Robust, Secure and Private Bayesian Inference

Christos Dimitrakakis  Blaine Nelson
Aikaterini Mitrokotsa  Benjamin Rubinstein

June 28, 2013

Abstract

This paper examines the robustness and privacy properties of Bayesian estimators under a general set of assumptions. These assumptions generalise the concept of differential privacy to arbitrary outcome spaces and distribution families. We demonstrate our results with a number of examples where they hold. We then prove general bounds on the change of the posterior distribution due to changes in the data. Finally, we prove finite sample bounds for privacy under a strong adversarial model.

1 Introduction

A holy grail for research in statistical learning is developing algorithms that are simultaneously efficient under benign noise, robust to adversarial manipulation, and whose inferences preserve privacy of sensitive training data. Intuitively, learning algorithms that are continuous mappings—meaning that their output (e.g., a classifier) does not significantly vary with small perturbations to the input (e.g., similar datasets)—enjoy robustness. Likewise, as small variations to training data only result in slightly different learned models, it would be difficult for an adversary to leverage knowledge of the learning process to uncover an unknown datum. As such, there is hope that robustness and privacy may be simultaneously achieved and perhaps are deeply linked. In this paper, we investigate general conditions that guarantee robustness and privacy in a Bayesian setting and demonstrate the connection between them.

We show that under mild assumptions on the smoothness of the likelihood function, any resulting posterior distribution is largely insensitive to changes in the data. This entails both a simple assumption on uniform smoothness and a weaker assumption that our prior is concentrated on parameters for which smoothness is large. In addition, we show that this achieves privacy in a strong sense that generalises the Dwork et al. [7] notion of differential privacy. More specifically, the amount of information that an adversary requires to differentiate between two different datasets is linearly bounded by the dataset distance
and exponentially depends on the concentration of the prior distribution on parameters for which the likelihood is smooth. We complement our exposition with a set of example distribution families and corresponding distances for which the assumptions hold.

**Paper organisation.** Section 2 discusses related work. Section 3 introduces the setting and our main assumptions. Section 4 proves results on robustness and gives a number of examples. Section 5 bounds the discriminatory power of an adversary. We conclude the paper with Section 6. Proofs of the main theorems are given in the appendix.

## 2 Related work

The foundation of this work can be traced to statistical decision theory in a Bayesian setting [1, 2], in which learning is cast as a statistical inference problem and decision-theoretic criteria are used as a basis for assessing, selecting and designing procedures. Using expected risk as a foundation, the *empirical risk minimisation* framework forms the core of statistical machine learning and *Bayes risk minimisation* in a Bayesian setting [1, 16]. However, expected risk is a poor criterion under adversarial settings.

For adversarial settings, the usual approach is to select procedures according to a minimax risk criterion in which nature acts adversarially. In the field of *Robust Statistics*, the minimax asymptotic bias of a procedure incurred within an $\epsilon$-contamination neighbourhood is used as a robustness criterion giving rise to the notion of a procedure’s *influence function* and *breakdown point* to characterise their robustness [11, 13]. In a Bayesian context, robustness appears in several guises including minimax risk, robustness of the posterior within $\epsilon$-contamination neighbourhoods, and robust priors [1]. In this context Grünwald and Dawid [9] demonstrated the link between robustness in terms of the minimax expected score of the likelihood function and the (generalized) maximum entropy principle.

Differential privacy, first proposed by Dwork et al. [7], 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. A randomised inference preserves *differential privacy* if perturbing one training instance results in small pointwise multiplicative change to the inference likelihood. Our notion of privacy for the Bayesian setting focuses on the affect of observational perturbations on the posterior in the same way, subsuming and generalising differential privacy to allow for wider deviation of posteriors under larger data perturbations. Similar generalisations are possible for standard differential privacy when multiple observation instances are unknown to the adversary [7].

A popular approach to admitting differential privacy is the *exponential mechanism* [14] which generalises the *Laplace mechanism* of adding Laplace noise to released statistics [7]. This releases a response with probability exponential in
a score function measuring distance to the non-private response. In our setting with exponential families, concentration of the prior has a similar effect to the score, however unlike the exponential mechanism, it is already a fundamental aspect of non-private inference. More recently Duchi et al. [4] provided information-theoretic bounds for private learning, by modeling the protocol of our interaction with the adversary as an arbitrary conditional distribution, rather than restricting it to specific mechanisms. These bounds can be seen as complementary to ours.

Little research in differential privacy has focused on the Bayesian paradigm. Williams and McSherry [18] applied probabilistic inference to improve the utility of differentially private releases by computing posteriors in a noisy measurement model.

Smoothness of the learning map, achieved for Bayesian inference here 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. [15] used the $\gamma$-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. [10] extended this technique to adding Gaussian process noise for differentially private release of infinite-dimensional functions lying in an RKHS.

Finally, Dwork and Lei [5] 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 [6].

In summary, our paper generalises the concept of differential privacy and formalises a simple, but general link between robustness and privacy in a Bayesian setting. We prove bounds on both robustness and the leakage of information which depend on or generalised notion of smoothness and on a concentration characteristic of the chosen distribution. We show that our assumptions are meaningful with a number of examples.

3 The setting

Assume that our observations lie in a set $\mathcal{S}$ equipped with a semi-metric\footnote{Meaning that the triangle inequality may not hold.} $\rho : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$ and that they are drawn from some distribution $P_{\theta}$ in a family $\mathcal{F}$. More specifically, we define a parameter set $\Theta$ indexing a family of probability measures $\mathcal{F}$ on $(\mathcal{S}, \mathcal{G}_S)$, where $\mathcal{G}_S$ is an appropriate $\sigma$-algebra on $\mathcal{S}$:

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

and where we use $p_{\theta}$ to denote the corresponding densities. In order to perform inference in the Bayesian setting, we define $\zeta$ to be a prior measure on $(\Theta, \mathcal{G}_\Theta)$, reflecting our subjective beliefs about which $\theta$ is more likely to be...
true, i.e. for any set $B \in \mathcal{S}$, $\zeta(B)$ is our prior belief that $\theta^* \in B$. In general, a posterior distribution after observing an $x \in \mathcal{S}$ is:

$$
\zeta(B \mid x) = \frac{\int_B p_\theta(x) \, d\zeta(\theta)}{\phi(x)} \quad x \in \mathcal{S},
$$

(3.2)

where:

$$
\phi(X) \triangleq \int_{\Theta} p_\theta(X) \, d\zeta(\theta), \quad \forall X \in \mathcal{S},
$$

(3.3)

is the corresponding marginal measure. In addition, we define a quasi-metric\(^2\) $D(\cdot \mid \cdot)$ on distributions. In this paper, we focus on the well-known KL-divergence:

$$
D(P \mid Q) = \int_{\mathcal{S}} \ln \frac{dP}{dQ} \, dP.
$$

(3.4)

We wish to show that if our distribution family is such that for the probabilities of two similar observation pairs $x, y$ are somehow close, then this family is robust. In order to formalise this notion, we introduce the following assumptions, which rely on a distance $d$. The first assumption states that the likelihood is smooth for all members of the family:

**Assumption 1** (Lipschitz continuity). There exists $L > 0$ such that, for any $\theta \in \Theta$:

$$
d(p_\theta(x), p_\theta(y)) \leq L p(x, y), \quad \forall x, y \in \mathcal{S}.
$$

(3.5)

It may be difficult for this assumption to hold uniformly over $\Theta$. Consequently, we shall relax it by only requiring that our prior probability $\zeta$ is concentrated in the parts of the family for which the likelihood is smoothest:

**Assumption 2** (Relaxed Lipschitz continuity). Let

$$
\Theta_0 \triangleq \left\{ \theta \in \Theta : \sup_{x, y \in \mathcal{S}} \{d(p_\theta(x), p_\theta(y)) - L p(x, y)\} \leq 0 \right\}
$$

(3.6)

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

$$
\zeta(\Theta_0) \geq 1 - \exp(-cL).
$$

(3.7)

This assumption is rather weak. In fact, we shall give examples of distribution families for which a bound on the value of the concentration constant $c$ can be determined analytically.

\(^2\)Meaning that symmetry does not hold.
In the sequel, we often use the absolute log-ratio as our distance function for applying these assumptions. It is defined as

\[
d_{lr}(p, q) = \begin{cases} 
0 & \text{if } p = q = 0 \\
\ln \left| \frac{p}{q} \right| & \text{otherwise}
\end{cases},
\]

which is a proper metric on \( \mathbb{R}_+ \times \mathbb{R}_+ \); i.e., it is non-negative, symmetric and it obeys the triangle inequality.

4 Robustness of the posterior distribution

As we shall see, the above assumptions provide guarantees on the robustness of the posterior. That is, for any possible pair of observations \((x, y)\), the KL divergence between the resulting posteriors is bounded by the distance between \(x\) and \(y\). To prove this claim, we use the absolute log-ratio distance in Eq. (3.8) for which (3.5) becomes:

\[
\ln \left| \frac{p(x)}{p(y)} \right| \leq L \rho(x, y), \quad x, y \in \mathcal{S}, \theta \in \Theta.
\]

Remark 1. This inequality extends the definition of differential privacy [7] to a Lipschitz continuity of the log-likelihood ratio. When \( \rho \) is the Hamming distance, this reduces to standard differential privacy with privacy parameter \( L \).

We now introduce our first main result. This states that the KL divergence between the posteriors arising from two different observations is bounded by the distance between the observations. Consequently, any distribution family \( \mathcal{F} \) and prior \( \xi \) satisfying our assumptions is robust, in the sense that the posterior does not change significantly with small changes to the data.

Theorem 1. If \( d(\cdot, \cdot) = d_{lr}(\cdot, \cdot) \), the absolute log-ratio distance,

(i) Under Assumption 1,

\[
D(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)) \leq 2L \rho(x, y)
\]

(ii) Under Assumption 2,

\[
D(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)) \leq 2L \rho(x, y)e^{-c} (1 - e^{-c})^{-2}.
\]

Note that the second claim bounds the KL divergence in terms of our 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.
Remark 2. In what follows we study for different choices of likelihood and corresponding conjugate prior, what constraints must be placed on the prior’s concentration in order to guarantee a desired level of privacy $L$. These case-studies follow closely 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 $e$ of $e$-differential privacy.

4.1 Application to exponential distributions

Suppose that we have $n$ data points drawn from an exponential distribution with a rate parameter $\theta > 0$, using an exponential prior with rate parameter $c > 0$, where the exponential distribution is given by $p_\theta(x) = \theta \exp(-\theta x)$ for $x \geq 0$. Because this prior is conjugate to the likelihood, it has a closed form as a gamma distribution: $\text{Gamma}(n + 1, c + \sum x_i)$.

To apply our assumptions, we first compute the absolute log-ratio distance for any $x_1$ and $x_2$ according to the exponential likelihood function:

$$d_{lr}(p_{\theta,n}(x_1), p_{\theta,n}(x_2)) = \theta |x_1 - x_2| .$$

Thus, under Assumption 2, using $\rho(x, y) = |x - y|$, the set of feasible parameters for any $L > 0$ is $\Theta_L = (0, L)$. Therefore the assumption requires that the prior adequately supports this range, but because the CDF at $L$ of the exponential prior with parameter $c > 0$ is simply given by $1 - \exp(-c L)$, every such prior satisfies the assumption. It follows that $c$ can be selected freely to obtain the desired result for Theorem 1.

4.2 Application to the Laplace Distribution

Now suppose the data is drawn from a Laplace distribution with a likelihood $p_{\mu,s}(x) = \frac{1}{2s} \exp\{-\frac{1}{s} |x - \mu|\}$. For any $x_1$ and $x_2$, the absolute log-ratio distance for this distribution can be bounded as

$$d_{lr}(p_{\mu,s}(x_1), p_{\mu,s}(x_2)) = \frac{1}{s} \|x_1 - \mu\| - \|x_2 - \mu\| \leq \frac{1}{s} \|x_1 - x_2\| ,$$

where the inequality follows from the triangle inequality applied to $\| \cdot \|$. Thus, if we use $\rho(x, y) = \|x - y\|$, the set of feasible parameters for Assumption 2 is $\mu \in \mathbb{R}$ and $s \geq L$. Again we can use an exponential prior with rate parameter $c > 0$ for the inverse scale, $\frac{1}{s}$, and any prior on $\mu$ to obtain the second part of Assumption 2. These similarities are not surprising considering that if $X \sim \text{Laplace}(\mu, s)$ then $\|X - \mu\| \sim \text{Exponential}(\frac{1}{s})$.

4.3 Application to Beta-Binomial

Here we consider data drawn from a binomial distribution with a beta prior on its proportion parameter, $\theta$. Thus, the likelihood and prior functions are

$$p_{\theta,n}(X = k) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} \quad \text{and} \quad \pi_\theta(\theta) = \frac{1}{B(a, b)} \theta^{a-1} (1 - \theta)^{b-1} ,$$
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. Here we consider the application of Assumption 2 to this beta-binomial distribution. For this purpose, we must quantify the parameter sets \( \Theta_L \) 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

\[
d_{lr}(p_{\theta_1}(k_1), p_{\theta_2}(k_2)) = \left| \Delta_n(k_1, k_2) + (k_1 - k_2) \ln \frac{\theta}{1 - \theta} \right|
\]

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. (3.6), 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

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

However, by the definition of \( \Delta_n(k_1, k_2) \), the ratio \( \Delta_n(k_1, k_2) / (k_1 - k_2) \) 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. (4.4), we therefore have, for the case \( L \geq \ln n \):

\[
\Theta_L = \left[ \left( 1 + \frac{c}{\pi} \right)^{-1}, \left( 1 + \frac{n}{\pi} \right)^{-1} \right] .
\]

Hence, the feasible set of parameters is an interval that is symmetric about \( \frac{1}{2} \) for all \( L \geq \ln n \). Because of the shape of this interval, it is natural to choose a beta prior with \( a = b > 1 \), which concentrates around \( \frac{1}{2} \). Thus, for any \( n, L, \) and \( c \), one can find a prior parameter \( a \) that meets the second condition of Assumption 2.

### 4.4 Application to the normal distribution

For the normal distribution, the assumption becomes:

\[
L \rho(x, y) \geq \frac{2|\mu - x - y| |x - y|}{2\sigma^2} \quad (4.5)
\]

It is easy to see that if we select \( \rho(x, y) = f(|x - y|) \) for some increasing \( f \) then (4.5) cannot be satisfied for any fixed \( L \). Indeed, choose \( x = y + 1 \) and \( y \to \infty \), the left side remains bounded, by the right side goes to infinity. This perhaps unsurprising as inference with a normal distribution is generally not robust [12].

7
For this reason, we must use another semi-metric. A suitable one can be found as follows. Taking the absolute log ratio of the Gaussian densities we have
\[
\frac{1}{2\sigma^2} \left| (x - \mu)^2 - (y - \mu)^2 \right| = \frac{1}{2\sigma^2} \left| x^2 - y^2 - 2\mu x + 2\mu y \right| \quad (4.6)
\]
\[
\leq \frac{1}{2\sigma^2} \left( |x^2 - y^2| + 2|\mu| |x - y| \right). \quad (4.7)
\]
\[
\leq \max \{ |\mu|, 1 \} \left( |x^2 - y^2| + 2 |x - y| \right). \quad (4.8)
\]
Consequently, we can set \( \rho(x, y) = |x^2 - y^2| + 2 |x - y| \) and \( L(\mu, \sigma) = \frac{\max \{ |\mu|, 1 \}}{2\sigma^2} \).

4.5 Application to discrete Bayesian networks

Consider a finite family \( \mathcal{F} = \{ P_\theta : \theta \in \Theta \} \) of discrete Bayesian networks on \( K \) variables. More specifically, each member \( P_\theta \) is a distribution on a finite (or countable) outcome space \( \mathcal{S} = \prod_{k=1}^K S_k \) and we write \( P_\theta(x) \) for the probability of any outcome \( x = (x_1, \ldots, x_K) \) in \( \mathcal{S} \). We also let \( \rho(x, y) \triangleq \sum_{k=1}^K \mathbb{1} \{ x_k \neq y_k \} \) be the distance between \( x \) and \( y \).

It is instructive to first examine the case where all variables are independent and we have a single observation. Then \( P_\theta(x) = \prod_{k=1}^K \theta_{k,x_k} \) and
\[
\ln \frac{P_\theta(x)}{P_\theta(y)} = \ln \prod_{k=1}^K \frac{\theta_{k,x_k}}{\theta_{k,y_k}} \leq \sum_{k=1}^K \ln \frac{\theta_{k,x_k}}{\theta_{k,y_k}} \mathbb{1} \{ x_k \neq y_k \} \leq \max_{i,j,k} \ln \frac{\theta_{k,i}}{\theta_{k,j}} \rho(x, y).
\] (4.9)

Consequently, if \( \epsilon \) is the smallest probability assigned to any one sub-event, then \( L > \ln 1/\epsilon \) and we should be placing most of our probability mass in \( \theta \) where \( \epsilon \) is not small. To extend this to the general case, where we consider sequences of observations \( x_{k,l}, y_{k,l} \). To take the network connectivity into account, let \( v \in \mathbb{N}^K \) be such that \( v_k = 1 + \deg(k) \) and define:
\[
\rho(x, y) \triangleq v^\top \delta(x, y), \quad \delta_k(x, y) \triangleq \sum_{l} \mathbb{1} \{ x_{k,l} \neq y_{k,l} \}. \quad (4.10)
\]

Using a similar argument to (4.9), it is easy to see that in this case \( \ln \frac{P_\theta(x)}{P_\theta(y)} \leq \ln \frac{1}{\epsilon} \cdot \rho(x, y) \).

4.6 General exponential families

Finally, we consider the general case of likelihoods in the exponential family with the form \( p_\theta(x) = h(x) \exp \{ \eta_\theta^\top T(x) - A(\eta_\theta) \} \), where \( h(x) \) is the base measure, \( \eta_\theta \) is the natural parameter for the distribution corresponding to \( \theta \),
$T(x)$ 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 $\Theta_L$ of Assumption 2 must satisfy

$$\frac{|h(x) - h(y) + \eta_\theta^T (T(x) - T(y))|}{p(x, y)} \leq L \quad \forall x, y \in \mathcal{S}.$$ 

If the left-hand side has an amenable form, then we can quantify the set $\Theta_L$ for which this requirement holds—this is exactly what was done for the beta-binomial. For distributions for which $h(x)$ is constant and $T(x)$ is scalar (e.g., Bernoulli, exponential, and Laplace), this requirement simplifies to

$$\frac{|T(x) - T(y)|}{p(x, y)} \leq \frac{L}{h \eta_\theta^T}.$$ 

One can then find the supremum of the left-hand side independent of $\eta_\theta$ thus yielding a simple formula for the feasible $L$ for any $\theta$.

5 Bounding posterior query complexity

An adversary wishing to breach privacy, needs distinguish $x$ from $y$. To do so, he has to decide whether our posterior is $\xi(\cdot | x)$ or $\xi(\cdot | y)$. In this section, we lower 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 [7].

Let us consider the case where the adversary can query sample $\theta_k \sim \xi(\cdot | x)$. This is arguably a very powerful query model, since normally an adversary is limited to observing marginals and moments. Then the adversary need only 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 data $x$ and data $y$.

Due to the sampling model, we first require a finite sample bound on the quality of the empirical distribution. The adversary could try and distinguish different posteriors by forming the empirical distribution on any sub-algebra $\mathcal{G}$. However, in general he cannot improve on the rate given by the following lemma, apart from the leading constant.

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

$$\left\| \hat{P}^n|_{\mathcal{G}} - P|_{\mathcal{G}} \right\|_1 \leq \sqrt{\frac{3}{n} \ln \frac{1}{\delta}}, \quad \forall \delta \in (0, 1). \quad (5.1)$$

Theorem 1 bounds the KL divergence between posteriors resulting from similar data. Thus, we can combine it with the above bound on the adversary’s
empirical error 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 2.** Under Assumption 1, the adversary requires:

\[
 n \geq \frac{3}{2L} e \ln \frac{1}{\delta} \quad (5.2)
\]

samples to distinguish \( x \) from \( y \) such that \( \rho(x, y) \geq \epsilon \) with probability at least \( 1 - \delta \). Under Assumption 2, this becomes:

\[
 n \geq \frac{3}{2L} e^\epsilon (1 - e^{-\epsilon})^2 \ln \frac{1}{\delta}. \quad (5.3)
\]

Consequently, a larger concentration \( c \) both increases the effort required by the adversary and reduces the sensitivity of our prior.

6 Conclusion

We have provided 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 secure in a certain sense. In particular, our results establish that differential privacy (and its generalisation) can be achieved while using only existing constructs in robust Bayesian inference. Rather than adding noise by introducing a new stage to the inference process (before/during/after learning) as is common for differentially private algorithms, our results merely place concentration conditions on the prior.

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 Bayesian inference. As a second 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 sampling 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.
A Collected proofs

Proof of Theorem 1. First, we can break down the distance in two parts.
\[
D (\xi (\cdot | x) \parallel \xi (\cdot | y)) = \int_{\Theta} \ln \frac{d\xi (\theta | x)}{d\xi (\theta | y)} d\xi (\theta) = \int_{\Theta} \ln \left( \frac{p_\theta (x)}{p_\theta (y)} \times \frac{\phi (y)}{\phi (x)} \right) d\xi (\theta)
\]
(A.1)
\[
= \int_{\Theta} \ln \frac{p_\theta (x)}{p_\theta (y)} d\xi (\theta) + \int_{\Theta} \ln \frac{\phi (y)}{\phi (x)} d\xi (\theta)
\]
(A.2)
\[
\leq \int_{\Theta} \left| \ln \frac{p_\theta (x)}{p_\theta (y)} \right| d\xi (\theta) + \int_{\Theta} \left| \ln \frac{\phi (y)}{\phi (x)} \right| d\xi (\theta)
\]
(A.3)
\[
\leq L \rho (x, y) + \left| \ln \frac{\phi (y)}{\phi (x)} \right|.
\]
(A.4)

Let us now tackle claim (i). From Ass. 1 we have that \( p_\theta (y) \leq \exp (L \rho (x, y)) p_\theta (x) \) for all \( \theta \) so
\[
\phi (y) = \int_{\Theta} p_\theta (y) d\xi (\theta) \leq \exp (L \rho (x, y)) \int_{\Theta} p_\theta (x) d\xi (\theta) = \exp (L \rho (x, y)) \phi (x).
\]
(A.5)

Combining this with (A.4) we obtain
\[
D (\xi (\cdot | x) \parallel \xi (\cdot | y)) \leq 2L \rho (x, y).
\]
(A.6)

Claim (ii) is handled similarly. Once more, we can break down the distance in parts. Let \( \Theta_{(a,b)} = \Theta \setminus \Theta_a \). Then \( \bar{\xi} (\Theta_{(a,b)}) = \bar{\xi} (\Theta_b) - \bar{\xi} (\Theta_a) \leq e^{-ca} \), as \( \Theta_B \supset \Theta_a \), while \( \bar{\xi} \leq 1 \) and Ass 2. Now, let us sum over \( L \):
\[
D (\xi (\cdot | x) \parallel \xi (\cdot | y)) \leq \sum_{L=1}^{\infty} \left\{ \int_{\Theta_{[L-1,L]}} \left| \ln \frac{p_\theta (x)}{p_\theta (y)} \right| d\xi (\theta) + \int_{\Theta_{[L-1,L]}} \left| \ln \frac{\phi (y)}{\phi (x)} \right| d\xi (\theta) \right\}
\]
\[
\leq 2L \rho (x, y) \sum_{L=1}^{\infty} Le^{-cl} = 2L \rho (x, y) e^{-c} (1 - e^{-c})^{-2},
\]
(A.7)
via the geometric series.

Proof of Lemma 1. We use the inequality due to Weissman et al. [17] on the \( \ell_1 \) norm, which states that for any multinomial distribution \( p \) with \( m \) outcomes, the \( \ell_1 \) deviation of the empirical distribution \( \hat{p}_n \) satisfies:
\[
P (|\hat{p}_n - p|_1 \geq \epsilon) \leq \left( \frac{m}{n} - 2 \right) e^{-\frac{\epsilon^2 m}{2}}.
\]
(A.8)
The right hand side is bounded by \( e^{m \ln 2 - \frac{1}{2} n \epsilon^2} \). Substituting \( \epsilon = \sqrt{\frac{3}{n} \ln \frac{1}{\delta}} \):
\[
P (|\hat{p}_n - p|_1 \geq \sqrt{\frac{3}{n} \ln \frac{1}{\delta}}) \leq e^{m \ln 2 - \frac{3}{2} n \ln \frac{1}{\delta}} \leq e^{\log_2 \sqrt{\frac{3}{n} \ln 2} - \frac{3}{2} n \ln \frac{1}{\delta}} = e^{\frac{1}{2} \ln 2 - \frac{3}{2} n \ln \frac{1}{\delta}} = \delta.
\]
(A.9)
where the second inequality follows from \( m \leq \log_2 \sqrt{1/\delta} \).
Proof of Theorem 2. Firstly, we recall that Pinsker’s inequality [see 8] states that:

$$D\left( Q \parallel P \right) \geq \frac{1}{2} \| Q - P \|_1^2.$$  \hspace{1cm} (A.10)

In addition, the data processing inequality states that, for any sub-algebra \(\mathcal{G}\):

$$\left\| Q_{\mathcal{G}} - P_{\mathcal{G}} \right\|_1 \leq \| Q - P \|_1.$$  \hspace{1cm} (A.11)

It now follows that

$$2Lp(x,y)^a \geq 2Le^a \geq D\left( \hat{\xi}(\cdot | x) \|| \xi(\cdot | y) \right) \geq \frac{1}{2} \| \hat{\xi}(\cdot | x) - \hat{\xi}(\cdot | y) \|_1^2$$ \hspace{1cm} (A.12)

$$\geq \frac{1}{2} \| \hat{\xi}_{\mathcal{G}}(\cdot | x) - \hat{\xi}_{\mathcal{G}}(\cdot | y) \|_1^2,$$ \hspace{1cm} (A.13)

from Pinkev’s and the data processing inequalities. On the other hand, due to (5.1) the adversary’s \(\ell_1\) error is bounded by \(\sqrt{\frac{3}{n} \ln \frac{1}{\delta}}\) with probability \(1 - \delta\). Consequently, the adversary requires that \(\frac{3}{n} \ln \frac{1}{\delta} \leq 2Le\). Re-arranging, we obtain the stated bound.

The second case is identical, but now the adversary requires \(\frac{3}{n} \ln \frac{1}{\delta} \leq 2Le^{-c}(1 - e^{-c})^{-2}\). Re-arranging, we obtain the stated bound. \(\square\)

References


