Differentially Private Bootstrap: New Privacy Analysis and Inference Strategies

Zhanyu Wang [1] Guang Cheng [2] Jordan Awan [1]

[1]Department of Statistics, Purdue University [2]Department of Statistics, University of California, Los Angeles

Motivation and Contributions

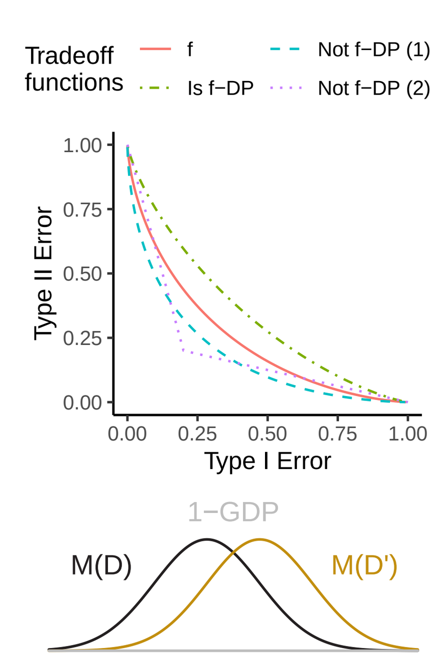
- Most of prior work focused on differential private (DP) point estimates of a parameter, but not general-purpose methods to quantify the uncertainty of a DP procedure.

We obtain a tight privacy analysis of a DP bootstrap and develop inference strategies.

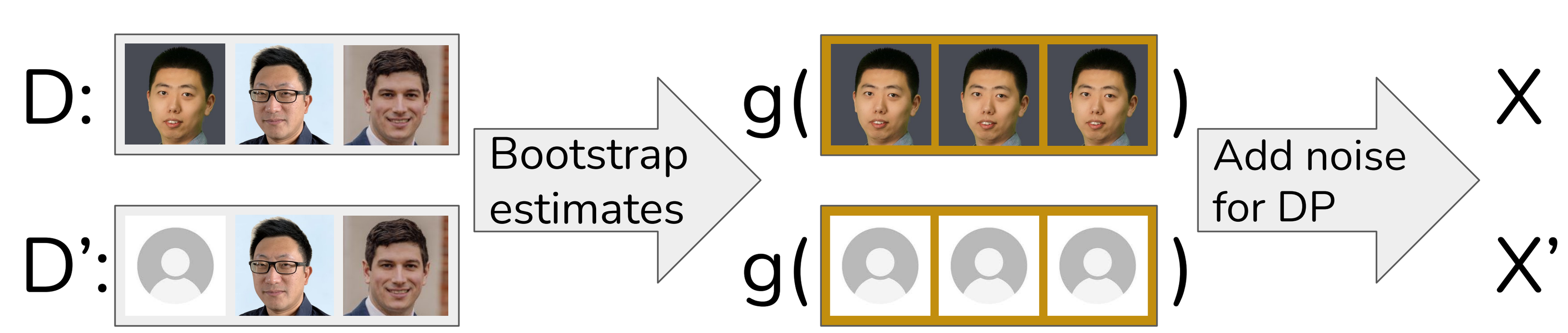
- We derive the privacy guarantee of the DP bootstrap for one bootstrap estimate.
- We quantify the asymptotic cumulative privacy cost of many DP bootstrap estimates.
- We use **deconvolution** on the DP bootstrap estimates to obtain a private estimate of the sampling distribution. For real-world experiments, our private CIs achieve the nominal **coverage** level and offer the first approach to private inference for quantile regression.

Background and f-DP (Dong et al., 2022)

- $\mathcal{M}$ inputs a dataset D and outputs a random variable. $d(\cdot,\cdot)$: Hamming distance.
- With one observation $X \sim \mathcal{M}(D_{\text{true}})$, consider a **hypothesis test** $H_0: X \sim \mathcal{M}(D), H_1: X \sim \mathcal{M}(D')$ where $d(D, D') \le 1$. Then $\mathcal{M}$ provides a stronger privacy guarantee if this test is harder.
- For $H_0: X \sim P$, $H_1: X \sim Q$, and any rejection rule $\phi(X)$, $T_{P,Q}(\alpha)$ is the **tradeoff function** which maps the **type I error** α to the smallest corresponding **type II error**.
- $\mathcal{M}$ is f-DP if $T_{\mathcal{M}(D),\mathcal{M}(D')} \ge f$ for any $d(D, D') \le 1$.
- If $f = T_{\mathcal{N}(0,1),\mathcal{N}(\mu,1)}$, f-DP is called μ-Gaussian DP (GDP).
- Gaussian Mechanism: $\mathcal{M}(D) = g(D) + \xi$ satisfies μ-GDP if $\xi \sim \mathcal{N}(0, \text{sensitivity}(g)^2/\mu^2)$, $\text{sensitivity}(g) = \sup_{d(D,D')\le 1} |g(D) - g(D')|$.



Difficulty in privacy analysis with bootstrap estimates



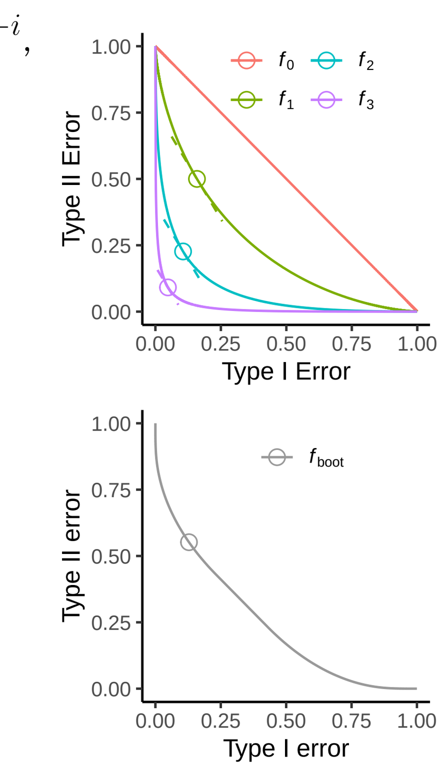
- **Bootstrap** (Efron, 1979) is sampling with replacement.
  → Sensitivity after Bootstrap can be n times larger than without Bootstrap.
- We need many bootstrap estimates for an accurate estimation of sampling distribution.
  → More observations make the hypothesis test easier therefore weaker privacy guarantee.

DP guarantee with single and multiple bootstrap estimates

- $\mathcal{M} \circ \texttt{boot}(D)$ indicates that the input of $\mathcal{M}$ is a bootstrap sample of dataset D.
- Given $\mathcal{M}$ being (ε,δ)-DP, Balle et al. (2018) proved a lower bound for $\mathcal{M} \circ \texttt{boot}$ in (ε,δ)-DP which can be converted to an f-DP bound using results in (Dong et al., 2022).
- However, the converted result is intractable and cannot be easily evaluated.

Theorem 1: Tractable result of one DP Bootstrap estimate in f-DP

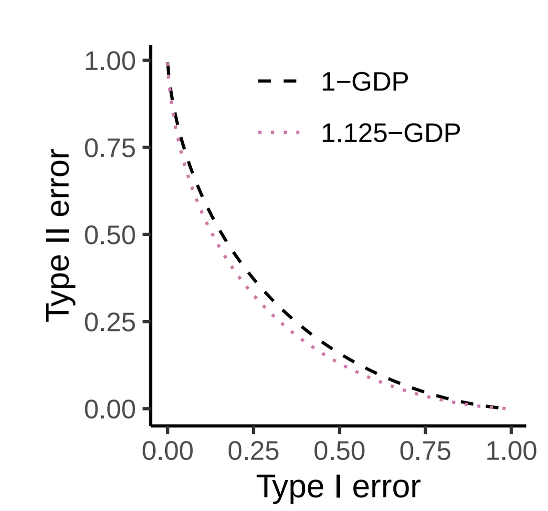
- For a bootstrap sample, with probability $p_i = \binom{n}{i}(1/n)^i(1-1/n)^{n-i}$, the different individual in D and D' is chosen i times.
- Assume $\mathcal{M}$ satisfies f_i-Group DP with group size i
  - i.e., $T_{\mathcal{M}(D),\mathcal{M}(D')} \ge f_i$ for all $d(D, D') \le i$.
  - e.g., if $\mathcal{M}$ is μ-GDP, then for group size i, it is also $i\mu$-GDP.
- For any **slope** $\lambda \in (-\infty, 0]$, find α_i such that $f_i'(\alpha_i) = \lambda$.
  Let $\underline{f} = (f_1, \dots, f_k)$. $\underline{p} = (p_1, \dots, p_k)$, $\alpha = \sum_{i=1}^k p_i\alpha_i$.
  Define $\text{mix}(\underline{p}, \underline{f}): \alpha \mapsto \sum_{i=1}^k p_i f_i(\alpha_i)$ by parameterization in λ.
- $\mathcal{M} \circ \texttt{boot}$ is $f_{\texttt{boot}}$-DP where $f_{\texttt{boot}} := \text{mix}\left((p_0, \underline{p}), (f_0, \underline{f})\right)$, $p_0 = (1-1/n)^n$, $\underline{p} = \frac{1}{1-p_0}(p_1, \dots, p_n)$, $f_0(\alpha) = 1 - \alpha$.
- A stronger result is $f_{\texttt{boot}} := \text{Symm}(p_0 f_0 + (1-p_0)\text{mix}(\underline{p}, \underline{f}))$ and $\text{Symm}(\cdot)$ maps asymmetric tradeoff functions to symmetric ones.



- If for $i = 1, \dots, k$, f_i' is monotonically increasing for every α in $[0,1]$, we have $\text{mix}(\underline{p}, \underline{f}) = (\sum_{i=1}^k (p_i f_i \circ (f_i')^{-1})) \circ (\sum_{i=1}^k p_i (f_i')^{-1})^{-1}$ since $(\sum_{i=1}^k p_i (f_i')^{-1})$ maps the slope λ to the **type I error**, and $(\sum_{i=1}^k (p_i f_i \circ (f_i')^{-1}))$ maps the slope λ to the **type II error**.
- Matching the slope follows the **Neyman-Pearson lemma** as the slope is the negative **likelihood ratio** between $M(D')$ and $M(D)$ on the boundary of the optimal rejection region.
- To achieve this lower bound, the adversary needs to separately design rejection rule ϕ_i with type I error α_i for the case that the bootstrap samples having i copies of the different individual, and the adversary also needs to optimize α_i.

Theorem 2: Asymptotic composition result of many DP Bootstrap estimates in GDP

- Assume $\mathcal{M}_{B,i}$ satisfies μ_B-GDP.
- $\mathcal{M}_i' = \mathcal{M}_{B,i} \circ \texttt{boot}$, $\mathcal{M}_{\texttt{boot}}^B = \{\mathcal{M}_1', \dots, \mathcal{M}_B'\}$.
- $\mathcal{M}_{\texttt{boot}}^B$ is **asymptotically** μ-GDP if $\lim_{B\to\infty} \mu_B\sqrt{(2-2/e)B} \to \mu$. $(\sqrt{(2-2/e)} < 1.125)$
- For Gaussian mechanism, if adding noises $\xi \sim \mathcal{N}(0, \sigma^2)$ on the non-private output guarantees μ-GDP, then for $(\mu/\sqrt{B})$-GDP, we only need to add $\xi \sim \mathcal{N}(0, B\sigma^2)$.

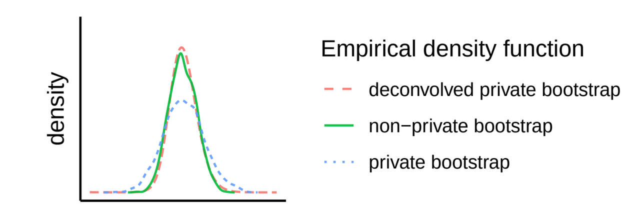


Deconvolution for estimating sampling distribution

To recover the sampling distribution from DP Bootstrap estimates, we use **additive noise mechanism** $\tilde{y} = y + \xi$ to guarantee DP and use **deconvolution** to recover the distribution of y (bootstrap estimates) from DP bootstrap estimates $\tilde{y}$ and the distribution of ξ (added noises).

- Let $Z = X + Y$; its PDF $f_Z(t) := \int_{-\infty}^{\infty} f_X(\tau) f_Y(t-\tau) d\tau$.
- Deconvolution is solving f_X given f_Y and f_Z.
- In practice, we use the `deconvolveR` (Narasimhan and Efron, 2020). It works well when $1 \le \text{Var}(X)/\text{Var}(Y)$; Otherwise, the estimate $\hat{f}_X$ is usually flatter than f_X.

Figure 1. DP Bootstrap and deconvolution.



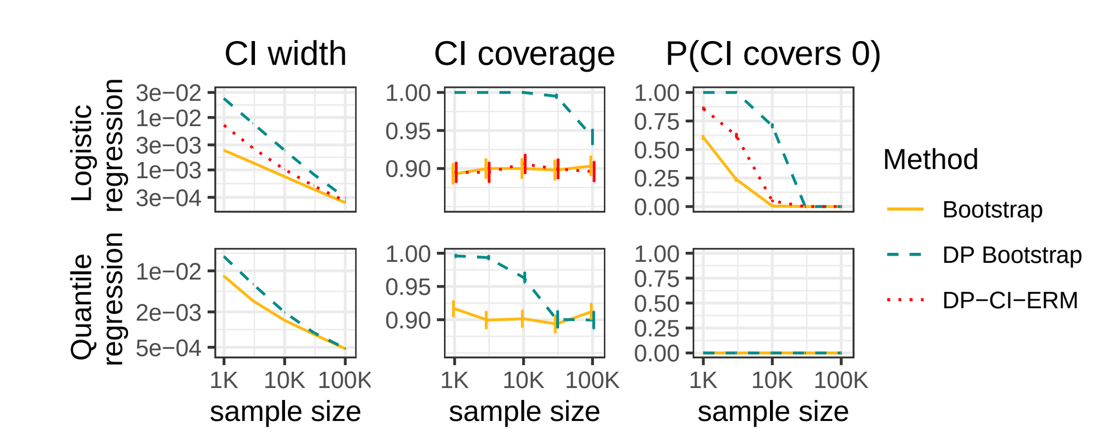
Private confidence intervals (CI)

- We construct private CIs using **quantiles** of the deconvolved sampling distribution. We compare our DP Bootstrap with NoisyVar (Du et al., 2020) and DP-CI-ERM (Wang et al., 2019) on the 2016 Canada Census dataset. We build CIs for the population mean income and for the slope parameter in the logistic regression and quantile regression between the market income and shelter cost. (The confidence level is 90%, and the privacy guarantee is 1-GDP.)

Table 1. 90% CIs for the mean income. $(n = 200,000, B = 100)$

Method	CI Coverage	CI Width
Bootstrap	0.910 (0.006)	279.4 (0.54)
DP Bootstrap	0.905 (0.007)	291.0 (0.54)
NoisyVar	0.857 (0.008)	253.6 (0.16)

Figure 2. Results of CIs for the regression coefficient. Note that DP-CI-ERM cannot be used on quantile regression.



Discussions and References

- New inference techniques are needed when using non-additive noise mechanisms.
- New composition results may help quantify non-asymptotic cumulative privacy costs.

Balle, B., Barthe, G., and Gaboardi, M. (2018). Privacy amplification by subsampling: Tight analyses via couplings and divergences. Advances in Neural Information Processing Systems, 31.

Dong, J., Roth, A., and Su, W. J. (2022). Gaussian differential privacy. Journal of the Royal Statistical Society: Series B, 84(1).

Du, W., Foot, C., Moniot, M., Bray, A., and Groce, A. (2020). Differentially private confidence intervals. arXiv:2001.02285.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7(1).

Narasimhan, B. and Efron, B. (2020). deconvolveR: A g-modeling program for deconvolution and empirical bayes estimation. Journal of Statistical Software, 94.

Wang, Y., Kifer, D., and Lee, J. (2019). Differentially private confidence intervals for empirical risk minimization. Journal of Privacy and Confidentiality, 9(1).