Robust and Private Bayesian Inference

Christos Dimitrakakis[1], Blaine Nelson[2⋆],
Aikaterini Mitrokotsa[1], and Benjamin I. P. Rubinstein[3]

[1] Chalmers University of Technology, Sweden
[2] University of Potsdam, Germany
[3] The University of Melbourne, Australia

Abstract. We examine the robustness and privacy of Bayesian inference, under assumptions on the prior, and with no modifications to the Bayesian framework. First, we generalise the concept of differential privacy to arbitrary dataset distances, outcome spaces and distribution families. We then prove bounds on the robustness of the posterior, introduce a posterior sampling mechanism, show that it is differentially private and provide finite sample bounds for distinguishability-based privacy under a strong adversarial model. Finally, we give examples satisfying our assumptions.

1 Introduction

Significant research challenges for statistical learning include efficiency, robustness to noise (stochasticity) and adversarial manipulation, and preserving training data privacy. In this paper we study techniques for meeting these challenges simultaneously, through a simple unification of Bayesian inference, differential privacy and distinguishability. In particular, we examine the following problem.

Summary of setting. A Bayesian statistician ($\mathscr{B}$) wants to communicate results about some data x to a third party ($\mathscr{A}$), but without revealing the data x itself. (x could be a single datum, or a sample of data.) More specifically:

(i) $\mathscr{B}$ selects a model family ($\mathcal{F}_\Theta$) and a prior (ξ).
(ii) $\mathscr{A}$ is allowed to see $\mathcal{F}_\Theta$ and ξ and is computationally unbounded.
(iii) $\mathscr{B}$ observes data x and calculates the posterior $\xi(\theta|x)$ but does not reveal it. Instead, $\mathscr{B}$ responds to queries at times $t = 1, \ldots$ as follows.
(iv) $\mathscr{A}$ sends a query q_t to $\mathscr{B}$.
(v) $\mathscr{B}$ responds $q_t(\theta_t)$ where θ_t is drawn from the posterior: $\theta_t \sim \xi(\theta|x)$.

We show that if $\mathcal{F}_\Theta$ or ξ are chosen appropriately, the resulting posterior-sampling mechanism satisfies generalized differential privacy and indistinguishability properties. The intuition 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

⋆ Blaine Nelson is now at Google, Mountain View

easily discover unknown information about the data. This suggests that robustness and privacy can be simultaneously achieved and perhaps are deeply linked. We show that under mild assumptions this is indeed true for the posterior distribution, suggesting a differentially-private mechanism for Bayesian inference.

Our contributions. (i) We generalise differential privacy to arbitrary dataset distances, outcome spaces, and distribution families. (ii) 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 dataset result in small posterior changes; (iii) We introduce a novel *posterior sampling mechanism* that is private. Unlike other common mechanisms, our approach sits squarely in the non-private (Bayesian) learning framework without modification; (iv) We introduce the notion of *dataset distinguishability* for which we provide finite-sample bounds for our mechanism (v) We provide examples of conjugate-pair distributions where our assumptions hold.

Paper organisation. Section 1.1 discusses related work. Section 2 specifies the setting and our assumptions. Section 3 proves results on robustness of Bayesian learning. Section 4 proves privacy results. Examples where our assumptions hold are given in Section 5. We present a discussion of our results in Section 6. Appendix A contains proofs of the main theorems. Proofs of the examples and a discussion on matching lower bounds are given in a technical report [8].

1.1 Related Work

In Bayesian statistical decision theory [1, 2, 7], 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 cost function, the Bayes-optimal procedure minimises the *Bayes risk* under a particular prior distribution.

In an adversarial setting, this is extended to a minimax risk, 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 notion of a procedure's *influence function* and *breakdown point* to characterise robustness [17, 18]. In a Bayesian context, robustness appears in several guises including minimax risk, robustness of the posterior within ϵ-contamination neighbourhoods, and robust priors [1]. In this context Grünwald and Dawid [15] demonstrated the link between robustness in terms of the minimax expected score of the likelihood function and the (generalized) maximum entropy principle, whereby nature is allowed to select a worst-case prior.

Differential privacy, first proposed by Dwork et al. [12], has achieved prominence in the theory of computer science, databases, and more recently learning communities. Its success is largely due to the semantic guarantee of privacy it formalises. Differential privacy is normally defined with respect to a randomised

mechanism for responding to queries. Informally, a mechanism preserves differential privacy if perturbing one training instance results in a small change to the mechanism's response distribution. Differential privacy is detailed in Section 2.

A popular approach for differential privacy is the *exponential mechanism* [19] which generalises the *Laplace mechanism* of adding Laplace noise to released statistics [12]. This mechanism releases a response with probability exponential in a score function measuring distance to the non-private response. An alternate approach, employed for privatising regularised ERM [6], is to alter the inferential procedure itself, in that case by adding a random term to the primal objective. Further results on the accuracy of the exponential mechanism with respect to the Kolmogorov-Smirnov distance are given in [23]. Unlike previous studies, our mechanisms do not require modification to the underlying learning framework.

In a different direction, Duchi et al. [9] provided information-theoretic bounds for private learning, by modelling the protocol for interacting with an adversary as an arbitrary conditional distribution, rather than restricting it to specific mechanisms. In a similar vein Chaudhuri and Hsu [5] 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.

Little research in differential privacy focuses on the Bayesian paradigm, and to our knowledge, none has established differentially-private Bayesian inference. Williams and McSherry [25] 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 [26] 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 [20] showed that the Gibbs estimator [19] 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.

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 [3]. Rubinstein et al. [22] used the γ-uniform stability of the SVM to calibrate the level of noise for using the Laplace mechanism to achieve differential privacy for the SVM. Hall et al. [16] extended this technique to adding Gaussian process noise for differentially private release of infinite-dimensional functions lying in an RKHS.

Finally, Dwork and Lei [11] 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 [13]. Little further work has explored robustness and privacy, and no general connection is known.

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 information to the queries of $\mathscr{A}$ (*e.g.,* how 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 good 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

We begin with our notation. Let $\mathcal{S}$ be the set of all possible datasets. 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 $\mathcal{X}$.

Comparing datasets. Central to notions of privacy and robustness, is the concept of distance between datasets. Firstly, the effect of dataset perturbation on learning depends on the amount of noise as quantified by some distance. Secondly, the amount that an attacker can learn from queries can be quantified in terms of the distance of his guesses to the true dataset. To model these situations, we equip $\mathcal{S}$ with a pseudo-metric[4] $\rho : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$. Using pseudo-metrics, we considerably generalise previous work on differential privacy, which considers only the special case of Hamming distance. We note that a similar generalisation has been developed in parallel and independently by Chatzikokolakis et al. [4].

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 dataset x. More precisely, we assume that data $x \in \mathcal{S}$ have been drawn from some distribution P_{θ^*} on $\mathcal{S}$, parametrised by θ^*, 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 $\mathcal{S}$:

$$\mathcal{F}_\Theta \triangleq \{ P_\theta : \theta \in \Theta \}, \tag{1}$$

and where we use p_θ to denote the corresponding densities[5] 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 $\theta^* \in B$. In general, the posterior distribution after observing $x \in \mathcal{S}$ is:

$$\xi(B \mid x) = \frac{\int_B p_\theta(x) \, \mathrm{d}\xi(\theta)}{\phi(x)} \ , \tag{2}$$

[4] Meaning that $\rho(x, y) = 0$ does not necessarily imply $x = y$.
[5] *I.e.,* the Radon-Nikodym derivative of P_θ relative to some dominating measure ν.

where ϕ is the corresponding marginal density given by:

$$\phi(x) \triangleq \int_{\Theta} p_\theta(x) \, \mathrm{d}\xi(\theta) \ . \tag{3}$$

While the choice of the prior is generally arbitrary, this paper shows that its careful selection can yield good privacy guarantees.

Privacy. We first recall the idea of differential privacy [10]. This states that on similar datasets, 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.

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, there is at most one $i \in \{1, \ldots, n\}$ such that $x_i \neq y_i$.*

As a first step, we generalise this definition to arbitrary dataset spaces $\mathcal{S}$ that are not necessarily product spaces. To do so, we introduce the notion of differential privacy under a pseudo-metric ρ on the space of all datasets.

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 $x \in \mathcal{S}$, then:*

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

Remark 1. 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.,* they only 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 $P(B \mid x) \leq P(B \mid x) + \delta\rho(x, y)$.

Definition 1 allows for privacy against a very strong attacker $\mathscr{A}$, who attempts to match the empirical distribution induced by the true dataset by querying

[6] Making the definition wholly equivalent is possible, but results in an unnecessarily complex definition.

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 dataset 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. Our Definition 2 allows for a much broader encoding of the attacker's knowledge via the selected pseudo-metric.

2.2 Our Main Assumptions

In the sequel, we show that if the distribution family $\mathcal{F}_\Theta$ or prior ξ is such that close datasets $x, y \in \mathcal{S}$, 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 dataset x from similar datasets.

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 parameterizations of the family:

Assumption 1 (Lipschitz continuity) *Let $d(\cdot,\cdot)$ be a metric on $\mathbb{R}$. There exists $L > 0$ such that, for any $\theta \in \Theta$:*

$$d(p_\theta(x), p_\theta(y)) \leq L\rho(x, y), \qquad \forall x, y \in \mathcal{S} \ . \tag{4}$$

However, it may be difficult for this assumption to hold uniformly over Θ. This can be seen by a counterexample for the Bernoulli family of distributions. Consequently, we relax it by only requiring that $\mathscr{B}$'s *prior* probability ξ is concentrated in the parts of the family for which the likelihood is smoothest:

Assumption 2 (Stochastic Lipschitz continuity[21]) *Let $d(\cdot,\cdot)$ be a metric on $\mathbb{R}$ and let*

$$\Theta_L \triangleq \left\{\theta \in \Theta : \sup_{x,y\in\mathcal{S}} \left\{d(p_\theta(x), p_\theta(y)) - L\rho(x, y)\right\} \leq 0\right\} \tag{5}$$

be the set of parameters for which Lipschitz continuity holds with Lipschitz constant L. Then there is some constant $c > 0$ such that, for all $L \geq 0$:

$$\xi(\Theta_L) \geq 1 - \exp(-cL) \ . \tag{6}$$

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.

To make our assumptions concrete, we now fix the distance function d to be the absolute log-ratio,

$$d(a, b) \triangleq \begin{cases} 0 & \text{if } a = b = 0 \\ \left|\ln \frac{a}{b}\right| & \text{otherwise} \end{cases} , \tag{7}$$

which is a proper metric on $\mathbb{R}_+ \times \mathbb{R}_+$. This particular choice of distance yields guarantees on differential privacy and indistinguishability.

We next 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 1. *If p_Θ satisfies Assumption 1 (resp. Assumption 2) with respect to pseudo-metric ρ and constant L (or c), then, for any fixed $n \in \mathbb{N}$, $p_\Theta^n(\{ x_i \}) = \prod_{i=1}^n p_\Theta(x_i)$ satisfies the same assumption with respect to:*

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

and constant $L \cdot n$ (or $\frac{c}{n}$). Further, if $\{ x_i \}$ and $\{ y_i \}$ differ in at most k items, the assumption holds with the same pseudo-metric but with constant $L \cdot k$ (or $\frac{c}{k}$) instead.

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 datasets 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:

$$D\left(P \parallel Q\right) = \int_S \ln \frac{\mathrm{d}P}{\mathrm{d}Q} \, \mathrm{d}P \ . \tag{8}$$

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 dataset. It is notable that our mechanisms are simply tuned through the choice of prior.

Theorem 1. *When $d : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ is the absolute log-ratio distance* (7)*, ξ is a prior distribution on Θ and $\xi(\cdot \mid x)$ and $\xi(\cdot \mid y)$ are the respective posterior distributions for datasets $x, y \in \mathcal{S}$, the following results hold:*

(i) Under a metric ρ and $L > 0$ satisfying Assumption 1,

$$D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) \leq 2L\rho(x, y) \tag{9}$$

(ii) Under a metric ρ and $c > 0$ satisfying Assumption 2,

$$D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) \leq \frac{\kappa}{c} \cdot \rho(x, y) \tag{10}$$

where κ is constant (see Appendix A); $\kappa \approx 4.91081$.

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.

4 Privacy Properties of the Posterior Distribution

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, which is described in Section 4.2. Finally, Section 4.3 introduces an alternative notion of privacy: *dataset distinguishability*. We prove a high-probability bound on the sample complexity of distinguishability under our assumptions.

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 privacy exhibited by the posterior depends on which assumption holds.

Theorem 2. *Using the log-ratio distance (as in Theorem 1),*

(i) Under Assumption 1, for all $x, y \in \mathcal{S}$, $B \in \mathfrak{S}_\Theta$*:*

$$\xi(B \mid x) \leq \exp\{2L\rho(x,y)\}\xi(B \mid y) \tag{11}$$

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

(ii) Under Assumption 2, for all $x, y \in \mathcal{S}$, $B \in \mathfrak{S}_\Theta$*:*

$$|\xi(B \mid x) - \xi(B \mid y)| \leq \sqrt{\frac{\kappa}{2c}\rho(x,y)}$$

i.e., *the posterior* ξ *is* $\left(0, \sqrt{\frac{\kappa}{2c}}\right)$*-differentially private under pseudo-metric* $\sqrt{\rho}$.

4.2 Posterior Sampling Query Model

Given that we have a full posterior distribution, we use it to define an algorithm achieving privacy. In this framework, we allow the adversary to submit a set of queries $\{ q_k \}$ which are mappings from parameter space Θ to some arbitrary answer set Ψ; *i.e.,*, $q_k : \Theta \to \Psi$. If we know the true parameter θ, then we would reply to any query with $q_k(\theta)$. However, since θ is unknown, we must select a method for conveying the required information. There are three main approaches that we are aware of. The first is to marginalise θ out. 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 q_k using a θ_k sampled from the posterior.

This sample-based interactive query model is presented in Algorithm 1. First, the algorithm calculates the posterior distribution $\xi(\cdot \mid x)$. Then, for the k^{th} received query q_k, the algorithm draws a sample θ_k from the posterior distribution and responds with $q_k(\theta_k)$.

Algorithm 1 Posterior sampling query model

1: **Input** prior ξ, data $x \in \mathcal{S}$
2: Calculate posterior $\xi(\cdot \mid x)$.
3: **for** $k = 1, \ldots$ **do**
4: Observe query $q_k : \Theta \to \Psi$.
5: Sample $\theta_k \sim \xi(\cdot \mid x)$.
6: Return $q_k(\theta_k)$.
7: **end for**

In this context, Theorem 2 can be interpreted as proving differential privacy for the posterior sampling mechanism for the case when the response set is the parameter set; *i.e.,* $\Psi = \Theta$ and $q_k(\theta) = \theta$. Due to the data-processing inequality, this also holds for all query functions. As an example, consider querying conditional expectations:

Example 1. Let each model P_θ in the family define a distribution on the product space $\mathcal{S} = \bigcup_{n=1}^{\infty} \mathcal{X}^n$, such for any $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$, $P_\theta(x) = \prod_i P_\theta(x_i)$. In addition, let $\mathcal{X} = \mathcal{Y} \times \mathcal{Z}$ (with appropriate algebras $\mathfrak{S}_\mathcal{X}, \mathfrak{S}_\mathcal{Y}, \mathfrak{S}_\mathcal{Z}$) and write $x_i = (x_{i,\mathcal{Y}}, x_{i,\mathcal{Z}})$ for point x_i and its two components. A conditional expectation query would require an answer to the question:

$$\mathbb{E}_\theta(x_{|\mathcal{Y}} \mid x_{|\mathcal{Z}}),$$

where the parameter θ is unknown to the questioner. In this case, the answer set Ψ would be identical to $\mathcal{Y}$, while k would index the values in $\mathcal{Z}$.

4.3 Distinguishability of Datasets

A limitation of the differential privacy framework is that it does not give us insight on the amount of effort required by an adversary to obtain private information. In fact, an adversary wishing to breach privacy, needs to distinguish x from alternative datasets 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 his error in determining the posterior in terms of the number of queries he performs. This is analogous to the dataset-size bounds on queries in interactive models of differential privacy [12].

Let us consider an adversary querying to sample $\theta_k \sim \xi(\cdot \mid x)$. This is the most powerful query possible under the model shown in Algorithm 1. 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 2. *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 $1-\delta$:*

$$\left\| \hat{P}^n_{|\mathfrak{S}} - P_{|\mathfrak{S}} \right\|_1 \leq \sqrt{\frac{3}{n} \ln \frac{1}{\delta}}. \tag{12}$$

Of course, the adversary could choose any arbitrary estimator ψ to guess x. The accompanying technical report [8] describes how to apply Le Cam's method to obtain matching lower bound rates in this case, by defining *dataset estimators*. This is however is not essential for the remainder of the paper.

We can combine this bound on the adversary's estimation error with Theorem 1's bound on the KL divergence between posteriors resulting from similar data to obtain a measure of how fine a distinction between datasets the adversary can make after a finite number of draws from the posterior:

Theorem 3. *Under Assumption 1, the adversary can distinguish between data x, y with probability $1-\delta$ if:*

$$\rho(x,y) \geq \frac{3}{4Ln} \ln \frac{1}{\delta}. \tag{13}$$

Under Assumption 2, this becomes:

$$\rho(x,y) \geq \frac{3c}{2\kappa n} \ln \frac{1}{\delta}. \tag{14}$$

Consequently, either smoother likelihoods (*i.e.,* decreasing L), or a larger concentration on smoother likelihoods (*i.e.,* increasing c), both 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.

5 Examples satisfying our assumptions

In what follows we study, for different choices of likelihood and corresponding conjugate prior, what constraints must 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 are 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 $p_\theta(x) = h(x) \exp\left\{\eta_\theta^\top T(x) - A(\eta_\theta)\right\}$, where $h(x)$ is the base measure, η_θ is the distribution's natural parameter corresponding

to θ, $T(x)$ 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 $x, y \in \mathcal{S}$: $\left|\ln \frac{h(x)}{h(y)} + \eta_\theta^\top \left(T(x) - T(y)\right)\right| \leq L\rho(x,y)$. 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 $T(x)$ is scalar (*e.g.,* Bernoulli, exponential, and Laplace), this requirement simplifies to $\frac{|T(x)-T(y)|}{\rho(x,y)} \leq \frac{L}{\eta_\theta}$. 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. Here are some examples, whose proofs can be found in [8].

Lemma 3 (Exponential conjugate prior). *For the case of an exponential distribution $\mathcal{Exp}(\theta)$ with exponential conjugate prior $\theta \sim \mathcal{Exp}(\lambda)$, $\lambda > 0$ satisfies Assumption 2 with parameter $c = \lambda$ and metric $\rho(x, y) = |x - y|$.*

Lemma 4 (Laplace conjugate prior). *The Laplace distribution Laplace(θ) and Laplace conjugate prior $\theta \sim$ Laplace(μ, s, λ), $\mu \in \mathbb{R}$, $s \geq L$, $\lambda > 0$ satisfies Assumption 2 with parameters $c = \lambda$ and metric $\rho(x, y) = |x - y|$*

Lemma 5 (Beta-Binomial conjugate prior). *The Binomial distribution $\mathcal{Binom}(\theta, n)$, with Binomial prior $\theta \sim Beta(\alpha, \beta)$, $\alpha = \beta > 1$ satisfies Assumption 2 for $c = O(\alpha)$ and metric $\rho(x, y) = |x - y|$.*

Lemma 6 (Normal distribution). *The normal distribution $N(\mu, \sigma^2)$ with an exponential prior $\sigma^2 \sim \mathcal{Exp}(\lambda)$ satisfies Assumption 2 with parameter $c = \lambda$ and metric $\rho(x, y) = \left|x^2 - y^2\right| + 2\,|x - y|$.*

Lemma 7 (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 $\mathcal{S}$. We also let $\rho(x, y) \triangleq \sum_{k=1}^{K} \mathbb{I}\{x_k \neq y_k\}$ be the distance between x and y. If ϵ is the smallest probability assigned to any one sub-event, then Assumption 1 is satisfied with $L = \ln 1/\epsilon$.*

The above examples demonstrate that our assumptions are reasonable. In fact, for several of them we recover standard choices of prior distributions.

6 Conclusion

We have presented a unifying framework for private and secure inference in a Bayesian setting. Under simple but general assumptions, we have shown that Bayesian inference is both robust and private in a certain formal sense. In particular, our results establish that generalised differential privacy can be achieved while using only existing constructs in Bayesian inference. Our results merely

place concentration conditions on the prior. This allows us to use a general posterior sampling mechanism for responding to queries.

Due to its relative simplicity on top of non-private inference, our framework may thus serve as a fundamental building block for more sophisticated, general-purpose 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.

We have not examined how privacy concerns relate to learning. While larger c improves privacy, it also concentrates the prior so much that learning would be inhibited. Thus, c should be chosen to optimise the trade-off between privacy and learning. However, we leave this issue for future work.

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. This work was partially supported by the Marie Curie Project "Efficient Sequential Decision Making Under Uncertainty", grant No: 237816 and the FP7 STREP project "BEAT: Biometric Evaluation & Testing" grant No: 284989.

A Proofs of main theorems

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

$$\begin{aligned} d(p^n_\Theta(\{\, x_i \,\}), p^n_\Theta(\{\, y_i \,\})) &= n \textstyle\sum_{i=1}^n d(p_\Theta(x_i), p_\Theta(y_i)) \\ &\leq L \cdot n \textstyle\sum_{i=1}^n d(x_i, y_i) \ . \end{aligned}$$

This can be reduced from n to k if only k items differ since $d(p_\Theta(x_i), p_\Theta(y_i)) = 0$ if $x_i = y_i$.

For Assumption 2, the same argument shows that the Θ_L from Eq. (5) becomes $\Theta_{L \cdot n}$ (or $\Theta_{L \cdot k}$ for the k differing items case) for the product distribution. Hence, the same prior can be used to give the bound required by Eq. (6) if parameter $\frac{c}{n}$ (or $\frac{c}{k}$) is used.

Proof (Proof of Theorem 1). Let us now tackle claim (1.i). First, we can decompose the KL-divergence $D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right)$ into two parts:

$$\begin{aligned} \int_\Theta \ln \frac{\mathrm{d}\xi(\theta \mid x)}{\mathrm{d}\xi(\theta \mid y)} \, \mathrm{d}\xi(\theta) &= \int_\Theta \ln \frac{p_\theta(x)}{p_\theta(y)} \, \mathrm{d}\xi(\theta) + \int_\Theta \ln \frac{\phi(y)}{\phi(x)} \, \mathrm{d}\xi(\theta) \\ \leq \int_\Theta \left| \ln \frac{p_\theta(x)}{p_\theta(y)} \right| \mathrm{d}\xi(\theta) + \int_\Theta \ln \frac{\phi(y)}{\phi(x)} \, \mathrm{d}\xi(\theta) &\leq L\rho(x, y) + \left| \ln \frac{\phi(y)}{\phi(x)} \right| . \end{aligned} \tag{15}$$

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

$$\phi(y) = \int_\Theta p_\theta(y)\,\mathrm{d}\xi(\theta) \leq \exp(L\rho(x,y)) \int_\Theta p_\theta(x)\,\mathrm{d}\xi(\theta) = \exp(L\rho(x,y))\phi(x). \tag{16}$$

Combining this with (15) we obtain $D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) \leq 2L\rho(x,y)$.

Claim (1.ii) is dealt with similarly. Once more, we can break down the distance in parts. Let $\Theta_{[a,b]} \triangleq \Theta_b \setminus \Theta_a$. Then $\xi(\Theta_{[a,b]}) = \xi(\Theta_b) - \xi(\Theta_a) \leq e^{-ca}$, as $\Theta_b \supset \Theta_a$, while $\xi(\Theta_b) \leq 1$ and $\xi(\Theta_a) \geq 1 - e^{-ca}$ from Ass 2. We can partition Θ into uniform intervals $[(L-1)\alpha, L\alpha)$ of size $\alpha > 0$ indexed by L. We bound the divergence on each partition and sum over L.

$$\begin{aligned}
&D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) \\
\leq &\sum_{L=1}^{\infty} \left\{ \int_{\Theta_{[(L-1)\alpha, L\alpha)}} \left| \ln \frac{p_\theta(x)}{p_\theta(y)} \right| \mathrm{d}\xi(\theta) + \int_{\Theta_{[(L-1)\alpha, L\alpha]}} \ln \frac{\phi(y)}{\phi(x)}\,\mathrm{d}\xi(\theta) \right\} \\
\leq &2\rho(x,y)\alpha \sum_{L=1}^{\infty} L e^{-c(L-1)\alpha} = 2\rho(x,y)\alpha \left(1 - e^{-c\alpha}\right)^{-2},
\end{aligned} \tag{17}$$

via the geometric series. This holds for any size parameter $\alpha > 0$ and is convex for $\alpha > 0$, $c > 0$. Thus, there is an optimal choice for α that minimizes this bound. Differentiating w.r.t α and setting the result to 0 yields $\alpha^\star = \frac{\omega}{c}$ where ω is the unique non-zero solution to $e^\omega = 2\omega + 1$. The optimal bound is then $D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) \leq \frac{2\omega}{(1-e^{-\omega})^2} \cdot \frac{\rho(x,y)}{c}$ As the $\omega \approx 1.25643$ is the unique positive solution to $e^\omega = 2\omega + 1$, and we define $\kappa = \frac{2\omega}{(1-e^{-\omega})^2} \approx 4.91081$.

Proof (Proof of Theorem 2). For part (2.i), 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 1, we showed that $\phi(y) \leq \exp\{L\rho(x,y)\}\phi(x)$ for all $x, y \in \mathcal{S}$. From Eq. 2, 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.ii), note that from Theorem (1.ii) that the KL divergence of the posteriors under assumption is bounded by $\kappa\rho(x,y)/c$. Now, recall Pinsker's inequality [cf. 14]:

$$D\left(Q \| P\right) \geq \frac{1}{2} \left\| Q - P \right\|_1^2. \tag{18}$$

Using it, this bound yields: $|\xi(B \mid x) - \xi(B \mid y)| \leq \sqrt{\frac{1}{2} D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right)} \leq \sqrt{\kappa\rho(x,y)/2c}$

Proof (Proof of Lemma 2). We use the inequality due to Weissman et al. [24] on the ℓ_1 norm, which states that for any multinomial distribution p with m outcomes, the ℓ_1 deviation of the empirical distribution $\hat{p}_n$ satisfies: $\mathbb{P}(\|\hat{p}_n - p\|_1 \geq \epsilon) \leq (2^m - 2)e^{-\frac{1}{2}n\epsilon^2}$. The right hand side is bounded by $e^{m \ln 2 - \frac{1}{2}n\epsilon^2}$. Substituting $\epsilon = \sqrt{\frac{3}{n} \ln \frac{1}{\delta}}$:

$$\begin{aligned}\mathbb{P}(\|\hat{p}_n - p\|_1 \geq \sqrt{\frac{3}{n} \ln \frac{1}{\delta}}) &\leq e^{m \ln 2 - \frac{3}{2} \ln \frac{1}{\delta}} \\ &\leq e^{\log_2 \sqrt{\frac{1}{\delta}} \ln 2 - \frac{3}{2} \ln \frac{1}{\delta}} = e^{\frac{1}{2} \ln \frac{1}{\delta} - \frac{3}{2} \ln \frac{1}{\delta}} = \delta. \end{aligned} \tag{19}$$

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

Proof (Proof of Theorem 3). Recall that the data processing inequality states that, for any sub-algebra $\mathfrak{S}$:

$$\left\|Q_{|\mathfrak{S}} - P_{|\mathfrak{S}}\right\|_1 \leq \|Q - P\|_1 . \tag{20}$$

Using this and Pinsker's inequality (18) we get:

$$\begin{aligned}2L\rho(x, y) \geq 2L\epsilon &\geq D\left(\xi(\cdot \mid x) \| \xi(\cdot \mid y)\right) \\ &\geq \frac{1}{2} \|\xi(\cdot \mid x) - \xi(\cdot \mid y)\|_1^2 \geq \frac{1}{2} \left\|\xi_{|\mathfrak{S}}(\cdot \mid x) - \xi_{|\mathfrak{S}}(\cdot \mid y)\right\|_1^2 . \end{aligned} \tag{21}$$

On the other hand, due to (12) the adversary's ℓ_1 error in the posterior distribution is bounded by $\sqrt{\frac{3}{n} \ln \frac{1}{\delta}}$ with probability $1 - \delta$. Using the above inequalities, we can bound the error in terms of the distinguishability of the real dataset 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: $2\kappa\rho(x, y)/c \geq \frac{3}{n} \ln \frac{1}{\delta}$.

References

[1] Berger, J.O.: Statistical Decision Theory and Bayesian Analysis. Springer-Verlag (1985)

[2] Bickel, P.J., Doksum, K.A.: Mathematical Statistics: Basic Ideas and Selected Topics, vol. 1. Holden-Day Company (2001)

[3] Bousquet, O., Elisseeff, A.: Stability and generalization. Journal of Machine Learning Research 2(Mar), 499–526 (2002)

[4] Chatzikokolakis, K., Andres, M.E., Bordenabe, N.E., Palamidessi, C.: Broadening the scope of differential privacy using metrics. In: PET. pp. 82–102 (2013)

[5] Chaudhuri, K., Hsu, D.: Convergence rates for differentially private statistical estimation. In: ICML. (2012)

[6] Chaudhuri, K., Monteleoni, C., Sarwate, A.D.: Differentially private empirical risk minimization. Journal of Machine Learning Research 12(Mar), 1069–1109 (2011)

[7] DeGroot, M.H.: Optimal Statistical Decisions. John Wiley & Sons (1970)
[8] Dimitrakakis, C., Nelson, B., Mitrokotsa, A., Rubinstein, B.: Robust and private Bayesian inference. Technical report, arXiv:1306.1066 (2014)
[9] Duchi, J.C., Jordan, M.I., Wainwright, M.J.: Local privacy and statistical minimax rates. Technical report, arXiv:1302.3203 (2013)
[10] Dwork, C.: Differential privacy. In: ICALP 2006. pp. 1–12 (2006)
[11] Dwork, C., Lei, J.: Differential privacy and robust statistics. In: STOC. pp. 371–380 (2009)
[12] Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: TCC. pp. 265–284 (2006)
[13] Dwork, C., Smith, A.: Differential privacy for statistics: What we know and what we want to learn. Journal of Privacy and Confidentiality 1(2), 135–154 (2009)
[14] Fedotov, A.A., Harremoës, P., Topsoe, F.: Refinements of Pinsker's inequality. IEEE Transactions on Information Theory 49(6), 1491–1498 (2003)
[15] Grünwald, P.D., Dawid, A.P.: Game theory, maximum entropy, minimum discrepancy, and robust bayesian decision theory. The Annals of Statistics 32(4), 1367–1433 (2004)
[16] Hall, R., Rinaldo, A., Wasserman, L.: Differential privacy for functions and functional data. Journal of Machine Learning Research 14(Feb), 703–727 (2013)
[17] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A.: Robust Statistics: The Approach Based on Influence Functions. John Wiley and Sons (1986)
[18] Huber, P.J.: Robust Statistics. John Wiley and Sons (1981)
[19] McSherry, F., Talwar, K.: Mechanism design via differential privacy. In: FOCS. pp. 94–103 (2007)
[20] Mir, D.: Differentially-private learning and information theory. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops. pp. 206–210. ACM (2012)
[21] Norkin, V.: Stochastic Lipschitz functions. Cybernetics and Systems Analysis 22(2), 226–233 (1986)
[22] Rubinstein, B.I.P., Bartlett, P.L., Huang, L., , Taft, N.: Learning in a large function space: Privacy-preserving mechanisms for SVM learning. Journal of Privacy and Confidentiality 4(1) (2012)
[23] Wasserman, L., Zhou, S.: A statistical framework for differential privacy. Journal of the American Statistical Association 105(489), 375–389 (2010)
[24] Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., Weinberger, M.J.: Inequalities for the L1 deviation of the empirical distribution. Technical report, Hewlett-Packard Labs (2003)
[25] Williams, O., McSherry, F.: Probabilistic inference and differential privacy. In: NIPS. pp. 2451–2459 (2010)
[26] Xiao, Y., Xiong, L.: Bayesian inference under differential privacy. Technical report, arXiv:1203.0617 (2012)