Better Private Distribution Testing by Leveraging Unverified Auxiliary Data

Maryam Aliakbarpour MARYAMA@RICE.EDU
Department of Computer Science & Ken Kennedy Institute, Rice University

Arnav Burudgunte ABURUDGU@PURDUE.EDU
Purdue University

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

Ronitt Rubinfeld RONITT@CSAIL.MIT.EDU
Computer Science and Artificial Intelligence Laboratory, MIT

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We extend the framework of augmented distribution testing (Aliakbarpour, Indyk, Rubinfeld, and Silwal, NeurIPS 2024) to the differentially private setting. This captures scenarios where a data analyst must perform hypothesis testing tasks on sensitive data, but is able to leverage prior knowledge (public, but possibly erroneous or untrusted) about the data distribution.

We design private algorithms in this augmented setting for three flagship distribution testing tasks, *uniformity*, *identity*, and *closeness* testing, whose sample complexity smoothly scales with the claimed quality of the auxiliary information. We complement our algorithms with information-theoretic lower bounds, showing that their sample complexity is optimal (up to logarithmic factors).

Keywords: distribution testing, identity testing, closeness testing, differential privacy, learning-augmented algorithms

1. Introduction

Accurately analyzing data while preserving individual privacy is a fundamental challenge in statistical inference. Since its formulation nearly two decades ago, Differential Privacy (DP) (Dwork et al., 2006) has emerged as the leading framework for privacy-preserving data analysis, providing strong mathematical privacy guarantees and gaining adoption by major entities such as the U.S. Census Bureau, Amazon (Amazon Web Services, 2024), Google (Erlingsson et al., 2014), Microsoft (Ding et al., 2017), and Apple (Differential Privacy Team, Apple, 2017; Thakurta et al., 2017).

Unfortunately, DP guarantees often come at the cost of increased data requirements or computational resources, which has limited the widespread adoption of differential privacy in spite of its theoretical appeal. To address this issue, a recent line of work has investigated whether access to even small amounts of additional *public* data could help mitigate this loss of performance. Promising results for various tasks have been shown, both experimentally (Kerrigan et al., 2020; Lowy et al., 2024; Bu et al., 2024; Daum et al., 2024) and theoretically (Bie et al., 2022; Ben-David et al., 2023). The use of additional auxiliary information is very enticing, as such access is available in many real-world applications: for example, hospitals handling sensitive patient data might leverage public datasets, records from different periods or locations, or synthetic data generated by machine learning models to improve analysis. Similarly, medical or socio-econonomic studies focusing on a minority or protected group can leverage statistical data from the overall population.

However, integrating public data introduces its own challenges, as it often lacks guarantees regarding its accuracy or relevance to private datasets. If the external data distribution deviates significantly from the target population, then it becomes at best useless for the task at hand, and at worst misleading; yet assessing its reliability prior to using it can be as complex as the original inference task. This leads to the following question, which is the focus of our work:

How can we design private algorithms which optimally utilize auxiliary information in a way that seamlessly adapts to its (unknown, arbitrary) quality?

We formalize the question by extending the recently proposed framework of "augmented testing" (Aliakbarpour et al., 2024) to the differentially private setting (see Section 2 for formal definitions). In our new framework, the testing algorithm is provided with the auxiliary information as an "advice" probability distribution $\hat{p}$, purported to be a good approximation of the unknown data distribution p, along with a claimed accuracy α of this approximation. The algorithm must correctly solve the inference task when the advice is indeed α-accurate, but it is allowed to abort and return a "failure" symbol when it detects that the advice does not meet the claimed accuracy α. The name of the game is to achieve better data utilization as a function of this new accuracy parameter α by leveraging the advice when it is good, yet not being fooled by it when it is bad enough to derail the algorithm.

Our new model must take into account the privacy constraints of the samples. While there is no need to consider the privacy of sample data $\hat{p}$ that is already public, we must ensure that any other sample data remains private, even if our algorithm decides to abort due to bad advice.

Finally, for the purpose of this paper, we primarily focus on distribution testing (which can be viewed as a finite-sample, computational take on hypothesis testing), the cornerstone of many scientific endeavors. However, we believe that differentially private inference using (unverified) auxiliary public data is a promising avenue for work beyond the specific tasks we consider here.

1.1. Related work

Distribution testing has been extensively studied since its first formulation in Goldreich et al. (1998); Batu et al. (2000); we only refer below to the papers most relevant to us, and refer the reader to the expositions, textbooks, and surveys (Rubinfeld, 2012; Goldreich, 2017; Canonne, 2020, 2022) for a more extensive coverage. *Uniformity testing*, that is, the task of deciding if an unknown distribution over a given domain of size n is uniform, or at distance at least ε from it (in total variation), was shown to have sample complexity $\Theta(\sqrt{n}/\varepsilon^2)$ in Goldreich and Ron (2000, 2011); Paninski (2008); Acharya et al. (2015); Diakonikolas et al. (2018). It was shown to be equivalent to the more general task of *identity* testing, where the reference distribution is allowed to be arbitrary (instead of uniform) in Diakonikolas (2016); Goldreich (2020). The sample complexity of *closeness* testing, where the algorithm has access to samples from two unknown distributions and must decide whether they are equal or at total variation distance at least ε, was settled in a series of works (Batu et al., 2000; Valiant, 2011; Chan et al., 2014; Diakonikolas et al., 2021; Canonne and Sun, 2022), culminating in the tight sample complexity $\Theta(\sqrt{n}/\varepsilon^2 + n^{2/3}/\varepsilon^{4/3})$.

Beyond this standard distribution testing setting, the work closest to ours is the recent paper of Aliakbarpour et al. (2024), which introduced the (non-private) framework of augmented testing for distributions. While there has been before a large body of work on learning-augmented

algorithms (and previously, on algorithms with advice) (e.g., Hsu et al. (2019); Indyk et al. (2019); Jiang et al. (2020); Aamand et al. (2023); Bhattacharyya et al. (2024)), the specific formulation whereby the algorithm is provided with advice in the form of a hypothesis probability distribution, *and is allowed to abort if it detects this advice is incorrect*, is crucial to obtaining non-vacuous formulations of the problem. While Aliakbarpour et al. (2024) do address identity (and uniformity) as well as closeness testing of distributions in their augmented testing framework, they only consider the non-private versions of these tasks: introducing the constraint of differential privacy on the data (but not on the advice itself, considered public) is the main novelty of our work, and comes with a host of technical and conceptual challenges to overcome. We refer the reader to the website https://algorithms-with-predictions.github.io/, for more on learning-augmented algorithms.

On the privacy-preserving side, there is a large literature on DP distribution testing, where the samples provided to the algorithm are considered sensitive and the algorithm must satisfy the constraint of (central) DP, notably Cai et al. (2017); Acharya et al. (2018); Aliakbarpour et al. (2018, 2019). The tight private sample complexity of uniformity, identity,[1] and closeness testing under ξ-DP are now known to be $\Theta(\sqrt{n}/\varepsilon^2 + \sqrt{n}/(\varepsilon\sqrt{\xi}) + n^{1/3}/(\varepsilon^{4/3}\xi^{2/3}) + 1/(\varepsilon\xi))$ and $\Theta(\sqrt{n}/\varepsilon^2 + n^{2/3}/\varepsilon^{4/3} + \sqrt{n}/(\varepsilon\sqrt{\xi}) + n^{1/3}/(\varepsilon^{4/3}\xi^{2/3}) + 1/(\varepsilon\xi))$, respectively (Zhang, 2021).

As mentioned earlier, a recent line of work has investigated the use of additional public data for DP inference questions, and in particular for distribution learning questions (Bie et al., 2022; Ben-David et al., 2023). We emphasize that while these results can accommodate a (small) amount of *distribution shift* between the private and public data distributions, they assume that the publicly available data *is* accurate (comes from the same distribution as, or one very close to, the the sensitive data). This is a crucial limiting assumption, and one our paper does not make: in our setting, we do not have to trust the quality of the auxiliary data. Our framework for distribution testing bears some resemblance to the framework of Khodak et al. (2023) for learning-augmented private quantile estimation, in which the algorithm must be correct (and efficient) even if the prediction is poor.

Another important difference is in the parameter regime: (Bie et al., 2022; Ben-David et al., 2023) consider the setting where the quantity of public data is small compared to the amount of public data: that is, they investigate whether the algorithm can benefit from a few "useful" public data points, in order to perform its analysis on a much larger sensitive dataset. In contrast, we focus on the setting where the public data abounds (so that even using it to get a full probability distribution $\hat{p}$ as advice is possible), but the private data is much scarcer.

The broader concept of learning from both public and private data has had some success beyond distribution learning. On the experimental side, prior work has pre-trained models on public data and then used private data for finetuning (Li et al., 2022; Bu et al., 2024). On the theoretical side, there has been work on using public data to improve the efficiency of learning algorithms (Block et al., 2024) and statistical upper and lower bounds for learning when only some of the data is private (Alon et al., 2019).

1.2. Our results

Our main results are sample-efficient private algorithms for identity and closeness testing in the augmented setting, complemented by (nearly)-matching information-theoretic lower bounds.

1. Note that the equivalence between uniformity and identity testing of Diakonikolas (2016); Goldreich (2020) carries over to the differentially private setting.

Theorem 1 (Informal version of Theorem 16 and Theorem 19) *There is an algorithm for private augmented identity testing of distributions over $[n]$ which, given a reference distribution q, privacy parameter $\xi > 0$, distance parameter $\varepsilon \in (0, 1]$, purported accuracy $\alpha \in (0, 1]$, as well as advice distribution $\hat{p}$, takes*

$$\Theta\left(\frac{\sqrt{n}}{\varepsilon^2} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} + \frac{n^{1/3}}{\varepsilon^{4/3}\xi^{2/3}} + \frac{1}{\varepsilon\xi}\right)$$

samples when the distance $\eta := \mathrm{d}_{\mathrm{TV}}(\hat{p}, q)$ satisfies $\eta \leq \alpha$, and

$$\Theta\left(\min\left(\frac{1}{(\eta-\alpha)^2} + \frac{1}{(\eta-\alpha)\xi}, \frac{\sqrt{n}}{\varepsilon^2} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} + \frac{n^{1/3}}{\varepsilon^{4/3}\xi^{2/3}} + \frac{1}{\varepsilon\xi}\right)\right)$$

when $\eta > \alpha$. Moreover, this is optimal.

Observe that in the "interesting regime" – i.e., when the advice suggests to the algorithm that it should reject, as the claimed distance between $\hat{p}$ and the unknown distribution p is strictly smaller than the distance between $\hat{p}$ and the reference distribution q – and, in particular, for vanishing α, the resulting sample complexity can become as small as $\Theta(1/\eta^2 + 1/(\eta\xi))$, leading to stark savings in the amount of private data required to conduct the task.

Note that the above theorem also applies to uniformity testing in a straightforward manner. In particular, by setting q as the uniform distribution, our identity testing algorithm immediately applies to uniformity testing, and our lower bound—proven under the assumption of uniform q—establishes the complete equivalence of these two problems.
Our results for the task of closeness testing are summarized in the next theorem:

Theorem 2 (Informal version of Theorem 6 and Theorem 38) *There is an algorithm for private augmented closeness testing of distributions over $[n]$ which, given privacy parameter $\xi > 0$, distance parameter $\varepsilon \in (0, 1]$, purported accuracy $\alpha \in (0, 1]$, as well as advice distribution $\hat{p}$ for one of the two distributions, takes*

$$\tilde{O}\left(\frac{n^{2/3}\alpha^{1/3}}{\varepsilon^{4/3}} + \frac{\sqrt{n}}{\varepsilon^2} + \frac{n^{1/3}}{\varepsilon^{4/3}\xi^{2/3}} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} + \frac{1}{\varepsilon\xi}\right)$$

samples. Moreover, this is optimal up to logarithmic factors in n.

Again, observe that the advice influences the first term of the sample complexity. As the claimed advice quality α goes to 0 (and ignoring logarithmic factors), our results show that, surprisingly, the task becomes essentially as simple as differentially private *identity* testing.

1.3. Overview of our technical contributions

Upper bounds. A popular technique for testing closeness of distributions p and q in total variation (or ℓ_1) distance is to first test ℓ_2 distance and translate the result into an ℓ_1 guarantee. Chan et al. (2014) use this technique to give an algorithm whose sample complexity is proportional to the minimum of the ℓ_2 norms of p and q. On a domain of size n, the ℓ_2 norm of a distribution can vary between $1/\sqrt{n}$ and 1, leading to a large difference between the best and worst case costs of testing. The *flattening technique*, introduced in Diakonikolas and Kane (2016), aims to shrink this gap by

reducing every instance of closeness testing to the aforementioned best case. The technique maps a distribution p to a "flattened distribution" p' with a low ℓ_2-norm. It accomplishes this by breaking any high-probability element into multiple new domain elements (buckets) and assigning its probability mass equally among them. The key problem is to determine the number of buckets required for each element i, which can vary depending on the problem's structure.

The original approach proposed in Diakonikolas and Kane (2016) is as follows. Suppose we have k samples from an unknown distribution p, referred to as *flattening samples*. Let k_i denote the frequency of element i among these samples. For each element i, we assign $k_i + 1$ buckets. Elements which appear more frequently (i.e., the large elements) are split into more buckets, reducing their individual probability mass. More formally, one can show that the resulting distribution p' has an ℓ_2-norm of at most $1/\sqrt{k}$ in expectation. (Note that we can simulate a sample from p' by drawing a fresh sample i from p and assigning it to a random bucket in $[k_i + 1]$.)

Why is this result significant for closeness testing (or related problems)? One can show that if the same flattening is applied to two distributions, their ℓ_1-distance remains unchanged. Thus, testing the closeness of p and q reduces to testing the closeness of their flattened versions p' and q'. Since the flattened distributions have a reduced ℓ_2-norm, we can leverage the ℓ_2-tester from Chan et al. (2014), whose sample complexity is proportional to the ℓ_2-norm of p, thereby achieving more sample-efficient testing. This technique is particularly appealing because it allows us to exploit problem-specific structure to design an effective flattening scheme. For example, Aliakbarpour et al. (2024) use a predicted distribution to guide the flattening: if the prediction assigns a higher probability to an element, more buckets are allocated to it, further reducing its ℓ_2-contribution. After flattening, they test whether the ℓ_2-norm of the new distribution p' is sufficiently small. If not, they conclude that the prediction quality was poor, and abort early. If the norm is small, the ℓ_2-tester from Chan et al. (2014) can be used with significantly fewer samples, yielding an augmented closeness tester with an optimal dependence on the prediction quality. The prediction allows the algorithm to achieve a useful flattening with fewer samples, leading to significant savings over the optimal non-augmented algorithm.

To achieve our goals, a natural approach is thus to privatize the algorithm from Aliakbarpour et al. (2024). Unfortunately, the sample-based flattening step introduces privacy challenges, because flattening via samples is highly sensitive: changing a single sample can alter the number of buckets for two elements, drastically changing the structure of the resulting distribution. To ensure privacy, the outcomes of these very different cases must be made similar, typically by adding a large amount of noise. Consequently, statistics computed on the flattened distribution would require a large number of samples to be privatized, leading to suboptimal algorithms.

To circumvent this issue, we need to reduce the sensitivity of the flattening step. To do so, we leverage the framework of Aliakbarpour et al. (2019), which focuses on privatizing a flattening-based tester – exactly what we need for our approach to work! However, this is not as straightforward as it seems. Because the advice distribution $\hat{p}$ used as a guide in the flattening step may be wildly inaccurate, we must verify that the outcome of the flattening step is satisfactory; that is, that the ℓ_2 norm of the flattened distribution is sufficiently small (as it should be, were the advice correct). To further complicate matters, this verification must be done *privately*.

Our solution is to apply flattening in two steps. First, we use the prediction $\hat{p}$ to flatten p into p'. Then we use samples drawn from p (more precisely, from p') to further flatten the distribution into p''. If $\hat{p}$ is α-close to the true distribution, we expect that the ℓ_2-norm of p'' is roughly less than α/k. We separate the two flattening steps due to our privacy requirements; since the prediction is public

information, there is no privacy concern in the first step, which in turn simplifies the analysis of the second step.

The standard statistic for computing the ℓ_2-norm is the number of collisions in a sample set (i.e., the number of pairs of equal samples) (Goldreich and Ron, 2011). This statistic is highly sensitive to changes in one sample. For example, consider estimating the ℓ_2-norm using ℓ samples. Suppose element i appears $\ell/2$ times. If we assign one bucket to i, the number of collisions is $\binom{\ell/2}{2}$. However, if we assign two buckets to i, we expect each bucket to receive roughly $\ell/4$ samples, so the total number of collisions is $2\binom{\ell/4}{2}$. This change reduces the number of collisions by roughly $\ell^2/16$, whereas ideally we would like this number to be a constant.

Thankfully, this worst-case only occurs in highly atypical scenarios. In general, if we draw k samples for flattening, we expect about kp_i buckets for element i. Then if we draw ℓ samples for estimation, we expect approximately ℓp_i samples from i. Distributing these equally among the buckets yields roughly ℓ/k samples per bucket. Assuming $k = \ell$, a typical dataset should have a constant number of samples per bucket. Consequently, altering either a flattening sample or an estimation sample would change the overall number of collisions by at most a constant factor. Hence, the sensitivity of the statistic should be low for a "typical" dataset.

The challenge is thus to pay only a privacy cost proportional to the sensitivity of a typical dataset, rather than the worst-case dataset. To this end, we identify the "bad events" that cause our dataset to have high sensitivity, which correspond to having a high unbalance between samples either within buckets, or between flattening and estimation sets. That is, we must ensure that the following holds for every element $i \in [n]$:

1. When the samples of element i are distributed among buckets, no bucket receives an unreasonably high number of samples.

2. The number of instances of i in the estimation sample set is not much larger than the number of instances of i in the flattening sample set.

To address the first issue, we mitigate the randomness inherent in distributing samples among buckets. The original statistic sums the collisions across buckets, where if $\ell_{i,j}$ denotes the number of samples in the j-th bucket of element i, the contribution is $\binom{\ell_{i,j}}{2}$. We replace this quantity with its expectation under the assumption of a uniform distribution of samples among buckets.

The second issue is more complex. We need to ensure that the number of instances of element i is approximately the same in both the flattening and estimation sample sets. To achieve this, we employ a technique from Aliakbarpour et al. (2019). Roughly speaking, each high-sensitivity sample set X is randomly mapped to another sample set X' such that any imbalance in the number of samples for an element is corrected. The mapping has the key property of preserving the Hamming distance between datasets. More precisely, if two datasets X_1 and X_2 (differing in one element) are mapped to low-sensitivity datasets Y_1 and Y_2, then there exists a coupling between Y_1 and Y_2 such that their Hamming distance is bounded by a constant. This property ensures that applying a private algorithm to the Y's (with low sensitivity) yields privacy guarantees for the original X's.

In our context, we ensure that the transformed dataset has a balanced number of instances for each element across both the estimation and flattening sets. We then run our private algorithm on this new sample set. If the original sample set already had low sensitivity, it remains unchanged, and we obtain the correct answer with high probability. If the dataset had high sensitivity, it is transformed

to ensure privacy; although this may render the statistical result unreliable, such high-sensitivity datasets occur with low probability.[2]

With these technical components, we obtain a low-sensitivity statistic for the ℓ_2-norm and apply the Laplace mechanism to ensure privacy. Plugging this into the above outline yields our main algorithmic result. A complete proof is given in Appendix C, with a short version is presented in Section 3.

The upper bound for identity and uniformity testing can be obtained via simple prioritization of the upper bound in Aliakbarpour et al. (2024) and the existing private algorithm for identity testing Acharya et al. (2018).

Lower bounds. For the lower bounds, the difficulties are reversed. The lower bound for private augmented closeness testing is derived by combining two existing lower bounds; further details are provided in Appendix D.

For uniformity testing, and consequently for identity testing, to show a lower bound, we must combine the statistical inference and privacy constraints on a private augmented tester. When $\hat{p}$ and q are closer than the suggested accuracy α, we show that private identity testing reduces to augmented private identity testing using a similar reduction to Aliakbarpour et al. (2024). In this case, known lower bounds (Zhang, 2021) automatically apply to our setting.

However, if $\hat{p}$ and q are more than α-far apart, we must combine several tools for proving information-theoretic lower bounds. In the non-augmented setting, a lower bound can be shown using a version of Le Cam's method for differential privacy (Acharya et al., 2018). Informally, the idea is to construct a coupling between two distributions (one close to the uniform distribution, one far) whose expected Hamming distance is low given too few samples. Since no answer will succeed for both, and no private algorithm can answer too differently on the two sets of samples, no algorithm can succeed with high probability. In our case, there are *three* possible answers, which requires us to construct three distributions such that no single answer works for all three. It also does not suffice to directly follow the approach of Aliakbarpour et al. (2024) and construct a multivariate coupling between the three distributions such that all three are close in Hamming distance. This would prove the privacy terms of our lower bound, but would not give the correct combination of statistical and privacy terms.

To avoid this problem, we use a general version of Le Cam's method under privacy and statistical constraints. Informally, our result states that if any algorithm requires s_1 samples to statistically distinguish two distributions with high probability and s_2 samples to guarantee that they are far in Hamming distance, it requires $\max(s_1, s_2)$ samples to distinguish them privately. We then construct a hard distribution which is ε-far from the uniform distribution, and a hard distribution which is α-close to the prediction $\hat{p}$. Any private augmented uniformity tester cannot answer similarly for all three distributions, so it must distinguish the uniform distribution from at least one of the hard distributions. Therefore, the sample complexity must satisfy at least one of the constraints imposed by our version of Le Cam's method. This explains why the optimal sample complexity is the minimum of two expressions, which matches our upper bound. For more details, see Appendix B.

2. For readers familiar with differential privacy, note that our approach differs from the well-known "Propose-Test-Release" mechanism, where noise is added proportional to the typical sensitivity (see Section 3 in Vadhan (2017)). In that mechanism, the probability of a high-sensitivity (bad) event is absorbed into an approximate DP guarantee via an additive error δ. However, our approach is more suitable for pure DP: our privacy guarantee remains intact, and the bad event only introduces a small probability of error in accuracy.

2. Preliminaries and Problem Setup

We use $[n]$ to indicate the set of integers $\{1, 2, \ldots, n\}$, and standard asymptotic notation ($O(\cdot), \Omega(\cdot), \Theta(\cdot)$) as well as the (slightly) less standard $\tilde{O}(\cdot)$, which omits polylogarithmic factors in its argument. In this paper, we focus on discrete probability distributions with a finite domain, which we identity with their probability mass functions (pmf). Let p denote a probability distribution over $[n]$: for any $i \in [n]$, we denote by p_i the probability of the element i according to p. For a subset $S \subset [n]$, $p(S)$ then denotes the probability of observing an element in S according to p: $p(S) = \sum_{i \in S} p_i$. The distribution of s i.i.d. samples from p is written as $p^{\otimes s}$. If we know the probability p_i of each element, we say p is a *known distribution*. We also use the following standard probability distributions: $\mathbf{Poi}(\lambda)$ denotes a Poisson distribution with mean λ; $\mathbf{Bern}(x)$ denotes a Bernoulli distribution with success probability x; and $\mathbf{Lap}(b)$ denotes a Laplace distribution with scale b.

The total variation distance between p and q is defined as the maximum possible difference in probabilities assigned to any event (subset of outcomes):

$$\mathrm{d}_{\mathrm{TV}}(p, q) \coloneqq \max_{S \subseteq [n]} |p(S) - q(S)| = \frac{1}{2}\|p - q\|_1 \in [0, 1].$$

where for the last expression we identify the pmf of p, q with their vector of probabilities. For two distributions p and q, and any $\varepsilon \in [0, 1]$, we say p is ε-close to q iff $\mathrm{d}_{\mathrm{TV}}(p, q)$ is at most ε. We say p and q are ε-far from each other iff $\mathrm{d}_{\mathrm{TV}}(p, q)$ is greater than ε. A property $\mathcal{P}$ is a set of distributions. We say a distribution p has property $\mathcal{P}$ if $p \in \mathcal{P}$, and p is ε-far from property $\mathcal{P}$ if $\mathrm{d}_{\mathrm{TV}}(p, q) > \varepsilon$ for all $q \in \mathcal{P}$.

Problem setup: The standard distribution testing problem is to decide, given samples from an unknown distribution p, whether p has a property $\mathcal{P}$ or whether p is ε-far from property $\mathcal{P}$. In our case, the tester receives three additional inputs: a privacy parameter ξ, which it must respect, a predicted distribution (or "advice") $\hat{p}$, and a suggested accuracy value α. If $\hat{p}$ is α-close to p, the tester must either output accept or reject with high probability; otherwise, it may output $\perp$, which denotes inaccurate information.

Definition 3 (Augmented private tester) *Suppose we are given three parameters $\varepsilon, \alpha \in (0, 1)$, and $\xi > 0$, along with distributions p and $\hat{p}$ over the same domain. Suppose $\mathcal{A}$ receives samples from p and returns a value in $\{\mathsf{accept}, \mathsf{reject}, \perp\}$. We say algorithm $\mathcal{A}$ is a $(\xi, \varepsilon, \alpha, \delta)$-private augmented tester for property $\mathcal{P}$ iff the following holds:*

- *$\mathcal{A}$ is a ξ-differentially private algorithm.*
- *If $p \in \mathcal{P}$, $\mathcal{A}$ does not output* reject *with probability more than δ.*
- *If p is ε-far from $\mathcal{P}$, $\mathcal{A}$ does not output* accept *with probability more than δ.*
- *If p and $\hat{p}$ are α-close, $\mathcal{A}$ does not output $\perp$ with probability more than δ.*

Remark 4 (On knowing the parameter α) *While it may seem difficult to suggest an accuracy level α for an unverified prediction, this is not an actual limitation, as one can avoid this in practice using the search algorithm of Aliakbarpour et al. (2024). The algorithm iteratively increases α until the tester either returns* accept *or* reject. *If the true accuracy is α^* and the tester uses $f(\alpha)$ samples for a single run with input α, the search algorithm uses $O(f(\alpha^*))$ total samples in expectation.*

While we do not provide the full proof here, careful analysis shows that a private version of this claim can be established under certain constraints. The main challenge is determining the stopping point at which the tester privately returns either accept *or* reject. *This can be accomplished using the well-known Sparse Vector Technique, which does not increase the privacy cost by more than a constant factor, provided that the stopping point can be replaced with a threshold query to a low-sensitivity function. Notably, this condition holds for all the algorithms in this paper.*

We will refer to a non-private augmented tester as a $(\varepsilon, \alpha, \delta)$-augmented tester. We will refer to a non-augmented private tester as a (ξ, α, δ)-private tester. Finally, a (ε, δ)-tester is simply a standard, non-private, non-augmented tester.

We focus on two common variants of the distribution testing problem in which $\mathcal{P}$ contains a single distribution q, and the problem is to distinguish between the cases $p = q$ and $\mathrm{d}_{\mathrm{TV}}(p, q) > \varepsilon$. In the case of *identity testing*, q is a known distribution. In the case of *closeness testing*, q is unknown, and the algorithm receives samples from both p and q.

Differential privacy: For two datasets X and X', we write the Hamming distance of X and X' – the number of samples in which the two datasets differ – as $\mathrm{Ham}(X, X')$. We consider an algorithm to be private if it satisfies the following definition.

Definition 5 (Differential privacy) *Fix parameter $\xi > 0$. Let $X, X' \in \mathcal{X}$ be two datasets with $\mathrm{Ham}(X, X') = 1$; that is, X and X' differ in exactly one sample. A randomized algorithm $\mathcal{A} : \mathcal{X} \to \{\mathsf{accept}, \mathsf{reject}, \perp\}$ is ξ-differentially private if for all such X and X' and any $O \subseteq \{\mathsf{accept}, \mathsf{reject}, \perp\}$, $\mathbf{Pr}[\mathcal{A}(X) \in O] \leq e^{\xi} \cdot \mathbf{Pr}[\mathcal{A}(X') \in O]$.*

Our algorithms use the Laplace mechanism (Dwork and Roth, 2014), which works as follows. Let $\mathcal{X}$ be the set of all datasets and $f \colon \mathcal{X} \to \mathbb{R}$ be a real-valued function. Define the sensitivity of f, denoted $\Delta(f)$, as $\Delta(f) \coloneqq \max_{X, X' \in \mathcal{X} : \mathrm{Ham}(X, X') = 1} |f(X) - f(X')|$. The quantity $f(X) + \mathbf{Lap}(\Delta(f)/\xi)$ is then ξ-differentially private.

In this paper, we focus on pure differential privacy in the central setting, and implicitly focus on the "high-privacy regime" (range $\xi \in (0, 1]$). We leave other variants (e.g., local DP or approximate DP) as open problems.

Organization of the paper: A detailed yet succinct description of the upper bound for private augmented closeness testing is presented in Section 3 of the main paper. Due to space constraints, some proofs have been omitted here; a complete version is available in Appendix C. The lower bound for private augmented closeness testing is given in Appendix D. Additionally, the upper bound for identity testing (and, equivalently, uniformity testing) is presented in Appendix A, while the lower bound for uniformity testing (and hence identity testing) appears in Appendix B.

3. Upper bound for Private Augmented Closeness Testing

We establish the algorithmic part of Theorem 2, restated below:

Theorem 6 *Let $\varepsilon, \alpha \in (0, 1]$ and $\xi > 0$. Let p and q both be unknown distributions over $[n]$. Then Algorithm 1 is a ξ-DP augmented $(\varepsilon, \alpha, 0.32)$-closeness tester which takes s samples each from p and q, where*

$$s = \tilde{O}\left(\frac{n^{2/3}\alpha^{1/3}}{\varepsilon^{4/3}} + \frac{\sqrt{n}}{\varepsilon^2} + \frac{1}{\varepsilon\xi} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} + \frac{n^{1/3}}{\varepsilon^{4/3}\xi^{2/3}}\right) \tag{1}$$

Algorithm 1: PRIVATE-AUGMENTED-CLOSENESS-TESTER

Input: $p, q, \hat{p}, \alpha, n, \delta = 0.32$

Output: accept, reject, or $\perp$

if $\varepsilon = o(n^{-1/4})$ *or* $\varepsilon^2\xi = o(n^{-1})$ **then**

  Run the tester of Lemma 29 and return its answer

$\hat{k}_p, \hat{k}_q \leftarrow \mathbf{Poi}(k)$, $\hat{\ell} \leftarrow \mathbf{Poi}(\ell)$ where k, ℓ are determined as described in the proof of Theorem 6

if $\hat{k}_p + \hat{k}_q > 200k$ *or* $\hat{\ell} > 100\ell$ **then**

  return reject

$p', q' \leftarrow$ Flatten p and q according to Eq. (24)

$p'', q'' \leftarrow$ Flatten p' and q' according to Eq. (25) using $\hat{k}_p$ samples from p and $\hat{k}_q$ from q

$\overline{L} \leftarrow$ compute the statistic in Eq. (38) using $\hat{\ell}$ samples

$\tilde{L} \leftarrow \overline{L} + \mathbf{Lap}\left(\frac{\Delta(\overline{L})}{\xi}\right)$

if $\tilde{L} > 30 \cdot \left(\frac{2\alpha}{k} + \frac{4}{n}\right)$ **then**

  return $\perp$

Run the tester of Lemma 25 and return its answer

Roadmap. We describe our flattening procedure in Section 3.2, drawing on the flattening procedures of Aliakbarpour et al. (2024) and Aliakbarpour et al. (2019) which we review in Appendix C.1; and analyze our algorithm in Section 3.3. Finally, in Section 3.4, we show how to privately estimate the ℓ_2 norm to verify our flattening procedure. Due to space constraints, all proofs and details are deferred to Appendix C.

3.1. Overview of Flattening

Flattening is motivated by a result of Chan et al. (2014), in which the authors give a closeness tester for distributions p and q whose sample complexity is $O(bn/\varepsilon^2)$, where $b = \min(\|p\|_2, \|q\|_2)$. The flattening technique described by Diakonikolas and Kane (2016) works as follows: draw a multiset F of $\mathbf{Poi}(k)$ samples from p. Let k_i denote the number of elements in F equal to i. The new distribution p' will be supported on domain $\{(i, j) : i \in [n], j \in [k_i]\}$. To draw a sample from p', draw i from p and choose j uniformly at random from the set $[k_i]$. This effectively divides the probability mass of each element i evenly into k_i buckets. Diakonikolas and Kane (2016) show that the expected ℓ_2 norm of p' is low: $\mathbf{E}[\|p'\|_2^2] \leq \frac{1}{k}$.

In the rest of this section, we describe two variations on the original technique which will be useful in constructing our augmented private tester.

Private Flattening: The private tester of Aliakbarpour et al. (2019) attempts to derandomize the original flattening of Diakonikolas and Kane (2016). When a flattening-based tester divides its samples into flattening and test sets, it introduces randomness based on the permutation π used to partition the samples. This results in high sensitivity of the test statistic Z (we treat the statistic itself as a black box, analyzing its sensitivity using results from Aliakbarpour et al. (2019)). The private tester reduces the sensitivity by taking the expectation of Z over every permutation π of the data and the randomness r of the flattening procedure:

$$\overline{Z} := \mathbf{E}_{\pi,r}[Z]$$

The authors show that $\overline{Z}$ has low sensitivity and preserves the correctness of any flattening procedure which satisfies certain properties. In Appendix C.1, we give a formal definition of these properties

and establish that the flattening of Diakonikolas (2016) satisfies these properties, allowing it to be made private.

Augmented Flattening: In the augmented setting of Aliakbarpour et al. (2024), the estimate $\hat{p}$ guides the choice of buckets for each element i. Let $\nu \in (0, 1]$ be a flattening parameter. Rather than dividing elements into $k_i + 1$ buckets, the flattener creates $b_i = \left\lfloor \frac{\hat{p}_i}{\nu} \right\rfloor + k_i + 1$ buckets in flattened distributions p' and q'. If $\hat{p}$ is α-close to p, the expected ℓ_2^2 norm after flattening is shown to be $\frac{2\alpha}{k} + 4 \cdot \nu$. The tester then draws $O(\sqrt{n})$ samples each from p' and q' to estimate their ℓ_2^2 norm. If at least one distribution has been sufficiently flattened, the algorithm proceeds to the test phase. Otherwise, it returns $\perp$.

3.2. Our Two-Step Flattening

To simplify the analysis of our augmented private closeness tester, we split the flattening into two steps. The first step flattens p and q into distributions p' and q' using only on the prediction $\hat{p}$. The second step draws samples from p' and q' and uses these samples to flatten them into a third set of distributions p'' and q''. This two-step flattening makes it much easier to guarantee that the procedure is private. Since the first step draws no samples, it can be performed without any modifications for privacy. Since the second step does not use $\hat{p}$, it is precisely the same procedure as the flattening described in Lemma 24, and can be made private by the same mechanism. We show that if $\mathrm{d}_{\mathrm{TV}}(\hat{p}, p) \leq \alpha$, then $\|p''\|_2^2$ is small in expectation, allowing us to test efficiently with fewer flattening samples than the optimal non-augmented tester.

Flattening procedure: Suppose we are given parameter k, which determines the expected number of flattening samples. The two-step flattening works as follows.

1. From the original distributions p and q, estimate $\hat{p}$, and suggested accuracy α, construct p' and q' by flattening each element $i \in [n]$ into b_i buckets, where

$$b_i = \lceil n\hat{p}_i \rceil + 1 \tag{2}$$

   (This is analogous to setting $\nu = 1/n$ in Equation 22). Since p, $\hat{p}$, and α are not private, this can be done as a preprocessing step before running the private portion of the algorithm. After this step, p' is defined by having, for all $i \in [n]$ and $j \in [b_i]$, $p'_{i,j} = p_i/b_i$. Sampling from p' can be done as follows: sample $i \sim p$ and $j \sim \mathrm{Unif}([b_i])$, and return (i, j). Sampling from q' is equivalent (note that both p and q use the same bucketing, so p' and q' have the same domain).

2. Sample $\hat{k}_p, \hat{k}_q \sim \mathbf{Poi}(k)$ independently. Draw a multiset $F^{(p)}$ of $\hat{k}_p$ samples from p' and a multiset $F^{(q)}$ of $\hat{k}_q$ samples from q'. Let $F = F^{(p)} \cup F^{(p)}$, and $k_{i,j}$ be the total number of instances of element (i, j) in F. Then divide each element of the domain of p', q' into $b_{i,j} = k_{i,j} + 1$ buckets.

   The domain of p'' and q'' is $\Omega = \{(i, j, m) : i \in [n], j \in [b_i], m \in [b_{i,j}]\}$, and

$$p''_{i,j,m} = \frac{p_{i,j}}{b_{i,j}}, \qquad q''_{i,j,m} = \frac{q_{i,j}}{b_{i,j}} \tag{3}$$

   for all $(i, j, m) \in \Omega$. To sample from p'', we sample $(i, j) \sim p'$ and $m \sim \mathrm{Unif}([b_{i,j}])$, then return (i, j, m).

Effect of flattening: We now show that the domain size after flattening remains $\Theta(n+k)$ with high probability. After the second flattening step, the number of elements is $n'' = |\Omega|$ (itself a random variable), where $n'' = \sum_{i=1}^{n}\sum_{j=1}^{b_i} b_{i,j} = \sum_{i=1}^{n}\sum_{j=1}^{b_i}(k_{i,j}+1) \leq \hat{k}_p + \hat{k}_q + 3n$. Note that $\hat{k}_p + \hat{k}_q \sim \mathbf{Poi}(2k)$, so $\hat{k}_p + \hat{k}_q \leq 200k$ with probability at least 0.99 by Markov's inequality. Therefore, $n'' = O(n+k)$ with probability at least 0.99. Next, we bound the expected ℓ_2^2 norm of the flattened distribution, assuming the suggested accuracy level α is valid:

Lemma 7 (see Lemma 28) *Let p and $\hat{p}$ be distributions over $[n]$ with $\mathrm{d_{TV}}(p,\hat{p}) \leq \alpha$. Then the two-step flattening produces a distribution p'' with $\mathbf{E}_F\left[\|p''\|_2^2\right] \leq \frac{2\alpha}{k} + \frac{4}{n}$.*

3.3. The Algorithm

At a high level, our algorithm works in three steps. The first step flattens p and q into distributions p'' and q'' as described above. The second step tests the ℓ_2^2 norm of p''. If the norm is not within a constant multiplicative factor of the bound guaranteed by Lemma 28, then we have $\mathrm{d_{TV}}(\hat{p},p) > \alpha$ with high probability, and the tester returns $\perp$. If the norm *is* close to the bound, our algorithm can perform an efficient closeness test. The final step runs the private closeness tester of Aliakbarpour et al. (2019) and returns its result. We give pseudocode for our algorithm in Algorithm 1.

By our lower bound (Theorem 38), there are some regimes of parameters in which an augmented tester cannot use asymptotically fewer samples than a non-augmented private tester. In these cases, we simply invoke the optimal (private) non-augmented tester of Zhang (2021), whose sample complexity is given by Lemma 29. In all other regimes, we show that our tester is optimal up to $\log n$ factors. We are now ready to prove our upper bound.

Proof [Proof Outline of Theorem 6] For the analysis, we will condition on two assumptions:

1. Poissonization did not cost too many samples: $\hat{k} = \hat{k}_p + \hat{k}_q < 200k$, $\hat{\ell} < 100\ell$, and $\hat{s} < 100s$.
2. Ratios are roughly what they should be: for all $i \in [n]$, $\frac{\ell_i}{k_i+1} < 12\log\left(\frac{n}{0.05}\right)\cdot\frac{\ell}{k}$ This holds with probability at least 0.95 by Lemma 37. If this assumption does not hold in our original dataset, we enforce it via the differentially private mapping of Lemma 35.

Privacy: If $\varepsilon = o(n^{-1/4})$ or $\varepsilon^2\xi = o(n^{-1})$, we run the tester of Lemma 29, which is private by the same result. If $\varepsilon = \Omega(n^{-1/4})$ and $\varepsilon^2\xi = \Omega(n^{-1})$, the algorithm's outputs depend only on the statistic $\tilde{L}$, and the choice of $\hat{k}_p$, $\hat{k}_q$ and $\hat{\ell}$. The latter two do not depend on the sampled data. If assumption (2) is satisfied, then $\tilde{L}$ is ξ-DP by Lemma 12. Finally, the second flattening step is a equivalent to the non-augmented flattening procedure described by Aliakbarpour et al. (2019). Therefore, if the algorithm reaches Line 12, its output is private by Lemma 25.

Correctness: If $\varepsilon = o(n^{-1/4})$ or $\varepsilon^2\xi = o(n^{-1})$, the algorithm returns a valid answer with probability at least 0.95 by Lemma 29. Thus, we can hereafter assume that $\varepsilon = \Omega(n^{-1/4})$ and $\varepsilon^2\xi = \Omega(n^{-1})$.

- First, we show that if $\mathrm{d_{TV}}(p,\hat{p}) \leq \alpha$, the algorithm does not output $\perp$ with high probability. The only case in which the algorithm returns $\perp$ is when $\tilde{L} > 30(2\alpha/k + 4/n)$. By Lemma 12, the estimate $\tilde{L}$ is within a constant factor of $\|p''\|_2^2$ with probability at least 0.94: $\frac{\|p''\|_2^2}{2} \leq \tilde{L} \leq 3\frac{\|p''\|_2^2}{2}$ Combining Lemma 28 with Markov's inequality, with high probability $\|p''\|_2^2 \leq 20\left(\frac{2\alpha}{k}+\frac{4}{n}\right)$. Therefore, by union bound, we have that $\tilde{L}$ is not more than 20 times the expected value of $\|p''\|_2^2$ with probability at least 0.91, and the probability of returning $\perp$ in this case is at most 0.09.

- Next, we show that the algorithm does not output reject with high probability when $p = q$. Conditioned on assumption (1), the algorithm only rejects if the tester of Lemma 25 rejects. This occurs with probability at most 0.25 when $p = q$.

- Finally, we show that the algorithm does not return accept when $\mathrm{d}_{\mathrm{TV}}(p, q) > \varepsilon$. In this case, the algorithm only accepts if the tester of Lemma 25 accepts, so with probability at most 0.25.

Therefore, the algorithm outputs a valid answer with probability 0.75 conditioned on the assumptions. Taking a union bound, the algorithm outputs a valid answer with probability 0.68 overall.

Sample Complexity: By assumption (1), we draw $O(k)$ flattening samples, $O(\ell)$ estimation samples, and $O(s)$ test samples from both p and q. Let $n'' = O(n + k)$ be the domain size of p''. Then the parameters must satisfy the following constraints: (1) $k \cdot \min\left(\frac{k}{\log^2 n}, \frac{\ell}{\log n}\right) = \Omega\left(\frac{n''}{\xi}\right)$ and (2) $\ell \geq \Theta(\sqrt{n''})$ for ℓ_2 norm estimation, and (3) $s \geq \Theta(n'' \cdot \sqrt{\mathbf{E}_F\left[\|p''\|_2^2\right]}/\varepsilon^2 + \sqrt{n''\Delta(\overline{Z})}/(\varepsilon\sqrt{\xi}))$ for testing. By Lemma 25, the sensitivity of the test statistic is $\Delta(\overline{Z}) = \Theta\left(\frac{s+k}{k}\right)$, while by Lemma 28 we have $\mathbf{E}_F[\|p''\|_2^2] \leq 2\alpha/k + 4/n$. A careful case analysis and choice of k, s then yields the claimed sample complexity. ■

Remark 8 *Algorithm 1 can be implemented in polynomial time in the number of samples taken. The test statistic $\tilde{L}$ can be computed efficiently by counting the number of instances of each element in the flattening and test sets, then computing the closed form expression in Eq. (38). We emphasize that this expectation is only over the random choice of buckets, not over permutations of samples as in Aliakbarpour et al. (2019).*

To complete the analysis, note that both non-augmented private testers used in our algorithm (see Lemmas 25 and 29) run in polynomial time. Note that the tester of Lemma 25 computes the expected value of an expression over every permutation of samples; this expression appears to take exponential time to compute. However, as described by Aliakbarpour et al. (2019), the computation can be performed efficiently. For each domain element, (more precisely, for each element that appears at least once in the sample set), the expression consists of three terms: a, b, and c. Here, a and b are the numbers of test samples drawn from p and q, respectively, and c is the number of samples in the flattening set. By iterating over all possible values a, b, and c for each element (each has at most $O(k + s + 1)$ possible values, the total number of samples), we can compute in polynomial time the probability of observing any given combination under random permutations.

3.4. Private Testing of ℓ_2 Norm

To determine whether the two-step flattening worked, we must (privately) test the ℓ_2 norm of the flattened distribution p''. The standard ℓ_2 tester (Goldreich and Ron, 2011; Aliakbarpour et al., 2024) computes the number of collisions in its sample set, and has the following guarantee:

Lemma 9 (Goldreich and Ron (2011); Aliakbarpour et al. (2024)) *Let $\delta \in (0, 1)$ and p be a distribution over $[n]$. Let E be a multiset of $\ell = O\left(\sqrt{n}\log(1/\delta)\right)$ samples from p. For each $i \in [n]$, define ℓ_i as the number of instances of element i in E. Let $L(E) = \binom{\ell}{2}^{-1} \sum_{i=1}^{n} \binom{\ell_i}{2}$. Then with probability at least $1 - \delta$, $\frac{\|p\|_2^2}{2} \leq L(E) \leq \frac{3\|p\|_2^2}{2}$.*

The statistic above has high sensitivity; intuitively, if a large number of samples fall in the same bin (i, j, m), changing a single instance of element i could affect a large number of collisions. To avoid this, we will derandomize this tester by taking

$$\overline{L}(E) \coloneqq \mathbf{E}_r[L \mid E] \tag{4}$$

where r is the string of random bits used to choose the bucket when generating a sample from p'' given a sample from p', and E is a multiset (the *estimation set*) of $\mathbf{Poi}(\ell)$ flattening samples from p'. Each of these samples is a pair (i, j). Rather than sample the third coordinate to generate a sample from p'', we will consider the expected number of collisions over all such samplings, conditioned on E. In the result below, we give a closed-form expression for $\overline{L}$.

Lemma 10 *Let $k_{i,j}$ and $\ell_{i,j}$ be the number of instances of element (i, j) in the flattening and estimation sets F and E respectively. For each $i \in [n]$, let b_i be the number of buckets created for element i by the first flattening step. Then*

$$\overline{L} = \frac{1}{\binom{\ell}{2}} \sum_{i=1}^{n} \mathbf{E}_r \left[\sum_{j=1}^{b_i} \sum_{k=1}^{b_{i,j}} \binom{\ell_{i,j,k}}{2} \;\middle|\; k_i, \ell_i \right] = \frac{1}{\binom{\ell}{2}} \sum_{i=1}^{n} \sum_{j=1}^{b_i} \frac{\binom{\ell_{i,j}}{2}}{k_{i,j}+1} \tag{5}$$

In general, the sensitivity of $\overline{L}$ might be quite high. In the worst case, we might draw no instances of element (i, j) in F but $\Theta(\ell)$ instances in E, giving a sensitivity of $\Theta(1)$. However, the sensitivity is considerably lower whenever the flattening and estimation sets are similar (i.e., no element has a much higher frequency in E than in F). The following lemma formalizes this intuition.

Lemma 11 *Suppose that there exists $A \geq 0$ such that for all $i \in [n]$, $\frac{\ell_{i,j}}{k_{i,j}+1} \leq A \cdot \frac{\ell}{k}$. Then the sensitivity of $\overline{L}$ is bounded by $\Delta(\overline{L}) \leq O(\frac{A^2}{k^2} + \frac{A}{k\ell})$.*

Assuming the sensitivity obeys this bound (in Appendix C.5, we show how to transform any dataset into one that satisfies this property for $A = \Theta(\log n)$), the last remaining task is to show that we can efficiently estimate $\|p''\|_2^2$ to within a constant multiplicative factor:

Lemma 12 *Let E be a set of $\mathbf{Poi}(\ell)$ estimation samples from p, and $\overline{L}$ be the derandomized ℓ_2^2 norm estimator defined above. Let $A = 12\log(n/0.05)$. For a given value of $\xi > 0$, define*

$$\tilde{L} = \overline{L} + \mathbf{Lap}\big(\Delta(\overline{L})/\xi\big)\,.$$

Suppose $\hat{k} = \hat{k}_p + \hat{k}_q \leq 100k$, that Eq. (43) holds, and that $k \cdot \min(k/A^2, \ell/A) \geq C_1 \cdot \frac{k+n}{\xi}, \ell \geq C_2\sqrt{k+n}$ for sufficiently large constants C_1, C_2. Then $\mathbf{Pr}\Big[|\tilde{L} - \|p''\|_2^2| > \|p''\|_2^2/2\Big] \leq 0.06$.

Putting all these elements together as outlined establishes Theorem 6.

Acknowledgments

CC is supported by an ARC DECRA (DE230101329). RR is supported by the NSF TRIPODS program (award DMS-2022448) and CCF-2310818. This work was initiated while MA was serving as a research fellow and CC and RR were visiting the Simons Institute for the Theory of Computing as part of the Sublinear Algorithms program.

References

Anders Aamand, Justin Y. Chen, Huy Lê Nguyen, Sandeep Silwal, and Ali Vakilian. Improved frequency estimation algorithms with and without predictions. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/2e49934cac6cb8604b0c67cfa0828718-Abstract-Conference.html.

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*, pages 6879–6891, 2018.

Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *NIPS*, pages 3591–3599, 2015.

Maryam Aliakbarpour, Ilias Diakonikolas, and Ronitt Rubinfeld. Differentially private identity and equivalence testing of discrete distributions. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, pages 169–178, 2018.

Maryam Aliakbarpour, Ilias Diakonikolas, Daniel Kane, and Ronitt Rubinfeld. Private testing of distributions via sample permutations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 10877–10888, 2019. URL http://papers.nips.cc/paper/9270-private-testing-of-distributions-via-sample-permutations.

Maryam Aliakbarpour, Piotr Indyk, Ronitt Rubinfeld, and Sandeep Silwal. Optimal algorithms for augmented testing of discrete distributions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Noga Alon, Raef Bassily, and Shay Moran. Limits of private learning with access to public data, 2019. URL https://arxiv.org/abs/1910.11519.

Amazon Web Services. Aws clean rooms announces differential privacy capability for secure data collaboration, April 2024. URL https://aws.amazon.com/about-aws/whats-new/2024/04/aws-clean-rooms-differential-privacy-generally-available/. Accessed: 2024-11-13.

Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *41st Annual Symposium on Foundations of Computer Science, FOCS 2000, 12-14 November 2000, Redondo Beach, California, USA*, pages 259–269, 2000. doi: 10.1109/SFCS.2000.892113. URL https://doi.org/10.1109/SFCS.2000.892113.

Shai Ben-David, Alex Bie, Clément L. Canonne, Gautam Kamath, and Vikrant Singhal. Private distribution learning with public data: The view from sample compression. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors,

Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1687466683649e8bdcdec0e3f5c8de64-Abstract-Conference.html.

Arnab Bhattacharyya, Davin Choo, Philips George John, and Themis Gouleakis. Learning multivariate gaussians with imperfect advice. *CoRR*, abs/2411.12700, 2024.

Alex Bie, Gautam Kamath, and Vikrant Singhal. Private estimation with public data. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/765ec49952dd0140ac754d6d3f9bc899-Abstract-Conference.html.

Adam Block, Mark Bun, Rathin Desai, Abhishek Shetty, and Steven Wu. Oracle-efficient differentially private learning with public data, 2024. URL https://arxiv.org/abs/2402.09483.

Zhiqi Bu, Xinwei Zhang, Mingyi Hong, Sheng Zha, and George Karypis. Pre-training differentially private models with limited public data. *CoRR*, abs/2402.18752, 2024. doi: 10.48550/ARXIV.2402.18752. URL https://doi.org/10.48550/arXiv.2402.18752.

B. Cai, C. Daskalakis, and G. Kamath. Priv'it: Private and sample efficient identity testing. In *International Conference on Machine Learning, ICML*, pages 635–644, 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.

Clément L. Canonne and Yucheng Sun. Optimal closeness testing of discrete distributions made (complex) simple. *CoRR*, abs/2204.12640, 2022.

Siu-on Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014.

Deniz Daum, Richard Osuala, Anneliese Riess, Georgios Kaissis, Julia A. Schnabel, and Maxime Di Folco. On differentially private 3d medical image synthesis with controllable latent diffusion models. In Anirban Mukhopadhyay, Ilkay Öksüz, Sandy Engelhardt, Dorit Mehrof, and Yixuan Yuan, editors, *Deep Generative Models - 4th MICCAI Workshop, DGM4MICCAI 2024, Held in Conjunction with MICCAI 2024, Marrakesh, Morocco, October 10, 2024, Proceedings*, volume 15224 of *Lecture Notes in Computer Science*, pages 139–149. Springer, 2024. doi: 10.1007/978-3-031-72744-3_14. URL https://doi.org/10.1007/978-3-031-72744-3_14.

Ilias Diakonikolas. Learning structured distributions. In *CRC Handbook of Big Data*, pages 267–283. 2016.

Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 685–694, 2016. doi: 10.1109/FOCS.2016.78. URL https://doi.org/10.1109/FOCS.2016.78.

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, ICALP 2018, July 9-13, 2018, Prague, Czech Republic*, pages 41:1–41:14, 2018. doi: 10.4230/LIPIcs.ICALP.2018.41. URL https://doi.org/10.4230/LIPIcs.ICALP.2018.41.

Ilias Diakonikolas, Themis Gouleakis, Daniel M. Kane, John Peebles, and Eric Price. Optimal testing of discrete distributions with high probability. In *STOC*, pages 542–555. ACM, 2021.

Differential Privacy Team, Apple. Learning with privacy at scale. https://machinelearning.apple.com/research/learning-with-privacy-at-scale, 2017. Accessed: 2024-05-22.

Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30, 2017.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL http://dx.doi.org/10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 1054–1067, 2014.

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

Oded Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. In *Computational Complexity and Property Testing - On the Interplay Between Randomness and Computation*, pages 152–172. Springer, 2020. doi: 10.1007/978-3-030-43662-9_10. URL https://doi.org/10.1007/978-3-030-43662-9_10.

Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Electron. Colloquium Comput. Complex.*, TR00-020, 2000. URL https://eccc.weizmann.ac.il/eccc-reports/2000/TR00-020/index.html.

Oded Goldreich and Dana Ron. *On testing expansion in bounded-degree graphs*, pages 68–75. Springer-Verlag, Berlin, Heidelberg, 2011. ISBN 9783642226694.

Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM (JACM)*, 45(4):653–750, 1998.

Chen-Yu Hsu, Piotr Indyk, Dina Katabi, and Ali Vakilian. Learning-based frequency estimation algorithms. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=r1lohoCqY7.

Piotr Indyk, Ali Vakilian, and Yang Yuan. Learning-based low-rank approximations. In *Advances in Neural Information Processing Systems*, pages 7400–7410, 2019.

Tanqiu Jiang, Yi Li, Honghao Lin, Yisong Ruan, and David P. Woodruff. Learning-augmented data stream algorithms. In *International Conference on Learning Representations*, 2020.

Gavin Kerrigan, Dylan Slack, and Jens Tuyls. Differentially private language models benefit from public pre-training. *CoRR*, abs/2009.05886, 2020. URL https://arxiv.org/abs/2009.05886.

Mikhail Khodak, Kareem Amin, Travis Dick, and Sergei Vassilvitskii. Learning-augmented private algorithms for multiple quantile release. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners, 2022. URL https://arxiv.org/abs/2110.05679.

Andrew Lowy, Zeman Li, Tianjian Huang, and Meisam Razaviyayn. Optimal differentially private model training with public data. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=NFEJQn7vX0.

Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inf. Theory*, 54(10):4750–4755, 2008. doi: 10.1109/TIT.2008.928987. URL https://doi.org/10.1109/TIT.2008.928987.

Ronitt Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

Arjun G. Thakurta, Adam H. Vyrros, Umesh S. Vaishampayan, Gaurav Kapoor, Julien Freudiger, Vimal R. Sridhar, and David Davidson. Learning new words, 2017.

Salil Vadhan. The complexity of differential privacy. *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich*, pages 347–450, 2017.

Paul Valiant. Testing symmetric properties of distributions. *SIAM J. Comput.*, 40(6):1927–1968, 2011. doi: 10.1137/080734066. URL https://doi.org/10.1137/080734066.

Huanyu Zhang. Statistical inference in the differential privacy model. *CoRR*, abs/2108.05000, 2021.

Appendix A. Upper bound for Private Augmented Identity Testing

Our upper bound follows directly from the upper bound of Aliakbarpour et al. (2024) for non-private augmented identity testing. The tester aims to detect the case where p is far from q using the Scheffé set S of $\hat{p}$ and q, where $S \coloneqq \{i \in [n] : \hat{p}_i < q_i\}$. The difference in the probability of S under $\hat{p}$ and q is precisely the total variation distance:

$$\eta \coloneqq \mathrm{d}_{\mathrm{TV}}(q, \hat{p}) = |q(S) - \hat{p}(S)|$$

The algorithm draws samples from p and computes the test statistic σ, the fraction of samples which fall in S. If σ is sufficiently far from $q(S)$, the tester rejects; otherwise, it returns $\perp$. The full pseudocode of our private tester is given in Algorithm 2.

The sample complexity of the non-private augmented tester depends on η and α:

Lemma 13 (Aliakbarpour et al. (2024)) *Let $\varepsilon \in (0, 1], \alpha \in [0, 1)$. Suppose we are given known distributions q and $\hat{p}$, and unknown distribution p, all over $[n]$. Let $\eta \coloneqq \mathrm{d}_{\mathrm{TV}}(q, \hat{p})$ and $\delta_0 \coloneqq 1/10$. There is an $(\varepsilon, \alpha, \delta_0)$-augmented identity testing algorithm which takes s samples from p, where*

$$s = \begin{cases} \Theta\left(\frac{\sqrt{n}}{\varepsilon}\right) & \text{if } \eta \leq \alpha \\ \Theta\left(\min\left(\frac{1}{(\eta-\alpha)^2}, \frac{\sqrt{n}}{\varepsilon^2}\right)\right) & \text{if } \eta > \alpha \end{cases}$$

This result shows an improvement in sample complexity if $\hat{p}$ is a sufficiently good estimate of p, and does no worse than a standard identity tester if not. To privatize the tester, we simply add Laplace noise to the test statistic σ.

Lemma 14 *Fix privacy parameter $\xi > 0$. Let $\hat{p}$ and q be known distributions, and p be an unknown distribution, over $[n]$. Let $S \coloneqq \{i \in [n] : \hat{p}_i < q_i\}$. For a set $\{x_1, ..., x_s\}$ of s i.i.d. samples from p, let*

$$\sigma \coloneqq \frac{1}{s} \sum_{j=1}^{s} \mathbb{1}\left[x_j \in S\right]$$

Then the statistic $\hat{\sigma} = \sigma + \mathbf{Lap}(1/(s\xi))$ is ξ-differentially private.

Proof The sensitivity of σ is

$$\Delta(\sigma) = \max_{X, X' : |X - X'| < 1} |\sigma(X) - \sigma(X')| = \frac{1}{s}$$

Therefore, adding $\mathbf{Lap}(1/(s\xi))$ noise suffices to make the statistic private by the Laplace mechanism. ■

In the case where augmentation does not decrease the sample complexity, we must run a non-augmented private identity tester. The following lemma gives the sample complexity of private identity testing.

Algorithm 2: PRIVATE-AUGMENTED-IDENTITY-TESTER

Input: $p, q, \hat{p}, n, \varepsilon, \alpha, \xi, \delta = 0.1$

Output: accept, reject, or $\perp$

$\eta \leftarrow \mathrm{d}_{\mathrm{TV}}(\hat{p}, q)$

if $\eta \leq \alpha$ *or* $\frac{\sqrt{n}}{\varepsilon^2} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} + \frac{n^{1/3}}{\varepsilon^{4/3}\xi^{2/3}} + \frac{1}{\varepsilon\xi} \leq \frac{1}{(\eta-\alpha)^2} + \frac{1}{(\eta-\alpha)\xi}$ **then**

  Run the tester of Lemma 15 and output its answer

$S \leftarrow \{i \in [n] : \hat{p}_i < q_i\}$

Draw $s = \Theta\left(\frac{1}{(\eta-\alpha)^2} + \frac{1}{(\eta-\alpha)\xi}\right)$ samples $x_1, ...x_s$ from p

$\sigma \leftarrow \frac{1}{s}\sum_{j=1}^{s} \mathbb{1}[x_j \in S]$

$\hat{\sigma} \leftarrow \sigma + \mathbf{Lap}\left(\frac{1}{s\xi}\right)$

if $|\hat{\sigma} - q(S)| > \frac{\eta-\alpha}{4}$ **then**

  return reject

else

  return $\perp$

Lemma 15 (Acharya et al. (2018)) *Fix parameters $\varepsilon \in (0, 1]$ and $\xi > 0$. Let q be a known distribution and p be an unknown distribution over $[n]$. There is a $(\xi, \varepsilon, \delta = 0.1)$-private identity tester which takes s samples from p, where*

$$s = \Theta\left(\frac{\sqrt{n}}{\varepsilon^2} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} + \frac{n^{1/3}}{\varepsilon^{4/3}\xi^{2/3}} + \frac{1}{\varepsilon\xi}\right) \tag{6}$$

Additionally, any $(\xi, \varepsilon, \delta = 0.1)$-private identity tester requires $\Omega(s)$ samples.

Combining the results above, we have the following upper bound.

Theorem 16 *Fix parameters $\varepsilon \in (0, 1]$, $\alpha \in [0, 1)$, $\xi > 0$, and $\delta_0 := 1/10$. Let $\hat{p}$ and q be known distributions, and p be an unknown distribution, over $[n]$. Then Algorithm 2 is a $(\xi, \varepsilon, \alpha, \delta_0)$-private augmented identity tester which takes s samples from p, where*

$$s = \begin{cases} \Theta\left(\frac{\sqrt{n}}{\varepsilon^2} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} + \frac{n^{1/3}}{\varepsilon^{4/3}\xi^{2/3}} + \frac{1}{\varepsilon\xi}\right), & \eta \leq \alpha \\ \Theta\left(\min\left(\frac{1}{(\eta-\alpha)^2} + \frac{1}{(\eta-\alpha)\xi}, \frac{\sqrt{n}}{\varepsilon^2} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} + \frac{n^{1/3}}{\varepsilon^{4/3}\xi^{2/3}} + \frac{1}{\varepsilon\xi}\right)\right), & \eta > \alpha \end{cases} \tag{7}$$

Moreover, the algorithm can be implemented in polynomial time in s.

Proof Note that the choice to use one algorithm or the other only depends on the (public) parameters, not on the samples: therefore, the decision of which algorithm to use does not compromise privacy.

If $\eta \leq \alpha$ or $\sqrt{n}/\varepsilon^2 + \sqrt{n}/(\varepsilon\sqrt{\xi}) + n^{1/3}/(\varepsilon^{4/3}\xi^{2/3}) + 1/(\varepsilon\xi) \leq 1/(\eta-\alpha)^2 + 1/(\xi(\eta-\alpha))$, we run the non-augmented private identity tester, which is ξ-differentially private and returns the correct answer with probability at least 0.9. The sample complexity in this case is given by Lemma 15.

Next, we consider the case where $\eta > \alpha$ and $1/(\eta-\alpha)^2 + 1/(\xi(\eta-\alpha)) < \sqrt{n}/\varepsilon^2 + \sqrt{n}/(\varepsilon\sqrt{\xi}) + n^{1/3}/(\varepsilon^{4/3}\xi^{2/3}) + 1/(\varepsilon\xi)$. We will separately prove privacy, correctness, and the sample complexity.

Privacy: In the case where $\eta \leq \alpha$ or the non-augmented tester has lower sample complexity than the augmented version, we run the non-augmented private identity tester, which is ξ-differentially private by Lemma 15. If we run the augmented tester, the algorithm is ξ-differentially private by Lemma 14.

Correctness: Note that the augmented tester can never return accept. We will only consider the other two possibilities. First, we show that with high probability, the algorithm does not return reject when $p = q$. If $s = \Omega(1/(\eta - \alpha)^2)$, then by Hoeffding's inequality,

$$\mathbf{Pr}\left[|\sigma - p(S)| \geq \frac{\eta - \alpha}{8}\right] \leq 2\exp\left(-2\left(\frac{\eta - \alpha}{8}\right)^2 \cdot s\right) \leq 0.05 \tag{8}$$

By the cdf of the Laplace distribution, if $s = \Omega(1/((\eta - \alpha)\xi))$,

$$\mathbf{Pr}\left[|\hat{\sigma} - \sigma| \geq \frac{\eta - \alpha}{8}\right] = \exp\left(\frac{-(\eta - \alpha)\xi}{8\Delta(\sigma)}\right) = \exp\left(\frac{-(\eta - \alpha)\xi \cdot s}{8}\right) \leq 0.05 \tag{9}$$

Combining Eq. (8) and Eq. (9) and taking a union bound, the following holds with probability at least 0.9:

$$|\hat{\sigma} - p(S)| \leq |\hat{\sigma} - \sigma| + |\sigma - p(S)| \leq \frac{\eta - \alpha}{4}$$

If $p = q$, then $p(S) = q(S)$, and the algorithm returns reject with probability at most 0.1.

Finally, we show that if $\mathrm{d_{TV}}(p, \hat{p}) \leq \alpha$, then the algorithm does not return $\perp$ with high probability. In this case, with probability at least 0.9,

$$\begin{aligned}
\alpha &\geq \mathrm{d_{TV}}(p, \hat{p}) \geq |p(S) - \hat{p}(S)| \\
&\geq |p(S) - q(S)| - |q(S) - \hat{p}(S)| \\
&= \mathrm{d_{TV}}(p, q) - |q(S) - \hat{p}(S)| \\
&= \eta - |q(S) - \hat{p}(S)| \\
&\geq \eta - |\hat{\sigma} - p(S)| - |\hat{\sigma} - q(S)| \\
&\geq \eta - \frac{\eta - \alpha}{4} - |\hat{\sigma} - q(S)|
\end{aligned}$$

Therefore,

$$|\hat{\sigma} - q(S)| \geq \frac{3(\eta - \alpha)}{4}$$

and the probability that the algorithm returns $\perp$ is at most 0.1.

Sample Complexity: We have used the fact that $s = \Omega(1/(\eta - \alpha)^2)$ and $s = \Omega(1/((\eta - \alpha)\xi))$. Therefore, it suffices to have

$$s = \Theta\left(\frac{1}{(\eta - \alpha)^2} + \frac{1}{(\eta - \alpha)\xi}\right).$$

We can efficiently compute σ by checking (in constant time) whether $x_j \in S$ for each sample x_j, yielding a tester which takes linear time in the number of samples. ■

Appendix B. Lower bound for Private Uniformity Testing

Our lower bound can be split into two parts based on the regime of parameters. When $\alpha \geq \eta$, we cannot improve on private identity testing, and we obtain our lower bound from known results. When $\alpha < \eta$, augmented private testing may be able to achieve a lower sample complexity. In this case, we prove a lower bound using a version of Le Cam's method.

B.1. Lower Bound When $\alpha \geq \eta$

When the predicted distribution $\hat{p}$ is further from the known distribution q than the desired accuracy (that is, $\eta \leq \alpha$), standard identity testing can be reduced to augmented identity testing. It immediately follows from this that private identity testing can be reduced to private augmented identity testing. Therefore, the lower bound when $\eta \leq \alpha$ is precisely the lower bound for private identity testing. We formalize this argument by combining two known results.

Lemma 17 (Aliakbarpour et al. (2024)) *Fix testing parameters $\varepsilon, \alpha \in [0, 1/2)$ and privacy parameter $\xi > 0$. Let $\hat{p}$ and q be known distributions over $[n]$, and p be an unknown distribution over $[n]$. Let $\eta := \mathrm{d}_{\mathrm{TV}}(q, \hat{p})$. Then if there exists an $(\varepsilon, \alpha, \delta = 1/3)$-augmented identity tester which takes s samples, there exists an $(\varepsilon, \delta = 1/3)$ standard identity tester which takes s samples.*

Theorem 18 *Fix testing parameters $\varepsilon, \alpha \in [0, 1/2)$ and privacy parameter $\xi > 0$. Let $\hat{p}$ and q be known distributions over $[n]$, and p be a known distribution over $[n]$. Let $\eta := \mathrm{d}_{\mathrm{TV}}(q, \hat{p})$. Then if $\eta < \alpha$, any $(\xi, \alpha, \varepsilon, \delta = 0.2)$-private augmented identity tester for testing identity of p and q with prediction $\hat{p}$ requires s samples, where*

$$s \geq \Omega\left(\frac{\sqrt{n}}{\varepsilon^2} + \frac{n^{1/3}}{\varepsilon^{4/3}\xi^{2/3}} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} + \frac{1}{\varepsilon\xi}\right) \tag{10}$$

Proof By Lemma 17, identity testing can be reduced to augmented identity testing when $\eta \leq \alpha$. Since the privacy constraint does not reduce the sample complexity necessary for correctness, private identity testing can be reduced to private augmented identity testing. Therefore, the lower bound for private identity testing given in Lemma 15 is also a lower bound for augmented testing in this case. ■

B.2. Lower Bound When $\alpha < \eta$

When $\alpha < \eta$, it is sometimes possible for the augmented private tester to achieve a lower sample complexity than the non-augmented tester. To account for this possibility, our lower bound uses an extension of Le Cam's method to the augmented setting. At a high level, the idea is to construct three distributions such that no single answer is valid for all three. We show that, with too few samples, (1) the distributions are statistically indistinguishable from the uniform distribution and (2) under differential privacy constraints, no algorithm can answer differently for all three distributions with high probability. We conclude that any algorithm which uses too few samples must have high error probability on at least one distribution. Formally, our result is the following:

Theorem 19 *Fix testing parameters $\varepsilon, \alpha \in [0, 1/2)$ and privacy parameter $\xi > 0$. Let $\hat{p}$ and q be known distributions, and p be an unknown distribution over $[n]$. Let $\eta := \mathrm{d}_{\mathrm{TV}}(q, \hat{p})$. Then if $\eta > \alpha$,*

any ξ-differentially private $(\alpha, \varepsilon, \delta = 0.2)$-augmented testing algorithm for testing identity of p and q with prediction $\hat{p}$ requires s samples, where

$$s \geq \Omega\left(\min\left(\frac{1}{(\eta-\alpha)^2} + \frac{1}{\xi(\eta-\alpha)}, \frac{\sqrt{n}}{\varepsilon^2} + \frac{n^{1/3}}{\varepsilon^{4/3}\xi^{2/3}} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} + \frac{1}{\varepsilon\xi}\right)\right) \tag{11}$$

We defer the proof to the end of this section. We begin by establishing a general tool to combine statistical and privacy constraints. In particular, we show that if two sets of samples look similar with respect to either statistical indistinguishability or Hamming distance, any algorithm must behave similarly on those inputs.

Lemma 20 *Suppose we are given three positive integers n, s_1, and s_2, along with parameters $\xi > 0$ and $\tau \in (0, 1)$. Consider two distributions (or families of distributions) p_1 and p_2 over $[n]$. Let T_1^s and T_2^s indicate two sample set of size s from p_1 and p_2 respectively. Assume we have the following properties:*

- *Suppose we pick $p \in \{p_1, p_2\}$ uniformly at random, and generate a sample set of size s from p. Upon receiving the sample set, no algorithm can distinguish whether the sample set is generated according to p_1 or p_2 with probability more than $(1+\tau)/2$.*
- *For any $s \leq s_2$, there exists a coupling C between T_1^s and T_2^s such that*

$$\mathbf{E}_{(T_1^s, T_2^s)\sim C}[\mathrm{Ham}(T_1^s, T_2^s)] \leq \frac{0.4\,\tau}{\xi}\,.$$

Let $\mathcal{M} : [n]^s \to \mathcal{R}$ be a ξ-private algorithm that receives a sample set of size s from an unknown underlying distribution and produces an outcome in a finite set R. Then, for any such $\mathcal{M}$, $s \leq \max(s_1, s_2)$, and any subset of outcomes $R \subseteq \mathcal{R}$, we have:

$$\mathbf{Pr}_{X\sim p_1^{\otimes s}}[\mathcal{M}(X) \in R] \leq 1.5 \cdot \mathbf{Pr}_{X\sim p_2^{\otimes s}}[\mathcal{M}(X) \in R] + \tau.$$

Proof We consider two cases based on whether $s \leq s_1$. If $s \leq s_1$, we use the statistical indistinguishability of p_1 and p_2. The behavior of $\mathcal{M}$ should be almost identical for sample sets drawn from these two families due to the following standard trick. Consider an algorithm $\mathcal{A}$ that runs as follows: Upon receiving a sample set X, it invokes $\mathcal{M}(X)$. If the outcome of $\mathcal{M}$ is in R it declares p_1; otherwise, it declares p_2. Since this algorithm cannot distinguish p_1 from p_2, we have:

$$\begin{aligned}
\frac{1+\tau}{2} \geq \mathbf{Pr}_{X\sim p^{\otimes s}}[\mathcal{A}(X) = p] &= \mathbf{Pr}_{X\sim p^{\otimes s}}[\mathcal{A}(X) = p_1 | p = p_1] \cdot \mathbf{Pr}[p = p_1] \\
&\quad + \mathbf{Pr}_{X\sim p^{\otimes s}}[\mathcal{A}(X) = p_2 | p = p_2] \cdot \mathbf{Pr}[p = p_2] \\
&= \frac{1}{2}\left(\mathbf{Pr}_{X\sim p_1^{\otimes s}}[\mathcal{M}(X) \in R] + \mathbf{Pr}_{X\sim p_2^{\otimes s}}[\mathcal{M}(X) \notin R]\right) \\
&= \frac{1}{2}\left(\mathbf{Pr}_{X\sim p_1^{\otimes s}}[\mathcal{M}(X) \in R] + 1 - \mathbf{Pr}_{X\sim p_2^{\otimes s}}[\mathcal{M}(X) \in R]\right)
\end{aligned}$$

Thus, we obtain:

$$\mathbf{Pr}_{X\sim p_1^{\otimes s}}[\mathcal{M}(X) \in R] \leq \mathbf{Pr}_{X\sim p_2^{\otimes s}}[\mathcal{M}(X) \in R] + \tau$$

Now, consider the case where $s > s_1$. Since $s \le \max(s_1, s_2)$, we must have $s \le s_2$. Thus, we use the coupling C_2. Let W_2 denote the event that $\mathrm{Ham}(T_1^s, T_2^s) \ge 0.4/\xi$ for (T_1^s, T_2^s) drawn from C_2. By assumption and Markov's inequality, the probability of W_2 is at most τ.

Using the definition of a ξ-private algorithm $\mathcal{M}$, for every pair of datasets T_1^s and T_2^s with $\mathrm{Ham}(T_1^s, T_2^s) \le 0.4/\xi$, we have

$$\begin{aligned}\mathbf{Pr}[\mathcal{M}(T_1^s) \in R] &\le e^{\xi \cdot \mathrm{Ham}\left(T_1^s, T_2^s\right)} \mathbf{Pr}[\mathcal{M}(T_2^s) \in R] \\ &\le e^{0.4} \cdot \mathbf{Pr}[\mathcal{M}(T_2^s) \in R] \le 1.5 \cdot \mathbf{Pr}[\mathcal{M}(T_2^s) \in R].\end{aligned}$$

Now, we obtain

$$\begin{aligned}\mathbf{Pr}_{X \sim p_1^{\otimes s}}[\mathcal{M}(X) \in R] &= \mathbf{Pr}_{(T_1^s, T_2^s) \sim C_1}[\mathcal{M}(T_1^s) \in R] \\ &= \mathbf{Pr}_{(T_1^s, T_2^s) \sim C_1}\left[\mathcal{M}(T_1^s) \in R \mid \overline{W_2}\right] \cdot \mathbf{Pr}_{(T_1^s, T_2^s) \sim C_1}\left[\overline{W_2}\right] \\ &+ \mathbf{Pr}_{(T_1^s, T_2^s) \sim C_1}[\mathcal{M}(T_1^s) \in R \mid W_2] \cdot \mathbf{Pr}_{(T_1^s, T_2^s) \sim C_1}[W_2] \\ &\le 1.5 \cdot \mathbf{Pr}_{(T_1^s, T_2^s) \sim C_1}\left[\mathcal{M}(T_2^s) \in R \mid \overline{W_2}\right] \cdot \mathbf{Pr}_{(T_1^s, T_2^s) \sim C_1}\left[\overline{W_2}\right] \\ &+ \Big(1.5 \cdot \mathbf{Pr}_{(T_1^s, T_2^s) \sim C_1}[\mathcal{M}(T_2^s) \in R \mid T_1^s \ne T_2^s] + 1\Big) \cdot \mathbf{Pr}_{(T_1^s, T_2^s) \sim C_1}[W_2] \\ &\le 1.5 \cdot \mathbf{Pr}_{(T_1^s, T_2^s) \sim C_1}[\mathcal{M}(T_2^s) \in R] + \tau \\ &= 1.5 \cdot \mathbf{Pr}_{X \sim p_2^{\otimes s}}[\mathcal{M}(X) \in R] + \tau\,.\end{aligned}$$

Hence, the proof is complete. ■

B.3. Construction of Hard Distributions

It remains to construct three distributions which satisfy the indistinguishability and closeness assumptions of Lemma 20. We will use a similar construction to Aliakbarpour et al. (2024). Some of the necessary properties are well-known; we will prove the rest.

Choice of distributions: First, we describe the three distributions used in the proof. For the following constructions, we will assume without loss of generality that n is even. (For odd n, we can set the probability of the last element to 0 and choose the remaining elements using the construction for $n-1$.) Let $q = U_n$ be the uniform distribution on $[n]$. We will define the predicted distribution $\hat{p}$ as follows:

$$\hat{p}_i := \begin{cases} \frac{1+2\eta}{n}, & i \text{ is even} \\ \frac{1-2\eta}{n}, & i \text{ is odd} \end{cases} \tag{12}$$

Note that $\mathrm{d_{TV}}(\hat{p}, q) = \eta$.

We will generate two other distributions $p^\bullet$ and $p^\diamond$. Set $\varepsilon' \in (0, 1/2]$ to be $\Theta(\varepsilon)$ and $\alpha' \in (0, \alpha)$ such that $\eta - \alpha' = \Theta(\eta - \alpha)$. Let $\mathbf{Z} \sim \mathrm{Unif}(\{-1, 1\}^{n/2})$. Define $p^\bullet$ and $p^\diamond$ as follows:

$$p_i^\bullet = \begin{cases} \frac{1+2\mathbf{Z}_{i/2} \cdot \varepsilon'}{n}, & i \text{ is even} \\ \frac{1-2\mathbf{Z}_{i/2} \cdot \varepsilon'}{n}, & i \text{ is odd} \end{cases} \tag{13}$$

$$p_i^{\diamond} = \begin{cases} \frac{1+2\cdot(\eta-\alpha')}{n}, & i \text{ is even} \\ \frac{1-2\cdot(\eta-\alpha')}{n}, & i \text{ is odd} \end{cases} \tag{14}$$

Note that $p^{\bullet}$ is ε'-far from U_n, so accept is an incorrect answer given samples from $p^{\bullet}$. Similarly, $p^{\diamond}$ is α-close to $\hat{p}$, so $\perp$ is an invalid answer given samples from $p^{\diamond}$. Finally, reject is an invalid answer given samples from U_n. Therefore, no single answer works for all three distributions U_n, $p^{\bullet}$, and $p^{\diamond}$.

To prove the privacy assumption of Lemma 20, we will need two couplings: a coupling $\mathcal{C}^{\bullet}$ between U_n and $p^{\bullet}$, and a coupling $\mathcal{C}^{\diamond}$ between U_n and $p^{\diamond}$. The existence of $\mathcal{C}^{\bullet}$ follows from a known result:

Lemma 21 (Acharya et al. (2018)) *Let $\varepsilon' \in (0, 1/2)$, and $n > 0$ be an even number. Let U_n be the uniform distribution on $[n]$, and let $p^{\bullet}$ be the distribution defined in Eq. (13). Then there exists a coupling $\mathcal{C}^{\bullet}$ between $U_n^{\otimes s}$ and $p^{\bullet\otimes s}$ such that for any $(T_1^s, T_2^s) \sim \mathcal{C}^{\bullet}$, we have*

$$\mathbf{E}[\mathrm{Ham}(T_1^s, T_2^s)] \leq C(\varepsilon')^2 \cdot \min\left(\frac{s^2}{n}, \frac{s^{3/2}}{n^{1/2}}\right) \tag{15}$$

for some sufficiently large constant $C > 0$.

Finally, we will show that there exists a coupling between $T_1^s \sim U_n^{\otimes s}$ and $T_3^s \sim \mathcal{C}^{\diamond\otimes s}$ such that the Hamming distance between the two is small.

Lemma 22 *Let $\eta, \alpha' > 0$ be parameters such that $\eta - \alpha' \in (0, 1/2)$, and $n > 0$ be an even number. Let U_n be the uniform distribution on $[n]$, and let $p^{\diamond}$ be the distribution defined in Eq. (14). Then there exists a coupling $\mathcal{C}^{\diamond}$ such that for any $(T_1^s, T_3^s) \sim \mathcal{C}^{\diamond}$, we have*

$$\mathbf{E}[\mathrm{Ham}(T_1^s, T_3^s)] = s(\eta - \alpha') \tag{16}$$

Proof We will construct the coupling by the following process:

1. Draw $P_1, ..., P_s \sim \mathrm{Unif}([n/2])$.
2. For each $j \in [s]$, draw $Z_j \sim \mathbf{Bern}(1/2)$ and $Z_j' \sim \mathbf{Bern}(1 - 2(\eta - \alpha'))$. Let $X_j = 2P_j - Z_j$. Then define Y_j as
$$Y_j = \begin{cases} 2P_j - Z_j, & Z_j = 0 \\ 2P_j - Z_j', & Z_j = 1 \end{cases}$$
3. Let $T_1^s = \{X_1, ..., X_n\}$ and $T_3^s = \{Y_1, ..., Y_n\}$. Return (T_1^s, T_3^s).

First, we will show that this is a valid coupling: that is, the marginal distributions on T_1^s and T_3^s are $U_n^{\otimes s}$ and $p^{\diamond\otimes s}$ respectively. For all $i \in [n]$ and $j \in [s]$,

$$\mathbf{Pr}[X_j = i] = \frac{1}{2} \cdot \mathbf{Pr}\left[P_j = \left\lceil \frac{i}{2} \right\rceil\right] = \frac{1}{n}$$

Since the values X_j are independent, the marginal on T_1^s is $U_n^{\otimes s}$. For all even $i \in [n]$ and $j \in [s]$, we also have

$$\begin{aligned}\mathbf{Pr}[Y_j = i] &= \mathbf{Pr}\left[P_j = \frac{i}{2}\right] \cdot \left(\mathbf{Pr}[Z_j = 0] + \mathbf{Pr}\left[Z'_j = 0 \mid Z_j = 1\right] \cdot \mathbf{Pr}[Z_j = 1]\right) \\ &= \frac{2}{n}\left(\frac{1}{2} + 2(\eta - \alpha) \cdot \frac{1}{2}\right) \\ &= \frac{1 + 2(\eta - \alpha')}{n}\end{aligned}$$

For all odd i,

$$\begin{aligned}\mathbf{Pr}[Y_j = i] &= \mathbf{Pr}\left[P_j = \left\lceil \frac{i}{2} \right\rceil\right] \cdot \mathbf{Pr}\left[Z'_j = 1 \mid Z_j = 1\right] \cdot \mathbf{Pr}[Z_j = 1] \\ &= \frac{2}{n} \cdot (1 - 2(\eta - \alpha')) \cdot \frac{1}{2} \\ &= \frac{1 - 2(\eta - \alpha')}{n}\end{aligned}$$

which proves that $\mathcal{C}^\diamond$ is a valid coupling.

Finally, we bound the expected Hamming distance between the two sample sets. For all j, $X_j \neq Y_j$ if and only if $Z_j = 1$ and $Z'_j = 0$. Therefore,

$$\begin{aligned}\mathbf{E}[\mathrm{Ham}(T_1^s, T_3^s)] &= \mathbf{E}\left[\sum_{j=1}^{s} \mathbb{1}[X_j \neq Y_j]\right] = \sum_{j=1}^{s} \mathbf{Pr}[X_j \neq Y_j] \\ &= \sum_{j=1}^{s} \mathbf{Pr}[Z_j = 1] \cdot \mathbf{Pr}\left[Z'_j = 0\right] \\ &= \sum_{j=1}^{s} \frac{1}{2} \cdot 2(\eta - \alpha') \\ &= s(\eta - \alpha')\end{aligned}$$

as desired. ■

B.4. Proof of Theorem 19

Proof Suppose $\mathcal{A}\colon [n]^s \to \{\mathsf{accept}, \mathsf{reject}, \perp\}$ is an $(\alpha, \varepsilon, \delta = 0.2)$-augmented identity tester which uses

$$s < \min\left(\max\left(\frac{0.004\sqrt{n}}{(\varepsilon')^2}, \frac{C_1\sqrt{n}}{\varepsilon'\sqrt{\xi}}, \frac{C_2 n^{1/3}}{(\varepsilon')^{4/3}\xi^{2/3}}\right), \max\left(\frac{0.00005}{(\eta - \alpha')^2}, \frac{0.4}{\xi(\eta - \alpha')}\right)\right) \tag{17}$$

for some constants C_1, C_2. We will use Lemma 20 and the distributions constructed above to show that $\mathcal{A}$ must have a high error probability if its sample complexity is too low. Let $T_1^s \sim U_n^{\otimes s}$, $T_2^s \sim p^{\bullet \otimes s}$, and $T_3^s \sim p^{\diamond \otimes s}$. We have the following facts about the distributions:

- By a result of Paninski (2008), no algorithm can distinguish U_n from $p^\bullet$ with probability greater than 0.505 with fewer than $0.004\sqrt{n}/(\varepsilon')^2$ samples.
- By Lemma 21, for any $s \le \max(C_1\sqrt{n}/(\varepsilon'\sqrt{\xi}), C_2 n^{1/3}/((\varepsilon')^{4/3}\xi^{2/3}))$ there exists a coupling $\mathcal{C}^\bullet$ between T_1^s and T_2^s such that

$$\mathbf{E}_{(T_1^s,T_2^s)\sim\mathcal{C}^\bullet}[\mathrm{Ham}(T_1^s,T_2^s)] \le \frac{0.4}{\xi}$$

- No algorithm can distinguish U_n from $p^\bullet$ with probability greater than 0.505 with fewer than $0.00005/(\eta-\alpha')^2$ samples. This follows by noting that distinguishing a distribution which is biased toward even elements from the uniform distribution is equivalent to distinguishing a $\eta-\alpha'$-biased coin from a fair one. It is a folklore fact that this requires at least $\Omega(1/(\eta-\alpha')^2$ samples. (See, e.g., Aliakbarpour et al. (2024)).
- By Lemma 22, if $s \le 0.4/(\xi(\eta-\alpha'))$, there exists a coupling $\mathcal{C}^\diamond$ between T_1^s and T_3^s such that

$$\mathbf{E}_{(T_1^s,T_3^s)\sim\mathcal{C}^\diamond}[\mathrm{Ham}(T_1^s,T_3^s)] \le \frac{0.4}{\xi}$$

Since $\mathcal{A}$ has only three possible outputs, we have:

$$1 = \mathbf{Pr}_{T_1\sim U_n^{\otimes s}}[\mathcal{A}(T_1)=\mathsf{accept}] + \mathbf{Pr}_{T_1\sim U_n^{\otimes s}}[\mathcal{A}(T_1)=\mathsf{reject}] + \mathbf{Pr}_{T_1\sim U_n^{\otimes s}}[\mathcal{A}(T_1)=\perp]$$

Using Lemma 20, we have:

$$1 \le \mathbf{Pr}_{T_1\sim U_n^{\otimes s}}[\mathcal{A}(T_1)=\mathsf{accept}] + 1.5\cdot\mathbf{Pr}_{T_2\sim p^{\dot\otimes s}}[\mathcal{A}(T_2)=\mathsf{reject}] + 1.5\cdot\mathbf{Pr}_{T_3\sim p^{\diamond\otimes s}}[\mathcal{A}(T_3)=\perp] + 0.02$$

Based on our definition of U_n, $p^\bullet$ and $p^\diamond$, the three probabilities above are at most the probability that $\mathcal{A}$ makes a mistake. Thus, we have:

$$1 \le 4\delta + 0.02$$

Hence, $\delta \ge 0.245$, which contradicts our assumption that $\delta = 0.2$. Therefore, s cannot be lower than the bound in Eq. (17), which immediately implies the lower bound in Eq. (11). ■

Appendix C. Upper Bound for Private Augmented Closeness Testing

Our result for closeness testing combines the flattening-based closeness testers of Aliakbarpour et al. (2024) and Aliakbarpour et al. (2019). The flattening technique, introduced by Diakonikolas and Kane (2016), uses samples to map the original distributions p and q onto new distributions p' and q' such that p' has a low ℓ_2 norm, but $\mathrm{d_{TV}}(p,q) = \mathrm{d_{TV}}(q',q')$. The main result of Aliakbarpour et al. (2019) is that any flattening procedure which satisfies certain technical conditions can be made private. Since the augmented closeness tester of Aliakbarpour et al. (2024) also relies on flattening, we show that it too can be made private with some modifications. In Appendix C.1, we review the flattening procedures of Aliakbarpour et al. (2024) and Aliakbarpour et al. (2019). We describe our own procedure in Appendix C.2, and present our algorithm in Appendix C.3. Finally, in Appendix C.4, we show how to privately estimate the ℓ_2 norm to verify the success of our flattening.

C.1. Overview of Flattening

In this section, we formally state the guarantees of previous flattening-based testers described in Section 3.1. We restate the descriptions of both algorithms for convenience, followed by the technical results we need for our analysis.

Private Flattening: The private tester of Aliakbarpour et al. (2019) attempts to derandomize the original flattening of Diakonikolas and Kane (2016). When a flattening-based tester divides its samples into flattening and test sets, it introduces randomness based on the permutation π used to partition the samples. This results in high sensitivity of the test statistic Z, as swapping one of the samples used for the flattening can wildly affect the resulting value of Z. The private tester reduces the sensitivity by taking the expectation of Z over every permutation π of the data and the randomness r of the flattening procedure:

$$\overline{Z} \coloneqq \mathbf{E}_{\pi,r}[Z]$$

The authors show that $\overline{Z}$ has low sensitivity and preserves the correctness of any flattening procedure which satisfies two technical conditions. Such a flattening is called a *proper procedure*.

Definition 23 *Let $\mathcal{A}$ be an algorithm which draws a multiset X from distribution p and reduces testing property $\mathcal{P}$ of p to testing closeness between two distributions p' and q'. We say $\mathcal{A}$ is a proper procedure if there exist constants $c_0 < 1$ and $c_1 \geq 1$ such that the following holds for any p' and q' produced by $\mathcal{A}$:*

$$\mathbf{Pr}_X\left[\mathbf{E}_\pi\left[\left\|p'-q'\right\|_2^2 \,\middle|\, X\right] \geq 4c_0 \cdot \mathbf{E}_{\mathcal{A}}\left[\left\|p'-q'\right\|_2^2\right]\right] \geq 9/10 \tag{18}$$

$$\mathbf{E}_{\mathcal{A}}\left[\left\|p'-q'\right\|_4^4\right] \leq c_1 \cdot \left(\mathbf{E}_{\mathcal{A}}\left[\left\|p'-q'\right\|_2^2\right]\right)^2 \tag{19}$$

As we show in Appendix C.2, a crucial step in our flattening procedure is identical to one proven to be proper in Aliakbarpour et al. (2019): the flattening of Diakonikolas and Kane (2016) when the flattening set is drawn from both p and q.

Lemma 24 *Let $\mathcal{A}$ be an algorithm for testing closeness of distributions p and q with the following flattening procedure: draw a multiset $F^{(p)}$ from p and a multiset $F^{(q)}$ from q. Let k_i denote the number of instances of element i in $F^{(p)} \cup F^{(q)}$. Divide the probability mass of element i evenly into k_i buckets. Then $\mathcal{A}$ is a proper procedure.*

The sample complexity of any proper procedure is given by the following result.

Lemma 25 (Aliakbarpour et al. (2019)) *Let $\mathcal{A}$ be a proper procedure for testing property $\mathcal{P}$ which reduces the problem to testing closeness of p' and q' over a domain of size n' using $\mathbf{Poi}(k)$ flattening samples. Then there exists an $(\xi, \varepsilon, 3/4)$-private tester for property $\mathcal{P}$ which draws $\mathbf{Poi}(s)$ test samples, for any*

$$s \geq \Theta\left(\frac{n' \cdot \sqrt{\mathbf{E}_{\mathcal{A}}\left[\min(\|p'\|_2^2, \|q'\|_2^2)\right]}}{\varepsilon^2} + \frac{\sqrt{n\Delta(\overline{Z})}}{\varepsilon\sqrt{\xi}}\right) \tag{20}$$

Additionally, the sensitivity of the test statistic $\overline{Z}$ is bounded by

$$\Delta(\overline{Z}) = O\left(\frac{s+k}{k}\right) \tag{21}$$

Augmented Flattening: In the augmented setting of Aliakbarpour et al. (2024), the estimate $\hat{p}$ guides the choice of buckets for each element i. Let $\nu \in (0, 1]$ be a flattening parameter. Rather than dividing elements into $k_i + 1$ buckets, the flattener creates

$$b_i = \left\lfloor \frac{\hat{p}_i}{\nu} \right\rfloor + k_i + 1 \tag{22}$$

buckets in flattened distributions p' and q'. If $\hat{p}$ is α-close to p, the expected ℓ_2 norm after flattening is shown to be

$$\mathbf{E}\left[\|p'\|_2^2\right] \leq \frac{2\alpha}{k} + 4 \cdot \nu \tag{23}$$

The tester then draws $O(\sqrt{n})$ samples each from p' and q' to estimate their ℓ_2^2 norm. If at least one distribution has been sufficiently flattened, the algorithm proceeds to the test phase. Otherwise, it returns $\perp$. The sample complexity of this tester is given below.

Lemma 26 (Aliakbarpour et al. (2024)) *Fix parameters $\alpha, \varepsilon \in (0, 1]$. Let p and q be unknown distributions, and $\hat{p}$ be a known distribution over $[n]$. There exists an $(\varepsilon, \alpha, \delta = 0.3)$-augmented closeness tester for p and q which uses $\Theta\left(\frac{n^{2/3}\alpha^{1/3}}{\varepsilon^{4/3}} + \frac{\sqrt{n}}{\varepsilon^2}\right)$ samples from both p and q.*

As discussed in the introduction, to mimic this approach we need to perform this last step – the estimation of the ℓ_2^2 norm to verify that the flattening was successful – in a *differentially private* fashion.

C.2. Two-Step Flattening

To simplify the analysis of our augmented private closeness tester, we split the flattening into two steps. The first step flattens p and q into distributions p' and q' using only the prediction $\hat{p}$. The second step draws samples from p' and q' and uses these samples to flatten them into a third set of distributions p'' and q''. This two-step flattening makes it much easier to guarantee that the procedure is private. Since the first step draws no samples, it is already differentially private. Since the second step does not use $\hat{p}$, it is precisely the same procedure as the flattening described in Lemma 24, and can be made private by the same mechanism. We show that if $\mathrm{d_{TV}}(\hat{p}, p) \leq \alpha$, then $\|p''\|_2^2$ is small in expectation, allowing us to test efficiently with fewer flattening samples than the optimal non-augmented tester.

Flattening procedure: Let k be a parameter which will determine the expected number of flattening samples. The two-step flattening works as follows.

1. From the original distributions p and q, estimate $\hat{p}$, and suggested accuracy α, construct p' and q' by flattening each element $i \in [n]$ into b_i buckets, where

$$b_i = \lceil n\hat{p}_i \rceil + 1 \tag{24}$$

(This is analogous to setting $\nu = 1/n$ in Equation 22). Since $\hat{p}$ and α are not private, this can be done as a preprocessing step before running the private portion of the algorithm. After this step, p' is defined by having, for all $i \in [n]$ and $j \in [b_i]$, $p'_{i,j} = p_i/b_i$. Sampling from p' can be done as follows: sample $i \sim p$ and $j \sim \mathrm{Unif}([b_i])$, and return (i, j). Sampling from q' is analogous (note that both p and q use the same bucketing, so p' and q' have the same domain).

2. Sample $\hat{k}_p, \hat{k}_q \sim \mathbf{Poi}(k)$ independently. Draw a multiset $F^{(p)}$ of $\hat{k}_p$ samples from p' and a multiset $F^{(q)}$ of $\hat{k}_q$ samples from q'. Let $F = F^{(p)} \cup F^{(p)}$, and $k_{i,j}$ be the total number of instances of element (i,j) in F. Then divide each element of the domain of p', q' into $b_{i,j} = k_{i,j} + 1$ buckets.

   The domain of p'' and q'' is $\Omega = \{(i,j,m) : i \in [n], j \in [b_i], m \in [b_{i,j}]\}$, and

$$p''_{i,j,m} = \frac{p_{i,j}}{b_{i,j}}, \qquad q''_{i,j,m} = \frac{q_{i,j}}{b_{i,j}} \tag{25}$$

   for all $(i,j,m) \in \Omega$. To sample from p'', we sample $(i,j) \sim p'$, $m \sim \mathrm{Unif}([b_{i,j}])$, then return (i,j,m).

Effect of flattening: We now show that the domain size after flattening remains $\Theta(n+k)$ with high probability. After the second flattening step, the number of elements is $n'' = |\Omega|$ (itself a random variable), where

$$n'' = \sum_{i=1}^{n}\sum_{j=1}^{b_i} b_{i,j} = \sum_{i=1}^{n}\sum_{j=1}^{b_i}(k_{i,j}+1) = \hat{k}_p + \hat{k}_q + \sum_{i=1}^{n} b_i \le \hat{k}_p + \hat{k}_q + \sum_{i=1}^{n}(n\hat{p}_i + 2) \le \hat{k}_p + \hat{k}_q + 3n$$

Note that $\hat{k}_p + \hat{k}_q \sim \mathbf{Poi}(2k)$, so $\hat{k}_p + \hat{k}_q \le 200k$ with probability at least 0.99 by Markov's inequality. Therefore, $n'' = O(n+k)$ with probability at least 0.99. Next, we bound the expected ℓ_2^2 norm of the flattened distribution, assuming the suggested accuracy level α is valid.

Lemma 27 (See, e.g., (Diakonikolas and Kane, 2016, Lemma 2.6)) *Fix $\lambda > 0$. If $X \sim \mathbf{Poi}(\lambda)$, then $\mathbf{E}\left[\frac{1}{1+X}\right] \le \frac{1}{\lambda}$.*

Proof Follows from manipulating the series corresponding to the expectation:

$$\mathbf{E}\left[\frac{1}{1+X}\right] = e^{-\lambda}\sum_{k=0}^{\infty}\frac{1}{k+1}\frac{\lambda^k}{k!} = \frac{1}{\lambda}e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^{k+1}}{(k+1)!} = \frac{1}{\lambda}e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{k!} = \frac{1-e^{-\lambda}}{\lambda}.$$

■

Lemma 28 *Let p and $\hat{p}$ be distributions over $[n]$ with $\mathrm{d_{TV}}(p,\hat{p}) \le \alpha$. Then the two-step flattening produces a distribution p'' with*

$$\mathbf{E}_F\left[\|p''\|_2^2\right] \le \frac{2\alpha}{k} + \frac{4}{n} \tag{26}$$

Proof The proof is inspired by Aliakbarpour et al. (2024). Let $\Delta_i = p_i - \hat{p}_i$ for all $i \in [n]$, so that

$$p_i = \Delta_i + \hat{p}_i \le 2\max(\Delta_i, \hat{p}_i)$$

Letting $A \coloneqq \{i \in [n] : \Delta_i \ge \hat{p}_i\}$, this gives the following bound on p_i:

$$p_i \le \begin{cases} 2\Delta_i, & i \in A \\ 2\hat{p}_i, & i \in [n] \setminus A \end{cases}$$

By definition of p' and p'', we have

$$\begin{aligned}
\|p''\|_2^2 &= \sum_{i=1}^{n}\sum_{j=1}^{b_i}\sum_{m=1}^{b_{i,j}}(p''_{i,j,m})^2 = \sum_{i=1}^{n}\sum_{j=1}^{b_i}\frac{(p'_{i,j})^2}{b_{i,j}} \\
&= \sum_{i\in A}\sum_{j=1}^{b_i}\frac{(p'_{i,j})^2}{b_{i,j}} + \sum_{i\in[n]\setminus A}\sum_{j=1}^{b_i}\frac{(p'_{i,j})^2}{b_{i,j}} \\
&\le \sum_{i\in A}\sum_{j=1}^{b_i}\frac{2\Delta_i p'_{i,j}}{b_i b_{i,j}} + \sum_{i\in[n]\setminus A}\sum_{j=1}^{b_i}\frac{p_i^2}{b_i^2 b_{i,j}} \\
&\le \sum_{i\in A}\sum_{j=1}^{b_i}\frac{2\Delta_i p'_{i,j}}{b_i b_{i,j}} + \sum_{i\in[n]\setminus A}\sum_{j=1}^{b_i}\frac{p_i^2}{b_i^2} && (b_{i,j}\ge 1) \\
&\le \sum_{i\in A}\sum_{j=1}^{b_i}\frac{2\Delta_i p'_{i,j}}{b_i(k_{i,j}+1)} + \sum_{i\in[n]\setminus A}\sum_{j=1}^{b_i}\frac{(2\hat{p}_i)^2}{b_i^2} \\
&\le \sum_{i\in A}\sum_{j=1}^{b_i}\frac{2\Delta_i p'_{i,j}}{b_i(k_{i,j}+1)} + 4\sum_{i\in[n]\setminus A}\frac{\hat{p}_i^2}{b_i} \\
&\le \sum_{i\in A}\sum_{j=1}^{b_i}\frac{2\Delta_i p'_{i,j}}{b_i(k_{i,j}+1)} + 4\sum_{i\in[n]\setminus A}\frac{\hat{p}_i^2}{n\hat{p}_i} && (b_i \ge n\hat{p}_i) \\
&\le \sum_{i\in A}\sum_{j=1}^{b_i}\frac{2\Delta_i p'_{i,j}}{b_i(k_{i,j}+1)} + \frac{4}{n}
\end{aligned}$$

Note that we can write $k_{i,j} = k_{i,j}^{(p)} + k_{i,j}^{(q)}$, where $k_{i,j}^{(p)} \sim \mathbf{Poi}\left(kp'_{i,j}\right)$ and $k_{i,j}^{(q)} \sim \mathbf{Poi}\left(kq'_{i,j}\right)$. Applying Lemma 27,

$$\mathbf{E}_F\left[\frac{1}{k_{i,j}+1}\right] \le \mathbf{E}_F\left[\frac{1}{k_{i,j}^{(p)}+1}\right] \le \frac{1}{kp'_{i,j}}$$

Combining the two results, the expected norm is bounded by

$$\mathbf{E}_F\left[\|p''\|_2^2\right] \le \sum_{i\in A}\left(\frac{2\Delta_i}{k}\right) + \frac{4}{n} \le \frac{2\alpha}{k} + \frac{4}{n}$$

the last inequality since $\sum_{i\in A}\Delta_i = p(A) - \hat{p}(A) \le \sup_S(p(S) - \hat{p}(S)) = \mathrm{d_{TV}}(p,\hat{p}) \le \alpha$. ■

C.3. The Algorithm

At a high level, our algorithm works in three steps. The first step flattens p and q into distributions p'' and q'' as described in Appendix C.2. The second step tests the ℓ_2^2 norm of p''. If the norm is not within a constant multiplicative factor of the bound guaranteed by Lemma 28, then we have $\mathrm{d_{TV}}(\hat{p},p) > \alpha$ with high probability, and the tester returns $\perp$. If the norm *is* close to the bound, our

algorithm can perform an efficient closeness test. The final step runs the private closeness tester of Aliakbarpour et al. (2019) and returns its result. We give pseudocode for our algorithm in Algorithm 1.

By our lower bound (Theorem 38), there are some regimes of parameters in which an augmented tester cannot use asymptotically fewer samples than a non-augmented private tester. In these cases, we simply invoke the non-augmented tester, whose sample complexity is given by Lemma 29. In all other regimes, we show that our tester is optimal up to $\log n$ factors.

Lemma 29 (Zhang (2021)) *Let $\varepsilon \in (0,1]$ and $\xi > 0$. Let p and q both be unknown distributions over $[n]$. Then there is a $(\xi, \varepsilon, 0.05)$-private closeness tester for p and q which takes $\Theta\left(\frac{n^{2/3}}{\varepsilon^{4/3}} + \frac{\sqrt{n}}{\varepsilon^2} + \frac{1}{\varepsilon\xi} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} + \frac{n^{1/3}}{\varepsilon^{4/3}\xi^{2/3}}\right)$ samples.*

We are now ready to prove our upper bound.

Theorem 30 *[Theorem 6, restated] Let $\varepsilon, \alpha \in (0,1]$ and $\xi > 0$. Let p and q both be unknown distributions over $[n]$. Then Algorithm 1 is a ξ-differentially private augmented $(\varepsilon, \alpha, 0.32)$-closeness tester which takes s samples each from p and q, where*

$$s = \tilde{O}\left(\frac{n^{2/3}\alpha^{1/3}}{\varepsilon^{4/3}} + \frac{\sqrt{n}}{\varepsilon^2} + \frac{1}{\varepsilon\xi} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} + \frac{n^{1/3}}{\varepsilon^{4/3}\xi^{2/3}}\right) \tag{27}$$

Proof For the analysis, we will condition on two assumptions:

1. $\hat{k} = \hat{k}_p + \hat{k}_q < 200k$, $\hat{\ell} < 100\ell$, and $\hat{s} < 100s$. This holds with probability at least 0.97 by Markov's inequality and a union bound.

2. For all $i \in [n]$,
$$\frac{\ell_i}{k_i + 1} < 12 \log\left(\frac{n}{0.05}\right) \cdot \frac{\ell}{k} \tag{28}$$
This holds with probability at least 0.95 by Lemma 37. If this assumption does not hold in our original dataset, we enforce it via the mapping of Lemma 35, which preserves privacy.

Privacy: If $\varepsilon = o(n^{-1/4})$ or $\varepsilon^2\xi = o(n^{-1})$, we run the tester of Lemma 29, which is private by the same result.

If $\varepsilon = \Omega(n^{-1/4})$ and $\varepsilon^2\xi = \Omega(n^{-1})$, the algorithm's outputs depend only on the statistic $\tilde{L}$, and the choice of $\hat{k}_p$, $\hat{k}_q$ and $\hat{\ell}$. The latter two do not depend on the sampled data. If assumption (2) is satisfied, then $\tilde{L}$ is ξ- differentially private by Lemma 12 . Finally, the second flattening step is a proper procedure by Lemma 24. Therefore, if the algorithm reaches Line 12, its output is private by Lemma 25.

Correctness: If $\varepsilon = o(n^{-1/4})$ or $\varepsilon^2\xi = o(n^{-1})$, the algorithm returns a valid answer with probability at least 0.95 by Lemma 29. Therefore, we need only consider the case where $\varepsilon = \Omega(n^{-1/4})$ and $\varepsilon^2\xi = \Omega(n^{-1})$.

First, we show that if $\mathrm{d}_{\mathrm{TV}}(p, \hat{p}) \le \alpha$, the algorithm does not output $\perp$ with high probability. The only case in which the algorithm returns $\perp$ is when $\tilde{L} > 30(2\alpha/k + 4/n)$. By Lemma 12 , the estimate $\tilde{L}$ is within a constant factor of $\|p''\|_2^2$ with probability at least 0.94:

$$\frac{\|p''\|_2^2}{2} \le \tilde{L} \le 3\frac{\|p''\|_2^2}{2} \tag{29}$$

Combining Lemma 28 with Markov's inequality, the following holds with probability at least 0.95:

$$\|p''\|_2^2 \le 20\left(\frac{2\alpha}{k} + \frac{4}{n}\right) \tag{30}$$

Therefore, by union bound, we have that $\tilde{L}$ is not more than 20 times the expected value of $\|p''\|_2^2$ with probability at least 0.91, and the probability of returning $\perp$ in this case is at most 0.09.

Next, we show that the algorithm does not output reject with high probability when $p = q$. Conditioned on assumption (1), the algorithm only rejects if the tester of Lemma 25 rejects. This occurs with probability at most 0.25 when $p = q$.

Finally, we show that the algorithm does not return accept when $\mathrm{d_{TV}}(p, q) > \varepsilon$. In this case, the algorithm only accepts if the tester of Lemma 25 accepts, which happens with probability at most 0.25.

Therefore, the algorithm outputs a valid answer with probability 0.75 conditioned on the assumptions. Taking a union bound, the algorithm outputs a valid answer with probability 0.68 overall.

Sample Complexity: By assumption (1), we draw $O(k)$ flattening samples, $O(\ell)$ estimation samples, and $O(s)$ test samples from both p and q. Let $n'' = O(n + k)$ be the domain size of p''. Then the parameters must satisfy the following constraints:

$$k \cdot \min\left(\frac{k}{\log^2 n}, \frac{\ell}{\log n}\right) = \Omega\left(\frac{n''}{\xi}\right) \quad \text{(for } \ell_2 \text{ norm estimation)} \tag{31}$$

$$\ell \ge \Theta(\sqrt{n''}) \quad \text{(for } \ell_2 \text{ norm estimation)} \tag{32}$$

$$s \ge \Theta\left(\frac{n'' \cdot \sqrt{\mathbf{E}_F\left[\|p''\|_2^2\right]}}{\varepsilon^2} + \frac{\sqrt{n''\Delta(\overline{Z})}}{\varepsilon\sqrt{\xi}}\right) \quad \text{(for testing)} \tag{33}$$

By Lemma 25, the sensitivity of the test statistic is

$$\Delta(\overline{Z}) = \Theta\left(\frac{s+k}{k}\right).$$

And by Lemma 28, we have

$$\mathbf{E}_F\left[\|p''\|_2^2\right] \le \sqrt{\frac{2\alpha}{k} + \frac{4}{n}}.$$

Finally, to simplify the constraint in Eq. (31), we note that the equation implies two constraints on k:

$$\frac{k^2}{\log^2 n} = \Omega\left(\frac{k}{\xi}\right), \text{ so } k = \Omega\left(\frac{\log^2 n}{\xi}\right)$$
$$\frac{k^2}{\log^2 n} = \Omega\left(\frac{n}{\xi}\right), \text{ so } k = \Omega\left(\frac{\sqrt{n}\log n}{\sqrt{\xi}}\right).$$

If both of the above hold and $\ell = k$, the parameters always satisfy Eq. (31). Note that the first constraint dominates if and only if $\xi = o(\log n/\sqrt{n})$.

We are now ready to describe the sample complexity. We will consider the following cases:

Case 1: $\varepsilon = o(n^{-1/4})$. In this case we run the non-augmented private tester, whose sample complexity is given by Lemma 29. We have

$$\frac{n^{2/3}}{\varepsilon^{4/3}} = \frac{n^{2/3}\varepsilon^{2/3}}{\varepsilon^2} = O\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$$

Therefore, $n^{2/3}/\varepsilon^{4/3}$ is not the dominating term, and the sample complexity can be written as

$$O\left(\frac{\sqrt{n}}{\varepsilon^2} + \frac{n^{1/3}}{\varepsilon^{4/3}\xi^{2/3}} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} + \frac{1}{\varepsilon\xi}\right)$$

which matches our lower bound.

Case 2: $\varepsilon = \Omega(n^{-1/4})$ and $\varepsilon^2\xi = o(n^{-1})$. Once again, we run the non-augmented tester. By our assumption,

$$\varepsilon^2\xi = o(n^{-1}) = o\left(\frac{n^{-2/3}}{n^{1/3}}\right) = o\left(\frac{\varepsilon^{8/3}}{n^{1/3}}\right)$$

Rearranging terms, we have $n/\varepsilon^2 = o(1/\xi^3)$. Consequently,

$$\frac{n^{2/3}}{\varepsilon^{4/3}} = \left(\frac{n}{\varepsilon^2}\cdot\frac{n^3}{\varepsilon^6}\right)^{1/6} = O\left(\left(\frac{1}{\xi^3}\cdot\frac{n^3}{\varepsilon^6}\right)^{1/6}\right) = O\left(\frac{\sqrt{n}}{\varepsilon\sqrt{\xi}}\right)$$

Therefore, $n^{2/3}/\varepsilon^{4/3}$ is not the dominating term. The sample complexity is

$$O\left(\frac{\sqrt{n}}{\varepsilon^2} + \frac{n^{1/3}}{\varepsilon^{4/3}\xi^{2/3}} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} + \frac{1}{\varepsilon\xi}\right)$$

which matches our lower bound.

Case 3: $\varepsilon = \Omega(n^{-1/4})$, $\varepsilon^2\xi = \Omega(n^{-1})$, and

$$\frac{\sqrt{n}\log n}{\sqrt{\xi}} = O\left(\frac{n^{2/3}\alpha^{1/3}}{\varepsilon^{4/3}} + \frac{\sqrt{n}}{\varepsilon^2} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}}\right) \tag{34}$$

In this case, we run the augmented tester. We set

$$\ell = k = \Theta\left(\frac{n^{2/3}\alpha^{1/3}}{\varepsilon^{4/3}} + \frac{\sqrt{n}}{\varepsilon^2} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}}\right) = O(n) \tag{35}$$

and

$$s = \Theta\left(\frac{n}{\varepsilon^2}\sqrt{\frac{2\alpha}{k} + \frac{4}{n}} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}}\right) = O\left(\frac{n^{2/3}\alpha^{1/3}}{\varepsilon^{4/3}} + \frac{\sqrt{n}}{\varepsilon^2} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}}\right) = O(k)$$

Therefore, $\Delta(\overline{Z}) = \Theta(1)$, and s satisfies Eq. (33). By our assumptions, we have the following facts:

$$\frac{n^{2/3}\alpha^{1/3}}{\varepsilon^{4/3}} \leq n\alpha^{1/3} = O(n)$$
$$\frac{\sqrt{n}}{\varepsilon^2} = O(n)$$
$$\frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} = O\left(n\right)$$

We also have $\ell = k = \Omega(\sqrt{n}/\varepsilon^2) = \Omega(\sqrt{n})$, which satisfies Eq. (32). Finally, $k = \Omega(\sqrt{n}\log n/\sqrt{\xi}) = \Omega(\sqrt{n+k}\log n/\sqrt{\xi})$, which satisfies Eq. (31). In this case, the overall sample complexity is equal to Eq. (35). Each of the terms in this upper bound appear in our lower bound, so the sample complexity is optimal.

Case 4: $\varepsilon = \Omega(n^{-1/4})$, $\varepsilon^2\xi = \Omega(n^{-1})$, and the following holds:

$$\begin{cases} \frac{n^{2/3}\alpha^{1/3}}{\varepsilon^{4/3}} + \frac{\sqrt{n}}{\varepsilon^2} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} = O\left(\frac{\sqrt{n}\log n}{\sqrt{\xi}}\right) \\ O\left(\frac{\sqrt{n}\log n}{\sqrt{\xi}}\right) = O(n) \end{cases} \quad \text{and,}$$

In this case, we run the augmented tester. The $\sqrt{n}\log n/\sqrt{\xi}$ term dominates each term in Eq. (35). Therefore, we set $\ell = k = \Theta(\sqrt{n}\log n/\sqrt{\xi})$, and

$$s = \Theta\left(\frac{n}{\varepsilon^2}\sqrt{\frac{2\alpha}{k} + \frac{4}{n}} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}}\right) = O(k)$$

Similar to the previous case, this setting satisfies all three constraints. The sample complexity is $\Theta(\sqrt{n}\log n/\sqrt{\xi})$. Since a $\sqrt{n}/(\varepsilon\sqrt{\xi})$ term appears in our lower bound, our sample complexity will be off by a factor of $O(\varepsilon \cdot \log n)$ compared to the lower bound.

Case 5: $\varepsilon = \Omega(n^{-1/4})$, $\varepsilon^2\xi = \Omega(n^{-1})$, and $\sqrt{n}\log n/\sqrt{\xi} = \Omega(n)$. In this case, we run the augmented tester. We set $\ell = k = \Theta(\log^2 n/\xi)$. Since $k = \Omega(n)$, we have $n'' = O(k)$ and $\Delta(\overline{Z}) = \Theta((s+k)/k)$. Therefore, we require

$$s = \Omega\left(\frac{k}{\varepsilon^2}\sqrt{\frac{\alpha}{k} + \frac{1}{n}} + \frac{\sqrt{s+k}}{\varepsilon\sqrt{\xi}}\right)$$

Notice that $\alpha/k = O(1/n)$, so solving for s yields

$$s = \Theta\left(\frac{k}{n\varepsilon^2} + \frac{1}{\varepsilon^2\xi} + \frac{\sqrt{k}}{\varepsilon\sqrt{\xi}}\right)$$

We can bound each of these terms using our assumptions. Since $\sqrt{n}\log n/\xi = \Omega(n)$, we must have $\xi = O(\log n/\sqrt{n})$. Using these facts, we can bound s:

$$\frac{k}{n\varepsilon^2} = O\left(\frac{\log^2 n}{n\varepsilon^2\xi}\right) = O(\log^2 n) = O(k)$$
$$\frac{1}{\varepsilon^2\xi} = O(n) = O(k)$$
$$\frac{\sqrt{k}}{\varepsilon\sqrt{\xi}} = \Theta\left(\frac{\log n}{\varepsilon\xi}\right)$$

Therefore, the overall sample complexity is $O(\log^2 n/\xi + \log n/(\varepsilon\xi))$. Once again, this is within a factor of $O(\varepsilon \cdot \log^2 n + \log n)$ of our lower bound.

■

C.4. Private Testing of ℓ_2 Norm

To determine whether the two-step flattening worked, we must (privately) test the ℓ_2 norm of the flattened distribution p''. The standard ℓ_2 tester (Goldreich and Ron, 2011; Aliakbarpour et al., 2024) works by computing the number of collisions between elements in its sample set. This algorithm has the following guarantee:

Lemma 31 (Goldreich and Ron (2011); Aliakbarpour et al. (2024)) *Let $\delta \in (0,1)$ and p be a distribution over $[n]$. Let E be a multiset of $\ell = O\left(\sqrt{n}\log(1/\delta)\right)$ samples from p. For each $i \in [n]$, define ℓ_i as the number of instances of element i in E. Let*

$$L(E) = \frac{1}{\binom{\ell}{2}} \sum_{i=1}^{n} \binom{\ell_i}{2} \tag{36}$$

Then with probability at least $1-\delta$,

$$\frac{\|p\|_2^2}{2} \le L(E) \le \frac{3\|p\|_2^2}{2} \tag{37}$$

The statistic above has high sensitivity; intuitively, if a large number of samples fall in the same bucket (i,j,m), changing a single instance of element i could affect a large number of collisions. To avoid this, we will derandomize this tester by taking

$$\overline{L}(E) \coloneqq \mathbf{E}_r[L \mid E] \tag{38}$$

where r is the string of random bits used to choose the bucket when generating a sample from p'' given a sample from p', and E is a multiset (the *estimation set*) of $\mathbf{Poi}(\ell)$ flattening samples from p'. Each of these samples is a pair (i,j). Rather than sample the third coordinate to generate a sample from p'', we will consider the expected number of collisions over all such samplings, conditioned on E. In the result below, we give a closed-form expression for $\overline{L}$.

Lemma 32 *[Lemma 10, restated] Let $k_{i,j}$ and $\ell_{i,j}$ be the number of instances of element (i,j) in the flattening and estimation sets F and E respectively. For each $i \in [n]$, let b_i be the number of buckets created for element i by the first flattening step. Then*

$$\overline{L} = \frac{1}{\binom{\ell}{2}} \sum_{i=1}^{n} \mathbf{E}_r \left[\sum_{j=1}^{b_i} \sum_{k=1}^{b_{i,j}} \binom{\ell_{i,j,k}}{2} \;\middle|\; k_i, \ell_i \right] = \frac{1}{\binom{\ell}{2}} \sum_{i=1}^{n} \sum_{j=1}^{b_i} \frac{\binom{\ell_{i,j}}{2}}{k_{i,j}+1} \tag{39}$$

Proof Fix $k_{i,j}$ and $\ell_{i,j}$. Recall that element (i,j) is divided into $b_{i,j} = k_{i,j} + 1$ buckets. Define $\ell_{i,j,m}$ as the number of instances of bucket (i,j,m) in the estimation set. Since $b_{i,j}$ and $\ell_{i,j}$ are both fixed, $\ell_{i,j,m} \sim \mathbf{Bin}(\ell_{i,j}, 1/b_{i,j})$, where the randomness comes only from r. We have the following:

$$\begin{aligned} \mathbf{E}_r[\ell_{i,j,m}] &= \frac{\ell_{i,j}}{b_{i,j}} \\ \mathbf{E}_r[\ell_{i,j,m}^2] &= \mathbf{Var}[\ell_{i,j,m}] + \mathbf{E}_r[\ell_{i,j,m}]^2 = \frac{\ell_{i,j}}{b_{i,j}}\left(1 - \frac{1}{b_{i,j}}\right) + \frac{\ell_{i,j}^2}{b_{i,j}^2} = \frac{\ell_{i,j}}{b_{i,j}} + \frac{\ell_{i,j}^2 - \ell_{i,j}}{b_{i,j}^2} \end{aligned} \tag{40}$$

We can now write the expectation as

$$\begin{aligned} \frac{1}{\binom{\ell}{2}} \sum_{i=1}^{n} \mathbf{E}_r \left[\sum_{j=1}^{b_i} \sum_{m=1}^{b_{i,j}} \binom{\ell_{i,j,m}}{2} \;\middle|\; k_i, \ell_i \right] &= \frac{1}{\binom{\ell}{2}} \sum_{i=1}^{n} \sum_{j=1}^{b_i} \sum_{m=1}^{b_{i,j}} \mathbf{E}_r \left[\frac{\ell_{i,j,m}^2 - \ell_{i,j,m}}{2} \;\middle|\; k_i, \ell_i \right] \\ &= \frac{1}{\binom{\ell}{2}} \sum_{i=1}^{n} \sum_{j=1}^{b_i} \sum_{m=1}^{b_{i,j}} \left(\frac{\ell_{i,j}}{2b_{i,j}} + \frac{\ell_{i,j}^2 - \ell_{i,j}}{2b_{i,j}^2} - \frac{\ell_{i,j}}{2b_{i,j}} \right) \\ &= \frac{1}{\binom{\ell}{2}} \sum_{i=1}^{n} \sum_{j=1}^{b_i} b_{i,j} \cdot \frac{\ell_{i,j}^2 - \ell_{i,j}}{2b_{i,j}^2} \\ &= \frac{1}{\binom{\ell}{2}} \sum_{i=1}^{n} \sum_{j=1}^{b_i} \frac{\ell_{i,j}^2 - \ell_{i,j}}{2b_{i,j}} \\ &= \frac{1}{\binom{\ell}{2}} \sum_{i=1}^{n} \sum_{j=1}^{b_i} \frac{\binom{\ell_{i,j}}{2}}{k_{i,j}+1} \end{aligned} \tag{41}$$

■

In general, the sensitivity of $\overline{L}$ might be quite high. In the worst case, we might draw no instances of element (i,j) in F but $\Theta(\ell)$ instances in E, giving a sensitivity of $\Theta(1)$. However, note that the sensitivity is considerably lower if the flattening and estimation sets are similar (i.e., no element has a much higher frequency in E than in F). The following lemma formalizes this intuition.

Lemma 33 *Suppose that there exists $A \geq 0$ such that for all $i \in [n]$ and $j \in [b_i]$,*

$$\frac{\ell_{i,j}}{k_{i,j}+1} \leq A \cdot \frac{\ell}{k} \tag{42}$$

Then the sensitivity of $\overline{L}$ is bounded by

$$\Delta(\overline{L}) \leq O\left(\frac{A^2}{k^2} + \frac{A}{k\ell}\right) \tag{43}$$

Proof Assume that X and X' are fixed datasets which differ in only one sample (an instance of (i,j) in X is replaced by an instance of (i',j') in X'). The extra instance can only change the contribution from (i,j) and (i',j'). Let $k_{i,j}$ and $\ell_{i,j}$ represent the number of instances of (i,j) in the the flattening and estimation sets of X respectively. Let $k'_{i,j}$ and $\ell'_{i,j}$ represent the number of instances of (i,j) in the flattening and estimation sets of X'. We will bound the difference for a single term; by symmetry, doubling the bound gives the overall sensitivity.

Case 1: The extra instance of (i,j) falls in the flattening set F. In this case, $\ell_{i,j} = \ell'_{i,j}$ and $k_{i,j} = k'_{i,j} + 1 \geq 1$. As $\ell_{i,j}/(k_{i,j}+1) \leq A \cdot \ell/k$ and $k_{i,j} \geq 1$, $\ell_{i,j}/k_{i,j} \leq 2A \cdot \ell/k$. Therefore,

$$\begin{aligned}
\left| \frac{\binom{\ell_{i,j}}{2}}{\binom{\ell}{2} \cdot (k_{i,j}+1)} - \frac{\binom{\ell'_{i,j}}{2}}{\binom{\ell}{2} \cdot (k'_{i,j}+1)} \right| &= \left| \frac{\binom{\ell_{i,j}}{2}}{\binom{\ell}{2} \cdot (k_{i,j}+1)} - \frac{\binom{\ell_{i,j}}{2}}{\binom{\ell}{2} \cdot (k'_{i,j}+1)} \right| \\
&= \frac{\ell_{i,j}^2 - \ell_{i,j}}{\ell^2 - \ell} \left| \frac{1}{k_{i,j}+1} - \frac{1}{k'_{i,j}+1} \right| \\
&= \frac{\ell_{i,j}^2 - \ell_{i,j}}{\ell^2 - \ell} \left| \frac{k'_{i,j}+1-(k_{i,j}+1)}{(k_{i,j}+1)(k'_{i,j}+1)} \right| \\
&= \frac{\ell_{i,j}^2 - \ell_{i,j}}{\ell^2 - \ell} \cdot \frac{1}{k_{i,j}(k_{i,j}+1)} \\
&\leq \frac{\ell_{i,j}^2}{\ell^2 - \ell} \cdot \frac{1}{k_{i,j}^2} \\
&\leq \frac{1}{\ell^2 - \ell} \cdot \left(2A \cdot \frac{\ell}{k} \right)^2 \\
&\leq O\left(\frac{A^2}{k^2} \right)
\end{aligned} \tag{44}$$

Case 2: The extra instance of i falls in the estimation set E, in which case $k_{i,j} = k'_{i,j}$ and $\ell_{i,j} = \ell'_{i,j} + 1$. Then we have

$$\begin{aligned}
\left| \frac{\binom{\ell_{i,j}}{2}}{\binom{\ell}{2} \cdot (k_{i,j}+1)} - \frac{\binom{\ell'_{i,j}}{2}}{\binom{\ell}{2} \cdot (k'_{i,j}+1)} \right| &= \left| \frac{\binom{\ell_{i,j}}{2}}{\binom{\ell}{2} \cdot (k_{i,j}+1)} - \frac{\binom{\ell'_{i,j}}{2}}{\binom{\ell}{2} \cdot (k_{i,j}+1)} \right| \\
&= \frac{\left| (\ell'_{i,j}+1)^2 - (\ell'_{i,j}+1) - \ell'^2_{i,j} + \ell'_{i,j} \right|}{(\ell^2 - \ell)(k_{i,j}+1)} \\
&= \frac{2\ell'_{i,j}}{(\ell^2 - \ell)(k_{i,j}+1)} \\
&\leq \frac{2\ell_{i,j}}{k_{i,j}(\ell^2 - \ell)} \\
&\leq \frac{2A\ell}{k(\ell^2 - \ell)} \\
&\leq O\left(\frac{A}{\ell k} \right)
\end{aligned} \tag{45}$$

Combining the two cases proves the lemma. ■

For now, we assume Eq. (42) holds, and thus that the sensitivity obeys the bound of Eq. (43) (in Appendix C.5, we show how to transform any dataset into one that satisfies this property for $A = \Theta(\log n)$). The next task is to show that under this assumption, we can estimate $\|p''\|_2^2$ to within a constant multiplicative factor.

Lemma 34 *[Lemma 12, restated] Let E be a set of $\mathbf{Poi}(\ell)$ estimation samples from p, and $\overline{L}$ be the derandomized ℓ_2^2 norm estimator given in Eq. (38). Let $A = 12\log(n/0.05)$. For a given value of $\xi > 0$, define*

$$\tilde{L} = \overline{L} + \mathbf{Lap}\left(\frac{\Delta(\overline{L})}{\xi}\right) \tag{46}$$

Suppose $\hat{k} = \hat{k}_p + \hat{k}_q \le 100k$, that Eq. (43) holds, and that

$$\begin{aligned} k \cdot \min(k/A^2, \ell/A) &\ge C_1 \cdot \frac{k+n}{\xi} \\ \ell &\ge C_2\sqrt{k+n} \end{aligned} \tag{47}$$

for some sufficiently large constants C_1, C_2. Then

$$\mathbf{Pr}\left[\left|\tilde{L} - \|p''\|_2^2\right| > \frac{\|p''\|_2^2}{2}\right] \le 0.06$$

Proof First, we show that the magnitude of the Laplace noise is not too large. Since $\hat{k} \le 100k$, the domain size of p'' is at most $n'' \le 200k + 3n$, and $\|p''\|_2 \ge 1/\sqrt{200k+3n}$. Using the pdf of the Laplace distribution and Eq. (47), we have

$$\mathbf{Pr}\left[\left|\tilde{L} - \overline{L}\right| > \frac{\|p''\|_2^2}{4}\right] = \exp\left(\frac{-\|p''\|_2^2\xi}{4\Delta(\overline{L})}\right) \le \exp\left(\frac{-\xi}{4(200k+3n)\Delta(\overline{L})}\right) \le 1/100 \tag{48}$$

Next, we show that the derandomized statistic $\overline{L}$ concentrates around the true ℓ_2^2 norm of p''. From Goldreich and Ron (2011), we have

$$\mathbf{E}\left[\overline{L}\right] = \mathbf{E}_X\left[\mathbf{E}_r[L]\right] = \mathbf{E}_{X,r}\left[L\right] = \|p''\|_2^2 \tag{49}$$

The variance of $\overline{L}$ can also be bounded using the result of Goldreich and Ron (2011) and the law of total variance:

$$\mathbf{Var}\left[\overline{L}\right] = \mathbf{Var}[L] - \mathbf{E}[\mathbf{Var}[L \mid r]] \le \mathbf{Var}[L] \le 2\left(\mathbf{E}[L]\right)^{3/2} = \frac{1}{\binom{\ell}{2}^{1/2}} \cdot \left(\mathbf{E}\left[\overline{L}\right]\right)^{3/2} = \frac{\|p''\|_2^3}{\binom{\ell}{2}^{1/2}} \tag{50}$$

By Chebyshev's inequality, we obtain the following bound:

$$\mathbf{Pr}\left[\left|\overline{L} - \|p''\|_2^2\right| > \frac{\|p''\|_2^2}{4}\right] \le \frac{\mathbf{Var}\left[\overline{L}\right]}{\|p''\|_2^4/16} \le \frac{32}{\binom{\ell}{2}^{1/2} \cdot \|p''\|_2} \le \frac{64\sqrt{200k+3n}}{\ell} \le 0.05 \tag{51}$$

(The second to last inequality is true for all $\ell > 2$, and the last is true for $\ell \ge 1280\sqrt{200k+3n}$.)

Applying a union bound, the total error is bounded by

$$\left|\tilde{L} - \|p''\|_2^2\right| \leq \left|\tilde{L} - \overline{L}\right| + \left|\overline{L} - \|p''\|_2^2\right| \leq \frac{\|p''\|_2^2}{4} + \frac{\|p''\|_2^2}{4} = \frac{\|p''\|_2^2}{2} \tag{52}$$

with probability at least 0.94, as desired. ■

C.5. Extending the Domain of The Private Closeness Tester

From Lemma 33, it is clear that as long as Equation 42 holds, the sensitivity is low. We will use the mapping technique given in Aliakbarpour et al. (2019) to ensure that our dataset always has this property. Let $A \geq 2$. Define $\mathcal{X}$ as the set of all datasets (over domain $[n]$) and $\mathcal{X}^*$ as the subset of $\mathcal{X}$ which satisfies the property:

$$\mathcal{X}^* = \left\{ X \in \mathcal{X} : \forall i \in [n], \frac{\ell_i}{k_i + 1} \leq A \cdot \frac{\ell}{k} \right\} \tag{53}$$

where, as before, ℓ_i and k_i represent the number of occurrences of $i \in [n]$ in the testing and flattening parts of X, respectively. If $X \in \mathcal{X}^*$, we can add minimal noise to $\overline{L}$ and release a private statistic $\tilde{L}$ which is close to L with high probability. The issue arises when $X \notin \mathcal{X}^*$. In this case, we use a mapping to transform X into another dataset $Y \in \mathcal{X}$ such that the mapping from X to Y is differentially private. We do not worry about the tester's correctness in this case; the goal is merely to release a private statistic.

Lemma 35 *There exists a randomized mapping that takes $X, X' \in \mathcal{X}$ to $Y, Y' \in \mathcal{X}^*$ respectively with the following properties:*

- *If $X \in \mathcal{X}^*$, then $Y = X$*
- *If the Hamming distance between X and X' is 1, there exists a coupling $\mathcal{C}$ between the random outputs of the mapping Y and Y', where for any $(Y, Y') \sim \mathcal{C}$, the Hamming distance between Y and Y' is at most 4.*

Proof The proof follows the idea of (Aliakbarpour et al., 2019, Lemma C.5), with the required modifications to match our definition of $\mathcal{X}^*$. Given a flattening set F and an estimation set L, the goal is to replace elements of F until $\ell_i/(k_i + 1) \leq A \cdot \ell/k$ for all i. For each element i, we need $r_i(X)$ extra copies in F, where

$$r_i(X) = \max\left\{ \left\lceil \frac{k \cdot \ell_i(X)}{A \cdot \ell} \right\rceil - k_i(X) - 1, 0 \right\} \tag{54}$$

Construct a multiset R with $r_i(X)$ copies of i. To avoid violating the property, we must not remove more than $s_i(X)$ copies of each i, where

$$s_i(X) = \max\left\{ k_i(X) + 1 - \left\lceil \frac{k \cdot \ell_i(X)}{A \cdot \ell} \right\rceil, 0 \right\} \tag{55}$$

To find these slots, mark s_i instances of each i as "available" in F. We have at least $|R|$ available slots in total:

$$|R| = \sum_{i=1}^{n} r_i(X) \leq \sum_{i=1}^{n} \frac{k \cdot \ell_i(X)}{A \cdot \ell} = \frac{k}{A} \leq (A-1) \cdot \frac{k}{A} \leq k - \sum_{i=1}^{n} \frac{k_i(X)}{A} = k - \left(\sum_{i=1}^{n} \frac{k_i(X)}{A} \right)$$
$$\leq k - \left(\sum_{i=1}^{n} \left\lceil \frac{k_i(X)}{A} \right\rceil - 1 \right) \leq k - \sum_{i=1}^{n} k_i(X) - s_i(X) = \sum_{i=1}^{n} s_i(X) \tag{56}$$

We choose the first $|R|$ available slots in F and insert elements of R in those slots, randomly assigning elements to slots. The resulting dataset is in $\mathcal{X}^*$. If $|R| = 0$, then the original dataset was already in $\mathcal{X}^*$, and we change nothing.

Note that there are $|R|$ slots and $|R|$ elements to be assigned to them, so there are $|R|!$ possible assignments of elements to slots. If we choose an assignment uniformly at random, each assignment has probability $1/|R|!$ of being chosen. The existence of the coupling then follows by a result shown as part of the proof of (Aliakbarpour et al., 2019, Lemma C.5):

Lemma 36 *Let $\mathcal{M}$ be the mapping above, and let X, X' be two datasets which differ in exactly one sample. Then there exists a coupling $\mathcal{C}$ between the outputs $\mathcal{M}(X)$ and $\mathcal{M}(X')$ such that for any $(\mathcal{M}(X), \mathcal{M}(X')) \sim \mathcal{C}$, $\mathrm{Ham}(\mathcal{M}(X), \mathcal{M}(X')) \leq 4$.*

This completes the proof. ■

Given that our mapping satisfies Lemma 35, there exists a choice of A which decreases the error probability of the tester by at most δ' while preserving privacy.

Lemma 37 (Aliakbarpour et al. (2019)) *Let $\mathcal{A}$ be a $\xi/4$-differentially private algorithm over $\mathcal{X}^*$ with parameter $A \geq 12 \log(n/\delta')$ which tests the ℓ_2 norm of p and returns the correct answer with probability at least $1 - \delta$. Let $\mathcal{B}$ be a randomized mapping satisfying the conditions in Lemma 35. Then the algorithm which returns $\mathcal{A}(\mathcal{B}(X))$ is ξ-differentially private and returns the correct answer with probability at least $1 - \delta - \delta'$.*

Appendix D. Lower Bound for Private Augmented Closeness Testing

We now state our lower bound for private augmented closeness testing. This result follows from the known lower bound for augmented closeness testing and a simple reduction from private identity testing.

Theorem 38 *Let $\varepsilon, \alpha \in (0, 1]$. Let p and q both be unknown distributions over $[n]$. Then any $(\varepsilon, \alpha, \delta = 0.2)$-augmented closeness tester for p and q requires*

$$\Omega\left(\frac{n^{2/3}\alpha^{1/3}}{\varepsilon^{4/3}} + \frac{\sqrt{n}}{\varepsilon^2} + \frac{n^{1/3}}{\varepsilon^{4/3}\xi^{2/3}} + \frac{\sqrt{n}}{\varepsilon\sqrt{\xi}} + \frac{1}{\varepsilon\xi} \right)$$

samples.

Proof The first two terms in the lower bound come directly from the lower bound for the non-private version of this problem:

Lemma 39 (Aliakbarpour et al. (2024)) *Let $\varepsilon, \alpha \in (0, 1]$. Let p and q both be unknown distributions over $[n]$. Then, for every $\delta \leq 11/24$, any $(\varepsilon, \alpha, \delta)$-augmented closeness tester for p and q requires*

$$\Omega\left(\frac{n^{2/3}\alpha^{1/3}}{\varepsilon^{4/3}} + \frac{\sqrt{n}}{\varepsilon^2}\right)$$

samples.

The proof of the remaining terms is via a standard reduction from the identity testing problem to the augmented closeness testing problem. A similar statement is provided in Aliakbarpour et al. (2024) in the non-private setting. Here we include the proof for the private version for the sake of completeness.

Fix $\delta \leq 1/5$. Given a $(\xi, \varepsilon, \alpha, \delta)$-private augmented closeness tester and an instance of identity testing between unknown distribution p and known distribution q, we set $\hat{p} = q$ and run the closeness tester with samples from p and q, predicted distribution $\hat{p}$, and suggested accuracy $\alpha = \varepsilon$. If the closeness tester returns reject or $\perp$, we output reject; if the closeness tester returns accept, we output accept.

Next, we show that with high probability our output is correct. If $\mathrm{d_{TV}}(p, q) > \varepsilon$, the closeness tester will not return accept with probability greater than $1/5$. If $p = q$, since $q = \hat{p}$, the closeness tester will not return $\perp$ or reject with probability greater than $2\delta \leq 2/5$ by a union bound. Finally, the closeness tester is ξ-differentially private, so any function of its output is similarly private. Therefore, the closeness tester can be used to construct a ξ-private $(\varepsilon, 2\delta)$-identity tester, and the lower bound of Acharya et al. (2018) stated in Lemma 15 gives a lower bound for private augmented closeness testing. ■