Differential Privacy for Bayesian Inference through Posterior Sampling

Christos Dimitrakakis, Blaine Nelson, Zuhe Zhang, Aikateirni Mitrokotsa, Benjamin Rubinstein

Journal of Machine Learning Research 18 (2017) 1-39 Submitted 5/15; Revised 2/17; Published 3/17

Differential Privacy for Bayesian Inference through Posterior Sampling*

Christos Dimitrakakis CHRISTOS.DIMITRAKAKIS@GMAIL.COM
University of Lille, F-59650 Villeneuve-d'Ascq, France
SEAS, Harvard University, Cambridge MA-02138, USA
DIT, Chalmers University of Technology, SE-412 96, Gothenburg, Sweden

Blaine Nelson BLAINE.NELSON@GOOGLE.COM
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043, USA

Zuhe Zhang ZHANG.ZUHE@GMAIL.COM
School of Mathematics & Statistics
The University of Melbourne
Parkville, VIC 3010, Australia

Aikaterini Mitrokotsa AIKMITR@CHALMERS.SE
Department of Computer Science & Engineering
Chalmers University of Technology
SE-412 96, Gothenburg, Sweden

Benjamin I. P. Rubinstein BRUBINSTEIN@UNIMELB.EDU.AU
School of Computing & Information Systems
The University of Melbourne
Parkville, VIC 3010, Australia

Editor: Charles Elkan

Abstract

Differential privacy formalises privacy-preserving mechanisms that provide access to a database. Can Bayesian inference be used directly to provide private access to data? The answer is yes: under certain conditions on the prior, sampling from the posterior distribution can lead to a desired level of privacy and utility. For a uniform treatment, we define differential privacy over arbitrary data set metrics, outcome spaces and distribution families. This allows us to also deal with non-i.i.d or non-tabular data sets. We then prove bounds on the sensitivity of the posterior to the data, which delivers a measure of robustness. We also show how to use posterior sampling to provide differentially private responses to queries, within a decision-theoretic framework. Finally, we provide bounds on the utility of answers to queries and on the ability of an adversary to distinguish between data sets. The latter are complemented by a novel use of Le Cam's method to obtain lower bounds on distinguishability. Our results hold for arbitrary metrics, including those for the common definition of differential privacy. For specific choices of the metric, we give a number of examples satisfying our assumptions.

*. A preliminary version of this paper appeared in *Algorithmic Learning Theory 2014* (Dimitrakakis et al., 2014). This version corrects proofs, constant factors in the upper bounds and introduces new material on utility analysis, lower bounds and examples.

Keywords: Bayesian inference, differential privacy, robustness, adversarial Learning

1. Introduction

The Bayesian framework for statistical decision theory incorporates uncertainty into decision making in a probabilistic manner. This makes the framework attractive, as predictions and modelling can all be made with the machinery of probability. More specifically, a Bayesian statistician begins by assuming that the world is described by a probabilistic model within some family, and he assigns a prior belief to each one of the models. After observing data, this belief is adjusted through Bayes's theorem to the so called posterior belief. This expresses the statistician's conclusion given the data and the prior assumptions. The statistician can then release the posterior to the world, for others to build upon, or use for principled decision making under uncertainty.

Unfortunately, it is frequently the case that the data acquired by the statistician is sensitive. Consequently, there is a fear that any information released by the statistician that depends on the data—be that the posterior distribution itself or any decisions that follow from the calculated posterior—may reveal sensitive information in the original data. Recently, the framework of differential privacy has been proposed to codify this leakage of information. If an algorithm is differentially private, then its output can only leak a bounded amount of information about its input.

We are interested in the question of how one can build differentially-private algorithms within the Bayesian framework. More precisely, we examine when the choice of prior is sufficient to guarantee differential privacy for decisions that are derived from the posterior distribution. Our work develops a unified understanding of privacy and learning in adversarial environments, under a decision-theoretic framework. We show that under suitable assumptions, standard Bayesian inference and posterior sampling can achieve uniformly good utility with a fixed privacy budget in the differential privacy setting. We also indicate strong connections between robustness and privacy. Under the base level of data privacy provided by the posterior distribution, the statistician can safely respond to external queries using samples from the posterior. When estimating a linear model from sensitive data, for example, samples from the posterior correspond to different possible fits. The more samples used, the more privacy is leaked, while query responses may be more accurate.

Our proposed approach complements existing mechanisms rather well, and may be particularly useful in situations where Bayesian inference is already in use. For this reason, we provide illustrative examples in the exponential family. However, our setting is wholly general and not limited to specific distribution families, or i.i.d. observations. Any family could be chosen: so long as it either satisfies our assumptions directly, or can be restricted so that it does. For example, our framework applies to families of discrete Bayesian networks with directed-acyclic topologies (*e.g.*, Markov chains; see Lemma 24 on page 21) and multivariate Gaussians (see Lemma 23), where the observations may not satisfy the i.i.d. assumption.

Summary of setting. A Bayesian statistician ($\mathscr{B}$) wishes to communicate results about data x to a third party ($\mathscr{A}$), but without revealing the data x itself. We make no assumptions on the data x, which could be a single observation, an i.i.d. sample, or a sequence of observations. The protocol of interaction between $\mathscr{B}$ and $\mathscr{A}$ is summarised below.

1. $\mathscr{B}$ selects a model family ($\mathcal{F}_\Theta$) and a prior (ξ).
2. $\mathscr{A}$ is allowed to see $\mathcal{F}_\Theta$ and ξ and is computationally unbounded.
3. $\mathscr{B}$ observes data x and calculates the posterior $\xi(\theta \mid x)$ but does not reveal it. Then, for steps $t = 1, 2, \ldots$, repeat the following:
4. $\mathscr{A}$ sends his utility function u and a query q_t to $\mathscr{B}$.
5. $\mathscr{B}$ responds with the response r_t maximising u, in a manner that depends on the query and the posterior.

Let us now elaborate. In this framework, the choice of the model family $\mathcal{F}_\Theta$ is dictated by the problem. The choice of ξ is normally determined by the prior knowledge of $\mathscr{B}$, but we show that this also affects what level of privacy is achieved. Informally speaking, informative priors achieve better privacy, as the posterior has a weaker dependency on the data. It is natural to assume that the prior itself is public, as it should reflect publicly available information. The statistician's conclusion from the observed data x is then summarised in the posterior distribution $\xi(\theta \mid x)$, which remains private.

The second part of the process is the interaction with $\mathscr{A}$. We adopt a decision-theoretic viewpoint to characterise what the optimal responses to queries should be. More specifically, we assume the existence of a "true" parameter $\theta \in \Theta$, and that $\mathscr{A}$ has a utility function $u_\theta(q_t, r_t)$, which he wishes to maximise. For example, consider the case where $\theta = (\mu, \Sigma)$ are the parameters of a normal distribution. An example query q_t is *"what is the expected value $\mathbb{E}_\theta\, x_i = \mu$ of the distribution?"*. The optimal response r_t, would then be a real vector that depends on the utility function. A possible utility function is the negative squared L_2 distance:

$$u_\theta(q_t = \text{"what is the mean?"}, r_t) = -\| \mathbb{E}_\theta\, x_i - r_t \|_2^2.$$

While θ is unknown, $\mathscr{B}$ has information about it in the form of a posterior distribution. Using standard decision-theoretic notions, the optimal response of $\mathscr{B}$ would maximise the expected utility $\mathbb{E}_\xi(u \mid q_t, r_t, x)$, where the expectation is taken over the posterior distribution. However, this deterministic response cannot be differentially private.

In this paper, we promote the use of *posterior sampling* to respond to queries. The posterior sampling mechanism draws a set $\hat{\Theta}$ of i.i.d. samples from the posterior distribution. Then, all the responses only depend on the posterior through $\hat{\Theta}$. Since our algorithm only takes a single sample set $\hat{\Theta}$, further queries by the adversary reveal nothing more about the data than what can be inferred from $\hat{\Theta}$. The empirical distribution induced by $\hat{\Theta}$ serves as a private surrogate for the exact (non-private) posterior. Consequently, we can respond to an arbitrary number of queries with a bounded privacy budget, while guaranteeing good utility for all responses.

We show that if $\mathcal{F}_\Theta$ and ξ are chosen appropriately, this results in differentially-private responses, as well as robustness of the posterior.[1] In addition, we prove upper and lower bounds on how easy it is for an adversary to distinguish two ϵ-close data sets. Finally, we bound the loss in utility incurred due to privacy. The intuition behind our results

1. More specifically, that small changes in the data result in small changes in the posterior in terms of the KL divergence.

is that robustness and privacy are linked via smoothness. Learning algorithms that are smooth mappings—their output (*e.g.*, a spam filter) varies little with perturbations to input (*e.g.*, similar training corpora)—are robust: outliers have reduced influence, and adversaries cannot easily discover unknown information about the data. This suggests that robustness and privacy can be simultaneously achieved and are in fact deeply linked.

We provide a uniform mathematical treatment of the privacy and robustness properties of Bayesian inference based on generalised differential privacy to arbitrary data set distances, outcome spaces, and distribution families. This paper can be summarised as making the following distinct contributions:

- Under certain regularity conditions on the prior distribution ξ or likelihood family $\mathcal{F}_\Theta$, we show that the posterior distribution is *robust*: small changes in the data set result in small posterior changes.
- We introduce a novel *posterior sampling mechanism* that is private.[2] Unlike other common mechanisms in differential privacy, our approach sits squarely in the non-private (Bayesian) learning framework without modification.
- We provide necessary and sufficient conditions for differentially private Bayesian inference.
- We introduce the notion of *data set distinguishability* for which we provide finite-sample bounds for our mechanism: how large would $\hat{\Theta}$ need to be for $\mathscr{A}$ to distinguish two data sets with high probability.
- We provide examples of conjugate-pair distributions where our assumptions hold, including discrete Bayesian networks. We find that even though Bayesian posterior sampling does provide privacy guarantees directly, those appear to be very weak for standard conjugate families. However, with a small modification of the prior, it is easy to obtain good privacy guarantees.

Paper organisation. Section 2 specifies the setting and our assumptions. Section 3 proves results on robustness of Bayesian learning. Section 4 proves our main privacy results. In particular, Section 4.1 shows that the posterior distribution is differentially private, Section 4.2 describes our posterior sampling query response algorithm, Section 4.3 derives bounds on data set indistinguishability, Section 4.5 shows how to obtain matching lower bounds for distinguishability, while Section 4.4 shows how utility and privacy can be traded off within our framework. Examples where our assumptions hold are given in Section 5. We present a discussion of our results, related work and links to the exponential mechanism and robust Bayesian inference in Section 6. Appendix A contains proofs of the main theorems. Finally, Appendix B details proofs of the examples demonstrating our assumptions.

2. Problem Setting

We consider the problem of a Bayesian statistician ($\mathscr{B}$) communicating with an untrusted third party ($\mathscr{A}$). $\mathscr{B}$ wants to convey useful responses to the queries of $\mathscr{A}$ (*e.g.*, how

2. Although previously used *e.g.*, for efficient exploration in reinforcement learning (Thompson, 1933; Osband et al., 2013), posterior sampling has not previously been employed for privacy.

many people suffer from a disease or vote for a particular party) without revealing private information about the original data (*e.g.*, whether a particular person has cancer). This requires communicating information in a way that strikes a balance between utility and privacy. In this paper, we study the inherent privacy and robustness properties of Bayesian inference and explore the question of whether $\mathscr{B}$ can select a prior distribution so that a computationally unbounded $\mathscr{A}$ cannot obtain private information from queries.

2.1 Definitions and Notation

We begin with our notation. Let $\mathcal{S}$ be the set of all possible data sets. For example, if $\mathcal{X}$ is a finite alphabet, then we might have $\mathcal{S} = \bigcup_{n=0}^{\infty} \mathcal{X}^n$, *i.e.*, the set of all possible observation sequences over $\mathcal{X}$. However, $\mathcal{S}$ can have arbitrary structure and so social network or mobility trace data are also handled in this framework. Probability measures on parameters θ are usually denoted by ξ, while measures and densities on data are denoted by P_θ or p_θ respectively. Expectations are denoted by $\mathbb{E}_\xi\, g \triangleq \int_\Theta g(\theta)\, \mathrm{d}\xi(\theta)$, where the subscript denotes the underlying distribution with respect to which we are taking expectations. In case of ambiguity, we explicitly write *e.g.*, $\mathbb{E}_{x\sim P_\theta}\, f(x) = \int_{\mathcal{S}} f(x)\, \mathrm{d}P_\theta(x)$ to denote which variables are drawn from which distributions. Finally, we use $\mathbb{I}\{\pi\}$ to be the identity function, taking the value 1 when the predicate π is true, and 0 otherwise.

2.1.1 Distances Between Data Sets

Central to the notions of privacy and robustness, is the concept of distance between data sets. Firstly, the effect of data set perturbation on learning depends on the amount of noise as quantified by some distance. This is useful for characterising robustness to noise or adversarial manipulation of the data. Secondly, the amount that an attacker can learn from queries can be quantified in terms of the distance of his guesses to the true data set. Finally, it allows for a unified mathematical treatment, as it permits different types of neighbourhoods to be defined. To model these situations, we equip $\mathcal{S}$ with a pseudo-metric[3] $\rho : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$. This generalisation has also been used by Chatzikokolakis et al. (2013), which has laid the groundwork for metric-based differential privacy. While this concept has many applications in the context of geographical information systems, we apply this generalisation of differential privacy without necessarily referring to some underlying physical distance.

2.1.2 Bayesian Inference

This paper focuses on the *Bayesian inference* setting, where the statistician $\mathscr{B}$ constructs a posterior distribution from a prior distribution ξ and a training data set x. More precisely, we assume that data $x \in \mathcal{S}$ have been drawn from some distribution $P_{\theta^\star}$ on $\mathcal{S}$, parameterised by $\theta^\star$, from a family of distributions $\mathcal{F}_\Theta$. $\mathscr{B}$ defines a parameter set Θ indexing a family of distributions $\mathcal{F}_\Theta$ on $(\mathcal{S}, \mathfrak{S}_\mathcal{S})$, where $\mathfrak{S}_\mathcal{S}$ is an appropriate σ-algebra on $\mathcal{S}$:

$$\mathcal{F}_\Theta \triangleq \{\, P_\theta : \theta \in \Theta \,\},$$

3. Meaning that $\rho(x, y) = 0$ does not necessarily imply $x = y$.

and where we use p_θ to denote the corresponding densities[4] when necessary. To perform inference in the Bayesian setting, $\mathscr{B}$ selects a prior measure ξ on $(\Theta, \mathfrak{S}_\Theta)$ reflecting $\mathscr{B}$'s subjective beliefs about which θ is more likely to be true, *a priori*; *i.e.*, for any measurable set $B \in \mathfrak{S}_\Theta$, $\xi(B)$ represents $\mathscr{B}$'s prior belief that $\theta^\star \in B$. In general, the posterior distribution after observing $x \in \mathcal{S}$ is:

$$\xi(B \mid x) = \frac{\int_B p_\theta(x) \, \mathrm{d}\xi(\theta)}{\phi(x)} \ , \tag{1}$$

where ϕ is the corresponding marginal density given by:

$$\phi(x) \triangleq \int_\Theta p_\theta(x) \, \mathrm{d}\xi(\theta) \ .$$

While the choice of the prior is generally arbitrary, this paper shows that its careful selection can yield good privacy guarantees. Throughout the paper, we shall use the following simple example to ground our observations and theory. This consists of a finite family of distributions, on a finite alphabet. Consequently, calculation of the posterior distribution is always simple. It is also easy to verify our assumptions on this model.

Example 1 (Finite Bernoulli family.) *Consider a finite family of distributions $\mathcal{F}_\Theta = \{ P_\theta : \theta \in \Theta \}$ on alphabet $\mathcal{X} = \{ 0, 1 \}$, with $\theta \in [0, 1]$, such that for any model in the family and any observation x*

$$P_\theta(x) = \theta \, \mathbb{I}\{x = 1\} + (1 - \theta) \, \mathbb{I}\{x = 0\} .$$

For any sequence of observations $x_1, \ldots, x_T$, we have, with some abuse of notation,

$$P_\theta(x_1, \ldots, x_T) = \prod_{t=1}^{T} P_\theta(x_t),$$

i.e., *P_θ defines an i.i.d. distribution on the alphabet. This family corresponds to a set of Bernoulli models. The set of parameters Θ will be chosen to discretise the parameter space of Bernoulli distributions over Δ-sized intervals. For this, the k-th model's parameter will be $\theta_k = \Delta k$, with $\Delta \in (0, 1)$ and $k \leq 1/\Delta$.*

For the above family, we can use a uniform prior distribution $\xi(\theta_k) = \Delta$. The posterior distribution is easily calculated, since we need only sum over a finite number of parameters.

2.1.3 Privacy

We now recall the concept of differential privacy (Dwork, 2006). This states that on *neighbouring* data sets, a randomised query response mechanism yields (pointwise) similar distributions. We adopt the view of mechanisms as conditional distributions under which differential privacy can be seen as a measure of smoothness. In our setting, conditional distributions conveniently correspond to posterior distributions. These can also be interpreted as the distribution of a mechanism that uses posterior sampling, to be introduced in Section 4.2. The precise definition depends on the notion of neighbourhood, with the following choice being common:

4. *I.e.,* the Radon-Nikodym derivative of P_θ relative to some dominating measure ν.

Definition 1 ((ϵ, δ)-differential privacy) *A conditional distribution $P(\cdot \mid x)$ on $(\Theta, \mathfrak{S}_\Theta)$ is* (ϵ, δ)-differentially private *if, for all $B \in \mathfrak{S}_\Theta$ and for any $x \in \mathcal{S} = \mathcal{X}^n$*

$$P(B \mid x) \leq e^\epsilon P(B \mid y) + \delta,$$

for all y in the Hamming-1 neighbourhood *of x. That is, y may differ in at most one entry from x: there is at most one $i \in \{1, \ldots, n\}$ such that $x_i \neq y_i$.*

A typical situation where this definition is employed, is when x, y are matrices and x_i is a single row in the matrix. Then, the data sets are neighbours if a matrix row is changed.[5]

In our setting, it is reasonable to generalise this to arbitrary data set spaces $\mathcal{S}$ that are not necessarily product spaces. To do so, we use the notion of differential privacy under a pseudo-metric ρ on the space of all data sets, which allows for more subtle representations of attacker knowledge and for a more general treatment:

Definition 2 ((ϵ, δ)-differential privacy under ρ.) *A conditional distribution $P(\cdot \mid x)$ on $(\Theta, \mathfrak{S}_\Theta)$ is* (ϵ, δ)-differentially private under *a pseudo-metric $\rho : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$ if, for all $B \in \mathfrak{S}_\Theta$ and for any $x, y \in \mathcal{S}$,*

$$P(B \mid x) \leq e^{\epsilon \rho(x,y)} P(B \mid y) + \delta \rho(x, y) \ .$$

In our setting, ρ replaces the notion of neighbourhood. It is of course possible to use ρ that corresponds to the usual meaning of neighbourhood in differential privacy:

Remark 3 *If $\mathcal{S} = \mathcal{X}^n$ and $\rho(x, y) = \sum_{i=1}^n \mathbb{I}\{x_i \neq y_i\}$ is the Hamming distance, this definition is analogous to standard (ϵ, δ)-differential privacy. When considering only $(\epsilon, 0)$-differential privacy or $(0, \delta)$-privacy, it is an equivalent notion.*[6]

Proof For $(\epsilon, 0)$-DP, let $\rho(x, z) = \rho(z, y) = 1$; *i.e.*, the data differ in one element. Then, from standard DP, we have $P(B \mid x) \leq e^\epsilon P(B \mid z)$ and so obtain $P(B \mid x) \leq e^{2\epsilon} P(B \mid y) = e^{\rho(x,y)\epsilon} P(B \mid y)$. By induction, this holds for any x, y pair. Similarly, for $(0, \delta)$-DP, by induction we obtain $P(B \mid x) \leq P(B \mid y) + \delta \rho(x, y)$. ■

Definition 1 allows for privacy against a powerful attacker $\mathscr{A}$, who attempts to match the empirical distribution induced by the true data set, by querying the learned mechanism and comparing its responses to those given by distributions simulated using knowledge of the mechanism and knowledge of all but one datum—narrowing the data set down to a Hamming-1 ball. Indeed the requirement of differential privacy is sometimes *too strong* since it may come at the price of utility. Definition 2 allows for a much broader encoding of the attacker's knowledge via the selected pseudo-metric. It also allows a more fine-grained notion of privacy. This is quite useful for geographical information systems, as proposed by Chatzikokolakis et al. (2013), to which we refer the reader for a broader discussion of the use of metrics in differential privacy.

Finally, we can show that this generalisation of differential privacy satisfies the standard composition property.

5. Another common choice for neighbourhoods is to say that two data sets are neighbours if one results from the other by addition of a row.
6. Making the definition wholly equivalent is possible, but results in an unnecessarily complex definition.

Theorem 4 (Composition) *Let conditional distributions $P(\cdot \mid x)$ on $(\Theta, \mathfrak{S}_\Theta)$ be (ϵ, δ)-*differentially private under *a pseudo-metric $\rho : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$ and $P'(\cdot \mid x)$ on $(\Theta', \mathfrak{S}'_{\Theta'})$ be (ϵ', δ')-*differentially private under *the same pseudo-metric. Then the conditional distribution on the product space $(\Theta \times \Theta', \mathfrak{S}_\Theta \otimes \mathfrak{S}'_{\Theta'})$ given by*

$$Q(B \times B' \mid x) = P(B \mid x)P'(B' \mid x), \forall B \times B' \in \mathfrak{S}_\Theta \otimes \mathfrak{S}'_{\Theta'}$$

*satisfies $(\epsilon + \epsilon', \delta + \delta')$-*differentially private under *the pseudo-metric ρ. Here $\mathfrak{S}_\Theta \otimes \mathfrak{S}'_{\Theta'}$ is the product σ-algebra on $\Theta \times \Theta'$.*

Proof For any $y \in \mathcal{S}$

$$\begin{aligned} Q(B \times B' \mid x) &\leq \left[e^{\epsilon\rho(x,y)}P(B \mid y) + \delta\rho(x,y)\right] P'(B' \mid x) \\ &\leq e^{\epsilon\rho(x,y)}P(B \mid y)\left[e^{\epsilon'\rho(x,y)}P'(B' \mid y) + \delta'\rho(x,y)\right] + \delta\rho(x,y) \\ &\leq e^{(\epsilon+\epsilon')\rho(x,y)}P(B \mid y)P'(B' \mid y) + (\delta + \delta')\rho(x,y) \end{aligned}$$

■

2.2 Our Main Assumptions

In the sequel, we show that if the distribution family $\mathcal{F}_\Theta$ or prior ξ satisfies certain assumptions, then close data sets $x, y \in \mathcal{S}$ result in posterior distributions that are close. In that case, it is difficult for a third party to use such a posterior to distinguish the true data set x from similar data sets.

To formalise these notions, we introduce two possible assumptions one could make on the smoothness of the family $\mathcal{F}_\Theta$ with respect to some metric d on $\mathbb{R}_+$. The first assumption states that the likelihood is smooth for all parameterisations of the family. First, we define our notion of smoothness. Let $f(x, \theta) \triangleq \ln p_\theta(x)$ be the log probability of x under θ. The Lipschitz constant for a parameter value θ is:

$$\ell(\theta) \triangleq \inf \{\, u : |f(x,\theta) - f(y,\theta)| \leq u\rho(x,y) \forall x, y \in \mathcal{S} \,\}. \tag{2}$$

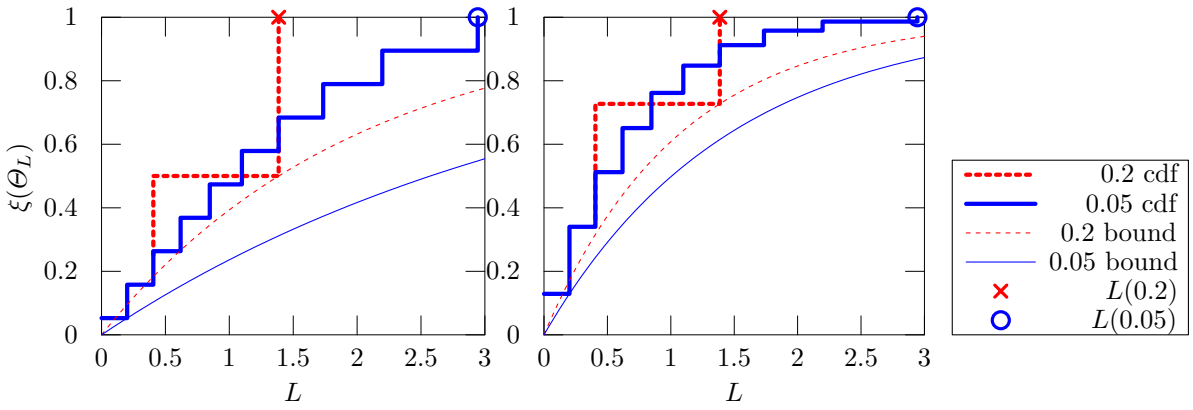
Our first assumption is uniform smoothness for all parameters.

Assumption 1 (Lipschitz continuity) *We assume there exists some $L < \infty$ such that:*

$$\ell(\theta) \leq L, \qquad \theta \in \Theta. \tag{3}$$

In other words, this assumption says that the log probability is Lipschitz with respect to ρ for any parameter value. Consider Example 1 for the Bernoulli model. It is easy to see that a model with Δ-sized intervals satisfies the above assumption with $L = \ln 1/\Delta$.

However, it may be difficult for this assumption to hold uniformly over Θ in general. This can be seen by the following counterexample for the Bernoulli family of distributions: when the parameter is 0, then any sequence $x = 0, 0, \ldots$ has probability 1, while any



(a) Uniform prior (b) Exponential prior

Figure 1: The mass of L-Lipschitz parameters, for two finite familes of Bernoulli distributions with $\Delta \in \{0.2, 0.05\}$ (thick lines) together with their respective stochastic Lipschitz bounds (thin lines) and the corresponding uniform Lipschitz constant L.

sequence containing a 1 has probability 0. The same thing occurs when we take $\Delta \to 0$ in Example 1. To avoid such problems, we relax the assumption by only requiring that $\mathscr{B}$'s *prior* probability ξ is concentrated in the regions of the family for which the likelihood is smoothest:

Assumption 2 (Stochastic Lipschitz continuity; Norkin, 1986) *First, define the subset of parameter values*

$$\Theta_L \triangleq \{\theta \in \Theta : \ell(\theta) \leq L\} \tag{4}$$

to be those parameters for which Lipschitz continuity holds with Lipschitz constant L. Then, there are some constants $c, L_0 > 0$ such that, for all $L \geq L_0$:

$$\xi(\Theta_L) \geq 1 - \exp(-c(L - L_0)) \ . \tag{5}$$

By not requiring uniform smoothness, this weaker assumption is easier to meet but still yields useful guarantees. In fact, in Section 5, we demonstrate that this assumption is satisfied by many important example distribution families. However, it will be illustrative to consider the discrete Bernoulli family example at this point.

Example 2 (Continuation of Example 1) *These conditions can be examined in terms of the finite family of Example 1. Figure 1 demonstrates the assumptions for $\Delta = 0.2$ (red dashed lines) and $\Delta = 0.05$ (blue solid lines).*

In particular, the two thick lines Figure 1a show the probability mass of L-Lipschitz parameters for the two families. They are both step functions, as the families are discrete.[7]

7. The $\Delta = 0.2$ family only has two steps, as the Lipschitz constant is symmetric about $\theta = 0.5$.

The $\times$ and $\circ$ symbols show the corresponding Lipschitz constants for the two families respectively, and we can clearly see $\Delta = 0.2$ has about half the Lipschitz constant of $\Delta = 0.05$. The thinner curves depict the highest lower bound on the probability mass defined in Assumption 2. There we see that the higher curve is achieved by $\Delta = 0.2$.

In order to improve the lower bound, we need to modify our prior distribution on the family members so as to place less mass on the more sensitive parameters. The result of this operation is shown in Figure 1b, which uses the prior $\xi(\theta) \propto \exp(-\ell(\theta))$, i.e., it places exponentially smaller weight in more sensitive parameters. This results in both lower bounds being shifted upwards, corresponding to a higher c constant in Assumption 2. Of course, this has no effect on Assumption 1.

For completeness, we now show that verifying our assumptions for a distribution of a single random variable lifts to a corresponding property for the product distribution on i.i.d. samples.

Lemma 5 *If $\mathcal{F}_\Theta$ satisfies Assumption 1 (resp. Assumption 2) with respect to pseudo-metric ρ and constant L (or c), then, for any fixed $n \in \mathbb{N}$, the product family $\mathcal{F}_\Theta^n$ with densities (sim. measures) $p_\Theta^n(\{ x_i \}) = \prod_{i=1}^n p_\Theta(x_i)$ satisfies the same assumption with respect to:*

$$\rho^n(\{ x_i \}, \{ y_i \}) = \sum_{i=1}^n \rho(x_i, y_i)$$

and constant L (or c).

2.2.1 Necessary Conditions

Finally, let us discuss whether the above conditions are necessary to achieve differential privacy. In fact, either the first condition must be true, or a similar condition must hold on the marginals for every possible data set pair (x, y). Our second condition can be seen as a specific case of the necessary condition for the marginals, as explained below.

Theorem 6 *For a prior ξ to be $2L$-differentially private for a family $\mathcal{F}_\Theta$, either*

$$\sup_{\theta \in \Theta} \ln \frac{P_\theta(x)}{P_\theta(y)} \leq L\rho(x, y), \qquad \text{or} \qquad \ln \frac{\phi(y)}{\phi(x)} \leq L\rho(x, y) \tag{6}$$

for all $x, y \in \mathcal{X}$.

Proof If neither condition holds for some pair (x, y) then there is θ such that $\ln \frac{P_\theta(x)}{P_\theta(y)} > L\rho(x, y)$ and $\ln \frac{\phi(y)}{\phi(x)} > L\rho(x, y)$. Simply adding the two, we obtain $\ln \frac{\xi(\theta|x)}{\xi(\theta|y)} > 2L\rho(x, y)$, and so the resulting posterior is not L-differentially private. ■

In our main results, we show that the first part of the conditions, which is equivalent to our first assumption, is also sufficient. However, the second part is too weak to imply differential privacy on its own.

2.2.2 The Choice of Metric and Sufficient Statistics

The extent to which our assumptions hold for a particular family of distributions $\mathcal{F}_\Theta$ depends mainly on ρ.[8] The choice of metric is also important for achieving differential privacy with

8. Although our results are stated in terms of metrics, it is easy to translate them to neighbourhood-based results, simply by bounding the ρ-distance of any neighbouring data sets. See also the discussion in Section 6.

respect to it. Let us specifically consider metrics defined in terms of a difference in statistics:

$$\rho(x, y) \triangleq \|\tau(x) - \tau(y)\| \quad ,$$

where $\tau : \mathcal{S} \to \mathcal{V}$ is a statistic mapping from data sets to a normed vector space.

In that case, our assumptions imply that τ must be a *sufficient* statistic, since if $\tau(x) = \tau(y)$ then $\rho(x, y) = 0$ and it follows that $P_\theta(x) = P_\theta(y)$. More generally, ρ must be such that if the distance between x, y is zero, then their probabilities should be equal. We will see some examples of such statistics for conjugate distributions in the exponential family in Section 5. That means that we cannot use a metric which simply ignores part of the data, for example.

Similarly, the very definition of differential privacy (Definition 2) implies that τ must be a *Bayes-sufficient* statistic. That means that for any x, y, it holds

$$\tau(x) = \tau(y) \quad \Rightarrow \quad \xi(B \mid x) = \xi(B \mid y) \ , \quad \forall B \in \mathfrak{S}_\Theta \ .$$

Note that this is a slightly weaker condition than a sufficient statistic, which is necessary for our assumptions to hold.

2.3 Summary of Results

Given the above assumptions, we show: firstly, that if we choose an informative prior ξ, the resulting posterior is robust in terms of KL-divergence to small changes in the data. Secondly, that the posterior distribution is differentially private. Thirdly, that this implies that sampling from the posterior can be used as part of a differentially-private mechanism. We complement these with results on how easily an adversary can distinguish two similar data sets from posterior samples. Finally, we characterise the trade-off between utility and privacy, stated here informally for ease of exposition:

Claim 1 *If $\mathscr{A}$ prefers to use the prior $\xi^\star$, but $\mathscr{B}$ uses a prior ξ satisfying Assumption 1, and $\mathscr{A}$'s utility is bounded in $[0, 1]$, the following is true for the posterior sampling mechanism with N samples:*

- *The mechanism is $2NL$-differentially private.*
- *$\mathscr{A}$'s utility loss is $O\left([1 - \xi^\star(\Theta_L)] + \sqrt{1/N}\right)$ w.h.p., where Θ_L is the support of ξ.*

The following sections discuss our main results in detail. We begin by proving that our assumptions result in robust posteriors, in the sense that the KL divergence between posteriors arising from similar data sets is small. Then we show that they also result in differentially private posterior distributions, and analyse the resulting posterior sampling mechanism. We conclude with some examples and a discussion of related work.

3. Robustness of the Posterior Distribution

We now show that the above assumptions provide guarantees on the robustness of the posterior. That is, if the distance between two data sets x, y is small, then so too is the distance between the two resulting posteriors, $\xi(\cdot \mid x)$ and $\xi(\cdot \mid y)$. We prove this result for

the case where we measure the distance between the posteriors in terms of the well-known KL-divergence:

$$D\left(P \parallel Q\right) = \int_S \ln \frac{\mathrm{d}P}{\mathrm{d}Q} \mathrm{d}P \ .$$

The following theorem shows that any distribution family $\mathcal{F}_\Theta$ and prior ξ satisfying one of our assumptions is robust, in the sense that the posterior does not change significantly with small changes to the data set. It is notable that our mechanisms are simply tuned through the choice of prior.

Theorem 7 *When ξ is a prior distribution on Θ and $\xi(\cdot \mid x)$ and $\xi(\cdot \mid y)$ are the respective posterior distributions for data sets $x, y \in \mathcal{S}$, the following results hold:*

1. *Under a pseudo-metric ρ and $L > 0$ satisfying Assumption 1,*

$$D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) \leq 2L\rho(x, y) \ . \tag{7}$$

2. *Under a pseudo-metric ρ and $c > 1$ satisfying Assumption 2,*

$$D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) \leq C_\xi^{\mathcal{F}_\Theta} \left(1 + 2L_0 + c^{-1}\right) \rho(x, y) \ , \tag{8}$$

where $C_\xi^{\mathcal{F}_\Theta}$ is the ratio between the maximum and marginal likelihoods (9)*, and assuming there exists $\chi \in (0, 1]$ such that: $\forall x, y \in \mathcal{S}$ there is a sequence $\{z_k\} \subset \mathcal{S}$, with $z_0 = x, z_n = y$, satisfying $\chi\rho(z_k, z_{k+1}) \leq c - 1 \ \forall z_k$.*

Note that the second claim bounds the KL divergence in terms of $\mathscr{B}$'s prior belief that L is small, which is expressed via the constant c. The larger c is, the less prior mass is placed in large L and so the more robust inference becomes. Of course, choosing c to be too large may decrease efficiency.

It is important to also discuss the constant $C_\xi^{\mathcal{F}_\Theta}$. To get a better intuition, consider the case where $\Theta, \mathcal{X}$ are finite. Let $\theta^\star_{\mathrm{ML}}(x)$ be the maximum-likelihood estimate for x. Then we have that:

$$C_\xi^{\mathcal{F}_\Theta} \ = \ \max_x \frac{P_{\theta^\star_{\mathrm{ML}}(x)}(x)}{\sum_\Theta P_\theta(x)\xi(\theta)} \ \leq \ \max_x \frac{1}{\xi(\theta^\star_{\mathrm{ML}}(x))} \ , \tag{9}$$

there is therefore a natural dependency on the prior mass placed on maximum-likelihood estimators.

Finally, χ is going to be 1 for most metric spaces of interest. A notable exception is when the Hamming distance is used, which requires $\chi < 1$ as an additional technical condition for $c < 2$. However, this only affects our results under the second assumption.

4. Privacy and Utility

We next examine the differential privacy of the posterior distribution. We show in Section 4.1 that this can be achieved under either of our assumptions. The result can also be

interpreted as the differential privacy of a *posterior sampling mechanism* for responding to queries (described in Section 4.2), for which we prove a bound on the utility depending on the number of samples taken. Section 4.3 examines an alternative notion of privacy, *data set distinguishability*, similar to Wasserman and Zhou (2010). For this, we prove a bound on privacy, that also depends on the number of samples taken. Together, these exhibit a trade off between utility and privacy controlled by choosing the number of samples appropriately, in a manner described in Section 4.4.

4.1 Differential Privacy of Posterior Distributions

We consider our generalised notion of differential privacy for posterior distributions (Definition 2); and show that the type of differential privacy exhibited by the posterior depends on which assumption holds.

Theorem 8 *1. Under a pseudo-metric ρ and $L > 0$ satisfying Assumption 1, for all $x, y \in \mathcal{S}$, $B \in \mathfrak{S}_\Theta$:*

$$\xi(B \mid x) \leq \exp\{2L\rho(x,y)\}\xi(B \mid y) \; .$$

i.e., *the posterior ξ is $(2L, 0)$-differentially private under pseudo-metric ρ.*

2. Under a pseudo-metric ρ and $c > 1$ satisfying Assumption 2, $C_\xi^{\mathcal{F}_\Theta}$ defined in (9), for all $x, y \in \mathcal{S}$, $B \in \mathfrak{S}_\Theta$:

$$|\xi(B \mid x) - \xi(B \mid y)| \leq \sqrt{\frac{C_\xi^{\mathcal{F}_\Theta}}{2}\left(1 + 2L_0 + c^{-1}\right)\rho(x,y)},$$

i.e., *the posterior ξ is $\left(0, O(\sqrt{C_\xi^{\mathcal{F}_\Theta}(L_0 + 1/c)})\right)$-differentially private*[9] *under pseudo-metric $\sqrt{\rho}$.*

The difference between the two bounds' form is due to the fact that while the first claim has a direct proof, the second claim arises from the KL divergence bound in Theorem 7.

Finally, we show that posterior distributions are also randomly differentially private.

Corollary 9 *Under pseudo-metric ρ, $c > 1$ and $L \geq L_0 > 0$ satisfying Assumption 2:*

$$\mathbb{P}\left[\forall B \in \mathfrak{S}_\Theta : \xi(B \mid x) \leq \exp\left\{2L\rho(x,y)\right\}\xi\left(B \mid y\right), \forall x, y \in \mathcal{S}\right] \geq 1 - \exp(-c(L - L_0)) \; .$$

i.e., *the posterior ξ is $(2L, 0, \exp(-c(L - L_0)))$-randomly differentially private (Hall et al., 2011) under pseudo-metric ρ.*

This is a conceptually different definition from the original RDP, as the measure over which the randomness is defined is not the data distribution, but the prior measure ξ.

This property of the posterior distribution directly leads to the definition of a posterior sampling mechanism which will be differentially private. This is explained in the following section.

9. This holds, for example, for hamming distance as in the Beta-Binomial example presented in Lemma 21.

4.2 Posterior Sampling Mechanism

Given that we have a full posterior distribution which is differentially private, we can use it to define a private mechanism. We may allow the adversary to submit an arbitrary set of queries $\{ q_t \}$ with each $q_t \in \mathcal{Q}$. Each query warrants a response r_t in a set of possible responses $\mathcal{R}$. The adversary is allowed to condition the queries on our previous responses.

We extend our original approach (Dimitrakakis et al., 2014) to take some utility function u into account, which scores preferences of responses given a query. The algorithm requires a prior ξ to be defined on a family $\mathcal{F}_\Theta$ of probability distributions, whose members do not necessarily generate i.i.d. observations. They could be Markov chains for example. The first step is to simply draw a number of samples from the posterior, as in the original approach (Algorithm 2). After the algorithm calculates the posterior distribution $\xi(\cdot \mid x)$, N parameter samples are drawn from it, producing a parameter set $\hat{\Theta}$. Thereafter, responses depend only on the utility function and the sample $\hat{\Theta}$, and we do not draw new samples after every query. This allows us to work with a fixed privacy budget.

Algorithm 1 BAPS: Bayesian Posterior Sampling

1: **input** prior ξ, data $x \in \mathcal{S}$
2: Calculate posterior $\xi(\theta \mid x)$.
3: **for** $k = 1, \ldots, N$ **do**
4: Sample $\theta^{(k)} \sim \xi(\theta \mid D)$.
5: **end for**
6: **return** $\hat{\Theta} = \left\{ \theta^{(k)} : k = 1, \ldots, N \right\}$.

Corollary 10 *Algorithm 1 is differentially private under the conditions of Theorem 8, namely:*

1. *Under a pseudo-metric ρ and $L > 0$ satisfying Assumption 1, the algorithm is $(2NL, 0)$-differentially private under pseudo-metric ρ; or*
2. *Under a pseudo-metric ρ and $c > 1$ satisfying Assumption 2, $C_\xi^{\mathcal{F}_\Theta}$ defined in (9), the algorithm is $\left(0, O(N\sqrt{C_\xi^{\mathcal{F}_\Theta}(L_0 + 1/c)})\right)$-differentially private under pseudo-metric $\sqrt{\rho}$.*

Proof This follows directly from Theorems 8 and 4 (composition), as the algorithm samples from the posterior distribution, which is differentially private. ■

Utility and optimal responses. We assume the collection of a set of utility functions $\mathcal{U} = \{ u_\theta : \theta \in \Theta \}$, such that the optimal response for a given parameter θ is the one maximising a utility function $u_\theta : \mathcal{Q} \times \mathcal{R} \to [0, 1]$. If we know the true parameter θ, then we should respond to any query q with $r \in \arg\max_r u_\theta(q, r)$. However, since θ is unknown, we must select a method for conveying the required information. In a Bayesian setting, there are three main approaches we could employ. The standard methodology is to maximise

expected utility with respect to the posterior. This corresponds to marginalising out θ, and responding with:

$$r_t \quad \in \quad \arg\max_r \int_\Theta u_\theta(q_t, r) \, \mathrm{d}\xi(\theta \mid x) \ .$$

The second is to use the *maximum a posteriori* value of θ. The final, which we employ here, is to use sampling; *i.e.*, to reply to each query using parameters sampled from the posterior. This allows us to reply to arbitrary queries without compromising privacy, since the most information an adversary could obtain is the set of sampled parameters. By adjusting the number of samples used, we can easily trade off between privacy and utility.

After this we respond to a series of queries. For the t-th received query q_t, the algorithm returns the optimal response over the sampled parameter set $\hat{\Theta}$, in the manner shown in Algorithm 2. Since we allow arbitrary queries, the third party could simply ask for $\hat{\Theta}$ with a suitable choice of the utility function. Then if u is bounded, it is easy to show that the loss due to sampling is bounded.

Algorithm 2 PSAQR: Posterior Sample Query Response

1: **input** Parameter sample $\hat{\Theta}$.
2: **for** $t = 1, \ldots$ **do**
3: Observe query $q_t \in \mathcal{Q}$, perhaps depending on $r_1, \ldots, r_{t-1}$ and $q_1, \ldots, q_{t-1}$.
4: **return** $r_t \in \arg\max_r \sum_{\theta \in \hat{\Theta}} u_\theta(q_t, r)$
5: **end for**

Lemma 11 *The returned responses of the PSAQR mechanism have a utility which is within* $O\left(\sqrt{\ln(1/\delta)/N}\right)$ *of the optimal value with probability at least* $1 - \delta$ *for any* $\delta > 0$.

Now that we have demonstrated bounds on the utility for the algorithm above, we turn to the issue of how utility and privacy can be optimally tuned. First, we try and quantify the amount of samples an adversary needs to distinguish two data sets.

4.3 Distinguishability of Data Sets

In this section, we wish to relate the size of the sample $\hat{\Theta}$ to the amount of information about x that can be obtained by the adversary $\mathscr{A}$. More precisely, we need to bound how well $\mathscr{A}$ can distinguish x from all alternative data sets y. Within the posterior sampling query model, $\mathscr{A}$ has to decide whether $\mathscr{B}$'s posterior is $\xi(\cdot \mid x)$ or $\xi(\cdot \mid y)$. However, he can only do so within some neighbourhood ϵ of the original data. In this section, we bound $\mathscr{A}$'s error in determining the posterior in terms of the number of samples used. This is analogous to the data set-size bounds on queries in interactive models of differential privacy (Dwork et al., 2006), as well as the point of view of privacy as hypothesis testing (Kairouz et al., 2015; Wasserman and Zhou, 2010) where an adversary wishes to distinguish the data set from two alternatives.

For this section, we consider a utility function whose optimal response is $\hat{\Theta}$. This corresponds to the most powerful query possible under the model shown in Algorithm 2.

Then, the adversary needs only to construct the empirical distribution to approximate the posterior up to some sample error. By bounds on the KL divergence between the empirical and actual distributions we can bound his power in terms of how many samples he needs in order to distinguish between x and y.

Due to the sampling model, we first require a finite sample bound on the quality of the empirical distribution. The adversary could attempt to distinguish different posteriors by forming the empirical distribution on any sub-algebra $\mathfrak{S}$.

Lemma 12 *For any $\delta \in (0,1)$, let $\mathscr{M}$ be a finite partition of the sample space $\mathcal{S}$, of size $m \leq \log_2 \sqrt{1/\delta}$, generating the σ-algebra $\mathfrak{S} = \sigma(\mathscr{M})$. Let $x_1, \ldots, x_n \sim P$ be i.i.d. samples from a probability measure P on $\mathcal{S}$, let $P_{|\mathfrak{S}}$ be the restriction of P on $\mathfrak{S}$ and let $\hat{P}^n_{|\mathfrak{S}}$ be the empirical measure on $\mathfrak{S}$. Then, with probability at least $1-\delta$:*

$$\left\| \hat{P}^n_{|\mathfrak{S}} - P_{|\mathfrak{S}} \right\|_1 \leq \sqrt{\frac{3}{n} \ln \frac{1}{\delta}} \,. \tag{10}$$

We can combine this bound on the adversary's estimation error with Theorem 7's bound on the KL divergence between posteriors resulting from similar data to obtain a measure of how fine a distinction between data sets the adversary can make after a finite number of draws from the posterior:

Theorem 13 *Under Assumption 1, the adversary can distinguish between data x, y with probability $1-\delta$ if:*

$$\rho(x,y) \geq \frac{3}{4Ln} \ln \frac{1}{\delta} \;.$$

Under Assumption 2, this becomes:

$$\rho(x,y) \geq \frac{3}{2n\left(C_\xi^{\mathcal{F}_\Theta} + 2L_0 c^{-1}\right)} \ln \frac{1}{\delta} \;.$$

Consequently, either smoother likelihoods (*i.e.*, decreasing L), or a larger concentration on smoother likelihoods (*i.e.*, increasing c), increases the effort required by the adversary and reduces the sensitivity of the posterior. Note that, unlike the results obtained for differential privacy of the posterior sampling mechanism, these results have the same algebraic form under both assumptions.

4.4 Trading off Utility and Privacy

By construction, in our setting there are three ways with which to tune privacy. The first is the choice of family; the second is the choice of prior; and the third is how many samples N to draw. The choice of family is usually fixed due to other considerations. However, we have the choice of either tuning the prior, so that we can satisfy our assumptions with some suitable constants L or c, or by tuning the number of samples N in the posterior sampling framework.

The following lemma bounds the regret we suffer in terms of utility when the private posterior we use is ξ, in the case where the posterior we would like to use (assuming no privacy constraints) was $\xi^\star$.

Lemma 14 *If our utility is bounded in* $[0,1]$*, the private posterior we use is* ξ*, while the ideal posterior is* $\xi^\star$*, then the regret suffered is bounded by* $2\|\xi - \xi^\star\|_1$*.*

Finally, consider the case where $\mathscr{B}$, being a true Bayesian, is convinced that $\xi^\star$ is the correct prior distribution to use, but needs to use the prior ξ in order to achieve privacy. The following theorem bounds the expected KL divergence between the two resulting posteriors.

Lemma 15 *If* $\forall \theta \in \Theta$*,* $|\ln \xi^\star(\theta)/\xi(\theta)| \leq \eta$ *then the expected KL divergence is*

$$\mathbb{E}_{x\sim\phi^\star} D(\xi^\star(\cdot \mid x) \| \xi(\cdot \mid x)) \leq 2\eta \,,$$

where $\phi^\star$ *is the* $\xi^\star$ *marginal distribution.*

We can now combine Lemmas 11 and 14 with Lemma 15, to obtain the following result:

Corollary 16 *If* $\mathscr{A}$ *has a preferred prior* $\xi^\star$*, while the private prior used by* $\mathscr{B}$ *is* ξ *and it satisfies the conditions of Lemma 15, then the loss of* $\mathscr{A}$ *in terms of the* $\xi^\star$*-expected utility is* $O\left(\eta + \sqrt{\ln(1/\delta)/N}\right)$*, with probability at least* $1-\delta$*.*

Consequently, if $\mathscr{A}$ believes the correct prior should be $\xi^\star$, he can use the private posterior sample to make decisions, incurring a small loss. Finally, we already showed that $\mathscr{A}$ cannot distinguish between data that are closer than $O(1/N)$ with high probability. Hence, in this setting we can tune N to trade off utility and privacy.

The following theorem characterises the link between the choice of prior, the number of samples, privacy and utility directly. This connects several of our results in one place.

Theorem 17 *If, instead of using a non-private prior* $\xi^\star$*, we use a prior* ξ *restricted on* Θ_L *(such that it satisfies Assumption 1 with constant* L*) and generate* N *samples from the posterior, then (a) the sample is* $2LN$*-differentially private and (b) the loss of* $\mathscr{A}$ *in terms of the* $\xi^\star$*-expected utility is* $O\left([1-\xi^\star(\Theta_L)] + \sqrt{\ln(1/\delta)/N}\right)$*, with probability at least* $1-\delta$ *for any* $\delta > 0$*.*

Proof For (a) note that due to composition, N repetitions give $2LN$-differential privacy. For (b), let Θ_L be the support of ξ. Then, because ξ is the restriction of $\xi^\star$ on Θ_L, it holds that:

$$\begin{aligned}\|\xi - \xi^\star\|_1 &= \xi(\Theta_L) - \xi^\star(\Theta_L) + \xi^\star(\Theta \setminus \Theta_L) - \xi(\Theta \setminus \Theta_L) \\ &= 2[1 - \xi^\star(\Theta_L)] \ .\end{aligned}$$

We now just need to couple this with Lemmas 14 and 11 to directly obtain the stated bound on the utility. ■

In practice, our choice of ξ gives us a base amount of privacy that depends only on L. By keeping ξ fixed and increasing N, we can easily trade off privacy and utility.

Finally, we should note that the adversary could choose any arbitrary estimator ψ to guess x. Section 4.5 below describes how to apply Le Cam's method to obtain matching lower bounds in this case, by defining *data set estimators* as a model for the adversary.

4.5 Lower Bounds

It is possible to apply standard minimax theory to obtain lower bounds on the rate of convergence of the adversary's estimate to the true data. In order to do so, we can for example apply the method due to LeCam (1973), which places lower bounds on the expected distance between an estimator and the true parameter. In order to apply it in our case, we simply replace the parameter space with the data set space.

Le Cam's method assumes the existence of a family of probability measures indexed by some parameter, with the parameter space being equipped with a pseudo-metric. In our setting, we use Le Cam's method in a slightly unorthodox, but very natural manner. Define the family of probability measures on Θ to be:

$$\Xi \triangleq \{ \xi(\cdot \mid x) : x \in \mathcal{S} \},$$

the family of posterior measures in the parameter space, for a specific prior ξ. Consequently, now $\mathcal{S}$ plays the role of the parameter space, while ρ is used as the pseudo-metric. The original family $\mathcal{F}_\Theta$ plays no further role in this construction, other than a way to specify the posterior distributions from the prior.

Now let ψ be an arbitrary estimator of the unknown data x. As in (LeCam, 1973), we extend ρ to subsets of $\mathcal{S}$ via

$$\rho(A, B) \triangleq \inf \{ \rho(x, y) : x \in A, y \in B \} \ , \quad A, B \subset \mathcal{S} \ .$$

Now we can re-state the following well-known lemma for our specific setting.

Lemma 18 (Le Cam's method) *Let ψ be an estimator of x on Ξ taking values in the metric space $(\mathcal{S}, \rho)$. Suppose that there are well-separated subsets $\mathcal{S}_1, \mathcal{S}_2$ such that $\rho(\mathcal{S}_1, \mathcal{S}_2) \geq 2\delta$. Suppose also that Ξ_1, Ξ_2 are subsets of Ξ such that $x \in \mathcal{S}_i$ for $\xi(\cdot \mid x) \in \Xi_i$. Then:*

$$\sup_{x \in \mathcal{S}} \mathbb{E}_\xi(\rho(\psi, x) \mid x) \geq \delta \sup_{\xi_i \in co(\Xi_i)} \| \xi_1 \wedge \xi_2 \| \ .$$

This lemma has an interesting interpretation in our case. The quantity

$$\mathbb{E}_\xi(\rho(\psi, x) \mid x) = \int_\Theta \rho(\psi(\theta), x) \, \mathrm{d}\xi(\theta \mid x) \ ,$$

is the expected distance between the real data x and the guessed data $\psi(\theta)$ when θ is drawn from the posterior distribution. Consequently, it is possible to apply this method directly to obtain results for specific families of posteriors. These would of course be dependent on the family, the prior and the metric. While we shall not engage in this exercise, we point the interested reader to (Yu, 1997), which provides two simple examples with minimax rates of $O(n^{-4/9})$ and $O(n^{-4/5})$.

5. Examples Satisfying our Assumptions

In what follows we study, for different choices of likelihood and corresponding conjugate prior, what constraints can be placed on the prior's concentration to guarantee a desired level of privacy. These case studies closely follow the pattern in differential privacy research

where the main theorem for a new mechanism is a set of sufficient conditions on (*e.g.*, Laplace) noise levels to be introduced to a response in order to guarantee a level ϵ of ϵ-differential privacy.

For exponential families, we have the canonical form $p_\theta(x) = h(x) \exp\left\{\eta_\theta^\top \tau(x) - A(\eta_\theta)\right\}$, where $h(x)$ is the base measure, η_θ is the distribution's natural parameter corresponding to θ, $\tau(x)$ is the distribution's sufficient statistic, and $A(\eta_\theta)$ is its log-partition function. For distributions in this family, under the absolute log-ratio distance, the family of parameters Θ_L of Assumption 2 must satisfy, for all $x, y \in \mathcal{S}$: $\left|\ln \frac{h(x)}{h(y)} + \eta_\theta^\top \left(\tau(x) - \tau(y)\right)\right| \leq L\rho(x, y)$. If the left-hand side has an amenable form, then we can quantify the set Θ_L for which this requirement holds. Particularly, for distributions where $h(x)$ is constant and $\tau(x)$ is scalar (*e.g.*, Bernoulli, exponential, and Laplace), this requirement simplifies to $\frac{|\tau(x)-\tau(y)|}{\rho(x,y)} \leq \frac{L}{\eta_\theta}$. One can then find the supremum of the left-hand side independent from θ, yielding a simple formula for the feasible L for any θ. For each example, a detailed proof can be found in Appendix B. Note that in the following examples, we are making the conventional assumption in machine learning that data are bounded ($||x|| \leq B$). Also we use $\xi(\theta)\mathbb{1}_{[c_1,c_2]}$ to denote the trimmed density function obtained by setting the density outside $[c_1, c_2]$ to zero and renormalising the density.

We begin with a few simple examples for single observations, that are nevertheless illustrative.

Lemma 19 (Exponential-Exponential conjugate prior) *The exponential distribution $\mathcal{Exp}(x; \theta)$ with a trimmed exponential conjugate prior $\theta \sim \mathcal{Exp}(\theta; \lambda)\mathbb{1}_{[c_1,c_2]}$, $\lambda > 0$, satisfies Assumption 2 with parameter $c = \lambda$, $L_0 = c_1$, $C_\xi^{\mathcal{F}_\Theta} = c_2 / \min\left\{c_1 e^{-c_1 B}, c_2 e^{-c_2 B}\right\}$ and metric $\rho(x, y) = |x - y|$.*

Consequently, the trimmed-exponential prior results in a posterior sampling mechanism that is $(0, \delta)$-DP under ρ, with $\delta = \sqrt{\frac{1}{2} C_\xi^{\mathcal{F}_\Theta}(1 + 2c_1 + 1/\lambda)}$. It is also $(0, \delta)$-DP under the classical definition if $x, y \in [0, 1]$.

Lemma 20 (Laplace-Exponential conjugate prior) *The distribution Laplace$(x; s, \mu)$ with a trimmed exponential conjugate prior $1/s = \theta \sim Exp(\theta; \lambda)\mathbb{1}_{[c_1,c_2]}$, $\mu \in \mathbb{R}$, $s \geq 1/L$, $\lambda > 0$ satisfies Assumption 2 with parameters $c = \lambda$, $L_0 = c_1$,*

$$C_\xi^{\mathcal{F}_\Theta} = \begin{cases} \frac{c_2}{2\min\left\{\frac{1}{2c_2}, \frac{1}{2c_1}\exp\left(\frac{-B-\mu}{c_1}\right)\right\}} & , \quad x < \mu \\ \frac{c_2}{2\min\left\{\frac{1}{2c_2}, \frac{1}{2c_1}\exp\left(\frac{\mu-B}{c_1}\right)\right\}} & , \quad x \geq \mu \end{cases},$$

and metric $\rho(x, y) = |x - y|$.

It should come as no surprise that the same type of $(0, \delta)$-privacy is achieved for the Laplace distribution with a trimmed exponential prior. Now we move on to an example from which we draw multiple samples.

Lemma 21 (Beta-Binomial conjugate prior) *The Binomial distribution $\mathcal{Binom}(\theta, n)$, with prior $\theta \sim \mathcal{Beta}(\alpha, \beta)$, $\alpha = \beta > 1$ satisfies Assumption 2 for $L_0 = \ln n$, $c = 2^{-2\alpha+1}/B(\alpha)$,*

where $B(\alpha)$ denotes the beta function with parameters $\alpha = \beta$,

$$C_\xi^{\mathcal{F}_\Theta} = B(\alpha)/B\left(\frac{n+2\alpha-1}{2}, \frac{n+2\alpha+1}{2}\right)$$

and metric $\rho(x,y) = \|x-y\|_1$, where $x, y \in \{0,1\}^n$.

This is an example of a conjugate prior pair that is $(0,\delta)$-DP without trimming the prior, with $\delta = \sqrt{\frac{1}{2} C_\xi^{\mathcal{F}_\Theta}(1 + 2\ln n + 2^{2\alpha-1}B(\alpha))}$. Unfortunately, δ is increasing with n, and as Zheng (2015) shows, this result is essentially unimprovable with direct posterior sampling unless the prior is trimmed.

We next present two results on normal distributions.

Lemma 22 (Normal distribution with known mean and unknown variance) *The normal distribution $N(x;\mu,\sigma^2)$ with a trimmed exponential prior $1/\sigma^2 = \theta \sim \mathcal{E}\chi p(\theta;\lambda)\mathbb{1}_{[c_1,c_2]}$ satisfies Assumption 2 with parameter $c = \frac{2\lambda}{\max\{|\mu|,1\}}$, $L_0 = \frac{c_1 \max\{|\mu|,1\}}{2}$,*

$$C_\xi^{\mathcal{F}_\Theta} = \min\left\{\sqrt{c_2/c_1}\exp\left(\frac{c_1 c_2^2}{2}\right), \exp\left(\frac{c_2^3}{2}\right)\right\}$$

and metric $\rho(x,y) = \left|x^2 - y^2\right| + 2|x-y|$.

This example is interesting, because privacy is achieved under a rather unusual metric. However, note that the posterior is classically $(0, 3\delta)$-DP for data in $[0,1]$.

Unbounded observation spaces are generally a problem for privacy, even for finite parameter spaces, generally because likelihoods become vanishingly small, thus making log likelihood ratios arbitrarily large. However, the following two examples circumvent this problem. In the first example, we consider a general multivariate extension of Lemma 22. In the second we consider the case of discrete Bayesian networks, where privacy depends on the network connectivity and the probability of rare events—we have also considered posterior sampling of networks under complementary conditions, and output perturbation applied to posterior updates, in recent work (Zhang et al., 2016). In these examples, data is usually not i.i.d. (depending on the choice of network or covariance matrix) and the observation space is not a product space.

Lemma 23 (Multivariate normal distribution) *The multivariate normal distribution $N(x;\mu,A^{-1})$ satisfies our Assumption 1 with $L = \frac{1}{2}(\sum_{i=1}^n \lambda_i^2)^{\frac{1}{2}} \max\{1, ||\mu||_2\}$ under metric $\rho(x,y) = ||xx^\top - yy^\top||_F + 2||x-y||_2$. When $\mu = 0$, Assumption 1 is satisfied with $L = \frac{1}{2}(\sum_{i=1}^n \lambda_i^2)^{\frac{1}{2}}$ under metric $\rho(x,y) = ||(xx^\top - yy^\top)||_F$.*

Once more, we achieved $(\epsilon, 0)$-DP under our metric, which implies a $(3\epsilon, 0)$ classical DP for bounded data.

Lemma 24 (Discrete Bayesian networks) *Consider a family of discrete Bayesian networks on K variables, $\mathcal{F}_\Theta = \{P_\theta : \theta \in \Theta\}$. More specifically, each member P_θ, is a distribution on a finite space $\mathcal{S} = \prod_{k=1}^K \mathcal{S}_k$ and we write $P_\theta(x)$ for the probability of any outcome*

$x = (x_1, \ldots, x_K)$ in $\mathcal{S}$. Let $\varepsilon \triangleq \min_{\theta, x_k, x_{\mathcal{P}(k)}} P_\theta(x_k \mid x_{\mathcal{P}(k)})$, be the smallest conditional probability in the graph, where $\mathcal{P}(k)$ are the parents of node k.

Our observations can be independent samples $\{ x^t : t \in [T] \}$ of dependent variables $x_1^t, \ldots, x_k^t$. Define the connectivity vector $v \in \mathbb{N}^K$ such that $v_k = 1 + \deg(k)$, where $\deg(k)$ is the out-degree of node K. We now define the distance between two data sets x, y to be

$$\rho(x, y) \triangleq v^\top \delta(x, y), \qquad \delta_k(x, y) \triangleq \sum_{t=1}^{T} \mathbb{I}\{x_{k,t} \neq y_{k,t}\}.$$

Then Assumption 1 is satisfied with $L = \ln 1/\varepsilon$.

Consequently, discrete Bayesian networks, endowed with any prior on the family given in the above example, are $(2 \ln 1/\varepsilon, 0)$-DP under ρ. This also implies that they are $2\|v\|_\infty \ln 1/\varepsilon$-DP under the classical definition.

A simple application of this example is to data drawn from a Markov model on a finite state space. In particular, consider a time-homogeneous family of transition matrices $\theta_{i,j} \triangleq P_\theta(x_{t+1} = i \mid x_t = j)$. Then a prior consisting of product of truncated Dirichlet distributions that bound all multinomial probabilities above ε satisfies our assumptions and results in a $4 \ln 1/\varepsilon$-DP mechanism.

The above examples demonstrate that our assumptions are reasonable. In fact, for several of them we recover standard choices of prior distributions. However, for the privacy guarantees to be reasonable, it is best to restrict the prior to a set of parameters that is not very sensitive.

6. Discussion

We have presented a unifying framework for private and secure inference in a Bayesian setting. Under concentration conditions on the prior, we have shown that Bayesian inference is both robust and private. Firstly, we prove that similar data sets result in posterior distributions with small KL divergence. Secondly, we establish that the posterior is differentially private. This allows us to use a general posterior sampling mechanism for responding to queries, where privacy and utility are easy to trade off by adjusting the number of samples taken.

Owing to the fact that no additional machinery is required, this framework may serve as a fundamental building block for more sophisticated, private Bayesian inference. As an additional step towards this goal, we have demonstrated the application of our framework to deriving analytical expressions for well-known distribution families, and for discrete Bayesian networks. Finally, we bounded the amount of effort required of an attacker to breach privacy when observing samples from the posterior. This serves as a principled guide for how much access can be granted to querying the posterior, while still guaranteeing privacy.

Conversion of our results to the neighbourhood formulation. We state most of our results on specific models using a distance based on a sufficient statistic. Hence, to convert these to standard differential privacy, we only need to bound the ρ-distance of any neighbouring data sets. A good example are DBNs, where the case $\rho(x, y) = 1$ corresponds exactly to that of one record changing in a databse.

Practical application of our results. In general, it is hard to verify whether an existing model family will satisfy DP, because it implies checking whether the log-likelihood function is Lipschitz. Some parametric conjugate families, like the ones we examined in the examples, are amenable to analytic treatment. In practice, though, this might not be possible. It is for this reason that we propose to use rejection sampling in order to sample from the truncated posterior distribution. In particular, it is possible to resample from the posterior distribution, until a sample within the allowed interval of parameters is obtained. This is an approach we recently used in an application paper successfully (Zhang et al., 2016).

6.1 Related Work

In the past, little research in differential privacy focused on the Bayesian paradigm, with Dimitrakakis et al. (2014) being the first to establish conditions for differentially-private Bayesian inference. Nevertheless, our paper has many interesting links with both previous and follow up work, with respect to differential privacy, robustness and Bayesian inference, which we outline below. First, we discuss relations to other mechanisms achieving differential privacy and theoretical works about differential privacy; secondly, we discuss related work on the connection between robustness and privacy; and we conclude the related work section with a discussion of previous versions of this paper and follow-up work.

6.1.1 Differential Privacy

In our paper, we employ a Bayesian framework whereby optimal responses are characterised by the fact that they maximise expected utility. In Bayesian statistical decision theory (Berger, 1985; Bickel and Doksum, 2001; DeGroot, 1970), learning is cast as a statistical inference problem and decision-theoretic criteria are used as a basis for assessing, selecting and designing procedures. In particular, for a given utility function, the Bayes-optimal procedure maximises the expected utility under the posterior distribution.

In our setting, however, decisions using the data are not taken by the statistician $\mathscr{B}$. Instead, $\mathscr{A}$ provides a utility function, and trusts $\mathscr{B}$ to give him responses to queries that maximise expected utility. However $\mathscr{B}$ must also balance the need for privacy of the data provider, which results in some utility loss for $\mathscr{A}$. This is naturally captured by the difference in utility by making the decision private. This idea had already been explored in the exponential mechanism by McSherry and Talwar (2007), which connected differential privacy to mechanism design.

The exponential mechanism can be seen as a generalisation of the *Laplace mechanism*, which adds Laplace noise to released statistics (Dwork et al., 2006). The exponential mechanism releases a response with probability exponential in a utility function describing the usefulness of each response, with the best response having maximal utility. An alternate approach, employed for privatising regularised empirical-risk minimisation (Chaudhuri et al., 2011), is to alter the inferential procedure itself, in that case by adding a random term to the primal objective. We view our posterior sampling mechanism as a Bayesian counterpart. Further results on the accuracy of the exponential mechanism with respect to the Kolmogorov-Smirnov distance are given in (Wasserman and Zhou, 2010), which introduced the concept of privacy as hypothesis testing where an adversary wishes to distinguish two data sets. This is similar to our notion of data set distinguishability.

Learning from private data. In a different direction, Duchi et al. (2013) provided information-theoretic bounds for private learning. This essentially represents the protocol for interacting with an adversary as an arbitrary conditional distribution, rather than restricting it to specific mechanisms or models. In this way, they obtain fundamental bounds on rates of convergence from differentially-private views of data.

Bayesian inference and privacy. Other work at the intersection of privacy and Bayesian inference includes that of Williams and McSherry (2010) who applied Bayesian inference to improve the utility of differentially-private releases by computing posteriors in a noisy measurement model. In a similar vein, Xiao and Xiong (2012) used Bayesian credible intervals to respond to queries with as high utility as possible, subject to a privacy budget. In the PAC-Bayesian setting, Mir (2012) showed that the Gibbs estimator (McSherry and Talwar, 2007) is differentially private. While their algorithm corresponds to a posterior sampling mechanism, it is a posterior found by minimising risk bounds; by contrast, our results are purely Bayesian and come from conditions on the prior. It is also worthwhile noting that our Assumption 1 can in some cases be made equivalent to the definition of Pufferfish privacy (Kifer and Machanavajjhala, 2014), a privacy concept with Bayesian semantics. Thus, our results imply that in some cases Pufferfish privacy also results in differential privacy. Finally, independently to our preliminary work (Dimitrakakis et al., 2014), Wang et al. (2015) later proved differential privacy results for Gaussian processes under similar assumptions.

6.1.2 Robustness and Privacy

Dwork and Lei (2009) made the first connection between (frequentist) robust statistics and differential privacy, developing mechanisms for the interquartile, median and B-robust regression. While robust statistics are designed to operate near an ideal distribution, they can have prohibitively high global, worst-case sensitivity. In this case privacy was still achieved by performing a differentially-private test on local sensitivity before release (Dwork and Smith, 2009). In later work, Dwork et al. (2015) show that differentially-private views of the data result in good generalisation abilities. We discuss this more extensively in Section 6.1.3.

In a similar vein Chaudhuri and Hsu (2012) drew a quantitative connection between robust statistics and differential privacy by providing finite-sample convergence rates for differentially-private plug-in statistical estimators in terms of the *gross error sensitivity*, a common measure of robustness. These bounds can be seen as complementary to ours because our Bayesian estimators do not have private views of the data but use a suitably-defined prior instead.

Smoothness of the learning map, achieved here for Bayesian inference by appropriate concentration of the prior, is related to *algorithmic stability* which is used in statistical learning theory to establish error rates (Bousquet and Elisseeff, 2002). Rubinstein et al. (2012) used γ-uniform stability to calibrate the level of noise when using the Laplace mechanism to achieve differential privacy for the SVM. Hall et al. (2013) extended this technique to adding Gaussian process noise for differentially private release of infinite-dimensional functions lying in an RKHS.

In the Bayesian setting, robustness is typically handled through maximin policies. This is done by assuming that the prior distribution is selected arbitrarily by nature. In the field of robust statistics, the minimax asymptotic bias of a procedure incurred within an ε-contamination neighbourhood is used as a robustness criterion giving rise to the notions of a procedure's *influence function* and *breakdown point* to characterise robustness (Hampel et al., 1986; Huber, 1981). In a Bayesian context, robustness appears in several guises including minimax risk, robustness of the posterior within ε-contamination neighbourhoods, and robust priors (Berger, 1985). In this context Grünwald and Dawid (2004) demonstrated the link between robustness in terms of the minimax expected score of the likelihood function and the (generalised) maximum entropy principle, whereby nature is allowed to select a worst-case prior.

6.1.3 Previous Versions and Follow Up Work

Finally, we note that preliminary versions of this work appeared on arXiv (Dimitrakakis et al., 2013. Latest version 2015.) and ALT (Dimitrakakis et al., 2014). This version corrects technical issues with one proof, which affected the leading constants. We also replaced the original mechanism with one taking a fixed sample, which allows us to maintain a fixed privacy budget for an arbitrary number of queries. We make a novel use of Le Cam's method to prove lower bounds on indistinguishability, and we complement our original bounds with bounds for the utility of the mechanism. Finally, we discuss the relationship between posterior sampling, the exponential mechanism and the *safe Bayesian* generalisation of Bayesian inference. Follow-up work includes: Wang et al. (2015) who, under similar assumptions proved differential privacy results for Gibbs samplers; Zheng (2015) who improved some of our original bounds and also presented new results for other members of the exponential family; and Zhang et al. (2016) who recently initiated the exploration of the posterior sampler in probabilistic graphical models on multiple random variables.

Another important follow up work is that of Dwork et al. (2015). They have shown that *any* differentially private algorithm results in robustness, in the sense that the divergence between posterior distribution arising from similar data is small. This has a direct impact on the generalisation ability of statistical models and inferences drawn, and consequently allows for what they call the "re-usable hold-out". In our work, on the other hand, we have shown that with the right choice of prior, Bayesian inference is both private and robust. We have also shown that if the posterior distribution is robust, then it is also differentially private. In conclusion, robustness and privacy appear to be deeply linked, as our works have jointly shown conditions when one implies the other in three different ways: not only the same sufficient conditions can achieve both privacy and robustness, but privacy can also imply robustness, and robustness implies privacy. Further links between the two concepts are likely, as explained in the next section.

6.2 Future Directions

Although we have shown how Bayesian inference can already be differentially private by appropriately setting the prior, we have not examined how this affects learning. While larger c improves privacy, it also concentrates the prior so much that learning would be

inhibited. Thus, c could be chosen to optimise the trade-off between privacy and learning. However, we believe that the choice of the number of samples is easier to control.

From the theoretical side, we believe that the constant $C_\xi^{\mathcal{F}_\Theta}$ could be substantially improved, since right now it seems to be rather loose. It is also possible that its existence is only an artefact of the analysis, since it only appears for Assumption 2. However, we thought it crucial to include the results from this assumption in the paper, since they are connected to the second necessary condition. Hopefully, future work will uncover improved bounds for Assumption 2, or a similar condition to it.

Other future directions include investigating the links between posterior sampling and the exponential mechanism, as well as with the safe Bayesian approach (Grünwald, 2012) to inference. Consider an exponential mechanism which, given a utility function $u : \Theta \times \mathcal{Q} \to \mathbb{R}$ and a base measure μ on Θ returns $\theta \in \Theta$ sampled from the density

$$f(\theta) \propto e^{\epsilon u(\theta, q)} \frac{\mathrm{d}\mu(\theta)}{\mathrm{d}\lambda} \ .$$

As also noted by Wang et al. (2015), this has a similar form to the posterior distribution, by setting $u(\theta, q) = \ln p_\theta(x)$ and setting $\mu = \xi$ to the prior. This idea was used independently by Zhang et al. (2016) for releasing MAP point estimates. In this framework, privacy is achieved by setting ϵ to a sufficiently small value. However, it is interesting to note that this is how Grünwald (2012) obtains robustness results for modified Bayesian inference. This implies that in some cases we can gain both privacy and efficiency. We note that in our case, we have proven that privacy is attainable by altering the prior, which corresponds to the base measure in the exponential mechanism. Consequently, we believe it is worthwhile examining settings where adjusting both ϵ and the prior measure may be advantageous.

Acknowledgments

We gratefully thank Aaron Roth, Kamalika Chaudhuri, and Matthias Bussas for their discussion and insights as well as the anonymous reviewers for their comments on the paper, which helped to improve it significantly. This work was partially supported by the Marie Curie Project "Efficient Sequential Decision Making Under Uncertainty", Grant Number 237816; the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement n 608743; the SNSF Project, "SwissSenseSynergia"; and the Australian Research Council (DE160100584).

Appendix A. Proofs of Main Results

Proof of Lemma 5 For Assumption 1, the proof follows directly from the definition of the absolute log-ratio distance; namely,

$$\begin{aligned} |\ln p_\theta^n(\{x_i\}) - \ln p_\theta^n(\{y_i\})| &\leq \textstyle\sum_{i=1}^n |\ln p_\theta(x_i) - \ln p_\theta(y_i)| \\ &\leq L \textstyle\sum_{i=1}^n \rho(x_i, y_i) \ . \end{aligned}$$

For Assumption 2, consider sub-family Θ_L from Eq. (4) for marginal p_θ and pseudo-metric ρ, and define the corresponding sub-family Θ_L^n in terms of product distribution p_θ^n

and pseudo-metric ρ^n. Then the same argument as above shows that $\Theta_L \subseteq \Theta_L^n$. Hence, the same prior and parameter c yield the lower bound of Eq. (5), for Θ_L^n. ■

Proof of Theorem 7 Let us now tackle claim 1. First, we can decompose the KL-divergence into two parts.

$$\begin{aligned} D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) &= \int_\Theta \ln \frac{\mathrm{d}\xi(\theta \mid x)}{\mathrm{d}\xi(\theta \mid y)} \,\mathrm{d}\xi(\theta \mid x) \\ &= \int_\Theta \ln \frac{p_\theta(x)}{p_\theta(y)} \,\mathrm{d}\xi(\theta \mid x) + \int_\Theta \ln \frac{\phi(y)}{\phi(x)} \,\mathrm{d}\xi(\theta \mid x) \\ &\leq \int_\Theta \left| \ln \frac{p_\theta(x)}{p_\theta(y)} \right| \mathrm{d}\xi(\theta \mid x) + \int_\Theta \ln \frac{\phi(y)}{\phi(x)} \,\mathrm{d}\xi(\theta \mid x) \\ &\leq L\rho(x,y) + \left| \ln \frac{\phi(y)}{\phi(x)} \right| \ . \end{aligned} \tag{11}$$

From Assumption 1, $p_\theta(y) \leq \exp(L\rho(x,y)) p_\theta(x)$ for all θ so:

$$\begin{aligned} \phi(y) &= \int_\Theta p_\theta(y) \,\mathrm{d}\xi(\theta) \\ &\leq \exp(L\rho(x,y)) \int_\Theta p_\theta(x) \,\mathrm{d}\xi(\theta) = \exp(L\rho(x,y))\phi(x) \ . \end{aligned}$$

Combining this with (11) we obtain

$$D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) \leq 2L\rho(x,y) \ .$$

Claim 2 is dealt with similarly. Once more, we can break down the distance in parts. In more detail, we first write:

$$D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) \leq \underbrace{\int_\Theta \left| \ln \frac{p_\theta(x)}{p_\theta(y)} \right| \mathrm{d}\xi(\theta \mid x)}_{A} + \underbrace{\int_\Theta \ln \frac{\phi(y)}{\phi(x)} \,\mathrm{d}\xi(\theta \mid x)}_{B} \ ,$$

as before. Now, let us re-write the A term as

$$\int_\Theta \left| \ln \frac{p_\theta(x)}{p_\theta(y)} \right| \frac{p_\theta(x)}{\phi(x)} \,\mathrm{d}\xi(\theta) \leq \sup_{\theta'} \frac{p_{\theta'}(x)}{\phi(x)} \int_\Theta \left| \ln \frac{p_\theta(x)}{p_\theta(y)} \right| \mathrm{d}\xi(\theta) \ ,$$

so that the left-hand side term is the ratio between the maximal likelihood and marginal likelihood. Using the same steps, we can bound B in the same manner.

Now, let us define a data-dependent and a data-independent bound:

$$C_\xi^{\mathcal{F}_\Theta}(x) \triangleq \sup_\theta \frac{p_\theta(x)}{\phi(x)} \ , \qquad\qquad C_\xi^{\mathcal{F}_\Theta} \triangleq \sup_x C_\xi^{\mathcal{F}_\Theta}(x) \ .$$

Replacing, we obtain:

$$D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) \leq C_\xi^{\mathcal{F}_\Theta} \underbrace{\int_\Theta \left| \ln \frac{p_\theta(x)}{p_\theta(y)} \right| \mathrm{d}\xi(\theta)}_{A} + \underbrace{\int_\Theta \ln \frac{\phi(y)}{\phi(x)} \,\mathrm{d}\xi(\theta \mid x)}_{B} \ .$$

Now, to bound the individual terms, we start from A and note that theorem 3 of (Norkin, 1986) on the Lipschitz property of the expectation of stochastic Lipschitz functions applies.

Theorem 25 *(Norkin, 1986) If ξ is a probability measure on Θ and $f : \mathcal{S} \times \Theta \to \mathbb{R}$ is a ξ-measurable function, such that for any $\theta \in \Theta$, $f(\cdot, \theta)$ is $\ell(\theta)$-Lipschitz, then the function $f_\xi(x) \triangleq \mathbb{E}_\xi f(x, \theta)$ is L_ξ-Lipschitz, where $L_\xi = \mathbb{E}_\xi \ell(\theta)$.*

Recall that the expectation of a non-negative random variable can be written in terms of its CDF F as $\int_0^\infty [1 - F(t)] \, \mathrm{d}t$. In our case, $\ell(\theta)$ is a random variable on Θ, and we can write its cumulative distribution function as

$$F(t) \triangleq \xi\left(\{\, \theta \in \Theta : \ell(\theta) \leq t \,\}\right) = \xi(\Theta_t) \ ,$$

by the definition of Θ_t. It follows that $\ln p_\theta(x)$ is L_ξ-Lipschitz, where through the formula for the expectation of positive variables:

$$L_\xi = \int_0^\infty [1 - \xi(\Theta_t)] \, \mathrm{d}t \leq L_0 \xi(\Theta_{L_0}) + [1 - \xi(\Theta_{L_0})] \int_0^\infty e^{-ct} \, \mathrm{d}t \leq L_0 + c^{-1} \ . \tag{12}$$

So, term A becomes $C_\xi^{\mathcal{F}_\Theta} \left(L_0 + c^{-1}\right) \rho(x, y)$.

Now let us move on to term B. For technical reasons, we start by considering a pair x, y such that $\rho(x, y) \leq c - 1$. This also implies that $c > 1$, since the distance cannot be negative.

$$\frac{\phi(x)}{\phi(y)} \overset{(a)}{=} \int_\Theta \frac{p_\theta(x)}{\phi(y)} \, \mathrm{d}\xi(\theta) \overset{(b)}{\leq} \int_\Theta \frac{p_\theta(y) e^{\ell(\theta)\rho(x,y)}}{\phi(y)} \, \mathrm{d}\xi(\theta) \overset{(c)}{\leq} C_\xi^{\mathcal{F}_\Theta} \int_\Theta e^{\ell(\theta)\rho(x,y)} \, \mathrm{d}\xi(\theta) \ . \tag{13}$$

Note that $\left\{\, \theta \in \Theta : e^{\ell(\theta)\rho(x,y)} \leq t \,\right\} = \left\{\, \theta \in \Theta : \ell(\theta) \leq \rho(x, y)^{-1} \ln t \,\right\} = \Theta_{\rho(x,y)^{-1} \ln t}$. So the CDF of the random variable $e^{\ell(\theta)}$ is $F(t) = \xi(\Theta_{\rho(x,y)^{-1} \ln t})$. Then:

For positive random variables, $\mathbb{E}\, X^\rho = \rho \int_0^\infty t^{\rho-1}[1 - F(t)]dt$. Applying this to our case, we get:

$$\begin{aligned}
\mathbb{E}_\xi \, e^{\ell(\theta)\rho(x,y)} &= \mathbb{E}_\xi[e^{\ell(\theta)\rho(x,y)} \mid \ell \leq L_0]\xi(\Theta_{L_0}) + \mathbb{E}_\xi[e^{\ell(\theta)\rho(x,y)} \mid \ell > L_0][1 - \xi(\Theta_{L_0})] \\
&\leq e^{L_0\rho(x,y)} + \rho(x, y) \int_{t_0}^\infty t^{\rho(x,y)-1}[1 - \xi(\Theta_{\ln t})] \, \mathrm{d}t \\
&\leq e^{L_0\rho(x,y)} + \rho(x, y) \int_{t_0}^\infty e^{\ln t[\rho(x,y)-1]} e^{-c(\ln t - L_0)} \, \mathrm{d}t \qquad \text{(where } t_0 = e^{L_0}\text{)} \\
&= e^{L_0\rho(x,y)} + \rho(x, y) \int_{t_0}^\infty e^{\ln t[\rho(x,y)-c-1]+cL_0} \, \mathrm{d}t \\
&= e^{L_0\rho(x,y)} + \rho(x, y) e^{cL_0} \int_{t_0}^\infty t^{\rho(x,y)-c-1} \, \mathrm{d}t \\
&= e^{L_0\rho(x,y)} + \rho(x, y) e^{cL_0} \frac{t_0^{\rho(x,y)-c}}{c - \rho(x, y)} \\
&= e^{L_0\rho(x,y)} + \rho(x, y) e^{cL_0} \frac{e^{L_0(\rho(x,y)-c)}}{c - \rho(x, y)} \\
&\leq e^{L_0\rho(x,y)} + \rho(x, y) e^{cL_0} e^{L_0(\rho(x,y)-c)} \\
&= e^{L_0\rho(x,y)} + \rho(x, y) e^{L_0\rho(x,y)} = (1 + \rho(x, y)) e^{L_0\rho(x,y)} \leq e^{(1+L_0)\rho(x,y)}.
\end{aligned}$$

Consequently, $\ln \phi(x)/\phi(y) \leq C_\xi^{\mathcal{F}_\Theta}(1+L_0)\rho(x,y)$.

To handle larger distances ρ, we can simply apply the above result repeatedly between k data sets $z_1, \ldots, z_k$, where $z_1 = x$, $z_k = y$ and such that $\rho(z_i, z_{i+1}) < c - 1$. By chaining logarithmic ratios, *i.e.*, using the fact that $\ln \phi(x)/\phi(y) = \ln \phi(x)/\phi(z) + \ln \phi(z)/\phi(y)$ we can now extend our result to general pairs for term B. Replacing those terms, we obtain:

$$D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) \leq C_\xi^{\mathcal{F}_\Theta}\left(1 + 2L_0 + c^{-1}\right)\rho(x,y) \ .$$

If the intermediate points do not exist under ρ, we can simply scale it by $\chi \leq 1$, thus obtaining the final result. ■

Proof of Theorem 8 For part 1, we assumed that there is an $L > 0$ such that $\forall x, y \in \mathcal{S}$, $\left|\log \frac{p_\theta(x)}{p_\theta(y)}\right| \leq L\rho(x,y)$, thus implying $\frac{p_\theta(x)}{p_\theta(y)} \leq \exp\{L\rho(x,y)\}$. Further, in the proof of Theorem 7, we showed that $\phi(y) \leq \exp\{L\rho(x,y)\}\phi(x)$ for all $x, y \in \mathcal{S}$. From Eq. (1), we can then combine these to bound the posterior of any $B \in \mathfrak{S}_\Theta$ as follows for all $x, y \in \mathcal{S}$:

$$\xi(B \mid x) = \frac{\int_B \frac{p_\theta(x)}{p_\theta(y)} p_\theta(y)\,\mathrm{d}\xi(\theta)}{\phi(y)} \cdot \frac{\phi(y)}{\phi(x)} \leq \exp\{2L\rho(x,y)\}\xi(B \mid y) \ .$$

For part 2, note that from Theorem 7 part 2 that the KL divergence of the posteriors under assumption is bounded by (8). Now, recall Pinsker's inequality (cf. Fedotov et al., 2003):

$$D\left(Q \| P\right) \geq \frac{1}{2}\left\|Q - P\right\|_1^2 = 2\|Q - P\|_{\mathrm{TV}}^2 \triangleq 2 \sup_B |Q(B) - P(B)|^2 \tag{14}$$

This yields: $|\xi(B \mid x) - \xi(B \mid y)| \leq \sqrt{\frac{1}{2} D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right)} \leq \sqrt{\frac{1}{2} C_\xi^{\mathcal{F}_\Theta}\left(1 + 2L_0 c^{-1}\right)\rho(x,y)}$. ■

Proof of Lemma 11 Sampling N times from the posterior, gives us the following estimate of the utility function $\hat{u}_\xi(q,r) = \frac{1}{N}\sum_{\theta \in \hat{\Theta}} u_\theta(q,r)$, which with probability at least $1 - \delta$ satisfies $|\hat{u}_\xi(q,r) - u(q,r)| < \sqrt{\frac{\ln(2/\delta)}{2N}} = \epsilon$, $\forall r, q$, via Hoeffding's inequality and the boundedness of u. Consequently, we can be at most 2ϵ-away from the optimal. ■

Proof of Lemma 12 (Note that in this proof, ε, δ do not refer to the privacy parameters.) We use the inequality due to Weissman et al. (2003) on the ℓ_1 norm, which states that for any multinomial distribution P with m outcomes, the ℓ_1 deviation of the empirical distribution $\hat{P}_n$ after n draws from the multinomial satisfies:

$$\mathbb{P}\left(\left\|\hat{P}_n - P\right\|_1 \geq \varepsilon\right) \leq (2^m - 2)e^{-\frac{1}{2}n\varepsilon^2}, \qquad \forall \varepsilon > 0 \ .$$

The right hand side is bounded by $e^{m \ln 2 - \frac{1}{2} n \varepsilon^2}$. Substituting $\varepsilon = \sqrt{\frac{3}{n} \ln \frac{1}{\delta}}$:

$$\begin{aligned}
\mathbb{P}\left(\left\| \hat{P}_n - P \right\|_1 \geq \sqrt{\frac{3}{n} \ln \frac{1}{\delta}} \right) &\leq e^{m \ln 2 - \frac{3}{2} \ln \frac{1}{\delta}} \\
&\leq e^{\log_2 \sqrt{\frac{1}{\delta}} \ln 2 - \frac{3}{2} \ln \frac{1}{\delta}} \\
&= e^{\frac{1}{2} \ln \frac{1}{\delta} - \frac{3}{2} \ln \frac{1}{\delta}} \\
&= \delta \ .
\end{aligned}$$

where the second inequality follows from $m \leq \log_2 \sqrt{1/\delta}$. ■

Proof of Theorem 13 Recall that the data processing inequality states that, for any sub-algebra $\mathfrak{S}$:

$$\left\| Q_{|\mathfrak{S}} - P_{|\mathfrak{S}} \right\|_1 \leq \| Q - P \|_1 \ .$$

Using this and Pinsker's inequality (14) we obtain:

$$\begin{aligned}
2L\rho(x,y) &\geq D\left(\xi(\cdot \mid x) \| \xi(\cdot \mid y) \right) \\
&\geq \frac{1}{2} \left\| \xi(\cdot \mid x) - \xi(\cdot \mid y) \right\|_1^2 \\
&\geq \frac{1}{2} \left\| \xi_{|\mathfrak{S}}(\cdot \mid x) - \xi_{|\mathfrak{S}}(\cdot \mid y) \right\|_1^2 \ .
\end{aligned}$$

On the other hand, due to (10) the adversary's ℓ_1 error in the posterior distribution is bounded by $\sqrt{\frac{3}{n} \ln \frac{1}{\delta}}$ with probability $1 - \delta$. In order for him to be able to distinguish the two different posteriors, it must hold that

$$\left\| \xi_{|\mathfrak{S}}(\cdot \mid x) - \xi_{|\mathfrak{S}}(\cdot \mid y) \right\|_1 \geq \sqrt{\frac{3}{n} \ln \frac{1}{\delta}} \ .$$

Using the above inequalities, we can bound the error in terms of the distinguishability of the real data set x from an arbitrary set y as:

$$4L\rho(x,y) \geq \frac{3}{n} \ln \frac{1}{\delta} \ .$$

Rearranging, we obtain the required result. The second case is treated similarly to obtain:

$$2C_\xi^{\mathcal{F}_\Theta} \left(1 + 2L_0 + c^{-1} \right) \rho(x,y) \geq \frac{3}{n} \ln \frac{1}{\delta} \ .$$

■

Proof of Lemma 14 Let $r, r^\star$ be the optimal responses under $\xi, \xi^\star$ respectively. For notational convenience, let $u_\xi = \int_\Theta u_\theta \, \mathrm{d}\xi(\theta)$ denote the expected utility under a belief ξ.

Then our regret is

$$\begin{aligned} u_\xi(q,r) - u_\xi(q,r^\star) &= u_\xi(q,r) - u_{\xi^\star}(q,r) \\ &\quad + u_{\xi^\star}(q,r) - u_{\xi^\star}(q,r^\star) \\ &\quad + u_{\xi^\star}(q,r^\star) - u_\xi(q,r^\star) \\ &\leq 2 \left\| \xi - \xi^\star \right\|_1 \ . \end{aligned}$$

This follows from the fact that

$$\begin{aligned} u_\xi(q,r) - u_{\xi^\star}(q,r) &= \int_\Theta u_\theta(q,r) \, \mathrm{d}[\xi - \xi^\star](\theta) \\ &\leq \|u\|_\infty \|\xi - \xi^\star\|_1 \end{aligned}$$

and then using the boundedness of u. The third term is dealt with identically. For the second term, note that $u_{\xi^\star}(q,r) - u_{\xi^\star}(q,r^\star) \leq 0$ since $r^\star$ maximises $u_{\xi^\star}$. ■

Proof of Lemma 15 Let $\phi^\star(x) = \int_\Theta p_\theta(x) \, \mathrm{d}\xi^\star(x)$ be the prior marginal distribution. Then the $\xi^\star$-expected KL divergence between the two posteriors is

$$\begin{aligned} &\sum_x \int_\Theta \ln \frac{\mathrm{d}\xi^\star(\theta \mid x)}{\mathrm{d}\xi(\theta \mid x)} \, \mathrm{d}\xi^\star(\theta \mid x) \phi^\star(x) \\ &\leq \sum_x \int_\Theta \left(\left| \ln \frac{\mathrm{d}\xi^\star(\theta)}{\mathrm{d}\xi(\theta)} \right| + \left| \ln \frac{\phi(x)}{\phi^\star(x)} \right| \right) \mathrm{d}\xi^\star(\theta \mid x) \phi^\star(x) \\ &\leq 2\eta \ . \end{aligned}$$

The first term $\left| \ln \frac{\mathrm{d}\xi^\star(\theta)}{\mathrm{d}\xi(\theta)} \right|$ is bounded by η by assumption. From the same assumption, it follows that $\phi(x) = \int_\Theta p_\theta(x) \, \mathrm{d}\xi(\theta) \leq \int_\Theta p_\theta(x) e^\eta \, \mathrm{d}\xi^\star(\theta) = e^\eta \phi^\star(x)$, and so the second term is also bounded by η. ■

Appendix B. Proofs of Examples

Proof of Lemma 19 Since $Exp(x;\theta)$ is monotonic decreasing in x and concave as a function of θ, we have $\inf_{\{||x|| \leq B, \theta \in [c_1,c_2]\}} Exp(x;\theta) = \min\left\{c_1 e^{-c_1 B}, c_2 e^{-c_2 B}\right\} \leq \phi(x)$. Then we have

$$C_\xi^{\mathcal{F}_\Theta} = c_2 / \min\left\{c_1 e^{-c_1 B}, c_2 e^{-c_2 B}\right\} \ .$$

Next we compute the absolute log-ratio distance for any x_1 and x_2 according to the exponential likelihood function:

$$|\ln p_\theta(x_1) - \ln p_\theta(x_2)| = \theta |x_1 - x_2| \ .$$

Thus, for $\theta \in [c_1, c_2]$, under Assumption 2, using $\rho(x,y) = |x - y|$, the set of feasible parameters for any $L > c_1$ is $\Theta_L = (c_1, L)$. Note the density of the renormalized exponential prior on $[c_1, c_2]$ is given by $K\lambda e^{-\lambda\theta}$, where $K = (e^{-\lambda c_1} - e^{-\lambda c_2})^{-1}$. Thus the CDF at L of

this density is $K\left(e^{-\lambda c_1} - e^{-\lambda L}\right)$ for $L \in [c_1, c_2]$ and 1 for $L \geq c_2$. It is natural to choose L_0 to be c_1. Then we need to find c such that

$$\xi(\Theta_L) = \int_{c_1}^{L} K\lambda e^{-\lambda\theta} d\theta = K(e^{-\lambda c_1} - e^{-\lambda L}) \geq 1 - e^{-c(L-c_1)}$$

for $L \in (c_1, c_2)$. By plugging K into the inequality, we have

$$e^{-c(L-c_1)} \geq \frac{e^{-\lambda(L-c_2)} - 1}{e^{-\lambda(c_1-c_2)} - 1} \ .$$

Since $e^{-\lambda(L-c_2)} \leq e^{-\lambda(c_1-c_2)}$, it is sufficiency to find c such that $e^{-c(L-c_1)} \geq e^{-\lambda(L-c_1)}$. Therefore we can have $c = \lambda$. ■

Proof of Lemma 20 Note that $Laplace(x; s, \mu)$ is monotonic decreasing in x if $x < \mu$, and increasing in x if $x \geq \mu$. Since $Laplace(x; s, \mu)$ is concave as a function of s, we have $\phi(t) \geq \min\left\{\frac{1}{2c_2}, \frac{1}{2c_1}\exp\left(\frac{-B-\mu}{c_1}\right)\right\}$ if $x < \mu$ and $\phi(t) \geq \min\left\{\frac{1}{2c_2}, \frac{1}{2c_1}\exp\left(\frac{\mu-B}{c_1}\right)\right\}$ if $x \geq \mu$. Thus, we can take

$$C_\xi^{\mathcal{F}_\Theta} = \begin{cases} \frac{c_2}{2\min\left\{\frac{1}{2c_2}, \frac{1}{2c_1}\exp\left(\frac{-B-\mu}{c_1}\right)\right\}} \ , & x < \mu \\ \frac{c_2}{2\min\left\{\frac{1}{2c_2}, \frac{1}{2c_1}\exp\left(\frac{\mu-B}{c_1}\right)\right\}} \ , & x \geq \mu \end{cases} .$$

For any x_1 and x_2, the absolute log-ratio distance for this distribution can be bounded as

$$\begin{aligned} &|\ln p_{\mu,s}(x_1) - \ln p_{\mu,s}(x_2)| \\ =&\tfrac{1}{s}\left|\|x_1 - \mu\| - \|x_2 - \mu\|\right| \leq \tfrac{1}{s}\|x_1 - x_2\| \ , \end{aligned}$$

where the inequality follows from the triangle inequality on $\|\cdot\|$. Thus, if we use $\rho(x, y) = \|x - y\|$, the set of feasible parameters for Assumption 2 is $\mu \in \mathbb{R}$ and $\frac{1}{s} = \theta \leq L$. Again we can use the trimmed exponential prior with rate parameter $\lambda > 0$ for the inverse scale, $\frac{1}{s}$, and similar to the previous example, Assumption 2 is satisfied with $c = \lambda$ and $L_0 = c_1$. ■

Proof of Lemma 21 Here, we consider data drawn from a Binomial distribution with a beta prior on its proportion parameter, θ. Thus, the likelihood and prior functions are

$$\begin{aligned} p_{\theta,n}(X = k) &= \tbinom{n}{k}\theta^k(1-\theta)^{n-k} \\ \xi_0(\theta) &= \tfrac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} \ , \end{aligned}$$

where $k \in \{0, 1, 2, \ldots, n\}$, $a, b \in \mathbb{R}_+$ and $B(a, b)$ is the beta function. The resulting posterior is a Beta-Binomial distribution. Again we consider the application of Assumption 2 to this Beta-Binomial distribution. For this purpose, we must quantify the parameter sets Θ_L for a given $L > 0$ according to a distance function. The absolute log-ratio distance between the Binomial likelihood function for any pair of arguments, k_1 and k_2, is

$$|\ln p_{\theta,n}(k_1) - \ln p_{\theta,n}(k_2)| = \left|\Delta_n(k_1, k_2) + (k_1 - k_2)\ln\tfrac{\theta}{1-\theta}\right|$$

where $\Delta_n(k_1, k_2) \triangleq \ln \binom{n}{k_1} - \ln \binom{n}{k_2}$. By substituting this distance into the supremum of Eq. (4), we seek feasible values of $L > 0$ for which the supremum is non-negative; here, we explore the case where $\rho((n, k_1), (n, k_2)) \triangleq |k_1 - k_2|$. Without loss of generality, we assume $k_1 > k_2$, and thus require that

$$\sup_{k_1 > k_2} \left| \frac{\Delta_n(k_1,k_2)}{k_1 - k_2} + \ln \frac{\theta}{1-\theta} \right| \leq L \ . \tag{15}$$

However, by the definition of $\Delta_n(k_1, k_2)$, the ratio $\frac{\Delta_n(k_1,k_2)}{k_1-k_2}$ is in fact the slope of the chord from k_2 to k_1 on the function $\ln \binom{n}{k}$. Since the function $\ln \binom{n}{k}$ is concave in k, this slope achieves its maximum and minimum at its boundary values; *i.e.*, it is maximised for $k_1 = 1$ and $k_2 = 0$ and minimised for $k_1 = n$ and $k_2 = n - 1$. Thus, the ratio attains a maximum value of $\ln n$ and a minimum of $-\ln n$ for which the above supremum is simply $\ln n + \left| \ln \frac{\theta}{1-\theta} \right|$. From Eq. (15), we therefore have, for all $L \geq \ln n$:

$$\Theta_L = \left[\left(1 + \tfrac{e^L}{n}\right)^{-1}, \left(1 + \tfrac{n}{e^L}\right)^{-1} \right] \ .$$

We want to bound $\xi(\Theta_L)$. We know that: $\xi(\Theta_L) = 1 - \xi\left(\Theta_L^{\complement}\right)$ where $\Theta_L^{\complement}$ is the complement of Θ_L. so $\xi(\Theta_L^{\complement})$ is composed of two symmetric intervals: $\left[0, \left(1 + \frac{e^L}{n}\right)^{-1}\right)$ and $\left(\left(1 + \frac{n}{e^L}\right)^{-1}, 1\right]$. We selected $\alpha = \beta$, therefore the mass must concentrate at $\frac{1}{2}$, as we have $\alpha > 1$.

Due to symmetry, the mass outside of Θ_L is two times that is the first interval. This is:

$$\frac{2}{B(\alpha, \alpha)} \int_0^{\frac{p}{1+p}} x^{\alpha-1}(1-x)^{\alpha-1} \, \mathrm{d}x \ .$$

where p denotes $ne^{-L} \in [0, 1]$, Therefore c is upper bounded by

$$\ln \left(\frac{2A(p)}{B(\alpha, \alpha)} \right) / (L_0 - L) = \ln \left(\frac{2A(p)}{B(\alpha, \alpha)} \right) / \ln p,$$

where $A(p)$ denotes the incomplete Beta function $\int_0^{\frac{p}{1+p}} x^{\alpha-1}(1-x)^{\alpha-1} dx$. Note that we have

$$A'(p) = \frac{p^{\alpha-1}}{(1+p)^{2\alpha}} \ ,$$

$$A''(p) = \frac{p^{\alpha-2}[(\alpha-1)(1+p) - 2\alpha p]}{(1+p)^{2\alpha+1}} \ .$$

Claim 2 *$H(p) = \alpha A(p) - \frac{p^\alpha}{(1-p)(1+p)^{2\alpha-1}} \leq 0$ for all $p \in (0, 1)$.*

Proof Calculating derivatives and simplifying

$$\begin{aligned}
& H'(p) \\
= \; & \alpha A'(p) - \frac{\alpha p^{\alpha-1}(1-p)(1+p)^{2\alpha-1} - p^\alpha\left[(2\alpha-1)(1-p)(1+p)^{2\alpha-2} - (1+p)^{2\alpha-1}\right]}{[(1-p)(1+p)^{2\alpha-1}]^2} \\
= \; & \frac{\alpha p^{\alpha-1}}{(1+p)^{2\alpha}} - \frac{\alpha p^{\alpha-1}(1-p)(1+p) - p^\alpha[(2\alpha-1)(1-p) - (1+p)]}{(1-p)^2(1+p)^{2\alpha}} \\
= \; & \frac{p^{\alpha-1}}{(1+p)^{2\alpha}}\left(\alpha - \frac{\alpha(1-p^2) - 2p(\alpha-1-p\alpha)}{(1-p)^2}\right) \\
= \; & \frac{p^{\alpha-1}}{(1+p)^{2\alpha}(1-p)^2}\left(\alpha(1-2p+p^2) - \alpha(1-p^2) + 2p(\alpha-1-\alpha p)\right) \\
= \; & \frac{-2p^\alpha}{(1+p)^{2\alpha}(1-p)^2} < 0 \; .
\end{aligned}$$

Therefore $H(p)$ is strictly decreasing. Then combined with $H(0) = 0$, we claim follows. ■

Claim 3 *$G(p) = p\frac{A'(p)}{A(p)} \ln p - \ln \frac{2A(p)}{B(\alpha,\alpha)} < 0$ for all $p \in (0,1)$.*

Proof Again taking derivatives

$$\begin{aligned}
G'(p) &= \frac{A'(p)}{A(p)}(1+\ln p) + p\ln p \frac{A''(p)A(p) - A'(p)^2}{A(p)^2} - \frac{A'(p)}{A(p)} \\
&= \frac{\ln p}{A(p)^2}(A(p)A'(p) + pA''(p)A(p) - pA'(p)^2) \\
&= \frac{\ln p}{A(p)^2}\left[\frac{p^{\alpha-1}}{(1+p)^{2\alpha}}A(p)\left(1 + \frac{(\alpha-1)(1+p) - 2\alpha p}{1+p}\right) - \frac{p^{2\alpha-1}}{(1+p)^{4\alpha}}\right] \\
&= \frac{\ln p}{A(p)^2}\frac{p^{\alpha-1}}{(1+p)^{2\alpha+1}}\left[\alpha(1-p)A(p) - \frac{p^\alpha}{(1+p)^{2\alpha-1}}\right] \\
&= \frac{p^{\alpha-1}}{(p+1)^{2\alpha+1}A(p)^2}H(p)\ln p(1-p) > 0 \; .
\end{aligned}$$

So $G(p)$ is strictly increasing. Combined with $\lim_{p\to 1} G(p) = 0$, the claim follows. ■

Claim 4 *$F(p) = \ln\left(2I_{\frac{p}{1+p}}(\alpha)\right) / \ln p$ is decreasing in $p \in (0,1)$, where the incomplete Beta function $I_{\frac{p}{1+p}}(\alpha) = A(p)/B(\alpha,\alpha)$.*

Proof Taking derivatives

$$\begin{aligned}
F'(p) &= \frac{1}{\ln^2 p}\left(\frac{A'(p)}{A(p)}\ln p - \frac{1}{p}\ln\frac{2A(p)}{B(\alpha,\alpha)}\right) \\
&= \frac{1}{p\ln^2 p}\left(\frac{A'(p)}{A(p)}p\ln p - \ln\frac{2A(p)}{B(\alpha,\alpha)}\right) \\
&= \frac{1}{p\ln^2 p}G(p) < 0 \; .
\end{aligned}$$

■

Therefore $\ln\left(2I_{\frac{p}{1+p}}(\alpha)\right)/\ln p$ is monotonic decreasing in p. Thus the minimum value of $F(p)$ is $\frac{1}{B(\alpha)2^{2\alpha-1}}$ as $p \to 1$, which we can take as our c in this example.

Let us consider $C_\xi^{\mathcal{F}_\Theta}$ for this example. We have

$$\frac{p_\theta(x)}{\phi(x)} = \frac{B(\alpha,\beta)\theta^x(1-\theta)^{n-x}}{B(\alpha+x, n+\beta-x)},$$

where $\theta \in [0,1]$ and $x \in [0,1,\ldots,n]$. Note that

$$\frac{B(\alpha+x+1, n+\beta-x-1)}{B(\alpha+x, n+\beta-x)} = \frac{\Gamma(\alpha+x+1)\Gamma(n+\beta-x-1)}{\Gamma(\alpha+x)\Gamma(n+\beta+1)} = \frac{\alpha+x}{n+\beta-x-1}.$$

So $B(\alpha+x+1, n+\beta-x-1) \le B(\alpha+x, n+\beta-x)$ if $x \le \frac{n+\beta-\alpha-1}{2}$; $B(\alpha+x+1, n+\beta-x-1) > B(\alpha+x, n+\beta-x)$ otherwise. Thus

$$B(\alpha+x, n+\beta-x) \ge B\left(\frac{n+\alpha+\beta-1}{2}, \frac{n+\alpha+\beta+1}{2}\right).$$

Hence we can take $C_\xi^{\mathcal{F}_\Theta} = B(\alpha,\beta)/B\left(\frac{n+\alpha+\beta-1}{2}, \frac{n+\alpha+\beta+1}{2}\right)$. ■

Proof of Lemma 22 Since $N(x;\mu,\theta)$ is decreasing in x^2 and concave as a function of θ. We have $\phi(t) \ge \inf_{\{x|||x||\le B\},\theta\in[c_1,c_2]} N(x;\mu,\theta) = \min\left\{\sqrt{\frac{c_1}{2\pi}}e^{\frac{-c_1c_2^2}{2}}, \sqrt{\frac{c_2}{2\pi}}e^{\frac{-c_2^3}{2}}\right\}$. Then we can take

$$C_\xi^{\mathcal{F}_\Theta} = \min\left\{\sqrt{c_2/c_1}e^{\frac{c_1c_2^2}{2}}, e^{\frac{c_2^3}{2}}\right\}$$

For the normal distribution, (4) requires: $2L\rho(x,y)\sigma^2 \ge |2\mu - x - y|\,|x-y|$. Taking the absolute log ratio of the Gaussian densities we have

$$\begin{aligned}&\frac{1}{2\sigma^2}\left|\left((x-\mu)^2 - (y-\mu)^2\right)\right| \\ \le &\frac{\max\{|\mu|,1\}}{2\sigma^2}\left(\left|x^2-y^2\right| + 2|x-y|\right).\end{aligned}$$

Consequently, we can set $\rho(x,y) = \left|x^2-y^2\right| + 2|x-y|$ and $L(\mu,\sigma) = \frac{\max\{|\mu|,1\}}{2\sigma^2}$. Again, the trimmed exponential prior is given by $K\lambda e^{-\lambda\theta}$, where $K = (e^{-\lambda c_1} - e^{-\lambda c_2})^{-1}$. Thus the CDF at L of this density is $K\left(e^{-\lambda c_1} - e^{-\lambda L}\right)$ for $L \in [\frac{c_1\max\{|\mu|,1\}}{2}, \frac{c_2\max\{|\mu|,1\}}{2}]$ and 1 for $L \ge \frac{c_2\max\{|\mu|,1\}}{2}$. Thus the CDF at L of this density is $K\left(e^{-\lambda c_1} - e^{\frac{-2\lambda L}{\max\{|\mu|,1\}}}\right)$. We choose L_0 to be $\frac{c_1\max\{|\mu|,1\}}{2}$. Then we need to find c such that

$$\xi(\Theta_L) = \int_{c_1}^{L} K\lambda e^{-\lambda\theta}d\theta = K(e^{-\lambda c_1} - e^{-\lambda L}) \ge 1 - e^{-c\left(L - \frac{c_1\max\{|\mu|,1\}}{2}\right)}.$$

By plugging K into the inequality, we have

$$e^{-c\left(L-\frac{c_1 \max\{\,|\mu|,1\,\}}{2}\right)} \geq \frac{e^{\frac{-2\lambda L}{\max\{\,|\mu|,1\,\}}+\lambda c_2}-1}{e^{-\lambda(c_1-c_2)}-1}.$$

Since $e^{-\lambda\left(\frac{2\lambda L}{\max\{\,|\mu|,1\,\}}-c_2\right)} \leq e^{-\lambda(c_1-c_2)}$, it is sufficiency to find c such that

$$e^{-c\left(L-\frac{c_1 \max\{\,|\mu|,1\,\}}{2}\right)} \geq e^{-\lambda\left(\frac{2L}{\max\{\,|\mu|,1\,\}}-c_1\right)}.$$

This is equivalent to have c satisfying

$$c\left(L-\frac{c_1 \max\{\,|\mu|,1\,\}}{2}\right) \leq \lambda\left(\frac{2L}{\max\{\,|\mu|,1\,\}}-c_1\right).$$

Then we can take $c=\frac{2\lambda}{\max\{\,|\mu|,1\,\}}$ to satisfy the above inequality. ■

Proof of Lemma 23 Consider the likelihood log-ratio distance of multivariate normal distributions with precision matrix A:

$$\frac{1}{2}|x^\top Ax - y^\top Ay| \ ,$$

where A is positive definite with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n > 0$). For simplicity, assume the mean to be a zero vector then

$$\begin{aligned}
|x^\top Ax - y^\top Ay| &= \left|\sum_{i,j} x_i x_j A_{i,j} - \sum_{i,j} y_i y_j A_{i,j}\right| \\
&= \left|\sum_{i,j} A_{i,j}(x_i x_j - y_i y_j)\right| \\
&= |Tr(A(xx^\top - yy^\top)')| \\
&\leq [Tr(A^2)Tr((xx^\top - yy^\top)(xx^\top - yy^\top)')]^{\frac{1}{2}} \\
&= \left(\sum_{i=1}^n \lambda_i^2\right)^{\frac{1}{2}} ||(xx^\top - yy^\top)||_F \ .
\end{aligned}$$

For mean equal to μ, we have

$$\frac{1}{2}|(x^\top-\mu)A(x-\mu) - (y^\top-\mu)A(y-\mu)| \ .$$

By the above analysis we have the difference being bounded by

$$\frac{1}{2}\left(\sum_{i=1}^n \lambda_i^2\right)^{\frac{1}{2}} ||(x-\mu)(x-\mu)' - (y-\mu)(y-\mu)')||_F \ .$$

Note that

$$\begin{aligned}||(x-\mu)(x-\mu)' - (y-\mu)(y-\mu)'||_F =& ||xx^\top - \mu(x^\top - y^\top) - (x-y)\mu' - yy^\top||_F \\ \leq& ||xx^\top - yy^\top||_F + 2||\mu(x-y)'||_F \\ =& ||xx^\top - yy^\top||_F + 2||\mu||_2||(x-y)'||_2 \\ \leq& \max\{1, ||\mu||_2\}(||xx^\top - yy^\top||_F + 2||x-y||_2) \ .\end{aligned}$$

■

Proof of Lemma 24 It is instructive to first examine the case where all variables are independent and we have a single draw from P_θ. Then $P_\theta(x) = \prod_{k=1}^K \theta_{k,x_k}$ and

$$\left|\ln \frac{P_\theta(x)}{P_\theta(y)}\right| = \left|\ln \prod_{k=1}^K \frac{\theta_{k,x_k}}{\theta_{k,y_k}}\right| \leq \sum_{k=1}^K \left|\ln \frac{\theta_{k,x_k}}{\theta_{k,y_k}}\right| \mathbb{I}\left\{x_k \neq y_k\right\} \leq \max_{i,j,k} \left|\ln \frac{\theta_{k,i}}{\theta_{k,j}}\right| \rho(x,y) \ . \tag{16}$$

Consequently, if $\varepsilon \triangleq \min_{k,j} \theta_{k.j}$ is the smallest probability assigned to any one sub-event, then $L > \ln 1/\varepsilon$, since $\theta_{k,j} \leq 1$.

In the general case, we have independent draws x^t, y^t, where $x^t \sim P_\theta(x)$ and the variables x_k^t have dependences defined through a graphical model, such that $P_\theta(x) = \prod_k P_\theta(x_k \mid x_{\mathcal{P}(k)})$, where $\mathcal{P}(k)$ are the parents of node k. Similarly to (16), we write

$$\begin{aligned}\left|\ln \frac{P_\theta(x)}{P_\theta(y)}\right| &= \left|\ln \prod_t \frac{P_\theta(x^t)}{P_\theta(y^t)}\right| = \left|\ln \prod_t \prod_k \frac{P_\theta(x_k^t \mid x_{\mathcal{P}(k)}^t)}{P_\theta(y_k^t \mid y_{\mathcal{P}(k)}^t)}\right| \\ &\leq \sum_{t.k} \left|\ln \frac{P_\theta(x_k^t \mid x_{\mathcal{P}(k)}^t)}{P_\theta(y_k^t \mid y_{\mathcal{P}(k)}^t)}\right| \leq \ln \frac{1}{\epsilon} \sum_{t.k} \mathbb{I}\left\{x_k^t \neq y_k^t \vee x_{\mathcal{P}(k)}^t \neq y_{\mathcal{P}(k)}^t\right\}. \end{aligned} \tag{17}$$

The last term is the number of times a value is different in x and y times one plus the number of variables it affects. To model this, let $v \in \mathbb{N}^K$ be such that $v_k = 1 + \deg(k)$ and define: $\rho(x,y) \triangleq v^\top \delta(x,y)$ and $\delta_k(x,y) \triangleq \sum_t \mathbb{I}\left\{x_{k,t} \neq y_{k,t}\right\}$. Rewriting (17) in terms of ρ, we obtain $\left|\ln \frac{P_\theta(x)}{P_\theta(y)}\right| \leq \ln \frac{1}{\varepsilon} \cdot \rho(x,y)$ as desired. ■

References

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, 1985.

Peter J. Bickel and Kjell A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*, volume 1. Holden-Day Company, 2001.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

Konstantinos Chatzikokolakis, Miguel E. Andres, Nicolas Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In *Privacy Enhancing Technologies*, pages 82–102, 2013.

Kamalika Chaudhuri and Daniel Hsu. Convergence rates for differentially private statistical estimation. In *Proceedings of the 29th International Conference on Machine Learning*, ICML, pages 1327–1334, 2012.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

Morris H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, 1970.

Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin Rubinstein. Robust and private Bayesian inference. Technical Report 1306.1066, arXiv, 2013. Latest version 2015.

Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin Rubinstein. Robust and private Bayesian inference. In *25th Conference on Algorithmic Learning Theory (ALT)*, volume 8776 of *Lecture Notes in Computer Science*, pages 291–305. Springer, 2014.

John C Duchi, Michael Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 429–438. IEEE, 2013.

Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming (ICALP)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC, pages 371–380, 2009.

Cynthia Dwork and Adam Smith. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2):135–154, 2009.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference*, TCC, pages 265–284, 2006.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349 (6248):636–638, 2015.

Alexei A. Fedotov, Peter Harremoës, and Flemming Topsoe. Refinements of Pinsker's inequality. *IEEE Transactions on Information Theory*, 49(6):1491–1498, 2003.

Peter Grünwald. The safe Bayesian: Learning the learning rate via the mixability gap. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, (ALT), pages 169–183, 2012.

Peter D. Grünwald and A. Philip Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.

Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Random differential privacy. *Journal of Privacy and Confidentiality*, 4(2), 2011.

Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Differential privacy for functions and functional data. *Journal of Machine Learning Research*, 14(Feb):703–727, 2013.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley and Sons, 1986.

Peter J. Huber. *Robust Statistics*. John Wiley and Sons, 1981.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *Proceedings of The 32nd International Conference on Machine Learning*, ICML, pages 1376–1385, 2015.

Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)*, 39(1):3, 2014.

Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS, pages 94–103, 2007.

Darakhshan Mir. Differentially-private learning and information theory. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 206–210. ACM, 2012.

VI Norkin. Stochastic Lipschitz functions. *Cybernetics and Systems Analysis*, 22(2):226–233, 1986.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, NIPS, pages 3003–3011, 2013.

Benjamin I. P. Rubinstein, Peter L. Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: Privacy-preserving mechanisms for SVM learning. *Journal of Privacy and Confidentiality*, 4(1), 2012.

W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Yu-Xiang Wang, Stephen E. Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *Proceedings of The 32nd International Conference on Machine Learning*, ICML, pages 2493–2502, 2015.

Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the L_1 deviation of the empirical distribution. Technical report, Hewlett-Packard Labs, 2003.

Oliver Williams and Frank McSherry. Probabilistic inference and differential privacy. In *Advances in Neural Information Processing Systems*, NIPS, pages 2451–2459, 2010.

Yonghui Xiao and Li Xiong. Bayesian inference under differential privacy. arXiv preprint arXiv:1203.0617, 2012.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Zuhe Zhang, Benjamin I. P. Rubinstein, and Christos Dimitrakakis. On the differential privacy of Bayesian inference. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, AAAI, pages 2365–2371, 2016.

Shijie Zheng. The differential privacy of Bayesian inference, 2015. Bachelor's thesis, Harvard College `http://nrs.harvard.edu/urn-3:HUL.InstRepos:14398533`.