Differentially Private Testing of Identity and Closeness of Discrete Distributions

Jayadev Acharya\textsuperscript{*} \hspace{1cm} Ziteng Sun\textsuperscript{*} \hspace{1cm} Huanyu Zhang\textsuperscript{*}
Cornell University \hspace{1cm} Cornell University \hspace{1cm} Cornell University
acharya@cornell.edu \hspace{1cm} zs335@cornell.edu \hspace{1cm} hz388@cornell.edu

July 18, 2017

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 $\varepsilon$-differential privacy. Our results improve over the best known algorithms for identity testing, and are the first results for differentially private closeness testing.

Our bounds are tight up to a constant factor whenever the number of samples is $O(k)$, a regime which has garnered much attention over the last decade since it allows property testing even when only a fraction of the domain is observed.

Our upper bounds are obtained by converting (and sometimes combining) the known non-private testing algorithms into differentially private algorithms. We propose a simple, yet general procedure based on coupling of distributions, to establish sample complexity lower bounds for differentially private algorithms.

\textsuperscript{*}The authors are listed in alphabetical order. This research is supported by the National Science Foundation NSF-CRII-1657471, and start-up grant from Cornell University
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 [NP33, LR06], 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 [BFR+00, BFF+01, GR00, Bat01, Pan08, AJOS13, AJOS14, CDVV14, DKN15, BV15, CDGR16, DK16, BC17], references therein, and [Can15] 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 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 [NS08], and individual information from census-related data [Swe02]. 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., [War65, Dal77, DN03, WZ10, DJW13, WJD12, CMS11]). While the motivation is clear, even a formal notion of privacy is not straightforward. We use differential privacy [DMNS06], a notion which rose from database and cryptography literature, and has emerged as one of the most popular privacy measures (See [DMNS06, Dwo08, WZ10, DRV10, BLR13, MT07, LMH+15, KOV17], references therein, and the recent book [DR14]). 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 to understand how the data requirement grows to ensure privacy, along with the same accuracy level. In this paper, we study the sample size requirements for differentially private discrete distribution testing.

1.1 Prior Work and Our Contributions

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 constraint \(\varepsilon\), we provide \(\varepsilon\)-differentially private algorithms for both these tasks. We also provide lower bounds that are optimal up to constant factors 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 in the upper bound argument is to ensure that the test statistic has a small sensitivity to ensure privacy. We also prove a general theorem based on a coupling argument that can be used to prove sample complexity lower bounds on differentially private algorithms, which can be of independent interest.
A summary of the results is presented in Table 1, which we now describe in detail.

1. **Binary Testing.** To study the problem of identity testing, we warm up with the binary testing problem, where \( k = 2 \). The sample complexity of this problem is \( \Theta\left(\frac{1}{\alpha^2} + \frac{1}{\alpha^2\epsilon}\right) \). The upper bound is extremely simple, and can be obtained by the Laplace mechanism [DR14], and the lower bound follows as an application of our general lower bound argument. The result is stated in Theorem 2.

2. **Reduction from identity to uniformity.** We reduce the problem of \( \epsilon-DP \) identity testing of distributions over \([k]\) to \( \epsilon-DP \) uniformity testing over distributions over \([6k]\). Such a reduction, without privacy constraints was shown in [GoI16], 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 4.

3. **Identity Testing.** It was recently shown that \( O\left(\frac{\sqrt{k}}{\alpha^2}\right) \) [Pan08, VV14, DKN15, ADK15] 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, we consider the statistic in [Pan08] for uniformity testing. This statistic has a small sensitivity, and gives an \( O\left(\sqrt{k}/\alpha^2\right) \) sample algorithm when \( \alpha^2 > 1/\sqrt{k} \), which is the range when \( O\left(\sqrt{k}/\alpha^2\right) = O(k) \). In this sparse regime, where \( m = O(k) \), we modify their test and make it \( \epsilon \)-differentially private and with sample complexity \( O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha^2\epsilon}\right) \). [CDK17] studied identity testing and obtained an algorithm with complexity \( O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{k\log k}{\alpha^2\epsilon} + \frac{(k\log k)^{1/3}}{\alpha^{2/3}\epsilon^{2/3}}\right) \). Our bounds improve their bounds significantly.

In the non-sparse setting we use our closeness testing algorithm discussed below to obtain an algorithm with complexity \( O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha^2\epsilon}\right) \). The results are stated in Theorem 3.

4. **Closeness Testing.** Closeness testing problem was proposed by [BFR+00], and optimal bound of \( \Theta\left(\max\left\{\frac{k^{2/3}}{\alpha^{4/3}}, \frac{k^{1/2}}{\alpha^2}\right\}\right) \) was shown in [CDVV14]. 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 \( \Theta\left(\frac{k^{2/3}}{\alpha^{4/3}} + \frac{\sqrt{k}}{\alpha^2\epsilon}\right) \), and in the dense regime, we obtain a bound of \( O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha^2\epsilon}\right) \). These results are stated in Theorem 5.

5. **Lower Bounds using Coupling.** We propose a general method to prove lower bounds for distribution testing problems. We show that if there exists a coupling between a pair of distributions with the expected hamming distance bounded by \( D \) between the coupled samples, then if \( D = o\left(1/\epsilon\right) \), there is no \( \epsilon \)-differentially private algorithm to distinguish between the two distributions. This result is stated in Theorem 1.

We construct coupling between binary distributions, and apply Theorem 1 to obtain a lower bound of \( \Omega\left(\frac{1}{\alpha^2} + \frac{1}{\alpha^2\epsilon}\right) \) samples for binary testing problem, which is tight up to a constant factor. For uniformity testing, we design a coupling between the distribution \( u[k] \), and the mixture of distributions used in Paninski [Pan08]. We show that the expected hamming distance of the coupling is at most \( O\left(\frac{m^2\alpha^2}{k}\right) \) giving a lower bound of \( \Omega\left(\frac{\sqrt{k}}{\alpha^2}\right) \). Combining with the lower bound of \( \Omega\left(\sqrt{k}/\alpha^2\right) \) without privacy constraints, we obtain the lower bound in Theorem 3, which is also tight in sparse sample regime.

Since closeness testing is a harder problem than identity testing, all the lower bounds port over to closeness testing. The closeness testing lower bounds are given in Theorem 5.
| Problem | Sample Complexity Bounds |
|---------|--------------------------|
| Is $p = B(\frac{1}{2})$? | Non-private algorithms $\Theta\left(\frac{1}{\alpha^2}\right)$ [folklore]  
$\varepsilon$-DP algorithms $\Theta\left(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)$ [DR14], and Theorem 2 |
| Identity Testing | Non-private algorithms $\Theta\left(\frac{\sqrt{k}}{\alpha^2}\right)$ [Pan08]  
$\varepsilon$-DP algorithms $O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k\log k}}{\alpha\sqrt{\varepsilon}}\right)$ [CDK17]  
**IF** $\alpha^2 = \Omega\left(\frac{1}{\sqrt{k}}\right)$ and $\alpha^2\varepsilon = \Omega\left(\frac{1}{k}\right)$  
$S(\text{IT}, k, \alpha, \varepsilon) = \Theta\left(\frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}}\right)$ [Theorem 3]  
**ELSE**  
$\Omega\left(\frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}} + \frac{1}{\alpha\varepsilon}\right) \leq S(\text{IT}, k, \alpha, \varepsilon) \leq O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)$ [Theorem 3] |
| Closeness Testing | Non-private algorithms $\Theta\left(\frac{k^{2/3}}{\alpha^{4/3}} + \frac{k^{1/2}}{\alpha^2}\right)$ [CDVV14]  
$\varepsilon$-DP algorithms  
**IF** $\alpha^2 = \Omega\left(\frac{1}{\sqrt{k}}\right)$ and $\alpha^2\varepsilon = \Omega\left(\frac{1}{k}\right)$  
$S(\text{CT}, k, \alpha, \varepsilon) = \Theta\left(\frac{k^{2/3}}{\alpha^{4/3}} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}}\right)$ [Theorem 5]  
**ELSE**  
$\Omega\left(\frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}} + \frac{1}{\alpha\varepsilon}\right) \leq S(\text{CT}, k, \alpha, \varepsilon) \leq O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)$ [Theorem 5] |

Table 1: Summary of the sample complexity bounds for the problems of identity testing, and closeness testing of discrete distributions.

### 1.2 Other Related Works

A number of papers have recently studied hypothesis testing problems under differential privacy guarantees [WLK15, GLRV16, RK17]. Some works analyze the distribution of the test statistic in the asymptotic regime. The work most closely related to ours is in [CDK17], 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 work analyzing various statistical testing and estimation tasks under the notion of local differential privacy [WJD12, DJW13, EPK14, PG16, KBR16, YB17]. These papers study some of the most basic statistical problems and also provide minimax lower bounds using Fano’s inequality. [DHS15] study structured distribution estimation under differential privacy.

Information theoretic approaches to data privacy have also been studied in the past few years using quantities like mutual information, and guessing probability to quantify privacy [Mir12, SRP13, CY16, WYZ16, IW17].
1.3 Organization of the paper

The paper is organized as follows. In Section 2, we provide the problem definitions, and notations. A general technique for proving lower bounds for differentially private algorithms is described in Section 3. In Section 4, we study differentially private binary hypothesis testing as a warm-up. Section 5 gives upper and lower bounds for identity testing, and closeness testing is studied in Section 6.

2 Preliminaries.

We consider discrete distributions over a domain of size $k$, which we assume without loss of generality to be $[k] \overset{\text{def}}{=} \{1, \ldots, k\}$. We denote length-$m$ samples $X_1, \ldots, X_m$ by $X^m$. For $x \in [k]$, let $p_x$ be the probability of $x$ under $p$. Let $M_x(X^m)$ be the number of times $x$ appears in $X^m$. 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 $b$.

**Definition 1.** The total variation distance between distributions $p$, and $q$ over a discrete set $[k]$ is

$$d_{TV}(p,q) = \sup_{A \subseteq [k]} p(A) - q(A) = \frac{1}{2} \| p - q \|_1.$$  

**Definition 2.** Let $p$, and $q$ be distributions over $X$, and $Y$ respectively. A coupling between $p$ and $q$ is a distribution over $X \times Y$ whose marginals are $p$ and $q$ respectively.

**Definition 3.** The Hamming distance between two sequences $X^m_1$ and $Y^m_1$, is

$$d(X^m_1, Y^m_1) \overset{\text{def}}{=} \sum_{i=1}^m I\{X_i \neq Y_i\},$$

the number of positions where $X^m_1$, and $Y^m_1$ differ.

$(\varepsilon, \delta)$-differential privacy is defined below.

**Definition 4.** A randomized algorithm $A$ on a set $X^m \rightarrow S$ is said to $(\varepsilon, \delta)$-differentially private if for any $S \subseteq \text{range}(A)$, and all pairs of $X^m_1$, and $Y^m_1$ with $d(X^m_1, Y^m_1) \leq 1$

$$\Pr(A(X^m_1) \in S) \leq e^{\varepsilon} \cdot \Pr(A(Y^m_1) \in S) + \delta.$$  

In this paper, we study the case when $\delta = 0$, called pure differential privacy. For simplicity, we denote pure differential privacy as $\varepsilon$-differential privacy ($\varepsilon$-DP). A notion that is often useful in establishing bounds for differential privacy is that of sensitivity, defined below.

**Definition 5.** The sensitivity of a function $f : [k]^m \rightarrow \mathbb{R}$ is

$$\Delta(f) \overset{\text{def}}{=} \max_{d(X^m_1, Y^m_1) \leq 1} |f(X^m_1) - f(Y^m_1)|.$$  

**Definition 6.** The sigmoid-function $\sigma : \mathbb{R} \rightarrow (0, 1)$ is

$$\sigma(x) \overset{\text{def}}{=} \frac{1}{1 + \exp(-x)} = \frac{\exp(x)}{1 + \exp(x)}.$$
We need the following result for the sigmoid function.

**Lemma 1.** For all $x, \gamma \in \mathbb{R}$, $\exp(-|\gamma|) \leq \frac{\sigma(x+\gamma)}{\sigma(x)} \leq \exp(|\gamma|)$.

**Proof.** Since $\sigma(x)$ is an increasing function, it suffice to assume that $\gamma > 0$. In this case,
\[
\frac{\sigma(x+\gamma)}{\sigma(x)} = \exp(\gamma) \cdot \frac{1 + \exp(x)}{1 + \exp(x+\gamma)} < \exp(\gamma).
\]

### Identity Testing (IT)

Given description of a probability distribution $q$ over $[k]$, parameters $\alpha$, and $\varepsilon$, and $m$ independent samples $X_1^m$ from an unknown distribution $p$. An algorithm $A$ is an $(k, \alpha)$ identity testing algorithm for $q$, if
- when $p = q$, $A$ outputs “$p = q$” with probability at least 0.9, and
- when $d_{TV}(p, q) \geq \alpha$, $A$ outputs ”$p \neq q$” with probability at least 0.9.

Furthermore, if $A$ is $\varepsilon$-differentially private, we say $A$ is an $(k, \alpha, \varepsilon)$-identity testing algorithm.

**Definition 7.** The sample complexity of DP-identity testing problem, denoted $S(IT, k, \alpha, \varepsilon)$, is the smallest $m$ for which there exists an $(k, \alpha, \varepsilon)$-identity testing algorithm $A$ that uses $m$ samples. When privacy is not a concern, we denote the sample complexity as $S(IT, k, \alpha)$. When $q = u[k]$, the problem reduces to uniformity testing, and the sample complexity is denoted $S(UT, k, \alpha, \varepsilon)$.

### Closeness Testing (CT)

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

Furthermore, if $A$ is $\varepsilon$-differentially private, we say $A$ is an $(k, \alpha, \varepsilon)$-closeness testing algorithm. Similar to identity testing, the sample complexity of closeness testing is defined below.

**Definition 8.** The sample complexity of an $(k, \alpha, \varepsilon)$-closeness testing problem, denoted $S(CT, k, \alpha, \varepsilon)$, is the least values of $m$ for which there exists an $(k, \alpha, \varepsilon)$-closeness testing algorithm $A$. When privacy is not a concern, we denote the sample complexity of closeness testing as $S(CT, k, \alpha)$.

## 3 Privacy Bounds Via Coupling

Recall that coupling between distributions $p$ and $q$ over $X$, and $Y$, is a distribution over $X \times Y$ whose marginal distributions are $p$ and $q$ (Definition 2). For simplicity, we treat coupling as a randomized function $f : X \to Y$ 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 ε-DP 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 notion, and will be used to prove sample complexity bounds of differentially private algorithms.

**Theorem 1.** Suppose there is a coupling between distributions $p$ and $q$ over $\mathcal{X}^m$, such that $\mathbb{E}[d(X_1^m,Y_1^m)] \leq D$. Then, any $\varepsilon$-differentially private algorithm that distinguishes $p$, and $q$ with error probability at most $1/10$ must satisfy $D = \Omega\left(\frac{1}{\varepsilon}\right)$.

**Proof.** Let $A : \mathcal{X}^m \rightarrow \{p,q\}$ be an $\varepsilon$-DP hypothesis testing algorithm with error probability at most $0.1$. Let $(X_1^m,Y_1^m)$ be distributed according to the coupling of the Theorem. Then, we have $X_1^m \sim p$, implying

$$\Pr(A(X_1^m) = p) \geq 0.9,$$

and $Y_1^m \sim q$, implying

$$\Pr(A(Y_1^m) = q) \geq 0.9.$$

Then,

$$\Pr(A(X_1^m) = p \cap A(Y_1^m) = q) \geq 0.9 + 0.9 - 1 = 0.8.$$

Applying Markov’s inequality to the random variable $d(X_1^m,Y_1^m)$,

$$\Pr(d(X_1^m,Y_1^m) > 10D) < \Pr(d(X_1^m,Y_1^m) > 10 \cdot \mathbb{E}[d(X_1^m,Y_1^m)]) < 0.1.$$

Therefore,

$$\Pr(A(X_1^m) = p \cap A(Y_1^m) = q \cap d(X_1^m,Y_1^m) < 10D) \geq 0.8 + 0.9 - 1 = 0.7. \quad (1)$$

The condition of differential privacy states that for any $X_1^m$, and $Y_1^m$,

$$e^{-\varepsilon \cdot d(X_1^m,Y_1^m)} \leq \frac{\Pr(A(X_1^m) = p)}{\Pr(A(Y_1^m) = p)} < e^{\varepsilon \cdot d(X_1^m,Y_1^m)}.$$

Consider one sequence pair $X_1^m$, and $Y_1^m$ that satisfies (1). Then, we know that $\Pr(A(X_1^m) = p) > 0.7$, and $\Pr(A(Y_1^m) = q) > 0.7$. By the condition of differential privacy,

$$0.3 \geq \Pr(A(Y_1^m) = p) \geq \Pr(A(X_1^m) = p) \cdot e^{-\varepsilon \cdot d(X_1^m,Y_1^m)} = 0.7 \cdot e^{-10\varepsilon D}.$$

Taking logarithm we obtain

$$D \geq \frac{\ln(7/3)}{10} \frac{1}{\varepsilon} = \Omega\left(\frac{1}{\varepsilon}\right),$$

completing the proof.}\]
4 Binary Identity Testing

We start with the simplest testing problem. Given \(b_0, \alpha > 0\), and \(\varepsilon > 0\), and sample \(X_1^m \in \{0,1\}^m\) to \(B(b)\), the goal is to distinguish between \(b = b_0\), and \(|b - b_0| \geq \alpha\).

Simple bias and variance arguments show that the sample complexity of this problem is \(\Theta(1/\alpha^2)\).

In this section, we study the sample complexity for \(\varepsilon\)-DP algorithms.

We will prove the following theorem. We note that the upper bound can be simply the well known Laplace mechanism in differential privacy. We add a \(\text{Lap}(1/\varepsilon)\) random variable to the number of 1’s in \(X_1^m\), and then threshold the output appropriately. The privacy is guaranteed by privacy guarantees of the Laplace mechanism. A small bias variance computation also gives the second term. For completeness, we provide a proof of the upper bound using our techniques in Section A. The lower bound is proved using the coupling defined in Example 1 with Theorem 1.

**Theorem 2.** Given \(b_0 \in [0,1]\), and \(\varepsilon > 0\), and \(\alpha > 0\). There is an \(\varepsilon\)-DP algorithm that takes \(O\left(\frac{1}{\alpha^2} + \frac{1}{\alpha \varepsilon}\right)\) samples from a distribution \(B(b)\) and distinguishes between \(b = b_0\), and \(|b - b_0| \geq \alpha\) with probability at least \(9/10\). Moreover, any algorithm for this task requires \(\Omega\left(\frac{1}{\alpha^2} + \frac{1}{\alpha \varepsilon}\right)\) samples.

4.1 Binary Testing Lower Bound via Coupling

Suppose \(b_0 = 0.5\). Then it is well known that at least \(\Omega(1/\alpha^2)\) samples are necessary to test whether \(b = b_0\), or \(|b - b_0| > \alpha\). We will prove the second term, namely a lower bound of \(\Omega\left(\frac{1}{\alpha \varepsilon}\right)\) using a coupling.

Consider the special case of Example 1 with \(b_2 = \frac{1}{2} + \alpha\), and \(b_1 = \frac{1}{2}\). Then, \(D = (b_2 - b_1)m = \alpha m\), and \(E[d(X_1^m, Y_1^m)] = \alpha m\). Applying Theorem 1, we know that any \(\varepsilon\)-DP algorithm must satisfy,

\[
E[d(X_1^m, Y_1^m)] \geq \Omega\left(\frac{1}{\varepsilon}\right),
\]

which implies that

\[
m \geq \Omega\left(\frac{1}{\alpha \varepsilon}\right).
\]

5 Identity Testing

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

**Theorem 3.** If \(\alpha > 1/k^{1/4}\), and \(\varepsilon \alpha^2 > 1/k\),

\[
S(\text{IT}, k, \alpha, \varepsilon) = O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha \sqrt{\varepsilon}}\right),
\]

otherwise,

\[
S(\text{IT}, k, \alpha, \varepsilon) = O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha^2 \varepsilon}\right).
\]

Moreover, for all values of \(k, \alpha, \varepsilon\),

\[
S(\text{IT}, k, \alpha, \varepsilon) = \Omega\left(\frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha \sqrt{\varepsilon}} + \frac{1}{\alpha \varepsilon}\right).
\]
This theorem shows that in the sparse regime, when \( m = O(k) \), our bounds are tight up to constant factors in all parameters.

In Theorem 4 we will show a reduction from identity to uniformity testing. Using this, it will be enough to design algorithms for uniformity testing, which is done in Section 5.2 where we will prove the upper bound. 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 5.3.

### 5.1 Uniformity Testing implies Identity Testing

The sample complexity of testing identity of any distribution is \( O(\sqrt{k\alpha^2}) \), a bound that is tight for the uniform distribution. Recently [Gol16] proposed a scheme to reduce the problem of testing identity of distributions over \([k]\) for total variation distance \( \alpha \) to the problem of testing uniformity over \([6k]\) with total variation parameter \( \alpha/3 \). In other words, they show that \( S(\text{IT}, k, \alpha) \leq S(\text{UT}, 6k, \alpha/3) \).

Building up on their construction, we show that such a bound also holds for differentially private algorithms.

**Theorem 4.**

\[
S(\text{IT}, k, \alpha, \epsilon) \leq S(\text{UT}, 6k, \alpha/3, \epsilon).
\]

The theorem is proved in Section B.

### 5.2 Identity Testing – Upper Bounds

By Theorem 4, any upper bound on uniformity testing is a bound on identity testing. To obtain differentially private algorithms, we need test statistic with small sensitivity. In the sparse regime, when \( m = O(k) \), [Pan08] gave such a statistic. However, in the dense regime, when \( m = \Omega(k) \), the tight upper bounds based on chi-squared, or \( \ell_2 \) tests could have a high sensitivity. A simple argument (Lemma 2) shows that any algorithm for closeness testing can be used for identity testing. We also prove that the test statistic used for proving optimal closeness testing bounds have a small sensitivity. Moreover, in the dense regime, we will show that the sample complexity of closeness testing is equal to that of identity testing.

With these arguments, we propose Algorithm 1 for testing uniformity.

**Algorithm 1 Uniformity Testing**

**Input:** \( \epsilon, \alpha, k \), i.i.d. samples \( X^m \) from \( p \) over \([k]\)

1: if \( \alpha^2 > 1/\sqrt{k} \), and \( \alpha^2 \epsilon > 1/k \) then
2: Run Algorithm 2 (sparse uniformity testing) on \( \epsilon, \alpha \), and \( X^m \)
3: else
4: Generate \( Y^m \), i.i.d. from \( u[k] \)
5: Run Algorithm 3 (closeness testing) on \( \epsilon, \alpha \), and \( X^m \), and \( Y^m \)
6: end if

We now show that Algorithm 1 achieves the upper bounds of Theorem 3.

#### 5.2.1 Upper Bounds in the sparse regime.

Consider the case when \( \alpha > 1/k^{1/4} \), and \( \alpha^2 \epsilon > 1/k \). We will prove an upper bound of \( O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha \sqrt{\epsilon}}\right) \).

Note that in this case, Algorithm 1 calls Algorithm 2.
Recall that $M_x(X_1^m)$ is the number of appearances of $x$ in $X_1^m$, and let

$$\Phi_j(X_1^m) = \{x : M_x(X_1^m) = j\}$$

be the number of symbols appearing $j$ times in $X_1^m$. [Pan08] used $\Phi_1(X_1^m)$ as the statistic. For $X_1^m \sim u[k],$

$$\mathbb{E} [\Phi_1(X_1^m)] = k \cdot \left( m \cdot \left( \frac{1}{k} \right)^{m-1} \right) = m \cdot \left( \frac{1}{k} \right)^{m-1}. \quad (2)$$

For $X_1^m$ generated from a distribution $p$ with $d_{TV}(p, u) \geq \alpha$, [Pan08, Lemma 1] showed that:

$$\mathbb{E} [\Phi_1(X_1^m)] \leq m \cdot \left( \frac{1}{k} \right)^{m-1} - \frac{m^2 \alpha^2}{k}. \quad (3)$$

They also showed that

$$\text{Var}(\Phi_1(X_1^m)) = O\left( \frac{m^2}{k} \right), \quad (4)$$

and used Chebychev’s inequality to obtain the sample complexity upper bound of $O\left( \sqrt{k/\varepsilon^2} \right)$ without privacy constraints. We modify their algorithm slightly to obtain a differentially private algorithm. Let

$$Z(X_1^m) = m \cdot \left( \frac{1}{k} \right)^{m-1} - \Phi_1(X_1^m) - \frac{m^2 \alpha^2}{2k} \quad (5)$$

Suppose $X_1^m \sim p$, then

$$\mathbb{E} [Z(X_1^m)] = -\frac{m^2 \alpha^2}{2k}, \text{ if } p = u[k],$$

and $X_1^m \sim p,$

$$\mathbb{E} [Z(X_1^m)] \geq \frac{m^2 \alpha^2}{2k}, \text{ if } d_{TV}(p, u[k]) \geq \alpha.$$

**Algorithm 2 Uniformity testing in the sparse sample regime**

**Input:** $\varepsilon, \alpha$, i.i.d. samples $X_1^m$ from $p$

1: Let $Z$ be the value of the statistic in (5).
2: Generate $Y \sim B(\sigma(\varepsilon \cdot Z))$, $\sigma$ is the sigmoid function.
3: if $Y = 0$ then
4: return $p = u[k]$
5: end if
6: return $p \neq u[k] > \alpha$

We first prove the privacy bound. If we change one symbol in $X_1^m$, $\Phi_1(X_1^m)$ can change by at most 2, and therefore $\varepsilon \cdot Z$ changes by at most $2\varepsilon$. Invoking Lemma 1, the probability of any output changes by a multiplicative $\exp(2\varepsilon)$, and the algorithm is $2\varepsilon$-differentially private.
The error probability proof is along the lines of binary testing. We first consider when \( p = u[k] \). Using (4) let \( \text{Var}(\Phi_1(X^m_1)) \leq cm^2/k \) for a constant \( c \). By the Chebychev’s inequality

\[
\Pr \left( Z(X^m_1) > -\frac{m^2\alpha^2}{6k} \right) \leq \Pr \left( \mathbb{E}[\Phi_1(X^m_1)] - \Phi_1(X^m_1) > \frac{m^2\alpha^2}{3k} \right) \\
\leq \Pr \left( \mathbb{E}[\Phi_1(X^m_1)] - \Phi_1(X^m_1) > \frac{cm\alpha^2}{\sqrt{k}} \cdot \frac{m^2\alpha^2}{3c\sqrt{k}} \right) \\
\leq 9c^2 \cdot \frac{k}{m^2\alpha^4}.
\] (6)

Therefore, there is a \( C_1 \) such that if \( m \geq C_1 \sqrt{k}/\alpha^2 \), then under the uniform distribution \( \Pr \left( Z(X^m_1) > -\frac{m^2\alpha^2}{6k} \right) \) is at most \( 1/100 \). Now furthermore, if \( \varepsilon \cdot m^2\alpha^2/6k > \log(25) \), then for all \( X^m_1 \) with \( Z(X^m_1) < -\frac{m^2\alpha^2}{6k} \) with probability at least 0.95, the algorithm outputs the uniform distribution. Combining the conditions, we obtain that there is a constant \( C_2 \) such that for \( m = C_2 \left( \frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}} \right) \), with probability at least 0.9, the algorithm outputs uniform distribution when the input distribution is indeed uniform. The case of non-uniform distribution is similar since the variance bounds hold for both the cases, and is omitted.

5.2.2 Upper Bound in the dense regime.

We now consider the case when \( \alpha < 1/k^{1/4} \), or \( \alpha^2\varepsilon < 1/k \). In this case, the Algorithm 1 calls Algorithm 3. We first prove the following simple result.

Lemma 2. \( S(\text{IT}, k, \alpha, \varepsilon) \leq S(\text{CT}, k, \alpha, \varepsilon) \).

Proof. Suppose we want to test identity with respect to \( q \). Given \( X^m_1 \) from \( p \), generate \( Y^m_1 \) 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 \( \varepsilon \) far in total variation. Therefore, we can simply return the output of an \( (k, \alpha, \varepsilon) \)-closeness testing algorithm on \( X^m_1 \), and \( Y^m_1 \). \( \square \)

In Section 6.1, we will prove that when \( \alpha < 1/k^{1/4} \), or \( \alpha^2\varepsilon < 1/k \), the complexity of \( (k, \alpha, \varepsilon) \)-closeness testing is \( O \left( \frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha\varepsilon} \right) \), proving our upper bound.

5.3 Sample Complexity Lower bounds for Uniformity Testing

We will show that for any value of \( k, \alpha, \varepsilon \),

\[
S(\text{IT}, k, \alpha, \varepsilon) = \Omega \left( \frac{\sqrt{k}}{\alpha^2} + \frac{\sqrt{k}}{\alpha\sqrt{\varepsilon}} + \frac{1}{\alpha\varepsilon} \right).
\]

The first term is the lower bound without privacy constraints, proved in [Pan08]. The final term is a lower bound for testing identity to \( B(0.5) \), which is easier than uniformity testing by Theorem 4. We need to prove the middle term.

To this end, we invoke LeCam’s two point theorem. We will consider a collection of \( 2^{k/2} \) distributions proposed by [Pan08], each of which has a total variation distance of \( \alpha \) from \( u[k] \). Consider a distribution \( Q_1 \) over \([k]^m\) that first picks one of the \( 2^{m/2} \) distributions uniformly at
random and then generates \( m \) independent samples from it. Let \( Q_2 \) denote the distribution of \( m \) independent samples from the uniform distribution. Then by LeCam’s theorem, a lower bound on \( m \) to distinguish \( Q_1 \), and \( Q_2 \) is a lower bound on \( S(\text{IT}, k, \alpha, \varepsilon) \). To obtain a sample complexity lower bound, we will design a coupling between \( Q_1 \), and \( Q_2 \), and bound its expected Hamming distance. This is proved in the following lemma.

**Lemma 3.** There is a coupling between \( X_{m_1} \) generated by \( Q_2 \), and \( Y_{m_1} \) by \( Q_1 \) such that

\[
E[d(X_{m_1}, Y_{m_1})] \leq 8 \frac{m^2 \alpha^2}{k}.
\]

The construction of \( Q_2 \), and the proof of Lemma 3 is slightly technical and is given in Section C. Assuming this lemma, by Theorem 1, we need \( E[d(X_{m_1}, Y_{m_1})] = \Omega(\frac{1}{\varepsilon}) \) to distinguish between \( Q_1 \), and \( Q_2 \). For this to hold, we need \( \frac{m^2 \alpha^2}{k} = \Omega(\frac{1}{\varepsilon}) \), which gives us \( m \geq \Omega(\frac{\sqrt{\varepsilon}}{\alpha \sqrt{\varepsilon}}) \).

### 6 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 5.** If \( \alpha > 1/k^{1/4} \), and \( \varepsilon \alpha^2 > 1/k \),

\[
S(\text{CT}, k, \alpha, \varepsilon) = \Theta\left(\frac{k^{2/3}}{\alpha^{1/3}} + \frac{k^{1/2}}{\alpha \sqrt{\varepsilon}}\right),
\]

otherwise,

\[
S(\text{CT}, k, \alpha, \varepsilon) = O\left(\frac{k^{1/2}}{\alpha^2} + \frac{1}{\alpha \varepsilon}\right).
\]

This theorem shows that in the sparse regime, when \( m = O(k) \), our bounds are tight up to constant factors in all parameters.

#### 6.1 Closeness Testing – Upper Bounds

To prove the upper bounds, we privatize the closeness testing algorithm of [CDVV14]. To reduce the strain on the readers, we drop the sequence notations explicitly and let

\[
\mu_i^\text{def} = M_i(X_{m_1}), \text{ and } \nu_i^\text{def} = M_i(Y_{m_1}).
\]

Variants of the chi-squared test have been used to test closeness of distributions in the recent years [ADJ+12, CDVV14]. In particular, the statistic used by [CDVV14] is

\[
Z(X_{m_1}, Y_{m_1})^\text{def} = \sum_{i \in [k]} \frac{(\mu_i - \nu_i)^2 - \mu_i - \nu_i}{\mu_i + \nu_i},
\]
where we assume that 

\[ ((\mu_i - \nu_i)^2 - \mu_i - \nu_i)/(\mu_i + \nu_i) = 0, \text{ when } \mu_i + \nu_i = 0. \]

The results in [CDVV14] 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:

\[ \mathbb{E}[Z(X^m_1, Y^m_1)] = 0 \text{ when } p = q, \tag{7} \]
\[ \text{Var}(Z(X^m_1, Y^m_1)) \leq 2 \min\{k, m\} \text{ when } p = q, \tag{8} \]
\[ \mathbb{E}[Z(X^m_1, Y^m_1)] \geq \frac{m^2 \alpha^2}{4k + 2m} \text{ when } d_{TV}(p, q) \geq \alpha, \tag{9} \]
\[ \text{Var}(Z(X^m_1, Y^m_1)) \leq \frac{1}{1000} \mathbb{E}[Z(X^m_1, Y^m_1)]^2 \text{ when } p \neq q, \text{ and } m = \Omega\left(\frac{1}{\alpha^2}\right). \quad (10) \]

We use the same approach with the test statistic as with binary testing and uniformity testing to obtain a differentially private closeness testing method, described in Algorithm 3.

**Algorithm 3**

**Input:** \(\varepsilon, \alpha\), sample access to distribution \(p\) and \(q\)

1. \(Z' \leftarrow Z(X^m_1, Y^m_1) - \frac{1}{2} \left(\frac{m^2 \alpha^2}{4k + 2m}\right)\)
2. Generate \(Y \sim B(\sigma(\exp(\varepsilon \cdot Z')))\)
3. if \(Y = 0\) then
4. \(\text{return } p = q\)
5. else
6. \(\text{return } p \neq q\)
7. end if

We will show that Algorithm 3 satisfies sample complexity upper bounds described in theorem 5. We first bound the sensitivity (Definition 5) of the test statistic to prove privacy bounds.

**Lemma 4.** \(\Delta(Z(X^m_1, Y^m_1)) \leq 14\).

**Proof.** Since \(Z(X^m_1, Y^m_1)\) is symmetric, without loss of generality assume that one of the symbols is changed in \(Y^m_1\). This would cause at most two of the \(\nu_i\)'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 \(i\)th term of the statistic is

\[
\left|\frac{(\mu_i - \nu_i)^2 - (\mu_i - \nu_i + 1)^2}{\mu_i + \nu_i - 1}\right| = \left|\frac{(\mu_i + \nu_i)(2\mu_i - 2\nu_i + 1) + (\mu_i - \nu_i)^2}{(\mu_i + \nu_i)(\mu_i + \nu_i - 1)}\right|
\leq \frac{2\mu_i - 2\nu_i + 1}{\mu_i + \nu_i - 1} + \frac{\mu_i - \nu_i}{\mu_i + \nu_i - 1}
\leq \frac{3|\mu_i - \nu_i| + 1}{\mu_i + \nu_i - 1}
\leq 3 \frac{4}{\mu_i + \nu_i - 1} \leq 7. \quad (11)
\]

When \(\mu_i + \nu_i = 1\), the change can again be bounded by 7. Since at most two of the \(\nu_i\)'s change, we obtain the desired bound.
Since the sensitivity of the statistic is at most 14, the input to the sigmoid changes by at most 14ε when any input sample is changed. Invoking Lemma 1, the probability of any output changes by a multiplicative $\exp(14\varepsilon)$, and the algorithm is 14ε-differentially private.

We now prove the correctness of the algorithm:

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

In this case, $\frac{k^{2/3}}{\alpha \sqrt{\varepsilon}} + \frac{k^{1/2}}{\alpha \sqrt{\varepsilon}} \leq 2k$.

We consider the case when $p = q$, then $\text{Var}(Z(X^m, Y^m)) \leq 2 \min\{k, m\}$. Let $\text{Var}(Z(X^m, Y^m)) \leq cm$ for some constant $c$. By the Chebychev’s inequality,

$$\Pr\left(Z' > -\frac{1}{6} \cdot \frac{m^2 \alpha^2}{4k + 2m}\right) \leq \Pr\left(Z(X^m, Y^m) - \mathbb{E}[Z(X^m, Y^m)] > \frac{1}{3} \cdot \frac{m^2 \alpha^2}{4k + 2m}\right) \leq \Pr\left(Z(X^m, Y^m) - \mathbb{E}[Z(X^m, Y^m)] > \frac{1}{3} \cdot \frac{m^2 \alpha^2}{8k}\right) \leq \Pr\left(Z(X^m, Y^m) - \mathbb{E}[Z(X^m, Y^m)] > cm^{1/2} \cdot \frac{m^{3/2} \alpha^2}{24ck}\right) \leq 256c^2 \cdot \frac{k^2}{m^3 \alpha^4},$$

(12)

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

Therefore, there is a $C_1$ such that if $m \geq C_1 k^{2/3} / \alpha^{4/3}$, then under $p = q$, $\Pr\left(Z' > -\frac{1}{6} \cdot \frac{m^2 \alpha^2}{4k + 2m}\right)$ is at most 1/100. Now furthermore, if $\varepsilon \cdot m^2 \alpha^2 / 36k > \log(20)$, then for all $Z' < -\frac{1}{6} \cdot \frac{m^2 \alpha^2}{4k + 2m}$, 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 $m = C_2 \left(\frac{k^{2/3}}{\alpha \sqrt{\varepsilon}} + \frac{k^{1/2}}{\alpha \sqrt{\varepsilon}}\right)$, 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 $O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha \sqrt{\varepsilon}}\right)$ on the sample complexity. We still consider the case when $p = q$. We first note that when $\varepsilon \cdot m^2 \alpha^2 / 36 > \log(20)$, then either $\sqrt{k} / \alpha^2 + \frac{1}{\alpha \sqrt{\varepsilon}} > k$. Hence we can assume that the sample complexity bound we aim for is at least $\Omega(k)$. So $\text{Var}(Z(X^m, Y^m)) \leq ck$ for constant $c$. By the Chebychev’s inequality,

$$\Pr\left(Z' > -\frac{1}{6} \cdot \frac{m^2 \alpha^2}{4k + 2m}\right) \leq \Pr\left(Z(X^m, Y^m) - \mathbb{E}[Z(X^m, Y^m)] > \frac{1}{3} \cdot \frac{m^2 \alpha^2}{4k + 2m}\right) \leq \Pr\left(Z(X^m, Y^m) - \mathbb{E}[Z(X^m, Y^m)] > \frac{1}{3} \cdot \frac{m^2 \alpha^2}{6}\right) \leq \Pr\left(Z(X^m, Y^m) - \mathbb{E}[Z(X^m, Y^m)] > ck^{1/2} \cdot \frac{m^2 \alpha^2}{18ck^{1/2}}\right) \leq 144c^2 \cdot \frac{k}{m^2 \alpha^4}.\quad (13)$$

Therefore, there is a $C_1$ such that if $m \geq C_1 k^{1/2} / \alpha^2$, then under $p = q$, $\Pr\left(Z' > -\frac{1}{6} \cdot \frac{m^2 \alpha^2}{4k + 2m}\right)$ is at most 1/100. In this situation, if $\varepsilon \cdot m^2 \alpha^2 / 36 > \log(20)$, then for all $Z' < -\frac{1}{6} \cdot \frac{m^2 \alpha^2}{4k + 2m}$, 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 $m = C_2 \left( \frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha \varepsilon} \right)$, 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.

### 6.2 Closeness Testing – Lower Bounds

To show the lower bound part of Theorem 5, we use Lemma 2 in Section 5.2.2.

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

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

When $\alpha > \frac{1}{k^{1/4}}$, $\frac{k^{2/3}}{\alpha^{4/3}}$ is the dominating term in the sample complexity $S(\text{CT}, k, \alpha) = \Theta \left( \frac{k^{2/3}}{\alpha^{4/3}} + \frac{\sqrt{k}}{\alpha^2} \right)$, giving us the first term. By Lemma 2 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 3 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

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

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.

### References

[ADJ+12] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, and Ananda Theertha Suresh. Competitive classification and closeness testing. In COLT, 2012.

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam C Kamath. Optimal testing for properties of distributions. In NIPS, 2015.

[AJOS13] 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 (AISTATS), 2013.

[AJOS14] 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 (ISIT), 2014.

[Bat01] Tugkan Batu. Testing properties of distributions. PhD thesis, Cornell University, 2001.
[BC17] Tugkan Batu and Clément L. Canonne. Generalized uniformity testing. In FOCS, 2017.

[BFF+01] Tugkan Batu, Lance Fortnow, Eldar Fischer, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In FOCS, pages 442–451, 2001.

[BFR+00] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In FOCS, pages 259–269, 2000.

[BLR13] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to noninteractive database privacy. Journal of the ACM (JACM), 60(2):12, 2013.

[BV15] Bhaswar Bhattacharya and Gregory Valiant. Testing closeness with unequal sized samples. In Advances in Neural Information Processing Systems, pages 2611–2619, 2015.

[Can15] Clément L. Canonne. A survey on distribution testing: Your data is big, but is it blue? Electronic Colloquium on Computational Complexity (ECCC), 22:63, 2015.

[CDGR16] Clément L Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. In 33rd Symposium on Theoretical Aspects of Computer Science, 2016.

[CDK17] Bryan Cai, Constantinos Daskalakis, and Gautam Kamath. Priv’it: Private and sample efficient identity testing. In ICML, 2017.

[CDVV14] Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In SODA, pages 1193–1203, 2014.

[CMS11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. Journal of Machine Learning Research, 12(Mar):1069–1109, 2011.

[CY16] Paul Cuff and Lanqing Yu. Differential privacy as a mutual information constraint. In ACM SIGSAC Conference on Computer and Communications Security, pages 43–54, 2016.

[Dal77] Tore Dalenius. Towards a methodology for statistical disclosure control. Statistisk Tidskrift, 15:429–444, 1977.

[DHS15] Ilias Diakonikolas, Moritz Hardt, and Ludwig Schmidt. Differentially private learning of structured discrete distributions. In NIPS, pages 2566–2574, 2015.

[DJW13] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In FOCS, pages 429–438. IEEE, 2013.

[DK16] Ilias Diakonikolas and Daniel M Kane. A new approach for testing properties of discrete distributions. arXiv preprint arXiv:1601.05557, 2016.

[DKN15] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In SODA, pages 1841–1854, 2015.
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.

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.

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

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In FOCS, pages 51–60, 2010.

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.

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. ACM, 2014.

Marco Gaboardi, Hyun Woo Lim, Ryan Rogers, and Salil P Vadhan. Differentially private chi-squared hypothesis testing: goodness of fit and independence testing. In ICML, pages 2111–2120, 2016.

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.

Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. Electronic Colloquium on Computational Complexity (ECCC), 7(20), 2000.

Ibrahim Issa and Aaron B. Wagner. Operational definitions for some common information leakage metrics. In ISIT, 2017.

Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. arXiv preprint arXiv:1602.07387, 2016.

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

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.

Erich Lehmann and Joseph Romano. Testing statistical hypotheses. Springer Science & Business Media, 2006.

Darakhshan J Mir. Information-theoretic foundations of differential privacy. In International Symposium on Foundations and Practice of Security, pages 374–381, 2012.
[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In FOCS, pages 94–103. IEEE, 2007.

[NP33] J. Neyman and E. S. 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.

[NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In Security and Privacy.IEEE Symposium on, pages 111–125, 2008.

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

[PG16] Adriano Pastore and Michael Gastpar. Locally differentially-private distribution estimation. In Information Theory (ISIT), IEEE International Symposium on, pages 2694–2698, 2016.

[RK17] 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.

[SRP13] 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.

[Swe02] Latanya Sweeney. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05):557–570, 2002.

[VV14] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In FOCS, pages 51–60. IEEE, 2014.

[War65] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60(309):63–69, 1965.

[WJD12] Martin J Wainwright, Michael I Jordan, and John C Duchi. Privacy aware learning. In Advances in Neural Information Processing Systems, pages 1430–1438, 2012.

[WLK15] Yue Wang, Jaewoo Lee, and Daniel Kifer. Differentially private hypothesis testing, revisited. arXiv preprint arXiv:1511.03376, 2015.

[WYZ16] 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.

[WZ10] Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. Journal of the American Statistical Association, 105(489):375–389, 2010.

[YB17] Min Ye and Alexander Barg. Optimal schemes for discrete distribution estimation under locally differential privacy. CoRR, abs/1702.00610, 2017.
A Upper Bound for Binary Testing

Our \( \varepsilon \)-DP algorithm simply estimates the empirical bias, and decides if it close to \( b_0 \). Let \( M_1(X_1^m) \) be the number of one’s in \( X_1^m \). Then,

\[
\mathbb{E}[M_1(X_1^m)] = mb, \text{ and } \text{Var}(M_1(X_1^m)) = mb(1-b).
\]

We compute the deviation of \( M_1(X_1^m) \) from its expectation, and use it as our statistic:

\[
Z(X_1^m) = M_1(X_1^m) - mb.
\]

Algorithm 4 Binary Testing

Input: \( \varepsilon, \alpha, b_0 \) i.i.d. samples \( X_1^m \) from \( B(b) \)

Generate \( Y \sim B(\sigma(\varepsilon \cdot (|Z(X_1^m)| - \frac{am}{2}))) \)

if \( Y = 0 \)
return \( p = B(b_0) \)
else
return \( p \neq B(b_0) \)

Lemma 5. Algorithm A is an \( \varepsilon \)-differentially private algorithm for testing if a binary distribution is \( B(b_0) \). Moreover, it has error probability at most 0.1, with sample complexity \( O\left(\frac{1}{\alpha^2} + \frac{1}{\alpha \varepsilon}\right) \).

Proof. We first prove the correctness. Consider the case when \( b = b_0 \). By Chebychev’s inequality,

\[
\Pr\left(|Z(X_1^m)| \geq \beta \cdot \frac{\sqrt{m}}{2}\right) \leq \Pr\left(Z(X_1^m)^2 \geq \beta^2 mb(1-b)\right) \leq \frac{1}{\beta^2}.
\]

For \( \beta = 10 \), we have \( \Pr(|Z(X_1^m)| \geq \beta \cdot \sqrt{m}/2) < 1/100. \) Suppose \( m \) satisfies

\[
\varepsilon \frac{am - \beta \sqrt{m}}{2} > \log \frac{1}{0.02},
\]

then with probability at least 99/100, \( \varepsilon (\frac{am}{2} - |Z(X_1^m)|) > \log 50 \). Under this condition, the algorithm outputs \( p \neq B(b_0) \) with probability at most 1/20. Therefore, the total probability of error is at most .01+.05 <0.1. Note that there is a constant \( C \), such that for \( m \geq C\left(\frac{1}{\alpha^2} + \frac{1}{\alpha \varepsilon}\right) \), \( \varepsilon (\frac{am}{2} - |Z(X_1^m)|) \geq \log(20) \). The case when \( |b - b_0| > \alpha \) follows from similar arguments.

We now prove the privacy guarantee. When one of the samples is changed, \( Z(X_1^m) \) changes by at most one, and \( \varepsilon \cdot (|Z(X_1^m)| - \frac{am}{2}) \) changes by at most \( \varepsilon \). Invoking Lemma 1, the probability of any output changes by at most a multiplicative \( \exp(\varepsilon) \), and the algorithm is \( \varepsilon \)-differentially private.

B Proof of Theorem 4

Proof. We first briefly describe the essential components of the construction of [Gol16]. Given an explicit distribution \( q \) over \([k]\), there exists a randomized function \( F_q : [k] \rightarrow [6k] \) 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_{i}^{s}$ from a distribution $p$ over $[k]$. Apply $F_{q}$ independently to each of the $X_{i}$ to obtain a new sequence $Y_{i}^{s} = F_{q}(X_{i}^{s}) \overset{\text{def}}{=} F_{q}(X_{1}) \ldots F_{q}(X_{s})$. 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, $S(\mathcal{IT}, k, \alpha) \leq S(\mathcal{UT}, 6k, \alpha/3)$ (See [Gol16] for details).

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

$$\mathcal{A}(Y_{i}^{s} = u[6k]) \leq \mathcal{A}(Y_{i}^{s'} = u[6k]) \cdot e^{\varepsilon}. \tag{15}$$

Moreover, since $F_{q}$ is applied independently to each coordinate,

$$\Pr(F_{q}(X_{i}^{s}) = Y_{i}^{s}) = \Pr(F_{q}(X_{1}^{s-1}) = Y_{1}^{s-1}) \Pr(F_{q}(X_{s}) = Y_{s}).$$

Then,

$$\Pr(\mathcal{A}'(X_{i}^{s}) = q) = \Pr(\mathcal{A}(F_{q}(X_{i}^{s})) = u[6k]) = \sum_{Y_{i}^{s}} \Pr(\mathcal{A}(Y_{i}^{s}) = u[6k]) \Pr(F_{q}(X_{i}^{s}) = Y_{i}^{s})$$

$$= \sum_{Y_{1}^{s-1}} \sum_{Y_{s} \in [k]} \Pr(\mathcal{A}(Y_{i}^{s}) = u[6k]) \Pr(F_{q}(X_{1}^{s-1}) = Y_{1}^{s-1}) \Pr(F_{q}(X_{s}) = Y_{s})$$

$$= \sum_{Y_{1}^{s-1}} \Pr(F_{q}(X_{1}^{s-1}) = Y_{1}^{s-1}) \left[ \sum_{Y_{s} \in [k]} \Pr(\mathcal{A}(Y_{i}^{s}) = u[6k]) \Pr(F_{q}(X_{s}) = Y_{s}) \right]. \tag{16}$$

Similarly,

$$\Pr(\mathcal{A}'(X_{i}^{s}) = q) = \sum_{Y_{i}^{s-1}} \Pr(F_{q}(X_{1}^{s-1}) = Y_{1}^{s-1}) \left[ \sum_{Y_{s} \in [k]} \Pr(\mathcal{A}(Y_{i}^{s}) = u[6k]) \Pr(F_{q}(X_{s}) = Y_{s}') \right]. \tag{17}$$

For a fixed $Y_{1}^{s-1}$, the term within the bracket in (16), and (17) are both expectations over the final coordinate. However, by (15) these expectations differ at most by a multiplicative $e^{\varepsilon}$ factor. This implies that

$$\Pr(\mathcal{A}'(X_{i}^{s}) = q) \leq \Pr(\mathcal{A}'(X_{i}^{s}) = q) e^{\varepsilon}. \tag{18}$$

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 Lower Bound Instance, and Proof of Lemma 3

We first describe the $2^k/2$ distributions considered by [Pan08]. For $z \in \{\pm1\}^{k/2}$, define a distribution $P_z$ over $[k]$ such that

$$P_z(2i - 1) = \frac{1 + z_i \cdot 2\alpha}{k}, \text{ and } P_z(2i) = \frac{1 - z_i \cdot 2\alpha}{k}.$$  

Then for any $z$, $d_{TV}(P_z, u[k]) = \alpha$. Let $Q_1$ be the following distribution over $[k]^m$:

- Select $z \in \{\pm1\}^{k/2}$ uniformly at random.
- Output $m$ independent samples from $P_z$.

Let $Q_2$ be the distribution that generates $m$ i.i.d. samples from $u[k]$.

Proof of Lemma 3. We now prove that there is a coupling between $Q_1$, and $Q_2$ satisfying Lemma 3.

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 $P_1$ 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 $P_2$ 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 6. 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,

$$P_2(X_1^t) = \binom{t}{t_0} \frac{1}{2^t},$$

and

$$P_1(X_1^t) = \binom{t}{t_0} \frac{1}{2^t} \left( \frac{(1 - 2\alpha)^{t_0}(1 + 2\alpha)^{t_1} + (1 + 2\alpha)^{t_0}(1 - 2\alpha)^{t_1}}{2} \right).$$

The term in the parantheses above is minimised when either $t_0 = t_1 = t/2$. In this case,

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

Therefore,

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

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

20
Proof. The coupling is as follows. Note that \( \sum_{X_1} \min \{ P_1(X_1), P_2(X_1) \} = 1 - d_{TV}(P_1, P_2) \). Consider the following coupling:

1. Let \( R \) be a \( U[0, 1] \) random variable.
2. If \( R < 1 - d_{TV}(P_1, P_2) \) then output \((X_1', X_1')\) with probability \( \frac{\min \{ P_1(X_1'), P_2(X_1') \}}{1 - d_{TV}(P_1, P_2)} \).

Then \( \mathbb{E} [d(X_1', Y_1')] \leq t \cdot d_{TV}(P_1, P_2) = 2t^2 \alpha^2 \leq 4(t^2 - t)\alpha^2 \) when \( t \geq 2 \), and when \( t = 1 \), the distributions are identical and the Hamming distance of the coupling is equal to zero. \( \square \)

We now have the tools to prove Lemma 3.

Proof of Lemma 3. 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 \( Q_2 \):
   - 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 \( Q_1 \):
   - 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 \( \ell \) 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 7, after replacing \( \{ 0, 1 \} \) with \( \{ 2j - 1, 2j \} \).

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

\[
\mathbb{E} [d(X_1^m, Y_1^m)] = \sum_{j=1}^{k/2} \mathbb{E} [d(X_1^{T_j}, Y_1^{T_j})] \\
= \frac{k}{2} \mathbb{E} [d(X_1^R, Y_1^R)] \\
\leq \frac{k}{2} \cdot \mathbb{E} [4\alpha^2(R^2 - R)],
\]

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

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

Plugging this, we obtain

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

proving the claim. \( \square \)