Private goodness-of-fit(s) of discrete distributions (a very short review)

Clément L. Canonne

April, 2020

Abstract

The goal of this short note is to provide a short overview of the sample complexity of *identity testing* (also known as one-sample testing, or goodness-of-fit) under various types of privacy constraints, and map the current landscape in view of the flurry of recent works on this question.

The main focus of this document is the question of *identity testing* (i.e., one-sample testing, or goodness-of-fit) of probability distributions over a known discrete domain of size k[1] under various privacy constraints. The reader interested in discovering more on this formulation and similar questions absent the privacy component is referred to the recent surveys [Can15, BW18] on distribution testing.

Our identity problem. Identity testing is the question of deciding, based on observing a sequence of i.i.d. observations from some unknown probability distribution, whether this distribution conforms to a purported (and fixed in advance) model – or, in the contrary, is statistically quite far from this model. Formally, it is defined as follows:

Definition 1 (Identity Testing)**.** Given a fixed, known distribution $\mathbf{q}$ over $[k]$, an *identity testing algorithm for* $\mathbf{q}$ *with sample complexity* n takes as input a parameter $\alpha \in (0, 1]$ and n i.i.d. samples from an unknown distribution $\mathbf{p}$ over $[k]$, and outputs either accept or reject. The algorithm must satisfy the following, where the probability is over the randomness of the samples:

- If $\mathbf{p} = \mathbf{q}$, then the algorithm outputs accept with probability at least $2/3$;
- If $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) > \alpha$, then the algorithm outputs reject with probability at least $2/3$.

The sample complexity of identity testing to $\mathbf{q}$ is then the minimum sample complexity over all identity testing algorithms for $\mathbf{q}$; and the *sample complexity of identity testing* is the maximum sample complexity, over all reference distributions $\mathbf{q}$.

A couple remarks are in order: first, the above can be rephrased as a composite hypothesis testing (in a minimax setting), where $\mathcal{H}_0 = \{\mathbf{q}\}$ and $\mathcal{H}_1 = \{ \mathbf{p} : \mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) > \alpha \}$. Second, for simplicity, we focused in the above on a constant error probability (equal for both Type I and Type II), set to $1/3$. By standard arguments, one can in all settings considered here decrease this to an arbitrarily small $\beta \in (0, 1]$ at the price of a mere multiplicative $\log(1/\beta)$ factor in the sample complexity,[2] by repeating the test independently and taking the majority outcome.

[1] Without loss of generality, the set $[k] = \{1, 2, \dots, k\}$

[2] Which is not optimal, as a $\sqrt{\log(1/\beta)}$ is achievable instead [DGPP18]; but is good enough.

Our different Differential Privacies. We will consider this problem in seven different settings of privacy,[3] outlined (informally) below. All are parameterized by a *privacy parameter* $\varepsilon > 0$, which we will think of as being in $(0, 1]$: the smaller the ε, the better the privacy guarantee.

No privacy: The n samples from the unknown distribution $\mathbf{p}$ are held by n different users, who send their data to a central server running a testing algorithm whose output is then revealed to the world, no constraints enforced. *Everyone fully trusts everyone.*

(Central) differential privacy: The n samples from the unknown distribution $\mathbf{p}$ are held by n different users, who send their data to a central server running a testing algorithm whose output is then revealed to the world, under the constraint that this output does not reveal too much about any single user's data. *Users fully trust the server, but not the outside world.* Introduced in [DMNS06].

Local differential privacy: The n samples from the unknown distribution $\mathbf{p}$ are held by n different users, who based on their sample send a message to a central server running a testing algorithm whose output is then revealed to the world. The constraint that any user's message to the server does not reveal too much about this user's data. *Users trust neither the server nor the outside world.* Introduced in [KLN+11]; later reformulated in [DJW13].
We will further consider three subsettings, depending on whether the users share some additional common random seed (e.g., broadcast ahead of time by the central server) or send all their messages simultaneously or sequentially:[4]

- **Private-coin:** The users only have their own personal (trusted) randomness, and don't have anything in common. All send their message in parallel.
- **Public-coin:** The users have their own personal (trusted) randomness, as well as a common shared (not necessarily trusted) random seed. All send their message in parallel.
- **Interactive:** The users have their own personal (trusted) randomness. They send their message sequentially, so that the i-th user is aware of the messages sent by users $1, 2, \dots, i-1$ (in particular, they can use this to also have a common shared random seed).

Pan-privacy against one intrusion: The n samples from the unknown distribution $\mathbf{p}$ are held by n different users, who independently send their data one by one to a central server running a testing algorithm whose output is then revealed to the world. The constraint is that any adversary that breaches the server to look at the internal state of the algorithm *at most once* does not learn too much about any single user's data. *Users trust the server at the time they send their data*, but are not too sure the server will not be compromised in the future; and *definitely do not trust the outside world.* Introduced in [DNP+10].

Shuffle privacy: The n samples from the unknown distribution $\mathbf{p}$ are held by n different users, who based on their sample send some messages to a trusted third party which shuffles all the messages. The third party then sends the (permuted) messages to a central server running a testing algorithm whose output is then revealed to the world. The constraint is that the list of shuffled messages does not reveal too much about any single user's data, even when a small fraction of users are corrupted and deviate adversarially from the protocol.[5] *Users trust the third party, but not the server nor the outside world.* Introduced in [CSU+19]; see [Che20] for a useful timeline.

Of all these notions, aside from the "no privacy" one, the central differential privacy (DP) is the least stringent in terms of privacy while the local differential privacy (LDP) is the most. Pan-privacy and shuffle

[3] There are quite a few other notions, or variants of the ones listed here, which we will not touch upon in this note.

[4] We will not discuss here the setting of *fully interactive* Local DP, which allows for arbitrary (not just sequential) rounds of messages. Indeed, to the best of our knowledge there is no identity testing result specific to this setting (except, of course, the upper bounds from the models mentioned here, which of course carry over).

[5] We note that the earlier definitions of shuffle privacy did not include that last, arguably very natural, robustness requirement. As discussed in [BCJM20] (and in said earlier literature), this is, however, a natural condition and how one ought to think of shuffle privacy: if one user goes amok, the privacy of everyone else should not immediatey be threatened.

privacy are somewhere in the middle. Of course, better privacy comes at a price: DP typically allows for much more sample-efficient algorithms, while LDP requires a rather enormous sample size for the same task/utility.

The lay of the land. We now summarize what is known about identity testing in the above privacy models, and point to the papers where the bounds were established.

	Upper bound (UB)	**Lower bound (LB)**	**References**
No privacy	$\frac{k^{1/2}}{\alpha^2}$		[VV17] (UB), [Pan08] (LB)
Central DP	$\frac{k^{1/2}}{\alpha^2}+\frac{k^{1/2}}{\alpha\varepsilon^{1/2}}+\frac{k^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}}+\frac{1}{\alpha\varepsilon}$		[ASZ18, ADR18]
Local DP, private-coin	$\frac{k^{3/2}}{\alpha^2\varepsilon^2}$		[ACFT19, ACT19b, ACH$^+$19][6]
Local DP, public-coin	$\frac{k}{\alpha^2\varepsilon^2}$		[ACFT19] (UB, LB) [ACT19b] (LB)
Local DP, interactive	$\frac{k}{\alpha^2\varepsilon^2}$		[ACFT19] (UB), [AJM19] (LB)
Pan-privacy	$\frac{k^{1/2}}{\alpha^2}+\frac{k^{2/3}}{\alpha^{4/3}\varepsilon^{2/3}}+\frac{k^{1/2}}{\alpha\varepsilon}$	$\frac{k^{1/2}}{\alpha^2}+\frac{k^{2/3}}{\alpha^{4/3}\varepsilon^{2/3}}+\frac{1}{\alpha\varepsilon}$	[AJM19]
Shuffle privacy	$\left(\frac{k^{1/2}}{\alpha^2}+\frac{k^{2/3}}{\alpha^{4/3}\varepsilon^{4/3}}+\frac{k^{1/2}}{\alpha\varepsilon^2}\right)\log^{1/2}k$	$\frac{k^{1/2}}{\alpha^2}+\frac{k^{2/3}}{\alpha^{4/3}\varepsilon^{2/3}}+\frac{1}{\alpha\varepsilon}$	[BCJM20]

Table 1: The current landscape of identity testing, in the various models of privacy outlined above. 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.

It is worth noting that some of the papers referenced above do not claim to address the general case of identity testing, focusing instead on the special case of *uniformity* testing, wher the reference distribution is uniform. However, an argument of Diakonikolas and Kane [DK16] and Goldreich [Gol20], generalized to various settings by Acharya, Canonne, and Tyagi [ACT19a, Appendix A], shows that identity testing is essentialy equivalent to this special case of uniformity testing.

References

[ACFT19] Jayadev Acharya, Clément L. Canonne, Cody Freitag, and Himanshu Tyagi. Test without trust: Optimal locally private distribution testing. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pages 2067–2076. PMLR, 2019.

[6] [ACFT19] establishes the upper bound, and shows the matching lower bound for some special cases of protocols (algorithms). [ACT19b] proves the lower bound for all private-coin LDP protocols. [ACH$^+$19] provides an alternative protocol achieving the upper bound, which significantly improves on the amount of communication (message length) required.

[ACH$^+$19] Jayadev Acharya, Clément L. Canonne, Yanjun Han, Ziteng Sun, and Himanshu Tyagi. Domain compression and its application to randomness-optimal distributed goodness-of-fit. *CoRR*, abs/1907.08743, 2019.

[ACT19a] Jayadev Acharya, Clement Canonne, and Himanshu Tyagi. Communication-constrained inference and the role of shared randomness. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 30–39, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[ACT19b] Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Inference under information constraints: Lower bounds from chi-square contraction. In *COLT*, volume 99 of *Proceedings of Machine Learning Research*, pages 3–17. PMLR, 2019.

[ADR18] Maryam Aliakbarpour, Ilias Diakonikolas, and Ronitt Rubinfeld. Differentially private identity and equivalence testing of discrete distributions. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 169–178, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[AJM19] Kareem Amin, Matthew Joseph, and Jieming Mao. Pan-private uniformity testing. *CoRR*, abs/1911.01452, 2019.

[ASZ18] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private testing of identity and closeness of discrete distributions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6878–6891. Curran Associates, Inc., 2018.

[BCJM20] Victor Balcer, Albert Cheu, Matthew Joseph, and Jieming Mao. Connecting robust shuffle privacy and pan-privacy. *CoRR*, abs/2004.09481, 2020.

[BW18] Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: a selective review. *Ann. Appl. Stat.*, 12(2):727–749, 2018.

[Can15] Clément L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, 2015. To appear as a graduate student survey in Theory of Computing.

[Che20] Albert Cheu. The Sudden Surge of Shuffling in the Privacy Literature, 2020. Online; accessed 24 April 2020.

[CSU$^+$19] Albert Cheu, Adam Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. In *Advances in cryptology—EUROCRYPT 2019. Part I*, volume 11476 of *Lecture Notes in Comput. Sci.*, pages 375–403. Springer, Cham, 2019.

[DGPP18] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *45th International Colloquium on Automata, Languages, and Programming*, volume 107 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages Art. No. 41, 14. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2018.

[DJW13] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *54th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2013*, pages 429–438. IEEE Computer Society, 2013.

[DK16] Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*, pages 685–694. IEEE Computer Soc., Los Alamitos, CA, 2016.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer.

[DNP+10] Cynthia Dwork, Moni Naor, Toniann Pitassi, Guy N. Rothblum, and Sergey Yekhanin. Pan-private streaming algorithms. In *ICS*, pages 66–80. Tsinghua University Press, 2010.

[Gol20] Oded Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. In *Computational Complexity and Property Testing*, volume 12050 of *Lecture Notes in Computer Science*, pages 152–172. Springer, 2020.

[KLN+08] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. What can we learn privately? In *49th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2008*, pages 531–540. IEEE Computer Society, 2008.

[KLN+11] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826, 2011. Journal version of [KLN+08].

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inform. Theory*, 54(10):4750–4755, 2008.

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

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