Privacy for Free: Posterior Sampling and Stochastic Gradient Monte Carlo

Yu-Xiang Wang[†] YUXIANGW@CS.CMU.EDU
Stephen E. Fienberg[♯,†] FIENBERG@STAT.CMU.EDU
Alexander J. Smola[†,‡] ALEX@SMOLA.ORG
[†]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[♯]Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[‡]Marianas Labs Inc., Pittsburgh, PA 15213, USA

Abstract

We consider the problem of Bayesian learning on sensitive datasets and present two simple but somewhat surprising results that connect Bayesian learning to "differential privacy", a cryptographic approach to protect individual-level privacy while permitting database-level utility. Specifically, we show that under standard assumptions, getting one sample from a posterior distribution is differentially private "for free"; and this sample as a statistical estimator is often consistent, near optimal, and computationally tractable. Similarly but separately, we show that a recent line of work that use stochastic gradient for Hybrid Monte Carlo (HMC) sampling also preserve differentially privacy with minor or no modifications of the algorithmic procedure at all, these observations lead to an "anytime" algorithm for Bayesian learning under privacy constraint. We demonstrate that it performs much better than the state-of-the-art differential private methods on synthetic and real datasets.

1. Introduction

Bayesian models have proven to be one of the most successful classes of tools in machine learning. It stands out as a principled yet conceptually simple pipeline for combining expert knowledge and statistical evidence, modeling with complicated dependency structures and harnessing uncertainty by making probabilistic inferences (Geman & Geman, 1984; Gelman et al., 2014). In the past few decades, the Bayesian approach has been intensively used in modeling speeches (Rabiner, 1989), text documents (Blei et al., 2003), images/videos (Fei-Fei & Perona, 2005), social networks (Airoldi et al., 2009), brain activity (Penny et al., 2011), and is often considered gold standard in many of these application domains. Learning a Bayesisan model typically involves sampling from a posterior distribution, therefore the learning process is inherently randomized.

Differential privacy (DP) is a cryptography-inspired notion of privacy (Dwork, 2006; Dwork et al., 2006). It is designed to provide a very strong form of protection of individual user's private information and at the same time allow data analyses to be conducted with proper utility. Any algorithm that preserves differential privacy must be appropriately randomized too. For instance, one can differential-privately release the average salary of Californian males by adding a Laplace noise proportional to the sensitivity of this figure upon small perturbation of the data sample.

In this paper, we connect the two seemingly unrelated concepts by showing that under standard assumptions, the intrinsic randomization in the Bayesian learning can be exploited to obtain a degree of differential privacy. In particular, we show that:

- Any algorithm that produces a single sample from the exact (or approximate) posterior distribution of a Bayesian model with bounded log-likelihood is ϵ (or (ϵ, δ))-differentially private[1]. By the classic results in asymptotic statistics (Le Cam, 1986; Van der Vaart, 2000), we show that this posterior sample is a consistent estimator whenever the Bayesian model is consistent; and near optimal whenever standard regularity conditions of the maximum likelihood estimate hold.
- The popular large-scale sampler Stochastic Gradient Langevin Dynamics (Welling & Teh, 2011) and extensions, e.g. Ahn et al. (2012); Chen et al. (2014); Ding et al. (2014) obey (ϵ, δ)-differentially private with no algorithmic changes when the stepsize is chosen to

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

[1] The same observation appeared earlier in Mir (2013) and Dimitrakakis et al. (2014) under slightly different regimes and assumptions (see Appendix F for details).

be small. This gives us a procedure that can potentially output many (correlated) samples from an approximate posterior distribution.

These simple yet interesting findings make it possible for differential privacy to be explicitly considered when designing Bayesian models, and for Bayesian posterior sampling to be used as a valid DP mechanism. We demonstrate empirically that these methods work as well as or better than the state-of-the-art differential private empirical risk minimization (ERM) solvers using objective perturbation (Chaudhuri et al., 2011; Kifer et al., 2012).

The results presented in this paper are closely related to a number of previous work, e.g., McSherry & Talwar (2007); Mir (2013); Bassily et al. (2014); Dimitrakakis et al. (2014). We invite readers to refer to our full paper (Wang et al., 2015), or refer to Appendix F in the supplementary document.

2. Notations and Preliminary

Throughout the paper, we assume data point $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\theta} \in \Theta$ is the model. This can be the finite dimensional parameter of a single exponential family model or a collection of these in a graphical model, or a function in a Hilbert space or other infinite dimensional objects if the model is nonparametric. $\pi(\boldsymbol{\theta})$ denotes a prior belief of the model parameters and $p(\boldsymbol{x}|\boldsymbol{\theta})$ and $\ell(\boldsymbol{x}|\boldsymbol{\theta})$ are the likelihood and log-likelihood of observing data point x given model parameter $\boldsymbol{\theta}$. If we observe $X = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, the posterior

$$\pi(\boldsymbol{\theta}|X) = \frac{\pi(\boldsymbol{\theta}) \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\theta})}{\int \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\pi}$$

denotes the updated belief conditioned on the observed data. Learning Bayesian models correspond to finding the mean or mode of the posterior distribution, but often, the entire distribution is treated as the output, which provides much richer information than just a point estimator. In particular, we get error bars of the estimators for free (credibility intervals).

Ignoring the philosophical disputes of Bayesian methods for the moment, practical challenges of Bayesian learning are often computational. As the models get more complicated, often there is not a closed-form expression for the posterior. Instead, we often rely on Markov Chain Monte Carlo methods, e.g., Metropolis-Hastings algorithm (Hastings, 1970) to generate samples. This is often prohibitively expensive when the data is large. One recent approach to scale up Bayesian learning is to combine stochastic gradient estimation as in Robbins & Monro (1951) and Monte Carlo methods that simulates stochastic differential equations, e.g. Neal (2011). These include Stochastic Gradient Langevin dynamics (SGLD) (Welling & Teh, 2011), Stochastic Gradient Fisher scoring (SGFS) (Ahn et al., 2012), Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) (Chen et al., 2014) as well as more recent Stochastic Gradient Nosé-Hoover Thermostat (SGNHT) (Ding et al., 2014). We will describe them with more details and show that these series of tools provide differential privacy as a byproduct of using stochastic gradient and requiring the solution to not collapse to a point estimate.

2.1. Differential privacy

Now we will talk about what we need to know about differential privacy. Let the space of data be $\mathcal{X}$ and data points $X, Y \in \mathcal{X}^n$. Define $d(X, Y)$ to be the edit distance or Hamming distance between data set X and Y, for instance, if X and Y are the same except one data point, then $d(X, Y) = 1$.

Definition 1. (Differential Privacy) We call a randomized algorithm $\mathcal{A}$ (ϵ, δ)-differentially private with domain $\mathcal{X}^n$ if for all measurable set $S \subset \text{Range}(\mathcal{A})$ and for all $X, Y \in \mathcal{X}^n$ such that $d(X, Y) \leq 1$, we have

$$\mathbb{P}(\mathcal{A}(X) \in S) \leq \exp(\epsilon)\mathbb{P}(\mathcal{A}(Y) \in S) + \delta.$$

If $\delta = 0$, then $\mathcal{A}$ is the called ϵ-differential private.

This definition naturally prevents linkage attacks and the identification of individual data from adversaries having arbitrary side information and infinite computational power.

There are several interesting properties of differential privacy that we will exploit here. Firstly, the definition is closed under post-processing.

Lemma 1 (Post-processing immunity)**.** *If $\mathcal{A}$ is an (ϵ, δ)-DP algorithm, $\mathcal{B} \circ \mathcal{A}$ is also (ϵ, δ)-DP algorithm $\forall \mathcal{B}$.*

This is natural because otherwise the whole point of differential privacy will be forfeited. Also, the definition automatically allows for cases when the sensitive data are accessed more than once.

Lemma 2 (Composition rule)**.** *If $\mathcal{A}_1$ is (ϵ_1, δ_1)-DP, and $\mathcal{A}_2$ is (ϵ_2, δ_2)-DP then $(\mathcal{A}_1 \circ \mathcal{A}_2)$ is $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-DP.*

We will describe more advanced properties of DP as we need in Section 4.

3. Posterior sampling and differential privacy

In this section, we make a simple observation that under boundedness condition of a log-likelihood, getting one single sample from the posterior distribution (denoted by "OPS mechanism" from here onwards) preserves a degree of differential privacy for free. Then we will cite classic results in statistics and show that this sample is a consistent estimator in a Frequentist sense and near-optimal in many cases.

Algorithm 1 One-Posterior Sample (OPS) estimator

```
input Data $X$, log-likelihood function $\ell(\cdot|\cdot)$ satisfying $\sup_{\boldsymbol{x},\boldsymbol{\theta}} \|\ell(\boldsymbol{x}|\boldsymbol{\theta})\| \leq B$ a prior $\pi(\cdot)$. Privacy loss $\epsilon$.
  1. Set $\rho = \min\{1, \frac{\epsilon}{4B}\}$.
  2. Re-define log-likelihood function and the prior $\ell'(\cdot|\cdot) := \rho\ell(\cdot|\cdot)$ and $\pi'(\cdot) := (\pi(\cdot))^{\rho}$.
output $\hat{\boldsymbol{\theta}} \sim P(\boldsymbol{\theta}|X) \propto \exp\left(\sum_{i=1}^{N} \ell'(\boldsymbol{\theta}|\boldsymbol{x}_i)\right)\pi'(\boldsymbol{\theta})$.
```

3.1. Implicitly Preserving Differential Privacy

To begin with, we show that sampling from the posterior distribution is intrinsically differentially private.

Theorem 1. *If $\sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\theta}\in\Theta} |\log p(\boldsymbol{x}|\boldsymbol{\theta})| \leq B$, releasing one sample from the posterior distribution $p(\boldsymbol{\theta}|X^n)$ with any prior preserves $4B$-differential privacy. Alternatively, if $\mathcal{X}$ is a bounded domain (e.g., $\|x\|_* \leq R\ \forall \boldsymbol{x} \in \mathcal{X}$) and $\log p(\boldsymbol{x}|\boldsymbol{\theta})$ is an L-Lipschitz function in $\|\cdot\|_*$ for any $\boldsymbol{\theta} \in \Theta$, then releasing one sample from the posterior distribution preserves $4LR$-differential privacy.*

The proof is provided in the Appendix. Readers familiar with differential privacy must have noticed that this is actually an instance of the exponential mechanism (McSherry & Talwar, 2007), a general procedure that preserves privacy while making outputs with higher utility exponentially more likely. If one sets the utility function to be the log-likelihood and the privacy parameter being $4B$, then we get exactly the one-posterior sample mechanism. This exponential mechanism point of view provides an an simple extension which allows us to specify ϵ by simply scaling the log-likelihood (see Algorithm 1). We will overload the notation OPS to also represent this mechanism where we can specify ϵ. The nice thing about this algorithm is that there is almost zero implementation effort to extend all posterior sampling-based Bayesian learning models to have differentially privacy of any specified ϵ.

Assumption on the boundedness. The boundedness on the loss-function (log-likelihood here) is a standard assumption in many DP works (Chaudhuri et al., 2011; Bassily et al., 2014; Song et al., 2013; Kifer et al., 2012). Lipschitz constant L is usually small for continuous distributions (at least when the parameter space Θ is bounded). This is a bound on $\log p(\boldsymbol{x}|\boldsymbol{\theta})$) so as long as $p(\boldsymbol{x}|\boldsymbol{\theta})$ does not increase or decrease super exponentially fast at any point, L will be a small constant. R can also be made small by a simple preprocessing step that scales down all data points. In the aforementioned papers that assume L, it is typical that they also assume $R = 1$ for convenience. So we will do the same. In practice, we can algorithmically remove large data points from the data by some predefined threshold or using the "Propose-Test-Release" framework in (Dwork & Lei, 2009) or perform weighted training where we can assign lower weight to data points with large magnitude. Note that this is a desirable step for the robustness to outliers too. Exponential families (in Hilbert space) are an example, see e.g. Bialek et al. (2001); Hofmann et al. (2008); Wainwright & Jordan (2008).

3.2. Consistency and Near-Optimality

Now we move on to study the consistency of the OPS estimator. In great generality, we will show that the one-posterior sample estimator is consistent whenever the Bayesian model is posterior consistent. Since the consistency in Bayesian methods can have different meanings, we briefly describe two of them according to the nomenclature in Orbanz (2012).

Definition 2 (Posterior consistency in the Bayesian Sense)**.** For a prior π, we say the model is posterior consistent in the Bayesian sense, if $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$, $\boldsymbol{x}_1, ..., \boldsymbol{x}_n \sim p_{\boldsymbol{\theta}}$, and the posterior

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \overset{\text{weakly}}{\longrightarrow} \delta_{\boldsymbol{\theta}} \text{ a.s. } \pi.$$

$\delta_{\boldsymbol{\theta}}$ is the Dirac-delta function at $\boldsymbol{\theta}$.

In great generality, Doob's well-known theorem guarantees posterior consistency in the Bayesian sense for a model with any prior under no conditions except identifiability and measurability. A concise statement of Doob's result can be found in Van der Vaart (2000, Theorem 10.10)).

An arguably more reasonable definition is given below. It applies to the case when the statistician who chooses the prior π does not know about the true parameter.

Definition 3 (Posterior consistency in the Frequentist Sense)**.** For a prior π, we say the model is posterior consistent in the Frequentist sense, if for every $\boldsymbol{\theta}_0 \in \Theta$, $\boldsymbol{x}_1, ..., \boldsymbol{x}_n \sim p_{\boldsymbol{\theta}}$, the posterior

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \overset{\text{weakly}}{\longrightarrow} \delta_{\boldsymbol{\theta}_0} \text{ a.s. } p_{\boldsymbol{\theta}_0}.$$

This type of consistency is much harder to satisfy especially when Θ is an infinite dimensional space, in which case the consistency often depends on the specific priors to use (Ghosal, 2010).

Regardless which definition one favors, the key notion of consistency is that the posterior distribution to concentrates around the true underlying $\boldsymbol{\theta}$ that generates the data.

Proposition 1. *The one-posterior sample estimator is consistent* if and only if *the Bayesian model is posterior consistent (in either Definition 2 or 3).*

Proof. The equivalence follows from the standard equivalence of convergence weakly and convergence in probability when a random variable converges weakly to a point mass. □

How about the rate of convergence? In the low dimensional setting when $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is suitably differentiable and the prior is supported at the neighborhood of the true parameter, then by the Bernstein-von Mises theorem (Le Cam, 1986), the posterior mean is an asymptotically efficient estimator and the posterior distribution converges in L_1-distance to a normal distribution with covariance being the inverse Fisher Information.

Proposition 2. *Under the regularity conditions where Bernstein-von Mises theorem holds, the One-Posterior sample $\hat{\boldsymbol{\theta}} \sim \pi(\boldsymbol{\theta}|\boldsymbol{x}_1, .., \boldsymbol{x}_n)$ obeys*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{weakly}{\longrightarrow} \mathcal{N}(0, 2\mathbb{I}^{-1}),$$

i.e., the One-Posterior sample estimator has an asymptotic relative efficiency of 2.

Proof. Let the One-Posterior sample $\hat{\boldsymbol{\theta}} \sim \pi(\boldsymbol{\theta}|\boldsymbol{x}_1, .., \boldsymbol{x}_n)$. By Bernstein-von Mises theorem $\sqrt{n}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) \overset{\text{weakly}}{\to} \mathcal{N}(0, \mathbb{I}^{-1})$. By the asymptotic normality and efficiency of the posterior mean estimator $\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{\text{weakly}}{\to} \mathcal{N}(0, \mathbb{I}^{-1})$. The proof is complete by taking the sum of the two asymptotically independent Gaussian vectors ($\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}$ are asymptotically independent). □

The above proposition suggests that in many interesting classes of parametric Bayesian models, the One-Posterior Sample estimator is asymptotically near optimal. Similar statements can also be obtained for some classes of semi-parametric and nonparametric Bayesian models (Ghosal, 2010), which we leave as future work.The drawback of the above two propositions is that it is only stated for the version of the OPS when $\epsilon = 4B$. Using results in De Blasi & Walker (2013) and Kleijn et al. (2012) for misspecified models, we can prove consistency, asymptotic normality for any ϵ and parameterize the asymptotic relative efficiency of the OPS estimator as a function of ϵ. Details are given in the appendix.

3.3. (Efficient) sampling from approximate posterior

The privacy guarantee in Theorem 1 requires sampling from the exact posterior. In practice, however, exact samplers are rare. As Bayesian models get more and more complicated, often the only viable option is to use Markov Chain Monte Carlo (MCMC) samplers which are almost never exact. There are exceptions, e.g., Propp & Wilson (1998) but they only apply to problems with very special structures. A natural question to ask is whether we can still say something meaningful about privacy when the posterior sampling is approximate. It turns out that we can, and the level of approximation in privacy is the same as the level of approximation in the sampling distribution.

Proposition 3. *If $\mathcal{A}$ that sampling from distribution P_X preserves ϵ-differential privacy, then any approximate sampling procedures $\mathcal{A}'$ that produces a sample from P'_X such that $\|P_X - P'_X\|_{L_1} \leq \delta$ for any X preserves $(\epsilon, (1+e^{\epsilon})\delta)$-differential privacy.*

We are using L_1 distance of the distribution because it is a commonly accepted metric to measure the convergence rate MCMC (Rosenthal, 1995), and Proposition 3 leaves a clean interface for computational analysis in determining the number of iterations needed to attain a specific level of privacy protection.

A note on computational efficiency. The (unsurprising) bad news is that even approximate sampling from the posterior is NP-Hard in general, see, e.g. Sontag & Roy (2011, Theorem 8). There are however interesting results on when we can (approximately) sample efficiently. Approximation is easy for sampling LDA when $\alpha > 1$ while NP-Hard when $\alpha < 1$. A more general result in Applegate & Kannan (1991) suggests that we can get a sample with arbitrarily close approximation in polynomial time for a class of near log-concave distributions. The log-concavity of the distributions would imply convexity in the log-likelihood, thus, this essentially confirms the computational efficiency of all convex empirical risk minimization problems under differential privacy constraint (see Bassily et al. (2014)).

The nice thing is that since we do not modify the form of the sampling algorithm at all, the OPS algorithm is going to be a computationally tractable DP method whenever the Bayesian learning model of interest is proven to be computationally tractable.

3.4. Discussions and comparisons

OPS has a number of advantages over the state-of-the-art differentially private ERM method: objective perturbation (Chaudhuri et al., 2011; Kifer et al., 2012) (OBJPERT from here onwards). OPS works with arbitrary bounded loss functions and priors while OBJPERT needs a number of restrictive assumptions including twice differentiable loss functions, strongly convexity parameter to be greater than a threshold and so on. These restrictions rule out many commonly used loss functions, e.g., ℓ_1-loss, hinge loss, Huber function just to name a few.

Also, OBJPERT 's privacy guarantee holds only for the exact optimal solution, which is often hard to get in practice. In contrast, OPS works when the sample is drawn from an approximate posterior distribution. From a practical point of view, since OPS stems from the intrinsic privacy protection of Bayesian learning, it requires very little implementation effort to deploy it for practical applications. It also requires the problem to be strong convexity with a minimum strong convexity parameter. When the condition is not satisfied, ObjPert will need to add additional quadratic regularization to make it so, which may bias the problem unnecessarily.

4. Stochastic Gradient MCMC and (ϵ, δ)-Differential privacy

Given a fixed privacy budget, we saw that the single posterior sample produces an nearly optimal point estimate, but what if we want multiple samples? Trivially, we can run OPS multiple times but the privacy loss will aggregate linearly. Can we use the privacy budget in a different way that produces many approximate posterior samples?

In this section we will provide an answer to it by looking at a class of Stochastic Gradient MCMC techniques developed over the past few years. We will show that they are also differentially private for free if the parameters are chosen appropriately. The idea is to simply privately release an estimate of the gradient (as in Song et al. (2013); Bassily et al. (2014)) and leverage upon the following two celebrated lemmas in differential privacy in the same way as Bassily et al. (2014) does in deriving the near-optimal (ϵ, δ)-differentially private SGD.

The first lemma is the advanced composition which allows us to trade off a small amount of δ to get a much better bound for the privacy loss due to composition.

Lemma 3 (Advanced composition, c.f.,Theorem 3.20 in (Dwork & Roth, 2013))**.** *For all $\epsilon, \delta, \delta' \geq 0$, the class of (ϵ, δ)-DP mechanisms satisfy $(\epsilon', k\delta + \delta')$-DP under k-fold adaptive composition for:*

$$\epsilon' = \sqrt{2k \log(1/\delta')}\epsilon + k\epsilon(e^{\epsilon} - 1).$$

Remark 1. When $\epsilon = \frac{c}{\sqrt{2k\log(1/\delta')}} < 1$ for some constant $c < \sqrt{\log(1/\delta')}$, we can simplify the above expression into $\epsilon' \leq 2c$. To see this, apply the inequality $e^{\epsilon} - 1 \leq 2\epsilon$ (easily shown via Taylor's theorem and the assumption that $\epsilon \leq 1$).

In addition, we will also make use of the following lemma due to Beimel et al. (2014).

Lemma 4 (Privacy for subsampled data. Lemma 4.4 in Beimel et al. (2014).)**.** *Over a domain of data sets $\mathcal{X}^N$, if an algorithm $\mathcal{A}$ is (ϵ, δ) differentially private (with $\epsilon < 1$), then for any data set $X \in \mathcal{X}^N$, running $\mathcal{A}$ on a uniform random γN-entries of X ensures $(2\gamma\epsilon, \delta)$-DP.*

To make sense of the above lemma, notice that we are subsampling uniform randomly and the probability of any single data point being sampled is only γ. Thus, if we arbitrarily perturb one of the data points, its impact is evenly spread across all data points thanks to random sampling.

Let $f : \mathcal{X}^n \to \mathbb{R}^d$ be an arbitrary d-dimensional function. Define the ℓ_2 sensitivity of f to be

$$\Delta_2 f = \sup_{Y: d(X,Y) \leq 1} \|f(X) - f(Y)\|_2.$$

Suppose we want to output $f(X)$ differential privately, "Gaussian Mechanism" output $\hat{f}(X) = f(X) + \mathcal{N}(0, \sigma^2 I_d)$ for some appropriate σ.

Theorem 2 (Gaussian Mechanism, c.f. Dwork & Roth (2013))**.** *Let $\epsilon \in (0, 1)$ be arbitrary. "Gaussian Mechanism" with $\sigma \geq \Delta_2 f\sqrt{2\log(1.25/\delta)}/\epsilon$ is (ϵ, δ)-differentially private.*

This will be the main workhorse that we use here.

4.1. Stochastic Gradient Langevin Dynamics

SGLD iteratively update the parameters to by running a perturbed version of the minibatch stochastic gradient descent on the negative log-posterior objective function

$$-\sum_{i=1}^{N} \log p(\boldsymbol{x}_i|\boldsymbol{\theta}) - \log \pi(\boldsymbol{\theta}) =: \sum_{i=1}^{N} \ell(\boldsymbol{x}_i; \boldsymbol{\theta}) + r(\boldsymbol{\theta})$$

where $\ell(\boldsymbol{x}_i; \boldsymbol{\theta})$ and $r(\boldsymbol{\theta})$ are loss-function and regularizer under the empirical risk minimization.

If one were to run stochastic gradient descent or any other optimization tools on this, one would eventually a deterministic maximum a posteriori estimator. SGLD avoids this by adding noise in every iteration. At iteration t SGLD first samples uniform randomly τ data points $\{\boldsymbol{x}_{t_1}, ..., \boldsymbol{x}_{t_2}\}$ and then updates the parameter using

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \left(\nabla r(\boldsymbol{\theta}) + \frac{N}{\tau}\sum_{i=1}^{\tau} \nabla \ell(\boldsymbol{x}_{ti}|\boldsymbol{\theta}) \right) + \boldsymbol{z}_t, \tag{4.1}$$

where $\boldsymbol{z}_t \sim \mathcal{N}(0, \eta_t)$ and τ is the mini-batch size.

For the ordinary stochastic gradient descent to converge in expectation, the stepsize η_t can be chosen as anything that $\sum_{i=1}^{\infty} \eta_t = \infty$ and $\sum_{i=1}^{\infty} \eta_t^2 < \infty$ (Robbins & Monro, 1951). Typically, one can chooses stepsize $\eta_t = a(b + t)^{-\gamma}$ with $\gamma \in (0.5, 1]$. In fact, it is shown that for general convex functions and μ-strongly convex functions $\frac{1}{\sqrt{t}}$ and $\frac{1}{\mu t}$ can be used to obtain the minimax optimal $O(1/\sqrt{t})$ and $O(1/t)$ rate of convergence. These results substantiate the first phase of SGLD: a convergent algorithm to the optimal solution. Once it gets closer, however, it transforms into a posterior sampler. According to Welling & Teh (2011) and later formally proven in Sato & Nakagawa (2014), if we choose $\eta_t \to 0$, the random iterates $\boldsymbol{\theta}_t$ of SGLD converges in distribution to the $p(\boldsymbol{\theta}|X)$. The idea is that as the stepsize gets smaller, the stochastic error from the true gradient due to the random sampling of the minibatch converges to 0 faster than the injected Gaussian noise.

Algorithm 2 Differentially Private Stochastic Gradient Langevin Dynamics (DP-SGLD)

```
Require: Data $X$ of size $N$, Size of minibatch $\tau$, number of data passes $T$, privacy parameter $\epsilon, \delta$, Lipschitz constant $L$ and initial $\boldsymbol{\theta}_1$. Set $t = 1$.
  for $t = 1 : \lfloor NT/\tau \rfloor$ do
    1. Random sample a minibatch $S \subset [N]$ of size $\tau$.
    2. Sample each coordinate of $\boldsymbol{z}_t$ iid from $\mathcal{N}\left(0, \frac{128NTL^2}{\tau\epsilon^2}\log\left(\frac{2.5NT}{\tau\delta}\right)\log(2/\delta)\eta_t^2 \vee \eta_t\right)$.
    3. Update $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta_t\left(\nabla r(\boldsymbol{\theta}) + \frac{N}{\tau}\sum_{i\in S}\nabla\ell(\boldsymbol{x}_i|\boldsymbol{\theta})\right) + \boldsymbol{z}_t$,
    4. Return $\boldsymbol{\theta}_{t+1}$ as a posterior sample (after a predefined burn-in period).
    5. Increment $t \leftarrow t + 1$.
  end for
```

Algorithm 3 Hybrid Posterior Sampling Algorithm

```
Require: Data $X$ of size $N$, log-likelihood function $\ell(\cdot|\theta)$ with Lipschitz constant $L$ in the first argument, assume $\sup_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{x}\|$, a prior $\pi$. Privacy requirement $\epsilon$.
  1. Run OPS estimator: Algorithm 1 with $\epsilon/2$. Collect sample point $\theta_0$
  2. Run DP-SGLD (Algorithm 2) or other Stochastic Gradient Monte Carlo algorithms and collect samples.
output : Return all samples.
```

In addition, if we use some fixed stepsize lower bound, such that $\eta_t = \max\{1/(t+1), \eta_0\}$ (to alleviate the slow mixing problem of SGLD), the results correspond to a discretization approximation of a stochastic differential equation (Fokker-Planck equation), which obeys the following theorem due to Sato & Nakagawa (2014) (simplified and translated to our notation).

Theorem 3 (Weak convergence (Sato & Nakagawa, 2014))**.** *Assume $f(\boldsymbol{\theta}|X)$ is differentiable, $\nabla f(\boldsymbol{\theta}|X)$ is gradient Lipschitz and bounded* [2]*. Then*

$$\left|\mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|X)}[h(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}\sim SGLD}[h(\boldsymbol{\theta}(t))]\right| = O(\eta_t),$$

for any continuous and polynomial growth function h.

This theorem implies that one can approximate the posterior mean (and other estimators) using SGLD. Finite sample properties of SGLD is studied in Vollmer et al. (2015).

Now we will show that with a minor modification to just the "burn-in" phase of SGLD, we will be able to make it differentially private (see Algorithm 2).

Theorem 4 (Differentially private Minibatch SGLD)**.** *Assume initial $\boldsymbol{\theta}_1$ is chosen independent of the data, also assume $\ell(\boldsymbol{x}|\boldsymbol{\theta})$ is L-smooth in $\|\cdot\|_2$ for any $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\theta} \in \Theta$. In addition, let $\epsilon, \delta, \tau, T$ be chosen such that $T \geq \frac{\epsilon^2 N}{32\tau\log(2/\delta)}$. Then Algorithm 2 preserves (ϵ, δ)-differential privacy.*

The proof is provided in the Appendix.

α-Phase transition. For any $\alpha \in (0, 1)$, if we choose $\eta_t = \frac{\alpha\epsilon^2}{128L^2\log(2.5NT/(\tau\delta))\log(2/\delta)t}$, then whenever $t > \alpha NT/\tau$, then we are essentially running SGLD for the last $(1-\alpha)NT/\tau$ iterations, and we can collect approximate posterior samples from there.

Small constant η_0. Instead of making η_t to converge to 0 as t increases, we may alternatively use constant η_0 after t is larger than a threshold. This is a suggested heuristic in Welling & Teh (2011) and is inline with the analysis in Sato & Nakagawa (2014) and Vollmer et al. (2015).

Choice of T and τ By Bassily et al. (2014), it takes at least N data passes to converge in expectation to a point near the minimizer, so taking $T = 2N$ is a good choice. The variance of both random components in our stochastic gradient is smaller when we use larger τ. Smaller variances would improve the convergence of the stochastic gradient methods and make the SGLD a better approximation to the full Langevin Dynamics. The trade-off is that when τ is too large, we will use up the allowable T datapasses with just $O(T)$ iterations and the number of posterior samples we collect from the algorithm will be small.

Overcoming the large-noise in the "Burn-in" phase When the stepsize η_t is not small enough initially, we need to inject significantly more noise than what SGLD would have to ensure privacy. We can overcome this problem by initializing the SGLD sampler with a valid output of the OPS estimator, modified according to the exponential mechanism so that the privacy loss is calibrated to $\epsilon/2$. As the initial point is already in the high probability region of the posterior distribution, we no longer need to "Burn-in" the Monte Carlo sampler so we can simply choose a sufficiently small constant stepsize so that it remains a valid SGLD. This algorithm is summarized in Algorithm 3.

Comparing to OPS The privacy claim of DP-SGLD is very different from OPS . It does not require sampling to be nearly correct to ensure differential privacy. In fact, DP-SGLD privately releases the entire sequence of parameter updates, thus ensures differential privacy even if the internal state of the algorithm gets hacked. However, the quality of the samples is usually worse than OPS due to the random-walk like behavior. The interesting fact, however, is that if we run SGLD indefinitely without worrying about the stronger notion of internal privacy, it leads to a valid

[2] We use boundedness to make the presentation simpler. Boundedness trivially implies the linear growth condition in Sato & Nakagawa (2014, Assumption 2).

posterior sample, which is private by our first part. We can potentially use SGLD to sample from a "scaled" version so as to balancing the two ways of getting privacy.

4.2. Hamiltonian Dynamics, Fisher Scoring and Nosé-Hoover Thermostat

One of the practical drawback of SGLD is its random walk-like behavior which slows down the mixing significantly. In this section, we describe three extensions of SGLD that attempts to resolve the issue by either using auxiliary variables to counter the noise in the stochastic gradient(Chen et al., 2014; Ding et al., 2014), or to exploit second order information so as to use Newton-like updates with large stepsize (Ahn et al., 2012).

We note that in all these methods, stochastic gradients are the only form of data access, therefore similar results like what we described for SGLD follow nicely. We briefly describe each method and how to choose their parameters for differential privacy.

Stochastic Gradient Hamiltonian Monte Carlo. According to Neal (2011), Langevin Dynamics is a special limiting case of Hamiltonian Dynamics, where one can simply ignore the "momentum" auxiliary variable. In its more general form, Hamiltonian Monte Carlo (HMC) is able to generate proposals from distant states and hence enabling more rapid exploration of the state space. Chen et al. (2014) extends the full "leap-frog" method for HMC in Neal (2011) to work with stochastic gradient and add a "friction" term in the dynamics to "de-bias" the noise in the stochastic gradient.

$$\begin{cases} \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + h_t \boldsymbol{r}_{t-1} \\ \boldsymbol{p}_t = \boldsymbol{p}_{t-1} - h_t \widehat{\nabla} - \eta_t A \boldsymbol{p}_{t-1} + \mathcal{N}(0, 2(A - \widehat{B})h_t). \end{cases} \tag{4.2}$$

where $\widehat{B}$ is a guessed covariance of the stochastic gradient (the authors recommend restricting $\hat{B}$ to a single number or a diagonal matrix) and A can be arbitrarily chosen as long as $A \succ \widehat{B}$. If the stochastic gradient $\widehat{\nabla} \sim \mathcal{N}(\nabla, B)$ for some B and $\widehat{B} = B$, then this dynamics is simulating a dynamic system that yields the correct distribution. Note that even if the normal assumption holds and we somehow set $\widehat{B} = B$, we still requires h_t to go to 0 to sample from the actual posterior distribution, and as h_t converges to 0 the additional noise we artificially inject dominates and we get privacy for free. All we need to do is to set A, $\widehat{B}$ and h_t so that $2(A - \widehat{B})/h_t \succ \frac{128NTL^2}{\tau\epsilon^2} \log\left(\frac{2.5NT}{\tau\delta}\right) \log(2/\delta) I_n$. Note that as $h_t \to 0$ this quickly becomes true.

Stochastic Gradient Nosé-Hoover Thermostat As we discussed, the key issue about SGHMC is still in choosing $\widehat{B}$. Unless $\widehat{B}$ is chosen exactly as the covariance of true stochastic gradient, it does not sample from the correct distribution even as $h_t \to 0$ unless we trivially set $\hat{B} = 0$. The Stochastic Gradient Nosé-Hoover Thermostat (SGNHT) overcomes the issue by introducing an additional auxiliary variable ξ, which serves as a thermostat to absorb the unknown noise in the stochastic gradient. The update equations of SGNHT are given below

$$\begin{cases} \boldsymbol{p}_t = \boldsymbol{p}_{t-1} - \xi_{t-1} \boldsymbol{p}_{t-1} h_t - \widehat{\nabla} h_t + \mathcal{N}(0, 2Ah_t); \\ \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + h_t \boldsymbol{p}_{t-1}; \\ \xi_t = \xi_{t-1} + (\frac{1}{n} \boldsymbol{p}_t^T \boldsymbol{p}_t - 1) h_t. \end{cases} \tag{4.3}$$

Similar to the case in SGHMC, appropriately selected discretization parameter h_t and the friction term A will imply differential privacy.

Chen et al. (2014); Ding et al. (2014) both described a reformulation that can be interpret as SGD with momentum. This is by setting parameters $\eta = h^2, a = hA, \hat{b} = h\widehat{B}$ for SGHMC:

$$\begin{cases} \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{v}_{t-1} \\ \boldsymbol{v}_t = \boldsymbol{v}_{t-1} - \eta_t \widehat{\nabla} - a\boldsymbol{v} + \mathcal{N}(0, 2(a - \widehat{b})\eta_t I); \end{cases} \tag{4.4}$$

and $\boldsymbol{v} = \boldsymbol{p}h, \eta_t = h_t^2, \alpha = \xi h$ and $a = Ah$ for SGNHT:

$$\begin{cases} \boldsymbol{v}_t = \boldsymbol{v}_{t-1} - \alpha_{t-1} \boldsymbol{v}_{t-1} - \widehat{\nabla}(\boldsymbol{\theta}_{t-s}) \eta_t + \mathcal{N}(0, 2a\eta_t I); \\ \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{u}_{t-1}; \\ \alpha_t = \alpha_{t-1} + (\frac{1}{n} \boldsymbol{v}_t^T \boldsymbol{v}_t - \eta_t). \end{cases} \tag{4.5}$$

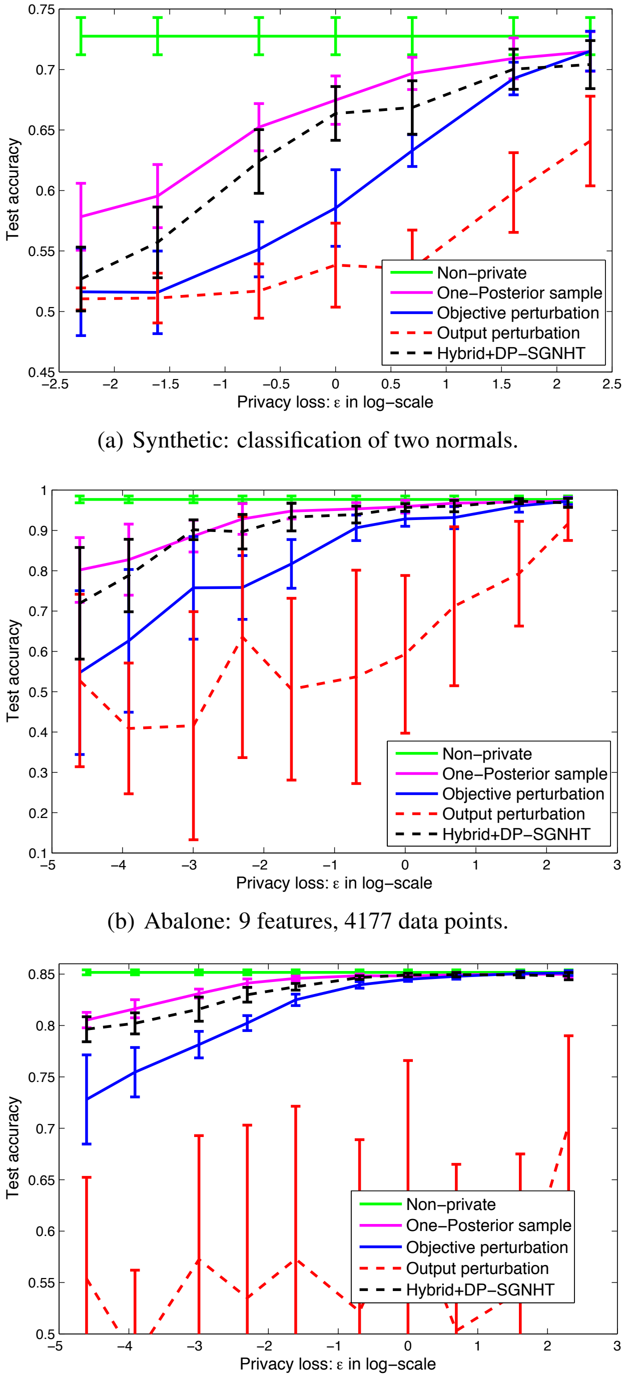
where $1 - a$ is the momentum parameter and η is the learning rate in the SGD with momentum. Again note that to obtain privacy, we need $\frac{2a}{\eta_t} \geq \frac{128NTL^2}{\tau\epsilon^2} \log(\frac{2NT}{\tau\delta}) \log(1/\delta)$.

Note that as η_t gets smaller, we have the flexibility of choosing a and η_t within a reasonable range.

Stochastic Gradient Fisher Scoring Another extension of SGLD is Stochastic Gradient Fisher Scoring (SGFS), where Ahn et al. (2012) proposes to adaptively interpolate between a preconditioned SGLD (see preconditioning (Girolami & Calderhead, 2011)) and a Markov Chain that samples from a normal approximation of the posterior distribution. For parametric problem where Bernstein-von Mises theorem holds, this may be a good idea. The heuristic used in the SGFS is that the covariance matrix of $\boldsymbol{\theta}|X$, which is also the inverse Fisher information I_N^{-1} is estimated on the fly. The key features of SGFS is that one can use the stepsize to trade off speed and accuracy, when the stepsize is large, it mixes rapidly to the normal approximation, as the stepsize gets smaller the stationary distribution converges to the true posterior. Further details of SGFS and ideas to privatize it is described in the appendix.

5. Experiments

To evaluate how our proposed methods work in practice, we selected two binary classification datasets: Abalone and



(a) Synthetic: classification of two normals.

(b) Abalone: 9 features, 4177 data points.

(c) Adult: 109 features, 32561 data points.

Figure 1. Comparison of Differential Private methods.

Adult, from the first page of UCI Machine Learning Repository and performed privacy constrained logistic regression on them. Specifically, we compared two of our proposed methods, OPS mechanism and hybrid algorithm against the state-of-the-art empirical risk minimization algorithm OBJPERT (Chaudhuri et al., 2011; Kifer et al., 2012) under varying level of differential privacy protection. The results are shown in Figure 1. As we can see from the figure, in both problems, OPS significantly improves the classification accuracy over OBJPERT . The hybrid algorithm also works reasonably well, given that it collected N samples after initializing it from the output of a run of OPS with privacy parameter $\epsilon/2$. For fairness, we used the (ϵ, δ)-DP version of the objective perturbation (Kifer et al., 2012) and similarly we used Gaussian mechanism (rather than Laplace mechanism) for output perturbation. All optimization based methods are solved using BFGS algorithm to high numerical accuracy. OPS is implemented using SGNHT and we ran it long enough so that we are confident that it is a valid posterior sample. Minibatch size and number of data passes in the hybrid DP-SGNHT are chosen to be both $\sqrt{N}$.

We note that the plain DP-SGLD and DP-SGNHT without an initialization using OPS does not work nearly as well. In our experiments, it often performs equally or slightly worse than the output perturbation. This is due to the few caveats (especially “the curse of numerical constant”) we described earlier.

6. Conclusion and future work

In this paper, we described two simple but conceptually interesting examples that Bayesian learning can be inherently differentially private. Specifically, we show that getting one sample from the posterior is a special case of exponential mechanism and this sample as an estimator is near-optimal for parametric learning. On the other hand, we illustrate that the algorithmic procedures of stochastic gradient Langevin Dynamics (and variants) that attempts to sample from the posterior also guarantee differential privacy as a byproduct. Preliminary experiments suggests that the One-Posterior-Sample mechanism works very well in practice and it substantially outperforms earlier privacy mechanism in logistic regression. While suffering from a large constant, our second method is also theoretically and practically meaningful in that it provides privacy protection in intermediate steps.

To carry the research forward, we think it is important to identify other cases when the existing randomness can be exploited for privacy. Randomized algorithms such as hashing and sketching, dropout and other randomization used in neural networks might be another thing to look at. More on the application end, we hope to explore the one-posterior sample approach in differentially private movie recommendation. Ultimately, the goal is to make differential privacy more practical to the extent that it can truly solve the real-life privacy problems that motivated its very advent.

Acknowledgments

The research was partially supported by NSF Award BCS-0941518 to the Department of Statistics at Carnegie Mellon University, a grant of Adobe Research, and a grant by Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

References

Ahn, Sungjin, Korattikara, Anoop, and Welling, Max. Bayesian posterior sampling via stochastic gradient fisher scoring. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 2012.

Airoldi, Edoardo M, Blei, David M, Fienberg, Stephen E, and Xing, Eric P. Mixed membership stochastic blockmodels. In *Advances in Neural Information Processing Systems*, pp. 33–40, 2009.

Applegate, David and Kannan, Ravi. Sampling and integration of near log-concave functions. In *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, pp. 156–163. ACM, 1991.

Bassily, Raef, Smith, Adam, and Thakurta, Abhradeep. Private empirical risk minimization, revisited. *arXiv preprint arXiv:1405.7085*, 2014.

Beimel, Amos, Brenner, Hai, Kasiviswanathan, Shiva Prasad, and Nissim, Kobbi. Bounds on the sample complexity for private learning and private data release. *Machine learning*, 94(3):401–437, 2014.

Bialek, William, Nemenman, Ilya, and Tishby, Naftali. Predictability, complexity, and learning. *Neural computation*, 13(11):2409–2463, 2001.

Blei, David M, Ng, Andrew Y, and Jordan, Michael I. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

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

Chen, Tianqi, Fox, Emily B., and Guestrin, Carlos. Stochastic Gradient Hamiltonian Monte Carlo. In *Proceeding of 31th International Conference on Machine Learning (ICML'14)*, 2014.

De Blasi, Pierpaolo and Walker, Stephen G. Bayesian asymptotics with misspecified models. *Statistica Sinica*, 23:169–187, 2013.

Dimitrakakis, Christos, Nelson, Blaine, Mitrokotsa, Aikaterini, and Rubinstein, Benjamin IP. Robust and private bayesian inference. In *Algorithmic Learning Theory*, pp. 291–305. Springer, 2014.

Ding, Nan, Fang, Youhan, Babbush, Ryan, Chen, Changyou, Skeel, Robert D, and Neven, Hartmut. Bayesian sampling using stochastic gradient thermostats. In *Advances in Neural Information Processing Systems*, pp. 3203–3211, 2014.

Dwork, Cynthia. Differential privacy. In *Proceedings of the 33rd international conference on Automata, Languages and Programming-Volume Part II*, pp. 1–12. Springer-Verlag, 2006.

Dwork, Cynthia and Lei, Jing. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 371–380. ACM, 2009.

Dwork, Cynthia and Roth, Aaron. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2013.

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

Fei-Fei, Li and Perona, Pietro. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pp. 524–531. IEEE, 2005.

Gelman, Andrew, Carlin, John B, and Stern, Hal S. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

Geman, Stuart and Geman, Donald. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

Ghosal, Subhashis. *The Dirichlet process, related priors and posterior asymptotics*, volume 2. Chapter, 2010.

Girolami, Mark and Calderhead, Ben. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Hastings, W Keith. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1): 97–109, 1970.

Hofmann, Thomas, Schölkopf, Bernhard, and Smola, Alexander J. Kernel methods in machine learning. *The annals of statistics*, pp. 1171–1220, 2008.

Kasiviswanathan, Shiva P and Smith, Adam. On the'semantics' of differential privacy: A bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1): 1, 2014.

Kifer, Daniel, Smith, Adam, and Thakurta, Abhradeep. Private convex empirical risk minimization and high-dimensional regression. *Journal of Machine Learning Research*, 1:41, 2012.

Kleijn, BJK, van der Vaart, AW, et al. The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

Le Cam, Lucien Marie. *On the Bernstein-von Mises theorem*. Department of Statistics, University of California, 1986.

Lehmann, Erich L and Romano, Joseph P. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

McSherry, Frank and Talwar, Kunal. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pp. 94–103. IEEE, 2007.

Mir, Darakhshan J. *Differential privacy: an exploration of the privacy-utility landscape*. PhD thesis, Rutgers University, 2013.

Neal, Radford. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2, 2011.

Orbanz, Peter. Lecture notes on bayesian nonparametrics. *Journal of Mathematical Psychology*, 56:1–12, 2012.

Penny, William D, Friston, Karl J, Ashburner, John T, Kiebel, Stefan J, and Nichols, Thomas E. *Statistical parametric mapping: the analysis of functional brain images: the analysis of functional brain images*. Academic press, 2011.

Propp, James and Wilson, David. Coupling from the past: a users guide. *Microsurveys in Discrete Probability*, 41: 181–192, 1998.

Rabiner, Lawrence. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Rajkumar, Arun and Agarwal, Shivani. A differentially private stochastic gradient descent algorithm for multiparty classification. In *International Conference on Artificial Intelligence and Statistics*, pp. 933–941, 2012.

Robbins, Herbert and Monro, Sutton. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Rosenthal, Jeffrey S. Minorization conditions and convergence rates for markov chain monte carlo. *Journal of the American Statistical Association*, 90(430):558–566, 1995.

Sato, Issei and Nakagawa, Hiroshi. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 982–990, 2014.

Smith, Adam. Efficient, differentially private point estimators. *arXiv preprint arXiv:0809.4794*, 2008.

Song, Shuang, Chaudhuri, Kamalika, and Sarwate, Anand D. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing*, 2013.

Sontag, David and Roy, Dan. Complexity of inference in latent dirichlet allocation. In *Advances in neural information processing systems*, pp. 1008–1016, 2011.

Van der Vaart, Aad W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Vollmer, Sebastian J, Zygalakis, Konstantinos C, et al. (non-) asymptotic properties of stochastic gradient langevin dynamics. *arXiv preprint arXiv:1501.00438*, 2015.

Wainwright, Martin J and Jordan, Michael I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2): 1–305, 2008.

Wang, Yu-Xiang, Fienberg, Stephen E, and Smola, Alex. Privacy for free: Posterior sampling and stochastic gradient monte carlo. *arXiv preprint arXiv:1502.07645*, 2015.

Welling, Max and Teh, Yee W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 681–688, 2011.

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

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