[Z] + 8\frac{k^{1/2}}{n^{1/2}}\mathbb{E}[Z]^{3/2} \tag{2.55}$$

Up to some constant factors, this is the same as (2.23). The only other difference is that, after transformation to ensure $\min_{i\in[k]} \mathbf{q}(i) \ge 1/(2k)$, the distance has become $\varepsilon/2$; so we also lose a constant factor in the expectation gap, which by Eq. (2.53) will be at least $\Delta := nk \cdot \frac{4(\varepsilon/2)^2}{k} = n\varepsilon^2$. The rest, as they say, is Chebyshev: in the $\mathbf{p} = \mathbf{q}$ case, we will need

$$\mathrm{Var}_{\mathbf{p}}[Z] \le 4k \ll n^2\varepsilon^4 \le \Delta^2\,,$$

which holds whenever $n \gg \sqrt{k}/\varepsilon^2$. In the "far" case ($\mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon/2$), we want $\mathrm{Var}_{\mathbf{p}}[Z] \ll \Delta(\mathbf{p})^2 = \mathbb{E}_{\mathbf{p}}[Z]^2$, which by Eq. (2.55) will require

$$\max\left(k, \frac{k}{n}\mathbb{E}_{\mathbf{p}}[Z], k^{1/2}\mathbb{E}_{\mathbf{p}}[Z], \frac{k^{1/2}}{n^{1/2}}\mathbb{E}_{\mathbf{p}}[Z]^{3/2}\right) \ll \mathbb{E}_{\mathbf{p}}[Z]^2\,.$$

Using $\mathbb{E}_{\mathbf{p}}[Z] \geq n\varepsilon^2$, we see that this also holds for $n \gg \sqrt{k}/\varepsilon^2$. By an argument nearly identical to the proof of Theorem 2.5, we obtain:

Theorem 2.21. The χ^2-based tester (Algorithm 10) is a testing algorithm for identity with sample complexity $n(k, \varepsilon, 1/3) = O(\sqrt{k}/\varepsilon^2)$ and time complexity $O(n \log k + k)$ in the Poissonized setting.

This was relatively painless, and shows that the general case of identity testing is no harder than uniformity testing – up to constant factors. The key insight, here, was to use this "mixture trick" (Eq. (2.52)) to replace $\mathbf{q}$ by a slightly modified version, $\mathbf{q}'$ with some nice guarantees. (Of course, if in practice the known $\mathbf{q}$ already satisfies $\min_i \mathbf{q}(i) \geq 1/(2k)$, this step is not necessary.)

One can interpret this trick as applying a suitable randomized transformation to the samples (*i.e.*, to the input distribution), which maps the reference distribution $\mathbf{q}$ to something a little closer to uniform, while somewhat preserving the pairwise distances between distributions. That is, coming up with a pair of mappings (Φ, Ψ) with $\Phi\colon \Delta_k \to \Delta_{k'}$ and $\Psi\colon [k] \to [k']$, such that (1) $\Phi(\mathbf{q}) \approx \mathbf{u}_{k'}$, (2) $\mathrm{dist}(\Phi(\mathbf{p}), \Phi(\mathbf{q})) \approx \mathrm{dist}(\mathbf{p}, \mathbf{q})$ for all $\mathbf{p}$, and (3) $x \sim \Psi(\mathbf{p})$ whenever $x \sim \mathbf{p}$.

Here, we chose $k = k'$ and $\Phi(\mathbf{p}) = \frac{1}{2}\mathbf{p} + \frac{1}{2}\mathbf{u}_k$ (and dist being the total variation distance, for a rather liberal interpretation of $\approx$), which worked for the χ^2-based tester; in the next two subsections, we will see variants of this idea, applying more generally to any tester.

2.2.2 Reduction to (near)-uniformity testing: ℓ_2 distance

The first idea is, given how convenient ℓ_2 distance has proven itself as a proxy for total variation distance when testing uniformity, to attempt to perform *identity* testing in ℓ_2 distance as well. This will not quite work, unfortunately, as the ℓ_2 guarantee will only translate to a good ℓ_1/total variation one when the reference distribution $\mathbf{q}$ is somewhat "close to uniform" (in a specific sense: when it has small ℓ_2 norm $\|\mathbf{q}\|_2$). Which is when the type of randomized mapping we just discussed will come in handy, providing a principled way to ensure this condition holds.

A good ℓ_2 tester. Our first step is thus to get a good ℓ_2 testing algorithm; for simplicity of analysis, we will throughout this section work in the Poissonized sampling setting, and establish the following theorem:

Theorem 2.22. Given the explicit description of a probability distribution $\mathbf{q} \in \Delta_k$, parameter $\varepsilon \in (0, 1]$, and Poisson(n) i.i.d. samples from an unknown $\mathbf{p} \in \Delta_k$, Algorithm 11 runs in time $O(n \log n + k)$ and distinguishes between $\|\mathbf{p} - \mathbf{q}\|_2 \leq \varepsilon$ and $\|\mathbf{p} - \mathbf{q}\|_2 \geq 2\varepsilon$ with probability at least 2/3 as long as $n \geq 100 \max(\|\mathbf{q}\|_2/\varepsilon^2, 2/\varepsilon)$.

Algorithm 11 Robust ℓ_2 Tester (for Identity)

Require: Multiset of n samples $x_1, \ldots, x_n \in \mathcal{X}$, parameters $\varepsilon \in (0, 1]$ and explicit description of $\mathbf{q} \in \Delta_k$. ▷ Assumes Poissonization

1: Set $\tau \leftarrow 2n^2\varepsilon^2$

2: Compute

$$Z = \sum_{j \in \mathcal{X}} \Big((N_j - n\mathbf{q}(i))^2 - N_j \Big)$$

where $N_j \leftarrow \sum_{t=1}^{n} \mathbb{1}\{x_t = j\}$.

3: **if** $Z \geq \tau$ **then return** 0 ▷ Far from $\mathbf{q}$ (in ℓ_2)

4: **else return** 1 ▷ Close to $\mathbf{q}$ (in ℓ_2)

Proof. Before computing the expectation and variance of the statistic Z from Algorithm 11, which does look a lot like a simpler version of Algorithms 3 and 10 (we "just" omitted the denominators!), we will make a simple observations observations. Namely, that by the reverse triangle inequality, if $\|\mathbf{p}\|_2 \geq 2\|\mathbf{q}\|_2$ then $\|\mathbf{p} - \mathbf{q}\|_2 \geq \frac{1}{2}\|\mathbf{p}\|_2$.

Turning to the analysis of the moments of Z, we can simply recall Claim 2.2 and capitalize on all our previous hard work:

$$\mathbb{E}_{\mathbf{p}}[Z] = n^2 \|\mathbf{p} - \mathbf{q}\|_2^2 \tag{2.56}$$

and

$$\begin{aligned} \mathrm{Var}_{\mathbf{p}}[Z] &= \sum_{i=1}^{k} \Big(2n^2 \mathbf{p}(i)^2 + 4n^3 \mathbf{p}(i)(\mathbf{p}(i) - \mathbf{q}(i))^2 \Big) \\ &\leq 2n^2 \|\mathbf{p}\|_2^2 + 4n^3 \|\mathbf{p}\|_2 \|\mathbf{p} - \mathbf{q}\|_2^2 \end{aligned} \tag{2.57}$$

where the second inequality just uses $\mathbf{p}(i) \le \|\mathbf{p}\|_\infty \le \|\mathbf{p}\|_2$. We are now ready to conclude:

- If $\|\mathbf{p}-\mathbf{q}\|_2 \le \varepsilon$, then $\mathbb{E}_{\mathbf{p}}[Z] \le n^2\varepsilon^2$ and by Markov's inequality

$$\Pr\left[Z \ge 3n^2\varepsilon^2\right] \le \frac{\mathbb{E}_{\mathbf{p}}[Z]}{3n^2\varepsilon^2} \le \frac{1}{3}$$

  (we did not even need Chebyshev!).

- If $\|\mathbf{p}-\mathbf{q}\|_2 \ge 2\varepsilon$, then $\mathbb{E}_{\mathbf{p}}[Z] \ge 4n^2\varepsilon^2$ and by Chebyshev's

$$\Pr\left[Z < 3n^2\varepsilon^2\right] \le \frac{16\operatorname{Var}_{\mathbf{p}}[Z]}{\mathbb{E}_{\mathbf{p}}[Z]^2} \le \frac{32\|\mathbf{p}\|_2^2}{n^2\|\mathbf{p}-\mathbf{q}\|_2^4} + \frac{64\|\mathbf{p}\|_2}{n\|\mathbf{p}-\mathbf{q}\|_2^2}$$

  (this time we did). This is where our earlier observation will come in handy: indeed, it guarantees that $\|\mathbf{p}\|_2 \le 2\max(\|\mathbf{q}\|_2, \|\mathbf{p}-\mathbf{q}\|_2)$, and so

$$\begin{aligned}\Pr\left[Z < 3n^2\varepsilon^2\right] &\le 64\max\left(\frac{\|\mathbf{q}\|_2^2}{16n^2\varepsilon^4}, \frac{1}{4n^2\varepsilon^2}\right) + 128\max\left(\frac{\|\mathbf{q}\|_2}{4n\varepsilon^2}, \frac{1}{2n\varepsilon}\right)\\ &= \max\left(\frac{4\|\mathbf{q}\|_2^2}{n^2\varepsilon^4} + \frac{32\|\mathbf{q}\|_2}{n\varepsilon^2}, \frac{16}{n^2\varepsilon^2} + \frac{64}{n\varepsilon}\right)\end{aligned}$$

  which can be seen to be at most $1/3$ for $n \ge \max\left(100\|\mathbf{q}\|_2/\varepsilon^2, 200/\varepsilon\right)$.

This concludes the proof of the theorem. □

To see how useful Theorem 2.22 is, let us apply it to uniformity testing to rederive the optimal sample complexity. After all, as mentioned above, in the particular case $\mathbf{q} = \mathbf{u}_k$ the statistic Z of Theorem 2.22 is simply a rescaling of Z_3 from Algorithm 3, and its analysis is eerily similar (relying on Claim 2.2) – so it should not be surprising that the guarantees they provide match.

Applying Theorem 2.22 with $\|\mathbf{q}\|_2 = 1/\sqrt{k}$ and distance parameter $\varepsilon' := \varepsilon/\sqrt{k}$ (from the relation TV/ℓ_2, Eq. (2.1)), we can distinguish between $\|\mathbf{p}-\mathbf{u}_k\|_2 \le \varepsilon/\sqrt{k}$ and $\|\mathbf{p}-\mathbf{u}_k\|_2 \ge 2\varepsilon/\sqrt{k}$ with

$$n \asymp \max\left(\frac{1/\sqrt{k}}{\varepsilon'^2}, \frac{1}{\varepsilon'}\right) \asymp \frac{\sqrt{k}}{\varepsilon^2}$$

samples, as promised. This then implies a uniformity testing algorithm with a small extra guarantee in the "close" case: distinguishing between $\|\mathbf{p} - \mathbf{u}_k\|_2 \le \varepsilon/\sqrt{k}$ and $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) \ge \varepsilon$. More generally, Theorem 2.22 will allow us to obtain this type of sample complexity for identity testing *as long as the reference* $\mathbf{q}$ *has* ℓ_2 *norm* $\|\mathbf{q}\|_2 \lesssim 1/\sqrt{k}$.

From arbitrary $\mathbf{q}$ to small ℓ_2 norm. Since the theorem we just established (Theorem 2.22) is particularly well suited to reference distributions $\mathbf{q}$ which are "close to uniform" in the sense of having very small ℓ_2 norm, it would be very convenient if we could assume this "without loss of generality" of every reference $\mathbf{q}$. Of course, this is not true – but, as we will see, there is a neat transformation Φ which will essentially give us this, at nearly no cost.

Namely, define, for every $i \in [k]$, the value $k_i := \lfloor k\mathbf{q}(i) \rfloor + 1$; and, accordingly,

$$k' := \sum_{i=1}^{k} k_i \le \sum_{i=1}^{k} (k\mathbf{q}(i) + 1) = 2k\,. \tag{2.58}$$

The tranformation $\Phi_{\mathbf{q}}$ will map a distribution $\mathbf{p} \in \Delta_k$ to a distribution $\Phi_{\mathbf{q}}(\mathbf{p}) \in \Delta_{k'}$ which, for each $i \in [k]$, has $k_i \ge 1$ elements with probability $\mathbf{p}(i)/k_i$:

$$\Phi_{\mathbf{q}}(\mathbf{p})(j) = \sum_{i=1}^{k} \frac{\mathbf{p}(i)}{k_i} \mathbb{1}\{j \in S_i\}, \qquad j \in [k'] \tag{2.59}$$

where $S_i := \{1 + \sum_{\ell=1}^{i-1} k_\ell, \ldots, \sum_{\ell=1}^{i} k_\ell\}$ for all $i \in [k]$.

Example 2.1. If $k = 2$ and $\mathbf{q}(1) = 1/3, \mathbf{q}(2) = 2/3$, then $k_1 = 1$, $k_2 = 2$, $k' = 3$, and for any $\mathbf{p} \in \Delta_2$ we have $\Phi_{\mathbf{q}}(\mathbf{p}) \in \Delta_3$ with

$$\Phi_{\mathbf{q}}(\mathbf{p})(1) = \mathbf{p}(1), \quad \Phi_{\mathbf{q}}(\mathbf{p})(2) = \Phi_{\mathbf{q}}(\mathbf{p})(3) = \frac{\mathbf{p}(2)}{2}\,,$$

and $S_1 = \{1\}$, $S_2 = \{2, 3\}$.

One advantage of the transformation $\Phi_{\mathbf{q}}$ described in Eq. (2.59) is its efficiency: knowing $\mathbf{q}$, one can compute all k_i in time linear in k. Moreover, given samples from any (unknown) distribution $\mathbf{p} \in \Delta_k$, one can easily simulate samples from $\Phi_{\mathbf{q}}(\mathbf{p})$ with the following randomized mapping $\Psi_{\mathbf{q}} \colon [k] \to [k']$:

$\Psi_{\mathbf{q}}$: Given $i \in [k]$, return one of the k_i elements from S_i uniformly at random.

One can check that, if $x \sim \mathbf{p}$, then $\Psi_{\mathbf{q}}(x) \sim \Phi_{\mathbf{q}}(\mathbf{p})$. Thus, so far we have defined a pair of mappings $(\Phi_{\mathbf{q}}, \Psi_{\mathbf{q}})$, which depend on our reference distribution $\mathbf{q}$ and allow us to go from distributions and samples over $[k]$ to distributions and samples over a slightly larger domain $[k']$. What did we gain in doing so?

The next two lemmas will provide the answer, by showing that (1) the transformation $\Phi_{\mathbf{q}}$ *preserves distances between distributions*, and that (2) after applying $\Phi_{\mathbf{q}}$ to our reference distribution $\mathbf{q}$, *the "new" reference distribution* $\Phi_{\mathbf{q}}(\mathbf{q})$ *has small* ℓ_2 *norm*. Which is great, as these two properties are exactly what we need to apply the ℓ_2 tester from Theorem 2.22! Specifically, we have the following guarantees:

Lemma 2.23 (Distances are preserved)**.** For any $\mathbf{p}_1, \mathbf{p}_2 \in \Delta_k$, we have

$$\mathrm{d}_{\mathrm{TV}}(\Phi_{\mathbf{q}}(\mathbf{p}_1), \Phi_{\mathbf{q}}(\mathbf{p}_2)) = \mathrm{d}_{\mathrm{TV}}(\mathbf{p}_1, \mathbf{p}_2)\,.$$

Proof. Unrolling the definitions, since $[k'] = \biguplus_{i=1}^{k} S_i$ we have

$$\begin{aligned} 2\mathrm{d}_{\mathrm{TV}}(\Phi_{\mathbf{q}}(\mathbf{p}_1), \Phi_{\mathbf{q}}(\mathbf{p}_2)) &= \sum_{i=1}^{k} \sum_{j \in S_i} |\Phi_{\mathbf{q}}(\mathbf{p}_1)(j) - \Phi_{\mathbf{q}}(\mathbf{p}_2)(j)| \\ &= \sum_{i=1}^{k} \sum_{j \in S_i} \frac{|\mathbf{p}_1(i) - \mathbf{p}_2(i)|}{k_i} \\ &= \sum_{i=1}^{k} |\mathbf{p}_1(i) - \mathbf{p}_2(i)| \\ &= 2\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_1, \mathbf{p}_2)\,, \end{aligned}$$

where we used the fact that $k_i = |S_i|$ for all $i \in [k]$. □

Lemma 2.24 ($\Phi_{\mathbf{q}}(\mathbf{q})$ is nice)**.** We have

$$\|\Phi_{\mathbf{q}}(\mathbf{q})\|_2 \leq \frac{\sqrt{2}}{\sqrt{k'}}\,.$$

Proof. This is again a game of unrolling:

$$\|\Phi_{\mathbf{q}}(\mathbf{q})\|_2^2 = \sum_{i=1}^{k} \sum_{j \in S_i} \Phi_{\mathbf{q}}(\mathbf{q})(j)^2 = \sum_{i=1}^{k} \sum_{j \in S_i} \frac{\mathbf{q}(i)^2}{k_i^2} = \sum_{i=1}^{k} \frac{\mathbf{q}(i)^2}{k_i} \leq \sum_{i=1}^{k} \frac{\mathbf{q}(i)}{k} = \frac{1}{k}$$

where the inequality follows from $k_i = 1 + \lfloor k\mathbf{q}(i) \rfloor \geq k\mathbf{q}(i)$. To conclude, recall from Eq. (2.58) that $k' \leq 2k$. □

With these two lemmas, we can combine the transformation $(\Phi_{\mathbf{q}}, \Psi_{\mathbf{q}})$ with the ℓ_2 testing algorithm from the previous subsection to get an efficient identity testing algorithm in total variation distance:

Algorithm 12 Identity Tester via ℓ_2 Reduction

Require: Multiset of n samples $x_1, \dots, x_n \in \mathcal{X}$, parameters $\varepsilon \in (0, 1]$ and explicit description of $\mathbf{q} \in \Delta_k$. ▷ Assumes Poissonization

1: Compute the values $k_1, \dots, k_k$, k', the corresponding disjoint sets $S_1, \dots, S_k$, and the distribution $\Phi_{\mathbf{q}}(\mathbf{q})$, as in Eq. (2.59).
2: Convert samples $x_1, \dots, x_n \in \mathcal{X}$ to samples $x'_1, \dots, x'_n \in \mathcal{X}' := [k']$, where $x'_i \leftarrow \Psi_{\mathbf{q}}(x_i)$. ▷ Requires randomness: $\Psi_{\mathbf{q}}$ is randomized.
3: Set $\varepsilon' := \varepsilon/\sqrt{k'}$, and invoke the robust ℓ_2 tester (Algorithm 11) on $x'_1, \dots, x'_n$, ε', and reference $\Phi_{\mathbf{q}}(\mathbf{q})$.
4: **if** Algorithm 11 returns 0 **then return** 0 ▷ Not $\mathbf{q}$
5: **else return** 1 ▷ Equal to $\mathbf{q}$

Theorem 2.25. The ℓ_2-reduction-based tester (Algorithm 12) is a (time-efficient) testing algorithm for identity with sample complexity $n(k, \varepsilon, 1/3) = O(\sqrt{k}/\varepsilon^2)$ in the Poissonized setting.[20]

Proof. The proof follows quite readily from what we have done already: given $N \sim \text{Poisson}(n)$ i.i.d. samples from some $\mathbf{p}$, we get, after passing them through $\Psi_{\mathbf{q}}$, a set of N i.i.d. samples from $\Psi_{\mathbf{q}}(\mathbf{p})$. Of course, if $\mathbf{p} = \mathbf{q}$ then $\Phi_{\mathbf{q}}(\mathbf{p}) = \Phi_{\mathbf{q}}(\mathbf{q})$; but also, by Lemma 2.23, if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$ then $\mathrm{d}_{\mathrm{TV}}(\Phi_{\mathbf{q}}(\mathbf{p}), \Phi_{\mathbf{q}}(\mathbf{q})) > \varepsilon$, and so

$$\|\Phi_{\mathbf{q}}(\mathbf{p}) - \Phi_{\mathbf{q}}(\mathbf{q})\|_2 > 2\varepsilon/\sqrt{k'} = 2\varepsilon' .$$

By Theorem 2.22, Algorithm 11 will then distinguish between our two cases $\|\Phi_{\mathbf{q}}(\mathbf{p}) - \Phi_{\mathbf{q}}(\mathbf{q})\|_2 = 0 \leq \varepsilon'$ and $\|\Phi_{\mathbf{q}}(\mathbf{p}) - \Phi_{\mathbf{q}}(\mathbf{q})\|_2 \geq 2\varepsilon'$ with probability at least 2/3, as long as

$$n \geq 100 \max\left(\|\Phi_{\mathbf{q}}(\mathbf{q})\|_2/\varepsilon'^2, 2/\varepsilon'\right) = 100 \max\left(k'\|\Phi_{\mathbf{q}}(\mathbf{q})\|_2/\varepsilon^2, \sqrt{k'}/\varepsilon\right).$$

[20] We do not state its exact time complexity here, as this would involve annoying considerations on how to efficiently represent $\Phi_{\mathbf{q}}, \Psi_{\mathbf{q}}$, and the complexity of sampling from the randomized function $\Psi_{\mathbf{q}}$.

Since $k'\|\Phi_{\mathbf{q}}(\mathbf{q})\|_2 \le \sqrt{2k'}$ by Lemma 2.24 and $k' \le 2k$ by Eq. (2.58), the right-hand-side is at most $200\sqrt{k}/\varepsilon^2$, and so having $n \ge 200\sqrt{k}/\varepsilon^2$ suffices. □

To conclude this subsection, it is worth highlighting the connection between this ℓ_2 reduction and the χ^2-based tester from Theorem 2.21. Indeed, while we focused on the fact that the tranformation $\Phi_{\mathbf{q}}$ preserved the total variation distance between distributions (Lemma 2.23), one can also see it as *converting χ^2 divergences (to $\mathbf{q}$) to ℓ_2 distances*: $\chi^2(\mathbf{p} \,||\, \mathbf{q}) \approx k\|\Phi_{\mathbf{q}}(\mathbf{p}) - \Phi_{\mathbf{q}}(\mathbf{q})\|_2^2$. Thus, in this sense, Algorithm 10 can be seen as an "unrolled" version of Algorithm 12, where the mapping $\Phi_{\mathbf{q}}$ is explicitly expanded into Algorithm 11, converting ℓ_2 testing of $\Phi_{\mathbf{q}}(\mathbf{p}), \Phi_{\mathbf{q}}(\mathbf{q})$ into χ^2 testing of $\mathbf{p}, \mathbf{q}$.

E: See Exercise 2.14.

2.2.3 Reduction to uniformity testing

So far, we have seen in Section 2.2.1 a transformation which brought the reference distribution $\mathbf{q}$ (and, actually, *all* distributions) a little closer to uniform, while preserving the total variation distances up to a factor 2 (Eq. (2.52)); and, in Section 2.2.2, a different transformation which brought down the ℓ_2 norm of the reference $\mathbf{q}$ to "near-uniform levels" while exactly preserving the total variation distances. In this section, we will describe a third mapping, which transforms the reference $\mathbf{q}$ to the *actual* uniform distribution (on a slightly larger domain), while *nearly* preserving total variation distances.

The advantage of this particular transformation is that it provides an actual *reduction*, in the formal computer science sense, from identity testing to uniformity testing. After applying this mapping, one can use any uniformity testing algorithm (*e.g.*, any of those we saw in Section 2.1) *as a blackbox*, without any modification, to perform the testing. As an algorithm designer, or (in the case of the author of this survey) a somewhat lazy person, this is particularly satisfying.

This transformation will be done in three steps:[21] (1) a mapping

[21] Recall that $\Phi_{\mathbf{q}}$ denotes a mapping between probability distributions, while $\Psi_{\mathbf{q}}$ is the corresponding (randomized) mapping between samples: for any distribution $\mathbf{p}$, applying $\Psi_{\mathbf{q}}$ on a sample $x \sim \mathbf{p}$ yields a sample $\Psi_{\mathbf{q}}(x)$ distributed according to $\Phi_{\mathbf{q}}(\mathbf{p})$. Importantly, computing $\Psi_{\mathbf{q}}(x)$ only requires knowledge of $\mathbf{q}$, and *not* of $\mathbf{p}$.

$(\Phi_{\mathbf{q}}^{(1)}, \Psi_{\mathbf{q}}^{(1)})$ which transforms any "well-behaved" reference distribution $\mathbf{q}$ over $[k+1]$ into the uniform distribution over some larger domain $[k']$, and exactly preserves total variation distances; (2) a mapping $(\Phi_{\mathbf{q}}^{(2)}, \Psi_{\mathbf{q}}^{(2)})$, which transforms any "not-too-badly behaved" reference $\mathbf{q}$ over $[k]$ into a "well-behaved" (in the sense of (1)) distribution over $[k+1]$, roughly preserving the distances; and, finally, a simple mapping $(\Phi_{\mathbf{q}}^{(3)}, \Psi_{\mathbf{q}}^{(3)})$ which transforms an arbitrary reference distribution $\mathbf{q}$ over $[k]$ into a "not-too-badly behaved" (in the sense of (2)) distribution over the same domain, also roughly preserving the distances. Combining the three will give us the overall mappings

$$\begin{aligned} \Phi_{\mathbf{q}} &:= \Phi_{\mathbf{q}}^{(1)} \circ \Phi_{\mathbf{q}}^{(2)} \circ \Phi_{\mathbf{q}}^{(3)} \colon \Delta_k \to \Delta_k \to \Delta_{k+1} \to \Delta_{k'} \\ \Psi_{\mathbf{q}} &:= \Psi_{\mathbf{q}}^{(1)} \circ \Psi_{\mathbf{q}}^{(2)} \circ \Psi_{\mathbf{q}}^{(3)} \colon [k] \to [k] \to [k+1] \to [k'] \end{aligned} \tag{2.60}$$

such that $\Phi_{\mathbf{q}}(\mathbf{q}) = \mathbf{u}_{k'}$, and $\mathrm{d_{TV}}(\Phi_{\mathbf{q}}(\mathbf{p}_1), \Phi_{\mathbf{q}}(\mathbf{p}_2)) \approx \mathrm{d_{TV}}(\mathbf{p}_1, \mathbf{p}_2)$ for all $\mathbf{p}_1, \mathbf{p}_2 \in \Delta_k$. Moreover, we will also have $k' \asymp k$: the domain size does not increase by more than a constant factor.

From well-behaved to uniform. In Section 2.2.2, the main idea was to associate each element $i \in [k]$ with a set of $k_i = 1 + \lfloor k\mathbf{q}(i) \rfloor$ elements S_i, so that upon sampling i we would return an element of S_i uniformly at random. This reduction has almost all properties we wanted, except that it did not map $\mathbf{q}$ to the uniform distribution on $[k'] = \Pi_i S_i$: only to something with small ℓ_2 norm. The reason for this boils down to our choice of k_i, and, more specifically looking at Eq. (2.59), to the sad fact that

$$\frac{\mathbf{q}(i)}{1 + \lfloor k\mathbf{q}(i) \rfloor} \neq \frac{1}{k}$$

but instead the LHS is some quantity which depends on i. The deeper reason for this being that $k\mathbf{q}(i)$ is not in general a positive integer (why would it be?), as otherwise we would just set $k_i := k\mathbf{q}(i)$ instead, and end up with $\mathbf{q}(i)/k_i = 1/k$ for all i. *But we can dream,* and this motivates the following definition of "well-behaved" distribution $\mathbf{q}$:

Definition 2.4 (Grained distribution)**.** Given a parameter $\gamma > 0$, we say that a probability distribution $\mathbf{q} \in \Delta_k$ is γ-*grained* if every probability is a positive multiple of γ, that is, if $\mathbf{q}(i) \in \gamma\mathbb{N}$ for every $i \in [k]$.

Now, suppose our reference distribution $\mathbf{q}$ over $[k]$[22] is $(1/k')$-grained, for some suitable integer $k' \geq k$ (we will take $k' = 4k$ or so). Then, we can, for every $i \in [k]$, set

$$k_i := k'\mathbf{q}(i) \in \mathbb{N} \tag{2.61}$$

and then define accordingly the disjoint sets $S_1, \ldots, S_k$ and the mappings $\Phi_{\mathbf{q}}^{(1)}, \Psi_{\mathbf{q}}^{(1)}$ as in Section 2.2.2, with this different choice of k_i's:

$$\Phi_{\mathbf{q}}^{(1)}(\mathbf{p})(j) = \sum_{i=1}^{k} \frac{\mathbf{p}(i)}{k_i} \mathbb{1}\{j \in S_i\} = \frac{1}{k'} \sum_{i=1}^{k} \frac{\mathbf{p}(i)}{\mathbf{q}(i)} \mathbb{1}\{j \in S_i\}, \qquad j \in [k'] \tag{2.62}$$

(note that $k' = \sum_{i=1}^{k} k'\mathbf{q}(i) = \sum_{i=1}^{k} k_i$). We directly have the analogue of Lemma 2.23, with the same proof: for any $\mathbf{p}_1, \mathbf{p}_2 \in \Delta_k$,

$$\mathrm{d}_{\mathrm{TV}}\left(\Phi_{\mathbf{q}}^{(1)}(\mathbf{p}_1), \Phi_{\mathbf{q}}^{(1)}(\mathbf{p}_2)\right) = \mathrm{d}_{\mathrm{TV}}(\mathbf{p}_1, \mathbf{p}_2)\,. \tag{2.63}$$

(The interested reader can also verify that the χ^2 divergences to $\mathbf{q}$ are now exactly mapped to ℓ_2 distances, not just approximately as in Exercise 2.14). However, we now have a much stronger version of Lemma 2.26: $\Phi_{\mathbf{q}}(\mathbf{q})$ is not just "nice," it is *the* uniform distribution!

E: Check it!

Lemma 2.26 ($\Phi_{\mathbf{q}}(\mathbf{q})$ is uniform)**.** We have $\Phi_{\mathbf{q}}(\mathbf{q}) = \mathbf{u}_{k'}$ (and, in particular, $\|\Phi_{\mathbf{q}}(\mathbf{q})\|_2 = 1/\sqrt{k'}$).

This follows readily from Eq. (2.62), which implies that $\Phi_{\mathbf{q}}^{(1)}(\mathbf{q})(j) = 1/k'$ for all $j \in [k']$. This completes the first step of our reduction; in the next two, we will see how to go from an arbitrary reference distribution $\mathbf{q}$ to one which satisfies our very strong assumptions – that is, one which is $(1/k')$-grained for some reasonable parameter k'.

From not-too-badly behaved to well-behaved. Let us assume we have a reference distribution $\mathbf{q} \in \Delta_k$, and we want to "convert" it to a $(1/k')$-grained distribution. A very natural idea is to simply shave off the extra probability mass from each point, to make each probability be a multiple of $1/k'$; and then to somehow move all that remaining

[22] As discussed above, we will later apply this to a reference distribution over $[k+1]$; but for ease of notation here we keep k as the domain size parameter.

probability mass on a single new element, say $k+1$. That is, for every $i \in [k]$, we go from $\mathbf{q}(i)$ to

$$\frac{\lfloor k'\mathbf{q}(i) \rfloor}{k'}$$

and put probability $1 - \sum_{i=1}^{k} \lfloor k'\mathbf{q}(i) \rfloor / k'$ (which is also a multiple of $1/k'$) on $k+1$. Rewriting the above $\frac{\lfloor k'\mathbf{q}(i) \rfloor}{k'\mathbf{q}(i)} \cdot \mathbf{q}(i)$, this leads to defining $\Phi_{\mathbf{q}}^{(2)} \in \Delta_{k+1}$ as

$$\Phi_{\mathbf{q}}^{(2)}(\mathbf{p})(j) := \begin{cases} \frac{\lfloor k'\mathbf{q}(j) \rfloor}{k'\mathbf{q}(j)} \cdot \mathbf{p}(j) & j \in [k] \\ 1 - \sum_{i=1}^{k} \frac{\lfloor k'\mathbf{q}(i) \rfloor}{k'\mathbf{q}(i)} \mathbf{p}(i) & j = k+1 \end{cases} \tag{2.64}$$

for $\mathbf{p} \in \Delta_k$. This *almost* works: the issue is that some of the $\mathbf{q}(j)$'s could be arbitrarily small, and for those $\Phi_{\mathbf{q}}^{(2)}(\mathbf{q})(j) = 0$: this violates the requirement that a grained distribution only takes positive values (no 0), and also could lead to some other issues about distances not being well preserved: for instance, one could have some distributions $\mathbf{p}_1, \mathbf{p}_2$ such that $\mathrm{d_{TV}}\left(\Phi_{\mathbf{q}}^{(2)}(\mathbf{p}_1), \Phi_{\mathbf{q}}^{(2)}(\mathbf{p}_2)\right) = 0$, yet $\mathrm{d_{TV}}(\mathbf{p}_1, \mathbf{p}_2) = 1$!

E: Can you see why?

This will not happen, fortunately, if our reference distribution $\mathbf{q}$ is not *too* badly behaved: say, if it puts probability mass at least $1/(2k)$ on every $i \in [k]$. Then, as long as $k' \geq 4k$, we will have $k'\mathbf{q}(i) \geq 2$ for all i, and so

$$\min_{i \in [k]} \frac{\lfloor k'\mathbf{q}(i) \rfloor}{k'\mathbf{q}(i)} \geq \frac{1}{2}\,.$$

This will be enough to approximately preserve total variation distances: since then, for any two $\mathbf{p}_1, \mathbf{p}_2 \in \Delta_k$, we will have

$$\begin{aligned} \mathrm{d_{TV}}\left(\Phi_{\mathbf{q}}^{(2)}(\mathbf{p}_1), \Phi_{\mathbf{q}}^{(2)}(\mathbf{p}_2)\right) &= \frac{1}{2} \sum_{i=1}^{k+1} \left| \Phi_{\mathbf{q}}^{(2)}(\mathbf{p}_1)(i) - \Phi_{\mathbf{q}}^{(2)}(\mathbf{p}_2)(i) \right| \\ &\geq \frac{1}{2} \sum_{i=1}^{k} \left| \Phi_{\mathbf{q}}^{(2)}(\mathbf{p}_1)(i) - \Phi_{\mathbf{q}}^{(2)}(\mathbf{p}_2)(i) \right| \\ &= \frac{1}{2} \sum_{i=1}^{k} \frac{\lfloor k'\mathbf{q}(i) \rfloor}{k'\mathbf{q}(i)} |\mathbf{p}_1(i) - \mathbf{p}_2(i)| \quad \text{(Eq. (2.64))} \\ &\geq \frac{1}{4} \sum_{i=1}^{k} |\mathbf{p}_1(i) - \mathbf{p}_2(i)| \qquad \text{(good behavior)} \\ &= \frac{1}{2} \mathrm{d_{TV}}(\mathbf{p}_1, \mathbf{p}_2) \end{aligned} \tag{2.65}$$

(Of course, as usual, $\mathrm{d_{TV}}\left(\Phi_{\mathbf{q}}^{(2)}(\mathbf{p}_1), \Phi_{\mathbf{q}}^{(2)}(\mathbf{p}_2)\right) \leq \mathrm{d_{TV}}(\mathbf{p}_1, \mathbf{p}_2)$ by the data processing inequality (Fact 1.1)). So, under this "not-too-badly behaved" assumption on the reference distribution $\mathbf{q}$ and for $k' \geq 4k$, our transformation $\Phi_{\mathbf{q}}^{(2)}$ from Eq. (2.64) (1) maps $\mathbf{q}$ to an $(1/k')$-grained distribution $\Phi_{\mathbf{q}}^{(2)}(\mathbf{q})$ over $[k+1]$, and (2) preserves total variation distance between distributions up to a factor 2.

To conclude, it only remains to point out that $\Psi_{\mathbf{q}}^{(2)}$ can be easily implemented as follows:

$\Psi_{\mathbf{q}}^{(2)}$**:** Given $i \in [k]$, return i with probability $\frac{\lfloor k'\mathbf{q}(j) \rfloor}{k'\mathbf{q}(j)}$ and $k+1$ otherwise.

The last step is thus to show how to go from a completely arbitrary reference distribution $\mathbf{q}$ to a "not-too-badly behaved" one. Fortunately, this last step will not be too hard: we are almost there!

From arbitrary to not-too-badly behaved. This last mapping is the simplest: recall that its goal is to ensure that the reference distribution $\mathbf{q}$ puts at least probability $1/(2k)$ on each element of the domain, while still roughly preserving the total variation distance between distributions. But we have seen this one already: this is just the "mixture trick" of Eq. (2.52), where the mappings are actually independent of $\mathbf{q}$:

$$\Phi_{\mathbf{q}}^{(3)}(\mathbf{p}) = \frac{1}{2}\mathbf{p} + \frac{1}{2}\mathbf{u}_k$$

and $\Psi_{\mathbf{q}}^{(3)}$ is just the randomized function which returns a uniformly random value with probability $1/2$:

$\Psi_{\mathbf{q}}^{(3)}$**:** Given $i \in [k]$, return i with probability $\frac{1}{2}$ and a uniformly random element of $[k]$ otherwise.

We then have, as in Section 2.2.1, that $\mathrm{d_{TV}}\left(\Phi_{\mathbf{q}}^{(3)}(\mathbf{p}_1), \Phi_{\mathbf{q}}^{(3)}(\mathbf{p}_2)\right) \geq \frac{1}{2}\mathrm{d_{TV}}(\mathbf{p}_1, \mathbf{p}_2)$ for all $\mathbf{p}_1, \mathbf{p}_2$, and that $\min_i \Phi_{\mathbf{q}}^{(3)}(\mathbf{q})(i) \geq 1/(2k)$.

Putting it together. Combining the 3 mappings described and analyzed above for $k' := 4k$, we get the following:

Theorem 2.27. Given an arbitrary reference distribution $\mathbf{q} \in \Delta_k$, the pair $(\Phi_{\mathbf{q}}, \Psi_{\mathbf{q}})$ defined in Eq. (2.60) maps distributions and samples over

$[k]$ to distributions and samples over $[4k]$, and satisfies (1) $\Phi_{\mathbf{q}}(\mathbf{q}) = \mathbf{u}_{4k}$; (2) for every $\mathbf{p}_1, \mathbf{p}_2 \in \Delta_k$,

$$\mathrm{d}_{\mathrm{TV}}(\Phi_{\mathbf{q}}(\mathbf{p}_1), \Phi_{\mathbf{q}}(\mathbf{p}_2)) \geq \frac{1}{4}\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_1, \mathbf{p}_2)\,.$$

As a direct consequence, we get the following general theorem:

Theorem 2.28. The uniformity-reduction-based tester (Algorithm 13) is a testing algorithm for identity with sample complexity $n(k, \varepsilon, 1/3) = n_U(4k, \varepsilon/4, 1/3)$, where n_U denotes the sample complexity of the chosen uniformity testing algorithm. In particular, by choosing any optimal uniformity testing algorithm, one obtains $n(k, \varepsilon, 1/3) = O(\sqrt{k}/\varepsilon^2)$.

Algorithm 13 Identity Tester via Uniformity Reduction

Require: Multiset of n samples $x_1, \ldots, x_n \in \mathcal{X}$, parameters $\varepsilon \in (0, 1]$ and explicit description of $\mathbf{q} \in \Delta_k$. ▷ Assumes Poissonization if the uniformity testing algorithm used in Line 3 does.

1: Compute the mappings $\Psi_{\mathbf{q}}$ as in Eq. (2.60), for $k' \leftarrow 4k$
2: Convert samples $x_1, \ldots, x_n \in \mathcal{X}$ to samples $x'_1, \ldots, x'_n \in \mathcal{X}' := [k']$, where $x'_i \leftarrow \Psi_{\mathbf{q}}(x_i)$. ▷ Requires randomness: $\Psi_{\mathbf{q}}$ is randomized.
3: Invoke *any* uniformity tester over $[k']$ on $x'_1, \ldots, x'_n$, with distance parameter $\varepsilon/4$.
4: **if** the uniformity tester returns 0 **then return** 0 ▷ Not $\mathbf{q}$
5: **else return** 1 ▷ Equal to $\mathbf{q}$

To close this subsection, let us discuss some aspects of this reduction. The first is its generality: since the reduction only requires knowledge of the reference distribution $\mathbf{q}$ and defers the testing to any uniformity testing of one's choosing, one can apply it in other settings than the "standard" one where samples are fully available: for instance, in some of the constrained measurements settings discussed in Chapter 4. This means that in these cases as well, one can focus on getting a good uniformity testing algorithm, and then see it extended immediately to identity testing "for free"!

The second is its cost. The above argument transforms the identity testing question with parameters (k, ε) to uniformity testing with

parameters $(4k, \varepsilon/4)$. Since the cost of the latter scales as $\sqrt{k}/\varepsilon^2$, the "blowup" in sample complexity between uniformity and identity testing is $\sqrt{4}/(1/4)^2 = 32$. This might seem a lot! Yet, we did not make here any attempt at optimizing the parameters: one can check (see Exercise 2.15) that choosing better parameters reduces this blowup to "only" a factor ≈ 12.2.

2.2.4 Bonus: bucketing

We would be remiss to end this section without at least mentioning a simple, yet powerful technique which, too, *almost* reduces identity to uniformity testing. This technique, *bucketing*, was one of the first proposed; and while it does not quite lead to the optimal sample complexity, it *almost* gets us there. We hereafter provide only an outline of the main ideas and argument; the reader is encouraged to fill in the details.

Given our reference distribution $\mathbf{q} \in \Delta_k$ and a parameter $\varepsilon \in (0, 1]$, the starting point is to partition the domain into a logarithmic number of "buckets" $B_1 \dots, B_L \subseteq [k]$ such that $\mathbf{q}$ is roughly constant (up to a multiplicative factor) on each B_j:

$$B_j := \left\{ i \in [k] \,:\, \frac{1}{2^j} < \mathbf{q}(i) \le \frac{1}{2^{j-1}} \right\}, \qquad j \in [L] \tag{2.66}$$

where $L := \log \frac{2k}{\varepsilon}$. We also define the "leftover bucket"

$$B_0 := \left\{ i \in [k] \,:\, \mathbf{q}(i) \le \frac{1}{2^L} \right\}$$

which by our choice of L only contains elements with probability at most $\varepsilon/(2k)$ under $\mathbf{q}$, and so $\mathbf{q}(B_0) \le \varepsilon/2$. Thus, this leftover bucket will not contribute too much to the distance as long as $\mathbf{p}(B_0) \le 3\varepsilon/4$ as well, which can be checked separately with $O(1/\varepsilon)$ samples from $\mathbf{p}$ (*cf.* Fact 2.2). We can therfore effectively ignore it in the rest of the analysis.

One nice property of this bucketing is that, for every $j \in [L]$, we have

$$\frac{|B_j|}{2^j} \le \mathbf{q}(B_j) = \sum_{i \in B_j} \mathbf{q}(i) \le \frac{|B_j|}{2^{j-1}}\,. \tag{2.67}$$

and $|B_j| < 2^j$. In particular, denoting by $\mathbf{q}_j$ the conditional distribution of $\mathbf{q}$ on B_j, it is a simple matter to check that

$$\|\mathbf{q}_j\|_2^2 = \sum_{i \in B_j} \frac{\mathbf{q}(i)^2}{\mathbf{q}(B_j)^2} \leq \frac{4}{|B_j|^2}\,. \tag{2.68}$$

so, great! Each of the conditional distributions $\mathbf{q}_j$ has very small ℓ_2 norm, and we should be able to use the ℓ_2 testing algorithm of Theorem 2.22.

Moreover, for any $\mathbf{p} \in \Delta_k$, we have

$$\begin{aligned}
2\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{q}) &= \sum_{j=0}^{L} \sum_{i \in B_j} |\mathbf{p}(i) - \mathbf{q}(j)| \\
&\leq \sum_{j=0}^{L} \sum_{i \in B_j} \mathbf{q}(B_j) \left(\left| \frac{\mathbf{p}(i)}{\mathbf{q}(B_j)} - \frac{\mathbf{p}(i)}{\mathbf{p}(B_j)} \right| + \left| \frac{\mathbf{p}(i)}{\mathbf{p}(B_j)} - \frac{\mathbf{q}(i)}{\mathbf{q}(B_j)} \right| \right) \\
&= \sum_{j=0}^{L} \sum_{i \in B_j} \mathbf{p}(i) \left| \frac{\mathbf{q}(B_j)}{\mathbf{p}(B_j)} - 1 \right| + \sum_{j=0}^{L} \mathbf{q}(B_j) \sum_{i \in B_j} |\mathbf{p}_j(i) - \mathbf{q}_j(i)| \\
&= \sum_{j=0}^{L} |\mathbf{p}(B_j) - \mathbf{q}(B_j)| + 2 \sum_{j=0}^{L} \mathbf{q}(B_j) \mathrm{d}_{\mathrm{TV}}(\mathbf{p}_j, \mathbf{q}_j)\,,
\end{aligned}$$

where $\mathbf{p}_j$ is as before the conditional distribution of $\mathbf{p}$ on B_j. Letting $\bar{\Phi}_{\mathbf{q}}(\mathbf{p}), \bar{\Phi}_{\mathbf{q}}(\mathbf{q})$ denote the "flattened" distributions on $[L]$ induced by $\mathbf{p}, \mathbf{q}$, what this does is relating the total variation distance between $\mathbf{p}$ and $\mathbf{q}$ to the weighted distance between conditionals and distance between "flattenings:"

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{q}) \leq \mathrm{d}_{\mathrm{TV}}\Big(\bar{\Phi}_{\mathbf{q}}(\mathbf{p}), \bar{\Phi}_{\mathbf{q}}(\mathbf{q})\Big) + \sum_{j=0}^{L} \mathbf{q}(B_j) \mathrm{d}_{\mathrm{TV}}(\mathbf{p}_j, \mathbf{q}_j)\,. \tag{2.69}$$

What does it tell us? Define $H(\mathbf{q}) := \{\, j \in [L] :\ \mathbf{q}(B_j) \geq \varepsilon/(4L) \,\}$. On the one hand, if $\mathbf{p} = \mathbf{q}$ then $\mathrm{d}_{\mathrm{TV}}\Big(\bar{\Phi}_{\mathbf{q}}(\mathbf{p}), \bar{\Phi}_{\mathbf{q}}(\mathbf{q})\Big) = 0$, and $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_j, \mathbf{q}_j) = 0$ for all $j \in [L]$. On the other hand, if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{q}) > \varepsilon$ then

$$\frac{\varepsilon}{4} \leq \mathrm{d}_{\mathrm{TV}}\Big(\bar{\Phi}_{\mathbf{q}}(\mathbf{p}), \bar{\Phi}_{\mathbf{q}}(\mathbf{q})\Big) + \sum_{j \in H(\mathbf{q})} \mathbf{q}(B_j) \mathrm{d}_{\mathrm{TV}}(\mathbf{p}_j, \mathbf{q}_j)\,. \tag{2.70}$$

where we used that $\mathbf{q}(B_0) \leq \varepsilon/2$ and $\sum_{j \notin H(\mathbf{q})} \mathbf{q}(B_j) < L \cdot \varepsilon/(4L) = \varepsilon/4$. So if $\mathbf{p}$ is ε-far from $\mathbf{q}$, then one of the two terms of the RHS must

be at least $\varepsilon/8$; even more, in the latter case this implies that at least one of the $|H(\mathbf{q})|$ terms of the sum $\mathrm{d_{TV}}(\mathbf{p}_j, \mathbf{q}_j)$ must be at least $\varepsilon_j := \varepsilon/(8L\mathbf{q}(B_j))$! If $\mathbf{p} = \mathbf{q}$, however, all $|H(\mathbf{q})| + 1$ terms are 0.

E: Can you see why?

So this leads to the following natural testing idea, which we will refer to as the "bucketing-based tester":

- test $\mathrm{d_{TV}}\left(\bar{\Phi}_{\mathbf{q}}(\mathbf{p}), \bar{\Phi}_{\mathbf{q}}(\mathbf{q})\right) = 0$ vs. $\mathrm{d_{TV}}\left(\bar{\Phi}_{\mathbf{q}}(\mathbf{p}), \bar{\Phi}_{\mathbf{q}}(\mathbf{q})\right) > \varepsilon/8$: Now, this is itself another identity testing task, so it seems like we are back to square one! But not quite: since the domain of the distributions is only logarithmic, we can afford to do this with the baseline approach (Lemma 1.2) by *learning* the distribution $\bar{\Phi}_{\mathbf{q}}(\mathbf{p})$. This only costs us $O(L/\varepsilon^2) = O(\log(k/\varepsilon)/\varepsilon^2))$ samples.

- test, for every $j \in H(\mathbf{q})$, $\mathrm{d_{TV}}(\mathbf{p}_j, \mathbf{q}_j) = 0$ vs. $\mathrm{d_{TV}}(\mathbf{p}_j, \mathbf{q}_j) > \varepsilon_j$: another identity testing task, but now where the reference $\mathbf{q}_j$ has small ℓ_2 norm (by Eq. (2.68)). of course, to do this, we need to get enough samples from $\mathbf{p}_j$, *i.e.*, enough samples from $\mathbf{p}$ must fall in each B_j. What is "enough"? Since $\|\mathbf{q}_j\|_2 \asymp 1/\sqrt{B_j}$, using Theorem 2.22 as in the proof of Theorem 2.25 we obtain that

$$\max\left(\frac{|B_j|\|\mathbf{q}_j\|_2}{\varepsilon_j^2}, \frac{1}{\varepsilon_j}\right) \asymp \frac{\sqrt{|B_j|}}{\varepsilon_j^2} \asymp \frac{|B_j|^{5/2}L^2}{\varepsilon^2 2^{2j}}$$

  samples (up to constants) are enough (where we relied on Eq. (2.67)). Moreover, since we restrict ourselves to $j \in H(\mathbf{q})$, if $\mathbf{p} = \mathbf{q}$ we will get on expectation at least

$$n\mathbf{q}(B_j) \geq \frac{n|B_j|}{2^j}$$

  and, since we restrict ourselves to $j \in H(\mathbf{q})$ where $n\mathbf{q}(B_j) \geq n/(4L)$, when $\mathbf{p} = \mathbf{q}$ the actual number of samples observed in B_j will be at least half of this with overwhelming probability by a Chernoff bound. If we do not get that many samples for some j, then we immediately reject: this is sufficient evidence that $\mathbf{p} \neq \mathbf{q}$. So, based on the above, all we need for this to work is to have

E: Check it!

$$\frac{n|B_j|}{2^j} \gg \frac{|B_j|^{5/2}L^2}{\varepsilon^2 2^{2j}},$$

that is,

$$n \gg \frac{|B_j|^{3/2} L^2}{\varepsilon^2 2^j} .$$

Recalling first that $|B_j| < 2^j$ and then that $|B_j| \leq k$ for all $j \in [L]$, the RHS is at most $\sqrt{k}L^2/\varepsilon^2 = \tilde{O}\left(\sqrt{k}/\varepsilon^2\right)$, as desired.

Some final bookkeeping: we lose an extra factor $\log L = \log\log(k/\varepsilon)$ for a union bound over all $L+1$ tests performed, by actually running each of them with error probability $\delta := 1/(3(L+1))$. Overall, modulo to the missing details "left to the reader," we established the following:

E: That's you!

Theorem 2.29. The bucketing-based tester is a testing algorithm for identity with sample complexity $n(k, \varepsilon, 1/3) = \tilde{O}(\sqrt{k}/\varepsilon^2)$.

2.2.5 Discussion

This concludes this section on identity testing: to summarize, we saw 4 different approaches. The first, in Section 2.2.1, defines and analyzes directly a χ^2-type test statistic, which generalizes the one we had in the case of uniformity testing (Section 2.1.4). The second , in Section 2.2.2, takes a different route, by first obtaining a simple "ℓ_2 tester" (which implies, by the usual ℓ_2/ℓ_1 relation, a total variation one) which works well whenever the distribution $\mathbf{q}$ has small ℓ_2 norm; and then providing a reduction to this case, via a randomized transformation of both the input distribution $\mathbf{q}$ and the samples (from the unknown $\mathbf{p}$). The third, which we covered in Section 2.2.3, goes even further, and *only* provides a reduction via a sequence of such ranodmized transformations: showing that one can leverage any uniformity testing algorithm as a blackbox to solve the more general identity testing problem. As for the fourth, it relies on partitioning the domain into a small number of buckets, on each of which the reference distribution is not uniform – but close enough.

All four approaches have their pros and cons: the first might be more efficient in practice, while the second can generalize to other problems than identity testing (namely, any testing problem where bringing down the ℓ_2 norm of some probability distribution helps). The third, besides being intellectually satisfying, naturally extends to other settings such

as the "constrained measurements" ones discussed in Chapter 4. Finally, the fourth, albeit not optimal, not only introduces the very useful bucketing technique, but is often "good enough" to get nearly optimal results. Of course, more broadly, we may be tempted to say that the more techniques at our disposal, the better; and that four is greater than one.

2.3 Historical notes

From the computer science perspective, distribution testing begun with the influential work of Goldreich *et al.* (1998), who define it as a variant of the general paradigm of property testing. Goldreich and Ron (2000) then implicitly relied on uniformity testing (or, more specifically, uniformity testing with respect to ℓ_2 distance) to test whether the endpoints of short random walks of a graph were uniformly distributed. To do so, they introduced and analyzed the collision-based tester, although their analysis did not lead to the optimal sample complexity, but instead to a quartic dependence on ε. A systematic study of distribution testing was then initiated in Batu *et al.* (2000), focusing on closeness (two-sample) testing; identity testing was first considered in Batu *et al.* (2001). These papers, and many which followed, introduced several important and versatile algorithmic ideas, such as the *bucketing* technique from Section 2.2.4, and reductions between distribution testing questions. However, it is only with the work of Paninski (2008) that the tight bound of $\Theta\left(\sqrt{k}/\varepsilon^2\right)$ for uniformity testing was established through both an information-theoretic lower bound and a matching upper bound via the unique-elements tester we covered in Section 2.1.3.[23] Interestingly, the optimality of first uniformity tester proposed, the collision-based tester of Goldreich and Ron (2000), was only established nearly two decades later, by Diakonikolas *et al.* (2019b).

The right dependence on the error probability δ was shown to be $\sqrt{\log(1/\delta)}$ (*cf.* Eq. (2.50)), instead of the "obvious" logarithmic dependence, by Huang and Meyn (2013) (for a restricted range of parameters) and Diakonikolas *et al.* (2018) (for the general case). The

[23] With some caveats: the upper bound is restricted to the regime $\varepsilon \gtrsim 1/k^{1/4}$, and its original analysis had a flaw, discussed in p. 96.

latter established this result by analyzing the empirical-distance tester (Section 2.1.5), showing as a byproduct that it *did*, contrary to the common belief, not only work, but in fact achieve the optimal sample complexity.

The bipartite collision tester (Section 2.1.7) was proposed by Diakonikolas *et al.* (2019a) in the context of uniformity testing in the so-called streaming setting (as it allows to trade the memory required to store a small set of samples for the size of the second set of samples, which arrive one by one as a "stream"). The empirical-subset-weighing tester of Section 2.1.8 can be found in Acharya *et al.* (2022), where it was developed to establish a separation between adaptive and non-adaptive testing algorithms under some type of measurement constraints (where the algorithms only have partical access to the samples; a setting we will discuss at length in Chapter 4). Finally, the binary hashing technique of Section 2.1.6 is due to Acharya *et al.* (2020d), and the general form of the domain compression lemma (Theorem 2.12) as well as some of its generalizations can be found in Acharya *et al.* (2020d), Acharya *et al.* (2020a), and Amin *et al.* (2020).

The optimal bound for *identity* testing was then established separately by Chan *et al.* (2014) (where it is implicit, from their result on ℓ_2 testing), (Acharya *et al.*, 2015), which analyzes the χ^2-based tester of Section 2.2.1 (as a key routine to obtain a flurry of testing results, for various properties), and Valiant and Valiant (2017), which we will discuss in more detail in a moment. The ℓ_2-based reduction for identity testing covered in Section 2.2.2 is due to Diakonikolas and Kane (2016), where it is used as the main building block for a general testing framework. The identity-to-uniformity reduction detailed in Section 2.2.2 was then obtained by Goldreich (2016). We slightly departed from the original presentation of these results, in order to provide a more unified view.

Going back to Valiant and Valiant (2017), their work actually introduces and addresses a refinement of identity testing, which they term "instance-optimal identity testing" and which statisticians may be more familiar with under the name *local* minimax testing. That is, they parameterize the sample complexity of the testing problem in terms

of the distance parameter ε and a suitable functional of the reference distribution $\mathbf{q}$, instead of the domain size k. Namely, they provide upper and lower bounds (which match in many cases) of a quantity related to the 2/3 "norm" $\|\mathbf{q}\|_{2/3}$ of the reference distribution $\mathbf{q}$. Note that, when maximizing this quantity over all possible reference distributions $\mathbf{q} \in \Delta_k$, we retrieve the usual $\sqrt{k}$ dependence; yet, this may be significantly smaller, for particular choices of $\mathbf{q}$, leading to better sample complexity for identity testing to $\mathbf{q}$ in those cases. For more on this, we refer the reader to Blais *et al.* (2019) (where a discussion and alternative characterization as a function of $\mathbf{q}$ are provided) and Diakonikolas and Kane (2016), as well as the excellent survey of Balakrishnan and Wasserman (2018).

To conclude, the case of uniformity (or, more broadly, identity testing) for *continuous* densities has been considered by Ingster (see Ingster (1997), and references within), under smoothness assumptions on the unknown density. As mentioned in Chapter 1, such assumptions are necessary to obtain non-trivial bounds: as an example, for Lipschitz densities over $[0,1]$ (corresponding to the Sobolev space $W^{1,\infty}$), Ingster's results establish a tight sample complexity of $\Theta\left(1/\varepsilon^{5/2}\right)$, attained by a χ^2 test. He further showed that in the adaptive setting (as discussed on p.10), the sample complexity scales as $\Theta\left(\log\log(1/(\varepsilon(\mathbf{p}) \vee \varepsilon))/(\varepsilon(\mathbf{p}) \vee \varepsilon)^{5/2}\right)$.

Acknowledgment. Exercise 2.7 (and thus the failure of stochastic dominance for the bipartite collision tester) is due to Moritz Schauer, who provided a counterexample to stochastic dominance in this case (Schauer, 2021).

2.4 Exercises

Exercise 2.1. Prove the monotonicity of ℓ_p norms: if $1 \leq r \leq s \leq \infty$, then $\|x\|_s \leq \|x\|_r$ for every $x \in \mathbb{R}^n$.

Exercise 2.2. Prove Eq. (2.14): that is, the "unique elements" statistic Z_2 from Section 2.1.3 has expectation $\mathbb{E}_{\mathbf{p}}[Z_2] = \sum_{i \in \mathcal{X}} \mathbf{p}(i)(1-\mathbf{p}(i))^{n-1}$.

Exercise 2.3. Establish Claim 2.2, using (or computing) the expression for the first 4 moments of a Poisson(λ) random variable.

Exercise 2.4. Establish the upper bound part of Fact 2.1, by proving *via* an Hoeffding or Chernoff bound that the empirical estimator achieves the stated sample complexity. (The lower bound can be shown by considering the case $\alpha = 1/2$, but we have not seen in this chapter the information-theoretic tools to establish it: this will be in Chapter 3.)

Exercise 2.5. Establish the upper bound part of Fact 2.2, by proving *via* a Chernoff bound that appropriately thresholding the empirical estimator achieves the stated sample complexity. (For the lower bound, same remark as for Exercise 2.4.)

Exercise 2.6. Follow the analysis of Theorem 2.1 to derive, for the bipartite collisions tester, the guarantee Eq. (2.38) from the variance bound Eq. (2.37).

Exercise 2.7. Show that, in contrast to what we did in the empirical-distance tester case (Section 2.1.5), one cannot invoke stochastic dominance in the analysis of the bipartite collision tester to obtain the wishful variance bound Eq. (2.39) instead of Eq. (2.41). Specifically, show that it fails even for $k = 2$: if $M \sim \mathrm{Bin}(n_1, p)$, $N \sim \mathrm{Bin}(n_2, p)$ and $M' \sim \mathrm{Bin}(n_1, q)$, $N' \sim \mathrm{Bin}(n_2, q)$ (all independent) with $1/2 \leq q < p \leq 1$, it is *not* always true that

$$MN + (n_1 - M)(n_2 - N) \succeq M'N' + (n_1 - M')(n_2 - N')$$

Hint: consider the case $n_1 = 1$*, and* $\Pr[MN + (n_1 - M)(n_2 - N) \geq 1]$ *as a function of* p*.*

Exercise 2.8. It is known that $x \preceq y$ if, and only if, $x = Ay$ for some doubly stochastic matrix A (Arnold, 1987, Theorem 2.1). Check that the averaging from Lemma 2.7 indeed corresponds to multiplying the pmf $\mathbf{p}$ (seen as a vector) by such a matrix.

Exercise 2.9 ($\star$). Generalize Lemma 2.13 to relax the condition $n_3 \leq k^{2/3}$ to $n_3 \leq k^{(s-1)/s}$, for any fixed (constant) integer $s \geq 3$, by considering s-collisions instead of 3-collisions in Algorithm 8. How does the ℓ_∞ guarantee bound in (ii) change with s?

Exercise 2.10 ($\star$). Recall that our χ^2-based statistic (Eq. (2.19)) was analyzed under the Poissonized sampling model, which led us to define

it with a $-N_i$ term in the numerator. We will show that this term is necessary: that is, under the Poissonization assumption, consider the "simpler" statistic

$$Z_3' := \sum_{i=1}^{k} \frac{(N_i - n/k)^2}{n/k}\,.$$

Show that its expectation is $nk\|\mathbf{p} - \mathbf{u}_k\|_2^2 + k$ (so the expectation *gap* remains the same), but that the variance now contains an extra term $\frac{k^2}{n}$. What sample complexity does this yield?

Exercise 2.11 ($\star$). Combine the doubling search technique discussed in Section 1.1 with the sample complexity of uniformity testing given in Eq. (2.50) to prove the following. There is an adaptive uniformity testing algorithm which, on input k and $\varepsilon \in (0, 1]$, and access to samples from an unknown distribution $\mathbf{p} \in \Delta_k$:

- correctly distinguishes between (1) $\mathbf{p} = \mathbf{u}_k$ and (2) $\varepsilon(\mathbf{p}) := \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) > \varepsilon$, with probability at least 2/3;
- always takes at most

$$O\left(\frac{1}{\varepsilon^2}\left(\sqrt{k \log\log \frac{1}{\varepsilon}} + \log\log \frac{1}{\varepsilon}\right)\right)$$

  samples; but also
- if $\varepsilon(\mathbf{p}) > \varepsilon$, takes at most

$$O\left(\frac{1}{\varepsilon(\mathbf{p})^2}\left(\sqrt{k \log\log \frac{1}{\varepsilon(\mathbf{p})}} + \log\log \frac{1}{\varepsilon(\mathbf{p})}\right)\right)$$

  samples, with probability at least 2/3; and, finally,
- show that this constant-probability bound on the number of samples also holds *in expectation.*

That is, in the "far" case this algorithm never does much worse (up to a $\log\log$ factor) than an ideal algorithm provided with the exact value $\varepsilon(\mathbf{p})$ and asked to distinguish between $\mathbf{p} = \mathbf{u}_k$ and $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) = \varepsilon(\mathbf{p})$.

Exercise 2.12. Given two probability distributions $\mathbf{p}, \mathbf{q}$, an integer $n \geq 1$, and a parameter $\alpha \in [0,1]$, consider the following two sampling processes:

- Sample $N \sim \text{Poisson}(n)$, and draw N i.i.d. samples from the mixture $(1-\alpha)\mathbf{p} + \alpha\mathbf{q}$.
- Sample $N \sim \text{Poisson}(n)$, and draw N i.i.d. samples from $\mathbf{p}$. Then, for each $1 \leq i \leq N$, independently sample $B_i \sim \text{Bern}(\alpha)$: if $B_i = 1$, replace the i-th sample by a new (and independent from everything else) sample drawn from $\mathbf{q}$.

Show that these two processes result in the same distribution.

Exercise 2.13 ($\star$). Establish the analogue of Theorem 2.22 for the *two-distribution* case (when both $\mathbf{p}, \mathbf{q}$ are unknown, and you are given n i.i.d. samples from each). Specifically, consider the statistic $Z' = \sum_{i=1}^{k}((X_i - Y_i)^2 - X_i - Y_i)$ for which you will have to establish the following counterpart of Claim 2.2:

Claim 2.3. If $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ are independent, then $\mathbb{E}\left[(X-Y)^2 - X - Y\right] = (\lambda-\mu)^2$ and $\mathbb{E}\left[((X-Y)^2 - X - Y)^2\right] = (\lambda-\mu)^4 + 2(\lambda+\mu)^2 + 4(\lambda+\mu)(\lambda-\mu)^2$.

Show that the sample complexity is $O(\max(\|\mathbf{p}\|_2, \|\mathbf{q}\|_2)/\varepsilon^2)$. Try to establish the (incomparable) bound $O(\min(\|\mathbf{p}\|_2, \|\mathbf{q}\|_2)/\varepsilon^2 + 1/\varepsilon)$.

Exercise 2.14. Show that the transformation Φ from Section 2.2.2 (Eq. (2.59)) "maps χ^2 divergence to ℓ_2 distance" in the following, approximate way: for any $\mathbf{p}, \mathbf{q} \in \Delta_k$,

$$\|\Phi_{\mathbf{q}}(\mathbf{p}) - \Phi_{\mathbf{q}}(\mathbf{q})\|_2^2 = \sum_{i \in \mathcal{X}} \frac{(\mathbf{p}(i) - \mathbf{q}(i))^2}{1 + \lfloor k\mathbf{q}(i) \rfloor}.$$

Conclude that, assuming $\min_i \mathbf{q}(i) \geq 1/(2k)$ (as we could in Section 2.2.1 after using the "mixture trick" of Eq. (2.52)),

$$\frac{1}{2}\chi^2(\mathbf{p} \,||\, \mathbf{q}) \leq k\|\Phi_{\mathbf{q}}(\mathbf{p}) - \Phi_{\mathbf{q}}(\mathbf{q})\|_2^2 \leq \chi^2(\mathbf{p} \,||\, \mathbf{q})$$

for every $\mathbf{p} \in \Delta_k$.

Exercise 2.15 (⋆⋆). Generalize the transformation Φ from Section 2.2.3 in two ways: first, by replacing the mixture $\Phi^{(3)}_{\mathbf{q}}(\mathbf{p}) = \frac{1}{2}\mathbf{p} + \frac{1}{2}\mathbf{u}_k$ by $\alpha\mathbf{p}+(1-\alpha)\mathbf{u}_k$, where $\alpha \in (0,1)$. Second, by replacing the choice $k' = 4k$ in $\Phi^{(2)}(\mathbf{p})$ by $k' = \beta k$, for some integer β such that $\beta(1-\alpha) \geq 1$.

1. By tracking down the various restrictions on α, β and their use across $\Phi^{(1)}$, $\Phi^{(2)}$, and $\Phi^{(3)}$, show that doing so now maps identity testing with parameters (k, ε) to uniformity testing with parameters
$$\left(\beta k, \alpha\left(1 - \frac{1}{\beta(1-\alpha)}\right)\varepsilon\right)$$
2. Check that setting $(\alpha, \beta) = (1/2, 4)$ as in Section 2.2.3 recovers Theorem 2.28, and the blowup factor of 32 discussed at the end of the section.
3. Recalling that the sample complexity scales as $\sqrt{k}/\varepsilon^2$, optimize over (α, β) to find the optimal choice of parameters, and prove that the resulting blowup is ≈ 12.2.
4. What would be the optimal choice of (α, β), and the corresponding blowup, in a setting where the sample complexity of uniformity testing scales as k/ε^2 instead of $\sqrt{k}/\varepsilon^2$? (This is not that far-fetched: we will see in Section 4.3 an example of such a setting.)

2.5 Deferred proofs

We here provide the omitted proofs from the chapter. The first, from Section 2.1.3, is due to Nazarov (2021).

Lemma 2.30 (Lemma 2.3, restated). Fix $m \geq 1$ and $k \in \mathbb{N}$. For any $x_1, \ldots, x_k \geq 0$ such that $\sum_{i=1}^k x_i = 1$, we have
$$\frac{m \sum_{1 \leq i < j \leq k} x_i x_j \left((1 - x_i - x_j)^{m-1} - (1-x_i)^m (1-x_j)^m\right)}{\sum_{i=1}^k x_i (1 - (1-x_i)^m)} \leq 1$$

Proof. Define $(y_i)_{1 \leq i \leq k}$ by $y_i := 1 - (1-x_i)^m$, and note that Bernoulli's inequality implies that $y_i \leq m x_i$ for all i. In particular, $\sum_{i=1}^k y_i \leq m$. Further, since $x_i + x_j \leq 1$ for all $i \neq j$, we have $0 \leq 1 - x_i - x_j \leq (1-x_i)(1-x_j)$, which implies $(1-x_i-x_j)^{m-1} \leq (1-x_i)^{m-1}(1-x_j)^{m-1}$.

This lets us bound the numerator as

$$\begin{aligned} m\sum_{i<j} x_i x_j \Big((1-x_i-x_j)^{m-1} - (1-x_i)^m(1-x_j)^m\Big) \\ &= \frac{m}{2}\sum_{i\neq j} x_i x_j \Big((1-x_i-x_j)^{m-1} - (1-x_i)^m(1-x_j)^m\Big) \\ &\le \frac{m}{2}\sum_{i\neq j} x_i x_j (1-x_i)^{m-1}(1-x_j)^{m-1}(1-(1-x_i)(1-x_j)) \\ &\le \frac{m}{2}\sum_{i,j} x_i x_j (1-x_i)^{m-1}(1-x_j)^{m-1}(x_i+x_j) \end{aligned}$$

To relate this to the denominator, which can be rewritten as $\sum_{i=1}^{k} x_i y_i$, we rely on the following inequality: for every $x \in [0,1]$

$$1-(1-x)^m = m\int_0^x (1-u)^{m-1}\,\mathrm{d}m \ge mx(1-x)^{m-1}$$

and so $x_i(1-x_i)^{m-1} \le \frac{1}{m} y_i$ for all i. It follows that

$$\begin{aligned} m\sum_{i<j} x_i x_j \Big((1-x_i-x_j)^{m-1} - (1-x_i)^m(1-x_j)^m\Big) \\ &\le \frac{1}{2m}\sum_{i,j} y_i y_j (x_i+x_j) = \frac{1}{m}\sum_{i=1}^{k} x_i y_i \sum_{j=1}^{k} y_j \\ &\le \sum_{i=1}^{k} x_i y_i\,, \end{aligned}$$

concluding the proof. □

Note: The variance analysis from Paninski (2008). The proof of Theorem 2.4 presented in this survey departs from the original one from Paninski (2008). The argument analyzing the expectation gap (Lemma 2.2) is similar, although we tried to make it a little simpler and intuitive (which, admittedly, is very subjective). The main difference is in bounding the variance in the "far" case; indeed, while (Paninski, 2008) relies for this on the Efron–Stein inequality, the argument given is flawed, and the claimed variance bound does not follow.

Specifically, the proof of Paninski (2008, Lemma 2) claims the bound

$$\mathrm{Var}_{\mathbf{p}}[Z_2] \le n\sum_{i=1}^{k} p(i)(1-(1-\mathbf{p}(i))^{n-1})$$

which, if true, would imply (after renormalizing to match our notation)

$$\mathrm{Var}_{\mathbf{u}_k}[Z_2] \leq \frac{1}{n}(1-(1-1/k)^{n-1}) \underset{k\to\infty}{\sim} \frac{1}{k}$$

while the (exact) variance in the uniform case can be computed explicitly, and is asymptotically $\mathrm{Var}_{\mathbf{u}_k}[Z_2] \sim_{k\to\infty} \frac{2}{k}$. This shows that the former upper bound cannot hold as stated. The issue arises in the first step, when bounding the quantity $\frac{1}{2}\sum_{t=1}^{n}\mathbb{E}\left[(S-S^{(t)})^2\right]$ after applying the Efron–Stein inequality, as $(S-S^{(t)})^2$ can take values 0, 1,or 4 (not just 0 or 1), and some events leading to these have been overlooked. It is not clear to us how to fix their proof, which is why we chose to provide an alternative argument to bound the variance. Interestingly, the analysis from Paninski (2008, Lemma 2) *can* be used (partially) to bound the variance of a different statistic, Z_4, and we did so to derive Eq. (2.31) in Section 2.1.5.

3

Information-theoretic lower bounds

In this chapter, we will cover several techniques and "ready-to-use" theorems allowing us to easily (or, rather, not too painfully) establish sample complexity lower bounds. As a guiding example, we will show that many of the uniformity testing algorithms we saw Section 2.1 are sample-optimal, by establishing (in several ways) the $\Omega(\sqrt{k}/\varepsilon^2)$ sample complexity lower bound for uniformity testing.

3.1 Indistinguishability, Le Cam, and Ingster's method

To establish a lower bound on the sample complexity of testing a given property $\mathcal{P} = \bigcup_{k=1}^{\infty} \mathcal{P}_k$, we want to come up (for a given k) with a "simple" family of (pairs of) distributions such that, when chosing such a pair $(\mathbf{p}_1, \mathbf{p}_0)$ at random, (i) $\mathbf{p}_1$ has the property, but $\mathbf{p}_0$ is far from it, so a tester *should* distinguish between the two; but (ii) unless n is large enough, no algorithm taking n samples can actually distinguish between the two. The "simple" here is somewhat fuzzy, but can be rephrased as "a family simple enough to make proving (ii) as painless as possible."

In a slightly more formal way, the goal is to define two *priors*[1] ζ_0, ζ_1

[1] That is, two probability distributions over probability distributions.

over probability distributions, such that

- ζ_1 is supported on $\mathcal{P}_k$ (yes-instances); and
- ζ_0 is supported on $\overline{\mathcal{P}}_k^{\varepsilon} := \{ \mathbf{p} \in \Delta_k : \mathrm{d_{TV}}(\mathbf{p}, \mathcal{P}_k) > \varepsilon \}$ (no-instances);

and then, to show that it is impossible to distinguish two randomly chosen $\mathbf{p}_1 \sim \zeta_1$, $\mathbf{p}_0 \sim \zeta$ from n samples unless n is large enough as a function of k, ε, and possibly δ. One can also relax a little the above to ask that ζ_0 only be *mostly* supported on $\overline{\mathcal{P}}^{\varepsilon}$ (*i.e.*, no-instances are only far from $\mathcal{P}_k$ with high probability); and, similarly, one can possibly only require that ζ_1 only be mostly supported on $\mathcal{P}_k$. One can even allow the random choices of $\mathbf{p}_0$ and $\mathbf{p}_1$ to depend on each other (*i.e.*, coupling ζ_1 and ζ_0). For simplicity, we will not worry too much about this in this chapter.

Suppose now we came up with such a prior, and also that we chose ζ_1, ζ_0 such that every $\mathbf{p}_1, \mathbf{p}_0$ are at distance exactly ε. By Lemma 1.4, to distinguish with probability at least $1-\delta$ between two such distributions $\mathbf{p}_0$ and $\mathbf{p}_1$ from n samples, we must have

$$1 - 2\delta \le \mathrm{d_{TV}}\left(\mathbf{p}_0^{\otimes n}, \mathbf{p}_1^{\otimes n}\right). \tag{3.1}$$

To obtain a lower bound on n, we need to somehow relate this quantity to $\mathrm{d_{TV}}(\mathbf{p}_0, \mathbf{p}_1)$, which we know is ε. A first natural attempt is to use the fact that total variation distance is subadditive: $\mathrm{d_{TV}}\left(\mathbf{p}_0^{\otimes n}, \mathbf{p}_1^{\otimes n}\right) \le n \cdot \mathrm{d_{TV}}(\mathbf{p}_0, \mathbf{p}_1)$, which gives

$$1 - 2\delta \le n\varepsilon \tag{3.2}$$

leading to a sample complexity lower bound of $\Omega(1/\varepsilon)$. This is, quite frankly, underwhelming. The issue here is that "total variation distance does not tensorize," *i.e.*, that metric does not play nice when you take products of probability distributions. Alright, so maybe we can use a different metric (or "distance") between distributions which *does* tensorize, as a proxy?

Two choices come to mind: the first, Hellinger distance, or rather its square, and Kullback–Leibler divergence. Both are related to TV distance (Lemmas B.1 and B.3), and behave much better:

$$\mathrm{d_{TV}}\left(\mathbf{p}_0^{\otimes n}, \mathbf{p}_1^{\otimes n}\right)^2 \le 2\mathrm{d_H}\left(\mathbf{p}_0^{\otimes n}, \mathbf{p}_1^{\otimes n}\right)^2 \le 2n\mathrm{d_H}(\mathbf{p}_0, \mathbf{p}_1)^2 \tag{3.3}$$

and

$$\mathrm{d_{TV}}\left(\mathbf{p}_0^{\otimes n}, \mathbf{p}_1^{\otimes n}\right)^2 \leq \frac{1}{2}\mathrm{D}\left(\mathbf{p}_0^{\otimes n} \| \mathbf{p}_1^{\otimes n}\right)^2 = \frac{1}{2} n \mathrm{D}(\mathbf{p}_0 \| \mathbf{p}_1)\,, \tag{3.4}$$

respectively. If it so happens that $\mathrm{d_H}(\mathbf{p}_0, \mathbf{p}_1) \asymp \varepsilon$ or $\mathrm{D}(\mathbf{p}_0 \| \mathbf{p}_1) \asymp \varepsilon^2$ (both stronger statements than our assumption $\mathrm{d_{TV}}(\mathbf{p}_0, \mathbf{p}_1) = \varepsilon$, but both reasonable, and easy to check when coming up with $\mathbf{p}_0, \mathbf{p}_1$), then we get a sample complexity lower bound of $\Omega(1/\varepsilon^2)$. Which is much better, but still rather underwhelming: there is no dependence on k!

Why? We did not really take advantage of our "priors," really, we just fixed two $\mathbf{p}_1, \mathbf{p}_0$ and completely ignored whatever "the *random* choice of a yes- and no-instance" could bring. Instead of analyzing $\mathrm{d_{TV}}\left(\mathbf{p}_0^{\otimes n}, \mathbf{p}_1^{\otimes n}\right)$ for fixed $\mathbf{p}_0, \mathbf{p}_1$, we *could* instead bound

$$\mathbb{E}_{\zeta_0, \zeta_1}\left[\mathrm{d_{TV}}\left(\mathbf{p}_0^{\otimes n}, \mathbf{p}_1^{\otimes n}\right)\right], \tag{3.5}$$

which conceivably could lead to some improvement. That will not be enough, though: if $\mathrm{d_{TV}}\left(\mathbf{p}_0^{\otimes n}, \mathbf{p}_1^{\otimes n}\right) \asymp n\varepsilon^2$ for every random choice of $\mathbf{p}_0, \mathbf{p}_1$ – or even most of them – then taking this extra expectation will not buy us anything. That is unfortunate, but we can do better! We can put the expectation *inside* the total variation distance.

This follows from the following (simple) observation: if we have an algorithm $\mathcal{A}$ which correctly outputs 1 with probability at least $1 - \delta$ upon seeing n samples from *any* distribution $\mathbf{p}$ – that is, when given one sample from $\mathbf{p}^{\otimes n}$, then that algorithm must also be correct when given a sample from any mixture $\mu = \mathbb{E}_\zeta[\mathbf{p}^{\otimes n}] = \sum_{\mathbf{p}} \zeta(\mathbf{p})\mathbf{p}^{\otimes n}$ of those $\mathbf{p}^{\otimes n}$'s:

$$\begin{aligned}
\Pr_{x\sim\mu}[\mathcal{A}(x) = 1] &= \mathbb{E}_{x\sim\mu}[\mathbb{1}\{\mathcal{A}(x) = 1\}] = \mathbb{E}_{\mathbf{p}\sim\zeta}\left[\mathbb{E}_{x\sim\mathbf{p}^{\otimes n}}[\mathbb{1}\{\mathcal{A}(x) = 1\}]\right] \\
&= \mathbb{E}_{\mathbf{p}\sim\zeta}\left[\Pr_{x\sim\mathbf{p}^{\otimes n}}[\mathcal{A}(x) = 1]\right] \\
&\geq \mathbb{E}_{\mathbf{p}\sim\zeta}[1 - \delta] = 1 - \delta\,;
\end{aligned}$$

and, similarly, for mixtures of no-instances and probability to output 0. It is worth enphasizing that here, the mixture is over n-fold distributions $\mathbf{p}^{\otimes n}$, *not* over the distributions themselves: that is, we first pick $\mathbf{p}$ according to our prior ζ, then take all n samples from the same $\mathbf{p}$ – this

is very different from taking n samples from the mixture $\mathbb{E}_\zeta[\mathbf{p}]$, which would mean we pick a new $\mathbf{p}$ for each sample! That is,

$$\mathbb{E}_{\mathbf{p}\sim\zeta}[\mathbf{p}^{\otimes n}] \neq \mathbb{E}_{\mathbf{p}\sim\zeta}[\mathbf{p}]^{\otimes n} \tag{3.6}$$

in general, and the argument above corresponds to the quantity on the left. That is, we get

$$1 - 2\delta \leq \mathrm{d_{TV}}\left(\mathbb{E}_{\zeta_0}\left[\mathbf{p}_0^{\otimes n}\right], \mathbb{E}_{\zeta_1}\left[\mathbf{p}_1^{\otimes n}\right]\right), \tag{3.7}$$

and the right-hand side can end up being *much* smaller than the RHS of Eq. (3.1), or even than the expected distance from Eq. (3.5).[2]

Eq. (3.7) (and what leads to it) is often referred to as *Le Cam's two-point method*. Once we are there, however, we are left with the task of bounding this total variation distance between two mixtures, which seems anything but easy. In some cases, still, we *can* proceed further: an important example is when one of the two mixtures (say, ζ_1) is actually concentrated on a single distribution $\mathbf{p}_1$ (as we will see soon, this is the same for uniformity testing, for instance, where $\mathcal{P}_k$ is a singleton).

Let us further assume our family of no-instances are parameterized as $\{\mathbf{p}_\theta\}_{\theta\in\Theta}$, and rewrite our mixture of no-instances ζ_1 as a probability distribution π over Θ. Picking a no-instance $\mathbf{p}_0$ from ζ_1 becomes equivalent to picking $\theta \sim \pi$, and returning π_θ. Then, with this extra assumption and this rewriting, Eq. (3.7) becomes

$$1 - 2\delta \leq \mathrm{d_{TV}}\left(\mathbb{E}_{\theta\sim\pi}\left[\mathbf{p}_\theta^{\otimes n}\right], \mathbf{p}_1^{\otimes n}\right) \leq \frac{1}{2}\sqrt{\chi^2\left(\mathbb{E}_{\theta\sim\pi}\left[\mathbf{p}_\theta^{\otimes n}\right] \mid\mid \mathbf{p}_1^{\otimes n}\right)}, \tag{3.8}$$

where we invoked the relation between total variation distance and chi-square divergence (Lemma B.2) for the second inequality. At this point, the reason for using this inequality might appear quite mysterious: why χ^2 instead of, say, KL divergence, or Hellinger distance, or basically *anything else*? The reason lies in the following lemma, which shows how to bound the chi-square divergence between a mixture and a single, honest-to-goodness product distribution:

[2] It is worth pointing out that we always have $\mathrm{d_{TV}}\left(\mathbb{E}_{\zeta_0}\left[\mathbf{p}_0^{\otimes n}\right], \mathbb{E}_{\zeta_1}\left[\mathbf{p}_1^{\otimes n}\right]\right) \leq \mathbb{E}_{\zeta_0,\zeta_1}\left[\mathrm{d_{TV}}\left(\mathbf{p}_0^{\otimes n}, \mathbf{p}_1^{\otimes n}\right)\right]$, by joint convexity of total variation distance (which is a consequence of it being an f-divergence). The key is that, if we play our cards right when defining ζ_0, ζ_1, we will actually have $\mathrm{d_{TV}}\left(\mathbb{E}_{\zeta_0}\left[\mathbf{p}_0^{\otimes n}\right], \mathbb{E}_{\zeta_1}\left[\mathbf{p}_1^{\otimes n}\right]\right) \ll \mathbb{E}_{\zeta_0,\zeta_1}\left[\mathrm{d_{TV}}\left(\mathbf{p}_0^{\otimes n}, \mathbf{p}_1^{\otimes n}\right)\right]$.

Lemma 3.1 (Ingster's method)**.** Consider a random variable θ such that, for each realization $\theta = \vartheta$, $\mathbf{p}_\vartheta^{(n)} = \mathbf{p}_{1,\vartheta} \otimes \cdots \otimes \mathbf{p}_{n,\vartheta}$ is a product distribution. Further, let $\mathbf{p}^{(n)} = \mathbf{p}_1 \otimes \cdots \otimes \mathbf{p}_n$ be a fixed product distribution. Then,

$$\chi^2\Big(\mathbb{E}_\theta\Big[\mathbf{p}_\theta^{(n)}\Big] \mid\mid \mathbf{p}^{(n)}\Big) = \mathbb{E}_{\theta,\theta'}\left[\prod_{j=1}^{n}(1 + H_j(\theta,\theta'))\right] - 1,$$

where θ' is an independent copy of θ, and

$$H_j(\vartheta,\vartheta') := \mathbb{E}_{x\sim\mathbf{p}_j}\left[\frac{(\mathbf{p}_{j,\vartheta}(x) - \mathbf{p}_j(x))(\mathbf{p}_{j,\vartheta'}(x) - \mathbf{p}_j(x))}{\mathbf{p}_j(x)^2}\right]$$

is the "chi-square inner product" of $\mathbf{p}_{j,\vartheta}$ and $\mathbf{p}_{j,\vartheta'}$ with respect to $\mathbf{p}_j$.[3]

This statement is still slightly too general for our purposes, as in our case all marginals of any fixed $\mathbf{p}_\vartheta^{(n)}$ (and those of $\mathbf{p}^{(n)}$) are the same, since we take n i.i.d. samples. Applying Lemma 3.1 to Eq. (3.8) (after squaring it to get rid of the square root) leads to

$$4(1-2\delta)^2 \le \mathbb{E}_{\theta,\theta'\sim\pi}\big[(1 + H(\theta,\theta'))^n\big] - 1\,, \tag{3.9}$$

where

$$H(\theta,\theta') := \mathbb{E}_{x\sim\mathbf{p}_1}\left[\frac{(\mathbf{p}_\theta(x) - \mathbf{p}_1(x))(\mathbf{p}_{\theta'}(x) - \mathbf{p}_1(x))}{\mathbf{p}_1(x)^2}\right] \tag{3.10}$$

At this point, the reader is probably quietly wondering why, exactly, this is an improvement over what we had before, and where this is going. Before going through an application which, hopefully, will answer both of these questions and establish a lower bound for our running example of uniformity testing, let us try to give a heuristic argument.

Suppose that we pick our prior over no-instance well, so that $\mathbb{E}_\theta[\mathbf{p}_\theta] = \mathbf{p}_1$: that is, the average of our no-instances is the yes-instance $\mathbf{p}_1$,

[3] To explain the name, observe that

$$\mathbb{E}_{x\sim\mathbf{p}}\left[\frac{(\mathbf{q}(x) - \mathbf{p}(x))(\mathbf{q}'(x) - \mathbf{p}(x))}{\mathbf{p}(x)^2}\right] = \sum_{x\in\mathcal{X}}\frac{(\mathbf{q}(x) - \mathbf{p}(x))(\mathbf{q}'(x) - \mathbf{p}(x))}{\mathbf{p}(x)}\,,$$

and when $\mathbf{q} = \mathbf{q}'$ this is equal to $\chi^2(\mathbf{q} \mid\mid \mathbf{p})$.

which sounds like a reasonable thing to go for. If we expand the RHS of Eq. (3.9), we get

$$\begin{aligned}\mathbb{E}_{\theta,\theta'}\left[(1+H(\theta,\theta'))^n\right]-1 &= \mathbb{E}_{\theta,\theta'}\left[1+nH(\theta,\theta')+\ldots\right]-1\\ &= n\mathbb{E}_{\theta,\theta'}\left[H(\theta,\theta')\right]+(\text{high-order terms})\end{aligned}$$

Recalling the definition of $H(\theta,\theta')$ and by independence of θ,θ', we can then rewrite this first-order term as

$$\begin{aligned}n\mathbb{E}_{\theta,\theta'}\left[H(\theta,\theta')\right] &= n\mathbb{E}_{\theta,\theta'}\left[\mathbb{E}_{x\sim\mathbf{p}_1}\left[\frac{(\mathbf{p}_\theta(x)-\mathbf{p}_1(x))(\mathbf{p}_{\theta'}(x)-\mathbf{p}_1(x))}{\mathbf{p}_1(x)^2}\right]\right]\\ &= n\mathbb{E}_{x\sim\mathbf{p}_1}\left[\frac{(\mathbb{E}_\theta[\mathbf{p}_\theta(x)]-\mathbf{p}_1(x))(\mathbb{E}_{\theta'}[\mathbf{p}_{\theta'}(x)]-\mathbf{p}_1(x))}{\mathbf{p}_1(x)^2}\right]=0\end{aligned}$$

the last equality since $\mathbb{E}_\theta[\mathbf{p}_\theta(x)]=\mathbb{E}_{\theta'}[\mathbf{p}_{\theta'}(x)]=\mathbf{p}_1(x)$ (we chose our prior well!). This means that the first-order term (at least) of that expansion cancels out entirely: which is good news, as the more cancellations we get the more likely the overall thing will be small, and the better the upper bound we can hope for. That being said, it is worth illustrating this with an example.

Uniformity testing lower bound. Let us get back to our running example, uniformity testing. We saw in Chapter 2 (in various ways) that the sample complexity of uniformity testing over a known domain of size k is $O\left(\sqrt{k}/\varepsilon^2\right)$. Can we show that this is optimal (up to constant factors)?

To do so, we need to come up with a prior over yes-instances, and one over no-instances. The former is immediate: since there is literally only one yes-instance, *the* uniform distribution $\mathbf{u}_k$, our prior ζ_1 will just be concentrated on $\mathbf{u}_k$. For the latter, recalling what the "hard cases" for our uniformity testing algorithms were in Section 2.1, a good starting point would be to look at probability distributions "nearly uniform" locally, but still far overall: something where each probability is something like $(1\pm\varepsilon)/k$.

We also want as many of those distributions as possible in the support of ζ_0, as otherwise the testing problem would be too easy: if we only have N no-instances, then a testing algorithm could do N pairwise tests of the form *"is it this specific* $\mathbf{p}$*, or* $\mathbf{u}_k$*?*: since each $\mathbf{p}$ is ε-far

from $\mathbf{u}_k$, each test would cost $O(\log(1/\delta)/\varepsilon^2)$. By a union bound, doing this with error probability $\delta := 1/(3N)$ would suffice for the whole algorithm to succeed, and so the sample complexity overall would be $O((\log N)/\varepsilon^2)$. To have any chance of proving our $\sqrt{k}$ lower bound, we thus need N to be *exponential* in k.

Finally, we need each no-instance to be a *bona fide* probability distribution, so it should sum to one: the simplest way to achieve this is to pair elements of the domain together (assuming, without loss of generality, that k is even) to perturb them jointly: if one element of a pair has probability $(1+\varepsilon)/k$, the other gets $(1-\varepsilon)/k$. Keeping all this in mind, we define $2^{k/2}$ distributions, each parameterized by some $\theta \in \Theta := \{-1,+1\}^{k/2}$, as follows:

$$\mathbf{p}_\theta(i) = \frac{1 + (-1)^i \theta_{\lceil i/2 \rceil} \cdot 3\varepsilon}{k}, \qquad i \in [k] \tag{3.11}$$

that is, where $\theta_i \in \{-1,+1\}$ determines whether the perturbation for elements $2i-1, 2i$ is $1 \pm 3\varepsilon$ ("up-down") or $1 \mp 3\varepsilon$ ("down-up"). For every fixed $\theta \in \{-1,+1\}^{k/2}$, it is easy to check that the corresponding $\mathbf{p}_\theta$ satisfies

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_\theta, \mathbf{u}_k) = \frac{3}{2}\varepsilon > \varepsilon \tag{3.12}$$

since each element of the domain contributes exactly $3\varepsilon/k$ to the ℓ_1 distance; and that we have $N := 2^{k/2}$ such no-instances (so, indeed, exponentially many in k). To define our prior ζ_0, since we do not have any reason to do anything more complicated we will simply take the uniform distribution over all these $2^{k/2}$ no-instances; equivalently, in terms of parameter θ we will define our prior π as the uniform distribution over $\Theta = \{-1,+1\}^{k/2}$, so the $k/2$ values θ_i are i.i.d. Rademacher random variables.

One can then check that $\mathbb{E}_\theta[\mathbf{p}_\theta] = \mathbf{u}_k$, fitting our previous discussion. In view of bounding the RHS of Eq. (3.9), we need to first get a hold on the quantity $H(\theta,\theta')$ defined in Eq. (3.10). Plugging in our definition of $\mathbf{p}_\theta$ from Eq. (3.11) and recalling that $\mathbf{p}_1$ is simply $\mathbf{u}_k$, we get

$$H(\theta,\theta') = \frac{1}{k}\sum_{i=1}^{k}(k\mathbf{p}_\theta(i)-1)(k\mathbf{p}_{\theta'}(i)-1) = \frac{18\varepsilon^2}{k}\sum_{i=1}^{k/2}\theta_i\theta'_i\,. \tag{3.13}$$

This is going well: using $1+u \le e^u$ and the fact that the $\theta_i\theta'_i$'s are i.i.d. Rademacher random variables, we can write

$$\begin{aligned}\mathbb{E}_{\theta,\theta'}\left[(1+H(\theta,\theta'))^n\right] - 1 &\le \mathbb{E}_{\theta,\theta'}\left[e^{nH(\theta,\theta')}\right] - 1 \\ &= \mathbb{E}_{\theta,\theta'}\left[e^{\frac{18\varepsilon^2 n}{k}\sum_{i=1}^{k/2}\theta_i\theta'_i}\right] - 1 \\ &= \prod_{i=1}^{k/2}\mathbb{E}_{\theta,\theta'}\left[e^{\frac{18\varepsilon^2 n}{k}\theta_i\theta'_i}\right] - 1 \\ &\le e^{\frac{81\varepsilon^4 n^2}{k}} - 1 \end{aligned} \tag{3.14}$$

where the second inequality follows from Hoeffding's Lemma.[4]

Combining this with Eq. (3.9) and massaging the inequality, we get that n must satisfy

$$4(1-2\delta)^2 \le e^{\frac{81\varepsilon^4 n^2}{k}} - 1 \tag{3.15}$$

i.e., $\varepsilon^4 n^2/k \gtrsim 1$ (for $\delta \le 1/2$). We proved our lower bound!

Theorem 3.2. Every testing algorithm for uniformity must have sample complexity $n(k,\varepsilon,1/3) = \Omega(\sqrt{k}/\varepsilon^2)$.

What about δ? To conclude this chapter, it is worth pointing out that the above approach will *not* let us directly obtain any meaningful dependence on the error probability δ. This is painfully apparent from Eq. (3.15), as the LHS remains $\Theta(1)$ regardless of how small δ becomes. At a higher level, this comes from our use of Lemma B.2 in Eq. (3.8), which by directly bounding the TV distance (always in $[0,1]$!) by the chi-square divergence, destroys any hope to have a lower bound on the latter which grows as δ vanishes. Using Pinsker's inequality (Lemma B.3) and then bounding KL by chi-square would lead to the same issue; does this mean there is no hope in getting the dependence on δ in our lower bound?

[4] Since the random variable $\theta_i\theta'_i$ is uniform on $\{-1,+1\}$, we could have computed that expectation exactly; but this would have involved some unwieldy cosh which we would have to upper bound later on anyway.

Not quite. Instead of Lemma B.2, one could use the stronger bound given by Lemma B.4 to relate total variation distance and Kullback–Leibler divergence, and combine this with the refined upper bound between KL and χ^2 divergences given in Lemma B.5 to obtain

$$\frac{1}{4\delta} \leq 1 + \chi^2\left(\mathbb{E}_{\theta\sim\pi}\left[\mathbf{p}_\theta^{\otimes n}\right] \mid\mid \mathbf{p}_1^{\otimes n}\right). \tag{3.16}$$

(See Exercise 3.1 for the details.) As an example, combining this with the upper bound

$$\chi^2\left(\mathbb{E}_\theta\left[\mathbf{p}_\theta^{\otimes n}\right] \mid\mid \mathbf{u}_k^{\otimes n}\right) \leq e^{\frac{81\varepsilon^4 n^2}{k}} - 1$$

derived in Eq. (3.14) for uniformity testing, this readily gives the (improved, and optimal) lower bound, essentially for free.

Theorem 3.3. Every testing algorithm for uniformity must have sample complexity $n(k,\varepsilon,\delta) = \Omega(\sqrt{k\log(1/\delta)}/\varepsilon^2)$.

We did not even have to start everything from scratch. Not too bad!

3.2 Indistinguishability via Fano: a bit of mutual information

We will now cover another lower bound method, based on bounding an information-theoretic quantity, the mutual information $I(\mathbf{b}\wedge X)$ between two suitable random variables. As in the previous session, the first step is to come up with our two priors ζ_1, ζ_0 over yes- and no-instances (and, in particular, to show that either all or most of the no-instances are indeed ε-far from the property).

The second step is to look at the following process: (1) pick a uniformly random bit $\mathbf{b} \in \{0,1\}$; (2) draw a distribution $\mathbf{p}$ from $\zeta_{\mathbf{b}}$; (3) take n i.i.d. samples from $\mathbf{p}$. The key idea is that if one has a testing algorithm for our property (say, correct with probability 2/3), then one can use it to distinguish between ζ_0 and ζ_1 – but then, one can use *this* to guess the value of $\mathbf{b}$ with probability at least 2/3 based on the n samples observed. Intuitively, this must mean those n samples $X = (X_1, \ldots, X_n)$ carry some non-trivial amount of information about $\mathbf{b}$, that is, that the mutual information $I(\mathbf{b}\wedge X)$ is "large." This is made formal in the next lemma:

Fact 3.1. Let $\mathbf{b} \in \{0,1\}$ be a uniformly random bit, and X be a random variable taking values in some set S. If there exists an algorithm $\mathcal{A}$ such that $\Pr[\mathcal{A}(X) = \mathbf{b}] \geq 2/3$ (where the randomness is over $\mathbf{b}, X$, and the internal randomness of $\mathcal{A}$), then $I(\mathbf{b} \wedge X) \geq 2/25$.

Proof. We have

$$\begin{aligned} I(\mathbf{b} \wedge X) &= H(\mathbf{b}) - H(\mathbf{b} \mid X) = 1 - H(\mathbf{b} \mid X) \\ &\geq 1 - h(\Pr[\mathcal{A}(X) \neq \mathbf{b}]) \geq 1 - h(1/3) \\ &> 2/25\,, \end{aligned}$$

where the first inequality is Fano's (using that $\mathbf{b}$ takes values in $\{0,1\}$) and the second is monotonicity of the binary entropy function $h(x) = -x \log x - (1-x)\log(1-x)$. □

The third step is to upper bound $I(\mathbf{b} \wedge X)$ as a function of k, ε, and n, relying on the "nice" properties of mutual information (*e.g.*, chain rule, data-processing inequality, and whatever sticks after repeated readings of Cover and Thomas (2006, Chapter 2)). To make the best use of these nice properties, it is typically useful to make two additional simplifications:

- First, instead of exactly n samples, in stage (3) of the sampling process above we pick $N \sim \text{Poisson}(n)$ samples (where N is independent of $\mathbf{b}$ and X). This is the "Poissonized sampling" setting we saw in Chapter 2, and it will allow us to argue that the number of times N_i each element i of the domain is seen are independent random variables (conditioned on the draw of $\mathbf{p}$), which will *considerably* simplify our analysis of $I(\mathbf{b} \wedge X)$. This Poissonization trick can be done roughly without loss of generality, as discussed in Appendix C.

- Second, we will want to relax our requirement that the no-instances are *actual* probability distributions, and only ask that they be *measures* (not necessarily summing to one!), as long as they sum to something close to one with high probability: say, $1/2 \leq \|\mathbf{p}\|_1 \leq 3/2$ with probability $1 - o(1)$ (over the choice of $\mathbf{p} \sim \zeta_0$). This one appears to be with a little more loss of generality, but it turns out this can be assumed without losing much. (See Exercise 3.2.)

But again, let us go through an example to make this concrete.

Uniformity testing lower bound. Our single yes-instance is again the uniform distribution $\mathbf{u}_k$, which takes care of defining ζ_1. However, we will use a slightly different prior ζ_0 than in the previous section for our no-instances: the reason is that, for the $2^{k/2}$ no-instances defined in Eq. (3.11), the individual probabilities of elements $2i$ ans $2i-1$ were not independently chosen (this was the whole point: they compensated each other to have a total probability mass of one overall). As a result, when choosing a no-instance, we would have some annoying dependencies between N_{2i-1} and N_{2i}, even after conditioning on $\mathbf{b}$. We *could* handle these dependencies here, but this introduces some extra bean counting: the reader is encouraged to do so in Exercise 3.3, in this section we will do something slightly different.

Namely, we will use the second relaxation discussed above, and just define our prior ζ_0 as the uniform prior over 2^k distributions indexed by $\theta \in \{-1,+1\}^k$, with

$$\mathbf{p}_\theta(i) = \frac{1+\theta_i \cdot 3\varepsilon}{k}, \qquad i \in [k] \tag{3.17}$$

(*i.e.*, compared to Eq. (3.11), we do not "pair" elements together). As a result, we only can say that any given $\mathbf{p}_\theta$ will have

$$\sum_{i=1}^{k} \mathbf{p}_\theta(i) = 1 + \frac{3\varepsilon}{k}\sum_{i=1}^{k}\theta_i \in [1-3\varepsilon, 1+3\varepsilon] \tag{3.18}$$

(and actually have $\sum_{i=1}^{k} \mathbf{p}_\theta(i) = 1 + O(\varepsilon/\sqrt{k})$ with high probability over the choice of θ); but that is enough for us. What this *does* buy us is that, combined with the Poissonization sampling assumption discussed above, *conditioned on* $\mathbf{b}$ the random variables $N_1, \dots, N_k$ are mutually independent, with either $N_i \sim \mathrm{Poisson}(n/k)$ (if $\mathbf{b}=1$) or $N_i \sim \frac{1}{2}\mathrm{Poisson}((1+3\varepsilon)/k) + \frac{1}{2}\mathrm{Poisson}((1-3\varepsilon)/k)$ (if $\mathbf{b}=0$).

Thus, all that remains is getting a good upper bound on $I(\mathbf{b} \wedge X)$, where X is the set of $\mathrm{Poisson}(n)$ samples from our 3-step process. Since the order of the N samples in X does not matter, we have

$I(\mathbf{b} \wedge X) = I(\mathbf{b} \wedge (N_1, \ldots, N_k))$, and can write

$$\begin{aligned} I(\mathbf{b} \wedge X) &= H(N_1, \ldots, N_k) - H((N_1, \ldots, N_k) \mid \mathbf{b}) \\ &= H(N_1, \ldots, N_k) - \sum_{i=1}^{k} H(N_i \mid \mathbf{b}) \\ &\qquad\qquad \text{(conditional independence)} \\ &\leq \sum_{i=1}^{k} H(N_i) - \sum_{i=1}^{k} H(N_i \mid \mathbf{b}) \qquad \text{(subadditivity)} \\ &= \sum_{i=1}^{k} I(\mathbf{b} \wedge N_i)\,, \end{aligned}$$

which means that to bound $I(\mathbf{b} \wedge X)$ it suffices to analyze each $I(\mathbf{b} \wedge N_i)$ separately. In our particular case, by symmetry of the uniform distribution and of our prior (uniform prior over all our "Paninski" no-instances), all those $I(\mathbf{b} \wedge N_i)$'s are equal:

$$I(\mathbf{b} \wedge X) \leq k I(\mathbf{b} \wedge N_1)\,. \tag{3.19}$$

To analyse this, we first replace the RHS by something a little bit more manageable, to get the following:

$$I(\mathbf{b} \wedge X) \leq \frac{k}{2} \sum_{j=0}^{\infty} \Pr[N_1 = j \mid \mathbf{b} = 1] \left(1 - \frac{\Pr[N_1 = j \mid \mathbf{b} = 0]}{\Pr[N_1 = j \mid \mathbf{b} = 1]}\right)^2 \tag{3.20}$$

To see where this comes from, recall that mutual information is expected KL divergence, and chi-square divergence upper bounds KL:

and Eq. (3.20) is actually a "chi-square type" quantity in disguise.

$$\begin{aligned}
I(\mathbf{b} \wedge N_1) &= \mathbb{E}_{\mathbf{b}}\Big[\mathrm{D}\Big(P_{N_1|\mathbf{b}} \| P_{N_1}\Big)\Big] \\
&\le \mathbb{E}_{\mathbf{b}}\Big[\chi^2\Big(P_{N_1|\mathbf{b}} \,||\, P_{N_1}\Big)\Big] \\
&= \frac{1}{2}\Big(\chi^2\Big(P_{N_1|\mathbf{b}=1} \,||\, P_{N_1}\Big) + \chi^2\Big(P_{N_1|\mathbf{b}=0} \,||\, P_{N_1}\Big)\Big) \\
&= \frac{1}{2}\Big(\sum_{j=0}^{\infty} \frac{(\Pr[N_1 = j \mid \mathbf{b} = 1] - \Pr[N_1 = j])^2}{\Pr[N_1 = j]} \\
&\qquad + \sum_{j=0}^{\infty} \frac{(\Pr[N_1 = j \mid \mathbf{b} = 0] - \Pr[N_1 = j])^2}{\Pr[N_1 = j]}\Big) \\
&= \frac{k}{2}\sum_{j=0}^{\infty} \frac{(\Pr[N_1 = j \mid \mathbf{b} = 0] - \Pr[N_1 = j \mid \mathbf{b} = 1])^2}{\Pr[N_1 = j \mid \mathbf{b} = 0] + \Pr[N_1 = j \mid \mathbf{b} = 1]} \qquad (3.21) \\
&= \frac{k}{2}\sum_{j=0}^{\infty} \Pr[N_1 = j \mid \mathbf{b} = 1] \frac{\Big(1 - \frac{\Pr[N_1=j|\mathbf{b}=0]}{\Pr[N_1=j|\mathbf{b}=1]}\Big)^2}{1 + \frac{\Pr[N_1=j|\mathbf{b}=0]}{\Pr[N_1=j|\mathbf{b}=1]}}
\end{aligned}$$

using for the second-to-last equality that, since $\mathbf{b}$ is a uniform bit, $\Pr[N_1 = j] = \frac{1}{2}\Pr[N_1 = j \mid \mathbf{b} = 1] + \frac{1}{2}\Pr[N_1 = j \mid \mathbf{b} = 0]$ for all $j \ge 0$.

In view of bounding the RHS of Eq. (3.20), it is time to take advantage of what we know: after all, we know what our no-instances and yes-instance look like, and we chose them for a reason. Recalling that we work in the Poissonized sampling model, $N_1 \sim \mathrm{Poisson}(n\mathbf{p}(1))$, so we can compute

$$\begin{aligned}
\Pr[N_1 = j \mid \mathbf{b} = 1] &= e^{-n/k}\frac{(n/k)^j}{j!} \\
\Pr[N_1 = j \mid \mathbf{b} = 0] &= \frac{1}{2}\left(e^{-\frac{n(1+3\varepsilon)}{k}}\frac{(n(1+3\varepsilon)/k)^j}{j!} + e^{-\frac{n(1-3\varepsilon)}{k}}\frac{(n(1-3\varepsilon)/k)^j}{j!}\right)
\end{aligned}$$

and, in particular,

$$\frac{\Pr[N_1 = j \mid \mathbf{b} = 0]}{\Pr[N_1 = j \mid \mathbf{b} = 1]} = \frac{1}{2}\Big(e^{-\frac{3n\varepsilon}{k}}(1+3\varepsilon)^j + e^{\frac{3n\varepsilon}{k}}(1-3\varepsilon)^j\Big)\,. \qquad (3.22)$$

We can plug this back into Eq. (3.20), to get

$$I(\mathbf{b} \wedge X) \leq \frac{k}{2} e^{-n/k} \sum_{j=0}^{\infty} \frac{(n/k)^j}{j!} \left(1 - \frac{e^{-\frac{3n\varepsilon}{k}}(1+3\varepsilon)^j + e^{\frac{3n\varepsilon}{k}}(1-3\varepsilon)^j}{2}\right)^2$$
$$= \frac{k}{2}\left(\cosh\left(\frac{9n\varepsilon^2}{k}\right) - 1\right) \leq \frac{k}{2}\left(e^{\frac{81n^2\varepsilon^4}{2k^2}} - 1\right); \tag{3.23}$$

the last inequality being $\cosh u \leq e^{u^2/2}$, and the equality following from either a fastidious but manageable computation involving known converging series, or a quick computation involving Julia, Mathematica, or one's weapon of choice.

Since we know by Fact 3.1 that we must have $I(\mathbf{b} \wedge X) \gtrsim 1$, putting the two together implies that we must have

$$\frac{81n^2\varepsilon^4}{2k^2} \geq \ln\left(1 + \frac{4}{25k}\right) \asymp \frac{1}{k}$$

proving our lower bound:

Theorem 3.4. Every testing algorithm for uniformity must have sample complexity $n(k, \varepsilon, 1/3) = \Omega(\sqrt{k}/\varepsilon^2)$.

3.3 Indistinguishability via moment-matching

We will here describe a general, out-of-the-box theorem which applies to a broad class of distribution properties: namely, the class of *symmetric properties*, which are those which do not depend on the individual labels of the domain elements.

Definition 3.1. A property $\mathcal{P} = \cup_{k=1}^{\infty} \mathcal{P}_k$ of distributions is said to be *symmetric* if it is closed under permutations of the domain: for every k and every permutation $\sigma : \mathcal{X}_k \to \mathcal{X}_k$, if $\mathbf{p} \in \mathcal{P}_k$ then $\mathbf{p} \circ \sigma \in \mathcal{P}_k$.

A by now familiar example of symmetric property is uniformity, $\mathcal{P}_k = \{\mathbf{u}_k\}$, since the uniform distribution is invariant by relabeling: $\mathbf{u}_k(\sigma(i)) = \mathbf{u}_k(i)$ for every $i \in \mathcal{X}_k$ and every permutation σ of $\mathcal{X}_k$. Other notable examples include the property of "all distributions of support size at most s," that of "distributions of (Shannon) entropy

at least h," but, for instance, *not* the property of "distributions with a non-increasing pmf" (since it depends on the ordering of the domain).

The definition of symmetric properties can be extended to multiple distributions over the same domain: for instance, taking $\mathcal{X}_k = [k] \times [k]$, a property $\mathcal{P}_k$ of product distributions is symmetric if $\mathbf{p}_1 \otimes \mathbf{p}_2 \in \mathcal{P}_k$ implies $\mathbf{p}_1 \circ \sigma \otimes \mathbf{p}_2 \circ \sigma \in \mathcal{P}_k$ for all permutations σ. This is the case, *e.g.*, of the property corresponding to closeness testing, $\{ \mathbf{p}_1 \otimes \mathbf{p}_2 : \mathbf{p}_1 = \mathbf{p}_2 \}$, mentioned in Chapter 1.

Symmetric properties are nice in the sense that, when considering them, one can completely forget about the individual values of the n samples taken, and focus instead on the empirical histogram. That is, a sufficient statistic for symmetric properties is the *fingerprint* of the samples, which is just the tuple

$$F := (F_0, F_1, F_2, \dots, F_n) \in \mathbb{N}^{n+1}$$

where $F_j = \sum_{i=1}^{k} \mathbb{1}\{N_i = j\}$ is the number of elements of the domain which appear exactly j times among the n samples. In particular, we always have $\sum_{j=0}^{n} F_j = n$.

E: Check that you can express several of the algorithms in Section 2.1 as a function of F only: Exercise 3.5.

The main result discussed in this section is the "Wishful Thinking Theorem" of Valiant (2011), which applies to testing symmetric properties of distributions. Intuitively, this theorem ensures that "if the low-degree moments (ℓ_p norms) of two distributions match, then these distributions (up to relabeling) are hard to distinguish." To see why this is the case, and justify the name of the theorem, observe that since we focus on symmetric properties all which matters is the fingerprint F introduced about; that is, the number of j-collisions, for every $j \geq 0$.

Now, given a distribution $\mathbf{p}$, the number of j-collisions in n samples has expectation

$$\binom{n}{j} \|\mathbf{p}\|_j^j \asymp n^j \|\mathbf{p}\|_j^j$$

and variance, wishfully ignoring all dependencies, maybe something like $n^j \|\mathbf{p}\|_j^j$ as well (roughly what it would be if j-wise collisions were Binomial with parameters $\binom{n}{j}$ and $\|\mathbf{p}\|_j^j$ – again, wishful thinking). So, given two probability distributions $\mathbf{p}^{\mathsf{yes}}, \mathbf{p}^{\mathsf{no}}$, if the squared gap between the expected numbers of j-wise collisions was much smaller than the

maximum of the two variances

$$(n^j \|\mathbf{p}^{\mathsf{yes}}\|_j^j - n^j \|\mathbf{p}^{\mathsf{no}}\|_j^j)^2 \ll \max(n^j \|\mathbf{p}^{\mathsf{yes}}\|_j^j, n^j \|\mathbf{p}^{\mathsf{no}}\|_j^j)$$

for all $j \geq 1$; or, equivalently,

$$\frac{|n^j \|\mathbf{p}^{\mathsf{yes}}\|_j^j - n^j \|\mathbf{p}^{\mathsf{no}}\|_j^j|}{\sqrt{\max(n^j \|\mathbf{p}^{\mathsf{yes}}\|_j^j, n^j \|\mathbf{p}^{\mathsf{no}}\|_j^j)}} \ll 1$$

for all $j \geq 1$, then we could *hope* that the two distributions are indistinguishable from their fingerprints on n samples. Well, the reasoning above is flawed in a few ways, but can be made rigorous with enough work; and, luckily, someone else took care of this already:

Theorem 3.5 (Wishful Thinking Theorem (Valiant, 2011, Theorem 4.10))**.** Fix any *symmetric* property $\mathcal{P}$. Given a positive integer n, a distance parameter $\varepsilon \in (0, 1]$, and two distributions $\mathbf{p}^{\mathsf{yes}}, \mathbf{p}^{\mathsf{no}} \in \Delta_k$, suppose the following conditions hold:

1. $\|\mathbf{p}^{\mathsf{yes}}\|_\infty, \|\mathbf{p}^{\mathsf{no}}\|_\infty \leq \frac{1}{500n}$;
2. letting m^{yes}, m^{no} be the n-based moments of $\mathbf{p}^{\mathsf{yes}}, \mathbf{p}^{\mathsf{no}}$ (defined below),
$$\sum_{j=2}^{\infty} \frac{|m^{\mathsf{yes}}(j) - m^{\mathsf{no}}(j)|}{\lfloor j/2 \rfloor ! \sqrt{1 + \max(m^{\mathsf{yes}}(j), m^{\mathsf{no}}(j))}} < \frac{1}{24},$$
where $m^{\mathsf{yes}}(j) := n^j \|\mathbf{p}^{\mathsf{yes}}\|_j^j$, $m^{\mathsf{no}}(j) := n^j \|\mathbf{p}^{\mathsf{no}}\|_j^j$ for $j \geq 0$,
3. $\mathbf{p}^{\mathsf{yes}} \in \mathcal{P}_k$ and $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}^{\mathsf{no}}, \mathcal{P}_k) > \varepsilon$.

Then every testing algorithm for $\mathcal{P}$ must have sample complexity $n(k, \varepsilon, 1/3) > n$.

(Side remark: the term $j = 1$ does not appear in the sum, since $n\|\mathbf{p}\|_1 = n$ for every distribution $\mathbf{p}$, and so this term always cancels out.)

To see the strength of this theorem , let us use it to prove the $\Omega\left(\sqrt{k}/\varepsilon^2\right)$ lower bound for uniformity testing. Our distribution $\mathbf{p}^{\mathsf{yes}}$ will, of course, have to be the uniform distribution $\mathbf{u}_k$ itself; as for

$\mathbf{p}^{\mathsf{no}}$, let us take it to be any of the instances of "Paninski construction" (Eq. (3.11)), so that

$$\mathbf{p}^{\mathsf{no}}(2i) = \frac{1+3\varepsilon}{k}, \qquad \mathbf{p}^{\mathsf{no}}(2i-1) = \frac{1-3\varepsilon}{k}, \qquad 1 \le i \le k/2$$

(where we again assume without loss of generality that k is even, and $\varepsilon \in (0, 1/3]$). We then have $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}^{\mathsf{no}}, \mathcal{P}_k) = \mathrm{d}_{\mathrm{TV}}(\mathbf{p}^{\mathsf{no}}, \mathbf{u}_k) = \frac{3}{2}\varepsilon > \varepsilon$; so let's check the two conditions of the theorem hold.

The first condition, $\|\mathbf{p}^{\mathsf{yes}}\|_\infty, \|\mathbf{p}^{\mathsf{no}}\|_\infty \le \frac{1}{500n}$ will be satisfied as long as $n \le \frac{k}{1000}$, since $\|\mathbf{p}^{\mathsf{yes}}\|_\infty \le \|\mathbf{p}^{\mathsf{no}}\|_\infty \le 2/k$. This is a limitation which will limit the range of applicability of the lower bound, but we can live with it (and will get back to it later).

Turning to the second condition, we need to compute these n-based moments. Luckily, it is a simple matter to check that, for every $j \ge 2$,

$$m^{\mathsf{yes}}(j) = \frac{n^j}{k^{j-1}} \tag{3.24}$$

while

$$\begin{aligned} m^{\mathsf{no}}(j) &= n^j \sum_{i=1}^{k} \mathbf{p}^{\mathsf{no}}(i)^j = n^j \left(\frac{k}{2}\left(\frac{1+3\varepsilon}{k}\right)^j + \frac{k}{2}\left(\frac{1-3\varepsilon}{k}\right)^j \right) \\ &= \frac{n^j}{k^{j-1}} \left(\frac{(1+3\varepsilon)^j + (1-3\varepsilon)^j}{2} \right) \le \frac{2^j n^j}{k^{j-1}} \end{aligned} \tag{3.25}$$

For instance, for the special case of $j = 2$, the expression is a little nicer, and becomes

$$m^{\mathsf{no}}(2) = (1 + 9\varepsilon^2)\frac{n^2}{k}\,. \tag{3.26}$$

Without wanting to spoil the surprise, we "should" expect the term $j = 2$ of the series $\sum_{j=2}^{\infty} \frac{|m^{\mathsf{yes}}(j) - m^{\mathsf{no}}(j)|}{\lfloor j/2 \rfloor ! \sqrt{1 + \max(m^{\mathsf{yes}}(j), m^{\mathsf{no}}(j))}}$ to dominate (as the second moment $\|\mathbf{p}\|_2^2$ of the distribution is "what gives it away" in uniformity testing, as we saw now and again in Section 2.1), so we will want to make sure we handle that term as tightly as possible.

With the above expressions at our disposal, we can proceed: first, since the series decays quite fast already (at least geometrically) as $n/k \ll 1$, the factorial in the denominator does not look crucial and it

seems reasonable to ignore it. Moreover this maximum in the denominator seems annoying and will prevent us from easily computing the series, so let's get rid of it as well:

$$\begin{aligned}\sum_{j=2}^{\infty} \frac{|m^{\mathsf{yes}}(j) - m^{\mathsf{no}}(j)|}{\sqrt{1 + \max(m^{\mathsf{yes}}(j), m^{\mathsf{no}}(j))}} &\le \sum_{j=2}^{\infty} |m^{\mathsf{yes}}(j) - m^{\mathsf{no}}(j)| \\ &\le \frac{9\varepsilon^2 n^2}{k} + \sum_{j=3}^{\infty} \frac{(2^j - 1)n^j}{k^{j-1}} \\ &\le \frac{9\varepsilon^2 n^2}{k} + 2n \sum_{j=2}^{\infty} \frac{2^j n^j}{k^j} \\ &= \frac{9\varepsilon^2 n^2}{k} + \frac{8n^3}{k^2} \cdot \frac{1}{1 - 2n/k} \\ &\le \frac{9\varepsilon^2 n^2}{k} + \frac{9n^3}{k^2}\end{aligned}$$

where we used the assumption that $n \le k/1000$ to guarantee convergence of the geometric series, and bound its sum. Now, even ignoring the second term, we see that the RHS will only be less that $1/24$ (as required by the second condition of the theorem) if $n \ll \sqrt{k}/\varepsilon$, so the best lower bound we can hope for is $\Omega\left(\sqrt{k}/\varepsilon\right)$. But we wanted $\Omega\left(\sqrt{k}/\varepsilon^2\right)$!

Oops.

What went wrong? We were a little too eager to get rid of "this maximum in the denominator." It *is* annoying, and it *is* a good idea to get rid of it in order to be left with a geometric series (at least to get a sense of what is going on), but *not* in that way. Let's try again.

$$\begin{aligned}\sum_{j=2}^{\infty} \frac{|m^{\mathsf{yes}}(j) - m^{\mathsf{no}}(j)|}{\sqrt{1 + \max(m^{\mathsf{yes}}(j), m^{\mathsf{no}}(j))}} &\le \sum_{j=2}^{\infty} \frac{|m^{\mathsf{yes}}(j) - m^{\mathsf{no}}(j)|}{\sqrt{m^{\mathsf{yes}}(j)}} \\ &= \sum_{j=2}^{\infty} \frac{n^{j/2}}{k^{(j-1)/2}} \left(\frac{(1 + 3\varepsilon)^j + (1 - 3\varepsilon)^j}{2} - 1 \right) \\ &= \frac{1}{2\sqrt{k}} \sum_{j=2}^{\infty} (\alpha^j + \beta^j - 2\gamma^j)\end{aligned}$$

where $\alpha := (1 + 3\varepsilon)\sqrt{n/k}$, $\beta := (1 - 3\varepsilon)\sqrt{n/k}$, $\gamma := \sqrt{n/k}$, and we used Eqs. (3.24) and (3.25) for the first equality. Since all three are in

$(0,1)$ (recall that we have $n \ll k$), we can compute the geometric series to get

$$\sum_{j=2}^{\infty} \frac{|m^{\mathsf{yes}}(j) - m^{\mathsf{no}}(j)|}{\sqrt{1+\max(m^{\mathsf{yes}}(j), m^{\mathsf{no}}(j))}} \leq \frac{1}{2\sqrt{k}} \left(\frac{(1+3\varepsilon)^2\gamma^2}{1-(1+3\varepsilon)\gamma} + \frac{(1-3\varepsilon)^2\gamma^2}{1-(1-3\varepsilon)\gamma} - \frac{2\gamma^2}{1-\gamma} \right)$$

This looks better! Sure, this is quite ugly; but a Taylor expansion at 0 (since $\gamma = \sqrt{n/k} \ll 1$) tells us that the parenthesis of the RHS is

$$18\varepsilon^2\gamma^2 + o(\gamma^2) = \frac{18\varepsilon^2 n}{k} + o\left(\frac{n}{k}\right)$$

so we should be fine; and indeed, one can check that that parenthesis is equal to

$$\frac{18\varepsilon^2\gamma^2}{(1-\gamma)(1-(1+3\varepsilon)\gamma)(1-(1-3\varepsilon)\gamma)} \leq 144\varepsilon^2\gamma^2 \,.$$

From this, we get

$$\sum_{j=2}^{\infty} \frac{|m^{\mathsf{yes}}(j) - m^{\mathsf{no}}(j)|}{\sqrt{1+\max(m^{\mathsf{yes}}(j), m^{\mathsf{no}}(j))}} \leq \frac{1}{2\sqrt{k}} \cdot 144\varepsilon^2 \frac{n}{k} = \frac{72\varepsilon^2 n}{\sqrt{k}} \,. \tag{3.27}$$

This in turn will be less than $1/24$ for $n \leq \sqrt{k}/(1728\varepsilon^2)$. Success! Recalling finally the condition $n \ll k/$ (for the first condition of the Wishful Thinking theorem to hold) which imposes $\varepsilon \gg 1/k^{1/4}$, by invoking Theorem 3.5 we get the result we wanted:

Theorem 3.6. Every testing algorithm for uniformity must have sample complexity $n(k, \varepsilon, 1/3) = \Omega(\sqrt{k}/\varepsilon^2)$, provided that $\varepsilon \geq 1/k^{1/4}$.

The key aspect of this lower bound was how *painless* it was to obtain it. The main idea was to use the same Paninski construction as before, check a couple conditions, compute a geometric series, and then conclude by Theorem 3.5. (Sure, it might have felt a *little* longer than this, but this is mostly due to the author's choice of going through two consecutive attempts, instead of skipping directly to the second one.)

3.4 Indistinguishability on an instance-by-instance basis

We will now discuss and illustrate the use of a very convenient (albeit intimidating) result due to Valiant and Valiant (2017), which allows one to establish lower bounds tailored to any reference distribution $\mathbf{q}$.

Theorem 3.7. Given a distribution $\mathbf{q}$ over $[k]$, and associated values α_i such that $\alpha_i \in [0,1]$ for all $i \in [k]$, define the set of distributions $\mathcal{C} = \{\mathbf{p}_z\}_{z\in\{-1,+1\}^k}$ by setting, for every $z \in \{-1,+1\}^k$,

$$\mathbf{p}_z(i) := \frac{(1 + z_i\alpha_i)\mathbf{q}(i)}{\sum_{j=1}^k (1 + z_j\alpha_j)\mathbf{q}(j)}, \qquad i \in [k] \tag{3.28}$$

i.e., $\mathbf{p}_z(i) \propto (1 + z_i\alpha_i)\mathbf{q}(i)$. Then there exists an absolute constant $c > 0$ such that any algorithm which, given n i.i.d. samples from an arbitrary distribution $\mathbf{p}$, distinguishes with success probability at least $2/3$ between (i) $\mathbf{p} = \mathbf{q}$ and (ii) $\mathbf{p} \in \mathcal{C}$, must satisfy

$$n \geq \frac{c}{\sqrt{\sum_i \alpha_i^4 \mathbf{q}(i)^2}}\,. \tag{3.29}$$

Further, if $\max_i \alpha_i\mathbf{q}(i) \leq \frac{1}{2}\sum_{i=1}^k \alpha_i\mathbf{q}(i)$, then

$$\Pr_Z\left[\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_Z, \mathbf{q}) > \frac{1}{4}\sum_{i=1}^k \alpha_i\mathbf{q}(i)\right] \geq \frac{1}{2}\,, \tag{3.30}$$

where Z is uniformly random on $\{-1,+1\}^k$.

This is a bit of a mouthful, so let's break Theorem 3.7 down before seeing a few corollaries and applications. First, given a reference distribution $\mathbf{q}$, and a choice of "element-wise perturbations" values $\alpha_1, \ldots, \alpha_k$, Eq. (3.28) says we should define a "hard instance" by setting $\mathbf{p}(i) = (1 \pm \alpha_i)\mathbf{q}(i)$, choosing the sign independently and uniformly at random for every i, and then normalizing the resulting $\mathbf{p}$ to make it a true probability distribution. After doing this, Eq. (3.29) states that distinguishing $\mathbf{q}$ from a hard instance $\mathbf{p}$ chosen randomly this way requires $\Omega(1/\sqrt{\sum_i \alpha_i^4\mathbf{q}(i)^2})$ samples. Finally, Eq. (3.30) tells us that (until some mild condition on the α_i's), most of those hard instances are actually *far* from $\mathbf{q}$; in particular, we will typically want to choose

the α_i's so that the guaranteed distance $\frac{1}{4}\sum_{i=1}^{k}\alpha_i\mathbf{q}(i)$ is equal to our parameter ε.

But *how* should we choose these values $\alpha_1, \ldots, \alpha_k$? In view of what we just discussed, it seems natural to set $\alpha_i := 4\varepsilon$ for all i, ensuring that $\frac{1}{4}\sum_{i=1}^{k}\alpha_i\mathbf{q}(i) = \varepsilon$. This also immediately satisfies $\max_i \alpha_i\mathbf{q}(i) \leq \frac{1}{2}\sum_{i=1}^{k}\alpha_i\mathbf{q}(i)$, as long as $\|\mathbf{q}\|_\infty \leq 1/2$: a rather mild condition. As for Eq. (3.29), plugging in this choice of α_i's shows that it becomes

$$n \gtrsim \frac{1}{\varepsilon^2\|\mathbf{q}\|_2}$$

which seems... good? For instance, when $\mathbf{q}$ is the uniform distribution, then $\|\mathbf{q}\|_2 = 1/\sqrt{k}$, and we get an $\Omega(\sqrt{k}/\varepsilon^2)$ lower bound! Specifically, we (almost) proved:

Theorem 3.8. Every testing algorithm for identity with reference $\mathbf{q}$ must have sample complexity $n(k, \varepsilon, 1/3) = \Omega(1/(\varepsilon^2\|\mathbf{q}\|_2))$, provided that $\|\mathbf{q}\|_\infty \leq 1/2$.

In particular, every testing algorithm for uniformity must have sample complexity $n(k, \varepsilon, 1/3) = \Omega(\sqrt{k}/\varepsilon^2)$.

Proof. The above discussion, combined with Theorem 3.7, *almost* establishes the result we want, up to some annoying detail: the success probability is not exactly what we needed, due to the fact that Theorem 3.7 only guarantees a random distribution $\mathbf{p}_z$ is ε-far from $\mathbf{q}$ with probability 1/2. Namely, assume we have a testing algorithm $\mathcal{A}$ for identity with reference $\mathbf{q}$ with sample complexity $n = n(k, \varepsilon, 1/3)$. We can use it to distinguish between $\mathbf{p} = \mathbf{q}$ and a (uniformly randomly chosen) $\mathbf{p}$ from $\mathcal{C}$ (as defined with our choice $\alpha_i = 4\varepsilon$), giving that (1) if the distribution is $\mathbf{q}$, then we say so with probability at least 2/3 (good); (2) if the distribution is ε-far from $\mathbf{q}$, then we say $\mathbf{p} \in \mathcal{C}$ with probability at least 2/3 (good); *but* if (3) if the distribution is in $\mathcal{C}$ but *not* ε-far from $\mathbf{q}$, then we cannot say anything about being right or wrong (bad). So when $\mathbf{p}$ is chosen at random from $\mathcal{C}$, we can only guarantee we are correct with probability at least $2/3 \cdot (1/2) = 1/3$... not 2/3, which would be necessary to get the sample complexity lower bound from Theorem 3.7 (Eq. (3.29)) to apply.

Fortunately, there is a fix. Define $\mathcal{A}'$ as follows: given "enough" samples (but still $O(n)$), it runs $\mathcal{A}$ on disjoint subsets of n samples and takes a majority vote, to amplify its success probability from 2/3 to 3/4 (as described in Lemma 1.1). Looking at the output, it then does the following:

- if the output is 1 ($\mathbf{p} \neq \mathbf{q}$), then it outputs 1;
- if the output is 0 ($\mathbf{p} = \mathbf{q}$), then it outputs 0 with probability 17/24, and 1 with probability 7/24.

Why did we do this? If $\mathbf{p} = \mathbf{q}$, then the probability that $\mathcal{A}'$ correctly outputs 0 is now at least $\frac{3}{4} \cdot \frac{17}{24} > \frac{1}{2}$ (worse than before). But if $\mathbf{p} \in \mathcal{C}$, it will (correctly) output 1 with probability at least

$$\frac{3}{4} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{7}{24} > \frac{1}{2}$$

(better than before). Since both probabilities are constants greater than 1/2, we can again use the same amplification trick (Lemma 1.1) on $\mathcal{A}'$ to get $\mathcal{A}''$, which is correct in distinguishing $\mathbf{p} = \mathbf{q}$ from $\mathbf{p} \in \mathcal{C}$ with probability at least 2/3 in both cases. Moreover, since all these amplifications only required to blow up the sample complexity by a constant factor, $\mathcal{A}''$ still uses $n' = O(n)$ samples, and applying Theorem 3.7 on $\mathcal{A}''$ leads to

$$n' = \Omega(1/(\varepsilon^2 \|\mathbf{q}\|_2))$$

which implies $n = \Omega(1/(\varepsilon^2 \|\mathbf{q}\|_2))$. □

This corollary is very handy, and provides a non-trivial lower bound as a function of some easily interpretable function of the reference $\mathbf{q}$. One can even generalize it to distributions over $\mathbb{N}$ instead of $[k]$ (infinite discrete domains), keeping the same statement and proof!

E: Apply this to the Binomial distribution, to get $\Omega(k^{1/4}/\varepsilon^2)$.

This raises the question: is Theorem 3.8 always optimal? Or, to put things more bluntly: is allowing for different α_i's in Theorem 3.7 useful, or is it just for show, and unnecessarily complicated?

It is not just for show. Consider the following "Zipf" distribution $\mathbf{q}$ on $[k]$, where $\mathbf{q}(i) \propto 1/\sqrt{i}$:

$$\mathbf{q}(i) = \frac{1}{H_{k,1/2}\sqrt{i}}, \qquad i \in [k]\,, \tag{3.31}$$

where

$$H_{k,1/2} := \sum_{i=1}^{k} \frac{1}{\sqrt{i}} \underset{k\to\infty}{\sim} 2\sqrt{k}$$

is the generalized Harmonic number of order $1/2$ (and H_k will be the usual Harmonic number). A direct computation then shows that

$$\|\mathbf{q}\|_2 = \frac{\sqrt{H_k}}{H_{k,1/2}} \underset{k\to\infty}{\sim} \frac{\sqrt{\ln k}}{2\sqrt{k}},$$

and so Theorem 3.8 gives a lower bound of $\Omega\left(\frac{\sqrt{k}}{\varepsilon^2\sqrt{\log k}}\right)$ samples. This does not look too bad, especially since we know from Chapter 2 that an *upper* bound of $O\left(\sqrt{k}/\varepsilon^2\right)$ samples holds. But that leaves a gap of $\sqrt{\log k}$ between the two, which could go either way.

However, let us look at this $\mathbf{q}$, and what a typical "local perturbation" $\mathbf{p} \in \mathcal{C}$ looks like when we perturb each element by $1 \pm \varepsilon$:

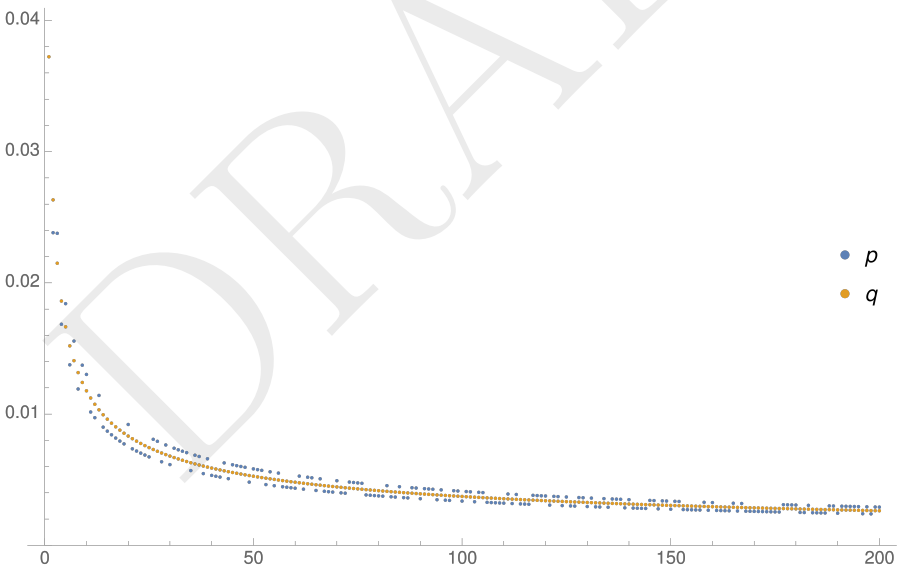

Figure 3.1: Our reference $\mathbf{q}$ (Zipf distribution), along with a randomly chosen perturbation $\mathbf{p}$, here depicted for $k = 200$, $\varepsilon = 1/10$.

While the second half ("tail") of the distribution looks somewhat uniform (all probabilities between $k/2$ and k are within a factor $\sqrt{2}$),

the first half is clearly not, with the first few elements having much higher probability. Perturbing those heavy elements by the same relative amount as the rest, "intuitively," is not a good idea, as it is easier to detect and can give away a lot more information. Instead, we will see what happens when we only perturb this somewhat-uniform tail of $\mathbf{q}$: define our α_i's by

$$\alpha_i := \begin{cases} 16\varepsilon & \text{if } i \geq \frac{k}{2} \\ 0 & \text{otherwise} \end{cases}$$

In view of applying Theorem 3.7, we first check that the distance of our hard instances $\mathbf{p}_z$ to $\mathbf{q}$ will be at least ε: $\max_i \alpha_i \mathbf{q}(i) = 16\varepsilon \mathbf{q}(k/2) \ll \varepsilon$, and

$$\frac{1}{4}\sum_{i=1}^{k} \alpha_i \mathbf{q}(i) = \frac{4\varepsilon}{H_{k,1/2}} \sum_{i=\frac{k}{2}}^{k} \frac{1}{\sqrt{i}} = 4\varepsilon\left(1 - \frac{H_{\frac{k}{2},1/2}}{H_{k,1/2}}\right) \geq 4\left(1 - \frac{1}{\sqrt{2}}\right)\varepsilon > \varepsilon\,,$$

so by Eq. (3.30) the distance is fine. Turning to the sample complexity lower bound this will lead to, we compute

$$\sum_{i=1}^{k} \alpha_i^4 \mathbf{q}(i)^2 = \frac{(16\varepsilon)^4}{H^2_{\frac{k}{2},1/2}} \sum_{i=\frac{k}{2}}^{k} \frac{1}{i} = \frac{(16\varepsilon)^4}{H^2_{\frac{k}{2},1/2}}\left(H_k - H_{\frac{k}{2}}\right) \asymp \frac{\varepsilon^4}{k}$$

since $H_k - H_{k/2} = \ln 2 + o(1)$. By Eq. (3.29), this means we get a (tight) lower bound of $\Omega(\sqrt{k}/\varepsilon^2)$ samples! All we needed to do was to restrict our perturbation to the "near-uniform" part of the reference distribution; crucially, this required the ability to choose different α_i's for different i, as allowed by Theorem 3.7.

To conclude this section, we will derive another corollary to Theorem 3.7, which provides the "best" perturbation possible for any given reference $\mathbf{q}$; that is, an optimal choice for the α_i's as a function of $\mathbf{q}$. Based on our previous example, we know that α_i somehow has to *adapt* to $\mathbf{q}(i)$, to differentiate between the heavy elements and the "near-uniform" part of the distribution $\mathbf{q}$.

Now, on the one hand we want to establish as large a lower bound as possible, and so Eq. (3.29) says we should maximize $\sum_{i=1}^{k} \alpha_i^4 \mathbf{q}(i)^2$. On the other hand, for our lower bound to be meaningful, we also need

our hard instances to be at distance ε from $\mathbf{q}$, and for that Eq. (3.30) imposes the condition $\sum_{i=1}^{k} \alpha_i \mathbf{q}(i) \asymp \varepsilon$. Since

$$\sum_{i=1}^{k} \alpha_i^4 \mathbf{q}(i)^2 \le \max_i \alpha_i^3 \mathbf{q}(i) \cdot \sum_{i=1}^{k} \alpha_i \mathbf{q}(i) \tag{3.32}$$

combining the two conditions leads us to try and maximize $\max_i \alpha_i^3 \mathbf{q}(i)$, and the condition for equality in Hölder's inequality further suggests we should have $\alpha_i^3 \mathbf{q}(i)$ constant; that is,

$$\alpha_i \propto 1/\mathbf{q}(i)^{1/3}\,. \tag{3.33}$$

We may not be able to do this exactly, as the theorem also requires that $\alpha_i \le 1$, so we will have to cap our α_i's; but this motivates the following idea. Without loss of generality, assume that $\mathbf{q}$ is non-increasing: $\mathbf{q}(1) \ge \mathbf{q}(2) \ge \cdots \ge \mathbf{q}(k)$ (we can, since we know $\mathbf{q}$, and can permute the domain if we want) and $\varepsilon \in (0, 1/8]$; and assume, *with* some loss of generality, that $\mathbf{q}(1) = \|\mathbf{q}\|_\infty \le 1/2$. First, define $\alpha \ge 0$ as a value such that

$$\frac{1}{4} \sum_{i=2}^{k} \left(1 \wedge \frac{\alpha}{\mathbf{q}(i)^{1/3}}\right) \mathbf{q}(i) = \varepsilon \tag{3.34}$$

where we start the summation at $i = 2$ to (later) handle the (annoying) condition from Theorem 3.7 on $\max_i \alpha_i \mathbf{q}(i)$, which will be $\alpha_1 \mathbf{q}(1)$. To see why such a choice of α always exists, note that the LHS of Eq. (3.34) is continuous and non-decreasing in α, equal to 0 for $\alpha = 0$, and goes to $\frac{1}{4}\sum_{i=2}^{k} \mathbf{q}(i) = \frac{1}{4}(1 - \|\mathbf{q}\|_\infty) \ge \frac{1}{8}$ for $\alpha \to \infty$.

One we have this value α, we can then set

$$\alpha_i = \begin{cases} 1 \wedge \frac{\alpha}{\mathbf{q}(i)^{1/3}} & \text{if } 1 \le i \le k \\ \alpha_2 \frac{\mathbf{q}(2)}{\mathbf{q}(1)} & \text{if } i = 1 \end{cases} \tag{3.35}$$

where the assumption that $\mathbf{q}$ is non-increasing implies that (i) $\alpha_1 \le \alpha_2 \le \alpha_3 \le \cdots \le \alpha_k$, and (ii) $\alpha_1 \mathbf{q}(1) = \alpha_2 \mathbf{q}(2) \ge \alpha_3 \mathbf{q}(3) \ge \cdots \ge \alpha_k \mathbf{q}(k)$ (since $\alpha_i \mathbf{q}(i) = \mathbf{q}(i) \wedge (\alpha \mathbf{q}(i)^{2/3})$).

Our (somewhat bizarre) choice of α_1 is a technicality to ensure that

$$\max_i \alpha_i \mathbf{q}(i) = \alpha_1 \mathbf{q}(1) = \frac{1}{2}(\alpha_1 \mathbf{q}(1) + \alpha_2 \mathbf{q}(2)) \le \frac{1}{2} \sum_{i=1}^{k} \alpha_i \mathbf{q}(i)$$

so that the condition of Theorem 3.7 preceding Eq. (3.30) is satisfied. Letting L to be the largest value $i \geq 2$ such that $\alpha \mathbf{q}(i)^{-1/3} \leq 1$, we then can rewrite Eq. (3.34) as

$$\alpha \sum_{i=2}^{L} \mathbf{q}(i)^{2/3} + \sum_{i=L+1}^{k} \mathbf{q}(i) = 4\varepsilon\,, \tag{3.36}$$

from which $\alpha \leq 4\varepsilon / \sum_{i=2}^{L} \mathbf{q}(i)^{2/3}$. Finally, recalling Eq. (3.29), we bound

$$\begin{aligned}
\sum_{i=1}^{k} \alpha_i^4 \mathbf{q}(i)^2 \leq \sum_{i=2}^{k} \alpha_i^4 \mathbf{q}(i)^2 &= \alpha^4 \sum_{i=2}^{L} \mathbf{q}(i)^{2/3} + \sum_{i=L+1}^{k} \mathbf{q}(i)^2 \\
&\leq \alpha^3 \left(\alpha \sum_{i=2}^{L} \mathbf{q}(i)^{2/3} + \sum_{i=L+1}^{k} \mathbf{q}(i) \right) \\
&= 4\varepsilon\alpha^3 \qquad \text{(By Eq. (3.36))} \\
&\leq \frac{(4\varepsilon)^4}{\left(\sum_{i=2}^{L} \mathbf{q}(i)^{2/3} \right)^3} \qquad (3.37)
\end{aligned}$$

where we used that $\mathbf{q}(i) \leq \alpha^3$ for all $i \geq L+1$ in the second inequality. Applying Theorem 3.7 (along with the similar arguments as in the proof of Theorem 3.8), what this shows is a lower bound of

$$\Omega\left(\frac{\left(\sum_{i=2}^{L} \mathbf{q}(i)^{2/3} \right)^{3/2}}{\varepsilon^2} \right) \tag{3.38}$$

samples to test identity to $\mathbf{q}$, where $L = L(\mathbf{q}, \varepsilon)$ is defined above, and $\mathbf{q}$ satisfies $\|\mathbf{q}\|_\infty \leq 1/2$ and is (without loss of generality) assumed non-decreasing. This bound can seem quite daunting, due to the way $L(\mathbf{q}, \varepsilon)$ is defined; fortunately, we can relax it a little to make it more interpretable (though technically looser).

Observe that Eq. (3.36) also implies $\sum_{i=L+1}^{k} \mathbf{q}(i) \leq 4\varepsilon$; thus, if we define $K = K(\mathbf{q}, \varepsilon)$ as the largest integer such that $\sum_{i=K}^{k} \mathbf{q}(i) > 4\varepsilon$, we are guaranteed that $K(\mathbf{q}, \varepsilon) \leq L(\mathbf{q}, \varepsilon)$ and therefore $\sum_{i=2}^{K} \mathbf{q}(i)^{2/3} \leq \sum_{i=2}^{L} \mathbf{q}(i)^{2/3}$. This leads to the following corollary:

Theorem 3.9. Given a probability distribution $\mathbf{q} \in \Delta_k$, let $\tilde{\mathbf{q}}_{-4\varepsilon}^{-\max}$ denote the vector obtained by seeing $\mathbf{q}$ as a vector in $[0,1]^k$, and

removing (i) its largest entry, and (ii) its smallest entries, stopping just before the total removed exceeds 4ε. Then every testing algorithm for identity with reference $\mathbf{q}$ must have sample complexity $n(k, \varepsilon, 1/3) = \Omega(\|\tilde{\mathbf{q}}_{-4\varepsilon}^{-\max}\|_{2/3}/\varepsilon^2)$, provided that $\|\mathbf{q}\|_\infty \leq 1/2$ and $\varepsilon \in (0, 1/8]$.

One can check that this does retrieve the $\Omega(\sqrt{k}/\varepsilon^2)$ testing lower bound both when (1) $\mathbf{q}$ is the uniform distribution, (2) $\mathbf{q}$ is the Zipf distribution we saw in our earlier example.

E: Check it! (Exercise 3.4)

Finally, it is worth mentioning that Theorem 3.7 is not restricted to identity testing (although this is the application we detailed in this section). One can use it to establish indistinguishability results for other questions, focusing on the sample complexity lower bound it provides (Eq. (3.29)) and possibly ignoring the distance part (Eq. (3.30)). For instance, it can be used to establish sample complexity lower bounds for testing in other norms than total variation distance, or to obtain lower bounds on estimating some parameter of interest.

3.5 Proving hardness by reductions

In this section, we will shift focus a little, and discuss how to *leverage work (other) people did* in order to establish new sample complexity lower bounds. To illustrate the idea, consider the property $\mathcal{P}^{\searrow} = \cup_{k=1}^{\infty} \mathcal{P}_k^{\searrow}$ of *monotone* (non-increasing) distributions, *i.e.*,

$$\mathcal{P}_k^{\searrow} := \{\, \mathbf{p} \in \Delta_k : \mathbf{p}(1) \geq \mathbf{p}(2) \geq \cdots \geq \mathbf{p}(k) \,\} \tag{3.39}$$

Say we want to test this property $\mathcal{P}^{\searrow}$: that is, given samples from a distribution over $\{1, 2, \ldots, k\}$, we want to test whether its pmf is non-increasing; more specifically, say we want to show it is not easy. We could try to use techniques from the previous subsections (though probably not directly those from Section 3.3, as $\mathcal{P}^{\searrow}$ is most definitely not a symmetric property) to establish a sample complexity lower bound: this will require quite a bit of thinking and at least of few non-trivial computations.

But we have already shown (in a few different ways by now) that *uniformity* testing had sample complexity $\Omega(\sqrt{k}/\varepsilon^2)$. Can we somehow show that testing monotonicity is *at least as hard*, and get the same lower

bound without working much more? Enters the concept of *reduction*, quite central to (theoretical) computer science. We already saw it in Section 2.2.3, when we used it to show how to use any uniformity testing algorithm for the more general problem of identity testing: that is, we described a reduction from uniformity to identity testing to get an algorithm (upper bound) for the latter, using an algorithm for the former. But every coin has two sides, and a reduction can be used to show lower bounds too!

Here's how. Imagine we had an "efficient" reduction from uniformity testing to monotonicity testing: given samples from an arbitrary distribution $\mathbf{p} \in \Delta_k$, we can obtain samples from a distribution $\Phi(\mathbf{p}) \in \Delta_{k'}$ such that (i) $\Phi(\mathbf{p}) \in \mathcal{P}^{\searrow}_{k'}$ whenever $\mathbf{p} = \mathbf{u}_k$, and $\mathrm{d}_{\mathrm{TV}}\left(\Phi(\mathbf{p}), \mathcal{P}^{\searrow}_{k'}\right) > \varepsilon'$ whenever $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) > \varepsilon$. We allow the domain to change a little, from k to some (not much larger) k'; and similarly for the distance parameter, originally ε, which can become some other (not much smaller) value ε'. See Fig. 3.2 for a depiction.

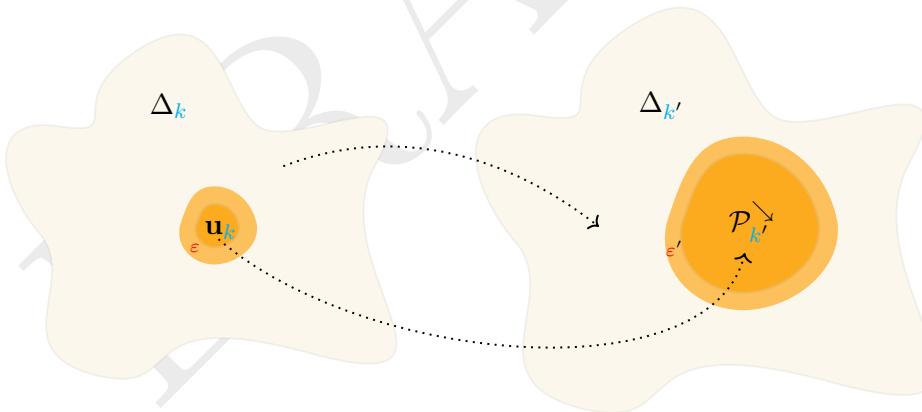

Figure 3.2: Illustration of what a reduction from uniformity testing to monotonicity testing is. The uniform distribution is mapped to some monotone distribution, while distributions far from uniform are mapped to distributions far from monotone. (Things that are neither uniform nor far from it can be mapped to anything.)

Then, any algorithm $\mathcal{T}$ for testing $\mathcal{P}^{\searrow}$ (with parameters k', ε') can be used to test uniformity (with parameters k, ε): convert the samples from

$\mathbf{p} \in \Delta_k$ to samples from $\Phi(\mathbf{p}) \in \Delta_{k'}$, run $\mathcal{T}$ on them, and output what this tester returns. Which is great... except that we know that $\Omega(\sqrt{k}/\varepsilon^2)$ samples are required for the latter task; so the sample complexity n of $\mathcal{T}$ must satisfy $n(k', \varepsilon', 1/3) = \Omega(\sqrt{k}/\varepsilon^2)$. We get a lower bound! Whose meaningfulness, of course, depends a lot on how k' and ε' are related to k and ε.

Let us look at an example, to make things more concrete. Consider the following (randomized) mapping $\Psi\colon [k] \to [2k]$:

Ψ: Given $i \in [k]$, return i with probability $\frac{1}{2}$ and $k + i$ otherwise.

Applying this to a sample from a probability distribution $\mathbf{p} \in \Delta_k$ results in a sample from $\Phi(\mathbf{p})$ over $k' := 2k$ (where $\Phi\colon \Delta_k \to \Delta_{k'}$), given by

$$\Phi(\mathbf{p})(i) = \frac{1}{2}\mathbf{p}(i \bmod k) \tag{3.40}$$

which "duplicates the domain and puts two contiguous copies of $\mathbf{p}$ next to each other." Why is it a good thing to consider? For a start, if we apply this to the uniform distribution $\mathbf{u}_k$, we end up with $\Phi(\mathbf{u}_k) = \mathbf{u}_{2k}$, the uniform distribution on our new domain, which is definitely in $\mathcal{P}_{2k}^{\searrow}$ (the uniform distribution may not be the most interesting monotone distribution, but it is a monotone distribution nonetheless). So this fits at least half the bill: as you will show in Exercise 3.6, it also fits the other half: distributions which are ε-far from uniform are mapped to distributions ε'-far from uniform, where $\varepsilon' := \varepsilon/2$.

E: Show this.

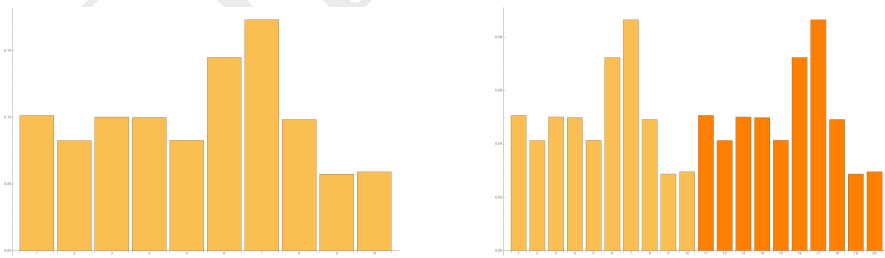

Figure 3.3: An example of the reduction outlined above, with a distribution $\mathbf{p}$ over a domain of size $k = 20$ (left) and the resulting $\Phi(\mathbf{p})$ on a domain of size $k' = 2k = 40$ on the right.

By the above discussion, the lower bound on uniformity testing directly implies that monotonicity testing with parameters $2k$ and $\varepsilon/2$

must have sample complexity $\Omega(\sqrt{k}/\varepsilon^2)$. Since constant factors are just constant factors, this leads to the following:

Theorem 3.10. Every testing algorithm for $\mathcal{P}^{\searrow}$ (monotonicity) must have sample complexity $n(k, \varepsilon, 1/3) = \Omega(\sqrt{k}/\varepsilon^2)$.

(As a side note, this is known to be tight, as least for $\varepsilon \gg \sqrt{\log k}/k^{1/4}$.) Crucially, reductions comes with another advantage: if we were to introduce another parameter (focusing on the dependence on δ, for instance), or work with constrained measurements as in Chapter 4, then as long as our reduction goes through in the setting considered, we only need to derive the uniformity testing lower bound in that setting to immediately get the corresponding lower bound for monotonicity testing as well. As said earlier, reductions are great to avoid unnecessary work!

A general result. The above reduction was quite specific to monotonicity testing: ideally, we would like more general statements, allowing us to derive as many lower bounds as possible while having to think as little as possible. In the rest of this section, we are going to describe such a result, which can be seen as some converse to the "testing-by-learning" baseline from Lemma 1.2. The high-level idea is as follows: suppose we have two properties $\mathcal{P}' \subseteq \mathcal{P}$, and we know that $\mathcal{P}'$ is hard to test. Then we want to conclude that $\mathcal{P}$, too, is hard to test – at least as hard as $\mathcal{P}'$. (In our earlier example, $\mathcal{P}'$ was uniformity, and $\mathcal{P}$ monotonicity: since the uniform distribution is monotone, the inclusion indeed holds.)

The issue with this conclusion, however, *is that it is not true.* One can come up with very simple examples showing it: taking $\mathcal{P}_k = \Delta_k$, for example (the trivial property containing all discrete distributions) and $\mathcal{P}'_k = \{\mathbf{u}_k\}$, we clearly have $\mathcal{P}'_k \subseteq \mathcal{P}_k$, yet $\mathcal{P}_k$ can be tested with exactly 0 samples – while $\mathcal{P}'_k$ requires $\Omega(\sqrt{k}/\varepsilon^2)$.

Still, in that counterexample, the property $\mathcal{P}$ itself is ginormous (all distributions!), which somewhat explains the issue: in contrast, the property $\mathcal{P}^{\searrow}$ of monotone distributions was much smaller. If we restrict the statement to "simple enough" properties, then maybe the statement will hold? *E.g.*, what about properties which can be *learned* efficiently? As we will see momentarily, this is indeed the case, albeit for a very specific sense of "learning" we first need to introduce.

Definition 3.2 (Agnostic learning)**.** A class $\mathcal{C} = \bigcup_{k=1}^{\infty} \mathcal{C}_k$ is said to be *agnostically learnable* with sample complexity $n(k, \varepsilon, \delta)$ if there is an algorithm which, given $n = n(k, \varepsilon, \delta)$ i.i.d. samples from an unknown *arbitrary* distribution $\mathbf{p} \in \Delta_k$, outputs a distribution $\hat{\mathbf{p}}$ such that

$$\mathrm{d_{TV}}(\mathbf{p}, \hat{\mathbf{p}}) \leq C \cdot \inf_{\mathbf{q} \in \mathcal{C}_k} \mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) + \varepsilon$$

with probability at least $1 - \delta$, where $C \geq 1$ is an absolute constant.

In other terms, an agnostic learning algorithm works even in the unrealizable setting (when $\mathbf{p} \notin \mathcal{C}_k$), and its output is "nearly as good" as the best candidate from $\mathcal{C}_k$ would be. The main result of this section is the following theorem, which roughly states that "if a property is easy to learn, then it is as hard to test as its hardest sub-property:"

Theorem 3.11 (Hardness by Reduction)**.** Fix any property $\mathcal{P}$, and suppose there exists $\mathcal{P}' \subseteq \mathcal{P}$ such that the following holds.

1. Agnostic learning of $\mathcal{P}$ (with a given "agnostic constant" $C \geq 1$) has sample complexity at most $n_{\mathcal{L}}(k, \varepsilon, \delta)$;
2. Testing $\mathcal{P}'$ has sample complexity at least $n_{\mathcal{T}}(k, \varepsilon, \delta)$;
3. There exists a range of parameters k, ε, δ for which learning $\mathcal{P}$ is easier than testing $\mathcal{P}'$:

$$n_{\mathcal{L}}(k, C\varepsilon, \delta) \leq \frac{1}{2} n_{\mathcal{T}}(k, 3C\varepsilon, 2\delta)$$

Then, for k, ε, δ in that range of parameters, every testing algorithm for $\mathcal{P}$ must have sample complexity at least $\frac{1}{2} n_{\mathcal{T}}(k, 3C\varepsilon, 2\delta)$.

Proof. The idea of the proof is quite simple: suppose we have a testing algorithm $\mathcal{A}$ for $\mathcal{P}$ with sample complexity $n_{\mathcal{A}}(k, \varepsilon, \delta)$, and let $\mathcal{L}$ be the agnostic learning algorithm for $\mathcal{P}$ with sample complexity $n_{\mathcal{L}}(k, \varepsilon, \delta)$ promised by Item 1. Then we can combine both to obtain a testing algorithm for $\mathcal{P}'$: but since $\mathcal{P}'$ is hard to test, this testing algorithm cannot be *too* sample-efficient (it must take at least $n_{\mathcal{T}}$ samples), giving us a lower bound $n_{\mathcal{A}} + n_{\mathcal{L}} \geq n_{\mathcal{T}}$. (Note that we only care about

sample complexity, and do not make any assumption on computational efficiency.)
In more detail, consider the following testing algorithm $\mathcal{A}'$ for $\mathcal{P}'$: on input k, ε, δ,

- Run $\mathcal{A}$ with parameters $k, \varepsilon/(3C), \delta/2$; let $b \in \{0,1\}$ be the result.
- Run $\mathcal{L}$ with parameters $k, \varepsilon/3, \delta/2$; let $\hat{\mathbf{p}}$ be the output.
- Check whether $\mathrm{d}_{\mathrm{TV}}(\hat{\mathbf{p}}, \mathcal{P}') \leq \varepsilon/3$; let $b' \in \{0,1\}$ indicate the result (this is purely computational and requires no samples from $\mathbf{p}$).
- Return $b \wedge b'$ (*i.e.*, 1 if, and only if, both b and b' were 1).

Since both $\mathcal{A}$ and $\mathcal{L}$ were run with error probability $\delta/2$, by a union bound they are simultaneously correct with probability at least $1 - \delta$. Assume hereafter this event holds.

- If $\mathbf{p} \in \mathcal{P}'$, then *a fortiori* $\mathbf{p} \in \mathcal{P}$, and $\mathcal{A}$ returns $b = 1$. Moreover, $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \hat{\mathbf{p}}) \leq C \cdot 0 + \varepsilon/3 = \varepsilon/3$, so $b' = 1$ as well, and overall $\mathcal{A}'$ returns 1.
- Let us now argue the "soundness." Suppose that $\mathcal{T}'$ returns 1: this means that $b = 1$, and so since $\mathcal{A}$ is a testing algorithm for $\mathcal{P}$ we must have $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}) \leq \varepsilon/(3C)$. But then, since $\mathcal{L}$ is an agnostic learner for $\mathcal{P}$, we get $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \hat{\mathbf{p}}) \leq C \cdot \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}) + \varepsilon/3 \leq 2\varepsilon/3$. And finally, since the last check was successful as well ($b' = 1$), we have that $\mathrm{d}_{\mathrm{TV}}(\hat{\mathbf{p}}, \mathcal{P}') \leq \varepsilon/3$. By the triangle inequality, it follows that
$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}') \leq \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \hat{\mathbf{p}}) + \mathrm{d}_{\mathrm{TV}}(\hat{\mathbf{p}}, \mathcal{P}') \leq \varepsilon\,.$$
By contrapositive, if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}') > \varepsilon$ then it must be the case that $\mathcal{T}'$ returns 0.

So $\mathcal{T}'$ is a *bona fide* testing algorithm for $\mathcal{P}'$, and its sample complexity is
$$n'(k, \varepsilon, \delta) = n_{\mathcal{A}}(k, \varepsilon/(3C), \delta/2) + n_{\mathcal{L}}(k, \varepsilon/3, \delta/2)$$
or, reparameterizing,
$$n'(k, 3C\varepsilon, 2\delta) = n_{\mathcal{A}}(k, \varepsilon, \delta) + n_{\mathcal{L}}(k, C\varepsilon, \delta)\,.$$

But since we know by Item 2 that $\mathcal{P}'$ is hard to test, we must then have

$$n_{\mathcal{A}}(k, \varepsilon, \delta) + n_{\mathcal{L}}(k, C\varepsilon, \delta) \geq n_{\mathcal{T}}(k, 3C\varepsilon, 2\delta)$$

which by Item 3 implies $n_{\mathcal{A}}(k, \varepsilon, \delta) \geq n_{\mathcal{T}}(k, 3C\varepsilon, 2\delta)$, and proves the theorem. □

With this theorem in hand, proving a sample complexity lower bound for a given property $\mathcal{P}$ boils down to scouring the literature to check if (1) $\mathcal{P}$ is easy to learn, and (2) something (anything!) inside $\mathcal{P}$ is known to be hard to test. Note that the result may not always be an *optimal* lower bound; but it often gets quite close to it, and is a simple and valuable starting point.

To conclude, let us see a direct application of this theorem to the property $\mathcal{P}^{\searrow}$. We can use the fact that monotone distributions can be agnostically learned with $O(\log(1+\varepsilon k)/\varepsilon^3)$ samples (Birgé, 1987; Daskalakis *et al.*, 2014) along with our uniformity testing lower bound (taking $\mathcal{P}'_k := \{\mathbf{u}_k\}$). This lets us derive the same $\Omega(\sqrt{k}/\varepsilon^2)$ sample complexity lower bound for testing $\mathcal{P}^{\searrow}$ as in Theorem 3.10, with the additional (small) restriction $\varepsilon \gg (\log k)/\sqrt{k}$. Here again, we relied on the uniformity testing lower bound: it is worth pointing out that we can (and sometimes must) use other "hard-to-test" sub-properties than uniformity! We will see such an example in Exercise 3.7, where you will be asked to prove a lower bound on testing "Poisson Binomial Distributions."

3.6 Historical notes

The contents of Section 3.1 (mostly) follow the exposition of Pollard (2003) of the celebrated work of Le Cam (*e.g.*, Le Cam (1973)). The χ^2 method described (Lemma 3.1) is, to the best of the author's knowledge, due to Ingster, and often referred to as the Ingster–Suslina method after (Ingster and Suslina, 2003). Section 3.2 (mostly) follows the Fano-based framework of Diakonikolas and Kane (2016), with some simplifications and (hopefully not) the occasional new typo.

The results discussed in Section 3.3 are due to Valiant (2011); those from Section 3.4 first appeared in Valiant and Valiant (2014), before the journal version (Valiant and Valiant, 2017).

The reduction-based method from Section 3.5 was introduced in Canonne *et al.* (2016) (the journal version appearing later as (Canonne *et al.*, 2017)), which also includes a variant for tolerant testing. This is not the only type of reduction-based method known; for a different flavour entirely (reduction from communication complexity), the reader is encouraged to consult Blais *et al.* (2019).

Importantly, all methods and results covered in this chapter focused on minimax *testing* lower bounds. While they can also be used to establish sample complexity lower bounds for estimation questions (since, essentially, "learning is harder than testing," as briefly discussed in Chapter 1), there are many situations where the bounds obtained by these methods will not be optimal for estimation tasks. The reader interested in estimation *versus* testing lower bound methods is referred to the foundational paper of Yu (1997).

3.7 Exercises

Exercise 3.1. Combine (the second part of) Lemma B.4 with (the first part of) Lemma B.5 to obtain Eq. (3.14) from Eq. (3.7). Use it to derive Theorem 3.3.

Exercise 3.2. Let $\mathbf{p}$ be a measure (not necessarily a probability measure) such that the ℓ_1 distance between $\mathbf{p}$ and $\mathcal{P}_k$ satisfies $\ell_1(\mathbf{p}, \mathcal{P}_k) > 2\varepsilon$, and $1/2 \leq \|\mathbf{p}\|_1 \leq 3/2$. Defining $\mathbf{p}' := \mathbf{p}/\|\mathbf{p}\|_1$ (an actual probability distribution), provide a lower bound on $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}_k)$. Moreover, show that obtaining n "samples" from the Poisson process with measure $\mathbf{p}$ is equivalent to getting $\mathrm{Poisson}(n\|\mathbf{p}\|_1)$ samples from the distribution $\mathbf{p}'$.

Conclude with how one could use a testing algorithm $\mathcal{A}$ for property $\mathcal{P}_k$ given $\mathrm{Poisson}(n)$ samples (*i.e.*, in the Poissonized sampling model) to distinguish between two families of measures (yes- and no-instances) far in ℓ_1 distance, thus justifying the relaxed assumption from Section 3.2.

Exercise 3.3 (⋆). Recall that we defined the no-instances in Section 3.2 by Eq. (3.17) (measures, instead of *bona fide* probability measures) in order to guarantee mutual independence of $N_1, \ldots, N_k$ (conditioned on $\mathbf{b}$. Check the argument to see which part of the argument would fail if we had used Eq. (3.11) instead. Then, modify the argument to fix this, and

obtain the same sample complexity lower bound. *(Hint: we still have mutual independence of the* $k/2$ *random variables* $(N_1, N_2), \ldots, (N_{k-1}, N_k)$ *conditioned on* $\mathbf{b}$*. Establish the analogue of Eq.* (3.20) *with* $N_1 = j, N_2 = \ell$ *instead of* $N_1 = j$*, and proceed from there.*

Exercise 3.4. Verify that applying Theorem 3.9 to (i) the uniform distribution $\mathbf{u}_k$ and (ii) the "Zipf" distribution $\mathbf{q} \in \Delta_k$ such that $\mathbf{q}(i) \propto 1/\sqrt{i}$ leads, in both cases, to an $\Omega(\sqrt{k}/\varepsilon^2)$ sample complexity lower bound for identity testing.

Exercise 3.5. Check that you can express several of the algorithms in Section 2.1 as a function of F only (as defined in Section 3.3). Specifically, verify this for Algorithms 1, 2 and 4. Verify this also for Algorithm 3, keeping in mind that this algorithm was stated and analyzed in the Poissonized setting: what does it change?

Exercise 3.6 ($\star$)**.** Prove that the mapping Φ defined in Eq. (3.40) does satisfy the requirements of a reduction, for $k' = 2k$ and $\varepsilon' = \varepsilon/2$. That is, if $\mathbf{p} \in \Delta_k$ is ε-far from $\mathbf{u}_k$, then $\Phi(\mathbf{p}) \in \Delta_{2k}$ is ε'-far from every distribution $\mathbf{q} \in \mathcal{P}_{2k}^{\searrow}$. *(Hint: for any given monotone* $\mathbf{q}$*, analyse the distance* $\mathrm{d}_{\mathrm{TV}}(\Phi(\mathbf{p}), \mathbf{q})$ *according to whether* $\mathbf{q}(k) > 1/(2k)$ *or not, relating this to the set* $S \subseteq [k]$ *on which* $\mathbf{p}$ *is greater than* $\mathbf{u}_k$*.)* Moreover, show that this loss by a factor $1/2$ in the distance is necessary.

Exercise 3.7. A *Poisson Binomial Distribution* (PBD) with parameters k and $\vec{p} = (p_1, \ldots, p_k)$ is the distribution of the sum of k independent Bernoulli random variables $X_1, \ldots, X_k$, where $X_i \sim \mathrm{Bern}(p_i)$. (This is a generalization of Binomial distributions, which correspond to $p_1 = \cdots = p_k$.) Let $\mathcal{P}_k^{\Join}$ denote the class of all PBDs with parameter k. Using the facts that (1) $\mathcal{P}_k^{\Join}$ can be agnostically learned with $O(\log^2(1/\varepsilon)/\varepsilon^2)$ samples (independent of k) (Daskalakis *et al.*, 2015), and (2) the "standard" Binomial distribution $\mathrm{Bin}(k, 1/2)$ is a PBD, show that testing $\mathcal{P}_k^{\Join}$ has sample complexity $\Omega(k^{1/4}/\varepsilon^2)$ (as long as $\varepsilon \geq 1/2^{O(k^{1/8})}$). *(Hint: combine the results of Sections 3.4 and 3.5.)*

4

Testing with Constrained Measurements

To conclude this survey, we will venture outside the usual sampling setting, and consider the following question: *what happens when the algorithm does not get to see the n i.i.d. samples?*

This may seem absurd at first: well, then, the algorithm is in trouble, isn't it? Yet, this type of question does in fact capture many natural (or interesting) settings. Among others:

Communication constraints: The data is divided among n users, each holding a single sample[1] (observation), and a central server seeks to perform the testing task. Unfortunately, the users each have a stringent *bandwidth constraint*, preventing them from sending their full data point to the server: instead, they are limited to only send ℓ of information.

Limited measurements: Data is hard to measure, and physical devices (or social incentive mechanisms) are imperfect or restricted. For instance, it may only be possible to perform a specific type of one-bit measurement to each data point: fix a threshold, and

[1] One could also, of course, consider scenarios where users hold multiple samples each; it is, however, a little more complicated to handle.

only learn whether the value is greater. Or it may be the case that sensors can be very accurate either for higher temperatures, or lower ones, but not both: which ones to choose to deploy?

Quantization: Very often, the underlying signal is continuous, but the measurement is intrinsically discrete. Which quantization scheme to choose? Should it be chosen once and for all, or should various quantization schemes be combined, different for distinct measurements?

Privacy: Sometimes the data is not only distributed across many users, but also sensitive: *e.g.*, medical data, location information, or financial records. The users, while willing to send some information to the central server in order to perform the testing task, seek to preserve the privacy of their personal data. This is captured, *e.g.*, by the framework of (local) differential privacy (Dwork *et al.*, 2006; Kasiviswanathan *et al.*, 2011; Duchi *et al.*, 2013), which guarantees (in a formal sense) that nobody – even the central server – can infer too much about any single user's data point.

Streaming and memory-limited devices: In some cases, the measurements are performed (or the data observed) sequentially by a device with limited working memory. The algorithm thus cannot store the totality of the dataset before performing computations on it, but instead must maintain a small "sketch" of the samples seen so far, and base its final output on this sketch only.

Noisy channels: We conclude this (non-exhaustive) list with the example of settings where the measurements are performed locally, and sent to a central entity for processing through a noisy communication channel. Each such transmission can be subject to data corruption, either through random noise or adversarially.

This is a *short* concluding chapter, and we will not cover all (or, indeed, most) of the above applications; the interested reader is referred to Section 4.4 for a few pointers. Instead, we will focus on the first setting, that of communication constraints, where each of n users observes a single (independent) sample from the same unknown distribution $\mathbf{p} \in \Delta_k$, but

must abide by a tight communication budget of $\ell \ll \log k$ bits. We will also focus on upper bounds (algorithms), and on our by-now-familiar example of uniformity testing.

4.1 Setting(s), and the devil lurking in the details

Before doing anything else, it is important to define – and discuss – this communication-limited setting. As mentioned, we have n i.i.d. samples $X_1, \dots, X_n$ drawn from the same unknown distribution $\mathbf{p} \in \Delta_k$. These samples are distributed among n users, where user i observes sample X_i and, from it, computes an ℓ-bit message $Y_i \in \{0,1\}^\ell$ and sends it to the central server.

Upon observing these n messages $Y_1, \dots, Y_n$, the central server (which does not have itself any sample from $\mathbf{p}$) runs an algorithm to test whether $\mathbf{p}$ is uniform, or ε-far from it, and must be correct with probability $1 - \delta$.

Note that we will implicitly assume afterwards that $\ell \leq \log k$,[2] since otherwise each user can simply send their full sample (which only takes $\log k$ bits to encode) to the server, and we are back in our familiar setting, with a tight $\Theta(\sqrt{k}/\varepsilon^2)$ sample complexity bound for uniformity testing. Given this new communication constraint, we know that the new bound on n, which is both the number of users and the number of samples, will be *at least* $\Theta(\sqrt{k}/\varepsilon^2)$; but, most likely, higher (and should depend on ℓ).

But what do the users know, exactly? We assume that the users are not given the parameters ε, δ, but know k: since they perform the measurement, they probably are aware of the domain of the data. They also know the details the protocol they must follow to compute the message to send (so, in particular, we can assume they know the total number of users n, if needed), and are given a way to identify themselves in this protocol (*i.e.*, each user has a unique ID).

[2] To avoid this implicit assumption, we could just replace ℓ by $\ell \wedge \log k$ in every statement, but that is neither nice to read nor to write.

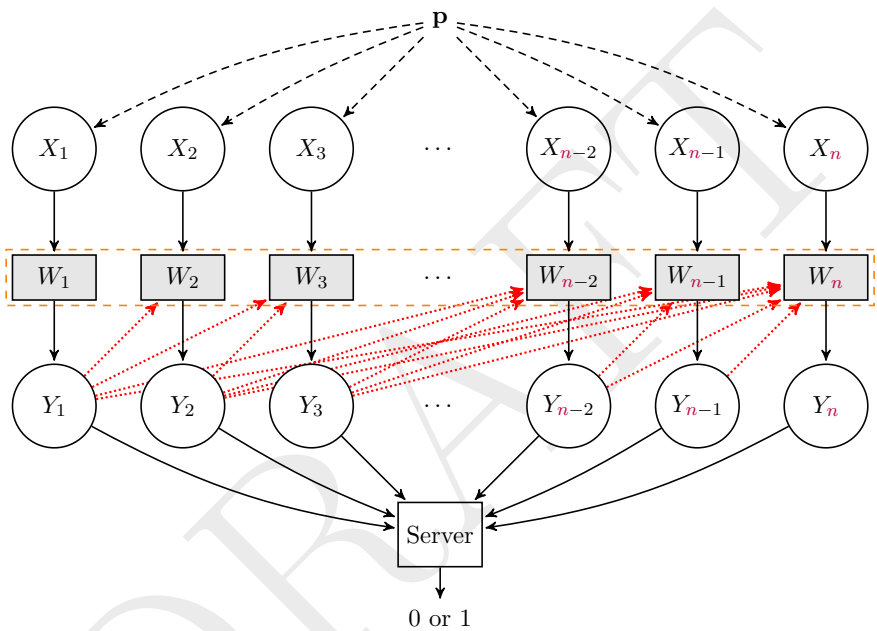

Figure 4.1: Depiction of the communication-constrained setting, in its (almost) full, glorious generality. The orange dashed box highlights that, in the public-coin setting, the users can jointly randomize their messages even though they do not directly communicate. The red dotted arrows indicate that, in the (sequentially) interactive setting, user i observes the messages $Y_1, \ldots, Y_{i-1}$, and can choose their own message Y_i based on those (as well as X_i).

But what do the users share, exactly? Now, we said that user i (for $1 \leq i \leq n$) computes their message $Y_i \in \{0,1\}^\ell$ from X_i. Let us make it a bit more formal: user i is equipped with a (possibly randomized) function $W_i \colon \mathcal{X} \to \{0,1\}^\ell$, which is decided ahead of time as part of the overall protocol the users and server follow; and sets $Y_i := W_i(X_i)$.

Note that we allow W_i to differ across users: we do not require that they all use the same mapping from observation to messages. We also allow it to be randomized, which, as we will see, can be quite useful: but this raises the question of *which random seed is used*. Are $W_1, \ldots, W_n$ randomized independently (*i.e.*, each W_i has its own, "private" random seed R_i, independent of both the inputs $X_1, \ldots, X_n$ and of the other R_j's)? Are they randomized jointly (*i.e.*, each W_i has its own, "private" random seed R_i as before, *and* a shared random seed U which all users and the server observe – still independent of the inputs $X_1, \ldots, X_n$, of course)? Or do we go even further, and do we allow W_i to depend on the messages previously sent, that is, we allow user i to observe $Y_1, \ldots, Y_{i-1}$ before sending their own message?

The answer is: any of the above, choose your own adventure. Each of the 3 settings above captures a different scenario, and has its own pros and cons:

- Independent randomization (no common random seed, only private randomness), and no seeing previous messages: this is the *private-coin* setting, sometimes called private-coin "simultaneous message-passing" (SMP). It is possibly the simplest to implement, which is a clear advantage; however, it is also the most restrictive, and thus we can expect the sample complexity to be the worst in this setting.

- Joint randomization (common random seed available, on top of private randomness), and no seeing previous messages: this is the *public-coin* setting. It makes sense in scenarios where the server can broadcast a message to all users, or when some earlier synchronization between devices has been performed ahead of time. More permissive than the private-coin setting, so we can hope to achieve better sample complexity.

- *Sequentially interactive:* common random seed available, on top of private randomness, *and* users get to see the messages sent by users before them. This is the most challenging to implement, and can come at the cost of delays and latencies; still, it might also allow for better sample complexity, so... maybe things balance out?

There are even more permissive settings (*e.g.*, the so-called "blackboard model," also known as *tree protocols*), but this is already quite a lot to absorb. The three settings discussed above are depicted (with more or less success) in Fig. 4.1.

But what can the users *achieve*, exactly? In the next section, Section 4.2, we will see a simple, yet powerful technique, *simulate-and-infer*, which leads to a $O(k^{3/2}/2^{\ell}\varepsilon^2)$ sample complexity for uniformity testing in the private-coin setting (Theorem 4.1). Viewed differently, this is

$$\underbrace{\frac{k}{2^{\ell}}}_{\substack{\text{Cost of distributed}\\ \text{setting}}} \cdot \underbrace{\frac{\sqrt{k}}{\varepsilon^2}}_{\substack{\text{Cost in the}\\ \text{centralized setting}}} \tag{4.1}$$

and happens to be optimal (no private-coin protocol for uniformity testing can do better, as a function of k, ε, ℓ). As a sanity check, the first factor disappears when $\ell = \log k$, which is comforting.

We will then see that, somewhat suprisingly, *public randomness helps a lot for testing*. In Section 4.3, using the *domain compression* technique introduced in Section 2.1.6, we will see that public-coin protocols can achieve the much better $O(k/2^{\ell/2}\varepsilon^2)$ sample complexity for uniformity testing; or, equivalently,

$$\underbrace{\sqrt{\frac{k}{2^{\ell}}}}_{\substack{\text{Cost of distributed}\\ \text{setting}}} \cdot \underbrace{\frac{\sqrt{k}}{\varepsilon^2}}_{\substack{\text{Cost in the}\\ \text{centralized setting}}} \tag{4.2}$$

What is perhaps even more surprising is that this is also optimal, *even* when one allows sequentially interactive protocols! So, public randomness helps; but interaction? Not so much.

Oh, one last detail: to make things slightly more confusing: since we are in a distributed setting, we now talk about testing *protocols*, not algorithms.

4.2 Simulate-and-Infer

We will focus here on proving the upper bound of the following theorem (the lower bound, unfortunately, would probably require another chapter, and more coffee than the author currently has at his disposal):

Theorem 4.1. There exists a private-coin testing protocol for uniformity under ℓ-bits communication constraints using $n(k, \varepsilon, \ell, 1/3) = O(k^{3/2}/(2^\ell \varepsilon^2))$ users. Moreover, this is optimal among all such private-coin protocols.

To establish this, we will describe a very simple technique, *simulate-and-infer*, which essentially allows users to simulate, *via* a distributed private-coin protocols, honest-to-goodness new i.i.d. samples from the actual (unknown) distribution $\mathbf{p}$, even though none of them has the communication budget to send their actual samples. Now, if it takes $k/2^\ell$ users to generate a single new sample at the server, then we can just repeat this in parallel on disjoint groups of users, until the server has enough samples to run they favorite non-distributed uniformity testing algorithm from Section 2.1. Hence the name of the technique: simulate samples, *then* infer from them... which leads to the sample complexity bound of Eq. (4.1) (and Theorem 4.1).

Let us formally state what this "simulation" technique is about.

Theorem 4.2 (Distributed Simulation)**.** For any $1 \leq \ell \leq \log k$, there exists a private-coin protocol which lets the server simulate an expected $n' \asymp n2^\ell/k$ i.i.d. samples from an unknown distribution $\mathbf{p} \in \Delta_k$, given the ℓ-bit messages from n users, each holding an independent sample from this $\mathbf{p}$.

We will not prove this theorem in detail, but instead give the main idea, starting with the case $\ell = 1$, showing how to generate *one* sample from $n = 2k$ users. The generalisation to $\ell \geq 2$ is then a little fastidious, but relatively straightforward: see Exercise 4.3.

Start by partitioning these $2k$ users in pairs: say, users $2i-1$ and $2i$, for $1 \le i \le k$. Pair i will be "assigned" element i of the domain, and the one-bit message they will send are just the indicators

$$Y_{2i-1} := \mathbb{1}\{X_{2i-1} = i\}, \qquad Y_{2i} := \mathbb{1}\{X_{2i} = i\}$$

of whether their respective sample fell on their assigned element i. The server, upon receiving these $n = 2k$ messages, will check the following two conditions:

- there exists one, and only one, pair $(2i-1, 2i)$ of users for which the "even" user sent 1 ($Y_{2i} = 1$); and
- for this pair $(2i-1, 2i)$, the "odd" user sent 0 ($Y_{2i-1} = 1$).

If those two conditions do not simultaneously hold (either at least two even users sent 1, or the odd user from the pair sent 1 as well), then the server aborts (does not output any sample, but outputs, say, $\perp$ instead). Otherwise, the server outputs i as its sample. This procedure may seem arbitrary, but it is then not too hard to check that the probability that $i \in [k]$ is output is then given by

$$\Pr[\text{output is } i] = \mathbf{p}(i) \prod_{\substack{1 \le j \le k \\ j \ne i}} (1-\mathbf{p}(j)) \cdot (1-\mathbf{p}(i)) = \mathbf{p}(i) \prod_{j=1}^{k} (1-\mathbf{p}(j)) \quad (4.3)$$

where (a) the first term, $\mathbf{p}(i)$, is the probability that user $2i$ sends 1, (b) the second term, $\prod_{j \ne i}(1-\mathbf{p}(j))$, is the probability that no other user $2j$ sends 1, and (c) the last term, $1-\mathbf{p}(i)$, is the probability that user $2i-1$ sends 0.

Squinting a bit at Eq. (4.3), we see that $\Pr[\text{output is } i] \propto \mathbf{p}(i)$ for every $i \in [k]$, since $\prod_{j=1}^{k}(1-\mathbf{p}(j))$ does not depend on i. This is encouraging! This means that, conditioned on outputting *something*, the server outputs a sample from the right distribution.

To conclude, it only remains to show that the probability to output *something* (which, by summing Eq. (4.3) over $i \in [k]$, is exactly this quantity $\prod_{j=1}^{k}(1-\mathbf{p}(j))$) is not too bad, say, at least a constant. So we need a good lower bound: using the rabbit-out-of-a-hat inequality

$1 - u \geq 1/4^u$ (which holds for $0 \leq u \leq 1/2$), we can write

$$\prod_{j=1}^{k} (1 - \mathbf{p}(j)) \geq \prod_{j=1}^{k} \frac{1}{4^{\mathbf{p}(j)}} = \frac{1}{4} \tag{4.4}$$

as long as $\|\mathbf{p}\|_\infty \leq 1/2$. *Which we do not know.* But we can enforce it, using private randomness from each user, and a factor 2 in the number of users: namely, start by mapping $\mathbf{p} \in \Delta_k$ to $\Phi(\mathbf{p}) \in \Delta_{2k}$, via the simple randomized mapping Ψ which, on input $i \in [k]$, returns either i or $i + k$, each with probability $1/2$. This only needs private randomness from each user, preserves the total variation distances, and does not require any knowledge of $\mathbf{p}$; but now, $\|\Phi(\mathbf{p})\|_\infty = \|\mathbf{p}\|_\infty/2 \leq 1/2$, so the above argument goes through – only replacing k by $2k$ (and so, $n = 2k$ users by $4k$ users).

What we showed can be summarized as follows:

Lemma 4.3 (Distributed Simulation, Baby Version)**.** There exists a private-coin protocol which lets the server simulate an expected $n' \geq \frac{1}{4}\lfloor \frac{n}{4k} \rfloor$ i.i.d. samples from an unknown distribution $\mathbf{p} \in \Delta_k$, given the 1-bit messages from n users, each holding an independent sample from this $\mathbf{p}$.

4.3 Random hashing and domain compression

We have now seen a general technique to obtain private-coin protocols under communication constraints. What about *public-coin* protocols? How do we take advantage of this common random seed the users have access to? Ironically, the answer to this can be found nearly a hundred pages ago, in Section 2.1.6; and, more specifically, Theorem 2.12, which will allow us to establish (the upper bound of) the following result:

Theorem 4.4. There exists a public-coin testing protocol for uniformity under ℓ-bits communication constraints using $n(k, \varepsilon, \ell, 1/3) = O(k/(2^{\ell/2}\varepsilon^2))$ users. Moreover, this is optimal, even among *interactive* protocols.

Recall that the "domain compression" technique described in Theorem 2.12 lets us trade *distance* for *domain size*: namely, we can replace

distance ε but domain size k by distance $\varepsilon' \asymp \varepsilon\sqrt{L/k}$ but domain size L, for any $2 \le L \le k$ of our choosing.[3] Since each user has ℓ bits to send, it is natural to set $L := 2^\ell$: this way, they have enough communication budget to send their full induced sample from this new domain, after which the server can run, again, its favorite (non-distributed) uniformity testing algorithm from Section 2.1 on the n samples over $[L]$, with distance parameter ε'.
Doing so, what we get is a sample complexity

$$n \asymp \frac{\sqrt{L}}{\varepsilon'^2} \asymp \frac{k}{\varepsilon^2\sqrt{L}} = \frac{k}{\varepsilon^2 2^{\ell/2}} \tag{4.5}$$

which is exactly what we were after. We are done!

4.4 Historical notes

Most of the material covered or hinted at in this last chapter can be found in the sequence of papers Acharya *et al.* (2020c), Acharya *et al.* (2020d), and Acharya *et al.* (2021a), as well as (for the interactive setting specifically) in Acharya *et al.* (2022) and Acharya *et al.* (2020b). Beyond this line of work (with which, for obvious reasons, the author of this survey is quite familiar), there is an extensive recent body of work addressing this type of information or measurement constraint in the context of estimation (learning) tasks. Focusing on distribution testing, Diakonikolas *et al.* (2019a) and Amin *et al.* (2020) and Berrett and Butucea (2020) consider goodness-of-fit testing under, respectively, communication or memory constraints, and local privacy constraints.

An extension of the Domain Compression Lemma (DCL) accounting for the amount of public randomness available (length of the common randomness seed), and providing a tight trade-off with respect to this extra parameter as well was obtained in Acharya *et al.* (2020a). The DCL itself found applications beyond the constrained measurements described here, and was recently used to obtain uniformity testing algorithms under two notions of differential privacy called *pan-privacy* (Amin *et al.*,

[3] Ignoring some technical details, as this is only guaranteed to hold with constant probability and thus requires some amplification by repetition (losing a constant factor in the sample complexity), as in Section 2.1.6.

2020) and *shuffle privacy* (Balcer *et al.*, 2021; Canonne and Lyu, 2022), enabling one to easily obtain public-coin algorithms from private-coin ones.

The Simulate-and-Infer technique discussed in Section 4.2 is not specific to testing, and was also recently leveraged for nonparametric estimation (*i.e.*,, learning of continuous Besov densities) in Acharya *et al.* (2021b).

4.5 Exercises

Exercise 4.1. Verify that the error amplification technique discussed in Lemma 1.1 still goes through in the communication-constrained distributed setting.

Exercise 4.2. Verify that the reduction from identity to uniformity testing discussed in Section 2.2.3 still goes through in the communication-constrained distributed setting, both in the private- and public-coin settings. Do the users need to know the reference distribution $\mathbf{q}$?

Exercise 4.3 ($\star$). Extend the argument of Lemma 4.3 to $\ell \geq 1$, to establish the more general Theorem 4.2. *(Hint: suppose that* $2^\ell - 1$ *divides* k*, and partition the domain in* $m := k/(2^\ell - 1)$ *sets. Each pair of users know is "assigned" one of these sets.)*

Exercise 4.4 ($\star\star$). Extend the argument of Theorem 4.2 further to apply to the case where user has a communication constraint ℓ_i (heterogeneous constraints among users). Establish an analogous bound, with 2^ℓ replaced by $\frac{1}{n}\sum_{j=1}^n 2^{\ell_j}$. *(Hint: consider a dyadic partition of the domain* $[k]$*. It* should *work.)*

Acknowledgements

The author is grateful to Yuhan Liu, Aditya Vikram Singh, and Sampson Wong for their valuable comments and feedback on this survey; and to Mike Casey, for his patience during the time it took to write it.

Appendices

A

Some good inequalities

We only mention here a few good bounds that we found to be useful, and sufficient in many or most settings. There are, of course, many others, and many refinements or variants of the bounds we present here. We refer the reader to, *e.g.*, Vershynin (2018, Chapter 2) or Boucheron *et al.* (2013) for a much more comprehensive and insightful coverage.

We start with the mother of all concentration inequalities, Markov's inequality:

Theorem A.1 (Markov's inequality)**.** Let X be a non-negative random variable with $\mathbb{E}[X] < \infty$. For any $t > 0$, we have

$$\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$$

Applying this to $(X - \mathbb{E}[X])^2$, we get

Theorem A.2 (Chebyshev's inequality)**.** Let X be a random variable with $\mathbb{E}\left[X^2\right] < \infty$. For any $t > 0$, we have

$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathrm{Var}[X]}{t^2}$$

By applying Markov's inequality to the moment-generating function (MGF) of $\sum_{i=1}^{n} X_i$ in various ways, one can also obtain the following statements:

Theorem A.3 (Hoeffding bound)**.** Let $X_1, \ldots, X_n$ be independent random variables, where X_i takes values in $[a_i, b_i]$. For any $t \geq 0$, we have

$$\Pr\left[\sum_{i=1}^{n} X_i > \sum_{i=1}^{n} \mathbb{E}[X_i] + t\right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \tag{A.1}$$

$$\Pr\left[\sum_{i=1}^{n} X_i < \sum_{i=1}^{n} \mathbb{E}[X_i] - t\right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \tag{A.2}$$

Corollary A.4 (Hoeffding bound)**.** Let $X_1, \ldots, X_n$ be i.i.d. random variables taking value in $[0, 1]$, with mean μ. For any $\gamma \in (0, 1]$ we have

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \gamma\right] \leq 2\exp\left(-2\gamma^2 n\right) \tag{A.3}$$

Theorem A.5 (Chernoff bound)**.** Let $X_1, \ldots, X_n$ be independent random variables taking value in $[0, 1]$, and let $P := \sum_{i=1}^{n} \mathbb{E}[X_i]$ For any $\gamma \in (0, 1]$ we have

$$\Pr\left[\sum_{i=1}^{n} X_i > (1+\gamma)P\right] < \exp\left(-\gamma^2 P/3\right) \tag{A.4}$$

$$\Pr\left[\sum_{i=1}^{n} X_i < (1-\gamma)P\right] < \exp\left(-\gamma^2 P/2\right) \tag{A.5}$$

In particular, if $X_1, \ldots, X_n$ are i.i.d. with mean μ, then for any $\gamma \in (0, 1]$ we have

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \gamma\mu\right] \leq 2\exp\left(-\gamma^2 n\mu/3\right) \tag{A.6}$$

As a rule of thumb, the "multiplicative" (Chernoff) from Theorem A.5 is preferable to the "additive" bound (Hoeffding) from Corollary A.4 whenever $\mu := P/n \ll 1$. In case one only has an upper or lower bound on the quantity $P = \sum_{i=1}^{n} \mathbb{E}[X_i]$, the following version of the Chernoff bound can come in handy:

Theorem A.6 (Chernoff bound (upper and lower bound version))**.** In the setting of Theorem A.5, suppose that $P_L \leq P \leq P_H$. Then for any

$\gamma \in (0, 1]$, we have

$$\Pr\left[\sum_{i=1}^{n} X_i > (1+\gamma)P_H\right] < \exp\Big(-\gamma^2 P_H/3\Big) \tag{A.7}$$

$$\Pr\left[\sum_{i=1}^{n} X_i < (1-\gamma)P_L\right] < \exp\Big(-\gamma^2 P_L/2\Big) \tag{A.8}$$

Theorem A.7 (Bernstein's inequality)**.** Let $X_1, \ldots, X_n$ be independent random variables taking values in $[-a, a]$, and such that $\mathbb{E}[X_i^2] \le v_i$ for all i. Then, for every $t \ge 0$, we have

$$\Pr\left[\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}[X_i]\right| \ge t\right] \le \exp\left(-\frac{t^2}{2(\sum_{i=1}^{n} v_i + \frac{a}{3}t)}\right).$$

In particular, if $X_1, \ldots, X_n$ are i.i.d. with mean μ and $\mathbb{E}[X_1^2] \le v$, then for any $\gamma \ge 0$ we have

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \ge \gamma\right] \le \exp\left(-\frac{\gamma^2 n}{2(v + \frac{a}{3}\gamma)}\right).$$

Observe that this tail bound exhibits both behaviours: it decays in a subgaussian fashion for small γ, before switching to a subexponential tail bound for large γ.

We conclude this section by providing a very convenient bound, specifically for Poisson random variables, which shares the same "two-tail" behaviour:

Theorem A.8 (Poisson concentration)**.** Let X be a Poisson(λ) random variable, where $\lambda > 0$. Then, for any $t > 0$, we have

$$\Pr[\, X \ge \lambda + t\,] \le e^{-\frac{t^2}{2\lambda}\psi\left(\frac{t}{\lambda}\right)} \le e^{-\frac{t^2}{2(\lambda+t)}} \tag{A.9}$$

and, for any $0 < t < \lambda$,

$$\Pr[\, X \le \lambda - t\,] \le e^{-\frac{t^2}{2\lambda}\psi\left(-\frac{t}{\lambda}\right)} \le e^{-\frac{t^2}{2(\lambda+t)}}, \tag{A.10}$$

where $\psi(u) := 2\frac{(1+u)\ln(1+u)-u}{u^2}$ for $u \ge -1$. In particular, for any $t \ge 0$,

$$\Pr[\, |X - \lambda| \ge t\,] \le 2e^{-\frac{t^2}{2(\lambda+t)}}. \tag{A.11}$$

B

Metrics and divergences between probability distributions

We here focus on distributions over discrete domains; all of the stated results do extend to the continuous settings, replacing ratios by Radon–Nikodym derivatives and sums by suitable integrals.

We briefly recall the definitions of the distance measures between probability distributions we will use here. This list is by no means exhaustive, of course: there be (many more) dragons.

Definition B.1. For two probability distributions $\mathbf{p}_1, \mathbf{p}_2$ over the same domain $\mathcal{X}$, the *Kullback–Leibler divergence* (in nats), *chi–square divergence*, and *Hellinger distance* are given by

$$\mathrm{D}(\mathbf{p}_1 \| \mathbf{p}_2) = \sum_{x \in \mathcal{X}} \mathbf{p}_1(x) \ln \frac{\mathbf{p}_1(x)}{\mathbf{q}_1(x)} \tag{B.1}$$

$$\chi^2(\mathbf{p}_1 \,||\, \mathbf{p}_2) = \sum_{x \in \mathcal{X}} \frac{(\mathbf{p}_1(x) - \mathbf{q}_1(x))^2}{\mathbf{q}_1(x)} \tag{B.2}$$

$$\mathrm{d}_{\mathrm{H}}(\mathbf{p}_1, \mathbf{p}_2) = \frac{1}{\sqrt{2}} \| \sqrt{\mathbf{p}} - \sqrt{\mathbf{q}} \|_2 \,, \tag{B.3}$$

with the convention that $0 \ln 0 = 0$. Note that the first two are not symmetric, do not satisfy the triangle inequality, and are unbounded.

Importantly, TV distance, squared Hellinger, KL divergence, and chi-square divergence are all instances of *f-divergences*, which directly endows them with many desirable properties – among which joint convexity and the data-processing inequality (Fact 1.1).

Squared Hellinger, KL divergence, and chi-square divergence also "tensorize" nicely: specifically, for any product probability distributions $\mathbf{p}_1 \otimes \cdots \otimes \mathbf{p}_n$ and $\mathbf{q}_1 \otimes \cdots \otimes \mathbf{q}_n$, we have

$$\mathrm{D}(\mathbf{p}_1 \otimes \cdots \otimes \mathbf{p}_n \| \mathbf{q}_1 \otimes \cdots \otimes \mathbf{q}_n) = \sum_{i=1}^{n} \mathrm{D}(\mathbf{p}_i \| \mathbf{q}_i) \tag{B.4}$$

$$\chi^2(\mathbf{p}_1 \otimes \cdots \otimes \mathbf{p}_n \,||\, \mathbf{q}_1 \otimes \cdots \otimes \mathbf{q}_n) = \prod_{i=1}^{n} \Big(1 + \chi^2(\mathbf{p}_i \,||\, \mathbf{q}_i)\Big) - 1 \tag{B.5}$$

and

$$\begin{aligned} \mathrm{d_H}(\mathbf{p}_1 \otimes \cdots \otimes \mathbf{p}_n, \mathbf{q}_1 \otimes \cdots \otimes \mathbf{q}_n)^2 &= 1 - \prod_{i=1}^{n} (1 - \mathrm{d_H}(\mathbf{p}_i, \mathbf{q}_i)^2) \\ &\leq \sum_{i=1}^{n} \mathrm{d_H}(\mathbf{p}_i, \mathbf{q}_i)^2 \,; \end{aligned} \tag{B.6}$$

while TV distance is much less cooperative, only giving the weaker

$$\mathrm{d_{TV}}(\mathbf{p}_1 \otimes \cdots \otimes \mathbf{p}_n, \mathbf{q}_1 \otimes \cdots \otimes \mathbf{q}_n) \leq \sum_{i=1}^{n} \mathrm{d_{TV}}(\mathbf{p}_i, \mathbf{q}_i) \tag{B.7}$$

(typically much looser, loosing up to a factor $\sqrt{n}$ compared to what one would get via, say, Hellinger).

We now state (and prove) several useful lemmas relating these various distance measures.

Lemma B.1. For every $\mathbf{p}, \mathbf{q}$ on $\mathcal{X}$,

$$\mathrm{d_H}(\mathbf{p}, \mathbf{q})^2 \leq \mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) \leq \sqrt{2}\mathrm{d_H}(\mathbf{p}, \mathbf{q})\,.$$

Proof. Let us first prove the left side. Using $a-b = (\sqrt{a}-\sqrt{b})(\sqrt{a}+\sqrt{b})$,

$$\begin{aligned}
\mathrm{d_H}(\mathbf{p},\mathbf{q})^2 &= \frac{1}{2}\sum_{x\in\mathcal{X}}\left(\sqrt{\mathbf{p}(x)}-\sqrt{\mathbf{q}(x)}\right)^2 \\
&\le \frac{1}{2}\sum_{x\in\mathcal{X}}\left|\sqrt{\mathbf{p}(x)}-\sqrt{\mathbf{q}(x)}\right|\left(\sqrt{\mathbf{p}(x)}+\sqrt{\mathbf{q}(x)}\right) \\
&= \frac{1}{2}\sum_{x\in\mathcal{X}}|\mathbf{p}(x)-\mathbf{q}(x)| = \mathrm{d_{TV}}(\mathbf{p},\mathbf{q})\,.
\end{aligned}$$

For the right side, we have, by Cauchy–Schwarz and then using the identity $2(a+b) = (\sqrt{a}+\sqrt{b})^2 + (\sqrt{a}-\sqrt{b})^2$,

$$\begin{aligned}
\mathrm{d_{TV}}(\mathbf{p},\mathbf{q}) &= \frac{1}{2}\sum_{x\in\mathcal{X}}\left|\sqrt{\mathbf{p}(x)}-\sqrt{\mathbf{q}(x)}\right|\left(\sqrt{\mathbf{p}(x)}+\sqrt{\mathbf{q}(x)}\right) \\
&\le \frac{1}{2}\sqrt{\sum_{x\in\mathcal{X}}\left(\sqrt{\mathbf{p}(x)}-\sqrt{\mathbf{q}(x)}\right)^2}\sqrt{\sum_{x\in\mathcal{X}}\left(\sqrt{\mathbf{p}(x)}+\sqrt{\mathbf{q}(x)}\right)^2} \\
&= \frac{1}{\sqrt{2}}\mathrm{d_H}(\mathbf{p},\mathbf{q})\sqrt{\sum_{x\in\mathcal{X}}\left(2(\mathbf{p}(x)+\mathbf{q}(x)) - \left(\sqrt{\mathbf{p}(x)}-\sqrt{\mathbf{q}(x)}\right)^2\right)} \\
&= \mathrm{d_H}(\mathbf{p},\mathbf{q})\sqrt{2-\mathrm{d_H}(\mathbf{p},\mathbf{q})^2}\,,
\end{aligned}$$

which implies the (slightly weaker) inequality we wanted to show. □

Lemma B.2. For every $\mathbf{p},\mathbf{q}$ on $\mathcal{X}$,

$$\mathrm{d_{TV}}(\mathbf{p},\mathbf{q})^2 \le \frac{1}{4}\chi^2(\mathbf{p}\,||\,\mathbf{q})\,.$$

Proof. By Cauchy–Schwarz,

$$\begin{aligned}
\mathrm{d_{TV}}(\mathbf{p},\mathbf{q}) &= \frac{1}{2}\sum_{x\in\mathcal{X}}|\mathbf{p}(x)-\mathbf{q}(x)| \\
&\le \frac{1}{2}\sqrt{\sum_{x\in\mathcal{X}}\frac{(\mathbf{p}(x)-\mathbf{q}(x))^2}{\mathbf{q}(x)}}\sqrt{\sum_{x\in\mathcal{X}}\mathbf{q}(x)} \\
&= \frac{1}{2}\sqrt{\chi^2(\mathbf{p}\,||\,\mathbf{q})}\,.
\end{aligned}$$

□

Lemma B.3 (Pinsker's Inequality)**.** For every $\mathbf{p}, \mathbf{q}$ on $\mathcal{X}$,

$$\mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) \leq \sqrt{\frac{1}{2}\mathrm{D}(\mathbf{p}\|\mathbf{q})}\,.$$

This inequality is "good enough" for most situations; nonetheless, we state here a lesser known, but stronger result, for when it is not:

Lemma B.4 (Bretagnolles–Huber Bound)**.** For every $\mathbf{p}, \mathbf{q}$ on $\mathcal{X}$,

$$\mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) \leq \sqrt{1 - e^{-\mathrm{D}(\mathbf{p}\|\mathbf{q})}}\,. \tag{B.8}$$

In particular, as $\sqrt{1 - e^{-x}} \leq 1 - \frac{1}{2}e^{-x}$ for $x \geq 0$, this implies

$$\mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) \leq 1 - \frac{1}{2}e^{-\mathrm{D}(\mathbf{p}\|\mathbf{q})}\,. \tag{B.9}$$

We refer the reader to Canonne (2022) or Tsybakov (2009, Section 2.4.1) for a proof and discussion of this inequality, due to Bretagnolle and Huber (1978).

Lemma B.5. For every $\mathbf{p}, \mathbf{q}$ on $\mathcal{X}$,

$$\mathrm{D}(\mathbf{p}\|\mathbf{q}) \leq \ln\Big(1 + \chi^2(\mathbf{p} \,||\, \mathbf{q})\Big) \leq \chi^2(\mathbf{p} \,||\, \mathbf{q})$$

Proof. The second inequality follows from the standard convexity inequality $\ln(1 + x) \leq x$ (for $x > -1$), so it suffices to prove the first. To do so, observe that

$$\begin{aligned}
\mathrm{D}(\mathbf{p}\|\mathbf{q}) &= \sum_{x \in \mathcal{X}} \mathbf{p}(x) \ln \frac{\mathbf{p}(x)}{\mathbf{q}(x)} \\
&\leq \ln \sum_{x \in \mathcal{X}} \frac{\mathbf{p}(x)^2}{\mathbf{q}(x)} && \text{(Jensen's inequality)} \\
&= \ln\Big(1 + \chi^2(\mathbf{p} \,||\, \mathbf{q})\Big)\,,
\end{aligned}$$

where we used concavity of the logarithm. □

Note that Lemmas B.3 and B.5 together imply a weaker version of Lemma B.2, losing a factor 2.

C

Poissonization

In the usual, standard sampling setting, the algorithm is given n i.i.d. samples from a distribution $\mathbf{p} \in \Delta_k$. This is sometimes called *multinomial* sampling setting, as then the vector of counts $(N_1, \ldots, N_k)$ (where N_i is the number of times we see element $i \in [k]$ among the n samples) follows a multinomial distribution with parameters n and $(\mathbf{p}(1), \ldots, \mathbf{p}(k))$.

An unfortunate aspect of this is that those $N_1, \ldots, N_k$ are not independent: each of them is marginally a Binomial random variable, with $N_i \sim \text{Bin}(n, \mathbf{p}(i))$, but those are dependent, since for instance $N_1 + \cdots + N_k = n$.[1] In turn, this can make many computations annoying or complicated.

A possible solution to this is to work instead in the *Poissonized sampling setting*, where the algorithm is given a *random* number of samples. Specifically, the sampling process is as follows. Given an integer n,

1. Draw $N \sim \text{Poisson}(n)$;
2. Draw N i.i.d. samples $X_1, \ldots, X_N$ from $\mathbf{p}$;

[1] More specifically, the N_i's are *negatively associated*; see Definition 2.3.

3. Provide $X_1, \ldots, X_N$ to the algorithm.

Equivalently, assume we have an infinite sequence $(X_i)_{i=1}^{\infty}$ of i.i.d. samples from $\mathbf{p}$, and the algorithm is provided the first N of them, where $N \sim \text{Poisson}(n)$ and $(X_i)_{i=1}^{\infty}$ are mutually independent. We can then define a *property testing in the Poissonized setting* exactly as in Definition 1.2, except for the fact that the "sample complexity" $n(k, \varepsilon, \delta)$ is now referring to the parameter of N (the Poisson random variable which is the number of samples actually given to the algorithm).

The reasons to do this are summarized in the following fact.

Fact C.1. Fix any $\mathbf{p} \in \Delta_k$, and let $(N_1, \ldots, N_k)$ denote the vector of counts among the samples in the Poissonized sampling setting with parameter n. Then (1) for every $i \in [k]$, $N_i \sim \text{Poisson}(n\mathbf{p}(i))$, and (2) $N_1, \ldots, N_k$ are mutually independent.

Moreover, tail bounds on Poisson concentration (Theorem A.8) imply that

$$\Pr\left[\frac{n}{2} \leq N \leq \frac{3n}{2}\right] \geq 1 - 2e^{-n/12} \tag{C.1}$$

which is at least $1 - \delta$ if $n \geq 12\ln(2/\delta)$. This can be used to show the following:

Lemma C.1. Suppose there exists a tester for property $\mathcal{P}$ in the Poissonized setting with sample complexity $n^{\text{Poi}}(k, \varepsilon, \delta)$. Then there exists a tester for property $\mathcal{P}$ (in the standard sampling setting) with sample complexity $n(k, \varepsilon, \delta) = \max\left(\frac{3}{2} \cdot n^{\text{Poi}}(k, \varepsilon, \delta/2), 18\ln(4/\delta)\right)$.

We also have a converse statement:

Lemma C.2. Suppose there exists a tester for property $\mathcal{P}$ (in the standard sampling setting) with sample complexity $n(k, \varepsilon, \delta)$. Then there exists a tester for property $\mathcal{P}$ in the Poissonized setting with sample complexity $n^{\text{Poi}}(k, \varepsilon, \delta) = \max(2 \cdot n(k, \varepsilon, \delta/2), 12\ln(4/\delta))$.

These two lemmas allow use to transfer upper and lower bounds establish the Poissonized sampling setting to the standard one, and vice versa. For more on Poissonization, see, *e.g.*, Valiant (2011, Section 4.3) and references within, or Canonne (2020b, Appendix D.3).

References

Acharya, J., C. L. Canonne, Y. Han, Z. Sun, and H. Tyagi. (2020a). "Domain Compression and its Application to Randomness-Optimal Distributed Goodness-of-Fit". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by J. Abernethy and S. Agarwal. Vol. 125. *Proceedings of Machine Learning Research*. PMLR. 3–40. URL: http://proceedings.mlr.press/v125/acharya20a.html.

Acharya, J., C. L. Canonne, C. Freitag, Z. Sun, and H. Tyagi. (2021a). "Inference Under Information Constraints III: Local Privacy Constraints". *IEEE J. Sel. Areas Inf. Theory*. 2(1): 253–267.

Acharya, J., C. L. Canonne, Y. Liu, Z. Sun, and H. Tyagi. (2022). "Interactive Inference Under Information Constraints". *IEEE Trans. Inf. Theory*. 68(1): 502–516. DOI: 10.1109/TIT.2021.3123905. URL: https://doi.org/10.1109/TIT.2021.3123905.

Acharya, J., C. L. Canonne, A. V. Singh, and H. Tyagi. (2021b). "Optimal Rates for Nonparametric Density Estimation under Communication Constraints". In: *NeurIPS*. URL: https://arxiv.org/abs/2107.10078.

Acharya, J., C. L. Canonne, Z. Sun, and H. Tyagi. (2020b). "Unified lower bounds for interactive high-dimensional estimation under information constraints". *CoRR*. abs/2010.06562.

Acharya, J., C. L. Canonne, and H. Tyagi. (2020c). "Inference under information constraints I: Lower bounds from chi-square contraction". *IEEE Trans. Inform. Theory*. 66(12): 7835–7855. ISSN: 0018-9448. DOI: 10.1109/TIT.2020.3028440. URL: https://doi.org/10.1109/TIT.2020.3028440.

Acharya, J., C. L. Canonne, and H. Tyagi. (2020d). "Inference under information constraints II: Communication constraints and shared randomness". *IEEE Trans. Inform. Theory*. 66(12): 7856–7877. ISSN: 0018-9448. DOI: 10.1109/TIT.2020.3028439. URL: https://doi.org/10.1109/TIT.2020.3028439.

Acharya, J., C. Daskalakis, and G. C. Kamath. (2015). "Optimal Testing for Properties of Distributions". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett, and R. Garnett. Curran Associates, Inc. 3577–3598.

Acharya, J., Z. Sun, and H. Zhang. (2018). "Differentially Private Testing of Identity and Closeness of Discrete Distributions". In: *NeurIPS*. 6879–6891.

Aliakbarpour, M., I. Diakonikolas, and R. Rubinfeld. (2018). "Differentially Private Identity and Equivalence Testing of Discrete Distributions". In: *ICML*. Vol. 80. *Proceedings of Machine Learning Research*. PMLR. 169–178.

Amin, K., M. Joseph, and J. Mao. (2020). "Pan-Private Uniformity Testing". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by J. Abernethy and S. Agarwal. Vol. 125. *Proceedings of Machine Learning Research*. PMLR. 183–218. URL: http://proceedings.mlr.press/v125/amin20a.html.

Arnold, B. C. (1987). *Majorization and the Lorenz order: a brief introduction*. Vol. 43. *Lecture Notes in Statistics*. Springer-Verlag, Berlin. vi+122. ISBN: 3-540-96592-0. DOI: 10.1007/978-1-4615-7379-1. URL: https://doi.org/10.1007/978-1-4615-7379-1.

Balakrishnan, S. and L. Wasserman. (2018). "Hypothesis testing for high-dimensional multinomials: a selective review". *Ann. Appl. Stat.* 12(2): 727–749. ISSN: 1932-6157. DOI: 10.1214/18-AOAS1155SF. URL: https://doi.org/10.1214/18-AOAS1155SF.

Balcer, V., A. Cheu, M. Joseph, and J. Mao. (2021). "Connecting Robust Shuffle Privacy and Pan-Privacy". In: *SODA*. SIAM. 2384–2403.

Batu, T. and C. L. Canonne. (2017). "Generalized uniformity testing". In: *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*. IEEE Computer Soc., Los Alamitos, CA. 880–889.

Batu, T., E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. (2001). "Testing random variables for independence and identity". In: *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001*. 442–451.

Batu, T., L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. (2000). "Testing that distributions are close". In: *41st Annual Symposium on Foundations of Computer Science, FOCS 2000*. 189–197.

Berrett, T. and C. Butucea. (2020). "Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms". In: *NeurIPS*.

Birgé, L. (1987). "On the Risk of Histograms for Estimating Decreasing Densities". *The Annals of Statistics*. 15(3): pp. 1013–1022. ISSN: 00905364. URL: http://www.jstor.org/stable/2241812.

Blais, E., C. L. Canonne, and T. Gur. (2019). "Distribution testing lower bounds via reductions from communication complexity". *ACM Trans. Comput. Theory*. 11(2): Art. 6, 37. ISSN: 1942-3454. DOI: 10.1145/3305270. URL: https://doi.org/10.1145/3305270.

Boucheron, S., G. Lugosi, and P. Massart. (2013). *Concentration inequalities*. Oxford University Press, Oxford. x+481. ISBN: 978-0-19-953525-5. DOI: 10.1093/acprof:oso/9780199535255.001.0001. URL: https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

Bretagnolle, J. and C. Huber. (1978). "Estimation des densités: risque minimax". In: *Séminaire de Probabilités, XII (Univ. Strasbourg, Strasbourg, 1976/1977)*. Vol. 649. *Lecture Notes in Math.* Springer, Berlin. 342–363.

Canonne, C. L. (2020a). "A short note on learning discrete distributions". arXiv: 2002.11457 [math.ST].

Canonne, C. L. (2020b). *A Survey on Distribution Testing: Your Data is Big. But is it Blue? Graduate Surveys*. No. 9. Theory of Computing Library. 1–100. DOI: 10.4086/toc.gs.2020.009. URL: http://www.theoryofcomputing.org/library.html.

Canonne, C. L. (2022). "A short note on an inequality between KL and TV". arXiv: 2202.07198 [math.PR].

Canonne, C. L., I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. (2016). "Testing Shape Restrictions of Discrete Distributions". In: *Proceedings of STACS*. DOI: 10.4230/LIPIcs.STACS.2016.25. URL: https://doi.org/10.4230/LIPIcs.STACS.2016.25.

Canonne, C. L., I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. (2017). "Testing Shape Restrictions of Discrete Distributions". *Theory of Computing Systems*: 1–59. DOI: 10.1007/s00224-017-9785-6. URL: http://dx.doi.org/10.1007/s00224-017-9785-6.

Canonne, C. L., A. Jain, G. Kamath, and J. Li. (2021). "The Price of Tolerance in Distribution Testing". *CoRR*. abs/2106.13414. arXiv: 2106.13414. URL: https://arxiv.org/abs/2106.13414.

Canonne, C. L. and H. Lyu. (2022). "Uniformity Testing in the Shuffle Model: Simpler, Better, Faster". In: *SIAM Symposium on Simplicity in Algorithms (SOSA)*.

Canonne, C. L., D. Ron, and R. A. Servedio. (2015). "Testing probability distributions using conditional samples". *SIAM Journal on Computing*. 44(3): 540–616. DOI: 10.1137/130945508.

Canonne, C. L. and R. Rubinfeld. (2014). "Testing Probability Distributions Underlying Aggregated Data". In: *Proceedings of ICALP*. 283–295.

Chakraborty, S., E. Fischer, Y. Goldhirsh, and A. Matsliah. (2013). "On the Power of Conditional Samples in Distribution Testing". In: *Proceedings of ITCS*. Berkeley, California, USA: ACM. 561–580. ISBN: 978-1-4503-1859-4. DOI: 10.1145/2422436.2422497.

Chan, S., I. Diakonikolas, G. Valiant, and P. Valiant. (2014). "Optimal Algorithms for Testing Closeness of Discrete Distributions". In: *Proceedings of SODA*. 1193–1203.

Cover, T. M. and J. A. Thomas. (2006). *Elements of information theory*. Second. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. xxiv+748. ISBN: 978-0-471-24195-9.

Daskalakis, C., I. Diakonikolas, and R. A. Servedio. (2014). "Learning k-Modal Distributions via Testing". *Theory of Computing*. 10(20): 535–570. DOI: 10.4086/toc.2014.v010a020.

Daskalakis, C., I. Diakonikolas, and R. A. Servedio. (2015). "Learning Poisson Binomial Distributions". *Algorithmica*. 72(1): 316–357.

Diakonikolas, I., T. Gouleakis, D. M. Kane, J. Peebles, and E. Price. (2021). "Optimal testing of discrete distributions with high probability". In: *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*. Ed. by S. Khuller and V. V. Williams. ACM. 542–555. DOI: 10.1145/3406325.3450997. URL: https://doi.org/10.1145/3406325.3450997.

Diakonikolas, I., T. Gouleakis, D. M. Kane, and S. Rao. (2019a). "Communication and Memory Efficient Testing of Discrete Distributions". In: *COLT*. Vol. 99. *Proceedings of Machine Learning Research*. PMLR. 1070–1106.

Diakonikolas, I., T. Gouleakis, J. Peebles, and E. Price. (2018). "Sample-optimal identity testing with high probability". In: *45th International Colloquium on Automata, Languages, and Programming*. Vol. 107. *LIPIcs. Leibniz Int. Proc. Inform.* Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. Art. No. 41, 14.

Diakonikolas, I., T. Gouleakis, J. Peebles, and E. Price. (2019b). "Collision-based testers are optimal for uniformity and closeness". *Chic. J. Theoret. Comput. Sci.*: Art. 1, 21. DOI: 10.4086/cjtcs.2019.001. URL: https://doi.org/10.4086/cjtcs.2019.001.

Diakonikolas, I. and D. M. Kane. (2016). "A New Approach for Testing Properties of Discrete Distributions". In: *57th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2016*. IEEE Computer Society.

Diakonikolas, I., D. M. Kane, and V. Nikishkin. (2015). "Testing Identity of Structured Distributions". In: *Proceedings of SODA*.

Dubhashi, D. and D. Ranjan. (1998). "Balls and bins: a study in negative dependence". *Random Structures Algorithms*. 13(2): 99–124. ISSN: 1042-9832. DOI: 10.1002/(SICI)1098-2418(199809)13:2<99::AID-RSA1>3.0.CO;2-M. URL: https://doi.org/10.1002/(SICI)1098-2418(199809)13:2%3C99::AID-RSA1%3E3.0.CO;2-M.

Dubhashi, D. P. and A. Panconesi. (2009). *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, Cambridge. xvi+196. ISBN: 978-0-521-88427-3. DOI: 10.1017/CBO9780511581274. URL: https://doi.org/10.1017/CBO9780511581274.

Duchi, J. C., M. I. Jordan, and M. J. Wainwright. (2013). "Local Privacy and Statistical Minimax Rates". In: *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013*. IEEE Computer Society. 429–438.

Dwork, C., F. McSherry, K. Nissim, and A. Smith. (2006). "Calibrating noise to sensitivity in private data analysis". In: *Theory of cryptography*. Vol. 3876. *Lecture Notes in Comput. Sci.* Springer, Berlin. 265–284.

Goldreich, O. (2016). "The uniform distribution is complete with respect to testing identity to a fixed distribution". *Electronic Colloquium on Computational Complexity (ECCC)*. 23: 15. URL: http://eccc.hpi-web.de/report/2016/015.

Goldreich, O., S. Goldwasser, and D. Ron. (1998). "Property Testing and Its Connection to Learning and Approximation". *Journal of the ACM*. 45(4): 653–750.

Goldreich, O. and D. Ron. (2000). "On Testing Expansion in Bounded-Degree Graphs". *Electronic Colloquium on Computational Complexity (ECCC)*. 7(20).

Huang, D. and S. Meyn. (2013). "Generalized error exponents for small sample universal hypothesis testing". *IEEE Transactions on Information Theory*. 59(12): 8157–8181.

Ingster, Y. I. (1986). "A minimax test of nonparametric hypotheses on the density of a distribution in L_p metrics". *Teor. Veroyatnost. i Primenen.* 31(2): 384–389. ISSN: 0040-361X.

Ingster, Y. I. (1997). "Adaptive chi-square tests". *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*. 244(Veroyatn. i Stat. 2): 150–166, 333. ISSN: 0373-2703. DOI: 10.1007/BF02673632. URL: https://doi.org/10.1007/BF02673632.

Ingster, Y. I. and I. A. Suslina. (2003). *Nonparametric goodness-of-fit testing under Gaussian models*. Vol. 169. *Lecture Notes in Statistics*. Springer-Verlag, New York. xiv+453. ISBN: 0-387-95531-3. DOI: 10.1007/978-0-387-21580-8. URL: https://doi.org/10.1007/978-0-387-21580-8.

Kamath, G. and J. R. Ullman. (2020). "A Primer on Private Statistics". *CoRR*. abs/2005.00010. arXiv: 2005.00010. URL: https://arxiv.org/abs/2005.00010.

Kamath, S., A. Orlitsky, D. Pichapati, and A. T. Suresh. (2015). "On Learning Distributions from their Samples". In: *Proceedings of the 28th Conference on Learning Theory, COLT 2015*. Vol. 40. *JMLR Workshop and Conference Proceedings*. JMLR.org. 1066–1100.

Kasiviswanathan, S. P., H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. (2011). "What can we learn privately?" *SIAM J. Comput.* 40(3): 793–826. ISSN: 0097-5397.

Le Cam, L. (1973). "Convergence of estimates under dimensionality restrictions". *Ann. Statist.* 1: 38–53. ISSN: 0090-5364. URL: http://links.jstor.org/sici?sici=0090-5364(197301)1:1%3C38:COEUDR%3E2.0.CO;2-V&origin=MSN.

Nazarov, F. (2021). "An inequality $k\frac{\sum_{i\neq j} x_i x_j((1-x_i-x_j)^{k-1}-(1-x_i)^k(1-x_j)^k)}{\sum_{i=1}^n x_i(1-(1-x_i)^k)} \leq 2$". Mathematics Stack Exchange. URL: https://math.stackexchange.com/q/4240571.

Onak, K. and X. Sun. (2018). "Probability-Revealing Samples". In: *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*. Ed. by A. J. Storkey and F. Pérez-Cruz. Vol. 84. *Proceedings of Machine Learning Research*. PMLR. 2018–2026. URL: http://proceedings.mlr.press/v84/onak18a.html.

Paninski, L. (2008). "A Coincidence-Based Test for Uniformity Given Very Sparsely Sampled Discrete Data". *IEEE Transactions on Information Theory*. 54(10): 4750–4755.

Parnas, M., D. Ron, and R. Rubinfeld. (2006). "Tolerant property testing and distance approximation". *Journal of Computer and System Sciences*. 72(6): 1012–1042.

Pearson, K. (1900). "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. 50(302): 157–175. DOI: 10.1080/14786440009463897.

Pollard, D. (2003). "Asymptopia". URL: http://www.stat.yale.edu/~pollard/Books/Asymptopia/ (accessed on 11/08/2016).

Rubinfeld, R. and R. A. Servedio. (2009). "Testing monotone high-dimensional distributions". *Random Structures and Algorithms*. 34(1): 24–44. ISSN: 1042-9832. DOI: 10.1002/rsa.v34:1.

Schauer, M. (2021). "Stochastic dominance between (products of) binomials". MathOverflow. eprint: https://mathoverflow.net/q/406217. URL: https://mathoverflow.net/q/406217.

Suzuki, K., D. Tonien, K. Kurosawa, and K. Toyota. (2006). "Birthday paradox for multi-collisions". In: *Information security and cryptology—ICISC 2006*. Vol. 4296. *Lecture Notes in Comput. Sci.* Springer, Berlin. 29–40.

Tsitsiklis, J. N. (1993). "Decentralized detection". In: *Advances in Statistical Signal Processing*. Ed. by H. V. Poor and J. B. Thomas. Vol. 2. JAI Press. 297–344.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. *Springer Series in Statistics*. Springer, New York. xii+214. ISBN: 978-0-387-79051-0. DOI: 10.1007/b13794. URL: https://doi.org/10.1007/b13794.

Valiant, G. and P. Valiant. (2011). "Estimating the Unseen: An $n/\log n$-sample Estimator for Entropy and Support Size, Shown Optimal via New CLTs". In: *Symposium on Theory of Computing Conference, STOC'11*. 685–694.

Valiant, G. and P. Valiant. (2014). "An Automatic Inequality Prover and Instance Optimal Identity Testing". In: *55th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2014*.

Valiant, G. and P. Valiant. (2017). "An Automatic Inequality Prover and Instance Optimal Identity Testing". *SIAM Journal on Computing*. 46(1): 429–455.

Valiant, P. (2011). "Testing symmetric properties of distributions". *SIAM Journal on Computing*. 40(6): 1927–1968.

Vershynin, R. (2018). *High-dimensional probability*. Vol. 47. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. xiv+284. ISBN: 978-1-108-41519-4. DOI: 10.1017/9781108231596. URL: https://doi.org/10.1017/9781108231596.

Wu, Y. and P. Yang. (2020). "Polynomial Methods in Statistical Inference: Theory and Practice". *Found. Trends Commun. Inf. Theory*. 17(4): 402–586. DOI: 10.1561/0100000095. URL: https://doi.org/10.1561/0100000095.

Yu, B. (1997). "Assouad, Fano, and Le Cam". In: *Festschrift for Lucien Le Cam*. Springer. 423–435. DOI: 10.1007/978-1-4612-1880-7_29. URL: http://dx.doi.org/10.1007/978-1-4612-1880-7_29.