Tailoring Differentially Private Bayesian Inference to Distance Between Distributions

Extended Abstract

ABSTRACT

Bayesian inference is a statistical method which allows one to derive a *posterior* distribution over a parameter, starting from a *prior* distribution and observed data. Settings where sensitive individual information is , different approaches have been taken to make this process differentially private. Dimitrakakis et al. [2], and Wang et al. [6], for instance, proved that, under specific conditions, sampling from the posterior distribution is already differentially private. Other works, e.g., [8], [3], designed differentially private mechanisms that output a representation of the full posterior distribution. Also, accuracy of these mechanisms was measured using metrics on the space of the numeric parameters of the posterior distribution, e.g., the ℓ_1-norm.

In this work we present a new differentially private algorithm to compute the posterior in a Beta-Binomial system and in a Dirichlet-Multinomial system. We apply the exponential mechanism to the smooth sensitivity of a metric on probability distributions; we measured the accuracy of the process by using the same metric. We compared the accuracy of this approach with the ones based on ℓ_1-norm. Experimental results show that when the data size is small or when the parameters of the prior distribution are large, the former outperforms the latter.

KEYWORDS

Differential privacy, Bayesian inference, Hellinger distance

1 AN INFORMAL MOTIVATION

Publishing the posterior distribution inferred from a sensitive dataset can leak information about individuals on the dataset. So, we need to noise the posterior belief in order to protect the data set before releasing it. The question of how much noise to add is usually answered with an ϵ, and a sensitivity value. Sensitivity can be computed in many different ways based on which metric space is imposed on the output set of the mechanism, and yet in literature ([7, 8]) only metrics over the numeric parameters, e.g. ℓ_1 norm, are used. Another approach, which we will explore here, is to impose a metric over probability measures. An orthogonal question is that of how to measure the accuracy. Again, that can be answered in different ways based on the metric imposed on the output space, and yet again only in few works in literature (e.g. [8]) distances over probability measures have been used for these purposes.

The question that this work aims at answering is whether an approach based on probability measures can outperform in term of accuracy approaches based on metric over the numeric parameters of the distributions. We will see that in some cases this can happen.

2 BAYESIAN INFERENCE BACKGROUND

Given a prior belief $\Pr(\theta)$ on some parameter θ, and an observation $\mathbf{x}$, the posterior distribution on θ given $\mathbf{x}$ is computed as:

$$\Pr(\theta|\mathbf{x}) = \frac{\Pr(\mathbf{x}|\theta) \cdot \Pr(\theta)}{\Pr(\mathbf{x})}$$

where the expression $\Pr(\mathbf{x}|\theta)$ denotes the *likelihood* of θ when $\mathbf{x}$ is observed. Since we consider $\mathbf{x}$ to be fixed the likelihood is a function of θ. For the same reason $\Pr(\mathbf{x})$ is a constant independent of θ. Usually in statistics the prior distribution $\Pr(\theta)$ is chosen so that it represents the initial belief on θ. That is, when no data has been observed. In practice though, prior distributions and likelihood functions are usually chosen so that the posterior belongs to the same *family* of distributions. In this case we say that the prior is conjugate of the likelihood function. Using conjugate priors simplify calculcations and allows for inference to be performed in a recursive fashion over the data. In this work we will consider a specific instance of Bayesian inference and one of its generalizations. Specifically, we will consider the situation where θ represents the parameter –informally called *bias*– of a Bernoulli distributed random variable, and its immediate generalization where the parameter $\boldsymbol{\theta}$ represents the the vector of parameters of a categorical distributed random variable. In the former case the prior distribution over $\theta \in [0, 1]$ is going to be a beta distribution, $\text{beta}(\alpha, \beta)$, with parameters $\alpha, \beta \in \mathbb{R}^+$, and with p.d.f:

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

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

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

In the latter case the prior distribution over $\boldsymbol{\theta} \in [0, 1]^k$ is given by a Dirichelet distribution, $\text{DL}(\boldsymbol{\alpha})$, for $k \in \mathbb{N}$, and $\boldsymbol{\alpha} \in (\mathbf{R}^+)^k$, with p.d.f:

$$\Pr(\boldsymbol{\theta}) \equiv \frac{1}{\text{B}(\boldsymbol{\alpha})} \cdot \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

where $\text{B}(\cdot)$ is the generealised beta function. The data $\mathbf{x}$ will be a sequence of $n \in \mathbb{N}$ values coming from a universe $\mathcal{X}$, such that $|\mathcal{X}| = k$. The likelihood function will be:

$$\Pr(\mathbf{x}|\boldsymbol{\theta}) \equiv \prod_{a_i \in \mathcal{X}} \theta_i^{\Delta\alpha_i},$$

with $\Delta\alpha_i = \sum_{j=1}^{n}[x_j = a_i]$, where in $[\cdot]$ we use Iverson bracket notation. Denoting by $\Delta\boldsymbol{\alpha}$ the vector $\langle \Delta\alpha_1, \dots \Delta\alpha_k \rangle$ the posterior distribution over $\boldsymbol{\theta}$ turns out to be

$$\Pr(\boldsymbol{\theta}|\mathbf{x}) = \mathsf{DL}(\boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}).$$

where + denotes the component wise sum of vectors of reals.

3 THE PROBLEM AND BASELINE APPROACH

We are interested in designing a mechanism for privately releasing *fully* the posterior distributions derived in section 2, and not a sample from them. It's worth noticing that the posterior distributions are fully characterized by their parameters, and the family (beta, Dirichlet) they belong to. Hence, in case of the Beta-Binomial model we are interested in releasing a private version of the pair of parameters $(\alpha', \beta') = (\alpha + \Delta\alpha, \beta + n - \Delta\alpha)$, and in the case of the Dirichlet-Multinomial model we are interested in a private version of $\boldsymbol{\alpha}' = (\boldsymbol{\alpha} + \Delta\boldsymbol{\alpha})$. Zhang et al. [8] and Xiao and Xiong [7] have already attacked this problem and their solution consisted in adding independent Laplacian noise to the parameters of the posteriors. That is, in the case of the Beta-Binomial system the value released would be: $(\tilde{\alpha}, \tilde{\beta}) = (\alpha + \Delta\tilde{\alpha}, \beta + n - \Delta\tilde{\alpha})$ where $\Delta\tilde{\alpha} \sim Lap(\Delta\alpha, \frac{2}{\epsilon})$, and where $Lap(\mu, \nu)$ denotes a Laplacian random variable with mean μ and scale ν. This mechanism is ϵ-differentially private, and the noise is calibrated w.r.t. to a sensitivity of 2 which is derived by using ℓ_1 norm over the pair of parameters. Indeed, considering two adjacent[1] data observations $\mathbf{x}, \mathbf{x}'$, that, from a unique prior, give rise to two posterior distributions, characterized by the pairs (α', β') and (α'', β'') then $|\alpha' - \alpha''| + |\beta' - \beta''| \leq 2$. This argument extends similarly to the Dirichelet-Multinomial system.
Also, in previous works, the accuracy has been measured again with respect to ℓ_1 norm. That is, an upper bound has been given to

$$\Pr[|\alpha - \tilde{\alpha}| + |\beta - \tilde{\beta}| \geq \gamma]$$

where $(\alpha, \beta), (\tilde{\alpha}, \tilde{\beta})$ are as defined above. In this work we will use a metric based on a different norm to compute the sensitivity and provide guarantees on the accuracy. In particular we will consider a metric over probability measures and not over the parameters that represent them. Specifically, we will use the Hellinger distance $\mathcal{H}(\cdot, \cdot)$. The choice of Hellinger distance was dictated by two facts, first of all it simplifies calculations in the case of the probabilistic models considered here and second of all it also provides straightforwardly bounds for the total variation distance, which represents also the maximum advantage an unbounded adversary can have in distingishing two distributions. Given two beta distributions $\boldsymbol{\beta}_1 = \text{beta}(\alpha_1, \beta_1)$, and $\boldsymbol{\beta}_2 = \text{beta}(\alpha_2, \beta_2)$ the following equality holds

$$\mathcal{H}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \sqrt{1 - \frac{\mathrm{B}(\frac{\alpha_1+\alpha_2}{2}, \frac{\beta_1+\beta_2}{2})}{\sqrt{\mathrm{B}(\alpha_1, \beta_1)\mathrm{B}(\alpha_2, \beta_2)}}}$$

The same change of metric will be applied to the experimental accuracy guarantees.

[1] Given $\mathbf{x}$, $\mathbf{x}'$ we say that $\mathbf{x}$ and $\mathbf{x}'$ are adjacent and we write, $\mathbf{adj}(\mathbf{x}, \mathbf{x}')$, iff $\sum_i^n [x_i = x'_i] \leq 1$.

4 OUR APPROACH - EXPONENTIAL MECHANISM WITH SMOOTH SENSITIVITY

Given a prior distribution $\boldsymbol{\beta}_{\text{prior}} = \text{beta}(\alpha, \beta)$ and a sequence of n observations $\mathbf{x} \in \{0, 1\}^n$, we define the follwing set:

$$\mathcal{R}_{\text{post}} \equiv \{\text{beta}(\alpha', \beta') \mid \alpha' = \alpha + \Delta\alpha, \beta' = \beta + n - \Delta\alpha\}$$

where $\Delta\alpha$ is as defined in Section 2. Notice that $\mathcal{R}_{\text{post}}$ has $n + 1$ elements, and the Bayesian Inference process will produce an element from $\mathcal{R}_{\text{post}}$ that we denote by $\mathsf{BI}(\mathbf{x})$ – we don't explicitely parametrize the result by the prior, which from now on we consider fixed and we denote it by $\boldsymbol{\beta}_{\text{prior}}$.

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

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

this mechanism is based on the basic exponential mechanism [4], with $\mathcal{R}_{\text{post}}$ as range and $\mathcal{H}(\cdot, \cdot)$ as scoring function. The difference is that in this mechanism we don't calibrate the noise w.r.t. to the global sensitivity of the scoring function but w.r.t. to the smooth sensitivity $S(\mathbf{x})$ – Nissim et al. [5]– of $\mathcal{H}(\mathsf{BI}(\mathbf{x}), \cdot)$. The smooth sensitivity is computed as follows:

$$S(\mathbf{x}) = \max_{\mathbf{x}' \neq \mathbf{x}, \mathbf{x}' \in \{0,1\}^n} \left\{ \Delta_I \left(\mathcal{H}(\mathsf{BI}(\mathbf{x}'), \cdot) \right) \cdot e^{-\gamma \cdot d(\mathbf{x}, \mathbf{x}')} \right\} \quad (1)$$

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 $\Delta_I \left(\mathcal{H}(\mathsf{BI}(\mathbf{x}'), \cdot) \right)$ denotes the local sensitivity at $\mathsf{BI}(\mathbf{x}')$, or equivalently at $\mathbf{x}'$ –w.r.t. ℓ_1 norm–of the scoring function used in our mechanism, that is:

$$\Delta_I \left(\mathcal{H}(\mathsf{BI}(\mathbf{x}'), \cdot) \right) = \max_{\mathbf{x}'' \in \mathcal{X}^n : \mathbf{adj}(\mathbf{x}', \mathbf{x}''), r \in \mathcal{R}_{\text{post}}} |\mathcal{H}(\mathsf{BI}(\mathbf{x}'), r) - \mathcal{H}(\mathsf{BI}(\mathbf{x}''), r)|$$

This mechanism also extends to the Dirichlet-Multinomial system $\mathsf{DL}(\boldsymbol{\alpha})$ by rewriting the Hellinger distance as:

$$\mathcal{H}(\mathsf{DL}(\boldsymbol{\alpha}_1), \mathsf{DL}(\boldsymbol{\alpha}_2)) = \sqrt{1 - \frac{\mathrm{B}(\frac{\boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2}{2})}{\sqrt{\mathrm{B}(\boldsymbol{\alpha}_1)\mathrm{B}(\boldsymbol{\alpha}_2)}}},$$

and by replacing the $\mathcal{R}_{\text{post}}$ with set of posterior Dirichlet distributions candidates. Also, the smooth sensitivity $S(\mathbf{x})$ in (1) will be computed by letting $\mathbf{x}'$ range over all the elements in $\mathcal{X}^n$ adjacent to $\mathbf{x}$. Notice that $\mathcal{R}_{\text{post}}$ has $\binom{n+1}{m-1}$ elements in this case. We will denote by $\mathcal{M}^D_{\mathcal{H}}$ the mechanism for the Dirichlet-Multinomial system.

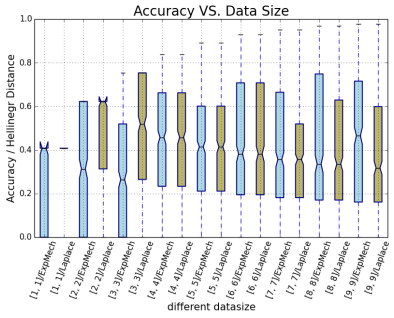
The following result guarantees that the mechanisms are indeed differentially private.

Theorem 4.1 (Privacy). *For $\gamma = \ln(1 - \frac{\epsilon}{2\ln(\frac{\delta}{2(n+1)})})$, both $\mathcal{M}^B_{\mathcal{H}}$ and $\mathcal{M}^D_{\mathcal{H}}$ are (ϵ, δ)-differentially private.*

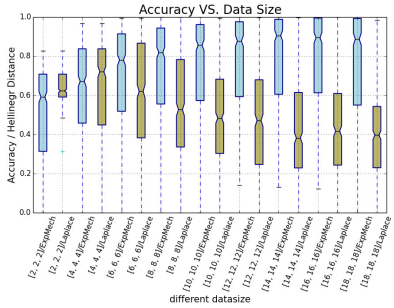
5 PRELIMINARY EXPERIMENTAL RESULTS

In this section, we evaluate the accuracy of the mechanisms defined in Section (4) w.r.t. four variables, including data size, dimensions, data variance, prior distribution, and some combinations thereof. Every plot is an average over 1000 runs. In all the experiments we set $\epsilon = 0.8$, and $\delta = 0.00000001$.

In the following whiskers-plots, the y-axis shows the accuracy (or equivalently, the error) of the mechanisms. Accuracy is measured using Hellinger distance. The x-axis, instead shows one of the previously mentioned independent variables –or a combination thereof. The boxes extend from the lower to the upper quartile values of the data, with a line at the median. A notch on the box around the median is also drawn to give a rough guide to the significance of difference of medians; The whiskers extend from the box to show the range of the data. A blue box in the plots represents the exponential mechanism behaviour, while the yellow box next to it represents the performance of a variation of the basic Laplace mechanism presented in Section (3) with the same settings: that is ϵ, δ, data, prior. The variation considered performs a post processing on the released parameters so that they are consistent. For instance when the sum of the noised parameters is greater than n we will truncate them so that they sum up to n.



(a) two dimensions with beta(1, 1) prior distribution



(b) three dimensions with DL(1, 1, 1) prior distribution

Figure 1: Balanced datasets

Increasing data size with balanced datasets. In Figure 1 we consider *balanced* datasets of observations. This means that in the Beta-Binomial setting the datasets will be half 1s and half 0s, while int the Dirichelet-Multinomial the data will be equally split in all the $k = 3$ bins. Figure 1 shows that when the data size increases, the accuracy of $\mathcal{M}^B_{\mathcal{H}}$ and $\mathcal{M}^D_{\mathcal{H}}$ decreases. In Figure 1(a), when the data size is smaller than 12, $\mathcal{M}^B_{\mathcal{H}}$ can outperform the Laplace mechanism but it fails to do so for bigger datasets. The same happens in as in Figure 1(b): the Laplace mechanism starts performing better for datasets bigger than 15.

Increasing dimensions and data size with balanced dataset. In Figure 2, in the x-axis are observed data sets of different sizes and different dimensions. The plot shows that increasing the number of dimensions have a similar pejorative effect on $\mathcal{M}^B_{\mathcal{H}}$, $\mathcal{M}^D_{\mathcal{H}}$, and the Laplace mechanism. Fixing the number of dimensions and increasing the data size shows that the Laplace mechanism is more accurate then both $\mathcal{M}^B_{\mathcal{H}}$ and $\mathcal{M}^D_{\mathcal{H}}$. In other words, dimensions have little influence on whether our mechanisms will beat the Laplace mechanism.

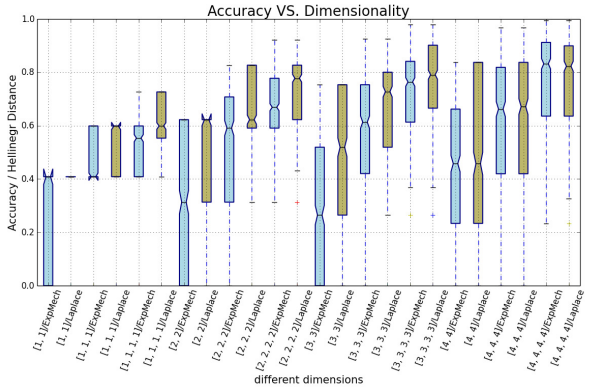


Figure 2: Increasing dimensions and data size with balanced dataset. Priors have 1s in every dimension.

Fixed data size and unbalaced datasets. In Figure 3, in the x-axis we considered different levels of balance of the datasets. We study this variable only under two-dimensions beta distribution in order to be more concise. The plot shows that $\mathcal{M}^B_{\mathcal{H}}$ accuracy is better than the one of Laplace when the dataset is balanced.

Fixed dataset varying balanced priors. In Figure 4, we fix the dataset to be $\langle 5, 5, 5 \rangle$. We also considered balanced priors with increasing values in their dimensions. The plot shows that in the beginning the Laplace mechanism performs better but it is outperformed after a while.

6 CONCLUSIONS AND FUTURE WORKS

From what we have seen in the previous sections we can obtain some preliminary conclusions. That is, the probabiliy measure approach outperforms the ℓ_1-norm approach in the following cases:

(1) When the data size is small
(2) When the observed data is balanced
(3) When priors parameters increase (eventually).

These results although very motivating, are still not enough for real world applications. Hence, we will continue our work in the follwoing directions:

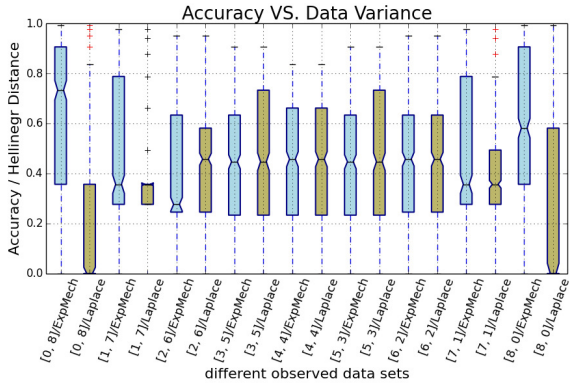


Figure 3: Unbalanced datasets. Prior distributions have 1s in every dimension.

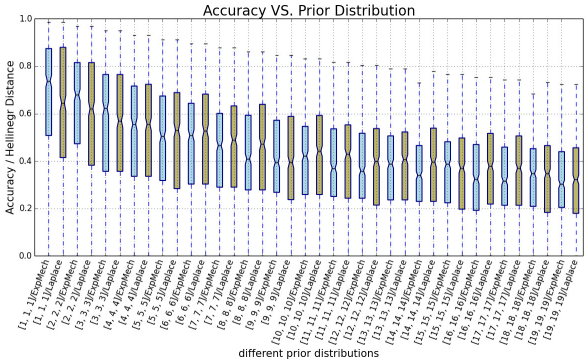


Figure 4: Observed data set is: $\langle 5, 5, 5 \rangle$, varying balanced priors

(1) For now, we just have a intuitive idea on the accuracy behavior of our mechanisms, and not a precise formula or bound on it. When do our mechanisms perform better than the baseline mechanism and when they don't? How much influence will elements in Section 5 have on the accuracy? Are there any other important factors we missed? These are all questions w.r.t. the accuracy that we are going to explore next, and in a more principled and formal way.
(2) Theorem 4.1 provides an upper bound to the privacy loss for $\mathcal{M}^B_{\mathcal{H}}$ and $\mathcal{M}^D_{\mathcal{H}}$ but not necessarily a tight one. Indeed, 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 – that is noise proportional to a higher value of ϵ– 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].

REFERENCES

[1] I. Csiszár and P.C. Shields. 2004. Information Theory and Statistics: A Tutorial. *Foundations and TrendsÂő in Communications and Information Theory* 1, 4 (2004), 417–528. https://doi.org/10.1561/0100000004
[2] Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin IP Rubinstein. 2014. Robust and private Bayesian inference. In *International Conference on Algorithmic Learning Theory*. Springer, 291–305.
[3] James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. 2016. On the theory and practice of privacy-preserving Bayesian data analysis. *arXiv preprint arXiv:1603.07294* (2016).
[4] Frank McSherry and Kunal Talwar. 2007. Mechanism Design via Differential Privacy, In Annual IEEE Symposium on Foundations of Computer Science (FOCS). https://www.microsoft.com/en-us/research/publication/mechanism-design-via-differential-privacy/
[5] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2007. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM, 75–84.
[6] Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. 2015. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*. 2493–2502.
[7] Yonghui Xiao and Li Xiong. 2012. Bayesian inference under differential privacy. *arXiv preprint arXiv:1203.0617* (2012).
[8] Zuhe Zhang, Benjamin IP Rubinstein, Christos Dimitrakakis, et al. 2016. On the Differential Privacy of Bayesian Inference.. In *AAAI*. 2365–2371.