Robust and Private Bayesian Inference

Christos Dimitrakakis Blaine Belson
Aikaterini Mitrokotsa Benjamin Rubinstein

April 1, 2014

Abstract

We examine the robustness and privacy properties of Bayesian inference under assumptions on the prior, but without any 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. 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. 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 $\xi(\theta|x)$. (iv) $\mathscr{A}$ performs repeated queries to $\mathscr{B}$. (v) $\mathscr{A}$ responds by sampling from the posterior $\xi(\theta|x)$.

We show that if $\mathcal{F}_\Theta$ or ξ is chosen appropriately, the resulting mechanism satisfies generalized differential privacy and indistinguishability properties. The main idea we pursue is that robustness and privacy are inherently linked through smoothness. Learning algorithms that are smooth

mappings—their output (*e.g.,* a spam filter) does not significantly vary with small perturbations to their input (*e.g.,* similar training corpora)—enjoy robustness. Intuitively under smoothness, training outliers have reduced influence while it is also difficult for an adversary to leverage knowledge of the learning process to discover unknown information about the data. This suggests that robustness and privacy may be simultaneously achieved and perhaps are deeply linked. We show that under mild assumptions, this is indeed true for the posterior distribution.

The study of learning, security, robustness and privacy, and their relationships, is timely. Interest in adversarial learning is accelerating Joseph et al. (2013) while differential privacy has brought data privacy onto firm theoretical footing Dwork et al. (2006); McSherry & Talwar (2007); Duchi et al. (2013). In practice, security and privacy online are in tension with learning and are of growing economic and societal concern. Our works aims towards a unified understanding of learning in adversarial environments.

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 some classical examples of conjugate distributions where our assumptions hold.

Paper organisation. Section 1.1 discusses related work. Section 2 specifies the problem setting and our main assumptions. Section 3 proves results on robustness of Bayesian learning with a number of examples. Section 4 bounds the ability of the adversary to discriminate datasets. Examples of distributions for which our assumptions hold are given in Section 5. We conclude the paper with Section 6. Proofs of the main theorems are given

in the appendix, while those of non-essential lemmas are given in the supplement.

1.1 Related Work

In Bayesian statistical decision theory DeGroot (1970); Berger (1985); Bickel & Doksum (2001), 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 Huber (1981); Hampel et al. (1986). 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 & Dawid (2004) 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. (2006), 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 small a change to the probabilities of the mechanism.

A popular approach for achieving differential privacy is the *exponential mechanism* McSherry & Talwar (2007) which generalises the *Laplace mechanism* of adding Laplace noise to released statistics Dwork et al. (2006). This 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 Chaudhuri et al. (2011), is to alter the inferential procedure itself, in that case by adding a random term to the primal objective. Unlike previous studies, our mechanisms do not require modification to the underlying learning framework.

In a different direction, Duchi et al. (2013) 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. These bounds can be seen as complementary to ours.

Little research in differential privacy has focused on the Bayesian paradigm. Williams & McSherry (2010) 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 Bousquet & Elisseeff (2002). Rubinstein et al. (2012) 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. (2013) extended this technique to adding Gaussian process noise for differentially private release of infinite-dimensional functions lying in an RKHS.

Finally, Dwork & 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 & Smith (2009). 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 statistical findings to an untrusted third party ($\mathscr{A}$). While $\mathscr{B}$ wants to convey useful statistical information to any queries, but without revealing private information about the original data (*e.g.*, how many people suffer from a disease or vote for a particular party). In so doing, $\mathscr{B}$ must also preserve the *local* privacy of users represented in the dataset. This requires finding a query response mechanism for communicating information 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[1] $\rho : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$. Using pseudo-metrics, we considerably generalise previous work on differential privacy, which considers only the special case of Hamming distance.

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 $\mathcal{S}$:

$$\mathcal{F}_\Theta \triangleq \{ P_\theta : \theta \in \Theta \}, \tag{2.1}$$

and where we use p_θ to denote the corresponding densities[2] 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:

$$\xi(B \mid x) = \frac{\int_B p_\theta(x) \, \mathrm{d}\xi(\theta)}{\phi(x)} \ , \tag{2.2}$$

where ϕ is the corresponding marginal density given by:

$$\phi(x) \triangleq \int_\Theta p_\theta(x) \, \mathrm{d}\xi(\theta) \ . \tag{2.3}$$

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

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

[2] *I.e.,* the Radon-Nikodym derivative of P_θ with respect to some dominating measure ν

Privacy. We first recall the idea of differential privacy Dwork (2006). 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 we use the Hamming distance $\rho(x, y) = \sum_{i=1}^n \mathbb{I}\{x_i \neq y_i\}$, this definition is analogous to standard (ϵ, δ)-differential privacy. In fact, when considering only $(\epsilon, 0)$- differential privacy or $(0, \delta)$-privacy, it is an equivalent notion.*[3]

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 of the true dataset by querying

[3]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 this requirement 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 $x, y \in \mathcal{S}$ have similar probabilities, then its posterior distributions 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 $\theta \in \Theta$:*

$$d(p_\theta(x), p_\theta(y)) \leq L\rho(x,y), \qquad \forall x, y \in \mathcal{S} \ . \tag{2.4}$$

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 (Relaxed Lipschitz continuity)**.** *Let $d(\cdot,\cdot)$ be a metric on $\mathbb{R}$ and let*

$$\Theta_L \triangleq \left\{\theta \in \Theta : \sup_{x,y\in\mathcal{S}} \left\{d(p_\theta(x), p_\theta(y)) - \frac{a}{b} L\rho(x,y)\right\} \leq 0\right\} \tag{2.5}$$

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 $L \geq 0$:

$$\xi(\Theta_L) \geq 1 - \exp(-cL). \tag{2.6}$$

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 several example distribution families.

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

$$d(a,b) \triangleq \begin{cases} 0 & \text{if } a = b = 0 \\ \left|\ln \frac{a}{b}\right| & \text{otherwise} \end{cases}, \tag{2.7}$$

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:*

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

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:

$$D\left(P \parallel Q\right) = \int_S \ln \frac{\mathrm{d}P}{\mathrm{d}Q} \, \mathrm{d}P \ . \tag{3.1}$$

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* (2.7)*, 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,

$$D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) \leq 2L\rho(x, y) \tag{3.2}$$

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

$$D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) \leq \frac{\kappa}{c} \cdot \rho(x, y) \tag{3.3}$$

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

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 $x, y \in \mathcal{S}$, $B \in \mathfrak{S}_\Theta$:

$$\xi(B \mid x) \leq \exp\{2L\rho(x, y)\}\xi(B \mid y) \tag{4.1}$$

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 different θ_k sampled from the posterior.

This sample-based 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 $q_k(\theta_k)$.

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 $q_k(\theta) = \theta$.

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**

As a further illustration, we provide the example of 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 Dwork et al. (2006).

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 $1 - \delta$:*

$$\left\| \hat{P}^n_{|\mathfrak{S}} - P_{|\mathfrak{S}} \right\|_1 \leq \sqrt{\frac{3}{n} \ln \frac{1}{\delta}}. \tag{4.2}$$

Of course, the adversary could choose any arbitrary estimator ψ to guess x. Appendix A describes how one might apply Le Cam's method to obtain lower bounds rates in this case. We defer a detailed discussion of this issue to future work.

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:*

$$\rho(x, y) \geq \frac{3}{4Ln} \ln \frac{1}{\delta}. \tag{4.3}$$

Under Assumption 2, this becomes:

$$\rho(x, y) \geq \frac{3c}{2\kappa n} \ln \frac{1}{\delta}. \tag{4.4}$$

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.

First consider exponential families, of the form

$$p_\theta(x) = h(x) \exp\left\{ \eta_\theta^\top T(x) - A(\eta_\theta) \right\},$$

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.

Lemma 3 (Exponential conjugate prior)**.** *The exponential distribution $\mathcal{E}\mathcal{X}\mathcal{P}(\theta)$ with exponential conjugate prior $\theta \sim \mathcal{E}\mathcal{X}\mathcal{P}(\lambda)$, $\lambda > 0$ satisfies Assumption 2 with parameter $c = \lambda$ and metric $\rho(x,y) = |x - y|$.*

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

Lemma 5 (Beta-Binomial conjugate prior)**.** *The Binomial distribution $\mathcal{B}\mathcal{I}\mathcal{N}\mathcal{O}\mathcal{M}(\theta, n)$, 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 $\sigma^2 \sim \mathcal{E}\mathcal{X}\mathcal{P}(\lambda)$ satisfies Assumption 2 with parameter $c = \lambda$ and metric $\rho(x,y) = \left|x^2 - y^2\right| + 2\left|x - y\right|$.*

Lemma 7 (Discrete Bayesian networks)**.** *Consider a family $\mathcal{F}_\Theta = \{ P_\theta : \theta \in \Theta \}$ of discrete Bayesian networks on K variables. 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 $L = \ln 1/\epsilon$.*

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 private in a certain 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 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.

A The Le Cam Method

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 dataset 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:

$$\Xi \triangleq \{ \xi(\cdot \mid x) : x \in \mathcal{S} \}, \tag{A.1}$$

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 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 Le Cam, we extend ρ to subsets of $\mathcal{S}$ so that

$$\rho(A, B) \triangleq \inf \{ \rho(x, y) : x \in A, y \in B \}, \qquad A, B \subset \mathcal{S}. \tag{A.2}$$

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

Lemma 8 (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:*

$$\sup_{x \in \mathcal{S}} \mathbb{E}_{\xi}(\rho(\psi, x) \mid x) \geq \delta \sup_{\xi_i \in co(\Xi_i)} \|\xi_1 \wedge \xi_2\|. \tag{A.3}$$

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 pretty much directly to obtain results for specific families of posteriors. As shown by *e.g.,* Yu (1997), even in simple scenarios the lower bound on the minimax estimation rate is $O(n^{-1/2})$.

B Proofs of examples

Proof of Lemma 3. We first compute the absolute log-ratio distance for any x_1 and x_2 according to the exponential likelihood function:

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

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 the prior to adequately support this range, but because the CDF at L of the exponential prior with parameter $\lambda > 0$ is simply given by $1 - \exp(-\lambda L)$, every such prior satisfies the assumption with $c = \lambda$. □

Proof of Lemma 4. For any x_1 and x_2, the absolute log-ratio distance for this distribution can be bounded as

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

where the inequality follows from the triangle inequality 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 an exponential prior with rate parameter $lambda > 0$ for the inverse scale, $\frac{1}{s}$, and any prior on μ to obtain the second part of Assumption 2. Every such prior satisfies the assumption with $c = \lambda$. These similarities are not surprising considering that if $X \sim Laplace(\mu, s)$ then $\|X - \mu\| \sim Exponential(\frac{1}{b})$. □

Proof of Lemma 5. 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

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

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. (2.5), 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

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

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. (B.1), 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: $\xi(\Theta_L) = 1 - \xi(\Theta_L^{\complement})$ where $\Theta_L^{\complement}$ is the complement of Θ_L. We selected $\alpha = \beta$, so $1 - \xi(\Theta')$ is composed of two symmetric intervals: $\left[0, (1 + \frac{e^L}{n})^{-1}\right)$ and $\left((1 + \frac{n}{e^L})^{-1}, 1\right]$. In addition, 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:

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

Since $\alpha > 1$ it holds that for all $x \in [0, z]$:

$$(1-x)^{\alpha-1} \leq 1, \qquad z < \frac{1}{2}.$$

This is bounded above by simply appyling the max bound for integrals.

$$\begin{aligned}\frac{2}{B(\alpha, \alpha)} \cdot \int_0^z x^{\alpha-1}(1-x)^{\alpha-1}dx &< \frac{2}{B(\alpha, \alpha)} \int_0^z x^{\alpha-1}dx \\ &= \frac{2}{B(\alpha, \alpha)} \frac{1}{\alpha} \cdot z^{\alpha}\end{aligned}$$

If we use $z = (1 + \frac{e^L}{n})^{-1}$ i.e. the desired upper limit we have:

$$\frac{2}{B(\alpha, \alpha)} \frac{1}{\alpha} \cdot z^{\alpha} = \frac{2}{B(\alpha, \alpha)} \cdot \frac{1}{\alpha} \cdot n^{\alpha} \cdot (n + e^L)^{-\alpha}$$

Finally, we have that $n > 0$ and hence, $(n + e^L)^{-\alpha} < e^{-\alpha L}$ so that we have: $\frac{2}{B(\alpha,\alpha)} \cdot \frac{1}{\alpha} \cdot n^{\alpha} \cdot (n + e^L)^{-\alpha} < \frac{2}{B(\alpha,\alpha)} \cdot \frac{1}{\alpha} \cdot n^{\alpha} \cdot e^{-\alpha L}$. We want to upper bound this by e^{-cL}. Solving for c, we obtain

$$c \leq \ln(1/v) + \alpha.$$

where $v = \frac{2}{B(\alpha,\alpha)} \cdot \frac{1}{\alpha} \cdot n^{\alpha}$.

□

Proof of Lemma 6. For the normal distribution (2.5) requires: $2L\rho(x,y)\sigma^2 \geq |2\mu - x - y|\,|x - y|$. Taking the absolute log ratio of the Gaussian densities we have

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

Consequently, we can set $\rho(x,y) = \left|x^2 - y^2\right| + 2\,|x-y|$ and $L(\mu,\sigma) = \frac{\max\{\,|\mu|,1\,\}}{2\sigma^2}$. It is easy to see that the normal distribution with an exponential prior on its variance satisfies the assumptions. □

Proof of Lemma 7. 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

$$\begin{aligned}\left|\ln\frac{P_\theta(x)}{P_\theta(y)}\right| = \left|\ln\prod_{k=1}^K\frac{\theta_{k,x_k}}{\theta_{k,y_k}}\right| &\leq \sum_{k=1}^K\left|\ln\frac{\theta_{k,x_k}}{\theta_{k,y_k}}\right|\mathbb{I}\left\{x_k\neq y_k\right\} \\ &\leq \max_{i,j,k}\left|\ln\frac{\theta_{k,i}}{\theta_{k,j}}\right|\rho(x,y). \qquad \text{(B.2)}\end{aligned}$$

Consequently, if ϵ is the smallest probability assigned to any one sub-event, then $L > \ln 1/\epsilon$.

In the general case, we have observations sequences $x_{k,t}, y_{k,t}$ and dependent variables. To take the network connectivity into account, let $v \in \mathbb{N}^K$ be such that $v_k = 1 + \deg(k)$ and define: $\rho(x,y) \triangleq v^\top \delta(x,y)$ and $\delta_k(x,y) \triangleq \sum_t \mathbb{I}\left\{x_{k,t} \neq y_{k,t}\right\}$. Using a similar argument to (B.2), it is easy to see that in this case $|\ln\frac{P_\theta(x)}{P_\theta(y)}| \leq \ln\frac{1}{\epsilon}\cdot\rho(x,y)$. □

C Collected Proofs

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

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

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. (2.5) 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. (2.6) if parameter $\frac{c}{n}$ (or $\frac{c}{k}$) is used. □

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

$$\begin{aligned} D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) &= \int_\Theta \ln \frac{\mathrm{d}\xi(\theta \mid x)}{\mathrm{d}\xi(\theta \mid y)} \,\mathrm{d}\xi(\theta) \\ &= \int_\Theta \ln \frac{p_\theta(x)}{p_\theta(y)} \,\mathrm{d}\xi(\theta) + \int_\Theta \ln \frac{\phi(y)}{\phi(x)} \,\mathrm{d}\xi(\theta) \\ &\leq \int_\Theta \left| \ln \frac{p_\theta(x)}{p_\theta(y)} \right| \,\mathrm{d}\xi(\theta) + \int_\Theta \ln \frac{\phi(y)}{\phi(x)} \,\mathrm{d}\xi(\theta) \\ &\leq L\rho(x,y) + \left| \ln \frac{\phi(y)}{\phi(x)} \right|. \end{aligned} \tag{C.1}$$

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

$$\begin{aligned} \phi(y) &= \int_\Theta p_\theta(y) \,\mathrm{d}\xi(\theta) \\ &\leq \exp(L\rho(x,y)) \int_\Theta p_\theta(x) \,\mathrm{d}\xi(\theta) \\ &= \exp(L\rho(x,y))\phi(x). \end{aligned} \tag{C.2}$$

Combining this with (C.1) we obtain

$$D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) \leq 2L\rho(x,y). \tag{C.3}$$

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 thus partition Θ into disjoint sets corresponding to uniformly sized intervals $[(L-1)\alpha, L\alpha)$ of size $\alpha > 0$ indexed by L. We bound the divergence on

each partition and sum over L.

$$
\begin{aligned}
& D\left(\xi(\cdot \mid x) \parallel \xi(\cdot \mid y)\right) \\
\leq & \sum_{L=1}^{\infty} \left\{ \int_{\Theta_{[(L-1)\alpha, L\alpha)}} \left| \ln \frac{p_\theta(x)}{p_\theta(y)} \right| \mathrm{d}\xi(\theta) \right. \\
& \qquad \left. + \int_{\Theta_{[(L-1)\alpha, L\alpha]}} \ln \frac{\phi(y)}{\phi(x)} \, \mathrm{d}\xi(\theta) \right\} \\
\leq & 2\rho(x,y)\alpha \sum_{L=1}^{\infty} L e^{-c(L-1)\alpha} \\
= & 2\rho(x,y)\alpha \left(1 - e^{-c\alpha}\right)^{-2}, \qquad \text{(C.4)}
\end{aligned}
$$

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 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. 2.2, we can then combine these to bound the posterior of any $B \in \mathfrak{S}_\Theta$ as follows for all $x, y \in \mathcal{S}$:

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

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. Fedotov et al., 2003):

$$
D\left(Q \| P\right) \geq \frac{1}{2} \left\| Q - P \right\|_1^2 . \qquad \text{(C.5)}
$$

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 of Lemma 2. 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 $\hat{p}_n$ satisfies:

$$\mathbb{P}(\|\hat{p}_n - p\|_1 \geq \epsilon) \leq (2^m - 2) e^{-\frac{1}{2} n \epsilon^2}. \tag{C.6}$$

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

$$\begin{aligned} \mathbb{P}(\|\hat{p}_n - p\|_1 \geq \sqrt{\frac{3}{n} \ln \frac{1}{\delta}}) &\leq e^{m \ln 2 - \frac{3}{2} \ln \frac{1}{\delta}} \\ &\leq e^{\log_2 \sqrt{\frac{1}{\delta}} \ln 2 - \frac{3}{2} \ln \frac{1}{\delta}} = e^{\frac{1}{2} \ln \frac{1}{\delta} - \frac{3}{2} \ln \frac{1}{\delta}} = \delta. \end{aligned} \tag{C.7}$$

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

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

$$\left\| Q_{|\mathfrak{S}} - P_{|\mathfrak{S}} \right\|_1 \leq \|Q - P\|_1 . \tag{C.8}$$

Using this and Pinsker's inequality (C.5) we get:

$$\begin{aligned} 2L\rho(x,y) \geq 2L\epsilon &\geq D\left(\xi(\cdot \mid x) \| \xi(\cdot \mid y)\right) \\ &\geq \frac{1}{2} \|\xi(\cdot \mid x) - \xi(\cdot \mid y)\|_1^2 \\ &\geq \frac{1}{2} \left\| \xi_{|\mathfrak{S}}(\cdot \mid x) - \xi_{|\mathfrak{S}}(\cdot \mid y) \right\|_1^2 . \end{aligned} \tag{C.9}$$

On the other hand, due to (4.2) 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:

$$4L\rho(x,y) \geq \frac{3}{n} \ln \frac{1}{\delta}. \tag{C.10}$$

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

$$2\kappa\rho(x,y)/c \geq \frac{3}{n} \ln \frac{1}{\delta}. \tag{C.11}$$

□

References

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

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

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

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

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

Duchi, John C., Jordan, Michael I., and Wainwright, Martin J. Local privacy and statistical minimax rates. Technical Report 1302.3203, arXiv, 2013.

Dwork, Cynthia. Differential privacy. In *ICALP*, pp. 1–12, 2006.

Dwork, Cynthia and Lei, Jing. Differential privacy and robust statistics. In *STOC*, pp. 371–380, 2009.

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

Dwork, Cynthia, McSherry, Frank, Nissim, Kobbi, and Smith, Adam. Calibrating noise to sensitivity in private data analysis. In *TCC*, pp. 265–284, 2006.

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

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

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

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

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

Joseph, Anthony D., Laskov, Pavel, Roli, Fabio, Tygar, J. Doug, and Nelson, Blaine. Machine Learning Methods for Computer Security (Dagstuhl Perspectives Workshop 12371). *Dagstuhl Manifestos*, 3(1):1–30, 2013. ISSN 2193-2433. doi: http://dx.doi.org/10.4230/DagMan.3.1.1. URL `http://drops.dagstuhl.de/opus/volltexte/2013/4356`.

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

McSherry, Frank and Talwar, Kunal. Mechanism design via differential privacy. In *FOCS*, pp. 94–103, 2007.

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

Weissman, Tsachy, Ordentlich, Erik, Seroussi, Gadiel, Verdu, Sergio, and Weinberger, Marcelo J. Inequalities for the L1 deviation of the empirical distribution. Technical report, Hewlett-Packard Labs, 2003.

Williams, Oliver and McSherry, Frank. Probabilistic inference and differential privacy. In *NIPS*, pp. 2451–2459, 2010.

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