This is the supplementary file of the paper: "Privacy for Free: Posterior Sampling and Stochastic Gradient Monte Carlo". In Appendix A, we provide deferred proofs of the results in the paper. In Appendix B, we describe the statistical analysis for OPS with general ϵ. In Appendix C, we discuss a differential private extension of Stochastic Gradient Fisher Scoring (SGFS). The subsequent appendices are about a qualitative experiment, additional discussions on the proposed methods and relationships to existing work.

A. Proofs

Proof of Theorem 1. The posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x}_1,...,\boldsymbol{x}_n) = \frac{\prod_{i=1}^n p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} \prod_{i=1}^n p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$. For any $\boldsymbol{x}_1,...,\boldsymbol{x}_n, x'_k$, The ratio can be factorized into

$$\frac{p(\boldsymbol{\theta}|\boldsymbol{x}_1,...,x'_k,...,\boldsymbol{x}_n)}{p(\boldsymbol{\theta}|\boldsymbol{x}_1,...,\boldsymbol{x}_k,...,\boldsymbol{x}_n)} = \underbrace{\frac{p(\boldsymbol{x}'_k|\boldsymbol{\theta})\prod_{i=1:n,i\neq k} p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\prod_{i=1}^n p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})}}_{\text{Factor 1}} \times \underbrace{\frac{\int_{\boldsymbol{\theta}} \prod_{i=1}^n p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} p(\boldsymbol{x}'_k|\boldsymbol{\theta})\prod_{i=1:n,i\neq k} p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}}_{\text{Factor 2}}.$$

It follows that

$$\begin{aligned}
\text{Factor 1} &= \frac{p(\boldsymbol{x}'_k|\boldsymbol{\theta})}{p(\boldsymbol{x}_k|\boldsymbol{\theta})} = e^{\log p(\boldsymbol{x}'_k|\boldsymbol{\theta}) - \log p(\boldsymbol{x}_k|\boldsymbol{\theta})} \leq e^{2B}, \\
\text{Factor 2} &= \frac{\int_{\boldsymbol{\theta}} \prod_{i\neq k} p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{x}_k)d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} p(\boldsymbol{x}'_k|\boldsymbol{\theta})\prod_{i\neq k} p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{\int_{\boldsymbol{\theta}} \prod_{i\neq k} p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{x}'_k|\boldsymbol{\theta})\frac{p(\boldsymbol{x}_k)}{p(\boldsymbol{x}'_k)}d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} p(\boldsymbol{x}'_k|\boldsymbol{\theta})\prod_{i\neq k} p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \\
&= \frac{\int_{\boldsymbol{\theta}} \prod_{i\neq k} p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{x}'_k|\boldsymbol{\theta})e^{\log p(\boldsymbol{x}_k|\boldsymbol{\theta}) - \log p(\boldsymbol{x}'_k|\boldsymbol{\theta})}d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} p(\boldsymbol{x}'_k|\boldsymbol{\theta})\prod_{i\neq k} p(\boldsymbol{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \\
&\leq e^{2B}\frac{m(\boldsymbol{x}_1,...,\boldsymbol{x}'_k,...,\boldsymbol{x}_n)}{m(\boldsymbol{x}_1,...,\boldsymbol{x}'_k,...,\boldsymbol{x}_n)} = e^{2B}.
\end{aligned}$$

where we use $m(X)$ to denote the marginal distribution. As a result, the whole thing is bounded by e^{4B}.

Alternatively, we can use the Lipschitz constant and boundedness to get $\log p(\boldsymbol{x}'_k|\boldsymbol{\theta}) - \log p(\boldsymbol{x}_k|\boldsymbol{\theta}) \leq L\|x'_k - \boldsymbol{x}_k\|_* \leq 2LR$. □

Proof of Proposition 3. For any $S \in \text{Range}(\mathcal{A}')$, and $d(X, X') \leq 1$

$$\begin{aligned}
\mathbb{P}\left(\mathcal{A}'(X) \in S\right) &= \int_S dP'_X \leq \int_S dP_X + \delta \\
&\leq e^\epsilon \int_S dP_{X'} + \delta \leq e^\epsilon \int_S dP_{X'} \\
&\leq e^\epsilon \int_S dP'_{X'} + (1+e^\epsilon)\delta \\
&= e^\epsilon \mathbb{P}\left(\mathcal{A}'(X') \in S\right) + (1+e^\epsilon)\delta,
\end{aligned}$$

This is $(\epsilon, (1+e^\epsilon)\delta)$-DP by definition. □

Proof of Theorem 4. In every iteration, the only data access is $\sum_{i\in S}\nabla\ell(\boldsymbol{x}_i|\boldsymbol{\theta})$ and by the L-Lipschitz condition, the sensitivity of $\sum_{i\in S}\nabla\ell(\boldsymbol{x}_i|\boldsymbol{\theta})$ is at most $2L$. Get the essential noise that is added to $\sum_{i\in S}\nabla\ell(\boldsymbol{x}_i|\boldsymbol{\theta})$ by removing the $\frac{N^2\eta_t^2}{\tau^2}$ factor from the variance σ^2 in the algorithm, and Gaussian mechanism, ensures the privacy loss to be smaller than $\frac{\epsilon\sqrt{N}}{\sqrt{32\tau T\log(2/\delta)}}$ with probability $> 1 - \frac{\tau\delta}{2NT}$.

Using the same technique in Bassily et al. (2014), we can further exploit the fact that the subset S that we use to compute the stochastic gradient is chosen uniformly randomly. By Lemma 4, the privacy loss for this iteration is in fact

$$\frac{\epsilon\sqrt{N}}{\sqrt{32\tau T\log(2/\delta)}} \cdot \frac{2\tau}{N} = \frac{\epsilon/2}{\sqrt{2(NT/\tau)\log(2/\delta)}}.$$

Verify that we can indeed do that as $\frac{\epsilon\sqrt{N}}{\sqrt{32\tau T\log(2/\delta)}} < 1$ from the assumption on T. Note that to get T data passes with minibatches of size τ, we need to go through at most $\lfloor \frac{NT}{\tau} \rfloor \leq \frac{NT}{\tau}$ iterations. Apply the advanced composition theorem (Remark 1), we get an upper bound of the total privacy loss ϵ and failure probability $\delta = \frac{\delta}{2} + \frac{\tau\delta}{2NT} \cdot \frac{NT}{\tau}$ accordingly.

The proof is complete by noting that choosing a larger noise level when η_t is bigger can only reduces the privacy loss under the same failure probability. □

B. Analysis for general OPS

In this section, we prove consistency, asymptotic normality for any ϵ and parameterize the asymptotic relative efficiency of the OPS estimator as a function of ϵ. The key idea is that when scaling the log-likelihood and sample from a different distribution, we are essentially fitting a model that may not include the data-generating true distribution. De Blasi & Walker (2013) shows that under mild conditions, when the model is misspecified, the posterior distribution will converge to a point mass $\boldsymbol{\theta}^*$ that minimizes the KL-divergence between between the true distribution and the corresponding distribution in the misspecified model. $\boldsymbol{\theta}^*$ is essentially MLE and in our case, since we only scaled the distribution, the MLE will remain exactly the same. De Blasi & Walker (2013)'s result is quite general and covers both parametric and nonparametric Bayesian models and whenever their assumptions hold, the OPS estimator is consistent. Using a similar argument and the modified Bernstein-Von-Mises theorem in Kleijn et al. (2012), we can prove asymptotic normality and near optimality for the subset of problems where regularities of MLE hold.

Proposition 4. *Under the same assumption as Proposition 2, if we set a different ϵ by rescaling the log-likelihood by a factor of $\frac{\epsilon}{4B}$, then the the One-Posterior sample estimator obeys*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{\textit{weakly}}{\longrightarrow} \mathcal{N}\left(0, (1 + \frac{4B}{\epsilon})\mathbb{I}^{-1}\right),$$

in other word, the estimator has an ARE of $(1 + \frac{4B}{\epsilon})$.

Proof. By scaling the log-likelihood, we are essentially changing the correct model $p_{\boldsymbol{\theta}}$ to a misspecified model $(p_{\boldsymbol{\theta}})^{\frac{\epsilon}{4B}}$. Let the true log-likelihood be ℓ and the misspecified log-likelihood be $\tilde{\ell} = \frac{\epsilon}{4B}\ell$, in addition, define

$$V(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}\nabla\tilde{\ell}(\boldsymbol{\theta})\nabla\tilde{\ell}(\boldsymbol{\theta})^T = \frac{\epsilon^2}{16B^2}\mathbb{E}_{\boldsymbol{\theta}}\nabla\ell(\boldsymbol{\theta})\nabla\ell(\boldsymbol{\theta})^T = \frac{\epsilon^2}{16B^2}\mathbb{I}(\boldsymbol{\theta})$$

$$J(\boldsymbol{\theta}) := -\mathbb{E}_{\boldsymbol{\theta}}\nabla^2\tilde{\ell}(\boldsymbol{\theta}) = -\frac{\epsilon}{4B}\mathbb{E}_{\boldsymbol{\theta}}\nabla^2\ell(\boldsymbol{\theta}) = -\frac{\epsilon}{4B}\mathbb{I}(\boldsymbol{\theta}).$$

The last equality holds under the standard regularity conditions. By the sandwich formula, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ under the misspecified model is asymptotically normal:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \overset{\text{weakly}}{\to} \mathcal{N}(0, J^{-1}VJ(-1)) = \mathcal{N}(0, \mathbb{I}^{-1})$$

where $\boldsymbol{\theta}^*$ defines the closest (in terms of KL-divergence) model in the misspecified class of distributions to the true distribution that generates the data. Since the difference is only in scaling, the minimum KL-divergence is obtained at $\boldsymbol{\theta}^* = \boldsymbol{\theta}$. Now under the same regularity conditions, we can invoke the modified Bernstein-Von-Mises theorem for misspecified models (Kleijn et al., 2012, Lemma 2.2), which says that the posterior distribution $p(\theta|X^n)$ (of the misspecified model) converges in distribution to $\mathcal{N}(\hat{\boldsymbol{\theta}}, (nJ)^{-1})$. In our case, $(nJ)^{-1} = \frac{4B}{n\epsilon}\mathbb{I}^{-1}$. The proof is concluded by noting that the posterior sample is an independent draw. □

We make a few interesting remarks about the result.

1. Proposition 4 suggests that for models with bounded log-likelihood, OPS is only a factor of $(1 + 4B/\epsilon)$ away from being optimal. This is in sharp contrast to most previous statistical analysis of DP methods that are only tight up to a numerical constant (and often a logarithmic term). In ℓ_2-norm, the convergence rate is $O(\frac{\sqrt{1+4B/\epsilon}\|I^{-1}\|_F}{\sqrt{n}})$. The bound depends on the dimension through the Frobenius norm which is usually $O(\sqrt{d})$. The bound can be further

sharpened using assumptions on the intrinsic rank, incoherence conditions or the rate of decays in eigenvalues of the Fisher information. In ℓ_∞-norm, the convergence rate is $\frac{\sqrt{1+4B/\epsilon}\|I^{-1}\|_2}{\sqrt{n}}$, which does not depend on the dimension of the problem.

2. Another implication is on statistical inference. Proposition 4 essentially generalizes that classic results in hypothesis testing and confidence intervals, e.g., Wald test, generalized likelihood ratio test, can be directly adopted for the private learning problems, with an appropriate calibration using ϵ. We can control the type I error in an asymptotically exact fashion. In addition, the trade-off with ϵ and the test power is also explicitly described, so in cases where the power of the tests are well-studied (Lehmann & Romano, 2006), the same handle can be used to analyze the most-powerful-test under privacy constraints.

3. Lastly, Kleijn et al. (2012)'s result is much more general. It is easy to extend the guarantee for OPS to handle private Bayesian learning in a fully agnostic setting and in non-iid cases. We will leave the formalization of these claims as future directions.

C. Stochastic Gradient Fisher Scoring

C.1. Fisher Scoring and Stochastic Gradient Fisher Scoring

Fisher scoring is simply the Newton's method for solving maximum likelihood estimation problem. The score function $S(\theta)$ is the gradient of the log-likelihood. So intuitively, if we solve the equation $S(\theta) = 0$, we can obtain the maximum likelihood estimate. Often this equation is highly non-linear, so we consider the an iterative update for the linearized score function (or a quadratic approximation of the likelihood) by Taylor expand it at the current point θ_0

$$S(\theta) \approx S(\theta_0) + I(\theta_0)(\theta - \theta_0)$$

where $I(\theta_0) = -\sum_{i=1}^n \nabla\nabla^T \ell(Z_i; \theta)$ is the observed Fisher information evaluated at θ_0.

By the fact that $S(\theta^*) = 0$, and plug in the above equation, we get $\theta^* = \theta_0 + I^{-1}(\theta_0)S(\theta_0)$ Note that this is a fix point iteration and it gives us an iterative update rule to search for θ^* via

$$\theta_{k+1} = \theta_k + I^{-1}(\theta_k)S(\theta_k).$$

Recall that S is the gradient of the score function and I^{-1} is the covariance of the score function and (under mild regularity conditions) the Hessian of the log-likelihood. As a result, this is often the same as Newton iterations.

An intuitive idea to avoid passing the entire dataset in every iteration is to simply replacing the gradient (the score function) with stochastic gradient. Then Fisher Scoring can be thought of as a quasi-newton method.

C.2. Privacy extension

By invoking a more advanced version of the Gaussian Mechanism, we will show that similar privacy guarantee can be obtained for a modified version of SGFS (described in Algorithm 4) while preserving its asymptotic properties. Specifically, under the assumption that I_N is given, when η_t is big, it also samples from a normal approximation (with larger variance), when η_t is small, the private algorithm becomes exactly the same as SGFS. Moreover, for a sequence of samples from the posterior, the online estimate in the Fisher Information converges an $O(1/N)$ approximation of true Fisher Information as in Ahn et al. (2012, Theorem 1).

The privacy result relies on a more specific smoothness assumption. Assume that for any parameter $\theta \in \mathbb{R}^d$, and $X \in \mathcal{X}^N$ the ellipsoid $E = FB^d$ defined by transforming the unit ball B^d using a linear map F contains the symmetric polytope spanned by $\{\pm\nabla\ell(x_1, \theta), ..., \pm\nabla\ell(x_N, \theta)\}$. From a differential private point of view, this implies that $\nabla_\theta \ell(x, \theta)$'s sensitivity is different towards different direction. Then the non-spherical gaussian mechanism states

Lemma 5 (Non-Spherical Gaussian Mechanism)**.** *Output* $\sum_{i=1}^N \nabla\ell(x_i, \theta) + Fw$ *where* $w \sim \mathcal{N}(0, \frac{(1+\sqrt{1\log(1/\delta)})^2}{\epsilon^2} I_d)$ *obeys* (ϵ, δ)*-DP.*

Theorem 5. *Let* F *be that* $\ell(x; \theta') \leq \ell_\theta + \nabla\ell(x; \theta)^T(\theta' - \theta) + \frac{1}{2}\|F(\theta' - \theta)\|^2$ *for any* $x \in \mathcal{X}, \theta \in \Theta$*. Moreover, let* $\epsilon, \delta, \tau, T$ *be chosen such that* $T \geq \frac{\epsilon^2 N}{32\tau\log(2/\delta)}$*. Then Algorithm 4 guarantees* $(2\epsilon, 2\delta)$*-differential privacy.*

Algorithm 4 Differentially Private Stochastic Gradient Fisher Scoring (DP-SGFS)

Require: Data X of size N, Size of minibatch τ, number of data passes T, stepsize η_t for $t = 1, ..., \lfloor NT/\tau \rfloor$, a public Lipschitz matrix F, and initial θ_1. Set $t = 1$, $\sigma^2 = \frac{32T \log(2.5NT/\tau\delta) \log(2/\delta)}{N\tau\epsilon^2}$

for $t = 1 : \lfloor NT/\tau \rfloor$ **do**

1. Random sample a minibatch $S \subset [N]$ of size τ, compute $\bar{g} = \frac{1}{\tau} \sum_{i \in S} \nabla \ell(x_i|\theta)$.
2. Sample $Z_t \sim \mathcal{N}(0, \sigma^2 \vee \frac{1}{N^2 \eta_t} I_d)$, $W_{ij} \sim \mathcal{N}(0, 49\|F\|^4 \sigma^2)$.
3. Compute private stochastic gradient and sample covariance matrix

$$\tilde{g} = \bar{g} + FZ_t, \qquad \text{and} \qquad V = \mathcal{P}_{S_+^d} \left\{ \frac{1}{\tau - 1} \sum_{i \in S} \{\nabla \ell_i(\theta_t) - \bar{g}\} \{\nabla \ell_i(\theta_t) - \bar{g}\}^T + W \right\}.$$

4. Update the guessed Fisher Information estimate $\hat{I}_t = (1 - \kappa_t)\hat{I}_{t-1} + \kappa_t V$.
5. Update and return $\theta_{t+1} \leftarrow \theta_t + 2 \left(\frac{(\tau+N)N}{\tau} \hat{I}_t + \frac{4FF^T}{\eta_t} \right)^{-1} (\nabla r(\theta_t) + N\tilde{g})$.
6. Increment $t \leftarrow t + 1$.

end for

Proof. First of all, $\|F\|_2$ is an upper bound for any $\nabla \ell(x|\theta)$, so by applying Lemma 6 on the every set of subsamples in each iteration, by Gaussian mechanism (Lemma 2) and the invariance to post-processing, we know that V is a private release. Then the proof follows by the same line of argument (subsampling and advanced composition) as in Theorem 4 for $\tilde{g}$ and V respectively, then the result follows by applying the simple composition theorem. □

Lemma 6 (Sensitivity of the sample covariance operator)**.** *Let $\|x\| \leq L$ for any $x \in \mathcal{X}$, $n > 4$, then*

$$\sup_{k, x_1, ..., x_n, x'_k} \|\widehat{\mathrm{Cov}}(x_1, ..., x_k, ..., x_n) - \widehat{\mathrm{Cov}}(x_1, ..., x'_k, ..., x_n)\|_F \leq \frac{7L^2}{n-1}.$$

Proof. We prove by taking the difference of two adjacent covariance matrices and bound the residual.

$$\begin{aligned} \widehat{\mathrm{Cov}}(X') =& \widehat{\mathrm{Cov}}(X) + \frac{1}{n-1}(xx^T - x'[x']^T) + \frac{1}{n(n-1)}(xx^T + x'[x']^T - x[x']^T - x'x^T) \\ & - \frac{1}{n-1}\mu(x - x')^T - \frac{1}{n-1}(x - x')\mu^T. \end{aligned}$$

Now assume $n > 4$ and take the upper bound of every term, we get $\Delta_2(\mathrm{Cov}(X)) \leq \frac{7L^2}{n-1}$. □

D. Additional Experiments

Figure 2 is a plain illustration of how these stochastic gradient sampler works using a randomly generated linear regression model (note the its posterior distribution will be normal, as the contour illustrates). On the left, it shows how these methods converge like stochastic gradient descent to the basin of convergence. Then it becomes a posterior sampler. The figure on the right shows that the stochastic gradient thermostat is able to produce more accurate/unbiased result and differential privacy at the level of $\epsilon = 10$ becomes negligible.

E. Additional discussions and caveats.

So far, we have proposed a differentially private Bayesian learning algorithm that is memory efficient, statistically near optimal for a large class of problems, and we can release many intermediate iterates to construct error bars. Given that differential privacy is usually very restrictive, some of these results may appear too good to be true. This is a reasonable suspicion due to the following caveats.

Small η helps both privacy and accuracy. It is true that as η goes to 0, the stationary distribution that these method samples from gets closer to the target distribution. On the other hand, since the variance of the noise we need to add for privacy scales in $O(\eta^2)$ and that for posterior sampling scales like $O(\eta)$, privacy and accuracy benefits from the same

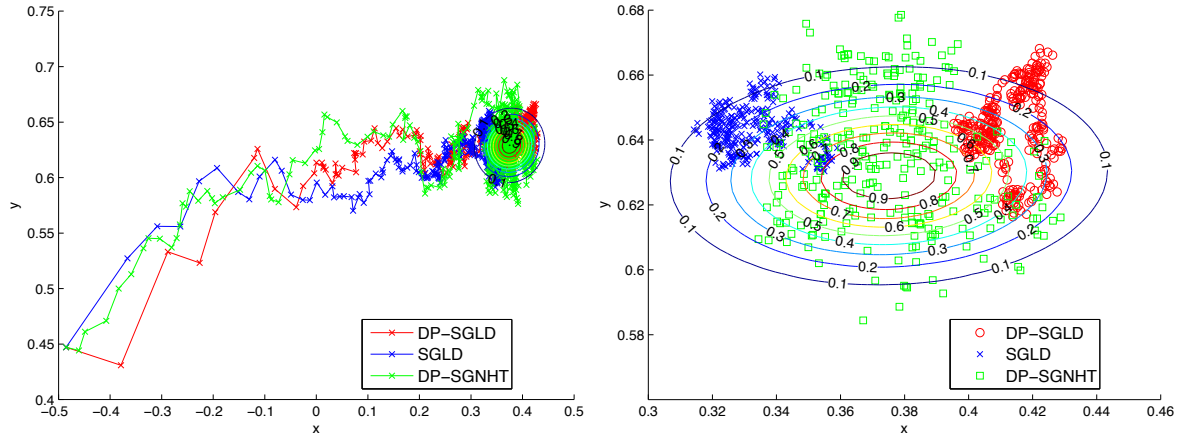

Figure 2. Illustration of stochastic gradient langevin dynamics and its private counterpart at $\epsilon = 10$.

underlying principle. The caveat is that we also have a budget on how many samples can we collect. Also the smaller the stepsize η is, the slower it mixes, as a result, the samples we collect from the monte carlo sampler is going to be more correlated to each other.

Adaptivity of SGNHT. While SGNHT is able to adaptively adjust the temperature so that the samples that it produces remain "unbiased" in some sense as $\eta \to 0$. The reality is that if the level of noise is too large, either we adjust the stepsize to be too small to search the space at all, or the underlying stochastic differential equation becomes unstable and quickly diverges. As a result, the adaptivity of SGNHT breaks down if the privacy parameter gets to small.

Computationally efficiency. For a large problem, it is usually the case that we would like to train with only one pass of data or very small number of data passes. However, due to the condition in Lemma 4, our result does not apply to one pass of data unless τ is chosen to be as large as N. While we can still choose T to be sufficiently large and stop early, but we amount of noise that we add in each iteration will remain the same.

The Curse of Numerical constant. The analysis of algorithms often involves larger numerical constants and polylogarithmic terms in the bound. In learning algorithms these are often fine because there are more direct ways to evaluate and compare methods' performances. In differential privacy however, constants do matter. This is because we need to use these bounds (including constants) to decide how much noise or perturbation we need to inject to ensure a certain degree of privacy. These guarantees are often very conservative, but it is intractable to empirically evaluate the actual ϵ of differential privacy due to its "worst" case definition. Our stochastic gradient based differentially private sampler suffers from exactly that. For moderate data size, the product of the constant and logarithmic terms can be as large as a few thousands. That is the reason why it does not perform as well as other methods despite the theoretically being optimal in scaling (the optimality result is due to SGD (Bassily et al., 2014)).

F. Related work

We briefly discuss related work here. For the first part, we become aware recently that Mir (2013) and Dimitrakakis et al. (2014) independently developed the idea of using posterior sampling for differential privacy. Mir (2013, Chapter 5) used a probabilistic bound of the log-likelihood to get $(\epsilon, \delta) - DP$ but focused mostly on conjugate priors where the posterior distribution is in closed-form. Dimitrakakis et al. (2014) used Lipschitz assumption and bounded data points (implies our boundedness assumption) to obtain a generalized notion of differential privacy. Our results are different in that we also studied the statistical and computational properties. Bassily et al. (2014) used exponential mechanism for empirical risk minimization and the procedure is exactly the same as OPS . Our difference is to connect it to Bayesian learning and to provide results on limiting distribution, statistical efficiency and approximate sampling. We are not aware of a similar asymptotic distribution with the exception of Smith (2008), where a different algorithm (the subsample-and-aggregate scheme) is proven to give an estimator that is asymptotically normal and efficient (therefore, stronger than our result) under a different set of assumptions. Specifically, Smith (2008)'s method requires boundedness of the parameter space while ours

method can work with potentially unbounded space so long as the log-likelihood is bounded.

Related to the general topic, Kasiviswanathan & Smith (2014) explicitly modeled the "semantics" of differential privacy from a Bayesian point of view, Xiao & Xiong (2012) developed a set of tools for performing Bayesian inference under differential privacy, e.g., conditional probability and credibility intervals. Williams & McSherry (2010) studied a related but completely different problem that uses posterior inference as a meta-post-processing procedure, which aims at "denoising" the privately obfuscated data when the private mechanism is known. Integrating Williams & McSherry (2010) with our procedure might lead to some further performance boost, but investigating its effect is beyond the scope of the current paper.

For the second part, the idea to privately release stochastic gradient has been well-studied. Song et al. (2013); Bassily et al. (2014) explicitly used it for differentially private stochastic gradient descent. And Rajkumar & Agarwal (2012) used it for private multi-party training. Our Theorem 4 is a simple modification of Theorem 2.1 in Bassily et al. (2014). Bassily et al. (2014) also showed that the differential private SGD using Gaussian mechanism with $\tau = 1$ matches the lower bound up to constant and logarithmic, so we are confident that not many algorithms can do significantly better than Algorithm 2. Our contribution is to point out the interesting algorithmic structures of SGLD and extensions that preserves differential privacy. The method in Song et al. (2013) requires disjoint minibatches in every data pass, and it requires adding significantly more noise in settings when Lemma 4 applies. Song et al. (2013) are however applicable when we are doing only a small number of data passes and for these cases, it gets a much better constant. Rajkumar & Agarwal (2012)'s setting is completely different as it injects a fixed amount of noise to the gradient corresponds to each data point exactly once. In this way, it replicates objective perturbation (Chaudhuri et al., 2011) (assuming the method actually finds the optimal solution).

Objective perturbation is originally proposed in Chaudhuri et al. (2011) and the (ϵ, δ) version that we refer to first appears in Kifer et al. (2012). Comparing to our two mechanisms that attempts to sample from the posterior, their privacy guarantee requires the solution to be exact while ours does not. In comparison, OPS estimator is differentially private allows the distribution it samples from to be approximate, DP-SGLD on the other hand releases all intermediate results and every single iteration is public.