Tailoring Differentially Private Bayesian Inference to Distance Between Distributions

Mark Bun[†], Gian Pietro Farina*, Marco Gaboardi*, Jiawen Liu*

[†]Princeton University, *University at Buffalo, SUNY

Objectives

Design a tool for differentially private probabilistic programming featuring:

1. programming constructs to describe bayesian models and perform probabilistic inference,
2. programming constructs useful to ensure differential privacy,
3. type-checking as a method to ensure that the actual programs are differentially private.

Bayesian Inference Background

beta distribution, $\mathbf{beta}(\alpha, \beta)$, with parameters $\alpha, \beta \in \mathbb{R}^+$, and with p.d.f:

$$\Pr(\theta) \equiv \frac{\theta^{\alpha}(1-\theta)^{\beta}}{\mathrm{B}(\alpha, \beta)}$$

where $\mathrm{B}(\cdot,\cdot)$ is the beta function. The data $\mathbf{x}$ will be a sequence of $n \in \mathbb{N}$ binary values, that is $\mathbf{x} = (x_1, \dots x_n)$, $x_i \in \{0, 1\}$, and the likelihood function is:

$$\Pr(\mathbf{x}|\theta) \equiv \theta^{\Delta\alpha}(1-\theta)^{n-\Delta\alpha}$$

where $\Delta\alpha = \sum_{i=1}^{n} x_i$. From this it can easily be derived that the posterior distribution is:

$$\Pr(\theta|\mathbf{x}) = \mathbf{beta}(\alpha + \Delta\alpha, \beta + n - \Delta\alpha)$$

Differentially private Bayesian inference

Release a private version of posterior distribution $(\tilde{\alpha}, \tilde{\beta}) = (\alpha + \widetilde{\Delta\alpha}, \beta + n - \widetilde{\Delta\alpha})$ where $\widetilde{\Delta\alpha} \sim Lap(\Delta\alpha, \frac{2}{\epsilon})$, and where $Lap(\mu, \nu)$ denotes a Laplace random variable with mean μ and scale ν.

Our Approach - Exponential Mechanism with Smooth Sensitivity

define the mechanism $\mathcal{M}^B_{\mathcal{H}}$ which, given in input a sequence of observations $\mathbf{x}$ and parameters $\epsilon > 0$ and $\delta > 0$, produces an element r in $\mathcal{R}_{\text{post}}$ with probability:

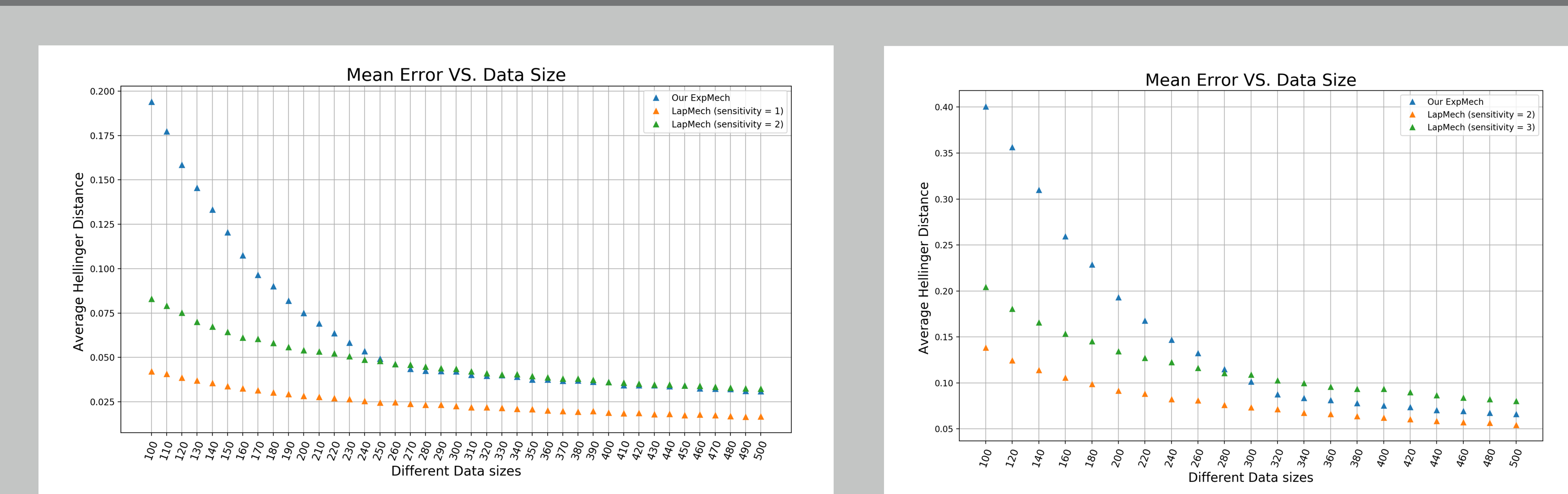
$$\Pr_{z \sim \mathcal{M}^B_{\mathcal{H}}}[z = r] = \frac{exp\left(\frac{-\epsilon \cdot \mathcal{H}(\mathrm{BI}(\mathbf{x}), r)}{2 \cdot S(\mathbf{x})}\right)}{\sum_{r \in \mathcal{R}_{\text{post}}} exp\left(\frac{-\epsilon \cdot \mathcal{H}(\mathrm{BI}(\mathbf{x}), r)}{2 \cdot S(\mathbf{x})}\right)}.$$

The smooth sensitivity is computed as follows:

$$S(\mathbf{x}) = \max_{\mathbf{x}' \in \{0,1\}^n} \left\{ \Delta_l\Big(\mathcal{H}(\mathrm{BI}(\mathbf{x}'), \cdot)\Big) \cdot e^{-\gamma \cdot d(C(\mathbf{x}), C(\mathbf{x}'))} \right\}, \tag{1}$$

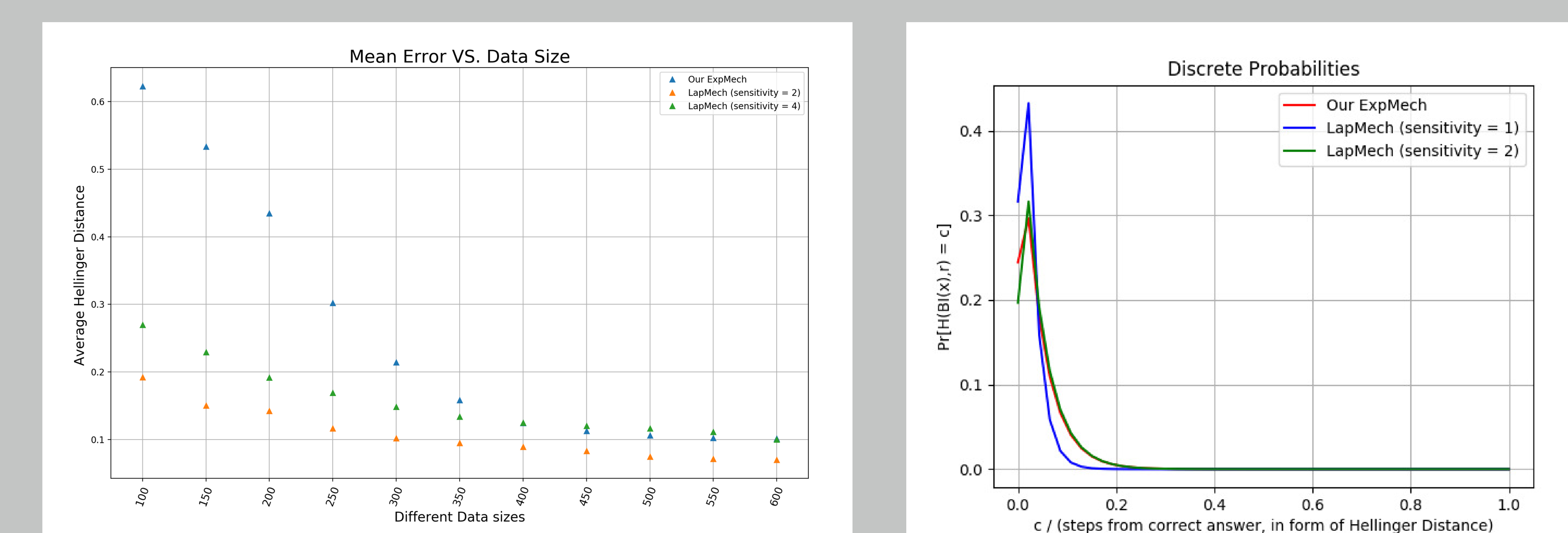
where d is the Hamming distance between two datasets, $\gamma = \gamma(\epsilon, \delta)$ is a function of ϵ and δ to be determined later, and where $\Delta_l\Big(\mathcal{H}(\mathrm{BI}(\mathbf{x}'), \cdot)\Big)$ denotes the local sensitivity at

Some Experimental Results



(a) Data set size from **300** to **800** (b) Data set size from **14000** to **20000**

Figure 1: Increasing data size with fixed prior $\mathbf{beta}(1, 1)$. Unbalanced datasets of mean $(0.1, 0.9)$ and parameters $\epsilon = 0.8$ and $\delta = 10^{-8}$



(a) Data set size from **300** to **800** (b) Data set size from **14000** to **20000**

Figure 2: Increasing data size with fixed prior $\mathbf{beta}(1, 1)$. Unbalanced datasets of mean $(0.1, 0.9)$ and parameters $\epsilon = 0.8$ and $\delta = 10^{-8}$

Conclusion and Future Work

- Our the probabiliy measure approach outperforms the ℓ_1-norm approach when the Laplace noise cannot recognize the data to be protected is histogram and data size grow large.
- 1. The accuracy that we are going to explore next, and in a more principled and formal way.
  2. Experiments have shown that the actual privacy loss in the experiments can be smaller than ϵ. This means that we could improve accuracy, by adding less noise but still achieve (ϵ, δ)-dp.
  3. The choice of the Hellinger distance might seem quite ad-hoc. Hence, it is worth exploring other distances over distributions. An interesting class of probability metrics is the family of f-divergences [1].
  4. Other application of our scheme are going to be explored.

References

[1] I. Csiszár and P.C. Shields. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528, 2004.