Proceedings of Machine Learning Research vol 196:1–5, 2024 37th Annual Conference on Learning Theory

Open Problem: Tight Characterization of Instance-Optimal Identity Testing

Clément Canonne CLEMENT.CANONNE@SYDNEY.EDU.AU
University of Sydney

Editors: Shipra Agrawal and Aaron Roth

Abstract

In the *instance-optimal* identity testing introduced by Valiant and Valiant (2014), one is given the (succinct) description of a discrete probability distribution $\mathbf{q}$, as well as a a parameter $\varepsilon \in (0,1]$ and i.i.d. samples from an (unknown, arbitrary) discrete distribution $\mathbf{p}$. The goal is to distinguish with high probability between the cases (i) $\mathbf{p} = \mathbf{q}$ and (ii) $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$, using the minimum number of samples possible as a function of (some simple functional of) $\mathbf{q}$ and ε. This is in contrast with the standard formulation of identity testing, where the sample complexity is taken as worst-case over all possible reference distributions $\mathbf{q}$. Valiant and Valiant provided upper and lower bounds on this question, where the sample complexity is expressed in terms of the "$\ell_{2/3}$ norm" of some (truncated version) of the reference distribution $\mathbf{q}$. However, these upper and lower bounds do not always match up to constant factors, and can differ by an arbitrary multiplicative gap for some choices of $\mathbf{q}$. The question then is: what is the tight characterization of the sample complexity of instance-optimal identity testing? What is the "right" functional $\Phi(\mathbf{q})$ for it?

Keywords: distribution testing, hypothesis testing, statistics, finite-sample guarantees

1. Introduction

In the identity testing problem, one is given the (succinct) description of a discrete probability distribution $\mathbf{q}$ over a domain of size k, as well as a a parameter $\varepsilon \in (0,1]$ and i.i.d. samples from an (unknown, arbitrary) discrete distribution $\mathbf{p}$. The goal is to distinguish with high probability between the cases (i) $\mathbf{p} = \mathbf{q}$ and (ii) $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$,[1] using the minimum number of samples possible as a function of (some simple functional of) k and ε. Specifically, the *sample complexity* $n = n(k, \varepsilon)$ is the least number such that there exists a (possibly randomized) algorithm $T \colon \Delta(\mathcal{X}) \times \mathcal{X}^n \to \{0,1\}$ such that, for every $\mathbf{q}$,

- $\Pr_{X \sim \mathbf{q}^{\otimes n}}[T(\mathbf{q}, X) = 1] \leq 1/3$.
- If $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$, then $\Pr_{X \sim \mathbf{p}^{\otimes n}}[T(\mathbf{q}, X) = 0] \leq 1/3$.

This question, as well as many variants (including one where the Type I and Type II errors, here set to $1/3$, are allowed to be any probability $\delta \in (0,1]$), have been thoroughly studied in the theoretical computer science literature (and, in some form or the other, in the statistics literature), specifically in the broader area of *distribution testing*; and its sample complexity, as a function of all relevant parameters, is known up to constant factors to be

$$n(k, \varepsilon) = \Theta(\sqrt{k}/\varepsilon^2)$$

[1] Recall that the total variation distance between two distributions $\mathbf{p}, \mathbf{q}$ over domain $\mathcal{X}$ is given by $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) = \sup_S(\mathbf{p}(S) - \mathbf{q}(S))$, where the supremum is taken over all (measurable, but here we only deal with finite domains and the counting measure) subsets $S \subseteq \mathcal{X}$.

which is achieved by several different, efficient algorithms. See, e.g., Canonne (2020); Balakrishnan and Wasserman (2018); Goldreich (2017) and Canonne (2022) for surveys and a detailed monograph about this question.

However, this formulation is rather coarse, as it does not allow the sample complexity to depend on the specific reference distribtuion $\mathbf{q}$ (but instead is worse-case over all possible distributions on a domain of size k)[2]. This leaves out a lot of nuance and possible improvements: intuitively, if $\mathbf{q}$ has small support, or is concentrated over very few elements (say, $k' \ll k$), then testing identity to $\mathbf{q}$ should depend on k', not on k! Moreover, this rules out testing identity over infinite (but discrete) domains, such as $\mathcal{X} = \mathbb{N}$. To address this, Valiant and Valiant (2014) introduced *instance-optimal* identity testing (or, as Goldreich (2017) prefers to term it, *massively parameterized* identity testing), where the sample complexity n is allowed to be a function of the reference distribution $\mathbf{q}$ itself, not of the domain size $k = |\mathcal{X}|$. For this formulation to make sense, of course, one needs to identify a suitable, "succinct" functional[3] $\Phi\colon \Delta(\mathcal{X}) \times (0,1] \to \mathbb{N}$ of the reference distribution (and the distance ε), to avoid tautological statements of the kind *"the sample complexity is the least number of i.i.d. samples required to solve the question with respect to reference* $\mathbf{q}$*."*

1.1. The Valiant–Valiant result.

Valiant and Valiant, in their influential paper, provided upper and lower bounds on the massively parameterized identity testing question, where they expressed the sample complexity in terms of the "$\ell_{2/3}$ norm" of some (truncated version) of the reference distribution $\mathbf{q}$. Specifically, identifying any discrete probability distribution $\mathbf{p}$ over $\mathcal{X}$ (without loss generality, $\mathcal{X} = \mathbb{N}$) with its probability mass function, a sequence $\mathbf{p} \in [0,1]^{\mathbb{N}}$ such that $\|\mathbf{p}\|_1 = 1$, consider the following functional $\Phi_{\mathrm{VV}}(\mathbf{p}, \varepsilon)$:

1. Let $v(\mathbf{p}) \in [0,1]^{\mathbb{N}}$ be the non-increasing sequence obtained by sorting $\mathbf{p}$ (breaking ties arbitrarily);

2. Let $v(\mathbf{p})^{-\max} \in [0,1]^{\mathbb{N}}$ be the sequence obtained from $v(\mathbf{p})$ by removing the first element (highest value) of the sequence;

3. Let $v(\mathbf{p})^{-\max}_{-\varepsilon} \in [0,1]^{\mathbb{N}}$ be the sequence obtained from $v(\mathbf{p})^{-\max}$ by zeroing out all elements $v(\mathbf{p})^{-\max}_i$ for $i > i(\varepsilon)$, where $i(\varepsilon) := \min\{j : \sum_{i=j+1}^{\infty} v(\mathbf{p})^{-\max}_i \leq \varepsilon\}$.

4. Set $\Phi_{\mathrm{VV}}(\mathbf{p}, \varepsilon) := \|v(\mathbf{p})^{-\max}_{-\varepsilon}\|_{2/3}$, where $\|x\|_{2/3} = \left(\sum_{i=1}^{\infty} |x_i|^{2/3}\right)^{3/2}$.

In other words, given a probability distribution $\mathbf{p}$ over $\mathbb{N}$ and a value $\varepsilon \in (0,1]$, $\Phi_{\mathrm{VV}}(\mathbf{p}, \varepsilon)$ is obtained by (1) sorting $\mathbf{p}$, (2) removing its largest element and its "ε-tail, and (3) computing the $2/3$-quasinorm of the resulting vector.

Valiant and Valiant then provided upper and lower bounds[4] for the question, showing that the sample complexity $n(\mathbf{q}, \varepsilon)$ of massively parameterized identity testing satisfies

$$\Omega\left(\max\left(\frac{1}{\varepsilon}, \frac{\Phi_{\mathrm{VV}}(\mathbf{q}, \varepsilon)}{\varepsilon^2}\right)\right) \leq n(\mathbf{q}, \varepsilon) \leq O\left(\max\left(\frac{1}{\varepsilon}, \frac{\Phi_{\mathrm{VV}}(\mathbf{q}, \varepsilon/16)}{\varepsilon^2}\right)\right) \tag{1}$$

[2] This worst case is achieved for the uniform distribution over k elements.

[3] Here, $\Delta(\mathcal{X})$ denotes the probability simplex, that is, the set of probability distributions over $\mathcal{X}$.

[4] We believe there is a small gap in the proof of their lower bound, in cases where the reference distribution $\mathbf{q}$ is nearly degenerate (namely, near-point mass); see Section 4 for more details.

So, are we done? Not quite. While the above upper and lower bounds do look very similar, and indeed for many natural choices of $\mathbf{q}$ will indeed be within constant factors, the difference between $\Phi_{\mathrm{VV}}(\mathbf{q}, \varepsilon)$ and $\Phi_{\mathrm{VV}}(\mathbf{q}, \varepsilon/16)$ (due to the factor 16) can be made arbitrarily large! For instance, as observed in Blais et al. (2017), for every $k \geq 2$ there exists a relatively simple reference distribution $\mathbf{q}$ over k elements[5] such that

$$\Phi_{\mathrm{VV}}(\mathbf{q}, \varepsilon) = k^{\frac{1-\varepsilon}{2}+o(1)}$$

That is, the ε is *in the exponent*, and so $\Phi_{\mathrm{VV}}(\mathbf{q}, \varepsilon/16)$ and $\Phi_{\mathrm{VV}}(\mathbf{q}, \varepsilon)$ differ by a $k^{\Theta(\varepsilon)}$ factor.

2. A different characterization: the Blais–Canonne–Gur result

In later work, Blais et al. (2017) obtained a different set of upper and lower bounds for massively parameterized identity testing, in terms of the so-called K-function between ℓ_1 and ℓ_2 spaces. Namely, for a sequence $a \in \ell_1 + \ell_2$, define $\kappa_a \colon (0, \infty) \to [0, \infty)$ by

$$\kappa_a(t) = \inf_{\substack{(a',a'') \in \ell_1 \times \ell_2 \\ a'+a''=a}} \|a'\|_1 + t\|a''\|_2 \tag{2}$$

and let $\Phi_{\mathrm{BCG}}(\mathbf{p}, \varepsilon) := \kappa_{\mathbf{p}}^{-1}(1 - 2\varepsilon)$ for $\varepsilon \in (0, 1/2]$. Then Blais et al. (2017, Theorem 30,Theorem 33) establish that the sample complexity $n(\mathbf{q}, \varepsilon)$ of massively parameterized identity testing satisfies

$$\Omega\left(\frac{\Phi_{\mathrm{BCG}}(\mathbf{q}, \varepsilon)}{\varepsilon}\right) \leq n(\mathbf{q}, \varepsilon) \leq O\left(\max\left(\frac{1}{\varepsilon}, \frac{\Phi_{\mathrm{BCG}}(\mathbf{q}, \varepsilon/18)}{\varepsilon^2}\right)\right) \tag{3}$$

and the upper and lower bounds are incomparable to those of (1) (see Blais et al. (2017, Section 6.3)). What's more, the upper and lower bounds do exhibit the same type of quantitative gaps as the Valiant–Valiant result (also shown, for instance, for the same "Harmonic distribution"). Thus, Φ_{BCG} does not fully characterize the sample complexity of the question either!

3. The work of Chhor and Carpentier

In Chhor and Carpentier (2022), the authors consider a generalization of the massively parameterized identity testing to consider other distance metrics (i.e., all ℓ_p norms for $p \in [1, 2]$, instead of the total variation distance which corresponds to $p = 1$), and establish results analogous to Valiant and Valiant (2014) for this range of metrics. Their focus, instead of the sample complexity, is the dual quantity: namely, the optimal separation radius $\varepsilon = \varepsilon(\mathbf{q}, n, p)$ as a function of the reference distribution $\mathbf{q}$, the number of samples n, and the distance parameter p. The functional they obtain for $p = 1$ is closely related to that of Valiant and Valiant, as it involves the $2/3$-quasinorm of a truncated probability vector obtained from $\mathbf{q}$. The main difference is that the "tail" removed from $\mathbf{q}$ is with respect to the square of the probabilities, not the probabilities themselves (see Chhor and Carpentier (2022, Eq. (5)). We note that due to their focus on the separation radius, the result obtained in Chhor and Carpentier (2022, Theorem 1) presents, once translated into sample complexity, the same shortcoming as that of Valiant and Valiant and Blais, Canonne, and Gur: namely, they only characterize $\varepsilon = \varepsilon(\mathbf{q}, n, 1)$ up to constant factors, which leads to the same "gap" in the sample complexity as before (constant factors in the dependence on ε, which can affect the sample complexity by arbitrary amounts).

[5] Namely, the "Harmonic distribution" $\mathbf{q}$ such that $\mathbf{q}(i) \propto \frac{1}{i}$ for $1 \leq i \leq k$.

4. A small gap in the lower bound proof of Valiant and Valiant (for corner cases)

The lower bound argument given in Valiant and Valiant (2017, Proposition 2) (full version of Valiant and Valiant (2014)) applies their Corollary 1, stating (using our notation, where $\mathbf{q}$ is the reference distribution)

> "Letting m be the index at which $\mathbf{q}_i$ is maximized, consider the value of α for which $\frac{1}{2}\sum_{i\neq m}\min(\mathbf{q}_i, \alpha\mathbf{q}_i^{2/3}) = \varepsilon$, and [...]"

However, such a value of α is only guaranteed to exist when the maximum probability value of $\mathbf{q}$, namely $\mathbf{q}_m$, satisfies $\mathbf{q}_m \leq 1-2\varepsilon$. Indeed, if $\mathbf{q}_m > 1-2\varepsilon$, then

$$\frac{1}{2}\sum_{i\neq m}\min(\mathbf{q}_i, \alpha\mathbf{q}_i^{2/3}) \leq \frac{1}{2}\sum_{i\neq m}\mathbf{q}_i = \frac{1}{2}(1-\mathbf{q}_m) < \varepsilon$$

and no setting of $\alpha > 0$ can achieve the desired equality. Now, this is only a corner case, and does not mean that the end result is incorrect – just that the proof as stated does not cover that case! *We do believe the theorem holds, even for such corner cases: proving it (Open Question 3) would be nice.*

Note that this is not quite a trivial case, whenever $\|\mathbf{q}\|_\infty \in [1-2\varepsilon, 1-\varepsilon]$, and in particular this is *not* always captured by the $\Omega(1/\varepsilon)$ part of the stated lower bound of Valiant and Valiant (2017, Proposition 2). As a small example: fr any given $\varepsilon \in (0,1/2)$ and $k \geq 2$, consider the distribution $\mathbf{q} = \mathbf{q}(\varepsilon)$ supported on $\{1,2,\dots,k\}$ such that

$$\mathbf{q}_1 = 1-\frac{3}{2}\varepsilon, \qquad \mathbf{q}_2 = \dots = \mathbf{q}_k = \frac{3\varepsilon}{2(k-1)}$$

Then $\Phi_{\mathrm{VV}}(\mathbf{q},\varepsilon) = \Theta(\varepsilon\sqrt{k})$. Now, one can (relatively) easily check that testing identity to $\mathbf{q}$ with distance parameter ε can be done with (and requires) $\Theta(\sqrt{k}/\varepsilon)$ samples; and the sample complexity lower bound promised by Proposition 2 is, indeed, $\Omega\Big(\max\Big(\frac{1}{\varepsilon}, \frac{\Phi_{\mathrm{VV}}(\mathbf{q},\varepsilon)}{\varepsilon^2}\Big)\Big) = \Omega\Big(\frac{\sqrt{k}}{\varepsilon}\Big)$.

5. The open questions

This leads us to our (three) open questions:

Open Question 1 *Is there a "succinct" functional $\Phi\colon \Delta(\mathcal{X})\times(0,1] \to \mathbb{N}$ which* fully *characterizes the sample complexity of massively parameterized identity testing, (or, say, up to constant factors)? If so, what is it? If not, can we prove such a quantity cannot exist?*

(Note that the "succinctness" requirement here is a little vague. Ideally, Φ should either have a closed form, or be efficiently computable given the explicit description of the reference distribution $\mathbf{q}$.)

Open Question 2 *Put together, (1) and (3) imply a relation (and some inequalities) between the "2/3-quasinorm" involved in Φ_{VV} and the K-functional appearing in Φ_{BCG}. However, this relation goes through a very contrived argument, involved a statistical hypothesis testing problem! Is there a direct, non-contrived proof that these two quantities are related?*

Open Question 3 *Fix the (small) gap in the lower bound proof of Valiant and Valiant (2014), described in Section 4.*

Prizes. For a resolution of Open Question 1, we offer \$200 (AUD). For a resolution of Open Question 2, we offer an extra large wombat plush toy. For a resolution of Open Question 3, we offer four boxes of Tim Tams and a jar of Vegemite.

References

Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: a selective review. *Ann. Appl. Stat.*, 12(2):727–749, 2018. ISSN 1932-6157. doi: 10.1214/18-AOAS1155SF. URL https://doi.org/10.1214/18-AOAS1155SF.

Eric Blais, Clément L. Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. In *CCC*, volume 79 of *LIPIcs*, pages 28:1–28:40. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

Clément L. Canonne. *A Survey on Distribution Testing: Your Data is Big. But is it Blue?* Number 9 in Graduate Surveys. Theory of Computing Library, 2020. doi: 10.4086/toc.gs.2020.009. URL http://www.theoryofcomputing.org/library.html.

Clément L. Canonne. Topics and techniques in distribution testing: A biased but representative sample. *Found. Trends Commun. Inf. Theory*, 19(6):1032–1198, 2022.

Julien Chhor and Alexandra Carpentier. Sharp local minimax rates for goodness-of-fit testing in multivariate binomial and Poisson families and in multinomials. *Math. Stat. Learn.*, 5(1-2):1–54, 2022. ISSN 2520-2316. doi: 10.4171/msl/32. URL https://doi.org/10.4171/msl/32.

Oded Goldreich. *Introduction to property testing*. Cambridge University Press, Cambridge, 2017. ISBN 978-1-107-19405-2. doi: 10.1017/9781108135252. URL https://doi.org/10.1017/9781108135252.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *55th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2014*, pages 51–60. IEEE Computer Soc., Los Alamitos, CA, 2014. doi: 10.1109/FOCS.2014.14. URL https://doi.org/10.1109/FOCS.2014.14.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM J. Comput.*, 46(1):429–455, 2017. ISSN 0097-5397. doi: 10.1137/151002526. URL https://doi.org/10.1137/151002526.