Uncertainty Quantification and Inference under Differential Privacy

by

Ogonnaya Michael Romanus

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 9, 2025

Keywords: Differential privacy, sample proportions, simulation-based inference, conformal prediction

Approved by

Roberto Molinari, Chair, Assistant Professor of Mathematics and Statistics
Aleksandra (Sesa) Slavkovic, Professor of Statistics, Penn State University
Mark Carpenter, Professor of Mathematics and Statistics
Jingyi (Ginny) Zheng, Associate Professor of Mathematics and Statistics
Yang Zhou, Associate Professor of Computer Science and Software Engineering
Cheryl Seals, Professor of Computer Science and Software Engineering

Abstract

Advances in technology have led to the proliferation of adversarial techniques capable of undermining traditional data security protocols designed to ensure data privacy. Such breaches have resulted in significant financial and reputational costs for companies, particularly in addressing privacy violations and complying with regulatory penalties. Strict privacy regulations, including the General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and the Health Insurance Portability and Accountability Act (HIPAA), further necessitate the integration of privacy-preserving mechanisms throughout the data modeling pipeline to avoid legal repercussions. Differential privacy (DP) and its variants have emerged as the gold standard for mathematically quantifying and mitigating privacy loss in sensitive data analysis. However, the noise introduced to achieve DP often distorts data structure and statistical properties, rendering traditional inferential methods inapplicable.

This dissertation develops efficient and robust methods for privacy-preserving data analysis under DP guarantees. Chapter 1 introduces a flexible simulation-based framework that constructs the distribution of a DP estimator by matching observed DP statistics with simulated counterparts. The method is applied to one- and two-sample confidence interval estimation, hypothesis testing, chi-square tests of independence, and logistic regression with categorical predictors. Extensive simulations demonstrate that the proposed approach performs comparably to state-of-the-art methods. Applications to real-world datasets further confirm that inferences drawn from the DP statistical tests align with those obtained via conventional non-private methods.

Chapter 2 presents a binary search-based DP algorithm for conformal prediction in classification tasks. Extensive empirical evaluations reveal that this method outperforms the only existing alternative in both efficiency and predictive performance. Evaluations on benchmark datasets (CIFAR-10, ImageNet, and CoronaHack) demonstrate its superiority over the current state-of-the-art method across realistic and practical scenarios.

Acknowledgments

First and foremost, I give all glory to the Almighty God, my Lord and Savior Jesus Christ, the keeper of my soul and my eternal hope. I am profoundly grateful for the gift of life and the divine strength that sustained me through this Ph.D. journey.

My deepest gratitude goes to my advisor, Dr. Roberto Molinari, whose invaluable guidance, patience, and encouragement shaped this research. His insightful feedback and steadfast support were instrumental in bringing this work to fruition.

I extend sincere appreciation to my committee members: Dr. Aleksandra (Sesa) Slavkovic, Dr. Mark Carpenter, Dr. Jingyi (Ginny) Zheng, Dr. Yang Zhou, Dr. Cheryl Seals, and Dr. George Flowers for their thoughtful critiques and constructive suggestions that elevated this dissertation.

This work was enriched by my brilliant collaborators, Dr. Younes Boulaguiem and Dr. Stéphane Guerrier, whose expertise and camaraderie fueled our stimulating discussions on simulation-based indirect inference.

To my fellow graduate students (Ukamaka, Chinedu, Monday, Hewan, Yagmur, Masuzyo, and others) and the staff of the Department of Mathematics and Statistics—thank you for creating such a vibrant intellectual community.

My family's unconditional love and sacrifices formed the bedrock of my perseverance. To my parents, siblings, and loved ones—your belief in me never wavered, even in my darkest moments.

I am blessed by the steadfast friendship of Peter, Jerry, Kelechi, and Francis, whose support and prayers carried me forward.

To Ginika, my beloved fiancée, your unwavering love, ceaseless prayers, and boundless encouragement were my sanctuary through every challenge. This achievement is as much yours as it is mine, and I cherish the journey ahead as we build our future together.

I dedicate this dissertation to my beloved mother, Ijeago Alice Okoro, whose unwavering prayers, boundless love, endless sacrifices, enduring patience, and profound kindness have been my guiding light.

Table of Contents

List of Figures

List of Tables

Chapter 1

Simulation-Based Differentially Private Statistical Inference For Categorical Data

1.1 Introduction

Differential Privacy (DP), along with its various adaptations in recent years, aims to safeguard individual privacy by providing a mathematically rigorous guarantee of protection [19]. This guarantee is achieved in practice by injecting probabilistic (random) noise into different stages of the data processing and analysis pipeline. The privacy mechanisms implementing DP can act directly on the raw data itself or on the outputs derived from the data, such as summary statistics, model parameters, or query results. By quantifying the privacy risk in the presence of arbitrary external information, DP ensures strong protection against re-identification attacks. However, this privacy-preserving approach comes at a cost: the introduction of noise inevitably affects the utility of the data output. The extent of this impact varies depending on the analysis task and the specific privacy mechanism used, but generally, DP induces statistical bias, increases variability, and complicates statistical inference by affecting key tasks such as hypothesis testing and uncertainty quantification [68, 71].

A fundamental characteristic of DP is that it implies that the data is affected by two distinct sources of randomness: the inherent randomness from sampling of the original dataset and the additional randomness imposed by the privacy mechanism. This dual source of randomness poses significant challenges for statistical inference. Once noise is introduced to satisfy privacy constraints, the original data become inaccessible to the analyst, meaning that standard inferential procedures based on classical likelihood theory may no longer be applicable. Moreover, the sampling distribution of privatized data can become highly complex and difficult to characterize analytically, particularly when noise is introduced in a non-trivial manner [74]. As a result,

standard statistical methods may not account for the distortions introduced by DP, leading to biased estimators, inaccurate confidence intervals, and unreliable hypothesis testing results.

To illustrate these challenges, consider a setting in which we have a probabilistic model f_θ that describes the data-generating process for the sensitive dataset $\mathcal{D}$, where $\theta \in \Theta \subset \mathbb{R}^p$ represents the parameter of interest. Furthermore, suppose that a privacy mechanism $m(s|\mathcal{D})$ is applied to privatize the data, producing the observed privatized summary s. A natural approach to make an inference on θ in this setting would be to compute the marginal likelihood of θ given s. However, computing this marginal likelihood is often infeasible in practice, as it requires integrating over all possible realizations of the original dataset $\mathcal{D}$ [7]. This intractability increases when dealing with high-dimensional data and with the complexity of the privacy mechanism, particularly when the noise is non-Gaussian or correlated. Consequently, conventional likelihood-based inference methods may be impractical within the DP framework.

Given these limitations, inference methods that bypass the computation of the likelihood have been increasingly investigated for DP-constrained statistical analysis. Among these, simulation-based inference techniques, such as parametric bootstrap and approximate Bayesian computation, have been widely explored to handle uncertainty quantification in DP settings [61, 23, 25, 36]. These methods leverage the ability to sample from the generative model $f_\theta(\mathcal{D})$ and the privacy mechanism $m(s|\mathcal{D})$ to approximate posterior distributions or construct valid confidence intervals without requiring explicit likelihood evaluation. Moreover, simulation-based inference methods align naturally with the DP framework, since the privacy mechanism $m(s|\mathcal{D})$ can be made publicly available without compromising individual privacy. This transparency allows for the development of robust inference procedures while ensuring that privacy constraints are rigorously maintained.

However, these methods also introduce computational burdens as they typically require extensive sampling to achieve accurate approximations. With this in mind, also in this work we approach the problem of statistical inference on privatized categorical data through a different paradigm than the likelihood-based one, and notably we use the simulated moment-matching (extremum estimator) framework [see e.g. 32]. In fact, we make use of the approach proposed

in [55] which is based on the statistical properties of moment-matching simulation-based approaches that were studied and refined in [33], [34] and [76]. In this work, we therefore adopt this moment-matching approach to deliver valid DP confidence intervals (and hypothesis tests) for parameters of interest. In particular, given the discrete nature of the matching solution, we use a fiducial argument to solve the optimization problem [36]. In contrast to existing methods which are specific to certain types of statistical tasks, this proposed approach can be applied to a wider variety of applications on categorical data, such as, among others, two sample tests for proportions, chi-squared tests or (saturated) logistic models with categorical predictors [see e.g. 1].

1.2 Related Work

Over the past decade, significant research has focused on statistical inference within the framework of DP. For example, the Bayesian framework offers a highly flexible approach for analyzing privatized data. Williams and McSherry [74] were the first to highlight that, as mentioned earlier, the marginal likelihood is computationally intractable. To address this challenge, various data augmentation MCMC methods have been introduced. For example, Bernstein and Sheldon [9, 10] demonstrated that when the distribution of a sufficient statistic is approximated as a normal distribution, a simple Gibbs sampler can be used to estimate the posterior distribution. Meanwhile, Ju et al. [40] developed a Metropolis-within-Gibbs sampler that directly targets the exact private posterior distribution, although its effectiveness is limited unless conjugate priors and small sample sizes are used. For simpler cases, Gong [31] introduced an independent and identically distributed (i.i.d.) sampling method based on rejection sampling. Alternatively, Karwa [42]proposed a variational approximation specifically tailored for a naive Bayes classifier.

Beyond Bayesian methods, frequentist approaches have also been explored for statistical inference on privatized data. Wang et al. [72] argued that conventional asymptotic approximations often perform poorly in finite samples when applied to privatized data. To address this, they proposed a modified asymptotic framework in which only the sufficient statistics are subject to asymptotic approximations, while the noise introduced by the privacy mechanism

remains unaltered. Another widely used frequentist technique involves the parametric bootstrap, which approximates the sampling distribution based on privatized data [24, 2]. However, Awan and Wang [8] demonstrated that the parametric bootstrap can lead to biased inference, resulting in inadequate coverage and incorrect type I error rates. As a more reliable alternative, they introduced a simulation-based inference approach that ensures conservative control over coverage and type I errors, although at the cost of increased computational complexity.

Specifically, discussing categorical (nominal) data, Vu and Slavkovic [69] were among the first to discuss statistical hypothesis testing within a DP framework in these settings. In particular, using the Laplace DP mechanism, they studied the performance of standard statistical tests (e.g., tests for proportions and independence in 2×2 tables) under the constraint of DP showing how these are highly affected by DP mechanisms and that large sample sizes are needed to achieve reasonable performance. Additionally, they utilized normal approximation to the sampling distribution to calibrate the type I error. Building on this work, Gaboardi et al. [27] expanded these techniques to multinomial data, employing Monte Carlo methods to control and estimate the type I error. In contrast, Wang et al. [73] formulated asymptotic distributions for their tests and validated the reliability of the type I error through simulations. Kifer and Rogers [45] devised private chi-square tests for goodness-of-fit and identity problems, ensuring that the asymptotic distributions of the test statistics, after accounting for added noise for privacy, align with those of classical chi-square tests. Furthermore, Uhlerop et al. [66] developed DP p-values for chi-squared tests of GWAS data and derived the exact sampling distribution of the noisy statistics.

To our knowledge, the first formal inference framework for categorical data was the uniformly most powerful (UMP) hypothesis test under DP for binomial data proposed in [6]. Their testing procedure relies on the Tulap distribution, which is a combination of discrete Laplace and uniform distributions. They also present an algorithm for generating a p-value to access one-sided hypotheses, demonstrating that the resulting p-value is the smallest ϵ-DP p-value for this test. Simultaneously, Kazan et al. [43] and Peña and Barrientos [58] introduced a similar framework for constructing a differentially private version of any non-private hypothesis test in a black-box manner, employing the Subsample-and-Aggregate technique proposed by

Nissim et al. [53]. Despite both methods starting with subsampling, the aggregation strategies differ. Kazan et al. [43] used the algorithm of Awan and Slavkovic [6] as the aggregation function, while Peña and Barrientos [58] designed a custom binomial test based on a randomized response-type method.

1.3 Preliminaries

Let $X \sim f_\theta^K$, where f_θ^K is a multinomial distribution with K categories (classes) and, as mentioned earlier, θ is the parameter of interest that characterizes the distribution f. Typically $\theta \in \Delta^{K-1} \subset \mathbb{R}^K$ is a vector defining the probabilities for each of the K classes and belonging to the $(K-1)$-dimensional simplex:

$$\Theta := \left\{ \theta \in \mathbb{R}_+^K : \sum_{k=1}^{K} \theta_k = 1, \theta_k \geq 0 \text{ for } k = 1, \ldots, K \right\}.$$

More specifically, we have that $\theta_k = \mathbb{P}(X = k)$ for $k = 1, \ldots, K$, which can also be expressed through other parameterizations based on known deterministic transformations of θ (for example, log-odds in logistic models). Based on this, letting $\mathcal{D} := \{x_1, \ldots, x_n\}$ represent the data with n samples, the maximum-likelihood (best unbiased) estimator for θ_k is given by the sample proportion:

$$\bar{\theta}_k = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i = k), \tag{1.1}$$

which therefore represents the estimated probability of X falling into class k. Without loss of generality, for the following sections, let us consider the case of $K = 2$ such that $X \in \{0, 1\}$ which implies that f_θ is a Binomial distribution with parameter θ corresponding to the probability of success (i.e. $\theta = \mathbb{P}(X = 1)$) with estimator $\bar{\theta}$. We will then extend our study to the case $K > 2$ further on.

Once we have the estimator $\bar{\theta}$, a common task is to build Confidence Intervals (CIs) for the true parameter, say θ_0, or to run some hypothesis tests on the true parameter using this estimator. To do so, the sampling distribution of the estimator is required, and, in the case of the sample proportion, there are exact as well as asymptotic distributions that are commonly

employed to achieve appropriate coverage and type I errors. However, under the constraint of pure ϵ-DP we cannot release the observed sample proportion, and a common mechanism used to guarantee ϵ-DP in these settings is to release the statistic delivered from the following additive mechanism:

$$\hat{\pi} := \bar{\theta} + \frac{\Delta_{\bar{\theta}}}{\epsilon} Z, \tag{1.2}$$

where Z is a zero-mean and unit variance random variable independent of the data; ϵ is the privacy budget and $\Delta_{\bar{\theta}} = \sup_{\mathcal{D},\mathcal{D}'} \|\bar{\theta}(\mathcal{D}) - \bar{\theta}(\mathcal{D}')\|$ is the sensitivity of the sample proportion $\bar{\theta}$ with respect to a norm $\|\cdot\|$ computed on n samples [see e.g. 20, 21]. More in detail, $\mathcal{D}$ and $\mathcal{D}'$ are two neighboring databases of size n (i.e. databases that differ only by one entry) and $\bar{\theta}(\mathcal{D})$ represents the sample proportion computed on database $\mathcal{D}$. Hence, the sensitivity $\Delta_{\bar{\theta}}$ measures the maximum amount by which the sample proportion can change by arbitrarily changing one observation from the database, which in the case of the sample proportion is notably $\Delta_{\bar{\theta}} = 1/n$. On this note, in this work we only consider *additive* DP mechanisms of the form in (1.2) and assume that there are no privacy constraints on the data size n, as is the case for the US Census and other surveys, for example. To satisfy ϵ-DP, the L_1 norm is used and Z is a vector with entries coming from Laplace$(0, 1)$. To satisfy ϵ-Gaussian Differential Privacy (GDP) the L_2 norm is used and $Z \sim N(0, 1)$. As a side note, it was shown in [29] that adding discrete Laplace noise is optimal under ϵ-DP; however, in this work we want to allow for more general definitions of DP. Hereinafter, for notational simplicity, we denote the privacy noise as $Y := \Delta_{\bar{\theta}} Z/\epsilon$ (for any appropriate choice of Z). Based on this, we have our first assumption for our approach proposed later in Section 1.4.1.

Assumption 1. *The privacy mechanism to ensure DP is an additive noise mechanism as in* (1.2).

In this work we will use notation that aligns with the definition of ϵ-DP. However, as long as the mechanism remains an additive one, any form of DP can be addressed by the proposed methodology described further on. This being said, with the addition of the privacy noise it is straightforward to see that $\hat{\pi} = \bar{\theta} + Y$ is an unbiased and consistent estimator for the true

parameter of interest θ_0. However, it may not be straightforward to derive or approximate the non-asymptotic sampling distribution of $\hat{\pi}$ given the convolution of two sources of randomness (sampling and privacy noise). If using discrete Laplace noise for the one-sample proportion setting, then Awan and Slavkovic [6] derived the "Tulap" distribution which allowed them to deliver the UMP test for binomial data (along with CIs and similar inferential tools). Under other privacy mechanisms, alternative solutions can be conceived, but they have different limitations, as highlighted earlier. For example, the parametric bootstrap [22] could generally be a potential solution but it would not work if $\hat{\pi} \notin [0, 1]$, which is indeed possible for different choices of Y, and it may not deliver the correct distribution since this depends on the point in which $\hat{\pi}$ lands on the interval $[0, 1]$.

To overcome the above issues, we consider an alternative simulation-based framework proposed in [55] which is inspired by indirect inference [32] and akin to the repro sampling technique in [75]. More specifically, we define a new DP estimator based on a matching criterion and use tools from the fiducial inference literature to deliver solutions for it [35]. Under certain regularity conditions, the bootstrapped solutions for this estimator are valid when employing the percentile approach.

1.4 Methodology

To describe the proposed approach taken from [55], we first abuse notation slightly by defining the categorical variable as $X := X(\theta_0)$ to underline its implicit dependence on the true parameter θ_0 through the distribution f_{θ_0}. This being said, our goal is to construct valid CIs for $\theta_0 \in [0, 1]$ knowing that the initial estimator $\hat{\pi}$, computed on $X(\theta_0)$, converges in probability to the non-stochastic limit θ_0. To build such intervals, let us express the data-generating mechanism as $X(\theta) = g(\theta, W)$, where $W \in \mathbb{R}^m$ is a random vector whose distribution does not depend on θ. For example, W can be seen as a vector of iid standard uniform random variables such as those used to simulate samples from parametric distributions through their inverse cumulative distribution functions. Following this expression, we can also express *simulated* data as $X^*(\theta) = g(\theta, W^*)$ where W^* represents an independent copy of W. As a consequence, $\hat{\pi}$ represents the DP estimator computed on the observed data $X(\theta_0, w)$, where w is fixed and

unknown for the observed data, while we let $\hat{\pi}^*(\theta)$ represent the DP estimator computed on the simulated data $X^*(\theta, w^*)$, where w^* is a sample of the random variable W^*.

Based on the setup of the above problem, a first possible framework that one may consider to find a distribution with respect to the parameter θ would be that of (generalized) Fiducial inference [35]. The latter aims to find a distribution by matching the observed data with simulated data or, under certain conditions, by conditioning on sufficient statistics for θ. However, ignoring ongoing conceptual debates regarding this framework, in our DP setting we do not have access to the data, and it is not clear if the DP estimator $\hat{\pi}$ qualifies as a sufficient statistic once privacy noise has been added (especially if $\hat{\pi} \notin (0, 1)$). We therefore consider an alternative framework which defines an estimator for the quantity of interest starting from auxiliary estimators (including biased and/or inconsistent estimators for θ_0) and employs tools from the fiducial literature to deliver solutions for this estimator and valid (frequentist) statistical coverage of θ_0. Indeed, this framework employs general statistics related to the parameter θ_0 and the proposed estimator for this parameter is given by the solution to the optimization problem:

$$\hat{\theta} \in \underset{\theta \in [0,1]}{\operatorname{argmin}} \|\hat{\pi} - \hat{\pi}^*(\theta)\|, \tag{1.3}$$

where $\|\cdot\|$ denotes the Euclidean norm. With this definition, recalling our previous notation, it is clear that if $w^* = w$, then θ_0 is guaranteed to be a solution of (1.3). In practice, the distribution of $\hat{\theta}$ is approximated by generating H independent realizations of the random variable W^*, with H generally being large, which simply amounts to simulating H datasets of size n from the distribution f_θ. For each of these, the DP estimator $\hat{\pi}^*(\theta)$ is computed by also generating new realizations of the privacy noise Y^* (whose parameters are known), while $\hat{\pi}$ is fixed from the observed data for all $h = 1, \ldots, H$. Based on each of these replicates, the matching problem in (1.3) is solved H times thereby generating H samples of the random variable $\hat{\theta}$, that is, $\{\hat{\theta}_h\}_{h=1,\ldots,H}$, on which the percentile method can be used to obtain one- or two-sided CIs [more details in 55].

However, two major issues remain in obtaining this distribution: (i) the matching problem in (1.3) may not be computationally efficient to solve and (ii) it is not clear how to sample

$\hat{\theta}$ given the discrete nature of the data $\mathcal{D}$. For this purpose, by reformulating the matching problem in (1.3), we present a solution which relies on tools from the fiducial framework.

1.4.1 FIMA: Fiducial Matching

Let us start by rewriting the matching problem in (1.3) as an "argzero" problem, i.e.

$$\hat{\theta} \in \underset{\theta \in [0,1]}{\text{argzero}} \, [\hat{\pi} - \hat{\pi}^*(\theta)].$$

Essentially, this implies that we want to find the value of θ such that

$$\hat{\pi} = \hat{\pi}^*(\theta). \tag{1.4}$$

We can expand the second term of this matching as follows:

$$\hat{\pi}^*(\theta) := \bar{\theta}^*(\theta) + Y^*,$$

where $\bar{\theta}^*(\theta)$ is the simulated sample proportion implied by θ and Y^* is an independent copy of the privacy noise Y (by definition independent of the distribution f_θ). This allows us to rewrite (1.4) as follows:

$$\hat{\pi} = \bar{\theta}^*(\theta) + Y^* \implies \bar{\theta}^*(\theta) = \hat{\pi} - Y^*. \tag{1.5}$$

Let us now focus on the second equation in (1.5) where it can be noted that the simulated sample proportion $\bar{\theta}^*(\theta)$ can also be expressed as

$$\bar{\theta}^*(\theta) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(U_i^* \leq \theta),$$

with $U_i^* \overset{iid}{\sim} U(0,1)$ being an iid standard uniform random variable, for $i = 1, \ldots, n$ [35]. Hence, $\bar{\theta}^*(\theta)$ corresponds to an empirical cumulative distribution function (CDF) based on the sample $U_1^*, \ldots, U_n^*$, denoted as $F_{U^*}(\theta)$. As a result, the second equality in (1.5) can be

re-expressed as

$$F_{U^*}(\theta) = \hat{\pi} - Y^*,$$

such that, defining $\tilde{\theta}^* := \hat{\pi} - Y^*$, we have that the solution to (1.4) is given by

$$\hat{\theta} = F_{U^*}^{-1}(\tilde{\theta}^*). \tag{1.6}$$

Therefore, to obtain this solution we need to inverse the empirical distribution function F_U^* which is discrete by nature, implying that we do not obtain point solutions but intervals. Indeed, the interval solutions are given by:

$$I(\hat{\pi}, Y^*, \{U_i^*\}) = \begin{cases} \delta & \text{if } \tilde{\theta}^* \leq 0 \\ [\delta, U_{[1]}^*) & \text{if } 0 \leq \tilde{\theta}^* < \frac{1}{n} \\ [U_{[\lfloor n\tilde{\theta}^* \rfloor]}^*, U_{[\lfloor n\tilde{\theta}^* + 1 \rfloor]}^*) & \text{if } \frac{1}{n} \leq \tilde{\theta}^* < 1 - \frac{1}{n} \\ [U_{[n]}^*, 1 - \delta) & \text{if } 1 - \frac{1}{n} \leq \tilde{\theta}^* < 1 \\ 1 - \delta & \text{if } \tilde{\theta}^* \geq 1 \end{cases},$$

where $U_{[i]}^*$ represents the i^{th} ordered value of the sequence of standard uniform variables $\{U_i^*\} \in [0,1]^n$, $\lfloor v \rfloor$ represents the largest integer smaller than or equal to v and $0 < \delta \ll 1$ is a small constant close to zero (for example, the machine epsilon). As a result, sampling a value from the interval solution $I(\hat{\pi}, Y^*, \{U_i^*\})$ consists in a solution to the matching problem in (1.3). Indeed, following Hannig [35], we know that a *generalized fiducial quantity* for this problem is given by:

$$\hat{\theta} = U_{[\lfloor n\tilde{\theta}^* \rfloor]}^* + D\left(U_{[\lfloor n\tilde{\theta}^* \rfloor]}^* - U_{[\lfloor n\tilde{\theta}^* + 1 \rfloor]}^*\right),$$

where D is a random variable on the support $[0,1]$. Based on Hannig [35], any choice of distribution for D (under the stated support constraint) provides a fiducial quantity consisting in a solution to (1.6). The final procedure, justified on the basis of the above steps, is presented in Algorithm 1 and we refer to it as Fiducial Matching (FIMA).

Algorithm 1 Fiducial Matching (FIMA)

Input: $\hat{\pi}$: DP proportion/probability; n: data size; ϵ: privacy budget; H: number of solutions; q: distribution for Y; g: distribution for D; $\delta > 0$: small quantity close to zero.
for $h = 1, \ldots, H$ **do**
    $\tilde{\theta}_h^* = \hat{\pi} - Y_h^*$, where $Y_h^* \sim q\left(\frac{1}{n\epsilon}\right)$.
    if $\tilde{\theta}_h^* > 1$ **then**
        Return $\hat{\theta}_h = 1 - \delta$
    else if $\tilde{\theta}_h^* < 0$ **then**
        Return $\hat{\theta}_h = \delta$
    else
        Generate $\{U_{h,i}^*\}$, for $i =, 1, \ldots, n$ and $U_{h,i}^* \overset{iid}{\sim} U(0,1)$
        Return $\hat{\theta}_h = U^*_{[\lfloor n\tilde{\theta}_h^* \rfloor]} + D_h\,(U^*_{[\lfloor n\tilde{\theta}_h^* \rfloor]} - U^*_{[\lfloor n\tilde{\theta}_h^* + 1 \rfloor]})$, where $D_h \sim g$
    end if
end for
Output: A sequence $\{\hat{\theta}_h\}$ for $h = 1, \ldots, H$

Remark 1. *An interesting question is whether the FIMA sequence $\{\hat{\theta}_h\}$ (and quantities derived from it) delivers additional privacy if one were to directly release the latter instead of the original DP proportion $\hat{\pi}$. Indeed, while the original level of privacy is guaranteed based on the post-processing properties of DP, for finite n and H a small level of additional privacy could be achieved. More specifically, on top of the original DP level for $\hat{\pi}$, by composition we have that the FIMA sequence is $(\epsilon\, H)$-DP which obviously guarantees almost no additional privacy since in general we want H to be large. However, even if $H \to \infty$ such that $1/H \sum_{h=1}^{H} \tilde{\theta}_h^* \to \mathbb{E}[\tilde{\theta}_h^*] = \hat{\pi}$, for finite n we know that there is not a one-to-one relationship between $\tilde{\theta}_h^*$ and $\hat{\theta}_h$ due to the discrete nature of the CDF F_{U^*}. This implies that $F_{U^*}(\hat{\theta}_h) \nrightarrow \tilde{\theta}_h^*$ since it only allows to identify an interval within which $\tilde{\theta}_h^*$ lies therefore ensuring some additional privacy on the original DP proportion $\hat{\pi}$. However, as $n \to \infty$ the binomial CDF becomes smoother allowing to narrow the relationship between $\hat{\theta}_h$ and $\tilde{\theta}_h^*$ and therefore guaranteeing basically no additional privacy.*

The FIMA described in Algorithm 1 is a logical representation of the procedure but can of course be made computationally more efficient by avoiding the loop and generating all random values at the first step (i.e. Y^*, $\{U_i^*\}$ and D for all n and H). Thereafter, all that is needed is to perform indexing on predefined objects since $\hat{\pi}$ is fixed, and consequently $\{\tilde{\theta}_h^*\}$ can be computed in one vectorized operation. Moreover, in this work we will make the choice of $D \sim \text{Beta}(1/2, 1/2)$ which directly delivers the FIMA solution:

$$\hat{\theta} \sim \text{Beta}(n\tilde{\theta}^* + 1/2, n - n\tilde{\theta}^* + 1/2),$$

consisting in a posterior distribution based on Jeffrey's prior [see again 35]. Hence, using this solution, there is no need to generate the realizations D or the sequences $\{U_h^*\}$, thus also avoiding the ordering operation on the latter.

Remark 2. *An alternative solution to sample a point estimate $\hat{\theta}$ could be to approximate the distribution of $\hat{\pi}^*(\theta) = \bar{\theta}^*(\theta) + Y$ (and hence also $\hat{\pi}$) using a central limit theorem for $\bar{\theta}^*(\theta)$ thereby obtaining*

$$\hat{\pi}^* \overset{d}{\approx} \theta + \sqrt{\frac{\theta(1-\theta)}{n}} V^* + Y^*,$$

where $V^* \sim \mathcal{N}(0, 1)$. This would be similar in spirit to the approximation of [69]. For this approximation, there would exist explicit solutions for $\hat{\theta}$ since one would need to find the roots to a quadratic equation. Once these solutions are found, then one would pick $\hat{\theta}$ such that it lies in $[0, 1]$ and minimizes the loss in (1.3)*. While this solution is comparable in terms of empirical level and coverage, it has shown less power compared to FIMA (the properties of this approximation are however left for future work).*

Let us now confirm the validity of the FIMA when generating inferential solutions for CIs (and hypothesis tests). To do so, we follow the results in [55] and make use of the tools to determine the properties of extremum estimators [52] such as the one represented in (1.3). Based on these properties, we can then determine the validity of the studentized bootstrap (based on the consistency of the estimator $\hat{\pi}$) and, provided that $\hat{\pi} - \pi(\theta)$ converges to a zero-mean symmetric distribution (which is indeed a zero-mean Gaussian in our case), consequently ensures the

validity of the percentile bootstrap conditional on $\hat{\pi}$ [see 67, 16]. In this sense, we establish a required assumption on the true parameter θ_0, on which we aim to perform statistical inference.

Assumption 2 (Interior Point)**.** *The true parameter θ_0 is in the interior of the parameter space* $[0, 1]$.

This is a common assumption for matching-based (or extremum) estimators to ensure, among others, differentiability and relative Taylor expansions around the true parameter, as well as the validity of bootstrap solutions. However, to state the validity of FIMA we also define the quantile of the estimator $\hat{\theta}$ as:

$$\hat{\theta}_\alpha := \inf \left\{ v \in \mathbb{R} \ : \ \mathbb{P}\left(\hat{\theta} \leq v \,\middle|\, \hat{\pi} \right) \geq \alpha \right\}. \tag{1.7}$$

Based on these premises, the following result confirms that the FIMA produces valid CIs.

Theorem 1.1. *Letting $\alpha \in (0, 1)$, under Assumptions 1-2 we have*

$$\mathbb{P}\left(\theta_0 \leq \hat{\theta}_\alpha \right) = \alpha + o_p(1).$$

Proof. We verify the conditions for the validity of the percentile bootstrap [67, 16] and therefore use the results in [52] for extremum estimators such as the estimator proposed in (1.3). The first condition is a common regularity condition and requires the parameter space to be convex and compact, in addition to the true parameter being an interior point of this space: indeed the parameter space $[0, 1]$ is obviously convex and compact and under Assumption 2 we trivially respect this condition. The second condition requires $\pi(\theta)$ to be continuously differentiable and injective (one-to-one) on $[0, 1]$: we have that $\pi(\theta) := \mathbb{E}[\hat{\pi}(\theta)] = \mathbb{E}[\bar{\theta} + Y] = \mathbb{E}[\bar{\theta}] + \mathbb{E}[Y] = \theta$, and hence $\pi(\theta)$ is injective and continuously differentiable, therefore trivially respecting this condition as well. Defining $\sigma^2(\theta)$ as the asymptotic variance of the DP estimator $\hat{\pi}$, the final condition requires the scaled and centered DP estimator $\sqrt{n}\, \sigma^{-2}(\theta)\, (\hat{\pi} - \pi(\theta))$ to converge, uniformly in $\theta \in [0, 1]$, to a zero-mean random variable with a continuous distribution that

does not depend on θ: since the DP noise is $O_p(1/n)$, we have that

$$\sqrt{n}\,[\theta_0(1-\theta_0)]^{-1/2}\,(\hat{\pi}-\theta_0) \xrightarrow{D} \mathcal{N}(0,1),$$

thereby respecting this condition and thus concluding the proof. □

Remark 3. *The FIMA can easily be adapted to the case where, instead of the sample proportion, the privatized quantity is the count itself where, multiplying by n on both sides, Equation* (1.1) *simply becomes:*

$$\bar{X}_k = \sum_{i=1}^{n} \mathbb{1}(X_i = k),$$

i.e. the sum of observations in class k. In this case, continuing with the binomial (two-class) example and defining $\hat{X} = \bar{X} + Y$ as the DP count for one class (where the sensitivity for the noise is now $\Delta_{\bar{X}} = 1$) it is straightforward to see that:

$$\check{\pi} = \frac{1}{n}\,\hat{X},$$

is also an unbiased and consistent estimator for θ_0 with asymptotically normal distribution thereby respecting the conditions for Theorem 1.1 to hold. Hence, we have that:

$$\underset{\theta\in[0,1]}{\textit{argzero}}\,[\hat{X} - \hat{X}^*(\theta)] = \underset{\theta\in[0,1]}{\textit{argzero}}\,[\check{\pi} - \check{\pi}^*(\theta)],$$

where $\hat{X}^*(\theta)$ and $\check{\pi}^*(\theta)$ are the simulated counterparts in our matching problem. Therefore we have that the solution to the latter is given by $\hat{\theta} = F^{-1}_{U^*}(\check{\theta}^*)$ where $\check{\theta}^*$ is obtained by developing the matching problem as follows:

$$\begin{aligned}
\underbrace{\frac{1}{n}\hat{X}}_{\check{\pi}} &= \underbrace{\frac{1}{n}\hat{X}^*(\theta)}_{\check{\pi}^*(\theta)} \\
&= \frac{1}{n}\bar{X}(\theta) + \frac{1}{n}Y^* = \bar{\theta}^*(\theta) + \frac{1}{n}Y^* \\
\implies \underbrace{\bar{\theta}^*(\theta)}_{F_{U^*}(\theta)} &= \underbrace{\frac{1}{n}(\hat{X} - Y^*)}_{\check{\theta}^*},
\end{aligned}$$

where $\bar{X}(\theta)$ is the simulated non-private count.

Let us now consider the more general multinomial case for $K > 2$, i.e. $X \sim f_\theta^K$ where $\theta \in \Delta^{K-1} \subset \mathbb{R}^K$. In this case, the non-private estimator of the latter is represented by $\bar{\theta} := [\bar{\theta}_1, \ldots, \bar{\theta}_K]$ where $\bar{\theta}_k$ is the sample proportion for the k^{th} class defined in (1.1). Although this estimator could be privatized in calibrated manners for the specific problem [see e.g. 41, 11, 70], these approaches would not allow to find an explicit solution for the FIMA. Hence, we preserve the component-wise additive mechanism in (1.2) which allows us to define the DP proportion for the k^{th} class as $\hat{\pi}_k$, with $\hat{\pi} := [\hat{\pi}_1, \ldots, \hat{\pi}_K]$. As a result, we can now apply Algorithm 1 to each component of the vector $\hat{\pi}$ independently to obtain the vector of FIMA solutions $\hat{\theta} := [\hat{\theta}_1, \ldots, \hat{\theta}_K]$. However, while for the binomial case the FIMA solution lies within the $(0, 1)$ interval, as mentioned earlier, this multinomial approach would not guarantee that the vector $\hat{\pi}$ lives within the required $(K-1)$-dimensional simplex. As shown further on, this does not necessarily represent a problem for various statistical inferential procedures for categorical data. Therefore let us first consider the setting where the FIMA solutions $\hat{\theta}$ are used to deliver a single test statistic of interest: this can be the χ^2 test statistic (where estimated proportions are used to compute expected counts) or simply selecting a single component of the vector $\hat{\theta}$ to perform inference. In this sense, let us consider a function $\varphi : \mathbb{R}^K \to \mathbb{R}$ that defines such a statistic and make the following assumption on it.

Assumption 3 (Differentiable Stat)**.** *The function φ is continuously differentiable with respect to θ.*

This assumption is needed since it allows us to use the continuous mapping theorem (and the Delta method) to ensure that $\varphi(\hat{\theta})$ converges to $\varphi(\theta)$, based on the uniform convergence of $\hat{\theta}$, as well as to its asymptotic Gaussian distribution. Based on this, we also define $\vartheta := \varphi(\theta)$ and, consequently, let $\hat{\vartheta}_\alpha$ represent the α-level quantile of the distribution of $\hat{\vartheta} := \varphi(\hat{\theta})$ defined similarly to that in (1.7). For completeness, we also adapt Assumption 2 to this multinomial setting.

Assumption 4 (Interior Point (Multinomial))**.** *The true parameter θ_0 is in the interior of the parameter space Θ.*

Having stated the above assumptions, we can now give the following result.

Proposition 1. *Letting $\alpha \in (0, 1)$, under Assumptions 1, 3 and 4 we have*

$$\mathbb{P}\left(\vartheta_0 \leq \hat{\vartheta}_\alpha\right) = \alpha + o_p(1).$$

Proof. We follow the same steps as the proof of Theorem 1.1 and verify the conditions by adapting some definitions. Specifically, since the parameter space Θ is convex and compact, under Assumption 4 we respect the first condition. Similarly to Theorem 1.1, we also see that $\pi(\theta)$ is injective (since $\pi(\theta) = \theta$) and continuously differentiable which, along with Assumption 3 on the differentiability of φ, allows us to verify the second condition. Also, since the DP noise is $O_p(1/n)$, we have that

$$\sqrt{n}\,\Sigma(\theta_0)^{-1/2}(\hat{\pi} - \theta_0) \xrightarrow{D} \mathcal{N}_K(0, I),$$

where $\Sigma(\theta) = \text{diag}(\theta) - \theta\theta^\top$, thereby respecting the third condition and concluding the proof. □

Notice that this result implies the validity of the percentile bootstrap for the quantity $\hat{\vartheta}$ which represents a continuously differentiable transform of $\hat{\theta}$. The majority of the standard inferential tools for categorical data respect this assumption such as, for example, the χ^2-statistic or logit transforms for the components of $\hat{\theta}$ (which allow to estimate the parameters of logistic models with categorical predictors). An exception may be represented by functions (operators) that project the FIMA solution $\hat{\theta}$ onto the $(K-1)$-dimensional simplex. More specifically, projections onto the simplex are often not differentiable everywhere and may not be asymptotically Gaussian when close to the boundaries of the simplex. Although this remains problematic in finite samples, we can, however, observe that based on Assumption 4 and on the consistency of $\hat{\theta}$, the result of Proposition 1 still holds for such projections as $n \to \infty$.

1.5 Experiments

We investigate the performance of FIMA in different categorical settings, starting from the basic binomial data with one-sample proportion and two-sample proportion tests, and going to the χ^2-test (logistic models are investigated in Appendix A). For all settings, we will use $Z \sim \text{Laplace}(0, 1)$, therefore ensuring pure ϵ-DP, and a privacy budget of $\epsilon = 1$. Moreover, for each experimental condition, we run each method $B = 10^4$ times to evaluate its statistical properties and, for the FIMA, we set $H = 10^3$ which is generally reasonable to have a good approximation for the bootstrap distribution [16]. It must be noted that there are few methods that deliver statistical inference guarantees in all these settings, hence we compare FIMA to different approaches across the different settings.

1.5.1 One-Sample Proportion

We first study the one-sample proportion setting for which an optimal DP inferential solution already exists and was put forward in Awan and Slavkovic [6]: we refer to this method as Tulap since it is based on the Truncated-Uniform-Laplace (Tulap) distribution derived in their work. As mentioned in the introduction, this approach delivers the UMP under DP constraints for binomial data [1]. Therefore, we compare the FIMA to this method and, as benchmarks, we compare it also to two non-DP options consisting of the standard z-test for proportions (NP) and the exact binomial test (Exact). We focus our attention on the performance of CIs derived from all these methods in small sample settings: we choose to study coverage on samples of size $n = 30$ and for a level $1 - \alpha = 0.95$. We produce B CIs for each method and for different values of θ_0, i.e. $\theta_0 \in \{0.1, 0.11, 0.12, \ldots, 0.985\}$, also to check coverage when moving close to the boundaries of the parameter space. The left plot of Fig. 1.1 shows the results of this simulation: the non-private benchmarks show the well-known behavior in terms of CI coverage where (i) the z-test suffers from the normal asymptotic approximation for a discrete-natured problem (especially in finite-samples like this simulation) and generally under-covers while (ii) the exact CIs guarantee at least the $1-\alpha$ coverage but also suffers from inconsistent jumps along

[1] Tulap code: `https://github.com/Zhanyu-Wang/Simulation-based_Finite-sample_Inference_for_Privatized_Data/blob/main/table1/Awan_binomialDP.R`

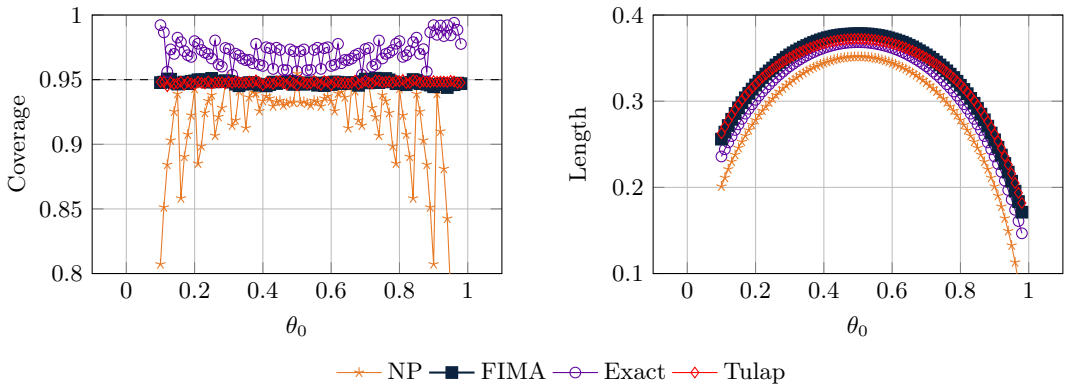


Figure 1.1: Comparison of the coverage (left) and length (right) of 95% CIs for each method and for different parameter values of θ_0.

the range of θ_0 values and tends to over-cover at the boundaries. On the other hand, the two DP CIs both appear to have an exact coverage across all values of θ_0 including the boundaries and are therefore counterintuitively preferable to the non-private benchmarks when it comes to CI coverage. If we focus on the length of the CIs the right plot in Fig. 1.1 shows that the non-private CIs obviously have shorter intervals across all the range of values θ_0 and therefore are more precise, although only the Exact CIs guarantee coverage. As expected, the two DP approaches have larger CI lengths, with Tulap generally having slightly shorter intervals across the different values of θ_0 compared to FIMA, except towards the boundaries of the parameter space, where the FIMA appears to be marginally more precise.

We also investigated the performance of hypothesis tests based on these approaches. For this purpose, we consider the set of hypotheses: $H_0 : \theta_0 \geq \gamma$ (null hypothesis) versus $H_A : \theta_0 < \gamma$ (alternative hypothesis), where $0 \leq \gamma \leq 1$ is a constant that we are interested in testing against evidence from the sample. For this experiment, we first study the level (type-I error) of the tests based on these different methods and, for this purpose, set $\gamma = \theta_0$ for different values of $\theta_0 = 0.1, 0.2, \ldots, 0.9$. Fixing the sample size to $n = 30$ to evaluate finite sample performance, in the left plot of Fig. 1.2 we can see that all methods are slightly liberal (i.e., reject H_0 more than needed) but are all within a reasonable range of the threshold $\alpha = 0.05$ considering the sample size (with Tulap showing the best overall performance). In the right graph, we can observe the power of these methods by fixing $\theta_0 = 0.2$ and varying the value of γ from 0.2 to

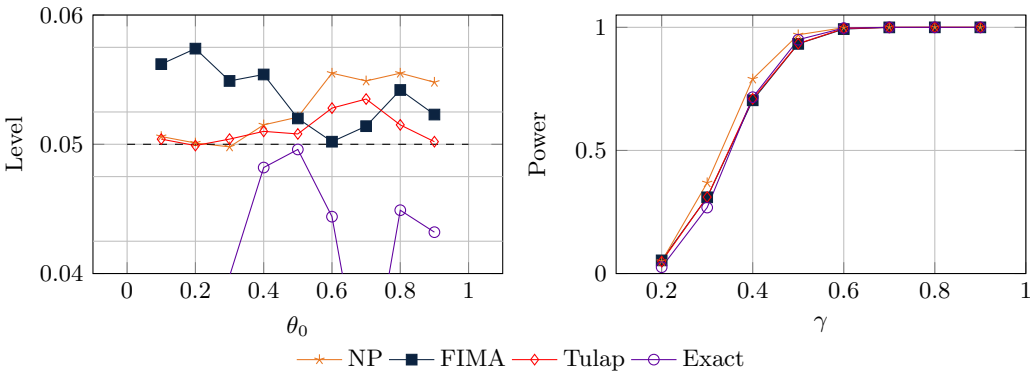


Figure 1.2: Comparison of the level (left) and power (right) of the different test procedures under the one-sided hypothesis test at $\alpha = 0.05$ for different parameter values of θ_0.

1. It can be observed that the non-private method is obviously the most powerful but is closely followed by the Tulap and FIMA approaches, which have comparable performance in terms of power. In the context of binomial data, our empirical results highlight how the proposed FIMA, while comparing favorably with respect to the standard non-DP approaches, also has a performance comparable to the optimal DP method based on the Tulap distribution of [6]. Additional results for this setting, with larger and varying sample sizes, can be found in Fig. A.1 of Appendix A where the FIMA shows good performance (often better) than its current alternatives.

1.5.2 Two-Sample Proportions

We now consider the two-sample (independent) proportions test where we are interested in testing $H_0 : \theta_1 - \theta_2 \geq 0$ versus $H_A : \theta_1 - \theta_2 < 0$, where θ_1 represents the proportion from the first population and θ_2 that from the second. In this setting, for the FIMA we consider the statistic $\lambda_h = \hat{\theta}_{1,h} - \hat{\theta}_{2,h}$ where $\hat{\theta}_{i,h}$ is the h^{th} solution of Algorithm 1 for the i^{th} sample. It can be noted that λ_h respects Assumption 3 and therefore fulfills the conditions of Proposition 1 (assuming Assumption 4 holds). As a DP alternative method, we make use of a recently proposed general hypothesis testing framework under DP put forward by Kazan et al. [43], called the Test of Tests (ToT), which compared favorably to other existing alternatives. The latter relies on the sample-and-aggregate approach discussed in [14] and we make use of the

optimization procedure proposed in [43] to find the best m (number of sub-samples) and α_0 (the sub-test significance threshold) [2]. Of course, we also use the non-private z-test (NP) as a general benchmark for the DP options considered here. Fig. 1.3 can be interpreted similarly to the plots in Fig. 1.2. Indeed, the plot on the left shows the level of the considered approaches when the sample size is $n = 30$ and where the x-axis represents different values of $\theta_0 = \theta_1 = \theta_2$ (under H_0), while the right plot shows the power of the methods when fixing $\theta_1 = 0.2$ and increasing θ_2 from 0.2 to 0.9. We can see that the ToT performs very well (indeed it is the best) in terms of type-I error while the NP and FIMA are reasonable when the two parameters are close to 0.5 but worsen (reject more than needed) when getting closer to the boundaries of the parameter space. The latter is generally expected in small samples but it can be seen that the FIMA is generally worse in these regions, although it still lies within an acceptable level considering the sample size. On the other hand, the ToT performs well across all parameter ranges. When looking at power we see a different picture: the NP test is obviously the most powerful option while the FIMA follows a similar progress below the NP power curve (as expected for a DP approach). For the ToT we consider two versions: one is the actual (empirical) ToT while the other is the theoretical power expected for the ToT [see 43]. We can observe that the theoretical power has a good performance and is better than the FIMA, while the empirical ToT appears to suffer the small sample setting and is not performing as it should do theoretically. Putting aside the small sample (which makes subsampling more complicated), it is probably the case that the default optimal parameter approach provided in [43] doesn't fit for these categorical settings; hence it may be necessary to also adapt the hyper-parameter tuning procedure for the ToT and therefore these results may not be considered as conclusive for this method for these experiments. However, given also the excellent performance in terms of level, we chose to keep the ToT as a potential alternative to consider in these experiments. Similarly to the one-sample proportion experiments, Fig. A.3 in Appendix A reports further results in this setting with larger sample sizes where the FIMA compares favorably to the considered alternatives and the power of ToT appear to have improved.

[2] Code can be found at `https://github.com/diff-priv-ht/test-of-tests`

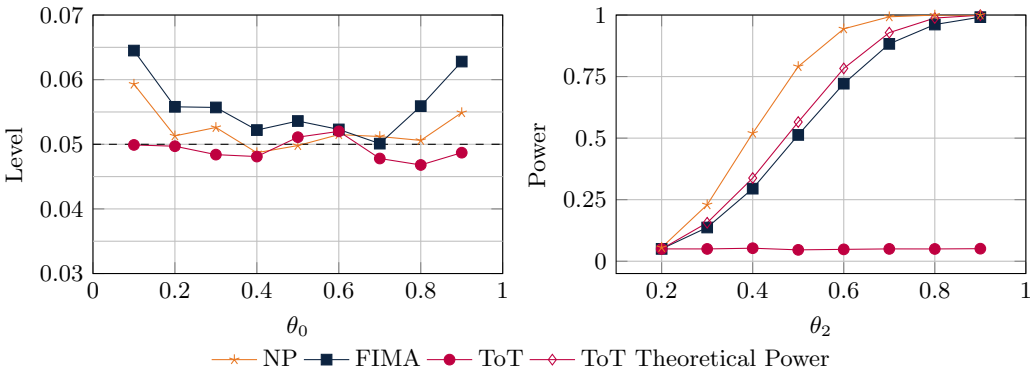


Figure 1.3: Comparison of the level (left) and power (right) of the different test procedures for two-sample hypothesis test at $\alpha = 0.05$ and for different parameter values of θ_0.

1.5.3 χ^2 Test

We also performed an experiment for the setting where we may be interested in performing a χ^2-test between two categorical variables $X \sim f_{\theta_1}^{K_1}$ and $Q \sim f_{\theta_2}^{K_2}$ with K_1 and K_2 classes respectively. In this case, we have to slightly adapt the FIMA to generate a bootstrap distribution of the χ^2 statistics under the null hypothesis of independence between the two categorical variables. Taking into account the distribution of the n total counts across different combinations of the two variables X and Q, these can be modeled by a multinomial distribution f_{θ_0} with $K = K_1 \cdot K_2$ categories in total and corresponding vector of probabilities $\theta_0 = \texttt{vectorize}(\theta_1 \theta_2^T)$ under independence, with $\texttt{vectorize}(A)$ representing the vectorization/flattening of the matrix A. With this framework in mind, we assume that we obtain privatized counts $\hat{G}_{ij}$ in the form of a contingency table with i and j representing the row and column respectively, with $i = 1, \ldots, K_1$ and $j = 1, \ldots, K_2$. In this case, to perform the χ^2-test we need to slightly adapt the reasoning in Remark 3. Indeed, we now are computing marginal counts for each variable X and Q which consist in sums of privatized counts $\hat{G}_{ij}$ across the rows and the columns. Hence, taking a marginal count across the rows as an example (i.e. targeting the marginal distribution $f_{\theta_2}^{K_2}$ of Q), we have that

$$\hat{Q}_j = \sum_{i=1}^{K_1} \hat{G}_{ij} = \sum_{i=1}^{K_1} \left(\bar{G}_{ij} + Y_{ij} \right),$$

where $\bar{G}_{ij}$ represents the non-privatized count of the contingency table ($\hat{X}_i$ therefore represents the sum across columns of privatized counts for the i^{row}). As a result, following the same reasoning as Remark 3, we have that the FIMA solution for θ_{2j} (i.e. the j^{th} element of θ_2) is given by $\hat{\theta}_{2j} = F^{-1}_{U^*}(\check{\theta}^*_{2j})$, with $\check{\theta}^*_{2j}$ now being defined as:

$$\check{\theta}^*_{2j} = \frac{1}{n}\left(\hat{Q}_j - \sum_{i=1}^{K_1} Y^*_{ij}\right),$$

thereby defining the vector of quantities $\check{\theta}_2 = [\check{\theta}_{21}, \ldots, \check{\theta}_{2K_2}]$ (with $\check{\theta}_1$ denoting the same vector of the other margin). Algorithm 2 provides the overview of this adaptation of the FIMA to this specific test. It must be noted that this is an adaptation specifically for this multinomial context (with sums of privacy noise terms) which would not be required if directly targeting the K-dimensional vector θ_0 characterizing the individuals counts in the table, in which case the strategy in Remark 3 can directly be applied component-wise. This is in fact the case for logistic regression with categorical predictors in Subsection 1.5.4.

Algorithm 2 FIMA for χ^2-test

1: **Input:** $\hat{G}$: table of DP counts; n: data size; ϵ: privacy budget; H: number of solutions; q: distribution for Y; g: distribution for D; $\delta > 0$: small quantity close to zero.
2: Compute marginal counts $\hat{\mathbf{X}} = \sum_{j=1}^{K_2} \hat{G}_{ij}$ and $\hat{\mathbf{Q}} = \sum_{i=1}^{K_1} \hat{G}_{ij}$
3: **for** $h = 1, \ldots, H$ **do**
4: Compute $\check{\theta}^*_{1h} = {}^1\!/_n(\hat{\mathbf{X}} - \mathbf{Y}^*_{1h})$ and $\check{\theta}^*_{2h} = {}^1\!/_n(\hat{\mathbf{Q}} - \mathbf{Y}^*_{2h})$, where $\mathbf{Y}^*_{1,h} := [\sum_{j=1}^{K_2} Y^*_{1,h,j}, \ldots, \sum_{j=1}^{K_2} Y^*_{K_1,h,j}]$ and $\mathbf{Y}^*_{2,h} := [\sum_{i=1}^{K_1} Y^*_{1,h,j}, \ldots, \sum_{i=1}^{K_1} Y^*_{K_2,h,j}]$, with $Y_{k,h,j} \sim q\left(\frac{2}{n\epsilon}\right)$.
5: Run lines 4-11 of Algorithm 1 on each element of $\check{\theta}^*_{1h}$ and $\check{\theta}^*_{2h}$ to obtain matrix $\hat{\theta}_h = \hat{\theta}^*_{1h}\hat{\theta}^{*T}_{2h}$
6: Generate contingency table $\bar{\mathbf{G}}^*_h$ from mulitnomial distribution with parameters n and `vectorize`($\hat{\theta}_h$)
7: Apply DP noise to $\bar{\mathbf{G}}^*_h$ using additive noise distribution q to obtain new contingency table of DP counts $\hat{\mathbf{G}}^*_h$
8: Compute χ^2_h (the χ^2-statistic) based on contingency table $\hat{\mathbf{G}}^*_h$
9: **end for**
10: **Output:** A sequence $\{\chi^2_h\}$ for $h = 1, \ldots, H$

Remark 4. *In the case of Algorithm 2, we must underline that the requirements of Proposition 1 only hold when the DP noise comes from a continuous distribution (and not from a discrete distribution). Indeed, in this case the χ^2-statistic is continuously differentiable with respect to*

θ_0 and thus respects Assumption 3. Underlining that the results presented in this section are similar also in the discrete case, the theory nevertheless does not currently support this setting.

We compare the FIMA in Algorithm 2 to the non-private χ^2-test (NP) as the standard benchmark, and also compare it to DP alternatives which include the ToT used in our previous experiment, as well as the specifically tailored DP χ^2-test put forward by Gaboardi et al. [27] (we refer to this as Gabo). We consider the same experimental setup as that of Gaboardi et al. [27] which consists of a simple 2×2 contingency table generated for different sample sizes ranging from $n = 10$ to $n = 20^4$ and with privacy budget $\epsilon = 0.1$ (in [27] they use $(\epsilon, \delta) - DP$ but we set $\delta = 0$). More specifically, to study the level of the methods, we specify $\theta_0 = [0.25, 0.25, 0.25, 0.25]$ under the null H_0 (independence) and, for the power under the alternative H_A, we set $\theta_0 = [0.25, 0.25, 0.25, 0.25] + 0.01 \cdot [1, -1, -1, 1]$. Also, for the FIMA, we use $H = 10^4$ given the higher number of parameters (counts) in this model and, hence, the need to better approximate the FIMA distribution.

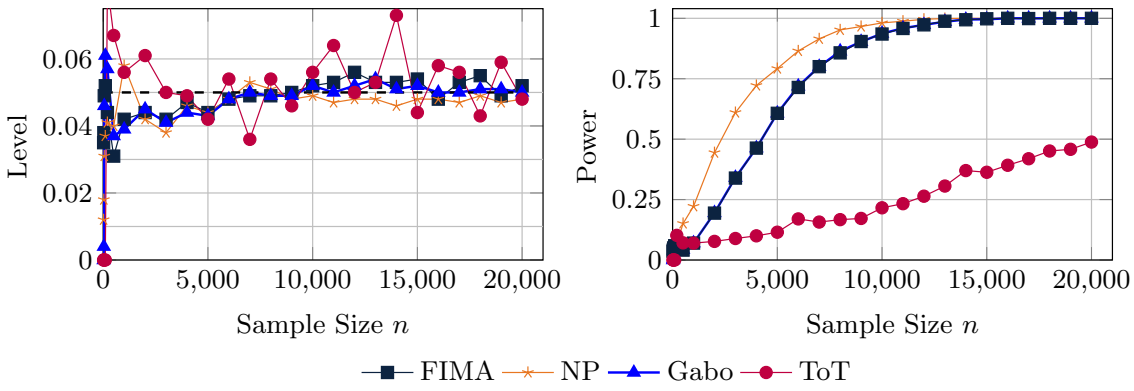


Figure 1.4: Comparison of the level (left) and power (right) of the different test procedures for χ^2-test at $\alpha = 0.05$ for different sample sizes n.

It can be seen in the left plot of Fig. 1.4 how the NP and Gabo fluctuate with more amplitude in smaller sample sizes and then start fluctuating more closely to the desired $\alpha = 0.05$ level as the sample size increases. However, the ToT and FIMA appear to have a lower fluctuation in small samples, while the FIMA appears to be more stable (from the conservative side) around the required level as the sample size increases, the ToT appears to fluctuate with more amplitude. With regard to power, putting aside the NP as the most powerful benchmark,

we can see that Gabo and FIMA have very similar performance in terms of power (with Gabo being slightly better overall), whereas the ToT also sees increasing power but less than its alternative, possibly being affected by similar issues as in Sec. 1.5.2 thereby probably requiring a better tuning of the hyperparameters m and α_0 (among others).

1.5.4 Logistic Models with Categorical Predictors

We run experiments using FIMA for inference on the parameters of logistic regression models with categorical predictors. In particular, following Agresti [1] we know that for a multinomial model with K classes and choosing the last as the reference class, then we have the following relationships:

$$\begin{aligned}
&\text{class } 1 && \text{logit}(\theta_1) = \beta_0 + \beta_1 \\
&\dots \quad \dots && \dots \\
&\text{class } K-1 && \text{logit}(\theta_{K-1}) = \beta_0 + \beta_{K-1} \\
&\text{class } K && \text{logit}(\theta_K) = \beta_0,
\end{aligned}$$

where β_k is the logistic model coefficient for the k^{th} class. As a consequence we have that:

$$\beta_0 = \text{logit}(\theta_K) \quad \text{and} \quad \beta_k = \text{logit}(\theta_k) - \text{logit}(\theta_K), \tag{1.8}$$

which allows us to make use of a plug-in approach; thus replacing the FIMA solution $\hat{\theta}_k$ in the above equations to obtain FIMA solutions $\hat{\beta}_k$. Indeed, the logit function is continuously differentiable therefore respecting Assumption 3. To the best of our knowledge, there are currently no existing DP methods that can directly be applied to this setting, and therefore we limit ourselves to showing the results for FIMA.

1.5.4.1 One Binary Predictor

Consider the simple logistic model regression model for binary response X and with one binary predictor T such that the model can be written as

$$\text{logit}(\theta_{k_T}) = \beta_0 + \beta_1 T,$$

where $\theta_{k_1} := \mathbb{P}(X = 1 \,|\, T = 1)$ and $\theta_{k_0} := \mathbb{P}(X = 1 \,|\, T = 0)$. Therefore, we apply Algorithm 1 to obtain FIMA solutions $\hat{\theta}_{k_1}$ and $\hat{\theta}_{k_0}$ respectively and plug these into the formula in (1.8). We maintain the same experimental parameters as in our previous analyses, fixing $\alpha = 0.05$, $\epsilon = 1$, $H = 10^4$, and $B = 10^4$. The predictor T is drawn from a Bernoulli distribution, $T \sim \text{Bern}(\theta_0)$ with $\theta_0 = 0.5$, while the true model parameters are set to $\beta_0 = 0.5$ and $\beta_1 = 2$. Table 1.1 presents the corresponding confidence interval coverage rates and lengths for both β_0 and β_1.

n	β_0		β_1	
	Coverage	Length	Coverage	Length
15	0.9518	5.6602	0.9761	9.4099
30	0.9493	3.9519	0.9724	6.9174
100	0.9503	1.1586	0.9693	3.8286
200	0.9495	0.8799	0.9464	2.0628
500	0.9499	0.5377	0.9483	1.1215
1000	0.9547	0.3696	0.9496	0.7670
2000	0.9491	0.2607	0.9453	0.5299

Table 1.1: CI coverage and lengths of the FIMA for logistic regression with one binary predictor for different sample sizes.

We can see that for all sample sizes the coverage is close to the nominal level of 0.95, or at least approaches it from the conservative side in the case of β_1.

1.5.4.2 Two Binary Predictors

We apply the same FIMA mechanism for the case of two binary predictors, T_1 and T_2 such that the model is defined as:

$$\text{logit}(\theta_{k_{\mathbf{T}}}) = \beta_0 + \beta_1 T_1 + \beta_2 T_2,$$

with $\mathbf{T} := [T_1\, T_2]$. We simulated the binary predictors as $T_1 \sim \text{Bern}(\theta_1)$ and $T_2 \sim \text{Bern}(\theta_2)$, with equal success probabilities $\theta_1 = \theta_2 = 0.5$. The true regression coefficients were specified as $\beta_0 = 0.5$ for the intercept, $\beta_1 = 2$ for the first predictor, and $\beta_2 = -2$ for the second predictor. The empirical results under this configuration are summarized in Table 1.2.

n	β_0		β_1		β_2	
	Coverage	Length	Coverage	Length	Coverage	Length
15	0.9867	9.52	0.9858	13.94	0.9885	13.36
30	0.9504	4.68	0.9654	10.05	0.9766	9.94
100	0.9491	1.51	0.9747	4.93	0.9629	4.50
200	0.9495	1.24	0.9592	3.42	0.9481	2.22
500	0.9525	0.77	0.9472	1.70	0.9540	1.31
1000	0.9513	0.53	0.9452	1.19	0.9601	0.86
2000	0.9512	0.37	0.9413	0.77	0.9583	0.62

Table 1.2: CI coverage and lengths of the FIMA for logistic regression with two binary predictors for different sample sizes.

In addition, in this case, the coverage and lengths for the β_i parameters appear to be good with the CI lengths decreasing with sample size.

In Appendix A.0.4, we examine the sensitivity of confidence interval properties to variations in the privacy budget ϵ. Consistent with theoretical expectations, our analysis demonstrates that increasing the privacy budget yields progressively tighter confidence intervals while maintaining nominal coverage rates.

1.5.5 Running Times

We complete our empirical experiments by highlighting the mean running times of the different approaches with varying sample sizes. Let us therefore consider the one- and two-sample proportion tests (sample sizes $n = 16, \ldots, 500$) as well as the χ^2-test (sample sizes $n = 20, \ldots, 20,000$). Fig. 1.5 shows the log mean running times (in milliseconds) in the different settings for increasing sample sizes. We generally observe that the DP approaches have longer running times and that, as for most of the alternatives in the different settings, the FIMA running times do not appear to increase with sample size across settings, indicating the scalability of the proposed approach. This is mainly due to the choice of $D \sim \text{Beta}(1/2, 1/2)$ to sample from the intervals, which allows us to directly sample FIMA solutions from the Beta distribution defined in Sec. 1.4.1.

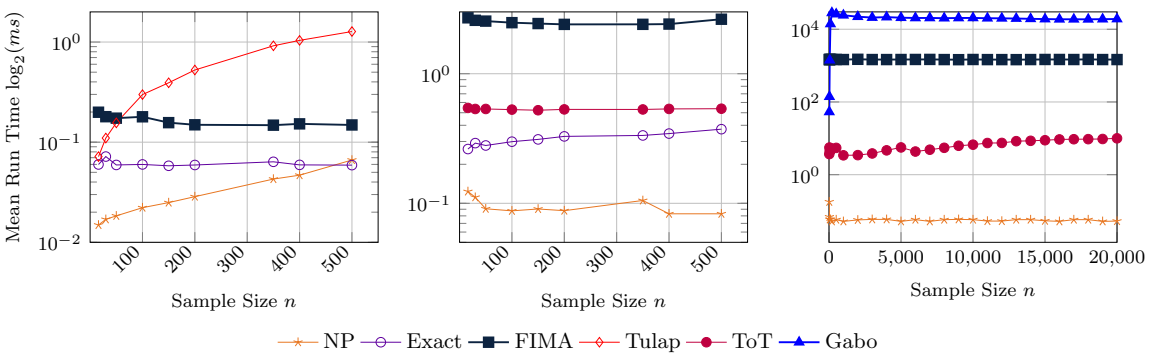


Figure 1.5: Mean log running times (in milliseconds) for all considered methods on different samples sizes for one-proportion hypothesis test (left plot), two-sample hypothesis test (center plot) and χ^2-test (right plot).

1.5.6 Applications

We apply the FIMA to real-world datasets, validating its practical utility. First, we examine the HIV infection data from the United States Centers for Disease Control and Prevention (CDC) for 2022[3]. Publicly reporting these data is crucial for guiding interventions, yet individuals represented in summary statistics may be re-identified, particularly in small or rural populations. We apply the two-sample test of proportions with FIMA to compare HIV infection rates

[3] `https://www.cdc.gov/hiv/data-research/facts-stats/index.html`

between sub-populations in this dataset, while ensuring individual privacy. In each application, we compare the results of FIMA with those obtained from a standard non-private (public) approach.

1.5.6.1 HIV Diagnoses in the US

In 2022, the CDC reported 37,981 new HIV diagnoses among individuals aged 13 and above in the U.S., with 70% (26,749) among gay, bisexual, and other men reporting male-to-male sexual contact. Among these, 9,374 are Hispanic/Latino and 8,831 are Black/African American. For targeted interventions, it is essential to assess which of these groups is more affected by HIV. We conduct a private test to determine if the proportion of HIV-positive Hispanic/Latino individuals reporting male-to-male sexual contact (θ_1) significantly exceeds that of their Black/African American counterparts (θ_2), leading to the set of hypotheses $H_0 : \theta_1 \leq \theta_2$ and $H_A : \theta_1 > \theta_2$.

We applied the Laplace mechanism to privatize observed proportions with a budget allocation of $\epsilon/2$. These privatized proportions, the sample size ($n = 37,981$), ϵ, and H were inputs for our FIMA two-sample test. With $\epsilon = 0.1$ and $H = 10^3$, we obtained p-value $= 0$. Results from the non-private asymptotic test (p-value $= 9.5 \times 10^{-7}$) and Fisher's exact binomial test (p-value $= 9.9 \times 10^{-7}$) aligned with our findings. To evaluate consistency, we repeated the private test 10^4 times with different random seeds, observing a minimum p-value of 0 and a maximum of 0.0062. We further analyzed the effect of ϵ on test outcomes. Table 1.3 shows the p-values for different privacy budgets, ϵ, and $H = 10^3$, demonstrating that conclusions align with non-private tests even for small ϵ values, down to $\epsilon = 0.01$.

ϵ	0.001	0.01	0.1	0.5	1	3	5	10
p-value	0.7634	0.0006	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 1.3: Values of ϵ and corresponding FIMA p-values of two-sample proportion tests on HIV infections data for Hispanic/Latino origin and Black/African American individuals who reported male-to-male sexual contact.

1.5.6.2 HIV Diagnoses

The CDC aims to reduce new HIV infections to 9,300 by 2025 and 3,000 by 2030. Effective interventions are required at multiple levels, especially for highly affected groups, such as those reporting male-to-male sexual contact. In Alabama, the Office of HIV Prevention and Care reported 374 new HIV cases in the third quarter of 2024, with 232 involving male-to-male sexual contact[4]. To evaluate progress toward the CDC's goals, we tested if the HIV infection rate in this subpopulation in Alabama is significantly lower than the 2022 national rate (0.7).

Using the FIMA one-sample test of proportions we tested $H_0 : \theta_0 \geq 0.7$ versus $H_1 : \theta_0 < 0.7$ with a privacy budget of $\epsilon = 1$ and $H = 10^4$. This yielded a p-value of 0.0002, compared to 0.0004 from the non-private test. A 95% differentially private confidence interval for the HIV infection rate among those reporting male-to-male sexual contact in Alabama is (0.5652, 0.6637), closely aligning with the non-private interval of (0.5711, 0.6695). Our results therefore indicate substantial progress toward the CDC's target for reducing HIV diagnoses, with the private and non-private tests producing consistent conclusions.

1.5.6.3 Poverty Status and Race in the US

We use 2022 American Community Survey (ACS) data[5] to test for dependence between race and poverty status in the US. Our FIMA χ^2-test of independence (see Sec. 1.5.3) ensures that this analysis can be conducted without compromising individual privacy. Testing this hypothesis on national data with a privacy budget of $\epsilon = 0.00001$ and $H = 10^3$ yielded a p-value of 0, consistent with the non-private test. Table 1.4 displays p-values for the FIMA χ^2-test across different privacy budgets ϵ. With $\epsilon = 0.00001$, we repeated the private test 10^4 times, obtaining a minimum p-value of 0 and a maximum of 0.004.

ϵ	0.000001	0.00001	0.0001	0.001	0.01	0.1	0.5	1	3	5	10
p-value	0.532	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 1.4: χ^2-test p-value for different privacy budget values (ϵ).

[4] `https://www.alabamapublichealth.gov/hiv/statistics.html`
[5] See URL in Appendix.

We extended this analysis to state and zipcode levels, using data from Alabama and the zipcode 36830. For Alabama, the non-private test returned a p-value of 0; however, the zipcode-level data had cells with zero counts, preventing the non-private test from running. With $\epsilon = 0.001$ and $H = 1000$, our private test yielded p-values of 0 for both state and zipcode levels. Thus, the FIMA χ^2-test aligns with the non-private test where applicable and remains robust even when the non-private test is infeasible.

1.6 Conclusion

The FIMA provides an efficient solution for statistical inference under DP constraints for categorical data characterized by class probabilities θ. While providing validity guarantees in terms of its bootstrap distributions, overall it can be observed that it provides a flexible alternative to existing DP approaches in these settings (without the need to specify additional hyper-parameters). Indeed, we cannot claim that the FIMA is the best overall approach based on our experiments, but it generally provides a computationally efficient alternative with stable performance in terms of type-I error and power across a variety of inferential tasks in categorical data settings. In particular, it can be observed that empirically it has a performance comparable to the UMP test of [6] while in the other settings it provides a generally safe choice (based solely on our experiments) since it produces good type-I error and is in line with (or better than) the power of some existing DP alternatives. Moreover, based on our theoretical results, different inferential tasks can be performed using FIMA whenever their statistics are a continuously differentiable function of the underlying parameter θ, which allows the method to be easily adapted to other categorical data settings. Subsequent studies could aim to broaden the applicability of the FIMA method by adapting it to non-categorical data types, such as those arising from normal and Poisson distributions. The complete implementation code for all experiments in this chapter is available in our public Github repository. A preprint of the associated manuscript is accessible on ArXiv. Furthermore, the `fima` package, which implements the FIMA method, can be installed directly from its dedicated Github repository.

Chapter 2

Differentially Private Conformal Prediction Via Quantile Binary Search

2.1 Introduction

In high-stakes domains such as health care, criminal justice, and financial risk assessment, machine learning models are increasingly being used to guide decision-making processes. However, traditional approaches often produce single-point predictions without quantifying uncertainty, which could lead to overconfident and unreliable outcomes. Many uncertainty quantification techniques have been proposed to address this need, including (among others) Bayesian neural networks, ensemble methods, and dropout-based approximations [see e.g. 12, 47, 56, 28, 44]. While these methods can offer meaningful uncertainty estimates, they often rely on strong modeling assumptions, are computationally intensive, or lack rigorous coverage guarantees (especially under model misspecification or when applied out-of-distribution). Conformal Prediction (CP) addresses various of these limitations by generating *prediction sets*, that is, subsets of the response space $\mathcal{Y}$ that, under the assumption of exchangeability, are guaranteed to contain the true response with a user-specified probability $1-\alpha$, where $\alpha \in (0, 1)$. These sets provide a reliable and model-agnostic framework for uncertainty quantification, making them increasingly attractive in safety-critical applications [see e.g. 65, 5, 64, 54]. While conformal methods require i.i.d. data for exact guarantees, recent extensions have begun to explore how to adapt them to distribution shift settings with approximate validity [see e.g. 63, 30, 39]. However, while CP provides statistical reliability, its reliance on calibration data introduces privacy risks, as any publicly released statistic based on these data could be exploited by an adversary and used for malicious purposes. Moreover, when training models on data that collect

individual-level information, these models tend to preserve individual-specific information, and some of them, such as deep neural networks, are known to *memorize* training examples [26, 15], enabling adversaries to extract sensitive information via model inversion [26] or membership inference attacks [62]. This poses a significant challenge also in regulated domains (e.g., under GDPR, HIPAA, or CCPA), where models must not leak information about individuals in the training or calibration sets.

To address these issues, Differential Privacy (DP) [20] has emerged as one of the main frameworks for privacy-preserving machine- and statistical-learning, ensuring that model outputs do not reveal whether any individual's data was included in the training set. However, existing DP frameworks focus primarily on protecting the *training* phase, leaving the calibration phase (and thus the CP step) vulnerable to privacy breaches. For example, an attacker could exploit the calibration scores to infer whether a specific individual was included in the calibration data [4]. Given this problem, there are nevertheless very few existing approaches that ensure that CP procedures can provide DP guarantees while maintaining reasonable performance in terms of uncertainty quantification metrics and computational efficiency. Hence, this work aims to further contribute to the goal of uncertainty quantification under privacy constraints by adapting an existing DP quantile selection approach based on binary search [37] and using it as a subroutine to deliver CP sets that guarantee privacy. We confirm the theoretical properties of the resulting CP procedure and highlight its favorable experimental performance compared to the few existing alternatives, both in terms of uncertainty quantification and computational efficiency.

2.2 Related Work

The idea behind CP was first described in [45] followed by different papers that developed its theory and applications (see [68], [6], [53], [43], and the references therein). Based on these, considerable attention has been paid in recent years to adapt CP to different settings and deliver additional applied and theoretical results (see [3] for a recent detailed review). Despite these different research outputs, little attention has been devoted to guaranteeing privacy for the calibration data through protection of the derived *non-conformity scores* used to deliver the

prediction sets for uncertainty quantification. The only existing work related to the central DP approach we present in this research is that of [4] where they privatize the CP procedure by discretizing the non-conformity scores into bins and applying the *exponential mechanism* [50] to select a DP quantile which is then used to form the prediction sets. More specifically, to maintain coverage guarantees under ϵ-DP, the target quantile is adjusted (inflated) to account for added noise. This ensures both privacy (via DP) and valid coverage, albeit with potentially larger prediction sets due to the privacy-utility trade-off. A possible practical limitation of this proposed approach lies in the choice of the optimal number of bins, as well as a hyperparameter that needs to be tuned to minimize the inflated quantile to achieve the smallest possible prediction set. Although procedures have been determined to compute these in an optimal manner, they can potentially increase the time complexity of this method. In recent years there have been few other CP approaches proposed in the setting of federated learning (see [38]) and *local DP* [59], however, these do not fit in the central DP framework of this paper.

2.3 Preliminaries

2.3.1 Conformal Prediction

To define the framework of the proposed method, let us consider n exchangeable data points $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ with $X_i \in \mathcal{X}^d$ being a d-dimensional input vector and $Y_i \in \mathcal{Y}$ being the target variable. Using the basic *split* CP [57, 27], we can partition these points into a training set with n_{train} observations (to fit a model of choice f) and a calibration set with $n_{\text{cal}} = n - n_{\text{train}}$ observations to quantify prediction uncertainty (assuming test data is taken from another sample). If we let $\mathcal{I} = \{1, \ldots, n\}$ represent the index set of the observations, then we can define $\mathcal{I}_{\text{cal}} \subset \mathcal{I}$ as the index set of observations in the calibration set. Assuming that we have trained a model $\hat{f}$ on the training set, the goal of CP is to compute *non-conformity scores* $s_i = s(X_i, Y_i, \hat{f})$ for each calibration point (X_i, Y_i), $\forall i \in \mathcal{I}_{\text{cal}}$. These scores measure the discrepancy between the model's prediction and the true label such that lower scores suggest better agreement, that is, better predictions from the trained model $\hat{f}$. Therefore we now have a set of non-conformity scores $\mathcal{S} = \{s_1, \ldots, s_{\text{cal}}\}$ on which we can obtain α-quantiles with

$\alpha \in (0, 1)$ being the significance level which represents the amount of errors (in percentage) that the model is allowed to make. We denote this quantile as q and let its dependence on the significance level α be implicit. Once the scores are computed, we can therefore sort them in increasing order (that is, $\{s_{[1]} \leq \ldots \leq s_{[n_{\text{cal}}]}\}$) and estimate this quantile by selecting the score in the r^{th} position where we define $r = \lceil (1-\alpha)(n_{\text{cal}}+1) \rceil$, with $\lceil v \rceil$ representing the smallest integer greater than x, leading to $\hat{q} = s_{[r]}$. Suppose that we now have a new test data point X_{test}, the resulting prediction set for the target Y_{test} is defined as:

$$\mathcal{C}(X_{\text{test}}) = \left\{ Y \in \mathcal{Y} : s(X_{\text{test}}, Y, \hat{f}) \leq \hat{q} \right\}.$$

Under the assumption of exchangeability of the data points, this set guarantees the coverage probability $\mathbb{P}(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}})) \geq 1-\alpha$, where the probability is taken over the $n+1$ data points $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{\text{test}}, Y_{\text{test}})$. For further details and extensions, we refer the reader to [3].

2.3.2 Differential Privacy

Differential Privacy (DP) currently represents one of the main frameworks for privacy protection with mathematically provable guarantees. More in detail, a randomized mechanism $\mathcal{M}$ is said to be (ϵ, δ)-DP if, for all neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ (that is, datasets with Hamming distance $d_{ham}(\mathcal{D}, \mathcal{D}') = 1$) and for all measurable subsets $\mathcal{Q}$ of the output space, we have that

$$\mathbb{P}[\mathcal{M}(\mathcal{D}) \in \mathcal{Q}] \leq e^{\epsilon} \mathbb{P}[\mathcal{M}(\mathcal{D}') \in \mathcal{Q}] + \delta.$$

There are many variations of this framework, including Gaussian-DP which benefits from a series of nice additional properties compared to the above (ϵ, δ)-DP [20, 18]. For the purpose of this work, we rely on a specific version of DP defined below.

Definition 1 (Zero-Concentrated Differential Privacy, [13])**.** *For $\rho > 0$, a randomized mechanism $\mathcal{M}$ satisfies $(\rho, 0)$-Zero-Concentrated Differential Privacy (zCDP) if for all neighboring*

datasets $\mathcal{D}$ and $\mathcal{D}'$ and for all $\alpha > 1$, the Rényi divergence D of order α satisfies:

$$D_\alpha(\mathcal{M}(\mathcal{D})\|\mathcal{M}(\mathcal{D}')) \leq \rho\alpha.$$

Unlike (ϵ, δ)-DP, ρ-zCDP is *tight* under sequential composition without the need for the *advanced composition theorem*, which may consume extra privacy budget. Nevertheless, there is a direct relationship between the (ϵ, δ)-DP framework and the zCDP framework: indeed, if $\mathcal{M}$ satisfies ρ-zCDP, then for $\delta > 0$, $\mathcal{M}$ satisfies (ϵ, δ)-DP for $\epsilon = \rho + 2\sqrt{\rho \log(1/\delta)}$ [see 51]. Hence, the latter implies the former. In particular, a randomized mechanism $\mathcal{M}$ satisfies $(\epsilon, 0)$-DP if and only if it satisfies ϵ-zCDP [13].

2.4 P-COQS: Private Conformity via Quantile Search

As highlighted in [4], the release of the non-conformity score quantile or of the prediction sets based on this quantile can compromise the privacy of individuals in the calibration set. Therefore, a solution to this problem is to directly release a DP quantile or to release the corresponding prediction sets, which would also preserve DP based on the privatized quantile. While the existing alternative in [4] employs the exponential mechanism to sample a DP quantile, this work relies on a direct adaptation of the DP binary search algorithm presented in [37]. More specifically, the latter algorithm releases a DP quantile by employing a noisy count function (denoted as `NoisyRC`) within a binary search procedure run on a sequence of ordered integers. In particular, the function $\texttt{NoisyRC}([a, b], \mathcal{D})$ returns a noisy count of $\mathcal{D} \cap [a, b]$, where $\mathcal{D}$ is a set containing ordered integers that are considered sensitive and hence must be accessed privately [see 37]. However, since in this work our sequences are not integers (given that the non-conformity scores are continuous), we adapt this DP quantile selection approach in a straightforward manner and, for completeness, we reproduce their adapted approach in Algorithm 3 with some modification to suit our application.

Broadly speaking, Algorithm 3 performs a binary search over the interval $[a, b]$ (often corresponding to $[0, 1]$ in the case of classification tasks) by iteratively narrowing the search space to localize the quantile of interest while preserving privacy through a randomized counting

Algorithm 3 Adapted `PrivQuant` Algorithm of [37]

Require: Non-conformity scores $\mathcal{S} = \{s_1, \ldots, s_{n_{\text{cal}}}\}$, significance level $\alpha \in (0, 1)$, lower and upper bounds for non-conformity scores $[a, b]$, $\delta > 0$ (small positive value, default $\delta = 10^{-10}$), privacy parameter ρ

Ensure: DP quantile q^{DP}

Fix $r = \lceil (1 - \alpha)(n_{\text{cal}} + 1) \rceil$ and $N = \lceil \log_2 ({}^{b-a}/_{\delta}) \rceil$
Initialize: left $\leftarrow a$, right $\leftarrow b$, $i \leftarrow 0$
while $i \leq N$ **do**
    mid $\leftarrow \frac{\text{left+right}}{2}$
    $c \leftarrow \texttt{NoisyRC}([a, \text{mid}], \mathcal{S})$
    if $c < r$ **then**
        left $\leftarrow$ mid $+ \delta$
    else
        right $\leftarrow$ mid
    end if
    $i \leftarrow i + 1$
end while
return $q^{\text{DP}} = \frac{\text{left+right}}{2}$

mechanism. More specifically, at each iteration, the algorithm evaluates a midpoint and calls the function `NoisyRC`, which returns a DP count of the number of non-conformity scores (in $\mathcal{S}$) less than or equal to the midpoint. This count guides the search by indicating whether the true quantile lies to the left or right of the midpoint. The process continues until it reaches iteration N which is the maximum number of iterations allowed to ensure ρ-zCDP. Indeed, as stated further on, the DP noise used for the function `NoisyRC` needs to be scaled to the number of iterations in Algorithm 3. Since we would ideally want the algorithm to terminate when the interval length is smaller than a small value δ (i.e. right $-$ left $\leq \delta$), the fixed number of iterations N needed to reach this point is such that ${}^{b-a}/_{2^N} \leq \delta$ (since the interval is divided by two after each iteration) which, solving for N, gives us

$$N = \lceil \log_2 ({}^{b-a}/_{\delta}) \rceil.$$

With this representing the exact number of iterations to guarantee the required condition (i.e. right $-$ left $\leq \delta$), Algorithm 3 returns the midpoint as the private quantile estimate q^{DP}. Once this quantile is obtained, following the definition in Section 2.3.1, the DP conformal prediction

set is simply defined as:

$$\mathcal{C}^{DP}(X_{\text{test}}) = \left\{ Y \in \mathcal{Y} : s(X_{\text{test}}, Y, \hat{f}) \leq q^{DP} \right\}. \tag{2.1}$$

We refer to the above procedure as Private Conformity via Quantile Search (P-COQS) which therefore is built on the adaptation described in Algorithm 3 and the consequent use of the resulting DP quantile q^{DP} in Equation (2.1). The entire P-COQS procedure is summarized in Algorithm 4.

Algorithm 4 P-COQS

Require: Training data $\mathcal{D}_{\text{train}} = \{(X_1, Y_1), \ldots, (X_{n_{\text{train}}}, Y_{n_{\text{train}}})\}$, calibration data $\mathcal{D}_{\text{cal}} = \{(X_1, Y_1), \ldots, (X_{n_{\text{cal}}}, Y_{n_{\text{cal}}})\}$, test data point X_{test}, significance level $\alpha \in (0, 1)$, lower and upper bounds for non-comformity scores $[a, b]$, $\delta > 0$ (small positive value, default $\delta = 10^{-10}$), privacy parameter ρ

Ensure: DP prediciton set $\mathcal{C}^{\text{DP}}$

1: Train model on training data $\mathcal{D}_{\text{train}}$ to obtain $\hat{f}$
2: Obtain non-conformity scores $\mathcal{S} = \{s_1, \ldots, s_{n_{\text{cal}}}\}$ from calibration data $\mathcal{D}_{\text{cal}}$ using $\hat{f}$
3: Run Algorithm 3 to obtain q^{DP}
4: Compute $\mathcal{C}^{DP}$ based on Equation (2.1) using q^{DP}
5: **return** $\mathcal{C}^{DP}$

Let us now focus on the properties of P-COQS and, for completeness, we start by studying the error bound and DP properties of Algorithm 3 which directly follow from the properties of the `PrivQuant` algorithm in [37]. Indeed, while the latter algorithm is tailored to search over positive integers, as mentioned earlier, Algorithm 3 adapts the latter to any ordered data type and any set of bounds $[a, b]$ on the real line. Let us therefore discuss (confirm) the DP properties of this adaptation and, to do so, let us first define the noisy count function as

$$\texttt{NoisyRC}([a, \text{mid}], \mathcal{S}) = \texttt{card}\big(\mathcal{S} \cap [a, \text{mid}]\big) + \mathcal{N}\left(0, \frac{\lceil \log_2(b-a/\delta) \rceil}{2\rho}\right), \tag{2.2}$$

where $\rho > 0$ is the privacy parameter under zCDP, with the sensitivity scaled to the maximum number of iterations N for Algorithm 3 (considering that count query l_1-sensitivity is 1), and `card` denotes the cardinality of set. Let us now denote $u = {}^{b-a}/_{\delta}$ and let Φ represent the standard normal CDF. Then, adapting directly from [37], the following proposition holds.

Proposition 2. *Algorithm 3 using the noisy count function in* (2.2) *is* ρ*-zCDP and returns a quantile with rank error*

$$\tau = \sqrt{\frac{\lceil \log_2(u) \rceil}{\rho} \log\left(2\frac{\lceil \log_2(u) \rceil}{\beta}\right)}$$

with probability at least $1 - \beta$*.*

Proof. The proof is straightforward. Indeed, since the sensitivity of the range query $\texttt{card}\big(\mathcal{S} \cap [a, \text{mid}]\big)$ is 1, then adding noise sampled from $\mathcal{N}\big(0, {}^{\lceil \log_2(u) \rceil}/_{2\rho}\big)$ will guarantee ${}^{\rho}/_{\lceil \log_2(u) \rceil}$-zCDP for each call of $\texttt{NoisyRC}([a, \text{mid}], \mathcal{S})$. Since there are $N = \lceil \log_2(u) \rceil$ calls before the algorithm terminates, then by the composition property of zCDP [13], Algorithm 3 is ρ-zCDP. Now let $\sigma^2 = {}^{N}/_{2\rho}$ denote the variance of the Gaussian noise added to guarantee ρ-zCDP in (2.2). Hence, the deviation between the true count $\texttt{card}(\mathcal{S} \cap [a, \text{mid}])$ and the noisy count $\texttt{NoisyRC}([a, \text{mid}], \mathcal{S})$, which we respresent as the random variable W, follows a Gaussian distribution $\mathcal{N}(0, \sigma^2)$. We want to bound the error for the N calls with high probability $1 - \beta$, for $\beta \in (0, 1)$. For each call of the algorithm, we have

$$\mathbb{P}(\big|\text{NoisyRC}([L, \text{mid}], D) - \texttt{card}\big(D \cap [L, \text{mid}]\big)\big| > \tau) = \mathbb{P}(|W| > \tau).$$

Since there are exactly N calls before the algorithm terminates, then by union bound and Gaussian tail probability we have that:

$$\mathbb{P}\Big(\bigcup_{i=1}^{N} \{|W_i| > \tau\}\Big) \leq N\mathbb{P}\big(|W| > \tau\big) \leq 2Ne^{-\frac{\tau^2}{2\sigma^2}}.$$

To ensure that this probability be at most β, we thus estimate the error bound as

$$2Ne^{-\frac{\tau^2}{2\sigma^2}} \leq \beta \implies \tau \geq \sigma\sqrt{2\ln\left(\frac{2N}{\beta}\right)}.$$

Substituting the value of σ and N and simplifying we obtain:

$$\tau \geq \sqrt{\frac{\lceil \log_2(u) \rceil}{\rho} \log\left(2\frac{\lceil \log_2(u) \rceil}{\beta}\right)}.$$

□

Remark 5. *Assuming* $\lceil \log_2(u) \rceil = \log_2(u)$ *and using change-of-base we have that*

$$\sqrt{\frac{\lceil \log_2(u) \rceil}{\rho} \log\left(\frac{2\lceil \log_2(u) \rceil}{\beta}\right)} = \sqrt{\frac{\log(u)}{\log(2)\rho} \log\left(\frac{2\log(u)}{\log(2)\beta}\right)}.$$

Looking at the inner logarithm we have

$$\log\left(\frac{2\log(u)}{\beta\log(2)}\right) = \log\left(\frac{\log(u)}{\beta}\right) + \log\left(\frac{2}{\log(2)}\right),$$

which, using asymptotic approximations, leads to

$$\tau \approx \frac{\log(u)}{\rho\log(2)} \cdot \log\left(\frac{\log(u)}{\beta}\right),$$

since the second term is constant and the first term therefore dominates in u*. This gives the same polylogarithmic order of* τ *as in [37] which is near optimal.*

These results allow us to quantify to what extent the rank associated with the private quantile q^{DP} from Algorithm 3 differs from the non-private rank in the calibration data. Firstly, obtaining a ρ-zCDP quantile ensures that Algorithm 4 preserves the desired privacy level. Moreover, Proposition 2 allows us to explicitly define this difference τ with high probability (at least $1 - \beta$). We will see that this information can be used and is indeed helpful when determining the coverage guarantees of P-COQS in Algorithm 4 which are stated in Theorem 2.1. For this purpose, we also define $s_{\text{test}} = s(X_{\text{test}}, Y, \hat{f})$.

Theorem 2.1 (Coverage Guarantee). *Let* $(s_1, \ldots, s_{n_{cal}}, s_{test})$ *be an exchangeable sequence of non-conformity scores. Then, the P-COQS prediction set* $\mathcal{C}^{DP}(X_{test})$ *in Algorithm 4 satisfies:*

$$1 - \alpha - \frac{\tau}{n_{cal} + 1} \leq \mathbb{P}\left[Y_{test} \in \mathcal{C}^{DP}(X_{test})\right] \leq 1 - \alpha + \frac{\tau + 1}{n_{cal} + 1}.$$

Proof. By Proposition 2 we have that Algorithm 3 outputs a private quantile q^{DP} with rank error τ ensuring that $s_{r-\tau} \leq s_r \leq s_{r+\tau}$. By the exchangeability condition of the non-conformity scores, the rank of $s_{\text{test}} = s(X_{\text{test}}, Y, \hat{f})$ among $(s_1, \ldots, s_{n_{\text{cal}}}, s_{\text{test}})$ is uniformly distributed over $\{1, \ldots, n_{\text{cal}} + 1\}$. Thus, recalling that $r = \lceil (1 - \alpha)(n_{\text{cal}} + 1) \rceil$ and that

$$Y \in \mathcal{C}^{DP}(X_{\text{test}}) \iff s_{\text{test}} \leq q^{DP},$$

we consequently have:

$$\underbrace{\frac{r - \tau}{n_{\text{cal}} + 1}}_{L} \leq \mathbb{P}\left(s_{\text{test}} \leq q^{DP}\right) \leq \underbrace{\frac{r + \tau}{n_{\text{cal}} + 1}}_{U},$$

where:

$$L = \frac{\lceil (1 - \alpha)(n_{\text{cal}} + 1) \rceil - \tau}{n_{\text{cal}} + 1} \geq 1 - \alpha - \frac{\tau}{n_{\text{cal}} + 1} \quad \text{and}$$
$$U = \frac{\lceil (1 - \alpha)(n_{\text{cal}} + 1) \rceil + \tau}{n_{\text{cal}} + 1} \leq \frac{(1 - \alpha)(n_{\text{cal}} + 1) + 1 + \tau}{n_{\text{cal}} + 1} = 1 - \alpha + \frac{\tau + 1}{n_{\text{cal}} + 1}.$$

Therefore this allows us to conclude that

$$1 - \alpha - \frac{\tau}{n_{\text{cal}} + 1} \leq \mathbb{P}\left[Y_{\text{test}} \in \mathcal{C}^{DP}(X_{\text{test}})\right] \leq 1 - \alpha + \frac{\tau + 1}{n_{\text{cal}} + 1}.$$

□

Although Theorem 2.1 does not ensure the coverage probability $(1 - \alpha)$, as opposed to the method in [4] that preserves this standard CP guarantee, it establishes an approximate coverage bound with an error term of order $\mathcal{O}(\tau/n_{\text{cal}})$. While this indeed implies an error margin for the

actual coverage, it nevertheless also provides an upper bound to the coverage guarantee which allows to understand how conservative and how efficient the resulting prediction sets can be [see e.g. 48, 49]. More in detail, this error depends not only on the size of the calibration set but also on the width of the interval $[a, b]$ and the privacy level ρ under zCDP. Specifically, as shown in Proposition 2, the coverage error increases with the length of the interval and decreases with the privacy budget. Therefore, achieving accurate coverage requires a sufficiently large calibration set when the parameters of the error bound τ are fixed. Nevertheless, all the experimental results in Section 2.5 indicate that P-COQS is able to achieve coverage comparable to the non-private conformal prediction method, despite this weaker theoretical guarantee.

Remark 6. *As highlighted in [37], the quantile method adapted in Algorithm 3 is nearly optimal, implying that our results are suboptimal (polylogarithmic) in terms of CP guarantees. Nevertheless, using the bounds in Proposition 2 and following remarks, it delivers a practical way to determine the true differentially private coverage guarantee with high probability (i.e.* $1-\beta$*) without having to inflate the coverage to mitigate the impact of privacy noise (which can negatively impact the informativeness of the CP sets). For example, suppose that we are dealing with a classification task where the non-conformity scores lie in the interval* $[0, 1]$*, then by fixing the other parameters* δ *and* ρ *(e.g.,* $\delta = 10^{-10}$ *and* $\rho = 0.1$*) and choosing* $\beta = 0.01$*, following Proposition 2 the true probability of DP coverage of P-COQS is at least* $1-\alpha-0.01503$ *and at most* $1 - \alpha + 0.0154$ *guaranteed with probability at least* $1 - \beta = 0.99$ *and calibration size* $n_{cal} = 3000$*.*

2.5 Experiments

In this section, we evaluate the proposed CP approach (P-COQS) under the same experimental setting as [4]. Additionally, we extend this evaluation with controlled simulation experiments. Specifically, we compare P-COQS with the method introduced in [4] (henceforth referred to as EXPONQ), examining scenarios where both approaches employ either non-private models ($\hat{f}$) or DP models ($\hat{f}^{\mathrm{DP}}$). For differentially private models, we train Naïve Bayes and Random

Forest models using IBM's `Diffprivlib` library in Python,[1] ensuring pure ϵ-DP guarantees. For deep neural networks, we adopt the same methodology as [4], utilizing the `Opacus` library to obtain (ϵ, δ)-DP models. For DP conformal prediction methods, we leverage the exact equivalence between ϵ-DP and $(\epsilon, 0)$-zCDP established in Lemma 3.2 of [13]. This allows us to analyze both P-COQS and EXPONQ under either privacy framework while maintaining identical privacy guarantees. We specifically examine how these methods behave under varying privacy parameters (ϵ) and different sample sizes. Consistent with the experimental design of [4], we restrict our investigation to classification tasks. The non-conformity measure we used is given by the *hinge loss* function:

$$s(x, y, \hat{f}) = 1 - \hat{f}(x)_y,$$

where $\hat{f}(x)_y$ represents the predicted probability of the model for the true class y. We evaluate the performance of CP methods using the metrics defined below.

Definition 2 (Prediction Set Quality Metrics)**.** *Given test points $\{(x_i, y_i)\}_{i=1}^{n_{test}}$ with corresponding CP sets $\{\mathcal{C}(x_i)\}_{i=1}^{n_{test}}$:*

- ***(Marginal) Coverage**: proportion of prediction sets that cover the true label:*

$$\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbb{1}\{y_i \in \mathcal{C}(x_i)\}$$

- ***Efficiency**: average size of the prediction sets:*

$$\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \texttt{card}(\mathcal{C}(x_i))$$

- ***Informativeness** proportion of prediction sets of size 1:*

$$\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbb{1}\{\texttt{card}(\mathcal{C}(x_i)) = 1\}$$

[1] DP implementations based on IBM's Diffprivlib: `https://diffprivlib.readthedocs.io/en/latest/modules/models.html#classification-models`

Therefore, if a CP approach performs well, we expect (i) the coverage to be close to $1-\alpha$, (ii) the efficiency to be closest to 1 from above as possible, and (iii) the informativeness to be closest to 1 from below as possible. For all settings (simulations and benchmark datasets), where possible, we replicate all experimental settings and parameters used in [4].

2.5.1 Simulated Data

We consider a low-dimensional binary classification problem with eight features. The feature matrix for class 1 is sampled from a Gaussian $\mathcal{N}(\mu_1, \Sigma_1)$, with $\mu_1 = [0.8, \ldots, 0.8]^T \in \mathbb{R}^8$ and $\Sigma_1 = \mathbb{I} \cdot 7 \in \mathbb{R}^{8\times 8}$, while class 2 follows $\mathcal{N}(\mu_2, \Sigma_2)$ with $\mu_2 = [-1, \ldots, -1]$ and $\Sigma_2 = \mathbb{I} \cdot 8$, where $\mathbb{I}$ denotes the identity matrix. This setting delivers a reasonable overlap between the two classes and implies a slightly non-linear decision boundary due to the small difference in the covariance matrices. The generated data is class-balanced, with 60% used for the training set, 24% for the calibration set, and 16% reserved for evaluating CP performance. As for the models, considering the slight non-linearity of the decision boundary (and assuming, however, that in reality we would not know the underlying data-generating process), we consider the two classifiers mentioned earlier: Naïve Bayes (NB) and Random Forest (RF) (we will see applications with deep neural networks in Section 2.5.2). In particular, in this section we will focus on CP performance when employing the DP versions of these models (i.e. $\hat{f}^{DP}$), whereas the results using the non-DP models $\hat{f}$ can be found in the appendix. This being said, we evaluate coverage, efficiency, and informativeness with a fixed privacy budget $\epsilon_f = 2$ when privatizing the models and repeat experimental runs $H = 1000$ times in each setting (unless otherwise specified). Moreover, when varying one of the parameters to study the sensitivity of the methods, the other parameters are fixed at: $\epsilon_{CP} = 1$ (privacy budget for CP); $n = 10,000$ (total sample size to be split); $\alpha = 0.1$ (significance level for coverage). For clarity of presentation, here we only discuss the results for the NB classifier, while we provide the same results for the RF classifier in Appendices B.1 and B.2.

As mentioned earlier, in the following sections, we study the effects of the privacy budget ϵ_{CP} for CP in the EXPONQ and P-COQS approaches, as well as of the total sample size (which is then split into training, calibration, and test) and significance level α. However, it must

be stressed that all results are obviously limited to the above data-generating mechanism and therefore are to be interpreted within this setting.

2.5.1.1 Effect of CP Privacy Budget (ϵ_{CP})

We examine the sensitivity of both P-COQS and EXPONQ to variations in the CP privacy parameter ϵ_{CP}, using private NB model while maintaining a fixed sample size of $n = 10,000$. For the DP model (achieving 75% accuracy), the results in Table 2.1 indicate that P-COQS shows significantly lower sensitivity to ϵ_{CP} compared to EXPONQ, while comparing favorably to it over the different evaluation metrics across the considered privacy budgets. Similar behavior is observed for the private RF model (see Table B.4 in Appendix B.2.1) and also when using the non-DP variants of the NB and RF models reported in Table B.9 and Table B.1 respectively, where as expected, efficiency and informativeness results are slightly better for both approaches.

ϵ_{CP}	**Coverage**		**Efficiency**		**Informativeness**	
	ExponQ	**P-COQS**	**ExponQ**	**P-COQS**	**ExponQ**	**P-COQS**
0.1	1.0000 (0.0002)	0.9000 (0.0107)	1.9995 (0.0030)	1.3991 (0.0350)	0.0005 (0.0030)	0.6009 (0.0350)
0.5	0.9402 (0.0141)	0.8999 (0.0100)	1.5692 (0.0694)	1.3985 (0.0326)	0.4308 (0.0694)	0.6015 (0.0326)
1	0.9217 (0.0116)	0.8999 (0.0098)	1.4841 (0.0457)	1.3985 (0.0322)	0.5159 (0.0457)	0.6015 (0.0322)
1.5	0.9148 (0.0104)	0.9000 (0.0098)	1.4553 (0.0385)	1.3986 (0.0319)	0.5447 (0.0385)	0.6014 (0.0319)
3	0.9076 (0.0098)	0.9000 (0.0097)	1.4272 (0.0337)	1.3987 (0.0318)	0.5728 (0.0337)	0.6013 (0.0318)
5	0.9046 (0.0097)	0.8999 (0.0097)	1.4160 (0.0325)	1.3986 (0.0320)	0.5840 (0.0325)	0.6014 (0.0320)
10	0.9024 (0.0096)	0.8999 (0.0097)	1.4073 (0.0321)	1.3986 (0.0320)	0.5927 (0.0321)	0.6014 (0.0320)

Table 2.1: Average effect of CP privacy budget ϵ_{CP} on conformal prediction with DP Naïve Bayes model. The numbers are metric averages over 1000 runs (per method and privacy budget) and in parentheses is the corresponding variance of the metrics.

2.5.1.2 Sample Size Effect

We investigate the impact of sample size on the performance of the CP methods considered here. For each sample size, the sample is split as mentioned at the start of this section (i.e. 60% training, 24% calibration and 16% test). We recall that we discuss the results with the DP version of NB (see Appendix B.3.2 for the result of the non-DP model $\hat{f}$) and also recall that, as in other experiments, the privacy budget for the models is fixed at $\epsilon_f = 2$. The results in Table 2.2 show that in general P-COQS targets the correct $1 - \alpha$ coverage except in smaller

sample settings (that is, $n \leq 200$) and overall performs well compared to EXPONQ. More in detail, as discussed in the previous section, it can be seen how the sample size has an important effect on the P-COQS precision (measured in terms of metric variability) where, aside from visibly under-covering with respect to the $1 - \alpha$ target-level, smaller sample sizes show lower precision of P-COQS compared to EXPONQ (similar conclusions hold for RF in Appendices B.2.2 and B.1.2). This confirms the observations and conclusions made from Theorem 2.1. Nevertheless, it can be observed that the P-COQS under-covering appears to be solved already starting from $n = 200$ and does not appear to be an issue when using a non-DP model (see Table B.3.2 in the Appendix). In addition, the variability of the P-COQS metrics tend to converge towards that of EXPONQ (as also seen in the other simulation settings of this section).

n	**Coverage**		**Efficiency**		**Informativeness**		**Model Accuracy**
	ExponQ	**P-COQS**	**ExponQ**	**P-COQS**	**ExponQ**	**P-COQS**	
100	0.9791 (0.0447)	0.8529 (0.1472)	1.9266 (0.1180)	1.5958 (0.3184)	0.0734 (0.1180)	0.4042 (0.3184)	0.5867 (0.1248)
200	0.9954 (0.0148)	0.9295 (0.0731)	1.9896 (0.0287)	1.8077 (0.1663)	0.0104 (0.0287)	0.1923 (0.1663)	0.5543 (0.1039)
500	0.9969 (0.0074)	0.9043 (0.0474)	1.9830 (0.0274)	1.6153 (0.1206)	0.0170 (0.0274)	0.3847 (0.1206)	0.6289 (0.0612)
1000	0.9994 (0.0034)	0.9296 (0.0544)	1.9963 (0.0147)	1.7446 (0.1858)	0.0037 (0.0147)	0.2554 (0.1858)	0.6266 (0.0478)
2000	0.9761 (0.0130)	0.9014 (0.0220)	1.8469 (0.0577)	1.5316 (0.0644)	0.1531 (0.0577)	0.4684 (0.0644)	0.6972 (0.0268)
6000	0.9340 (0.0178)	0.9185 (0.0414)	1.6877 (0.0859)	1.6216 (0.1890)	0.3123 (0.0859)	0.3784 (0.1890)	0.6949 (0.0227)
10000	0.9217 (0.0116)	0.8999 (0.0098)	1.4841 (0.0457)	1.3985 (0.0322)	0.5159 (0.0457)	0.6015 (0.0322)	0.7479 (0.0125)

Table 2.2: Average effect of sample size on private conformal prediction with DP Naïve Bayes model. The numbers are metric averages over 1000 runs (per method and sample size) and in parentheses is the corresponding variance of the metrics.

2.5.1.3 Effect of Significance Level (α)

We examine the performance across varying α values (fixing $n = 10,000$, $\epsilon_f = 2$, and $\epsilon_{CP} = 1$). As seen in Table 2.3, P-COQS appears to better target the required coverage level on average, but has a higher variability than EXPONQ in this metric with a small $\alpha = 0.01$ level. This variability then becomes in line with (or smaller than) that of EXPONQ with larger values of α. With respect to the other metrics, the results indicate that P-COQS generally performs better than EXPONQ with the variability of the performance in line with the latter (similar conclusions holds for the DP RF model in Appendix B.2.3 as well as when using non-DP NB and RF models in Appendices B.3.3 and B.1.3 respectively).

α	Coverage		Efficiency		Informativeness	
	ExponQ	**P-COQS**	**ExponQ**	**P-COQS**	**ExponQ**	**P-COQS**
0.01	1.0000 (0.0002)	0.9988 (0.0045)	1.9995 (0.0030)	1.9857 (0.0511)	0.0005 (0.0030)	0.0143 (0.0511)
0.05	0.9732 (0.0088)	0.9500 (0.0076)	1.7579 (0.0604)	1.6152 (0.0373)	0.2421 (0.0604)	0.3848 (0.0373)
0.10	0.9217 (0.0116)	0.8999 (0.0098)	1.4841 (0.0457)	1.3985 (0.0322)	0.5159 (0.0457)	0.6015 (0.0322)

Table 2.3: Average effect of α on private conformal prediction using DP Naïve Bayes model. The numbers are metric averages over 1000 runs (per method and significance level) and in parentheses is the corresponding variance of the metrics.

2.5.1.4 Computational Efficiency

We compare the runtime of the considered methods under two settings which considered the non-DP and DP versions of each model (i.e. $\hat{f}$ and $\hat{f}^{DP}$ respectively). Confirming the model privacy budget to be $\epsilon_f = 2$, conformal privacy budget $\epsilon_{CP} = 1$, total sample size $n = 10,000$ and significance level $\alpha = 0.1$, we record the average runtime over 1000 trials. Table 2.4 highlights the considerable computational gain that P-COQS delivers with respect to EXPONQ while remaining comparable (or often better) across all performance metrics considered above (similar conclusions holds for the RF model in Appendix B.2.4).

	Coverage	Efficiency	Informativeness	Model Accuracy	Ave. time (secs)
$\hat{f}$ and EXPONQ	0.9227 (0.0116)	1.2509 (0.0360)	0.7491 (0.0360)	0.8253 (0.0092)	0.7562 (0.0111)
$\hat{f}$ and P-COQS	0.9006 (0.0010)	1.1788 (0.0201)	0.8212 (0.0201)	0.8253 (0.0092)	0.0072 (0.0002)
$\hat{f}^{DP}$ and EXPONQ	0.9217 (0.0116)	1.4841 (0.0457)	0.5159 (0.0457)	0.7479 (0.0125)	0.7534 (0.0152)
$\hat{f}^{DP}$ and P-COQS	0.8999 (0.0098)	1.3985 (0.0322)	0.6015 (0.0322)	0.7479 (0.0125)	0.0072 (0.0003)

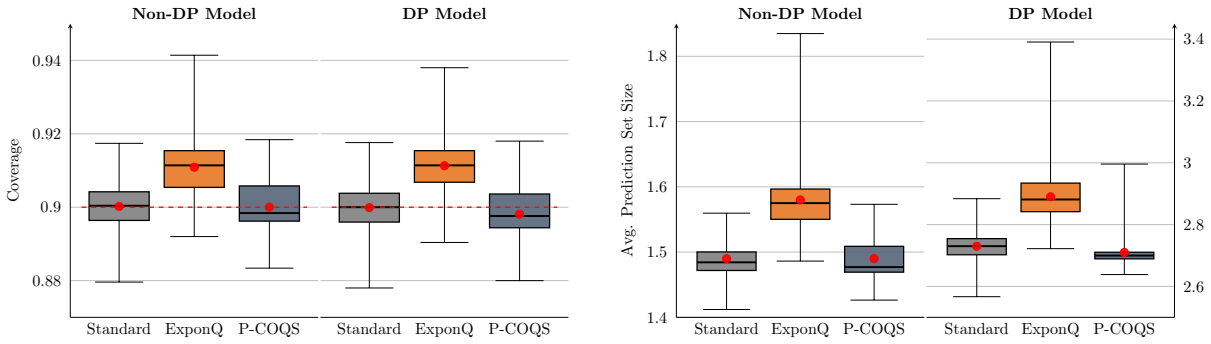
Table 2.4: Average computational time of P-COQS and EXPONQ combined with non-DP and DP versions of the NB classifier.

2.5.2 Datasets

In this section, we take the same settings as in [4] and study the performance of P-COQS when applied to three benchmark datasets: CIFAR-10 [46], ImageNet [17] and CoronaHack [60]. As in the simulation experiment, we compare the performance of P-COQS with that of EXPONQ proposed in [4] and follow the same experimental setup as in the latter. Unless otherwise stated, the significance level is fixed at $\alpha = 0.1$, CP privacy budget at $\epsilon_{CP} = 1$ and the DP models are trained to achieve (ϵ, δ)-DP using the `Opacus` library with ($\epsilon = 8$ and $\delta = 10^{-5}$).

2.5.2.1 CIFAR-10 Benchmark Analysis

We first evaluate P-COQS on the CIFAR-10 dataset, comparing its performance to EXPONQ under different privacy settings. Following [4], we consider two scenarios for model training: (i) one where a non-DP model is used ($\hat{f}$) and (ii) one where a DP model is used ($\hat{f}^{DP}$). Both DP and non-DP models share the same convolutional neural network architecture. The DP model achieved an accuracy of 60% compared to 77% for the non-DP model. The evaluation is based on 1000 random splits of the CIFAR-10 validation set, each of size $n = 5000$. Figure 2.1 displays the empirical coverage and prediction set sizes across these settings. Our findings suggest that P-COQS generally produces smaller prediction sets (Figure 2.1b), while achieving empirical coverage levels that are comparable to those of the standard (non-private) CP baseline (Figure 2.1a), except for the case where a DP model is used where P-COQS appears to slightly undercover with respect to the standard CP baseline. On the other hand, compared to P-COQS, by inflating the quantile EXPONQ guarantees the coverage level but tends to have larger prediction sets for this reason. In the latter case, for all methods it can be seen that prediction sets are larger when using a DP model (as expected).



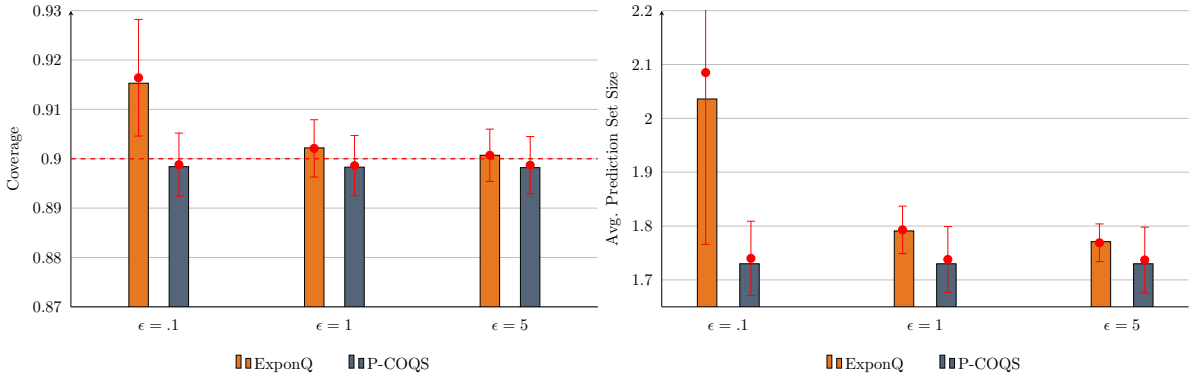
(a) Coverage

(b) Avg. Prediction Set Size

Figure 2.1: Results of Standard, EXPONQ and P-COQS methods applied to the CIFAR-10 dataset. (a) Boxplots with coverage distributions of the three methods under the non-DP model (left) and DP model (right) scenarios; (b) Boxplots with distributions of average set sizes of the three methods under the non-DP model (left) and DP model (right) scenarios.

2.5.2.2 Large-Scale Evaluation on ImageNet

We apply both considered private CP methods to the ImageNet dataset using a pre-trained (non-DP) ResNet-152 model. We evaluated the performance of P-COQS and EXPONQ under varying values of the conformal privacy budget ϵ_{CP}. The evaluation uses $n = 30,000$ samples for calibration and $20,000$ samples for validation. For each budget value ϵ_{CP}, performance metrics are computed over 100 random splits of the ImageNet validation set, following the procedure in [4]. As shown in Figure 2.2, P-COQS appears to slightly under-cover but this does not appear to be significant based on the 95% confidence intervals (red vertical whiskers in the plots) and remains stable in performance across the different values of conformal privacy budget, whereas EXPONQ appears to significantly over-cover for the small budgets (see Figure 2.2a). With this in mind, P-COQS also appears to be stable with respect to prediction set sizes while EXPONQ approaches these sizes only as ϵ_{CP} increases (as highlighted also in the simulation results in Section 2.5.1).



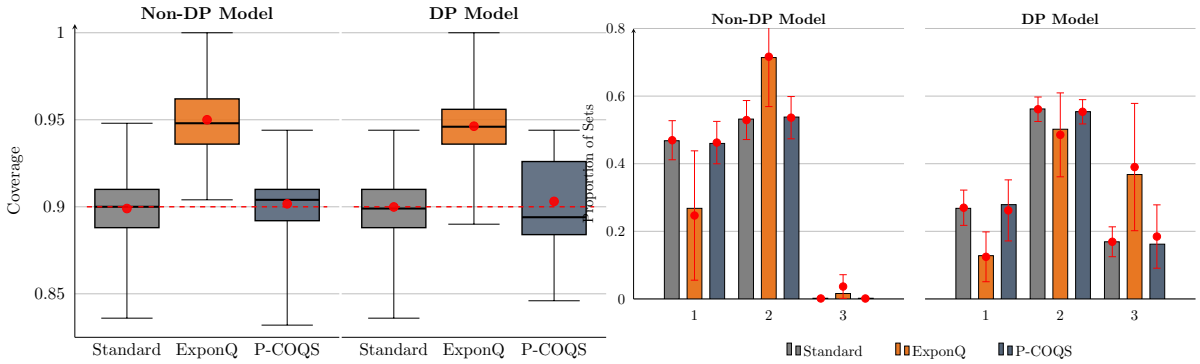
(a) Coverage (b) Avg. Prediction Set Size

Figure 2.2: Results of EXPONQ and P-COQS methods applied to the ImageNet dataset. (a) Median coverage (bar height), mean coverage (red dots) and 95% confidence intervals for coverage (red vertical whiskers) for each method under different privacy budgets ϵ (red-dashed horizontal line is the target coverage); (b) Average prediction set sizes (average set sizes in 100 splits): median (bar height), mean (red dots) and 95% confidence intervals (red vertical whiskers) for each method under different privacy budgets ϵ.

2.5.2.3 Medical Imaging Analysis with CoronaHack Dataset

Finally, we evaluate both methods on the CoronaHack dataset, a publicly available chest X-ray dataset comprising 5908 images categorized into three classes: normal, viral pneumonia

(primarily COVID-19), and bacterial pneumonia. We fine-tuned the final layer of a ResNet-18 model over 14 epochs using 4408 training samples, under both DP and non-DP settings. The DP model achieved 64% accuracy compared to 70% for the non-DP model. CP was calibrated on 1000 samples and validated on 500 samples, across 1000 random splits of the data. Figure 2.3 presents the results for different configurations of the DP and non-DP models. While EXPONQ tends to produce overly conservative prediction sets, P-COQS appears to better target the nominal coverage level of 0.9 (see Figure 2.3a). Moreover, looking at the prediction set sizes, the distribution of the P-COQS appears to deliver roughly similar set sizes to the non-private CP approach since the proportions of sets of specific sizes (1, 2 or 3 on the x-axis) are very close to each other (see Figure 2.3b). Also in this case EXPONQ produces larger sets (i.e. higher proportions for sets of size 2 or 3), where again the sizes tend to increase overall when using the DP model.



(a) Coverage

(b) Distribution of Prediction Set Sizes

Figure 2.3: Results of Standard, EXPONQ and P-COQS methods applied to the CoronaHack dataset. (a) Boxplots with coverage distributions of the three methods under the non-DP model (left) and DP model (right) scenarios; (b) Distribution of prediction set sizes where the y-axis represents the empirical proportion of sets of size 1, 2 or 3 (represented on the x-axis) for each method: median proportion (bar height), mean proportion (red dots) and 95% confidence intervals (red vertical whiskers) for each method under the non-DP model (left) and DP model (right) scenarios.

2.6 Conclusion

The proposed P-COQS provides a computationally efficient DP alternative to existing privacy-preserving approaches. It does so by adapting an existing DP quantile method and trading off some theoretical guarantees on the lower bound for the coverage probability, while however

guaranteeing quantifiable lower and upper error bounds for this probability. In this way, it is possible to determine (tight) error bounds which can then be used to assess the true minimum coverage level with high probability. Despite this theoretical trade-off, which however remains small (if not negligible) in common settings, the P-COQS shows an empirical performance over different metrics that is in line with standard non-private CP, indicating that it targets the correct coverage level in all settings while delivering more precise CP sets compared to the existing DP alternative (that we denoted as EXPONQ). The theoretical trade-off of the P-COQS can be observed mainly in settings with small sample sizes and/or privacy budgets in which the variability of P-COQS metrics is indeed larger compared to EXPONQ. However, the latter approach also makes a trade-off by generally over-covering with respect to the required coverage level in smaller samples and for small privacy budgets. As a consequence, the latter tends to deliver less efficient and informative prediction sets in these settings and appears to only approach those of P-COQS for larger privacy budgets and/or sample sizes. In general, considering these different trade-offs and without claiming the existence of a better approach, the proposed P-COQS appears to provide a valid privacy-preserving (and computationally efficient) CP approach also based on its generally good performance compared to EXPONQ in the different experimental settings considered. The implementation code for all experiments presented in this chapter is publicly available in our Github repository. Additionally, the associated research paper is available as a preprint on ARXIV. Furthermore, the `pcoqs` R package, which implements the P-COQS methodology, can be installed from its GitHub repository. Although both EXPONQ and P-COQS can be applied within different forms of CP (e.g. full CP), an important direction for improving the P-COQS method is to refine its theoretical coverage bounds to achieve the minimal CP $(1-\alpha)$-coverage guarantees, without overly compromising the statistical efficiency and informativeness of the prediction sets or the computational efficiency of the method, particularly when the calibration sample size and/or privacy budget is small. Additionally, extending P-COQS to more general settings, where the exchangeability assumption is relaxed, presents an important avenue for future research.

References

[1] Alan Agresti. *Categorical data analysis*. John Wiley & Sons, 2013.

[2] Daniel Alabi and Salil Vadhan. Hypothesis testing for differentially private linear regression. *Advances in Neural Information Processing Systems*, 35:14196–14209, 2022.

[3] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[4] Anastasios Nikolas Angelopoulos, Stephen Bates, Tijana Zrnic, and Michael I Jordan. Private prediction sets. *Harvard Data Science Review*, 4(2), 2022.

[5] Mario Astigarraga, Andrés Sánchez-Ruiz, and Gonzalo Colmenarejo. Conformal prediction-based machine learning in cheminformatics: Current applications and new challenges. *Artificial Intelligence in the Life Sciences*, page 100127, 2025.

[6] Jordan Awan and Aleksandra Slavković. Differentially private uniformly most powerful tests for binomial data. *Advances in Neural Information Processing Systems*, 31, 2018.

[7] Jordan Awan and Zhanyu Wang. Simulation-based, finite-sample inference for privatized data. *arXiv preprint arXiv:2303.05328*, 2023.

[8] Jordan Awan and Zhanyu Wang. Simulation-based, finite-sample inference for privatized data. *Journal of the American Statistical Association*, (just-accepted):1–27, 2024.

[9] Garrett Bernstein and Daniel R Sheldon. Differentially private Bayesian inference for exponential families. *Advances in Neural Information Processing Systems*, 31:2919–2929, 2018.

[10] Garrett Bernstein and Daniel R Sheldon. Differentially private Bayesian linear regression. *Advances in Neural Information Processing Systems*, 32:525–535, 2019.

[11] Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

[12] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.

[13] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *TCC*, 2016.

[14] Clément L Canonne, Gautam Kamath, Audra McMillan, Adam Smith, and Jonathan Ullman. The structure of optimal private tests for simple hypotheses. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 310–321, 2019.

[15] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[16] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Number 1. Cambridge university press, 1997.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[18] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022.

[19] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[20] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[21] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[22] Bradley Efron. Bayesian inference and the parametric bootstrap. *The annals of applied statistics*, 6(4):1971, 2012.

[23] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

[24] Cecilia Ferrando, Shufan Wang, and Daniel Sheldon. Parametric bootstrap for differentially private confidence intervals. In *International Conference on Artificial Intelligence and Statistics*, pages 1598–1618. PMLR, 2022.

[25] George Fishman. *Monte Carlo: concepts, algorithms, and applications*. Springer Science & Business Media, 2013.

[26] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[27] Marco Gaboardi, Hyun Lim, Ryan Rogers, and Salil Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *International conference on machine learning*, pages 2111–2120. PMLR, 2016.

[28] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016.

[29] Quan Geng and Pramod Viswanath. The optimal noise-adding mechanism in differential privacy. *IEEE Transactions on Information Theory*, 62(2):925–951, 2015.

[30] Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. *arXiv preprint arXiv:2106.00170*, 2021.

[31] Ruobin Gong. Exact inference with approximate computation for differentially private data via perturbations. *arXiv preprint arXiv:1909.12237*, 2019.

[32] Christian Gourieroux, Alain Monfort, and Eric Renault. Indirect inference. *Journal of applied econometrics*, 8(S1):S85–S118, 1993.

[33] Stéphane Guerrier, Elise Dupuis-Lozeron, Yanyuan Ma, and Maria-Pia Victoria-Feser. Simulation-based bias correction methods for complex models. *Journal of the American Statistical Association*, 114(525):146–157, 2019.

[34] Stéphane Guerrier, Mucyo Karemera, Samuel Orso, Maria-Pia Victoria-Feser, and Yuming Zhang. A general approach for simulation-based bias correction in high dimensional settings. *arXiv preprint arXiv:2010.13687*, 2020.

[35] Jan Hannig. On generalized fiducial inference. *Statistica Sinica*, pages 491–544, 2009.

[36] Jan Hannig, Hari Iyer, Randy CS Lai, and Thomas CM Lee. Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, 111(515):1346–1361, 2016.

[37] Ziyue Huang, Yuting Liang, and Ke Yi. Instance-optimal mean estimation under differential privacy. *Advances in Neural Information Processing Systems*, 34:25993–26004, 2021.

[38] Pierre Humbert, Batiste Le Bars, Aurélien Bellet, and Sylvain Arlot. One-shot federated conformal prediction. In *International Conference on Machine Learning*, pages 14153–14177. PMLR, 2023.

[39] Jef Jonkers, Glenn Van Wallendael, Luc Duchateau, and Sofie Van Hoecke. Conformal predictive systems under covariate shift. In *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, pages 406–423. PMLR, 2024.

[40] Nianqiao Ju, Jordan Awan, Ruobin Gong, and Vinayak Rao. Data augmentation MCMC for Bayesian inference from privatized data. *Advances in Neural Information Processing Systems*, 35:12732–12743, 2022.

[41] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *Advances in neural information processing systems*, 27, 2014.

[42] Vishesh Karwa, Dan Kifer, and Aleksandra B Slavković. Private posterior distributions from variational approximations. *arXiv preprint arXiv:1511.07896*, 2015.

[43] Zeki Kazan, Kaiyan Shi, Adam Groce, and Andrew P Bray. The test of tests: A framework for differentially private hypothesis testing. In *International Conference on Machine Learning*, pages 16131–16151. PMLR, 2023.

[44] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, pages 5574–5584, 2017.

[45] Daniel Kifer and Ryan Rogers. A new class of private chi-square tests. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 17, pages 991–1000, 2016.

[46] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[47] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[48] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[49] Miao Li, Michael Klamkin, Mathieu Tanneau, Reza Zandehshahvar, and Pascal Van Hentenryck. Conformal prediction with upper and lower bound models. *arXiv preprint arXiv:2503.04071*, 2025.

[50] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[51] Joseph P Near and Chiké Abuah. Programming differential privacy. *URL: https://uvm*, 2021.

[52] Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

[53] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84, 2007.

[54] Henrik Olsson, Kimmo Kartasalo, Nita Mulliqi, Marco Capuccini, Pekka Ruusuvuori, Hemamali Samaratunga, Brett Delahunt, Cecilia Lindskog, Emiel AM Janssen, Anders Blilie, et al. Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. *Nature communications*, 13(1):7761, 2022.

[55] Samuel Orso, Mucyo Karemera, Maria-Pia Victoria-Feser, and Stéphane Guerrier. An accurate percentile method for parametric inference based on asymptotically biased estimators. *arXiv preprint arXiv:2405.05403*, 2024.

[56] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In *Advances in Neural Information Processing Systems*, pages 4026–4034, 2016.

[57] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European*

conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13, pages 345–356. Springer, 2002.

[58] Víctor Peña and Andrés F Barrientos. Differentially private hypothesis testing with the subsampled and aggregated randomized response mechanism. *arXiv preprint arXiv:2208.06803*, 2022.

[59] Coby Penso, Bar Mahpud, Jacob Goldberger, and Or Sheffet. Privacy-preserving conformal prediction under local differential privacy. *arXiv preprint arXiv:2505.15721*, 2025.

[60] Juan Caballero Pérez, José Manuel Caballero Contreras, Carlos de Blas Pérez, and Felipe Leza Alvarez. Databiology lab coronahack: Collection of public covid-19 data. *bioRxiv*, pages 2020–10, 2020.

[61] Christian P Robert. Approximate bayesian computation: a survey on recent results. In *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Leuven, Belgium, April 2014*, pages 185–205. Springer, 2016.

[62] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[63] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel J Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. *arXiv preprint arXiv:1904.06019*, 2019.

[64] Rudraksh Tuwani and Andrew Beam. Safe and reliable transport of prediction models to new healthcare settings without the need to collect new labeled data. *medRxiv*, 2023.

[65] Nasir Uddin and Tuwe Lofstrom. Applications of conformal regression on real-world industrial use cases using crepes and mapie. In *Conformal and Probabilistic Prediction with Applications*, pages 147–165. PMLR, 2023.

[66] Caroline Uhlerop, Aleksandra Slavković, and Stephen E Fienberg. Privacy-preserving data sharing for genome-wide association studies. *The Journal of privacy and confidentiality*, 5(1):137, 2013.

[67] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[68] Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28, 2015.

[69] Duy Vu and Aleksandra Slavkovic. Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, pages 138–143. IEEE, 2009.

[70] Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd international conference on artificial intelligence and statistics*, pages 1226–1235. PMLR, 2019.

[71] Yue Wang, Daniel Kifer, Jaewoo Lee, and Vishesh Karwa. Statistical approximating distributions under differential privacy. *Journal of Privacy and Confidentiality*, 8(1), 2018.

[72] Yue Wang, Daniel Kifer, Jaewoo Lee, and Vishesh Karwa. Statistical approximating distributions under differential privacy. *Journal of Privacy and Confidentiality*, 8(1), 2018.

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

[74] Oliver Williams and Frank McSherry. Probabilistic inference and differential privacy. *Advances in Neural Information Processing Systems*, 23:2451–2459, 2010.

[75] Min-ge Xie and Peng Wang. Repro samples method for finite-and large-sample inferences. *arXiv preprint arXiv:2206.06421*, 2022.

[76] Yuming Zhang, Yanyuan Ma, Samuel Orso, Mucyo Karemera, Maria-Pia Victoria-Feser, and Stéphane Guerrier. Just identified indirect inference estimator: Accurate inference through bias correction. *arXiv preprint arXiv:2204.07907*, 2022.

Appendices

Appendix A

Additional Experimental Results on Chapter 1

A.0.1 One sample proportion: sample size effect on coverage of CI

We explore a few more experimental settings in addition to those in Sec. 1.5, in particular we investigate the behavior of the different methods with larger and varying sample sizes. We first study the empirical performance of the approaches for the one-sample proportion case to understand how CI coverage (at level 0.95) and lengths behave with increasing sample sizes. We look at sample sizes in the following range: $n = \{8 \times 2^k : k = 0, \ldots, 6\} \cup \{700, 900, 1200\}$. The results of our experiments are shown in Fig. A.1 where the left plot shows coverage and the right one shows lengths.

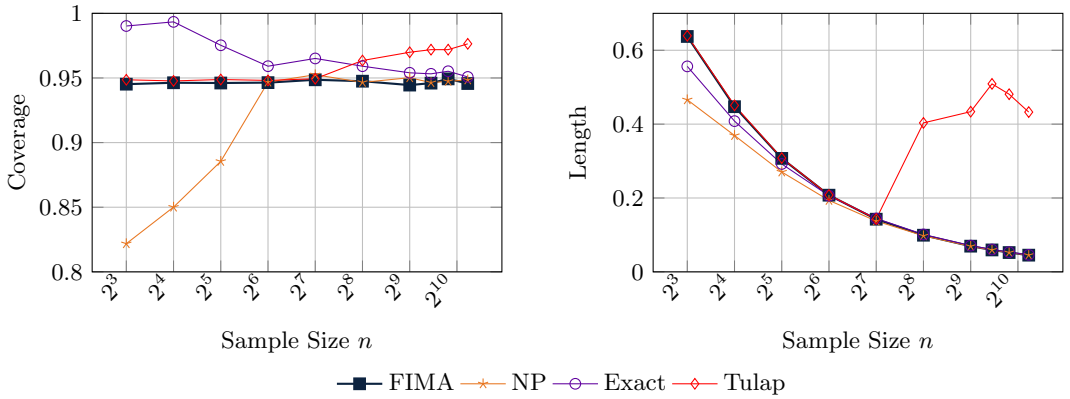


Figure A.1: One-sample CI coverage (left) and length (right) with different sample sizes

We can see how the FIMA shows a consistent and good coverage across all sample sizes, while the non-private CIs converge to the nominal level (from above and below respectively) as the sample size increases. We can see however that, as the sample size increases, at a certain

point the Tulap starts to over-cover. The latter corresponds to the point in the right plot where the Tulap CI lengths start to increase as well. This behavior of the Tulap, as for the ToT in other experiments, could be due to numerical issues in their implementation and should therefore not be considered as conclusive on the validity of these alternatives. This being said, in these experimental settings the FIMA CI lengths instead do indeed converge to the lengths of the non-private CIs as the sample size increases while maintaining good coverage.

A.0.2 Two sample proportions: level and power for $n_1 = n_2 = 100$

We also investigate the behaviour of the approaches considered for the two-sample proportion setting with a larger sample size $n_1 = n_2 = 100$ (instead of 30). The results for level and power of the same test as in Sec. 1.5.2 are presented in Fig. A.3.

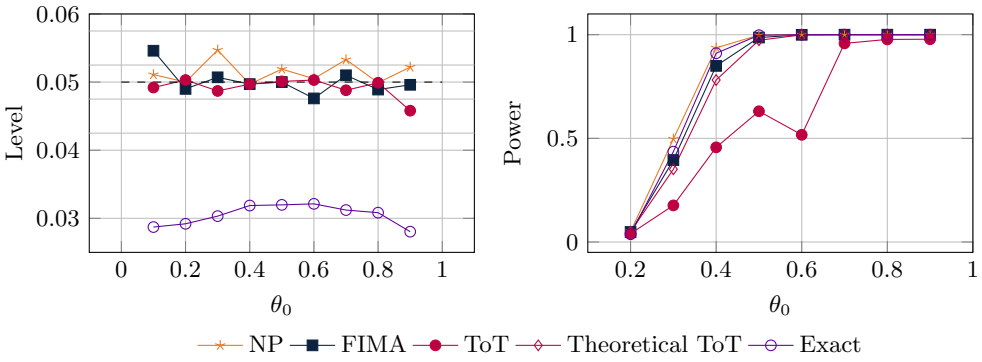


Figure A.2: Level (left plot) and power (right plot) evaluation of different test procedures under a one-sided hypothesis test for two proportions with $n = 100$.

It can be observed how all tests improve their level (left plot), especially FIMA and NP (non-private) while the ToT remains as good as in the smaller sample case. However, the right plot shows that the power of the FIMA improves and gets closer to the NP power compared to the theoretical ToT. The empirical ToT can be seen to greatly improve with respect to the small sample setting of $n = 30$ (see Sec. 1.5.2), although it still has lower power than what it should achieve theoretically (it would therefore probably require a better calibration of its hyper-parameters for these problems).

A.0.3 One and two sample proportions: sample size effect on power

We finally look at the behavior of hypothesis testing with increasing sample sizes (compared to the results when varying parameter values). Specifically we study the performance of the different approaches considered for one- and two-sample proportion tests. In the one-sample case, we fix $\theta_0 = 0.95$ and test the hypotheses $H_0 : \theta_0 \leq 0.9$ versus $H_A : \theta_0 > 0.9$, while for the two-sample setting we fix $\theta_1 = 0.8$ and $\theta_2 = 0.9$ and test the hypotheses $H_0 : \theta_1 \geq \theta_2$ versus $H_A : \theta_1 < \theta_2$.

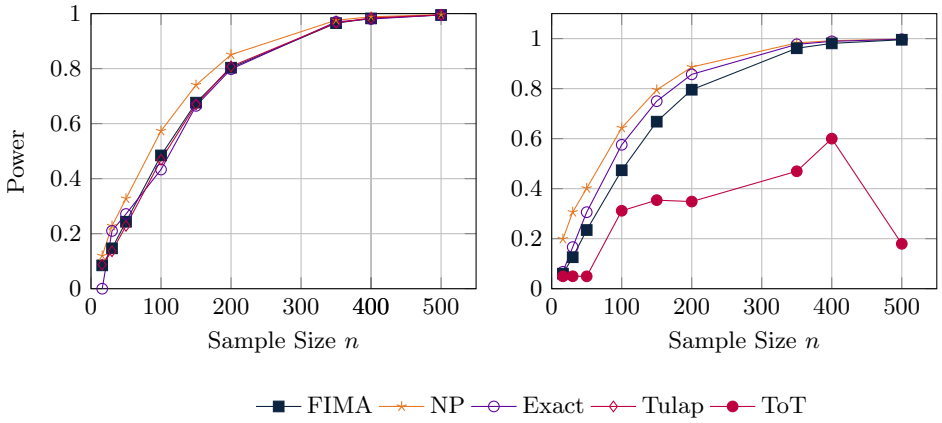


Figure A.3: Power with increasing sample sizes for one-sample proportion test (left plot) and two-sample proportion test (right plot) of different test procedures under a one-sided hypothesis test.

A.0.4 Logistic regression

A.0.4.1 Effect of privacy budget ϵ

Using the same parameter configuration as in Subsection 1.5.4, we analyze the influence of the privacy budget ϵ on the statistical properties of the confidence intervals for the model parameters. The empirical results, presented in Tables A.1 and A.2, demonstrate these effects for the cases with one and two predictors, respectively.

n	$\epsilon = 0.1$				$\epsilon = 0.3$				$\epsilon = 1$				$\epsilon = \infty$			
	β_0		β_1		β_0		β_1		β_0		β_1		β_0		β_1	
	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.
15	0.956	12.83	0.990	18.08	0.957	10.74	0.985	15.00	0.952	5.66	0.976	9.41	0.937	3.20	0.971	7.14
30	0.953	12.47	0.987	16.84	0.952	9.36	0.980	11.95	0.949	3.95	0.972	6.92	0.931	2.50	0.963	4.92
100	0.951	6.88	0.970	10.57	0.951	2.02	0.970	5.84	0.950	1.16	0.969	3.83	0.959	1.07	0.958	2.80
200	0.951	4.05	0.965	6.65	0.951	1.25	0.956	3.72	0.950	0.88	0.946	2.06	0.944	0.84	0.944	1.73
500	0.952	1.25	0.949	3.74	0.952	0.63	0.947	1.75	0.950	0.54	0.948	1.12	0.950	0.53	0.946	1.06
1000	0.953	0.63	0.951	2.32	0.952	0.40	0.946	0.96	0.955	0.37	0.950	0.77	0.953	0.37	0.949	0.75
2000	0.951	0.36	0.948	1.07	0.951	0.27	0.942	0.59	0.949	0.26	0.945	0.53	0.950	0.26	0.945	0.52

Table A.1: Coverage probabilities and interval lengths of FIMA for logistic regression with a single binary predictor. Results are reported for varying sample sizes n and privacy budgets ϵ, with significance level $\alpha = 0.05$.

n	$\epsilon = 0.1$						$\epsilon = 0.3$						$\epsilon = 1$						$\epsilon = 10$					
	β_0		β_1		β_2		β_0		β_1		β_2		β_0		β_1		β_2		β_0		β_1		β_2	
	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len	Cov	Len
15	.987	12.6	.995	17.0	.994	17.9	.979	11.8	.992	16.3	.993	16.5	.979	8.6	.986	11.9	.987	12.8	.921	3.9	.970	8.6	.975	7.8
30	.975	12.9	.993	18.1	.992	18.1	.987	11.9	.994	15.4	.992	14.9	.968	7.0	.981	10.0	.981	9.6	.944	3.5	.970	7.2	.943	5.3
100	.950	11.1	.982	15.5	.984	14.3	.950	5.9	.972	9.3	.959	8.0	.946	1.6	.973	4.7	.956	4.3	.961	1.7	.964	3.8	.950	2.8
200	.953	7.8	.971	10.2	.965	10.6	.954	2.6	.970	5.8	.947	4.6	.950	1.4	.963	3.7	.950	2.3	.955	1.1	.948	2.5	.950	1.9
500	.951	3.0	.961	5.5	.948	5.3	.950	1.0	.951	3.2	.950	1.9	.951	0.7	.944	1.9	.953	1.2	.954	0.7	.950	1.5	.958	1.2
1000	.953	1.1	.951	3.9	.949	2.4	.949	0.6	.947	1.8	.956	1.1	.949	0.5	.947	1.2	.953	0.8	.950	0.6	.949	1.1	.960	0.9
2000	.953	0.6	.949	2.3	.942	1.2	.952	0.4	.948	1.0	.956	0.7	.951	0.4	.951	0.8	.958	0.6	.947	0.4	.944	0.8	.960	0.6

Table A.2: Coverage probabilities and interval lengths of FIMA for logistic regression with two binary predictors. Results are reported for varying sample sizes n and privacy budgets ϵ, with significance level $\alpha = 0.05$.

A.0.4.2 Effect of significance level α

Assuming the same parameter setting as in Subsection 1.5.4, we vary the significance level to see how that impacts the coverage and length of the confidence intervals for the parameters. Tables A.3 and A.4 show the result for one and two predictors, respectively.

n	β_0						β_1					
	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.
15	0.994	9.26	0.952	5.66	0.898	4.36	0.996	14.26	0.976	9.41	0.945	7.43
30	0.991	6.31	0.949	3.95	0.899	3.09	0.996	10.53	0.972	6.92	0.938	5.48
100	0.988	1.54	0.950	1.16	0.901	0.97	0.995	5.70	0.969	3.83	0.919	3.06
200	0.989	1.16	0.950	0.88	0.899	0.74	0.990	2.92	0.946	2.06	0.898	1.69
500	0.990	0.71	0.950	0.54	0.901	0.45	0.988	1.49	0.948	1.12	0.898	0.94
1000	0.992	0.49	0.955	0.37	0.903	0.31	0.990	1.01	0.950	0.77	0.894	0.64
2000	0.990	0.34	0.949	0.26	0.897	0.22	0.988	0.70	0.945	0.53	0.887	0.44

Table A.3: Empirical coverage probabilities and confidence interval lengths of the FIMA for logistic regression with a single binary predictor. Results are shown across different sample sizes n and significance levels α, with fixed privacy budget $\epsilon = 1$.

n	β_0						β_1						β_2					
	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.
15	1.000	16.8	0.987	9.5	0.985	8.0	1.000	19.2	0.986	13.9	0.985	10.2	1.000	22.2	0.989	13.4	0.974	10.7
30	0.994	8.4	0.950	4.7	0.958	7.3	0.996	13.0	0.965	10.0	0.980	9.4	0.997	13.7	0.977	9.9	0.971	9.1
100	0.991	2.8	0.949	1.5	0.897	1.7	0.995	8.1	0.975	4.9	0.947	4.5	0.991	6.3	0.963	4.5	0.897	3.1
200	0.990	1.8	0.950	1.2	0.904	1.0	0.994	5.5	0.959	3.4	0.918	3.0	0.988	3.0	0.948	2.2	0.898	1.9
500	0.991	1.0	0.953	0.8	0.902	0.7	0.990	2.8	0.947	1.7	0.895	1.5	0.991	1.7	0.954	1.3	0.904	1.1
1000	0.991	0.7	0.951	0.5	0.894	0.4	0.989	1.4	0.945	1.2	0.898	1.0	0.994	1.2	0.960	0.9	0.909	0.7
2000	0.991	0.5	0.951	0.4	0.900	0.3	0.989	1.0	0.941	0.8	0.895	0.7	0.993	0.8	0.958	0.6	0.912	0.5

Table A.4: Empirical coverage probabilities and confidence interval lengths of the FIMA for logistic regression with two binary predictors. Results are shown across different sample sizes n and significance levels α, with fixed privacy budget $\epsilon = 1$.

The data used in Subsection 1.5.6.3 can be accessed at the following URL: `https://data.census.gov/table/ACSST5Y2022.S1701?t=Income%20and%20Poverty&g=010XX00US_040XX00US01,04,11,29,36,48_050XX00US01081_160XX00US0103076_860XX00US36830&y=2022`

Appendix B

Additional Experimental Results on Chapter 2

B.1 Simulation Results With Non-DP Random Forest Model

B.1.1 Effect of ϵ_{CP} with non-DP model

The parameter settings are as in the non-DP NB model and the model's accuracy is 81%.

$\epsilon_{\text{conformal}}$	**Coverage**		**Efficiency**		**Informativeness**	
	ExponQ	**P-COQS**	**ExponQ**	**P-COQS**	**ExponQ**	**P-COQS**
0.1	0.9998 (0.0004)	0.9024 (0.0104)	1.9885 (0.0123)	1.2222 (0.0243)	0.0115 (0.0123)	0.7778 (0.0243)
0.5	0.9464 (0.0188)	0.9025 (0.0098)	1.4094 (0.1213)	1.2223 (0.0220)	0.5906 (0.1213)	0.7777 (0.0220)
1	0.9228 (0.0123)	0.9026 (0.0098)	1.2936 (0.0408)	1.2224 (0.0218)	0.7064 (0.0408)	0.7776 (0.0218)
1.5	0.9154 (0.0107)	0.9026 (0.0098)	1.2660 (0.0294)	1.2225 (0.0217)	0.7340 (0.0294)	0.7775 (0.0217)
3	0.9084 (0.0099)	0.9026 (0.0098)	1.2415 (0.0235)	1.2225 (0.0220)	0.7585 (0.0235)	0.7775 (0.0220)
5	0.9058 (0.0099)	0.9026 (0.0099)	1.2327 (0.0224)	1.2225 (0.0220)	0.7673 (0.0224)	0.7775 (0.0220)
10	0.9041 (0.0100)	0.9026 (0.0099)	1.2273 (0.0224)	1.2226 (0.0219)	0.7727 (0.0224)	0.7774 (0.0219)

Table B.1: Average effect of CP privacy budget ϵ_{CP} on conformal prediction with non-DP Random Forest model. The numbers are metric averages over 1000 runs (per method and privacy budget) and in parentheses is the corresponding variance of the metrics.

B.1.2 Sample size effect with non-DP model

Samples	Coverage		Efficiency		Informativeness	
	ExponQ	**P-COQS**	**ExponQ**	**P-COQS**	**ExponQ**	**P-COQS**
100	0.9892 (0.0326)	0.9280 (0.0973)	1.8658 (0.1716)	1.5599 (0.3114)	0.1342 (0.1716)	0.4401 (0.3114)
200	0.9931 (0.0183)	0.9298 (0.0678)	1.8925 (0.1308)	1.4749 (0.2301)	0.1075 (0.1308)	0.5251 (0.2301)
500	0.9973 (0.0076)	0.9083 (0.0455)	1.9274 (0.0872)	1.3108 (0.1210)	0.0726 (0.0872)	0.6892 (0.1210)
1000	0.9987 (0.0036)	0.9056 (0.0304)	1.9520 (0.0577)	1.2706 (0.0786)	0.0480 (0.0577)	0.7294 (0.0786)
2000	0.9865 (0.0126)	0.9054 (0.0227)	1.7402 (0.1492)	1.2501 (0.0561)	0.2598 (0.1492)	0.7499 (0.0561)
6000	0.9393 (0.0181)	0.9026 (0.0123)	1.3745 (0.0986)	1.2268 (0.0293)	0.6255 (0.0986)	0.7732 (0.0293)
10000	0.9229 (0.0120)	0.9025 (0.0098)	1.2939 (0.0406)	1.2224 (0.0222)	0.7061 (0.0406)	0.7776 (0.0222)

Table B.2: Average effect of sample size on DP CP ($\epsilon_{CP} = 1$) with non-DP Random Forest model. The numbers are metric averages over 1000 runs (per method and sample size) and in parentheses is the corresponding variance of the metrics. Average effect of sample size on private CP with Non-DP Random Forest model.

B.1.3 Effect of α with non-DP model

α	Coverage		Efficiency		Informativeness	
	ExponQ	**P-COQS**	**ExponQ**	**P-COQS**	**ExponQ**	**P-COQS**
0.01	0.9998 (0.0004)	0.9909 (0.0033)	1.9885 (0.0123)	1.7554 (0.0370)	0.0115 (0.0123)	0.2446 (0.0370)
0.05	0.9800 (0.0120)	0.9516 (0.0072)	1.6512 (0.1426)	1.4204 (0.0275)	0.3488 (0.1426)	0.5796 (0.0275)
0.10	0.9229 (0.0120)	0.9025 (0.0098)	1.2939 (0.0406)	1.2224 (0.0222)	0.7061 (0.0406)	0.7776 (0.0222)

Table B.3: Average effect of α on DP CP ($\epsilon_{CP} = 1$) with non-DP Random Forest model. The numbers are metric averages over 1000 runs (per method and α value) and in parentheses is the corresponding variance of the metrics.

B.2 Simulation Results With DP Random Forest Model

The parameter settings are as in the DP NB model and the model's accuracy is 79%.

B.2.1 Effect of ϵ_{CP} with DP model

$\epsilon_{\text{conformal}}$	Coverage		Efficiency		Informativeness	
	ExponQ	**P-COQS**	**ExponQ**	**P-COQS**	**ExponQ**	**P-COQS**
0.1	1.0000 (0.0002)	0.8982 (0.0112)	1.9970 (0.0087)	1.2552 (0.0438)	0.0030 (0.0087)	0.7448 (0.0438)
0.5	0.9575 (0.0272)	0.8977 (0.0107)	1.5621 (0.2378)	1.2535 (0.0438)	0.4379 (0.2378)	0.7465 (0.0438)
1	0.9242 (0.0123)	0.8977 (0.0106)	1.3459 (0.0555)	1.2533 (0.0438)	0.6541 (0.0555)	0.7467 (0.0438)
1.5	0.9171 (0.0105)	0.8976 (0.0106)	1.3183 (0.0465)	1.2532 (0.0439)	0.6817 (0.0465)	0.7468 (0.0439)
3	0.9096 (0.0112)	0.8977 (0.0106)	1.2921 (0.0471)	1.2534 (0.0442)	0.7079 (0.0471)	0.7466 (0.0442)
5	0.9058 (0.0121)	0.8976 (0.0105)	1.2801 (0.0505)	1.2531 (0.0440)	0.7199 (0.0505)	0.7469 (0.0440)
10	0.9034 (0.0129)	0.8975 (0.0104)	1.2726 (0.0526)	1.2530 (0.0438)	0.7274 (0.0526)	0.7470 (0.0438)

Table B.4: Average effect of CP privacy budget ϵ_{CP} on conformal prediction with DP Random Forest model and $\epsilon_f = 2$. The numbers are metric averages over 1000 runs (per method and privacy budget) and in parentheses is the corresponding variance of the metrics.

B.2.2 Sample size effect with DP model

Sample Size	**Coverage**		**Efficiency**		**Informativeness**		**Model Accuracy**
	ExponQ	**P-COQS**	**ExponQ**	**P-COQS**	**ExponQ**	**P-COQS**	
100	0.9979 (0.0142)	0.9669 (0.0800)	1.9977 (0.0146)	1.9534 (0.1166)	0.0022 (0.0146)	0.0466 (0.1166)	0.4201 (0.1141)
200	0.9993 (0.0060)	0.9597 (0.0653)	1.9990 (0.0078)	1.9306 (0.1092)	0.0010 (0.0078)	0.0694 (0.1092)	0.4503 (0.0975)
500	0.9998 (0.0019)	0.9098 (0.0484)	1.9983 (0.0094)	1.6637 (0.1346)	0.0017 (0.0094)	0.3364 (0.1346)	0.6183 (0.0776)
1000	0.9999 (0.0011)	0.9060 (0.0321)	1.9977 (0.0110)	1.5422 (0.1067)	0.0023 (0.0110)	0.4578 (0.1067)	0.6808 (0.0604)
2000	0.9978 (0.0065)	0.9066 (0.0226)	1.9678 (0.0756)	1.4203 (0.0800)	0.0322 (0.0756)	0.5797 (0.0800)	0.7371 (0.0455)
6000	0.9467 (0.0258)	0.9009 (0.0133)	1.5064 (0.1986)	1.2913 (0.0519)	0.4936 (0.1986)	0.7087 (0.0519)	0.7836 (0.0239)
10000	0.9242 (0.0123)	0.8977 (0.0106)	1.3459 (0.0555)	1.2533 (0.0438)	0.6541 (0.0555)	0.7467 (0.0438)	0.7941 (0.0173)

Table B.5: Average effect of sample size on DP CP ($\epsilon_{CP} = 1$) with DP Random Forest model ($\epsilon_f = 2$). The numbers are metric averages over 1000 runs (per method and sample size) and in parentheses is the corresponding variance of the metrics.

B.2.3 Effect of α with DP model

α	Coverage		Efficiency		Informativeness	
	ExponQ	**P-COQS**	**ExponQ**	**P-COQS**	**ExponQ**	**P-COQS**
0.01	1.0000 (0.0002)	0.9904 (0.0036)	1.9970 (0.0087)	1.7604 (0.0422)	0.0030 (0.0087)	0.2396 (0.0422)
0.05	0.9914 (0.0121)	0.9484 (0.0077)	1.8486 (0.1751)	1.4475 (0.0435)	0.1514 (0.1751)	0.5525 (0.0435)
0.10	0.9242 (0.0123)	0.8977 (0.0106)	1.3459 (0.0555)	1.2533 (0.0438)	0.6541 (0.0555)	0.7467 (0.0438)

Table B.6: Average effect of α on DP CP ($\epsilon_{CP} = 1$) with DP Random Forest model ($\epsilon_f = 2$). The numbers are metric averages over 1000 runs (per method and α value) and in parentheses is the corresponding variance of the metrics.

B.2.4 Average CP computational time with DP and non-DP models

Description	Empirical coverage	Efficiency	Informativeness	Model Accuracy	Ave. time (secs)
$\hat{f}$ and EXPONQ	0.9229 (0.0120)	1.2939 (0.0406)	0.7061 (0.0406)	0.8125 (0.0093)	0.7353 (0.0098)
$\hat{f}$ and P-COQS	0.9025 (0.0098)	1.2224 (0.0222)	0.7776 (0.0222)	0.8125 (0.0093)	0.0074 (0.0002)
$\hat{f}^{DP}$ and EXPONQ	0.9242 (0.0123)	1.3459 (0.0555)	0.6541 (0.0555)	0.7941 (0.0173)	0.7390 (0.0088)
$\hat{f}^{DP}$ and P-COQS	0.8977 (0.0106)	1.2533 (0.0443)	0.7467 (0.0443)	0.7941 (0.0173)	0.0074 (0.0002)

Table B.7: Average computational time of P-COQS and EXPONQ combined with non-DP and DP versions of the Random Forest classifier. The numbers are metric averages over 1000 runs (per setting) and in parentheses is the corresponding variance of the metrics.

B.2.5 Effect of Model Privacy Budget (ϵ_f)

ϵ_f	Coverage		Efficiency		Informativeness		Model Accuracy
	ExponQ	**P-COQS**	**ExponQ**	**P-COQS**	**ExponQ**	**P-COQS**	
0.1	0.9256 (0.0142)	0.9021 (0.0167)	1.5647 (0.0841)	1.4891 (0.0723)	0.4353 (0.0841)	0.5109 (0.0723)	0.6870 (0.0591)
0.5	0.9248 (0.0126)	0.8999 (0.0112)	1.3839 (0.0674)	1.2956 (0.0530)	0.6161 (0.0674)	0.7044 (0.0530)	0.7803 (0.0242)
1	0.9241 (0.0127)	0.8982 (0.0104)	1.3575 (0.0611)	1.2664 (0.0454)	0.6425 (0.0611)	0.7336 (0.0454)	0.7898 (0.0187)
2	0.9242 (0.0123)	0.8977 (0.0106)	1.3459 (0.0555)	1.2533 (0.0438)	0.6541 (0.0555)	0.7467 (0.0438)	0.7941 (0.0173)
5	0.9241 (0.0123)	0.8973 (0.0104)	1.3405 (0.0546)	1.2481 (0.0407)	0.6595 (0.0546)	0.7519 (0.0407)	0.7957 (0.0167)
10	0.9241 (0.0122)	0.8972 (0.0104)	1.3405 (0.0542)	1.2479 (0.0409)	0.6595 (0.0542)	0.7521 (0.0409)	0.7958 (0.0168)

Table B.8: Average effect of model privacy budget ϵ_f on DP CP ($\epsilon_{CP} = 1$) with DP Random Forest model. The numbers are metric averages over 1000 runs (per method and privacy budget) and in parentheses is the corresponding variance of the metrics.

B.3 Simulation Results With Non-DP Naïve Bayes Model

The parameter settings are as described in Section 2.5.1 and the accuracy of the model is 83%.

B.3.1 Effect of Privacy Budget (ϵ_{CP})

ϵ_{CP}	Coverage		Efficiency		Informativeness	
	ExponQ	**P-COQS**	**ExponQ**	**P-COQS**	**ExponQ**	**P-COQS**
0.1	0.9999 (0.0004)	0.9005 (0.0104)	1.9678 (0.0289)	1.1787 (0.0223)	0.0322 (0.0289)	0.8213 (0.0223)
0.5	0.9444 (0.0174)	0.9005 (0.0101)	1.3465 (0.0972)	1.1788 (0.0206)	0.6535 (0.0972)	0.8212 (0.0206)
1	0.9227 (0.0116)	0.9006 (0.0099)	1.2509 (0.0360)	1.1788 (0.0201)	0.7491 (0.0360)	0.8212 (0.0201)
1.5	0.9154 (0.0103)	0.9006 (0.0099)	1.2254 (0.0267)	1.1788 (0.0201)	0.7746 (0.0267)	0.8212 (0.0201)
3	0.9081 (0.0097)	0.9006 (0.0099)	1.2019 (0.0213)	1.1788 (0.0200)	0.7981 (0.0213)	0.8212 (0.0200)
5	0.9051 (0.0097)	0.9006 (0.0099)	1.1925 (0.0203)	1.1789 (0.0200)	0.8075 (0.0203)	0.8211 (0.0200)
10	0.9029 (0.0097)	0.9006 (0.0099)	1.1857 (0.0198)	1.1789 (0.0201)	0.8143 (0.0198)	0.8211 (0.0201)

Table B.9: Average effect of CP privacy budget ϵ_{CP} on conformal prediction with non-DP Naïve Bayes model. The numbers are metric averages over 1000 runs (per method and privacy budget) and in parentheses is the corresponding variance of the metrics.

B.3.2 Sample size effect

Samples	Coverage		Efficiency		Informativeness	
	ExponQ	**P-COQS**	**ExponQ**	**P-COQS**	**ExponQ**	**P-COQS**
100	0.9794 (0.0434)	0.9040 (0.1004)	1.7300 (0.2179)	1.3677 (0.2713)	0.2700 (0.2179)	0.6323 (0.2713)
200	0.9905 (0.0207)	0.9206 (0.0695)	1.7901 (0.1638)	1.3394 (0.2004)	0.2099 (0.1638)	0.6606 (0.2004)
500	0.9967 (0.0078)	0.9053 (0.0440)	1.8566 (0.1158)	1.2222 (0.1071)	0.1434 (0.1158)	0.7778 (0.1071)
1000	0.9984 (0.0042)	0.9018 (0.0315)	1.8951 (0.0876)	1.1954 (0.0669)	0.1049 (0.0876)	0.8046 (0.0669)
2000	0.9836 (0.0129)	0.9030 (0.0212)	1.6272 (0.1483)	1.1891 (0.0466)	0.3728 (0.1483)	0.8109 (0.0466)
6000	0.9382 (0.0169)	0.9008 (0.0123)	1.3149 (0.0795)	1.1798 (0.0268)	0.6851 (0.0795)	0.8202 (0.0268)
10000	0.9227 (0.0116)	0.9006 (0.0099)	1.2509 (0.0360)	1.1788 (0.0200)	0.7491 (0.0360)	0.8212 (0.0200)

Table B.10: Average effect of sample size on DP CP ($\epsilon_{CP} = 1$) with non-DP Naïve Bayes model. The numbers are metric averages over 1000 runs (per method and sample size) and in parentheses is the corresponding variance of the metrics. Average effect of sample size on private CP with Non-DP Random Forest model.

B.3.3 Effect of α

α	Coverage		Efficiency		Informativeness	
	ExponQ	**P-COQS**	**ExponQ**	**P-COQS**	**ExponQ**	**P-COQS**
0.01	0.9999 (0.0004)	0.9901 (0.0033)	1.9677 (0.0289)	1.6703 (0.0349)	0.0323 (0.0289)	0.3297 (0.0349)
0.05	0.9785 (0.0115)	0.9500 (0.0073)	1.5670 (0.1306)	1.3610 (0.0244)	0.4330 (0.1306)	0.6390 (0.0244)
0.10	0.9227 (0.0116)	0.9006 (0.0099)	1.2509 (0.0360)	1.1788 (0.0200)	0.7491 (0.0360)	0.8212 (0.0200)

Table B.11: Average effect of α on DP CP ($\epsilon_{CP} = 1$) with non-DP Naïve Bayes model ($\epsilon_f = 2$). The numbers are metric averages over 1000 runs (per method and α value) and in parentheses is the corresponding variance of the metrics.

B.3.4 Effect of Model Privacy Budget (ϵ_f)

ϵ_f	Coverage		Efficiency		Informativeness		Model Accuracy
	ExponQ	**P-COQS**	**ExponQ**	**P-COQS**	**ExponQ**	**P-COQS**	
0.1	0.9218 (0.0115)	0.8994 (0.0101)	1.6580 (0.0335)	1.5905 (0.0233)	0.3420 (0.0335)	0.4095 (0.0233)	0.6319 (0.0139)
0.5	0.9222 (0.0110)	0.9000 (0.0094)	1.5361 (0.0389)	1.4532 (0.0263)	0.4639 (0.0389)	0.5468 (0.0263)	0.7233 (0.0110)
1	0.9405 (0.0366)	0.9812 (0.0409)	1.7647 (0.1443)	1.9207 (0.1718)	0.2353 (0.1443)	0.0793 (0.1718)	0.6559 (0.0303)
2	0.9217 (0.0116)	0.8999 (0.0098)	1.4841 (0.0457)	1.3985 (0.0322)	0.5159 (0.0457)	0.6015 (0.0322)	0.7479 (0.0125)
5	0.9224 (0.0117)	0.9000 (0.0098)	1.2784 (0.0372)	1.2036 (0.0216)	0.7216 (0.0372)	0.7964 (0.0216)	0.8156 (0.0095)
10	0.9225 (0.0116)	0.9002 (0.0098)	1.2570 (0.0361)	1.1840 (0.0204)	0.7430 (0.0361)	0.8160 (0.0204)	0.8230 (0.0093)

Table B.12: Average effect of model privacy budget ϵ_f on conformal prediction with DP Naïve Bayes model. The numbers are metric averages over 1000 runs (per method and privacy budget) and in parentheses is the corresponding variance of the metrics.