# 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

1. Design a differentially private Bayesian inference mechanism.
2. Improve accuracy by calibrating noise to the sensitivity of a metric over distributions (e.g. Hellinger distance ($\mathcal{H}$), $f$-divergences, etc. . . ).

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

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

## Differentially private Bayesian inference and motivations

1. Baseline approach:
   - Release $\mathbf{beta}(\alpha + \lfloor\widetilde{\Delta\alpha}\rfloor_0^n, \beta + n - \lfloor\widetilde{\Delta\alpha}\rfloor_0^n)$,
   - $\widetilde{\Delta\alpha} \sim \mathcal{L}(\Delta\alpha, \frac{S}{\epsilon})$
   - $[[\ S \propto ||\cdot||_1\ ]]$.
   - Measure accuracy with a metric over distributions, e.g. $\mathcal{H}$.

   But $S$ grows linearly with the dimension: too noisy when we generalize to Dirichlet-Multinomial ($\mathbf{DL}(\cdot)$) model.
2. Another approach:
   - Calibrate noise w.r.t *global* sensitivity of $\mathcal{H}$: but global sensitivity is still too big.
   - Fig. 1 shows that there is a gap between global and local sensitivity of $\mathcal{H}$.
3. A better approach:
   - Calibrate noise w.r.t. the *smooth* sensitivity of $\mathcal{H}$.

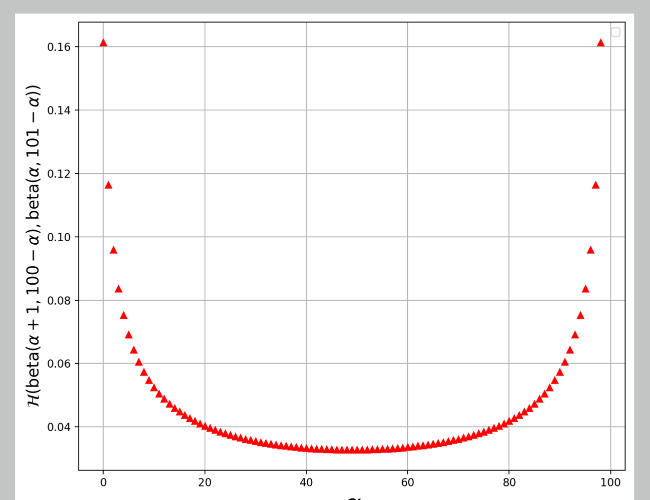



Figure 1: Sensitivity of $\mathcal{H}$. There is a gap between Global and Local sensitivity.

## Our approach: smoothed Hellinger distance based exponential mechanism

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

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

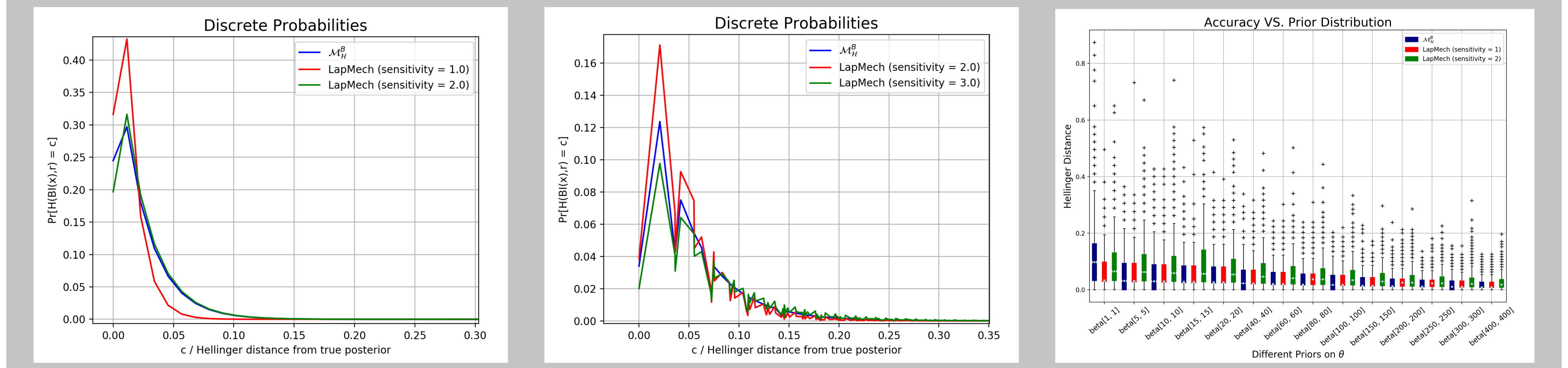
where:

- $\mathcal{R}_{\text{post}} \equiv \{\mathbf{beta}(\alpha', \beta') \mid \alpha' = \alpha + \Delta\alpha, \beta' = \beta + n - \Delta\alpha\}$. With prior distribution $\beta_{\text{prior}} = \mathbf{beta}(\alpha, \beta)$.
- $-\mathcal{H}(\mathbf{BI}(x), r)$ denotes the scoring function.
- $S(x) \equiv \mathbf{max}_{x'\in\{0,1\}^n}\left\{LS(x') \cdot e^{-\gamma\cdot d(x,x')}\right\}$: smooth sensitivity[1], $d$ is the Hamming distance.
- $LS(x') \equiv \max_{y\in\mathcal{X}^n:\mathbf{adj}(y,x'),r\in\mathcal{R}} |\mathcal{H}(\mathbf{BI}(y), r) - \mathcal{H}(\mathbf{BI}(x'), r)|$ is the local sensitivity of $x', \gamma = \ln(1 - \frac{\epsilon}{2\ln(\frac{\delta}{2(n+1)})})$.

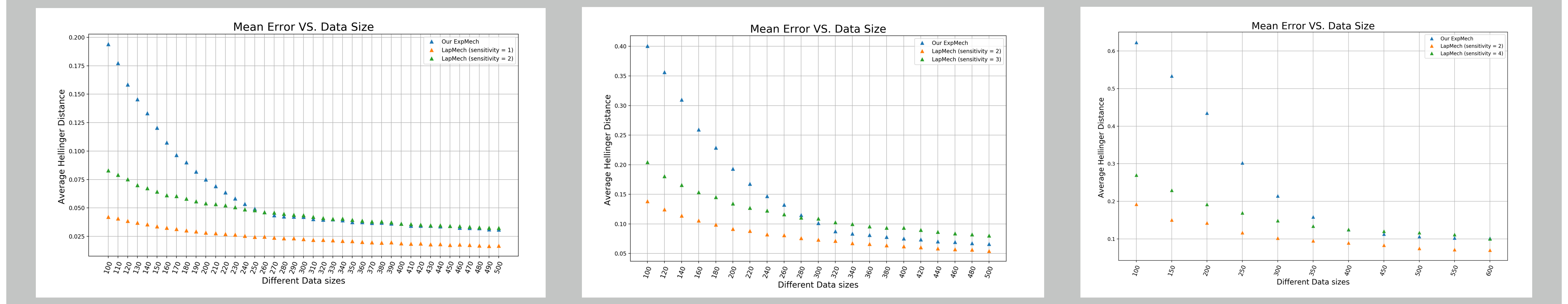
## Preliminary Experimental Results

Experiments are on three mechanisms and plotted as follows:

- **Green**: Baseline approach.
- **Red**: Improved baseline approach 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**: $\mathcal{M}^B_{\mathcal{H}}$.




(a) 2 dimensions with data size **600** (b) 3 dimensions with data size **600** (c) 2 dimensions with data size **100**




(d) 2 dimensions, data size $\in [100, 500]$ (e) 3 dimensions, data size $\in [100, 500]$ (f) 4 dimensions, data size $\in [100, 600]$

Figure 2: Priors are $\mathbf{beta(1, 1)}, \mathbf{DL(1, 1, 1)}$ and $\mathbf{DL(1, 1, 1, 1)}$ (except for Figure 2(c)) , balanced datasets, $\epsilon = 1.0$ and $\delta = 10^{-8}$.

## Conclusion

- The smoothed Hellinger distance based exponential mechanism outperforms asymptotically the baseline approach but not the improved one.
- Under the same data set size, $\mathcal{M}^B_{\mathcal{H}}$ 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.


mbun@cs.princeton.edu, {gianpiet,gaboardi,jliu223}@buffalo.edu