Master's thesis
Master's Programme in Data Science

Differentially Private Markov Chain Monte Carlo

Ossi Räisä

December 3, 2020

Supervisor(s):	Associate Professor Antti Honkela
Examiner(s):	Associate Professor Antti Honkela Dr. Antti Koskela

University of Helsinki
Faculty of Science

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty

Faculty of Science

Koulutusohjelma — Utbildningsprogram — Degree programme

Master’s Programme in Data Science

Tekijä — Författare — Author

Ossi Räisä

Työn nimi — Arbetets titel — Title

Differentially Private Markov Chain Monte Carlo

Työn laji — Arbetets art — Level

Master’s thesis

Aika — Datum — Month and year

December 3, 2020

Sivumäärä — Sidantal — Number of pages

40

Tiivistelmä — Referat — Abstract

ACM Computing Classification System (CCS):

Avainsanat — Nyckelord — Keywords

Differential Privacy, Markov Chain Monte Carlo

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

1. Introduction

2. Background

2.1 Differential Privacy

Differential privacy [DR14] is a property of an algorithm that quantifies the amount of information about private data an adversary can gain from the publication of the algorithm's output. The most commonly used definition uses two real numbers, ϵ and δ, to quantify the information gain, or, from the perspective of a data subject, the privacy loss of the algorithm.

The most common definition is called (ϵ, δ)-DP, approximate DP or ADP [DR14]. The case where $\delta = 0$ is called ϵ-DP or pure DP.

Definition 1. *An algorithm $\mathcal{M} \colon \mathcal{X} \to \mathcal{U}$ is (ϵ, δ)-ADP if for all neighbouring inputs $x \in \mathcal{X}$ and $x' \in \mathcal{X}$ and all measurable sets $S \subset \mathcal{U}$*

$$P(\mathcal{M}(x) \in S) \leq e^{\epsilon} P(\mathcal{M}(x') \in S) + \delta.$$

The neighbourhood relation in the definition is domain specific. With tabular data the most common definitions are the add/remove neighbourhood and substitute neighbourhood.

Definition 2. *Two tabular datasets are said to be add/remove neighbours if they are equal after adding or removing at most one row to or from one of them. The datasets are said to be in substitute neighbours if they are equal after changing at most one row in one of them.*

The neighbourhood relation is denoted by $\sim$. The definitions and theorems of this section are valid for all neighbourhood relations.

There many other definitions of differential privacy that are mostly used to compute (ϵ, δ)-bounds for ADP. This thesis uses two of them: Rényi-DP (RDP) [Mir17] and zero-concentrated differential privacy (zCDP) [BS16]. Both are based on Rényi divergence [Mir17], which is a particular way of measuring the difference between random variables.

Definition 3. *For random variables with density or probability mass functions P and Q the Rényi divergence of order $1 < \alpha < \infty$ is*

$$D_\alpha(P \mid\mid Q) = \frac{1}{\alpha - 1} \ln E_{x \sim Q} \left(\frac{P(x)^\alpha}{Q(y)^\alpha} \right).$$

Orders $\alpha = 1$ and $\alpha = \infty$ are defined by continuity:

$$D_1(P \mid\mid Q) = \lim_{\alpha \to 1-} D_\alpha(P \mid\mid Q),$$

$$D_\infty(P \mid\mid Q) = \lim_{\alpha \to \infty} D_\alpha(P \mid\mid Q).$$

Both Rényi-DP and zCDP can be expressed as bounds on the Rényi divergence between the outputs of an algorithm with neighbouring inputs:

Definition 4. *An algorithm $\mathcal{M}$ is (α, ϵ)-Rényi DP if for all $x \sim x'$*

$$D_\alpha(\mathcal{M}(x) \mid\mid \mathcal{M}(x')) \leq \epsilon.$$

$\mathcal{M}$ is ρ-zCDP if for all $\alpha > 1$ and all $x \sim x'$

$$D_\alpha(\mathcal{M}(x) \mid\mid \mathcal{M}(x')) \leq \rho\alpha.$$

Rényi-DP and zCDP bounds can be converted to ADP bounds [Mir17, BS16]:

Theorem 1. *If $\mathcal{M}$ is (α, ϵ)-RDP, $\mathcal{M}$ is also $(\epsilon - \frac{\ln \delta}{\alpha - 1}, \delta)$-ADP for any $0 < \delta < 1$. If $\mathcal{M}$ is ρ-zCDP, $\mathcal{M}$ is also $(\rho + \sqrt{-4\rho \ln \delta}, \delta)$-ADP for any $0 < \delta < 1$.*

A very useful property of all of these definitions is composition [DR14]: if algorithms $\mathcal{M}$ and $\mathcal{M}'$ are DP, the algorithm first computing $\mathcal{M}$ and then $\mathcal{M}'$, outputting both results, is also DP, although with worse bounds. More precisely

Definition 5. *Let $\mathcal{M} \colon \mathcal{X} \to \mathcal{U}$ and $\mathcal{M}' \colon \mathcal{X} \times \mathcal{U} \to \mathcal{U}'$ be algorithms. Their composition is the algorithm outputting $(\mathcal{M}(x), \mathcal{M}'(x, \mathcal{M}(x)))$ for input x.*

Theorem 2. *Let $\mathcal{M} \colon \mathcal{X} \to \mathcal{U}$ and $\mathcal{M} \colon \mathcal{X} \times \mathcal{U} \to \mathcal{U}'$ be algorithms. Then*

1. *If $\mathcal{M}$ is (ϵ, δ)-ADP and $\mathcal{M}'$ is (ϵ', δ')-ADP, then their composition is $(\epsilon+\epsilon', \delta+\delta')$-ADP [DR14]*

2. *If $\mathcal{M}$ is (α, ϵ)-RDP and $\mathcal{M}'$ is (α, ϵ')-RDP, then their composition is $(\alpha, \epsilon + \epsilon')$-RDP [Mir17]*

3. *If $\mathcal{M}$ is ρ-zCDP and $\mathcal{M}'$ is ρ'-zCDP, then their composition is $(\rho + \rho')$-zCDP [BS16]*

All of the composition results can be extended to any number of compositions by induction. Note that any step of the composition can depend on the results of the previous steps, not only on the private data. There are also other composition theorems for ADP that trade increased δ for decreased ϵ or vice-versa, but this thesis does not apply them directly.

As any algorithm that does not use private data in any way is $(0, 0)$-ADP, 0-zCDP and $(\alpha, 0)$-RDP with all α, Theorem 2 has the following corollary, called post-processing immunity:

Theorem 3. *Let $\mathcal{M}\colon \mathcal{X} \to \mathcal{U}$ be an ADP, RDP or zCDP algorithm with some privacy parameters. Let $f\colon \mathcal{U} \to \mathcal{U}'$ be any algorithm not using the private data. Then the composition of $\mathcal{M}$ and f is ADP, RDP or zCDP with the same privacy parameters.*

There are many different DP algorithms that are commonly used, which are also called mechanisms [DR14]. This thesis only requires one of the most commonly used ones: the Gaussian mechanism [DR14].

Definition 6. *The Gaussian mechanism with parameter σ^2 is an algorithm that, with input x, outputs a sample from $\mathcal{N}(x, \sigma^2)$, where $\mathcal{N}$ denotes the normal distribution.*

The RDP and zCDP bounds for the Gaussian mechanism are quite simple. The ADP bound is more complicated:

Theorem 4. *If for all inputs x and x', $||x - x'||_2 \leq \Delta$, the Gaussian mechanism is*

1. *$(\alpha, \frac{\alpha\Delta^2}{2\sigma^2})$-RDP [Mir17]*
2. *$\frac{\Delta^2}{2\sigma^2}$-zCDP [BS16]*
3. *n compositions of the Gaussian mechanism are $(\epsilon, \delta(\epsilon))$-ADP [SMM19] with*

$$\delta(\epsilon) = \frac{1}{2}\left(\operatorname{erfc}\left(\frac{\sigma(\epsilon - n\mu)}{\sqrt{2n}\Delta}\right) - e^{\epsilon}\operatorname{erfc}\left(\frac{\sigma(\epsilon + n\mu)}{\sqrt{2n}\Delta}\right)\right),$$

where $\mu = \frac{\Delta^2}{2\sigma^2}$ and erfc is the complementary error function.

The most common use case for the Gaussian mechanism is computing a function $f\colon \mathcal{X} \to \mathbb{R}$ of private data and feeding the result into the Gaussian mechanism to privately release the function value. The condition that the inputs of the Gaussian mechanism cannot vary too much leads into the concept of sensitivity of a function

Definition 7. *The l_p-sensitivity Δ_p, with neighbourhood relation $\sim$, of a function $f\colon \mathcal{X} \to \mathbb{R}^n$ is*

$$\Delta_p f = \sup_{x \sim x'} ||f(x) - f(x')||_p.$$

Theorem 4 implies that the value of any function with finite l_2-sensitivity can be privately released using the Gaussian mechanism with appropriate noise variance σ^2. Of course, the usefulness of the released value depends on the magnitude of σ^2 compared to the actual value.

2.2 Bayesian Inference and Markov Chain Monte Carlo

In Bayesian inference, the parameters of a statistical model are inferred from observed data using Bayes' theorem [GCS+14]. The result is not just a point estimate of the parameters, but a probability distribution describing the likelihood of different values of the parameters.

Bayes' theorem relates the *posterior* belief of the parameters $p(\theta \mid X)$ to the *prior* belief $p(\theta)$ through the observed data X and the likelihood of the data $p(X \mid \theta)$ as follows:

$$p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{\int p(X \mid \theta)p(\theta)d\theta}.$$

It is theoretically possible to compute $p(\theta \mid X)$ given any likelihood, prior and data, but the integral in the denominator is in many cases difficult to compute [GCS+14]. In such cases the posterior cannot be feasibly computed. However, many of the commonly used summary statistics of the posterior, such as the mean, variance and credible intervals, can be approximated from a sample of the posterior. *Markov chain Monte Carlo* (MCMC) is a widely used algorithm to obtain such samples [GCS+14].

MCMC algorithms sequentially sample values of θ with the goal of eventually having the chain of sampled values converge to a given distribution [GCS+14]. While this can be done in many ways, this thesis focuses on a particular MCMC algorithm: *Metropolis-Hastings* (MH).

At each iteration i, the Metropolis-Hastings algorithm samples θ_i from a distribution π of the parameters by first picking a proposal θ' from a proposal distribution $q(\theta_{i-1})$ [GCS+14], where θ_{i-1} is the previously sampled value*. We shorten θ_{i-1} to θ in the following. The ratio of posterior and proposal densities is calculated

$$r(\theta, \theta') = \frac{\pi(\theta')}{\pi(\theta)} \frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)},$$

and the proposal is accepted with probability $\min\{1, r\}$. If the proposal is accepted, $\theta_i = \theta$, otherwise $\theta_i = \theta$.

*The value of θ_0 for the first iteration is given as input to the algorithm.

It can be shown that, with a suitable proposal distribution, the chain of θ_i values converges to π [GCS+14]. The Gaussian distribution centered at the current value is a commonly used proposal.

When MCMC is used in Bayesian inference, the distribution to approximate is

$$\pi(\theta) = p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{\int p(X \mid \theta)p(\theta)d\theta}.$$

The difficult integral $\int p(X \mid \theta)p(\theta)d\theta$ in the denominator cancels out when computing r, so only the likelihood and the prior are needed. For numerical stability, r is usually computed in log space, which makes the acceptance probability $\min\{1, e^{\lambda(\theta,\theta')}\}$ where

$$\lambda(\theta, \theta') = \ln \frac{p(X \mid \theta')}{p(X \mid \theta)} + \ln \frac{p(\theta')}{p(\theta)} + \ln \frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)}.$$

The dataset X is typically a table with n independent rows. The likelihood is given as $p(x_j \mid \theta)$ for row x_j. Independence of the rows means that

$$p(X \mid \theta) = \prod_{j=1}^{n} p(x_j \mid \theta)$$

which means that the log likelihood ratio term of λ is

$$\ln \frac{p(X \mid \theta')}{p(X \mid \theta)} = \sum_{j=1}^{n} \ln \frac{p(x_j \mid \theta')}{p(x_j \mid \theta)}$$

Algorithm 1 puts all of this together to summarise the MH algorithm used for Bayesian inference.

Algorithm 1: Metropolis-Hastings: number of iterations k, proposal distribution q and initial value θ_0 and dataset X as input

```
for $1 \leq i \leq k$ do
    denote $\theta = \theta_{i-1}$
    sample $\theta' \sim q(\theta)$
    $\ln p(X \mid \theta) = \sum_{j=1}^{n}(\ln p(x_j \mid \theta') - \ln p(x_j \mid \theta))$
    $\lambda = \ln p(X \mid \theta) + \ln p(\theta') - \ln p(\theta) + \ln q(\theta \mid \theta') - \ln q(\theta' \mid \theta)$
    $\theta_i = \begin{cases} \theta' & \text{with probability } \min\{1, e^{\lambda}\} \\ \theta & \text{otherwise} \end{cases}$
end
return $(\theta_1, \ldots, \theta_k)$
```

2.3 The Banana Distribution

The banana distribution [TPK14] is a banana-shaped probability distribution that is a challenging target for MCMC algorithms. For this reason it has been used to test MCMC algorithms in the literature [TPK14].

Definition 8. *Let X have a d-variate Gaussian distribution with mean μ and covariance matrix Σ. Let*

$$g(x) = (x_1, x_2 - a(x_1 - m)^2 - b, x_3, \dots, x_d),$$

with $a, b, m \in \mathbb{R}$. The banana distribution with parameters μ, Σ, a, b and m is the distribution of $g(X)$. It is denoted by $\mathrm{Ban}(\mu, \Sigma, a, b, m)$.

In the literature, the banana distribution is simply used as the target to sample from, and is not the posterior in a Bayesian inference problem [TPK14]. To test differentially private MCMC algorithms, the target distribution must be the posterior of some inference problem, as otherwise there is no data to protect with differential privacy. Theorem 5 gives a suitable inference problem for testing DP MCMC algorithms.

Theorem 5. *Let*

$$\begin{aligned}
\theta = (\theta_1, \dots, \theta_d) &\sim \mathrm{Ban}(0, \sigma_0^2 I, a, b, m)\\
X_1 &\sim \mathcal{N}(\theta_1, \sigma_1^2)\\
X_2 &\sim \mathcal{N}(\theta_2 + a(\theta_1 - m)^2 + b, \sigma_2^2)\\
X_3 &\sim \mathcal{N}(\theta_3, \sigma_3^2)\\
&\vdots\\
X_d &\sim \mathcal{N}(\theta_d, \sigma_d^2)
\end{aligned}$$

Given data $x_1, \dots, x_d \in \mathbb{R}^n$ and denoting $\tau_i = \frac{1}{\sigma_i^2}$, the posterior of θ tempered with T is the banana distribution $\mathrm{Ban}(\mu, \Sigma, a, b, m)$ *with*

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ji} \quad i \in \{1, 2\}$$

$$\begin{aligned}
\mu &= \left(\frac{Tn\tau_1\bar{x}_1}{Tn\tau_1 + \tau_0}, \dots, \frac{Tn\tau_d\bar{x}_d}{Tn\tau_d + \tau_0}\right),\\
\Sigma &= \mathrm{diag}\left(\frac{1}{Tn\tau_1 + \tau_0}, \dots, \frac{1}{Tn\tau_d + \tau_0}\right).
\end{aligned}$$

Proof. Because

$$g^{-1}(y) = (y_1, y_2 + a(y_1 - m)^2 + b, y_3 \dots, y_d)$$

and the Jacobian determinant of g^{-1} is 1, for a positive-definite Σ the banana distribution has density proportional to

$$\exp\left(-\frac{1}{2}(g^{-1}(x) - \mu)^T \Sigma^{-1} (g^{-1}(x) - \mu)\right)$$

With $\Sigma = \operatorname{diag}(\sigma_1^2, \sigma_2^2)$ the density is proportional to

$$\exp\left(-\frac{1}{2}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 + a(x_1-m)^2 + b - \mu_2}{\sigma_2}\right)^2 + \sum_{i=3}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right)\right)$$

Denote $u = \theta_2 + a(\theta_1 - m)^2 + b$. The tempered posterior of θ is

$$\begin{aligned}
p(\theta \mid X) &\propto p(X \mid \theta)^T p(\theta) \\
&= p(X_1 \mid \theta_1)^T p(X_2 \mid \theta_1, \theta_2)^T \prod_{i=3}^{d} p(X_i \mid \theta_i)^T p(\theta) \\
&= p(X_1 \mid \theta_1)^T p(X_2 \mid \theta_1, \theta_2)^T \prod_{i=3}^{d} p(X_i \mid \theta_i)^T \\
&\cdot \exp\left(-\frac{1}{2}\left(\tau_0\theta_1^2 + \tau_0(\theta_2 + a(\theta_1 - m)^2 + b)^2 \sum_{i=3}^{d} \tau_0\theta_i^2\right)\right) \\
&= p(X_1 \mid \theta_1)^T p(X_2 \mid \theta_1, \theta_2)^T \exp\left(-\frac{1}{2}\left(\tau_0\theta_1^2 + \tau_0(\theta_2 + a(\theta_1 - m)^2 + b)^2\right)\right) \\
&\cdot \prod_{i=3}^{d} p(X_i \mid \theta_i)^T \exp\left(-\frac{1}{2}\sum_{i=3}^{d} \tau_0\theta_i^2\right)
\end{aligned}$$

Considering the upper and lower part of the last expression separately

$$
\begin{aligned}
&p(X_1 \mid \theta_1)^T p(X_2 \mid \theta_1, \theta_2)^T \exp\left(-\frac{1}{2}\left(\tau_0\theta_1^2 + \tau_0(\theta_2 + a(\theta_1 - m)^2 + b)^2\right)\right) \\
&\propto \left(\prod_{i=1}^{n} \exp\left(-\frac{(x_{i1} - \theta_1)^2\tau_1}{2}\right)\right)^T \cdot \left(\prod_{i=1}^{n} \exp\left(-\frac{(x_{i2} - \theta_2 - a(\theta_1 - m)^2 - b)^2\tau_2}{2}\right)\right)^T \\
&\cdot \exp\left(-\frac{1}{2}\left(\tau_0\theta_1^2 + \tau_0(\theta_2 + a(\theta_1 - m)^2 + b)^2\right)\right) \\
&= \exp\left(-\frac{1}{2}\Big(T\tau_1\sum_{i=1}^{n}(x_{i1} - \theta_1)^2 + T\tau_2\sum_{i=1}^{n}(x_{i2} - u)^2 + \tau_0\theta_1^2 + \tau_0 u^2\Big)\right) \\
&= \exp\Bigg(-\frac{1}{2}\Big(T\tau_1\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 + T\tau_1 n(\bar{x}_1 - \theta_1)^2 \\
&+ T\tau_2\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2 + T\tau_2 n(\bar{x}_2 - u)^2 + \tau_0\theta_1^2 + \tau_0 u^2\Big)\Bigg) \\
&\propto \exp\left(-\frac{1}{2}\Big(T\tau_1 n(\bar{x}_1 - \theta_1)^2 + T\tau_2 n(\bar{x}_2 - u)^2 + \tau_0\theta_1^2 + \tau_0 u^2\Big)\right) \\
&= \exp\Bigg(-\frac{1}{2}\Big(T\tau_1 n\bar{x}_1^2 - 2T\tau_1 n\bar{x}_1\theta_1 + nT\tau_1\theta_1^2 + \tau_0\theta_1^2 \\
&+ T\tau_2 n\bar{x}_2^2 - 2T\tau_2 n\bar{x}_2 u + nT\tau_2 u^2 + \tau_0 u^2\Big)\Bigg) \\
&\propto \exp\left(-\frac{1}{2}\Big((Tn\tau_1 + \tau_0)\theta_1^2 - 2T\tau_1 n\bar{x}_1\theta_1 + (Tn\tau_2 + \tau_0)u^2 - 2T\tau_2 n\bar{x}_2 u\Big)\right) \\
&= \exp\left(-\frac{1}{2}\Big((Tn\tau_1 + \tau_0)\left(\theta_1^2 - \frac{2T\tau_1 n\bar{x}_1\theta_1}{Tn\tau_1 + \tau_0}\right) + (Tn\tau_2 + \tau_0)\left(u^2 - \frac{2T\tau_2 n\bar{x}_2 u}{Tn\tau_2 + \tau_0}\right)\Big)\right) \\
&\propto \exp\left(-\frac{1}{2}\Big((Tn\tau_1 + \tau_0)\left(\theta_1 - \frac{T\tau_1 n\bar{x}_1}{Tn\tau_1 + \tau_0}\right)^2 + (Tn\tau_2 + \tau_0)\left(u - \frac{T\tau_2 n\bar{x}_2}{Tn\tau_2 + \tau_0}\right)^2\Big)\right)
\end{aligned}
$$

and

$$
\begin{aligned}
&\prod_{i=3}^{d} p(X_i \mid \theta_i)^T \cdot \exp\left(-\frac{1}{2}\sum_{i=3}^{d}\tau_0\theta_i^2\right) \\
&\propto \exp\left(-\frac{1}{2}T\sum_{j=3}^{d}\tau_j\sum_{i=1}^{n}(x_{ij}-\theta_j)^2 - \frac{1}{2}\sum_{j=3}^{d}\tau_0\theta_j^2\right) \\
&= \exp\left(-\frac{1}{2}\sum_{j=3}^{d}\left(T\tau_j\sum_{i=1}^{n}(x_{ij}-\theta_j)^2 + \tau_0\theta_j^2\right)\right) \\
&= \exp\left(-\frac{1}{2}\sum_{j=3}^{d}\left(T\tau_j\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2 + T\tau_j n(\bar{x}_j-\theta_j)^2 + \tau_0\theta_j^2\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\sum_{j=3}^{d}\left(T\tau_j n(\bar{x}_j-\theta_j)^2 + \tau_0\theta_j^2\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\sum_{j=3}^{d}\left(-2T\tau_j n\bar{x}_j\theta_j + T\tau_j n\theta_j^2 + \tau_0\theta_j^2\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\sum_{j=3}^{d}\left((Tn\tau_j+\tau_0)\theta_j^2 - 2Tn\tau_j\bar{x}_j\theta_j + \frac{(Tn\tau_j\bar{x}_j)^2}{Tn\tau_j+\tau_0}\right)\right) \\
&= \exp\left(-\frac{1}{2}\sum_{j=3}^{d}(Tn\tau_j+\tau_0)\left(\theta_j - \frac{Tn\tau_1\bar{x}_i}{Tn\tau_j+\tau_0}\right)^2\right)
\end{aligned}
$$

Multiplying the resulting expression above gives a density proportional to the banana distribution. As $p(\theta \mid X)$ is proportional to the density of a banana distribution, the posterior is the banana distribution $\mathrm{Ban}(\mu, \Sigma, a, b, m)$ with

$$
\begin{aligned}
\mu &= \left(\frac{Tn\tau_1\bar{x}_1}{Tn\tau_1+\tau_0}, \dots, \frac{Tn\tau_d\bar{x}_d}{Tn\tau_d+\tau_0}\right), \\
\Sigma &= \mathrm{diag}\left(\frac{1}{Tn\tau_1+\tau_0}, \dots, \frac{1}{Tn\tau_d+\tau_0}\right).
\end{aligned}
$$

□

3. Differentially Private MCMC

As seen in Section 2.1, an algorithm can be made differentially private by adding Gaussian noise the its output. The noise could also be added to any intermediate value calculated by the algorithm, and post processing immunity will guarantee that the same DP bounds that hold for releasing the intermediate value also hold for releasing the final result of the algorithm.

In 2019, Yildirim and Ermis [YE19] realised that if Gaussian noise is added to the exact value of λ, the noise can be corrected for yielding a differentially private MCMC algorithm which converges to the correct distribution. In the same year, Heikkilä et. al. [HJDH19] developed another DP MCMC algorithm, called DP Barker, which uses subsampling to amplify privacy.

3.1 DP Penalty

In 1999, Ceperley and Dewing [CD99] developed a variant of Metropolis-Hastings called the penalty algorithm, where only a noisy approximation of λ is known. They developed the algorithm for simulations in physics where computing λ requires computing energies of complex systems, which can only be approximated. The penalty algorithm modifies the acceptance probability to account for the noise added to λ and still converges to the correct distribution if the noise is Gaussian with known variance.

The DP penalty algorithm adds Gaussian noise to the value of λ, and uses the penalty algorithm to correct the acceptance probability so that the algorithm still converges to the correct distribution [YE19]. The corrected acceptance probability for Gaussian noise with variance σ^2 is

$$\min\{1, e^{\lambda(\theta,\theta')-\frac{1}{2}\sigma^2}\}$$

Theorem 6 gives the number of iterations DP penalty can be run for when the privacy cost is computed through zCDP, which is what Yildirim and Ermis prove in their paper [YE19]. A tighter, but harder to use, bound can be reached without using zCDP. This is given by Theorem 7.

Theorem 6. *Let $\epsilon > 0$, $0 < \delta < 1$, $\alpha > 0$ and $\tau > 0$. Let*

$$\rho = (\sqrt{\epsilon - \ln\delta} - \sqrt{-\ln\delta})^2$$

$$c(\theta, \theta') = \sup_{x_j, x_j'} (p(x_j \mid \theta') - p(x_j \mid \theta) - (p(x_j' \mid \theta') - p(x_j' \mid \theta)))$$

$$\sigma^2(\theta, \theta') = \tau^2 n^{2\alpha} c^2(\theta, \theta')$$

Then DP penalty can be run for

$$k = \lfloor 2\tau^2 n^{2\alpha}\rho \rfloor$$

iterations when using σ^2 as the variance of the Gaussian noise.

Theorem 7. *Let $\epsilon > 0$ and $\tau > 0$. Define c and σ as in Theorem 6. The DP penalty algorithm, after running for k iterations using σ as the noise variance, is $(\epsilon, \delta(\epsilon))$-DP for*

$$\delta(\epsilon) = \frac{1}{2}\left(\operatorname{erfc}\left(\frac{\tau n^\alpha(\epsilon - k\mu)}{\sqrt{2k}}\right) - e^\epsilon \operatorname{erfc}\left(\frac{\tau n^\alpha(\epsilon + k\mu)}{\sqrt{2k}}\right)\right)$$

where $\mu = \frac{1}{\tau^2 n^{2\alpha}}$.

Proof. DP penalty is an adaptive composition of Gaussian mechanisms that release noisy values of $\lambda(\theta, \theta')$. The sensitivity of $\lambda(\theta, \theta')$ is $c(\theta, \theta')$. For the tight ADP bound used here, the sensitivity must be constant in each iteration. This is achieved by releasing $\frac{\lambda(\theta,\theta')}{c(\theta,\theta')}$ instead, which has sensitivity 1. $c(\theta, \theta')$ does not depend on X, so $\lambda(\theta, \theta')$ can be obtained from $\frac{\lambda(\theta,\theta')}{c(\theta,\theta')}$ by post processing.

Adding Gaussian noise with variance σ_n^2 to $\frac{\lambda(\theta,\theta')}{c(\theta,\theta')}$ is equivalent to adding Gaussian noise with variance $\sigma_n^2 c^2(\theta, \theta')$ to $\lambda(\theta, \theta')$. Setting $\sigma_n^2 = \tau^2 n^{2\alpha}$ and plugging into the ADP bound of Theorem 4 proves the claim. □

Theorem 7 is harder to use than Theorem 6 because the number of iteration DP penalty can be run for given an (ϵ, δ)-bound cannot be computed analytically for the former. However, the maximum number of iterations can be solved for numerically.

Theorems 6 and 7 require a bound on sensitivity of the log likelihood ratio. If there is a bound

$$|\ln p(x_j \mid \theta') - \ln p(x_j \mid \theta)| \leq L||\theta - \theta'||_2$$

for all D_j, θ and θ' then

$$c(\theta, \theta') \leq 2L||\theta - \theta'||_2$$

The former bound is true in some model, such as logistic regression. In other models it can be forced by clipping the log likelihood ratios to the interval $[-L||\theta - \theta'||_2, L||\theta - \theta'||_2]$. This will remove the guarantee of eventually converging to the correct posterior, but if L is chosen to be large enough, the clipping will not affect the acceptance decision frequently. As a tradeoff, picking a large L will increase the variance of the Gaussian noise and slow down convergence through it.

Yildirim and Ermis [YE19] propose two potential ways to improve the performance of the penalty algorithm. The first improvement is only proposing changes in one dimension in a multidimensional problem. This decreases $||\theta - \theta'||_2$, which means that it decreases the noise variance.

The second improvement is called *guided random walk* (GRW) [YE19]. In GRW, proposals change only one dimension, as above. Additionally, a direction is associated with each dimension, and proposals are only made the current direction of the chosen dimension. After an accepted proposal, the direction is kept the same, but after a reject it is switched. This means that the chain can move towards areas of higher probability faster because, after some initial proposals are rejected, the directions for each dimension point towards the area of high probability, so all proposals are towards it. Without GRW, most proposals would move the chain away from the area of high probability, and would likely be rejected.

3.2 DP Barker

The DP Barker algorithm of Heikkilä et. al. [HJDH19] is based on the Barker acceptance test [Bar65] instead of the Metropolis-Hastings test. Instead of using the MH acceptance probability, the Barker acceptance test samples $V_{log} \sim \text{Logistic}(0, 1)$ and accepts if

$$\lambda + V_{log} > 0$$

If Gaussian noise with variance σ^2 is added to λ, there exists a correction distribution V_{corr} such that $\mathcal{N}(0, \sigma^2) + V_{corr}$ has the same distribution as V_{log}. Because the variance of V_{log} is $\frac{\pi^2}{3}$ [HJDH19], the variance of V_{corr} must be $\frac{\pi^2}{3} - \sigma^2$ which means that there is an upper bound to the noise variance: $\sigma^2 < \frac{\pi^2}{3}$. Testing whether $\lambda + \mathcal{N}(0, \sigma^2) + V_{corr} > 0$ is equivalent to testing whether $\lambda + V_{log} > 0$, which means that it is possible to derive a DP MCMC algorithm based on the Barker acceptance test if the correction distribution can be sampled from.

However, the analytical form of V_{corr} is not known [HJDH19]. Heikkilä et. al. approximate the distribution with a Gaussian mixture model. This means that their algorithm only converges to an approximately correct distribution, but the approximation error can be made very small.

If the sum in λ was only computed over a subset of the data, the algorithm would take less computation to run, and would be less sensitive to changes in the data.

The latter property is called *subsampling amplification* of differential privacy [WBK19]. Using the λ computed with subsampling instead of the full data λ introduces an additional error that must be corrected for to have the algorithm converge to the correct distribution.

The *central limit theorem* (CLT) states that the distribution of a sum of random variables approaches a Gaussian distribution as more random variables are summed, if some conditions on the independence and variance of the random variables are met [HJDH19]. With the CLT, it can be argued that the error from using the subsampled λ instead of the full data λ has an approximately Gaussian distribution, if the subsample is large enough [HJDH19].

The variance of the error from subsampling can be estimated by the sample variance of the individual terms in the sum in λ [HJDH19]. This allows combining the errors from subsampling and the Gaussian noise from the Gaussian mechanism to a single Gaussian noise value. The V_{corr} distribution can then be used to approximate the Barker acceptance test as above. See algorithm 2 for the DP Barker algorithm. [*]

Heikkilä et. al. [HJDH19] do not directly bound the sensitivity of λ as is done in DP penalty, because the sample variance also depends on input data. Instead they directly bound the Rényi divergence between $\mathcal{N}(0, \sigma^2 - \sigma_b^2)$, where σ_b^2 is the batch sample variance, for two adjacent inputs. Subsampling amplification is accounted for with an amplification theorem for Rényi DP [WBK19].

Theorem 8. *If*

$$|\ln p(x_j \mid \theta') - \ln p(x_j \mid \theta)| \leq \frac{\sqrt{|B|}}{n}$$

$$\alpha < \frac{|B|}{5}, \alpha \in \mathbb{N}$$

for all $\theta, \theta' \in \Theta$, all X and $1 \leq j \leq n$, running k iterations of DP Barker is $(\alpha, k\epsilon(\alpha))$-RDP, with

$$\epsilon(\alpha) = \frac{1}{\alpha - 1} \ln \left(1 + q^2 \binom{\alpha}{2} \min\{4(e^{\epsilon'(2)} - 1), 2e^{\epsilon'(2)}\} + 2 \sum_{j=3}^{\alpha} q^j \binom{\alpha}{j} e^{(j-1)\epsilon'(j)} \right)$$

and

$$\epsilon'(\alpha) = \frac{5}{2|B|} + \frac{1}{2(\alpha - 1)} \ln \frac{2|B|}{|B| - 5\alpha} + \frac{2\alpha}{|B| - 5\alpha}$$

where n is the number of rows in D, $|B|$ is the size of the minibatch and $q = \frac{|B|}{n}$.

Like DP penalty, DP Barker requires a bound on the log likelihood ratio for one row of data. The bound can be forced through clipping if the model does not meet it, but because of the n in the denominator of the bound, it can get very tight for large

[*] See [HJDH19] for the sampling procedure of V_{corr}.

values of n. As a result, clipping may be needed for almost all log likelihood ratios, which may cause the algorithm to converge to a very different distribution from the posterior.

To alleviate the tight bound on log likelihood sensitivity, DP Barker is best used with a tempered likelihood [HJDH19]. In tempering, the log likelihood is multiplied by a number $T = \frac{n_0}{n} < 1$. This increases the variance of the resulting posterior and may lower modeling error in some cases [HJDH19].

Using the tempered likelihood, the log likelihood bound becomes

$$T|\ln p(x_j \mid \theta') - \ln p(x_j \mid \theta)| \leq \frac{\sqrt{|B|}}{n}$$

which is equivalent to

$$|\ln p(x_j \mid \theta') - \ln p(x_j \mid \theta)| \leq \frac{\sqrt{|B|}}{n_0}$$

Typically $n_0 \ll n$ for large datasets, so using a tempered likelihood requires significantly less clipping than a nontempered likelihood.

Algorithm 2: DP Barker

```
for $1 \leq i \leq k$ do
    denote $\theta = \theta$
    sample $\theta' \sim q(\theta)$
    sample $B \subset \{1, \ldots, n\}$
    for $j \in B$ do
        $r_j = \ln p(\theta' \mid x_j) - \ln p(\theta \mid x_j)$
    end
    $\sigma_b^2 = \mathrm{Var}\{r_j \mid j \in B\}$
    $\lambda = \frac{n}{|B|} \sum_{j \in B} r_j + \ln \frac{p(\theta')}{p(\theta)} + \ln \frac{q(\theta|\theta')}{q(\theta'|\theta)}$
    sample $s \sim \mathcal{N}(0, \sigma^2 - \sigma_b^2)$
    sample $c \sim V_{corr}^{\sigma^2}$
    $\theta_i = \begin{cases} \theta' & \text{if } \lambda + s + c > 0 \\ \theta & \text{otherwise} \end{cases}$
end
return $(\theta_1, \ldots, \theta_k)$
```

3.3 Comparing DP Penalty and DP Barker

4. Variations of the Penalty Algorithm

4.1 The Penalty Algorithm with Subsampling

In the DP Barker algorithm, the log likelihood ratio is computed using only a subsample of the dataset to amplify privacy. Subsampling can also be used with the penalty algorithm in the same way, if the acceptance test is corrected for the subsampling.

As with DP Barker, the error from subsampling is approximately normally distributed by the central limit theorem. The variance of the subsampling error can be estimated from the sample variance of individual terms of the sum in the log likelihood ratio. This means that the penalty method can be used to correct for the subsampling error.

The acceptance probability with subsampling is

$$\min\{1, e^{\lambda^*(\theta,\theta') - \frac{1}{2}(\sigma^2+\sigma_b^2)}\},$$

where

$$\lambda^*(\theta,\theta') = \frac{nT}{|B|}\sum_{j\in B} \ln\frac{p(x_j \mid \theta')}{p(x_j \mid \theta)} + \ln\frac{p(\theta')q(\theta \mid \theta')}{p(\theta)q(\theta' \mid \theta)},$$

and σ_b^2 is the sample variance of the log likelihood ratios in batch B. Denote

$$r_j = \ln\frac{p(x_j \mid \theta')}{p(x_j \mid \theta)},$$

$$R = \sum_{x\in B} r_j.$$

Then σ_b^2 can be estimated from the sample variance of r_j:

$$\begin{aligned}\sigma_b^2 &= \mathrm{Var}\left(\frac{nT}{|B|}\sum_{j\in B} r_j\right) = \frac{nT^2}{|B|^2}\sum_{j\in B}\mathrm{Var}(r_j) = \frac{nT^2}{|B|}\mathrm{Var}(r_j)\\ &\approx \frac{(nT)^2}{|B|^2}\sum_{j\in B}\left(r_j - \frac{R}{|B|}\right)^2 = \frac{(nT)^2}{|B|^2}\left(\sum_{j\in B} r_j^2 - \frac{R^2}{|B|}\right).\end{aligned}$$

Because σ_b^2 depends on the data, releasing λ privately is not enough, $\lambda - \frac{1}{2}\sigma_b^2$ must be released privately. This means that using subsampling requires adding additional noise to account for the sensitivity of $\frac{1}{2}\sigma_b^2$.

The sensitivity of $\lambda - \frac{1}{2}\sigma_b^2$ is

$$\Delta\lambda + \frac{1}{2}\Delta\sigma_b^2.$$

With the bound $r_j \leq L||\theta - \theta'||_2$ used in DP penalty, the bound sensitivity of λ is the same as without subsampling. The sensitivity of σ_b^2 must be bounded separately.

Lemma 1. *The sensitivity of $\frac{1}{2}\sigma_b^2$, with $r_j \leq L||\theta - \theta'||_2$, has upper bound*

$$\frac{1}{2}\Delta\sigma_b^2 \leq \left(\frac{nT}{b}\right)^2 \left|1 - \frac{1}{b}\right| L^2||\theta - \theta'||_2^2 + \frac{2(b-1)}{b}\left(\frac{nT}{b}\right)^2 L^2||\theta - \theta'||_2^2.$$

Proof. For datasets $X \sim X'$, that only differ in one element, denote the common part they have by X^*, and the differing element by $x \in X$ and $x' \in X'$

$$\begin{aligned}
\Delta\sigma_b^2 &= \sup_{D \sim D'} |\sigma_b^2(X) - \sigma_b^2(X')| \\
&= \left(\frac{nT}{b}\right)^2 \sup_{X \sim X'} \left| \sum_{x \in X} r^2(x) - \sum_{x \in X'} r^2(x) + \frac{1}{b}R^2(X') - \frac{1}{b}R^2(X) \right| \\
&= \left(\frac{nT}{b}\right)^2 \sup_{x, x', X^*} \left| r^2(x) - r^2(x') + \frac{1}{b}(R(X^*) + r(x'))^2 - \frac{1}{b}(R(X^*) + r(x))^2 \right| \\
&= \left(\frac{nT}{b}\right)^2 \sup_{x, x', X^*} \Big| r^2(x) - r^2(x') + \frac{1}{b}(R^2(X^*) + 2R(X^*)r(x') + r^2(x')) \\
&- \frac{1}{b}(R^2(X^*) + 2R(X^*)r(x) + r^2(x)) \Big| \\
&= \left(\frac{nT}{b}\right)^2 \sup_{x, x', X^*} \left| \left(1 - \frac{1}{b}\right)(r^2(x) - r^2(x')) + \frac{2}{b}D(X^*)(r(x') - r(x)) \right| \\
&\leq \left(\frac{nT}{b}\right)^2 \left|1 - \frac{1}{b}\right| \sup_{x, x'} \left| (r^2(x) - r^2(x')) \right| + \frac{2}{b}\left(\frac{nT}{b}\right)^2 \sup_{x, x', X^*} \left| R(X^*)(r(x') - r(x)) \right| \\
&= \left(\frac{nT}{b}\right)^2 \left|1 - \frac{1}{b}\right| \sup_{x, x'} \left| (r^2(x) - r^2(x')) \right| + \frac{2}{b}\left(\frac{nT}{b}\right)^2 \sup_{x, x'} |r(x') - r(x)| \sup_{X^*} |R(X^*)| \\
&\leq \left(\frac{nT}{b}\right)^2 \left|1 - \frac{1}{b}\right| \sup_{x, x'} \left| (r^2(x) - r^2(x')) \right| + \frac{2}{b}\left(\frac{nT}{b}\right)^2 \sup_{x, x'} |r(x') - r(x)| (b-1) \sup_{d} |r(x)|.
\end{aligned}$$

Plugging the bound $\sup_x |r(x)| \leq L||\theta - \theta'||_2$ into the last expression proves the claim. □

Theorem 9. *Let*

$$\Delta_\lambda = \frac{2nTL}{|B|}||\theta - \theta'||_2,$$

$$\Delta_\sigma = \left(\frac{nT}{b}\right)^2 \left|1 - \frac{1}{b}\right| L^2 ||\theta - \theta'||_2^2, + \frac{2(b-1)}{b} \left(\frac{nT}{b}\right)^2 L^2 ||\theta - \theta'||_2^2.$$
$$c(\theta, \theta') = \Delta_\lambda + \Delta_\sigma,$$
$$\sigma^2(\theta, \theta') = \tau c^2(\theta, \theta').$$

Then running DP penalty with subsampling for k iterations is $(\alpha, k\epsilon(\alpha))$-RDP, with

$$\epsilon(\alpha) = \frac{1}{\alpha - 1} \ln \left(1 + q^2 \binom{\alpha}{2} \min\{4(e^{\epsilon'(2)} - 1), 2e^{\epsilon'(2)}\} + 2\sum_{j=3}^{\alpha} q^j \binom{\alpha}{j} e^{(j-1)\epsilon'(j)}\right),$$

and

$$\epsilon'(\alpha) = \frac{\alpha}{2\tau},$$

where n is the number of rows in D, $|B|$ is the size of the minibatch and $q = \frac{|B|}{n}$.

Proof. By Lemma 1, $\Delta_\sigma(\theta, \theta')$ an upper bound to the sensitivity of $\frac{1}{2}\sigma_b^2$, therefore $c(\theta, \theta')$ is an upper bound to the sensitivity of $\lambda - \frac{1}{2}\sigma_b^2$.

This means that a Gaussian mechanism taking a subsample B of the data as input and uses $\sigma(\theta, \theta')$ as the noise variance is $(\alpha, \epsilon'(\alpha))$-RDP with

$$\epsilon'(\alpha) = \frac{\alpha}{2\tau}.$$

By the subsampling amplification theorem [WBK19, Theorem 9] and the composition theorem of RDP (Theorem 2), the combination of subsampling and Gaussian mechanism is $(\alpha, k\epsilon(\alpha))$-RDP with

$$\epsilon(\alpha) = \frac{1}{\alpha - 1} \ln \left(1 + q^2 \binom{\alpha}{2} \min\{4(e^{\epsilon'(2)} - 1), 2e^{\epsilon'(2)}\} + 2\sum_{j=3}^{\alpha} q^j \binom{\alpha}{j} e^{(j-1)\epsilon'(j)}\right)$$

when run for k iterations for integer $\alpha \geq 2$. □

The appearance of n as a multiplier of Δ_λ and n^2 as a multiplier of Δ_σ causes the noise variance to increase with n, without a corresponding decrease in ϵ'. This makes subsampled DP penalty unsuitable for problems with large datasets, unless tempering is used. Like with DP Barker, using tempering $T = \frac{n_0}{n}$ cancels n out of the noise variance.

4.2 DP Metropolis-Adjusted Langevin Algorithm

Metropolis adjusted Langevin algorithm (MALA) [RR98] is Metropolis-Hastings with the proposal distribution

$$q(\theta' \mid \theta, X) = \mathcal{N}\left(\theta + \frac{1}{2}\sigma_p^2 \nabla \ln p(X \mid \theta), \sigma_p^2 I\right)$$

The gradient in the proposal is meant to guide the random walk to regions of high probability. Because of it, MALA can achieve a higher acceptance rate than standards random walk MH, and requires fewer iterations to converge [RR98].

MALA can be made DP with using the penalty algorithm. Because MALA is just the MH algorithm with a particular proposal, the penalty algorithm using the MALA proposal will converge to the correct distribution. Proving any DP bound for MALA requires computing the sensitivities of the different data-dependent values that the algorithm computes.

Like random walk MH, the acceptance probability for non-DP MALA is

$$p = \min\left\{1, \exp(\lambda(X, \theta, \theta'))\right\},$$

with

$$\lambda(X, \theta, \theta') = \sum_{x \in X} \ln \frac{p(x \mid \theta')}{p(x \mid \theta)} + \ln \frac{q(\theta \mid \theta', X)}{q(\theta' \mid \theta, X)} + \ln \frac{p(\theta')}{p(\theta)}.$$

For DP MALA, the penalty correction is applied, and the acceptance probability is

$$p = \min\left\{\exp\left(\lambda^*(X, \theta, \theta') - \frac{1}{2}\sigma^2(\theta, \theta')\right)\right\}$$

where $\lambda^*(X, \theta, \theta') \sim \mathcal{N}(\lambda(X, \theta, \theta'), \sigma^2(\theta, \theta'))$ and σ^2 is symmetric in θ and θ'.

Algorithm 3: DP MALA

```
for $1 \leq i \leq k$ do
    sample $\theta'$ from $\mathcal{M}_g(\theta_{i-1} + \frac{1}{2}\sigma_p^2 \nabla \ln p(X \mid \theta_{i-1}))$
    sample $\lambda^*$ from $\mathcal{M}_l(\lambda)$
    $p = \min\left\{1, \exp\left(\lambda^* - \frac{1}{2}\sigma^2\right)\right\}$
    $\theta_i = \begin{cases} \theta^* & \text{with probability } p \\ \theta_{i-1} & \text{otherwise} \end{cases}$
end
```

On each iteration, the proposal θ' is sampled from a Gaussian mechanism $\mathcal{M}_g(\theta + \frac{1}{2}\sigma_p^2 \nabla \ln p(X \mid \theta))$ with variance $a\sigma_p^2$. Then $\lambda^*(X, \theta, \theta')$ is sampled from a Gaussian mechanism $\mathcal{M}_l(\lambda(X, \theta, \theta'))$ with variance σ^2. The reason for using variance $a\sigma_p^2$ for the proposal instead of σ_p^2 that non-DP MALA uses is to allow better adjustment of the privacy cost of the proposal.

Lemma 2. *$\mathcal{M}_g$ is ρ_g-zCDP with*

$$\rho_g(\theta) = \frac{\sigma_p^2 c_g^2}{8a}$$

where

$$c_g \geq \Delta_2 \nabla \ln p(X \mid \theta)$$

Proof. The sensitivity of

$$\frac{1}{2}\sigma_p^2 \nabla \ln p(X \mid \theta)$$

is

$$\frac{1}{2}\sigma_p^2 \nabla \Delta_2 \nabla \ln p(X \mid \theta)$$

The zCDP bound follows from Theorem 4. □

c_g cannot depend on θ as the privacy cost of each iteration in a composition must be determinable ahead of time to apply Theorem 2.

Lemma 3. *$\mathcal{M}_l$ is ρ_l-zCDP with*

$$\rho_l = \frac{1}{2\tau},$$

when

$$c_l(\theta, \theta') = \Delta_2^2 \lambda(X, \theta, \theta'),$$

and

$$\sigma^2(\theta, \theta') = \tau c_l^2(\theta, \theta').$$

Proof. By Theorem 4, $\mathcal{M}_l$ is ρ-zCDP with

$$\rho_l = \frac{c_l^2(\theta, \theta')}{2\sigma^2(\theta, \theta')}.$$

The claim follows from plugging in $\sigma(\theta, \theta')$. □

Theorem 10. *With given ADP bounds ϵ and δ, DP MALA can be run for*

$$k = \left\lfloor \frac{\rho}{\rho_g + \rho_l} \right\rfloor$$

iterations, where ρ_g is from Lemma 2, ρ_l is from Lemma 3, and

$$\rho = \left(\sqrt{\log \frac{1}{\delta} + \epsilon} - \sqrt{\log \frac{1}{\delta}} \right)^2$$

Proof. A single iteration is $(\rho_l + \rho_g)$-zCDP. The claim then follows from Theorems 1 and 2. □

The zCDP bounds of DP MALA can be computed with Theorem 10 if the sensitivities c_g and c_l are known. As in DP penalty, they are specific the likelihood, but they can be forced with clipping. Clipping the gradients does not affect the distribution the algorithm converges to, but can diminish the amount of utility the algorithm can get from the gradients. Clipping the likelihood does change the invariant distribution, but if the clipping is limited the algorithm should remain accurate.

Clipping the gradients of each individual iteration results in

$$||\nabla \log p(X \mid \theta) - \nabla \log p(X' \mid \theta)||_2 = ||\nabla \log p(x \mid \theta) - \nabla \log p(x' \mid \theta)||_2 \leq 2L_g$$

so

$$c_g(\theta) = 2L_g$$

and

$$\rho_g = \frac{\sigma_p^2 L_g^2}{2a}.$$

c_l consists of two term that depend on the data, the log likelihood term

$$\sum_{i=1}^{n} \ln \frac{p(x_i \mid \theta')}{p(x_i \mid \theta)}$$

and the log proposal ratio

$$\ln \frac{q(\theta \mid \theta', X)}{q(\theta' \mid \theta, X)}$$

The log likelihood term is the same that appears in DP penalty, and can be bound in the same way. If it is clipped such that $\left|\log \frac{p(x|\theta')}{p(x|\theta)}\right| \leq L||\theta - \theta'||_2$ then

$$\Delta \left(\sum_{x_i n X} \log \frac{p(x \mid \theta')}{p(x \mid \theta)} \right) \leq 2L||\theta - \theta'||_2.$$

The log proposal term does not appear in DP penalty, because it uses a symmetric proposal where the term is always 0. In DP MALA, the proposal is not symmetric, and depends on the data through $\nabla \ln p(X \mid \theta)$, so the term must be included in the sensitivity calculation.

Let d be the dimensionality of θ. Then

$$q(\theta \mid \theta', X) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi a \sigma_p^2}} \exp \left(\frac{(\theta_i - \theta'_i - \frac{1}{2}\sigma_p^2 \frac{\partial}{\partial \theta'_i} \log p(X, \theta'))^2}{2a\sigma_p^2} \right)$$

$$\begin{aligned}
\log \frac{q(\theta \mid \theta', X)}{q(\theta' \mid \theta, X)} &= \sum_{i=1}^{d} \left(C + \frac{(\theta_i - \theta'_i - \frac{1}{2}\sigma_p^2 \frac{\partial}{\partial \theta'_i} \log p(X \mid \theta'))^2}{2a\sigma_p^2} \right) - \sum_{i=1}^{d} \Bigg(C + \frac{(\theta'_i - \theta_i - \frac{1}{2}\sigma_p^2 \frac{\partial}{\partial \theta_i} \log p(X \mid \theta))^2}{2a\sigma_p^2} \\
&= \frac{1}{2a\sigma_p^2} \sum_{i=1}^{d} \left((\theta_i - \theta'_i - \frac{1}{2}\sigma_p^2 \frac{\partial}{\partial \theta'_i} \log p(X \mid \theta'))^2 - (\theta'_i - \theta_i - \frac{1}{2}\sigma_p^2 \frac{\partial}{\partial \theta_i} \log p(X \mid \theta))^2 \right) \\
&= \frac{1}{2a\sigma_p^2} \sum_{i=1}^{d} \left((\theta_i - \theta'_i - G_i(X \mid \theta'))^2 - (\theta'_i - \theta_i - G_i(X \mid \theta))^2 \right)
\end{aligned}$$

Let

$$G_i(X \mid \theta) = \frac{1}{2}\sigma_p^2 \frac{\partial}{\partial \theta_i} \log p(X \mid \theta)$$

As

$$\log p(X \mid \theta) = \sum_{x \in X} \log p(x \mid \theta)$$

if X and X' only differ in one row, those being x and x', then

$$\begin{aligned} G_i(X \mid \theta) - G_i(X' \mid \theta) &= \frac{1}{2}\sigma_p^2 \sum_{j=1}^{n} \frac{\partial}{\partial \theta_i}(\log p(x_j \mid \theta) - \log p(x'_j \mid \theta)) \\ &= \frac{1}{2}\sigma_p^2 \frac{\partial}{\partial \theta_i}(\log p(x \mid \theta) - \log p(x' \mid \theta)) \\ &= G_i(x \mid \theta) - G_i(x' \mid \theta) \end{aligned}$$

Let X^* be the equal part of X and X'. Then

$$G_i(X \mid \theta) = G_i(X^* \mid \theta) + G_i(x \mid \theta)$$

and

$$G_i(X' \mid \theta) = G_i(X^* \mid \theta) + G_i(x' \mid \theta)$$

Additionally

$$(a+b)^2 - (a+c)^2 = 2a(b-c) + b^2 - c^2$$

for any $a, b, c \in \mathbb{R}$.

Let

$$\Delta(X) = \left| \log \frac{q(\theta \mid \theta', X)}{q(\theta' \mid \theta, X)} - \log \frac{q(\theta \mid \theta', X)}{q(\theta' \mid \theta, X)} \right|$$

$$
\begin{aligned}
\Delta(X) = {} & \frac{1}{2a\sigma_p^2} \Bigg| \sum_{i=1}^{d} (\theta_i - \theta'_i - G_i(X \mid \theta'))^2 - (\theta'_i - \theta_i - G_i(X \mid \theta))^2 \\
& - (\theta_i - \theta'_i - G_i(X' \mid \theta'))^2 + (\theta'_i - \theta_i - G_i(X' \mid \theta))^2 \Bigg| \\
= {} & \frac{1}{2a\sigma_p^2} \Bigg| \sum_{i=1}^{d} \Big((\theta_i - \theta'_i)^2 - 2(\theta_i - \theta'_i) G_i(X \mid \theta') + G_i(X \mid \theta')^2 \\
& - (\theta'_i - \theta_i)^2 + 2(\theta'_i - \theta_i) G_i(X \mid \theta) - G_i(X \mid \theta)^2 \\
& - (\theta_i - \theta'_i)^2 + 2(\theta_i - \theta'_i) G_i(X' \mid \theta') - G_i(X' \mid \theta')^2 \\
& + (\theta'_i - \theta_i)^2 - 2(\theta'_i - \theta_i) G_i(X' \mid \theta) + G_i(X' \mid \theta)^2 \Big) \Bigg| \\
= {} & \frac{1}{2a\sigma_p^2} \Bigg| \sum_{i=1}^{d} \Big(2(\theta'_i - \theta_i)(G_i(X \mid \theta) - G_i(X' \mid \theta)) + 2(\theta_i - \theta'_i)(G_i(X' \mid \theta') - G_i(X \mid \theta')) \\
& + G_i(X \mid \theta')^2 - G_i(X \mid \theta)^2 - G_i(X' \mid \theta')^2 + G_i(X' \mid \theta)^2 \Big) \Bigg| \\
= {} & \frac{1}{2a\sigma_p^2} \Bigg| \sum_{i=1}^{d} \Big(2(\theta'_i - \theta_i)(G_i(x \mid \theta) - G_i(x' \mid \theta) + G_i(x \mid \theta') - G_i(x' \mid \theta')) \\
& + (G_i(X^* \mid \theta') + G_i(x \mid \theta'))^2 - (G_i(X^* \mid \theta') + G_i(x' \mid \theta'))^2 \\
& + (G_i(X^* \mid \theta) + G_i(x' \mid \theta))^2 - (G_i(X^* \mid \theta) + G_i(x \mid \theta))^2 \Big) \Bigg| \\
= {} & \frac{1}{2a\sigma_p^2} \Bigg| \sum_{i=1}^{d} \Big(2(\theta'_i - \theta_i)(G_i(x \mid \theta) - G_i(x' \mid \theta) + G_i(x \mid \theta') - G_i(x' \mid \theta')) \\
& + 2G_i(X^* \mid \theta')(G_i(x \mid \theta') - G_i(x' \mid \theta')) + G_i(x \mid \theta')^2 - G_i(x' \mid \theta')^2 \\
& + 2G_i(X^* \mid \theta)(G_i(x' \mid \theta) - G_i(x \mid \theta)) + G_i(x' \mid \theta)^2 - G_i(x \mid \theta)^2 \Big) \Bigg| \\
\leq {} & \frac{1}{2a\sigma_p^2} \Bigg(2 \left| (\theta' - \theta)^T (G(x \mid \theta) - G(x' \mid \theta) + G(x \mid \theta') - G(x' \mid \theta')) \right| \\
& + 2 \left| G(X^* \mid \theta')^T (G(x \mid \theta') - G(x' \mid \theta')) \right| + 2 \left| G(X^* \mid \theta)^T (G(x' \mid \theta) - G(x \mid \theta)) \right| \\
& + \big| ||G(x \mid \theta')||_2^2 - ||G(x' \mid \theta')||_2^2 + ||G(x' \mid \theta)||_2^2 - ||G(x \mid \theta)||_2^2 \big| \Bigg)
\end{aligned}
$$

Continued:

$$\begin{aligned}
\Delta(X) \leq{}& \frac{1}{2a}\left|(\theta' - \theta)^T(\nabla \log p(x \mid \theta) - \nabla \log p(x' \mid \theta) + \nabla \log p(x \mid \theta') - \nabla \log p(x' \mid \theta'))\right| \\
&+ \frac{\sigma_p^2}{4a}\left|\nabla \log p(X^* \mid \theta')^T(\nabla \log p(x \mid \theta') - \nabla \log p(x' \mid \theta'))\right| \\
&+ \frac{\sigma_p^2}{4a}\left|\nabla \log p(X^* \mid \theta)^T(\nabla \log p(x' \mid \theta) - \nabla \log p(x \mid \theta))\right| \\
&+ \frac{\sigma_p^2}{8a}\left|||\nabla \log p(x \mid \theta')||_2^2 - ||\nabla \log p(x' \mid \theta')||_2^2 + ||\nabla \log p(x' \mid \theta)||_2^2 - ||\nabla \log p(x \mid \theta)||_2^2\right| \\
\leq{}& \frac{1}{2a}||\theta' - \theta||_2||\nabla \log p(x \mid \theta) - \nabla \log p(x' \mid \theta) + \nabla \log p(x \mid \theta') - \nabla \log p(x' \mid \theta')||_2 \\
&+ \frac{\sigma_p^2}{4a}||\nabla \log p(X^* \mid \theta')||_2||\nabla \log p(x \mid \theta') - \nabla \log p(x' \mid \theta')||_2 \\
&+ \frac{\sigma_p^2}{4a}||\nabla \log p(X^* \mid \theta)||_2||\nabla \log p(x' \mid \theta) - \nabla \log p(x \mid \theta)||_2 \\
&+ \frac{\sigma_p^2}{8a}\left|||\nabla \log p(x \mid \theta')||_2^2 - ||\nabla \log p(x' \mid \theta')||_2^2 + ||\nabla \log p(x' \mid \theta)||_2^2 - ||\nabla \log p(x \mid \theta)||_2^2\right| \\
\leq{}& \frac{2}{a}L_g||\theta' - \theta||_2 + \frac{\sigma_p^2}{a}(n-1)L_g^2 + \frac{\sigma_p^2 L_g^2}{2a}
\end{aligned}$$

with bound $||\nabla \log p(x \mid \theta)||_2 \leq L_g$. There may be possibility for improving the bound with additional information on $\nabla \log p(x \mid \theta) - \nabla \log p(x \mid \theta')$.

Together with the bound on the log likelihood ration this implies that

$$\Delta\lambda(X, \theta, \theta') \leq 2L||\theta - \theta'||_2 + \frac{2}{a}L_g||\theta - \theta'||_2 + \frac{\sigma_p^2}{a}(n-1)L_g^2 + \frac{1}{2a}\sigma_p^2 L_g^2$$

If there is an additional bound $|| \log p(X \mid \theta)||_2 \leq L_S$, then $|| \log p(X^* \mid \theta)||_2 \leq L_S + L_g$ and

$$\Delta\lambda(X, \theta, \theta') \leq 2L||\theta - \theta'||_2 + \frac{2}{a}L_g||\theta - \theta'||_2 + \frac{\sigma_p^2}{a}(L_S + L_g)L_g + \frac{1}{2a}\sigma_p^2 L_g^2$$

With large n this results in a much tighter sensitivity bound than the previous bound.

5. Differentially Private Hamiltonian Monte Carlo

6. Limitations and Alternatives of DP MCMC

6.1 Dataset Size

6.2 Alternatives to DP MCMC

6.2.1 DP Variational Inference

6.2.2 Releasing Summary Statistics Privately

7. Experiments

7.1 Maximum Mean Discrepancy

The convergence of non-DP MCMC algorithms is typically assessed using $\hat{R}$ [GCS+14], which measures how well multiple chains started from different points have mixed together. The utility of a sample produced by an MCMC algorithm can be evaluated using effective sample size (ESS) [GCS+14], which is an estimate of the size of an uncorrelated sample of the posterior with the same estimation utility as the MCMC sample.

Both $\hat{R}$ and ESS require that the MCMC algorithm asymptotically targets the true posterior, as they cannot detect an algorithm that has converged to the wrong distribution. Because some of the DP MCMC algorithms use approximations that may cause the algorithms to not converge to the true posterior, such as clipping the log likelihood ratios, $\hat{R}$ and ESS are not suitable for assessing the performance of the algorithms.

Because the DP MCMC algorithms may not converge to the correct distribution, their performance should be evaluated with a metric that measures how close to the true distribution they are. A very general such metric is maximum mean discrepancy (MMD) [GBR+12]. MMD between distributions p and q is defined as

$$\mathrm{MMD}(p, q) = \sup_{f \in \mathcal{F}} (E_{x\sim p} f(x) - E_{y\sim q} f(y))$$

where $\mathcal{F}$ is some class of functions. By choosing a suitable $\mathcal{F}$, $\mathrm{MMD}(p, q)$ can be estimated from a sample from p and q. The suitable classes $\mathcal{F}$ can be characterised by a kernel function $k\colon P \times Q \to \mathbb{R}$, where P and Q are the supports of p and q, respectively. The Gaussian radial basis function (RBF) kernel

$$k(x, y) = \exp\left(\frac{||x - y||_2^2}{2\sigma^2}\right)$$

is particularly well suited, as it has the property that $\mathrm{MMD}(p, q) = 0$ if and only if $p = q$. After choosing a kernel, $\mathrm{MMD}(p, q)$ may be estimated from finite samples of p and q. To evaluate MCMC algorithms, one of the samples is the output of the algorithm to be evaluated, and the other is a sample from the true posterior.

The choice of the σ parameter of k affects the way MMD evaluates different kinds of differences in p and q. For some preliminary experiments in this thesis, σ was chosen to be 1, which penalised error in the mean much more than errors in higher moments in the experiments. The σ used for the final experiments chosen by picking 50 subsamples from both samples with replacement and setting σ to be the median between distances of points of the subsamples. This is following the procedure of Gretton et. al. [GBR+12], with the addition of the subsampling step to handle samples of different sizes.

7.2 The Effects of Clipping

The first experiment evaluates the effect of clipping log likelihood ratios. Both HMC and random walk Metropolis-Hastings (RWMH)[*] algorithms were run on 2 and 10 dimensional banana models. RWMH did not converge in reasonable time in 10 dimensions so it was excluded from these results. DP was not used so that error from the extra noise would not affect the results.

The banana model used hyperparameter values $a = 20$, $b = m = 0$, $\sigma_1^2 = 20$, $\sigma_2^2 = 2.5$, $\sigma_3^2 = 1$, $\sigma_0^2 = 1000$ and $n = 100000$. In both experiments, 500 samples from HMC and 3000 samples from RWMH were taken and the latter half[†] of them were compared to 2000 samples from the true posterior. The reference posterior sample was also compared to other samples from the posterior to obtain a baseline

Figure 7.1 shows the results of the clipping experiment. The top left and bottom left panels show MMD as a function of the clip bound and the fraction of log likelihoods that was actually clipped for both HMC and RWMH in the 2-dimensional model. The effect of clipping on MMD is nonexistent for all but the lowest clip bounds. The top and bottom right panels show results for the 10-dimensional model. This time there are chains that did not converge correctly with most clip bounds, but the chains with the higher bounds converged. Based on these results, if the clip fraction is less than 10%, clipping is likely undetectable without a large sample.

[*] Metropolis-Hastings using the Gaussian distribution as the proposal distributions

[†] As the start of an MCMC chain depends heavily on the starting point, samples at the start should be discarded as they are not representative of the target distribution

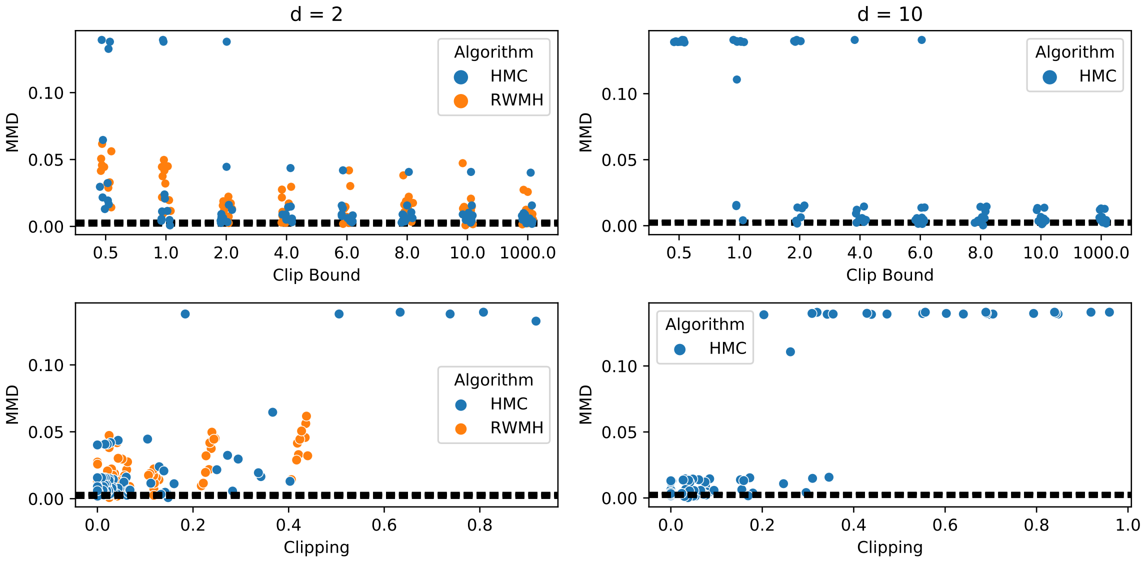

Figure 7.1: The effect of log likelihood ratio clipping on the posterior of the banana model for random walk Metropolis-Hastings and HMC. The top row shows posterior MMD as a function of the clip bound, and the bottom row shows MMD as a function of the fraction of log likelihoods that were clipped. The left columns used a 2-dimensional posterior while the right columns had a 10-dimensional posterior. The black lines show the MMDs ten different samples of the true posterior compared to the reference sample. All of the lines are close the each other and appear as a single line. Random walk Metropolis-Hastings was not included with the 10-dimensional experiment as it was unable to converge and provide meaningful results there.

7.3 Accuracy of CLT With Subsampling

7.4 Banana Distribution

8. Conclusions

Bibliography

[Bar65] Av A Barker. Monte carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics*, 18(2):119–134, 1965.

[BS16] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography - 14th International Conference, TCC 2016-B, Beijing, China, October 31 - November 3, 2016, Proceedings, Part I*, pages 635–658, 2016.

[CD99] DM Ceperley and Mark Dewing. The penalty method for random walks with uncertain energies. *The Journal of chemical physics*, 110(20):9812–9820, 1999.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[GBR$^+$12] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

[GCS$^+$14] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman & Hall/CRC texts in statistical science series. CRC Press, Boca Raton, third edition, 2014.

[HJDH19] Mikko A. Heikkilä, Joonas Jälkö, Onur Dikmen, and Antti Honkela. Differentially private markov chain monte carlo. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 4115–4125, 2019.

[Mir17] Ilya Mironov. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium, CSF 2017, Santa Barbara, CA, USA, August 21-25, 2017*, pages 263–275, 2017.

[RR98] Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

[SMM19] David M. Sommer, Sebastian Meiser, and Esfandiar Mohammadi. Privacy loss classes: The central limit theorem in differential privacy. *PoPETs*, 2019(2):245–269, 2019.

[TPK14] Minh-Ngoc Tran, Michael K. Pitt, and Robert Kohn. Adaptive metropolis-hastings sampling using reversible dependent mixture proposals. *Statistics and Computing*, 26(1-2):361–381, 2014.

[WBK19] Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled renyi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 1226–1235, 2019.

[YE19] Sinan Yildirim and Beyza Ermis. Exact MCMC with differentially private moves - revisiting the penalty algorithm in a data privacy framework. *Statistics and Computing*, 29(5):947–963, 2019.