Topics and Techniques in Distribution Testing

Other titles in Foundations and Trends® in Communications and Information Theory

Rank-Metric Codes and Their Applications
Hannes Bartz, Lukas Holzbaur, Hedongliang Liu, Sven Puchinger, Julian Renner and Antonia Wachter-Zeh
ISBN: 978-1-63828-000-2

Common Information, Noise Stability, and Their Extensions
Lei Yu and Vincent Y. F. Tan
ISBN: 978-1-63828-014-9

Information-Theoretic Foundations of DNA Data Storage
Ilan Shomorony and Reinhard Heckel
ISBN: 978-1-68083-956-2

Asymptotic Frame Theory for Analog Coding
Marina Haikin, Matan Gavish, Dustin G. Mixon and Ram Zamir
ISBN: 978-1-68083-908-1

Modeling and Optimization of Latency in Erasure-coded Storage Systems
Vaneet Aggarwal and Tian Lan
ISBN: 978-1-68083-842-8

An Algebraic and Probabilistic Framework for Network Information Theory
S. Sandeep Pradhan, Arun Padakandla and Farhad Shirani
ISBN: 978-1-68083-766-7

Topics and Techniques in Distribution Testing

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

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Communications and Information Theory

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

C. L. Canonne. *Topics and Techniques in Distribution Testing*. Foundations and Trends® in Communications and Information Theory, vol. 19, no. 6, pp. 1032–1198, 2022.

ISBN: 978-1-63828-101-6

Foundations and Trends® in Communications and Information Theory

Volume 19, Issue 6, 2022

Editorial Scope

Topics

Foundations and Trends® in Communications and Information Theory publishes survey and tutorial articles in the following topics:

- Coded modulation
- Coding theory and practice
- Communication complexity
- Communication system design
- Cryptology and data security
- Data compression
- Data networks
- Demodulation and Equalization
- Denoising
- Detection and estimation
- Information theory and statistics
- Information theory and computer science
- Joint source/channel coding
- Modulation and signal design
- Multiuser detection
- Multiuser information theory
- Optical communication channels
- Pattern recognition and learning
- Quantization
- Quantum information processing
- Rate-distortion theory
- Shannon theory
- Signal processing for communications
- Source coding
- Storage and recording codes
- Speech and Image Compression
- Wireless Communications

Information for Librarians

Foundations and Trends® in Communications and Information Theory, 2022, Volume 19, 4 issues. ISSN paper version 1567-2190. ISSN online version 1567-2328 . Also available as a combined paper and online subscription.

Contents

1 What Is Distribution Testing? 2
1.1 Formulation, and relation to Hypothesis Testing 4
1.2 Why total variation distance? 10
1.3 The road not taken: tolerant testing 12
1.4 Historical notes 13

2 Testing Goodness-of-fit of Univariate Distributions 16
2.1 Uniformity testing 18
2.2 Identity testing 69
2.3 Historical notes 89
2.4 Exercises 92
2.5 Deferred proofs 96

3 Information-theoretic Lower Bounds 98
3.1 Indistinguishability, Le Cam, and Ingster's method 98
3.2 Indistinguishability via Fano: a bit of mutual information 106
3.3 Indistinguishability via moment-matching 111
3.4 Indistinguishability on an instance-by-instance basis . . . 117
3.5 Proving hardness by reductions 124
3.6 Historical notes 131
3.7 Exercises 132

4 Testing with Constrained Measurements 134
4.1 Setting(s), and the devil lurking in the details 136
4.2 Simulate-and-Infer 139
4.3 Random hashing and domain compression 142
4.4 Historical notes 143
4.5 Exercises 144

Acknowledgements 146

Appendices 147

A Some Good Inequalities 148

B Metrics and Divergences Between Probability Distributions 151

C Poissonization 155

References 157

Topics and Techniques in Distribution Testing

Clément L. Canonne

University of Sydney, Australia; 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 (2022), "Topics and Techniques in Distribution Testing", Foundations and Trends® in Communications and Information Theory: Vol. 19, No. 6, pp 1032–1198. DOI: 10.1561/010000114.

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 on 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 mean 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.

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 (see Figure 1.1 for an illustration). 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,

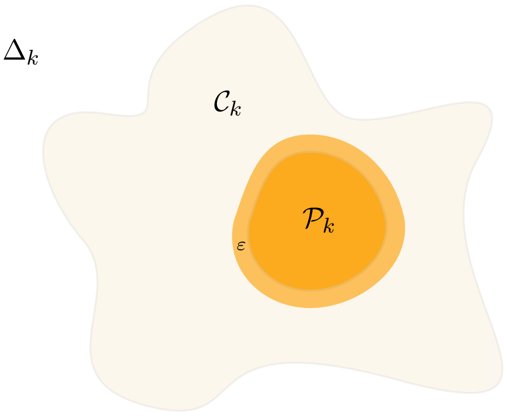

Figure 1.1: An example of property to test. Here, $\mathcal{P}_k \subseteq \mathcal{C}_k \subseteq \Delta_k$, where the property $\mathcal{P}_k$ is depicted as the inner orange area ("yolk"), and the "egg white" is the area of rejection, *i.e.*, the subset of $\mathcal{C}_k$ at total variation distance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

and we may want to choose $\mathcal{C}_k$ differently to perform hypothesis testing *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.

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

Dependence on the error probability δ. Our definition of testing algorithm leaves the error probability δ as a free parameter; however, it 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 Section 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 $\mathrm{d_{TV}}(\hat{\mathbf{p}}, \mathbf{p}) \leq \varepsilon/2$ with probability at least $1 - \delta$. Assuming this is the case, then (i) if $\mathbf{p} \in \mathcal{P}_k$, then of course $\mathrm{d_{TV}}(\hat{\mathbf{p}}, \mathcal{P}) \leq \varepsilon/2$; while (ii) if $\mathrm{d_{TV}}(\mathbf{p}, \mathcal{P}) > \varepsilon$, by the triangle inequality (since total variation distance is a metric) we must have $\mathrm{d_{TV}}(\hat{\mathbf{p}}, \mathcal{P}) > \varepsilon/2$.

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 Section 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.*, Canonne [28] and Kamath *et al.* [65]. 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 $n\left(k, \varepsilon(\mathbf{p}), \frac{\delta}{\log(1/\varepsilon(\mathbf{p}))}\right)$, 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 distributions: the total variation distance (Definition 1.1). It is natural to wonder if 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 be 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 [67]. 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

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

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 $\text{Type I} + \text{Type II} = 1 + \mathbf{p}_0(S^*) - \mathbf{p}_1(S^*) = 1 - \mathrm{d}_{\mathrm{TV}}(\mathbf{p}_0, \mathbf{p}_1)$. □

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* [72], and is formalized by introducing two parameters $0 \le \varepsilon' < \varepsilon \le 1$, 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 $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{P}_k) \le \varepsilon'$, then $\Pr_{S,\mathcal{A}}[\,\mathbf{b} = 1\,] \ge 1 - \delta$;

(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 [81], 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 *et al.* [86] for a primer on some of those techniques, and to Canonne *et al.* [32] 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 [73] 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),

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

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 *et al.* [14] 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 perhaps shares most similarity with the work of Ingster [61], [62].

Distribution testing was first introduced in an influential paper by Goldreich *et al.* [58], which formally defined the broader field of property testing; Goldreich *et al.* [59] 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.* [19] 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.* [37], 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 [41], [43]. 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 [34], [36], or the ability to ask for, or observe, the probability of individual elements of the domain [35], [70], [75]. 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 [29].

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* [52] and its variants) – see, *e.g.*, [64], memory constraints, and bandwidth constraints [79]; of which we will cover a fraction in Section 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.

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 [85, Chapter 2] or Boucheron *et al.* [23] 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

$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq \frac{\operatorname{Var}[X]}{t^2}$$

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

$$\Pr\left[\sum_{i=1}^{n} X_i > (1+\gamma)P_H\right] < \exp\left(-\gamma^2 P_H/3\right) \tag{A.7}$$

$$\Pr\left[\sum_{i=1}^{n} X_i < (1-\gamma)P_L\right] < \exp\left(-\gamma^2 P_L/2\right) \tag{A.8}$$

Theorem A.7 (Bernstein's inequality). Let $X_1, \dots, X_n$ be independent random variables taking values in $[-a,a]$, and such that $\mathbb{E}\left[X_i^2\right] \leq v_i$ for all i. Then, for every $t \geq 0$, we have

$$\Pr\left[\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}[X_i]\right| \geq t\right] \leq \exp\left(-\frac{t^2}{2(\sum_{i=1}^{n} v_i + \frac{a}{3}t)}\right).$$

In particular, if $X_1, \dots, X_n$ are i.i.d. with mean μ and $\mathbb{E}\left[X_1^2\right] \leq v$, then for any $\gamma \geq 0$ we have

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \gamma\right] \leq \exp\left(-\frac{\gamma^2 n}{2(v + \frac{a}{3}\gamma)}\right).$$

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

$$\Pr[\, X \geq \lambda + t \,] \leq e^{-\frac{t^2}{2\lambda}\psi\left(\frac{t}{\lambda}\right)} \leq e^{-\frac{t^2}{2(\lambda+t)}} \tag{A.9}$$

and, for any $0 < t < \lambda$,

$$\Pr[\, X \leq \lambda - t \,] \leq e^{-\frac{t^2}{2\lambda}\psi\left(-\frac{t}{\lambda}\right)} \leq e^{-\frac{t^2}{2(\lambda+t)}}, \tag{A.10}$$

where $\psi(u) := 2\frac{(1+u)\ln(1+u)-u}{u^2}$ for $u \geq -1$. In particular, for any $t \geq 0$,

$$\Pr[\, |X - \lambda| \geq t \,] \leq 2e^{-\frac{t^2}{2(\lambda+t)}}. \tag{A.11}$$

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

$$\mathrm{D}(\mathbf{p}_1 \| \mathbf{p}_2) = \sum_{x \in \mathcal{X}} \mathbf{p}_1(x) \ln \frac{\mathbf{p}_1(x)}{\mathbf{p}_2(x)} \tag{B.1}$$

$$\chi^2(\mathbf{p}_1 \mid\mid \mathbf{p}_2) = \sum_{x \in \mathcal{X}} \frac{(\mathbf{p}_1(x) - \mathbf{p}_2(x))^2}{\mathbf{p}_2(x)} \tag{B.2}$$

$$\mathrm{d}_{\mathrm{H}}(\mathbf{p}_1, \mathbf{p}_2) = \frac{1}{\sqrt{2}} \|\sqrt{\mathbf{p}_1} - \sqrt{\mathbf{p}_2}\|_2 , \tag{B.3}$$

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

$$\mathrm{D}(\mathbf{p}_1 \otimes \cdots \otimes \mathbf{p}_n \| \mathbf{q}_1 \otimes \cdots \otimes \mathbf{q}_n) = \sum_{i=1}^{n} \mathrm{D}(\mathbf{p}_i \| \mathbf{q}_i) \tag{B.4}$$

$$\chi^2(\mathbf{p}_1 \otimes \cdots \otimes \mathbf{p}_n \mid\mid \mathbf{q}_1 \otimes \cdots \otimes \mathbf{q}_n) = \prod_{i=1}^{n} \Big(1 + \chi^2(\mathbf{p}_i \mid\mid \mathbf{q}_i)\Big) - 1 \tag{B.5}$$

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 $a-b=(\sqrt{a}-\sqrt{b})(\sqrt{a}+\sqrt{b})$,

$$\begin{aligned}
\mathrm{d_H}(\mathbf{p},\mathbf{q})^2 &= \frac{1}{2}\sum_{x\in\mathcal{X}}\left(\sqrt{\mathbf{p}(x)}-\sqrt{\mathbf{q}(x)}\right)^2 \\
&\leq \frac{1}{2}\sum_{x\in\mathcal{X}}\left|\sqrt{\mathbf{p}(x)}-\sqrt{\mathbf{q}(x)}\right|\left(\sqrt{\mathbf{p}(x)}+\sqrt{\mathbf{q}(x)}\right) \\
&= \frac{1}{2}\sum_{x\in\mathcal{X}}|\mathbf{p}(x)-\mathbf{q}(x)| = \mathrm{d_{TV}}(\mathbf{p},\mathbf{q})\,.
\end{aligned}$$

For the right side, we have, by Cauchy–Schwarz and then using the identity $2(a+b)=(\sqrt{a}+\sqrt{b})^2+(\sqrt{a}-\sqrt{b})^2$,

$$\begin{aligned}
\mathrm{d_{TV}}(\mathbf{p},\mathbf{q}) &= \frac{1}{2}\sum_{x\in\mathcal{X}}\left|\sqrt{\mathbf{p}(x)}-\sqrt{\mathbf{q}(x)}\right|\left(\sqrt{\mathbf{p}(x)}+\sqrt{\mathbf{q}(x)}\right) \\
&\leq \frac{1}{2}\sqrt{\sum_{x\in\mathcal{X}}\left(\sqrt{\mathbf{p}(x)}-\sqrt{\mathbf{q}(x)}\right)^2}\sqrt{\sum_{x\in\mathcal{X}}\left(\sqrt{\mathbf{p}(x)}+\sqrt{\mathbf{q}(x)}\right)^2} \\
&= \frac{1}{\sqrt{2}}\mathrm{d_H}(\mathbf{p},\mathbf{q})\sqrt{\sum_{x\in\mathcal{X}}\left(2(\mathbf{p}(x)+\mathbf{q}(x))-\left(\sqrt{\mathbf{p}(x)}-\sqrt{\mathbf{q}(x)}\right)^2\right)} \\
&= \mathrm{d_H}(\mathbf{p},\mathbf{q})\sqrt{2-\mathrm{d_H}(\mathbf{p},\mathbf{q})^2}\,,
\end{aligned}$$

which implies the (slightly weaker) inequality we wanted to show. □

Lemma B.2. For every $\mathbf{p},\mathbf{q}$ on $\mathcal{X}$,

$$\mathrm{d_{TV}}(\mathbf{p},\mathbf{q})^2 \leq \frac{1}{4}\chi^2(\mathbf{p}\parallel\mathbf{q})\,.$$

Proof. By Cauchy–Schwarz,

$$\begin{aligned}
\mathrm{d_{TV}}(\mathbf{p},\mathbf{q}) &= \frac{1}{2}\sum_{x\in\mathcal{X}}|\mathbf{p}(x)-\mathbf{q}(x)| \\
&\leq \frac{1}{2}\sqrt{\sum_{x\in\mathcal{X}}\frac{(\mathbf{p}(x)-\mathbf{q}(x))^2}{\mathbf{q}(x)}}\sqrt{\sum_{x\in\mathcal{X}}\mathbf{q}(x)} \\
&= \frac{1}{2}\sqrt{\chi^2(\mathbf{p}\parallel\mathbf{q})}\,.
\end{aligned}$$

□

Lemma B.3 (Pinsker's Inequality)**.** For every $\mathbf{p}, \mathbf{q}$ on $\mathcal{X}$,

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \le \sqrt{\frac{1}{2}\mathrm{D}(\mathbf{p}\|\mathbf{q})}\,.$$

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 $\mathcal{X}$,

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \le \sqrt{1 - e^{-\mathrm{D}(\mathbf{p}\|\mathbf{q})}}\,. \tag{B.8}$$

In particular, as $\sqrt{1-e^{-x}} \le 1 - \frac{1}{2}e^{-x}$ for $x \ge 0$, this implies

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \le 1 - \frac{1}{2}e^{-\mathrm{D}(\mathbf{p}\|\mathbf{q})}\,. \tag{B.9}$$

We refer the reader to Canonne [30] or Tsybakov [80, Section 2.4.1] for a proof and discussion of this inequality, due to Bretagnolle *et al.* [25].

Lemma B.5. For every $\mathbf{p}, \mathbf{q}$ on $\mathcal{X}$,

$$\mathrm{D}(\mathbf{p}\|\mathbf{q}) \le \ln\Big(1 + \chi^2(\mathbf{p} \mid\mid \mathbf{q})\Big) \le \chi^2(\mathbf{p} \mid\mid \mathbf{q})$$

Proof. The second inequality follows from the standard convexity inequality $\ln(1+x) \le x$ (for $x > -1$), so it suffices to prove the first. To do so, observe that

$$\begin{aligned}
\mathrm{D}(\mathbf{p}\|\mathbf{q}) &= \sum_{x\in\mathcal{X}} \mathbf{p}(x) \ln\frac{\mathbf{p}(x)}{\mathbf{q}(x)} \\
&\le \ln \sum_{x\in\mathcal{X}} \frac{\mathbf{p}(x)^2}{\mathbf{q}(x)} && \text{(Jensen's inequality)} \\
&= \ln\Big(1 + \chi^2(\mathbf{p} \mid\mid \mathbf{q})\Big)\,,
\end{aligned}$$

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 $(N_1, \dots, N_k)$ (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 $(\mathbf{p}(1), \dots, \mathbf{p}(k))$.

An unfortunate aspect of this is that those $N_1, \dots, N_k$ are not independent: each of them is marginally a Binomial random variable, with $N_i \sim \mathrm{Bin}(n, \mathbf{p}(i))$, but those are dependent, since for instance $N_1 + \dots + N_k = n$.[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 $N \sim \mathrm{Poisson}(n)$;
2. Draw N i.i.d. samples $X_1, \dots, X_N$ from $\mathbf{p}$;
3. Provide $X_1, \dots, X_N$ to the algorithm.

[1] More specifically, the N_i's are *negatively associated*; see Definition 2.3.

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 $(N_1, \dots, N_k)$ 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) $N_1, \dots, N_k$ 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 $n^{\Fish}(k, \varepsilon, \delta)$. Then there exists a tester for property $\mathcal{P}$ (in the standard sampling setting) with sample complexity $n(k, \varepsilon, \delta) = \max\Big(\frac{3}{2} \cdot n^{\Fish}(k, \varepsilon, \delta/2), 18\ln(4/\delta)\Big)$.

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 $n^{\Fish}(k, \varepsilon, \delta) = \max(2 \cdot n(k, \varepsilon, \delta/2), 12\ln(4/\delta))$.

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 [84, Section 4.3] and references within, or Canonne [29, Appendix D.3].

References

[1] J. Acharya, C. L. Canonne, Y. Han, Z. Sun, and H. Tyagi, "Domain compression and its application to randomness-optimal distributed goodness-of-fit," in *Proceedings of Thirty Third Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 125, pp. 3–40, PMLR, 2020, URL: http://proceedings.mlr.press/v125/acharya20a.html.

[2] J. Acharya, C. L. Canonne, C. Freitag, Z. Sun, and H. Tyagi, "Inference under information constraints III: local privacy constraints," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, 2021, pp. 253–267.

[3] J. Acharya, C. L. Canonne, Y. Liu, Z. Sun, and H. Tyagi, "Interactive inference under information constraints," *IEEE Trans. Inf. Theory*, vol. 68, no. 1, 2022, pp. 502–516. DOI: 10.1109/TIT.2021.3123905.

[4] J. Acharya, C. L. Canonne, A. V. Singh, and H. Tyagi, "Optimal rates for nonparametric density estimation under communication constraints," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 26 754–26 766, 2021, URL: https://proceedings.neurips.cc/paper/2021/hash/e1021d43911ca2c1845910d84f40aeae-Abstract.html.

[5] J. Acharya, C. L. Canonne, Z. Sun, and H. Tyagi, "Unified lower bounds for interactive high-dimensional estimation under information constraints," *CoRR*, vol. abs/2010.06562, 2020.

[6] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints I: Lower bounds from chi-square contraction," *IEEE Trans. Inform. Theory*, vol. 66, no. 12, 2020, pp. 7835–7855. DOI: 10.1109/TIT.2020.3028440.

[7] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints II: Communication constraints and shared randomness," *IEEE Trans. Inform. Theory*, vol. 66, no. 12, 2020, pp. 7856–7877. DOI: 10.1109/TIT.2020.3028439.

[8] J. Acharya, C. Daskalakis, and G. C. Kamath, "Optimal Testing for Properties of Distributions," in *Advances in Neural Information Processing Systems 28*, Curran Associates, Inc., 2015, pp. 3577–3598.

[9] J. Acharya, Z. Sun, and H. Zhang, "Differentially private testing of identity and closeness of discrete distributions," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 6879–6891, 2018, URL: https://proceedings.neurips.cc/paper/2018/hash/7de32147a4f1055bed9e4faf3485a84d-Abstract.html.

[10] J. Acharya, Z. Sun, and H. Zhang, "Hadamard response: Estimating distributions privately, efficiently, and with little communication," in *Proceedings of Machine Learning Research*, ser. Proceedings of Machine Learning Research, vol. 89, pp. 1120–1129, PMLR, 2019, URL: http://proceedings.mlr.press/v89/acharya19a.html.

[11] M. Aliakbarpour, I. Diakonikolas, and R. Rubinfeld, "Differentially private identity and equivalence testing of discrete distributions," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, vol. 80, pp. 169–178, PMLR, 2018, URL: http://proceedings.mlr.press/v80/aliakbarpour18a.html.

[12] K. Amin, M. Joseph, and J. Mao, "Pan-private uniformity testing," in *Proceedings of Thirty Third Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 125, pp. 183–218, PMLR, 2020, URL: http://proceedings.mlr.press/v125/amin20a.html.

[13] B. C. Arnold, *Majorization and the Lorenz order: a brief introduction*, vol. 43, ser. Lecture Notes in Statistics. Springer-Verlag, Berlin, 1987, pp. vi+122. DOI: 10.1007/978-1-4615-7379-1.

[14] S. Balakrishnan and L. Wasserman, "Hypothesis testing for high-dimensional multinomials: A selective review," *Ann. Appl. Stat.*, vol. 12, no. 2, 2018, pp. 727–749. DOI: 10.1214/18-AOAS1155SF.

[15] V. Balcer, A. Cheu, M. Joseph, and J. Mao, "Connecting robust shuffle privacy and pan-privacy," in *SODA*, pp. 2384–2403, SIAM, 2021.

[16] L. P. Barnes, Y. Han, and A. Özgür, "Lower bounds for learning distributions under communication constraints via Fisher information," *J. Mach. Learn. Res.*, vol. 21, 2020, Paper No. 236, 30.

[17] T. Batu and C. L. Canonne, "Generalized uniformity testing," in *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*, IEEE Computer Soc., Los Alamitos, CA, 2017, pp. 880–889.

[18] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White, "Testing random variables for independence and identity," in *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001*, pp. 442–451, 2001.

[19] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing that distributions are close," in *41st Annual Symposium on Foundations of Computer Science, FOCS 2000*, pp. 189–197, 2000.

[20] T. Berrett and C. Butucea, "Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms," in *NeurIPS*, 2020.

[21] L. Birgé, "On the risk of histograms for estimating decreasing densities," *The Annals of Statistics*, vol. 15, no. 3, 1987, pp. 1013–1022, URL: http://www.jstor.org/stable/2241812.

[22] E. Blais, C. L. Canonne, and T. Gur, "Distribution testing lower bounds via reductions from communication complexity," *ACM Trans. Comput. Theory*, vol. 11, no. 2, 2019, Art. 6, 37. DOI: 10.1145/3305270.

[23] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities*. Oxford University Press, Oxford, 2013, pp. x+481. DOI: 10.1093/acprof:oso/9780199535255.001.0001.

[24] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff, "Communication lower bounds for statistical estimation problems via a distributed data processing inequality," in *Symposium on Theory of Computing Conference, STOC'16*, pp. 1011–1020, ACM, 2016.

[25] J. Bretagnolle and C. Huber, "Estimation des densités: Risque minimax," in *Séminaire de Probabilités, XII (Univ. Strasbourg, Strasbourg, 1976/1977)*, ser. Lecture Notes in Math. Vol. 649, Springer, Berlin, 1978, pp. 342–363.

[26] C. Butucea, A. Dubois, M. Kroll, and A. Saumard, "Local differential privacy: Elbow effect in optimal density estimation and adaptation over Besov ellipsoids," *Bernoulli*, vol. 26, no. 3, 2020, pp. 1727–1764. DOI: 10.3150/19-BEJ1165.

[27] C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld, "Testing Shape Restrictions of Discrete Distributions," in *Proceedings of STACS*, 2016. DOI: 10.4230/LIPIcs.STACS.2016.25.

[28] C. L. Canonne, *A short note on learning discrete distributions*, 2020. arXiv: `2002.11457 [math.ST]`.

[29] C. L. Canonne, *A Survey on Distribution Testing: Your Data is Big. But is it Blue?* Ser. Graduate Surveys 9. Theory of Computing Library, 2020, pp. 1–100. DOI: 10.4086/toc.gs.2020.009.

[30] C. L. Canonne, *A short note on an inequality between KL and TV*, 2022. arXiv: `2202.07198 [math.PR]`.

[31] C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld, "Testing shape restrictions of discrete distributions," *Theory of Computing Systems*, 2017, pp. 1–59. DOI: 10.1007/s00224-017-9785-6.

[32] C. L. Canonne, A. Jain, G. Kamath, and J. Li, "The price of tolerance in distribution testing," in *Conference on Learning Theory, 2-5 July 2022, London, UK*, ser. Proceedings of Machine Learning Research, vol. 178, pp. 573–624, PMLR, 2022, URL: https://proceedings.mlr.press/v178/canonne22a.html.

[33] C. L. Canonne and H. Lyu, "Uniformity testing in the shuffle model: Simpler, better, faster," in *SIAM Symposium on Simplicity in Algorithms (SOSA)*, 2022.

[34] C. L. Canonne, D. Ron, and R. A. Servedio, "Testing probability distributions using conditional samples," *SIAM Journal on Computing*, vol. 44, no. 3, 2015, pp. 540–616. DOI: 10.1137/130945508.

[35] C. L. Canonne and R. Rubinfeld, "Testing probability distributions underlying aggregated data," in *Proceedings of ICALP*, pp. 283–295, 2014.

[36] S. Chakraborty, E. Fischer, Y. Goldhirsh, and A. Matsliah, "On the power of conditional samples in distribution testing," in *Proceedings of ITCS*, pp. 561–580, Berkeley, California, USA: ACM, 2013. DOI: 10.1145/2422436.2422497.

[37] S. Chan, I. Diakonikolas, G. Valiant, and P. Valiant, "Optimal algorithms for testing closeness of discrete distributions," in *Proceedings of SODA*, pp. 1193–1203, 2014.

[38] T. M. Cover and J. A. Thomas, *Elements of information theory*, Second. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006, pp. xxiv+748.

[39] C. Daskalakis, I. Diakonikolas, and R. A. Servedio, "Learning k-modal distributions via testing," *Theory of Computing*, vol. 10, no. 20, 2014, pp. 535–570. DOI: 10.4086/toc.2014.v010a020.

[40] C. Daskalakis, I. Diakonikolas, and R. A. Servedio, "Learning Poisson Binomial Distributions," *Algorithmica*, vol. 72, no. 1, 2015, pp. 316–357.

[41] I. Diakonikolas, T. Gouleakis, D. M. Kane, J. Peebles, and E. Price, "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*, pp. 542–555, ACM, 2021. DOI: 10.1145/3406325.3450997.

[42] I. Diakonikolas, T. Gouleakis, D. M. Kane, and S. Rao, "Communication and memory efficient testing of discrete distributions," in *COLT*, ser. Proceedings of Machine Learning Research, vol. 99, pp. 1070–1106, PMLR, 2019.

[43] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price, "Sample-optimal identity testing with high probability," in *45th International Colloquium on Automata, Languages, and Programming*, ser. LIPIcs. Leibniz Int. Proc. Inform. Vol. 107, Art. No. 41, 14, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2018.

[44] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price, "Collision-based testers are optimal for uniformity and closeness," *Chic. J. Theoret. Comput. Sci.*, 2019, Art. 1, 21. DOI: 10.4086/cjtcs.2019.001.

[45] I. Diakonikolas and D. M. Kane, "A new approach for testing properties of discrete distributions," in *57th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2016*, IEEE Computer Society, 2016.

[46] I. Diakonikolas, D. M. Kane, and V. Nikishkin, "Testing Identity of Structured Distributions," in *Proceedings of SODA*, 2015.

[47] D. Dubhashi and D. Ranjan, "Balls and bins: A study in negative dependence," *Random Structures Algorithms*, vol. 13, no. 2, 1998, pp. 99–124. DOI: 10.1002/(SICI)1098-2418(199809)13:2<99::AID-RSA1>3.0.CO;2-M.

[48] D. P. Dubhashi and A. Panconesi, *Concentration of measure for the analysis of randomized algorithms.* Cambridge University Press, Cambridge, 2009, pp. xvi+196. DOI: 10.1017/CBO9780511581274.

[49] J. Duchi and R. Rogers, "Lower bounds for locally private estimation via communication complexity," in *Proceedings of the Thirty-Second Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 99, pp. 1161–1191, Phoenix, USA: PMLR, 2019.

[50] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013*, pp. 429–438, IEEE Computer Society, 2013.

[51] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Minimax optimal procedures for locally private estimation," *J. Amer. Statist. Assoc.*, vol. 113, no. 521, 2018, pp. 182–201.

[52] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography*, ser. Lecture Notes in Comput. Sci. Vol. 3876, Springer, Berlin, 2006, pp. 265–284.

[53] O. Fischer, U. Meir, and R. Oshman, "Distributed uniformity testing," in *Proceedings of the 2018 ACM Symposium on Principles of Distributed Computing, PODC 2018, Egham, United Kingdom, July 23-27, 2018*, pp. 455–464, ACM, 2018, URL: https://dl.acm.org/citation.cfm?id=3212772.

[54] A. Garg, T. Ma, and H. L. Nguyen, "On communication cost of distributed statistical estimation and dimensionality," in *Advances in Neural Information Processing Systems 27*, pp. 2726–2734, 2014.

[55] O. Goldreich, "On Multiple Input Problems in Property Testing," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*, ser. Leibniz International Proceedings in Informatics (LIPIcs), vol. 28, pp. 704–720, Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2014. DOI: 10.4230/LIPIcs.APPROX-RANDOM.2014.704.

[56] O. Goldreich, "The uniform distribution is complete with respect to testing identity to a fixed distribution," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 23, 2016, p. 15, URL: http://eccc.hpi-web.de/report/2016/015.

[57] O. Goldreich, *Introduction to Property Testing*. Cambridge University Press, 2017, URL: http://www.wisdom.weizmann.ac.il/~oded/pt-intro.html.

[58] O. Goldreich, S. Goldwasser, and D. Ron, "Property testing and its connection to learning and approximation," *Journal of the ACM*, vol. 45, no. 4, 1998, pp. 653–750.

[59] O. Goldreich and D. Ron, "On testing expansion in bounded-degree graphs," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 7, no. 20, 2000.

[60] D. Huang and S. Meyn, "Generalized error exponents for small sample universal hypothesis testing," *IEEE Transactions on Information Theory*, vol. 59, no. 12, 2013, pp. 8157–8181.

[61] Y. I. Ingster, "A minimax test of nonparametric hypotheses on the density of a distribution in L_p metrics," *Teor. Veroyatnost. i Primenen.*, vol. 31, no. 2, 1986, pp. 384–389.

[62] Y. I. Ingster, "Adaptive chi-square tests," *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, vol. 244, no. Veroyatn. i Stat. 2, 1997, pp. 150–166, 333. DOI: 10.1007/BF02673632.

[63] Y. I. Ingster and I. A. Suslina, *Nonparametric goodness-of-fit testing under Gaussian models*, vol. 169, ser. Lecture Notes in Statistics. Springer-Verlag, New York, 2003, pp. xiv+453. DOI: 10.1007/978-0-387-21580-8.

[64] G. Kamath and J. R. Ullman, "A primer on private statistics," *CoRR*, vol. abs/2005.00010, 2020. [Online]. Available: https://arxiv.org/abs/2005.00010, URL: https://arxiv.org/abs/2005.00010.

[65] S. Kamath, A. Orlitsky, D. Pichapati, and A. T. Suresh, "On learning distributions from their samples," in *Proceedings of the 28th Conference on Learning Theory, COLT 2015*, ser. JMLR Workshop and Conference Proceedings, vol. 40, pp. 1066–1100, JMLR.org, 2015.

[66] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM J. Comput.*, vol. 40, no. 3, 2011, pp. 793–826.

[67] L. Le Cam, "Convergence of estimates under dimensionality restrictions," *Ann. Statist.*, vol. 1, 1973, pp. 38–53, URL: http://links.jstor.org/sici?sici=0090-5364(197301)1:1<38:COEUDR>2.0.CO;2-V&origin=MSN.

[68] U. Meir, D. Minzer, and R. Oshman, "Can distributed uniformity testing be local?" In *Proceedings of the 2019 ACM Symposium on Principles of Distributed Computing, PODC 2019, Toronto, ON, Canada, July 29 - August 2, 2019*, pp. 228–237, ACM, 2019. DOI: 10.1145/3293611.3331613.

[69] F. Nazarov, *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, 2021, URL: https://math.stackexchange.com/q/4240571.

[70] K. Onak and X. Sun, "Probability-revealing samples," in *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, ser. Proceedings of Machine Learning Research, vol. 84, pp. 2018–2026, PMLR, 2018, URL: http://proceedings.mlr.press/v84/onak18a.html.

[71] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data," *IEEE Transactions on Information Theory*, vol. 54, no. 10, 2008, pp. 4750–4755.

[72] M. Parnas, D. Ron, and R. Rubinfeld, "Tolerant property testing and distance approximation," *Journal of Computer and System Sciences*, vol. 72, no. 6, 2006, pp. 1012–1042.

[73] K. Pearson, "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*, vol. 50, no. 302, 1900, pp. 157–175. DOI: 10.1080/14786440009463897.

[74] D. Pollard, *Asymptopia*, 2003. (accessed on 11/08/2016), URL: http://www.stat.yale.edu/~pollard/Books/Asymptopia/.

[75] R. Rubinfeld and R. A. Servedio, "Testing monotone high-dimensional distributions," *Random Structures and Algorithms*, vol. 34, no. 1, 2009, pp. 24–44. DOI: 10.1002/rsa.v34:1.

[76] M. Schauer, *Stochastic dominance between (products of) binomials*, MathOverflow, 2021, URL: https://mathoverflow.net/q/406217.

[77] O. Shamir, "Fundamental limits of online and distributed algorithms for statistical learning and estimation," in *Advances in Neural Information Processing Systems 27*, pp. 163–171, 2014.

[78] K. Suzuki, D. Tonien, K. Kurosawa, and K. Toyota, "Birthday paradox for multi-collisions," in *Information security and cryptology—ICISC 2006*, ser. Lecture Notes in Comput. Sci. Vol. 4296, Springer, Berlin, 2006, pp. 29–40.

[79] J. N. Tsitsiklis, "Decentralized detection," in *Advances in Statistical Signal Processing*, vol. 2, pp. 297–344, JAI Press, 1993.

[80] A. B. Tsybakov, *Introduction to nonparametric estimation*, ser. Springer Series in Statistics. Springer, New York, 2009, pp. xii+214. DOI: 10.1007/b13794.

[81] G. Valiant and P. Valiant, "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*, pp. 685–694, 2011.

[82] G. Valiant and P. Valiant, "An automatic inequality prover and instance optimal identity testing," in *55th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2014*, 2014.

[83] G. Valiant and P. Valiant, "An automatic inequality prover and instance optimal identity testing," *SIAM Journal on Computing*, vol. 46, no. 1, 2017, pp. 429–455.

[84] P. Valiant, "Testing symmetric properties of distributions," *SIAM Journal on Computing*, vol. 40, no. 6, 2011, pp. 1927–1968.

[85] R. Vershynin, *High-dimensional probability*, vol. 47, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018, pp. xiv+284. DOI: 10.1017/9781108231596.

[86] Y. Wu and P. Yang, "Polynomial methods in statistical inference: Theory and practice," *Found. Trends Commun. Inf. Theory*, vol. 17, no. 4, 2020, pp. 402–586. DOI: 10.1561/0100000095.

[87] M. Ye and A. Barg, "Optimal schemes for discrete distribution estimation under locally differential privacy," *IEEE Trans. Inform. Theory*, vol. 64, no. 8, 2018, pp. 5662–5676. DOI: 10.1109/TIT.2018.2809790.

[88] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*, Springer, 1997, pp. 423–435. DOI: 10.1007/978-1-4612-1880-7_29.

[89] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," in *Advances in Neural Information Processing Systems 26*, pp. 2328–2336, 2013.