DIFFERENTIALLY PRIVATE NON-PARAMETRIC CONFIDENCE INTERVALS

Katrina Ligett, Moshe Shenfeld, Tomer Shoham, and Noa Velner-Harris

Department of Computer Science
The Hebrew University of Jerusalem

ABSTRACT

We propose a general framework for differentially private confidence interval estimation. Our approach takes any differentially private estimator of any arbitrary quantity as a black box, and from it constructs differentially private non-parametric confidence intervals of that quantity. In fact, our method produces a full CDF estimation of the private statistic. Our approach uses privacy amplification to leverage the randomness that naturally arises from common subsampling and bootstrapping techniques. Under mild assumptions, the quality of our estimator is asymptotically equivalent to its non-private counterpart, as long as the dataset size $n = \Omega\left(\frac{\ln(1/\delta)}{\varepsilon^2}\right)$, and we show promising initial numerical results that support that claim.

1 Introduction

Confidence interval (CI) estimation is a fundamental aspect of statistical analysis, crucial for quantifying estimator uncertainty and facilitating hypothesis testing. Given a significance level $\gamma \in (0, 1)$, a valid confidence interval includes the true quantity of interest w.p. $> 1 - \gamma$. In parametric scenarios where the underlying distribution of the statistic is known, CIs can be derived from distribution parameters, which are estimated from a sample. In such cases, differentially private CIs can be derived from a private estimator of this parameter in a relatively straightforward manner [9, 12, 18]. In the non-parametric setting, however, where minimal assumptions (if any) are made about the distribution (e.g., bounded range or moments), such approaches are not applicable. If the distribution of the statistic approaches a known parametric limiting distribution (e.g., a normal distribution), CIs can be derived by estimating its parameters, which can also be done privately [18]. In absence of a convenient limiting distribution, CIs can sometimes be constructed for specific quantities when their CI can be expressed as another parameter which can be empirically estimated. For example, Drechsler et al. [8] provide non-parametric DP CIs for the median by directly estimating other quantiles.

Another approach to non-private CIs relies on resampling methods, such as bootstrapping (sampling with replacement), initially introduced by Efron [10]. These methods offer convergence rates comparable (and even superior in some cases) to normal approximation and similar techniques. However, sampling with replacement implies each element might participate in the resampled dataset more than once, increasing the query's sensitivity to changing a single element, adversely affecting the privacy-accuracy tradeoff. Brawner and Honaker [5] try to get around this with a variant of bootstrapping where the maximal number of times each element can be sampled is bounded by $\ln(\ln(n))$ w.h.p., and so the statistical guarantees are nearly identical to classical bootstrapping. Unfortunately, while the bound on the number of appearances is nearly constant, this still leads to a blowup in the sensitivity, and a proportional increase in the privacy parameter ε.

An alternative technique for mitigating the effect of high sensitivity is the sub-sample and aggregate framework [14], where the dataset is split into disjoint subsets, each one is used to produce a non-private output, and those outputs are aggregated in a private manner. In recent years, a technique known as Bag-of-Little-Bootstrap (BLB), proposed in 2014 by Kleiner et al. [13], has emerged as a valuable tool for bootstrapping large databases and has found application in DP estimation [7, 6]. It relies on bootstrapping samples of the original size from each subset, to construct a non-private CI. Importantly, these methods do not require a private estimator of the target quantity.

1.1 Our proposal

Our method also relies on resampling, but takes a different path. We avoid the pitfall of increased sensitivity by using smaller subsets, and using a private estimator allows us to make use of all estimated statistics. Given a dataset of size n, we propose subsampling (without replacement) [15] or bootstrapping [4] T subsets of size $m < n$, and then using a DP mechanism on each subset to obtain T statistics. Combining DP's composition and amplification-by-subsampling properties, this results in a private estimation of the CDF of the (private estimation of the) statistics on a dataset of size m. Any post-processing of this CDF to a CI retains privacy, and unlike the BLB method it provides private CIs for all significance levels simultaneously.

While privacy amplification is known for sampling both with and without replacement, privacy guarantees for sampling with replacement are comparable to those of sampling without replacement only in the $m \ll n$ regime [2]. In this regime, statistical analysis of these two sampling methods is nearly identical, so this work focuses mainly on subsampling without replacement. Composition of the privacy guarantees over the T subsamples can be done using either basic or advanced composition. If the private estimator of the statistic is pure-DP ($\delta = 0$), basic composition results in pure-DP CIs.

Our statistical analysis relies on the commonly used assumptions that, under some standardization, the statistic's limiting distribution is normal. This applies, for example, to bounded linear queries (via the CLT). We also rely on the error induced by the private estimator being dominated by the statistical noise for sufficiently large sample sizes. This assumption holds for many mechanisms (e.g., Laplace/Gaussian noise addition to linear queries), where the sampling error of a statistic over a dataset of size n is $O\left(\frac{1}{\sqrt{n}}\right)$, while the error induced by the private mechanism is $O\left(\frac{1}{n\varepsilon}\right)$. By balancing the subset size m and number of subsets T and lower bounding $n = \Omega\left(\frac{\ln(1/\delta)}{\varepsilon^2}\right)$, we can ensure the asymptotic error of our technique approaches those of its non-private version as the sample size n grows.

2 Algorithm description

In this section, we present our main proposal for constructing confidence intervals. We have two sampling methods, with or without replacement: bootstrapping and subsampling. Denote by Y_i the output of the private mechanism applied on the ith subsample, and by $\bar{Y}$ the average of all Y_i. We propose to construct confidence intervals at a $1-\gamma$ significance level based on one of two methods. The first is normal approximation, that is,

$$CI_{Norm}((Y_1, ..., Y_T), \gamma, C)) = \bar{Y} \pm Z_{1-\gamma/2} \cdot \hat{\sigma} \text{ with } \hat{\sigma} = \sqrt{C \cdot \frac{1}{T-1} \sum_{i=1}^{T} (Y_i - \bar{Y})^2}, \tag{1}$$

where $Z_{1-\gamma/2}$ is the inverse CDF of the standard Gaussian distribution at the point $1-\gamma/2$, and C is some extrapolation factor that depends on the sampling method. For example, when using subsampling, Shao and Wu [16] show that taking $C = \frac{m}{n-m}$ gives a consistent estimator of the SD. The extrapolation factor is needed since we estimate from a sample of size m and not n, and different assumptions invite different extrapolations that depend on the non-private mechanism.

The second method, the quantile method, uses the quantiles of the empirical distribution of the estimator,

$$CI_{Quan}((Y_1, ..., Y_T), \gamma, C) = \left[\bar{Y} - C \cdot (\bar{Y} - \widetilde{Y}_{\lfloor T\gamma/2 \rfloor}), \bar{Y} + C \cdot (\widetilde{Y}_{\lceil T(1-\gamma/2) \rceil} - \bar{Y})\right], \tag{2}$$

where $\widetilde{Y}_1, ..., \widetilde{Y}_T$ are the sorted values of $Y_1, ..., Y_T$, and C is some extrapolation factor. $\bar{Y}$ can also be replaced by the median of $(Y_1, ..., Y_T)$ to get more stable results.

Definition 2.1 (PrivSub$_{n,m,T}(\mathcal{M})$)**.** *Given a dataset $X \in \mathbb{R}^{n \times *}$, a subsample size $m \in \mathbb{N}$, and number of subsamples $T \in \mathbb{N}$, the algorithm PrivSub$_{n,m,T}(\mathcal{M})$, gets a DP mechanism $M \in \mathcal{M} : \mathcal{X}^n \to \mathbb{R}$, samples T subsamples without replacement, and applies the DP mechanism on each subsample. Based on these estimates (which form a private CDF of the estimator), we can constructs a confidence interval using one of two approaches: normal approximation* (1) *or the quantile method* (2)*.*

3 Asymptotic analysis

In this section, we provide an initial asymptotic analysis of privacy, validity, and accuracy. The technical results are presented informally, emphasizing the intuition behind our analysis.

3.1 Privacy analysis

Our privacy analysis relies on privacy amplification by subsampling [2] to reduce privacy loss from each estimation by approximately the ratio of the subset size to the full dataset size. Amplification by subsampling, in a nutshell, is based on the idea that a differentially private mechanism run on a random subsample of a population provides higher privacy guarantees than when run on the entire population. Advanced composition enables us to bound the privacy loss growth by the square root of the number of subsamples.

Lemma 3.1. *(informal) Given $n, T \in \mathbb{N}$; $m \in [n]$; $\varepsilon, \delta \in (0, 1]$, and a (ε, δ)-DP $\mathcal{M} : \mathcal{X}^n \to \mathbb{R}$, then PrivSub$_{n,m,T}(\mathcal{M})$ is (ε', δ')-DP where $\varepsilon' = \Omega\left(\frac{m\sqrt{T\ln(1/\delta)}}{n}\varepsilon\right)$ and $\delta' = \Omega(T\delta)$ (by advanced composition).*

Additionally, if $\delta = 0$, PrivSub$_{n,m,T}(\mathcal{M})$ is $(\varepsilon'', 0)$-DP where $\varepsilon'' = \Omega\left(\frac{Tm}{n}\varepsilon\right)$ (by basic composition).

The lemma relies on the fact that for small enough ε, amplification by subsampling gives us that sampling a sample of size m out of n without replacement gives privacy amplification of $\varepsilon' = \Omega\left(\frac{m}{n}\varepsilon\right)$, where ε is the original privacy parameter [3, Theorems 9 and 10]. The analysis of δ is more complex, where for subsampling, we have privacy amplification of $\frac{m}{n}$, and for bootstrapping, it requires more subtle analysis of the algorithm. From advanced composition we have that for T samples, one pays $\Theta(\sqrt{T})$ in ε, and linearly in δ, for $\delta > 0$. For our algorithm to be pure DP, basic composition is required, and then ε scales linearly with T.

3.2 Statistical analysis

The first step is to show that our proposed algorithm produces valid confidence intervals, that is, when $n, m \to \infty$ and T is sufficiently large, then the coverage probability of the confidence interval is at least $1 - \gamma$, where γ is the required significance level. Notice that the distribution of the privately estimated statistic is a convolution of the underlying distribution and additional randomness of the private mechanism. Assuming the error induced by the additional randomness is negligible relative to the sampling error for sufficiently large sample size, it suffices to show that the CI is valid when using non-private estimations of the statistics.[1] This is standard for both standard-deviation-based CI using normal approximation [16, 11], and quantile-based CI using subsampling [15, 17, 19].

Unlike validity, there are several possible accuracy definitions for a CI, such as its (expected) width, or tightness, the extent to which the probability that the true quantity is in the CI exceeds $1 - \gamma$. The accuracy analysis also depends on the way the CI is constructed, and the sampling method. We provide a general claim that reduces the analysis of the private CI's accuracy to its non-private counterpart.

Lemma 3.2. *Given $n \in \mathbb{N}$, $\varepsilon > \frac{1}{\sqrt{n}}$, a statistic with a normal limiting distribution, and a ε-DP mechanism $\mathcal{M}$ for estimating this statistic from a dataset of size n, if the mechanism's sample error is $O(\frac{1}{n\varepsilon})$, then the distribution error of the mechanism's output scales like $O(1/\sqrt{n})$, which is asymptotically equivalent to the error of the non-private estimator.*

3.3 Combined analysis

Putting everything together, we have the following informal statement.

Theorem 3.3. *(informal) Let $\varepsilon', \delta' \in (0, 1)$. If $n = \Omega\left(\frac{\ln(1/\delta')}{\varepsilon'^2}\right)$, then for any $T \in \mathbb{N}$; $m \in [n]$ that satisfy $mT = O(n)$, and for any mechanism $\mathcal{M}$ satisfying (ε, δ)-DP for $\varepsilon = \Omega\left(\varepsilon'\sqrt{\frac{n}{m\ln(1/\delta')}}\right)$ and $\delta = \Omega(\delta'/T)$, for which the assumptions of Lemma 3.2 hold, we have that*

1. ***Privacy:** $PrivSub_{n,m,T}(\mathcal{M})$ is (ε', δ')-DP.*
2. ***Validity:** The confidence interval constructed from the output of $PrivSub_{n,m,T}(\mathcal{M})$ is asymptotically valid.*
3. ***Accuracy:** The accuracy of the private CI is asymptotically identical to the non-private CI.*

Proof. The first claim is a direct result of Lemma 3.1. The second claim can be proved with classical statistical methods, as explained in Subsection 3.2. The third claim is the result of noting that the conditions of Lemma 3.2 hold; that is, the perturbation error is at most the statistical error. □

Notice that while Part 3 of Theorem 3.3 bounds the error of the private CI by that of the non-private CI, that error depends on m and T, and we do not provide a bound for the general case, that is, we only ensure that the sampling

[1] In fact, weaker assumptions suffice. For normal approximation we only need to assume the added randomness can be normally approximated, and for the quantile method no additional assumption is needed.

error dominates the perturbation. For the non-private CI to be accurate, we typically choose $m = \Omega(\sqrt{n})$. We further argue that under normal approximation, it suffices that $T = O(\ln(1/\gamma))$, which allows for $m = \Theta(n)$, in which case the error of the non-private CI is $O\left(\frac{1}{\sqrt{n}}\right)$. We have not yet developed the full analysis of this point.

4 Initial numerical study

In this section we describe one of several numerical analyses we ran to support our results. Here we evaluate four CI methods for the median of a normal distribution $\mathcal{N}(0, 2^2)$ truncated to the range $[-6, 4]$, all relying on sampling m out of n without replacement. The methods differ by the way each of the T estimations were produced (using a private mechanism or non-private estimation of the statistic) and the way the CI was derived from those estimations (quantile or normal approximation). We additionally compare our results to the correct value (labeled "empirical results"), which in the case of the coverage is the desired confidence level (0.9), and in the case of the width is the CI of the empirical estimation using samples of size n.

As the private estimator we use the inverse sensitivity mechanism as implemented by Asi and Duchi [1], and its privacy parameter ε' was set such that $\text{PrivSub}_{n,m,T}(\mathcal{M})$ is (ε, δ)-DP for the choice of m and T. Optimizing these parameters to ensure statistical validity and minimize CI width requires further research. We set T to be a constant (200), and m to be $O(\sqrt{n})$.[2]

The left panel of the figure depicts the expected width of the CI in log-scale, to make it easier to compare the wide range of values. We can see that the two non-private methods closely match the empirical one, which decreases as $1/\sqrt{n}$, while the private yield significantly wider CIs for small sample size but approach the non-private ones as n grows. In the right panel we present the coverage of the four methods. We can see that the two non-private ones reach the desired coverage, while the private ones have a somewhat higher coverage for small values of n.

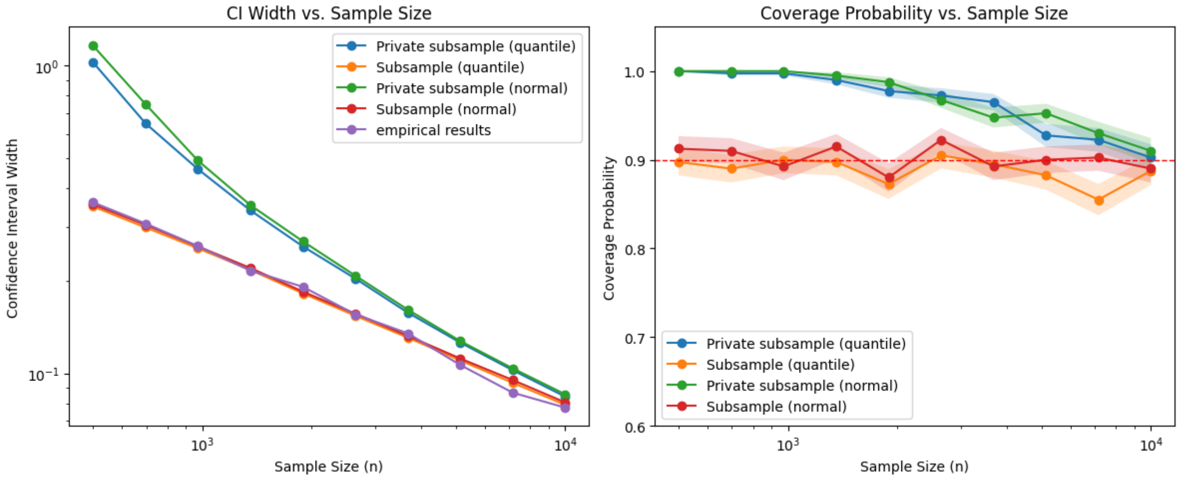

Figure 1: CI width and coverage of the four methods, with privacy budget $\varepsilon = 4, \delta = 10^{-6}$, n varying from 500 to $10,000$. The shaded area shows ± one standard deviation around the estimate.

5 Ongoing work

In this manuscript, we describe the main idea and some initial results of our framework for private non-parametric confidence interval estimation. We are actively developing both theoretical and empirical consequences of this framework. In particular, although our initial empirical results look promising and we have a sketch of a theoretical argument regarding accuracy, we have more work to do to establish rigorous accuracy guarantees. One of the directions we are still resolving is in optimizing the tradeoff between m and T to provide the best possible guarantees. Theorem sketch 3.3 shows that the sample size has to be asymptotically larger than $T \cdot m$, but this still leaves questions open about the rates of m and T. We also are working towards a more extensive numerical study, including comparison of our proposed method with those of Drechsler et al. [8] and Chadha et al. [6].

[2] The exact expression was $m = \max\{50, \min\{5 \cdot \sqrt{n}, 0.1 \cdot n\}\}$, since a very small m leads to large error estimations, and when m/n is close to 1 the samples are strongly correlated.

References

[1] Asi, H. and Duchi, J. C. (2020). Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. *Advances in neural information processing systems*, 33:14106–14117.

[2] 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.

[3] Balle, B. and Wang, Y.-X. (2018). Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pages 394–403. PMLR.

[4] Bickel, P. J., Götze, F., and van Zwet, W. R. (2012). *Resampling fewer than n observations: gains, losses, and remedies for losses*. Springer.

[5] Brawner, T. and Honaker, J. (2018). Bootstrap inference and differential privacy: Standard errors for free. *Unpublished Manuscript*.

[6] Chadha, K., Duchi, J., and Kuditipudi, R. (2024). Resampling methods for private statistical inference. *arXiv preprint arXiv:2402.07131*.

[7] Covington, C., He, X., Honaker, J., and Kamath, G. (2021). Unbiased statistical estimation and valid confidence intervals under differential privacy. *arXiv preprint arXiv:2110.14465*.

[8] Drechsler, J., Globus-Harris, I., Mcmillan, A., Sarathy, J., and Smith, A. (2022). Nonparametric differentially private confidence intervals for the median. *Journal of Survey Statistics and Methodology*, 10(3):804–829.

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

[10] Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer.

[11] Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596.

[12] Karwa, V. and Vadhan, S. (2018). Finite sample differentially private confidence intervals. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[13] Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(4):795–816.

[14] Nissim, K., Raskhodnikova, S., and Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84.

[15] Politis, D. N., Romano, J. P., Wolf, M., Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling in the IID Case*. Springer.

[16] Shao, J. and Wu, C. J. (1989). A general theory for jackknife variance estimation. *The annals of Statistics*, pages 1176–1197.

[17] Shi, X. (1991). Some asymptotic results for jackknifing the sample quantile. *The Annals of Statistics*, pages 496–503.

[18] Wang, Y., Kifer, D., Lee, J., and Karwa, V. (2018). Statistical approximating distributions under differential privacy. *Journal of Privacy and Confidentiality*, 8(1).

[19] Wu, C. F. (1990). On the asymptotic properties of the jackknife histogram. *The Annals of Statistics*, pages 1438–1452.