Debiased Parametric Bootstrap Inference on Privatized Data

Zhanyu Wang
Purdue University
wang4094@purdue.edu

Jordan Awan
Purdue University
jawan@purdue.edu

ABSTRACT

In this work, we design a debiased parametric bootstrap framework for statistical inference from differentially private data. Existing usage of the parametric bootstrap on privatized data ignored or avoided handling the effect of clamping, a technique employed by the majority of privacy mechanisms. We show that ignoring the impact of clamping often leads to under-coverage of confidence intervals and miscalibrated type I errors of hypothesis tests. The main reason for the failure of the existing methods is the inconsistency of the parameter estimate based on the privatized data. We propose using the indirect inference method to estimate the parameter values consistently and use the improved estimator for parametric bootstrap inference. To implement the indirect estimator, we present a novel simulation-based approach along with the theory establishing the consistency of the corresponding parametric bootstrap distribution. Our simulation studies show that our framework produces confidence intervals with the correct coverage and performs hypothesis testing with the correct type I error in finite sample settings.

KEYWORDS

differential privacy, confidence intervals, hypothesis tests, simulation-based inference, asymptotic statistics

1 INTRODUCTION

In the age of big data, utilizing diverse data sources to train models offers significant benefits but also leaves the data providers vulnerable to malicious attacks. To mitigate privacy concerns, Dwork et al. [9] introduced the concept of Differential Privacy (DP), which quantifies the level of privacy assurance offered by a specific data processing procedure. Following the advent of DP, numerous mechanisms have been developed to provide DP-guaranteed point estimates for parameters [10]. These mechanisms inject additional uncertainty into the output to manage the tradeoff between its utility and privacy guarantee.

Statistical inference not only aims to provide an accurate point estimator for a population parameter of interest but also attempts to quantify the inherent uncertainty of such an estimate due to the randomness of population sampling. The sampling distributions of the statistics based on random samples are commonly utilized in inference, such as in constructing confidence intervals (CI) or performing hypothesis tests (HT). As previously discussed, the output of a DP mechanism incorporates extra noise, resulting in the sampling distribution of DP outputs differing from that of non-private outputs. To accommodate this difference, several methods have been developed, supported by theoretical analysis in specific settings, to perform inference based on DP estimates. Our paper assumes a data-generating model and accepts any DP mechanism. We propose using this model to obtain an indirect estimator used in parametric bootstrap (PB) for statistical inference under the DP guarantee. The work most relevant to our study is that of Awan and Wang [2], who also employ a data-generating model in a simulation-based methodology, repro samples method [17], to generate finite-sample valid CIs and HTs from DP summary statistics. However, their approach suffers from over-conservativeness and extensive computation.

Using PB in DP statistical inference is not a novel approach. Du et al. [8] proposed various methods to construct DP confidence intervals, differing in parameter estimation techniques, and they all use PB to derive confidence intervals through simulation. Ferrando et al. [12] were the first to theoretically analyze the use of PB with DP guarantees. They validated the consistency of their confidence intervals in two private estimation settings: exponential families and linear regression via sufficient statistic perturbation (SSP). Alabi and Vadhan [1] leveraged PB to conduct DP hypothesis testing specifically for linear regression. However, these three methods all assume that clamping does not influence the data distribution, an assertion that is overly strong and can result in inaccurate inferences. To correct the clamping bias, using our novel adaptive indirect estimator, we have developed a debiased parametric bootstrap framework with theoretical guarantees of consistency and demonstrated the effectiveness through simulations.

2 BACKGROUND

For $p \in \mathbb{N}^+$, $x \in \mathbb{R}^p$, $\Omega \in \mathbb{R}^{p\times p}$, let $\|x\|_\Omega := x^\intercal \Omega x$. For $\Omega_1, \Omega_2 \in \mathbb{R}^{p\times p}$, we write $\Omega_1 \succ \Omega_2$ if $\Omega_1 - \Omega_2$ is positive definite, and $\Omega_1 \succeq \Omega_2$ if $\Omega_1 - \Omega_2$ is positive semidefinite. $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2, and $\Phi(\cdot)$ is the cumulative distribution function (CDF) of $N(0,1)$. We write $X_n = O_p(a_n)$ if for any $\epsilon > 0$, there are M and N such that $\mathbb{P}(|X_n/a_n| \geq M) < \epsilon$ for any $n > N$. We denote a dataset by $D := (x_1, \ldots, x_n)$ which has sample size $|D| = n$, and if $x_i \overset{\text{iid}}{\sim} F(x|\theta^*)$, we write $D \sim F^n(x|\theta^*)$.

2.1 Differential privacy

In this paper, we use ε-DP [9] and Gaussian DP (GDP) [7] in our examples although our results also apply to many other DP notions.

A mechanism is a randomized function M that takes a dataset D as input and outputs a random variable or vector S. The Hamming distance between two datasets with the same sample sizes is $d(D, D')$, the number of entries in which D and D' differ.

Definition 2.1. M satisfies ε-DP if for any $d(D, D') \leq 1$ and measurable set S, $P(M(D) \in S) \leq \exp(\varepsilon) P(M(D') \in S)$.

Definition 2.2. M satisfies μ-GDP if for any $d(D, D') \leq 1$, any hypothesis test between $H_0 : S \sim M(D)$ and $H_1 : S \sim M(D')$ has a type II error β bounded below by $\Phi(\Phi^{-1}(1-\alpha) - \mu)$ where α is the type I error.

Definition 2.2 means that it is harder to distinguish $M(D)$ from $M(D')$ than to distinguish $N(0,1)$ from $N(\mu, 1)$ if M is μ-GDP and $d(D, D') \leq 1$.

2.2 Parametric bootstrap

Definition 2.3. Let $D \sim F^n(x|\theta^*)$ where the true unknown parameter is θ^*. Given $\hat{\theta}(D)$ and $\hat{\tau}(D)$ as estimates of θ^* and $\tau^* := \tau(\theta^*)$ respectively, where τ^* is the parameter of interest, the parametric bootstrap estimator of τ^* is defined by $\hat{\tau}(D^b)$ where $D^b \sim F^n(x|\hat{\theta}(D))$.

We use the empirical CDF of $\{\hat{\tau}(D^b)\}_{b=1}^B$ to approximate the sampling distribution of $\hat{\tau}(D)$, and construct CIs or perform HTs for τ^*. If it is the case that $\hat{\theta}(D) - \theta^* = O_p(\frac{1}{\sqrt{n}})$, Beran [4] showed that the asymptotic equivariance of $\hat{\tau}$ guarantees the *consistency of PB*, which implies that the coverage of CIs and the type I error of HTs are asymptotically consistent with the nominal levels.

Definition 2.4 (Asymptotic equivariance [4])**.** Let $H_n(\theta^*)$ be the distribution of $\sqrt{n}(\hat{\tau}(D) - \tau(\theta^*))$ where $D \sim F^n(x|\theta^*)$. $\hat{\tau}$ is asymptotically equivariant if $H_n(\theta^* + h_n/\sqrt{n})$ converges to a limiting distribution $H(\theta^*)$ for all convergent sequences $h_n \to h$ and all θ^*.

Definition 2.5 (Bootstrap consistency [15])**.** Let $D \sim F^n(x|\theta^*)$ and $D^b \sim F^n(x|\hat{\theta}(D))$. $H_n(\theta^*)$ is the distribution of $\sqrt{n}(\hat{\tau}(D) - \tau(\theta^*))$, and $H_n(\hat{\theta}(D))$ is the random distribution of $\sqrt{n}(\hat{\tau}(D^b) - \hat{\tau}(D))$ depending on D. $\hat{\tau}(D^b)$ is consistent if $H_n(\hat{\theta}(D)) \xrightarrow{P} H_n(\theta^*)$.

Proposition 2.6 (Parametric bootstrap consistency [4, 12])**.** *Suppose* $\sqrt{n}(\hat{\theta}(D) - \theta^*) \xrightarrow{d} J(\theta^*)$ *and* $\hat{\tau}$ *is asymptotically equivariant with continuous* $H(\theta^*)$*. Then the parametric bootstrap estimator* $\hat{\tau}(D^b)$ *is consistent.*

2.3 Inaccurate inference due to clamping

As PB only requires a point estimator, $\hat{\theta}(D)$, for obtaining D^b and $\hat{\tau}(D^b)$, the privacy guarantee for the private statistical inference is the same as the DP point estimator because of the post-processing property [10]. However, DP estimators are often biased, which leads to inaccurate inference results from the naïve use of PB.

We use two examples in the existing literature to demonstrate the inaccurate inference results in the PB methods using estimators that are biased due to clamping. The first example is constructing a CI for the population mean of a normal distribution [16], where the results in Table 1 demonstrate the under-coverage problem of using NOISYVAR+SIM [8] to construct a DP CI. The second example is conducting a HT for one coefficient in linear regression [2] where the results in Figure 1 show the mis-calibrated type I error of using the framework of Alabi and Vadhan [1] which is based on PB.

In Section 5, we use these two settings to show the advantage of using our proposed method.

Privacy guarantee	1-GDP	0.5-GDP	0.3-GDP	0.1-GDP
Coverage	0.803	0.806	0.804	0.819

Table 1: Coverage of private CIs with nominal confidence level 0.9 for the population mean of $N(0.5, 1)$ by NOISYVAR+SIM [8]. Results are from [16, Table 2].

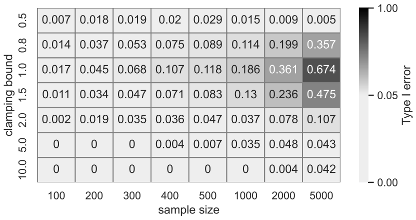

Figure 1: Type I error of private HTs on $H_0 : \beta_1^* = 0$ and $H_1 : \beta_1^* \neq 0$ with a linear regression model $Y = \beta_0^* + X\beta_1^* + \epsilon$ using DP Monte Carlo tests [1] with nominal significance level 0.05. Results are from [2, Figure 5].

3 DEBIASED PARAMETER ESTIMATION

In this section, we describe the indirect estimator [13], which can solve the bias issue in the clamping procedure of DP mechanisms. We also propose an adaptive indirect inference estimator, that automatically optimizes the covariance matrix of the indirect estimator.

The underlying principle of the indirect estimator is to fix the "random seeds" for synthetic data generation and find the parameter that makes the synthetic data most similar to the observed statistic. We describe the indirect estimator with additional consideration of the DP mechanisms used in releasing the observed statistic.

Let the true model be $Z = g(\theta^*; U)$ where $\theta^* \in \Theta \subseteq \mathbb{R}^q$ is the unknown parameter, $U \sim F_u$ is the source of uncertainty that is unobserved and following a known distribution F_u, and g is a deterministic function generating our observation Z. Note that F_u does not depend on θ. An example is $Z \sim N(\mu, \sigma^2)$ which can be represented as $Z = \mu + \sigma U$ where $\theta^* = (\mu, \sigma)$ and $U \sim F_u := N(0,1)$. For simplicity, we denote $\mathbf{z}_n := (z_1, \ldots, z_n)$, $\mathbf{u}_n := (u_1, \ldots, u_n)$; We write $\mathbf{z}_n = g(\theta^*; \mathbf{u}_n)$ if $z_i = g(\theta^*; u_i)$, and we write $\mathbf{u}_n \sim F_u^n$ if $u_i \overset{\text{iid}}{\sim} F_u$ for $i = 1, \ldots, n$. Let $\hat{\beta}_n \in \mathbb{B} \subseteq \mathbb{R}^p$ be a statistic calculated from $\mathbf{z}_n$ by maximizing a criterion, i.e., $\hat{\beta}_n := \arg\max_{\beta \in \mathbb{B}} Q_n(\beta, \mathbf{z}_n, \mathbf{u}_{\text{DP}})$ where $\mathbf{u}_{\text{DP}} \sim F_{\text{DP}}$ denotes the source of extra uncertainty introduced by DP mechanisms. While $\hat{\beta}_n$ is often an estimator for θ^*, it could also be a general set of summary statistics informative for θ^*. The optimization-form definition of $\hat{\beta}_n$ is useful for DP mechanisms such as objective perturbation [5]; It is also compatible with DP mechanisms like $\hat{\beta}_n := q_n(\mathbf{z}_n, \mathbf{u}_{\text{DP}})$ since it is equivalent to $\hat{\beta}_n := \arg\max_{\beta \in \mathbb{B}} Q_n(\beta, \mathbf{z}_n, \mathbf{u}_{\text{DP}}) := -\frac{1}{2}\|\beta - q_n(\mathbf{z}_n, \mathbf{u}_{\text{DP}})\|_2^2$.

As F_u and F_{DP} are known, for each θ, we simulate $\tilde{\beta}_n(\theta, \mathbf{u}_n^h, \mathbf{u}_{\text{DP}}^h) := \arg\max_{\beta \in \mathbb{B}} Q_n(\beta, \mathbf{z}_n^h(\theta), \mathbf{u}_{\text{DP}}^h)$ from synthetic data $\mathbf{z}_n^h(\theta) := g(\theta; \mathbf{u}_n^h)$ generated with $\mathbf{u}_n^h \overset{\text{iid}}{\sim} F_u^n$ and $\mathbf{u}_{\text{DP}}^h \overset{\text{iid}}{\sim} F_{\text{DP}}$, $h = 1, \ldots, H$. Let $\mathbf{u}_n^{[H]} := (\mathbf{u}_n^1, \ldots, \mathbf{u}_n^H)$, $\mathbf{u}_{\text{DP}}^{[H]} := (\mathbf{u}_{\text{DP}}^1, \cdots, \mathbf{u}_{\text{DP}}^H)$, and $\Omega_n > 0$.

Definition 3.1 (Indirect estimator [13])**.** The indirect estimator is

$$\hat{\theta}_{\text{IND}} := \underset{\theta\in\Theta}{\arg\min} \left\| \hat{\beta}_n - \frac{1}{H}\sum_{h=1}^{H} \tilde{\beta}_n\left(\theta, \mathbf{u}_n^h, \mathbf{u}_{\text{DP}}^h\right) \right\|_{\Omega_n}.$$

We write $\tilde{\beta}_n\left(\theta, \mathbf{u}_n^h, \mathbf{u}_{\text{DP}}^h\right)$ as $\tilde{\beta}_n^h(\theta)$ if it is clear from context.

In terms of the asymptotic variance of $\hat{\theta}_{\text{IND}}$, there are some choices of Ω_n better than others. However, the optimal Ω_n may depend on θ^* and require additional effort to find a good estimation. For a novel and computationally efficient approach, we propose to use the inverse of the sample covariance matrix of $\tilde{\beta}_n^h$ as an adaptive Ω_n and show its asymptotic optimality in Remark 4.8.

Definition 3.2 (Adaptive indirect estimator)**.** Let $m_\beta^H(\theta)$ and $S_\beta^H(\theta)$ be the sample mean and covariance matrix of $\{\tilde{\beta}_n(\theta, \mathbf{u}_n^h, \mathbf{u}_{\text{DP}}^h)\}_{h=1}^H$. The adaptive indirect estimator of θ^* is

$$\hat{\theta}_{\text{ADI}} := \underset{\theta\in\Theta}{\arg\min} \left\| \hat{\beta}_n - m_\beta^H(\theta) \right\|_{\left(S_\beta^H(\theta)\right)^{-1}}.$$

4 THEORY

In this section, we provide the theoretical guarantees for using the indirect estimators in parametric bootstrap.

To prove the consistency and asymptotic equivariance of $\hat{\theta}_{\text{IND}}$, we have the following assumptions.

(A1) $\sup_{\beta\in\mathbb{B}} |Q_n(\beta, \mathbf{z}_n, \mathbf{u}_{\text{DP}}) - Q_\infty(\beta, F_u, F_{\text{DP}}, \theta^*)| \xrightarrow{\text{P}} 0$ *where* $Q_\infty(\beta, F_u, F_{\text{DP}}, \theta^*)$ *is non-stochastic and continuous in* β*.*

(A2) Θ *and* $\mathbb{B}$ *are compact. Let* $b(\theta) := \arg\max_{\beta\in\mathbb{B}} Q_\infty(\beta, F_u, F_{\text{DP}}, \theta)$ *which is continuous and injective (one-to-one), and* $\beta^* := b(\theta^*)$*. The Jacobian matrix* $\frac{\partial b}{\partial\theta^\intercal}$ *exists and is full-column rank.*

(A3) $\sup_{\theta\in\Theta} \|\tilde{\beta}_n(\theta, \mathbf{u}_n^h, \mathbf{u}_{\text{DP}}^h) - b(\theta)\| \xrightarrow{\text{P}} 0.$

(A4) $\Omega_n \succ 0$ *are deterministic matrices converging to* $\Omega \succ 0$*.*

(A5) *For every* $(\theta, \mathbf{u}_n^h, \mathbf{u}_{\text{DP}}^h)$, $\frac{\partial\tilde{\beta}_n(\theta,\mathbf{u}_n^h,\mathbf{u}_{\text{DP}}^h)}{\partial\theta}$ *and* $\frac{\partial b(\theta)}{\partial\theta}$ *exist and are continuous in* θ; $\frac{\partial\tilde{\beta}_n(\theta,\mathbf{u}_n^h,\mathbf{u}_{\text{DP}}^h)}{\partial\theta} \xrightarrow{\text{P}} \frac{\partial b(\theta)}{\partial\theta}$. $B^* := \frac{\partial b(\theta^*)}{\partial\theta}$.

(A6) *For every* $(\beta, \mathbf{z}_n, \mathbf{u}_{\text{DP}})$, $\frac{\partial Q_n(\beta,\mathbf{z}_n,\mathbf{u}_{\text{DP}})}{\partial\beta}$ *and* $\frac{\partial^2 Q_n(\beta,\mathbf{z}_n,\mathbf{u}_{\text{DP}})}{(\partial\beta)(\partial\beta^\intercal)}$ *exist and are continuous in* β; $\sqrt{n}(\frac{\partial Q_n(\beta^*,\mathbf{z}_n^h(\theta^*),\mathbf{u}_{\text{DP}}^h)}{\partial\beta}) \xrightarrow{\text{d}} F_{Q,u,\text{DP}}^*$, $J^* := -\frac{\partial^2 Q_\infty(\beta^*,F_u,F_{\text{DP}},\theta^*)}{(\partial\beta)(\partial\beta^\intercal)} \xleftarrow{\text{P}} -\frac{\partial^2 Q_n(\beta^*,\mathbf{z}_n^h(\theta^*),\mathbf{u}_{\text{DP}}^h)}{(\partial\beta)(\partial\beta^\intercal)}$.

Remark 4.1. (A1, A3) are for the consistency of $\hat{\beta}_n$ and $\hat{\theta}_{\text{IND}}$ as they are M-estimators [15]. (A2) is for the identifiability of θ^* using $\hat{\beta}_n$. (A4) generalizes ℓ_2 norm for more efficient $\hat{\theta}$. (A5, A6) are for the Taylor expansion to obtain asymptotical distributions of $\hat{\beta}_n$ and $\hat{\theta}$ which requires us to have the true model $Z = g(\theta^*; U)$ continuous in θ^* given any U. Note that for Z following a discrete distribution such as a Binomial or Poisson distribution, we can transform Z to its continuous counterparts [14] as an approximation to the true data generating process.

Theorem 4.2. *Under (A1, A2, A3, A4, A5, A6),*
1) $\hat{\theta}_{\text{IND}}$ *is a consistent estimator of* θ^**, and* $\sqrt{n}(\hat{\theta}_{\text{IND}} - \theta^*)$ *converges in distribution to* $((B^*)^\intercal\Omega B^*)^{-1}(B^*)^\intercal\Omega(J^*)^{-1}(v_0 - \frac{1}{H}\sum_{h=1}^H v_h)$ *where* $v_h \overset{iid}{\sim} F_{Q,u,\text{DP}}^*$ *for* $h = 0, 1, \ldots, H$*;*
2) $\hat{\theta}_{\text{IND}}$ *is asymptotically equivariant which implies that the parametric bootstrap based on* $\hat{\theta}_{\text{IND}}$ *is consistent.*

Remark 4.3. We primarily use the asymptotic results of Theorem 4.2 to establish the consistency of the PB. While the asymptotic distribution could itself be used for statistical inference [13], it would require estimating $(B^*, J^*, F_{Q,u,\text{DP}}^*)$, and its finite-sample performance may be unsatisfactory. However, the asymptotic distribution is still helpful in constructing an approximate pivot for more efficient statistical inference [3, Table 1] as illustrated in Section 5.2.

Remark 4.4. The first part of Theorem 4.2 on the consistency and asymptotic distribution is inspired by [13, Propositions 1 and 3] while we give a more detailed and precise proof and focus on its application with DP. We generalize the asymptotic distribution of $\sqrt{n}(\frac{\partial Q_n}{\partial\beta}(\beta^*, \mathbf{z}_n^h(\theta^*), \mathbf{u}_{\text{DP}}^h))$ from normal distribution [13] to $F_{Q,u,\text{DP}}^*$. In Example 4.1, we show the necessity of such a generalization in DP settings.

Example 4.1. For $\varepsilon = \frac{1}{\sqrt{n}}$ and $X = (x_1, \ldots, x_n)$ where $x_i \in [0, 1]$, in order to estimate the population mean corresponding to X under ε-DP, we use the Laplace mechanism which releases $q_n := \frac{1}{n}\sum_{i=1}^n x_i + \mathbf{u}_{\text{DP}}$, $\mathbf{u}_{\text{DP}} \sim \text{Laplace}(\frac{1}{n\varepsilon})$. Let $Q_n(\beta) := -\frac{1}{2}\|\beta - q_n\|_2^2$, which indicates $Q_\infty(\beta) = -\frac{1}{2}\|\beta - \beta^*\|_2^2$ where $\beta^* = \mathbb{E}[x_i]$. If $x_i \overset{\text{iid}}{\sim} \text{Uniform}([0, 1])$, we have $\sqrt{n}(\frac{\partial Q_n(\beta^*)}{\partial\beta}) = \sqrt{n}(q_n - \beta^*)$, then $F_{Q,u,\text{DP}}^*$ is not normal but a convolution of $N(0, \frac{1}{12})$ and $\text{Laplace}(1)$.

Remark 4.5. If there is θ such that $\hat{\beta}_n - \frac{1}{H}\sum_{h=1}^H \tilde{\beta}_n^h(\theta) = 0$, the indirect estimator is equal to the Just Identified Indirect Inference estimator [18] which has been shown to enjoy nice properties, including consistency, asymptotic normality, and finite sample bias correction, better than the Bootstrap Bias Corrected estimator [18].

To prove the validity of using the adaptive indirect estimator, $\hat{\theta}_{\text{ADI}}$, in parametric bootstrap, we need additional assumptions.

(A7) *For any* θ *and* $\{H_n\}_{n=1}^\infty$ *where* $\lim_{n\to\infty} H_n = \infty$*, we have* $nS_\beta^{H_n}(\theta) \xrightarrow{\text{P}} \lim_{n\to\infty} \text{Var}(\sqrt{n}(\tilde{\beta}_n^h(\theta) - b(\theta))) = \text{Var}(\lim_{n\to\infty} \sqrt{n}(\tilde{\beta}_n^h(\theta) - b(\theta))) =: \Sigma(\theta)$ *is continuous in* θ, $\Sigma(\theta^*) \succ 0$*, and* $\frac{\partial(nS_\beta^{H_n}(\theta))}{\partial\theta} = O_p(1)$.

(A8) *For* $v \sim F_{Q,u,\text{DP}}^*$*, we have* $\mathbb{E}[v]$ *exists and is finite.*

Remark 4.6. (A7) is for the consistency of $nS_\beta^{H_n}(\theta)$ and $\hat{\theta}_{\text{ADI}}$. The choice of $\{H_n\}_{n=1}^\infty$ and (A8) are for the law of large numbers.

Theorem 4.7. *Under (A1, A2, A3, A5, A6, A7, A8), for* $H := H_n$*,*
1) $\hat{\theta}_{\text{ADI}}$ *is a consistent estimator of* θ^**, and* $\sqrt{n}(\hat{\theta}_{\text{ADI}} - \theta^*)$ *converges in distribution to* $((B^*)^\intercal\Omega^* B^*)^{-1}(B^*)^\intercal\Omega^*(J^*)^{-1}(v - \mathbb{E}(v))$ *where* $v \sim F_{Q,u,\text{DP}}^*$ *and* $\Omega^* = \Sigma(\theta^*)^{-1} = \text{Var}[(J^*)^{-1}v]^{-1}$*;*
2) $\hat{\theta}_{\text{ADI}}$ *is asymptotically equivariant; therefore, the parametric bootstrap based on* $\hat{\theta}_{\text{ADI}}$ *is consistent;*
3) $\text{Var}\left(\lim_{n\to\infty}\sqrt{n}(\hat{\theta}_{\text{IND}} - \theta^*)\right) \geq \text{Var}\left(\lim_{n\to\infty}\sqrt{n}(\hat{\theta}_{\text{ADI}} - \theta^*)\right)$ *for any choice of* $\{\Omega_n\}_{n=1}^\infty$ *in* $\hat{\theta}_{\text{IND}}$*.*

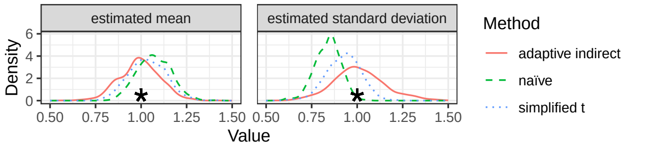

Figure 2: Comparison of the bias in different estimates. The estimator in [12] is the same as the simplified t estimator.

Remark 4.8. The last part of Theorem 4.7 shows that the adaptive indirect estimator is asymptotically the minimum variance estimator among the indirect estimators based on Z.

5 SIMULATION

In this section, we use simulations on DP statistical inference to demonstrate the performance of our debiased estimator used in PB. We construct CIs for the population mean and variance of normal distributions, and we conduct HTs with a linear regression model. All results are computed over 1000 replicates.

5.1 CI for parameters of a normal distribution

Consider $x_1, \ldots, x_n \overset{\text{iid}}{\sim} N(\mu^*, (\sigma^*)^2)$ where the true parameters are $\mu^* = 1$ and $\sigma^* = 1$. Given $n = 100$ and a dataset $D := (x_1, \ldots, x_n)$, we use the Gaussian mechanism to release two statistics for the inference of μ^* and σ^*: Let $x_i]_a^b = \max(a, \min(b, x_i))$, $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i]_a^b$, and $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i]_a^b - \hat{\mu})^2$; We release $\tilde{\mu} = \hat{\mu} + \frac{(b-a)\eta_1}{n}$ and $\tilde{\sigma}^2 = \hat{\sigma}^2 + \frac{(b-a)^2\eta_2}{n}$ where $\eta_1, \eta_2 \overset{\text{iid}}{\sim} N(0,1)$. Then $(\tilde{\mu}, \tilde{\sigma}^2)$ is $\sqrt{2}$-GDP.

In Table 2, we compare our adaptive indirect estimator with other estimators used in PB to construct CIs with level $1 - \alpha = 0.95$ and we set $a = 0, b = 3$. Let $\hat{\tau}(D) = (\tilde{\mu}, \tilde{\sigma}^2)$, and the PB estimators are $\{\hat{\tau}(D^b)\}_{b=1}^B$ where D^b is generated using $N(\tilde{\mu}, \tilde{\sigma}^2)$. The naïve percentile method uses the $\alpha/2$ and $1 - \alpha/2$ percentile of each dimension of $\{\hat{\tau}(D^b)\}_{b=1}^B$ to construct CIs; The simplified t method uses the $\alpha/2$ and $1 - \alpha/2$ percentile of $\{2\hat{\tau}(D) - \hat{\tau}(D^b)\}_{b=1}^B$ instead. Both results suffer from the under-coverage issue.

For the indirect estimator denoted by $\hat{\theta}(D)$, we use $\hat{\tau}(D)$ as the statistic $\hat{\beta}_n$ for the estimation of $\theta^* := (\mu^*, (\sigma^*)^2)$. The PB estimators are $\{\hat{\theta}(D^b)\}_{b=1}^B$ where D^b is generated using a normal distribution with parameter $\hat{\theta}(D)$. The results of percentile CIs using PB with the indirect estimator have satisfactory coverage.

We also compare our method with Repro [2], which is another simulation-based technique similar to the indirect estimator but does not use PB. The results by Repro are more conservative than ours by PB since they have much higher coverage but also larger average width. The other bias correction methods include Efron's bias-corrected (BC) percentile method [11], the automatic percentile method [6], and the method of Ferrando et al. [12].

Figure 2 illustrates the bias in the estimators. The naïve method is $\hat{\tau}(D)$, the simplified t method is $2\hat{\tau}(D) - \frac{1}{B}\sum_{b=1}^B \hat{\tau}(D^b)$ which is also used by [12], and the indirect estimator is $\hat{\theta}(D)$. The adaptive indirect estimator is the only one not having significant bias.

	Coverage		Average width	
	μ	σ	μ	σ
PB (naïve percentile)	0.697 (0.015)	0.006 (0.002)	0.311 (0.001)	0.293 (0.001)
PB (simplified t)	0.869 (0.011)	0.817 (0.012)	0.311 (0.001)	0.293 (0.001)
PB [12]	0.808 (0.012)	0.371 (0.015)	0.311 (0.001)	0.293 (0.001)
PB (Efron's BC)	0.854 (0.011)	0.042 (0.006)	0.298 (0.001)	0.139 (0.002)
PB (automatic percentile)	0.865 (0.011)	0.126 (0.010)	0.314 (0.001)	0.261 (0.001)
PB **(adaptive indirect)**	0.949 (0.007)	0.931 (0.008)	0.457 (0.003)	0.574 (0.003)
Repro [2]	0.989 (0.003)	0.998 (0.001)	0.599 (0.003)	0.758 (0.005)

Table 2: Results of the 95% CI for the inference of population mean μ and standard deviation σ in a normal distribution.

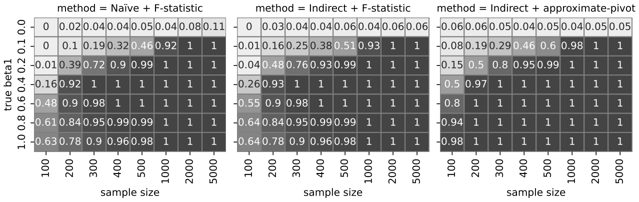

Figure 3: Comparison of the rejection probability on $H_0 : \beta_1^* = 0$ and $H_1 : \beta_1^* \neq 0$ for $Y = \beta_0^* + X\beta_1^* + \epsilon$ (significance level 0.05.)

5.2 HT for parameters in linear regression

We follow the experiment setting in [1]. Let the true model be $Y = \beta_0^* + X\beta_1^* + \epsilon$ where $\beta_0^* = -0.5$, $X \sim N(0.5, 1)$, $\varepsilon \sim N(0, 0.25)$. In Figure 3, we show the rejection probability of different methods that are 1-GDP for HTs on $H_0 : \beta_1^* = 0$ and $H_1 : \beta_1^* \neq 0$ under various settings of the parameter of interest, β_1^*, and sample size, n.

The results in the first subfigure of Figure 3 are from the method in [1] which uses Gaussian mechanism to obtain privatized sufficient statistics for PB and the F-statistic in the HT. There are two problems with the results in this subfigure: 1) The type I error, i.e., the first row in the subfigures, is sometimes larger than the significance level 0.05; 2) The rejection probability is not always larger for larger β_1^*. The first problem is caused by the bias in the naïve estimator from the privatized sufficient statistics, and the second problem may be due to the inefficiency of the F-statistic. For the results in the second subfigure, we replace the naïve estimator with the private adaptive indirect estimator in PB which solves the first problem. Furthermore, we construct an approximate pivot using the adaptive indirect estimator and its asymptotic distribution in Theorem 4.7, and the results are in the third subfigure where both problems are perfectly solved.

6 CONCLUSION

We propose a novel debiased estimator, the adaptive indirect estimator, used in parametric bootstrap for consistent private statistical inference, which solves the issue of clamping in DP mechanisms affecting statistical inference in a flexible and general way. Our estimator is based on the indirect estimator [13] where the classic indirect inference uses the asymptotic normality of the parameter estimation while we use parametric bootstrap instead to address the new challenge coming from the DP mechanisms.

One direction of future work is the analysis of using a combination of the indirect estimator and other debiased estimators in parametric bootstrap. Another direction is finding a more efficient statistic $\hat{\beta}_n$ used in the indirect estimator under various DP settings.

REFERENCES

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

[2] Jordan Awan and Zhanyu Wang. 2023. Simulation-based, Finite-sample Inference for Privatized Data. arXiv preprint arXiv:2303.05328 (2023).

[3] Rudolf Beran. 1987. Prepivoting to reduce level error of confidence sets. Biometrika 74, 3 (1987), 457–468.

[4] Rudolf Beran. 1997. Diagnosing bootstrap success. Annals of the Institute of Statistical Mathematics 49 (1997), 1–24.

[5] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. 2011. Differentially private empirical risk minimization. Journal of Machine Learning Research 12, 3 (2011).

[6] Thomas J Diciccio and Joseph P Romano. 1989. The automatic percentile method: Accurate confidence limits in parametric models. Canadian Journal of Statistics 17, 2 (1989), 155–169.

[7] Jinshuo Dong, Aaron Roth, and Weijie Su. 2021. Gaussian Differential Privacy. Journal of the Royal Statistical Society (2021).

[8] Wenxin Du, Canyon Foot, Monica Moniot, Andrew Bray, and Adam Groce. 2020. Differentially private confidence intervals. arXiv preprint arXiv:2001.02285 (2020).

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

[10] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. 9, 3-4 (2014), 211–407.

[11] Bradley Efron. 1981. Nonparametric standard errors and confidence intervals. canadian Journal of Statistics 9, 2 (1981), 139–158.

[12] Cecilia Ferrando, Shufan Wang, and Daniel Sheldon. 2022. Parametric Bootstrap for Differentially Private Confidence Intervals. In International Conference on Artificial Intelligence and Statistics. PMLR, 1598–1618.

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

[14] Andrii Ilienko. 2013. Continuous counterparts of Poisson and binomial distributions and their properties. Annales Univ. Sci. Budapest., Sect. Comp. 39 (2013), 137–147.

[15] Aad W Van der Vaart. 2000. Asymptotic statistics. Vol. 3. Cambridge university press.

[16] Zhanyu Wang, Guang Cheng, and Jordan Awan. 2022. Differentially Private Bootstrap: New Privacy Analysis and Inference Strategies. arXiv preprint arXiv:2210.06140 (2022).

[17] Min-ge Xie and Peng Wang. 2022. Repro Samples Method for Finite-and Large-Sample Inferences. arXiv preprint arXiv:2206.06421 (2022).

[18] Yuming Zhang, Yanyuan Ma, Samuel Orso, Mucyo Karemera, Maria-Pia Victoria-Feser, and Stéphane Guerrier. 2022. A Flexible Bias Correction Method based on Inconsistent Estimators. arXiv preprint arXiv:2204.07907 (2022).