Some Topics in Markov Chain Monte Carlo Theory

Zhang Rui-Yang

Preface

This thesis looks at several topics in the theory of Markov chain Monte Carlo. There are three parts to this thesis. The first part is a general introduction to the concept of the Markov chain Monte Carlo and introduces some basic results of the geometric ergodicity of a Markov chain. The second part focuses on various results about the Langevin algorithms, and studies some theoretical results of the unadjusted Langevin algorithm. Chapter 4 is mostly based on Dalalyan (2017).

The third part focuses on the recently introduced Barker proposal in Livingstone and Zanella (2022) and its various extensions. Chapter 5 introduces the Barker proposal, and most of its content is based on Livingstone and Zanella (2022), Hird et al. (2022), and Vogrinc et al. (2022). The derivations in Section 5.3 are partially original. The other two chapters of this part are original work done jointly by the author and his supervisor. Chapter 6 considers the Barker proposal when the Metropolis adjustment is removed which can then be viewed as a numerical scheme to solve stochastic differential equations, while Chapter 7 looks at an alternative way to extend the Barker proposal from one-dimensional to n-dimensional.

I would like to thank my supervisor Sam Livingstone for recommending me to do a thesis project, and for his guidance and encouragement along the way. I would also like to thank everyone in Sam's research group, including (in alphabetical order) Luke Hardcastle, Max Hird, Xitong Liang, and Giorgos Vasdekis, for their helpful and lively discussions. Special shout-out to Max for his careful reading and constructive comments on a draft of my thesis. This thesis would not be possible without the support of my parents who patiently listen to all my ramblings.

London, UK
April, 2023

Contents

Part I

Introduction

Chapter 1

Metropolis-Hastings Algorithms

Markov chain Monte Carlo (MCMC) algorithms have proved to be extremely effective in various computation-intensive settings, such as Bayesian statistics (Diaconis, 2009), statistical mechanics (Faulkner and Livingstone, 2022), and machine learning (Andrieu et al., 2003).

Usually, MCMC algorithms can be implemented to do two things: estimating and sampling. When we would like to compute an integral that is (almost) impossible to do by hand, say it is of high dimension and has a complicated form, we have to turn to an approximate solution and use Monte Carlo methods instead. Other times we might have a probability distribution in mind, and we would like to generate independent and identically distributed (i.i.d.) samples from this distribution, and MCMC algorithms are good at it, especially when the target distribution is too complex to be sampled from using standard methods (e.g. inverse CDF). The second goal is harder to achieve than the first goal, and we can compute good approximations once we have obtained good samples using ergodic averages. For example, if we would like to estimate an integral of the form

$$\int f(x)\pi(x)dx =: \mathbb{E}_\pi[f(X)],$$

where π is a probability distribution and f is an arbitrary function, we could generate i.i.d. samples $X_1, X_2, \ldots, X_n$ following π and estimate the integral by

$$\mathbb{E}_\pi[f(X)] \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

This approximation is justified by the strong law of large numbers (SLLN) (Williams, 1991) which says

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \mathbb{E}_\pi[f(X)] \qquad a.s.$$

as $n \to \infty$. This estimation scheme is known as the **Monte Carlo method**. Note that this scheme works for any integral, as we can have

$$\int f(x)dx = \int \frac{f(x)}{\pi(x)}\pi(x)dx = \mathbb{E}_\pi\left[\frac{f}{\pi}(X)\right]$$

for some function f and probability distribution function π.

The rough underlying idea of MCMC is as follows. An ergodic[1] Markov chain, after running for many steps, will converge to an equilibrium distribution regardless of its initial position. This means, once we have reached the equilibrium, every new step made by the chain can be viewed as samples from that equilibrium distribution, and thus we can easily obtain samples and compute estimations using the Monte Carlo method afterward. This also explains the name of the algorithm, i.e. we use a Markov chain to generate samples and then use the Monte Carlo method to approximate.

One of the first MCMC algorithms is the **Metropolis-Hastings Algorithm** (MH) (Hastings, 1970). Given a target distribution $\pi \in \mathbb{R}^n$ for some dimension n (we will take $\mathbb{R}^n$ as our default space for the rest of the thesis), a proposal kernel $Q(\cdot, \cdot)$ with $Q(x, A) = \int_A q(x, y)dy$ where $q(x, y)$ is the probability of moving from x to y, and a starting position x, we have

Algorithm 1 Metropolis-Hastings Algorithm

Require: Target distribution π, proposal kernel $q(\cdot, \cdot)$, starting position x

$X_0 = x$
for $i = 0, 1, 2, \cdots$ **do**
    $X_{curr} = X_i$
    $X_{prop} \sim q(X_{curr}, \cdot)$
    Draw $U \sim Unif[0, 1]$
    if $\alpha(X_{curr}, X_{prop}) < U$ **then**
        Accept X_{prop} and $X_{i+1} = X_{prop}$.
    else
        Reject X_{prop} and $X_{i+1} = X_{curr}$.
    end if
end for

Here, $\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right\}$ is the acceptance probability. The above algorithm will output a sequence $\{X_n\}$, and under some conditions on Q the distribution of X_n will converge to π.

The part of the above algorithm that decides whether or not we should keep the proposed move is commonly referred to as the **Metropolis adjustment**. Note that this adjustment is essential for an MH algorithm, but it is not always needed for a general MCMC algorithm. An algorithm that is Metropolis-Hastings but with the adjustment removed is commonly called an **unadjusted** algorithm, and an example of such an algorithm is the unadjusted Langevin algorithm (to be discussed in Chapter 3 and 6).

The overall transition kernel $P(x, dy)$ of the above distribution is

$$P(x, dy) = Q(x, dy)\alpha(x, y) + \delta_x(dy) \int (1 - \alpha(x, u))Q(x, du)$$

where the Dirac measure $\delta_a(A) = 1$ when $a \in A$ and 0 otherwise for any (measurable) set A. The above kernel consists of two parts. The first part is when we accept the proposal that moves us from x to y, and the second part is when our proposal is rejected but we have $x = y$ to begin with.

We would want the transition kernel to have π as its invariant measure / distribution. It turns out that if the kernel satisfies the detailed balance equation(s), the kernel will be π-reversible,

[1] ϕ-irreducible and aperiodic

or simply **reversible**, and the π-invariance is guaranteed. Recall that for a Markov chain with transition kernel P to be π-invariant, it means that we have

$$\int_x \pi(dx)P(x, dy) = \pi(dy).$$

Definition 1.1. *A Markov chain with transition kernel P is π-**reversible** for some distribution π when it satisfies the **detailed balance equation(s)***

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx)$$

for all possible x, y.

Proposition 1.2. *If a Markov chain with transition kernel P is π-reversible, then it is π-invariant.*

Proof. Using the detailed balance equation, we have

$$\int_x \pi(dx)P(x, dy) = \int_x \pi(dy)P(y, dx) = \pi(dy) \int_x P(y, dx) = \pi(dy),$$

as desired. □

So, if a Markov chain with transition kernel P satisfies the detailed balance equation, it would be π-reversible and therefore π-invariant, and this means, given that the chain is ergodic[2], the chain has π as its equilibrium measure / distribution, which is extremely desirable.

Theorem 1.3. *The Metropolis-Hastings algorithm, as constructed in Algorithm 1, produces a Markov chain $\{X_n\}_{n\in\mathbb{N}}$ that is π-reversible if target π and proposal kernel Q admit densities.*

Proof. We just need to show that the transition kernel $P(x, dy)$ of the algorithm satisfies the detailed balance equation

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx).$$

The equation is trivial when $x = y$, so we will only consider $x \neq y$. We have

$$\begin{aligned}
\pi(dx)P(x, dy) &= \pi(dx) \left[Q(x, dy)\alpha(x, y) + \delta_x(dy) \int (1 - \alpha(x, u))Q(x, du) \right] \\
&= [\pi(x)dx][q(x, y)dy]\alpha(x, y) \\
&= [\pi(x)dx][q(x, y)dy] \min \left\{ 1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} \right\} \\
&= \min \{\pi(x)q(x, y), \pi(y)q(y, x)\} \, dxdy \\
&= \pi(y)q(y, x) \min \left\{ \frac{\pi(x)q(x, y)}{\pi(y)q(y, x)}, 1 \right\} dxdy \\
&= \pi(dy)Q(y, dx)\alpha(y, x) = \pi(dy)P(y, dx),
\end{aligned}$$

as desired. □

[2] Almost always the case for Markov chains generated by MCMC algorithms, and is easy to check.

Even though the Metropolis-Hastings algorithm will generate a Markov chain that respects π as its invariant measure, we do not know when it will actually converge to π. This question of convergence (and the rate of convergence) is a big and active area of research in MCMC theory, and we direct the readers to Roberts and Rosenthal (2004) for a survey on various existing results. We have considered exactly this question in various forms in the later chapters, such as Chapter 3 and 6.

1.1 Metropolis Adjusted Langevin Algorithm

The Metropolis Adjusted Langevin Algorithm (MALA) is a specific kind of MH algorithm. It is an algorithm with a particular choice of proposal kernel that is based on the Langevin diffusion process. The Langevin diffusion $\{L_t\}$ can be characterised by the following stochastic differential equation (SDE):

$$dL_t = \frac{1}{2}\nabla \log \pi(L_t)dt + dB_t,$$

where the probability density π is differentiable and $\pi > 0$.

It can be verified that, using the Fokker-Planck equation, π is indeed an invariant distribution of the above SDE (Xifara et al., 2014). The proposal of MALA involves a discretisation of this SDE, and it would naturally incur a discretisation error which is then corrected via the Metropolis adjustment step.

The algorithm of k-dimensional MALA is as follows.

Algorithm 2 k-Dimensional Metropolis Adjusted Langevin Algorithm

Require: Target distribution π, starting position x

$X_0 = x$
for $i = 0, 1, 2, \cdots$ **do**
  $X_{curr} = X_i$
  $X_{prop} \sim N_k(X_{curr} + \frac{h}{2}\nabla \log \pi(X_{curr}), hI_k)$.
  Draw $U \sim Unif[0, 1]$
  if $\alpha(X_{curr}, X_{prop}) < U$ **then**
    Accept X_{prop} and $X_{i+1} = X_{prop}$.
  else
    Reject X_{prop} and $X_{i+1} = X_{curr}$.
  end if
end for

Here, $N_k(\mu, \Sigma)$ refers to a k-dimensional normal distribution with mean vector μ and variance matrix Σ. Note that the choice of $\nabla \log \pi$ instead of $\nabla \pi$ is not an arbitrary one. Many real-life applications of MCMC algorithms are for Bayesian inferences, where we would only know distributions proportionally, so we would only have $C\pi$ instead of π for some unknown constant C. This would not affect $\nabla \log \pi$ as we have

$$\nabla \log C\pi = \frac{C\pi'}{C\pi} = \frac{\pi'}{\pi} = \nabla \log \pi,$$

which is not the case for $\nabla \pi$.

Chapter 2

Geometric Ergodicity

This chapter is heavily adapted from the survey Roberts and Rosenthal (2004).

2.1 Total Variation Distance Basics

For a (time-homogeneous) Markov chain $\{X_n\}$ with state space $\mathcal{X}$, we let it have **stationary** distribution $\pi(\cdot)$ and define its **transition kernel** as

$$P(x, A) = \mathbb{P}[X_{n+1} \in A \mid X_n = x]$$

for any measurable set $A \subseteq \mathcal{X}$, and the **n-step transition kernel** as

$$P^n(x, A) = \mathbb{P}[X_n \in A \mid X_0 = x]$$

for any measurable set $A \subseteq \mathcal{X}$. If a transition distribution $P(x, \cdot)$ admits a density P, we will denote P as the **transition density**.

If we want to measure the convergence of a Markov chain, we would need a notion of distance between distributions. There are many possibilities and one of them is the total variation distance, defined as follows.

Definition 2.1. *The **total variation distance** between two probability measures $\nu_1(\cdot)$ and $\nu_2(\cdot)$ defined on the same measurable space $(\Sigma, \mathcal{F})$ is*

$$\|\nu_1(\cdot) - \nu_2(\cdot)\| = \sup_{A \in \mathcal{F}} |\nu_1(A) - \nu_2(A)|.$$

This notion allows us to formally describe the convergence of a Markov chain to its stationary distribution, i.e. the Markov chain converges to its stationary distribution $\pi(\cdot)$ if

$$\lim_{n \to \infty} \|P^n(x, \cdot) - \pi(\cdot)\| = 0$$

for any π-a.e. x in the state space.

This convergence is only qualitative, and we do not know the exact rate of this convergence. Later on, we will introduce the concepts of uniform ergodicity and geometric ergodicity, which are more refined types of convergence of a Markov chain.

There are several results about the total variation distance that we will use later.

Proposition 2.2. *We have*

(a) $\|\mu_1(\cdot) - \mu_2(\cdot)\| = \sup_{f:\mathcal{X}\to[0,1]} |\int f d\mu_1 - \int f d\mu_2|$.
(b) $\|\mu_1(\cdot) - \mu_2(\cdot)\| = (b-a)^{-1} \sup_{f:\mathcal{X}\to[a,b]} |\int f d\mu_1 - \int f d\mu_2|$ *for any* $b > a$.
(c) If $\pi(\cdot)$ *is stationary for a Markov chain kernel* P*, then* $\|P^n(x,\cdot) - \pi(\cdot)\|$ *is non-increasing in* n.
(d) Letting $(\mu_i P)(A) := \int P(x,A)\mu_i(dx)$*, we always have* $\|(\mu_1 P)(\cdot) - (\mu_2 P)(\cdot)\| \leq \|\mu_1(\cdot) - \mu_2(\cdot)\|$.
(e) Let $t(n) := 2\sup_{x\in\mathcal{X}} \|P^n(x,\cdot) - \pi(\cdot)\|$*, then,* $t(a+b) \leq t(a)t(b)$ *for all* $a, b \in \mathbb{N}$.

Proof. (a) Consider $\rho = \mu_1 + \mu_2$, and let $g := d\mu_1/d\rho$ and $h := d\mu_2/d\rho$ be the Radon-Nikodym derivatives (Williams, 1991). WLOG, we assume $\rho(\{g > h\}) > \rho(\{g \leq h\})$. We have

$$\left|\int f d\mu_1 - \int f d\mu_2\right| = \left|\int f(g-h)d\rho\right|.$$

Since $f : \mathcal{X} \to [0,1]$, this integral will be maximised when f takes 1 on $A := \{g > h\}$ and takes 0 on $\{g \leq h\}$. In that case, we have

$$\sup_{f:\mathcal{X}\to[0,1]} \left|\int f d\mu_1 - \int f d\mu_2\right| = \left|\int_A (g-h)d\rho\right| = |\mu_1(A) - \mu_2(A)| = \|\mu_1(\cdot) - \mu_2(\cdot)\|.$$

(b) Similar to before. Just consider f that takes b on $A := \{g > h\}$ and takes a on $\{g \leq h\}$. Then, we have the desired equality.

(c) We have for any measurable $A \subseteq \mathcal{X}$

$$\begin{aligned}
\|P^{n+1}(x,\cdot) - \pi(\cdot)\| &\leq |P^{n+1}(x,A) - \pi(A)| \\
&= \left|\int_y P^n(x,dy)P(y,A) - \int_y \pi(dy)P(y,A)\right| \\
&\leq \|P^n(x,\cdot) - \pi(\cdot)\|
\end{aligned}$$

where the first inequality follows from the definition of total variation distance, and the last inequality follows from (a) as $P(\cdot, A) : \mathcal{X} \to [0,1]$.

(d) Similar to (c). We have for any measurable $A \subseteq \mathcal{X}$

$$\begin{aligned}
\|(\mu_1 P)(\cdot) - (\mu_2 P)(\cdot)\| &\leq |(\mu_1 P)(A) - (\mu_2 P)(A)| \\
&= \left|\int P(x,A)\mu_1(dx) - \int P(x,A)\mu_2(dx)\right| \\
&\leq \|\mu_1(\cdot) - \mu_2(\cdot)\|
\end{aligned}$$

where the first inequality follows from the definition of total variation distance, and the last inequality follows from (a) as $P(\cdot, A) : \mathcal{X} \to [0,1]$.

(e) For any a, b, we define $\hat{P}(x,\cdot) := P^a(x,\cdot) - \pi(\cdot)$ and $\hat{Q}(x,\cdot) := P^b(x,\cdot) - \pi(\cdot)$. Consider some

$f : \mathcal{X} \to [0,1]$, we have

$$\begin{aligned}(\hat{P}\hat{Q}f)(x) &= \int_y f(y) \int_z [P^a(x,dz) - \pi(dz)][P^b(z,dy) - \pi(dy)] \\ &= \int_y f(y) \int_z [P^a(x,dz)P^b(z,dy) - \pi(dz)P^b(z,dy) - P^a(x,dz)\pi(dy) + \pi(dz)\pi(dy)] \\ &= \int_y f(y)[P^{a+b}(x,dy) - \pi(dy) - \pi(dy) + \pi(dy)] \\ &= \int_y f(y)[P^{a+b}(x,dy) - \pi(dy)],\end{aligned}$$

which means $t(a+b) \leq 2\sup_{x\in\mathcal{X}}(\hat{P}\hat{Q}f)(x)$ by (a).

Next, let $g(x) := (\hat{Q}f)(x) = \int_y \hat{Q}(x,dy)f(y) = \int_y f(y)[P^m(x,dy) - \pi(dy)]$, and $g^* := \sup_x |g(x)|$. Then, we have

$$\begin{aligned}g^* = \sup_x |g(x)| &= \sup_x \left| \int_y f(y)[P^m(x,dy) - \pi(dy)] \right| \\ &\leq \sup_x \|P^m(x,\cdot) - \pi(\cdot)\| = \frac{1}{2}t(m)\end{aligned}$$

where the last inequality follows from (a). So $2g^* \leq t(m)$.

If $g^* = 0$, then $g(x) = 0$ for all x and we have $(\hat{P}\hat{Q}f)(x) = 0$. If $g^* \neq 0$, we let $(g/g^*)(x) := g(x)/g^*$, so $g/g^* : \mathcal{X} \to [-1,1]$. We have $(\hat{Q}f)(x) = g^* \cdot (g/g^*)(x)$. This means

$$\begin{aligned}t(a+b) &\leq 2\sup_x |(\hat{P}\hat{Q}f)(x)| \\ &= 2g^* \sup_x |(\hat{P}[g/g^*])(x)| \\ &\leq t(b) \sup_x |(\hat{P}[g/g^*])(x)| \\ &\leq t(b)2\sup_x \|P^a(x,\cdot) - \pi(\cdot)\| \\ &\leq t(b)t(a)\end{aligned}$$

where the second last inequality follows from (b) as $g/g^* : \mathcal{X} \to [-1,1]$. When $g^* = 0$, $t(a+b) \leq t(a)t(b)$ holds trivially as both sides will become zero. Thus, we have obtained our desired inequality. □

2.2 Ergodicity

In this section, we will investigate the notion of ergodicity as a guarantee of Markov chain convergence.

Definition 2.3. *A Markov chain is* ϕ***-irreducible*** *if there exists a non-zero* σ*-finite measure* ϕ *on* $\mathcal{X}$ *such that for all measurable* $A \subseteq \mathcal{X}$ *with* $\phi(A) > 0$ *and all* $x \in \mathcal{X}$*, there exists a positive integer* n *such that* $P^n(x,A) > 0$*.*

Definition 2.4. *A Markov chain with stationary distribution* $\pi(\cdot)$ *is* ***aperiodic*** *if there does not exists* $d \geq 2$ *and disjoint subsets* $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_d \subseteq \mathcal{X}$ *with* $P(x, \mathcal{X}_{i+1}) = 1$ *for all* $x \in \mathcal{X}_i$ *for* $i = 1, 2, \ldots, d-1$ *and* $P(x, \mathcal{X}_1) = 1$ *for all* $x \in \mathcal{X}_d$*, such that* $\pi(\mathcal{X}_1) > 0$*.*

Theorem 2.5 (Roberts and Rosenthal (2004)). *If a Markov chain on a state space with countably generated σ-algebra is ϕ-irreducible and aperiodic with a stationary distribution $\pi(\cdot)$, then for π-a.e. $x \in \mathcal{X}$, we have*

$$\lim_{n\to\infty} \|P^n(x,\cdot) - \pi(\cdot)\| = 0.$$

In particular, we have $\lim_{n\to\infty} \|P^n(x,A) - \pi(A)\| = 0$ for all measurable $A \subseteq \mathcal{X}$.

Remark. *For most of the Markov chains for MCMC algorithms, ϕ-irreducibility and aperiodicity are almost always satisfied, since we usually can reach anywhere in the state space from any starting point via the $\mathbb{R}^n$-supported proposal kernel.*

2.2.1 Uniform Ergodicity

Ergodicity establishes convergence, yet it does not indicate the rate of convergence. The notion of uniform ergodicity, and the notion of geometric convergence introduced in the following subsection, shed light on this rate.

Definition 2.6. *A Markov chain having stationary distribution $\pi(\cdot)$ is **uniformly ergodic** if*

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq M\rho^n$$

for $n = 1, 2, 3, \ldots$ and for $\rho < 1$ and $M < \infty$.

Definition 2.7. *A subset $C \subseteq \mathcal{X}$ is **small** if there exists a positive integer n_0, $\varepsilon > 0$, and a probability measure $\nu(\cdot)$ on $\mathcal{X}$ such that the following **minorisation condition** holds:*

$$P^{n_0}(x,\cdot) \geq \varepsilon\nu(\cdot)$$

for all $x \in C$. In particular, we have $P^{n_0}(x,A) \geq \varepsilon\nu(A)$ for all measurable $A \subseteq \mathcal{X}$ for all $x \in C$.

It turns out that small sets are not hard to find for Markov chains generated by MH algorithms. In fact, all nonempty compact sets are small. Since we are working in $\mathbb{R}^n$, every compact set is simply closed and bounded, according to the Heine-Borel theorem.

Theorem 2.8 (Theorem 2.2 of Roberts and Tweedie (1996b)). *Suppose that π is bounded away from 0 and ∞ on compact sets, and there exist positive δ_q and ε_q such that for every x,*

$$|x - y| \leq \delta_q \implies q(x,y) \geq \varepsilon_q.$$

Then, the Markov chain produced by that MH algorithm with proposal kernel $q(x,y)dy$ is irreducible, aperiodic, and every nonempty compact set is small.

Theorem 2.9 (Roberts and Rosenthal (2004)). *A Markov chain with invariant distribution $\pi(\cdot)$ that satisfies the minorisation condition for some $n_0 \in \mathbb{N}$, $\varepsilon > 0$, probability measure $\nu(\cdot)$ with small set $C = \mathcal{X}$ is uniformly ergodic. Additionally, we have*

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq (1-\varepsilon)^{\lfloor n/n_0 \rfloor}$$

for all $x \in \mathcal{X}$.

2.2.2 Geometric Ergodicity

Definition 2.10. *A Markov chain having stationary distribution $\pi(\cdot)$ is **geometrically ergodic** if*

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq M(x)\rho^n$$

for $n = 1, 2, 3, \ldots$ and for $\rho < 1$ and $M(x) < \infty$ for π-a.e. $x \in \mathcal{X}$.

Definition 2.11. *A Markov chain is said to satisfy a* ***drift condition*** *if there are constants* $\lambda \in (0,1)$, $b < \infty$, *and a Lyapunov function* $V : \mathcal{X} \to [1, \infty]$ *such that*

$$PV \leq \lambda V + b1_C$$

i.e. $\int_{\mathcal{X}} P(x, dy)V(y) \leq \lambda V(x) + b1_C(x)$ *for all* $x \in \mathcal{X}$.

Remark. *The drift condition ensures that when the Markov chain leaves the small set* C, *it will return at a geometrical rate* λ.

Theorem 2.12 (Roberts and Rosenthal (2004))**.** *A* ϕ*-irreducible, aperiodic Markov chain with stationary distribution* $\pi(\cdot)$ *is geometrically ergodic if it satisfies a minorisation condition for some* $C \subseteq \mathcal{X}$, $\varepsilon > 0$, *probability measure* $\nu(\cdot)$ *as well as the drift condition for* $\lambda \in (0,1)$, $b < \infty$, *and a function* $V : \mathcal{X} \to [1, \infty]$ *with* $V(x) < \infty$ *for* π*-a.e.* $x \in \mathcal{X}$.

In fact, we could further simplify the conditions needed to be checked for geometric ergodicity. Since we have from Theorem 2.8 that all compact sets are small, we just need to check for the following version of the drift condition.

Proposition 2.13 (Proposition 3.1 of Roberts and Tweedie (1996b))**.** *If* π *satisfies Theorem 2.8, then the Markov chain produced by the MH algorithm with transition kernel* P *is geometrically ergodic if and only if there exists a real-valued function* $V > 1$ *such that*

$$\limsup_{\|x\| \to \infty} \frac{PV(x)}{V(x)} = \limsup_{\|x\| \to \infty} \int \frac{V(y)}{V(x)} P(x, dy) < 1.$$

Remark. *The norm for* $\|x\| \to \infty$ *is the standard Euclidean norm, instead of the total variation distance. We do not specify this difference here, as it is usually clear from the context which norm we are using.*

Proof. The minorisation condition needed for geometric ergodicity is satisfied by Theorem 2.8, so we just need to verify that the above inequality is an alternative form of the drift condition.

It is obvious that the drift condition implies the inequality. We have,

$$\begin{aligned} PV(x) &\leq \lambda V(x) + b1_C(x) \\ \frac{PV(x)}{V(x)} &\leq \lambda + b\frac{1_C(x)}{V(x)} \\ \limsup_{\|x\| \to \infty} \frac{PV(x)}{V(x)} &\leq \limsup_{\|x\| \to \infty} \left[\lambda + b\frac{1_C(x)}{V(x)}\right] = \lambda < 1 \end{aligned}$$

For the reverse direction, we have

$$\limsup_{\|x\| \to \infty} \frac{PV(x)}{V(x)} = \lambda < 1.$$

For any $\varepsilon > 0$, there exists a constant $M_\varepsilon > 0$ such that for all $\|x\| > M_\varepsilon$, we have

$$\frac{PV(x)}{V(x)} = \lambda + \varepsilon \iff PV(x) = (\lambda + \varepsilon)V(x).$$

Next, for $\|x\| \leq M_\varepsilon$, since this is a bounded set and

$$PV(x) - (\lambda + \varepsilon)V(x)$$

is continuous in x, we have

$$PV(x) - (\lambda + \varepsilon)V(x) < b$$

for some constant b. Therefore, we have the inequality

$$PV(x) - (\lambda + \varepsilon)V(x) < b1_{\{\|x\| \leq M_\varepsilon\}}(x) \iff PV(x) < (\lambda + \varepsilon)V(x) + b1_{\{\|x\| \leq M_\varepsilon\}}(x).$$

Since ε is arbitrary, we can take the limit of it going to zero, and get the desired inequality. □

2.3 Proofs of Theorems

In this section, we will present the proofs of the mentioned theorems in previous sections. The proofs employ a powerful technique called the **coupling** argument. Consider two random variables X and Y that are defined on the same space $\mathcal{X}$, with laws $L(X)$ and $L(Y)$ respectively.

Lemma 2.14 (Coupling Inequality)**.** $\|L(X) - L(Y)\| \leq \mathbb{P}(X \neq Y)$

Proof. We have

$$\begin{aligned}
&\|L(X) - L(Y)\| \\
&= \sup_A |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)| \\
&= \sup_A |\mathbb{P}(X \in A, X = Y) + \mathbb{P}(X \in A, X \neq Y) - \mathbb{P}(Y \in A, Y = X) - \mathbb{P}(Y \in A, Y \neq X)| \\
&= \sup_A |\mathbb{P}(X \in A, X \neq Y) - \mathbb{P}(Y \in A, Y \neq X)| \\
&= \sup_A \mathbb{P}(X \neq Y)|\mathbb{P}(X \in A \mid X \neq Y) - \mathbb{P}(Y \in A \mid X \neq Y)| \leq \mathbb{P}(X \neq Y).
\end{aligned}$$

□

The dependency of X and Y, or how they are coupled, is not fixed, and that gives us room to utilise the above inequality.

Example. *Consider $X, Y \sim N(0, 1)$, or any other distribution really. Clearly, we have $\|L(X) - L(Y)\| = 0$ as $L(X) = L(Y)$.*

- *If $X = Y$, we have $0 = \|L(X) - L(Y)\| \leq \mathbb{P}(X \neq Y) = 0$.*
- *If X, Y are independent so $X \neq Y$ a.s., we have $0 = \|L(X) - L(Y)\| \leq \mathbb{P}(X \neq Y) = 1$.*

The first inequality is much better than the second one, and the quality of the inequality, as illustrated in this example, is dictated by how we design the dependency of X and Y.

2.3.1 Coupling Construction

The ergodicity results are all about the quantity $\|P^n(x, \cdot) - \pi(\cdot)\|$. If we can construct two random variable sequences $\{X_n\}$ and $\{X'_n\}$ such that their respective laws are $P^n(x, \cdot)$ and $\pi(\cdot)$, we could obtain a bound on this total variation distance using the coupling inequality by simply computing the probability that the two random variables (for a fixed n) are different, i.e. $\mathbb{P}(X_n \neq X'_n)$.

As illustrated in the previous example, the quality of the bound by coupling inequality relies on how we design the dependency of the two random variables. Here, we present a neat coupling

construction of X_n and X'_n so they have the desired laws, and at the same time have a low probability of being different.

Recall that the minorisation condition gives us $\varepsilon, n_0, C, \nu(\cdot)$. We will use these things in the following construction.

ALGORITHM OF CONSTRUCTING A COUPLING

Start: $X_0 = x$, $X'_0 \sim \pi(\cdot)$.

Loop: Given the current states X_n and X'_n

1. If $X_n = X'_n$, then draw $X_{n+1} = X'_{n+1} \sim P(X_n, \cdot)$. Increase n by 1.
2. If $(X_n, X'_n) \in C \times C$, then
   (a) draw $X_{n+n_0} = X'_{n+n_0} \sim \nu(\cdot)$ with probability ε.
   (b) draw conditionally independently $X_{n+n_0} \sim 1/(1-\varepsilon)[P^{n_0}(X_n, \cdot) - \varepsilon\nu(\cdot)]$ and $X'_{n+n_0} \sim 1/(1-\varepsilon)[P^{n_0}(X'_n, \cdot) - \varepsilon\nu(\cdot)]$ with probability $1-\varepsilon$.
   The intermediate steps $X_{n+1}, \ldots, X_{n+n_0-1}$ and $X'_{n+1}, \ldots, X'_{n+n_0-1}$ are filled from the correct conditional distributions. Increase n by n_0.
3. Else, conditionally independently draw $X_{n+1} \sim P(X_n, \cdot)$ and $X'_{n+1} \sim P(X'_n, \cdot)$. Increase n by 1.

It would not be hard to notice that X_n and X'_n are marginally updated according to transition kernel P. Furthermore, since $X_0 = x$ and $X'_0 \sim \pi(\cdot)$, we have $X_n \sim P^n(x, \cdot)$ and $X'_n \sim \pi(\cdot)$. So, using the coupling inequality on X_n and X'_n, we have

$$\|P^n(x, \cdot) - \pi(\cdot)\| \leq \mathbb{P}(X_n \neq X'_n).$$

The probability on the right will be pretty small using our construction. The second step utilises the minorisation condition. And when we are outside C and with the two random variables not being equal, the drift condition (if applicable) ensures that X_n and X'_n will return to C very quickly.

Let us have a quick taste of the power of this construction by proving the uniform ergodicity result.

Proof. (of Uniform Ergodicity (Theorem 2.9)) Since we know $C = \mathcal{X}$ from the conditions of the theorem, the coupling construction will only use Step 1 and Step 2. Step 1 ensures that once the two random variables are the same, they will stay the same. Step 2 (a) ensures that, for every n_0 steps, the two random variables will be the same (drawn from $\nu(\cdot)$) with a probability of at least ε, as they could have been equal to start with. This means, if we have $n = n_0 m$ for some integer m, then we have

$$\|P^n(x, \cdot) - \pi(\cdot)\| \leq \mathbb{P}(X_n \neq X'_n) \leq (1-\varepsilon)^m = (1-\varepsilon)^{n/n_0}.$$

Now, n might not always be a multiple of n_0. For some m and $n \in (mn_0, (m_1)n_0 - 1]$, we have

$$\|P^n(x, \cdot) - \pi(\cdot)\| \leq \|P^{mn_0}(x, \cdot) - \pi(\cdot)\|$$

using Proposition 2.2 (c). Therefore, we have the desired inequality

$$\|P^n(x, \cdot) - \pi(\cdot)\| \leq (1-\varepsilon)^{\lfloor n/n_0 \rfloor}$$

for all n. □

2.3.2 Proof of Geometric Ergodicity (Theorem 2.12)

We will prove here the geometric ergodicity result. Notice that since we no longer have the condition that $C = \mathcal{X}$, we will inevitably use Step 3 of the coupling construction. The idea is that, when X_n and X'_n are not equal and not both in C, they will return to C quickly.

Recall from earlier the drift condition: A Markov chain is said to satisfy a **drift condition** if there are constants $\lambda \in (0,1)$, $b < \infty$, and a Lyapunov function $V : \mathcal{X} \to [1, \infty]$ such that

$$PV \leq \lambda V + b1_C$$

i.e. $\int_{\mathcal{X}} P(x, dy)V(y) \leq \lambda V(x) + b1_C(x)$ for all $x \in \mathcal{X}$.

We will require a modified condition. Consider $\overline{P}$ with

$$\overline{P}h(x, y) = \int_{\mathcal{X}} \int_{\mathcal{X}} h(z, w)P(x, dz)P(y, dw)$$

which represents the kernel of running two independent copies of the chain with transition kernel P. The **bivariate drift condition** states that

$$\overline{P}h(x, y) \leq \frac{h(x, y)}{\alpha}$$

for $(x, y) \notin C \times C$, some $h : \mathcal{X} \times \mathcal{X} \to [1, \infty)$ and some $\alpha > 1$.

This bivariate drift condition is closely related to the drift condition.

Proposition 2.15. *Suppose the drift condition is satisfied for some $V : \mathcal{X} \to [1, \infty]$, $C \subseteq \mathcal{X}$, $\lambda < 1$, and $b < \infty$. Let $d := \inf_{x \in \mathcal{X} \setminus C} V(x)$. Then, if*

$$d > \frac{b}{1 - \lambda} - 1,$$

then the bivariate drift condition is satisfied for the same C with $h(x, y) = [V(x) + V(y)]/2$ and $\alpha^{-1} = \lambda + b/(d + 1) < 1$.

Proof. If $(x, y) \notin C \times C$, then we have either $x \notin C$ or $y \notin C$ or both. So, $h(x, y) = [V(x) + V(y)]/2 \geq (d + 1)/2$ as $V \geq 1$, which means $1 \leq 2h(x, y)/(d + 1)$. Furthermore, using the drift condition, we have

$$PV(x) + PV(y) \leq \lambda V(x) + \lambda V(y) + b$$

as at most one of x and y is not in C. Then, we have

$$\begin{aligned}
\overline{P}h(x, y) &= \int_{\mathcal{X}} \int_{\mathcal{X}} \frac{1}{2}[V(x) + V(y)]P(x, dz)P(y, dw) \\
&= \frac{1}{2}\left[\int_{\mathcal{X}} \int_{\mathcal{X}} V(x)P(x, dz)P(y, dw) + \int_{\mathcal{X}} \int_{\mathcal{X}} V(y)P(x, dz)P(y, dw)\right] \\
&= \frac{1}{2}\left[\int_{\mathcal{X}} V(x)P(x, dz) + \int_{\mathcal{X}} V(y)P(y, dw)\right] \\
&= \frac{1}{2}[PV(x) + PV(y)] \leq \frac{1}{2}[\lambda V(x) + \lambda V(y) + b] \\
&= \lambda\left[\frac{1}{2}V(x) + \frac{1}{2}V(y)\right] + \frac{b}{2} = \lambda h(x, y) + \frac{b}{2} \\
&\leq \lambda h(x, y) + \frac{b}{2}2h(x, y)/(d + 1) = h(x, y)[\lambda + b/(d + 1)].
\end{aligned}$$

We are almost there. The final thing is to check if we really have $\lambda + b/(d+1) < 1$. We have

$$\begin{aligned} d &> \frac{b}{1-\lambda} - 1 \\ d+1 &> \frac{b}{1-\lambda} \\ 1-\lambda &> \frac{b}{d+1} \\ 1 &> \lambda + \frac{b}{d+1}, \end{aligned}$$

as desired. □

Recall that in Step 2(b) of the coupling construction, X_n and X'_n are drawn from

$$R(X_n, \cdot) := \frac{1}{1-\varepsilon}[P^{n_0}(X_n, \cdot) - \varepsilon\nu(\cdot)].$$

We define $\overline{R}$ similar to that of $\overline{P}$. For $h(x, y) : \mathcal{X} \times \mathcal{X} \to [1, \infty)$, we have

$$\overline{R}h(x, y) := \int_{\mathcal{X}} \int_{\mathcal{X}} h(z, w) R(x, dz) R(y, dw)$$

for any $(x, y) \in C \times C$. We further define

$$B_{n_0} := \max\left[1, \alpha^{n_0}(1-\varepsilon) \sup_{C\times C} \overline{R}h\right].$$

Theorem 2.16. *Consider a Markov chain on state space $\mathcal{X}$ with transition kernel P. Suppose the drift condition and the bivariate drift condition hold for some $C \subseteq \mathcal{X}$, $\alpha > 1$, $n_0 \in \mathbb{N}$, and $\varepsilon > 0$. Then, for any joint initial distribution $L(X_0, X'_0)$ and any integer $j \in [1, k]$, if $\{X_n\}$ and $\{X'_n\}$ are two copies of the Markov chain starting from $L(X_0, X'_0)$, then*

$$\|L(X_k) - L(X'_k)\| \leq (1-\varepsilon)^j + \alpha^{-k}(B_{n_0})^{j-1}\mathbb{E}[h(X_0, X'_0)].$$

Remark. *Since $j \leq k$, if we pick a small enough $r > 0$ and let $j = \lfloor rk \rfloor$, then the bound would be exponentially decaying.*

Proof. This theorem is proved using the coupling inequality and the coupling construction as well.

We first assume $n_0 = 1$ for the minorisation condition and prove the theorem under this additional condition. The modification required for the case $n_0 > 1$ is outlined in Section 4.4 of Roberts and Rosenthal (2004).

Let $N_k := \#\{m \ : \ 0 \leq m \leq k, (X_m, X'_m) \in C \times C\}$ be the number of times that the pair of chains visit C between time 0 and time k. Then, coupling inequality gives us

$$\|P^k(x, \cdot) - \pi(\cdot)\| \leq \mathbb{P}(X_k \neq X'_k) = \mathbb{P}(X_k \neq X'_k, N_{k-1} \geq j) + \mathbb{P}(X_k \neq X'_k, N_{k-1} < j).$$

A bound for each of the two terms is thus needed. It is easy to bound the first term. Notice that going for Step 2(b), instead of Step 2(a), for j times straight from the start implies that we have both $X_k \neq X'_k$ and $N_{k-1} \geq j$. The probability of going for Step 2(b) is $1 - \varepsilon$, so we have

$$\mathbb{P}(X_k \neq X'_k, N_{k-1} \geq j) \leq (1-\varepsilon)^j.$$

Now we bound the second term. Let $M_k := \alpha^k B^{-N_{k-1}} h(X_k, X'_k) 1_{\{X_k \neq X'_k\}}$ for $k = 0, 1, 2, \ldots$. We set $N_{-1} = 0$. We claim that $\{M_k\}$ is a supermartingale, i.e. we have

$$\mathbb{E}[M_{k+1} \mid X_0, \ldots, X_k, X'_0, \ldots, X'_k] \leq M_k.$$

A proof of this result can be found in Lemma 13 of Roberts and Rosenthal (2004). With this, note that since $B \geq 1$, we have

$$\begin{aligned}
\mathbb{P}(X_k \neq X'_k, N_{k-1} < j) &= \mathbb{P}(X_k \neq X'_k, N_{k-1} \leq j-1) \\
&\leq \mathbb{P}(X_k \neq X'_k, B^{-N_{k-1}} \leq B^{-(j-1)}) \\
&= \mathbb{P}(1_{\{X_k \neq X'_k\}} B^{-N_{k-1}} \leq B^{-(j-1)}) \\
&\leq B^{(j-1)} \mathbb{E}(1_{\{X_k \neq X'_k\}} B^{-N_{k-1}}) \qquad \text{Markov inequality} \\
&\leq B^{(j-1)} \mathbb{E}(1_{\{X_k \neq X'_k\}} B^{-N_{k-1}} h(X_k, X'_k) \qquad h \geq 1 \\
&= B^{(j-1)} \alpha^{-k} \mathbb{E}[M_k] \\
&\leq B^{(j-1)} \alpha^{-k} \mathbb{E}[M_0] \qquad \{M_k\} \text{ supermartingale} \\
&= B^{(j-1)} \alpha^{-k} \mathbb{E}[h(X_0, X'_0)].
\end{aligned}$$

So, with the extra condition that $n_0 = 1$, we have

$$\begin{aligned}
\|P^k(x, \cdot) - \pi(\cdot)\| &\leq \mathbb{P}(X_k \neq X'_k, N_{k-1} \geq j) + \mathbb{P}(X_k \neq X'_k, N_{k-1} < j) \\
&\leq (1-\varepsilon)^j + B^{(j-1)} \alpha^{-k} \mathbb{E}[h(X_0, X'_0)],
\end{aligned}$$

as desired for the theorem. □

So, if we would like to prove the geometric ergodicity (Theorem 2.12), we just need to make sure the conditions for quantitative convergence bound (Theorem 2.16) are satisfied when we have both the drift and minorisation conditions.

Notice that the difference between the conditions for geometric ergodicity and quantitative convergence bound is the additional bivariate drift condition. The bivariate drift condition can be obtained from the (univariate) drift condition with an extra condition that $d := \inf_{x \in \mathcal{X} \setminus C} V(x) > b/(1-\lambda) - 1$, where b, λ are constants of the (univariate) drift condition. This is established in Proposition 2.15.

However, we could not prove that this extra condition of d always holds. Extra work is needed for the case where $d < b/(1-\lambda) - 1$, and we direct the readers to Roberts and Rosenthal (2004) for a proof of that.

Part II

Langevin Algorithms

Chapter 3

Unadjusted Langevin Algorithm

Numerical solutions of differential equations have been extensively studied in the mathematical community. One of the early methods to approximate the solution of an ordinary differential equation (ODE) is that of the **Euler method** (Sauer, 2011). Consider the following ODE with initial condition

$$\frac{d}{dt}y = f(t, y), \qquad y(0) = y_0,$$

then the Euler method will yield the approximate solutions w_i at time t_i with these values defined by

$$\begin{aligned} w_0 &= y_0, \quad t_0 = 0, \\ w_{i+1} &= w_i + hf(t_i, w_i), \quad t_{i+1} = t_i + h \end{aligned}$$

for $i = 0, 1, 2, \ldots$. Here, h is a tuning parameter and it denotes the step size of the update. Naturally, the approximation will be better for smaller values of h.

When we have an SDE instead of an ODE, we will have both the deterministic drift term and a stochastic diffusion term. The scheme to approximate the solution of an SDE, therefore, needs to be adjusted. This is known as the **Euler-Maruyama Method** (Sauer, 2011). Consider the following SDE with initial condition

$$dX_t = a(X_t, t)dt + b(X_t, t)dW_t, \qquad X_0 = x_0,$$

where a is the drift function, b is the volatility function, and $\{W_t\}$ is a Wiener process, then the Euler-Maruyama method will yield the following approximate solutions w_i at time t_i with these values defined by

$$\begin{aligned} w_0 &= x_0, \quad t_0 = 0, \\ w_{i+1} &= w_i + ha(t_i, w_i) + z_i b(t_i, w_i), \quad t_{i+1} = t_i + h \end{aligned}$$

where $z_i \sim N(0, h)$ are i.i.d. noises and h is the tuning parameter denoting the step size.

So, using the Euler-Maruyama method, we can approximate the k-dimensional Langevin diffusion

$$dL_t = \frac{1}{2}\nabla \log \pi(L_t)dt + dB_t$$

with initial value x_0 using step size h :

$$U_0 = x_0$$
$$U_{n+1} = U_n + \frac{h}{2}\nabla \log \pi(U_n) + \epsilon_n, \quad \epsilon_n \sim N(0, hI_k)$$

for $n = 0, 1, 2, \ldots$. In fact, we could simply write $U_{n+1} \sim N(U_n + \frac{h}{2}\nabla \log \pi(U_n), hI_k)$. This yields a sequence $\{U_n\}$ which can be used as the output of the MCMC algorithm.

In the Statistics literature, this method is also known as the **Unadjusted Langevin Algorithm** (ULA), since this algorithm is essentially MALA but without the Metropolis adjustment step, i.e. every proposal is accepted. In the machine learning literature, this method is sometimes referred to as the Langevin Monte Carlo (LMC). Here, we will use ULA to address this algorithm.

Extensive research has been conducted on ULA to study its various theoretical properties. Roberts and Tweedie (1996a) studied the rate of convergence (if at all) of the approximation to the target distribution. Dalalyan (2017) obtained a non-asymptotic bound on the convergence of the approximation of ULA samples, assuming that the target distributions are smooth and log-concave. Durmus and Moulines (2019) provided further theoretical results on the convergence, as well as proposed a decaying over dimension scheme for the selection of tuning parameter h of the algorithm in order to have good convergence properties.

3.1 Geometric Ergodicity of ULA

Here, we will show the geometric ergodicity of ULA for a specific choice of the target distribution.

Proposition 3.1. *The Markov chain produced by ULA with one-dimensional target distribution $\pi \propto \exp(-x^2/2)$ and tuning parameter h with $h^2 < 2$ is geometric ergodic.*

Proof. By Proposition 2.13, we just need to check for

$$\limsup_{|x|\to\infty} \frac{PV(x)}{V(x)} = \limsup_{|x|\to\infty} \int \frac{V(y)}{V(x)} P(x, dy) < 1.$$

Consider $V(x) = e^{s|x|}$ for some $s > 0$. We have, where $\xi \sim N(0, h^2)$,

$$\begin{aligned}
\int \frac{V(y)}{V(x)} P(x, dy) &= \mathbb{E}_\xi \left[\frac{\exp[s|x + h^2/2(-x) + \xi|]}{e^{s|x|}} \right] \\
&= \mathbb{E}_\xi \left[\exp[s|x - h^2/2x + \xi| - s|x|] \right] \\
&\leq \mathbb{E}_\xi \left[\exp[s|x - h^2/2x| + s|\xi| - s|x|] \right] \\
&= \mathbb{E}_\xi \left[\exp[s|1 - h^2/2||x| + s|\xi| - s|x|] \right] \\
&= \mathbb{E}_\xi \left[e^{s|\xi|} \right] \exp[s|1 - h^2/2||x| - s|x|] \\
&= \mathbb{E}_\xi \left[e^{s|\xi|} \right] \exp[-sh^2/2|x|].
\end{aligned}$$

Notice that $\mathbb{E}_\xi[e^{s|\xi|}]$ is a positive constant for any fixed s, and $\exp[-sh^2/2|x|] \to 0$ as $|x| \to \infty$. Therefore, for large enough $|x|$, we have

$$\limsup_{|x|\to\infty} \int \frac{V(y)}{V(x)} P(x, dy) = \lim_{|x|\to\infty} \int \frac{V(y)}{V(x)} P(x, dy) = 0 < 1,$$

as desired. □

We also have the following more general geometric ergodicity result.

For some fix d, we define

$$S_d^+ := \lim_{x\to\infty} \frac{h}{2}\nabla \log \pi(x) x^{-d}$$

and

$$S_d^- := \lim_{x\to-\infty} \frac{h}{2}\nabla \log \pi(x) |x|^{-d}$$

Theorem 3.2 (Theorem 3.1 of Roberts and Tweedie (1996a)). *The ULA chain U_n is geometrically ergodic if one of the following holds:*

1. *for some $d \in [0, 1)$, both $S_d^+ < 0$ and $S_d^- > 0$ exist.*
2. *for $d = 1$, both $S_d^+ < 0$ and $S_d^- > 0$ exist, and $(1 + S_d^+)(1 - S_d^-) < 1$.*

In the above special case, we have $d = 1$, $S_d^+ = -h/2$, and $S_d^- = h/2$. Given $h^2 < 2$, we would certainly have $(1 + S_d^+)(1 - S_d^-) < 1$.

Chapter 4

ULA Convergence Rate for Smooth and Log-Concave Targets

4.1 Motivation

Existing convergence bounds of MCMC algorithms, at least before Dalalyan (2017), tend to be of two types: asymptotic bounds that depend on dimension derived from scaling arguments, and relatively loose non-asymptotic geometric ergodicity bounds using the drift and minorisation argument (Qin and Hobert, 2021). We would like to have tighter bounds in order to better understand and assess the quality of the various MCMC algorithms.

The non-asymptotic bounds using the drift and minorisation argument impose mild conditions on the target distribution. As a result, the bounds obtained are loose, especially when the dimension of the target distribution is high (Qin and Hobert, 2021). One could obtain much tighter bounds by imposing stronger conditions on the target distribution, and this is the approach taken in Dalalyan (2017).

Dalalyan (2017) borrowed insights from the (convex) optimisation literature and imposed smoothness conditions on the target distribution. Later works on the convergence bounds of MCMC algorithms, such as Dwivedi et al. (2018) and Andrieu et al. (2022), followed the same trend of imposing strong conditions on the target distribution.

In Dalalyan (2017), the author focused on one particular MCMC algorithm, the Unadjusted Langevin Algorithm (ULA), and aimed to study its quality when we try to sample from distributions with certain good and relatively realistic properties. By the quality of ULA, we really mean the rate of convergence of the approximation (in the form of the total variation distance between the sample and the target). In this chapter, we will summarise of the results of Dalalyan (2017) by highlighting the key results and completing with expanded and re-written proofs for clarity.

The ULA approximation takes two steps. First, we will approximate the target distribution using a continuous-time Langevin diffusion. Second, we will approximate the continuous-time Langevin diffusion using a discretised Langevin diffusion. The following diagram illustrates this

relationship.

Target Distribution $\longleftrightarrow$ Distribution of L. Diffusion $\longleftrightarrow$ Distribution of Discretised L.

The overall approximation error is bounded by the sum of two bounds from each step.

4.2 Setup

We assume the densities of target distribution π are of the form e^{-f} where f is a function that (i) is strongly convex, and (ii) has a Lipschitz continuous gradient. That is, for $f : \mathbb{R}^n \to \mathbb{R}$, there exists two positive constants m and M such that

$$\begin{cases} f(\theta) - f(\bar{\theta}) - \nabla f(\bar{\theta})^T(\theta - \bar{\theta}) \geq \frac{m}{2}||\theta - \bar{\theta}||_2^2, \\ ||\nabla f(\theta) - \nabla f(\bar{\theta})||_2 \leq M||\theta - \bar{\theta}||_2 \end{cases} \tag{4.1}$$

for all $\theta, \bar{\theta} \in \mathbb{R}^n$. We will call f convex and $\exp(-f)$ strongly log-concave if f satisfies the first inequality above, and $\exp(-f)$ simply as log-concave if it satisfies it with $m = 0$.

At this point, we establish a Lemma that will be used later on.

Lemma 4.1 (Lemma 1.2.3 in Nesterov (2003))**.** *If the function f satisfies the second inequality in (4.1), then*

$$f(\theta) - f(\bar{\theta}) - \nabla f(\bar{\theta})^\top(\theta - \bar{\theta}) \leq \frac{M}{2}||\theta - \bar{\theta}||_2^2$$

for all $\theta, \bar{\theta} \in \mathbb{R}^n$.

There are two notions of divergence that will be used later on. For two probability measures μ and ν defined on a space $\mathcal{X}$ such that μ is absolutely continuous with respect to ν (i.e. for any set E with $\nu(E) = 0$ we have $\mu(E) = 0$. This enables us to define Radon-Nikodym derivatives), the Kullback-Leibler and χ^2 divergences between μ and ν are respectively defined as (Dalalyan, 2017)

$$KL(\mu\|\nu) = \int_{\mathcal{X}} \log\left(\frac{d\mu}{d\nu}(x)\right)\mu(dx), \qquad \chi^2(\mu\|\nu) = \int_{\mathcal{X}} \left(\frac{d\mu}{d\nu}(x) - 1\right)^2 \nu(dx).$$

4.3 Error 1

The unadjusted Langevin algorithm (ULA) is similar to gradient descent, but involves an additional step of random perturbation.

Starting from an initial position $\theta^{(0)} \in \mathbb{R}^n$, the subsequent steps are updated based on the following update rule:

$$\theta^{(k+1,h)} = \theta^{(k,h)} - h\nabla f(\theta^{(k,h)}) + \sqrt{2h}\xi^{(k+1)} \tag{4.2}$$

for $k = 0, 1, 2, \cdots$, $h > 0$ be the tuning parameter that is often referred to as the step-size, and a sequence of i.i.d. $\xi^{(0)}, \xi^{(1)} \cdots$ that are multivariate standard normal and independent of $\theta^{(0)}$. The target distribution has a density proportional to $\exp(-f)$.

Recall that ULA is the Euler discretisation of the Langevin diffusion $\{L_t : t \in \mathbb{R}^+\}$ with invariant density π and characterised by the SDE

$$dL_t = -\nabla f(L_t)\, dt + \sqrt{2}dW_t. \tag{4.3}$$

When f satisfies condition (4.1), equation (4.3) has a unique strong solution which is a Markov process. The transition kernel of this process is denoted by $P_L^t(x,\cdot)$ with $P_L^t(x,A) = P(L_t \in A | L_0 = x)$ for all Borel sets $A \subset \mathbb{R}^n$ and any initial condition $x \in \mathbb{R}^n$. These results are shown in Roberts and Tweedie (1996a).

Condition (4.1) yields the spectral gap property of the semigroup $\{P_L^t : t \in \mathbb{R}^+\}$, which then implies that the process L_t is geometrically ergodic in the following sense:

Lemma 4.2. *Under condition 4.1, for any probability density ν,*

$$||\nu P_L^t - \pi||_{TV} \leq \frac{1}{2}\chi^2(\nu||\pi)^{1/2} e^{-tm/2}, \qquad \forall t \geq 0.$$

Remark. *This provides the bound for the first estimation error, between the target distribution and the distribution of $\{L_t\}$, as described at the beginning of this chapter. Also, $\|\cdot\|_{TV}$ is the total variation distance.*

Proof. First, notice that under condition (4.1), the process $\{L_t\}$ is geometrically ergodic in $L^2(\mathbb{R}^n, \pi)$, that is:

$$\int_{\mathbb{R}^n} (\mathbb{E}[\varphi(L_t)|L_0 = x] - \mathbb{E}_\pi[\varphi(\theta)])^2 \, d\pi \leq e^{-tm}\mathbb{E}_\pi[\varphi^2(\theta)]$$

for every $t > 0$ and every $\varphi \in L^2(\mathbb{R}^n, \pi)$. This is a well-known result (Dalalyan, 2017).

By the definition of the total variation distance and the fact that π is the invariant density of the semigroup $\{P_L^t\}$, we have

$$\begin{aligned}
||\nu P_L^t - \pi||_{\text{TV}} &= \sup_A \Big| \int_{\mathbb{R}^n} P_L^t(x,A) \, d\nu - \pi(A) \Big| \\
&= \sup_A \Big| \int_{\mathbb{R}^n} P_L^t(x,A) \, d\nu - \int_{\mathbb{R}^n} \pi(A) \, d\nu \Big| \\
&= \sup_A \Big| \int_{\mathbb{R}^n} (P_L^t(x,A) - \pi(A)) \, d\nu \Big| \\
&= \sup_A \Big| \int_{\mathbb{R}^n} (P_L^t(x,A) - \pi(A))\nu(x) \, dx - \int_{\mathbb{R}^n} (P_L^t(x,A)\pi(x) - \pi(A)\pi(x)) \, dx \Big| \\
&\quad \textit{since } \pi \textit{ is the invariant distribution of } P_L^t \\
&= \sup_A \Big| \int_{\mathbb{R}^n} (P_L^t(x,A) - \pi(A))(\nu(x) - \pi(x)) \, dx \Big| \\
&\leq \sup_A \int_{\mathbb{R}^n} \Big| P_L^t(x,A) - \pi(A) \Big| \Big| \frac{\nu(x)}{\pi(x)} - 1 \Big| \pi(x) \, dx \\
&\quad \textit{using integral form of triangle inequality} \\
&\leq \sup_A \Big(\int_{\mathbb{R}^n} \Big| P_L^t(x,A) - \pi(A) \Big|^2 \pi(x) \, dx \Big)^{1/2} \chi^2(\nu||\pi)^{1/2} \\
&\quad \textit{using Cauchy-Schwartz inequality.}
\end{aligned}$$

Here, the supremum is taken over all possible elements of the Borel set of $\mathbb{R}^n$.

Next, for every fixed Borel set A, if we set $\varphi(x) = \mathbf{1}_A(x) - \pi(A)$ and use the first inequality of this proof, we get

$$\begin{aligned}\int_{\mathbb{R}^n} \left|P_L^t(x,A) - \pi(A)\right|^2 \pi(x)\, dx &= \int_{\mathbb{R}^n} (\mathbb{E}[\varphi(L_t)|L_0 = x] - \mathbb{E}_\pi[\varphi(\theta)])^2\, d\pi \\ &\leq e^{-tm}\mathbb{E}_\pi[\varphi^2(\theta)] \\ &= e^{-tm}\pi(A)(1-\pi(A)) \\ &\leq \frac{1}{4}e^{-tm}.\end{aligned}$$

Combining what we have shown so far, we get

$$\begin{aligned}||\nu P_L^t - \pi||_{TV} &\leq \sup_A \Big(\int_{\mathbb{R}^n} \left|P_L^t(x,A) - \pi(A)\right|^2 \pi(x)\, dx\Big)^{1/2} \chi^2(\nu||\pi)^{1/2} \\ &\leq \sup_A \left(\frac{1}{4}e^{-tm}\right)^{1/2} \chi^2(\nu||\pi)^{1/2} \\ &= \frac{1}{2}\chi^2(\nu||\pi)^{1/2} e^{-tm/2},\end{aligned}$$

as desired. □

Lemma 4.2 shows that for large values of t, the distribution of L_t approaches exponentially fast to the target distribution π. Due to this, ULA aims to then approximate P_L^t by $\theta^{(k,h)}$ where $t = kh$, in order to reach π.

–

The first and probably the most influential work providing probabilistic analysis of asymptotic properties of the ULA is Roberts and Tweedie (1996a). However, one of the recommendations made by the authors of that paper is that the ergodicity of the Markov chain generated by the Langevin algorithm is very sensitive to the choice of the tuning parameter h and a bad choice will lead to the transience of the chain. However, as we will show in the following Proposition 4.3, under the strong convexity assumption on f as well as the Lipschitz continuity of its gradient, we can ensure the non-transience of the Markov chain $\{\theta^{(k,h)}\}$ as long as $h \leq 1/M$.

Proposition 4.3. *Let the function f be continuously differentiable on $\mathbb{R}^n$ and satisfy (4.1) with $f^* = \inf_{x\in\mathbb{R}^n} f(x)$. Then, for every $h \leq 1/M$, we have*

$$\mathbb{E}[f(\theta^{(k,h)}) - f^*] \leq (1-mh)^k \mathbb{E}[f(\theta^{(0)}) - f^*] + \frac{Mn}{m}.$$

This result is implied by a stronger result that we will prove next.

Proposition 4.4. *Let the function f be continuously differentiable on $\mathbb{R}^n$ and satisfy (4.1) with $f^* = \inf_{x\in\mathbb{R}^n} f(x)$. Then, for every $h \leq 1/M$, we have*

$$\mathbb{E}[f(\theta^{(k,h)}) - f^*] \leq (1-mh)^k \mathbb{E}[f(\theta^{(0)}) - f^*] + \frac{Mn}{m(2-Mh)}.$$

Proof. To simplify the proof, we will use the shorthand notation $f^{(k)} = f(\theta^{(k,h)})$ and $\nabla f^{(k)} = \nabla f(\theta^{(k,h)})$. Due to strong convexity of f, we have (Boyd and Vandenberghe (2004), p459)

$$f^{(k+1)} = f^{(k)} + (\nabla f^{(k)})^\top(\theta^{(k+1,h)} - \theta^{(k,h)}) + \frac{1}{2}(\theta^{(k+1,h)} - \theta^{(k,h)})^\top \nabla^2 f(z)(\theta^{(k+1,h)} - \theta^{(k,h)})$$

for some z between $\theta^{(k,h)}$ and $\theta^{(k+1,h)}$.

Next, (4.1) implies that $\nabla^2 f^{(x)} \preceq MI$, which means the above equality becomes

$$f^{(k+1)} \leq f^{(k)} + (\nabla f^{(k)})^\top(\theta^{(k+1,h)} - \theta^{(k,h)}) + \frac{1}{2}M||\theta^{(k+1,h)} - \theta^{(k,h)}||_2^2,$$

where $\|\cdot\|_2$ is the L^2 / Euclidean norm.

Using the ULA update rule $\theta^{(k+1,h)} = \theta^{(k,h)} - h\nabla f^{(k)} + \sqrt{2h}\ \varepsilon^{(k+1)}$ where the εs are standard Gaussian, we have

$$\begin{aligned} f^{(k+1)} &\leq f^{(k)} + (\nabla f^{(k)})^\top(-h\nabla f^{(k)} + \sqrt{2h}\ \varepsilon^{(k+1)}) + \frac{1}{2}M||-h\nabla f^{(k)} + \sqrt{2h}\ \varepsilon^{(k+1)}||_2^2 \\ &= f^{(k)} - h||\nabla f^{(k)}||_2^2 + \sqrt{2h}(\nabla f^{(k)})^\top\varepsilon^{(k+1)} + \frac{1}{2}M||h\nabla f^{(k)} - \sqrt{2h}\ \varepsilon^{(k+1)}||_2^2. \end{aligned}$$

Taking expectation on both sides, we get

$$\mathbb{E}(f^{(k+1)}) \leq \mathbb{E}(f^{(k)}) - h\mathbb{E}[||\nabla f^{(k)}||_2^2] + 0 + \frac{M}{2}\mathbb{E}[||h\nabla f^{(k)} - \sqrt{2h}\ \varepsilon^{(k+1)}||_2^2].$$

If we focus on the expectation of the last term in the above inequality, we have

$$\begin{aligned} \mathbb{E}[||h\nabla f^{(k)} - \sqrt{2h}\ \varepsilon^{(k+1)}||_2^2] &= \mathbb{E}[(h\nabla f^{(k)} - \sqrt{2h}\ \varepsilon^{(k+1)})^\top(h\nabla f^{(k)} - \sqrt{2h}\ \varepsilon^{(k+1)})] \\ &= \mathbb{E}[h^2(\nabla f^{(k)})^\top\nabla f^{(k)} - (\nabla f^{(k)})^\top\sqrt{2h}\ \varepsilon^{(k+1)}) - (\sqrt{2h}\ \varepsilon^{(k+1)})^\top\nabla f^{(k)} \\ &\quad + 2h\ \varepsilon^{(k+1)}\varepsilon^{(k+1)})] \\ &= h^2\mathbb{E}[||\nabla f^{(k)}||_2^2] + 2hn \end{aligned}$$

The last term is $2hn$ since the variance of each ε is 1 and there are n of them.

With this, we get

$$\begin{aligned} \mathbb{E}(f^{(k+1)}) &\leq \mathbb{E}(f^{(k)}) - h\mathbb{E}[||\nabla f^{(k)}||_2^2] + 0 + \frac{M}{2}\mathbb{E}[||h\nabla f^{(k)} - \sqrt{2h}\ \varepsilon^{(k+1)}||_2^2] \\ &\leq \mathbb{E}(f^{(k)}) - h\mathbb{E}[||\nabla f^{(k)}||_2^2] + \frac{M}{2}(h^2\mathbb{E}[||\nabla f^{(k)}||_2^2] + 2hn) \\ &= \mathbb{E}(f^{(k)}) - (h - \frac{h^2M}{2})\mathbb{E}[||\nabla f^{(k)}||_2^2] + Mhn. \end{aligned}$$

Due to strong convexity of f, we have (Boyd and Vandenberghe (2004), p460) $f^* - f(x) \geq -\frac{1}{2m}||\nabla f(x)||_2^2$ and thus

$$||\nabla f^{(k)}||_2^2 \geq 2m(f^{(k)} - f^*).$$

If we set $x = \theta^{(k,h)}$ of the above inequality and apply it to the inequality of $\mathbb{E}(f^{(k+1)})$, we get, when $h - h^2M/2 > 0$, or equivalently $h < 2/M$,

$$\begin{aligned} \mathbb{E}(f^{(k+1)}) &\leq \mathbb{E}(f^{(k)}) - (h - \frac{h^2M}{2})\mathbb{E}[||\nabla f^{(k)}||_2^2] + Mhn \\ &\leq \mathbb{E}(f^{(k)}) - (2mh - mh^2M)\mathbb{E}[(f^{(k)} - f^*)] + Mhn. \end{aligned}$$

Note that the condition on h is satisfied with the conditions of this proposition.

We set $\gamma := 2mh - mh^2M \in (0,1)$ for any $0 < h < 2/M$. Subtracting f^* from both sides of the above inequality gives us

$$\begin{aligned}
\mathbb{E}[f^{(k+1)}] - f^* &\leq E(f^{(k)}) - f^* - \gamma\mathbb{E}[f(x) - f^*] + Mhn \\
\mathbb{E}[f^{(k+1)} - f^*] &\leq E(f^{(k)} - f^*) - \gamma\mathbb{E}[f(x) - f^*] + Mhn \\
\mathbb{E}[f^{(k+1)} - f^*] &\leq (1-\gamma)\mathbb{E}[f^{(k)} - f^*] + Mhn.
\end{aligned}$$

Applying the inequality recursively gives us

$$\begin{aligned}
\mathbb{E}[f^{(k+1)} - f^*] &\leq (1-\gamma)\mathbb{E}[f^{(k)} - f^*] + Mhn \\
&\leq (1-\gamma)^2\mathbb{E}[f^{(k-1)} - f^*] + Mhn(1-\gamma) + Mhn \\
&\cdots \\
&\leq (1-\gamma)^{k+1}\mathbb{E}[f^{(0)} - f^*] + Mhn(1 + (1-\gamma) + \cdots + (1-\gamma)^k) \\
&\leq (1-\gamma)^{k+1}\mathbb{E}[f^{(0)} - f^*] + Mhn(\sum_{i=0}^{\infty}(1-\gamma)^i) \\
&\leq (1-\gamma)^{k+1}\mathbb{E}[f^{(0)} - f^*] + Mhn\gamma^{-1}.
\end{aligned}$$

If we swap $\gamma := 2mh - mh^2M$ back, we would thus get

$$\begin{aligned}
\mathbb{E}[f(\theta^{(k,h)}) - f^*] &\leq (1 - 2mh + mh^2M)^k\mathbb{E}[f^{(0)} - f^*] + \frac{Mhn}{2mh - mh^2M} \\
&\leq (1-mh)^k\mathbb{E}[f^{(0)} - f^*] + \frac{Mn}{m(2-Mh)}
\end{aligned}$$

since $h < 1/M$ as given in the proposition. The proof is completed. □

This proposition has a corollary, which will be used in the next section.

Corollary 4.5. *Let $h \leq 1/(\alpha M)$ with $\alpha \geq 1$ and integer $K \geq 1$. Under the conditions of Proposition 4.4, it holds that*

$$h\sum_{k=0}^{K-1}\mathbb{E}[||\nabla f(\theta^{(k,h)})||_2^2] \leq \frac{M\alpha}{2\alpha - 1}\mathbb{E}[||\theta^{(0)} - \theta^*||_2^2] + \frac{2\alpha MKhn}{2\alpha - 1}.$$

Proof. From the condition on h, we can get the following inequality after some simple manipulations.

$$\begin{aligned}
h &\leq \frac{1}{\alpha M} \\
Mh &\leq \frac{1}{\alpha} \\
2 - Mh &\geq 2 - \frac{1}{\alpha} = \frac{2\alpha - 1}{\alpha}
\end{aligned}$$

Then, using the inequality obtained in the proof of Proposition 4.4

$$\mathbb{E}(f^{(k+1)}) \leq \mathbb{E}(f^{(k)}) - (h - \frac{h^2M}{2})\mathbb{E}[||\nabla f^{(k)}||_2^2] + Mhn,$$

and the inequality with h obtained recently, we have

$$\begin{aligned}
\mathbb{E}(f^{(k+1)}) &\leq \mathbb{E}(f^{(k)}) - (h - \frac{h^2 M}{2})\mathbb{E}[||\nabla f^{(k)}||_2^2] + Mhn \\
\mathbb{E}[f^{(k)} - f^{(k+1)}] + Mhn &\geq \frac{1}{2}h(2 - Mh)\mathbb{E}[||\nabla f^{(k)}||_2^2] \\
&\geq \frac{1}{2}h\frac{2\alpha - 1}{\alpha}\mathbb{E}[||\nabla f^{(k)}||_2^2] \\
&= \frac{h(2\alpha - 1)}{2\alpha}\mathbb{E}[||\nabla f^{(k)}||_2^2]
\end{aligned}$$

for all $k \in \mathbb{N}$.

If we sum up $k = 0, 1, \cdots, K-1$, we get

$$\begin{aligned}
\sum_{k=0}^{K-1} \frac{h(2\alpha - 1)}{2\alpha}\mathbb{E}[||\nabla f^{(k)}||_2^2] &\leq \sum_{k=0}^{K-1} \mathbb{E}[f^{(k)} - f^{(k+1)}] + \sum_{k=0}^{K-1} Mhn \\
h\sum_{k=0}^{K-1} \frac{2\alpha - 1}{2\alpha}\mathbb{E}[||\nabla f^{(k)}||_2^2] &\leq \mathbb{E}[\sum_{k=0}^{K-1}(f^{(k)} - f^{(k+1)})] + KMhn \\
h\sum_{k=0}^{K-1} \mathbb{E}[||\nabla f^{(k)}||_2^2] &\leq \frac{2\alpha}{2\alpha - 1}\mathbb{E}[f^{(0)} - f^{(K)}] + \frac{2\alpha}{2\alpha - 1}KMhn \\
&\leq \frac{2\alpha}{2\alpha - 1}\mathbb{E}[f^{(0)} - f^*] + \frac{2\alpha}{2\alpha - 1}KMhn
\end{aligned}$$

with the last step due to the fact that $f^{(K)} \geq f^*$ by definition of f^*.

Using Lemma 4.1 with $\theta = \theta^{(0)}$ and $\bar{\theta} = \theta^*$, and taking the expectation, we get

$$\mathbb{E}[f^{(0)} - f^*] - \mathbb{E}[\nabla {f^*}^\top (f^{(0)} - f^*)] \leq \frac{M}{2}\mathbb{E}[||\theta^{(0)} - \theta^*||_2^2],$$

which is just

$$2\mathbb{E}[f^{(0)} - f^*] \leq M\mathbb{E}[||\theta^{(0)} - \theta^*||_2^2]$$

as f^* being the global minimum implies $\nabla f^* = 0$.

Substituting this result to our previously established inequality yields

$$h\sum_{k=0}^{K-1} \mathbb{E}[||\nabla f(\theta^{(k,h)})||_2^2] \leq \frac{M\alpha}{2\alpha - 1}\mathbb{E}[||\theta^{(0)} - \theta^*||_2^2] + \frac{2\alpha MKhn}{2\alpha - 1},$$

as desired. □

4.4 Error 2

To study the approximation error between the distribution of L_{Kh} from the continuous-time Langevin diffusion and $\theta^{(k,h)}$ of the discretised process generated by Equation (4.2), we need to find another continuous-time process that coincides with $\theta^{(k,h)}$ at certain times. We will then compare the two continuous-time processes to obtain Error 2.

We first introduce a continuous-time Markov process $\{D_t : t \geq 0\}$ such that the distribution of the random vectors $(\theta^{(0)}, \theta^{(1,h)}, \cdots, \theta^{(K,h)})$ and $(D_0, D_h, \cdots, D_{Kh})$ coincide. To be more precise, we introduce a diffusion-type continuous-time process D obeying the following stochastic differential equation:

$$dD_t = b_t(D)dt + \sqrt{2}dW_t, \quad t \geq 0, \quad D_0 = \theta^{(0)} \tag{4.4}$$

with the (nonanticipative) drift $b_t(D) = -\sum_{k=0}^{\infty} \nabla f(D_{kh}) \mathbf{1}_{[kh,(k+1)h]}(t)$. By integrating the last equation on the interval $[kh, (k+1)h]$, we check that the increments of this process satisfy $D_{(k+1)h} - D_{kh} = -h\nabla f(D_{kh}) + \sqrt{2h}\zeta^{(k+1)}$, where $\zeta^{(k)} = (W_{(k+1)h} - W_{kh})/\sqrt{h}$. Since the Brownian motion is a Gaussian process with independent increments, we conclude that $\{\zeta^{(k)} : k = 1, 2, \cdots, K\}$ is a sequence of i.i.d. standard Gaussian random vectors. This implies that the distribution of the random vectors $(\theta^{(0)}, \theta^{(1,h)}, \cdots, \theta^{(K,h)})$ and $(D_0, D_h, \cdots, D_{Kh})$ coincide.

If for some $B > 0$, the nonanticipative drift function $b : C(\mathbb{R}_+, \mathbb{R}^n) \times \mathbb{R}_+ \to \mathbb{R}^n$ satisfies the inequality $||b(D,t)||_2 \leq B(1 + ||D||_\infty)$ for every $t \in [0, Kh]$ and every $D \in C(\mathbb{R}_+, \mathbb{R}^n)$, then the Kullback-Leibler divergence between $\mathbb{P}_L^{x,Kh}$ and $\mathbb{P}_D^{x,Kh}$, the distributions of the processes $\{L_t : t \in [0, Kh]\}$ and $\{D_t : t \in [0, Kh]\}$ with the initial value $L_0 = D_0 = x$, is given by

$$KL(\mathbb{P}_L^{x,Kh} || \mathbb{P}_D^{x,Kh}) = \frac{1}{4} \int_0^{Kh} \mathbb{E}[||\nabla f(D_t) + b_t(D)||_2^2] \, dt. \tag{4.5}$$

The last equality remains valid even when the initial values of the processes are random but have the same distribution.

–

We are ready to show the bound of Error 2.

Lemma 4.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function satisfying the second inequality in (4.1) and $\theta^* \in \mathbb{R}^n$ be a stationary point (i.e. $\nabla f(\theta^*) = 0$). For any $T > 0$, let $\mathbb{P}_L^{x,T}$ and $\mathbb{P}_D^{x,T}$ be respectively the distributions of the Langevin diffusion (4.3) and its approximation (4.4) on the space of all continuous paths on $[0, T]$ with values in $\mathbb{R}^n$, with a fixed initial value x. Then, if $h \leq 1/(\alpha M)$ with $\alpha \geq 1$, it holds that*

$$KL(\mathbb{P}_L^{x,Kh} || \mathbb{P}_D^{x,Kh}) \leq \frac{M^3 h^2 \alpha}{12(2\alpha - 1)}(||x - \theta^*||_2^2 + 2Khn) + \frac{nKM^2h^2}{4}.$$

Proof. Setting $T = Kh$ and using (4.5), we get

$$\begin{aligned} KL(\mathbb{P}_L^{x,T} || \mathbb{P}_D^{x,T}) &= \frac{1}{4} \int_0^T \mathbb{E}[||\nabla f(D_t) + b_t(D)||_2^2] \, dt \\ &= \frac{1}{4} \sum_{k=0}^{K-1} \int_{kh}^{(k+1)h} \mathbb{E}[||\nabla f(D_t) - \nabla f(D_{kh})||_2^2] \, dt \end{aligned}$$

using the fact that $b_t(D) = -\sum_{k=0}^{\infty} \nabla f(D_{kh}) \mathbf{1}_{[kh,(k+1)h]}(t)$.

Since ∇f is Lipschitz continuous with Lipschitz constant M, i.e. $||\nabla f(\theta) - \nabla f(\bar{\theta})||_2^2 \leq M^2 ||\theta - \bar{\theta}||_2^2$, we have

$$KL(\mathbb{P}_L^{x,T} || \mathbb{P}_D^{x,T}) \leq \frac{M^2}{4} \sum_{k=0}^{K-1} \int_{kh}^{(k+1)h} \mathbb{E}[||D_t - D_{kh}]||_2^2] \, dt$$

From Equation (4.4) $dD_t = b_t(D)dt + \sqrt{2}dW_t$, we will integrate this SDE over t and kh where t is between kh and $(k+1)h$. This gives

$$D_t = D_{kh} + \int_{kh}^{t} b_s(D)ds + \sqrt{2}\int_{kh}^{t} dWs,$$

which, if we apply the definition of $b_s(D)$ as well as the computation rule of stochastic integral, is equivalent to the following:

$$D_t - D_{kh} = -(t-kh)\nabla f(D_{kh}) + \sqrt{2(t-kh)}\zeta.$$

Then, we have

$$\begin{aligned}
||D_t - D_{kh}||_2^2 &= [D_t - D_{kh}]^\top [D_t - D_{kh}] \\
&= (t-kh)^2||\nabla f(D_{kh})||_2^2 - (t-kh)\sqrt{2(t-kh)}\nabla f(D_{kh})^\top \zeta \\
&\quad - (t-kh)\sqrt{2(t-kh)}\zeta^\top \nabla f(D_{kh}) + 2(t-kh)\zeta^\top\zeta.
\end{aligned}$$

Taking expectation, we get

$$\begin{aligned}
\mathbb{E}[||D_t - D_{kh}||_2^2] &= (t-kh)^2\mathbb{E}[||\nabla f(D_{kh})||_2^2] - (t-kh)\sqrt{2(t-kh)}\mathbb{E}[\nabla f(D_{kh})^\top\zeta] \\
&\quad - (t-kh)\sqrt{2(t-kh)}\mathbb{E}[\zeta^\top\nabla f(D_{kh})] + 2(t-kh)\mathbb{E}[\zeta^\top\zeta] \\
&= (t-kh)^2\mathbb{E}[||\nabla f(D_{kh})||_2^2] + 2n(t-kh).
\end{aligned}$$

The above workings occurred similarly in the proof of Proposition 4.4. Substituting this equality into the inequality of KL divergence, we get

$$\begin{aligned}
KL(\mathbb{P}_L^{x,T}||\mathbb{P}_D^{x,T}) &\leq \frac{M^2}{4}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h}\mathbb{E}[||D_t - D_{kh}]||_2^2]\ dt \\
&= \frac{M^2}{4}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h}(t-kh)^2\mathbb{E}[||\nabla f(D_{kh})||_2^2] + 2n(t-kh)\ dt \\
&= \frac{M^2}{4}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h}(t-kh)^2\mathbb{E}[||\nabla f(D_{kh})||_2^2]\ dt + \frac{M^2}{4}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h}2n(t-kh)\ dt \\
&= \frac{M^2}{4}\sum_{k=0}^{K-1}\mathbb{E}[||\nabla f(D_{kh})||_2^2]\int_{kh}^{(k+1)h}(t-kh)^2\ dt + \frac{M^2}{4}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h}2n(t-kh)\ dt \\
&= \frac{M^2h^3}{12}\sum_{k=0}^{K-1}\mathbb{E}[||\nabla f(D_{kh})||_2^2] + \frac{nKM^2h^2}{4} \\
&= \frac{M^2h^3}{12}\sum_{k=0}^{K-1}\mathbb{E}[||\nabla f(\theta^{(k,h)})||_2^2] + \frac{nKM^2h^2}{4}.
\end{aligned}$$

Applying Corollary 4.5, we could obtain

$$\begin{aligned}
KL(\mathbb{P}_L^{x,T}||\mathbb{P}_D^{x,T}) &\leq \frac{M^2h^2}{12}\Big\{h\sum_{k=0}^{K-1}\mathbb{E}[||\nabla f(\theta^{(k,h)})||_2^2]\Big\} + \frac{nKM^2h^2}{4} \\
&\leq \frac{M^2h^2}{12}\Big\{\frac{M\alpha}{2\alpha-1}\mathbb{E}[||\theta^{(0)}-\theta^*||_2^2] + \frac{2\alpha MKhn}{2\alpha-1}\Big\} + \frac{nKM^2h^2}{4} \\
&\leq \frac{M^2h^2}{12}\Big\{\frac{M\alpha}{2\alpha-1}||x-\theta^*||_2^2 + \frac{2\alpha MKhn}{2\alpha-1}\Big\} + \frac{nKM^2h^2}{4} \\
&\leq \frac{M^3h^2\alpha}{12(2\alpha-1)}(||x-\theta^*||_2^2 + 2Khn) + \frac{nKM^2h^2}{4},
\end{aligned}$$

as desired. □

We can set $T = Kh$, and assume that the initial value of the ULA follows the distribution $\nu \sim N_n(\theta^*, M^{-1}I_n)$ where θ^* is a stationary point of f and M is the constant in (4.1). In this case, the equation of Lemma 4.6 becomes

$$\begin{aligned}
KL(\nu\mathbb{P}_L^{x,T}||\nu\mathbb{P}_D^{x,T}) &\leq \frac{M^3h^2\alpha}{12(2\alpha-1)}(\mathbb{E}[||\theta^{(0)}-\theta^*||_2^2] + 2Khn) + \frac{nKM^2h^2}{4} \\
&= \frac{M^3h^2\alpha}{12(2\alpha-1)}(M^{-1}n + 2Khn) + \frac{nKM^2h^2}{4} \\
&= \frac{nM^2h^2\alpha}{12(2\alpha-1)} + \frac{nM^3h^3\alpha K}{6(2\alpha-1)} + \frac{nKM^2h^2}{4} \\
&= \frac{nM^2h^2\alpha}{12(2\alpha-1)} + \frac{nM^3Th^2\alpha}{6(2\alpha-1)} + \frac{nM^2Th}{4} \\
&= \frac{nM^2Th}{4}\Big(\frac{\alpha}{3K(2\alpha-1)} + \frac{2Mh\alpha}{3(2\alpha-1)} + 1\Big).
\end{aligned}$$

For $K \geq \alpha$ and $h \leq 1/(\alpha M)$, we have

$$\frac{\alpha}{3K(2\alpha-1)} \leq \frac{1}{3(2\alpha-1)}$$

and

$$\frac{2Mh\alpha}{3(2\alpha-1)} \leq \frac{2}{3(2\alpha-1)},$$

so

$$\frac{\alpha}{3K(2\alpha-1)} + \frac{2Mh\alpha}{3(2\alpha-1)} + 1 \leq \frac{1}{3(2\alpha-1)} + \frac{2}{3(2\alpha-1)} - 1 = \frac{2\alpha}{(2\alpha-1)}.$$

Thus, we have

$$KL(\nu\mathbb{P}_L^{x,T}||\nu\mathbb{P}_D^{x,T}) \leq \frac{nM^2Th\alpha}{2(2\alpha-1)}. \tag{4.6}$$

This equation is going to be used later when we prove Theorem 4.8

4.5 Main Result

We are almost ready to establish the main Theorem. But first, we need one auxiliary lemma.

Lemma 4.7. *Let us denote by $\nu_{h,x}$ the conditional density of $\theta^{(1,h)}$ given $\theta^{(0)} = x$, where the sequence $\{\theta^{(k,h)}\}_{k\in\mathbb{N}}$ is defined by (4.2) with a function f satisfying (4.1). In other terms, $\nu_{h,x}$ is the density of the Gaussian distribution $N_n(x - h\nabla f(x), 2hI_n)$. If $h \leq 1/(2M)$, then*

$$\mathbb{E}\Big[\frac{\nu_{h,x}(\theta)^2}{\pi(\theta)^2}\Big] \leq \exp\Big\{\frac{1}{2m}||\nabla f(x)||_2^2 - \frac{n}{2}\log(2hm)\Big\}.$$

Theorem 4.8. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function satisfying (4.1) and $\theta^* \in \mathbb{R}^n$ be its global minimum point. Assume that for some $\alpha \geq 1$, we have $h \leq 1/(\alpha M)$ and $K \geq \alpha$. Then, for any time horizon $T = Kh$, the total variation distance between the target distribution P_π and the approximation νP_θ^K furnished by the ULA with initial distribution $\nu \sim N_n(\theta^*, M^{-1}I_n)$ satisfies*

$$||\nu P_\theta^K - P_\pi||_{TV} \leq \frac{1}{2}\exp\Big\{\frac{n}{4}\log(\frac{M}{m}) - \frac{Tm}{2}\Big\} + \Big\{\frac{nM^2Th\alpha}{4(2\alpha-1)}\Big\}^{1/2}.$$

Remark. *The second term in the right-hand side of the above inequality tends to infinity when $T = Kh$ tends to infinity and h remains fixed.*

Proof. The bound consists of two partial bounds.

$$\text{Target Distribution} \quad \longleftrightarrow \quad \text{Distribution of } \{L_t\} \quad \longleftrightarrow \quad \text{Distribution of } \{\theta^{(k,h)}\}$$

Using triangle inequality, we have

$$||\nu P_\theta^K - P_\pi||_{\mathrm{TV}} = ||\nu P_D^{Kh} - P_\pi||_{\mathrm{TV}} \leq ||\nu P_L^T - P_\pi||_{\mathrm{TV}} + ||\nu P_D^T - \nu P_L^T||_{\mathrm{TV}}.$$

The first term of the very right-hand side is the approximation of the first arrow (between Target Distribution and Distribution of $\{\theta^{(k,h)}\}$), and the second term is the approximation by the second arrow. Two approximation bounds have been established more or less previously, and a minimum amount of work is required here to connect the established dots.

The first term requires Lemma 4.2. This result gives us

$$||\nu P_L^T - P_\pi||_{TV} \leq \frac{1}{2}\chi^2(\nu||\pi)^{1/2}e^{-Tm/2}.$$

By definition of χ^2 divergence, we have

$$\begin{aligned}
\chi^2(\nu||\pi) &= \int \Big(\frac{d\nu}{d\pi}(x) - 1\Big)^2 \pi(dx) \\
&= \mathbb{E}_\pi\Big(\frac{d\nu}{d\pi}(x) - 1\Big)^2 \\
&= \mathbb{E}_\pi\Big(\frac{\nu}{\pi}(x) - 1\Big)^2 \\
&\quad \textit{since we can view } \nu = d\nu/dL \textit{ where } L \textit{ is the Lebesgue measure, same goes for } \pi \\
&= \mathbb{E}_\pi\Big(\frac{\nu^2}{\pi^2}(x) - 2\frac{\nu}{\pi}(x) + 1\Big) \\
&= \mathbb{E}_\pi\Big(\frac{\nu^2}{\pi^2}(x)\Big) - 1.
\end{aligned}$$

Notice that Lemma 4.7 involves $\nu_{h,x}$ which is the density of $N_n(x - h\nabla f(x), 2hI_n)$. Here, our ν is the density of $\nu \sim N_n(\theta^*, M^{-1}I_n)$, which is $\nu_{1/(2M),\theta^*}$ and this satisfies the condition of Lemma 4.7. So, applying it to the first term above, we get

$$\begin{aligned}\mathbb{E}_\pi\Big(\frac{\nu^2}{\pi^2}(x)\Big) &\leq \exp\Big\{\frac{1}{2m}||\nabla f(\theta^*)||_2^2 - \frac{n}{2}\log(2m/(2M))\Big\} \\ &= \exp\Big\{-\frac{n}{2}\log(m/M)\Big\} \\ &= \exp\Big\{\frac{n}{2}\log\Big(\frac{M}{m}\Big)\Big\}.\end{aligned}$$

Note that M/m in the last line above is known as the condition number for the distribution with density proportional to e^{-f}. This is a significant quantity in general that it is worth taking note of.

So, we have

$$\|\nu P_L^T - P_\pi\|_{TV} \leq \frac{1}{2}\chi^2(\nu||\pi)^{1/2}e^{-Tm/2} \leq \frac{1}{2}\exp\Big\{\frac{n}{4}\log\Big(\frac{M}{m}\Big) - \frac{Tm}{2}\Big\},$$

which settles the first error.

The next error is a simple application of the Pinsker inequality. The Pinsker inequality states that, for two probability distributions P and Q, we have

$$||P - Q||_{TV} \leq \Big(\frac{1}{2}\ KL(P||Q)\Big)^{1/2}.$$

Using this, we have

$$||\nu P_D^T - \nu P_L^T||_{TV} \leq \Big(\frac{1}{2}\ KL(\nu P_D^T||\nu P_L^T)\Big)^{1/2} \leq \Big(\frac{nM^2Th\alpha}{4(2\alpha - 1)}\Big)^{1/2},$$

where the last inequality is due to Equation (4.6).

Thus, combining the pieces, we have

$$\begin{aligned}||\nu P_\theta^K - P_\pi||_T V &\leq ||\nu P_L^T - P_\pi||_{\mathrm{TV}} + ||\nu P_D^T - \nu P_L^T||_{TV} \\ &\leq \frac{1}{2}\exp\Big\{\frac{n}{4}\log\Big(\frac{M}{m}\Big) - \frac{Tm}{2}\Big\} + \Big\{\frac{nM^2Th\alpha}{4(2\alpha - 1)}\Big\}^{1/2}\end{aligned}$$

as desired. □

4.6 Discussion

The above theorem provides a non-asymptotic bound on mixing time that can be used to definitively state whether the chain is sufficiently close to equilibrium or not.

There are two natural questions we could ask at this stage. First, under similar conditions, can we establish similar convergence bounds for other common MCMC algorithms? Second, how realistic are the assumptions imposed on the target distribution?

Many results have been established since Dalalyan (2017). Works such as Dwivedi et al. (2018) and Chewi et al. (2020) are all imposing the same type of conditions on the target distribution.

There have been a significant amount of work done in recent years in this area, and many of them are summarised in the unfinished draft of Chewi (2023).

Now, is log-concave a realistic assumption? There are common distributions that are log-concave, such as the normal distribution and the exponential distribution. However, some frequently used distributions are not log-concave, such as the Student's t-distribution. Therefore, we should certainly not stop at log-concavity. Recently, there have been attempts at moving towards non-log-concavity (Chewi et al., 2022).

Part III

Barker Algorithms

Chapter 5

The Barker Proposal

5.1 Background

Since the genesis of the Metropolis-Hastings algorithm, there have been a lot of discussions focusing on what a good proposal kernel could be. Among them, there are a few that stand out: random walk Metropolis (RWM), Metropolis adjusted Langevin algorithm (MALA), and Hamiltonian Monte Carlo (HMC). RWM employs a simple centred normal distribution as the proposal kernel and its standard deviation is the tuning parameter. MALA, on the other side, exploits the gradient information of the target distribution and uses that to drift towards the direction of the region of higher probability of the target distribution. HMC can be viewed as an extension of MALA, by doing multiple MALA proposals at each update step. To simplify the discussion, let us focus mostly on RWM and MALA.

In modern-day practices, the target distributions of MCMC algorithms tend to be high dimensional. So it is only natural for us to consider how well various MCMC algorithms scale over dimensions. It turns out that if we assess the quality of an MCMC algorithm using expected squared jump distance (ESJD), we could notice that MALA scales much better than RWM (Livingstone and Zanella, 2022).

Another concern in practice is about the tuning of the MCMC algorithm parameter. Some guidelines do exist for parameter tuning by looking at the value of the acceptance rate (see Roberts and Rosenthal (2001) for a survey on various scaling results), and they allow us to use the algorithms adaptively by adjusting the parameter on the fly (Andrieu and Thoms, 2008). The problem, however, is that a small perturbation of the tuning parameter may cause a large impact on the spectral gap, which is strongly related to the quality of the generated samples (Rosenthal, 2003). This problem results in difficulties in fine-tuning parameters, and it will become more apparent when the algorithms are implemented adaptively as the perturbation issue gets compounded. As illustrated theoretically in Livingstone and Zanella (2022), MALA and HMC have poor robustness to tuning (using the language of the mentioned paper) while RWM has better robustness.

Of course, there is always the elementary problem of all MCMC algorithms - does it converge quickly? We would be preliminarily satisfied with the algorithm's quality of convergence once we have established its geometric ergodicity.

Combining all the above issues, an algorithm with good scaling over dimension property, robustness to tuning, and geometric ergodicity all at once would be desirable. And one such algorithm is the Barker proposal, introduced in Livingstone and Zanella (2022).

5.2 Algorithm

In this section, we will present the algorithm of the Barker proposal. The Barker proposal is still under the framework of MH algorithms, but its proposal kernel is drastically different from the existing algorithms. The algorithm invokes the gradient information of the target distribution, similar to algorithms such as MALA and HMC. However, the way gradient information is incorporated into the algorithm is different from that of MALA and HMC. Instead of directly having a drift towards the gradient, the gradient information skews the symmetric noise (say $N(0, \sigma^2)$) towards the direction of the gradient. This circumvents the problem of MALA and HMC when the algorithm is at a place with a huge gradient, as in those cases the proposed step will move too far and skip over the entire region of high probability.

The following is the one-dimensional Barker algorithm. Let the target distribution be $\pi(\cdot)$, the current position be x, and $q(\cdot)$ be some symmetrical density. The following is how the Barker algorithm generates a proposal at each step. This proposal is then fed into the standard MH algorithm.

Algorithm 3 One-Dimensional Proposal of Barker

Require: Target distribution π, current position x, symmetric density q

1: Draw $z \sim q(\cdot)$

2:
$$b = \begin{cases} +1 & \text{with probability} \quad 1/[1 + \exp(-z\nabla \log \pi(x))] \\ -1 & \text{elsewise} \end{cases}$$

3: $y = x + b \cdot z$

Next, we will look at the n-dimensional Barker algorithm. There are two immediate ways of extending the above algorithm, and their difference lies in how we could extend b. The first way is to make no change with the way we pick b, so the probability will simply become $1/[1 + \exp(-z^\top \nabla \log \pi(x))]$. This means we are skewing q towards the average direction of gradient per component. The second way is to choose a b_i for each $i = 1, 2, \ldots, n$ so that we are deciding the direction of skewing for each component independently. The probability will then become $1/[1 + \exp(-z_i \partial_i \log \pi(x))]$ for component i, where each z_i is drawn independently from q. Figure 5.1, adapted from Hird et al. (2022), illustrates the difference between the two options in dimension 2.

The second option of deciding the skewing direction component-wise is the one used in Livingstone and Zanella (2022). It has also been proved via spectral gap that this is the superior choice in the paper as Proposition 5. Therefore, the proposal of the n-dimensional Barker algorithm is as follows. Let the target distribution be $\pi(\cdot)$, the current position be x, and $q(\cdot)$ be some symmetrical density.

It is natural to ask if it is possible to consider any other options for extending the algorithm into higher dimensions. Another option is considered in Chapter 7.

Algorithm 4 n-Dimensional Proposal of Barker

Require: Target distribution π, current position x, symmetric density q

1: Draw $z \sim q(\cdot)$

2: **for** $i = 1, 2, \ldots, n$ **do**

3:
$$b_i = \begin{cases} +1 & \text{with probability } \frac{1}{1+\exp(-z_i \partial_i \log \pi(x))} \\ -1 & \text{elsewise} \end{cases}$$

4: **end for**

5: $y = x + b \cdot z$

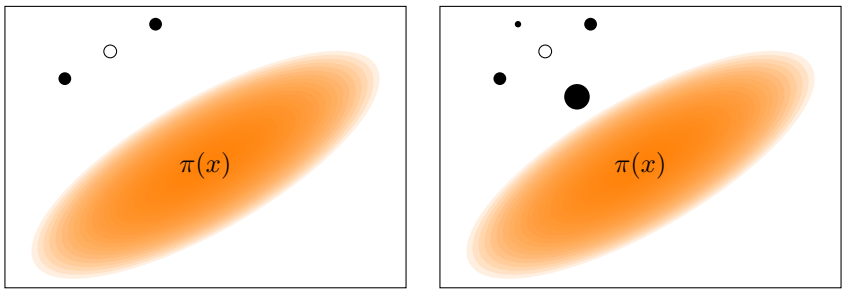


Figure 5.1: Illustrative diagrams for two options for proposal in higher dimensions. The white ball x is the current state, and the sizes of the black balls indicate the probability of moving to each candidate point.

5.3 Skew-Symmetric Distributions and Balancing Functions

It is therefore natural to ask, at this stage, if the introduced Barker proposal actually satisfies the desirable properties of good scaling over dimension property, robustness to tuning, and geometric ergodicity. The answer is a resounding yes. Detailed proofs of these things can be found in Livingstone and Zanella (2022), and further theoretical analysis can be found in Vogrinc et al. (2022).

In these papers, one can realise that the nice properties come from two ingredients of the algorithm - a balancing function and a skew-symmetric distribution. By using a balancing function, we can ensure that the algorithm satisfies the detailed balance equation even though it might not satisfy it originally (Zanella, 2020), and we call such an algorithm a **locally-balanced** one. This ensures the scaling over dimension property, as proved in Vogrinc et al. (2022). The use of a skew-symmetric distribution as the proposal kernel of the MH algorithm allows us to have the robustness to tuning property, as inferred from the proof of Theorem 5 of Livingstone and Zanella (2022).

In the rest of this section, we will restrain ourselves to the algorithm in one dimension and look at these two ingredients (the balancing function and the skew-symmetric distribution) in detail, and derive that one would get the Barker proposal naturally if we have the goal in mind to incorporate both these things in the algorithm.

5.3.1 Skew-Symmetric Distribution

Let us first consider the candidate transition kernel of the one-dimensional Barker algorithm. The candidate transition kernel describes the movement of the proposal, before passing through the Metropolis adjustment step of MH algorithms. We will denote the transition kernel from x to $y = z + x$ as $j_x(z)$. Also, for simplicity of notation, we write $\beta_x := \nabla \log \pi(x)$ and $F_L(x) := 1/[1 + e^{-x}]$. Also, q is a symmetric distribution. Using these notations, the proposal then becomes:

Algorithm One-Dimensional Proposal of Barker

Require: Target distribution π, current position x, symmetric density q

1: Draw $z \sim q(\cdot)$

2:
$$b = \begin{cases} +1 & \text{with probability} \quad F_L(\beta_x \xi) \\ -1 & \text{with probability} \quad 1 - F_L(\beta_x \xi) \end{cases}$$

3: $y = x + b \cdot z$

The candidate transition kernel $j_x(z)$ can then be derived as follows.

$$\begin{aligned} j_x(z) &= q(z)F_L(\beta_x z) + q(-z)[1 - F_L(-\beta_x z)] \\ &= q(z)F_L(\beta_x z) + q(z)[1 - F_L(-\beta_x z)] \\ &= q(z)\left[\frac{1}{1+\exp(-\beta_x z)} + 1 - \frac{1}{1+\exp(\beta_x z)}\right] \\ &= q(z)\frac{2}{1+\exp(-\beta_x z)} = 2q(z)F_L(\beta_x z). \end{aligned}$$

A skew-symmetric distribution has the general form of $f(z) = 2f_0(z)G(\beta z)$ where f_0 is a symmetric probability distribution function and $G(z)$ is a cumulative density function with symmetric derivative G' (Azzalini and Regoli, 2012). The symmetric derivative of G implies that we have $G(z) + G(-z) = 1$ for all z. All the requirements for a skew-symmetric distribution are satisfied for $j_x(z)$. Therefore, we could say that the one-dimensional proposal is simply drawn from the distribution $2q(z)F_L(\beta_x z)$.

For a skew-symmetric distribution, the distribution will have a positive skew if $\beta > 0$, and a negative skew for $\beta < 0$. So, by using $\beta_x := \nabla \log \pi(x)$, we will be skewing the distribution towards the gradient. This is how the gradient information is exploited for the Barker proposal.

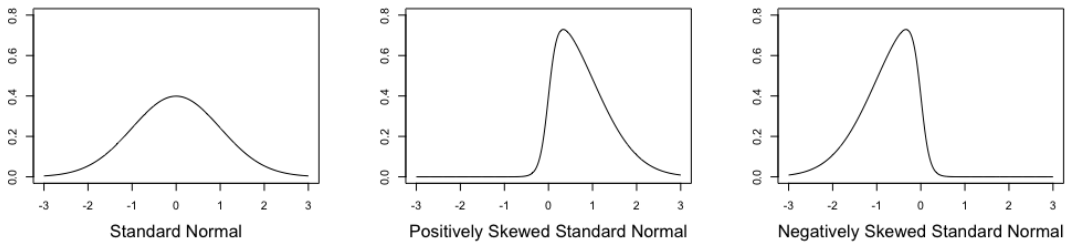


Figure 5.2: Skew Symmetric Distributions

If we re-examine the derivations above, we would notice that the choice of the logistic CDF F_L is not unique if we simply want to obtain a skew-symmetric distribution as our candidate transition kernel. Any other cumulative density function with a symmetric derivative could, in theory, work. However, the choice of F_L is in fact not arbitrary, as we will see right below that this is the unique choice if we would like our algorithm to have skew-symmetric proposals and be a locally-balanced algorithm.

5.3.2 Balancing Function and Locally-Balanced Algorithms

For a Markov chain with transition kernel A that admits a density q, if it satisfies the detailed-balanced equation with π, i.e.

$$\pi(x)q(x,y) = \pi(y)q(y,x)$$

for all x, y, then the Markov chain is π-reversible. If the chain is also aperiodic and ϕ-irreducible (which is almost always the case for chains of MH algorithms), then the Markov chain will have π as its equilibrium distribution. Thus, it is certainly desirable (and even required) for the MCMC algorithm-generated Markov chains to satisfy the detailed balance equation. However, Q might not always satisfy it, in which case we can remedy this problem by introducing a balancing function g. This approach is first proposed in Zanella (2020) in the context of discrete state space MCMC algorithms. The simplicity of this approach allows it to be easily transplanted to the general state space cases.

Let $t(x,y) := \pi(y)q(y,x)/[\pi(x)q(x,y)]$, and $t(x,y) := 0$ when $\pi(x)q(x,y) = 0$. If we further let $p(x,y) := q(x,y)g(t(x,y))$, then in order for this p to satisfy the detailed balance equation, we would have

$$\begin{aligned}
\pi(x)p(x,y) &= \pi(x)q(x,y)g(t(x,y)) \\
&= \pi(y)q(y,x)\frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}g(t(x,y)) \\
&= \pi(y)q(y,x)t^{-1}(x,y)g(t(x,y)) \\
\pi(y)p(y,x) &= \pi(y)q(y,x)g(t(y,x)) \\
&= \pi(y)q(y,x)g(1/t(x,y)),
\end{aligned}$$

so g must satisfies the condition: $g(t) = tg(1/t)$ for all t. Of course, we set $g(0) := 0$.

For simplicity, we can assume q is symmetric, i.e. $q(x,y) = q(y,x)$ for all x, y. In this case, $t(x,y) = \pi(y)/\pi(x)$, which means

$$p(x,y) = q(x,y)g(\pi(y)/\pi(x)).$$

We could make an approximation to the fraction $\pi(y)/\pi(x)$ by using a Taylor series expansion and keeping the first order term. This allows the balancing function to do its job of ensuring the satisfaction of the detailed balance equation **locally**. This is why such algorithms are called locally-balanced. The first order approximation in the one-dimensional case is as follows:

$$\pi(y)/\pi(x) = \exp[\log(\pi(y)) - \log(\pi(x))] \approx \exp[(y-x)\nabla \log \pi(x)].$$

Now, we can rewrite the kernel p using this approximation. By letting $z := y - x$ and $\beta_x := \nabla \log \pi(x)$, we have

$$p(x,y) := p_x(z) = q_x(z)g(e^{z\beta_x})/Z(x)$$

where $Z(x)$ is a normalising constant. We know from above that $g(t) = tg(1/t)$. Also, as we would want to enjoy all the benefits of a skew-symmetric distribution proposal, we would want this p to be a skew-symmetric distribution, i.e. $G(z) + G(-z) = 1$. Combining these two information, if we treat $G(z) = g(e^{z\beta_x})$, then we have

$$\begin{aligned} g(e^{z\beta_x}) + g(e^{-z\beta_x}) &= 1 \\ g(e^{z\beta_x}) + g(1/e^{z\beta_x}) &= 1 \\ g(e^{z\beta_x}) + e^{-z\beta_x} g(e^{z\beta_x}) &= 1 \\ g(e^{z\beta_x}) &= 1/[1 + e^{-z\beta_x}]. \end{aligned}$$

Thus, we have the following unique choice of g

$$g(e^t) = \frac{1}{1+e^{-t}} = \frac{e^t}{e^t+1},$$

and $g(e^{z\beta_x}) = F_L(\beta_x z)$ where $F_L(t) = 1/[1 + e^{-t}]$ as defined earlier. We are almost there. The final thing we need to check is the normalising constant $Z(x)$. We have

$$\begin{aligned} \int_{-\infty}^{\infty} q_x(z) g(e^{z\beta_x}) dz &= \int_0^{\infty} q_x(z) g(e^{z\beta_x}) dz + \int_{-\infty}^{0} q_x(z) g(e^{z\beta_x}) dz \\ &= \int_0^{\infty} q_x(z) g(e^{z\beta_x}) dz + \int_0^{\infty} q_x(z) g(e^{-z\beta_x}) dz \\ &= \int_0^{\infty} q_x(z) [g(e^{z\beta_x}) + g(e^{-z\beta_x})] dz \\ &= \int_0^{\infty} q_x(z) dz = 1/2. \end{aligned}$$

Thus, we have successfully derived $p_x(z) = 2q_x(z)F_L(z\beta_x)$, as desired.

An alternative derivation of the above results can be found in Hird et al. (2022).

Chapter 6

The Barker Scheme

In Chapter 3, we have drawn the link between the unadjusted Langevin algorithm (ULA) and the Euler-Maruyama scheme used in numerical solutions of SDEs. This link leads to the natural question: can we obtain a new numerical scheme for SDEs by considering the unadjusted versions of various MH algorithms?

The answer is no for RWM, as we would get a trivial Brownian motion in that case. However, the answer is yes for the Barker proposal. In this chapter, we will look at the Barker scheme, which is the unadjusted Barker proposal adapted into the framework of SDE numerical solutions. We will also denote the same algorithm as the Unadjusted Barker. The two terms "Barker scheme" and "Unadjusted Barker" are used interchangeably in this thesis. This scheme is not completely developed yet, and many results about it are still partial. We will look at this scheme in the case of numerically solving one-dimensional autonomous SDEs, and we have established a geometric ergodicity result with several relatively strong conditions (which we intend to weaken in future work). Several numerical studies are conducted and provided in this chapter.

6.1 Algorithm Setup

Consider the (one-dimensional) autonomous SDE

$$dY_t = \mu(Y_t)dt + \sigma(Y_t)dW_t, \qquad Y_0 = y_0. \tag{6.1}$$

Here, $\{W_t\}_{t\geq 0}$ is a standard Wiener process, $\mu : \mathbb{R} \to \mathbb{R}$ is a drift function and $\sigma : \mathbb{R} \to [0, \infty)$ is a volatility function.

Our numerical scheme with step-size $\delta > 0$ is

$$\begin{aligned} X_{(n+1)\delta} &= X_{n\delta} + b_{n+1}\xi_{n+1} \\ &= X_{n\delta} + \sqrt{\delta} b_{n+1}\sigma(X_{n\delta})\nu_{n+1} \\ X_0 &= x_0 = y_0 \end{aligned} \tag{6.2}$$

where $\xi_{n+1} \sim N(0, \sigma^2(X_{n\delta})\delta)$, $\nu_{n+1} \sim N(0, 1)$, and

$$b_{n+1} = \begin{cases} +1 & \text{w.p. } p(X_{n\delta}, \xi_{n+1}) \\ -1 & \text{otherwise.} \end{cases} \tag{6.3}$$

Here, the probability $p(X_{n\delta}, \xi_{n+1})$ and direction b_{n+1} inject skewness into the movement. The probability is chosen to be defined as

$$p(X_{n\delta}, \xi_{n+1}) = \left[1 + \exp\left(-2\frac{\xi_{n+1}\mu(Y_{n\delta})}{\sigma^2(Y_{n\delta})}\right)\right]^{-1}.$$

This choice of p is not unique, and we suspect any other probability function with similar properties can be used too.

6.2 Geometric Ergodicity of Unadjusted Barker

Theorem 6.1 (Original Result)**.** *Under the assumptions that* $\sigma(x) = C > 0$ *and* $\lim_{x\to\pm\infty} \mu(x) = \mp\infty$*, the Markov chain produced by the one-dimensional unadjusted Barker with parameter* δ *is geometrically ergodic.*

Proof. Without loss of generality, we let $\sigma(x) = \sqrt{2}$. It will become obvious that any arbitrary positive constant will still make the proof hold. This specific choice is only there to simplify the proof.

We need the drift condition and the minorisation condition to establish geometric ergodicity using a combination of Theorem 2.12 and Proposition 2.13. We will show the minorisation condition first.

Here, any compact set would be a small set C, and the minorisation measure ν is taken to be the uniform distribution restricted to C, so we have $\nu(A) = Leb(A \cap C)/Leb(C)$ for all $A \in \mathcal{B}$, where Leb is the Lebesgue measure on $\mathbb{R}$ (Williams, 1991).

Consider a small set $C = [a, b]$ with $a < b \in \mathbb{R}$, we have the transition density $p(x, y) \geq \varepsilon\nu(y)$ for any $x \in C$. This inequality holds trivially for $y \notin C$ as the $\nu(y) = 0$ in this case. For $x, y \in C$, we can then choose ε to be $\inf_{(x,y)\in C\times C}\{p(x, y)\} \cdot (b - a)$, in which case we would have the desired inequality. This value is a valid one as y is drawn from a distribution with support over the whole of $\mathbb{R}$. Thus, we have established the minorisation condition.

Next, we will establish the drift condition to show geometric ergodicity.

For the probability p for choosing the value of b, we have

$$p(x, \xi) = \frac{1}{1 + \exp[-2\xi\mu(x)/\sigma^2(x)]} = \frac{1}{1 + \exp[-\xi\mu(x)]}.$$

This means that $p(x, \xi) \to 1$ as $\xi\mu \to \infty$, and $p(x, \xi) \to 0$ as $\xi\mu \to -\infty$. Since we have $\lim_{x\to\pm\infty} \mu(x) = \mp\infty$, the limiting behaviour of p depends on the sign of ξ. In this proof, we will never consider the case where $\xi = 0$ as that is an event with probability zero. So, as $x \to \infty$, we have $p \to 1$ for $\xi < 0$ and $p \to 0$ for $\xi > 0$. Similarly, as $x \to -\infty$, we have $p \to 1$ for $\xi > 0$ and $p \to 0$ for $\xi < 0$. Furthermore, notice that $\xi = \sqrt{\delta}\sigma(x)\nu$ where $\sqrt{\delta}\sigma(x) > 0$, we would then have $\xi\nu > 0$, i.e. they have the same sign.

Consider the choice of Lyapunov function $V(x) = e^{s|x|}$ for some $s > 0$. We just need to establish the inequality

$$\limsup_{\|x\|\to\infty} \frac{PV(x)}{V(x)} = \limsup_{\|x\|\to\infty} \int \frac{V(y)}{V(x)} P(x, dy) < 1.$$

Here we will only discuss the case where $x \to \infty$. The case where $x \to -\infty$ follows naturally from the proof of the first case via symmetry, thus is omitted here. As a consequence, we consider $x > 0$ here. For the case of $x \to -\infty$, we will set $x < 0$. Thus, we can replace lim sup with lim.

Write $y = x + b\xi = x + b\sqrt{2\delta}\nu$. We have

$$\begin{aligned}
\frac{PV(x)}{V(x)} &= \mathbb{E}\left[\frac{V(y)}{V(x)}\right] = \mathbb{E}[e^{s|x+b\sqrt{2\delta}\nu|-s|x|}] \\
&= \mathbb{E}_\nu \mathbb{E}_{b|\nu}[e^{s|x+b\sqrt{2\delta}\nu|-s|x|}] \\
&= \mathbb{E}_\nu[e^{s|x+\sqrt{2\delta}\nu|-s|x|}p(x,\xi) + e^{s|x-\sqrt{2\delta}\nu|-s|x|}(1-p(x,\xi))] \\
&= \underbrace{\mathbb{E}_\nu[e^{s|x+\sqrt{2\delta}\nu|-s|x|}p(x,\xi)]}_{A} + \underbrace{\mathbb{E}_\nu[e^{s|x-\sqrt{2\delta}\nu|-s|x|}(1-p(x,\xi))]}_{B}
\end{aligned}$$

We will look at each of the two terms in the last line separately, and bound their limits individually. We will denote the two terms as A and B for easy reference.

Using the dominated convergence theorem (Williams, 1991), we can exchange the limit and the expectation. To see this, for A, we have

$$\begin{aligned}
|e^{s|x+\sqrt{2\delta}\nu|-s|x|}p(x,\xi)| &\le e^{s|x+\sqrt{2\delta}\nu|-s|x|} \\
&\le e^{s|x|+s|\sqrt{2\delta}\nu|-s|x|} \\
&= e^{\sqrt{2\delta}s|\nu|}.
\end{aligned}$$

Furthermore, if we use ϕ to denote the probability distribution function of the standard normal distribution, we have

$$\begin{aligned}
\mathbb{E}_\nu[e^{\sqrt{2\delta}s|\nu|}] &= \int_{-\infty}^{\infty} e^{\sqrt{2\delta}s|x|}\phi(x)dx \\
&= 2\int_0^{\infty} e^{\sqrt{2\delta}s|x|}\phi(x)dx \quad \text{as the integrand is an even function} \\
&= 2\int_0^{\infty} e^{\sqrt{2\delta}sx}\phi(x)dx \le 2\int_{-\infty}^{\infty} e^{\sqrt{2\delta}sx}\phi(x)dx \\
&= 2\mathbb{E}_\nu[e^{\sqrt{2\delta}s\nu}] = 2e^{s\delta^2} < \infty,
\end{aligned}$$

where the last equality is due to the moment generating function of the standard normal distribution.

We can provide a similar bound for B.

$$\begin{aligned}
|e^{s|x-\sqrt{2\delta}\nu|-s|x|}[1-p(x,\xi)]| &\le e^{s|x-\sqrt{2\delta}\nu|-s|x|} \\
&\le e^{s|x|+s|\sqrt{2\delta}\nu|-s|x|} \\
&= e^{\sqrt{2\delta}s|\nu|},
\end{aligned}$$

which makes this identical to the bound for A. Thus, we can exchange the limit and expectation for both A and B.

Now, let us bound $\lim A$. We have

$$\begin{aligned}
&\lim_{x\to\infty} \mathbb{E}_\nu[e^{s|x+\sqrt{2\delta}\nu|-s|x|}p(x,\xi)] \\
&= \mathbb{E}_\nu[\lim_{x\to\infty} e^{s|x+\sqrt{2\delta}\nu|-s|x|}p(x,\xi)] \\
&= \mathbb{E}_\nu[\lim_{x\to\infty} e^{s|x+\sqrt{2\delta}\nu|-s|x|}p(x,\xi)|\nu<0]\mathbb{P}(\nu<0) \\
&\quad + \mathbb{E}_\nu[\lim_{x\to\infty} e^{s|x+\sqrt{2\delta}\nu|-s|x|}p(x,\xi)|\nu>0]\mathbb{P}(\nu>0)
\end{aligned}$$

Intuitively speaking, the first term on the last line above denotes the inwards movement of the algorithm (in the sense that the Lyapunov function decreases and we are moving towards the minimum of V), whereas the second term denotes the outwards movement. The inwards movement is desirable while the outwards movement is not.

Notice that when $\nu > 0$, we have $s|x+\sqrt{2\delta}\nu| - s|x| = \sqrt{2\delta}s\nu$ and

$$\lim_{x\to\infty} e^{s|x+\sqrt{2\delta}\nu|-s|x|}p(x,\xi) = \lim_{x\to\infty} \frac{e^{\sqrt{2\delta}s\nu}}{1+e^{-\xi\mu(x)}} = 0$$

as $\xi > 0$ and $\lim_{x\to\infty}\mu(x) = -\infty$ by our condition. This means, the undesirable outwards movement will occur with probability zero as $x \to \infty$ under our assumptions here.

So, we have

$$\begin{aligned}
&\lim_{x\to\infty} \mathbb{E}_\nu[e^{s|x+\sqrt{2\delta}\nu|-s|x|}p(x,\xi)] \\
&= \mathbb{E}_\nu[\lim_{x\to\infty} e^{s|x+\sqrt{2\delta}\nu|-s|x|}p(x,\xi)|\nu<0]\mathbb{P}(\nu<0) + 0 \\
&\le \mathbb{E}_\nu[\lim_{x\to\infty} e^{s|x+\sqrt{2\delta}\nu|-s|x|}|\nu<0]\mathbb{P}(\nu<0) \\
&= \int_{-\infty}^{0} \lim_{x\to\infty} \exp\{s|x+\sqrt{2\delta}y| - s|x|\}\frac{e^{-y^2/2}}{\sqrt{2\pi}}dy.
\end{aligned}$$

For the integrand, for each fixed y, as $x \to \infty$, we have $\lim_{x\to\infty}\exp\{s|x+\sqrt{2\delta}y| - s|x|\} = \exp\{\sqrt{2\delta}sy\}$. So,

$$\begin{aligned}
&\mathbb{E}_\nu[\lim_{x\to\infty} e^{s|x+\sqrt{2\delta}\nu|-s|x|}|\nu<0]\mathbb{P}(\nu<0) \\
&= \int_{-\infty}^{0} \lim_{x\to\infty} \exp\{s|x+\sqrt{2\delta}y| - s|x|\}\frac{e^{-y^2/2}}{\sqrt{2\pi}}dy \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} \exp\{\sqrt{2\delta}sy - y^2/2\}/\sqrt{2\pi}dy = \frac{e^{\delta s^2}}{\sqrt{2\pi}}\int_{-\infty}^{0}\exp\{-(y-\sqrt{2\delta}s)^2/2\}dy \\
&= \frac{e^{\delta s^2}}{\sqrt{2\pi}}\int_{-\infty}^{-\sqrt{2\delta}s}\exp\{-w^2/2\}dw = \frac{e^{\delta s^2}}{\sqrt{2\pi}}\int_{\sqrt{2\delta}s}^{\infty}\exp\{-w^2/2\}dw \\
&< \frac{e^{\delta s^2}}{\sqrt{2\pi}}\int_{\sqrt{2\delta}s}^{\infty}\frac{w}{\sqrt{2\delta}s}\exp\{-w^2/2\}dw = \frac{e^{\delta s^2}}{\sqrt{2\pi}}\left[-\frac{1}{\sqrt{2\delta}s}e^{-w^2/2}\right]_{\sqrt{2\delta}s}^{\infty} \\
&= \frac{e^{\delta s^2}}{\sqrt{2\pi}}\frac{e^{-\delta s^2}}{\sqrt{2\delta}s} = \frac{1}{2s\sqrt{\pi\delta}}.
\end{aligned}$$

Therefore, we have

$$\lim_{x\to\infty} \mathbb{E}_\nu[e^{s|x+\sqrt{2\delta}\nu|-s|x|}p(x,\xi)] < \frac{1}{2s\sqrt{\pi\delta}}.$$

Now we bound the limit of B. This bound will be derived in a similar manner as that of A.

$$\begin{aligned}
&\lim_{x\to\infty} \mathbb{E}_\nu[e^{s|x-\sqrt{2\delta}\nu|-s|x|}(1-p(x,\xi))] \\
&= \mathbb{E}_\nu[\lim_{x\to\infty} e^{s|x-\sqrt{2\delta}\nu|-s|x|}(1-p(x,\xi))] \\
&= \mathbb{E}_\nu[\lim_{x\to\infty} e^{s|x-\sqrt{2\delta}\nu|-s|x|}(1-p(x,\xi))|\nu>0]\mathbb{P}(\nu>0) \\
&\quad + \mathbb{E}_\nu[\lim_{x\to\infty} e^{s|x-\sqrt{2\delta}\nu|-s|x|}(1-p(x,\xi))|\nu<0]\mathbb{P}(\nu<0)
\end{aligned}$$

When $\nu < 0$, we have

$$\lim_{x\to\infty} e^{s|x-\sqrt{2\delta}\nu|-s|x|}(1-p(x,\xi)) = \lim_{x\to\infty} \frac{e^{-\sqrt{2\delta}s\nu}}{1+e^{\xi\mu(x)}} = 0$$

as $\xi < 0$ and $\lim_{x\to\infty}\mu(x) = -\infty$. So,

$$\begin{aligned}
&\lim_{x\to\infty} \mathbb{E}_\nu[e^{s|x-\sqrt{2\delta}\nu|-s|x|}(1-p(x,\xi))] \\
&= \mathbb{E}_\nu[\lim_{x\to\infty} e^{s|x-\sqrt{2\delta}\nu|-s|x|}(1-p(x,\xi))|\nu>0]\mathbb{P}(\nu>0) \\
&\quad + \mathbb{E}_\nu[\lim_{x\to\infty} e^{s|x-\sqrt{2\delta}\nu|-s|x|}(1-p(x,\xi))|\nu<0]\mathbb{P}(\nu<0) \\
&= \mathbb{E}_\nu[\lim_{x\to\infty} e^{s|x-\sqrt{2\delta}\nu|-s|x|}(1-p(x,\xi))|\nu>0]\mathbb{P}(\nu>0) \\
&\le \mathbb{E}_\nu[\lim_{x\to\infty} e^{s|x-\sqrt{2\delta}\nu|-s|x|} \mid \nu>0]\mathbb{P}(\nu>0) \\
&= \int_0^\infty \lim_{x\to\infty} e^{s|x-\sqrt{2\delta}y|-s|x|}e^{-y^2/2}/\sqrt{2\pi}dy \\
&= \int_{-\infty}^0 \lim_{x\to\infty} e^{s|x+\sqrt{2\delta}y|-s|x|}e^{-y^2/2}/\sqrt{2\pi}dy \\
&= \mathbb{E}_\nu[\lim_{x\to\infty} e^{s|x+\sqrt{2\delta}\nu|-s|x|}|\nu<0]\mathbb{P}(\nu<0) < \frac{1}{2s\sqrt{\pi\delta}},
\end{aligned}$$

where the last inequality follows from the result above.

Combining all these, we have

$$\lim_{x\to\infty}\int \frac{V(y)}{V(x)}P(x,dy) < \frac{1}{2s\sqrt{\pi\delta}} + \frac{1}{2s\sqrt{\pi\delta}} = \frac{1}{s\sqrt{\pi\delta}} < 1$$

for any $s > (\pi\delta)^{-1/2} > 0$. As s is arbitrary, the desired inequality would always hold for a suitably chosen value of s. □

Remark. *Usually when we establish geometric ergodicity for MCMC generated chains, we only want to know the speed of convergence from that result as the ergodicity part is easy to obtain. However, in this case, the exact target distribution π is hard to find, therefore we could not use the usual π-invariance + irreducible + aperiodic argument to establish ergodicity. Therefore, it is quite hard to even obtain the existence of an equilibrium without proving this stronger geometric ergodicity result.*

6.3 Numerical Studies

Two numerical studies are conducted in this section. The first study is on comparing the weak convergence of the Barker scheme and that of the Euler-Maruyama scheme for basic SDEs. The second study looks at the Unadjusted Barker (UB), which is just the Barker proposal without the final Metropolis adjustment step. This study compares the performance of UB with several popular MH algorithms for simulating from a relatively complex Poisson random effects model that is frequently used in practice. The second study is closely related to the study conducted in Section 6.3 of Livingstone and Zanella (2022).

6.3.1 Simulations for Basic SDEs

Experiments in this section compare the performance of the Euler-Maruyama scheme and the Barker scheme while solving common SDEs numerically. The performance is assessed by considering the error, which is the absolute difference between the exact solution and the simulated solution, over varying step sizes. If we let X_T^δ be the simulated solution at time T using step size δ and let Y_T be the exact solution at time T, then if the numerical scheme converges weakly, we have, for any suitable f,

$$|\mathbb{E}[f(X_T^\delta)] - \mathbb{E}[f(Y_T)]| \leq C\delta^p$$

where C is a constant independent of δ and p is the weak convergence order (Platen and Kloeden, 1992). Taking log on both sides yield

$$\log|\mathbb{E}[f(X_T^\delta)] - \mathbb{E}[f(Y_T)]| \leq \log C + p\log\delta,$$

so log error will grow roughly linearly (since it is an inequality) with increasing step size.

The weak convergence order of the two numerical schemes has proved to be similar, which means we expect a similar gradient of the two log error plots.

6.3.1.1 Set Up

The SDE of our consideration is the Ornstein-Uhlenbeck process of the form

$$dX_t = \theta(\mu - X_t)dt + \sigma dW_t,$$

where θ, μ, σ are constants. The choice of this process is because it is ergodic with a known exact solution, which enables us to conduct the simulation smoothly.

The time T of the simulated path is 1000, with step sizes δ being one of $0.1, 0.2, 0.3, \ldots, 1$, and the number of steps N is calculated to be $\lfloor T/\delta \rfloor$. The parameters of the process are $\mu = 0$, $\theta = 1$, and $\sigma = \sqrt{2}$. The initial value of the SDE is $X_0 = 1$.

6.3.1.2 Log Error of Ornstein-Uhlenbeck

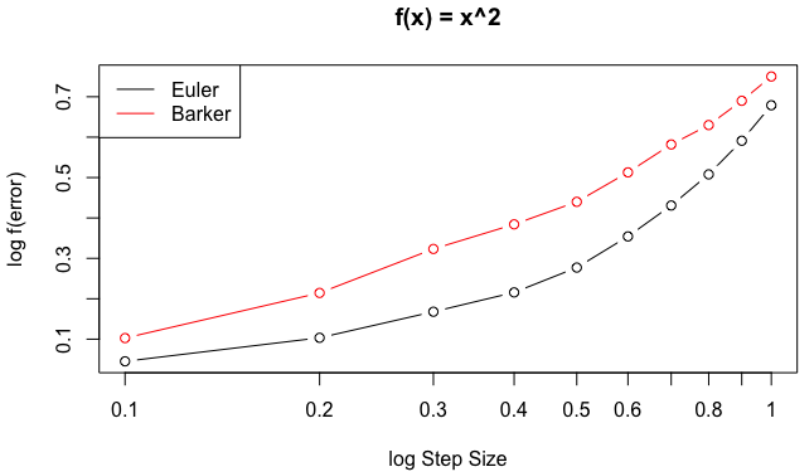


Figure 6.1: Log Error over Log Step Sizes

From the above diagram, we can notice an almost parallel relationship between the log error of simulated solution from the Euler-Maruyama scheme and the Barker scheme. This indicates the two schemes have similar weak convergence order, which is supported by our theory.

The code used for generating the above figure is included in the Appendix.

6.3.2 Poisson Random Effects Model

6.3.2.1 General Set Up

The following experiments compare the performance of ULA (the Euler-Maruyama scheme) and UB (the Barker scheme) while sampling from a Poisson random effects model, similar to the experiment in Section 6.3 of Livingstone and Zanella (2022).

The Poisson hierarchical model of consideration is of the following form:

$$\begin{aligned} y_{ij}|\eta_i &\overset{indep}{\sim} \text{Poi}(e^{\eta_i}) \qquad j = 1, \dots, n \\ \eta_i|\mu &\overset{indep}{\sim} \text{N}(\mu, 1) \qquad i = 1, \dots, 50 \\ \mu &\sim \text{N}(0, \sigma_\mu^2), \end{aligned}$$

where we test the two schemes on the task of sampling from the resulting posterior distribution $p(\mu, \eta_1, \dots, \eta_{50}|\boldsymbol{y})$ where $\boldsymbol{y} = (y_{ij})_{ij}$ is the observed data. We will then compare the two numerical schemes by measuring the qualities of samples of parameter μ using mean squared error (MSE) and (absolute) bias.

For our experiments using this model, we will assume the data-generating value of μ to be $\mu^* = 5$, and sample the η_i from their prior distributions.

Let N be the number of steps (after removing the burn-in) of the schemes, δ be the step size of the scheme, and S be the set of possible step sizes we are running the scheme at. We also denote K to be the number of iterations of the whole simulation.

For a fixed $\delta \in S$, we let μ_i^j, where $i = 1, 2, \ldots, N$ and $j = 1, 2, \cdots, K$, to denote the i-th sample of μ from the posterior for the j-th iteration, and $\overline{\mu}_N^j := \frac{1}{N}\sum_{i=1}^N \mu_i^j$ to be the sample mean of μ for the j-th iteration.

The MSE of the sample mean is MSE $= \frac{1}{K}\sum_{j=1}^K (\overline{\mu}_N^j - \mu^*)^2$, the (absolute) bias of the sample mean is Bias $= \frac{1}{K}\sum_{j=1}^K |\overline{\mu}_N^j - \mu^*|$.

We would also consider the sample variance and the sample quantile, and then we will study their respective MSEs. For sample variance of μ, it is $\overline{V}_{\mu N}^j := \frac{1}{N}\sum_{i=1}^N \left(\mu_i^j - \overline{\mu}_N^j\right)^2$. The MSEs are computed in a similar fashion as that of the sample mean.

6.3.2.2 MSE Comparison with Fixed Step Number

In this experiment, we let $S = \{0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04, 0.045, 0.05\}$,$N = 50000$, and $K = 100$. Then, we will compare the MSE of the two schemes at different step sizes.

We consider two ways to initialise the schemes. The first way is to start, for μ, right from the truth $\mu^* = 5$. This avoids the burn-in. The second way is to draw the starting point from $\mathrm{N}(5, 10^2)$, which is a warm start. This then requires us to remove the burn-in period of the simulated results. So, we run the schemes for 60000 steps and only take the last 50000 values. The choice to cut off the first 10000 is made after looking at the plot of samples at various step sizes and realising that all cases (approximately) are at stationarity after 10000 steps. Note that ULA tends to take a longer time to mix than UB for the same step size.

Start From Truth

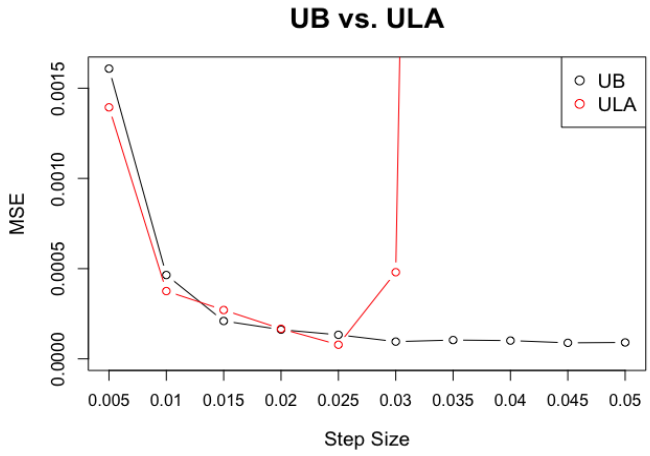


Figure 6.2: MSE Comparisons with Starting from Truth

Notice that ULA tends to behave badly for relatively large step sizes (0.035 onwards), while UB has stable performance for all possible step sizes in this simulation. The two schemes perform

similarly for small step sizes, and there is a decreasing trend of MSE as the step size increases when the step sizes are tiny.

Warm Start

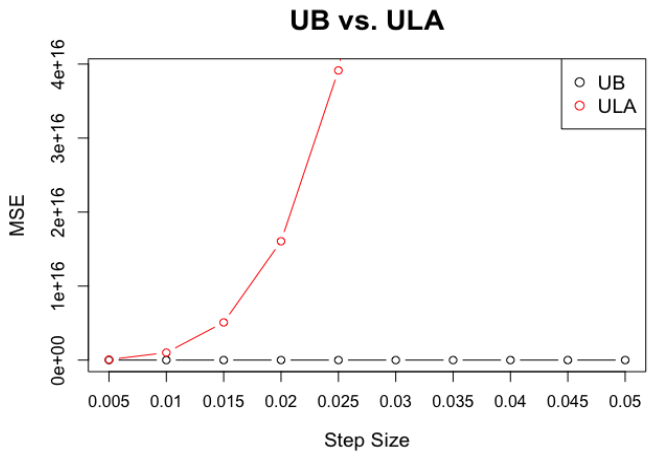


Figure 6.3: MSE Comparisons with Warm Start

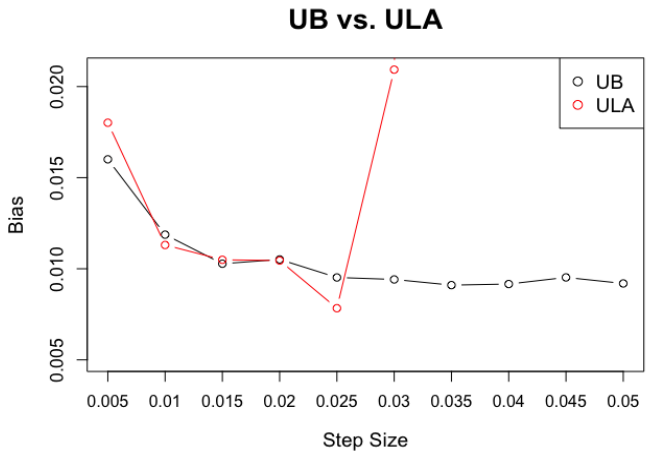
We see very similar behaviours of the schemes as the previous simulation that starts from the truth. In general, ULA has a much smaller range of possible step sizes for it to perform stably than UB. This 'robustness to tuning' property of the UB is a key feature of the Barker proposal algorithm too, as discussed in Livingstone and Zanella (2022).

6.3.2.3 Bias Comparison with Fixed Distance

In this experiment, we will consider the magnitude of error of the samples from the truth when the scheme has been run for a fixed length T with varying step sizes. Naturally, there would be a higher number of steps for smaller step sizes. Let $T = 1000$, $K = 100$, $S = \{0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04, 0.045, 0.05\}$, and $N = \lfloor T/\delta \rfloor$ for $\delta \in S$. Then, we will compare the (absolute) bias of the two schemes at different step sizes.

Similar to the previous set of simulations, we consider two ways to initialise the schemes - start from the truth and warm start. For the warm start, we draw the starting point from $\mathrm{N}(5, 3^2)$. We set $T = 1200$ instead, and remove the first one-sixth of the samples for the burn-in.

Start From Truth



The result for ULA aligns with the common strategy of using as small a step size as possible while running this scheme. The result for UB seems slightly different at a first glance, as the bias decreases at first and then stabilises. However, if we look at the actual values, there turns out to be very few differences for various step sizes. The comparison plot of the two schemes tells the same story as the comparison plots of previous simulations.

Warm Start

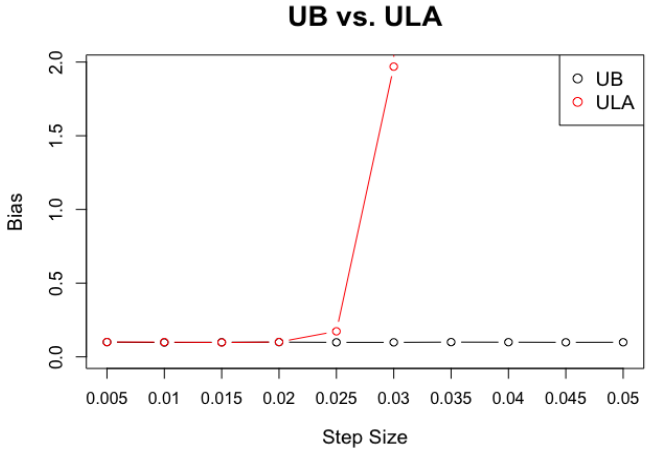


Figure 6.4: Bias Comparisons with Warm Start

The results are similar to the above case when we start from the truth. The unevenness of the UB result is only due to the tiny magnitudes of the biasses. This result indicates that it would be better to run the ULA with the smallest feasible step size, but this might not be the most efficient strategy for UB. Small step sizes would mean the schemes take more steps (thus more

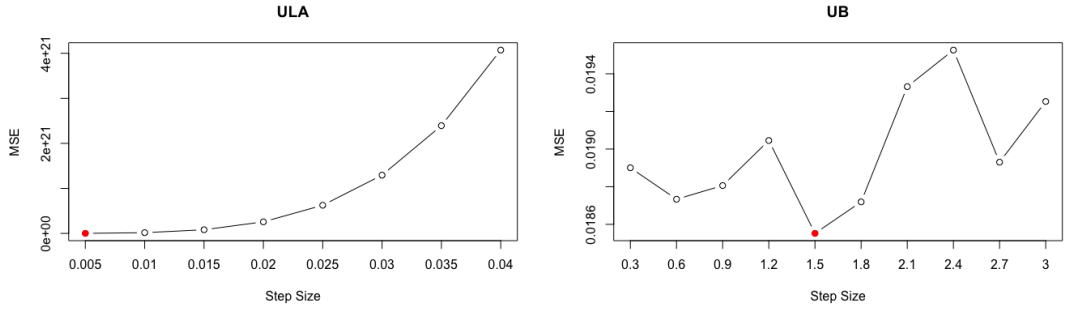
computationally costly) to travel the same distance. This increase in computational costs is taken usually in exchange for a reduced bias. For UB, the bias of the scheme remains low with larger step sizes (and therefore less computationally costly), so it would not be efficient to pick too small a step size.

6.3.2.4 MSE over Varying Step Size

The selection of step size (and parameter tuning in general) is a delicate matter in practice. Here, we compare the MSE of the two schemes for various step sizes. As hinted in the previous simulations, the range of reasonable step sizes for ULA is relatively small, compared to that of the UB. Thus, we consider two different sets of step sizes for this simulation. Additionally, we will only consider starting with a warm start.

For the UB simulation, we run the scheme with $T = 55000$ and remove the first 5000 samples for burn-in. Let $K = 100$ and $S = \{0.3, 0.6, 0.9, 1.2, 1.5, 1.8, 2.1, 2.4, 2.7, 3\}$. The starting point for μ is drawn from $\mathrm{N}(5, 10^2)$.

For the ULA simulation, we run the scheme with $T = 60000$ and remove the first 10000 samples for burn-in. Let $K = 100$ and $S = \{0.005, 0.010, 0.015, 0.020, 0.025, 0.030, 0.035, 0.040\}$. The starting point for μ is drawn from $\mathrm{N}(5, 10^2)$.



(a) ULA (b) UB

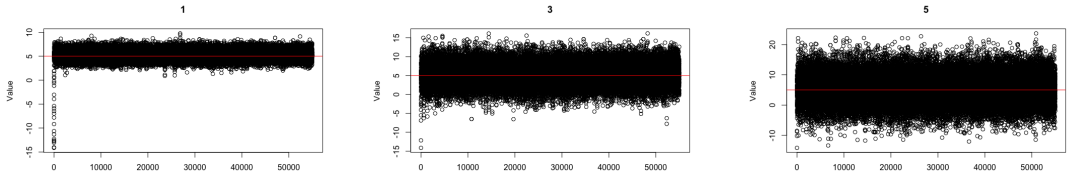
Figure 6.5: MSE over Varying Step Sizes

The result for ULA indicates that we should use the smallest possible step size, while the result for UB indicates that the choice of step size, within a sensible range, would all work very well. Here, the some of step sizes used for UB are rather outrageous, yet the scheme remains to perform well.

6.3.2.5 Variance and Quantile over Varying Step Sizes

In the previous experiment, we noticed that the MSE of the Barker proposal is extremely stable even when the step size is large. The stability over step sizes might be too good to be true at a first glance, and it is natural to suspect this is due to faulty codes rather than the algorithm itself. Here, we conduct further experiments on UB under the same setup to illustrate that this property is indeed valid.

Before looking at the variance and quantile of the estimation (of μ), we first take a look at the plot of the samples of UB for step sizes 1, 3, and 5.

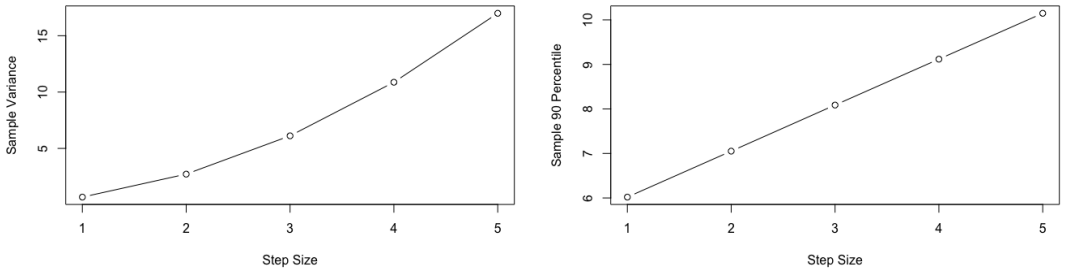


(a) Step Size = 1 (b) Step Size = 3 (c) Step Size = 5

Figure 6.6: Sample Plots for Different Step Sizes

The red horizontal line represents the true value of the parameter. Notice that the scheme managed to maintain sampling around the truth. As the step size increases, the fluctuation of the sample values begins to increase even though it is still centering around the truth. This reassures us that the scheme is correctly implemented.

Here, we consider the following two estimators - the variance of μ and the 90 percentile of μ.



(a) Sample Variance (b) Sample 90 Percentile

Figure 6.7: Variance and Quantile Varying Step Sizes

The above numerical results align with our observation from the sample path - as the step size increases the fluctuation of the sample increases.

Chapter 7

The Bouncy Barker

In this chapter, we will look at an alternative way to extend the one-dimensional Barker proposal to n-dimension. The extension is motivated by the recent work of Bouchard-Côté et al. (2018) where the movement of the samples is in the direction of the gradient, instead of moving in a Zig-Zag fashion (Bierkens and Roberts, 2017) as combinations of $e_1, e_2, \ldots, e_n$ where e_i is the n-vector with 1 at i-th coordinate and 0 elsewhere. Because of this motivation, we use the term 'bouncy' for this algorithm. We will denote the Barker proposal with a ZigZag-like update when the dimension is above one as **Barker**, and the alternative algorithm with a BPS-like update in high-dimension as **Bouncy Barker**.

7.1 Set-Up

When the dimension is one, Barker and Bouncy Barker provide the same proposal.

Algorithm One-Dimensional Proposal of Barker

Require: Target distribution π, current position x, symmetric density q

1: Draw $z \sim q(\cdot)$

2:

$$b = \begin{cases} +1 & \text{with probability } \frac{1}{1+\exp(-z\nabla \log \pi(x))} \\ -1 & \text{elsewise} \end{cases}$$

3: $y = x + b \cdot z$

Algorithm One-Dimensional Proposal of Bouncy Barker

Require: Target distribution π, current position x, symmetric density q

1: Draw $z \sim q(\cdot)$

2:

$$z' = \begin{cases} z & \text{with probability } \frac{1}{1+\exp(-z\nabla \log \pi(x))} \\ z - 2\frac{\langle \nabla \log \pi(x), z\rangle}{\|\nabla \log \pi(x)\|^2}\nabla \log \pi(x) & \text{elsewise} \end{cases}$$

3: $y = x + z'$

We can simplify the second step of Proposal of Bouncy Barker and get

$$z - 2\frac{\langle \nabla \log \pi(x), z\rangle}{\|\nabla \log \pi(x)\|^2}\nabla \log \pi(x) = z - 2\frac{\nabla \log \pi(x) \cdot z}{\|\nabla \log \pi(x)\|^2}\nabla \log \pi(x) = z - 2z = -z,$$

which coincides with that of the second step of the Proposal of Barker.

The two proposals, however, behave differently when the dimension is greater than 1.

Algorithm n-Dimensional Proposal of Barker

Require: Target distribution π, current position x, symmetric density q

1: Draw $z \sim q(\cdot)$

2: **for** $i = 1, 2, \ldots, n$ **do**

3:

$$b_i = \begin{cases} +1 & \text{with probability } \frac{1}{1+\exp(-z_i \partial_i \log \pi(x))} \\ -1 & \text{elsewise} \end{cases}$$

4: **end for**

5: $y = x + b \cdot z$

Algorithm n-Dimensional Proposal of Bouncy Barker

Require: Target distribution π, current position x, symmetric density q

1: Draw $z \sim q(\cdot)$

2:

$$z' = \begin{cases} z & \text{with probability } \frac{1}{1+\exp(-z^T \nabla \log \pi(x))} \\ z - 2\frac{\langle \nabla \log \pi(x), z\rangle}{\|\nabla \log \pi(x)\|^2}\nabla \log \pi(x) & \text{elsewise} \end{cases}$$

3: $y = x + z'$

7.2 Candidate Transition Kernel Derivation

When we try to implement the above algorithms, we would require a Metropolis adjustment step to decide if we will accept the proposal or not. The kernel of the proposal before passing through the Metropolis adjustment is called the **candidate transition kernel**. The candidate transition kernels of the proposals above are not symmetrical, so we could not cancel them out like in the case of RWM.

The candidate transition kernels of both one-dimensional and n-dimensional Barker are known. That of one-dimensional Bouncy Barker is known as well as it is the same as that of one-dimensional Barker. It turns out that under mild conditions on the distribution $q(\cdot)$ in Step 1 of the proposal, the candidate transition kernel of n-dimensional Bouncy Barker is very similar to that of n-dimensional Barker as well.

Let $j_{B_1}(x, y)$ denote the candidate transition kernel of one-dimensional Barker (and Bouncy

Barker). We have

$$\begin{aligned}
j_{B_1}(x,y) &= \frac{q(y-x)}{1+\exp(-(y-x)\nabla\log\pi(x))} + q(x-y)\left(1 - \frac{1}{1+\exp(-(x-y)\nabla\log\pi(x))}\right) \\
&= \frac{q(y-x)}{1+\exp(-(y-x)\nabla\log\pi(x))} + q(y-x)\left(\frac{\exp(-(x-y)\nabla\log\pi(x))}{1+\exp(-(x-y)\nabla\log\pi(x))}\right) \\
&= \frac{q(y-x)}{1+\exp(-(y-x)\nabla\log\pi(x))} + q(y-x)\left(\frac{1}{\exp(-(y-x)\nabla\log\pi(x))+1}\right) \\
&= \frac{2q(y-x)}{1+\exp(-(y-x)\nabla\log\pi(x))}.
\end{aligned}$$

So, the Metropolis adjustment $\alpha(x,y)$ is

$$\begin{aligned}
\alpha(x,y) &= \frac{\pi(y)}{\pi(x)} \cdot \frac{j_{B_1}(y,x)}{j_{B_1}(x,y)} \\
&= \frac{\pi(y)}{\pi(x)} \cdot \frac{2q(x-y)}{1+\exp(-(x-y)\nabla\log\pi(y))} \cdot \frac{1+\exp(-(y-x)\nabla\log\pi(x))}{2q(y-x)} \\
&= \frac{\pi(y)}{\pi(x)} \cdot \frac{1+\exp(-(y-x)\nabla\log\pi(x))}{1+\exp(-(x-y)\nabla\log\pi(y))}.
\end{aligned}$$

This can be easily adapted for the case of n-dimensional Barker. The candidate transition kernel and the Metropolis adjustment are

$$\begin{aligned}
j_{B_n}(x,y) &= \prod_i \frac{2q(y_i-x_i)}{1+\exp(-(y_i-x_i)\partial_i\log\pi(x))}, \\
\alpha(x,y) &= \frac{\pi(y)}{\pi(x)} \cdot \prod_i \frac{1+\exp(-(y_i-x_i)\partial_i\log\pi(x))}{1+\exp(-(x_i-y_i)\partial_i\log\pi(y))}.
\end{aligned}$$

To derive the candidate transition kernel of n-dimensional Bouncy Barker, we first define the function $b_x(z)$ that flips z in Step 2 of the proposal when the current position is x, i.e.

$$b_x(z) := z - 2\frac{\langle\nabla\log\pi(x), z\rangle}{\|\nabla\log\pi(x)\|^2}\nabla\log\pi(x).$$

Lemma 7.1. *b_x is an involution, i.e. $b_x(b_x(z)) = z$ for all z.*

Proof.

$$\begin{aligned}
&b_x(b_x(z)) \\
&= z - 2\frac{\langle\nabla\log\pi(x), z\rangle}{\|\nabla\log\pi(x)\|^2}\nabla\log\pi(x) - 2\frac{\left\langle\nabla\log\pi(x), \left(z - 2\frac{\langle\nabla\log\pi(x),z\rangle}{\|\nabla\log\pi(x)\|^2}\nabla\log\pi(x)\right)\right\rangle}{\|\nabla\log\pi(x)\|^2}\nabla\log\pi(x) \\
&= z - 2\frac{\langle\nabla\log\pi(x), z\rangle}{\|\nabla\log\pi(x)\|^2}\nabla\log\pi(x) - 2\frac{\langle\nabla\log\pi(x), z\rangle}{\|\nabla\log\pi(x)\|^2}\nabla\log\pi(x) \\
&\quad + 4\frac{\langle\nabla\log\pi(x), z\rangle}{\|\nabla\log\pi(x)\|^2} \cdot \frac{\langle\nabla\log\pi(x), \nabla\log\pi(x)\rangle}{\|\nabla\log\pi(x)\|^2}\nabla\log\pi(x) \\
&= z - 4\frac{\langle\nabla\log\pi(x), z\rangle}{\|\nabla\log\pi(x)\|^2}\nabla\log\pi(x) + 4\frac{\langle\nabla\log\pi(x), z\rangle}{\|\nabla\log\pi(x)\|^2}\nabla\log\pi(x) = z.
\end{aligned}$$

□

Lemma 7.2. *b_x is an isometry, i.e. for any $\alpha, \beta \in \mathbb{R}^n$, we have $\|\alpha - \beta\| = \|b_x(\alpha) - b_x(\beta)\|$.*

Remark. *An isometry $f : X \to Y$ with metrics d_X and d_Y for X and Y respectively is a map that satisfies $d_X(a, b) = d_Y(f(a), f(b))$ for all $a, b \in X$. Here, we are in the special case of $X = Y = \mathbb{R}^n$ with Euclidean metric.*

Proof. We first notice that b_x is linear, i.e. $b_x(\alpha) + b_x(\beta) = b_x(\alpha + \beta)$, as

$$\begin{aligned}
&b_x(\alpha) + b_x(\beta) \\
&= \alpha - 2\frac{\langle \nabla \log \pi(x), \alpha \rangle}{\|\nabla \log \pi(x)\|^2} \nabla \log \pi(x) + \beta - 2\frac{\langle \nabla \log \pi(x), \beta \rangle}{\|\nabla \log \pi(x)\|^2} \nabla \log \pi(x) \\
&= (\alpha + \beta) - 2\frac{\langle \nabla \log \pi(x), \alpha - \beta \rangle}{\|\nabla \log \pi(x)\|^2} \nabla \log \pi(x) \\
&= b_x(\alpha + \beta).
\end{aligned}$$

Moreover, $\|b_x(z)\|^2 = \|z\|^2$ for any z, as we have

$$\begin{aligned}
&\langle b_x(z), b_x(z) \rangle \\
&= \left\langle z - 2\frac{\langle \nabla \log \pi(x), z \rangle}{\|\nabla \log \pi(x)\|^2} \nabla \log \pi(x), z - 2\frac{\langle \nabla \log \pi(x), z \rangle}{\|\nabla \log \pi(x)\|^2} \nabla \log \pi(x) \right\rangle \\
&= \langle z, z \rangle - 2\langle \nabla \log \pi(x), z \rangle \frac{1}{\|\nabla \log \pi(x)\|^2} 2\langle \nabla \log \pi(x), z \rangle + 4\langle \nabla \log \pi(x), z \rangle^2 \frac{1}{\|\nabla \log \pi(x)\|^2} \\
&= \langle z, z \rangle.
\end{aligned}$$

Combining these two, we have $\|b_x(\alpha) - b_x(\beta)\|^2 = \|b_x(\alpha - \beta)\|^2 = \|\alpha - \beta\|^2$ for any α, β, meaning that b_x is indeed an isometry. □

So, if we let $j_{BB_n}(x, y)$ denote the candidate transition kernel of n-dimensional Bouncy Barker, we have

$$\begin{aligned}
j_{BB_n}(x, y) &= \frac{q(y - x)}{1 + \exp(-(y - x)^T \nabla \log \pi(x))} + q(b_x(y - x)) \left(1 - \frac{1}{1 + \exp(-b_x(y - x)^T \nabla \log \pi(x))}\right) \\
&= \frac{q(y - x)}{1 + \exp(-(y - x)^T \nabla \log \pi(x))} + q(b_x(y - x)) \frac{\exp(-b_x(y - x)^T \nabla \log \pi(x))}{1 + \exp(-b_x(y - x)^T \nabla \log \pi(x))}.
\end{aligned}$$

For any distribution q that is **spherically symmetric** about the origin and isometry i, we have $q(z) = q(i(z))$ for all z (Corollary of Theorem 4.1 of Fourdrinier et al. (2018)). This means, if q is spherically symmetric about the origin (e.g. a centred Gaussian distribution with covariance matrix being diagonal and having the same entries on the diagonal), we have $q(y - x) = q(b_x(y - x))$. This condition on q is not too outrageous in this scenario.

Additionally, we have $b_x(z)^T \nabla \log \pi(x) = -z^T \nabla \log \pi(x)$. To see this, we have

$$\begin{aligned}
&b_x(z)^T \nabla \log \pi(x) \\
&= z^T \nabla \log \pi(x) - \left[2\frac{\langle \nabla \log \pi(x), z \rangle}{\|\nabla \log \pi(x)\|^2} \nabla \log \pi(x)\right]^T \nabla \log \pi(x) \\
&= z^T \nabla \log \pi(x) - 2z^T \nabla \log \pi(x) \\
&= -z^T \nabla \log \pi(x),
\end{aligned}$$

as $\langle a, b\rangle = a^T b$ when $a, b \in \mathbb{R}^n$.

The two results above allow us to simplify the candidate transition kernel j_{BB_n} when q is spherically symmetric about the origin, and we have

$$\begin{aligned} j_{BB_n}(x,y) &= \frac{q(y-x)}{1+\exp(-(y-x)^T\nabla\log\pi(x))} + q(b_x(y-x))\frac{\exp(-b_x(y-x)^T\nabla\log\pi(x))}{1+\exp(-b_x(y-x)^T\nabla\log\pi(x))} \\ &= \frac{q(y-x)}{1+\exp(-(y-x)^T\nabla\log\pi(x))} + q(y-x)\frac{\exp((y-x)^T\nabla\log\pi(x))}{1+\exp((y-x)^T\nabla\log\pi(x))} \\ &= \frac{q(y-x)}{1+\exp(-(y-x)^T\nabla\log\pi(x))} + q(y-x)\frac{1}{\exp(-(y-x)^T\nabla\log\pi(x))+1} \\ &= \frac{2q(y-x)}{1+\exp(-(y-x)^T\nabla\log\pi(x))}. \end{aligned} \tag{7.1}$$

Consequently, the Metropolis adjustment of this proposal is

$$\alpha(x,y) = \frac{\pi(y)}{\pi(x)}\cdot\frac{j_{BB_n}(y,x)}{j_{BB_n}(x,y)} = \frac{\pi(y)}{\pi(x)}\cdot\frac{1+\exp(-(y-x)^T\nabla\log\pi(x))}{1+\exp(-(x-y)^T\nabla\log\pi(y))}.$$

7.3 Spectral Gap Bound

Using the transition kernels derived, we can obtain an upper bound for the spectral gap of the algorithm, which allows us to bound the variances of ergodic averages (Rosenthal, 2003). Consider the space the functions

$$L^2_{0,1}(\pi) = \{f : \mathbb{R}^d \to \mathbb{R} \mid \mathbb{E}_\pi[f] = 0, \mathrm{Var}_\pi[f] = 1\}.$$

A Markov chain of the Metropolis-Hastings type (i.e. generated by a Metropolis-Hastings algorithm) has the π-invariant kernel P, constructed by

$$P(x, dy) := \alpha(x,y)Q(x,dy) + r(x)\delta_x(dy),$$

where Q is a candidate kernel with density q, $\alpha(x,y)$ is the standard acceptance rate given by

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right),$$

and $r(x) := 1 - \int \alpha(x,y)Q(x,dy)$ is the average rejection probability.

The (right) **spectral gap** of a π-reversible Markov chain with transition kernel P is

$$\mathrm{Gap}(P) = \inf_{f\in L^2_{0,1}(\pi)} \frac{1}{2}\int [f(y)-f(x)]^2\pi(dx)P(x,dy).$$

The following result, established in the supplement of Livingstone and Zanella (2022) as Lemma 1.1, provides a way to obtain lower bounds on spectral gaps given point-wise lower bounds of the candidate kernel.

Lemma 7.3 (Lemma 1.1 in the supplement of Livingstone and Zanella (2022)). *Consider two Metropolis-Hastings kernels P_1 and P_2 with associated candidate kernels $Q_1(x, dy) = q_1(x, y)dy$ and $Q_2(x, dy) = q_2(x, y)dy$ and common target distribution π. If there is a $\gamma > 0$ such that $q_1(x, y) \geq \gamma q_2(x, y)$ for all fixed x, y with $x \neq y$, then*

$$Gap(P_1) \geq \gamma Gap(P_2).$$

With that, we could establish an upper bound on the bouncy Barker proposal. Let P^R be the candidate transition kernel of Random Walk Metropolis.

Proposition 7.4 (Original Result). *Let $\check{P}^{BB}$ denote the bouncy Barker proposal on $\mathbb{R}^d$ using Equation (7.1). Then, $Gap(P^R) \geq Gap(\check{P}^{BB})/2$.*

Proof. Let $q^R(x, x + z) = \mu_\sigma(z)$ be the candidate transition density of Random Walk Metropolis and μ_σ be spherically symmetric. The candidate density of bouncy Baker is, according to Equation (7.1),

$$\check{q}^{BB}(x, x + z) = \frac{2\mu_\sigma(z)}{1 + \exp(-z^T \nabla \log \pi(x))} = 2\mu_\sigma(z)\check{p}(x, z)$$

where $\check{p}(x, z) := 1/[1 + \exp(-z^T \nabla \log \pi(x))] \leq 1$. So, we have $q^R(x, x + z) \leq \check{q}^{BB}/2$, and thus $\text{Gap}(P^R) \geq \text{Gap}(\check{P}^{BB})/2$ using Lemma 7.3, as desired. □

Remark. *This result is identical to that of Proposition 5 in Livingstone and Zanella (2022) since the modified Barker has the same proposal kernel as bouncy Barker.*

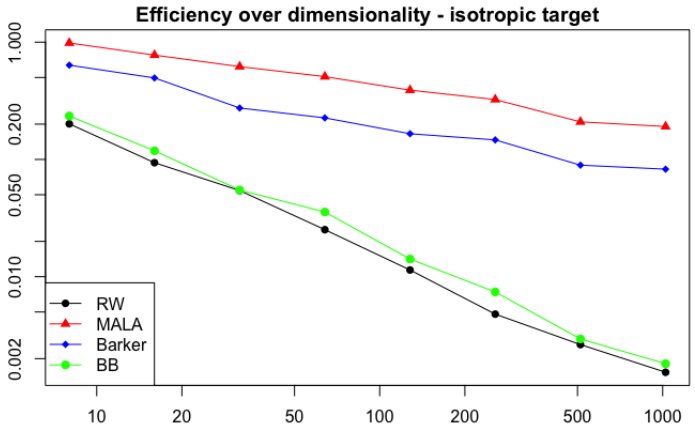


Figure 7.1: Efficiency Comparison Over Dimensionality

To visualise this result, we have the above diagram on the efficiency of various algorithms over dimensionality when we use spherically symmetric Gaussian as a proposal and an isotropic target. This is based on Figure 3 of Livingstone and Zanella (2022). The light green line, representing the bouncy Barker, is indeed similar to the Random Walk Metropolis, as established just now.

One possible explanation for the poor performance of bouncy Barker is that there are only two options at each proposal stage, regardless of the dimensionality of the target distribution. The

standard Barker, on the other hand, has 2^d options at each proposal stage when the dimension of the target is d. This larger number of options provides more flexible movements of the algorithm dynamics, implying a higher efficiency.

7.4 Discussion

Notice that the bouncy Barker, as well as the version of the Barker proposal with the same direction for each coordinate, has bad scaling over dimension property. Recall that the good scaling property is a result of the algorithm being locally-balanced, it would then not be too much of a stretch to believe that these two algorithms are no longer locally-balanced. And this is the case.

In Vogrinc et al. (2022), the authors proposed this general structure for the unnormalised proposal kernel for any first-order locally-balanced algorithm

$$\tilde{P}(x, dy) = \prod_{i=1}^{n} g(e^{(y_i - x_i)\partial_i \log \pi(x)}) \mu\left(\frac{dy_i - x}{\sigma}\right).$$

In the case of Barker with $g(t) = 1/(1+t^{-1})$, it is clear from the above equation that coordinates need to be independent. However, that requirement is not satisfied for the two failed attempts of extending the dimensions.

MALA is, in fact, a locally-balanced algorithm too, with $g(t) = \sqrt{t}$. This balancing function allows MALA to have good scaling properties even though the gradient information is not used component-wise. The derivation goes as follows:

$$\begin{aligned}
\prod_{i=1}^{n} g(e^{(y_i - x_i)\partial_i \log \pi(x)}) &= \prod_{i=1}^{n} e^{(y_i - x_i)\partial_i \log \pi(x)/2} \\
&= \exp\left[\sum (y_i - x_i)\partial_i \log \pi(x)/2\right] \\
&= \exp\left[\frac{1}{2}(y - x)^\top \nabla \log \pi(x)\right].
\end{aligned}$$

Chapter 8

Conclusion

The two main components of this thesis are on Langevin algorithms and Barker algorithms. We mostly focus on the theoretical properties of such algorithms and occasionally conduct numerical experiments to compare the performance of similar algorithms.

The part on Langevin algorithms revolves around the work of Dalalyan (2017), which established an explicit converge rate bound under the condition that the target distribution is log-concave and smooth. A natural extension would to be look at further work that improves on such bounds and similar bounds for other algorithms. Another natural extension would be to look at further work that weaken the condition on the target. There are more work conducted on the first direction, and less so on the second.

The part on Barker proposals consists mostly of original work. We look at two ways to extend the original Barker proposal introduced in Livingstone and Zanella (2022). The first way is to remove the Metropolis adjustment in the algorithm and link it to a scheme for solving SDEs numerically. This is the work of Chapter 6. The second way is to consider a seemingly new way (which we prove in this thesis that it is in fact identical to an existing way) to extend the Barker proposal to higher dimensions. This is the work of Chapter 7.

We intend to investigate both these two directions further in the future. For the work on the Barker scheme, we intend to extend the geometric ergodicity result (Theorem 6.1) to the case of n-dimension. We also aim to conduct more experiments on several complex models. Furthermore, we hope to look at other aspects (other than geometric ergodicity) of this numerical scheme, such as its weak convergence. For the work on Bouncy Barker, although the present extension is proved to be unsatisfactory, we intend to explore other ways for us to exploit both the bouncy-type moves as well as the skew-symmetric sampling of Barker proposal at the same time. This could bring us new insights into the Barker proposal.

Bibliography

Christophe Andrieu and Johannes Thoms. A tutorial on adaptive mcmc. *Statistics and Computing*, 18:343–373, 2008.

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50(1):5–43, 2003.

Christophe Andrieu, Anthony Lee, Sam Power, and Andi Q Wang. Explicit convergence bounds for metropolis markov chains: isoperimetry, spectral gaps and profiles. *arXiv preprint arXiv:2211.08959*, 2022.

Adelchi Azzalini and Giuliana Regoli. Some properties of skew-symmetric distributions. *Annals of the Institute of Statistical Mathematics*, 64:857–879, 2012.

Joris Bierkens and Gareth Roberts. A piecewise deterministic scaling limit of lifted metropolis–hastings in the curie–weiss model. *The Annals of Applied Probability*, 27(2):846–882, 2017.

Alexandre Bouchard-Côté, Sebastian J Vollmer, and Arnaud Doucet. The bouncy particle sampler: A nonreversible rejection-free markov chain monte carlo method. *Journal of the American Statistical Association*, 113(522):855–867, 2018.

Stephen P Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Sinho Chewi. *Log-Concave Sampling*. Unfinished Draft, 2023. https://chewisinho.github.io/main.pdf.

Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet, and Austin Stromme. Exponential ergodicity of mirror-langevin diffusions. *Advances in Neural Information Processing Systems*, 33:19573–19585, 2020.

Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Shunshi Zhang. Analysis of langevin monte carlo from poincare to log-sobolev. In *Conference on Learning Theory*, pages 1–2. PMLR, 2022.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

Persi Diaconis. The markov chain monte carlo revolution. *Bulletin of the American Mathematical Society*, 46(2):179–205, 2009.

Alain Durmus and Eric Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast! In *Conference on Learning Theory*, pages 793–797. PMLR, 2018.

Michael F Faulkner and Samuel Livingstone. Sampling algorithms in statistical physics: a guide for statistics and machine learning. *arXiv preprint arXiv:2208.04751*, 2022.

Dominique Fourdrinier, William E Strawderman, and Martin T Wells. *Shrinkage estimation*. Springer, 2018.

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

Max Hird, Samuel Livingstone, and Giacomo Zanella. A fresh take on 'barker dynamics' for mcmc. In *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC 2020, Oxford, United Kingdom, August 10–14*, pages 169–184. Springer, 2022.

Samuel Livingstone and Giacomo Zanella. The barker proposal: Combining robustness and efficiency in gradient-based mcmc. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 84(2):496, 2022.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2003.

Eckhard Platen and Peter Kloeden. *Numerical Solution of Stochastic Differential Equations*. Springer Berlin, Heidelberg, 1992.

Qian Qin and James P Hobert. On the limitations of single-step drift and minorization in markov chain convergence analysis. *The Annals of Applied Probability*, 31(4):1633–1659, 2021.

Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling for various metropolis-hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

Gareth O Roberts and Jeffrey S Rosenthal. General state space markov chains and mcmc algorithms. *Probability Surveys*, 1:20–71, 2004.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996a.

Gareth O Roberts and Richard L Tweedie. Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms. *Biometrika*, 83(1):95–110, 1996b.

Jeffrey S Rosenthal. Asymptotic variance and convergence rates of nearly-periodic markov chain monte carlo algorithms. *Journal of the American Statistical Association*, 98(461):169–177, 2003.

Timothy Sauer. *Numerical Analysis*. Addison-Wesley Publishing Company, 2011.

Jure Vogrinc, Samuel Livingstone, and Giacomo Zanella. Optimal design of the barker proposal and other locally-balanced metropolis-hastings algorithms. *arXiv preprint arXiv:2201.01123*, 2022.

David Williams. *Probability with Martingales*. Cambridge University Press, 1991.

Tatiana Xifara, Chris Sherlock, Samuel Livingstone, Simon Byrne, and Mark Girolami. Langevin diffusions and the metropolis-adjusted langevin algorithm. *Statistics & Probability Letters*, 91: 14–19, 2014.

Giacomo Zanella. Informed proposals for local mcmc in discrete spaces. *Journal of the American Statistical Association*, 115(530):852–865, 2020.

Appendix

The following is the R code used to generate Figure 6.1.

```
euler_scheme <- function(mu,sigma,delta,time, start){
  # the Euler-Maruyama scheme for a (one-dim) autonomous SDE
  # mu        drift term of the SDE
  # sigma     volatility term of the SDE
  # delta     step size
  # time      time of the simulation T = N / delta
  # start     starting point

  step <- floor(time * delta)
  curr <- start
  result <- c(curr)
  for(i in 1:step){
    new <- curr + delta * mu(curr) +
           sigma(curr) * rnorm(1,mean=0,sd=sqrt(delta))
    result <- c(result, new)
    curr <- new
  }
  return(result)
}

barker_scheme <- function(mu, sigma, delta, prob, time, start){
  # the Barker scheme for a (one-dim) autonomous SDE
  # mu        drift term of the SDE
  # sigma     volatility term of the SDE
  # delta     step size
  # prob      probability function for injection of skewness
  # time      time of the simulation T = N / delta
  # start     starting point

  step <- floor(time * delta)
  curr <- start
  result <- c(curr)
  for(i in 1:step){
    xi <- rnorm(1) * sqrt(delta) * sigma(curr)
    b_prob <- prob(mu=mu,sigma=sigma,delta=delta,curr=curr,xi=xi)
    if (runif(1) < b_prob){
      b <- 1
    }
    else{
```

```
      b <- -1
    }
    new <- curr + b*xi
    result <- c(result, new)
    curr <- new
  }
  return(result)
}

prob_cauchy <- function(mu, sigma, delta, curr, xi){
  # the Cauchy CDF probability function for injecting skewness
  # mu       drift term of the SDE
  # sigma    volatility term of the SDE
  # delta    step size
  # curr     current position
  if(sigma(curr) == 0){
    exponential <- "inf"
    prob <- 1
  }
  else{
    exp_power <- (2 * xi * mu(curr) / sigma(curr) )/ sigma(curr)
    prob <- 1/(1+exp(-exp_power))
  }
  return(prob)
}

# Ornstein-Uhlenbeck
# dX_t = theta(mu - X_t) dt + sigma dW_t
ornstein_uhlenbeck_drift <- function(curr){
  mu <- 0
  theta <- 1
  return(theta*(mu-curr))
}
ornstein_uhlenbeck_volatility <- function(curr){
  sigma <- sqrt(2)
  return(sigma)
}

mu <- 0
theta <- 1
sigma <- sqrt(2)
time <- 1000
delta_vec <- seq(0.01,0.1,0.01)
start <- 1

iter <- 10000
yt_barker <- c()
yt_euler <- c()

for (i in 1:iter){
  for (delta in delta_vec){
    barker_sample <- barker_scheme(mu = ornstein_uhlenbeck_drift,
        sigma = ornstein_uhlenbeck_volatility, delta = delta,
```

```
          prob = prob_cauchy, time = time, start = start)
    yt_barker <- c(yt_barker, barker_sample[length(barker_sample)])

    euler_sample <- euler_scheme(mu = ornstein_uhlenbeck_drift,
          sigma = ornstein_uhlenbeck_volatility, delta = delta,
          time = time, start = start)
    yt_euler <- c(yt_euler, euler_sample[length(euler_sample)])
  }
}
final_yt_barker <- t(matrix(yt_barker,ncol=iter))
final_yt_euler <- t(matrix(yt_euler,ncol=iter))

expected_solution <- start*exp(-theta * time) + mu * (1- exp(-theta*time))


f <- function(x){
  return(x^2)
}

log_error_barker <- log(colMeans(f(final_yt_barker)) - f(expected_solution))
log_error_euler <- log(colMeans(f(final_yt_euler)) - f(expected_solution))

# generate the plot
plot(log_error_barker,col="red",type="b",
     ylim=c(min(log_error_barker,log_error_euler),
     max(log_error_barker,log_error_euler)),xlab="log Step Size",
     ylab="log f(error)",xaxt="n",main="f(x) = x^2",log = 'x')
points(log_error_euler,col="black",type="b",log='x')
axis(1, at=1:10,labels=delta_vec)
legend("topleft", legend=c("Euler","Barker"), col=c("black", "red"),
     lty=c(1,1), cex=1)
```