Differential Privacy of Bayesian Inference

Joy Zheng, advised by Salil Vadhan

Undergraduate Researcher (Harvard College '15)

Privacy Tools for Sharing Research Data

A National Science Foundation Secure and Trustworthy Cyberspace Project

Motivation

Do **Bayesian inference** and related methods of **synthetic data generation** in statistics satisfy differential privacy?

While various differential-privacy specific methods for calculating statistical values have been developed [DL08, Smith08], we are interested in the degree to which standard techniques in statistics already fit the definition of differential privacy.

In particular, given that these techniques already involve:

1. randomness via sampling from various distributions
2. boiling down the dataset into a small number of (summary) parameters

they raise the possibility that we already get some privacy by default.

Differential Privacy

An algorithm A is (ε, δ)-differentially private if for all neighboring data sets X, X' which differ on only one data point, and for all sets S of outputs,

$$\Pr[A(X) \in S] \le e^{\varepsilon} \Pr[A(X') \in S] + \delta.$$

Inference Algorithms

Important parameters:

- A data set X of n points, whose points are numbers in some range
- The points are assumed to be drawn independently from some distribution $D(\theta)$, with θ unknown
- There is a prior distribution $\Theta(\theta_0)$ for θ, parameterized by value(s) θ_0

$$X \longrightarrow \theta \xrightarrow{D(\theta)} \text{Synthetic Data}$$

When we generate synthetic data, we do so by first determining a value of θ, and then drawing data points from $D(\theta)$. Unlucky draws of θ can be guarded against in IV (below) by drawing multiple values of θ and generating a few synthetic data points based upon each one.

We look at four types of information releases:

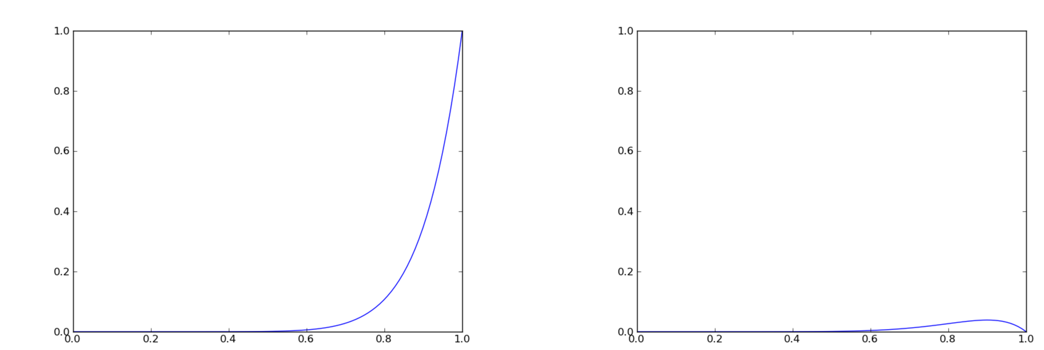
		Generation of θ	
		Most likely value	Drawn from the posterior $\theta \mid X, \theta_0$
Information released	θ only	I (never differentially private, because it is deterministic, but differentially private estimators exist [Smith08])	II
	Synthetic data	III	IV (known as multiple imputation [RRR 03])

Bernoulli/Binomial Distributions

- The data set X is in $\{0,1\}^n$.
- The parameter θ is a real number in $[0,1]$, and gives the probability that any given data point of X is equal to 1.
- The prior Θ is $Beta(\alpha,\beta)$, which is the conjugate prior distribution for the binomial; this can be interpreted as us having seen α-1 prior instances of a 1 and β-1 prior instances of a 0.

Results:

- II is not ε-differentially private.



Probability density functions of $\theta|X$ for an extreme example of neighboring data sets $X=1^{10}$ and $X=(1, 0^9)$, respectively. As we can see, the probabilities vary dramatically between these two data sets.

- There exists c where II is (ε, δ)-differentially private if

$$\alpha, \beta \ge c\left(\frac{1}{\varepsilon^2} \ln n \ln \frac{1}{\delta}\right).$$

Key point: the condition for ε-differential privacy is satisfied so long as the value of θ drawn lies close to its expectation. So long as α,β are large enough that we have seen a reasonable number of examples of both 0's and 1's in our prior and data, then the probability that θ is drawn far away from its expectation is exponentially small in the distance.

- There exists c where III is ε-differentially private if

$$\alpha, \beta \ge \frac{cm}{\varepsilon},$$

where m is the number of synthetic data points drawn.

- There exists c where IV is ε-differentially private if

$$\alpha, \beta \ge \frac{cm}{\varepsilon}.$$

If θ is not redrawn for each synthetic data point, then this is also a lower bound on the strength of the prior distribution needed to achieve privacy.

- There exists c, c_1 where IV is (ε, δ)-differentially private for sufficiently large n and $c_1 n$ synthetic data points if

$$\alpha, \beta \ge c\frac{1}{\varepsilon^2} \ln \frac{1}{\delta},$$

where θ is redrawn for each synthetic data point.

Normal Distributions

- The data set X is in $[-R,R]^n$ for some real number R. (If the data points are allowed to be unrestricted real numbers, then we effectively cannot obtain either ε or (ε, δ)-differential privacy.)
- The parameters $\theta=(\mu,\sigma^2)$ are real numbers, of which σ^2 is known while μ is unknown.
- The (conjugate) prior Θ on μ is another normal distribution $N(0,\sigma_0^2)$.

Results (preliminary):

- II is not ε-differentially private if we allow μ to be unrestricted.
- There exists c where II is ε-differentially private if we restrict μ to $[-R,R]$ and

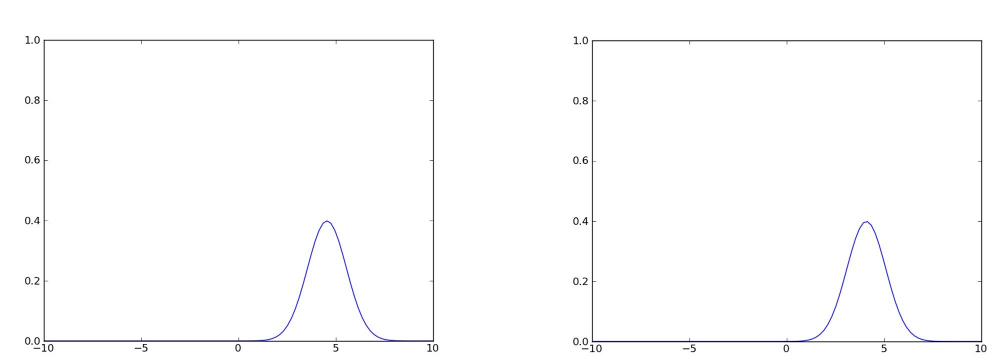
$$\frac{R}{\sigma} \le c\sqrt{\varepsilon}.$$

- There exists c where II is (ε, δ)-differentially private if

$$n \ge \frac{cR^2}{\sigma^2 \varepsilon}\left(\frac{1}{\varepsilon}\ln^2\frac{1}{\delta} + 1\right).$$

- Neither III nor IV are (ε, δ)-differentially private.
- There exists c where III and IV are ε-differentially private for a single synthetic data point z if we restrict z to $[-R,R]$ and

$$n \ge \frac{cR^2}{\sigma^2 \varepsilon}.$$



The probability density functions for z in the cases of neighboring data sets $X=(R^{10})$ and $X=(-R, R^9)$ with $R=5$ and $\sigma=\sigma_0=1$. The log-ratio of these two density functions is actually a linear function of z, which explains why we cannot achieve ε-differential privacy without restricting z.

- There exists c where III and IV are (ε, δ)-differentially private for sufficiently large n if we restrict z to $[-R,R]$ and

$$n \ge \frac{cR}{\sigma\sqrt{\varepsilon}}\left(\frac{1}{\sqrt{\varepsilon}}\ln\frac{1}{\delta} + 1\right).$$

Conclusions

- So far, the randomness involved in inference seems to provide privacy in some cases for the two distributions examined, with the bounds being polynomial in the relevant privacy parameters of $1/\varepsilon$ and $\ln(1/\delta)$.
- Generating synthetic data in these two cases appears more private than releasing the parameter, shown by the looser bounds on parameters needed to satisfy privacy.
- Some commonalities between the two distributions are:
  - We tend to lose ε-differential privacy whenever the parameter drawn lies far away from its expectation.
  - The range of "good" parameters decreases with the size of the data set, but the probability that it lies far away also decreases.
- However, it is not clear how the restrictions we needed to achieve privacy in the case of a normal distribution could be applied to other distributions.
- Moving forward, the goal is to generalize these results; we would like to pull out characteristics of the distributions which imply privacy, as has been done for a variant of differential privacy [DNMR 13].

References

[DL08] Dwork, C. and Lei, J. (2008). Differential privacy and robust statistics. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC).*

[DNMR 13] Dimitrakakis, C., Nelson, B., Mitrokotsa, A., and Rubinstein, B. (2013). Robust, secure, and private Bayesian inference. Available at http://arxiv.org/abs/1306.1066.

[RRR 03] Raghunathan, T.E., Reither, J.P., and Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* 19(1):1-16.

[Smith08] Smith, A. (2008). Efficient, differentially private point estimators. Available at http://arxiv.org/abs/0809.4794.

Contact

shijiezheng@college.harvard.edu

CRCS Center for Research on Computation and Society at Harvard School of Engineering and Applied Sciences

BERKMAN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

The Institute for Quantitative Social Science at Harvard University

DATA PRIVACY LAB