

# Tailoring Differentially Private Bayesian Inference to Distance Between Distributions

Mark Bun<sup>†</sup>, Gian Pietro Farina<sup>\*</sup>, Marco Gaboardi<sup>\*</sup>, Jiawen Liu<sup>\*</sup>

<sup>†</sup>Princeton University, <sup>\*</sup>University at Buffalo, SUNY

## Objectives

1. Designing a differentially private Bayesian inference mechanism.
2. Measuring accuracy with a metric over distributions (Hellinger distance ( $\mathcal{H}$ ),  $f$ -divergence, etc.).

## Bayesian inference (BI) (A Beta-Binomial model example):

- Prior on  $\theta$  :  $\text{beta}(\alpha, \beta)$ ,  $\alpha, \beta \in \mathbb{R}^+$ , observed data set  $\mathbf{x} = (x_1, \dots, x_n) \in \{0, 1\}^n$ ,  $n \in \mathbb{N}$ .
- Likelihood function:  $\mathbb{L}_{\mathbf{x}|\theta} = \theta^{\Delta\alpha} (1 - \theta)^{n - \Delta\alpha}$ , where  $\Delta\alpha = \sum_{i=1}^n x_i$ ;
- Posterior distribution over theta:  $\mathbb{P}_{\theta|\mathbf{x}} = \text{beta}(\alpha + \Delta\alpha, \beta + n - \Delta\alpha)$ .

## Differentially Private Bayesian Inference and Motivations

Releasing a differentially private posterior  $\text{beta}(\tilde{\alpha}, \tilde{\beta}) = (\alpha + \widetilde{\Delta\alpha}, \beta + n - \widetilde{\Delta\alpha})$ .

1. Baseline approach is Laplace mechanism (LapMech) with sensitivity proportional to  $\ell_1$  norm. But this sensitivity grows linear with the dimension, also we measure results by a different metric. Motivated by this, we calibrate the noise w.r.t sensitivity of the accuracy metric ( $\mathcal{H}$ ).
2. Maximum sensitivity of  $\mathcal{H}$  over  $\text{beta}$  distributions achieves at edges as in Fig. 1. But it's very low when move away from edge. Motivated by this, we apply smooth sensitivity in mechanism to improve accuracy.

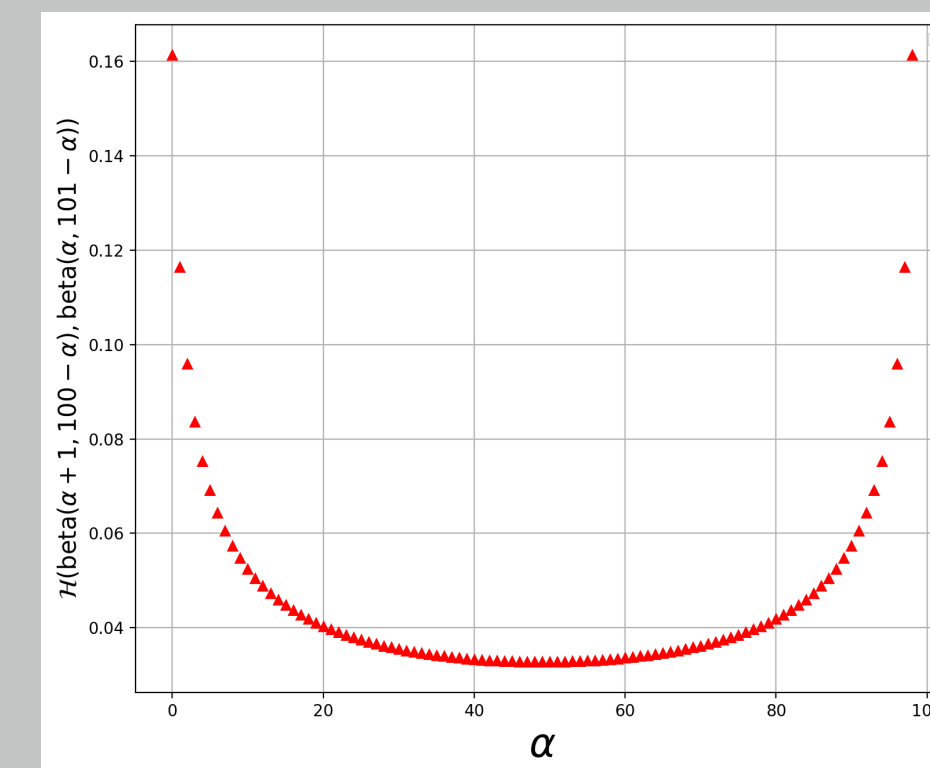


Figure 1: Sensitivities of  $\mathcal{H}$  over  $\text{beta}$

## Smoothed Hellinger Distance Based Exponential Mechanism

We define the mechanism  $\mathcal{M}_{\mathcal{H}}^B$  which produces an element  $r$  in  $\mathcal{R}_{\text{post}}$  with probaibility:

$$\mathbb{P}_{r \sim \mathcal{M}_{\mathcal{H}}^B} = \frac{\exp\left(\frac{-\epsilon \cdot \mathcal{H}(\text{BI}(\mathbf{x}), r)}{2 \cdot S(\mathbf{x})}\right)}{\sum_{r \in \mathcal{R}_{\text{post}}} \exp\left(\frac{-\epsilon \cdot \mathcal{H}(\text{BI}(\mathbf{x}), r)}{2 \cdot S(\mathbf{x})}\right)}$$

given in input an observations  $\mathbf{x}$ , parameters  $\epsilon > 0$  and  $\delta > 0$ , where:

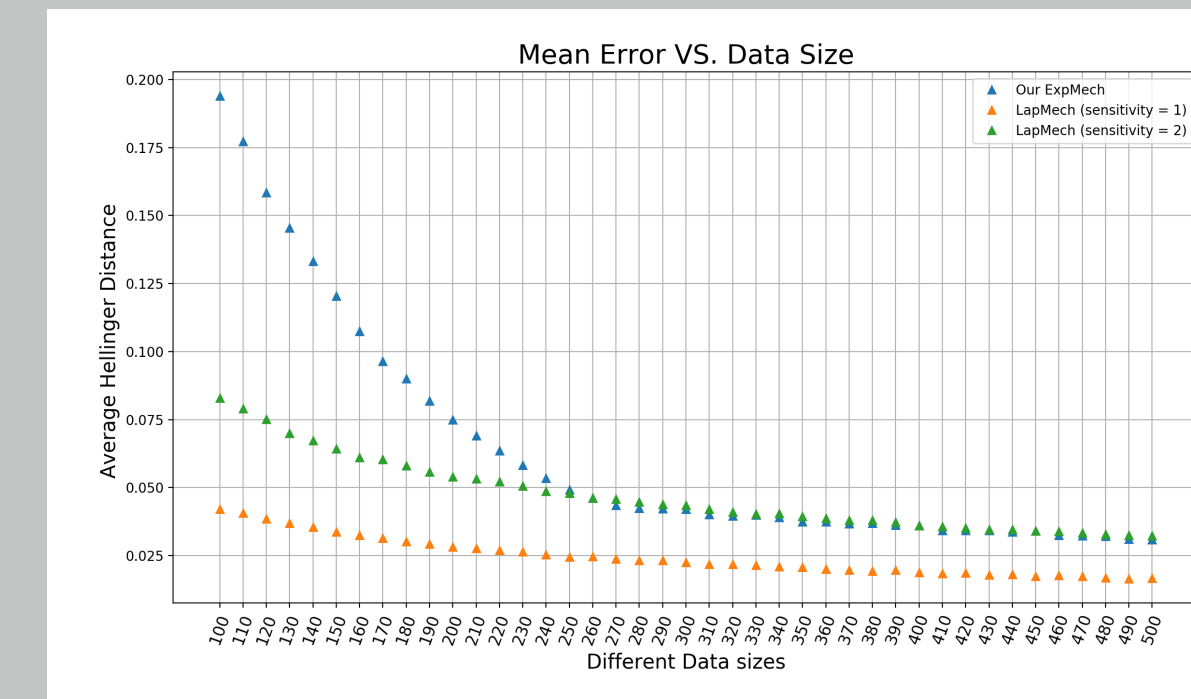
- $\mathcal{R}_{\text{post}}$ , the candidates set defined as  $\{\text{beta}(\alpha', \beta') \mid \alpha' = \alpha + \Delta\alpha, \beta' = \beta + n - \Delta\alpha\}$ , given the prior distribution  $\beta_{\text{prior}} = \text{beta}(\alpha, \beta)$  and observed data set size  $n$ .
- $\mathcal{H}(\text{BI}(\mathbf{x}), r)$  denotes the scoring function based on the Hellinger distance.
- $S(\mathbf{x})$ , the smooth sensitivity[1]:  $S(\mathbf{x}) = \max_{\mathbf{x}' \in \{0,1\}^n} \{LS(\mathbf{x}') \cdot e^{-\gamma \cdot d(\mathbf{x}, \mathbf{x}')}\}$ , where:
  - ▷  $d$ : Hamming distance between two data sets,
  - ▷  $LS(\mathbf{x}')$ , local sensitivity at  $\mathbf{x}'$ :  $LS(\mathbf{x}) = \max_{\mathbf{x}' \in \mathcal{X}^n: \text{adj}(\mathbf{x}, \mathbf{x}'), r \in \mathcal{R}} |\mathcal{H}(\text{BI}(\mathbf{x}), r) - \mathcal{H}(\text{BI}(\mathbf{x}'), r)|$ ,
  - ▷  $\gamma = \ln(1 - \frac{\epsilon}{2 \ln(\frac{\delta}{2(n+1)})})$  to ensure the  $(\epsilon, \delta)$ -differentially private.

## Preliminary Experimental Results

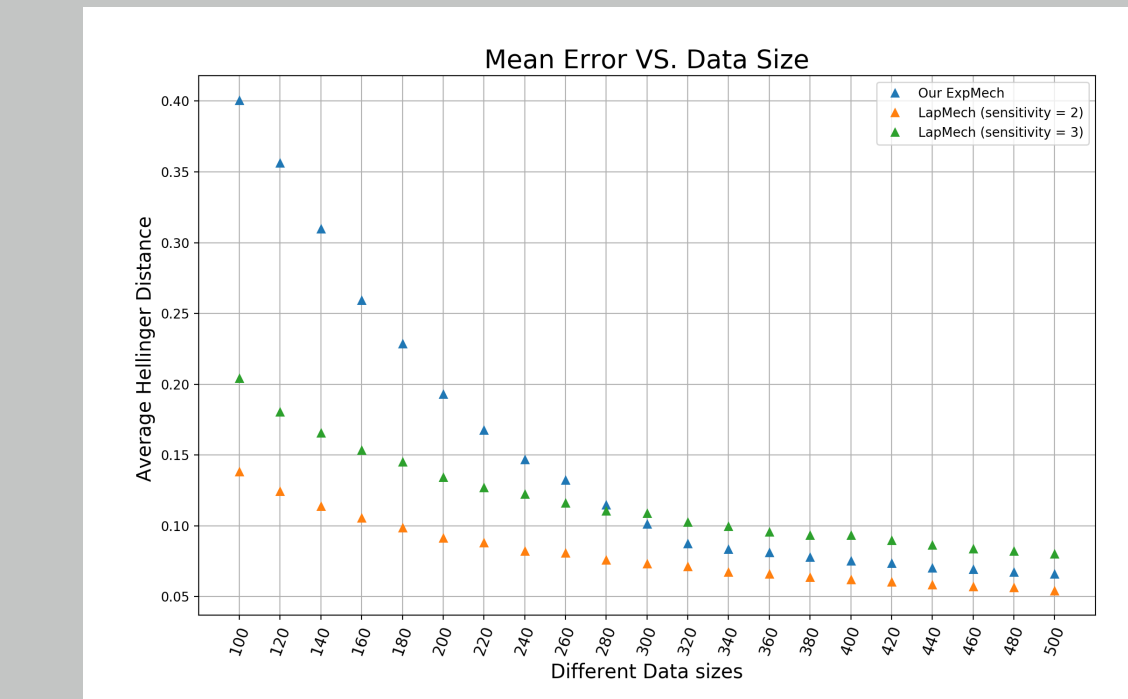
Experiments are on three mechanisms and plotted as follows:

- **Green**: Baseline approach. (noises being postprocessed to floor value)
- **Red**: LapMech with sensitivity **1** in 2 dimensions and **2** in higher dimensions, since it's equivalent to histogram problem (posteriors of adjacent data sets differ only in two dimensions).
- **Blue**:  $\mathcal{M}_{\mathcal{H}}^B$ .

Fig. 2 and Fig. 3(a) give us the average and 4-quantile of Hellinger distance between the sampled results and true posterior, by sampling for **10k** times under each data size or prior configuration.



(a) 2 dimensions, data size  $\in [100, 500]$



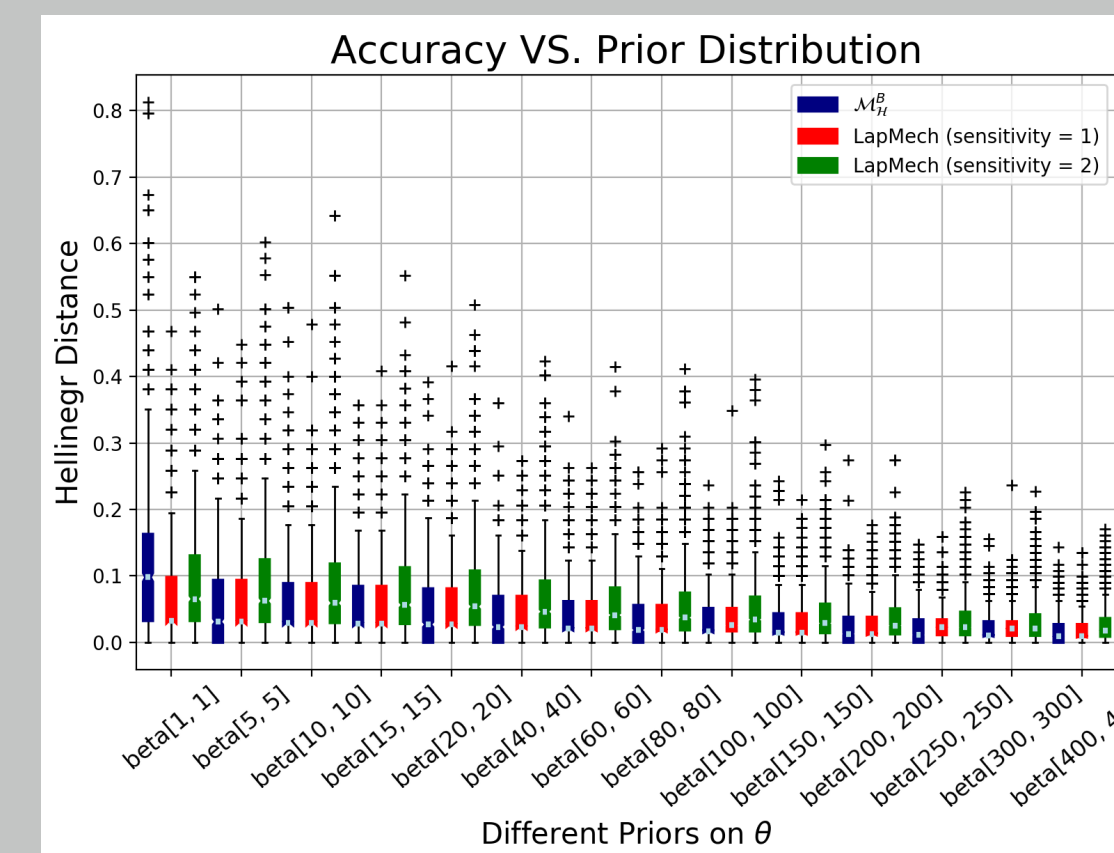
(b) 3 dimensions, data size  $\in [100, 500]$



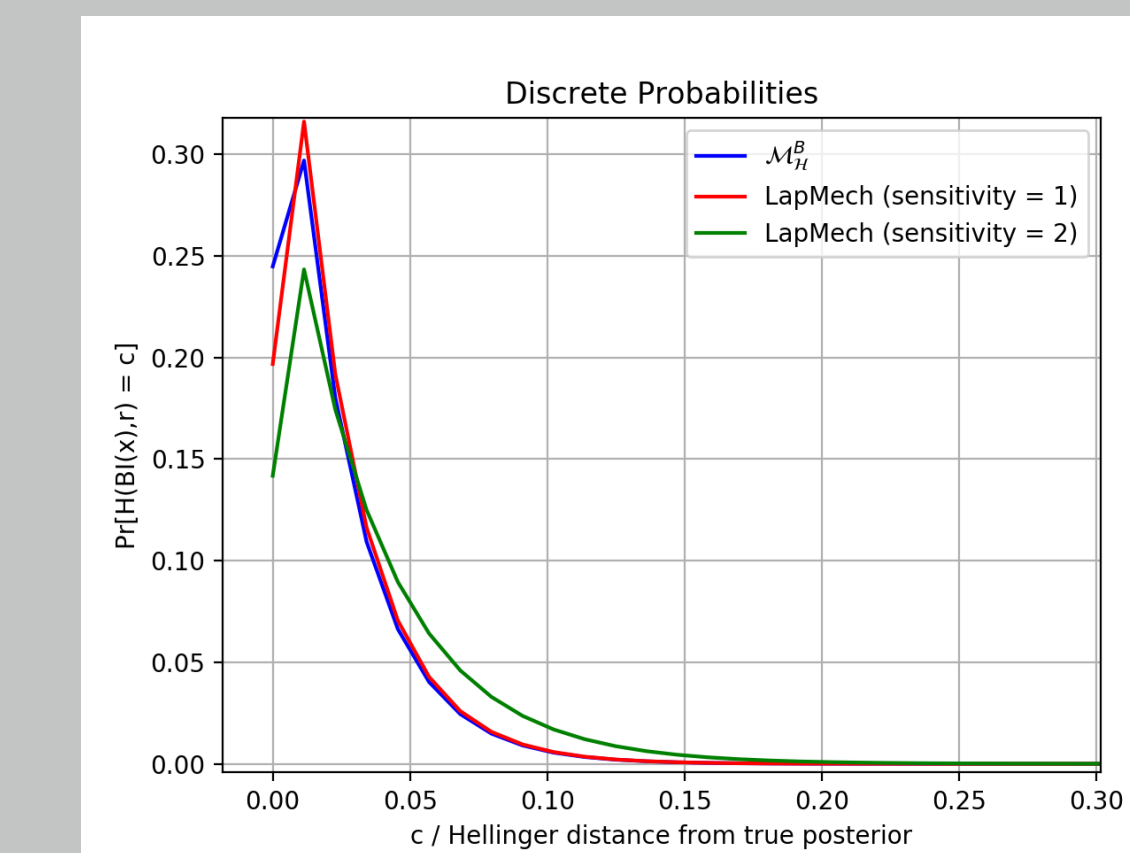
(c) 4 dimensions, data size  $\in [100, 600]$

Figure 2: Average accuracy by increasing data set size

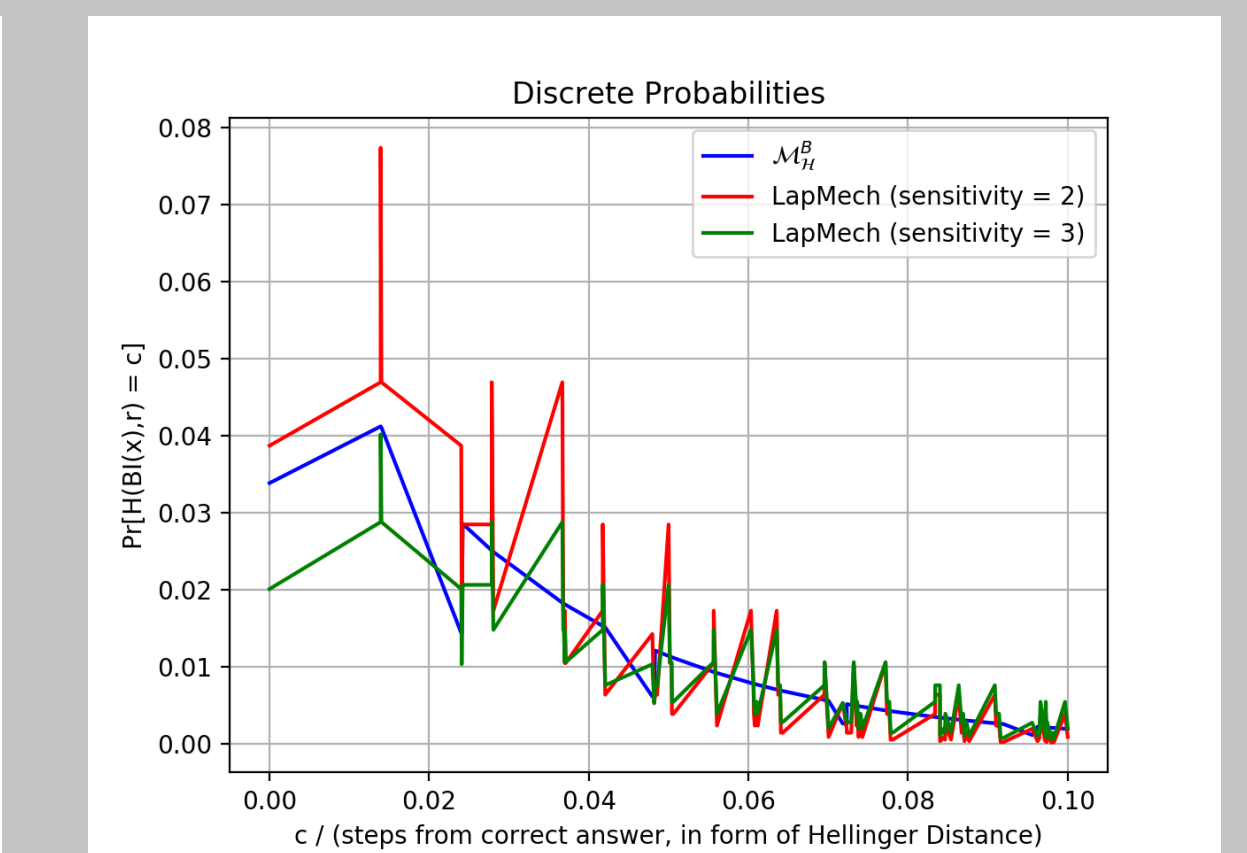
Fig. 3(b) and 3(c) give us the discrete probabilities. On the x-axis: the distance of a potential output  $r$  from the true answer and on y-axis the probability of  $r$  being output by the mechanisms.



(a) 2 dimensions with data size **100**



(b) 2 dimensions with data size **600**



(c) 3 dimensions with data size **600**

Figure 3: 4-quantile and discrete probability plots

Experiments above are with unit prior  $\text{beta}(1, 1)$ ,  $\text{beta}(1, 1, 1)$  and  $\text{beta}(1, 1, 1, 1)$  (except Fig. 3(a)), balanced datasets,  $\epsilon = 1.0$  and  $\delta = 10^{-8}$ .

## Conclusion

- The smoothed Hellinger distance based exponential mechanism outperforms asymptotically the baseline approach when the latter uses a sensitivity proportional to dimensionality.
- Under the same data set size,  $\mathcal{M}_{\mathcal{H}}^B$  can outperform LapMech by increasing the prior.

## References

- [1] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84. ACM, 2007.