Accuracy Gains from Privacy Amplification through Sampling for Differential Privacy

Jingchen (Monika) Hu

Vassar College
Joint work with Jörg Drechsler and Hang J. Kim
FCSM 2021

Background

For simple random sample without replacement:

- Population parameter: $g(\boldsymbol{Y}_N)$; sample estimator: $g(\mathcal{S}(\boldsymbol{Y}_N))$
- Goal: achieve (ϵ, δ)-DP for $g(\boldsymbol{Y}_N)$
  - Method 1: Add noise according to (ϵ, δ) to $g(\boldsymbol{Y}_N)$
  - Method 2: Add noise according to (ϵ_n, δ_n) to $g(\mathcal{S}(\boldsymbol{Y}_N))$
- Amplification effects: $\epsilon_n > \epsilon, \delta_n > \delta$, therefore less noise to be added (Li *et al.* (2012))

$$\epsilon_n = \log(1 + N/n(e^{\epsilon} - 1))$$

$$\delta_n = (N/n)\delta$$

Population mean

$\epsilon-$DP and add noise with Laplace Mechanism based on global sensitivity

- Method 1: Add noise according to ϵ to $g(\boldsymbol{Y}_N)$
  - Laplace Mechanism, global sensitivity $\Delta^G = R/N$
  - The variance of the privatized population parameter is

$$V_N = \text{Var}\left(\mathcal{A}(\bar{y}_N)\right) = 2\left(\frac{R}{\epsilon N}\right)^2 \tag{1}$$

- Method 2: Add noise according to ϵ_n to $g(\mathcal{S}(\boldsymbol{Y}_N))$
  - Laplace Mechanism, global sensitivity $\Delta_n^G = R/n$
  - The total variance of the privatized sample mean is

$$V_n = \text{Var}\left(\mathcal{A}(\bar{y}_n)\right) = \left(1 - \frac{n}{N}\right)\frac{S_N^2}{n} + 2\left(\frac{R}{\epsilon_n n}\right)^2 \tag{2}$$

- Difficult to get accuracy gains if the sensitivity strongly depends on the sample size; also need to consider sampling variance

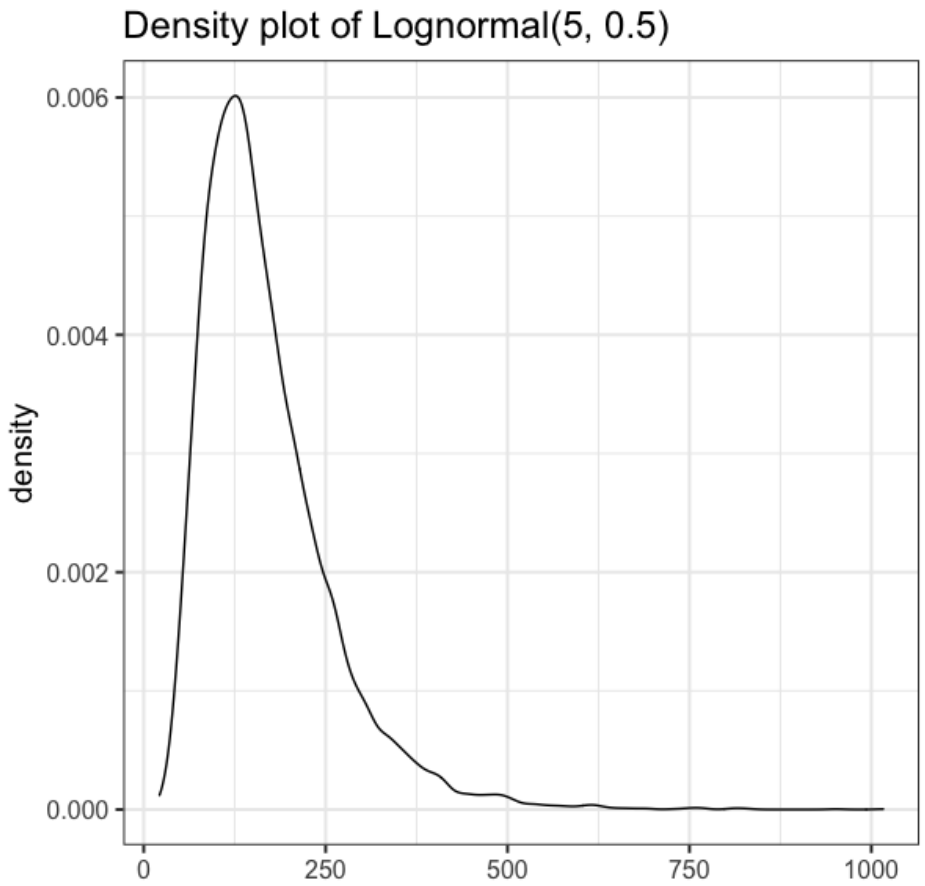
Population median

$(\epsilon, \delta)-$DP and add Laplace noise based on smooth sensitivity (Nissim *et al.* (2007))

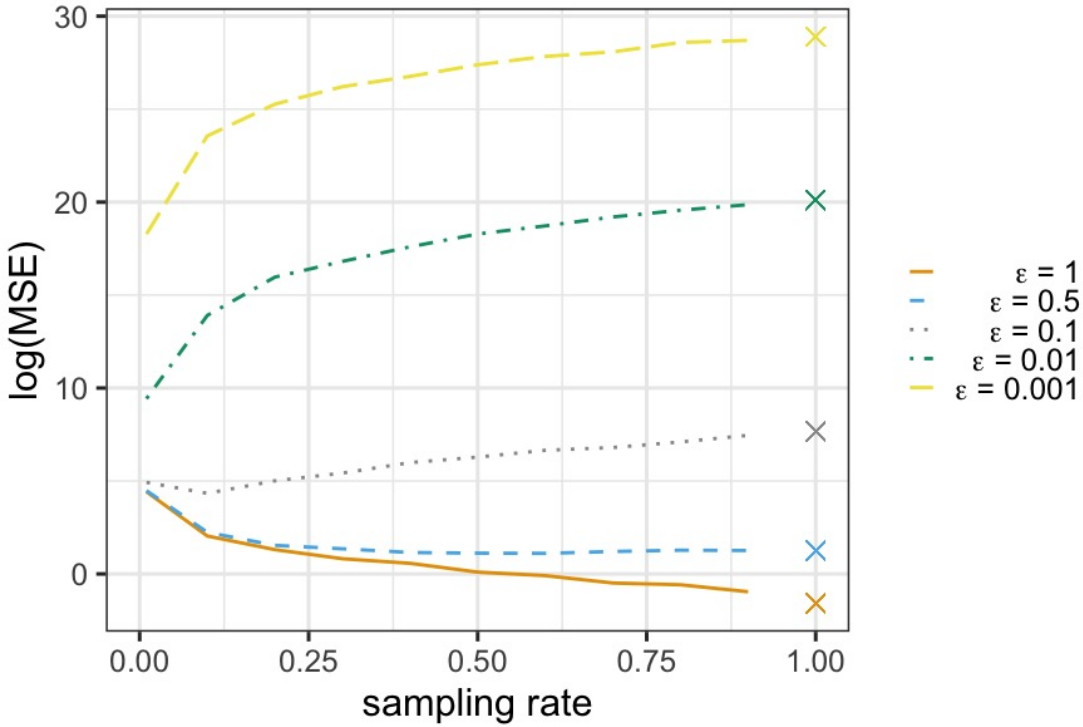
- Method 1: Add noise according to (ϵ, δ) to $g(\boldsymbol{Y}_N)$
  - Laplace noise, smooth sensitivity $\Delta^S_{\epsilon,\delta}(\boldsymbol{Y}_N)$
- Method 2: Add noise according to (ϵ_n, δ_n) to $g(\mathcal{S}(\boldsymbol{Y}_N))$
  - Laplace noise, smooth sensitivity $\Delta^S_{\epsilon_n,\delta_n}(\mathcal{S}(\boldsymbol{Y}_N))$
- Accuracy gains? No analytical solution...

Population median of a lognormal

$N = 10,001$



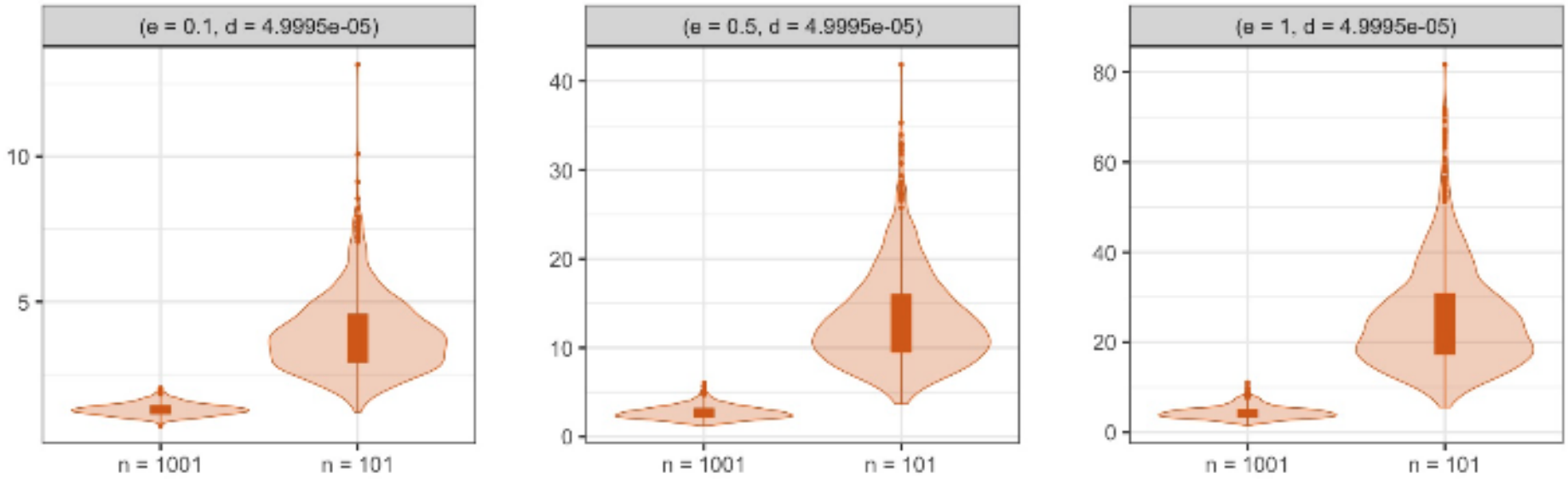
Population median of a lognormal



- log(MSE) of privatized sample medians (1000 repeated samples)
- Accuracy gains for $\epsilon \leq 0.1$ (too small to be used in practice)
- No gains for $\epsilon = 0.5$ and $\epsilon = 1$ (and larger values)
- Reason: no reduced sensitivity

Population median of a lognormal

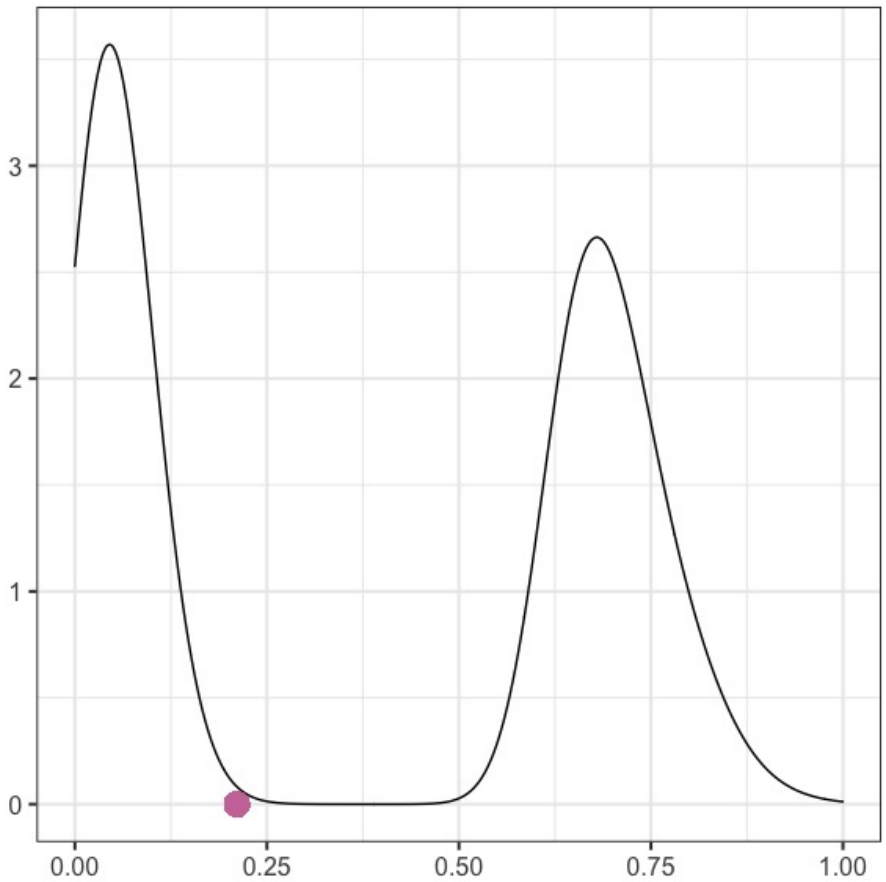
Ratio of the smooth sensitivity of the sample over the smooth sensitivity of the population



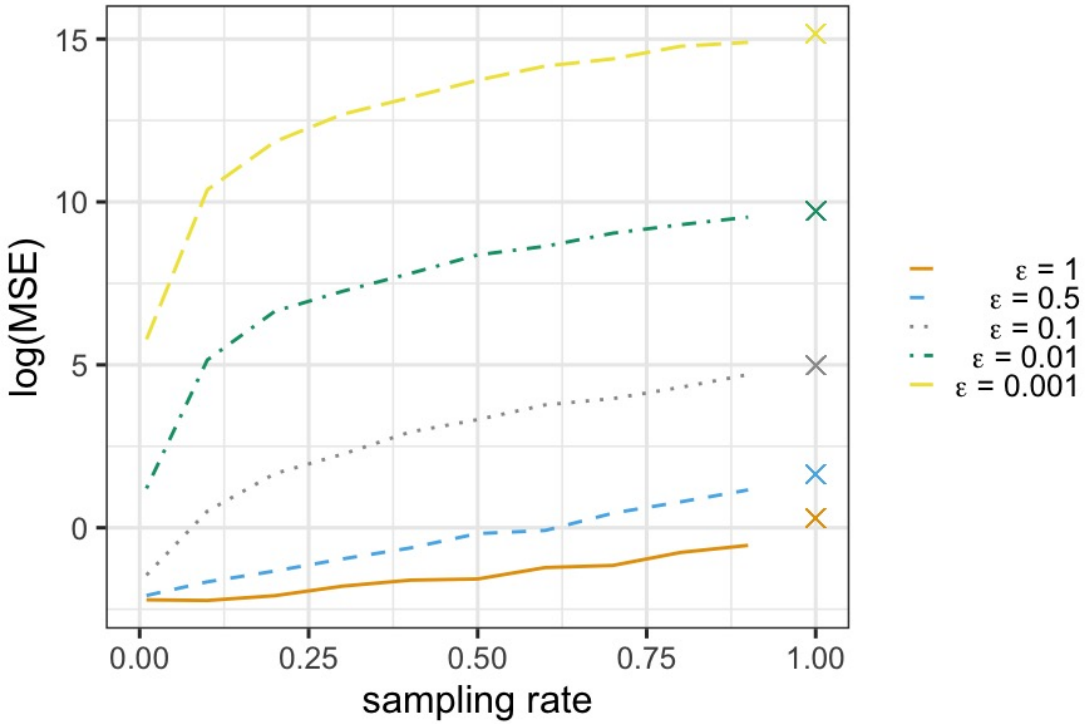
No reduced sensitivity and amplified values of (ϵ_n, δ_n): no accuracy gains

Population median of two scaled and shifted betas

$N = 10,001$



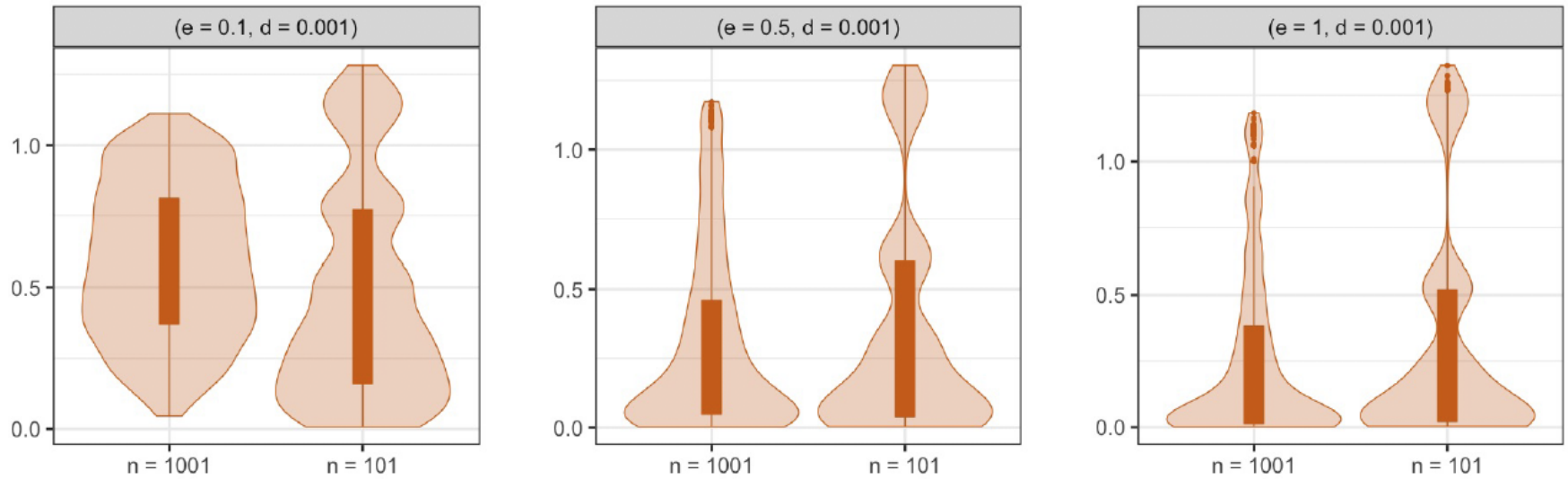
Population median of two scaled and shifted betas



- log(MSE) of privatized sample medians (1000 repeated samples)
- Accuracy gains for all ϵ values being considered
- Reason: reduced sensitivity

Population median of two scaled and shifted betas

Ratio of the smooth sensitivity of the sample over the smooth sensitivity of the population



Reduced sensitivity and amplified values of (ϵ_n, δ_n): accuracy gains

Summary and future work

- No accuracy gains for population mean with global sensitivity and $\epsilon-$DP
- For Lognormal(5, 0.5) with smooth sensitivity and $(\epsilon, \delta)-$DP, accuracy gains with small ϵ
- For mixture of two scaled and shifted betas with smooth sensitivity and $(\epsilon, \delta)-$DP, accuracy gains for a range of ϵ values (as large as $\epsilon = 1$)
- Key: Whether the sensitivity can reduced in the sample; if not, no accuracy gains to be expected; sampling variance also needs to be considered
- Future work: derive the optimal sampling rate from an accuracy perspective, extend to other mechanisms and other statistics, etc.

Summary and future work

- No accuracy gains for population mean with global sensitivity and $\epsilon-$DP
- For Lognormal(5, 0.5) with smooth sensitivity and $(\epsilon, \delta)-$DP, accuracy gains with small ϵ
- For mixture of two scaled and shifted betas with smooth sensitivity and $(\epsilon, \delta)-$DP, accuracy gains for a range of ϵ values (as large as $\epsilon = 1$)
- Key: Whether the sensitivity can reduced in the sample; if not, no accuracy gains to be expected; sampling variance also needs to be considered
- Future work: derive the optimal sampling rate from an accuracy perspective, extend to other mechanisms and other statistics, etc.

Summary and future work

- No accuracy gains for population mean with global sensitivity and $\epsilon-$DP
- For Lognormal(5, 0.5) with smooth sensitivity and $(\epsilon, \delta)-$DP, accuracy gains with small ϵ
- For mixture of two scaled and shifted betas with smooth sensitivity and $(\epsilon, \delta)-$DP, accuracy gains for a range of ϵ values (as large as $\epsilon = 1$)
- Key: Whether the sensitivity can reduced in the sample; if not, no accuracy gains to be expected; sampling variance also needs to be considered
- Future work: derive the optimal sampling rate from an accuracy perspective, extend to other mechanisms and other statistics, etc.

References

- Hu, J., Drechsler, J. and Kim, H. J., Accuracy gains from privacy amplification through sampling for differential privacy. arXiv: 2103.09705.
- Li, N., Qardaji, W., and Su, D. (2012), On sampling, anonymization, and differential privacy or k-anonymization meets differential privacy. In Proceedings of the 7th ACM Symposium on Information, Computer, and Communications Society, 32-33.
- Nissim, K., Raskhodnikova, S., and Smith, A. (2007), Smooth sensitivity and sampling in private data analysis. In Proceedings of the 39th Annual ACM Symposium on Theory of Computing, 75 - 83.