DP-Fast MH: Private, Fast, and Accurate Metropolis-Hastings for Large-Scale Bayesian Inference

Wanrong Zhang [*1] **Ruqi Zhang** [*2]

Abstract

Bayesian inference provides a principled framework for learning from complex data and reasoning under uncertainty. It has been widely applied in machine learning tasks such as medical diagnosis, drug design, and policymaking. In these common applications, data can be highly sensitive. Differential privacy (DP) offers data analysis tools with powerful worst-case privacy guarantees and has been developed as the leading approach in privacy-preserving data analysis. In this paper, we study Metropolis-Hastings (MH), one of the most fundamental MCMC methods, for large-scale Bayesian inference under differential privacy. While most existing private MCMC algorithms sacrifice accuracy and efficiency to obtain privacy, we provide the first exact and fast DP MH algorithm, using only a minibatch of data in most iterations. We further reveal, for the first time, a three-way trade-off among privacy, scalability (i.e. the batch size), and efficiency (i.e. the convergence rate), theoretically characterizing how privacy affects the utility and computational cost in Bayesian inference. We empirically demonstrate the effectiveness and efficiency of our algorithm in various experiments.

1. Introduction

Bayesian inference is one of the most powerful tools in data analytics and has been applied on a wide variety of data tasks (Gelman et al., 1995). In practice, computing the posterior is often intractable and thus requires approximation methods, such as Markov chain Monte Carlo (MCMC). The Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) is one of the most commonly used MCMC methods, where it works by conducting an accept/reject step to ensure the target distribution as its equilibrium distribution. By its construction, MH is guaranteed to converge asymptotically to the true posterior.

In many applications of Bayesian inference, privacy becomes a crucial consideration as the data can be highly sensitive. For example, doctors may use Bayesian inference to analyze past patient data and find a pattern to make accurate predictions in medical diagnosis; tech companies may use Bayesian methods to classify spam emails by looking at features in customers' emails. In these tasks, we need new Bayesian inference methods that can give useful inference results under the privacy constraint.

Differential privacy (DP) has emerged as the gold standard for privacy-preserving data analytics. It is a statistical notion of privacy, which ensures that the output distribution of a computation is robust when we arbitrarily change one person's data. DP algorithms hide individuals' data by introducing randomness. While there exists a few DP MCMC algorithms, they have significant limitations on their utility and applicability. In particular, most work has no guarantees on the asymptotic convergence or the convergence speed, thus could diverge far away from the target posterior in practice (Heikkilä et al., 2019; Li et al., 2019; Zhang et al., 2021). Some work significantly scarifies the convergence rate for privacy and can not scale to large datasets (Yıldırım & Ermiş, 2019). Moreover, most existing methods operate on the belief that there are trade-offs among privacy, efficiency (i.e. the convergence rate), and scalability (i.e. the batch size) in private Bayesian inference, but none of them formally characterize these trade-offs.

In this work, we solve all the above issues by providing a simple yet powerful DP MCMC algorithm, *DP-Fast MH*, with theoretically guarantees on privacy, utility, and scalability. To the best of our knowledge, our method is the *first* exact minibatch DP MCMC algorithm, which enables private and accurate Bayesian inference on large-scale datasets. Built upon TunaMH (Zhang et al., 2020) which is a non-private exact minibatch MH method, we developed DP-Fast MH by injecting an appropriate amount of Gaussian noise in

[*]Equal contribution [1]Harvard John A. Paulson School of Engineering and Applied Sciences [2]Department of Computer Science, Purdue University . Correspondence to: Wanrong Zhang <wanrongzhang@fas.harvard.edu>, Ruqi Zhang <ruqiz@purdue.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

the MH algorithm while still keeping the reversibility of the Markov chain. Notably, we show that due to the inherent randomness in the accept/reject step in MH, *privacy for free* can be achieved under certain cases without adding additional noise. This finding drastically improves the privacy-utility trade-off in previous DP MH methods. Theoretically, we give privacy guarantees (Theorem 2) following subsampling analysis and instantiated Gaussian mechanisms. We also show the exactness of our algorithm (Theorem 3), which guarantees the asymptotic convergence, meaning that the distribution is close to the target distribution when the number of iterations gets large. Our convergence rate theorem statement (Theorem 4) further quantifies the cost of privacy in terms of the convergence rate and batch size. Empirically, we demonstrate the accuracy of our estimated distribution on several common Bayesian inference tasks. We summarize our contributions as follows:

- We introduce DP-fast MH, the first private, asymptotically-exact, minibatch Metropolis-Hastings algorithm. Taking advantage of the inherent randomness in MH, we achieve privacy for free under certain conditions, which greatly improves the privacy-utility trade-off in existing differentially private MH methods.
- We prove (ϵ, δ)-DP guarantees, asymptotic convergence, and convergence rate bound for DP-fast MH, formally characterizing a three-way trade-off among privacy, scalability (i.e. the batch size), and efficiency (i.e. the convergence rate).
- We empirically validate our theorems, and demonstrate that our algorithm significantly outperforms previous methods in terms of both estimation accuracy and computational cost under various settings.

2. Related Work

Differentially private posterior sampling has recently gained attractions, with the methods falling into three categories: DP Metropolis-Hastings (Heikkilä et al., 2019; Yıldırım & Ermiş, 2019), DP Gradient-based MCMC (Li et al., 2019; Zhang et al., 2021; Räisä et al., 2021), and DP Exact Posterior Sampling (Dimitrakakis et al., 2014; Wang et al., 2015; Foulds et al., 2016; Zhang et al., 2016)

DP Metropolis-Hastings Differentially private Metropolis-Hastings has been developed in Heikkilä et al. (2019); Yıldırım & Ermiş (2019). Heikkilä et al. (2019) gives an algorithm based on the Barker acceptance test, which can be used with a minibatch of data. However, the algorithm has an asymptotic bias, making it unreliable in practice: the stationary distribution might be arbitrarily far away from the target distribution. Moreover, there is no theoretical convergence rate guarantee and setting the privacy budget upfront is hard. Yıldırım & Ermiş (2019) considers the penalty algorithm (Ceperley & Dewing, 1999) where Gaussian noise is added to the log-likelihood difference. This method converges to the correct distribution asymptotically, but the convergence speed decreases significantly due to adding a large amount of noise. Also, this method uses the entire dataset to evaluate the acceptance probability in each iteration, so it is computationally infeasible with large datasets.

DP Gradient-based MCMC Another line of work studies (stochastic) gradient-based MCMC methods (Li et al., 2019; Zhang et al., 2021; Räisä et al., 2021). Differentially private Stochastic Gradient MCMC (SGMCMC), including Stochastic Gradient Langevin Dynamics (SGLD) (Welling & Teh, 2011) and other variants, are shown to satisfy differential privacy guarantees by default. However, DP SGMCMC has an asymptotic bias and also lacks non-asymptotic convergence guarantees. It is worth noting that DP SGLD may have unbounded private loss even if the exact posterior is differentially private as desired (Heller & Fetaya, 2021). Räisä et al. (2021) develops a DP version of Hamiltonian Monte Carlo (Neal et al., 2011) whose MH step is based on Yıldırım & Ermiş (2019), thus inheriting the drawbacks of slow convergence and expensive computation.

DP Exact Posterior Sampling Dimitrakakis et al. (2014); Wang et al. (2015); Foulds et al. (2016); Zhang et al. (2016) consider differentially private posterior sampling, where the posterior is assumed to be obtained exactly. In turn, they only offer conditional privacy guarantees, which rely on the exact sampling assumption. The posterior is often intractable in practice, and it is usually impossible to validate the assumption in practical algorithms.

3. Preliminaries

This section provides necessary background on Metropolis-Hastings and differential privacy for understanding our algorithm and theoretical results.

3.1. Metropolis-Hastings

In Bayesian inference, given a dataset $D = \{x_i\}_{i=1}^N$ and a θ-parameterized model, we aim to compute the posterior distribution

$$\pi(\theta) \propto \exp\left(-\sum_{i=1}^N U_i(\theta)\right) = \exp\left(\sum_{i=1}^N \log p(x_i|\theta) + \log p(\theta)\right)$$

where $U_i(\theta)$ is the energy function, $p(x_i|\theta)$ is the likelihood

and $p(\theta)$ is the prior. The Metropolis-Hastings (MH) algorithm approximates the posterior by sampling from it. In each step, MH generates a proposal θ' from some distribution $q(\cdot \mid \theta)$ and accepts θ' with probability

$$a(\theta, \theta') = \min\left(1, \exp\big(\sum_{i=1}^{N}(U_i(\theta) - U_i(\theta'))\big) \cdot \frac{q(\theta|\theta')}{q(\theta'|\theta)}\right).$$

The construction of $a(\theta, \theta')$ ensures the reversibility of the Markov chain, which is a common condition to guarantee correct asymptotic convergence.

TunaMH The acceptance probability $a(\theta, \theta')$ in MH requires computing a sum over the entire dataset, which could be prohibitively expensive on large datasets. A common way to solve this issue is to use a minibatch approximation. In this paper, we build our DP MH algorithm upon a recently developed exact minibatch MH method, TunaMH (Zhang et al., 2020), which uses Poisson auxiliary variables to enable minibatching while still keeping the correct asymptotic convergence. The algorithm of TunaMH is outlined in Algorithm 1. At each iteration, TunaMH forms a minibatch $\mathcal{I}$ by sampling a batch size B from a Poisson distribution, and then sample (with replacement) a minibatch of size B to compute the acceptance probability. The expected batch size has been shown to be upper bounded by $\mathbf{E}[B] = \lambda + CM(\theta, \theta')$. Different from the original TunaMH, we keep the hyperparameter λ to control the batch size rather than setting $\lambda = \chi C^2 M^2(\theta, \theta')$ as suggested in the paper. The reason of keeping λ independent of $CM(\theta, \theta')$ is to bound the sensitivity of energy difference, which will be seen later in Section 4.

Algorithm 1 TunaMH

given: initial state $\theta \in \Theta$; proposal dist. q; hyperparameter λ; Asm. 1 parameters c_i, C, M
loop
  propose $\theta' \sim q(\cdot|\theta)$ and **compute** $M(\theta, \theta')$
  sample $B \sim$ Poisson $(\lambda + CM(\theta, \theta'))$
  obtain batch indices $\mathcal{I} \leftarrow \texttt{Batch}(B)$
  obtain MH Ratio $r \leftarrow \texttt{MHRatio}(\mathcal{I}, 0)$
  with probability $\min(1, r)$, set $\theta \leftarrow \theta'$
end loop

Function $\texttt{Batch}(B)$
initialize minibatch indices $\mathcal{I} \leftarrow \emptyset$
for $b \in \{1, \ldots, B\}$ **do**
  sample i_b such that $\mathbf{P}(i_b = i) = c_i/C$, for $i = 1 \ldots N$
  with probability $\frac{\lambda c_{i_b} + \frac{C}{2}(U_{i_b}(\theta') - U_{i_b}(\theta) + c_{i_b} M(\theta, \theta'))}{\lambda c_{i_b} + c_{i_b} CM(\theta, \theta')}$
  add i_b to $\mathcal{I}$
end for
return $\mathcal{I}$

Function $\texttt{MHRatio}(\mathcal{I}, \tau)$
$r \leftarrow \exp\left(2\sum_{i \in \mathcal{I}} \operatorname{artanh}\left(\frac{C\left(U_i(\theta) - U_i(\theta')\right)}{c_i(2\lambda + CM(\theta, \theta'))}\right) + \tau\right) \cdot \frac{q(\theta|\theta')}{q(\theta'|\theta)}$
return r

3.2. Differential Privacy

Differential privacy is a mathematical framework for privacy-preserving data analysis. It bounds the maximum amount that one person's data can affect the analysis performed on the dataset. We say two datasets D, D' are neighboring if they arbitrarily differ in the values at most one entry.

Definition 1 (Differential Privacy (Dwork et al., 2006))**.** *A randomized algorithm $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ is* (ϵ, δ)-differentially private *if for every pair of neighboring datasets $D, D' \in \mathcal{D}$, and for every subset of possible outputs $\mathcal{S} \subseteq \mathcal{R}$,* $\Pr[\mathcal{M}(D) \in \mathcal{S}] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in \mathcal{S}] + \delta$.

Differentially private mechanisms typically add noise that scales with the sensitivity of the function being evaluated and the idea is to use randomness to hide each individual's contribution. The sensitivity of a function f is defined as the maximum change in f between two neighboring sets: $\Delta f = \max_{D, D' \text{ neighbors}} |f(D) - f(D')|$ The Gaussian mechanism with parameters (ϵ, δ) takes in a function f, dataset D, and outputs $f(X) + \mathcal{N}(0, \sigma^2)$, where $\sigma = \sqrt{2\log(1.25/\delta)}\Delta f/\epsilon$.

Theorem 1 (Privacy of Gaussian Mechanism (Dwork & Roth, 2014))**.** *For any $\epsilon, \delta \in (0, 1)$, the Gaussian Mechanism with parameter $\sigma = \sqrt{2\log(1.25/\delta)}\Delta f/\epsilon$ is (ϵ, δ)-differentially private.*

Differential privacy is also robust under post-processing, which means once an output is differentially private, we can perform more computations based on the private output without leaking any more information than the privacy parameters. This property allows us to build our DP MH algorithm that privatizes the intermediate samples.

Differentially private algorithms compose adaptively. The advanced composition (Dwork et al., 2010) shows that the privacy parameter is $(O(\epsilon\sqrt{T\log 1/\delta}), T\delta)$ under T iterations. Recently, several relaxations of differential privacy including Rènyi differential privacy (RDP) (Mironov, 2017), and f-differential privacy (f-DP) (Dong et al., 2019) allow for tighter reasoning about composition. Moreover, reasoning the overall privacy guarantee of our entire algorithm is similar to that of the DP stochastic gradient descent (DP-SGD) algorithm (Abadi et al., 2016). The moments accountant technique in that paper was proposed specifically for analyzing DP-SGD, which gives a tighter bound: $(O(\epsilon\sqrt{T}), \delta)$ under T iterations.

Algorithm 2 DP-Fast MH

given: initial state $\theta \in \Theta$; proposal dist. q; hyperparameter λ; Asm. 1 parameters c_i, C, M, privacy parameters ϵ, δ, batch size upper bound K.
Set noise scale $\sigma_2 = \sqrt{2\log\frac{1.25}{\delta}}/\epsilon$ and $\sigma_1 = 6K\max_i c_i\sqrt{2\log\frac{2.5K\max_i c_i}{\delta C}}/(\epsilon C)$.
loop
  propose $\theta' \sim q(\cdot|\theta)$ and **compute** $M(\theta,\theta')$
  sample $B \sim$ Poisson $(\lambda + CM(\theta,\theta'))$
  if $B < K$ **then** {Run minibatch MH}
    obtain batch indices $\mathcal{I} \leftarrow \texttt{Batch}(B)$
    if $\Delta(\ell_1) \le \epsilon C/(6K\max_i c_i)$ **then** {Privacy for free}
      Let $\xi = 0$.
    else
      Sample $\xi \sim \mathcal{N}(0, \sigma_1^2\Delta(\ell_1)^2)$.
    end if
    obtain MH Ratio $r \leftarrow \texttt{MHRatio}(\mathcal{I}, \xi - \frac{\sigma_1^2\Delta(\ell_1)^2}{2})$
  else {Run full-batch MH}
    if $\Delta(\ell_2) \le \epsilon$ **then** {Privacy for free}
      Let $\xi = 0$.
    else
      Sample $\xi \sim \mathcal{N}(0, \sigma_2^2\Delta(\ell_2)^2)$
    end if
    $r \leftarrow \exp\left(\sum_{i=1}^N (U_i(\theta) - U_i(\theta')) + \xi - \frac{\sigma_2^2\Delta(\ell_2)^2}{2}\right) \cdot \frac{q(\theta|\theta')}{q(\theta'|\theta)}$
  end if
  with probability $\min(1, r)$, set $\theta \leftarrow \theta'$
end loop

4. DP-Fast MH: Private, Fast, and Accurate Metropolis-Hastings

In this section, we present our private exact minibatch Metropolis-Hastings algorithm, termed *DP-Fast MH*. Following previous work on exact minibatch MH methods (Cornish et al., 2019; Zhang et al., 2020), we have the following assumption on the posterior distribution.

Assumption 1. *For some constants $c_1, \ldots, c_N \in \mathbb{R}_+$, with $\sum_i c_i = C$, and symmetric function $M : \Theta \times \Theta \to \mathbb{R}_+$, for any $\theta, \theta' \in \Theta$, the energy difference is bounded by $|U_i(\theta) - U_i(\theta')| \le c_i M(\theta, \theta')$.*

This assumption holds for many common Bayesian inference problems. For example, if each energy function U_i is L_i-Lipschitz continuous, then we can set $c_i = L_i$ and $M(\theta, \theta') = \|\theta - \theta'\|$.

Assumption 1 assumes a local bound for the energy function, however, to ensure differential privacy, we need a global bound. Hence, we introduce Assumption 2, which has been used in all previous DP MH work (Heikkilä et al., 2019; Yıldırım & Ermiş, 2019).

Assumption 2. *We additionally assume $M(\theta, \theta') \le A, \forall \theta, \theta' \in \Theta$ for some constant A.*

Our algorithm is given formally in Algorithm 2 where the differences compared to the non-private baseline TunaMH are highlighted in blue color. In particular, there are three key differences to privatize MH while maintaining its original utility, which we discuss in details separately below.

Upper Bound on the Batch Size First, we add an additional hyperparameter K to decide whether to use a minibatch or the entire dataset. If the batch size B is less than the upper bound K, the algorithm will form a minibatch as in TunaMH; otherwise, it uses the entire dataset as in the standard MH. This modification is critical to guarantee privacy because the batch size B is a random variable ranging from 0 to ∞. Since we form the minibatch by sampling with replacement, it has two competing effects on privacy. On one hand, the possibility that an individual is sampled multiple times incurs additional privacy loss. On the other hand, the possibility that an individual is not sampled into the minibatch amplifies privacy. Intuitively, if the batch size is small, then the second effect dominates. Therefore, we limit the upper bound of the batch size to be a constant K.

In practice, K is often large enough, and most iterations will use minibatches. One may wonder whether we can use sampling without replacement to avoid the first effect. As shown in Zhang et al. (2020), doing this will make the Markov chain not satisfy the reversibility, thus the algorithm will not preserve the correct stationary distribution.

Added Gaussian Noise Second, DP-fast MH instantiates the Gaussian mechanism to preserve privacy, which adds Gaussian noise to the energy differences $\ell_1 = 2\sum_{i\in\mathcal{I}} \operatorname{artanh}\left(\frac{C(U_i(\theta)-U_i(\theta'))}{c_i(2\lambda+CM(\theta,\theta'))}\right)$ (or $\ell_2 = \sum_{i=1}^N (U_i(\theta) - U_i(\theta'))$ when using the standard MH). We note that similar idea of adding Gaussian noise appeared in the penalty algorithm Ceperley & Dewing (1999) in the non-private literature and was used as a DP MH algorithm later in Yıldırım & Ermiş (2019). The Gaussian noise is scaled with the sensitivity of the energy difference, which is given as follows.

Lemma 1. *The sensitivity of the energy difference is*

$$\Delta(\ell_1) = 2\log\left(1 + \frac{CM(\theta,\theta')}{\lambda}\right) \tag{1}$$

$$\Delta(\ell_2) = 2\max_i c_i M(\theta,\theta'). \tag{2}$$

Both $\Delta(\ell_1)$ and $\Delta(\ell_2)$ monotonically increase as $M(\theta,\theta')$ and C (or $\max_i c_i$) increase. This indicates that the added

noise per iteration is large when the proposed θ' is far away from the current θ or the Lipschitz constant of the energy function is large. Another interesting observation is that $\Delta(\ell_2)$ is generally smaller than $\Delta(\ell_1)$. This indicates that, to make one iteration the same level of privacy, we need to add *more* noise in the minibatch MH than in the standard MH and more noise means slower convergence. While using minibatches causes a decrease in the convergence rate of MCMC in the non-private setting (Zhang et al., 2020), here, we show that the privacy constraint further amplifies this decrease. To derive noise scale in Algorithm 2, we further show that the sampling with replacement step affects the privacy by a $6K \max_i c_i/C$ factor (see the proof of Theorem 2 in Appendix B for details). Thus, by setting the standard deviation of the Gaussian mechanism to $\epsilon C/(6K \max_i c_i)$ when using minibatches, we maintain the same level of privacy per iteration at ϵ as using full-batch data.

After obtaining a Gaussian noise ξ, we add it to the energy difference and additionally subtract a noise variance$/2$. The subtraction is to ensure that the private Markov chain still satisfies the reversibility and thus is guaranteed to be close to the target distribution along running (Ceperley & Dewing, 1999; Yıldırım & Ermiş, 2019).

Privacy for Free Third, we *only* add the additional Gaussian noise when the sensitities are larger than some constants. This is because we can achieve privacy for free due to the *intrinsic* randomness in the accept/reject step of the MH methods. Specifically, in the MH correction step, we `accept` or `reject` the proposed θ' with a probability depends on the energy functions U_i. For any two neighboring datasets D and D', we consider the following probability ratio

$$\frac{\Pr(\texttt{accept}\ \ \theta'|D)}{\Pr(\texttt{accept}\ \ \theta'|D')} = \frac{\min(0, r(D)}{\min(0, r(D')} \leq \exp(\Delta(\ell_1)).$$

Therefore, if $\Delta(\ell_1) \leq \epsilon C/(6K \max_i c_i)$, or similarly, $\Delta(\ell_2) \leq \epsilon$, then the inherent randomness in the MH correction step can achieve differential privacy guarantee, and no additional noise is needed! Since $\Delta(\ell_1)$ and $\Delta(\ell_2)$ depend on the proposed θ', we compare them against the privacy parameter at each iteration.

As a comparison, Yıldırım & Ermiş (2019) does not fully leverage the inherent privacy guarantees in the MH algorithm. Therefore, they added unnecessary noise given a target privacy level, which significantly slows down the convergence of the Markov chain as shown in Section 6. The Barker method Heikkilä et al. (2019) relies on the element of randomness in the accept-reject decision, while the level of privacy is directly tied to the quantity of available randomness. Our method does not suffer from this limitation as we can add more Gaussian noise for stronger privacy without sacrificing convergence to the target.

In summary, at each iteration, DP-fast MH first samples a minibatch size and checks if it uses a minibatch of data or full-batch data. Then it checks whether to require additional Gaussian noise. If so, it will instantiate the Gaussian mechanism which adds Gaussian noise to the energy difference function. Finally, it chooses `accept` or `reject` θ' based on the noisy acceptance probability. Compared to TunaMH, our algorithm only introduces one additional hyperparameter K, which upper bounds the minibatch size.

5. Theoretical Analysis on Privacy and Convergence

5.1. Privacy Analysis

As a starting point, when the sensitivity of the energy differences $\Delta(\ell_1)$ and $\Delta(\ell_2)$ are small, we get privacy for free from the randomness of accepting the proposed θ'. It suffices to analyze privacy that comes from the Gaussian mechanism, and privacy amplification by the sampled minibatch for each iteration. While the subroutine is similar to DP-SGD (Abadi et al., 2016), our privacy guarantee requires new analysis, since the batch size in our algorithm is a random variable (in contrast to a constant), the minibatch is formed by sampling with replacement (in contrast to without replacement), and different samples have different probabilities being chosen ((in contrast to uniform sampling). Finally, We also account for the possibility of using the standard MH, where the privacy is ensured by instantiating the Gaussian mechanism again. We provide the privacy guarantee for each iteration in Theorem 2, and the formal proof is deferred to Appendix B.

Theorem 2 (Privacy)**.** *Under Assumption 1 and 2, DP-fast MH (Algorithm 2) is (ϵ, δ)-differentially private per iteration.*

We remark that one can obtain the total privacy loss by any composition theorems over T iterations. Since these bounds are usually loose and the existing DP MH algorithms (Yıldırım & Ermiş, 2019; Heikkilä et al., 2019) use different approaches to account for the total privacy loss, we compare our DP-Fast MH algorithm against theirs by fixing the privacy budget for each iteration and the total number of iterations.

5.2. Convergence Analysis

Unlike previous DP MCMC work which typically do not come with convergence analysis, we provide both asymptotic and non-asymptotic convergence guarantees for our method. To analyze asymptotic convergence, we first notice that the way we add a Gaussian noise inside the exponential function does not affect the reversibility of the Markov chain (Ceperley & Dewing, 1999). Then, we show that the combination of the minibatch MH and full-batch MH still

results in a reversible transition kernel, since the probability of deciding whether to use minibatches $\Pr(B < K)$ is symmetric in θ and θ'. Based on the above two reasons, DP-fast MH still converges to the target distribution asymptotically. Theorem 3 formally describes our method's asymptotic accuracy.

Theorem 3 (Asymptotic Convergence)**.** *Under Assumption 1 and 2, DP-fast MH (Algorithm 2) is reversible with the stationary distribution π.*

The proof is given in Appendix C. This theorem guarantees that the posterior approximation from our algorithm will be closer to the target distribution when the number of iterations gets larger.

We now turn our attention to the non-asymptotic convergence of DP-fast MH. Although the correct asymptotic convergence of our method preserves, the convergence rate could be significantly affected by the privacy constraint, which may greatly slow down the algorithm's convergence in practice. Luckily, we show below that the convergence rate of DP-fast MH compared to the standard full-batch non-private MH can only be slowed down by at most a constant factor. We quantify the convergence rate using *spectral gap*, which has been widely used in the MCMC literature (Hairer et al., 2014; Levin & Peres, 2017; Zhang & De Sa, 2019; Zhang et al., 2020). The Markov chain with a larger spectral gap will converge faster.

Theorem 4 (Convergence Rate)**.** *Under Assumption 1 and 2, let $\bar{\gamma}$ denote the spectral gap of DP-fast MH (Algorithm 1), and let γ denote the spectral gap of the standard MH with the same target distribution and the proposal distribution. Then*

$$\frac{\bar{\gamma}}{\gamma} \geq \left(1 - \Phi\left(\frac{216K^2 \max_i c_i^2 \log \frac{2.5K \max_i c_i}{\delta C} \log^2\left(1 + \frac{CA}{\lambda}\right)}{\epsilon^2 C^2}\right)\right) \cdot \exp\left(-\frac{C^2A^2}{\lambda} - 2\sqrt{\frac{C^2A^2}{\lambda}\log 2}\right).$$

where Φ denotes the cumulative distribution function of the standard Gaussian distribution. This theorem shows the relative convergence rate of DP-fast MH with the standard MH. There are three main takeaways from the theorem. First, this theorem shows that when either the privacy hyperparameter ϵ or δ becomes small, the convergence rate becomes small, characterizing how much the privacy constraint slows down the convergence speed of the Markov chain.

Furthermore, it reveals a new *three-way* trade-off among privacy, batch size, and convergence rate: given a target privacy, a smaller batch size (i.e. larger K or smaller λ) results in a worse convergence rate (i.e. the lower bound for $\bar{\gamma}/\gamma$ becomes smaller). This trade-off shows that there exists a sweet spot for the hyperparameters K and λ such that the chain achieves an optimal performance considering both estimation accuracy and computational costs. It is generally difficult to obtain theoretical optimal values and we discuss how to tune them in practice in Section 6.

Finally, our bound shows that when C (the Lipschitz constant of the energy function) or A (the distance between θ and θ') becomes larger, the convergence rate becomes worse, indicating that DP-fast MH is more efficient when the target posterior has a small log-Lipschitz constant or when the proposed θ' is closer to the current position.

6. Experiments

We experimentally demonstrate the accuracy, efficiency and scalability of DP-fast MH under different privacy budget. In all experiments, we set the privacy parameter $\delta = 10^{-5}$. In Section 6.1, we compare DP-fast MH with existing methods under two settings: truncated Gaussian mixture and logistic regression on the MNIST dataset . In Section 6.2, we discuss the effect of the newly introduced hyperparameter K. For other hyperparameters in the algorithm which are already presented in the baseline method TunaMH, we set the values following Zhang et al. (2020). We released the code at `https://github.com/ruqizhang/dpfastmh`.

6.1. Comparison with Other Private Algorithms

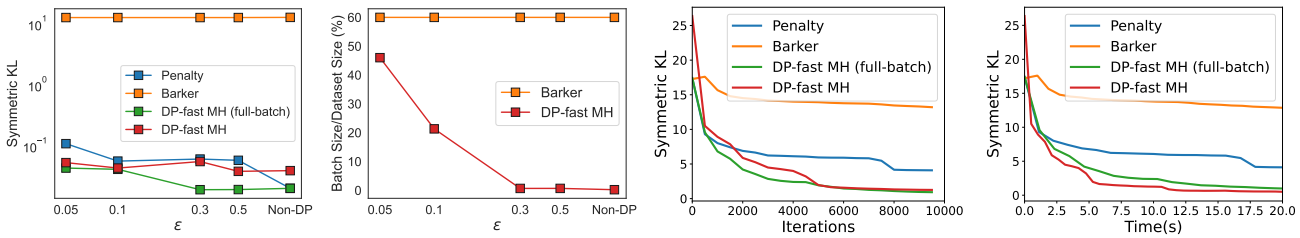
We compare DP-fast MH with the existing DP MH algorithms: the penalty algorithm (Yıldırım & Ermiş, 2019) which adds Gaussian noise to the energy difference and uses full-batch data in every iteration; and the Barker method (Heikkilä et al., 2019) which uses the Barker acceptance test (Barker, 1965) based on a minibatch approximation of the energy difference. We additionally show the results of the full-batch version of DP-fast MH (i.e. Algorithm 2 that always instantiates the full-batch MH). Since DP-fast MH trades convergence rate for scalability, DP-fast MH with full-batch data can be regarded as the upper limit of DP-fast MH in terms of performance, if ignoring the computational cost. We also include the results of the corresponding non-DP version for all DP MH methods as references.

Truncated Gaussian Mixture. We first test DP-fast MH on a two-dimensional truncated Gaussian mixture. Following previous work (Welling & Teh, 2011; Heikkilä et al., 2019; Zhang et al., 2020), the data is generated as follows

$$x_i \sim \frac{1}{2}\mathcal{N}(\theta_1, \sigma_x^2) + \frac{1}{2}\mathcal{N}(\theta_1 + \theta_2, \sigma_x^2)$$

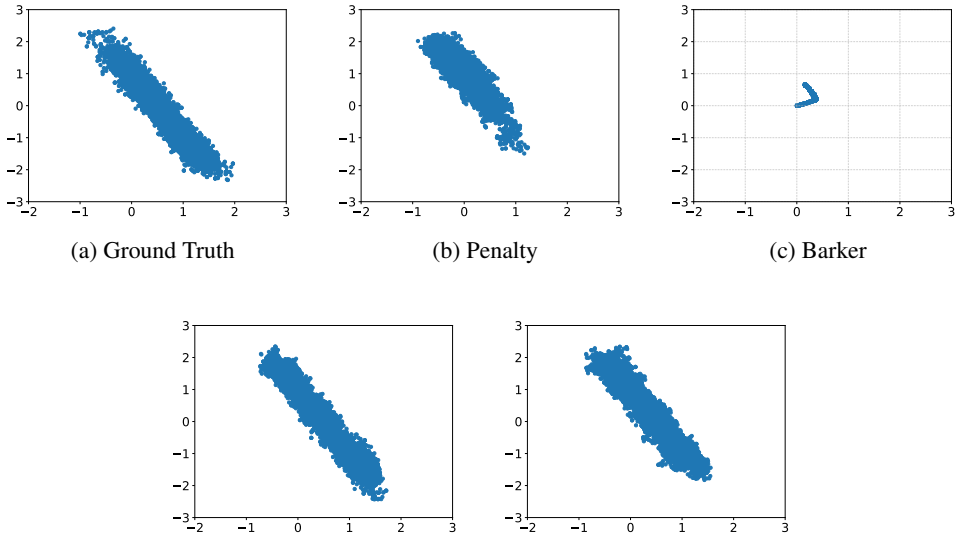
where $\theta_1 = 0, \theta_2 = 1$ and $\sigma^2 = 2$. The posterior θ has two modes at $(\theta_1, \theta_2) = (0, 1)$ and $(\theta_1, \theta_2) = (1, -1)$. We generate $N = 50000$ samples and use a temperature 500.

From Figure 1(a), we can see that DP-fast MH and DP-



(a) (b) (c) (d)

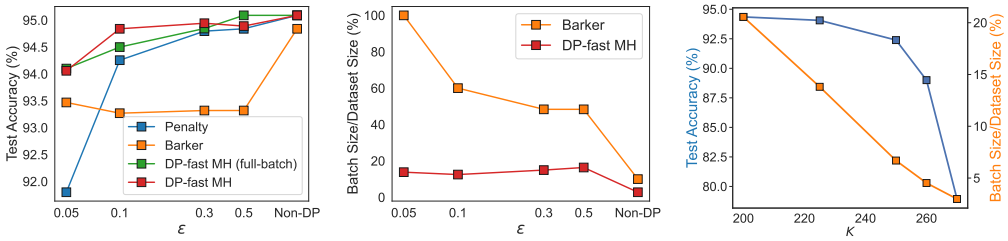
Figure 1. Truncated Gaussian mixture. (a) Symmetric KL vs. privacy. (b) Batch size vs. privacy. (c)&(d) When privacy $\epsilon = 0.05$, symmetric KL vs. iterations and time. DP-fast MH has the best results in terms of estimation accuracy, scalability, and efficiency.



(a) Ground Truth (b) Penalty (c) Barker

(d) DP-fast MH (full-batch) (e) DP-fast MH

Figure 2. Density estimation of Gaussian mixtures after 10^4 steps. The estimations of DP-fast MH and its full-bath version are close to the ground truth whereas previous methods are far away from the target distribution.



(a) (b) (c)

Figure 3. MNIST logistic regression. (a) Test accuracy vs. privacy. (b) Batch size vs. privacy. (c) Effect of hyperparameters K. DP-fast MH achieves the highest accuracy with the smallest batch size under all privacy levels.

fast MH (full-batch) have better estimation accuracy than previous methods under all privacy levels. Compared to the penalty algorithm, DP-fast MH (full-batch) takes advantage of privacy for free, which greatly improves the convergence of the chain. DP-fast MH (i.e. the minibatch version) has comparable performance with the full batch version, but drastically reduces the cost per step (see Figure(b)). We found that the Barker method is unable to converge to the target distribution because of using a very small step size to satisfy its assumptions[1]. Besides, these results verify our theoretical results on the privacy-efficiency trade-off in Theorem 4, which shows the efficiency (i.e. the convergence rate) becomes worse when the privacy level increases.

From Figure 1(b), we observe that DP-fast MH achieves a much better posterior approximation with a much smaller batch size compared with the existing minibatch method, Barker method. (Note that we do not include the Penalty method because it is a full-batch method.) This is because DP-fast MH inherits the nice properties from the non-DP baseline TunaMH, which is proved to have an asymptotically optimal batch size. These results also verify our theoretical results on the privacy-scalability trade-off in Theorem 4, which shows the batch size needs to be larger to keep a high efficiency when the privacy level increases.

Figure 1(c)&(d) show the convergence with respect to the number of iterations and the wall-clock time under $\epsilon = 0.05$ respectively. We see that DP-fast MH converges the fastest, since it converges after 4000 iterations (or 15 seconds), whereas the penalty algorithm has not fully converged after 10^4 iterations (or 20 seconds), and the Barker method still has large KL divergence after 10^4 iterations. Besides, DP-fast MH (full-batch) is also faster than the penalty algorithm, showing the benefits of utilizing the intrinsic randomness in MH to achieve privacy for free.

To illustrate the quality of estimation, we also visualize the density estimate after 10^4 steps in Figure 2. The estimated densities from both DP-fast MH and DP-fast MH (full-batch) are very close to the true posterior distribution (see Figure 2(d)&(e)), while the penalty algorithm cannot estimate the relative weights of the two modes accurately and the Barker method completely failed. These results verify the correct convergence of DP-fast MH and show that under the same privacy budget, DP-fast MH approximates the target distribution significantly better than existing methods.

Logistic Regression on MNIST Data. Next, we apply DP-fast MH to logistic regression on the MNIST image dataset. Following Zhang et al. (2020), the task is to classify 7s and 9s using the first 50 principal components as features. From Figure 3(a), we again find that minibatch and full-batch DP-fast MH both outperform previous methods significantly on test accuracy under all privacy levels, demonstrating that DP-fast MH achieves a better privacy-utility trade-off due to leveraging the inherent randomness in the MH algorithm. Similar to Figure 1(b), Figure 3(b) also demonstrates the better scalability of DP-fast MH compared to the previous minibatch method. When the privacy parameter ϵ goes smaller, the previous method has to use almost the entire dataset while our method only requires less than 20% data to obtain even better accuracy.

DP-fast MH vs. DP-fast MH (Full-Batch). DP-fast MH can be used with minibatch or full-batch of data. We now discuss their usage scenarios. The minibatch version of DP-fast MH will be preferred when the dataset size is large or the computational resource is limited. The full-batch version of DP-fast MH will be preferred if the dataset size is relatively small such that the number of convergence steps, rather than the cost per step, is the main concern.

6.2. Choice of K

DP-fast MH only introduces one new hyperparameter K, which is the minibatch upper bound, compared to the baseline non-DP method TunaMH. We now discuss how to choose this hyperparameter K in practice. Given a target privacy budget, K has two effects: (1) it affects the computation efficiency, since if the random batch size is greater than K, the standard MH will be invoked (as shown in Algorithm 2), which requires computation on the entire data. So the computational cost increases as K decreases. (2) Meanwhile, K affects the convergence rate of the chain. As shown in Theorem 4, a smaller K results in a larger spectral gap upper bound, thus a faster convergence rate.

We use the logistic regression task to investigate the performance under different choices of K with privacy parameter (in the Gaussian mechanism) $\epsilon = 0.05$. Figure 3(c) verifies our theoretical analysis on K, since the test accuracy and the batch size both increase when K decreases. We recommend setting K to be around $\epsilon C/(6 \max_i c_i)$ to achieve a good trade-off between efficiency and scalability, since the privacy parameter in the Gaussian mechanism of the minibatch MH (i.e. $\epsilon C/(6K \max_i c_i)$) is roughly a constant under this value of K.

7. Conclusions and Future Work

We present the first exact minibatch DP MCMC algorithm, DP-fast MH, for private large-scale Bayesian inference. Our method is developed by using Gaussian mechanisms in a way that maintains the reversibility of the Markov chain. Furthermore, we leverage the intrinsic randomness in the Metropolis-Hastings algorithm to achieve privacy for free under uncertain cases. We provide privacy, asymptotic

[1] In the Gaussian mixture experiment in Heikkilä et al. (2019), the barker method has used a pre-trained solution from variational inference as its initialization. Here, we train all methods from scratch.

convergence, and convergence rate guarantees, revealing a new three-way trade-off among privacy, scalability, and efficiency in Bayesian inference. We demonstrate on Gaussian mixture and logistic regression tasks that DP-fast MH significantly outperforms previous methods, indicating that our method has a much better privacy-utility trade-off. Moreover, DP-fast MH is simple to implement and can be broadly applied as a drop-in replacement for the standard MH when the target distribution has a bounded energy function.

While Bayesian inference is one of the major tools for analyzing data, a general inference algorithm that can guarantee privacy while retaining accuracy, scalability, and efficiency was largely unexplored. We believe this work fills an important gap and will accelerate the practical use of Bayesian inference for privacy-critical applications.

Acknowledgement

W.Z. is supported by a Computing Innovation Fellowship from the Computing Research Association (CRA) and the Computing Community Consortium (CCC).

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Barker, A. A. Monte carlo calculations of the radial distribution functions for a proton? electron plasma. *Australian Journal of Physics*, 18(2):119–134, 1965.

Ceperley, D. and Dewing, M. The penalty method for random walks with uncertain energies. *The Journal of chemical physics*, 110(20):9812–9820, 1999.

Cornish, R., Vanetti, P., Bouchard-Côté, A., Deligiannidis, G., and Doucet, A. Scalable metropolis-hastings for exact bayesian inference with large datasets. In *International Conference on Machine Learning*, pp. 1351–1360. PMLR, 2019.

Dimitrakakis, C., Nelson, B., Mitrokotsa, A., and Rubinstein, B. I. Robust and private bayesian inference. In *International Conference on Algorithmic Learning Theory*, pp. 291–305. Springer, 2014.

Dong, J., Roth, A., and Su, W. J. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pp. 265–284, 2006.

Dwork, C., Rothblum, G. N., and Vadhan, S. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 51–60. IEEE, 2010.

Foulds, J., Geumlek, J., Welling, M., and Chaudhuri, K. On the theory and practice of privacy-preserving bayesian data analysis. *arXiv preprint arXiv:1603.07294*, 2016.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Hairer, M., Stuart, A. M., and Vollmer, S. J. Spectral gaps for a metropolis–hastings algorithm in infinite dimensions. *The Annals of Applied Probability*, 24(6):2455–2490, 2014.

Hastings, W. K. Monte carlo sampling methods using markov chains and their applications. 1970.

Heikkilä, M., Jälkö, J., Dikmen, O., and Honkela, A. Differentially private markov chain monte carlo. *Advances in Neural Information Processing Systems*, 32, 2019.

Heller, G. and Fetaya, E. Can stochastic gradient langevin dynamics provide differential privacy for deep learning? *arXiv preprint arXiv:2110.05057*, 2021.

Levin, D. A. and Peres, Y. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Li, B., Chen, C., Liu, H., and Carin, L. On connecting stochastic gradient mcmc and differential privacy. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 557–566. PMLR, 2019.

Li, N., Qardaji, W., and Su, D. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, pp. 32–33, 2012.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Mironov, I. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275. IEEE, 2017.

Neal, R. M. et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Räisä, O., Koskela, A., and Honkela, A. Differentially private hamiltonian monte carlo. *arXiv preprint arXiv:2106.09376*, 2021.

Wang, Y.-X., Fienberg, S., and Smola, A. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pp. 2493–2502. PMLR, 2015.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.

Yıldırım, S. and Ermiş, B. Exact mcmc with differentially private moves. *Statistics and Computing*, 29(5):947–963, 2019.

Zhang, Q., Bu, Z., Chen, K., and Long, Q. Differentially private bayesian neural networks on accuracy, privacy and reliability. *arXiv preprint arXiv:2107.08461*, 2021.

Zhang, R. and De Sa, C. M. Poisson-minibatching for gibbs sampling with convergence rate guarantees. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhang, R., Cooper, A. F., and De Sa, C. M. Asymptotically optimal exact minibatch metropolis-hastings. *Advances in Neural Information Processing Systems*, 33:19500–19510, 2020.

Zhang, Z., Rubinstein, B. I., and Dimitrakakis, C. On the differential privacy of bayesian inference. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

A. Proof of Lemma 1

Proof. Fix any two neighboring datasets D, D' that differ on index i. We denote the energy difference for the minibatch as ℓ_1 and ℓ_1', and the energy difference for the entire data as ℓ_2 and ℓ_2'. By the definition of ℓ_1, we have $\ell_1' = \ell_1 + \Delta$, where

$$\begin{aligned}\Delta =& 2\left(\operatorname{artanh}\left(\frac{C\left(U_i(\theta|x_i') - U_i(\theta'|x_i')\right)}{c_i'(2\lambda + C'M(\theta,\theta'))}\right)\right).\\ &- \operatorname{artanh}\left(\frac{C\left(U_i(\theta|x_i) - U_i(\theta'|x_i)\right)}{c_i(2\lambda + CM(\theta,\theta'))}\right)\right).\end{aligned}$$

Because $\operatorname{artanh} x = \frac{1}{2}\log\frac{1+x}{1-x}$, the maximum $\max_{x_i} \operatorname{artanh}\left(\frac{C\left(U_i(\theta)-U_i(\theta')\right)}{c_i(2\lambda+CM(\theta,\theta'))}\right)$ must be obtained when $U_i(\theta|x_i) > U_i(\theta'|x_i)$, otherwise the log term is negative. Since $\frac{1+x}{1-x}$ is monotonically increasing when $0 \le x < 1$, and by Assumption 1 that $|U_i(\theta) - U_i(\theta')| \le c_i M(\theta,\theta')$, the maximum is reached when $U_i(\theta) - U_i(\theta') = c_i M(\theta,\theta')$. Thus, we have

$$\begin{aligned}&\max_{x_i} \operatorname{artanh}\left(\frac{C\left(U_i(\theta) - U_i(\theta')\right)}{c_i(2\lambda + CM(\theta,\theta'))}\right)\\ &\le \frac{1}{2}\log\left(\frac{2\lambda c_i + 2c_i CM}{2\lambda c_i}\right)\\ &= \frac{1}{2}\log(1 + \frac{CM}{\lambda}),\end{aligned}$$

Similarly, $\min_{x_i} \operatorname{artanh}\left(\frac{C\left(U_i(\theta)-U_i(\theta')\right)}{c_i(2\lambda+CM(\theta,\theta'))}\right)$ is bounded below by $-\frac{1}{2}\log(1+\frac{CM}{\lambda})$. Hence, $\Delta \le \Delta(\ell_1) := 2\log(1+\frac{CM}{\lambda})$.

For ℓ_2, we have

$$\ell_2' = \ell_2 + \Delta,$$

where $\Delta = U_i(\theta|x_i') - U_i(\theta'|x_i') - U_i(\theta|x_i) + U_i(\theta'|x_i)$. By by Assumption 1, we have that $\Delta \le \Delta(\ell_2) := 2\max_i c_i M(\theta,\theta')$.

□

B. Proof of Theorem 2

Proof. **Gaussian mechanism.** For the Gaussian noise that we use, if we choose σ in Algorithm 2 to be $\sqrt{2\log\frac{1.25}{\delta}}/\epsilon$, then each step is (ϵ,δ)-differentially private with respect to the minibatch. The Gaussian noise is also scaled with the sensitivity $\delta(\ell)$ which depends on the current θ and the proposed θ'.

Sampling (with replacement) with non-uniform probabilities. When forming a minibatch, we sample index i with probability $p_i = c_i/C$. Let D and D' be neighboring datasets of size n, and we assume that without loss of generality they differ in the last row. The worst case would be the last row has the highest probability of being sampled into the minibatch. Let $p = \max_i c_i/C$. We note that after the Poisson sampling step, for a given sample i_b, the additional step of determining whether to add i_b to the batch has a privacy amplification effect. Specifically, if the subsequent mechanism is (ϵ,δ)-differentially private, this step will result in (ϵ',δ')-differential privacy, where $\epsilon' = \log(1+\beta(\exp(\epsilon)-1))$ and $\delta' = \beta\delta$, with $\beta = \max_{i_b}\Pr[\text{add } i_b]$. This follows directly from Theorem 1 in (Li et al., 2012). Note that $\epsilon' \le \epsilon$ and $\delta' \le \delta$. We take (ϵ,δ)-differential privacy as a conservative privacy guarantee, since β depends on the data and is upper-bounded by 1.

For a given batch size B, let $\mathcal{M}_1$ denote the Poisson sampling step and let $\mathcal{M}_2$ be the (ϵ,δ)-DP mechanism on a minibatch. Let T be a random variable denoting the multiset of indices sampled by $\mathcal{M}_1$, and let $\ell(T)$ be the multiplicity of index n in T. Fixing a subset of possible log scale of the acceptance probability. For each $k = 0, 1, \ldots, B$, let

$$\begin{aligned}p_k &= \Pr[\ell(T) = k|B] = \binom{B}{k} p^k (1-p)^{B-k}\\ q_k &= \Pr[\mathcal{M}_2(D|_T) \in \mathcal{S}|\ell(T) = k, B]\\ q_k' &= \Pr[\mathcal{M}_2(D'|_T) \in \mathcal{S}|\ell(T) = k, B]\end{aligned}$$

By privacy of $\mathcal{M}_2$, we have $q_k \leq \exp(\epsilon)q_{k-1} + \delta$, so $q_k \leq \exp(k\epsilon)q_0 + \frac{\exp(k\epsilon)-1}{\exp(\epsilon)-1}\delta$. Hence,

$$\begin{aligned}\Pr[\mathcal{M}_1 \circ \mathcal{M}_2(D) \in \mathcal{S}|B] &= \sum_{k=0}^{B} p_k q_k \\ &\leq \sum_{k=0}^{B} \binom{B}{k} p^k (1-p)^{B-k} \left(\exp(k\epsilon)q_0 + \frac{\exp(k\epsilon)-1}{\exp(\epsilon)-1}\delta\right) \\ &= q_0(1-p+pe^{\epsilon})^B + \frac{\delta}{e^{\epsilon}-1}[(1-p+pe^{\epsilon})^B - 1]. \end{aligned} \tag{3}$$

Similarly, we have

$$\Pr[\mathcal{M}_1 \circ \mathcal{M}_2(D') \in \mathcal{S}|B] \geq q_0(1-p+p\exp(-\epsilon))^B - \frac{\delta}{\exp(-\epsilon)-1}[(1-p+p\exp(-\epsilon))^B - 1]. \tag{4}$$

Then we account for the randomness of the batch size B. To simplify the notation, we denote $\chi C^2 M^2(\theta,\theta') + CM(\theta,\theta')$ as λ, we have

$$\begin{aligned}\Pr[\mathcal{M}_1 \circ \mathcal{M}_2(D) \in \mathcal{S}] &\leq \sum_{B=0}^{K} \left(q_0(1-p+pe^{\epsilon})^B + \frac{\delta}{e^{\epsilon}-1}[(1-p+pe^{\epsilon})^B - 1]\right) \frac{\lambda^B \exp(-\lambda)}{B!} \\ &\leq q_0(1-p+pe^{\epsilon})^K + \frac{\delta}{e^{\epsilon}-1}[(1-p+pe^{\epsilon})^K - 1] \end{aligned} \tag{5}$$

Similarly, we also have that

$$\Pr[\mathcal{M}_1 \circ \mathcal{M}_2(D') \in \mathcal{S}] \geq \sum_{k=0}^{K} \left(q_0(1-p+pe^{-\epsilon})^B - \frac{\delta}{e^{-\epsilon}-1}[(1-p+pe^{-\epsilon})^B - 1]\right) \frac{\lambda^B \exp(-\lambda)}{B!} \tag{6}$$

$$\geq q_0(1-p+pe^{-\epsilon})^K \tag{7}$$

Combining inequalities (5) and (6), we obtain

$$\begin{aligned}\Pr[\mathcal{M}_1 \circ \mathcal{M}_2(D) \in \mathcal{S}] \leq (\frac{1-p+p\exp(\epsilon)}{1-p+p\exp(-\epsilon)})^K \\ \Pr[\mathcal{M}_1 \circ \mathcal{M}_2(D') \in \mathcal{S}] \\ + \frac{\delta}{\exp(\epsilon)-1}[(1-p+p\exp(\epsilon))^K - 1].\end{aligned}$$

Then $\mathcal{M}_1 \circ \mathcal{M}_2$ is (ϵ', δ')-differentially private for

$$\epsilon' \leq K \log \frac{1-p+p\exp(\epsilon)}{1-p+p\exp(-\epsilon)}) \leq 6Kp\epsilon,$$

and

$$\delta' \leq 2Kp\delta.$$

Account the possibility of using standard MH. With probability $q = \sum_{k=K+1}^{\infty} \frac{\lambda^k \exp(-\lambda)}{k!}$, our algorithm will invoke the private standard MH algorithm. We denote our algorithm per iteration be $\tilde{\mathcal{M}}$, and the private standard MH be $\mathcal{M}_3$. Then we have

$$\begin{aligned}\Pr[\tilde{\mathcal{M}}(D) \in \mathcal{S}] &= q \Pr[\mathcal{M}_1 \circ \mathcal{M}_2(D) \in \mathcal{S}] + (1-q) \Pr[\mathcal{M}_3(D) \in \mathcal{S}] \\ &\leq q(\exp(\epsilon') \Pr[\mathcal{M}_1 \circ \mathcal{M}_2(D') \in \mathcal{S}] + \delta') + (1-q)(\exp(\epsilon') \Pr[\mathcal{M}_3(D') \in \mathcal{S}] + \delta') \\ &= \exp(\epsilon') \Pr[\tilde{\mathcal{M}}(D') \in \mathcal{S}] + \delta'.\end{aligned}$$

Composition over T iterations. Last, we obtain the total privacy loss using the advanced composition over T iterations. Algorithm 2 is $(\epsilon, T\delta' + \delta)$-differentially private for $\epsilon = \sqrt{2T\log(1/\delta)}\epsilon' + T\epsilon'(e^{\epsilon'} - 1)$. We remark that we can obtain a tighter composition bound by using GDP (Dong et al., 2019).

□

C. Proof of Theorem 3

Proof. We prove that DP-fast MH is π-reversible. Let the transition operator be

$$T_{\text{DP-fast MH}}(\theta, \theta') = q(\theta'|\theta) a_{\text{DP-fast MH}}(\theta, \theta')$$

where $a_{\text{DP-fast MH}}(\theta, \theta')$ denotes the acceptance probability of DP-fast MH. DP-fast MH decides whether to use TunaMH or standard MH based on the minibatch size B and the size upper bound K, and it also decides whether to add Gaussian noise or not based on $\Delta(l_1), \Delta(l_2)$ and ϵ. Specifically, we have

$$\begin{aligned} a_{\text{DP-fast MH}}(\theta, \theta') = \mathbb{I}(B < K) &\left(\mathbb{I}(\Delta(l_1) > \frac{\epsilon C}{6K \max_i c_i}) a_{\text{DP-TunaMH}} + \mathbb{I}(\Delta(l_1) \leq \frac{\epsilon C}{6K \max_i c_i}) a_{\text{TunaMH}}(\theta, \theta') \right) \\ &+ \mathbb{I}(B \geq K) \left(\mathbb{I}(\Delta(l_2) > \epsilon) a_{\text{DP-MH}}(\theta, \theta') + \mathbb{I}(\Delta(l_2) \leq \epsilon) a_{\text{MH}}(\theta, \theta') \right). \end{aligned}$$

where $a_{\text{DP-TunaMH}}, a_{\text{DP-MH}}$ denote the acceptance probability of TunaMH and MH with added Gaussian noise.

Our key observation is that all indicator functions are symmetric in θ and θ' when $M(\theta, \theta')$ is symmetric in θ and θ' . Besides, following the analysis of the penalty algorithm (Ceperley & Dewing, 1999; Yıldırım & Ermiş, 2019), DP-TunaMH and DP-MH are both reversible. Combining all these results, we can show that

$$\pi(\theta) T_{\text{DP-fast MH}}(\theta, \theta') = \pi(\theta) q(\theta'|\theta) a_{\text{DP-fast MH}}(\theta, \theta')$$

is symmetric in θ and θ'. □

D. Proof of Theorem 4

Proof. We consider the acceptance probability ratio

$$\begin{aligned} \frac{a_{\text{DP-fast MH}}(\theta, \theta')}{a_{\text{MH}}(\theta, \theta')} &\geq \min\left(\frac{a_{\text{DP-TunaMH}}(\theta, \theta')}{a_{\text{MH}}(\theta, \theta')}, \frac{a_{\text{DP-MH}}(\theta, \theta')}{a_{\text{MH}}(\theta, \theta')} \right) \\ &= \min\left(\frac{a_{\text{DP-TunaMH}}(\theta, \theta')}{a_{\text{TunaMH}}(\theta, \theta')} \cdot \frac{a_{\text{TunaMH}}(\theta, \theta')}{a_{\text{MH}}(\theta, \theta')}, \frac{a_{\text{DP-MH}}(\theta, \theta')}{a_{\text{MH}}(\theta, \theta')} \right) \\ &\geq \min\left(\left(1 - \Phi\left(\frac{\sigma_1^2 \Delta(l_1)^2}{2}\right)\right) \cdot \exp\left(-\frac{C^2 A^2}{\lambda} - 2\sqrt{\frac{C^2 A^2}{\lambda} \log 2} \right), \left(1 - \Phi\left(\frac{\sigma_1^2 \Delta(l_1)^2}{2}\right)\right) \right), \end{aligned}$$

where the last equation is obtained using the results in Theorem 2 in Zhang et al. (2020) and Proposition 2 in Yıldırım & Ermiş (2019). By the assumption that $M(\theta, \theta') \leq A$, it follows that

$$\Delta(\ell_1) = 2\log\left(1 + \frac{CM(\theta, \theta')}{\lambda}\right) \leq 2\log\left(1 + \frac{CA}{\lambda}\right), \quad \Delta(\ell_2) = 2\max_i c_i M(\theta, \theta') \leq 2A \max_i c_i$$

Since $\sigma_1 = 6K \max_i c_i \sqrt{2\log\frac{2.5K \max_i c_i}{\delta C}}/(\epsilon C)$ and $\sigma_2 = \sqrt{2\log\frac{1.25}{\delta}}/\epsilon$, then

$$\sigma_1^2 \Delta(l_1)^2 \leq \frac{216K^2 \max_i c_i^2 \log\frac{2.5K \max_i c_i}{\delta C} \log^2\left(1 + \frac{CA}{\lambda}\right)}{\epsilon^2 C^2}, \quad \sigma_2^2 \Delta(l_2)^2 \leq \frac{8A^2 \max_i c_i^2 \log\frac{1.25}{\delta}}{\epsilon^2}.$$

The acceptance probability ratio thus has the following bound

$$\begin{aligned} &\frac{a_{\text{DP-fast MH}}(\theta, \theta')}{a_{\text{MH}}(\theta, \theta')} \\ &\geq \min\left(\left(1 - \Phi\left(\frac{32(\log 2)^2 \log\frac{1.25}{\delta}}{\epsilon^2}\right)\right) \cdot \exp\left(-\frac{C^2 A^2}{\lambda} - 2\sqrt{\frac{C^2 A^2}{\lambda} \log 2} \right), \left(1 - \Phi\left(\frac{2C^2 A^2 \log\frac{1.25}{\delta}}{9K^2 \epsilon^2}\right)\right) \right). \end{aligned}$$

Let $\bar{\gamma}$ denote the spectral gap of DP-fast MH, and let γ denote the spectral gap of the standard MH with the same target distribution π and the proposal distribution q. By a Dirichlet form argument, we have

$$\frac{\bar{\gamma}}{\gamma} \geq \left(1 - \Phi\left(\frac{216K^2 \max_i c_i^2 \log \frac{2.5K \max_i c_i}{\delta C} \log^2\left(1 + \frac{CA}{\lambda}\right)}{\epsilon^2 C^2}\right)\right) \cdot \exp\left(-\frac{C^2A^2}{\lambda} - 2\sqrt{\frac{C^2A^2}{\lambda}\log 2}\right).$$

□

E. Experimental Details

E.1. Truncated Gaussian Mixture

The data is generated from a mixture of two Gaussians. In order to get the bounds required by all methods, the Gaussian is truncated by setting $x_i \in [-3, 3]$. We assume a flat prior $p(\theta) = 1$. Then the energy is given by

$$U_i(\theta) = -\log p(x_i|\theta) = \log(2\sqrt{2\pi}\sigma_x) - \log\left[\exp\left(-\frac{(x_i-\theta_1)^2}{2\sigma_x^2}\right) + \exp\left(-\frac{(x_i-\theta_1-\theta_2)^2}{2\sigma_x^2}\right)\right].$$

We can set $M(\theta, \theta') = \|\theta - \theta'\|_2$ and

$$c_i = \sqrt{\left(\frac{2|x_i|+9}{\sigma_x^2}\right)^2 + \left(\frac{|x_i|+6}{\sigma_x^2}\right)^2}.$$

We tune the stepsize of each method to reach the acceptance rate 60% and set K to be around $\epsilon C / \max_i c_i$. The hyperparameters in baseline methods are set following their original papers (Yıldırım & Ermiş, 2019; Heikkilä et al., 2019).

E.1.1. Additional Experimental Results

We additionally show symmetric KL vs. the number of used data points in Figure 4, which does not depend on the implementation details. From this figure, we can clearly see that both minibatch and full-batch DP-fast MH converge significantly faster than previous methods.

E.2. Logistic Regression on MNIST

MNIST with only 7s and 9s images contains 12214 training data and 2037 test data. Let h be the sigmoid function. Let the label $y_i \in \{0, 1\}$, then the model in logistic regression (LR) is

$$p(y_i = 1) = h(\theta^\intercal x_i) = \frac{1}{1 + \exp(-\theta^\intercal x_i)}.$$

It follows that

$$U_i(\theta) = -y_i \log h(\theta^\intercal x_i) - (1 - y_i) \log h(-\theta^\intercal x_i).$$

It is easy to see that

$$|\partial_j U_i| = |(h(\theta^\intercal x_i) - y_i)x_{ij}| \leq 1 \cdot |x_{ij}|.$$

Thus we can set $M(\theta, \theta')$ to be $\|\theta - \theta'\|_2$ and c_i to be $\|x_i\|_2$. We set the target acceptance rate to be 60% and set K to be around $\epsilon C / \max_i c_i$. The hyperparameters in baseline methods are set following their original papers (Yıldırım & Ermiş, 2019; Heikkilä et al., 2019).

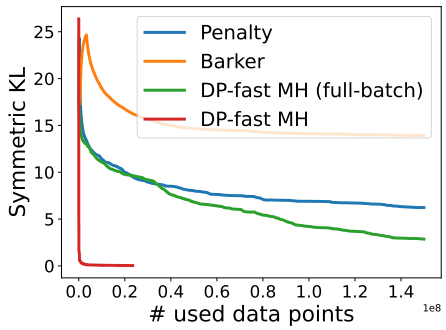


Figure 4. Symmetric KL vs. the number of used data points when privacy $\epsilon = 0.05$.