The Annals of Statistics
2023, Vol. 51, No. 2, 547–572
https://doi.org/10.1214/23-AOS2259

CANONICAL NOISE DISTRIBUTIONS AND PRIVATE HYPOTHESIS TESTS

BY JORDAN AWAN[1,a] AND SALIL VADHAN[2,b]

[1]*Department of Statistics, Purdue University,* [a]*jawan@purdue.edu*
[2]*Harvard John A. Paulson School of Engineering and Applied Sciences,* [b]*salil_vadhan@harvard.edu*

f-DP has recently been proposed as a generalization of differential privacy allowing a lossless analysis of composition, post-processing, and privacy amplification via subsampling. In the setting of f-DP, we propose the concept of a *canonical noise distribution* (CND), the first mechanism designed for an arbitrary f-DP guarantee. The notion of CND captures whether an additive privacy mechanism perfectly matches the privacy guarantee of a given f. We prove that a CND always exists, and give a construction that produces a CND for any f. We show that private hypothesis tests are intimately related to CNDs, allowing for the release of private p-values at no additional privacy cost, as well as the construction of uniformly most powerful (UMP) tests for binary data, within the general f-DP framework.

We apply our techniques to the problem of difference-of-proportions testing, and construct a UMP unbiased (UMPU) "semiprivate" test which upper bounds the performance of any f-DP test. Using this as a benchmark, we propose a private test based on the inversion of characteristic functions, which allows for optimal inference on the two population parameters and is nearly as powerful as the semiprivate UMPU. When specialized to the case of $(\epsilon, 0)$-DP, we show empirically that our proposed test is more powerful than any $(\epsilon/\sqrt{2})$-DP test and has more accurate type I errors than the classic normal approximation test.

1. Introduction.

The concept of differential privacy (DP) was introduced in Dwork et al. (2006), which offers a framework for the construction of private mechanisms and a rigorous notion of what it means to limit privacy loss when performing statistical releases on sensitive data. DP requires that the randomized algorithm M performing the release has the property that for any two datasets X and X' which differ in one individual's data (*adjacent datasets*), the distributions of $M(X)$ and $M(X')$ are "close." Since this seminal paper, many variants of differential privacy have been proposed; the variants primarily differ in how they formulate the notion of closeness. For example, pure and approximate DP are phrased in terms of bounding the probabilities of sets of outputs, according to $M(X)$ versus $M(X')$ (Dwork and Roth (2014)), whereas concentrated (Bun and Steinke (2016)) and Rényi (Mironov (2017)) DP are based on bounding a divergence between $M(X)$ and $M(X')$.

Wasserman and Zhou (2010) and Kairouz, Oh and Viswanath (2017) showed that pure and approximate DP can be expressed as imposing constraints on the type I and type II errors of hypothesis tests which seek to discriminate between two adjacent databases. Recently Dong, Roth and Su (2022) expanded this view, defining f-DP which allows for an arbitrary bound to be placed on the receiver–operator curve (ROC) or *tradeoff* function when testing between two adjacent databases. It is shown in Dong, Roth and Su (2022) that f-DP retains many of the useful properties of DP such as post-processing, composition, and subsampling and allows for lossless calculation of the privacy cost of each of these operations. Furthermore, as special cases, f-DP contains both pure and approximate DP, as well as relatives

Received November 2021; revised November 2022.
MSC2020 subject classifications. Primary 68P27; secondary 62F03.
Key words and phrases. Differential privacy, uniformly most powerful test, frequentist inference.

of divergence-based notions of DP (e.g., Gaussian DP (GDP) is slightly stronger than zero-concentrated DP).

In this paper, we study two fundamental privacy questions in the framework of f-DP. The first is based on optimizing the basic mechanism of adding independent noise to a real-valued statistic, and the second is about constructing hypothesis tests under the constraint of DP. We show that in fact, the two problems are intricately related, where the "canonical additive noise distribution" enables private p-values "for free," and gives a closed-form construction of certain optimal hypothesis tests.

One of the simplest and widely used type of privacy mechanism is noise addition, where independent noise is added to a real-valued statistic. Additive mechanisms are not only widely used by themselves, but are also often a key ingredient to more complex mechanisms such as functional mechanism (Zhang et al. (2012)), objective perturbation (Chaudhuri, Monteleoni and Sarwate (2011)), stochastic gradient descent (Abadi et al. (2016)), and the sparse vector technique (Dwork et al. (2009)), to name a few. The oldest and most widely used additive mechanisms are the Laplace and Gaussian mechanisms, but there have since been many proposed distributions which satisfy different definitions of DP. A natural question is what noise distributions are "optimal" or "canonical" for a given definition of privacy. The geometric mechanism/discrete Laplace mechanism is optimal for ϵ-DP counts, in terms of maximizing Bayesian utility (Ghosh, Roughgarden and Sundararajan (2012)), the staircase mechanism is optimal for ϵ-DP in terms of ℓ_1 or ℓ_2-error (Geng and Viswanath (2015)), and the truncated-uniform-Laplace (Tulap) distribution generalizes both the discrete Laplace and staircase mechanisms and is optimal for (ϵ, δ)-DP in terms of generating uniformly most powerful (UMP) hypothesis tests and uniformly most accurate (UMA) confidence intervals for Bernoulli data (Awan and Slavković (2018), Awan and Slavković (2020)). With divergence-based definitions of privacy, Gaussian noise is argued to be canonical for (zero) concentrated DP (Bun and Steinke (2016)), and the sinh-normal distribution is argued to be canonical for truncated concentrated DP (Bun et al. (2018)).

In this paper, we give the first formal definition of a *canonical noise distribution* (CND) which captures the notion of whether a distribution tightly matches a privacy guarantee f-DP. We show that the Gaussian distribution is canonical for Gaussian differential privacy (GDP), and the Tulap distribution is canonical for (ϵ, δ)-DP. We prove that a CND always exists for any nontrivial symmetric tradeoff function f, and give a general construction to generate a CND given any tradeoff function f. This construction results in the first general mechanism for an arbitrary f-DP guarantee. In the special case of (ϵ, δ)-DP, our construction results in the Tulap distribution.

Another privacy question is on the nature of DP hypothesis tests. Awan and Slavković (2018) showed that for independent Bernoulli data, there exists uniformly most powerful (UMP) (ϵ, δ)-DP tests which are based on the Tulap distribution, enabling "free" private p-values, at no additional cost to privacy.

We show that for an arbitrary tradeoff function f and any f-DP test, a free private p-value can always be generated in terms of a CND for f. We also extend the main results of Awan and Slavković (2018) from (ϵ, δ)-DP to f-DP as well as from i.i.d. Bernoulli variables to exchangeable binary data. This extension shows that the CND is the proper generalization of the Tulap distribution, and gives an explicit construction of the most powerful f-DP test for binary data, the first DP hypothesis test for a general f-DP guarantee.

We end with an extensive application to private difference-of-proportions testing. Testing two population proportions is a common hypothesis testing setting that arises when there are two groups with binary responses, such as A/B testing, clinical trials, and observational studies. As such, the techniques for testing these hypotheses are standardized and included in most introductory statistics textbooks. However, there currently lacks a theoretically-based private test with accurate sensitivity and specificity. Karwa and Vadhan (2018) were the first to attempt at tackling the private difference-of-proportions testing problem, and recently Awan and Cai (2020) used a novel asymptotic method to calibrate the type I errors of a related DP test in large sample sizes. Our application builds off of these prior works, with a much improved analysis and strong theoretical basis in the f-DP framework.

We show that in general, there does not exist a UMP unbiased f-DP test for this problem, but using our earlier results on most powerful f-DP tests for binary data, we show that there does exist a UMP unbiased "semiprivate" test, which satisfies a weakened version of f-DP. While this test does not satisfy f-DP, it does provide an upper bound on the power of any f-DP test, and gives intuition on the structure of a good f-DP test for this problem. We then design a novel f-DP test for the testing problem, based on using CNDs and an expression of the sampling distribution in terms of characteristic functions, enabling efficient computation via Gil-Pelaez inversion. Using theory of the parametric bootstrap, we argue that the test is asymptotically unbiased and has asymptotically accurate type I errors. Empirically, we show that the test has more accurate type I errors and p-values than the popularly used normal approximation test, and that the power of our proposed test is nearly as powerful as the semiprivate UMP unbiased test. In the case of ϵ-DP, we demonstrate through simulations that our test has higher power than any $(\epsilon/\sqrt{2})$-DP test, indicating that it is nearly optimal. Furthermore, our test has the benefit of allowing for optimal hypothesis tests and confidence intervals for each of the population proportions, using the techniques of Awan and Slavković (2020), as the proposed test is based on the same DP summary statistics.

Organization. In Section 2 we set the notation for the paper and review background differential privacy. In Section 3, we introduce the concept of a canonical noise distribution, give some basic properties of CNDs, and provide a general construction of a CND for any f-DP privacy notion. In Section 4, we show that any f-DP hypothesis test must satisfy constraints based on the function f, we give a general result for "free" DP p-values given an f-DP test function, and develop most powerful f-DP tests for exchangeable binary data. In Section 5, we consider the problem of privately testing the difference of population proportions. Specifically in Section 5.1, we develop a uniformly most powerful unbiased "semiprivate" test, which gives an upper bound on the power of any f-DP test, in Section 5.2 we propose an f-DP test based on the inversion of characteristic functions, and in Section 5.3 we evaluate the type I error and power of our two sample tests in simulations. Proofs and technical details are deferred to the Supplementary Material (Awan and Vadhan (2023)).

Related work. Vu and Slavković (2009) was the first work in private hypothesis testing, developing DP tests for population proportions as well as independence tests for 2×2 tables. These tests use additive Laplace noise, and use a normal approximation to the sampling distribution to calibrate the type I errors. Solea (2014) develop tests for normally distributed data using similar techniques. Wang, Lee and Kifer (2015) and Gaboardi et al. (2016) expanded on Vu and Slavković (2009), developing additional tests for multinomials. Wang, Lee and Kifer (2015) developed asymptotic sampling distributions for their tests, verifying the type I errors via simulations, whereas Gaboardi et al. (2016) use Monte Carlo methods to estimate and control the type I error. Uhler, Slavković and Fienberg (2013) develop DP p-values for chi-squared tests of GWAS data, and derive the exact sampling distribution of the noisy statistic. Kifer and Rogers (2016) develop private χ^2 tests for goodness-of-fit and identity problems which are designed to have the same asymptotic properties as the nonprivate tests.

Under "local differential privacy," a notion of DP where even the data curator does not have access to the original dataset, Gaboardi and Rogers (2018) develop multinomial tests based on asymptotic distributions.

The first uniformly most powerful hypothesis tests under DP for the testing of i.i.d. Bernoulli data were developed by Awan and Slavković (2018). Their tests are based on the Tulap distribution, an extension of the discrete Laplace and Staircase mechanisms. Awan and Slavković (2020) expanded on these results to offer UMP unbiased two-sided DP tests as well as optimal DP confidence intervals and confidence distributions for Bernoulli data.

Given a DP output, Sheffet (2017) and Barrientos et al. (2019) develop significance tests for regression coefficients. Wang et al. (2018) develop general approximating distributions for DP statistics, which can be used to construct hypothesis tests and confidence intervals, but which are only applicable to limited models. Awan and Cai (2020) also provide asymptotic techniques that can be used to conduct approximate hypothesis tests, given DP summary statistics, but which may have limited accuracy in finite samples.

Rather than the classical regime of fixing the type I error, and minimizing the type II error, there are several works on DP testing, where the goal is to optimize the sample complexity required to generate a test which places both the type I and type II errors below a certain threshold. Canonne et al. (2019) show that for simple hypothesis tests, a noisy clamped likelihood ratio test achieves optimal sample complexity. Cai, Daskalakis and Kamath (2017) and Kakizaki, Fukuchi and Sakuma (2017) both study the problem of ϵ-DP discrete identity testing from the sampling complexity perspective. Aliakbarpour, Diakonikolas and Rubinfeld (2018) also studies ϵ-DP identitiy testing as well as DP equivalence testing. Acharya, Sun and Zhang (2018) study identity and closeness testing of discrete distributions in the (ϵ, δ)-DP framework. Bun et al. (2019) derive sample complexity bounds for differentially privacy hypothesis selection, where the goal is to choose among a set of potential data generating distributions, which one has the smallest total variation distance to the true distribution. Suresh (2021) develop an alternative to the Neyman–Pearson lemma for simple hypotheses, which is robust to misspecification of the hypotheses; due to the connection between robustness and differential privacy (Dwork and Lei (2009)), this could be a promising tool for developing private tests.

Outside the hypothesis testing setting, there is some additional work on optimal population inference under DP. Duchi, Jordan and Wainwright (2018) give general techniques to derive minimax rates under local DP, and in particular give minimax optimal point estimates for the mean, median, generalized linear models, and nonparametric density estimation. Karwa and Vadhan (2017) develop nearly optimal confidence intervals for normally distributed data with finite sample guarantees, which could potentially be inverted to give approximately UMP unbiased tests.

Notable works that develop optimal DP mechanisms for general loss functions are Geng and Viswanath (2015) and Ghosh, Roughgarden and Sundararajan (2012), which give mechanisms that optimize symmetric convex loss functions, centered at a real-valued statistic. Similarly, Awan and Slavković (2021) derive optimal mechanisms among the class of K-Norm Mechanisms for a fixed statistic and sample size.

2. Background. In this section, we review some basic notation as well as background on differential privacy. Notation and terminology regarding hypothesis testing is deferred to Appendix A of the Supplementary Material (Awan and Vadhan (2023)).

We say that a real-valued function $f(x)$ is *increasing* (*decreasing*) if $a \leq b$ implies $f(a) \leq f(b)$ (resp. $f(a) \geq f(b)$). We say that f is *strictly increasing* (*strictly decreasing*) if $a < b$ implies $f(a) < f(b)$ (resp. $f(a) > f(b)$). Given an increasing function f, we define its inverse to be $f^{-1}(y) = \inf\{x \in \mathbb{R} | y \leq f(x)\}$. For a decreasing function f, the inverse is defined to be $f^{-1}(y) = \inf\{x \in \mathbb{R} | y \geq f(x)\}$.

For a real-valued random variable X, its *cumulative distribution function* (cdf) is defined as $F_X(t) = P(X \leq t)$, and its *quantile function* is F_X^{-1}. A real-valued random variable is *continuous* if its cdf $F_X(t)$ is continuous in t, and X is *symmetric about zero* if $F_X(t) = 1 - F_X(-t)$. For a random variable $X \sim P$, with cdf F, we use P and $F(\cdot)$ interchangeably to denote the distribution of X.

2.1. *Differential privacy.* In this section, we review the definition of f-DP which is formulated in terms of constraints on hypothesis tests and relate it to other notions of DP in the literature. A *mechanism* M is a randomized algorithm that takes as input a database D, and outputs a (randomized) statistic $M(D)$ in an abstract space $\mathcal{Y}$. Given two databases X and X' which differ in one person's contribution, a mechanism M satisfies differential privacy if given the output of M it is difficult to determine whether the original database was X or X'.

The notion "differing in one person's contribution" is often formalized in terms of a metric. In this paper, we use the Hamming metric, which is defined as follows: For any set $\mathcal{X}$, we write $\mathcal{X}^n = \{(x_1, x_2, \dots, x_n) | x_i \in \mathcal{X} \text{ for all } 1 \leq i \leq n\}$. The *Hamming metric* on $\mathcal{X}^n$ is defined by $H(X, X') = \#\{i | X_i \neq X_i'\}$. If $H(X, X') \leq 1$, we call X and X' *adjacent databases*. Note that by using the Hamming metric, we assume that the sample size n is a public value and does not require privacy protection.

All of the major variants of DP state that given a randomized algorithm M, for any two adjacent databases X, X', the distributions of $M(X)$ and $M(X')$ should be "similar." While many DP variants measure similarity in terms of divergences, recently Dong, Roth and Su (2022) proposed f-DP, which formalizes similarity in terms of constraints on hypothesis tests.

For two probability distributions P and Q, the *tradeoff function* $T(P, Q) : [0, 1] \to [0, 1]$ is defined as $T(P, Q)(\alpha) = \inf\{1 - \mathbb{E}_Q \phi | \mathbb{E}_P(\phi) \leq \alpha\}$, where the infinimum is over all measurable tests ϕ. The tradeoff function can be interpreted as follows: If $T(P, Q)(\alpha) = \beta$, then the most powerful test ϕ which is trying to distinguish between $H_0 = \{P\}$ and $H_1 : \{Q\}$ at type I error $\leq \alpha$ has type II error β. A larger tradeoff function means that it is harder to distinguish between P and Q. Note that the tradeoff function is closely related to the receiver–operator curve (ROC), and captures the difficulty of distinguishing between P and Q. A function $f : [0, 1] \to [0, 1]$ is a tradeoff function if and only if f is convex, continuous, decreasing, and $f(x) \leq 1 - x$ for all $x \in [0, 1]$ (Dong, Roth and Su (2022), Proposition 1). We say that a tradeoff function f is *nontrivial* if $f(\alpha) < 1 - \alpha$ for some $\alpha \in (0, 1)$; that is if f is not identically equal to $1 - \alpha$.

DEFINITION 2.1 (f-DP: Dong, Roth and Su (2022)). Let f be a tradeoff function. A mechanism M satisfies f-DP if for all $D, D' \in \mathcal{X}^n$ such that $H(D, D') \leq 1$, we have

$$T\big(M(D), M(D')\big) \geq f.$$

See Figure 1 for examples of tradeoff functions which do and do not satisfy f-DP for a particular f. In the above definition, the inequality $T(M(D), M(D')) \geq f$ is shorthand for $T(M(D), M(D'))(\alpha) \geq f(\alpha)$ for all $\alpha \in [0, 1]$. Without loss of generality, we can assume that f is symmetric: $f(\alpha) = f^{-1}(\alpha)$, where $f^{-1}(\alpha) = \inf\{t \in [0, 1] | f(t) \leq \alpha\}$. This is due to the fact that adjacency of databases is a symmetric relation (Dong, Roth and Su (2022), Proposition 2). For the remainder of the paper, we assume that f-DP also requires this symmetry.

Wasserman and Zhou (2010) and Kairouz, Oh and Viswanath (2017) both showed that (ϵ, δ)-DP can be expressed in terms of hypothesis testing, and in fact Dong, Roth and Su (2022) showed that (ϵ, δ)-DP can be expressed as a special case of f-DP.

DEFINITION 2.2 ((ϵ, δ)-DP: Dwork et al. (2006)). Let $\epsilon > 0$ and $\delta \geq 0$, and define $f_{\epsilon,\delta}(\alpha) = \max\{0, 1 - \delta - \exp(\epsilon)\alpha, \exp(-\epsilon)(1 - \delta - \alpha)\}$. Then we say that a mechanism M satisfies (ϵ, δ)-*DP* if it satisfies $f_{\epsilon,\delta}$-DP.

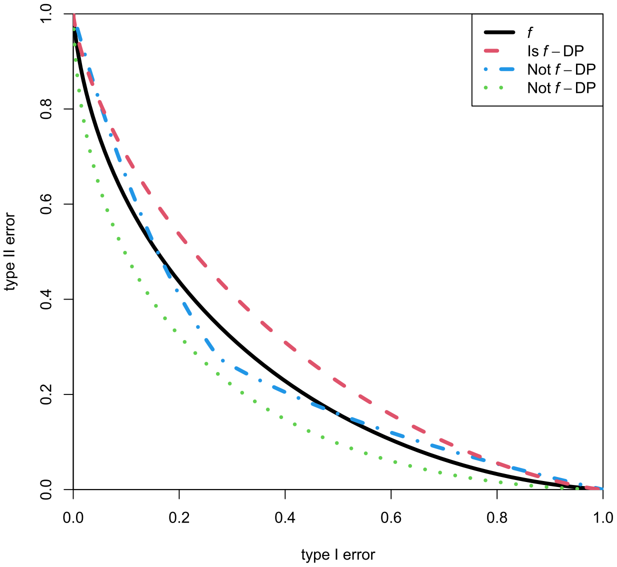


FIG. 1. *A plot of three examples of $T(M(D), M(D'))$. Only the red, dashed tradeoff curve satisfies f-DP.*

Another notable special case of f-DP is Gaussian DP (μ-GDP). Dong, Roth and Su (2022) showed that μ-GDP is perhaps the most natural single-parameter privacy definition, due to the central limit theorem for composition. Gaussian DP is closely related to zero-concentrated differential privacy (zCDP) (Bun and Steinke (2016)), a very popular relaxation of DP. GDP is slightly stronger than zCDP in that a mechanism satisfying GDP satisfies zCDP (Dong, Roth and Su (2022), Corollary B.6), but the converse is not true (Dong, Roth and Su (2022), Proposition B.7).

DEFINITION 2.3 (Gaussian differential privacy: Dong, Roth and Su (2022)). Let $\mu > 0$ and define

$$G_\mu(\alpha) = T\big(N(0, 1), N(\mu, 1)\big)(\alpha) = \Phi\big(\Phi^{-1}(1 - \alpha) - \mu\big),$$

where Φ is the cdf of $N(0, 1)$. We say that a mechanism M satisfies *μ-Gaussian differential privacy* (μ-GDP) if it is G_μ-DP.

3. Canonical noise distributions. One of the most basic techniques of designing a privacy mechanism is through adding data-independent noise. The earliest DP mechanisms add either Laplace or Gaussian noise, and there have since been several works developing optimal additive mechanisms including the geometric (discrete Laplace) (Ghosh, Roughgarden and Sundararajan (2012)), truncated-uniform-Laplace (Tulap) (Awan and Slavković (2018), Awan and Slavković (2020)), and staircase mechanisms (Geng and Viswanath (2015)). There have also been several works exploring multivariate and infinite-dimensional additive mechanisms such as K-norm (Awan and Slavković (2021), Hardt and Talwar (2010)), elliptical perturbations (Reimherr and Awan (2019)), and Gaussian processes (Hall, Rinaldo and Wasserman (2013), Mirshani, Reimherr and Slavković (2019)).

While there are many choices of additive mechanisms to achieve f-DP, we are interested in adding the least noise necessary in order to maximize the utility of the output. Rather than measuring the amount of noise by its variance or entropy, we focus on whether the privacy guarantee is tight.

In this section, we introduce the concept *canonical noise distribution* (CND), which captures whether a real-valued distribution is perfectly tailored to satisfy f-DP. We formalize this in Definition 3.1. We then show that for any symmetric f, we can always construct a CND, where the construction is given in Definition 3.7 and proved to be a CND in Theorem 3.9. We will see in Section 4 that CNDs are fundamental for understanding the nature of f-DP hypothesis tests, for constructing "free" DP p-values, and for the design of uniformly most powerful f-DP tests for binary data. We also see in Section 5 that CNDs are central to our application of difference-of-proportions tests as well.

Before we define canonical noise distribution, we must introduce the *sensitivity* of a statistic, a central concept of DP (Dwork et al. (2006)). A statistic $T : \mathcal{X}^n \to \mathbb{R}$ has *sensitivity* $\Delta > 0$ if $|T(X) - T(X')| \leq \Delta$ for all $H(X, X') \leq 1$. As the sensitivity measures how much a statistic can change when one person's data is modified, additive noise must be scaled proportionally to the sensitivity in order to protect privacy.

DEFINITION 3.1. Let f be a symmetric nontrivial tradeoff function. A continuous distribution function F is a *canonical noise distribution* (CND) for f if:

1. for every statistic $S : \mathcal{X}^n \to \mathbb{R}$ with sensitivity $\Delta > 0$, and $N \sim F(\cdot)$, the mechanism $S(X) + \Delta N$ satisfies f-DP. Equivalently, for every $m \in [0, 1]$, $T(F(\cdot), F(\cdot - m)) \geq f$,
2. $f(\alpha) = T(F(\cdot), F(\cdot - 1))(\alpha)$ for all $\alpha \in (0, 1)$,
3. $T(F(\cdot), F(\cdot - 1))(\alpha) = F(F^{-1}(1 - \alpha) - 1)$ for all $\alpha \in (0, 1)$,
4. $F(x) = 1 - F(-x)$ for all $x \in \mathbb{R}$; that is, F is the cdf of a random variable which is symmetric about zero.

The most important conditions of Definition 3.1 are 1 and 2, which state that the distribution can be used to satisfy f-DP and that the privacy bound is tight. For property 1, the value m can be interpreted as the quantity $|S(X) - S(X')|/\Delta$; then by the symmetry of F, it can be seen that $T(S(X) + \Delta N, S(X') + \Delta N) = T(F(\cdot), F(\cdot - m))$. Condition 3 of Definition 3.1 gives a closed form for the tradeoff function, and is equivalent to requiring that the optimal rejection set for discerning between $F(\cdot)$ and $F(\cdot - 1)$ is of the form (x, ∞) for some $x \in \mathbb{R}$. The last condition of Definition 3.1 enforces symmetry of the distribution, which makes CNDs much easier to work with.

Finally note that conditions 1 and 2 are not equivalent. Adding excessive noise would satisfy 1, but not 2, whereas a mechanism which fails $T(F(\cdot), F(\cdot - m)) \geq T(F(\cdot), F(\cdot - 1))$ for some $m \in (0, 1)$ would not satisfy property 1. The following example illustrates both cases.

EXAMPLE 3.2. Consider the discrete Laplace mechanism, which has cdf $F(t) = \frac{1-b}{1+b} b^{|t|}$ for $t \in \mathbb{Z}$ and $b \in (0, 1)$. Then it can be verified that the discrete Laplace distribution with $b = \exp(-\epsilon)$ satisfies $T(F(\cdot), F(\cdot - 1)) = f_{\epsilon,0}$, but not part 1 of Definition 3.1. For example, if $S(X) = 0$ and $S(X') = 0.1$, adding discrete Laplace noise $N \sim F$ results in distributions with disjoint support, since $S(X) + N$ takes values in $\mathbb{Z}$, whereas $S(X') + N$ takes values in $\mathbb{Z} + 0.1$. As the supports of the distributions are disjoint, we can have zero type I and type II error when testing between X and X', violating the $f_{\epsilon,0}$ bound.

It is well known that the continuous Laplace mechanism with scale parameter Δ/ϵ satisfies ϵ-DP, when added to a Δ-sensitivity statistic, and so satisfies property 1 of Definition 3.1 for $f_{\epsilon,0}$. However, as Dong, Roth and Su (2022) noted, it can be verified that the Laplace distribution does not satisfy property 2 of Definition 3.1, as there exists $\alpha \in (0, 1)$ such that the tradeoff function is strictly greater than $f_{\epsilon,0}$ at α.

REMARK 3.3. Note that property 2 of Definition 3.1 captures the intuition that a privacy mechanism should match the tradeoff function in the privacy guarantee to avoid introducing excessive noise. While this is indeed an intuitive idea, this has never previously been formalized into a precise criterion for a privacy mechanism, as we do in Definition 3.1. Furthermore, no prior work has attempted to build a mechanism that matches the tradeoff function for an arbitrary f-DP guarantee. In Theorem 3.9, we not only prove that a CND exists, but give a construction to build a CND for every f.

EXAMPLE 3.4 (CND for GDP). The distribution $N(0, 1/\mu)$, which has cdf $\Phi(1/\mu)$ (Φ is the cdf of a standard normal) is a CND for G_μ, defined in Definition 2.3. Property 1 is proved in (Dong, Roth and Su (2022)), properties 2 and 3 are easily verified, and the distribution is obviously symmetric. Dong, Roth and Su (2022) state that "GDP precisely characterizes the Gaussian mechanism." From the opposite perspective, we argue that this is because the normal distribution is a CND for G_μ.

PROPOSITION 3.5. *Let f be a symmetric nontrivial tradeoff function. Let F be a CND for f, and G be another cdf such that $T(G(\cdot), G(\cdot - 1)) \geq f$. Let $N \sim F$ and $M \sim G$. Then there exists a randomized function* $\text{Proc} : \mathbb{R} \to \mathbb{R}$ *which satisfies* $\text{Proc}(N) \overset{d}{=} M$ *and* $\text{Proc}(N + 1) \overset{d}{=} M + 1$, *where* "$\overset{d}{=}$" *means* equal in distribution.

Proposition 3.5 follows from property 2 in Definition 3.1 along with Dong, Roth and Su ((2022), Theorem 2), which is based on Blackwell's theorem (Blackwell (1950)). Proposition 3.5 shows that if we add noise from a CND to a statistic $S(X)$ versus $S(X) + 1$, we can post-process the result to obtain the same result as if we added noise from another distribution that achieves f-DP. This shows in some sense that a CND adds the least noise necessary to achieve f-DP. Note that Proposition 3.5 does not imply that a CND is optimal in every sense: for example, Geng and Viswanath (2015) derived the minimum variance additive $(\epsilon, 0)$-DP mechanism, which is not a CND for $f_{\epsilon,0}$. We will see in Section 4 that the properties of Definition 3.1 do lead to optimal properties of DP hypothesis tests.

In the remainder of this section, we show that given any tradeoff function f, we can always construct a canonical noise distribution (CND), but that a CND need not be unique.

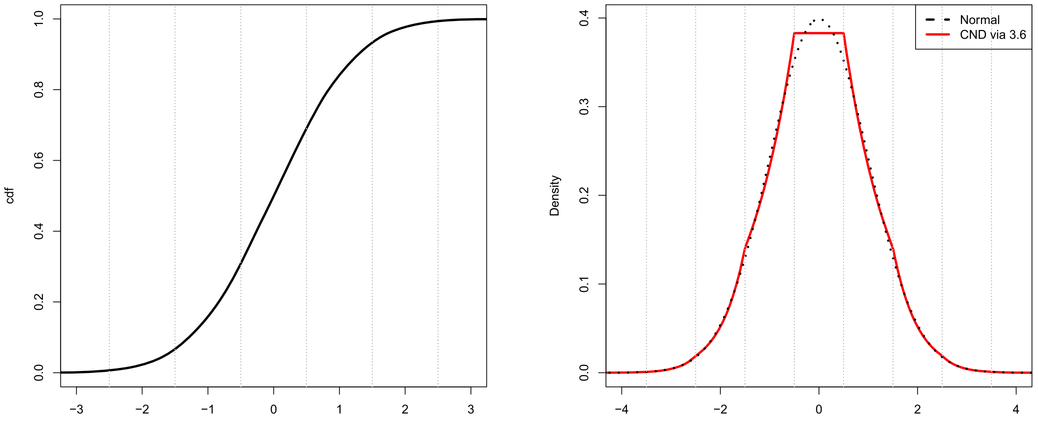
LEMMA 3.6. *Let f be a symmetric nontrivial tradeoff function and let F be a CND for f. Then $F(x) = 1 - f(F(x - 1))$ when $F(x - 1) > 0$ and $F(x) = f(1 - F(x + 1))$ when $F(x + 1) < 1$.*

PROOF SKETCH. The result follows from properties 2, 3, and 4 of Definition 3.1 along with some algebra of cdfs. □

In the Lemma 3.6, we see that a CND satisfies an interesting recurrence relation. If we know the value $F(x) = c$ for some $x \in \mathbb{R}$ and $c \in (0, 1)$, then we know the value of $F(y)$ for all $y \in \mathbb{Z} + x$. This means that if we specify F on an interval of length 1, such as $[-1/2, 1/2]$, then F is completely determined by the recurrence relation. While there are many choices to specify F on $[-1/2, 1/2]$, each of which may or may not lead to a CND. We show that using a particular linear function in $[-1/2, 1/2]$ does indeed give a CND. The remainder of this section is devoted to this construction of a CND and the proof that it has the properties of Definition 3.1.

DEFINITION 3.7. Let f be a symmetric nontrivial tradeoff function, and let $c \in [0, 1/2)$ be the unique fixed point of f: $f(c) = c$. We define $F_f : \mathbb{R} \to \mathbb{R}$ as

$$F_f(x) = \begin{cases} f\bigl(1 - F_f(x + 1)\bigr) & x < -1/2, \\ c(1/2 - x) + (1 - c)(x + 1/2) & -1/2 \leq x \leq 1/2, \\ 1 - f\bigl(F_f(x - 1)\bigr) & x > 1/2. \end{cases}$$



(a) Plot of cdf of the CND of Definition 3.7 corresponding to G_1. The function is linear between -1/2 and 1/2.

(b) Density plots of $N(0,1)$ as well as the CND of Definition 3.7 for the tradeoff function G_1.

FIG. 2. *Plots of CND construction of Definition* 3.7. *The vertical lines are at half-integer values.*

In Definition 3.7, the fact that there is a unique fixed point follows from the fact that f is convex and decreasing, and so intersects the line $y = \alpha$ at a unique value. In Lemma F.4 of the Supplementary Material (Awan and Vadhan (2023)), we establish that the fixed point c lies in the interval $[0, 1/2)$. Note that in Definition 3.7, the cdf corresponds to a uniform random variable on the interval $[-1/2, 1/2]$, but due to the recursive nature of F_f and the fact that f is in general nonlinear, the CND of Definition 3.7 need not be uniformly distributed on any other intervals. See Figure 2 for a plot of the pdf and cdf of the CND of Definition 3.7 corresponding to the tradeoff function G_1.

The following proposition verifies that F_f is a distribution function, as well as some other properties, such as continuity, symmetry, and concavity/convexity.

PROPOSITION 3.8. *Let f be a symmetric nontrivial tradeoff function, and let $F := F_f$. Then*:

1. *$F(x)$ is a cdf for a symmetric*, *continuous*, *real-valued random variable*,
2. *$F(x)$ satisfies $F(x) = 1 - f(F(x-1))$ whenever $F(x-1) > 0$ and $F(x) = f(1 - F(x+1))$ whenever $F(x+1) < 1$.*
3. *$F'(x)$ is decreasing on $(-1/2, \infty)$ and increasing on $(-\infty, 1/2)$*,
4. *$F(x)$ is strictly increasing on $\{x | 0 < F(x) < 1\}$.*

PROOF SKETCH. Most of the properties are proved by induction, checking that the properties hold on intervals of the type $[x - 1/2, x + 1/2]$ for $x \in \mathbb{Z}$ as well as at the break points at half-integer values. The full proof is found in Appendix F of the Supplementary Material (Awan and Vadhan (2023)). □

Theorem 3.9 below states that for any nontrivial tradeoff function, the construction of Definition 3.7 yields a canonical noise distribution. As we will see later, the existence (and construction) of a CND will enable us to prove that any f-DP test can be post-processed from a private test statistic, and this implies that we can always obtain hypothesis testing p-values at no additional privacy cost, a generalization of the result of Awan and Slavković (2018) which previously only held for (ϵ, δ)-DP and for Bernoulli data.

THEOREM 3.9. *Let f be a symmetric nontrivial tradeoff function and let F_f be as in Definition 3.7. Then F_f is a canonical noise distribution for f.*

PROOF SKETCH. F_f was already shown to be symmetric in Proposition 3.8. The two equalities, $f(\alpha) = T(F(\cdot), F(\cdot - 1))(\alpha) = F(F^{-1}(1-\alpha) - 1)$ can also be easily verified using the properties of Proposition 3.8. The main challenge is to show that $T(F(\cdot), F(\cdot - m)) \geq T(F(\cdot), F(\cdot - 1))$ for $m \in (0, 1)$. Lemma F.5 in the Supplementary Material (Awan and Vadhan (2023)) gives an alternative technical condition which makes it easier to verify property 1 of Definition 3.1. □

It turns out that the properties of Definition 3.1 do not uniquely determine a distribution. For instance, Φ the cdf of a standard normal is a CND for 1-GDP, but Φ is different from the construction in Definition 3.7. See Figure 2 for the cdf and pdf of these two CNDs. Note that the CND of Definition 3.7 is uniform in $[-1/2, 1/2]$ and has "kinks" at each half-integer value. On the other hand, the standard normal is smooth. This example shows that for some tradeoff functions there may be a more natural CND than the construction in Definition 3.7.

While there may be more natural CNDs in some settings, we emphasize the generality of the construction in Definition 3.7. In Proposition F.6 of the Supplementary Material (Awan and Vadhan (2023)), we present an exact method to sample from the CND of Definition 3.7 based on inverse transform sampling, allowing for straightforward implementation and application of our CND results.

3.1. *Canonical noise for (ϵ, δ)-DP.* So far, we have developed a constructive and general method of generating canonical noise distributions for f-DP. In the special case of (ϵ, δ)-DP, the CND of Definition 3.7 is equal to the cdf of the Tulap distribution, proposed in Awan and Slavković (2018), which is an extension of the Staircase mechanism (Geng and Viswanath (2015)) from $(\epsilon, 0)$-DP to (ϵ, δ)-DP.

COROLLARY 3.10. *The distribution* $\mathrm{Tulap}(0, b, q)$, *where* $b = \exp(-\epsilon)$ *and* $q = \frac{2\delta b}{1-b+2\delta b}$ *is a CND for $f_{\epsilon,\delta}$-DP, which agrees with the construction of Definition 3.7.*

PROOF SKETCH. The cdf of $\mathrm{Tulap}(0, b, q)$ is defined in the full proof. From the definition, it is easy to verify that the cdf of a Tulap random variable agrees with F_f on $[-1/2, 1/2]$. By Awan and Slavković ((2020), Lemma 2.8), the Tulap cdf also satisfies the recurrence relation of Definition 3.7. □

It was claimed in both Awan and Slavković (2018) and Awan and Slavković (2020) that adding Tulap noise satisfied (ϵ, δ)-DP, but their proof is actually incorrect and only holds for integer-valued statistics. The above Corollary along with Theorem 3.9 offers a complete and correct argument for Awan and Slavković ((2020), Theorem 2.11).

In Awan and Slavković (2018) and Awan and Slavković (2020), it was shown that the Tulap distribution could be used to design optimal hypothesis tests and confidence intervals for Bernoulli data. Our notion of a canonical noise distribution, and the fact that Tulap is a CND for (ϵ, δ)-DP sheds some light on why it had such optimality properties (even further explored in Section 4). The Tulap distribution is also closely related to discrete Laplace and the Staircase distributions, which were shown by Ghosh, Roughgarden and Sundararajan (2012) and Geng and Viswanath (2015) respectively to be optimal in terms of maximizing various definitions of utility in $(\epsilon, 0)$-DP.

While continuous Laplace noise is commonly used in $(\epsilon, 0)$-DP, Dong, Roth and Su (2022) pointed out that the tradeoff function for Laplace noise does not agree with $f_{\epsilon,\delta}$ for any values of ϵ and δ. From this observation, we conclude from Definition 3.1 that Laplace is not a CND for (ϵ, δ)-DP. From the perspective of CNDs, Tulap noise is preferable over the Laplace mechanism.

4. The nature of f-DP tests.

Recall that a test is a function $\phi : \mathcal{X}^n \to [0, 1]$, where $\phi(x)$ represents the probability of rejecting the null hypothesis given that we observed x. However, the mechanism corresponding to this test releases a random value drawn as $\mathrm{Bern}(\phi(x))$, where 1 represents "Reject" and 0 represents "Accept." we say that the test ϕ satisfies f-DP if the corresponding mechanism $\mathrm{Bern}(\phi(x))$ satisfies f-DP. Intuitively, Lemma 4.1 shows that a test satisfies f-DP if for adjacent databases x and x', the values $\phi(x)$ and $\phi(x')$ are close in terms of an inequality based on f.

LEMMA 4.1. *Let f be a symmetric tradeoff function. A test $\phi : \mathcal{X}^n \to [0, 1]$ satisfies f-DP if and only if $\phi(x) \leq 1 - f(\phi(x'))$ for all $x, x' \in \mathcal{X}^n$ such that $H(x, x') \leq 1$.*

PROOF SKETCH. If we take the rejection region to be the set $\{1\}$ then $\phi(x)$ is the type I error and $1 - \phi(x')$ is the type II error. The f-DP guarantee requires that $f(\phi(x)) \leq 1 - \phi(x')$, or equivalently, $\phi(x') \leq 1 - f(\phi(x))$. Using the rejection region $\{0\}$ and some algebra, we get $\phi(x) \leq 1 - f(\phi(x'))$. The full proof argues more precisely using the Neyman Pearson lemma, considering also randomized tests. □

Lemma 4.1 greatly simplifies the search for f-DP hypothesis tests and generalizes the bounds on private tests established in Awan and Slavković (2018).

EXAMPLE 4.2 ((ϵ, δ)-DP Tests). When we apply Lemma 4.1 to the setting of (ϵ, δ)-DP, we have the two inequalities: $(1 - \phi(x)) \geq 1 - \delta - \exp(\epsilon)\phi(x')$ and $(1 - \phi(x)) \geq \exp(-\epsilon)(1 - \delta - \phi(x'))$. Some algebra gives

$$\phi(x) \leq \begin{cases} \delta + \exp(\epsilon)\phi(x'), \\ 1 - \exp(-\epsilon)\big(1 - \delta - \phi(x')\big), \end{cases}$$

which agrees with the constraints derived in Awan and Slavković (2018).

The result of Lemma 4.1 can also be expressed in terms of canonical noise distributions in Corollary 4.3, giving the elegant relation that $F^{-1}(\phi(x))$ and $F^{-1}(\phi(x'))$ differ by at most 1 when x and x' are adjacent.

COROLLARY 4.3 (Canonical noise distributions). *Let f be a symmetric nontrivial tradeoff function and let F be a canonical noise distribution for f. Then a test ϕ satisfies f-DP if and only if $F^{-1}(\phi(x)) \leq F^{-1}(\phi(x')) + 1$ for all $x, x' \in \mathcal{X}^n$ such that $H(x, x') \leq 1$.*

PROOF SKETCH. The result follows from the fact that $f(\alpha) = F(F^{-1}(1 - \alpha) - 1)$, the symmetry of F, and some algebra of cdfs. □

Corollary 4.3 is also important for the construction of "free" DP p-values in Section 4.1.

4.1. *Free f-DP p-values.*

In Awan and Slavković (2018), it was shown that for Bernoulli data, the uniformly most powerful DP test could also be expressed as the post-processing of a privatized test statistic, offering p-values at no additional privacy cost. We generalize this result using canonical noise distributions and show that any f-DP test can be expressed as a post-processing threshold test based on a privatized test statistic, and that the test statistic can also be used to give private p-values.

Typically in statistics, it is preferred to report a p-value rather than an accept/reject decision at a single type I error. A p-value provides a continuous summary of how much evidence there is for the alternative hypothesis and allows for the reader to determine whether there is enough evidence to reject at the reader's personal type I error. Lower p-values give more evidence for the alternative hypothesis.

However, with privacy, one may wonder whether releasing a p-value rather than just the accept/reject decision would result in an increased privacy cost, or conversely whether a p-value at the same privacy level would have lower power. In fact, this question is related to fundamental concepts in differential privacy such as post-processing, privacy amplification, and composition. In Lemma 4.4, we recall the post-processing property of DP, which states that after a DP result is released, no post-processing can compromise the DP guarantee.

LEMMA 4.4 (Post-processing: Dong, Roth and Su (2022)). *Let M be an f-DP mechanism taking values in $\mathcal{Y}$. Let* Proc *be a mechanism from $\mathcal{Y}$ to $\mathcal{Z}$. Then* $\mathrm{Proc} \circ M$ *satisfies f-DP.*

Theorem 4.5 is the main result of this section, demonstrating that given an arbitrary f-DP hypothesis test, we can construct a summary statistic and p-value, with no additional privacy cost, using a CND.

THEOREM 4.5. *Let $\phi : \mathcal{X}^n \to [0, 1]$ be an f-DP test. Let F be a CND for f, and draw $N \sim F$. Then:*

1. *releasing $T = F^{-1}(\phi(x)) + N$ satisfies f-DP,*
2. *the variable $Z = I(T \geq 0)$, a post-processing of T, is distributed as $Z|X = x \sim \mathrm{Bern}(\phi(x))$,*
3. *the value $p = \sup_{\theta_0 \in H_0} \mathbb{E}_{X \sim \theta_0} F(F^{-1}(\phi(X)) - T)$ is also a post-processing of T and is a p-value for H_0,*
4. *if H_0 is a simple hypothesis and $\mathbb{E}_{H_0}\phi = \alpha$, then at type I error α, the p-value from part* 3 *is as powerful as ϕ at every alternative.*

PROOF SKETCH. Property 1 follows from Corollary 4.3, the observation that $F^{-1}(\phi(x))$ has sensitivity 1, and property 1 of Definition 3.1. Property 2 can be verified using algebra of cdfs. Property 3 is a standard construction of a p-value (Casella and Berger (2002), Theorem 8.3.27). Property 4 is a special case of Lemma F.8 in the Supplementary Material, a general lemma about p-values. □

We see from Theorem 4.5 that given an f-DP test ϕ, we can report both a summary statistic (namely, T) as well as a p-value (a post-processing of T) which contain strictly more information than only sampling $\mathrm{Bern}(\phi(x))$. This shows that for simple null hypotheses, there is no general privacy amplification when post-processing a p-value or test statistic to a binary accept/reject decision.

While in part 3 of Theorem 4.5 there are no assumptions on H_0, for some composite null hypotheses the resulting p-value may have very low power. Part 4 states that if the null hypothesis is a singleton, then the power is perfectly preserved.

We also remark that while the proof of Theorem 4.5 is not technical, it heavily relies on the properties of the CND, showing that the notion of CND has exactly the right properties for Theorem 4.5 to hold.

Note that Theorem 4.5 starts with an f-DP test, and shows how to get a private summary statistic and p-values. However, constructing a private test ϕ is another matter. In Section 4.2, we show that for exchangeable binary data, we can construct a most powerful f-DP test in terms of a CND.

REMARK 4.6. While recently there has been controversy around the use of p-values in scientific research (Colquhoun (2017), Wasserstein and Lazar (2016)), this is mostly due to the misuse or misinterpretation of a p-value. Many of the criticisms of p-values can be addressed by including additional statistical measures such as the effect size, confidence intervals, likelihood ratios, or Bayes factors. We view p-values as a valuable tool that is a component of a complete statistical analysis. Since the p-values of Theorem 4.5 are a post-processing of a private summary statistic, that statistic can also be potentially used for other statistical inference tasks, such as in Awan and Slavković (2020).

4.2. *Most powerful tests for exchangeable binary data.* In this section, we extend the main result of Awan and Slavković (2018), that of constructing most powerful DP tests, to general f-DP as well as exchangeable distributions on $\{0,1\}^n$. In contrast, the hypothesis tests of Awan and Slavković (2018) were limited to (ϵ,δ)-DP and i.i.d. Bernoulli data. A distribution P on a set $\mathcal{X}^n$ is *exchangeable* if given $\underline{X} \sim P$ and a permutation π, $\underline{X} \overset{d}{=} \pi(\underline{X})$. Note that i.i.d. data are always exchangeable, but there are exchangeable distributions that are not i.i.d. For example, sampling without replacement results in exchangeable but non-i.i.d. data.

In the next result, we extend Theorem 3.2 of Awan and Slavković (2018) from (ϵ,δ)-DP to the setting of general f-DP. The argument is essentially identical. We include the proof for completeness.

LEMMA 4.7 (Theorem 3.2 of Awan and Slavković (2018)). *Let $\mathcal{P}$ be a set of exchangeable distributions on $\mathcal{X}^n$. Let $\phi : \mathcal{X}^n \to [0,1]$ be a test satisfying f-DP. Then there exists a test $\phi' : \mathcal{X}^n \to [0,1]$ such that for all $\underline{x} \in \mathcal{X}^n$, $\phi'(\underline{x})$ only depends on the empirical distribution of $\underline{x}$, and $\int \phi'(\underline{x})\,dP = \int \phi(\underline{x})\,dP$ for all $P \in \mathcal{P}$.*

PROOF. Define $\phi'(\underline{x}) = \frac{1}{n!}\sum_{\pi\in\sigma(n)} \phi(\pi(\underline{x}))$, where $\sigma(n)$ is the symmetric group on n letters. Note that for any $\pi \in \sigma(n)$, $\phi(\pi(\cdot))$ satisfies f-DP (just rearranging the sample space). Furthermore, $\int \phi(\pi(\underline{x}))\,dP = \int \phi(\underline{x})\,dP$ by exchangeability. Finally, by the convexity of f, the set of tests ϕ which satisfy $\phi(x) \le 1 - f(\phi(x'))$ is a convex set, and so is closed under convex combinations. So, ϕ' defined above satisfies f-DP, and by the linearity of integrals, preserves the expectations. □

We work with the sample space $\mathcal{X} = \{0,1\}$. Note that by Lemma 4.7, because we are dealing with exchangeable distributions, the test need only depend on $X = \sum_{i=1}^n X_i$, so we define $\phi(x)$ for $x = 0,1,2,\dots,n$. Since changing one X_i only changes X by ±1, we need only relate $\phi(x)$ and $\phi(x-1)$.

The main result of this section, Theorem 4.8 constructs not only the first private hypothesis test in the general f-DP framework, but derives a most powerful f-DP test as well as a corresponding p-value in terms of the canonical noise distribution. The proof of Theorem 4.8 is similar to the proof of Awan and Slavković ((2018), Theorem 4.5), further demonstrating that the canonical noise distribution is the appropriate concept needed to extend their result from (ϵ,δ)-DP to arbitrary f-DP. Just like in Awan and Slavković (2018), we have the surprising result that the UMP DP test in this case only depends on the summary statistic $x+N$, where N is a CND. The extension from Bernoulli distributions to arbitrary exchangeable binary variables is simply an observation that the argument only depends on the likelihood ratio. However, the extension to exchangeable distributions will allow us to apply Theorem 4.8 to the difference-of-proportions problem in Section 5.

THEOREM 4.8. *Let f be a symmetric nontrivial tradeoff function and let F be a CND of f. Let $\mathcal{X} = \{0, 1\}$. Let P and Q be two exchangeable distributions on $\mathcal{X}^n$ with pmfs p and q such that $\frac{q}{p}$ is an increasing function of $x = \sum_{i=1}^n x_i$. Let $\alpha \in (0, 1)$. Then a most powerful f-DP test ϕ with level α for $H_0 : X \sim P$ versus $H_1 : X \sim Q$ can be expressed in any of the following forms*:

1. *There exists $y \in \{0, 1, 2, \dots, n\}$ and $c \in (0, 1)$ such that for all $x \in \{0, 1, 2, \dots, n\}$,*

$$\phi(x) = \begin{cases} 0 & x < y, \\ c & x = y, \\ 1 - f\bigl(\phi(x-1)\bigr) & x > y, \end{cases}$$

where if $y > 0$ then c satisfies $c \leq 1 - f(0)$, and c and y are chosen such that $\mathbb{E}_P\phi(x) = \alpha$. If $f(0) = 1$, then $y = 0$.

2. *$\phi(x) = F(x - m)$, where $m \in \mathbb{R}$ is chosen such that $\mathbb{E}_P\phi(x) = \alpha$.*
3. *Let $N \sim F$. Then $T = X + N$ satisfies f-DP. Then $p = \mathbb{E}_{X\sim P}F(X - T)$ is a p-value and $I(p \leq \alpha)|X = I(T \geq m)|X \sim \mathrm{Bern}(\phi(X))$, where $\phi(x)$ agrees with 1 and 2 above.*

PROOF SKETCH. Similar to the proof of Awan and Slavković ((2018), Theorem 4.5), we begin by establishing the equivalence of forms 1 and 2, and arguing that there exists a test of the form 2 by the Intermediate Value Theorem. Using Awan and Slavković ((2018), Lemma 4.4), a variation of the Neyman Pearson lemma, we argue that the proposed ϕ is most powerful. Statement 3 uses the expressions from Theorem 4.5 as well as some distributional algebra of CNDs to get the explicit formula. □

While Theorem 4.5 took an f-DP test and produced "free" private p-values, Theorem 4.8 constructs an optimal test from scratch beginning only with a CND.

EXAMPLE 4.9. Let us consider what distributions fit within the framework of Theorem 4.8. If the variables X_i are i.i.d., then they are distributed as Bernoulli. However, it is possible for the variables to be exchangeable and not independent. For example, the sum $X = \sum_{i=1}^n X_i$ could be distributed as a hypergeometric or Fisher's noncentral hypergeometric, which arises in two sample tests of proportions, see Section 5. For other exchangeable binary distributions, see Dang, Keeton and Peng (2009).

REMARK 4.10. Theorem 4.8 and Corollary 4.3 show that the results of Awan and Slavković (2020) extend to arbitrary f-DP. By simply modifying the Tulap distribution to a CND, all of the other results of Awan and Slavković (2020) carry over as well. In particular, for Bernoulli data, there exists a UMP one-sided test, a UMP unbiased two-sided test, UMA one sided confidence interval and UMA unbiased two-sided confidence interval. All of these quantities are a post-processing of the summary value $X + N$, where the noise N is drawn from a CND F of f.

5. Extension to semiprivate difference-of-proportions tests.

Testing two population proportions is a common hypothesis testing problem, which arises in clinical trials with control and test groups, A/B testing, and observation studies comparing two groups (such as men and women, students from two universities, or aspects of two different countries). As such, the techniques for testing such hypotheses are standard and taught in many statistics textbooks. However, there are limited techniques to test these hypotheses under f-DP.

In Appendix D of the Supplementary Material (Awan and Vadhan (2023)), we show that there does not exist a UMP (unbiased) f-DP test. Nevertheless, we use the techniques developed earlier in this paper to derive a "semiprivate" UMP unbiased test, which gives an upper bound on the power of any f-DP UMP unbiased test. The novel concept of "semiprivacy" enforces some of the DP constraints but not others, and this framework may be of independent interest when analyzing a combination of private and nonprivate releases (see Remark 5.4 for more details). We then construct an f-DP test which allows for optimal inference on the two population parameters, and which we show through simulations to have comparable power to the semiprivate UMP unbiased test. In the case of ϵ-DP, we show through simulations that the proposed DP test is similar to the semiprivate UMP unbiased test with privacy parameter $(\epsilon/\sqrt{2})$. We also demonstrate that the proposed test has more accurate p-values and type I error than commonly used Normal approximation tests.

5.1. *Semiprivate UMP unbiased test.* In this section, we simplify the search for an f-DP test for the difference of proportions, establishing a condition for the test to be *unbiased*. However, as demonstrated through an example in Appendix D of the Supplementary Material (Awan and Vadhan (2023)), there does not exist a UMP unbiased (UMPU) f-DP test. By weakening the privacy guarantee, we develop a "semiprivate" UMPU test which is efficiently implemented. While the "semiprivate" test does not satisfy f-DP, it gives an upper bound on the power of any other unbiased f-DP test, and serves as a useful baseline in Section 5.3.

We observe independent $X_i \overset{\text{iid}}{\sim} \text{Bern}(\theta_X)$ for $i = 1, \dots, n$ and $Y_j \overset{\text{iid}}{\sim} \text{Bern}(\theta_Y)$ for $j = 1, \dots, m$. For privacy, we consider two datasets *adjacent* if either one of the X_i is changed or one of the Y_i is changed (but only one total value). We consider m and n to be publicly known values. We wish to test $H_0 : \theta_X \geq \theta_Y$ versus $H_1 : \theta_X < \theta_Y$, subject to the constraint of differential privacy. Such one-sided tests can also be converted to two-sided tests using a Bonferroni correction, as discussed in Remark 5.9, at the end of Section 5.2.

By a similar argument as in Lemma 4.7, it is sufficient to consider tests which are functions of the empirical distributions of $\underline{X}$ and $\underline{Y}$. Equivalently, we may restrict to tests which are functions of $X = \sum_{i=1}^n X_i$ and $Y = \sum_{j=1}^m Y_j$. We consider two databases adjacent if either X changes by 1 or if Y changes by 1 (but not both). By Lemma 4.1, a test $\phi(x, y)$ satisfies f-DP if the following set of inequalities hold for all pairs of (x, y):

$$\begin{aligned} \phi(x, y) &\leq 1 - f\big(\phi(x + 1, y)\big), \\ \phi(x, y) &\leq 1 - f\big(\phi(x - 1, y)\big), \\ \phi(x, y) &\leq 1 - f\big(\phi(x, y + 1)\big), \\ \phi(x, y) &\leq 1 - f\big(\phi(x, y - 1)\big). \end{aligned} \tag{1}$$

Classically, it is known that even without privacy there is no uniformly most powerful test for this problem. Traditionally, attention is restricted to unbiased tests. Recall that a test is unbiased if for all $\theta_1 \in \Theta_1$ and $\theta_0 \in \Theta_0$, the power at θ_1 is higher than at θ_0 (here, θ represents the pair (θ_X, θ_Y)). Because the variables (X, Y) have distribution in the exponential family, the search for a UMP unbiased test can be restricted to tests which satisfy $\mathbb{E}_{\theta_X=\theta_Y}(\phi(X, Y)|X + Y = z) = \alpha$ (Schervish (2012), Proof of Theorem 4.124), since $X + Y$ is a complete sufficient statistic under H_0. When $\theta_X = \theta_Y = \theta_0$, $X + Y \sim \text{Binom}(m + n, \theta_0)$, and $Y|(X + Y = z) \sim \text{Hyper}(m, n, z)$, where $\text{Hyper}(m, n, z)$ is the hypergeometric distribution, where we draw m balls out of a total of $m + n$ balls, and where z balls are white, and the random variable counts the number of drawn white balls. This is equivalent to a permutation test where we shuffle the labels of the observations. Lemma 5.1 summarizes these observations.

LEMMA 5.1. *Let* $X \sim \text{Binom}(n, \theta_X)$ *and* $Y \sim \text{Binom}(m, \theta_Y)$ *be independent. Consider the test* $H_0 : \theta_X \geq \theta_Y$ *and* $H_1 : \theta_X < \theta_Y$. *Let* Φ *be a set of tests. If there exists a UMP test*

$\phi \in \Phi$ among those which satisfy

$$\mathbb{E}_{H\sim \mathrm{Hyper}(m,n,z)}\phi(z-H,H)=\alpha, \tag{2}$$

for all α, then ϕ is UMP unbiased size α among Φ.

PROOF. It is easy to verify that the power function is continuous, and that $X+Y$ is a boundedly complete sufficient statistic under H_0. By Schervish ((2012), Proposition 4.92) and Schervish ((2012), Lemma 4.122), the set of unbiased tests for this problem is a subset of the tests which satisfy Equation (2). It is also clear that Equation (2) implies that the test is size α. It follows that if a test is UMP among the tests in Φ satisfying Equation (2) then it is UMP unbiased size α among Φ. □

However, as demonstrated by an example given later in Appendix D of the Supplementary Material (Awan and Vadhan (2023)), in general there is no UMP test for the hypothesis $H_0 : \theta_X \geq \theta_Y$ versus $H_1 : \theta_X < \theta_Y$ among the set

$$\Phi_f = \{\phi(x,y) \mid \phi \text{ satisfies inequalities (1) and Equation (2)}\}. \tag{3}$$

The reason for this is that Lemma 5.1 suggests that a UMP unbiased test relies on being able to construct a UMP test, given $X+Y=z$. However, the inequalities (1) put constraints, relating $\phi(x,y)$ for different values of z.

Instead of requiring that all of the inequalities (1) hold, we weaken the requirement of differential privacy, to only include the constraints relating (x,y) with the same sum $x+y=z$. We call the following the set of "semiprivate" tests:

$$\Phi_f^{\text{semi}} = \left\{ \phi(x,y) \;\middle|\; \begin{array}{c} \text{for each } z\in\{0,1,\ldots,m+n\}, \\ \text{there exists } \psi \in \Phi_f, \\ \text{s.t. } \phi(x,y)=\psi(x,y) \text{ for all } x+y=z \end{array} \right\}.$$

Intuitively, Φ_f^{semi} is the set of tests, which satisfy the set of implied constraints of (1), which only relate (x,y) and $(x+1,y-1)$. So, the summary $z=X+Y$ is not protected at all, but for any $X+Y=z$, (X,Y) must satisfy f-DP. While these semiprivate tests are not necessarily intended for the purpose of privacy protection, by weakening the privacy requirement, they offer an upper bound on the performance of any DP test, as stated in Corollary 5.3.

THEOREM 5.2 (Semiprivate UMPU). *Let f be a symmetric nontrivial tradeoff function and let F be a CND for f. Let $X\sim \mathrm{Binom}(n,\theta_X)$ and $Y\sim \mathrm{Binom}(m,\theta_Y)$ be independent. Let $\alpha \in (0,1)$ be given. For the hypothesis $H_0:\theta_X\geq\theta_Y$ versus $H_1:\theta_X<\theta_Y$:*

1. *$\phi^*(x,y)=F(y-x-c(x+y))$ is the UMPU test of size α among Φ_f^{semi}, where $c(x+y)$ is chosen such that $\mathbb{E}_{H\sim\mathrm{Hyper}(m,n,x+y)}\phi^*((x+y)-H,H)=\alpha$.*
2. *Set $T=Y-X+N$, where $N\sim F$, and set $Z=X+Y$. Then*

$$p=\mathbb{E}_{H\sim\mathrm{Hyper}(m,n,Z)}F(2H-Z-T)$$

is the exact p-value corresponding to ϕ^*.

PROOF SKETCH. Lemma 5.1 reduced the problem to determining whether the test is UMP among those which satisfy Equation (2). The technical Lemmas F.10 and F.12 in the Supplementary Material (Awan and Vadhan (2023)), quantify the privacy of the semiprivate tests when viewed as a function of y (where z is fixed), and determine the CND of the derived tradeoff function. Conditional on z, the distribution of Y is a Fisher noncentral hypergeometric distribution (Fog (2008), Harkness (1965)). By Theorem 4.8, we can construct the most powerful DP test based on the CND. Finally, we verify a monotone likelihood ratio property of the noncentral hypergeometrics to argue that the test is uniformly most powerful. □

Corollary 5.3 shows that while the semiprivate UMPU test does not satisfy f-DP, we can use it as a benchmark to compare other tests, as it gives an upper bound on the highest possible power of any unbiased f-DP level α test.

COROLLARY 5.3. *Let $\phi^*(x,y)$ be the UMPU size α test among Φ_f^{semi}, and let $\phi(x,y)$ be any unbiased, level α test in Φ_f. Then for every pair of values of $\theta_X \leq \theta_Y$,*

$$\mathbb{E}_{\substack{X\sim\theta_X\\ Y\sim\theta_Y}}\phi^*(X,Y) \geq \mathbb{E}_{\substack{X\sim\theta_X\\ Y\sim\theta_Y}}\phi(X,Y).$$

REMARK 5.4. The semiprivate framework could potentially be of independent interest, as it is an example of a setting where some statistics are preserved exactly, whereas others are protected with privacy noise. For example, this is similar to the framework used for the 2020 Decennial Census, where certain counts are preserved without any privacy noise, and the other counts are sanitized by an additive noise mechanism. While they phrase their privacy guarantee in terms of post-processing, one could also view it as a "semiprivate" procedure, where their privacy guarantee only holds for the databases which agree with the preserved counts. This is an alternative perspective to *subspace differential privacy* (Gao, Gong and Yu (2022)), which restricts the output of a mechanism rather than the input database.

5.2. *Designing an f-DP test for difference-of-proportions.* Based on the negative result of Appendix D, we consider a different approach to building a well-performing DP test. A common nonprivate test for $H_0: \theta_X \geq \theta_Y$ versus $H_1: \theta_X < \theta_Y$ for $X \sim \mathrm{Binom}(n, \theta_X)$ and $Y \sim \mathrm{Binom}(m, \theta_Y)$ is based on the test statistic $Y/m - X/n$, which is intuitive as this quantity captures the sample evidence for the difference between θ_X and θ_Y. In fact this statistic has the important property that its expectation under the null does not depend on the parameter $\theta_X = \theta_Y$. If this were not the case, then tests based on this statistic would have limited power (Robins, van der Vaart and Ventura (2000)). However, the sampling distribution of this quantity depends on the parameter $\theta_0 = \theta_X = \theta_Y$ under the null (e.g., for $\theta_0 = 1/2$, the variance of $Y/m - X/n$ is higher than when θ_0 is larger or smaller). Typically, the central limit theorem is used to justify that

$$\frac{Y/m - X/n}{\sqrt{(1/m + 1/n)\hat{\theta}_0(1-\hat{\theta}_0)}} \approx N(0,1),$$

where $\hat{\theta}_0 = \frac{X+Y}{m+n}$ is the maximum likelihood estimator for θ_0 under the null. The central limit approximation works well in large samples, but for small samples this approximation can be inadequate as demonstrated in the simulations of Section 5.3.

5.2.1. *Inversion-based parametric bootstrap f-DP test.* In this section, we consider tests based on the following privatized summary quantities $X + N_1$ and $Y + N_2$, where $N_1, N_2 \overset{\text{iid}}{\sim} F$ where F is a CND of f. The vector $(X + N_1, Y + N_2)$ satisfies f-DP, since only one of X and Y changes by at most 1, between adjacent databases.

REMARK 5.5. Basing our test on these two noisy statistics has a few important benefits. As noted in Remark 4.10, given $X + N_1$ and $Y + N_2$ we can perform optimal hypothesis tests and confidence intervals for θ_X and θ_Y combining Theorem 4.8, Corollary 4.3 and the other results of Awan and Slavković (2020). In general this is not the case for an arbitrary

f-DP test of $H_0 : \theta_X \geq \theta_Y$ versus $H_1 : \theta_X < \theta_Y$. While 4.5 says that we can always get a summary statistic and p-value out of an arbitrary f-DP test, these values may not contain enough information to do inference (let alone optimal inference) for θ_X and θ_Y separately.

We consider the quantity $T = m^{-1}(Y + N_2) - n^{-1}(X + N_1)$. Asymptotics tells us that under the null hypothesis, $T/\sqrt{(1/m + 1/n)\theta_0(1 - \theta_0)} \xrightarrow{d} N(0, 1)$, which is the same sampling distribution as without privacy. However, as many other researchers have noted, while these approximations are serviceable in classical settings, the approximations are too poor when privacy noise is introduced (Wang et al. (2018)). One reason for this is that the noise introduced to achieve privacy, such as Laplace or Tulap, often has heavier tails than the limit distribution, which is often Gaussian.

We notice that T is a linear combination of independent random variables. So, we can use characteristic functions to derive the sampling distribution of T under a specific null parameter θ_0. We use $\psi_X(\cdot)$ to denote the characteristic function of a random variable X: $\psi_X(t) := \mathbb{E}_X e^{itX}$. Recall that for independent random variables $X_1, \dots, X_n$ and real values $a_1, \dots, a_n$, if $X = \sum_{i=1}^n a_i X_i$, then $\psi_X(t) = \prod_{i=1}^n \psi_{X_i}(a_i t)$. Then, the characteristic function of our test statistic T is given by

$$\psi_{T\sim\theta_0}(t) = \psi_{Y\sim\theta_0}(t/m)\psi_{N_2}(t/m)\psi_{X\sim\theta_0}(-t/n)\psi_{N_1}(-t/n).$$

We know the characteristic function for a binomial random variable, and for many common DP distributions N, we have formulas for ψ_N as well.

We can use the following inversion formula to evaluate the cdf of T.

LEMMA 5.6 (Inversion formula: Gil-Pelaez). *Let X be a real-valued continuous random variable, with characteristic function $\psi_X(t)$. Then the cdf of X can be evaluated as*

$$F_X(x) = \int_0^\infty \frac{\operatorname{Im}(e^{-itx}\psi_X(t))}{t}\,dt,$$

where $\operatorname{Im}(\cdot)$ *returns the imaginary component of a complex number*: $\operatorname{Im}(z) = (z - z^*)/(2i)$, *where z^* is the complex conjugate of z.*

Lemma 5.6 gives a computationally tractable method of evaluating the exact sampling distribution of T at a given null parameter. Since larger values of T give more evidence of the alternative hypothesis, $p(T) = 1 - F_{T\sim\theta_0}(T)$ is a p-value for the null hypothesis $H_0 : \theta_X = \theta_Y = \theta_0$ (Casella and Berger (2002), Theorem 8.3.27). However, this p-value depends on the null parameter θ_0, which we likely do not know. A solution is to substitute an estimator for θ_0 under the null hypothesis that $\theta_X = \theta_Y$, based on the privatized statistics $X + N_1$ and $Y + N_2$. A natural estimator is $\hat{\theta}_0 = \min\{\max\{\frac{X+N_1+Y+N_2}{m+n}, 0\}, 1\}$. Plugging this estimate in for θ_0 gives the approximate p-value:

$$\tilde{p}(T, \hat{\theta}_0) = 1 - F_{T_0\sim\hat{\theta}_0}(T).$$

This approximate p-value is our recommended f-DP test for the difference-of-proportions testing problem, and the procedure is summarized in Algorithm 1 for the cases of $(\epsilon, 0)$-DP and μ-GDP. While p-value is not exact, and is thus not guaranteed to have the intended type I error, the results of Robins, van der Vaart and Ventura (2000) imply that this p-value is asymptotically uniform under the null, implying that the test is asymptotically unbiased, with asymptotically accurate type I errors. Furthermore, as we demonstrate in Section 5.3, even with sample sizes as small as $n, m \geq 30$, the approximation is incredibly accurate, offering accuracy even higher than the classic normal approximation test, which is widely used and accepted. We also show in Section 5.3 that the power of the test is comparable to the semiprivate test of Section 5.1 indicating that it is nearly optimal.

Algorithm 1: ϵ-DP or μ-GDP approximate p-value, based on inversion

```
Let $X, Y, m$, and $n$ be given. Let either $\epsilon$ or $\mu$ be given.;
if $\epsilon$-DP then
    Draw $N_1, N_2 \overset{\text{iid}}{\sim} \text{Tulap}(0, \exp(-\epsilon), 0)$;
    Set $\psi_N(t) = \frac{[1-\exp(-\epsilon)]^2[\exp(-it/2)-\exp(it/2)]}{it[1-\exp(it-\epsilon)][1-\exp(-it-\epsilon)]}$;
end
if $\mu$-GDP then
    Draw $N_1, N_2 \overset{\text{iid}}{\sim} N(0, 1/\mu^2)$;
    Set $\psi_N(t) = \exp(-t^2/(2\mu^2))$;
end
Set $\psi_{Y\sim\theta}(t) = ((1-\theta) + \theta\exp(it))^m$ and $\psi_{X\sim\theta}(t) = ((1-\theta) + \theta\exp(it))^n$;
Set $\hat{X} = X + N_1$ and $\hat{Y} = Y + N_2$;
Set $T = \hat{Y}/m - \hat{X}/n$;
Set $\hat{\theta} = \min\{\max\{\frac{\hat{X}+\hat{Y}}{m+n}, 0\}, 1\}$;
Set $\psi_{T\sim\theta}(t) = \psi_{Y\sim\theta}(t/m)\psi_{X\sim\theta}(-t/n)\psi_N(t/m)\psi_N(-t/n)$;
Output $p$-value and summary values: $p = 1 - \int_0^\infty \frac{\text{Im}(\exp(itT)\psi_{T\sim\hat{\theta}}(t))}{t}\,dt$, $X + N_1$, $Y + N_2$
```

REMARK 5.7. While the p-value generated from Algorithm 1 may seem complex, it is relatively easy to implement. For instance in R, the command `integrate` can perform an accurate numerical integral. Another strength of Algorithm 1 is that the running time does not depend on the sample size m or n, whereas the semiprivate test runs in $O(m)$ time.

REMARK 5.8. Algorithm 1 can be viewed as an exact evaluation of a parametric bootstrap, where we by-pass the need for sampling by numerically computing the cdf. As such, we avoid the additional error and running time produced by the Monte Carlo sampling.

REMARK 5.9. While we focus on the one-sided hypothesis $H_0 : \theta_X \geq \theta_Y$ versus $H_1 : \theta_X < \theta_Y$, the test of Algorithm 1 can be easily modified to produce a "two-sided" test for $H_0 : \theta_X = \theta_Y$ versus $H_1 : \theta_X \neq \theta_Y$. Call p the one-sided p-value from Algorithm 1. Then $p_2 = 2\min\{p, 1-p\}$ is a p-value for the two-sided test. This procedure is an example of a *Bonferroni correction* or an *intersection-union test* (Casella and Berger (2002), Section 8.2.3).

5.3. *Simulations.* In this section, we perform several simulations to compare the performance of our proposed DP test to other competing DP tests, the semiprivate UMPU test, as well as popularly used nonprivate tests. While our results can be applied to arbitrary f-DP, we only run our simulations for $(\epsilon, 0)$-DP as this privacy definition is commonly used and introduces noise that is difficult to incorporate.

In Section 5.3.3, we consider the empirical power of the tests, and show that the inversion DP test out-performs other DP tests, and by comparing against the semiprivate test with privacy budget $\epsilon/\sqrt{2}$, show that it is observed to be more powerful than any $(\epsilon/\sqrt{2})$-DP test (see Remark 5.10 for the intuition behind the factor of $1/\sqrt{2}$). In Section 5.3.1, we consider the type I error of the various tests, and show that the observed type I error of the inversion test is more accurate than the commonly used nonprivate normal approximation test. We also show that naïve DP normal approximation tests have unacceptably inaccurate empirical type

I errors. In Section 5.3.2, we plot the empirical cumulative distribution functions of the p-values from the various tests demonstrating from another perspective that the proposed test has accurate type I error.

5.3.1. *Type I error.* The first simulation that we consider, and one of the most important, demonstrates the reliability of the type I error guarantees of our proposed test against alternative tests. Recall that in the best practices of scientific research, many approximate statistical tests are widely used and accepted. For instance, most hypothesis testing tools are based on asymptotic theory, such as the central limit theorem, which approximates the sampling distribution. As such, many widely used tests do not have exact type I error guarantees, but the error of these tests has been determined to be small enough for practical purposes. In Section 5.2, our proposed inversion-based test also involves an approximation to the sampling distribution. We demonstrate in the following simulation that the type I errors of this proposed test are more accurate than the widely accepted normal approximation test.

For the simulation, we measure the empirical type I error as the null θ_0 takes values in $\{0.05, 0.1, \dots, 0.95\}$ and sample sizes are set to $m = n = 30$, based on 20,000 replicates for each θ_0 value. We consider two values for the nominal type I error: in the left plot of Figure 3 we set $\alpha = 0.01$ and in the right plot of Figure 3 we set $\alpha = 0.05$. The dotted horizontal lines represent a 95% Monte Carlo confidence interval assuming that the true type I error is equal to the nominal level. As there are 19 unique theta values, if a curve crosses these thresholds more than once, this is evidence that the type I error is not appropriately calibrated. For this simulation, we only consider approximate tests as the nonprivate UMPU test and the semiprivate test have perfectly calibrated type I errors.

In red is the classic normal approximation test, described in Section 5.2. Such approximations are often considered accurate enough when the sample sizes n and m are greater than 30. Some rules of thumb for this problem require that there are at least 8 successes and failures in each group for the approximation to be accurate enough (Akritas (2015), page 321). We see in the left plot of Figure 3 that while this test has reasonable empirical type I error for moderate values of θ_0, the test is overly conservative for extreme values of θ_0. In the right plot of Figure 3, we see that the normal approximation test is much less reliable in this setting, with seven of the nineteen values outside of the 95% confidence region. We see that at extreme values of θ_0, the actual type I error rates are much higher than the nominal level, resulting

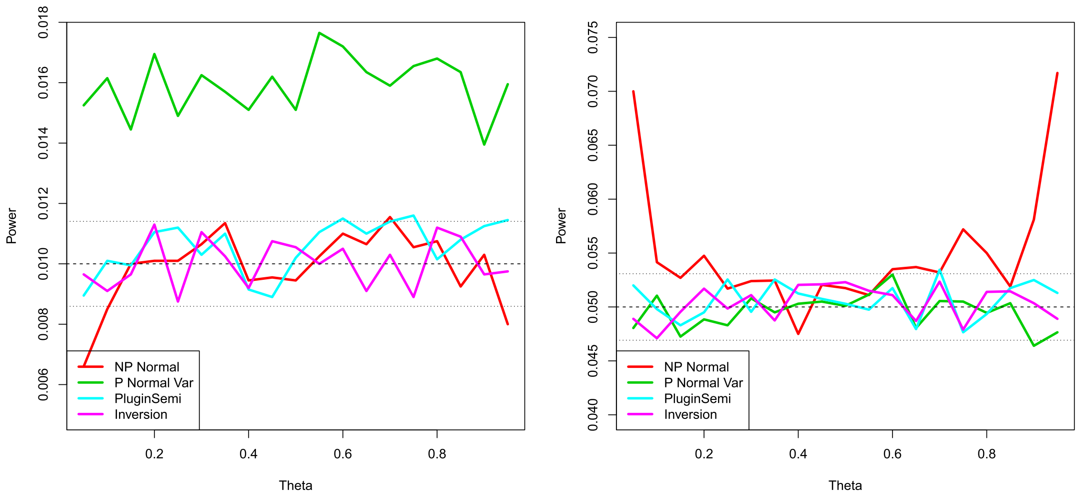


FIG. 3. *Empirical type I error as* θ_0 *varies in* $\{0.05, 0.1, \dots, 0.95\}$. *The nominal* α *level is* 0.01 (*left*) *and* 0.05 (*right*). $m = n = 30$, $\epsilon = 0.1$, *and results are over* 20,000 *replicates for each* θ_0.

in excessive false positives. It is interesting that the type I errors are over-conservative when $\alpha = 0.01$ and inflated when $\alpha = 0.05$. In general, it is hard to predict whether in a particular setting the type I errors will be too high or too low.

In green is an ϵ-DP normal approximation test, proposed by Karwa and Vadhan (2018) which is analogous to the one-sample test of Vu and Slavković (2009). See Appendix E of the Supplementary Material (Awan and Vadhan (2023)) for a description of the method. While the empirical type I errors of this test are acceptable when $\alpha = 0.05$, we see that for $\alpha = 0.01$, the empirical type I error is approximately 0.016 and is entirely outside the confidence region.

In light blue is an ϵ-DP, which splits the budget between privatizing $T = Y - X$ and $Z = X + Y$, and plugs in the results into the semiprivate test of Theorem 5.2. The test is described in Algorithm 3, which appears in Appendix E of the Supplementary Material (Awan and Vadhan (2023)). The empirical type I errors for the plugin test are slightly higher than expected, crossing the confidence band three times in the left plot and once in the right plot, but are much more reliable than the normal approximation tests discussed above.

Finally, in magenta is the inversion-based test of Algorithm 1. The empirical type I errors of the inversion-based test lie entirely within the confidence bands for both settings of α. This indicates that for the settings of these simulations, the type I errors of the inversion test are indistinguishable from the nominal level, and are much more accurate than the classic normal approximation test or a DP normal approximation test, such as in Vu and Slavković (2009).

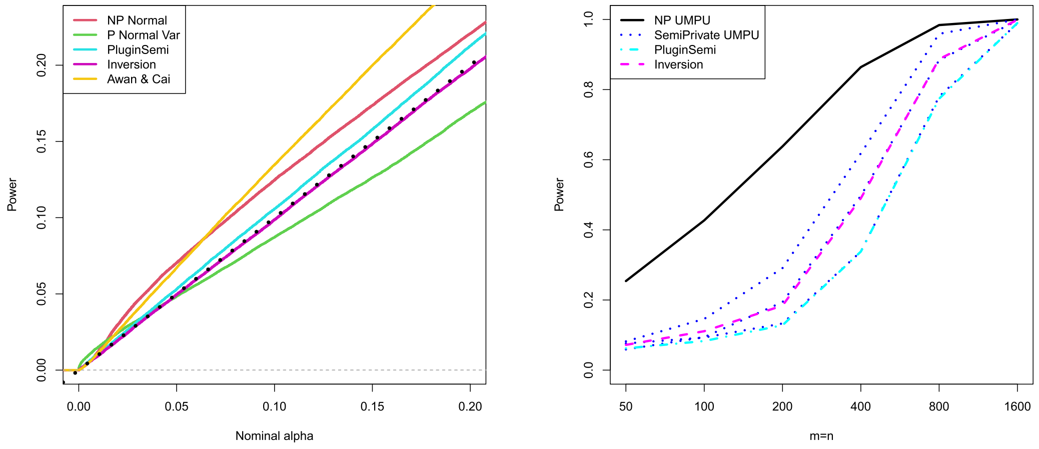
5.3.2. *P-Values.* In this section, we consider the empirical cumulative distribution function of the p-values, while holding θ_0 fixed. This can be interpreted as varying the nominal α value on the x-axis, with the empirical type I error on the y-axis. This differs from the previous simulation, where we varied the null value of θ along the x-axis, but left the nominal value of α fixed. Combined with the previous results, this simulation gives a more complete picture of how accurate the type I errors are, for a spectrum of nominal α values.

For the simulation, we set $\theta_0 = 0.95$, $n = 30$, $m = 40$, and $\epsilon = 0.1$. We chose to investigate $\theta_0 = 0.95$ since the type I errors in Section 5.3.1 were found to be more inaccurate for extreme values of θ_0. The results are based on 100,000 replicates with these settings. The simulation includes the same tests as in Section 5.3.1, marked with the same color scheme, as well as a test based on the simulation-based method of Awan and Cai (2020). Included is a dotted black line of intercept 0 and slope 1, which represents perfectly calibrated type I error rates.

We see that for these simulation settings, the nonprivate normal approximation test has inflated type I errors for nominal α values between 0.02 and 0.2. The DP normal approximation test has inflated type I error rates for nominal alpha values below 0.05, and deflated type I error rates for larger values of α. The plugin test also has inflated type I errors in this setting, while not as extreme as the normal approximate test. Finally, the curve for the inversion test is visually indistinguishable from the dotted black line, indicating that this tests has well-calibrated type I errors for this simulation setting, much improved over the other approximate tests considered here.

Awan and Cai (2020) tackled the same DP testing problem, and also based their test on adding Tulap noise to both X and Y. They implement their test a simulation-based algorithm, which they argue gives asymptotically accurate type I errors. We include their test in this section for comparison, and while Awan and Cai (2020) advocated this approach in large samples, we see in the left plot of Figure 4 that it has greatly inflated type I errors for the smaller sample sizes considered in this simulation.

5.3.3. *Power.* Finally, we compare the power of our candidate tests. We use the semiprivate UMPU test as a baseline for comparison: recall from Theorem 5.2 that the semiprivate test has perfectly calibrated type I errors, and is uniformly more powerful than any DP unbiased test. As such, it serves as an upper bound on the power of the other candidate tests. We



(a) Empirical cdf of the p-values. $n = 30$, $m = 40$, $\epsilon = .1$, $\theta_0 = .95$, and results are based on 100,000 replicates.

(b) Empirical power at $\theta_X = .5$ and $\theta_Y = .6$, while $m = n$ varies on the x-axis. The privacy parameter is $\epsilon = .1$, and the results are averaged over 1000 replicates for each sample size.

FIG. 4. *Simulation results comparing the p-values and power of various tests.*

will see that the inversion test (with $\epsilon = 0.1$) has power similar to the semiprivate UMPU with $\epsilon = (0.1/\sqrt{2})$, indicating that its power cannot be beaten by the most powerful $(\epsilon/\sqrt{2})$-DP unbiased test.

For the simulation, we vary the sample size $n = m$ along the x-axis and measure the empirical power on the y-axis, at a nominal α level of 0.05. The privacy parameter is set to $\epsilon = 0.1$ and the results are based on 1000 replicates for each sample size. In black is the nonprivate UMPU test, described Appendix C of the Supplementary Material, which is guaranteed to be more powerful than any of the private tests considered in this paper. The dotted dark blue curve is the semiprivate UMPU test of Section 5.1. Since the semiprivate UMPU has a weaker privacy guarantee than DP, this test gives an upper bound on the power of any DP test. We also include the semiprivate test implemented with $\epsilon = 0.1/\sqrt{2}$ and $\epsilon = 0.1/2$, with the same color and line scheme. We see that the plugin test, appearing in light blue, has similar power as the semiprivate test with $\epsilon = 0.1/2$, indicating that this test is more powerful than any $\epsilon/2$ test. In magenta, we have the inversion-based test, which has similar power as the semiprivate test with $\epsilon = 0.1/\sqrt{2}$, indicating that it is more powerful than any $\epsilon/\sqrt{2}$ test.

REMARK 5.10. That the inversion test has comparable power to the semiprivate test with $\epsilon/\sqrt{2}$ can be understood as follows: the semiprivate test is based on the test statistic $S = Y - X + N$, where N is a Tulap random variable. On the other hand, the inversion test is based on $\widetilde{X} = X + N_1$ and $\widetilde{Y} = Y + N_2$. If we tried to approximate the test statistic S using $\widetilde{X}$ and $\widetilde{Y}$, we end up with $\widetilde{S} = Y - X + (N_1 - N_2)$. If the same privacy parameters are used for N and N_1, N_2, then $\mathrm{Var}(N_1 - N_2) = 2\,\mathrm{Var}(N)$. By decreasing the privacy parameter of N to $\epsilon/\sqrt{2}$, we obtain equality of the variances.

6. Discussion.

In this paper, we proposed the new concept *canonical noise distribution*, which is an alternative to previous notions of an optimal noise adding mechanism for privacy. We showed that a CND is a fundamental concept in f-DP, with connections to optimal private hypothesis tests. Using CNDs and our results on f-DP hypothesis tests, we also developed a novel DP test for the difference-of-proportions, which was shown to have accurate type I errors and nearly optimal power. The introduction of CNDs also raises several questions:

It was noted in Section 3 that a CND is generally not unique for a given tradeoff function. While the construction in Definition 3.7 always results in a CND, and is easily sampled, it may not be the most natural CND. For example, when applied to the tradeoff function G_1, we see in Figure 2 that the CND constructed by Definition 3.7 has a nondifferentiable pdf. On the other hand, $N(0, 1)$ is also a CND for G_1 which has a smooth pdf. One may wonder if there is a more natural CND construction which recovers $N(0, 1)$ in the case of G_1, as well as whether there is a CND for $f_{\epsilon,\delta}$ which has a smooth pdf. A recent paper that builds upon the present work, Awan and Dong (2022), partially answers these questions, showing that in some cases it is possible to construct a *log-concave* CND, which recovers $N(0, 1)$ in the case of G_1; surprisingly, Awan and Dong (2022) also show that the Tulap distribution is the *unique* CND for $f_{\epsilon,0}$, ruling out the possibility of a smooth CND for $f_{\epsilon,0}$.

Another question is whether there is a natural and meaningful extension of CNDs to vector-valued statistics. The follow-up paper, Awan and Dong (2022), partially answers this question, giving a definition of a multivariate CND and general constructions under various assumptions. While they show that there exists multivariate CNDs for many general classes of tradeoff functions, including GDP, Laplace-DP, and (ϵ, δ)-DP, they also prove that there is *no* multivariate CND for $f_{\epsilon,0}$.

While this paper focused on the connection between CNDs and private hypothesis tests, it is an open question whether there are other fundamental optimality properties of CNDs. It was also noted in the Introduction that additive noise mechanisms often appear as a component of more complex DP mechanisms, and it is worth investigating whether CNDs can be used to optimize these other mechanisms for a particular f-DP guarantee.

The applications to DP hypothesis tests also raise many interesting questions. In general, there always exists a most powerful DP test for any composite null and simple alternative, as shown in Proposition B.1 of the Supplementary Material (Awan and Vadhan (2023)), which can be expressed as the solution to a convex optimization problem. However, solving the optimization problem is computationally burdensome for all but the simplest of problems. In Theorem 4.8, we derived a closed-form expression for the most powerful DP test. Do there exist closed-form expressions for other UMP DP tests to avoid computational optimization?

We also introduced the semiprivate framework which allowed us to derive an upper bound on the power of any unbiased f-DP test. Can this framework be applied to other DP testing problems to derive similar bounds? We also remarked that the semiprivate framework may be useful to better understand the privacy guarantee of mechanisms where certain statistics are privatized, whereas others are reported exactly, such as in the 2020 Decennial US Census—it remains to be seen whether the semiprivate framework can give new results or new understanding in these settings.

Acknowledgments. The authors thank the reviewers, the Associate Editor, and Editor Mammen for their careful reading and constructive comments, which greatly improved the quality of this paper. The first author is grateful for the hospitality of the Center for Research on Computation and Society at Harvard University, where part of this work was completed.

Funding. This work was supported by Cooperative Agreement CB16ADR0160001 from the U.S. Census Bureau. The first author was also supported in part by NSF Award Numbers SES-1534433, SES-1853209, and SES-2150615. The second author was also supported in part by a Simons Investigator Award.

SUPPLEMENTARY MATERIAL

Supplement to "Canonical noise distributions and private hypothesis tests" (DOI: 10.1214/23-AOS2259SUPPA; .pdf). This document contains additional background material, as well as the technical proofs and lemmas needed to establish the results of this paper.

Code for "Canonical noise distributions and private hypothesis tests" (DOI: 10.1214/23-AOS2259SUPPB; .zip). This file contains the R code used to implement the simulations and generate the figures of this paper.

REFERENCES

ABADI, M., CHU, A., GOODFELLOW, I., MCMAHAN, H. B., MIRONOV, I., TALWAR, K. and ZHANG, L. (2016). Deep learning with differential privacy. In *Proceedings of the* 2016 *ACM SIGSAC Conference on Computer and Communications Security* 308–318.

ACHARYA, J., SUN, Z. and ZHANG, H. (2018). Differentially private testing of identity and closeness of discrete distributions. *Adv. Neural Inf. Process. Syst.* **31**.

AKRITAS, M. (2015). *Probability and Statistics with R*. Pearson, New York.

ALIAKBARPOUR, M., DIAKONIKOLAS, I. and RUBINFELD, R. (2018). Differentially private identity and equivalence testing of discrete distributions. In *International Conference on Machine Learning* 169–178. PMLR, Stockholm, Sweden.

AWAN, J. and CAI, Z. (2020). One step to efficient synthetic data. arXiv preprint. Available at arXiv:2006.02397.

AWAN, J. and DONG, J. (2022). Log-concave and multivariate canonical noise distributions for differential privacy. *Adv. Neural Inf. Process. Syst.* **35**.

AWAN, J. and SLAVKOVIĆ, A. (2018). Differentially private uniformly most powerful tests for binomial data. *Adv. Neural Inf. Process. Syst.* **31** 4208–4218.

AWAN, J. and SLAVKOVIĆ, A. (2020). Differentially private inference for binomial data. *J. Priv. Confid.* **10**.

AWAN, J. and SLAVKOVIĆ, A. (2021). Structure and sensitivity in differential privacy: Comparing K-norm mechanisms. *J. Amer. Statist. Assoc.* **116** 935–954. MR4270035 https://doi.org/10.1080/01621459.2020.1773831

AWAN, J. and VADHAN, S. (2023). Supplement to "Canonical noise distributions and private hypothesis tests." https://doi.org/10.1214/23-AOS2259SUPPA, https://doi.org/10.1214/23-AOS2259SUPPB

BARRIENTOS, A. F., REITER, J. P., MACHANAVAJJHALA, A. and CHEN, Y. (2019). Differentially private significance tests for regression coefficients. *J. Comput. Graph. Statist.* **28** 440–453. MR3974892 https://doi.org/10.1080/10618600.2018.1538881

BLACKWELL, D. (1950). Comparison of experiments. Technical report, Howard Univ., Washington, United States.

BUN, M. and STEINKE, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography. Part I. Lecture Notes in Computer Science* **9985** 635–658. Springer, Berlin. MR3591832 https://doi.org/10.1007/978-3-662-53641-4_24

BUN, M., DWORK, C., ROTHBLUM, G. N. and STEINKE, T. (2018). Composable and versatile privacy via truncated CDP. In *STOC'*18*—Proceedings of the* 50*th Annual ACM SIGACT Symposium on Theory of Computing* 74–86. ACM, New York. MR3826235

BUN, M., KAMATH, G., STEINKE, T. and WU, S. Z. (2019). Private hypothesis selection. *Adv. Neural Inf. Process. Syst.* **32**.

CAI, B., DASKALAKIS, C. and KAMATH, G. (2017). Priv'it: Private and sample efficient identity testing. In *International Conference on Machine Learning* 635–644. PMLR, Sydney, Australia.

CANONNE, C. L., KAMATH, G., MCMILLAN, A., SMITH, A. and ULLMAN, J. (2019). The structure of optimal private tests for simple hypotheses. In *STOC'*19*—Proceedings of the* 51*st Annual ACM SIGACT Symposium on Theory of Computing* 310–321. ACM, New York. MR4003341

CASELLA, G. and BERGER, R. L. (2002). *Statistical Inference*. Duxbury, N. Scituate.

CHAUDHURI, K., MONTELEONI, C. and SARWATE, A. D. (2011). Differentially private empirical risk minimization. *J. Mach. Learn. Res.* **12** 1069–1109. MR2786918

COLQUHOUN, D. (2017). The reproducibility of research and the misinterpretation of p-values. *R. Soc. Open Sci.* **4** 171085. MR3744762 https://doi.org/10.1098/rsos.171085

DANG, X., KEETON, S. L. and PENG, H. (2009). A unified approach for analyzing exchangeable binary data with applications to developmental toxicity studies. *Stat. Med.* **28** 2580–2604. MR2750310 https://doi.org/10.1002/sim.3638

DONG, J., ROTH, A. and SU, W. J. (2022). Gaussian differential privacy. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 3–37. MR4400389

DUCHI, J. C., JORDAN, M. I. and WAINWRIGHT, M. J. (2018). Minimax optimal procedures for locally private estimation. *J. Amer. Statist. Assoc.* **113** 182–201. MR3803452 https://doi.org/10.1080/01621459.2017.1389735

DWORK, C. and LEI, J. (2009). Differential privacy and robust statistics. In *STOC'*09*—Proceedings of the* 2009 *ACM International Symposium on Theory of Computing* 371–380. ACM, New York. MR2780083

DWORK, C. and ROTH, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9** 211–407. MR3254020 https://doi.org/10.1561/0400000042

DWORK, C., MCSHERRY, F., NISSIM, K. and SMITH, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography. Lecture Notes in Computer Science* **3876** 265–284. Springer, Berlin. MR2241676 https://doi.org/10.1007/11681878_14

DWORK, C., NAOR, M., REINGOLD, O., ROTHBLUM, G. N. and VADHAN, S. (2009). On the complexity of differentially private data release. In *STOC'*09*—Proceedings of the* 2009 *ACM International Symposium on Theory of Computing* 381–390. ACM, New York. MR2780084

FOG, A. (2008). Sampling methods for Wallenius' and Fisher's noncentral hypergeometric distributions. *Comm. Statist. Simulation Comput.* **37** 241–257. MR2422884 https://doi.org/10.1080/03610910701790236

GABOARDI, M., LIM, H., ROGERS, R. and VADHAN, S. (2016). Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *Proceedings of the* 33*rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.). *Proceedings of Machine Learning Research* **48** 2111–2120. PMLR, New York.

GABOARDI, M. and ROGERS, R. (2018). Local private hypothesis testing: Chi-square tests. In *Proceedings of the* 35*th International Conference on Machine Learning* (J. Dy and A. Krause, eds.). *Proceedings of Machine Learning Research* **80** 1626–1635. PMLR, Stockholm, Sweden.

GAO, J., GONG, R. and YU, F.-Y. (2022). Subspace differential privacy. In *Proceedings of the AAAI Conference on Artificial Intelligence* **36** 3986–3995.

GENG, Q. and VISWANATH, P. (2015). The optimal noise-adding mechanism in differential privacy. *IEEE Trans. Inf. Theory* **62** 925–951. MR3455907 https://doi.org/10.1109/TIT.2015.2504967

GHOSH, A., ROUGHGARDEN, T. and SUNDARARAJAN, M. (2012). Universally utility-maximizing privacy mechanisms. *SIAM J. Comput.* **41** 1673–1693. MR3029267 https://doi.org/10.1137/09076828X

HALL, R., RINALDO, A. and WASSERMAN, L. (2013). Differential privacy for functions and functional data. *J. Mach. Learn. Res.* **14** 703–727. MR3033345

HARDT, M. and TALWAR, K. (2010). On the geometry of differential privacy. In *STOC'*10*—Proceedings of the* 2010 *ACM International Symposium on Theory of Computing* 705–714. ACM, New York. MR2743320

HARKNESS, W. L. (1965). Properties of the extended hypergeometric distribution. *Ann. Math. Stat.* **36** 938–945. MR0182073 https://doi.org/10.1214/aoms/1177700066

KAIROUZ, P., OH, S. and VISWANATH, P. (2017). The composition theorem for differential privacy. *IEEE Trans. Inf. Theory* **63** 4037–4049. MR3677761 https://doi.org/10.1109/TIT.2017.2685505

KAKIZAKI, K., FUKUCHI, K. and SAKUMA, J. (2017). Differentially private chi-squared test by unit circle mechanism. In *International Conference on Machine Learning* 1761–1770. PMLR, Sydney, Australia.

KARWA, V. and VADHAN, S. P. (2017). Finite sample differentially private confidence intervals. Available at arXiv:1711.03908.

KARWA, V. and VADHAN, S. (2018). Private correspondence.

KIFER, D. and ROGERS, R. (2016). A new class of private chi-square tests. In *Proceedings of the* 20*th International Conference on Artificial Intelligence and Statistics. AISTATS* **17** 991–1000.

MIRONOV, I. (2017). Rényi differential privacy. In 2017 *IEEE* 30*th Computer Security Foundations Symposium* (*CSF*) 263–275. IEEE Press, New York.

MIRSHANI, A., REIMHERR, M. and SLAVKOVIĆ, A. (2019). Formal privacy for functional data with Gaussian perturbations. In *International Conference on Machine Learning* 4595–4604. PMLR, Long Beach, CA.

REIMHERR, M. and AWAN, J. (2019). Elliptical perturbations for differential privacy. *Adv. Neural Inf. Process. Syst.* **32**.

ROBINS, J. M., VAN DER VAART, A. and VENTURA, V. (2000). Asymptotic distribution of *p* values in composite null models. *J. Amer. Statist. Assoc.* **95** 1143–1156. MR1804240 https://doi.org/10.2307/2669750

SCHERVISH, M. J. (2012). *Theory of Statistics*. Springer, Berlin.

SHEFFET, O. (2017). Differentially private ordinary least squares. In *Proceedings of the* 34*th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.). *Proceedings of Machine Learning Research* **70** 3105–3114. PMLR, Sydney, Australia.

SOLEA, E. (2014). Differentially private hypothesis testing for normal random variables. Master's thesis, The Pennsylvania State Univ.

SURESH, A. T. (2021). Robust hypothesis testing and distribution estimation in Hellinger distance. In *International Conference on Artificial Intelligence and Statistics* 2962–2970. PMLR, San Diego, CA.

UHLER, C., SLAVKOVIĆ, A. and FIENBERG, S. (2013). Privacy-preserving data sharing for genome-wide association studies. *J. Priv. Confid.* **5**.

VU, D. and SLAVKOVIĆ, A. (2009). Differential privacy for clinical trial data: Preliminary evaluations. In 2009 *IEEE International Conference on Data Mining Workshops* 138–143. IEEE Press, New York.

WANG, Y., KIFER, D., LEE, J. and KARWA, V. (2018). Statistical approximating distributions under differential privacy. *J. Priv. Confid.* **8**.

WANG, Y., LEE, J. and KIFER, D. (2015). Revisiting differentially private hypothesis tests for categorical data. Available at arXiv:1511.03376.

WASSERMAN, L. and ZHOU, S. (2010). A statistical framework for differential privacy. *J. Amer. Statist. Assoc.* **105** 375–389. MR2656057 https://doi.org/10.1198/jasa.2009.tm08651

WASSERSTEIN, R. L. and LAZAR, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose [Editorial]. *Amer. Statist.* **70** 129–133. MR3511040 https://doi.org/10.1080/00031305.2016.1154108

ZHANG, J., ZHANG, Z., XIAO, X., YANG, Y. and WINSLETT, M. (2012). Functional mechanism: Regression analysis under differential privacy. *Proc. VLDB Endow.* **5**.