Differentially Private Testing of Identity and Closeness of Discrete Distributions

Jayadev Acharya *
Cornell University
acharya@cornell.edu

Ziteng Sun *
Cornell University
zs335@cornell.edu

Huanyu Zhang *
Cornell University
hz388@cornell.edu

Abstract

We study the fundamental problems of identity testing (goodness of fit), and closeness testing (two sample test) of distributions over k elements, under differential privacy. While the problems have a long history in statistics, finite sample bounds for these problems have only been established recently.

In this work, we derive upper and lower bounds on the sample complexity of both the problems under (ε, δ)-differential privacy. We provide sample optimal algorithms for identity testing problem for all parameter ranges, and the first results for closeness testing. Our closeness testing bounds are optimal in the sparse regime where the number of samples is at most k.

Our upper bounds are obtained by privatizing non-private estimators for these problems. The non-private estimators are chosen to have small sensitivity. We propose a general framework to establish lower bounds on the sample complexity of statistical tasks under differential privacy. We show a bound on differentially private algorithms in terms of a coupling between the two hypothesis classes we aim to test. By carefully constructing chosen priors over the hypothesis classes, and using Le Cam's two point theorem we provide a general mechanism for proving lower bounds. We believe that the framework can be used to obtain strong lower bounds for other statistical tasks under privacy.

1 Introduction

Testing whether observed data conforms to an underlying model is a fundamental scientific problem. In a statistical framework, given samples from an unknown probabilistic model, the goal is to determine whether the underlying model has a property of interest.

This question has received great attention in statistics as hypothesis testing [1, 2], where it was mostly studied in the asymptotic regime when the number of samples $m \to \infty$. In the past two decades there has been a lot of work from the computer science, information theory, and statistics community on various distribution testing problems in the non-asymptotic (small-sample) regime, where the domain size k could be potentially larger than m (See [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15], references therein, and [16] for a recent survey). Here the goal is to characterize the minimum number of samples necessary (sample complexity) as a function of the domain size k, and the other parameters.

At the same time, preserving the privacy of individuals who contribute to the data samples has emerged as one of the key challenges in designing statistical mechanisms over the last few years. For example, the privacy of individuals participating in surveys on sensitive subjects

*The authors are listed in alphabetical order. This research was supported by NSF-CCF-CRII 1657471, and a grant from Cornell University.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

is of utmost importance. Without a properly designed mechanism, statistical processing might divulge the sensitive information about the data. There have been many publicized instances of individual data being de-anonymized, including the deanonymization of Netflix database [17], and individual information from census-related data [18]. Protecting privacy for the purposes of data release, or even computation on data has been studied extensively across several fields, including statistics, machine learning, database theory, algorithm design, and cryptography (See e.g., [19, 20, 21, 22, 23, 24, 25]). While the motivation is clear, even a formal notion of privacy is not straight forward. We use *differential privacy* [26], a notion which rose from database and cryptography literature, and has emerged as one of the most popular privacy measures (See [26, 27, 22, 28, 29, 30, 31, 32], references therein, and the recent book [33]). Roughly speaking, it requires that the output of the algorithm should be statistically close on two neighboring datasets. For a formal definition of differential privacy, see Section 2.

A natural question when designing a differentially private algorithm is to understand how the data requirement grows to ensure privacy, along with the same accuracy. In this paper, we study the sample size requirements for differentially private discrete distribution testing.

1.1 Results and Techniques

We consider two fundamental statistical tasks for testing distributions over $[k]$: (i) identity testing, where given sample access to an unknown distribution p, and a known distribution q, the goal is to decide whether $p = q$, or $d_{TV}(p, q) \geq \alpha$, and (ii) closeness testing, where given sample access to unknown distributions p, and q, the goal is to decide whether $p = q$, or $d_{TV}(p, q) \geq \alpha$. (See Section 2 for precise statements of these problems). Given differential privacy constraints (ε, δ), we provide (ε, δ)-differentially private algorithms for both these tasks. For identity testing, our bounds are optimal up to constant factors for all ranges of $k, \alpha, \varepsilon, \delta$, and for closeness testing the results are tight in the small sample regime where $m = O(k)$. Our upper bounds are based on various methods to privatize the previously known tests. A critical component is to design and analyze test statistic that have low sensitivity (see Definition 4), in order to preserve privacy.

We first state that any $(\varepsilon + \delta, 0)$-DP algorithm is also an (ε, δ) algorithm. [34] showed that for testing problems, any (ε, δ) algorithm will also imply a $(\varepsilon + c\delta, 0)$-DP algorithm. Please refer to Lemma 2 and Lemma 3 for more detail. Therefore, for all the problems, we simply consider $(\varepsilon, 0)$-DP algorithms (ε-DP), and we can replace ε with $(\varepsilon + \delta)$ in both the upper and lower bounds without loss of generality.

One of the main contributions of our work is to propose a general framework for establishing lower bounds for the sample complexity of statistical problems such as property estimation and hypothesis testing under privacy constraints. We describe this, and the other results below. A summary of the results is presented in Table 1, which we now describe in detail.

1. **DP Lower Bounds via Coupling.** We establish a general method to prove lower bounds for distribution testing problems. Suppose X_1^m, and Y_1^m are generated by two statistical sources. Further suppose there is a coupling between the two sources such that the expected hamming distance between the coupled samples is at most D, then if $\varepsilon + \delta = o(1/D)$, there is no (ε, δ)-differentially private algorithm to distinguish between the two sources. This result is stated precisely in Theorem 1. By carefully using designed coupling schemes, we provide lower bounds for identity testing, and closeness testing.
2. **Reduction from identity to uniformity.** We reduce the problem of ε-DP identity testing of distributions over $[k]$ to ε-DP uniformity testing over distributions over $[6k]$. Such a reduction, without privacy constraints was shown in [35], and we use their result to obtain a reduction that also preserves privacy, with at most a constant factor blow-up in the sample complexity. This result is given in Theorem 3.
3. **Identity Testing.** It was recently shown that $O(\frac{\sqrt{k}}{\alpha^2})$ [7, 36, 11, 37] samples are necessary and sufficient for identity testing without privacy constraints. The statistic used in these papers are variants of chi-squared tests, which could have a high global sensitivity.
   Given the reduction from identity to uniformity, it suffices to consider uniformity testing. We consider the test statistic studied by [38] which is simply the distance of the empirical distribution to the uniform distribution. This statistic also has a low sensitivity, and

futhermore has the optimal sample complexity in all parameter ranges, without privacy constraints. In Theorem 2, we state the optimal sample complexity of identity testing. The upper bounds are derived by privatizing the statistic in [38]. For lower bound, we use our technique in Theorem 1. We design a coupling between the uniform distribution $u[k]$, and a mixture of distributions, which are all at distance α from $u[k]$ in total variation distance. In particular, we consider the mixture distribution used in [7]. Much of the technical details go into proving the existence of couplings with small expected Hamming distance. [34] studied identity testing under pure differential privacy, and obtained an algorithm with complexity $O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k \log k}}{\alpha^{3/2}\varepsilon} + \frac{(k \log k)^{1/3}}{\alpha^{5/3}\varepsilon^{2/3}}\right)$. Our results improve their bounds significantly.

4. **Closeness Testing.** Closeness testing problem was proposed by [3], and optimal bound of $\Theta\left(\max\{\frac{k^{2/3}}{\alpha^{4/3}}, \frac{\sqrt{k}}{\alpha^2}\}\right)$ was shown in [10]. They proposed a chi-square based statistic, which we show has a small sensitivity. We privatize their algorithm to obtain the sample complexity bounds. In the sparse regime we prove a sample complexity bound of $\Theta\left(\frac{k^{2/3}}{\alpha^{4/3}} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}}\right)$, and in the dense regime, we obtain a bound of $O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha^2\varepsilon}\right)$. These results are stated in Theorem 4. Since closeness testing is a harder problem than identity testing, all the lower bounds from identity testing port over to closeness testing. The closeness testing lower bounds are given in Theorem 4.

Problem	Sample Complexity Bounds
Identity Testing	**Non-private :** $\Theta\left(\frac{\sqrt{k}}{\alpha^2}\right)$ **[7]**
	ε**-DP algorithms:** $O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k \log k}}{\alpha^{3/2}\varepsilon}\right)$ [34]
	$S(\mathtt{IT}, k, \alpha, \varepsilon) = \Theta\left(\frac{\sqrt{k}}{\alpha^2} + \max\left\{\frac{k^{1/2}}{\alpha\varepsilon^{1/2}}, \frac{k^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}}, \frac{1}{\alpha\varepsilon}\right\}\right)$[Theorem 2]
Closeness Testing	**Non-private:** $\Theta\left(\frac{k^{2/3}}{\alpha^{4/3}} + \frac{k^{1/2}}{\alpha^2}\right)$ [10]
	ε**-DP algorithms:**
	IF $\alpha^2 = \Omega\left(\frac{1}{\sqrt{k}}\right)$ **and** $\alpha^2\varepsilon = \Omega\left(\frac{1}{k}\right)$
	$S(\mathtt{CT}, k, \alpha, \varepsilon) = \Theta\left(\frac{k^{2/3}}{\alpha^{4/3}} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}}\right)$
	ELSE
	$\Omega\left(\frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}} + \frac{1}{\alpha\varepsilon}\right) \le S(\mathtt{CT}, k, \alpha, \varepsilon) \le O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha^2\varepsilon}\right)$ [Theorem 4]

Table 1: Summary of the sample complexity bounds for ε-DP identity, and closeness testing. For (ε, δ)-DP algorithms, we can simply replace ε in the sample complexity by $(\varepsilon + \delta)$.

1.2 Related Work

A number of papers have recently studied hypothesis testing problems under differential privacy guarantees [39, 40, 41]. Some works analyze the distribution of the test statistic in the asymptotic regime. The work most closely related to ours is [34], which studied identity testing in the finite sample regime. We mentioned their guarantees along with our results on identity testing in the previous section.

There has been a line of research for statistical testing and estimation problems under the notion of *local* differential privacy [24, 23, 42, 43, 44, 45, 46, 47, 48, 49]. These papers study some basic statistical problems and provide minimax lower bounds using Fano's inequality. [50] studies structured distribution estimation under differential privacy. Information theoretic approaches to data privacy have been studied recently using quantities like mutual information, and guessing probability to quantify privacy [51, 52, 53, 54, 55].

[56, 57] provide methods to prove lower bounds on DP algorithms via packing. Recently, [58] use coupling to prove lower bounds on the sample complexity for differentially private confidence intervals. Our results are more general, in that, we can handle mixtures of distributions, which can provide optimal lower bounds on identity testing. [59, 60] characterize

differential privacy through a coupling argument. [61] also uses the idea of coupling implicitly when designing differentially private partition algorithms. [62] uses our coupling argument to prove lower bounds for differentially private property estimation problems.

In a contemporaneous and independent work, [63], the authors study the same problems that we consider, and obtain the same upper bounds for the sparse case, when $m \leq k$. They also provide experimental results to show the performance of the privatized algorithms. However, their results are sub-optimal for $m = \Omega(k)$ for identity testing, and they do not provide any lower bounds for the problems. Both [34], and [63] consider only pure-differential privacy, which are a special case of our results.

Organization of the paper. In Section 2, we discuss the definitions and notations. A general technique for proving lower bounds for differentially private algorithms is described in Section 3. Section 4 gives upper and lower bounds for identity testing, and closeness testing is studied in Section 5.

2 Preliminaries

Let Δ_k be the class of all discrete distributions over a domain of size k, which wlog is assumed to be $[k] := \{1, \ldots, k\}$. We denote length-m samples $X_1, \ldots, X_m$ by X_1^m. For $x \in [k]$, let p_x be the probability of x under p. Let $M_x(X_1^m)$ be the number of times x appears in X_1^m. For $A \subseteq [k]$, let $p(A) = \sum_{x \in A} p_x$. Let $X \sim p$ denote that the random variable X has distribution p. Let $u[k]$ be the uniform distribution over $[k]$, and $B(b)$ be the Bernoulli distribution with bias b. The *total variation* distance between distributions p, and q over $[k]$ is $d_{TV}(p, q) := \sup_{A \subset [k]} \{p(A) - q(A)\} = \frac{1}{2}\|p - q\|_1$.

Definition 1. Let p, and q be distributions over $\mathcal{X}$, and $\mathcal{Y}$ respectively. A *coupling* between p and q is a distribution over $\mathcal{X} \times \mathcal{Y}$ whose marginals are p and q respectively.

Definition 2. The *Hamming distance* between two sequences X_1^m and Y_1^m is $d_H(X_1^m, Y_1^m) := \sum_{i=1}^m \mathbb{I}\{X_i \neq Y_i\}$, the number of positions where X_1^m, and Y_1^m differ.

Definition 3. A randomized algorithm $\mathcal{A}$ on a set $\mathcal{X}^m \to \mathcal{S}$ is said to be (ε, δ)-differentially private if for any $S \subset \text{range}(\mathcal{A})$, and all pairs of X_1^m, and Y_1^m with $d_H(X_1^m, Y_1^m) \leq 1$ such that $\Pr(\mathcal{A}(X_1^m) \in S) \leq e^{\varepsilon} \cdot \Pr(\mathcal{A}(Y_1^m) \in S) + \delta$.

The case when $\delta = 0$ is called *pure differential privacy*. For simplicity, we denote pure differential privacy as ε-differential privacy (ε-DP).

Next we state the group property of differential privacy. We give a proof in Appendix A.1.

Lemma 1. *Let $\mathcal{A}$ be a (ε, δ)-DP algorithm, then for sequences x_1^m, and y_1^m with $d_H(x_1^m, y_1^m) \leq t$, and $\forall S \subset \text{range}(\mathcal{A})$, $\Pr(\mathcal{A}(x_1^m) \in S) \leq e^{t\varepsilon} \cdot \Pr(\mathcal{A}(y_1^m) \in S) + \delta t e^{\varepsilon(t-1)}$.*

The next two lemmas state a relationship between (ε, δ) and ε-differential privacy. We give a proof of Lemma 2 in Appendix A.2. And Lemma 3 follows from [34].

Lemma 2. *Any $(\varepsilon + \delta, 0)$- differentially private algorithm is also (ε, δ)-differentially private.*

Lemma 3. *An (ε, δ)-DP algorithm for a testing problem can be converted to an $(\varepsilon + c\delta, 0)$ algorithm for some constant $c > 0$.*

Combining these two results, it suffices to prove bounds for $(\varepsilon, 0)$-DP, and plug in ε with $(\varepsilon + \delta)$ to obtain bounds that are tight up to constant factors for (ε, δ)-DP.

The notion of sensitivity is useful in establishing bounds under differential privacy.

Definition 4. The *sensitivity* of $f : [k]^m \to \mathbb{R}$ is

$$\Delta(f) := \max_{d_H(X_1^m, Y_1^m) \leq 1} |f(X_1^m) - f(Y_1^m)|.$$

For $x \in \mathbb{R}$, $\sigma(x) := \frac{1}{1+\exp(-x)} = \frac{\exp(x)}{1+\exp(x)}$ is the sigmoid function. The following properties follow from the definition of σ.

Lemma 4. *1. For all $x, \gamma \in \mathbb{R}$, $\exp(-|\gamma|) \leq \frac{\sigma(x+\gamma)}{\sigma(x)} \leq \exp(|\gamma|)$.*
2. Let $0 < \eta < \frac{1}{2}$. Suppose $x \geq \log \frac{1}{\eta}$. Then $\sigma(x) > 1 - \eta$.

Identity Testing (IT). Given description of $q \in \Delta_k$ over $[k]$, parameters α, and m independent samples X_1^m from unknown $p \in \Delta_k$. $\mathcal{A}$ is an (k, α)-identity testing algorithm for q, if when $p = q$, $\mathcal{A}$ outputs "$p = q$" with probability at least 0.9, and when $d_{TV}(p, q) \geq \alpha$, $\mathcal{A}$ outputs "$p \neq q$" with probability at least 0.9.

Definition 5. The sample complexity of DP-identity testing, denoted $S(\mathtt{IT}, k, \alpha, \varepsilon)$, is the smallest m for which there exists an ε-DP algorithm $\mathcal{A}$ that uses m samples to achieve (k, α)-identity testing. Without privacy concerns, $S(\mathtt{IT}, k, \alpha)$ denotes the sample complexity. When $q = u[k]$, the problem reduces to uniformity testing, and the sample complexity is denoted as $S(\mathtt{UT}, k, \alpha, \varepsilon)$.

Closeness Testing (CT). Given m independent samples X_1^m, and Y_1^m from unknown distributions p, and q. An algorithm $\mathcal{A}$ is an (k, α)-closeness testing algorithm if when $p = q$, $\mathcal{A}$ outputs $p = q$ with probability at least 0.9, and when $d_{TV}(p, q) \geq \alpha$, $\mathcal{A}$ outputs $p \neq q$ with probability at least 0.9.

Definition 6. The sample complexity of DP-closeness testing, denoted $S(\mathtt{CT}, k, \alpha, \varepsilon)$, is the smallest m for which there exists an ε-DP algorithm $\mathcal{A}$ that uses m samples to achieve (k, α)-closeness testing. When privacy is not a concern, we denote the sample complexity of closeness testing as $S(\mathtt{CT}, k, \alpha)$.

Hypothesis Testing (HT). Suppose we have distributions p and q over $\mathcal{X}^m$, and $X_1^m \sim p, Y_1^m \sim q$, we say an algorithm $\mathcal{A} : \mathcal{X}^m \to \{p, q\}$ can distinguish between p and q if $\Pr(\mathcal{A}(X_1^m) = q) < 0.1$ and $\Pr(\mathcal{A}(Y_1^m) = p) < 0.1$.

3 Privacy Bounds Via Coupling

Recall that *coupling* between distributions p and q over $\mathcal{X}$, and $\mathcal{Y}$, is a distribution over $\mathcal{X} \times \mathcal{Y}$ whose marginal distributions are p and q (Definition 1). For simplicity, we treat coupling as a randomized function $f : \mathcal{X} \to \mathcal{Y}$ such that if $X \sim p$, then $Y = f(X) \sim q$. Note that X, and Y are not necessarily independent.

Example 1. Let $B(b_1)$, and $B(b_2)$ be Bernoulli distributions with bias b_1, and b_2 such that $b_1 < b_2$. Let p, and q be distributions over $\{0, 1\}^m$ obtained by m *i.i.d.* samples from $B(b_1)$, and $B(b_2)$ respectively. Let X_1^m be distributed according to p. Generate a sequence Y_1^m as follows: If $X_i = 1$, then $Y_i = 1$. If $X_i = 0$, we flip another coin with bias $(b_2 - b_1)/(1 - b_1)$, and let Y_i be the output of this coin. Repeat the process independently for each i, such that the Y_i's are all independent of each other. Then $\Pr(Y_i = 1) = b_1 + (1 - b_1)(b_2 - b_1)/(1 - b_1) = b_2$, and Y_1^m is distributed according to q.

We would like to use coupling to prove lower bounds on differentially private algorithms for testing problems. Let p and q be distributions over $\mathcal{X}^m$. If there is a coupling between p and q with a small *expected* Hamming distance, we might expect that the algorithm cannot have strong privacy guarantees. The following theorem formalizes this intuition:

Theorem 1. *Suppose there is a coupling between p and q over $\mathcal{X}^m$, such that $\mathbb{E}[d_H(X_1^m, Y_1^m)] \leq D$ where $X_1^m \sim p, Y_1^m \sim q$. Then, any (ε, δ)-differentially private hypothesis testing algorithm $\mathcal{A} : \mathcal{X}^m \to \{p, q\}$ on p and q must satisfy $\varepsilon + \delta = \Omega\left(\frac{1}{D}\right)$*

Proof. Let (X_1^m, Y_1^m) be distributed according to a coupling of p, and q with $\mathbb{E}[d_H(X_1^m, Y_1^m)] \leq D$. By Markov's inequality, $\Pr(d_H(X_1^m, Y_1^m) > 10D) < \Pr(d_H(X_1^m, Y_1^m) > 10 \cdot \mathbb{E}[d_H(X_1^m, Y_1^m)]) < 0.1$. Let x_1^m and y_1^m be the realization of X_1^m and Y_1^m. Let $W = \{(x_1^m, y_1^m) | d_H(x_1^m, y_1^m) \leq 10D\}$. Then we have

$$0.1 \geq \Pr(\mathcal{A}(X_1^m) = q) \geq \sum_{(x_1^m, y_1^m) \in W} \Pr(X_1^m = x_1^m, Y_1^m = y_1^m) \cdot \Pr(\mathcal{A}(x_1^m) = q).$$

By Lemma 1, and $\Pr(d_H(X_1^m, Y_1^m) > 10D) < 0.1$, and $\Pr(\mathcal{A}(y_1^m) = q) \leq 1$,

$$\begin{aligned}\Pr\left(\mathcal{A}(Y_1^m)=q\right) &\le \sum_{(x_1^m,y_1^m)\in W}\Pr\left(x_1^m,y_1^m\right)\cdot\Pr\left(\mathcal{A}(y_1^m)=q\right)+\sum_{(x_1^m,y_1^m)\notin W}\Pr\left(x_1^m,y_1^m\right)\cdot 1\\ &\le \sum_{(x_1^m,y_1^m)\in W}\Pr\left(x_1^m,y_1^m\right)\cdot\left(e^{\varepsilon\cdot 10D}\Pr\left(\mathcal{A}(x_1^m)=q\right)+10D\delta\cdot e^{\varepsilon\cdot 10(D-1)}\right)+0.1\\ &\le 0.1e^{\varepsilon\cdot 10D}+10D\delta\cdot e^{\varepsilon\cdot 10D}+0.1.\end{aligned}$$

Since we know $\Pr\left(\mathcal{A}(Y_1^m)=q\right)>0.9$, then $0.9<\Pr\left(\mathcal{A}(Y_1^m)=q\right)<0.1e^{\varepsilon\cdot 10D}+10D\delta\cdot e^{\varepsilon\cdot 10D}+0.1$. Hence, either $e^{\varepsilon\cdot 10D}=\Omega(1)$ or $10D\delta=\Omega(1)$, which implies that $D=\Omega\big(\min\left\{\frac{1}{\varepsilon},\frac{1}{\delta}\right\}\big)=\Omega\Big(\frac{1}{\varepsilon+\delta}\Big)$, proving the theorem. □

Set $\delta=0$, we obtain the bound for pure differential privacy. In the next few sections, we use this theorem to get sample complexity bounds for differentially private testing problems.

4 Identity Testing

In this section, we prove the bounds for identity testing. Our main result is the following.

Theorem 2.

$$S(\mathtt{IT},k,\alpha,\varepsilon)=\Theta\Big(\frac{k^{1/2}}{\alpha^2}+\max\Big\{\frac{k^{1/2}}{\alpha\varepsilon^{1/2}},\frac{k^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}},\frac{1}{\alpha\varepsilon}\Big\}\Big).$$

Or we can write it according to the parameter range,

$$S(\mathtt{IT},k,\alpha,\varepsilon)=\begin{cases}\Theta\Big(\frac{\sqrt{k}}{\alpha^2}+\frac{k^{1/2}}{\alpha\varepsilon^{1/2}}\Big), & \text{when } k=\Omega\big(\frac{1}{\alpha^4}\big) \text{ and } k=\Omega\big(\frac{1}{\alpha^2\varepsilon}\big),\\ \Theta\Big(\frac{\sqrt{k}}{\alpha^2}+\frac{k^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}}\Big), & \text{when } k=\Omega\big(\frac{\alpha}{\varepsilon}\big) \text{ and } k=O\big(\frac{1}{\alpha^4}+\frac{1}{\alpha^2\varepsilon}\big),\\ \Theta\Big(\frac{\sqrt{k}}{\alpha^2}+\frac{1}{\alpha\varepsilon}\Big), & \text{when } k=O\big(\frac{\alpha}{\varepsilon}\big).\end{cases}$$

Our bounds are tight up to constant factors in all parameters. To get the sample complexity for (ε,δ)-differential privacy, we can simply replace ε by $(\varepsilon+\delta)$.

In Theorem 3 we will show a reduction from identity to uniformity testing under pure differential privacy. Using this, it will be enough to design algorithms for uniformity testing, which is done in Section 4.2.

Moreover since uniformity testing is a special case of identity testing, any lower bound for uniformity will port over to identity, and we give such bounds in Section 4.3.

4.1 Uniformity Testing implies Identity Testing

The sample complexity of testing identity of any distribution is $O(\frac{\sqrt{k}}{\alpha^2})$, a bound that is tight for the uniform distribution. Recently [35] proposed a scheme to reduce the problem of testing identity of distributions over $[k]$ for total variation distance α to the problem of testing uniformity over $[6k]$ with total variation parameter $\alpha/3$. In other words, they show that $S(\mathtt{IT},k,\alpha)\le S(\mathtt{UT},6k,\alpha/3)$. Building on [35], we prove that a similar bound also holds for differentially private algorithms. The proof is in Appendix B.

Theorem 3. $S(\mathtt{IT},k,\alpha,\varepsilon)\le S(\mathtt{UT},6k,\alpha/3,\varepsilon)$.

4.2 Identity Testing – Upper Bounds

In this section, we will show that by privatizing the statistic proposed in [38] we can achieve the sample complexity in Theorem 2 for all parameter ranges. The procedure is described in Algorithm 1.

Recall that $M_x(X_1^m)$ is the number of appearances of x in X_1^m. Let

$$S(X_1^m) := \frac{1}{2} \cdot \sum_{x=1}^{n} \left| \frac{M_x(X_1^m)}{m} - \frac{1}{k} \right|, \tag{1}$$

be the TV distance from the empirical distribution to the uniform distribution. Let $\mu(p) = \mathbb{E}[S(X_1^m)]$ when the samples are drawn from distribution p. They show the following separation result on the expected value of $S(X_1^m)$.

Lemma 5 ([38]). *Let p be a distribution over $[k]$ and $d_{TV}(p, u[k]) \geq \alpha$, then there is a constant c such that*

$$\mu(p) - \mu(u[k]) \geq c\alpha^2 \min\left\{ \tfrac{m^2}{k^2}, \sqrt{\tfrac{m}{k}}, \tfrac{1}{\alpha} \right\}.$$

[38] used this result to show that thresholding $S(X_1^m)$ at 0 is an optimal algorithm for identity testing. We first normalize the statistic to simplify the presentation of our DP algorithm. Let

$$Z(X_1^m) := \begin{cases} k\Big(S(X_1^m) - \mu(u[k]) - \frac{1}{2}c\alpha^2 \cdot \frac{m^2}{k^2}\Big), & \text{when } m \leq k, \\ m\Big(S(X_1^m) - \mu(u[k]) - \frac{1}{2}c\alpha^2 \cdot \sqrt{\frac{m}{k}}\Big), & \text{when } k < m \leq \frac{k}{\alpha^2}, \\ m\Big(S(X_1^m) - \mu(u[k]) - \frac{1}{2}c\alpha\Big), & \text{when } m \geq \frac{k}{\alpha^2}. \end{cases} \tag{2}$$

where c is the constant in Lemma 5, and $\mu(u[k])$ is the expected value of $S(X_1^m)$ when X_1^m are drawn from uniform distribution.

Algorithm 1 Uniformity testing

Input: ε, α, i.i.d. samples X_1^m from p

1: Let $Z(X_1^m)$ be evaluated from (1), and (2).
2: Generate $Y \sim B(\sigma(\varepsilon \cdot Z))$, σ is the sigmoid function.
3: **if** $Y = 0$, **return** $p = u[k]$, **else**, **return** $p \neq u[k]$.

We now prove that this algorithm is ε-DP. We need the following sensitivity result.

Lemma 6. $\Delta(Z) \leq 1$ *for all values of m, and k.*

Proof. Recall that $S(X_1^m) = \frac{1}{2} \cdot \sum_{x=1}^{n} \left| \frac{M_x(X_1^m)}{m} - \frac{1}{k} \right|$. Changing any one symbol changes at most two of the $M_x(X_1^m)$'s. Therefore at most two of the terms change by at most $\frac{1}{m}$. Therefore, $\Delta(S(X_1^m)) \leq \frac{1}{m}$, for any m. When $m \leq k$, this can be strengthened with observation that $M_x(X_1^m)/m \geq \frac{1}{k}$, for all $M_x(X_1^m) \geq 1$. Therefore, $S(X_1^m) = \frac{1}{2} \cdot \left(\sum_{x:M_x(X_1^m)\geq 1} \left(\frac{M_x(X_1^m)}{m} - \frac{1}{k} \right) + \sum_{x:M_x(X_1^m)=0} \frac{1}{k} \right) = \frac{\Phi_0(X_1^m)}{k}$, where $\Phi_0(X_1^m)$ is the number of symbols not appearing in X_1^m. This changes by at most one when one symbol is changed, proving the result. □

Using this lemma, $\varepsilon \cdot Z(X_1^m)$ changes by at most ε when X_1^m is changed at one location. Invoking Lemma 4, the probability of any output changes by a multiplicative $\exp(\varepsilon)$, and the algorithm is ε-differentially private.

To prove the sample complexity bound, we first show that the mean of the test statistic is well separated using Lemma 5. Then we use the concentration bound of the test statistic from [38] to get the final complexity. Due to lack of space, the detailed proof of sample complexity bound is given in Appendix C.

4.3 Sample Complexity Lower bounds for Uniformity Testing

In this section, we will show the lower bound part of Theorem 2. The first term is the lower bound without privacy constraints, proved in [7]. In this section, we will prove the terms associated with privacy.

The simplest argument is for $m \geq \frac{k}{\alpha^2}$, which hopefully will give you a sense of how coupling argument works. We consider the case of binary identity testing where the goal is to test whether the bias of a coin is $1/2$ or α-far from $1/2$. This is a special case of identity testing for distributions over $[k]$ (when $k-2$ symbols have probability zero). This is strictly harder than the problem of distinguishing between $B(1/2)$ and $B(1/2+\alpha)$. The coupling given in Example 1 has expected hamming distance of αm. Hence combing with Theorem 1, we get a lower bound of $\Omega(\frac{1}{\alpha\varepsilon})$.

We now consider the cases $m \leq k$ and $k < m \leq \frac{k}{\alpha^2}$.

To this end, we invoke LeCam's two point theorem, and design a hypothesis testing problem that will imply a lower bound on uniformity testing. The testing problem will be to distinguish between the following two cases.

Case 1: We are given m independent samples from the uniform distribution $u[k]$.

Case 2: Generate a distribution p with $d_{TV}(p, u[k]) \geq \alpha$ according to some prior over all such distributions. We are then given m independent samples from this distribution p.

Le Cam's two point theorem [64] states that any lower bound for distinguishing between these two cases is a lower bound on identity testing problem.

We now describe the prior construction for **Case 2**, which is the same as considered by [7] for lower bounds on identity testing without privacy considerations. For each $\mathbf{z} \in \{\pm 1\}^{k/2}$, define a distribution $p_{\mathbf{z}}$ over $[k]$ such that

$$p_{\mathbf{z}}(2i-1) = \frac{1 + \mathbf{z}_i \cdot 2\alpha}{k}, \text{ and } p_{\mathbf{z}}(2i) = \frac{1 - \mathbf{z}_i \cdot 2\alpha}{k}.$$

Then for any $\mathbf{z}$, $d_{TV}(P_{\mathbf{z}}, u[k]) = \alpha$. For **Case 2**, choose p uniformly from these $2^{k/2}$ distributions. Let Q_2 denote the distribution on $[k]^m$ by this process. In other words, Q_2 is a mixture of product distributions over $[k]$.

In **Case 1**, let Q_1 be the distribution of m *i.i.d.* samples from $u[k]$.

To obtain a sample complexity lower bound for distinguishing the two cases, we will design a coupling between Q_1, and Q_2, and bound its expected Hamming distance. While it can be shown that the Hamming distance of the coupling between the uniform distribution with any *one* of the $2^{k/2}$ distributions grows as αm, it can be significantly smaller, when we consider the mixtures. In particular, the following lemma shows that there exist couplings with bounded Hamming distance.

Lemma 7. *There is a coupling between X_1^m generated by Q_1, and Y_1^m by Q_2 such that*

$$\mathbb{E}\left[d_H(X_1^m, Y_1^m)\right] \leq C \cdot \alpha^2 \min\{\tfrac{m^2}{k}, \tfrac{m^{3/2}}{k^{1/2}}\}.$$

The lemma is proved in Appendix D. Now applying Theorem 1, we get the bound in Theorem 2.

5 Closeness Testing

Recall the closeness testing problem from Section 2, and the tight non-private bounds from Table 1. Our main result in this section is the following theorem characterizing the sample complexity of differentially private algorithms for closeness testing.

Theorem 4. *If $\alpha > 1/k^{1/4}$, and $\varepsilon\alpha^2 > 1/k$,*

$$S(\mathtt{CT}, k, \alpha, \varepsilon) = \Theta\left(\frac{k^{2/3}}{\alpha^{4/3}} + \frac{k^{1/2}}{\alpha\sqrt{\varepsilon}}\right),$$

otherwise,

$$\Omega\left(\frac{k^{1/2}}{\alpha^2} + \frac{k^{1/2}}{\alpha\sqrt{\varepsilon}} + \frac{1}{\alpha\varepsilon}\right) \leq S(\mathtt{CT}, k, \alpha, \varepsilon) \leq O\left(\frac{k^{1/2}}{\alpha^2} + \frac{1}{\alpha^2\varepsilon}\right).$$

This theorem shows that in the sparse regime, when $m = O(k)$, our bounds are tight up to constant factors in all parameters. To prove the upper bounds, we only consider the case when $\delta = 0$, which would suffice by lemma 2. We privatize the closeness testing algorithm of [10]. To reduce the strain on the readers, we drop the sequence notations explicitly and let

$$\mu_i := M_i(X_1^m), \text{ and } \nu_i := M_i(Y_1^m).$$

The statistic used by [10] is

$$Z(X_1^m, Y_1^m) := \sum_{i \in [k]} \frac{(\mu_i - \nu_i)^2 - \mu_i - \nu_i}{\mu_i + \nu_i},$$

where we assume that $((\mu_i - \nu_i)^2 - \mu_i - \nu_i)/(\mu_i + \nu_i) = 0$, when $\mu_i + \nu_i = 0$. It turns out that this statistic has a constant sensitivity, as shown in Lemma 8.

Lemma 8. $\Delta(Z(X_1^m, Y_1^m)) \leq 14$.

Proof. Since $Z(X_1^m, Y_1^m)$ is symmetric, without loss of generality assume that one of the symbols is changed in Y_1^m. This would cause at most two of the ν_i's to change. Suppose $\nu_i \geq 1$, and it changed to $\nu_i - 1$. Suppose, $\mu_i + \nu_i > 1$, the absolute change in the ith term of the statistic is

$$\begin{aligned}
\left| \frac{(\mu_i - \nu_i)^2}{\mu_i + \nu_i} - \frac{(\mu_i - \nu_i + 1)^2}{\mu_i + \nu_i - 1} \right| &= \left| \frac{(\mu_i + \nu_i)(2\mu_i - 2\nu_i + 1) + (\mu_i - \nu_i)^2}{(\mu_i + \nu_i)(\mu_i + \nu_i - 1)} \right| \\
&\leq \left| \frac{2\mu_i - 2\nu_i + 1}{\mu_i + \nu_i - 1} \right| + \left| \frac{\mu_i - \nu_i}{\mu_i + \nu_i - 1} \right| \\
&\leq \frac{3\,|\mu_i - \nu_i| + 1}{\mu_i + \nu_i - 1} \leq 3 + \frac{4}{\mu_i + \nu_i - 1} \leq 7.
\end{aligned}$$

When $\mu_i + \nu_i = 1$, the change can again be bounded by 7. Since at most two of the ν_i's change, we obtain the desired bound. □

We use the same approach with the test statistic as with uniformity testing to obtain a differentially private closeness testing method, described in Algorithm 2. Since the sensitivity of the statistic is at most 14, the input to the sigmoid changes by at most ε when any input sample is changed. Invoking Lemma 4, the probability of any output changes by a multiplicative $\exp(\varepsilon)$, and the algorithm is ε-differentially private.

Algorithm 2

Input: ε, α, sample access to distribution p and q

1: $Z' \leftarrow (Z(X_1^m, Y_1^m) - \frac{1}{2}\frac{m^2\alpha^2}{4k+2m})/14$
2: Generate $Y \sim B(\sigma(\exp(\varepsilon \cdot Z'))$
3: **if** $Y = 0$, **return** $p = q$
4: **else**, **return** $p \neq q$

The remaining part is to show that Algorithm 2 satisfies sample complexity upper bounds described in theorem 4. We will give the details in Appendix E, where the analysis of the lower bound is also given.

Acknowledgement

The authors thank Gautam Kamath for some very helpful suggestions about this work.

References

[1] Jerzy Neyman and Egon Sharpe Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

[2] Erich Leo Lehmann and George Casella. *Theory of Point Estimation*, volume 31. Springer, 2006.

[3] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '00, pages 259–269, Washington, DC, USA, 2000. IEEE Computer Society.

[4] Tuğkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, FOCS '01, pages 442–451, Washington, DC, USA, 2001. IEEE Computer Society.

[5] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, pages 68–75. Springer, 2011.

[6] Tugkan Batu. *Testing properties of distributions*. PhD thesis, Cornell University, 2001.

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

[8] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. A competitive test for uniformity of monotone distributions. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, 2013.

[9] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitksy, and Ananda Theertha Suresh. Sublinear algorithms for outlier detection and generalized closeness testing. In *Proceedings of the 2014 IEEE International Symposium on Information Theory*, 2014.

[10] Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 1193–1203, Philadelphia, PA, USA, 2014. SIAM.

[11] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 1841–1854, Philadelphia, PA, USA, 2015. SIAM.

[12] Bhaswar Bhattacharya and Gregory Valiant. Testing closeness with unequal sized samples. In *Advances in Neural Information Processing Systems*, NIPS '15, pages 2611–2619. Curran Associates, Inc., 2015.

[13] Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. In *Proceedings of the 33rd Symposium on Theoretical Aspects of Computer Science*, STACS '16, pages 25:1–25:14, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[14] Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 685–694, Washington, DC, USA, 2016. IEEE Computer Society.

[15] Tuğkan Batu and Clément L. Canonne. Generalized uniformity testing. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '17, pages 880–889, Washington, DC, USA, 2017. IEEE Computer Society.

[16] Clément L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22(63):63, 2015.

[17] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceesings of the 29th IEEE Symposium on Security and Privacy*, pages 111–125, 2008.

[18] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[19] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[20] Tore Dalenius. Towards a methodology for statistical disclosure control. *Statistisk Tidskrift*, 15:429–444, 1977.

[21] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, pages 202–210, New York, NY, USA, 2003. ACM.

[22] Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[23] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *Proceedings of the 54st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '13, pages 429–438. IEEE, 2013.

[24] Martin J Wainwright, Michael I Jordan, and John C Duchi. Privacy aware learning. In *Advances in Neural Information Processing Systems*, pages 1430–1438, 2012.

[25] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

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

[27] Cynthia Dwork. Differential privacy: A survey of results. In *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation*, TAMC '08, pages 1–19, Berlin, Heidelberg, 2008. Springer.

[28] Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '10, pages 51–60, Washington, DC, USA, 2010. IEEE Computer Society.

[29] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)*, 60(2):12, 2013.

[30] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science*, pages 94–103. IEEE, 2007.

[31] Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, and Vibhor Rastogi. The matrix mechanism: Optimizing linear counting queries under differential privacy. *The VLDB Journal*, 24(6):757–781, 2015.

[32] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017.

[33] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[34] Bryan Cai, Constantinos Daskalakis, and Gautam Kamath. Priv'it: Private and sample efficient identity testing. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 635–644. JMLR, Inc., 2017.

[35] Oded Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 23, 2016.

[36] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, pages 51–60. IEEE, 2014.

[37] Jayadev Acharya, Constantinos Daskalakis, and Gautam C Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems*, NIPS '15, pages 3577–3598. Curran Associates, Inc., 2015.

[38] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *Proceedings of the 45th International Colloquium on Automata, Languages, and Programming*, ICALP '18, pages 41:1–41:14, 2018.

[39] Yue Wang, Jaewoo Lee, and Daniel Kifer. Revisiting differentially private hypothesis tests for categorical data. *arXiv preprint arXiv:1511.03376*, 2015.

[40] Marco Gaboardi, Hyun-Woo Lim, Ryan M. Rogers, and Salil P. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *Proceedings of the 33rd International Conference on Machine Learning*, ICML '16, pages 1395–1403. JMLR, Inc., 2016.

[41] Ryan Rogers and Daniel Kifer. A New Class of Private Chi-Square Hypothesis Tests. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 991–1000, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

[42] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM Conference on Computer and Communications Security*, CCS '14, pages 1054–1067, New York, NY, USA, 2014. ACM.

[43] Adriano Pastore and Michael Gastpar. Locally differentially-private distribution estimation. In *Proceedings of the 2016 IEEE International Symposium on Information Theory*, pages 2694–2698, 2016.

[44] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 2436–2444, 2016.

[45] Shaowei Wang, Liusheng Huang, Pengzhan Wang, Yiwen Nie, Hongli Xu, Wei Yang, Xiang-Yang Li, and Chunming Qiao. Mutual information optimally local private discrete distribution estimation. *arXiv preprint arXiv:1607.08025*, 2016.

[46] Min Ye and Alexander Barg. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 64:5662–5676, 2018.

[47] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Hadamard response: Estimating distributions privately, efficiently, and with little communication. *arXiv preprint arXiv:1802.04705*, 2018.

[48] Or Sheffet. Locally private hypothesis testing. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 4612–4621. PMLR, 10–15 Jul 2018.

[49] Jayadev Acharya, Clément L Canonne, Cody Freitag, and Himanshu Tyagi. Test without trust: Optimal locally private distribution testing. *arXiv preprint arXiv:1808.02174*, 2018.

[50] Ilias Diakonikolas, Moritz Hardt, and Ludwig Schmidt. Differentially private learning of structured discrete distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 2566–2574. Curran Associates, Inc., 2015.

[51] Darakhshan J Mir. Information-theoretic foundations of differential privacy. In *International Symposium on Foundations and Practice of Security*, pages 374–381, 2012.

[52] Lalitha Sankar, S Raj Rajagopalan, and H Vincent Poor. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, 8(6):838–852, 2013.

[53] Paul Cuff and Lanqing Yu. Differential privacy as a mutual information constraint. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 43–54. ACM, 2016.

[54] Weina Wang, Lei Ying, and Junshan Zhang. On the relation between identifiability, differential privacy, and mutual-information privacy. *IEEE Transactions on Information Theory*, 62(9):5018–5029, 2016.

[55] Ibrahim Issa and Aaron B. Wagner. Operational definitions for some common information leakage metrics. In *Proceedings of the 2017 IEEE International Symposium on Information Theory*, ISIT '17, 2017.

[56] Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, pages 705–714. ACM, 2010.

[57] Salil Vadhan. The complexity of differential privacy. In Yehuda Lindell, editor, *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich*, chapter 7, pages 347–450. Springer International Publishing AG, Cham, Switzerland, 2017.

[58] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pages 44:1–44:9. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018.

[59] Gilles Barthe, Marco Gaboardi, Benjamin Grégoire, Justin Hsu, and Pierre-Yves Strub. Proving differential privacy via probabilistic couplings. In *Proceedings of the 31st Annual ACM/IEEE Symposium on Logic in Computer Science*, pages 749–758. ACM, 2016.

[60] Gilles Barthe, Noémie Fong, Marco Gaboardi, Benjamin Grégoire, Justin Hsu, and Pierre-Yves Strub. Advanced probabilistic couplings for differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 55–67. ACM, 2016.

[61] Cynthia Dwork, Moni Naor, Omer Reingold, and Guy N Rothblum. Pure differential privacy for rectangle queries via private partitions. In *International Conference on the Theory and Application of Cryptology and Information Security*, pages 735–751. Springer, 2015.

[62] Jayadev Acharya, Gautam Kamath, Ziteng Sun, and Huanyu Zhang. INSPECTRE: Privately estimating the unseen. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 30–39, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[63] 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*, pages 169–178, 2018.

[64] Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer New York, 1997.

[65] Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. Estimating Rényi entropy of discrete distributions. *IEEE Transactions on Information Theory*, 63(1):38–56, Jan 2017.

[66] Andreas Knoblauch. Closed-form expressions for the moments of the binomial probability distribution. *SIAM Journal on Applied Mathematics*, 69(1):197–204, 2008.

[67] Frank den Hollander. Probability theory: The coupling method. *Lecture notes available online (http://websites. math. leidenuniv. nl/probability/lecturenotes/CouplingLectures. pdf)*, 2012.

A Proof of Lemmas

A.1 Proof of Lemma 1

Proof. Let $d_H(x_1^m, y_1^m) = \hat{D}$. When $\hat{D} = 0$ or 1, the lemma is trivially true. Then, suppose $\hat{D} \geq 2$ we can find $\hat{D} - 1$ sequences $z_1^m, \ldots, z_{\hat{D}-1}^m$ over $\mathcal{X}^m$ with $d_H(x_1^m, z_1^m) = 1, d_H(z_{\hat{D}-1}^m, y_1^m) = 1$ and $d_H(z_i^m, z_{i+1}^m) = 1$ for $i \in \{1, 2, ..., \hat{D} - 2\}$. Hence, by the condition of (ε, δ) - differential privacy,

$$\begin{aligned}\Pr\left(\mathcal{A}(x_1^m) = q\right) &\leq e^{\varepsilon} \Pr\left(\mathcal{A}(z_1^m) = q\right) + \delta \leq e^{\varepsilon}(e^{\varepsilon} \Pr\left(\mathcal{A}(z_2^m) = q\right) + \delta) + \delta \leq \ldots \\ &\leq e^{\hat{D}\varepsilon} \Pr\left(\mathcal{A}(y_1^m) = q\right) + \delta \cdot \sum_{i=0}^{\hat{D}-1} e^{i\varepsilon} \leq e^{\hat{D}\varepsilon} \Pr\left(\mathcal{A}(y_1^m) = q\right) + \delta \hat{D} e^{(\hat{D}-1)\varepsilon} \\ &\leq e^{t\varepsilon} \Pr\left(\mathcal{A}(y_1^m) = q\right) + \delta t e^{\varepsilon(t-1)}.\end{aligned}$$

□

A.2 Proof of Lemma 2

Proof. Suppose $\mathcal{A}$ is a $(\varepsilon + \delta)$-differentially private algorithm. Then for any X_1^m and Y_1^m with $d_H(X_1^m, Y_1^m) \leq 1$ and any $S \subset \text{range}(\mathcal{A})$, we have

$$\Pr\left(\mathcal{A}(X_1^m) \in S\right) \leq e^{\varepsilon} \cdot \Pr\left(\mathcal{A}(Y_1^m) \in S\right) + (e^{\delta} - 1) \cdot e^{\varepsilon} \Pr\left(\mathcal{A}(Y_1^m) \in S\right).$$

If $e^{\varepsilon} \cdot \Pr\left(\mathcal{A}(Y_1^m) \in S\right) > 1 - \delta$, then $\Pr\left(\mathcal{A}(X_1^m) \in S\right) \leq 1 < e^{\varepsilon} \cdot \Pr\left(\mathcal{A}(Y_1^m) \in S\right) + \delta$. Otherwise, $e^{\varepsilon} \cdot \Pr\left(\mathcal{A}(Y_1^m) \in S\right) \leq 1 - \delta$. To prove $(e^{\delta} - 1) \cdot e^{\varepsilon} \cdot \Pr\left(\mathcal{A}(Y_1^m) \in S\right) < \delta$, it suffices to show $(e^{\delta} - 1)(1 - \delta) \leq \delta$, which is equivalent to $e^{-\delta} \geq 1 - \delta$, completing the proof. □

B Proof of Theorem 3

Proof. We first briefly describe the essential components of the construction of [35]. Given an explicit distribution q over $[k]$, there exists a randomized function $F_q : [k] \to [6k]$ such that if $X \sim q$, then $F_q(X) \sim u[6k]$, and if $X \sim p$ for a distribution with $d_{TV}(p, q) \geq \alpha$, then the distribution of $F_q(X)$ has a total variation distance of at least $\alpha/3$ from $u[6k]$. Given s samples X_1^s from a distribution p over $[k]$. Apply F_q independently to each of the X_i to obtain a new sequence $Y_1^s = F_q(X_1^s) := F_q(X_1) \ldots F_q(X_s)$. Let $\mathcal{A}$ be an algorithm that distinguishes $u[6k]$ from all distributions with total variation distance at least $\alpha/3$ from it. Then consider the algorithm $\mathcal{A}'$ that outputs $p = q$ if $\mathcal{A}$ outputs "$p = u[6k]$", and outputs $p \neq q$ otherwise. This shows that without privacy constraints, $S(\texttt{IT}, k, \alpha) \leq S(\texttt{UT}, 6k, \alpha/3)$ (See [35] for details).

We now prove that if further $\mathcal{A}$ was an ε-DP algorithm, then $\mathcal{A}'$ is also an ε-DP algorithm. Suppose X_1^s, and $X_1'^s$ be two sequences in $[k]^s$ that could differ only on the last coordinate, namely $X_1^s = X_1^{s-1} X_s$, and $X_1'^s = X_1^{s-1} X_s'$.

Consider two sequences $Y_1^s = Y_1^{s-1} Y_s$, and $Y_1'^s = Y_1^{s-1} Y_s'$ in $[6k]^s$ that could differ on only the last coordinate. Since $\mathcal{A}$ is ε-DP,

$$\mathcal{A}(Y_1^s = u[6k]) \leq \mathcal{A}(Y_1'^s = u[6k]) \cdot e^{\varepsilon}. \tag{3}$$

Moreover, since F_q is applied independently to each coordinate,

$$\Pr\left(F_q(X_1^s) = Y_1^s\right) = \Pr\left(F_q(X_1^{s-1}) = Y_1^{s-1}\right) \Pr\left(F_q(X_s) = Y_s\right).$$

Then,

$$\begin{aligned}
&\Pr\left(\mathcal{A}^{'}(X_1^s)=q\right)\\
&=\Pr\left(\mathcal{A}(F_q(X_1^s))=u[6k]\right)\\
&=\sum_{Y_1^s}\Pr\left(\mathcal{A}(Y_1^s)=u[6k]\right)\Pr\left(F_q(X_1^s)=Y_1^s\right)\\
&=\sum_{Y_1^{s-1}}\sum_{Y_s\in[6k]}\Pr\left(\mathcal{A}(Y_1^s)=u[6k]\right)\Pr\left(F_q(X_1^{s-1})=Y_1^{s-1}\right)\Pr\left(F_q(X_s)=Y_s\right)\\
&=\sum_{Y_1^{s-1}}\Pr\left(F_q(X_1^{s-1})=Y_1^{s-1}\right)\left[\sum_{Y_s\in[6k]}\Pr\left(\mathcal{A}(Y_1^s)=u[6k]\right)\Pr\left(F_q(X_s)=Y_s\right)\right]. \qquad (4)
\end{aligned}$$

Similarly,

$$\Pr\left(\mathcal{A}^{'}(X_1^{'s})=q\right)=\sum_{Y_1^{s-1}}\Pr\left(F_q(X_1^{s-1})=Y_1^{s-1}\right)\left[\sum_{Y_s^{'}\in[6k]}\Pr\left(\mathcal{A}(Y_1^{'s})=u[6k]\right)\Pr\left(F_q(X_s^{'})=Y_s^{'}\right)\right]. \qquad (5)$$

For a fixed Y_1^{s-1}, the term within the bracket in (4), and (5) are both expectations over the final coordinate. However, by (3) these expectations differ at most by a multiplicative e^{ε} factor. This implies that

$$\Pr\left(\mathcal{A}'(X_1^s)=q\right)\le\Pr\left(\mathcal{A}'(X_1^{'s})=q\right)e^{\varepsilon}.$$

The argument is similar for the case when the testing output is **not** $u[6k]$, and is omitted here. We only considered sequences that differ on the last coordinate, and the proof remains the same when any of the coordinates is changed. This proves the privacy guarantees of the algorithm. □

C Sample Complexity Bound of Algorithm 1

In this section, we prove the sample complexity bound of Algorithm 1 where we privatize the statistic proposed in [38] to achieve the sample complexity in Theorem 2 for all parameter ranges.

Because of the normalization in Equation 2 and lemma 5, for X_1^m drawn from $u[k]$

$$\mathbb{E}\left[Z(X_1^m)\right]\le\begin{cases}-\frac{1}{2}c\alpha^2\cdot\frac{m^2}{k}, & \text{when } m\le k,\\ -\frac{1}{2}c\alpha^2\cdot\frac{m^{3/2}}{k^{1/2}}, & \text{when } k<m\le\frac{k}{\alpha^2},\\ -\frac{1}{2}cm\alpha, & \text{when } m\ge\frac{k}{\alpha^2}\ .\end{cases} \qquad (6)$$

For X_1^m drawn from p with $d_{TV}(p,u[k])\ge\alpha$,

$$\mathbb{E}\left[Z(X_1^m)\right]\ge\begin{cases}\frac{1}{2}c\alpha^2\cdot\frac{m^2}{k}, & \text{when } m\le k,\\ \frac{1}{2}c\alpha^2\cdot\frac{m^{3/2}}{k^{1/2}}, & \text{when } k<m\le\frac{k}{\alpha^2},\\ \frac{1}{2}cm\alpha, & \text{when } m\ge\frac{k}{\alpha^2}\ .\end{cases} \qquad (7)$$

In order to prove the utility bounds, we also need the following (weak) version of the result of [38], which is sufficient to prove the sample complexity bound for constant error probability.

Lemma 9. *There is a constant $C>0$, such that when $m>C\sqrt{k}/\alpha^2$, then for $X_1^m\sim p$, where either $p=u[k]$, or $d_{TV}(p,u[k])\ge\alpha$,*

$$\Pr\left(\left|Z(X_1^m)-\mathbb{E}\left[Z(X_1^m)\right]\right|>\frac{2\mathbb{E}\left[Z(X_1^m)\right]}{3}\right)<0.01.$$

The proof of this result is in Appendix C.1.

We now proceed to prove the sample complexity bounds. Assume that $m > C\sqrt{k}/\alpha^2$, so Lemma 9 holds. Suppose ε be any real number such that $\varepsilon|\mathbb{E}\left[Z(X_1^m)\right]| > 3\log 100$. Let $\mathcal{A}(X_1^m)$ be the output of Algorithm 1. Denote the output by 1 when $\mathcal{A}(X_1^m)$ is "$p \neq u[k]$", and 0 otherwise. Consider the case when $X_1^m \sim p$, and $d_{TV}(p, u[k]) \geq \alpha$. Then,

$$\begin{aligned}
\Pr\left(\mathcal{A}(X_1^m) = 1\right) &\geq \Pr\left(\mathcal{A}(X_1^m) = 1 \text{ and } Z(X_1^m) > \frac{\mathbb{E}\left[Z(X_1^m)\right]}{3}\right) \\
&= \Pr\left(Z(X_1^m) > \frac{\mathbb{E}\left[Z(X_1^m)\right]}{3}\right) \cdot \Pr\left(\mathcal{A}(X_1^m) = 1 | Z(X_1^m) > \frac{\mathbb{E}\left[Z(X_1^m)\right]}{3}\right) \\
&\geq 0.99 \cdot \Pr\left(B\left(\sigma(\varepsilon \cdot \frac{\mathbb{E}\left[Z(X_1^m)\right]}{3})\right) = 1\right) \\
&\geq 0.99 \cdot 0.99 \geq 0.9,
\end{aligned}$$

where the last step uses that $\varepsilon\mathbb{E}\left[Z(X_1^m)\right]/3 > \log 100$, along with Lemma 4. The case of $p = u[k]$ follows from the same argument.

Therefore, the algorithm is correct with probability at least 0.9, whenever, $m > C\sqrt{k}/\alpha^2$, and $\varepsilon|\mathbb{E}\left[Z(X_1^m)\right]| > 3\log 100$. By (7), note that $\varepsilon|\mathbb{E}\left[Z(X_1^m)\right]| > 3\log 100$ is satisfied when,

$$\begin{aligned}
c\alpha^2 \cdot m^2/k &\geq (6\log 100)/\varepsilon, \quad \text{for } m \leq k, \\
c\alpha^2 \cdot m^{3/2}/k^{1/2} &\geq (6\log 100)/\varepsilon, \quad \text{for } k < m \leq k/\alpha^2, \\
c\alpha \cdot m &\geq (6\log 100)/\varepsilon, \quad \text{for } m \geq k/\alpha^2.
\end{aligned}$$

This gives the upper bounds for all the three regimes of m.

C.1 Proof of Lemma 9

In order to prove the lemma, we need the following lemma, which is proved in [38].

Lemma 10. *(Bernstein version of McDiarmid's inequality) Let Y_1^m be independent random variables taking values in the set $\mathcal{Y}$. Let $f : \mathcal{Y}^m \to \mathbb{R}$ be a function of Y_1^m so that for every $j \in [m]$, and $y_1, ...y_m, y_j' \in \mathcal{Y}$, we have that:*

$$\left|f(y_1, ...y_j, ...y_m) - f(y_1, ..., y_j', ...y_m)\right| \leq B,$$

Then we have

$$\Pr\left(f - \mathbb{E}\left[f\right] \geq z\right) \leq \exp\left(\frac{-2z^2}{mB^2}\right).$$

In addition, if for each $j \in [m]$ and $y_1, ...y_{j-1}, y_{j+1}, ...y_m$ we have that

$$\mathrm{Var}_{Y_j}[f(y_1, ...y_j, ...y_m)] \leq \sigma_j^2,$$

then we have

$$\Pr\left(f - \mathbb{E}\left[f\right] \geq z\right) \leq \exp\left(\frac{-z^2}{\sum_{j=1}^m \sigma_j^2 + 2Bz/3}\right).$$

The statistic we use $Z(X_1^m)$ has sensitivity at most 1, hence we can use $B = 1$ in Lemma 10.

We first consider the case when $k < m \leq \frac{k}{\alpha^2}$. When $p = u[k]$, we get $\mathbb{E}\left[Z(X_1^m)\right] = -\frac{1}{2}cm\alpha^2 \cdot \sqrt{\frac{m}{k}}$, then by the first part of Lemma 10,

$$\begin{aligned}
\Pr\left(Z(X_1^m) > \frac{\mathbb{E}\left[Z(X_1^m)\right]}{3}\right) &= \Pr\left(Z(X_1^m) > -\frac{1}{6}cm\alpha^2 \cdot \sqrt{\frac{m}{k}}\right) \\
&\leq \Pr\left(Z(X_1^m) - \mathbb{E}\left[Z(X_1^m)\right] > \frac{2}{3}cm\alpha^2 \cdot \sqrt{\frac{m}{k}}\right) \\
&\leq \exp\left(-\frac{8c^2m^2\alpha^4}{9k}\right). \qquad (8)
\end{aligned}$$

Therefore, there is a C_1 such that if $m \geq C_1\sqrt{k}/\alpha^2$, then under the uniform distribution $\Pr\left(Z(X_1^m) > \frac{\mathbb{E}[Z(X_1^m)]}{3}\right)$ is at most 1/100. The non-uniform distribution part is similar and we omit the case.

Then we consider the case when $\frac{k}{\alpha^2} < m$. When $p = u[k]$, we get $\mathbb{E}[Z(X_1^m)] = -\frac{1}{2}cm\alpha$, then also by the first part of Lemma 10,

$$\begin{aligned}\Pr\left(Z(X_1^m) > \frac{\mathbb{E}[Z(X_1^m)]}{3}\right) &= \Pr\left(Z(X_1^m) > -\frac{1}{6}cm\alpha\right)\\ &\leq \Pr\left(Z(X_1^m) - \mathbb{E}[Z(X_1^m)] > \frac{2}{3}cm\alpha\right)\\ &\leq \exp\left(-\frac{8c^2m\alpha^2}{9}\right).\end{aligned}$$

Using the same argument we can show that there is a constant C_2 such that for $m \geq C_2/\alpha^2$, then under the uniform distribution $\Pr\left(Z(X_1^m) > \frac{\mathbb{E}[Z(X_1^m)]}{3}\right)$ is at most 1/100. The case of non-uniform distribution is omitted because of the same reason.

At last we consider the case when $m \leq k$. In this case we need another result proved in [38]:

$$\mathrm{Var}_{X_j}[Z(x_1, x_2, ..., X_j, .., x_m)] \leq \frac{m}{k}, \forall j, x_1, x_2, ..., x_{j-1}, x_{j+1}, ...x_n.$$

When $p = u[k]$, we get $\mathbb{E}[Z(X_1^m)] = -\frac{1}{2}ck\alpha^2 \cdot \frac{m^2}{k^2}$, then by the second part of Lemma 10,

$$\begin{aligned}\Pr\left(Z(X_1^m) > \frac{\mathbb{E}[Z(X_1^m)]}{3}\right) &= \Pr\left(Z(X_1^m) > -\frac{1}{6}c\alpha^2 \cdot \frac{m^2}{k}\right)\\ &\leq \Pr\left(Z(X_1^m) - \mathbb{E}[Z(X_1^m)] > \frac{2}{3}c\alpha^2 \cdot \frac{m^2}{k}\right)\\ &\leq \exp\left(\frac{-\frac{4}{9}c^2\alpha^4\frac{m^4}{k^2}}{\frac{m^2}{k} + \frac{4}{9}c\alpha^2\frac{m^2}{k}}\right)\\ &\leq \exp\left(-\frac{2}{9}c\alpha^4\frac{m^2}{k}\right).\end{aligned}$$

Therefore, there is a C_3 such that if $m \geq C_3\sqrt{k}/\alpha^2$, then under the uniform distribution $\Pr\left(Z(X_1^m) > \frac{\mathbb{E}[Z(X_1^m)]}{3}\right)$ is at most 1/100. The case of non-uniform distribution is similar and is omitted.

Therefore, if we take $C = \max\{C_1, C_2, C_3\}$, we prove the result in the lemma.

D Proof of Lemma 7

D.1 $m \leq k, \min\{\frac{m^2}{k}, \frac{m^{3/2}}{k^{1/2}}\} = \frac{m^2}{k}$,

Before proving the lemma, we consider an example that will provide insights and tools to analyze the distributions Q_1, and Q_2. Let $t \in \mathbb{N}$. Let P_2 be the following distribution over $\{0,1\}^t$:

- Select $b \in \{\frac{1}{2} - \alpha, \frac{1}{2} + \alpha\}$ with equal probability.
- Output t independent samples from $B(b)$.

Let P_1 be the distribution over $\{0,1\}^t$ that outputs t independent samples from $B(0.5)$.

When $t = 1$, P_1 and P_2 both become $B(0.5)$. For t=2, $P_1(00) = P_1(11) = \frac{1}{4} + \alpha^2$, and $P_1(10) = P_1(01) = \frac{1}{4} - \alpha^2$, and $d_{TV}(P_1, P_2)$ is $2\alpha^2$. A slightly general result is the following:

Lemma 11. *For $t = 1$, $d_{TV}(P_1, P_2) = 0$ and for $t \geq 2$, $d_{TV}(P_1, P_2) \leq 2t\alpha^2$.*

Proof. Consider any sequence X_1^t that has t_0 zeros, and $t_1 = t - t_0$ ones. Then,

$$P_1(X_1^t) = \binom{t}{t_0}\frac{1}{2^t},$$

and

$$P_2(X_1^t) = \binom{t}{t_0}\frac{1}{2^t}\left(\frac{(1-2\alpha)^{t_0}(1+2\alpha)^{t_1} + (1+2\alpha)^{t_0}(1-2\alpha)^{t_1}}{2}\right).$$

The term in the parentheses above is minimized when $t_0 = t_1 = t/2$. In this case,

$$P_2(X_1^t) \geq P_1(X_1^t)\cdot(1+2\alpha)^{t/2}(1-2\alpha)^{t/2} = P_1(X_1^t)\cdot(1-4\alpha^2)^{t/2}.$$

Therefore,

$$d_{TV}(P_1,P_2) = \sum_{P_1>P_2} P_1(X_1^t) - P_2(X_1^t) \leq \sum_{P_1>P_2} P_1(X_1^t)\Big(1-(1-4\alpha^2)^{t/2}\Big) \leq 2t\alpha^2,$$

where we used the Weierstrass Product Inequality, which states that $1-tx \leq (1-x)^t$ proving the total variation distance bound. □

As a corollary this implies:

Lemma 12. *There is a coupling between X_1^t generated from P_1 and Y_1^t from P_2 such that $\mathbb{E}\left[d_H(X_1^t, Y_1^t)\right] \leq t \cdot d_{TV}(P_1,P_2) \leq 4(t^2-t)\alpha^2$.*

Proof. Observe that $\sum_{X_1^t} \min\{P_1(X_1^t), P_2(X_1^t)\} = 1 - d_{TV}(P_1,P_2)$. Consider the following coupling between P_1, and P_2. Suppose X_1^t is generated by P_1, and let R be a $U[0,1]$ random variable.

1. $R < 1 - d_{TV}(P_1,P_2)$ Generate X_1^t from the distribution that assigns probability $\frac{\min\{P_1(X_1^t),P_2(X_1^t)\}}{1-d_{TV}(P_1,P_2)}$ to X_1^t. Output (X_1^t, X_1^t).
2. $R \geq 1 - d_{TV}(P_1,P_2)$ Generate X_1^t from the distribution that assigns probability $\frac{P_1(X_1^t)-\min\{P_1(X_1^t),P_2(X_1^t)\}}{d_{TV}(P_1,P_2)}$ to X_1^t, and Y_1^t from the distribution that assigns probability $\frac{P_2(Y_1^t)-\min\{P_1(Y_1^t),P_2(Y_1^t)\}}{d_{TV}(P_1,P_2)}$ to Y_1^t independently. Then output (X_1^t, Y_1^t).

To prove the coupling, note that the probability of observing X_1^t is

$$(1-d_{TV}(P_1,P_2))\cdot\frac{\min\{P_1(X_1^t),P_2(X_1^t)\}}{1-d_{TV}(P_1,P_2)} + d_{TV}(P_1,P_2)\cdot\frac{P_1(X_1^t)-\min\{P_1(X_1^t),P_2(X_1^t)\}}{d_{TV}(P_1,P_2)} = P_1(X_1^t).$$

A similar argument gives the probability of Y_1^t to be $P_2(Y_1^t)$.

Then $\mathbb{E}\left[d_H(X_1^t, Y_1^t)\right] \leq t \cdot d_{TV}(P_1,P_2) = 2t^2\alpha^2 \leq 4(t^2-t)\alpha^2$ when $t \geq 2$, and when $t = 1$, the distributions are identical and the Hamming distance of the coupling is equal to zero. □

We now have the tools to prove Lemma 7 for $m \leq k$.

Proof of Lemma 7 for $m \leq k$. The following is a coupling between Q_1 and Q_2:

1. Generate m samples Z_1^m from a uniform distribution over $[k/2]$.
2. For $j \in [k/2]$, let $T_j \subseteq [m]$ be the set of locations where j appears. Note that $|T_j| = M_j(Z_1^m)$.
3. To generate samples from Q_1:
   - Generate $|T_j|$ samples from a uniform distribution over $\{2j-1, 2j\}$, and replace the symbols in T_j with these symbols.
4. To generate samples from Q_2:
   - Similar to the construction of P_1 earlier in this section, consider two distributions over $\{2j-1, 2j\}$ with bias $\frac{1}{2}-\alpha$, and $\frac{1}{2}+\alpha$.
   - Pick one of these distributions at random.
   - Generate $|T_j|$ samples from it over $\{2j-1, 2j\}$, and replace the symbols in T_j with these symbols.

From this process the coupling between Q_1, and Q_2 is also clear:

- Given X_1^m from Q_2, for each $j \in [k/2]$ find all locations ℓ such that $X_\ell = 2j-1$, or $X_\ell = 2j$. Call this set T_j.
- Perform the coupling between P_2 and P_1 from Lemma 12, after replacing $\{0,1\}$ with $\{2j-1, 2j\}$.

Using the coupling defined above, by the linearity of expectations, we get:

$$\begin{aligned}\mathbb{E}\left[d_H(X_1^m, Y_1^m)\right] &= \sum_{j=1}^{k/2} \mathbb{E}\left[d_H(X_1^{|T_j|}, Y_1^{|T_j|})\right] \\ &= \frac{k}{2}\mathbb{E}\left[d_H(X_1^R, Y_1^R)\right] \\ &\leq \frac{k}{2} \cdot \mathbb{E}\left[4\alpha^2(R^2 - R)\right],\end{aligned}$$

where R is a binomial random variable with parameters m and $2/k$. Now, a simple exercise computing Binomial moments shows that for $X \sim Bin(n,s)$, $\mathbb{E}\left[X^2 - X\right] = s^2(n^2-n) \leq n^2 s^2$. This implies that

$$\mathbb{E}\left[R^2 - R\right] \leq \frac{4m^2}{k^2}.$$

Plugging this, we obtain

$$\mathbb{E}\left[d_H(X_1^m, Y_1^m)\right] \leq \frac{k}{2} \cdot \frac{16\alpha^2 m^2}{k^2} = \frac{8m^2\alpha^2}{k},$$

proving the claim. □

D.2 $k \leq m \leq k/\alpha^2, \min\{\frac{m^2}{k}, \frac{m^{3/2}}{k^{1/2}}\} = \frac{m^{3/2}}{k^{1/2}}$

Lemma 11 holds for all values of t, and α. The lemma can be strengthened for cases where α is small.

Lemma 13. *Let P_1, and P_2 be the distributions over $\{0,1\}^t$ defined in the last section. There is a coupling between X_1^t generated by P_1, and Y_1^t by P_2 such that*

$$\mathbb{E}\left[d_H(X_1^t, Y_1^t)\right] \leq C \cdot (\alpha^2 t^{3/2} + \alpha^4 t^{5/2} + \alpha^5 t^3).$$

D.2.1 Proof of Lemma 7 assuming Lemma 13

Given the coupling we defined in Appendix D.2.2 for proving Lemma 13, the coupling between Q_1, and Q_2 uses the same technique in the last section for $m \leq k$.

- Given X_1^m from Q_2, for each $j \in [k/2]$ find all locations ℓ such that $X_\ell = 2j-1$, or $X_\ell = 2j$. Call this set T_j.
- Perform the coupling in Appendix D.2.2 between P_2 and P_1 on T_j, after replacing $\{0,1\}$ with $\{2j-1, 2j\}$.

Using the coupling defined above, by the linearity of expectations, we get:

$$\begin{aligned}\mathbb{E}\left[d_H(X_1^m, Y_1^m)\right] &= \sum_{j=1}^{k/2} \mathbb{E}\left[d_H(X_1^{|T_j|}, Y_1^{|T_j|})\right] \\ &= \frac{k}{2}\mathbb{E}\left[d_H(X_1^R, Y_1^R)\right] \\ &\leq \frac{k}{2} \cdot \mathbb{E}\left[64 \cdot \left(\alpha^4 R^{5/2} + \alpha^2 R^{3/2} + \alpha^5 R^3\right)\right],\end{aligned}$$

where $R \sim \mathrm{Bin}(m, 2/k)$.

We now bound the moments of Binomial random variables. The bound is similar in flavor to [65, Lemma 3] for Poisson random variables.

Lemma 14. *Suppose $\frac{m}{k} > 1$, and $Y \sim \mathrm{Bin}(m, \frac{1}{k})$, then for $\gamma \geq 1$, there is a constant C_γ such that*

$$\mathbb{E}\left[Y^\gamma\right] \leq C_\gamma \Big(\frac{m}{k}\Big)^\gamma.$$

Proof. For integer values of γ, this directly follows from the moment formula for Binomial distribution [66], and for other $\gamma \geq 1$, by Jensen's Inequality

$$\mathbb{E}\left[Y^\gamma\right] \leq \mathbb{E}\left[\Big(Y^{\lceil\gamma\rceil}\Big)^{\frac{\gamma}{\lceil\gamma\rceil}}\right] \leq \mathbb{E}\left[\Big(Y^{\lceil\gamma\rceil}\Big)\right]^{\frac{\gamma}{\lceil\gamma\rceil}} \leq \Big(C_{\lceil\gamma\rceil}\mathbb{E}\left[Y\right]^{\lceil\gamma\rceil}\Big)^{\frac{\gamma}{\lceil\gamma\rceil}} = C'(\mathbb{E}\left[Y\right])^\gamma,$$

proving the lemma. □

Therefore, letting $C = \max\{C_{5/2}, C_3, C_{3/2}\}$, we obtain

$$\mathbb{E}\left[d_H(X_1^m, Y_1^m)\right] \leq 32kC \cdot \left(\alpha^4\Big(\frac{m}{k}\Big)^{5/2} + \alpha^2\Big(\frac{m}{k}\Big)^{3/2} + \alpha^5\Big(\frac{m}{k}\Big)^3\right).$$

Now, notice $\alpha\sqrt{\frac{m}{k}} < 1$. Plugging this,

$$\begin{aligned}
\mathbb{E}\left[d_H(X_1^m, Y_1^m)\right] &\leq 32C \cdot k \cdot \left(\alpha^4\Big(\frac{m}{k}\Big)^{5/2} + \alpha^2\Big(\frac{m}{k}\Big)^{3/2} + \alpha^5\Big(\frac{m}{k}\Big)^3\right) \\
&= 32C \cdot k\alpha^2 \cdot \left(\alpha^2\frac{m}{k} \cdot \Big(\frac{m}{k}\Big)^{3/2} + \Big(\frac{m}{k}\Big)^{3/2} + \alpha^3\Big(\frac{m}{k}\Big)^{3/2}\Big(\frac{m}{k}\Big)^{3/2}\right) \\
&\leq 96C \cdot k\Big(\frac{m}{k}\Big)^{3/2},
\end{aligned}$$

completing the argument.

D.2.2 Proof of Lemma 13

To prove Lemma 13, we need a few lemmas first:

Definition 7. A random variable Y_1 is said to stochastically dominate Y_2 if for all t, $\Pr\left(Y_1 \geq t\right) \geq \Pr\left(Y_2 \geq t\right)$.

Lemma 15. *Suppose $N_1 \sim \mathrm{Bin}(t, \frac{1}{2}), N_2 \sim \frac{1}{2}\mathrm{Bin}(t, \frac{1+\alpha}{2}) + \frac{1}{2}\mathrm{Bin}(t, \frac{1-\alpha}{2})$. Then $Z_2 = \max\{N_2, t - N_2\}$ stochastically dominates $Z_1 = \max\{N_1, t - N_1\}$.*

Proof.

$$\Pr\left(Z_2 \geq l\right) = \sum_{i=0}^{t-l}\binom{t}{i}\left[\Big(\frac{1+\alpha}{2}\Big)^i\Big(\frac{1-\alpha}{2}\Big)^{t-i} + \Big(\frac{1-\alpha}{2}\Big)^i\Big(\frac{1+\alpha}{2}\Big)^{t-i}\right],$$

$$\Pr\left(Z_1 \geq l\right) = 2 \cdot \sum_{i=0}^{t-l}\binom{t}{i}\Big(\frac{1}{2}\Big)^t.$$

Define $F(l) = \Pr\left(Z_2 \geq l\right) - \Pr\left(Z_1 \geq l\right)$. What we need to show is $F(l) \geq 0, \forall l \geq \frac{t}{2}$. First we observe that $\Pr\left(Z_2 \geq \frac{t}{2}\right) = \Pr\left(Z_1 \geq \frac{t}{2}\right) = 1$ and $\Pr\left(Z_2 \geq t\right) = (\frac{1+\alpha}{2})^t + (\frac{1-\alpha}{2})^t \geq 2(\frac{1}{2})^t = \Pr\left(Z_1 \geq t\right)$. Hence $F(\frac{t}{2}) = 0, F(t) > 0$. Let

$$f(l) = F(l+1) - F(l) = -\binom{t}{l}\left[\Big(\frac{1+\alpha}{2}\Big)^l\Big(\frac{1-\alpha}{2}\Big)^{t-l} + \Big(\frac{1-\alpha}{2}\Big)^l\Big(\frac{1+\alpha}{2}\Big)^{t-l} - 2\Big(\frac{1}{2}\Big)^t\right].$$

Let $g(x) = \big(\frac{1+\alpha}{2}\big)^x\big(\frac{1-\alpha}{2}\big)^{t-x} + \big(\frac{1-\alpha}{2}\big)^x\big(\frac{1+\alpha}{2}\big)^{t-x} - 2\big(\frac{1}{2}\big)^t, x \in [t/2, t]$, then

$$\frac{dg(x)}{dx} = \ln\Big(\frac{1+\alpha}{1-\alpha}\Big) \cdot \left[\Big(\frac{1+\alpha}{2}\Big)^x\Big(\frac{1-\alpha}{2}\Big)^{t-x} - \Big(\frac{1-\alpha}{2}\Big)^x\Big(\frac{1+\alpha}{2}\Big)^{t-x}\right] \geq 0.$$

We know $g(t/2) < 0, g(t) > 0$, hence $\exists x^*, s.t. g(x) \leq 0, \forall x < x^*$ and $g(x) \geq 0, \forall x > x^*$. Because $f(l) = -\binom{t}{l}g(l)$, hence $\exists l^*, s.t. f(l) \leq 0, \forall l \geq l^*$ and $f(l) \geq 0, \forall l < l^*$. Therefore, $F(l)$ first increases and then decreases, which means $F(l)$ achieves its minimum at $\frac{t}{2}$ or t. Hence $F(l) \geq 0$, completing the proof. □

For stochastic dominance, the following definition [67] will be useful.

Definition 8. A coupling (X', Y') is a monotone coupling if $\Pr(X' \geq Y') = 1$.

The following lemma states a nice relationship between stochastic dominance and monotone coupling, which is provided as Theorem 7.9 in [67]

Lemma 16. *Random variable X stochastically dominates Y if and only if there is a monotone coupling between (X', Y') with $\Pr(X' \geq Y') = 1$.*

By Lemma 16, there is a monotone coupling between $Z_1 = \max\{N_1, t - N_1\}$ and $Z_2 = \max\{N_2, t - N_2\}$. Suppose the coupling is $P^c_{Z_1,Z_2}$, we define the coupling between X_1^t and Y_1^t as following:

1. Generate X_1^t according to P_1 and count the number of one's in X_1^t as n_1.
2. Generate n_2 according to $P^c[Z_2|Z_1 = \max\{n_1, t - n_1\}]$.
3. If $n_1 > t - n_1$, choose $n_2 - n_1$ of the zero's in X_1^t uniformly at random and change them to one's to get Y_1^t.
4. If $n_1 < t - n_1$, choose $n_2 - (t - n_1)$ of the one's in X_1^t uniformly at random and change them to zero's to get Y_1^t.
5. If $n_1 = t - n_1$, break ties uniformly at random and do the corresponding action.
6. Output (X_1^t, Y_1^t).

Since the coupling is monotone, and $d_H(X_1^t, Y_1^t) = Z_2 - Z_1$ for every pair of (X_1^t, Y_1^t), we get:

$$\mathbb{E}\left[d_H(X_1^t, Y_1^t)\right] = \mathbb{E}\left[\max\{N_2, t - N_2\}\right] - \mathbb{E}\left[\max\{N_1, t - N_1\}\right].$$

Hence, to show lemma 13, it suffices to show the following lemma:

Lemma 17. *Suppose $N_1 \sim \mathrm{Bin}(t, \frac{1}{2}), N_2 \sim \frac{1}{2}\mathrm{Bin}(t, \frac{1+\alpha}{2}) + \frac{1}{2}\mathrm{Bin}(t, \frac{1-\alpha}{2})$.*

$$\mathbb{E}\left[\max\{N_2, t - N_2\}\right] - \mathbb{E}\left[\max\{N_1, t - N_1\}\right] < C \cdot (\alpha^2 t^{3/2} + \alpha^4 t^{5/2} + \alpha^5 t^3)$$

Proof.

$$\begin{aligned}
&\mathbb{E}\left[\max\{N_2, t - N_2\}\right] \\
&= \sum_{0 \leq \ell \leq t/2} (t/2 + \ell) \binom{t}{\frac{t}{2} - \ell} \left(\left(\frac{1-\alpha}{2}\right)^{\frac{t}{2}-\ell} \left(\frac{1+\alpha}{2}\right)^{\frac{t}{2}+\ell} + \left(\frac{1+\alpha}{2}\right)^{\frac{t}{2}-\ell} \left(\frac{1-\alpha}{2}\right)^{\frac{t}{2}+\ell} \right) \\
&= \frac{t}{2} + \sum_{0 \leq \ell \leq t/2} \ell \binom{t}{\frac{t}{2} - \ell} \left(\left(\frac{1-\alpha}{2}\right)^{\frac{t}{2}-\ell} \left(\frac{1+\alpha}{2}\right)^{\frac{t}{2}+\ell} + \left(\frac{1+\alpha}{2}\right)^{\frac{t}{2}-\ell} \left(\frac{1-\alpha}{2}\right)^{\frac{t}{2}+\ell} \right).
\end{aligned}$$

Consider a fixed value of t. Let

$$f(\alpha) = \sum_{0 \leq \ell \leq t/2} \ell \binom{t}{\frac{t}{2} - \ell} \left(\left(\frac{1-\alpha}{2}\right)^{\frac{t}{2}-\ell} \left(\frac{1+\alpha}{2}\right)^{\frac{t}{2}+\ell} + \left(\frac{1+\alpha}{2}\right)^{\frac{t}{2}-\ell} \left(\frac{1-\alpha}{2}\right)^{\frac{t}{2}+\ell} \right).$$

The first claim is that this expression is minimized at $\alpha = 0$. This is because of the monotone coupling between Z_1 and Z_2, which makes $\mathbb{E}[Z_2] \geq \mathbb{E}[Z_1]$. This implies that $f'(0) = 0$, and by intermediate value theorem, there is $\beta \in [0, \alpha]$, such that

$$f(\alpha) = f(0) + \frac{1}{2}\alpha^2 \cdot f''(\beta). \tag{9}$$

We will now bound this second derivative. To further simplify, let

$$g(\alpha) = \left(\frac{1-\alpha}{2}\right)^{\frac{t}{2}-\ell} \left(\frac{1+\alpha}{2}\right)^{\frac{t}{2}+\ell} + \left(\frac{1+\alpha}{2}\right)^{\frac{t}{2}-\ell} \left(\frac{1-\alpha}{2}\right)^{\frac{t}{2}+\ell}.$$

Differentiating $g(\alpha)$, twice with respect to α, we obtain,

$$\begin{aligned}
g''(\alpha) = &\frac{1}{16} \cdot \left(\alpha^2(t^2 - t) - 4\alpha\ell(t-1) + 4\ell^2 - t\right) \left(\frac{1-\alpha}{2}\right)^{\frac{t}{2}-\ell-2} \left(\frac{1+\alpha}{2}\right)^{\frac{t}{2}+\ell-2} \\
&+ \frac{1}{16} \cdot \left(\alpha^2(t^2 - t) + 4\alpha\ell(t-1) + 4\ell^2 - t\right) \left(\frac{1+\alpha}{2}\right)^{\frac{t}{2}-\ell-2} \left(\frac{1-\alpha}{2}\right)^{\frac{t}{2}+\ell-2}.
\end{aligned}$$

Then $g''(\alpha)$ can be bound by,

$$g''(\alpha) \le \frac{1}{16} \cdot \left(\alpha^2 t^2 + 4\ell^2\right)\left(\left(\frac{1-\alpha}{2}\right)^{\frac{t}{2}-\ell-2}\left(\frac{1+\alpha}{2}\right)^{\frac{t}{2}+\ell-2} + \left(\frac{1+\alpha}{2}\right)^{\frac{t}{2}-\ell-2}\left(\frac{1-\alpha}{2}\right)^{\frac{t}{2}+\ell-2}\right).$$

When $\alpha < \frac{1}{4}$, $(1-\alpha^2)^2 > \frac{1}{2}$, and we can further bound the above expression by

$$g''(\alpha) \le 2 \cdot \left(\alpha^2 t^2 + 4\ell^2\right)\left(\left(\frac{1-\alpha}{2}\right)^{\frac{t}{2}-\ell}\left(\frac{1+\alpha}{2}\right)^{\frac{t}{2}+\ell} + \left(\frac{1+\alpha}{2}\right)^{\frac{t}{2}-\ell}\left(\frac{1-\alpha}{2}\right)^{\frac{t}{2}+\ell}\right).$$

Suppose X is a $\mathrm{Bin}(t, \frac{1+\beta}{2})$ distribution. Then, for any $\ell > 0$,

$$\Pr\left(\left|X - \frac{t}{2}\right| = \ell\right) = \binom{t}{\frac{t}{2}-\ell}\left(\left(\frac{1-\beta}{2}\right)^{\frac{t}{2}-\ell}\left(\frac{1+\beta}{2}\right)^{\frac{t}{2}+\ell} + \left(\frac{1+\beta}{2}\right)^{\frac{t}{2}-\ell}\left(\frac{1-\beta}{2}\right)^{\frac{t}{2}+\ell}\right).$$

Therefore, we can bound (9), by

$$f''(\beta) \le 2 \cdot \left(\beta^2 t^2 \mathbb{E}\left[\left|X - \frac{t}{2}\right|\right] + 4\mathbb{E}\left[\left|X - \frac{t}{2}\right|^3\right]\right).$$

For $X \sim \mathrm{Bin}(m, r)$,

$$\mathbb{E}\left[(X - mr)^2\right] = mr(1-r) \le \frac{m}{4}, \text{ and}$$
$$\mathbb{E}\left[(X - mr)^4\right] = mr(1-r)(3r(1-r)(m-2)+1) \le 3\frac{m^2}{4}.$$

We bound each term using these moments,

$$\mathbb{E}\left[\left|X - \frac{t}{2}\right|\right] \le \mathbb{E}\left[\left(X - \frac{t}{2}\right)^2\right]^{1/2} = \left(t\frac{(1-\beta^2)}{4} + \left(\frac{t\beta}{2}\right)^2\right)^{1/2} \le \sqrt{t} + t\beta.$$

We similarly bound the next term,

$$\begin{aligned}
\mathbb{E}\left[\left|X - \frac{t}{2}\right|^3\right] &\le \mathbb{E}\left[\left(X - \frac{t}{2}\right)^4\right]^{3/4} \\
&\le \mathbb{E}\left[\left(X - \frac{t(1+\beta)}{2} + \frac{t\beta}{2}\right)^4\right]^{3/4} \\
&\le 8\left(\mathbb{E}\left[\left(X - \frac{t(1+\beta)}{2}\right)^4\right]^{3/4} + \left(\frac{t\beta}{2}\right)^3\right) \\
&\le 8\left(t^{3/2} + \left(\frac{t\beta}{2}\right)^3\right),
\end{aligned}$$

where we use $(a+b)^4 \le 8(a^4 + b^4)$.

Therefore,

$$f''(\beta) \le 64 \cdot \left(\beta^2 t^{5/2} + t^{3/2} + (t\beta)^3\right) \le 64 \cdot \left(\alpha^2 t^{5/2} + t^{3/2} + (t\alpha)^3\right).$$

As a consequence,

$$\mathbb{E}\left[\max\{N_2, t - N_2\}\right] - \mathbb{E}\left[\max\{N_1, t - N_1\}\right] = \alpha^2 f''(\beta) \le 64 \cdot (\alpha^2 t^{3/2} + \alpha^4 t^{5/2} + \alpha^5 t^3).$$

completing the proof. □

E Proof of Theorem 4

E.1 Closeness Testing – Upper Bounds

In this section, we will show that Algorithm 2 satisfies sample complexity upper bounds described in Theorem 4.

The results in [10] were proved under Poisson sampling, and we also use Poisson sampling, with only a constant factor effect on the number of samples for the same error probability. They showed the following bounds:

$$\mathbb{E}\left[Z(X_1^m, Y_1^m)\right] = 0 \text{ when } p = q, \tag{10}$$

$$\mathrm{Var}(Z(X_1^m, Y_1^m)) \le 2\min\{k, m\} \text{ when } p = q, \tag{11}$$

$$\mathbb{E}\left[Z(X_1^m, Y_1^m)\right] \ge \frac{m^2\alpha^2}{4k + 2m} \text{ when } d_{TV}(p, q) \ge \alpha, \tag{12}$$

$$\mathrm{Var}(Z(X_1^m, Y_1^m)) \le \frac{1}{1000}\mathbb{E}\left[Z(X_1^m, Y_1^m)\right]^2 \text{ when } p \ne q, \text{ and } m = \Omega\left(\frac{1}{\alpha^2}\right). \tag{13}$$

Case 1: $\alpha^2 > \frac{1}{\sqrt{k}}$, and $\alpha^2\varepsilon > \frac{1}{k}$. In this case, we will show that $S(\mathtt{CT}, k, \alpha, \varepsilon) = O\left(\frac{k^{2/3}}{\alpha^{4/3}} + \frac{k^{1/2}}{\alpha\sqrt{\varepsilon}}\right)$. In this case, $\frac{k^{2/3}}{\alpha^{4/3}} + \frac{k^{1/2}}{\alpha\sqrt{\varepsilon}} \le 2k$.

We consider the case when $p = q$, then $\mathrm{Var}(Z(X_1^m, Y_1^m)) \le 2\min\{k, m\}$. Let $\mathrm{Var}(Z(X_1^m, Y_1^m)) \le cm$ for some constant c. By the Chebyshev's inequality,

$$\begin{aligned}
\Pr\left(Z' > -\frac{1}{84}\cdot\frac{m^2\alpha^2}{4k+2m}\right) &\le \Pr\left(Z(X_1^m, Y_1^m) - \mathbb{E}\left[Z(X_1^m, Y_1^m)\right] > \frac{1}{3}\cdot\frac{m^2\alpha^2}{4k+2m}\right) \\
&\le \Pr\left(Z(X_1^m, Y_1^m) - \mathbb{E}\left[Z(X_1^m, Y_1^m)\right] > \frac{1}{3}\cdot\frac{m^2\alpha^2}{8k}\right) \\
&\le \Pr\left(Z(X_1^m, Y_1^m) - \mathbb{E}\left[Z(X_1^m, Y_1^m)\right] > (cm)^{1/2}\cdot\frac{m^{3/2}\alpha^2}{24c^{1/2}k}\right) \\
&\le 576c\cdot\frac{k^2}{m^3\alpha^4},
\end{aligned}$$

where we used that $4k + 2m \le 8k$.

Therefore, there is a C_1 such that if $m \ge C_1 k^{2/3}/\alpha^{4/3}$, then under $p = q$, $\Pr\left(Z' > -\frac{1}{84}\cdot\frac{m^2\alpha^2}{4k+2m}\right)$ is at most 1/100. Now furthermore, if $\varepsilon\cdot m^2\alpha^2/(672k) > \log(20)$, then for all $Z' < -\frac{1}{84}\cdot\frac{m^2\alpha^2}{4k+2m}$, with probability at least 0.95, the algorithm outputs the $p = q$. Combining the conditions, we obtain that there is a constant C_2 such that for $m = C_2\left(\frac{k^{2/3}}{\alpha^{4/3}} + \frac{k^{1/2}}{\alpha\sqrt{\varepsilon}}\right)$, with probability at least 0.9, the algorithm outputs the correct answer when the input distributions satisfy $p = q$. The case of $d_{TV}(p, q) > \alpha$ distribution is similar and is omitted.

Case 2: $\alpha^2 < \frac{1}{\sqrt{k}}$, or $\alpha^2\varepsilon < \frac{1}{k}$. In this case, we will prove a bound of $O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha^2\varepsilon}\right)$ on the sample complexity. We still consider the case when $p = q$. We first note that when $\alpha^2 < \frac{1}{\sqrt{k}}$, or $\alpha^2\varepsilon < \frac{1}{k}$, then either $\frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha^2\varepsilon} > k$. Hence we can assume that the sample complexity bound we aim for is at least $\Omega(k)$. So $\mathrm{Var}(Z(X_1^m, Y_1^m)) \le ck$ for constant c. By the Chebyshev's inequality,

$$\begin{aligned}\Pr\left(Z' > -\frac{1}{84}\cdot\frac{m^2\alpha^2}{4k+2m}\right) \leq& \Pr\left(Z(X_1^m,Y_1^m) - \mathbb{E}\left[Z(X_1^m,Y_1^m)\right] > \frac{1}{3}\cdot\frac{m^2\alpha^2}{4k+2m}\right)\\ \leq& \Pr\left(Z(X_1^m,Y_1^m) - \mathbb{E}\left[Z(X_1^m,Y_1^m)\right] > \frac{1}{3}\cdot\frac{m\alpha^2}{6}\right)\\ \leq& \Pr\left(Z(X_1^m,Y_1^m) - \mathbb{E}\left[Z(X_1^m,Y_1^m)\right] > (ck)^{1/2}\cdot\frac{m\alpha^2}{18c^{1/2}k^{1/2}}\right)\\ \leq& 144\cdot c\cdot\frac{k}{m^2\alpha^4}.\end{aligned}$$

Therefore, there is a C_1 such that if $m \geq C_1 k^{1/2}/\alpha^2$, then under $p = q$, $\Pr\left(Z' > -\frac{1}{84}\cdot\frac{m^2\alpha^2}{4k+2m}\right)$ is at most $1/100$. In this situation, if $\varepsilon\cdot m\alpha^2/504 > \log(20)$, then for all $Z' < -\frac{1}{84}\cdot\frac{m^2\alpha^2}{4k+2m}$, with probability at least 0.95, the algorithm outputs the $p = q$. Combining with the previous conditions, we obtain that there also exists a constant C_2 such that for $m = C_2\left(\frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha^2\varepsilon}\right)$, with probability at least 0.9, the algorithm outputs the correct answer when the input distribution is $p = q$. The case of $d_{TV}(p,q) > \alpha$ distribution is similar and is omitted.

E.2 Closeness Testing – Lower Bounds

To show the lower bound part of Theorem 4, we need the following simple result.

Lemma 18. $S(\mathtt{IT}, k, \alpha, \varepsilon) \leq S(\mathtt{CT}, k, \alpha, \varepsilon)$.

Proof. Suppose we want to test identity with respect to q. Given X_1^m from p, generate Y_1^m independent samples from q. If $p = q$, then the two samples are generated by the same distribution, and otherwise they are generated by distributions that are at least ε far in total variation. Therefore, we can simply return the output of an (k, α, ε)-closeness testing algorithm on X_1^m, and Y_1^m. □

By Lemma 18 we know that a lower bound for identity testing is also a lower bound on closeness testing.

We first consider the sparse case, when $\alpha^2 > \frac{1}{\sqrt{k}}$, and $\alpha^2\varepsilon > \frac{1}{k}$. In this case, we show that

$$S(\mathtt{CT}, k, \alpha, \varepsilon) = \Omega\left(\frac{k^{2/3}}{\alpha^{4/3}} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}}\right).$$

When $\alpha > \frac{1}{k^{1/4}}$, $\frac{k^{2/3}}{\alpha^{4/3}}$ is the dominating term in the sample complexity $S(\mathtt{CT}, k, \alpha) = \Theta\left(\frac{k^{2/3}}{\alpha^{4/3}} + \frac{\sqrt{k}}{\alpha^2}\right)$, giving us the first term. By Lemma 18 we know that a lower bound for identity testing is also a lower bound on closeness testing giving the second term, and the lower bound of Theorem 2 contains the second term as a summand.

In the dense case, when $\alpha^2 < \frac{1}{\sqrt{k}}$, or $\alpha^2\varepsilon < \frac{1}{k}$, we show that

$$S(\mathtt{CT}, k, \alpha, \varepsilon, \delta) = \Omega\left(\frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}} + \frac{1}{\alpha\varepsilon}\right).$$

In the dense case, using the non-private lower bounds of $\Omega\left(\frac{k^{2/3}}{\alpha^{4/3}} + \frac{\sqrt{k}}{\alpha^2}\right)$ along with the identity testing bound of sample complexity lower bounds of note that $\frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}} + \frac{1}{\alpha\varepsilon}$ gives a lower bound of $\Omega\left(\frac{k^{2/3}}{\alpha^{4/3}} + \frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}} + \frac{1}{\alpha\varepsilon}\right)$. However, in the dense case, it is easy to see that $\frac{k^{2/3}}{\alpha^{4/3}} = O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}}\right)$ giving us the bound.