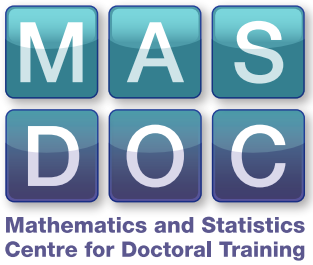


Stability and Examples of some Approximate MCMC Algorithms

by

Felipe Javier Medina Aguayo

Thesis

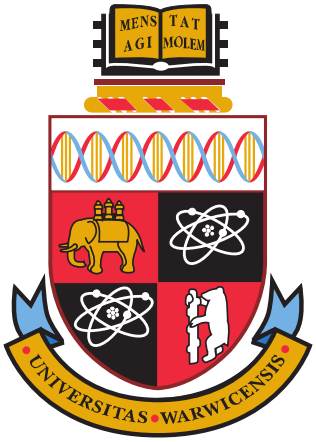
Submitted for the degree of

Doctor of Philosophy

Mathematics Institute

The University of Warwick

January 2017



Contents

List of Tables

List of Figures

Acknowledgments

First of all, I would like to thank my supervisors Anthony and Gareth for their invaluable guidance, endless patience and enthusiastic support throughout my years as a PhD student. I am indebted to them for all the fruitful discussions we had and for their continued encouragement. I consider myself very fortunate for having had the opportunity to work with them.

Many thanks also to my examiners Adam Johansen and Matti Vihola for all the time spent while reading and examining this thesis. Their useful comments and interest in my research have significantly improved some aspects of this work.

I am also grateful to all the people I met from Warwick who made my PhD an inspiring and memorable experience. Many thanks to my former professors and co-workers, without their encouragement and support I would not have pursued a postgraduate degree. Special thanks to all my friends, old and new, for all the enjoyable and amusing moments we spent during the last years.

I would also like to acknowledge the financial support received from the EPSRC-funded Centre for Research in Statistical Methodology (EP/D002060/1) and from Mexico's National Council for Science and Technology (CONACYT).

Asimismo, quisiera agradecer a toda mi familia por su apoyo incondicional. En especial, gracias a mi abuela por sus sabios consejos, a mi hermana por siempre creer en mí y a mis padres pues por ellos soy lo que soy. Finalmente, palabras no bastan para agradecer al amor de mi vida por todo su apoyo, paciencia y motivación durante este tiempo. Lo logramos!

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

The work presented here was carried out by myself, except where otherwise indicated. The contents of Chapter 3 and Section 4.1 appear in the following publication:

> F. J. Medina-Aguayo, A. Lee, and G. O. Roberts. Stability of noisy Metropolis–Hastings. *Statistics and Computing*, 26(6):1187–1211, 2016. ISSN 1573-1375. doi: 10.1007/s11222-015-9604-3.

Abstract

Approximate Monte Carlo algorithms are not uncommon these days, their applicability is related to the possibility of controlling the computational cost by introducing some noise or approximation in the method. We focus on the stability properties of a particular approximate MCMC algorithm, which we term noisy Metropolis-Hastings. Such properties have been studied before in tandem with the pseudo-marginal algorithm, but under fairly strong assumptions. Here, we examine the noisy Metropolis-Hastings algorithm in more detail and explore possible corrective actions for reducing the introduced bias. In this respect, a novel approximate method is presented, motivated by the class of exact algorithms with randomised acceptance. We also discuss some applications and theoretical guarantees of this new approach.

Chapter 1

Introduction

Monte Carlo algorithms are without doubt one of the most important class of methods that, together with modern computers, have modified the everyday practice of statistical inference. In particular, the appearance of Markov Chain Monte Carlo (MCMC) has been a milestone in several fields including physics, computer science, economics and of course statistics. The celebrated Metropolis-Hastings (MH) algorithm, originally conceptualised in Metropolis et al. (1953) and later refined in Hastings (1970), is possibly one of the most popular methods belonging to this class of algorithms. In words, an MCMC algorithm allows the construction of a Markov chain, by accepting or rejecting proposed moves, whose stationary probability distribution, say π, is some target of interest. In order to do this, the (almost) only requirement is the availability of the corresponding target density up to a constant of proportionality. Due to this, MCMC methods have found increasing application in Bayesian statistics, where the aforementioned target distribution is usually the posterior distribution of a parameter x with density

$$\pi(x) \equiv p(x|D) \propto p_0(x)\, l(x; D),$$

where $l(x; D)$ denotes the likelihood function given a set of observations D, and p_0 is some prior distribution for x.

However, in many statistical applications the target density $\pi(x)$ may be intractable or expensive to evaluate. By intractable we mean that an analytic expression is not available. Common sources of intractability arise when latent variables are used to model observed data, e.g. missing-data models or hierarchical models. For such cases, the likelihood may involve integrating out some unobserved process

Z, i.e.

$$l(x; D) = \int p(D|z, x)\, p(z|x)\, dz,$$

which is usually impossible to perform analytically. Furthermore, even for tractable likelihoods, the point-wise evaluation of $l(x; D)$ may be computationally expensive, e.g. when the size of D is huge. Therefore, in many of the previous settings standard Monte Carlo methods are infeasible and one must resort to more elaborate techniques.

One of these non-standard methods is the pseudo-marginal approach (see Beaumont, 2003 or Andrieu and Roberts, 2009), and for its implementation one requires non-negative unbiased estimates of the density $\pi(x)$. Similarly to a MH algorithm, the pseudo-marginal algorithm generates a Markov chain (now on an extended space), where the proposed moves are accepted or rejected using an appropriate probability ensuring π is again the stationary distribution (at least marginally). For this reason, the pseudo-marginal is considered an exact method. The remarkable property here is that point-wise evaluations of $\pi(x)$ are no longer needed, for this reason the pseudo-marginal has been object of recent study, see e.g. Andrieu et al. (2010), Andrieu and Vihola (2015), Andrieu and Vihola (2016), Doucet et al. (2015), Lyne et al. (2015), Maire et al. (2014) and Sherlock et al. (2015), Deligiannidis et al. (2015), Bardenet et al. (2015). Nevertheless, this approach comes with the trade-off of extra variability (due to the estimation of $\pi(x)$) that can affect drastically the mixing properties of the chain. In addition, it may be the case that unbiased estimates for the density of π are not at hand; hence the need for exploring alternative approaches, even if this means introducing some bias.

Approximate Monte Carlo algorithms are not uncommon these days, their applicability is related to the possibility of controlling the computational cost by introducing some noise or approximation in the method. One example is approximate Bayesian computation (ABC), belonging to the class of likelihood-free methods, which has been studied in some depth, see e.g. Marin et al. (2012). The theoretical study of approximate Markov chain methods is more recent, with a growing literature in the area. Examples include: Roberts et al. (1998) in the context of computer round-off error; Pillai and Smith (2014) and Rudolf and Schweizer (2015) for perturbations under the Wasserstein distance; the authors in Durmus and Moulines (2015), Dalalyan (2016), Alquier et al. (2014) and Teh et al. (2016) discuss approximate Langevin dynamics; approximations using sub-sampling ideas for large data sets are explored in Singh et al. (2012), Korattikara et al. (2014), Quiroz et al.

(2014), Maire et al. (2015) and Bardenet et al. (2015); more recently, the authors in Johndrow et al. (2015) investigate the computational time trade-off of some approximate MCMC methods.

A major part of this thesis focuses on the stability properties of a particular approximate MCMC algorithm, which we term noisy MH. In fact, the original pseudo-marginal algorithm (Beaumont, 2003) appeared as a modification of a specific noisy MH algorithm introduced in O'Neill et al. (2000). Nevertheless, the noisy MH algorithm has fundamentally different properties that have been studied in tandem with the pseudo-marginal by Beaumont (2003), Andrieu and Roberts (2009) and Alquier et al. (2014), but under fairly strong assumptions. The aim of this work is to examine the noisy MH algorithm in more detail and to explore possible corrective actions for reducing the introduced bias.

1.1 Outline and Contributions

The rest of this chapter contains a list of convention and notation used throughout the thesis. In Chapter 2 some of the background material essential for addressing the different chapters is presented. A brief review on Markov chains and geometric ergodicity is presented, together with some important aspects of the MH algorithm. We then discuss importance sampling and sequential Monte Carlo (SMC) algorithms, the latter in the context of hidden Markov models. Lastly, we introduce the pseudo-marginal and noisy MH algorithms, with a review of existing results and known properties of the corresponding chains.

Chapter 3 is joint work with Anthony Lee and Gareth Roberts and is based on the publication Medina-Aguayo et al. (2016). Fundamental stability properties of the noisy MH chain, like positive recurrence and geometric ergodicity, are discussed. Our main contributions are Theorems 3.1, 3.2 and 3.4, where we explore sufficient conditions for inheriting geometric ergodicity from a standard MH chain, as well as convergence of the approximate stationary distribution towards the true target. The relationship between our work and some results in Pillai and Smith (2014) and Rudolf and Schweizer (2015) is pointed out in subsequent remarks.

Chapter 4 is divided into 3 main sections. The first one, which is based also on the publication Medina-Aguayo et al. (2016), addresses a particular setting of the noisy MH algorithm arising from importance sampling. Sufficient conditions in this specific setting are explored leading to Proposition 4.5, which is a direct application of Theorem 3.4. The second section considers a different setting using SMC estimates for a specific hidden Markov model. Once again, sufficient conditions

for this particular case are investigated. The last section is devoted to the refinement and generalisation of some ideas presented in Chapter 3. Proposition 4.9 arises from these ideas, which is an improved version of Proposition 4.5.

In Chapter 5, we explore a novel method, described in Algorithm 5.1, which was motivated by a class of exact algorithms with randomised acceptance in an attempt to reduce the bias produced by the noisy MH algorithm. This new method is based on the penalty method (Ceperley and Dewing, 1999), which has become relevant in some recent work, e.g. in Andrieu and Vihola (2016), Sherlock et al. (2015), Deligiannidis et al. (2015) or Yıldırım (2016). We provide a section with some applications of the new approach, including sub-sampling for large data sets. In addition, theoretical guarantees are presented under some assumptions on the target and noise. Our main results are Theorem 5.4 and 5.6, which deal with a diffusion limit of the approximate chain and the convergence of the approximate stationary distribution.

Finally, in Chapter 6, we present a final discussion with a summary of our findings and with possible future research avenues.

1.2 Notation and Conventions

- For a measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, let $\mathcal{B}(\mathcal{X})$ denote the collection of Borel sets on $\mathcal{X}$. We write $\nu \ll \mu$ to say a measure ν is absolutely continuous with respect to μ. Throughout, we assume probability distributions have densities, denoted with the same letter, with respect to a dominating σ-finite measure μ. The Lebesgue measure is denoted by μ^{Leb}.

- For a probability distribution π on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, we denote expectations of functions $f : \mathcal{X} \to \mathbb{R}$ by

$$\pi(f) := \int_{\mathcal{X}} f(x)\, \pi(dx).$$

  Alternatively, if $X \sim \pi(\cdot)$ we also use the notation $\mathbb{E}[f(X)] \equiv \pi(f)$, and we define the variance of $f(X)$ as follows

$$\mathbb{V}[f(X)] := \mathbb{E}\left[(f(X) - \mathbb{E}[f(X)])^2\right] = \pi\left((f - \pi(f))^2\right).$$

- Let $\mathcal{N}\left(a, b^2\right)$ denote the Normal or Gaussian distribution with mean a and variance b^2. We use the letters ϕ and Φ for denoting the density function and cumulative distribution function, respectively, of a standard Gaussian

distribution $\mathcal{N}(0,1)$. Confusion should be avoided from the context when using $\mathbf{\Phi}$ for denoting a Markov chain.

- $Ber(p)$ denotes a Bernoulli distribution with success probability p, $Bin(N,p)$ denotes a binomial distribution of parameters N and p, $Mult(N,w)$ denotes a multinomial distribution of parameters N and w, $Exp(\lambda)$ denotes an exponential distribution of rate λ, $Beta(a,b)$ denotes a Beta distribution of parameters a and b, $\log\mathcal{N}\left(a,b^2\right)$ denotes a log-normal distribution of parameters a and b^2, and χ^2_d denotes a Chi-square distribution with d degrees of freedom.

- The Euclidean distance and the corresponding norm for vectors and matrices are denoted by $\|\cdot\|$. The total variation distance is denoted by $\|\cdot\|_{TV}$.

- Standard notation is used to denote the natural ($\mathbb{N}$), integer ($\mathbb{Z}$) and real ($\mathbb{R}$) numbers . The non-negative and positive real numbers are denoted by $\mathbb{R}_0^+$ and $\mathbb{R}^+$, respectively. Similarly, the non-negative integers are denoted by $\mathbb{N}_0$.

- We use $O(\cdot)$ and $o(\cdot)$ for the big-O and little-o notations, respectively. We also write for two functions f and g, defined on a common domain D,

$$f(x) \lesssim g(x),$$

  if there exists a constant $K > 0$ such that $f(x) \leq Kg(x)$ for any $x \in D$.

- Denote by C^2 the space of twice continuously differentiable functions, by C_c^∞ the set of smooth and compactly supported functions, and by $\mathbb{D}$ the space of càdlàg (right continuous with limits on the left) functions. For a function ψ defined on a topological space X, we define the support of ψ, denoted by $supp(\psi)$, to be the closure of the set of points $x \in X$ for which $\psi(x) = 0$.

- The gradient of a function f is denoted by ∇f, sometimes appearing with a subscript to avoid confusion with other variables. The corresponding Hessian matrix is denoted by $\nabla^2 f$.

- Let $\delta_x(\cdot)$ be the Dirac-delta distribution centred at x, and denote the indicator function over a set A by $\mathbb{1}(A)$.

- The n-dimensional zero vector is denoted by $0_n := (0,\ldots,0)$. The unit function over the product space $\mathcal{X}^T$ is denoted by $1_T : \mathcal{X}^T \to \mathbb{R}$, where $1_T(x) = 1$ for $x \in \mathcal{X}^T$.

Chapter 2

Preliminaries

In this chapter we introduce the probabilistic tools and background material required throughout the thesis. The first section is devoted to well-known results about Markov chains in general state spaces and the introduction of the Metropolis-Hastings algorithm. Geometric ergodicity and some important consequences are also discussed. The second section deals with importance sampling and sequential Monte Carlo methods, which will become relevant when dealing with intractability. In this respect, the final section introduces the pseudo-marginal and noisy MH chains, where existing work related to the ergodic properties of both chains is presented. This provides a stepping stone for the further analysis of the noisy Metropolis-Hastings chain in Chapters 3 and 4.

2.1 Markov Chain Monte Carlo

Suppose we want to simulate from a probability distribution π on a measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Since simulating directly from π is not always possible, the purpose of Markov Chain Monte Carlo (MCMC) is to create a Markov chain $\{\Phi_i\}_{i=0}^{\infty}$, on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, that converges in some sense towards π. In a Bayesian context, π is usually a posterior distribution whose density, also denoted by π and with respect to a dominating measure μ, only needs to be known up to a constant of proportionality.

In this section concepts and standard results about Markov chains are presented. These tools will be helpful for analysing properties of forthcoming MCMC algorithms and related approximations. In the last part of this section, the celebrated Metropolis-Hastings algorithm is introduced and discussed.

2.1.1 Markov Chains

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with filtration $(\mathcal{F}_n)_{n\geq 0}$ and consider a measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, where $\mathcal{X}$ is usually a subset of $\mathbb{R}^d$. A Markov Chain $\mathbf{\Phi} = \{\Phi_0, \Phi_1, \dots\}$ is an $\mathcal{X}$-valued stochastic process adapted to $(\mathcal{F}_n)_{n\geq 0}$ that satisfies, for every $A \in \mathcal{B}(\mathcal{X})$ and every $n \geq 0$,

$$\mathbb{P}\left[\Phi_{n+1} \in A \middle| \mathcal{F}_n\right] = \mathbb{P}\left[\Phi_{n+1} \in A \middle| \Phi_n\right], \qquad \mathbb{P}\text{-a.s.}$$

Our focus restricts to Markov chains with transitions that are independent of time. The definition of a transition kernel is required.

Definition 2.1 (Transition kernel)**.** Let $P : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0, 1]$ be a mapping such that:

(i). For each $A \in \mathcal{B}(\mathcal{X})$, $P(\cdot, A)$ is a measurable function on $\mathcal{X}$;

(ii). For each $x \in \mathcal{X}$, $P(x, \cdot)$ is a probability measure on $\mathcal{B}(\mathcal{X})$.

Then P is a transition probability kernel.

A Markov chain $\mathbf{\Phi}$ with transition kernel P and initial distribution μ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is time-homogeneous if, for any $n \geq 0$ and sets $A_0, \dots, A_n \in \mathcal{B}(\mathcal{X})$,

$$\begin{aligned}&\mathbb{P}\left[\Phi_0 \in A_0, \Phi_1 \in A_1, \dots, \Phi_n \in A_n\right] \\ &\qquad = \int_{A_0} \mu(dy_0) \int_{A_1} P(y_0, dy_1) \dots \int_{A_n} P(y_{n-1}, dy_n).\end{aligned}$$

By defining the n-step transition probability kernels

$$\begin{aligned}P^n(x, \cdot) &:= \int_{\mathcal{X}} P(x, dy)\, P^{n-1}(y, \cdot), \qquad \text{for } n \geq 1, \\ \text{and} \quad P^0(x, \cdot) &:= \delta_x(\cdot),\end{aligned}$$

a time-homogeneous Markov chain satisfies, for all $m, n \geq 0$ and any $A \in \mathcal{B}(\mathcal{X})$,

$$\mathbb{P}\left[\Phi_{n+m} \in A \middle| \Phi_n\right] = P^m(\Phi_n, A).$$

An important aspect about Markov chains, which is crucial to MCMC, is the long-term behaviour of the chain. Under some conditions, a Markov chain will converge to its unique invariant distribution, provided it admits one.

Definition 2.2 (Invariant distribution)**.** The probability measure π is an invariant distribution for a chain with transition kernel P if

$$\pi(A) = \int_{\mathcal{X}} \pi(dx) P(x, A), \tag{2.1}$$

for any $A \in \mathcal{B}(\mathcal{X})$.

A useful way for determining whether a Markov chain admits an invariant distribution is via reversibility, i.e. if it satisfies

$$\int_A \pi(dx) \int_B P(x, dy) = \int_B \pi(dy) \int_A P(y, dx), \tag{2.2}$$

for any $A, B \in \mathcal{B}(\mathcal{X})$. The above condition is commonly known as detailed balance and, by choosing $B = \mathcal{X}$, (2.2) implies (2.1).

The needed conditions for a Markov chain to converge are related to the concepts of aperiodicity and irreducibility.

Definition 2.3 (Aperiodicity and irreducibility)**.** A Markov chain with transition kernel P and invariant distribution π is:

(i). Aperiodic, if there are no $d \geq 2$ disjoint subsets $\mathcal{X}_1, \ldots, \mathcal{X}_d \subseteq \mathcal{X}$ such that $P\left(x, \mathcal{X}_{(i \pmod d)+1}\right) = 1$ for any $x \in \mathcal{X}_i$, where $\pi(\mathcal{X}_i) > 0$ for all $i \in \{1, \ldots, d\}$;

(ii). φ-irreducible, if there exists a non-zero measure φ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ such that for any $A \in \mathcal{B}(\mathcal{X})$ with $\varphi(A) > 0$, and all $x \in \mathcal{X}$ there exists $n \geq 1$ such that $P^n(x, A) > 0$.

Let $\|\cdot\|_{TV}$ denote the total variation norm given by

$$\|\mu\|_{TV} := \frac{1}{2} \sup_{|g| \leq 1} \left| \int \mu(dy) g(y) \right| = \sup_{A \in \mathcal{B}(\mathcal{X})} \mu(A), \tag{2.3}$$

where μ is any finite signed measure satisfying $\mu(\mathcal{X}) = 0$. The ergodic theorem for Markov chains can be now enunciated.

Theorem 2.1 (Ergodicity)**.** *Consider a φ-irreducible and aperiodic Markov chain $\mathbf{\Phi}$, with transition kernel P and invariant distribution π, then for π-a.e. initial state $\Phi_0 = x \in \mathcal{X}$*

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

A proof of the above result can be found in Roberts and Rosenthal (2004, Theorem 4). The idea is to show that φ-irreducibility implies the existence of sets where the chain regenerates (Meyn and Tweedie, 1994, Theorem 5.2.2), and then use a coupling argument for two chains with common transition kernel P, one with initial distribution π and the other starting at x. It is natural to ask whether the above result holds for any $x \in \mathcal{X}$ and not just π-a.e. A sufficient condition for this to happen is to impose Harris recurrence (see Meyn and Tweedie, 1994, Chapter 9). However, throughout the thesis, we restrict only to φ-irreducible and aperiodic chains.

Additionally, when using MCMC and other Monte Carlo methods, one is often interested in estimating $\pi(f) = \int_{\mathcal{X}} \pi(dx) f(x)$, for some π-integrable function $f : \mathcal{X} \to \mathbb{R}$. This is done by using ergodic averages of the form

$$S_n(f) := \frac{1}{n} \sum_{i=0}^{n-1} f(\Phi_i), \quad \text{for } n \geq 0.$$

Under the same conditions of Theorem 2.1, and provided $\pi(|f|) < \infty$, as $n \to \infty$

$$S_n(f) \xrightarrow{a.s.} \pi(f),$$

for π-a.e. starting point $\Phi_0 = x \in \mathcal{X}$. A proof of the above strong law of large numbers can be found in Meyn and Tweedie (1994, Theorem 17.1.2).

2.1.2 Geometric Ergodicity

Theorem 2.1 states the asymptotic convergence of a Markov chain, however this is merely a qualitative result. A powerful characterisation of the behaviour of a Markov chain is provided by geometric ergodicity, defined below. Geometrically ergodic Markov chains have a limiting invariant probability distribution, which they converge towards geometrically fast in total variation (Meyn and Tweedie, 2009, Chapter 15).

Definition 2.4 (Geometric ergodicity)**.** A φ-irreducible and aperiodic Markov chain $\mathbf{\Phi}$, with transition kernel P and invariant distribution π, is π-a.e. geometrically ergodic if there exists a π-a.e. finite function $V \geq 1$ and constants $\tau < 1$, $R < \infty$ such that

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{TV} \leq RV(x)\tau^n. \tag{2.4}$$

Remark 2.1. We say a Markov chain is uniformly ergodic if $\sup_x V(x) < \infty$ in (2.4).

Geometric ergodicity does not necessarily provide fast convergence in an absolute sense. For instance, consider cases in (2.4) where τ, or R, are "extremely" close to one, or "very large" respectively. Then, the decay of the total variation distance, though geometric, is not particularly fast (see Roberts and Rosenthal, 2004 for some examples). Nevertheless, geometric ergodicity is a useful tool when analysing reversible and non-reversible Markov chains, as will become apparent in further chapters. Additionally, geometric ergodicity is a desirable property since it can guarantee the existence of a Central Limit Theorem (CLT) for $S_n(f)$, see Roberts and Rosenthal (2004) for a general review. A proof of the following theorem can be found in Chan and Geyer (1994, Theorem 2) .

Theorem 2.2 (CLT)**.** *Let $\mathbf{\Phi}$ be a geometrically ergodic Markov chain with invariant distribution π and transition kernel P. Then, for π-a.e. initial state $\Phi_0 = x \in \mathcal{X}$ and any $f : \mathcal{X} \to \mathbb{R}$ with $\pi\left(|f|^{2+\delta}\right) < \infty$ for some $\delta > 0$, the asymptotic variance*

$$\sigma^2_{f,P} := \lim_{n\to\infty} n\mathbb{V}\left[S_n(f)\right] < \infty, \tag{2.5}$$

and

$$\sqrt{n}\left(S_n(f) - \pi(f)\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{f,P}\right).$$

Remark 2.2. For π-reversible Markov chains, the above result will hold for functions with $\pi\left(f^2\right) < \infty$ as proved in Roberts and Rosenthal (1997, Corollary 2.1).

Unfortunately, in many cases assessing geometric ergodicity is not a straightforward task. A possible approach is to assess whether a Markov chain satisfies a drift condition towards a small set, with the help of Foster-Lyapunov functions. For that, we first define the concept of small set.

Definition 2.5 (Small set)**.** Let P be the transition kernel of a Markov chain $\mathbf{\Phi}$. A subset $S \subseteq \mathcal{X}$ is small if there exists a positive integer n_0, $\varepsilon > 0$ and a probability measure ν on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ such that the following minorisation condition holds

$$P^{n_0}(x, \cdot) \geq \varepsilon\nu(\cdot), \qquad \text{for } x \in S. \tag{2.6}$$

Let $PV(x) := \int_{\mathcal{X}} P(x, dz)V(z)$, for a transition kernel P and function V. The following theorem, which is immediate from combining Roberts and Rosenthal (1997, Proposition 2.1) and Meyn and Tweedie (2009, Theorem 15.0.1), establishes the equivalence between geometric ergodicity and a geometric drift condition.

Theorem 2.3 (Geometric drift)**.** *Suppose that* $\mathbf{\Phi}$ *is a* φ*-irreducible and aperiodic Markov chain with transition kernel* P *and invariant distribution* π*. Then, the following statements are equivalent:*

(i). There exists a small set S*, constants* $\lambda < 1$ *and* $b < \infty$*, and a* π*-a.e finite function* $V \geq 1$ *satisfying the geometric drift condition*

$$PV(x) \leq \lambda V(x) + b\mathbb{1}(x \in S), \quad \text{for } x \in \mathcal{X}. \tag{2.7}$$

(ii). The chain is π*-a.e. geometrically ergodic, where the function* V *in* (2.4) *can be taken as in* (i)*.*

Finally, we introduce the concept of simultaneously geometrically ergodic chains, as in Roberts et al. (1998). This allows us to characterise a class of kernels satisfying a geometric drift condition as in (2.7) using the same V, S, λ and b. This concept will be crucial for addressing convergence of some chains in the following chapters.

Definition 2.6 (Simultaneous geometric ergodicity)**.** A class of Markov chain kernels $\{P_k\}_{k\in\mathcal{K}}$ is simultaneously geometrically ergodic if there exists a class of probability measures $\{\nu_k\}_{k\in\mathcal{K}}$, a measurable set $S \subseteq \mathcal{X}$, a real valued measurable function $V \geq 1$, a positive integer n_0 and positive constants ε, λ, b such that for each $k \in \mathcal{K}$:

(i). S is small for P_k, with $P_k^{n_0}(x, \cdot) \geq \varepsilon\nu_k(\cdot)$ for all $x \in S$;

(ii). the chain P_k satisfies the geometric drift condition in (2.7) with drift function V, set S and constants λ and b.

2.1.3 Metropolis-Hastings Algorithm

In general, MCMC methods require the introduction of an auxiliary kernel $q : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0, 1]$ for proposing moves along $\mathcal{X}$. Such moves will be accepted with certain probability that guarantees invariance under π. Assume π and q have densities (denoted with the same letters) with respect a dominating measure μ, and define the acceptance ratio by

$$r(x, y) := \frac{\pi(y)\, q(y, x)}{\pi(x)\, q(x, y)}.$$

The acceptance probability in the celebrated Metropolis-Hastings (MH) algorithm (see Metropolis et al., 1953 and Hastings, 1970) is given by

$$\alpha_{MH}(x,y) := \min\{1, r(x,y)\} = \min\left\{1, \frac{\pi(y)\, q(y,x)}{\pi(x)\, q(x,y)}\right\}, \tag{2.8}$$

implying the MH transition kernel can be expressed as follows

$$P_{MH}(x, dy) = q(x, dy)\alpha_{MH}(x,y) + \delta_x(dy)\rho_{MH}(x), \tag{2.9}$$

where

$$\rho_{MH}(x) := 1 - \int_{\mathcal{X}} q(x, dy)\alpha_{MH}(x,y)$$

denotes the rejection probability at state $x \in \mathcal{X}$.

It is straightforward to verify that π and P_{MH} satisfy detailed balance as in (2.2). A single draw from P_{MH} is described in Algorithm 2.1, noting that for its implementation the density of π needs to be known only up to a constant of proportionality.

Algorithm 2.1 Simulating from $P_{MH}(x, \cdot)$

1. Sample $Y \sim q(x, \cdot)$.
2. With probability $\alpha_{MH}(x, Y)$ defined in (2.8):
   return Y;
   otherwise:
   return x.

Aperiodicity and irreducibility are directly satisfied under specific conditions on π and q that are not uncommon in practice. The following result covers the case where the target and proposal have densities with respect the Lebesgue measure, denoted by μ^{Leb}.

Theorem 2.4. *Assume the density $\pi(x)$ is bounded away from 0 and ∞ on compact sets of $\mathcal{X}$. In addition, suppose there exist $\delta_q > 0$ and $\varepsilon_q > 0$ such that, for every $x \in \mathcal{X}$,*

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

Then, a MH chain targeting π with proposal q is μ^{Leb}-irreducible, aperiodic and

every non-empty compact set is small.

Remark 2.3. A proof of the above result can be found in Roberts and Tweedie (1996a, Theorem 2.2), noting in there that the minorisation condition in (2.6) is merely attained by the sub-kernel $q(x, dy)\alpha_{MH}(x, y)$.

Respecting the proposal distribution, several options are available. A special case, which will be relevant in Chapter 5, is to choose a symmetric proposal with density satisfying

$$q(x, y) = q(\|y - x\|), \tag{2.10}$$

e.g. a Gaussian random walk. For this case, the acceptance ratio reduces to

$$r(x, y) = \frac{\pi(y)}{\pi(x)},$$

and the resulting algorithm is commonly known as random-walk Metropolis, due to the seminal paper Metropolis et al. (1953). Other possibilities include the independence sampler, for which $q(x, y) = q(y)$; or the Langevin algorithm, where

$$Y \sim \mathcal{N}(x + \delta/2\nabla \log \pi(x), \delta), \qquad \text{for some } \delta > 0,$$

motivated by a discretisation of the Langevin diffusion (Roberts and Tweedie, 1996b).

Furthermore, the MH algorithm is not the only MCMC method that targets π. As noted in Hastings (1970), the acceptance probability in (2.8) can be replaced by any acceptance of the form

$$\alpha_s(x, y) = s(x, y)(1 + r(y, x))^{-1}, \tag{2.11}$$

where $s : \mathcal{X}^2 \to \mathbb{R}_0^+$ is a symmetric function satisfying $s(x, y) \leq 1 + r(y, x)$ for any $(x, y) \in \mathcal{X}^2$. Any Markov chain generated using α_s satisfies detailed balance in (2.2) since, for any $(x, y) \in \mathcal{X}^2$,

$$\frac{\alpha_s(x, y)}{\alpha_s(y, x)} = r(x, y). \tag{2.12}$$

Clearly, the MH acceptance falls into this category by choosing

$$s_{MH}(x, y) = 1 + \min\{r(x, y), r(y, x)\}.$$

If instead $s_B(x,y) \equiv 1$, the resulting acceptance α_B is commonly known as Barker's acceptance (Barker, 1965). Later on, in Chapter 5, we will discuss a different approach for obtaining α_B, involving randomised acceptance probabilities.

Finally, a useful way for assessing the optimality among two different π-reversible Markov chains is to consider Peskun's off-diagonal ordering (see e.g. Peskun, 1973 and Tierney, 1994). In the MCMC context, let α_{s_1} and α_{s_2} be two acceptance probabilities of the form (2.11) targeting π with common proposal kernel q, and let P_{s_1} and P_{s_2} be the corresponding π-reversible MCMC kernels. If $\alpha_{s_1}(x,y) \geq \alpha_{s_2}(x,y)$ for all $x \neq y$ where $(x,y) \in \mathcal{X}^2$, then the corresponding asymptotic variances as in (2.5), for $f : \mathcal{X} \to \mathbb{R}$ with $\pi\left(f^2\right) < \infty$, satisfy

$$\sigma^2_{f,P_{s_1}} \leq \sigma^2_{f,P_{s_2}}.$$

Therefore, for fixed target π and proposal q, the MH acceptance (denoted from now on simply by α) is the optimal choice in terms of asymptotic variance among other acceptance probabilities of the form (2.11).

2.2 Estimating intractable densities

We now look at a different approach for estimating the expectation of a function $f : \mathcal{X} \to \mathbb{R}$ under some probability distribution π on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, i.e.

$$\pi(f) = \int_{\mathcal{X}} f(x)\, \pi(dx). \tag{2.13}$$

In contrast with the previous section, where we constructed a Markov chain that admits π as invariant distribution, we consider appropriately weighted samples from some auxiliary distribution q.

Later on, the approaches discussed in this section will prove useful when dealing with an intractable distribution, i.e. when an analytical expression for its density is not available. The two main cases we consider are importance sampling and sequential Monte Carlo methods.

2.2.1 Importance Sampling

If simulating from π is possible, a natural estimator of (2.13) is given by the naive Monte Carlo estimator

$$MC_{\pi,N}(f) := \frac{1}{N}\sum_{i=1}^{N} f(X_i),$$

where $\{X_i\} \overset{i.i.d}{\sim} \pi(\cdot)$. This estimator is clearly unbiased and by the law of large numbers $MC_{\pi,N}(f) \overset{a.s.}{\to} \pi(f)$, provided f is π-integrable. If in addition $\pi\left(f^2\right) < \infty$, then

$$\mathbb{V}\left[MC_{\pi,N}(f)\right] = \frac{1}{N}\mathbb{V}\left[f(X_1)\right],$$

which guarantees the existence of a CLT and the construction of approximate confidence intervals for $MC_{\pi,N}(f)$.

However, as discussed in the previous section, simulating from π is not always plausible. Importance sampling may prove useful when it is possible to simulate from an auxiliary distribution q such that $\pi \ll q$. In such case, and assuming π and q have densities with respect to a dominating measure μ, (2.13) can be expressed as

$$\pi(f) = \int_{\mathcal{X}} f(x)\frac{\pi(x)}{q(x)}q(dx).$$

This leads to the importance sampling estimator

$$IS_{\frac{\pi}{q},N}(f) := MC_{q,N}\left(f\frac{\pi}{q}\right) = \frac{1}{N}\sum_{i=1}^{N} f(X_i)\frac{\pi(X_i)}{q(X_i)}, \tag{2.14}$$

where now $\{X_i\} \overset{i.i.d}{\sim} q(\cdot)$.

When the ratio π/q is only known up to a constant of proportionality, a self-normalised version of the above estimator should be preferred. Let $w_{\frac{\pi}{q}}(x)$, or simply $w(x)$, denote the computable and proportional version of π/q, the self-normalised importance sampling estimator is given by

$$\tilde{IS}_{\frac{\pi}{q},N}(f) := \frac{\sum_{i=1}^{N} f(X_i)\, w(X_i)}{\sum_{i=1}^{N} w(X_i)}.$$

Even though this self-normalised version is not unbiased, by the law of large numbers it still converges towards $\pi(f)$. Additionally, as noted in Robert and Casella (2013), for some settings the estimator $\tilde{IS}$ can perform better since its bias may be negligible

and its variance may be smaller when compared to IS.

Clearly, the choice of the auxiliary distribution q plays a fundamental role in the effectiveness of the importance sampling estimator. For instance, the authors in Robert and Casella (2013) show that its variance may be infinite if the ratio of densities π/q is unbounded. From a practical point of view, and for the estimator in (2.14), the authors suggest to look for distributions q for which, if $X \sim q(\cdot)$, then

$$|f(X)| \frac{\pi(X)}{q(X)}$$

is almost constant and with finite variance.

2.2.2 Sequential Monte Carlo

Estimating expectations of the form (2.13) may become difficult using importance sampling. In particular, coming up with good proposals may be challenging if π has a complicated form. Sequential Monte Carlo (SMC) methods (see e.g. Doucet et al., 2001) are a convenient approach when π can be decomposed in a sequence of distributions, for which importance sampling can be implemented at each step. We only develop this idea in the context of hidden Markov models, see e.g. Doucet et al. (2000) or Cappé et al. (2005) for a more comprehensive study. Nevertheless, SMC methods can be applied in more generality as described in the monograph Del Moral (2004).

A hidden Markov model (HMM) comprises two processes $\{X_n\}_{n\geq 0}$ and $\{Y_n\}_{n\geq 1}$ on measurable spaces $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$, respectively. The random variables $\{X_n\}_{n\geq 0}$ form a time-homogeneous Markov chain with transition kernel m_θ that depends on a set of parameters $\theta \in \Theta$, and for simplicity we assume throughout that the initial value $X_0 = x_0$ is known. The observed random variables $\{Y_n\}_{n\geq 1}$ are conditionally independent given the unobserved $\{X_n\}_{n\geq 1}$ and distributed according to

$$Y_n | X_n \sim g_\theta(\cdot | X_n),$$

which may also depend on θ. Figure 2.1 summarises a generic HMM.

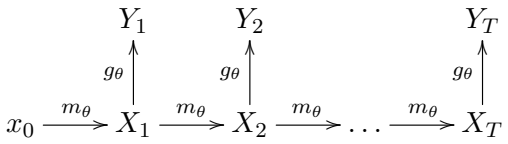


Figure 2.1: Hidden Markov Model

Let $\pi_{\theta,y_{1:T}}$ be the θ-dependent conditional distribution of $\{X_n\}_{n=1}^T$ given a fixed set of observations $y_{1:T} := \{y_n\}_{n=1}^T$. Usually, we are interested in estimating expectations of the form

$$\pi_{\theta,y_{1:T}}(f) = \frac{\int_{\mathcal{X}^T} f(x_{1:T}) \prod_{n=1}^T m_\theta(x_{n-1}, dx_n) \prod_{n=1}^T g_\theta(y_n|x_n)}{l(\theta; y_{1:T})}, \tag{2.15}$$

for some function $f : \mathcal{X}^T \to \mathbb{R}$, where $l(\theta; y_{1:T})$ denotes the likelihood for θ and given by

$$l(\theta; y_{1:T}) := \int_{\mathcal{X}^T} \prod_{n=1}^T m_\theta(x_{n-1}, dx_n) \prod_{n=1}^T g_\theta(y_n|x_n). \tag{2.16}$$

For simplicity, define the following measure on $\left(\mathcal{X}^T, \mathcal{B}\left(\mathcal{X}^T\right)\right)$

$$\gamma_{\theta,T}(A) := \int_A \prod_{n=1}^T m_\theta(x_{n-1}, dx_n) \prod_{n=1}^T g_\theta(y_n|x_n), \qquad \text{for } A \in \mathcal{B}\left(\mathcal{X}^T\right), \tag{2.17}$$

leading to a simplification of (2.15). This simply becomes

$$\pi_{\theta,y_{1:T}}(f) = \frac{\gamma_{\theta,T}(f)}{\gamma_{\theta,T}(1_T)},$$

where $1_T : \mathcal{X}^T \to \mathbb{R}$ is the unit function, noting also that the likelihood for θ reduces to

$$l(\theta; y_{1:T}) = \gamma_{\theta,T}\left(\mathcal{X}^T\right) = \gamma_{\theta,T}(1_T).$$

For the simple case where m_θ and g_θ follow Gaussian distributions, an explicit form for the likelihood $l(\theta; y_{1:T})$ is available since the integrals in $\gamma_{\theta,T}(1_T)$ can be computed analytically in a recursive manner. The involved recursions are usually known as the Kalman filter due to the seminal paper Kalman (1960). However, for non-Gaussian models or for complicated f, computing $\pi_{\theta,y_{1:T}}(f)$ will require solving complicated and high-dimensional integrals. One possible approach is to use self-normalised importance sampling with the help of an auxiliary distribution $q_{\theta,y_{1:T}}$ (depending possibly also on the initial state x_0), which leads to the following approximation

$$\tilde{IS}_{\frac{\pi_{\theta,y_{1:T}}}{q_{\theta,y_{1:T}}},N}(f) = \sum_{i=1}^N f\left(X_{1:T}^{(i)}\right) \frac{w\left(X_{1:T}^{(i)}\right)}{\sum_{j=1}^N w\left(X_{1:T}^{(j)}\right)},$$

where $\left\{X_{1:T}^{(i)}\right\} \overset{i.i.d}{\sim} q_{\theta,y_{1:T}}(\cdot)$ and

$$w(x_{1:T}) := \frac{\gamma_{\theta,T}(x_{1:T})}{q_{\theta,y_{1:T}}(x_{1:T})} = \frac{\prod_{n=1}^{T} m_\theta(x_{n-1},x_n)\prod_{n=1}^{T} g_\theta(y_n|x_n)}{q_{\theta,y_{1:T}}(x_{1:T})}$$
$$= \gamma_{\theta,T}(1_T)\frac{\prod_{n=1}^{T} m_\theta(x_{n-1},x_n)\prod_{n=1}^{T} g_\theta(y_n|x_n)}{\gamma_{\theta,T}(1_T)\, q_{\theta,y_{1:T}}(x_{1:T})} \propto \frac{\pi_{\theta,y_{1:T}}(x_{1:T})}{q_{\theta,y_{1:T}}(x_{1:T})}.$$

2.2.2.1 Bootstrap

Choosing an adequate $q_{\theta,y_{1:T}}$ may be a difficult task. A possibility is to construct such proposal sequentially, which may be useful as more observations become available. For doing this, as noted in Doucet and Johansen (2009), the auxiliary densities must satisfy

$$q_{\theta,y_{1:n}}(x_{1:n}) = q_{\theta,y_{1:n-1}}(x_{1:n-1})\, m'_{\theta,y_n}(x_{n-1},x_n), \qquad \text{for } n \geq 2,$$

where $m'_{\theta,y_n}(x_{n-1},\cdot)$ is an auxiliary proposal distribution. This implies that, for $n \geq 2$, the weights can be expressed as follows

$$w(x_{1:n}) = \frac{\gamma_{\theta,n}(x_{1:n})}{q_{\theta,y_{1:n}}(x_{1:n})}$$
$$= \frac{\prod_{m=1}^{n-1} m_\theta(x_{m-1},x_m)\prod_{m=1}^{n-1} g_\theta(y_m|x_m)}{q_{\theta,y_{1:n-1}}(x_{1:n-1})}\left(\frac{m_\theta(x_{n-1},x_n)\, g_\theta(y_n|x_n)}{m'_{\theta,y_n}(x_{n-1},x_n)}\right)$$
$$= w(x_{1:n-1})\underbrace{\left(\frac{m_\theta(x_{n-1},x_n)\, g_\theta(y_n|x_n)}{m'_{\theta,y_n}(x_{n-1},x_n)}\right)}_{=:\alpha_n\left(x_{(n-1):n}\right)}.$$

This method is commonly known as Sequential Importance Sampling (SIS) since the trajectories $\left\{X_{1:n}^{(i)}\right\}_i$, also known as particles, are sequentially propagated using the auxiliary proposals $\left\{m'_{\theta,y_n}\right\}_n$. For what follows, we restrict to the special case where $m'_{\theta,y_n} = m_\theta$, but more sensible choices for m'_{θ,y_n} should be preferred whenever possible. For this case, the incremental weight reduces to

$$\alpha_n\left(x_{(n-1):n}\right) = \alpha_n(x_n) = \frac{m_\theta(x_{n-1},x_n)\, g_\theta(y_n|x_n)}{m'_{\theta,y_n}(x_{n-1},x_n)} = g_\theta(y_n|x_n).$$

Although SIS may prove useful in some scenarios, it is a well known fact that it suffers from weight-degeneracy, meaning a small number of weights in $\left\{w\left(x_{1:n}^{(i)}\right)\right\}_i$ will dominate all the others as n increases. In order to overcome the degeneracy

behaviour, a resampling procedure for the particles may be introduced before the propagating step. The idea is to duplicate those trajectories whose weights are non-negligible. For this, we consider the number of offspring coming from the particle $X_{1:n}^{(i)}$, denoted by $N_n^{(i)}$, and obtain N equally-weighted resampled particles $\left\{\tilde{X}_{1:n}^{(i)}\right\}_i$. The offspring variables $\left\{N_n^{(i)}\right\}_i$ can be selected according to a multinomial distribution of parameters N and vector probabilities $\tilde{w}_n^{1:N} := \left(\tilde{w}_n^{(1)}, \ldots, \tilde{w}_n^{(N)}\right)$, where

$$\tilde{w}_n^{(i)} := \frac{\alpha_n\left(x_n^{(i)}\right)}{\sum_{j=1}^N \alpha_n\left(x_n^{(j)}\right)} = \frac{g_\theta\left(y_n \middle| x_n^{(i)}\right)}{\sum_{j=1}^N g_\theta\left(y_n \middle| x_n^{(j)}\right)}. \tag{2.18}$$

This modification to SIS, with the choice $m'_{\theta,y_n} = m_\theta$, is commonly known as the bootstrap filter (Gordon et al., 1993) and is summarised in Algorithm 2.2.

Algorithm 2.2 Bootstrap Filter

For each $n \in \{1, \ldots, T\}$:

1. For each $i \in \{1, \ldots, N\}$:
   (a) If $n = 1$:
   sample $X_1^{(i)} \sim m_\theta(x_0, \cdot)$;
   otherwise:
   sample $X_n^{(i)} \sim m_\theta\left(\tilde{X}_{n-1}^{(i)}, \cdot\right)$ and set $X_{1:n}^{(i)} = \left(\tilde{X}_{1:n-1}^{(i)}, X_n^{(i)}\right)$.
   (b) Compute $\tilde{w}_n^{(i)}$ using (2.18).
2. Resample N particles from $\left\{X_{1:n}^{(i)}\right\}_i$ according to a $Mult\left(N, \tilde{w}_n^{1:N}\right)$ to obtain $\left\{\tilde{X}_{1:n}^{(i)}\right\}_i$.

Estimate (2.15) using

$$\pi_{\theta,y_{1:T}}^N(f) := \sum_{i=1}^N f\left(X_{1:T}^{(i)}\right)\tilde{w}_T^{(i)}.$$

The resampling step comes with the cost of additional variance. In practice, one may implement variants of this algorithm, performing a resampling step only when the variance of the weights exceeds some predefined threshold (see e.g. Jun S. Liu, 1998). Additionally, the resampling process does not necessarily need to be done by means of the multinomial distribution. Other approaches are possible (see e.g. Douc and Cappé, 2005 or Li et al., 2015 for a more recent overview), as long as

the following condition is satisfied

$$\mathbb{E}\left[N_n^{(i)} \left| \left\{\tilde{w}_n^{(i)}\right\}_i \right.\right] = N\tilde{w}_n^{(i)}.$$

Such "unbiased" condition guarantees that unbiased estimators remain unbiased after the resampling step.

Algorithm 2.2 does not only provide a way for estimating (2.15) using $\pi^N_{\theta,y_{1:T}}$, it also provides unbiased estimators of the normalising constants $\gamma_{\theta,n}(1_n) = l(\theta; y_{1:n})$ at each step using

$$\gamma^N_{\theta,n}(1_n) := \prod_{m=1}^{n}\left\{\frac{1}{N}\sum_{j=1}^{N}\alpha_m\left(X_m^{(j)}\right)\right\} = \prod_{m=1}^{n}\left\{\frac{1}{N}\sum_{j=1}^{N} g_\theta\left(y_m \left| X_m^{(j)}\right.\right)\right\}, \tag{2.19}$$

see Del Moral (2004, Proposition 7.4.1) or Pitt et al. (2012).

Later on, we will pay special attention to the estimator in (2.19) since it provides an unbiased estimator for $l(\theta; y_{1:T})$ in intractable settings. For our purposes, bounding the relative variance of $\gamma^N_{\theta,T}(1_T)$ will be of particular interest. In Cérou et al. (2011), the authors show that under strong mixing conditions (basically requiring uniform upper and lower bounds on the densities of m_θ and g_θ), and for N sufficiently large

$$\mathbb{V}\left[\frac{\gamma^N_{\theta,T}(1_T)}{\gamma_{\theta,T}(1_T)}\right] \leq \frac{C_\theta(T)}{N}, \tag{2.20}$$

where the constant $C_\theta(T)$ is linear with T.

Finally, it is worth pointing out that there exists an extensive literature addressing error bounds and convergence results for $\pi^N_{\theta,y_{1:T}}$ and $\gamma^N_{\theta,n}$, see e.g. Del Moral (2004, Chapters 7 and 9), Chopin (2004), Douc et al. (2014) or Whiteley (2013). However, this is beyond the scope of this thesis.

2.3 The Pseudo-Marginal and Noisy Metropolis-Hastings

The aim is to simulate from an intractable probability distribution π for some random variable X, which takes values in a measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Suppose as before, π has a density $\pi(x)$ with respect to some reference measure μ, e.g. the counting or the Lebesgue measure. By intractable we mean that an analytical expression for the density $\pi(x)$ is not available and so implementation of a Markov chain Monte Carlo (MCMC) method targeting π is not straightforward.

One possible solution to this problem is to target a different distribution on the extended space $(\mathcal{X} \times \mathcal{W}, \mathcal{B}(\mathcal{X}) \times \mathcal{B}(\mathcal{W}))$, which admits π as marginal distribution. The pseudo-marginal algorithm (see Beaumont, 2003 or Andrieu and Roberts, 2009) falls into this category since it is a Metropolis–Hastings (MH) algorithm targeting a distribution $\bar{\pi}_N$ defined on the product space $(\mathcal{X} \times \mathcal{W}, \mathcal{B}(\mathcal{X}) \times \mathcal{B}(\mathcal{W}))$, where $\mathcal{W} \subseteq \mathbb{R}_0^+ := [0, \infty)$. It is given by

$$\bar{\pi}_N(dx, dw) := \pi(dx) Q_{x,N}(dw) w, \tag{2.21}$$

where $\{Q_{x,N}\}_{(x,N) \in \mathcal{X} \times \mathbb{N}}$ is a family of probability distributions on $(\mathcal{W}, \mathcal{B}(\mathcal{W}))$ satisfying for each $(x, N) \in \mathcal{X} \times \mathbb{N}$

$$\mathbb{E}\left[W_{x,N}\right] \equiv 1, \quad \text{for } W_{x,N} \sim Q_{x,N}(\cdot). \tag{2.22}$$

We restrict our attention to the case where, for each $x \in \mathcal{X}$, $W_{x,N}$ is $Q_{x,N}$-a.s. strictly positive, for reasons that will become clear.

The random variables $\{W_{x,N}\}_{x,N}$ are commonly referred as the weights. Formalising this algorithm using (2.21) and (2.22) was introduced by Andrieu and Vihola (2015), and "exactness" follows immediately: $\bar{\pi}$ admits π as a marginal. Given a proposal kernel $q : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0, 1]$, the respective proposal of the pseudo-marginal is given by

$$\bar{q}_N(x, w; dy, du) := q(x, dy) Q_{y,N}(du),$$

and, from (2.8), the acceptance probability can be expressed as

$$\bar{\alpha}_N(x, w; y, u) := \min\left\{1, \frac{\pi(y) u q(y, x)}{\pi(x) w q(x, y)}\right\}. \tag{2.23}$$

The pseudo-marginal algorithm defines a time-homogeneous Markov chain, with transition kernel $\bar{P}_N$ on the measurable space $(\mathcal{X} \times \mathcal{W}, \mathcal{B}(\mathcal{X}) \times \mathcal{B}(\mathcal{W}))$. A single draw from $\bar{P}_N(x, w; \cdot, \cdot)$ is presented in Algorithm 2.3.

Due to its exactness and straightforward implementation in many settings, the pseudo-marginal has gained recent interest and has been theoretically studied in some depth, see e.g. Andrieu and Roberts (2009), Andrieu and Vihola (2015), Andrieu and Vihola (2016), Doucet et al. (2015), Lyne et al. (2015), Maire et al. (2014), Sherlock et al. (2015) and Deligiannidis et al. (2015). These studies typically compare the pseudo-marginal Markov chain with a "marginal" Markov chain, arising in the case where all the weights are almost surely equal to 1, and (2.23) is then the

Algorithm 2.3 Simulating from $\bar{P}_N(x, w; \cdot, \cdot)$

1. Sample $Y \sim q(x, \cdot)$.
2. Draw $U \sim Q_{Y,N}(\cdot)$.
3. With probability $\bar{\alpha}_N(x, w; Y, U)$ defined in (2.23):
   return (Y, U);
   otherwise:
   return (x, w).

standard Metropolis–Hastings acceptance in (2.8).

2.3.1 Examples of pseudo-marginal algorithms

A common source of intractability for π occurs when a latent variable Z on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ is used to model observed data, as in HMMs presented in Section 2.2.2. Although the density $\pi(x)$ cannot be computed, in principle it can be approximated via importance sampling, using an auxiliary distribution h_x such that $\pi_x \ll h_x$, where π_x denotes the conditional distribution of $Z|X = x$. Therefore, for this setting, the weights are given by

$$W_{x,N} = \frac{1}{N} \sum_{k=1}^{N} \frac{\pi_x\left(Z_x^{(k)}\right)}{h_x\left(Z_x^{(k)}\right)}, \qquad \text{where } \left\{Z_x^{(k)}\right\}_k \overset{i.i.d.}{\sim} h_x(\cdot),$$

which motivates the following generic form when using averages of unbiased estimators

$$W_{x,N} = \frac{1}{N} \sum_{k=1}^{N} W_x^{(k)}, \qquad \text{where } \left\{W_x^{(k)}\right\}_k \overset{i.i.d.}{\sim} Q_x(\cdot) \text{ and } \mathbb{E}\left[W_x^{(k)}\right] \equiv 1. \tag{2.24}$$

It is clear that (2.24) describes only a special case of (2.22). Nevertheless, we will pay special attention to the former in Chapter 4. For similar settings to (2.24) see Andrieu and Roberts (2009).

Since (2.22) is more general, it allows $W_{x,N}$ to be any random variable with expectation 1. As discussed in Section 2.2.2, SMC methods involve the simulation of a system of particles and provide unbiased estimates of likelihoods associated with HMMs, irrespective of the size of the particle system. Looking back at the model given by Figure 2.1 on page 16, we recall from (2.16) and (2.17) that the likelihood

function for θ is given by

$$l\left(\theta; y_{1:T}\right) = \gamma_{\theta,T}\left(1_T\right) = \int_{\mathcal{X}^T} \prod_{n=1}^{T} m_\theta\left(x_{n-1}, dx_n\right) \prod_{n=1}^{T} g_\theta\left(y_n \left| x_n \right.\right),$$

and the unbiased SMC estimator for $l\left(\theta; y_{1:T}\right)$, based on N particles, is given by $\gamma^N_{\theta,T}\left(1_T\right)$ in (2.19) as a result of Algorithm 2.2. Therefore, we can then define

$$W_{\theta,N} := \frac{\gamma^N_{\theta,T}\left(1_T\right)}{\gamma_{\theta,T}\left(1_T\right)}, \tag{2.25}$$

noting that (2.22) is satisfied but (2.24) is not. The resulting pseudo-marginal algorithm has been developed and discussed in detail in Andrieu et al. (2010), where it and related algorithms are referred to as particle MCMC methods.

2.3.2 The noisy algorithm

Although the pseudo-marginal has the desirable property of exactness, it can suffer from "sticky" behaviour, exhibiting poor mixing and slow convergence towards the target distribution (Andrieu and Roberts, 2009 and Lee and Łatuszyński, 2014). The cause for this is well-known to be related with the value of the ratio between $W_{y,N}$ and $W_{x,N}$ at a particular iteration. Heuristically, when the value of the current weight (w in (2.23)) is large, proposed moves can have a low probability of acceptance. As a consequence, the resulting chain may not move after a considerable number of iterations.

In order to overcome this issue, a subtly different algorithm is performed in some practical problems (see, e.g., McKinley et al., 2014). The basic idea is to refresh, independently from the past, the value of the current weight at every iteration. The ratio of the weights between $W_{y,N}$ and $W_{x,N}$ still plays an important role in this alternative algorithm, but here refreshing $W_{x,N}$ at every iteration can improve mixing and the rate of convergence.

This alternative algorithm is commonly known as Monte Carlo within Metropolis (MCWM), as in O'Neill et al. (2000), Beaumont (2003) or Andrieu and Roberts (2009), since typically the weights are Monte Carlo estimates as in (2.24). From this point onwards it will be referred as the noisy MH algorithm or simply the noisy algorithm to emphasize that our main assumption is (2.22). Due to independence from previous iterations while sampling $W_{x,N}$ and $W_{y,N}$, the noisy algorithm also defines a time-homogeneous Markov chain with transition kernel $\tilde{P}_N$ (properly defined in the next chapter), but on the measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. A single draw

from $\tilde{P}_N(x, \cdot)$ is presented in Algorithm 2.4, and it is clear that we restrict our attention to strictly positive weights because the algorithm is not well-defined when both $W_{y,N}$ and $W_{x,N}$ are equal to 0.

Algorithm 2.4 Simulating from $\tilde{P}_N(x, \cdot)$

1. Sample $Y \sim q(x, \cdot)$.
2. Draw $W \sim Q_{x,N}(\cdot)$ and $U \sim Q_{Y,N}(\cdot)$, independently.
3. With probability $\bar{\alpha}_N(x, W; Y, U)$ defined in (2.23):
   return Y;
   otherwise:
   return x.

Even though these algorithms differ only slightly, the related chains have very different properties. In Algorithm 2.4, the value w is generated at every iteration whereas in Algorithm 2.3, it is treated as an input. As a consequence, Algorithm 2.3 produces a chain on $(\mathcal{X} \times \mathcal{W}, \mathcal{B}(\mathcal{X}) \times \mathcal{B}(\mathcal{W}))$ contrasting with a chain from Algorithm 2.4 taking values on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. However, the noisy chain is not invariant under π and it is not reversible in general. Moreover, it may not even have an invariant distribution as shown by some examples in Chapter 3.

From O'Neill et al. (2000) and Fernández-Villaverde and Rubio-Ramírez (2007), it is evident that the implementation of the noisy algorithm goes back even before the appearance of the pseudo-marginal, the latter initially conceptualised as Grouped Independence Metropolis–Hastings (GIMH) in Beaumont (2003). Theoretical properties, however, of the noisy algorithm have mainly been studied in tandem with the pseudo-marginal by Beaumont (2003), Andrieu and Roberts (2009) and more recently by Alquier et al. (2014).

The noisy chain generated by Algorithm 2.4 can be seen as a perturbed version of an idealised Markov chain where the weights $\{W_{x,N}\}_{x,N}$ are all equal to one. Perturbed Markov chains have been investigated in, e.g., Roberts et al. (1998), Breyer et al. (2001), Shardlow and Stuart (2000), Mitrophanov (2005), Ferré et al. (2013). More recently Pillai and Smith (2014) and Rudolf and Schweizer (2015) study such chains using the notion of Wasserstein distance. We focus on total variation distance introduce in (2.3), which is in fact a particular case of the Wasserstein distance.

2.3.3 Ergodic properties

As noted in Andrieu and Roberts (2009), if the weights $\{W_{x,N}\}_{x,N}$ are not essentially bounded then the pseudo-marginal chain cannot be geometrically ergodic; in such cases the "stickiness" may be more evident. In addition, if $\bar{P}_N$ has a left spectral gap (see Appendix A) and from Andrieu and Vihola (2015, Proposition 10) a sufficient, but not necessary (Lee and Łatuszyński, 2014), condition ensuring the pseudo-marginal inherits geometric ergodicity from the marginal is that the weights are uniformly bounded, i.e. for fixed N

$$\sup_{x\in\mathcal{X}} W_{x,N} < \infty.$$

More recently, the authors in Deligiannidis and Lee (2016) have showed that if the the weights have uniformly bounded second moments, i.e. for fixed N

$$\sup_{x\in\mathcal{X}} \mathbb{V}\left[W_{x,N}\right] < \infty,$$

then the asymptotic variance $\sigma^2_{f,\bar{P}_N}$ for functions of the primary variable $f(\cdot,u) : \mathcal{X} \to \mathbb{R}$ will be finite provided the marginal chain is variance bounding (see Appendix A).

The analyses in Andrieu and Roberts (2009) and Alquier et al. (2014) mainly study the noisy algorithm in the case where the marginal Markov chain is uniformly ergodic, see Remark 2.1. For this case, it has been shown that provided a noisy invariant distribution exists for the noisy chain, say $\tilde{\pi}_N$,

$$\lim_{N\to\infty} \left\|\tilde{\pi}_N(\cdot) - \pi(\cdot)\right\|_{TV} = 0.$$

However, there are many MH Markov chains for statistical estimation that cannot be uniformly ergodic, e.g. random walk Metropolis chains when π is not compactly supported (Mengersen and Tweedie, 1996). Our focus in the following chapter is therefore on inheritance of geometric ergodicity by the noisy chain, complementing existing results for the pseudo-marginal chain.

Chapter 3

Stability of Noisy Metropolis–Hastings

The objectives of this chapter can be illustrated using a simple example. Let π be a standard univariate Gaussian distribution $\mathcal{N}(0,1)$ and let the weights $\{W_{x,N}\}$ have an arithmetic average form as in (2.24) with

$$Q_x(\cdot) = \log \mathcal{N}\left(-\frac{1}{2}\sigma^2, \sigma^2\right) \quad \text{and} \quad \sigma^2 = 5,$$

where $\log \mathcal{N}(\mu, \sigma^2)$ denotes a log-normal distribution of parameters μ and σ^2. In addition, let the proposal q be a random walk given by $q(x, \cdot) = \mathcal{N}(x, 4)$. For this example, Figure 3.1 shows the estimated densities using the noisy chain for different values of N.

It appears that the noisy chain has an invariant distribution, and as N in-

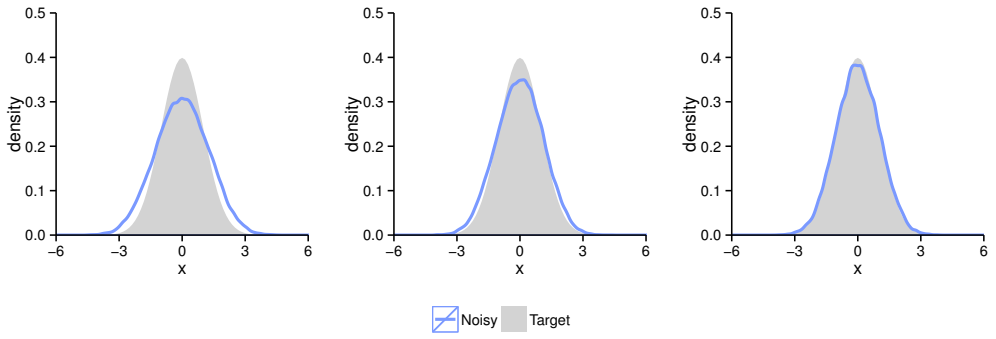


Figure 3.1: Estimated densities using the noisy chain with $100,000$ iterations for $N = 10$ (left), $N = 100$ (centre) and $N = 1,000$ (right)

creases it seems to approach the desired target π. Our objectives here are to answer the following types of questions about the noisy algorithm in general:

- Does an invariant distribution exist, at least for N large enough?
- Does the noisy Markov chain behave like the marginal chain for sufficiently large N?
- Does the invariant distribution, if it exists, converge to π as N increases?

We will see that the answer to the first two points is negative in general. However, all three questions can be answered positively when the marginal chain is geometrically ergodic and the distributions of the weights satisfy additional assumptions.

In order to formalise our analysis, let P denote the Markov transition kernel of a standard MH chain on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, targeting π with proposal q. We will refer to this chain and this algorithm using the term marginal (as in Andrieu and Roberts, 2009 and Andrieu and Vihola, 2015), which is the idealised version for which the noisy chain and corresponding algorithm are simple approximations. Therefore, as in (2.9),

$$P(x, dy) = \alpha(x, y)q(x, dy) + \delta_x(dy)\rho(x),$$

where α and ρ are the MH acceptance and rejection probabilities, respectively, which are given by

$$\begin{aligned} \alpha(x, y) &= \min\left\{1, \frac{\pi(y)\, q(y, x)}{\pi(x)\, q(x, y)}\right\} \\ \text{and} \quad \rho(x) &= 1 - \int_{\mathcal{X}} \alpha(x, y)\, q(x, dy). \end{aligned} \tag{3.1}$$

Similarly, for the transition kernel $\tilde{P}_N$ of the noisy chain, moves are proposed according to q but are accepted using $\bar{\alpha}_N$ (as in (2.23)) instead of α, once values for $W_{x,N}$ and $W_{y,N}$ are sampled. In order to distinguish the acceptance probabilities between the noisy and the pseudo-marginal processes, despite being the same after sampling values for the weights, define

$$\tilde{\alpha}_N(x, y) := \mathbb{E}\left[\bar{\alpha}_N(x, W_{x,N}; y, W_{y,N})\right]. \tag{3.2}$$

Here $\tilde{\alpha}_N$ is the expectation of a randomised acceptance probability, which permits defining the transition kernel of the noisy chain by

$$\tilde{P}_N(x, dy) := \tilde{\alpha}_N(x, y)\, q(x, dy) + \delta_x(dy)\, \tilde{\rho}_N(x),$$

where $\tilde{\rho}_N$ is the noisy rejection probability given by

$$\tilde{\rho}_N(x) := 1 - \int_{\mathcal{X}} \tilde{\alpha}_N(x, y)\, q(x, dy). \tag{3.3}$$

As noted in Section 2.3.2, the noisy kernel $\tilde{P}_N$ is just a perturbed version of P involving a ratio of weights in the noisy acceptance probability $\tilde{\alpha}_N$. When such weights are identically one, i.e. $Q_{x,N}(\{1\}) = 1$, the noisy chain reduces to the marginal chain, whereas the pseudo-marginal becomes the marginal chain with an extra component always equal to 1.

In Section 3.1, some simple examples are presented for which the noisy chain is positive recurrent, so it has an invariant probability distribution. This is perhaps the weakest stability property that one would expect a Monte Carlo Markov chain to have. However, other fairly surprising examples are presented for which the noisy Markov chain is transient even though the marginal and pseudo-marginal chains are geometrically ergodic. Section 3.2 is dedicated to inheritance of geometric ergodicity from the marginal chain, where we explore two different sets of sufficient conditions on the weights $\{W_{x,N}\}$ satisfying the general form in (2.22). The particular settings where the weights arise from arithmetic averages as in (2.24), or from a particle filter as in (2.25), are studied in the following chapter. Once geometric ergodicity is attained, it guarantees the existence of an invariant distribution $\tilde{\pi}_N$ for the noisy chain. Under the same sets of conditions, we show in Section 3.3 that $\tilde{\pi}_N$ and π can be made arbitrarily close in total variation as N increases. Moreover, explicit rates of convergence are possible to obtain in principle.

3.1 Motivating examples

3.1.1 Homogeneous weights with a random walk proposal

Assume a log-concave target distribution π on the positive integers, whose density with respect to the counting measure is given by

$$\pi(m) \propto \exp\{-h(m)\}\, \mathbb{1}(m \in \mathbb{N}),$$

where $h : \mathbb{N} \to \mathbb{R}$ is a convex function. In addition, let the proposal distribution be a symmetric random walk on the integers, i.e.

$$q(m, \{m+1\}) = \frac{1}{2} = q(m, \{m-1\}), \quad \text{for } m \in \mathbb{Z}. \tag{3.4}$$

From Mengersen and Tweedie (1996), it can be seen that the marginal chain is geometrically ergodic.

Now, assume the distribution of the weights $\{W_{m,N}\}_{m,N}$ is homogeneous with respect to the state space, meaning

$$W_{m,N} = W_N \sim Q_N(\cdot), \quad \text{for all } m \in \mathbb{N}. \tag{3.5}$$

In addition, assume $W_N > 0$ Q_N-a.s., then for $m \geq 2$

$$\tilde{P}_N(m, \{m-1\}) = \frac{1}{2}\mathbb{E}\left[\min\left\{1, \frac{\exp\{h(m)\}}{\exp\{h(m-1)\}}\left(\frac{W_N^{(1)}}{W_N^{(2)}}\right)\right\}\right]$$
$$\text{and} \quad \tilde{P}_N(m, \{m+1\}) = \frac{1}{2}\mathbb{E}\left[\min\left\{1, \frac{\exp\{h(m)\}}{\exp\{h(m+1)\}}\left(\frac{W_N^{(1)}}{W_N^{(2)}}\right)\right\}\right],$$

where $\left\{W_N^{(k)}\right\}_{k\in\{1,2\}} \overset{i.i.d.}{\sim} Q_N(\cdot)$. For this particular class of weights and using the fact that h is convex, the noisy chain is geometrically ergodic, implying the existence of an invariant probability distribution. The proof of the following proposition can be found in Section 3.4.

Proposition 3.1. *Consider a log-concave target density on the positive integers and a proposal density as in* (3.4)*. In addition, let the distribution of the weights be homogeneous as in* (3.5)*. Then, the chain generated by the noisy kernel $\tilde{P}_N$ is geometrically ergodic.*

It is worth noting that the distribution of the weights, though homogeneous with respect to the state space, can be taken arbitrarily, as long as the weights are positive. Homogeneity ensures that the distribution of the ratio of such weights is not concentrated near 0, due to its symmetry around one, i.e. for $z > 0$

$$\mathbb{P}\left[\frac{W_N^{(1)}}{W_N^{(2)}} \leq z\right] = \mathbb{P}\left[\frac{W_N^{(1)}}{W_N^{(2)}} \geq \frac{1}{z}\right].$$

In contrast, when the support of the distribution Q_N is unbounded, the corresponding pseudo-marginal chain cannot be geometrically ergodic.

3.1.2 Transient noisy chain with homogeneous weights

In contrast with example in Section 3.1.1, this one shows that the noisy algorithm can produce a transient chain even in simple settings. Let π be a geometric distribution on the positive integers, whose density with respect to the counting measure

is given by

$$\pi(m) = \left(\frac{1}{2}\right)^m \mathbb{1}\left(m \in \mathbb{N}\right). \tag{3.6}$$

In addition, assume the proposal distribution is a simple random walk on the integers, i.e.

$$q\left(m, \{m+1\}\right) = \theta = 1 - q\left(m, \{m-1\}\right), \quad \text{for } m \in \mathbb{Z}. \tag{3.7}$$

where $\theta \in (0,1)$. Under these assumptions, the marginal chain is geometrically ergodic, see Proposition 3.5 in Section 3.4.

Consider $N = 1$ and as in Section 3.1.1, let the distribution of weights be homogeneous and given by

$$W = (b - \varepsilon)\, Ber\left(s\right) + \varepsilon, \quad \text{for } b > 1 \text{ and } \varepsilon \in (0,1), \tag{3.8}$$

where $Ber\left(s\right)$ denotes a Bernoulli random variable of parameter $s \in (0,1)$. There exists a relationship between s, b and ε that guarantees the expectation of the weights is identically one. The following proposition, proven in Section 3.4 by taking $\theta > 1/2$, shows that the resulting noisy chain can be transient for certain values of b, ϵ and θ.

Proposition 3.2. *Consider a geometric target density as in* (3.6) *and a proposal density as in* (3.7)*. In addition, let the weights when* $N = 1$ *be given by* (3.8)*. Then, for some* b*,* ε *and* θ *the chain generated by the noisy kernel* $\tilde{P}_{N=1}$ *is transient.*

In contrast, since the weights are uniformly bounded by b, the pseudo-marginal chain inherits geometric ergodicity for any θ, b and ϵ. The left plot in Figure 3.2 shows an example.

3.1.3 Transient noisy chain with non-homogeneous weights

One could argue that the transient behaviour of the previous example is related to the large value of θ in the proposal distribution. However, as shown here, for any value of $\theta \in (0,1)$ one can construct weights satisfying (2.22) for which the noisy chain is transient. With the same assumptions as in the example in Section 3.1.2, except that now the distribution of weights is not homogeneous but given by

$$\begin{aligned} W_{m,1} &= (b - \varepsilon_m) Ber(s_m) + \varepsilon_m, \\ &\text{for } b > 1 \text{ and } \varepsilon_m = m^{-(3-(m \pmod 3))}, \end{aligned} \tag{3.9}$$

the noisy chain will be transient for b large enough. The proof can be found in Section 3.4.

Proposition 3.3. *Consider a geometric target density as in* (3.6) *and a proposal density as in* (3.7)*. In addition, let the weights when* $N = 1$ *be given by* (3.9)*. Then, for any* $\theta \in (0, 1)$ *there exists some* $b > 1$ *such that the chain generated by the noisy kernel* $\tilde{P}_{N=1}$ *is transient.*

The reason for this becomes apparent when looking at the behaviour of the ratios of weights. Even though $\varepsilon_m \to 0$ as $m \to \infty$, the non-monotonic behaviour of the sequence implies

$$\frac{\varepsilon_{m-1}}{\varepsilon_m} \in \begin{cases} O\left(m^2\right) & \text{if} \quad m \pmod 3 = 0, \\ O\left(m^{-1}\right) & \quad\quad m \pmod 3 \in \{1, 2\}, \end{cases}$$

and

$$\frac{\varepsilon_{m+1}}{\varepsilon_m} \in \begin{cases} O\left(m^{-2}\right) & \text{if} \quad m \pmod 3 = 2, \\ O\left(m\right) & \quad\quad m \pmod 3 \in \{0, 1\}. \end{cases}$$

Hence, the ratio of the weights can become arbitrarily large or arbitrarily close to zero with a non-negligible probability. This allows the algorithm to accept moves to the right more often, if m is large enough. Once again, the pseudo-marginal chain inherits the geometrically ergodic property from the marginal. See the central and right plots of Figure 3.2 for two examples using different proposals. In the following chapter, we will come back to this and the previous examples where we will look at the behaviour of the associated noisy chains as N increases.

3.2 Inheritance of ergodic properties

The inheritance of various ergodic properties of the marginal chain by pseudo-marginal Markov chains has been established using techniques that are powerful but suitable only for reversible Markov chains (see, e.g. Andrieu and Vihola, 2015). Since the noisy Markov chains treated here can be non-reversible, we approach the problem using the geometric drift condition stated in Section 2.1.2. From this point onwards, we assume that the marginal and noisy chains are φ-irreducible and aperiodic. In addition, for many of the following results we assume the marginal chain is geometrically ergodic and its kernel satisfies a specific minorisation condition as stated below.

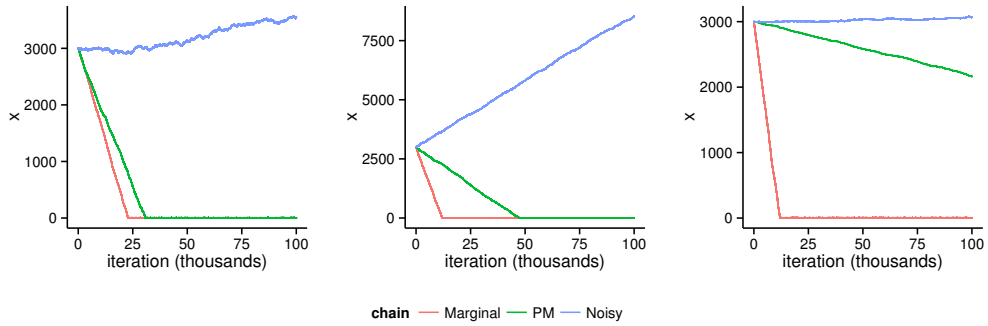


Figure 3.2: Runs of the marginal, pseudo-marginal and noisy chains. Left plot shows example in Section 3.1.2, where $\theta = 0.75$, $\varepsilon = 2 - \sqrt{3}$ and $b = 2\varepsilon\frac{\theta}{1-\theta}$. Central and right plots show example in Section 3.1.3, where $\theta = 0.5$ and $\theta = 0.25$ respectively, with $\varepsilon_m = m^{-(3-m \pmod 3)}$ and $b = 3 + \left(\frac{1-\theta}{\theta}\right)^3$

Assumption (P1)**.** *The marginal chain is geometrically ergodic, implying its kernel P satisfies the geometric drift condition in* (2.7) *for some constants $\lambda < 1$ and $b < \infty$, some function $V \geq 1$ and a small set $S \subseteq \mathcal{X}$.*

Assumption (P2)**.** *Let S be a small set for the marginal chain with transition kernel P. Then, for some $\varepsilon > 0$ and probability measure ν on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, the marginal acceptance probability α and the proposal kernel q satisfy*

$$\int_A \alpha(x,y)q(x,dy) \geq \varepsilon\nu(A), \qquad \text{for } x \in S \text{ and } A \subseteq \mathcal{X}.$$

Remark 3.1. Assumption P2 ensures the minorisation condition in (2.6) is attained by the sub-kernel $\alpha(x,y)q(x,dy)$. This occurs under fairly mild assumptions when S is compact, as noted in Remark 2.3.

3.2.1 Conditions involving a negative moment

From the examples of the previous section, it is clear that the weights play a fundamental role in the behaviour of the noisy chain. The following theorem states that the noisy chain will inherit geometric ergodicity from the marginal under some conditions on the weights, involving a uniform convergence in probability and convergence of negative moments.

Assumption (W1)**.** *For any $\delta > 0$, the weights $\{W_{x,N}\}_{x,N}$ satisfy*

$$\lim_{N\to\infty} \sup_{x\in\mathcal{X}} \mathbb{P}\left[\left|W_{x,N} - 1\right| \geq \delta\right] = 0.$$

Remark 3.2. The above assumption should be understood as uniform (in x) convergence in probability of $W_{x,N}$ towards 1.

Assumption (W2)**.** *The weights* $\{W_{x,N}\}_{x,N}$ *satisfy*

$$\lim_{N\to\infty} \sup_{x\in\mathcal{X}} \mathbb{E}\left[W_{x,N}^{-1}\right] = 1.$$

Theorem 3.1. *Assume P1, P2, W1 and W2. Then, there exists* $N_0 \in \mathbb{N}$ *such that for all* $N \geq N_0$*, the noisy chain with transition kernel* $\tilde{P}_N$ *is geometrically ergodic.*

The above result is obtained by controlling the dissimilarity of the marginal and noisy kernels. This is done by looking at the corresponding rejection and acceptance probabilities. The proofs of the following lemmas appear in Section 3.4.

Lemma 3.1. *For any* $\delta > 0$ *and* $(x, z) \in \mathcal{X}^2$

$$\mathbb{P}\left[\frac{W_{z,N}}{W_{x,N}} \leq 1-\delta\right] \leq 2 \sup_{x\in\mathcal{X}} \mathbb{P}\left[\left|W_{x,N}-1\right| \geq \frac{\delta}{2}\right].$$

Lemma 3.2. *Let* $\rho(x)$ *and* $\tilde{\rho}_N(x)$ *be the rejection probabilities from* (3.1) *and* (3.3) *respectively. Then, for any* $\delta > 0$

$$\tilde{\rho}_N(x) - \rho(x) \leq \delta + 2 \sup_{x\in\mathcal{X}} \mathbb{P}\left[\left|W_{x,N}-1\right| \geq \frac{\delta}{2}\right].$$

Lemma 3.3. *Let* $\alpha(x, y)$ *and* $\tilde{\alpha}_N(x, y)$ *be the acceptance probabilities from* (3.1) *and* (3.2) *respectively. Then,*

$$\tilde{\alpha}_N(x, y) \leq \alpha(x, y)\mathbb{E}\left[W_{x,N}^{-1}\right].$$

Notice that Assumptions W1 and W2 allow control on the bounds in the above lemmas. While Lemma 3.2 provides a bound for the difference of the rejection probabilities, Lemma 3.3 gives one for the ratio of the acceptance probabilities. The proof of Theorem 3.1 is now presented.

Proof of Theorem 3.1. Since the marginal chain P is geometrically ergodic, it satisfies the geometric drift condition in (2.7) for some $\lambda < 1$, $b < \infty$, some function

$V \geq 1$ and a small set $S \subseteq \mathcal{X}$. Now, using the above lemmas

$$\begin{aligned}\tilde{P}_N V(x) - PV(x) &= \int_{\mathcal{X}} q(x, dz) \left(\tilde{\alpha}_N(x, z) - \alpha(x, z)\right) V(z) \\ &+ V(x) \left(\tilde{\rho}_N(x) - \rho(x)\right) \leq \left(\sup_{x \in \mathcal{X}} \mathbb{E}\left[W_{x,N}^{-1}\right] - 1\right) PV(x) \\ &+ \left(\delta + 2 \sup_{x \in \mathcal{X}} \mathbb{P}\left[\left|W_{x,N} - 1\right| \geq \frac{\delta}{2}\right]\right) V(x).\end{aligned}$$

By Assumptions W1 and W2, for any ε, $\delta > 0$ there exists $N_1 \in \mathbb{N}$ such that

$$\sup_{x \in \mathcal{X}} \mathbb{P}\left[\left|W_{x,N} - 1\right| \geq \frac{\delta}{2}\right] < \frac{\varepsilon}{4} \quad \text{and} \quad \sup_{x \in \mathcal{X}} \mathbb{E}\left[W_{x,N}^{-1}\right] - 1 < \varepsilon,$$

whenever $N \geq N_1$, implying

$$\begin{aligned}\tilde{P}_N V(x) &\leq PV(x) + \varepsilon PV(x) + \left(\delta + \frac{\varepsilon}{2}\right) V(x) \\ &\leq \left(\lambda + \varepsilon\lambda + \delta + \frac{\varepsilon}{2}\right) V(x) + b\,(1 + \varepsilon)\, \mathbb{1}\left(x \in S\right).\end{aligned}$$

Taking $\delta = \frac{\varepsilon}{2}$ and $\varepsilon \in \left(0, \frac{1-\lambda}{1+\lambda}\right)$, the noisy chain with kernel $\tilde{P}_N$ also satisfies a geometric drift condition if $N \geq N_1$, for the same function V and set S.

It remains to prove that the set S is also small for $\tilde{P}_N$. Take $A \subseteq \mathcal{X}$, for any $\delta \in (0, 1)$

$$\begin{aligned}\tilde{P}_N(x, A) &\geq \int_A \tilde{\alpha}_N(x, z) q(x, dz) \geq \int_A \mathbb{E}\left[\min\left\{1, \frac{W_{z,N}}{W_{x,N}}\right\}\right] \alpha(x, z) q(x, dz) \\ &\geq (1 - \delta) \int_A \left(1 - \mathbb{P}\left[\frac{W_{z,N}}{W_{x,N}} \leq 1 - \delta\right]\right) \alpha(x, z) q(x, dz).\end{aligned}$$

Then, by Lemma 3.1

$$\tilde{P}_N(x, A) \geq (1 - \delta) \left(1 - 2 \sup_{x \in \mathcal{X}} \mathbb{P}\left[\left|W_{x,N} - 1\right| \geq \frac{\delta}{2}\right]\right) \int_A \alpha(x, z) q(x, dz).$$

By W1, there exists $N_2 \in \mathbb{N}$ such that for $N \geq N_2$

$$\sup_{x \in \mathcal{X}} \mathbb{P}\left[\left|W_{x,N} - 1\right| \geq \frac{\delta}{2}\right] \leq \frac{\delta}{2},$$

giving

$$\tilde{P}_N(x, A) \geq (1 - \delta)^2 \int_A \alpha(x, z) q(x, dz).$$

Due to P2,

$$\tilde{P}_N(x, A) \geq (1-\delta)^2 \varepsilon \nu(A), \qquad \text{for } x \in S.$$

The result then follows by considering $N_0 = \max\{N_1, N_2\}$. □

Remark 3.3. In fact, W1 and W2 together guarantee for any $\delta > 0$ and taking N sufficiently large

$$-\delta \leq \tilde{\alpha}_N(x, y) - \alpha(x, y) \leq \alpha(x, y)\delta,$$

which is the crucial assumption in Pillai and Smith (2014, Lemma 3.6) for obtaining a similar drift condition.

3.2.2 Conditions on the proposal distribution

In this subsection a different bound for the acceptance probabilities is provided, which allows dropping assumption W2 but imposes a different one on the proposal q instead.

Assumption (P1*)**.** *Assumption P1 holds and for the same drift function V in P1 there exists $K < \infty$ such that the proposal kernel q satisfies*

$$qV(x) \leq KV(x), \quad \text{for } x \in \mathcal{X}.$$

Theorem 3.2. *Assume P1*, P2 and W1. Then, there exists $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$, the noisy chain with transition kernel $\tilde{P}_N$ is geometrically ergodic.*

In order to prove Theorem 3.2 the following lemma is required. Its proof can be found in Section 3.4. In contrast with Lemma 3.3, this lemma provides a bound for the additive difference of the noisy and marginal acceptance probabilities.

Lemma 3.4. *Let $\alpha(x, y)$ and $\tilde{\alpha}_N(x, y)$ be the acceptance probabilities as defined in* (3.1) *and* (3.2)*, respectively. Then, for any $\eta > 0$*

$$\tilde{\alpha}_N(x, y) - \alpha(x, y) \leq \eta + 2 \sup_{x \in \mathcal{X}} \mathbb{P}\left[\left|W_{x,N} - 1\right| \geq \frac{\eta}{2\,(1+\eta)}\right].$$

Proof of Theorem 3.2. Using Lemma 3.2 and Lemma 3.4 with $\eta = \delta$

$$\begin{aligned}
\tilde{P}_N V(x) - PV(x) &= \int_{\mathcal{X}} q(x, dz) \left(\tilde{\alpha}_N(x, z) - \alpha(x, z)\right) V(z) \\
&+ V(x) \left(\tilde{\rho}_N(x) - \rho(x)\right) \leq \left(\delta + 2 \sup_{x \in \mathcal{X}} \mathbb{P}\left[\left|W_{x,N} - 1\right| \geq \frac{\delta}{2\left(1+\delta\right)}\right]\right) qV(x) \\
&+ \left(\delta + 2 \sup_{x \in \mathcal{X}} \mathbb{P}\left[\left|W_{x,N} - 1\right| \geq \frac{\delta}{2}\right]\right) V(x) \\
&\leq \left(\delta + 2 \sup_{x \in \mathcal{X}} \mathbb{P}\left[\left|W_{x,N} - 1\right| \geq \frac{\delta}{2\left(1+\delta\right)}\right]\right) \left(qV(x) + V(x)\right).
\end{aligned}$$

By W1, there exists $N_1 \in \mathbb{N}$ such that

$$\sup_{x \in \mathcal{X}} \mathbb{P}\left[\left|W_{x,N} - 1\right| \geq \frac{\delta}{2\left(1+\delta\right)}\right] < \frac{\varepsilon}{4},$$

whenever $N \geq N_1$. This implies

$$\tilde{P}_N V(x) \leq PV(x) + \left(\delta + \frac{\varepsilon}{2}\right) \left(qV(x) + V(x)\right),$$

and using P1*

$$\tilde{P}_N V(x) \leq \left(\lambda + \left(\delta + \frac{\varepsilon}{2}\right)\left(K+1\right)\right) V(x) + b\mathbb{1}\left(x \in S\right).$$

Taking $\delta = \frac{\varepsilon}{2}$ and $\varepsilon \in \left(0, \frac{1-\lambda}{1+K}\right)$, the noisy chain with kernel $\tilde{P}_N$ also satisfies a geometric drift condition if $N \geq N_1$, for the same function V and set S.

From the proof of Theorem 3.1, W1 and P2 imply the existence of $N_2 \in \mathbb{N}$ such that the set S is small for $\tilde{P}_N$, if $N \geq N_2$. The proof is completed by considering $N_0 = \max\left\{N_1, N_2\right\}$. □

Remark 3.4. By itself, W1 implies for any $\delta > 0$ and taking N sufficiently large

$$\left|\tilde{\alpha}_N(x, y) - \alpha(x, y)\right| \leq \delta,$$

but it needs to be paired with P1* to obtain the desired result. These assumptions are comparable to those in Pillai and Smith (2014, Lemma 3.6), taking f constant therein. Additionally, W1 and P1* imply the required conditions on $\mathcal{E}$ and λ in Rudolf and Schweizer (2015, Corollary 31), where a similar result is proved.

In general, assumption P1* may be difficult to verify as one must identify a particular function V, but it is easily satisfied when restricting to log-Lipschitz targets and when using a random walk proposal of the form (2.10). To see this the

following assumption is required, which is a particular case of P1 and is satisfied under some extra technical conditions (see, e.g., Roberts and Tweedie, 1996a).

Assumption (P1**). *$\mathcal{X} \subseteq \mathbb{R}^d$. The target π is log-Lipschitz, meaning that for some $L > 0$*

$$|\log \pi(z) - \log \pi(x)| \leq L\|z - x\|.$$

P1 holds taking the drift function $V = \pi^{-s}$, for any $s \in (0, 1)$. The proposal q is a random walk as in (2.10) *satisfying*

$$\int_{\mathbb{R}^d} \exp\{a\|u\|\}\, q(\|u\|)du < \infty,$$

for some $a > 0$.

See Section 3.4 for a proof of the following proposition.

Proposition 3.4. *P1** implies P1*.*

3.3 Convergence of the noisy invariant distribution

So far the only concern has been whether the noisy chain inherits the geometric ergodicity property from the marginal chain. As an immediate consequence, geometric ergodicity guarantees the existence of an invariant probability distribution $\tilde{\pi}_N$ for $\tilde{P}_N$, provided N is large enough. In addition, using the same conditions from Section 3.2, we can characterise and in some cases quantify the convergence in total variation of $\tilde{\pi}_N$ towards the desired target π, as $N \to \infty$.

3.3.1 Convergence in total variation

Provided N is large, the noisy kernels $\{\tilde{P}_{N+k}\}_{k\geq 0}$ together with the marginal P will be simultaneous geometrically ergodic, as defined in Definition 2.6. This will allow the use of coupling arguments for ensuring $\tilde{\pi}_N$ and π get arbitrarily close in total variation.

Theorem 3.3. *Assume P1, P2, W1 and W2. Alternatively, assume P1*, P2 and W1. Then,*

(i). There exists $N_0 \in \mathbb{N}$ such that the class of kernels $\left\{P, \tilde{P}_{N_0}, \tilde{P}_{N_0+1}, \dots\right\}$ is simultaneously geometrically ergodic;

(ii). For all $x \in \mathcal{X}$, $\lim_{N\to\infty} \|\tilde{P}_N(x,\cdot) - P(x,\cdot)\|_{TV} = 0$;

(iii). $\lim_{N\to\infty} \|\tilde{\pi}_N(\cdot) - \pi(\cdot)\|_{TV} = 0$.

Part (iii) of the above theorem is mainly a consequence of Roberts et al. (1998, Theorem 9) when parts (i) and (ii) hold. Indeed, by the triangle inequality,

$$\begin{aligned}\|\tilde{\pi}_N(\cdot) - \pi(\cdot)\|_{TV} \leq \|\tilde{P}_N^n(x,\cdot) - \tilde{\pi}_N(\cdot)\|_{TV} + \|P^n(x,\cdot) - \pi(\cdot)\|_{TV} \\ + \|\tilde{P}_N^n(x,\cdot) - P^n(x,\cdot)\|_{TV}.\end{aligned} \tag{3.10}$$

Provided $N \geq N_0$, the first two terms in (3.10) can be made arbitrarily small by increasing n. In addition, due to the simultaneous geometrically ergodic property, the first term in (3.10) is uniformly controlled regardless the value of N. Finally, by an inductive argument (explained in detail later on for deriving (3.12)), part (ii) implies that for all $x \in \mathcal{X}$ and all $n \in \mathbb{N}$

$$\lim_{N\to\infty} \|\tilde{P}_N^n(x,\cdot) - P^n(x,\cdot)\|_{TV} = 0.$$

Proof of Theorem 3.3. From the proofs of Theorem 3.1 and Theorem 3.2, under the stated assumptions there exists $N_0 \in \mathbb{N}$ such that the class of kernels $\left\{P, \tilde{P}_{N_0}, \tilde{P}_{N_0+1}, \dots\right\}$ satisfies conditions (i) and (ii) in Definition 2.6 for the same function V, small set S and constants λ, b; where the latter constants may differ from those considered in Assumption P1 or Assumption P1* .

To prove (ii), apply Lemma 3.2 and Lemma 3.4 to get

$$\begin{aligned}\sup_{A\in\mathcal{B}(\mathcal{X})} \left\{\tilde{P}_N(x,A) - P(x,A)\right\} &\leq \left(\eta + 2\sup_{x\in\mathcal{X}} \mathbb{P}_{Q_{x,N}}\left[\left|W_{x,N} - 1\right| \geq \frac{\eta}{2\,(1+\eta)}\right]\right) \\ &\quad \times \sup_{A\in\mathcal{B}(\mathcal{X})} q(x,A) + (\tilde{\rho_N}(x) - \rho(x)) \sup_{A\in\mathcal{B}(\mathcal{X})} \mathbb{1}_{x\in A} \\ &\leq \left(\eta + 2\sup_{x\in\mathcal{X}} \mathbb{P}_{Q_{x,N}}\left[\left|W_{x,N} - 1\right| \geq \frac{\eta}{2\,(1+\eta)}\right]\right) \\ &\quad + \left(\delta + 2\sup_{x\in\mathcal{X}} \mathbb{P}_{Q_{x,N}}\left[\left|W_{x,N} - 1\right| \geq \frac{\delta}{2}\right]\right)\end{aligned} \tag{3.11}$$

Finally, taking $N \to \infty$ and by W1

$$\lim_{N\to\infty} \sup_{A\in\mathcal{B}(\mathcal{X})} \left\{\tilde{P}_N(x,A) - P(x,A)\right\} \leq \eta + \delta.$$

The result follows since η and δ can be taken arbitrarily small.

For (iii), see Theorem 9 in Roberts et al. (1998) for a detailed proof. □

Remark 3.5. A Wasserstein distance variant of part (*iii*) in Theorem 3.3 has been proved in Rudolf and Schweizer (2015, Corollary 28), in which control of the difference between $\tilde{\alpha}_N$ and α is still required and can be obtained using W1.

3.3.2 Rate of convergence

Let $\{\tilde{\Phi}_n^N\}_{n\geq 0}$ denote the noisy chain and $\{\Phi_n\}_{n\geq 0}$ the marginal chain, which move according to the kernels $\tilde{P}_N$ and P, respectively and define

$$c_x := 1 - \|\tilde{P}_N(x,\cdot) - P(x,\cdot)\|_{TV}.$$

Using notions of maximal coupling for random variables defined on a Polish space (see Lindvall, 2002 and Thorisson, 2013), there exists a probability measure $\nu_x(\cdot)$ such that

$$P(x,\cdot) \geq c_x\nu_x(\cdot) \quad \text{and} \quad \tilde{P}_N(x,\cdot) \geq c_x\nu_x(\cdot).$$

Let $c := \inf_{x\in\mathcal{X}} c_x$, define a coupling in the following way

- If $\tilde{\Phi}_{n-1}^N = \Phi_{n-1} = y$, with probability c draw $\Phi_n \sim \nu_y(\cdot)$ and set $\tilde{\Phi}_n^N = \Phi_n$. Otherwise, draw independently $\Phi_n \sim R(y,\cdot)$ and $\tilde{\Phi}_n^N \sim \tilde{R}_N(y,\cdot)$, where

$$R(y,\cdot) := (1-c)^{-1}\left(P(y,\cdot) - c\nu_y(\cdot)\right)$$
$$\text{and} \quad \tilde{R}_N(y,\cdot) := (1-c)^{-1}\left(\tilde{P}_N(y,\cdot) - c\nu_y(\cdot)\right).$$

- If $\tilde{\Phi}_{n-1}^N \neq \Phi_{n-1}$, draw independently $\Phi_n \sim P(y,\cdot)$ and $\tilde{\Phi}_n^N \sim \tilde{P}_N(y,\cdot)$.

Since

$$\begin{aligned}\mathbb{P}\left[\tilde{\Phi}_n^N \neq \Phi_n \middle| \tilde{\Phi}_0^N = \Phi_0 = x\right] &\leq \mathbb{P}\left[\tilde{\Phi}_n^N \neq \Phi_n \middle| \tilde{\Phi}_{n-1}^N = \Phi_{n-1}, \tilde{\Phi}_0^N = \Phi_0 = x\right] \\ &+ \mathbb{P}\left[\tilde{\Phi}_{n-1}^N \neq \Phi_{n-1} \middle| \tilde{\Phi}_0^N = \Phi_0 = x\right] \leq 1 - c + \mathbb{P}\left[\tilde{\Phi}_{n-1}^N \neq \Phi_{n-1} | \tilde{\Phi}_0^N = \Phi_0 = x\right],\end{aligned}$$

and noting

$$\mathbb{P}\left[\tilde{\Phi}_1^N \neq \Phi_1 \middle| \tilde{\Phi}_0^N = \Phi_0 = x\right] \leq \sup_{x\in\mathcal{X}} \|\tilde{P}_N(x,\cdot) - P(x,\cdot)\|_{TV} = 1 - c,$$

an induction argument can be applied to obtain

$$\mathbb{P}\left[\tilde{\Phi}_n^N \neq \Phi_n \middle| \tilde{\Phi}_0^N = \Phi_0 = x\right] \leq n \sup_{x\in\mathcal{X}} \|\tilde{P}_N(x,\cdot) - P(x,\cdot)\|_{TV}.$$

Therefore, using the coupling inequality, the third term in (3.10) can be bounded by

$$\begin{aligned}\|\tilde{P}_N^n(x,\cdot) - P^n(x,\cdot)\|_{TV} &\leq \mathbb{P}\left[\tilde{\Phi}_n^N \neq \Phi_n | \tilde{\Phi}_0^N = \Phi_0 = x\right] \\ &\leq n \sup_{x\in\mathcal{X}} \|\tilde{P}_N(x,\cdot) - P(x,\cdot)\|_{TV}. \end{aligned} \tag{3.12}$$

On the other hand, using the simultaneous geometric ergodicity of the kernels and provided N is large enough, the noisy and marginal kernels will each satisfy a geometric drift condition as in (2.7) with a common drift function $V \geq 1$, small set S and constants λ, b. Therefore, by Theorem 2.3, there are $R > 0$ and $\tau < 1$ such that

$$\begin{aligned} \|\tilde{P}_N^n(x,\cdot) - \tilde{\pi}_N(\cdot)\|_{TV} &\leq RV(x)\tau^n \\ \text{and} \quad \|P^n(x,\cdot) - \pi(\cdot)\|_{TV} &\leq RV(x)\tau^n. \end{aligned} \tag{3.13}$$

Explicit values for R and τ are in principle possible, as done in Rosenthal (1995) or Meyn and Tweedie (1994), but this is not pursued here. For simplicity assume $\inf_{x\in\mathcal{X}} V(x) = 1$, then combining (3.12) and (3.13) in (3.10), for all $n \in \mathbb{N}$

$$\|\tilde{\pi}_N(\cdot) - \pi(\cdot)\|_{TV} \leq 2R\tau^n + n \sup_{x\in\mathcal{X}} \|\tilde{P}_N(x,\cdot) - P(x,\cdot)\|_{TV}. \tag{3.14}$$

So, if an analytic expression in terms of N is available for the second term on the right hand side of (3.14), it will be possible to obtain an explicit rate of convergence for $\tilde{\pi}_N$ and π.

Theorem 3.4. *Assume P1, P2, W1 and W2. Alternatively, assume P1*, P2 and W1. In addition, suppose*

$$\sup_{x\in\mathcal{X}} \|\tilde{P}_N(x,\cdot) - P(x,\cdot)\|_{TV} \leq \frac{1}{r(N)},$$

where $r : \mathbb{N} \to \mathbb{R}^+$ and $\lim_{N\to\infty} r(N) = +\infty$. Then, there exists $D > 0$ and $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$,

$$\|\tilde{\pi}_N(\cdot) - \pi(\cdot)\|_{TV} \leq D\frac{\log(r(N))}{r(N)}. \tag{3.15}$$

Proof. Let $R > 0, \tau \in (0,1)$ and $r > 0$. Pick r large enough, such that

$$\log\left(2Rr\log\left(\tau^{-1}\right)\right) \geq 1,$$

then the convex function $f : [1, \infty) \to \mathbb{R}^+$ where

$$f(s) = 2R\tau^s + \frac{s}{r},$$

is minimised at

$$s_* = \frac{\log\left(2Rr\log\left(\tau^{-1}\right)\right)}{\log\left(\tau^{-1}\right)}.$$

Restricting the domain of f to the positive integers and due to convexity, it is then minimised at either

$$n_1 = \lfloor s_* \rfloor \qquad \text{or} \qquad n_2 = \lceil s_* \rceil.$$

In any case

$$\min\left\{f(n_1), f(n_2)\right\} \leq f(s_* + 1) = \frac{1}{r}\left(1 + \frac{\tau + \log\left(2Rr\log\left(\tau^{-1}\right)\right)}{\log\left(\tau^{-1}\right)}\right).$$

Finally, take N large enough such that

$$\log\left(2Rr(N)\log\left(\tau^{-1}\right)\right) \geq 1,$$

and from (3.14)

$$\begin{aligned}\|\tilde{\pi}_N(\cdot) - \pi(\cdot)\|_{TV} &\leq \min\left\{f(n_1), f(n_2)\right\} \\ &\leq \frac{1}{r(N)}\left(1 + \frac{\tau + \log\left(2Rr(N)\log\left(\tau^{-1}\right)\right)}{\log\left(\tau^{-1}\right)}\right) = O\left(\frac{\log\left(r(N)\right)}{r(N)}\right),\end{aligned}$$

obtaining the result. □

Remark 3.6. A general result bounding the total variation between the law of a Markov chain and a perturbed version is presented in Rudolf and Schweizer (2015, Theorem 21). This is done using the connection between the V-norm distance and the Wasserstein distance introduced in Hairer and Mattingly (2011). With such a result, and considering the same assumptions in Theorem 3.4, one could in principle obtain an explicit value for D in (3.15).

So far, two different sets of sufficient conditions have been provided under which the noisy chain inherits geometric ergodicity from the marginal chain. In Chapter 4, the particular cases when the weights arise from an arithmetic average as in (2.24), or as a result of a bootstrap filter as in (2.25) are discussed. Also, some

examples of Section 3.1 will be revisited, where the behaviour as $N \to \infty$ of the corresponding chains is studied. We finish this chapter with the remaining proofs of the technical lemmas and propositions previously stated.

3.4 Proofs

The following proposition for state-dependent Markov chains on the positive integers will be useful for addressing some proofs. See Norris (1999) for a proof of parts (i) and (ii), for part (iii) see Callaert and Keilson (1973), which is proved within the birth-death process context.

Proposition 3.5. *Suppose we have a random walk on* $\mathbb{N}$ *with transition kernel* P. *Define for* $m \geq 1$

$$p_m := P(m, \{m+1\}) \quad \textit{and} \quad q_m := P(m, \{m-1\}),$$

with $q_1 = 0, p_1 \in (0,1]$ *and* $p_m, q_m > 0, p_m + q_m \leq 1$ *for all* $m \geq 2$. *The resulting chain is:*

(i). Recurrent if and only if

$$\sum_{m=2}^{\infty} \prod_{i=2}^{m} \frac{q_i}{p_i} \to \infty;$$

(ii). Positive recurrent if and only if

$$\sum_{m=2}^{\infty} \prod_{i=2}^{m} \frac{p_{i-1}}{q_i} < \infty; \tag{3.16}$$

(iii). Geometrically ergodic if

$$\lim_{m\to\infty} p_m < \lim_{m\to\infty} q_m. \tag{3.17}$$

Remark 3.7. Notice that (iii) is not an if and only if statement and, if the chain is not state-dependent, (3.16) implies (3.17).

3.4.1 Section 3.1

Proof of Proposition 3.1. Since h is convex

$$\begin{aligned} h(m) - h(m-1) &\geq h'(m-1) \quad \text{and} \\ h(m) - h(m+1) &\leq -h'(m), \end{aligned}$$

implying

$$\frac{\tilde{P}_N(m, \{m-1\})}{\tilde{P}_N(m, \{m+1\})} \geq \frac{\mathbb{E}\left[\min\left\{1, \exp\{h'(m-1)\}\frac{W_N^{(1)}}{W_N^{(2)}}\right\}\right]}{\mathbb{E}\left[\min\left\{1, \exp\{-h'(m)\}\frac{W_N^{(1)}}{W_N^{(2)}}\right\}\right]}.$$

Define $Z := \frac{W_N^{(1)}}{W_N^{(2)}}$, and since $\pi(m) \to 0$ it is true that

$$\log(k) := \lim_{m\to\infty} h'(m) > 0, \tag{3.18}$$

hence

$$\lim_{m\to\infty} \frac{\tilde{P}_N(m, \{m-1\})}{\tilde{P}_N(m, \{m+1\})} \geq \frac{\mathbb{E}\left[\min\left\{1, kZ\right\}\right]}{\mathbb{E}\left[\min\left\{1, k^{-1}Z\right\}\right]}. \tag{3.19}$$

If $k = +\infty$, it is clear that the limit in (3.19) diverges, consequently the noisy chain is geometrically ergodic according to Proposition 3.5. If $k < \infty$, the noisy chain will be geometrically ergodic if

$$\mathbb{E}\left[\min\left\{1, kZ\right\}\right] > \mathbb{E}\left[\min\left\{1, k^{-1}Z\right\}\right],$$

which can be translated to

$$k\mathbb{E}\left[Z\mathbb{1}\left(Z \leq k^{-1}\right)\right] + \mathbb{P}\left[Z > k^{-1}\right] > k^{-1}\mathbb{E}\left[Z\mathbb{1}\left(Z < k\right)\right] + \mathbb{P}\left[Z \geq k\right],$$

or equivalently to

$$k\mathbb{P}\left[k^{-1} < Z < k\right] + \left(k^2 - 1\right)\mathbb{E}\left[Z\mathbb{1}\left(Z \leq k^{-1}\right)\right] > \mathbb{E}\left[Z\mathbb{1}\left(k^{-1} < Z < k\right)\right]. \tag{3.20}$$

Now consider two cases, first if $\mathbb{P}\left[k^{-1} < Z < k\right] > 0$ then it is clear that

$$\mathbb{E}\left[(k - Z)\,\mathbb{1}\left(k^{-1} < Z < k\right)\right] > 0,$$

which satisfies (3.20). Finally, if $\mathbb{P}\left[k^{-1} < Z < k\right] = 0$ then

$$\mathbb{P}\left[Z \leq k^{-1}\right] = \frac{1}{2} = \mathbb{P}\left[Z \geq k\right],$$

implying from (3.18)

$$\left(k^2 - 1\right)\mathbb{E}\left[Z\mathbb{1}\left(Z \leq k^{-1}\right)\right] > 0,$$

and leading to (3.20). □

Proof of Proposition 3.2. For simplicity the subscript N is dropped. In this case,

$$Q_{m,1}\left(\cdot\right) = Q\left(\cdot\right) = (b - \varepsilon)Ber(s) + \varepsilon,$$

and the condition $\mathbb{E}\left[W\right] = 1$ implies

$$s = \frac{1 - \varepsilon}{b - \varepsilon}. \tag{3.21}$$

Let $\theta \in \left(\frac{1}{1+2\varepsilon}, 1\right)$ and set

$$b = \varepsilon\frac{2\theta}{1 - \theta}, \tag{3.22}$$

this implies $\bar{\alpha}(m, w; m - 1, u) \equiv 1$ and

$$\bar{\alpha}(m, w; m + 1, u) = \begin{cases} \frac{1-\theta}{2\theta} & u = w \\ 1 \quad \text{if} & u = b, w = \varepsilon \\ \left(\frac{1-\theta}{2\theta}\right)^2 & u = \varepsilon, w = b \end{cases}.$$

Therefore, for $m \geq 2$,

$$\begin{aligned} &\tilde{\alpha}(m, m - 1) = 1 \quad \text{and} \\ &\tilde{\alpha}(m, m + 1) = \frac{1 - \theta}{2\theta}\left(s^2 + (1 - s)^2\right) + \left(1 + \left(\frac{1 - \theta}{2\theta}\right)^2\right) s\left(1 - s\right). \end{aligned}$$

Consequently, $\tilde{P}(m, \{m-1\}) = 1-\theta$ and

$$\begin{aligned}\tilde{P}(m, \{m+1\}) &= \theta\left(\frac{1-\theta}{2\theta}\left(s^2+(1-s)^2\right)+\left(1+\left(\frac{1-\theta}{2\theta}\right)^2\right)s\,(1-s)\right)\\ &> \theta s(1-s).\end{aligned}$$

From Proposition 3.5, if

$$\tilde{P}(m, \{m+1\}) > \tilde{P}(m, \{m-1\}),$$

then the noisy chain will be transient. For this to happen, it is enough to pick θ and s such that

$$\theta s(1-s) - (1-\theta) \geq 0.$$

Let $s = \varepsilon$, then from (3.21) and (3.22)

$$\theta = \frac{(1-\varepsilon+\varepsilon^2)}{1-\varepsilon+3\varepsilon^2} = 1 - \frac{2\varepsilon^2}{1-\varepsilon+3\varepsilon^2}, \tag{3.23}$$

and if $\varepsilon \leq 2-\sqrt{3}$ then

$$\begin{aligned}\theta s(1-s) - (1-\theta) &= \frac{\varepsilon}{1-\varepsilon+3\varepsilon^2}\left((1-\varepsilon+\varepsilon^2)(1-\varepsilon)-2\varepsilon\right)\\ &\geq \frac{\varepsilon}{1-\varepsilon+3\varepsilon^2}\left((1-\varepsilon)^2-2\varepsilon\right) = \frac{\varepsilon}{1-\varepsilon+3\varepsilon^2}\left((2-\varepsilon)^2-3\right)\\ &\geq 0.\end{aligned}$$

Hence, for $\varepsilon \in \left(0, 2-\sqrt{3}\right)$ and setting $s = \varepsilon$, θ as in (3.23) and b as in (3.22), the resulting noisy chain is transient. □

Proof of Proposition 3.3. For simplicity the subscript N is dropped. In this case,

$$Q_{m,1}(\cdot) = Q_m(\cdot) = (b-\varepsilon_m)Ber(s_m) + \varepsilon_m,$$

and the condition $\mathbb{E}[W_m] = 1$ implies

$$s_m = \frac{1-\varepsilon_m}{b-\varepsilon_m}.$$

Then, for m large enough

$$\begin{aligned}\tilde{\alpha}(m,m-1) &= \mathbb{E}\left[\bar{\alpha}(m,W_m;m-1,W_{m-1})\right] \\ &= \min\left\{1,\frac{2\theta}{1-\theta}\right\} s_{m-1}s_m + s_{m-1}(1-s_m) \\ &\quad + (1-s_{m-1})(1-s_m)\mathbb{1}\left(m \pmod 3 = 0\right) + O\left(m^{-1}\right),\end{aligned}$$

and

$$\begin{aligned}\tilde{\alpha}(m,m+1) &= \mathbb{E}\left[\bar{\alpha}(m,W_m;m+1,W_{m+1})\right] \\ &= \min\left\{1,\frac{1-\theta}{2\theta}\right\} s_m s_{m+1} + (1-s_m)s_{m+1} \\ &\quad + (1-s_m)(1-s_{m+1})\mathbb{1}\left(m \pmod 3 \neq 2\right) + O\left(m^{-1}\right).\end{aligned}$$

Define

$$c_m := \frac{\tilde{P}(m,\{m-1\})}{\tilde{P}(m,\{m+1\})} = \frac{(1-\theta)\tilde{\alpha}(m,m-1)}{\theta\tilde{\alpha}(m,m+1)}.$$

Since $s_m \to \frac{1}{b}$ as $m \to \infty$,

$$\begin{aligned}c_{0,\infty} := \lim_{k\to\infty} c_{3k} &= \left(\frac{1-\theta}{\theta}\right)\frac{\left(\min\left\{1,\frac{2\theta}{1-\theta}\right\}-1\right)\frac{1}{b^2}+\frac{1}{b}+\left(1-\frac{1}{b}\right)^2}{\left(\min\left\{1,\frac{1-\theta}{2\theta}\right\}-1\right)\frac{1}{b^2}+\frac{1}{b}+\left(1-\frac{1}{b}\right)^2} \\ &\leq \left(\frac{1-\theta}{\theta}\right)\frac{1}{1-\frac{1}{b}} = \left(\frac{1-\theta}{\theta}\right)\frac{b}{b-1} =: l_0,\end{aligned}$$

$$\begin{aligned}c_{1,\infty} := \lim_{k\to\infty} c_{3k}c_{3k+1} &= c_{0,\infty}\left(\frac{1-\theta}{\theta}\right)\frac{\left(\min\left\{1,\frac{2\theta}{1-\theta}\right\}-1\right)\frac{1}{b^2}+\frac{1}{b}}{\left(\min\left\{1,\frac{1-\theta}{2\theta}\right\}-1\right)\frac{1}{b^2}+\frac{1}{b}+\left(1-\frac{1}{b}\right)^2} \\ &\leq l_0\left(\frac{1-\theta}{\theta}\right)\frac{\frac{1}{b}}{1-\frac{1}{b}} = \left(\frac{1-\theta}{\theta}\right)^2\frac{b}{(b-1)^2} =: l_1\end{aligned}$$

and

$$\begin{aligned}\lim_{k\to\infty} c_{3k}c_{3k+1}c_{3k+2} &= c_{1,\infty}\left(\frac{1-\theta}{\theta}\right)\frac{\left(\min\left\{1,\frac{2\theta}{1-\theta}\right\}-1\right)\frac{1}{b^2}+\frac{1}{b}}{\left(\min\left\{1,\frac{1-\theta}{2\theta}\right\}-1\right)\frac{1}{b^2}+\frac{1}{b}} \\ &\leq l_1\left(\frac{1-\theta}{\theta}\right)\frac{\frac{1}{b}}{\frac{1}{b}\left(1-\frac{1}{b}\right)} = \left(\frac{1-\theta}{\theta}\right)^3\frac{b^2}{(b-1)^3} =: l_2.\end{aligned}$$

Therefore, for any $\delta > 0$ there exists $k_0 \in \mathbb{N}$, such that whenever $k \geq k_0 + 1$

$$K := \prod_{j=k_0}^{k-1} c_{3j} c_{3j+1} c_{3j+2} < (l_2 + \delta)^{k-k_0},$$

implying

$$\begin{aligned} K c_{3k} &< (l_2 + \delta)^{k-k_0} (l_0 + \delta), \\ K c_{3k} c_{3k+1} &< (l_2 + \delta)^{k-k_0} (l_1 + \delta), \\ \text{and} \quad K c_{3k} c_{3k+1} c_{3k+2} &< (l_2 + \delta)^{k-k_0} (l_2 + \delta). \end{aligned}$$

Hence, for $i \in \{0, 1, 2\}$ and some $C > 0$

$$\prod_{j=2}^{3k+i} c_j < C(l_2 + \delta)^k.$$

Let $a_m := \prod_{j=2}^{m} c_j$, then a sufficient condition for the series $\sum_{m=2}^{\infty} a_m$ to converge, implying a transient chain according to Proposition 3.5, is $l_2 < 1$. This is the case for $b \geq 3 + \left(\frac{1-\theta}{\theta}\right)^3$, since

$$\begin{aligned} 1 - l_2 &= 1 - \left(\frac{1-\theta}{\theta}\right)^3 \frac{b^2}{(b-1)^3} = \frac{b^2}{(b-1)^3} \left(\frac{(b-1)^3}{b^2} - \left(\frac{1-\theta}{\theta}\right)^3 \right) \\ &= \frac{b^2}{(b-1)^3} \left(b - 3 + \frac{3}{b} - \frac{1}{b^2} - \left(\frac{1-\theta}{\theta}\right)^3 \right) \\ &> \frac{b^2}{(b-1)^3} \left(b - 3 - \left(\frac{1-\theta}{\theta}\right)^3 \right) \geq 0. \end{aligned}$$

Hence, the resulting noisy chain is transient if $b \geq 3 + \left(\frac{1-\theta}{\theta}\right)^3$, for any $\theta \in (0, 1)$. □

3.4.2 Section 3.2

Proof of Lemma 3.1. For any $\delta > 0$

$$\begin{aligned} \mathbb{P}\left[\frac{W_{z,N}}{W_{x,N}} \leq 1 - \delta\right] &\leq \mathbb{P}\left[W_{x,N} \geq 1 + \frac{\delta}{2}\right] + \mathbb{P}\left[W_{z,N} \leq 1 - \frac{\delta}{2}\right] \\ &\leq \mathbb{P}\left[\left|W_{x,N} - 1\right| \geq \frac{\delta}{2}\right] + \mathbb{P}\left[\left|W_{z,N} - 1\right| \geq \frac{\delta}{2}\right] \\ &\leq 2 \sup_{x \in \mathcal{X}} \mathbb{P}_{Q_{x,N}}\left[\left|W_{x,N} - 1\right| \geq \frac{\delta}{2}\right]. \end{aligned}$$

□

Proof of Lemma 3.2. Using the inequality

$$\min\{1, ab\} \geq \min\{1, a\} \min\{1, b\}, \quad \text{for } a, b \geq 0,$$

and applying Markov's inequality with $\delta > 0$,

$$\begin{aligned}
\tilde{\rho}_N(x) &= 1 - \int_{\mathcal{X}} q(x, dz)\tilde{\alpha}_N(x, z) \leq 1 - \int_{\mathcal{X}} q(x, dz)\alpha(x, z)\mathbb{E}\left[\min\left\{1, \frac{W_{z,N}}{W_{x,N}}\right\}\right] \\
&\leq 1 - (1 - \delta)\int_{\mathcal{X}} q(x, dz)\alpha(x, z)\mathbb{P}\left[\min\left\{1, \frac{W_{z,N}}{W_{x,N}}\right\} > 1 - \delta\right] \\
&= 1 - (1 - \delta)\int_{\mathcal{X}} q(x, dz)\alpha(x, z) + (1 - \delta)\int_{\mathcal{X}} q(x, dz)\alpha(x, z)\mathbb{P}\left[\frac{W_{z,N}}{W_{x,N}} \leq 1 - \delta\right] \\
&\leq 1 - (1 - \delta)(1 - \rho(x)) + \int_{\mathcal{X}} q(x, dz)\alpha(x, z)\mathbb{P}\left[\frac{W_{z,N}}{W_{x,N}} \leq 1 - \delta\right].
\end{aligned}$$

Finally, using Lemma 3.1

$$\begin{aligned}
\tilde{\rho}_N(x) &\leq \rho(x) + \delta(1 - \rho(x)) + 2\sup_{x \in \mathcal{X}} \mathbb{P}\left[\left|W_{x,N} - 1\right| \geq \frac{\delta}{2}\right](1 - \rho(x)) \\
&\leq \rho(x) + \delta + 2\sup_{x \in \mathcal{X}} \mathbb{P}\left[\left|W_{x,N} - 1\right| \geq \frac{\delta}{2}\right].
\end{aligned}$$

□

Proof of Lemma 3.3. For the first claim apply Jensen's inequality and the fact that

$$\min\{1, ab\} \leq \min\{1, a\} b, \quad \text{for } a \geq 0 \text{ and } b \geq 1,$$

hence

$$\begin{aligned}
\tilde{\alpha}_N(x, z) &\leq \min\left\{1, \frac{\pi(z)q(z, x)}{\pi(x)q(x, z)}\mathbb{E}\left[\frac{W_{z,N}}{W_{x,N}}\right]\right\} \leq \alpha(x, z)\mathbb{E}\left[W_{x,N}^{-1}\right]\mathbb{E}\left[W_{z,N}\right] \\
&= \alpha(x, z)\mathbb{E}\left[W_{x,N}^{-1}\right].
\end{aligned}$$

□

Proof of Lemma 3.4. Using the inequality

$$\min\{1, ab\} \leq \min\{1, a\} b, \quad \text{for } a \geq 0 \text{ and } b \geq 1,$$

$$
\begin{aligned}
\tilde{\alpha}_N(x,z) &= \mathbb{E}\left[\bar{\alpha}_N(x,W_{x,N};z,W_{z,N})\mathbb{1}_{\left\{\frac{W_{z,N}}{W_{x,N}}<1+\eta\right\}}\right] \\
&\quad + \mathbb{E}\left[\bar{\alpha}_N(x,W_{x,N};z,W_{z,N})\mathbb{1}_{\left\{\frac{W_{z,N}}{W_{x,N}}\geq 1+\eta\right\}}\right] \\
&\leq \alpha(x,z)\,(1+\eta)\,\mathbb{P}\left[\frac{W_{z,N}}{W_{x,N}}<1+\eta\right] + \mathbb{P}\left[\frac{W_{z,N}}{W_{x,N}}\geq 1+\eta\right] \\
&\leq \alpha(x,z)+\eta+\mathbb{P}\left[\frac{W_{z,N}}{W_{x,N}}\geq 1+\eta\right].
\end{aligned}
$$

Notice that

$$
\mathbb{P}\left[\frac{W_{z,N}}{W_{x,N}}\geq 1+\eta\right] = \mathbb{P}\left[\frac{W_{x,N}}{W_{z,N}}\leq \frac{1}{1+\eta}\right],
$$

then applying Lemma 3.1 taking $\delta = \frac{\eta}{1+\eta}$.

$$
\tilde{\alpha}_N(x,z) \leq \alpha(x,z)+\eta+2\sup_{x\in\mathcal{X}}\mathbb{P}\left[\left|W_{x,N}-1\right|\geq \frac{\eta}{2\,(1+\eta)}\right]. \qquad \square
$$

Proof of Proposition 3.4. Taking $V=\pi^{-s}$, where $0<s<\min\left\{1,\frac{a}{L}\right\}$,

$$
\begin{aligned}
\frac{qV(x)}{V(x)} &= \int_{\mathcal{X}}\frac{V(z)}{V(x)}q(x,dz) = \int_{\mathcal{X}}\left(\frac{\pi(x)}{\pi(z)}\right)^{s}q(x,dz) \\
&\leq \int_{\mathbb{R}^d}\exp\left\{a\|z-x\|\right\}q(\|z-x\|)dz.
\end{aligned}
$$

Finally, using the transformation $u=z-x$,

$$
\frac{qV(x)}{V(x)} \leq \int_{\mathbb{R}^d}\exp\left\{a\|u\|\right\}q(\|u\|)du,
$$

which implies P1*. $\square$

Chapter 4

Particular Cases and Generalisations

When weights $\{W_{x,N}\}$ are given by arithmetic averages, e.g. when using importance sampling for approximating $\pi(x)$, specific conditions can be imposed in order for the noisy chain to inherit geometric ergodicity from the marginal chain. Such conditions are explored in Section 4.1, where we also look back at some examples of Section 3.1, exploring the ergodic properties of the corresponding noisy chains as $N \to \infty$. We will see that the behaviour of the ratio of the weights, at least in the tails of the target π, plays an important role.

In Section 4.2, we discuss and provide an example of a noisy version of particle MCMC for the HMM in Figure 2.1 on page 16. Recall that the likelihood for θ, denoted by $l(\theta; y_{1:T})$, is unbiasedly estimated using $\gamma_{\theta,T}^{N}(1_T)$ given by (2.19) and as a result of the bootstrap filter described in Algorithm 2.2. Sufficient conditions on the weights $\{W_{\theta,N}\}$ are also provided, which will ensure the noisy chain inherits geometric ergodicity from the marginal chain.

The last part of this chapter, Section 4.3, is devoted to the relaxation of some assumptions on the weights $\{W_{x,N}\}$ for the general case. The ideas stated there will be of some use in the following chapter, where a modification of the noisy algorithm is presented.

4.1 Conditions for arithmetic averages

In the particular setting where the weights $\{W_{x,N}\}$ are arithmetic averages of unbiased estimators as in (2.24), i.e.

$$W_{x,N} = \frac{1}{N}\sum_{k=1}^{N} W_x^{(k)}, \qquad \text{where } \left\{W_x^{(k)}\right\}_k \overset{i.i.d.}{\sim} Q_x(\cdot) \text{ and } \mathbb{E}\left[W_x^{(k)}\right] \equiv 1,$$

specific conditions on these can be obtained to ensure geometric ergodicity is inherited from the marginal chain. For the simple case where the weights are homogeneous with respect to the state space, W1 is automatically satisfied. In order to attain W2, the existence of a negative moment for a single weight is required. In order to state this and other assumptions that follow, consider a prototypical weight W_x that has the same distribution as $W_x^{(k)}$ for any $k \in \{1, \ldots, N\}$.

Proposition 4.1. *Assume weights as in* (2.24). *If* $\mathbb{E}\left[W_x^{-1}\right] < \infty$ *then*

$$\lim_{N\to\infty} \mathbb{E}\left[W_{x,N}^{-1}\right] = 1. \tag{4.1}$$

Proof. By Jensen's inequality

$$\frac{1}{N}\sum_{k=1}^{N} \frac{1}{W_x^{(k)}} - \frac{N}{\sum_{k=1}^{N} W_x^{(k)}} \geq 0,$$

which implies, also by Jensen's inequality,

$$\mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} \frac{1}{W_x^{(k)}} - \frac{N}{\sum_{k=1}^{N} W_x^{(k)}}\right] \leq \mathbb{E}\left[W_x^{-1}\right] - 1.$$

Then, using Fatou's lemma and the law of large numbers

$$\begin{aligned}\mathbb{E}\left[W_x^{-1}\right] - 1 &\geq \limsup_{N\to\infty} \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} \frac{1}{W_x^{(k)}} - W_{x,N}^{-1}\right] \geq \liminf_{N\to\infty} \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} \frac{1}{W_x^{(k)}} - W_{x,N}^{-1}\right] \\ &\geq \mathbb{E}\left[\liminf_{N\to\infty}\left(\frac{1}{N}\sum_{k=1}^{N} \frac{1}{W_x^{(k)}}\right) - \limsup_{N\to\infty} W_{x,N}^{-1}\right] \geq \mathbb{E}\left[W_x^{-1}\right] - 1,\end{aligned}$$

hence

$$\lim_{N\to\infty} \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} \frac{1}{W_x^{(k)}} - W_{x,N}^{-1}\right] = E\left[W_x^{-1}\right] - 1. \tag{4.2}$$

Finally, since

$$\mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\frac{1}{W_x^{(k)}} - W_{x,N}^{-1}\right] = E\left[W_x^{-1}\right] - E\left[W_{x,N}^{-1}\right],$$

the expression in (4.2) becomes

$$\lim_{N\to\infty}\mathbb{E}\left[W_{x,N}^{-1}\right] = 1. \qquad \square$$

For homogeneous weights, (4.1) implies W2. When the weights are not homogeneous, stronger conditions are needed for W1 and W2 to be satisfied. An appropriate first assumption is that the weights are uniformly integrable.

Assumption (W3)**.** *The weights* $\{W_x\}_x$ *satisfy*

$$\lim_{K\to\infty}\sup_{x\in\mathcal{X}}\mathbb{E}\left[W_x\mathbb{1}\left(W_x > K\right)\right] = 0.$$

The second condition imposes an additional assumption on the distribution of the weights $\{W_x\}_x$ near 0.

Assumption (W4)**.** *There exists* $\gamma \in (0,1)$ *and constants* $M < \infty$, $\beta > 0$ *such that for* $w \in (0,\gamma)$ *the weights* $\{W_x\}_x$ *satisfy*

$$\sup_{x\in\mathcal{X}}\mathbb{P}\left[W_x \le w\right] \le Mw^{\beta}.$$

These new conditions ensure W1 and W2 are satisfied.

Proposition 4.2. *For weights as in* (2.24)*,*

(i). W3 implies W1;

(ii). W1 and W4 imply W2.

The following corollary is obtained as an immediate consequence of the above proposition, Theorem 3.1 and Theorem 3.2.

Corollary 4.1. *Let the weights be as in* (2.24)*. Assume W3 and either*

(i). P1, P2 and W4;

(ii). P1 and P2.*

Then, there exists $N_0 \in \mathbb{N}$ *such that for all* $N \ge N_0$*, the noisy chain with transition kernel* $\tilde{P}_N$ *is geometrically ergodic.*

The proof of Proposition 4.2 follows the statement and proof of Lemma 4.1. This lemma allows us to characterise the distribution of $W_{x,N}$ near 0 assuming W4 and also provides conditions for the existence and convergence of negative moments.

Lemma 4.1. *Let $\gamma \in (0,1)$ and $p > 0$.*

(i). Suppose Z is a positive random variable, and assume that for $z \in (0, \gamma)$

$$\mathbb{P}[Z \leq z] \leq M z^{\alpha}, \quad \text{where } \alpha > p, M < \infty.$$

Then,

$$\mathbb{E}\left[Z^{-p}\right] \leq \frac{1}{\gamma^p} + pM\frac{\gamma^{\alpha-p}}{\alpha - p}.$$

(ii). Suppose $\{Z_i\}_{i=1}^N$ is a collection of positive and independent random variables, and assume that for each $i \in \{1, \ldots, N\}$ and $z \in (0, \gamma)$

$$\mathbb{P}[Z_i \leq z] \leq M_i z^{\alpha_i}, \quad \text{where } \alpha_i > 0, M_i < \infty.$$

Then, for $z \in (0, \gamma)$

$$\mathbb{P}\left[\sum_{i=1}^N Z_i \leq z\right] \leq \left(\prod_{i=1}^N M_i\right) z^{\sum_{i=1}^N \alpha_i}.$$

(iii). Let the weights be as in (2.24). If for some $N_0 \in \mathbb{N}$

$$\mathbb{E}\left[W_{x,N_0}^{-p}\right] < \infty,$$

then for any $N \geq N_0$

$$\mathbb{E}\left[W_{x,N+1}^{-p}\right] \leq \mathbb{E}\left[W_{x,N}^{-p}\right].$$

(iv). Assume W1 and let $g : \mathbb{R}^+ \to \mathbb{R}$ be a function that is continuous at 1 and bounded on the interval $[\gamma, \infty)$. Then

$$\lim_{N\to\infty} \sup_{x\in\mathcal{X}} \mathbb{E}\left[|g(W_{x,N}) - g(1)| \mathbb{1}_{W_{x,N}\geq\gamma}\right] = 0.$$

Proof. The proof of (i) is motivated by Piegorsch and Casella (1985, Theorem 2.1) and Khuri and Casella (2002, Theorem 3), however the existence of a density func-

tion is not assumed here. Since Z^{-p} is positive,

$$\begin{aligned}\mathbb{E}\left[Z^{-p}\right] &= \int_{\mathbb{R}^+} \mathbb{P}\left[Z^{-p} \geq z\right] dz \leq \frac{1}{\gamma^p} + \int_{(\gamma^{-p},\infty)} \mathbb{P}\left[Z^{-p} \geq z\right] dz \\ &= \frac{1}{\gamma^p} + \int_{(0,\gamma)} pu^{-p-1}\mathbb{P}\left[Z \leq u\right] du \leq \frac{1}{\gamma^p} + pM\frac{\gamma^{\alpha-p}}{\alpha-p}.\end{aligned}$$

For part (ii), since the random variables $\{Z_i\}$ are positive, then for any $z > 0$

$$\mathbb{P}\left[\sum_{i=1}^{N} Z_i \leq z\right] = \mathbb{P}\left[\sum_{i=1}^{N} Z_i \leq z, \max_{i\in\{1,\ldots,N\}} \{Z_i\} \leq z\right].$$

Therefore, for $z \in (0, \gamma)$

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^{N} Z_i \leq z\right] &\leq \mathbb{P}\left[\max_{i\in\{1,\ldots,N\}} \{Z_i\} \leq z\right] = \prod_{i=1}^{N} \mathbb{P}\left[Z_i \leq z\right] \\ &\leq \left(\prod_{i=1}^{N} M_i\right) z^{\sum_{i=1}^{N}\alpha_i}.\end{aligned}$$

Part (iii) can be seen as a consequence of $W_{x,N}$ and $W_{x,N+1}$ being convex ordered and $g(x) = x^{-p}$ being a convex function for $x > 0$ and $p \geq 0$, (see, e.g., Andrieu and Vihola, 2016). We provide a self-contained proof by defining for $j \in \{1, \ldots, N+1\}$

$$S_{x,N}^{(j)} := \frac{1}{N} \sum_{k=1,k\neq j}^{N+1} W_x^{(k)},$$

and we have

$$W_{x,N+1} = \frac{1}{N+1} \sum_{j=1}^{N+1} S_{x,N}^{(j)}$$

and since the arithmetic mean is greater than or equal to the geometric mean

$$W_{x,N+1} \geq \left(\prod_{j=1}^{N+1} S_{x,N}^{(j)}\right)^{\frac{1}{N+1}}.$$

This implies for $p > 0$

$$\mathbb{E}\left[W_{x,N+1}^{-p}\right] \leq \mathbb{E}\left[\left(\prod_{j=1}^{N+1} S_{x,N}^{(j)}\right)^{-\frac{p}{N+1}}\right] \leq \prod_{j=1}^{N+1}\left(\mathbb{E}\left[\left(S_{x,N}^{(j)}\right)^{-p}\right]\right)^{\frac{1}{N+1}}$$
$$= \mathbb{E}\left[\left(S_{x,N}^{(1)}\right)^{-p}\right] = \mathbb{E}\left[W_{x,N}^{-p}\right],$$

where Hölder's inequality has been used and the fact that the random variables $\left\{S_{x,N}^{(j)} : j \in 1, \ldots, N+1\right\}$ are identically distributed according to $Q_{x,N}$.

For part (iv), let $M_\gamma = \sup_{y\in[\gamma,\infty)} |g(y)|$ and due to continuity at $y = 1$, for any $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\begin{aligned}\mathbb{E}\left[\left|g\left(W_{x,N}\right) - g(1)\right| \mathbb{1}_{W_{x,N}\in[\gamma,\infty)}\right] &\leq 2M_\gamma \mathbb{P}\left[\gamma \leq W_{x,N} \leq 1-\delta\right] \\ &+ 2M_\gamma \mathbb{P}\left[1+\delta \leq W_{x,N}\right] + \mathbb{E}\left[\left|g\left(W_{x,N}\right) - g(1)\right| \mathbb{1}_{W_{x,N}\in(1-\delta,1+\delta)}\right] \\ &\leq 2M_\gamma \mathbb{P}\left[\left|W_{x,N}-1\right| \geq \delta\right] + \varepsilon \mathbb{P}\left[\left|W_{x,N}-1\right| < \delta\right].\end{aligned}$$

Therefore, for fixed ε and by W1

$$\begin{aligned}\lim_{N\to\infty} \sup_{x\in\mathcal{X}} \mathbb{E}\left[\left|g\left(W_{x,N}\right) - g(1)\right| \mathbb{1}_{W_{x,N}\in[\gamma,\infty)}\right] &\\ \leq 2M_\gamma \lim_{N\to\infty} \sup_{x\in\mathcal{X}} \mathbb{P}\left[\left|W_{x,N}-1\right| \geq \delta\right] + \varepsilon &\\ \leq \varepsilon,&\end{aligned}$$

obtaining the result since ε can be picked arbitrarily small. □

Proof of Proposition 4.2. Part (i) is a consequence of Chandra (1989, Theorem 1). Assuming W3, it implies

$$\lim_{N\to\infty} \sup_{x\in\mathcal{X}} \mathbb{E}\left[\left|W_{x,N}-1\right|\right] = 0.$$

By Markov's inequality

$$\mathbb{E}\left[\left|W_{x,N}-1\right|\right] \geq \delta \mathbb{P}\left[\left|W_{x,N}-1\right| \geq \delta\right],$$

and the result follows.

To prove (ii), assume W4 and by part (ii) of Lemma 4.1, for $w \in (0, \gamma)$

$$\mathbb{P}\left[NW_{x,N} \leq w\right] \leq M^N w^{N\beta}.$$

Take $p > 1$ and define $N_0 := \lfloor \frac{p}{\beta} \rfloor + 1$, then using part (i) of Lemma 4.1 if $N \geq N_0$

$$\sup_{x \in \mathcal{X}} \mathbb{E}\left[W_{x,N}^{-p}\right] \leq \frac{N}{\gamma^p} + pNM^N \frac{\gamma^{N\beta - p}}{N\beta - p}.$$

Hence, by Hölder's inequality

$$\begin{aligned}\mathbb{E}\left[\left|W_{x,N}^{-1} - 1\right| \mathbb{1}_{W_{x,N} \in (0,\gamma)}\right] &\leq \mathbb{E}\left[W_{x,N}^{-1} \mathbb{1}_{W_{x,N} \in (0,\gamma)}\right] \\ &\leq \left(\mathbb{E}\left[W_{x,N}^{-p}\right]\right)^{\frac{1}{p}} \left(\mathbb{P}\left[W_{x,N} < \gamma\right]\right)^{\frac{p-1}{p}},\end{aligned}$$

and applying part (iii) of Lemma 4.1, for $N \geq N_0$

$$\mathbb{E}\left[\left|W_{x,N}^{-1} - 1\right| \mathbb{1}_{W_{x,N} \in (0,\gamma)}\right] \leq \left(\mathbb{E}\left[W_{x,N_0}^{-p}\right]\right)^{\frac{1}{p}} \left(\mathbb{P}\left[W_{x,N} < \gamma\right]\right)^{\frac{p-1}{p}}.$$

Therefore,

$$\sup_{x \in \mathcal{X}} \mathbb{E}\left[\left|W_{x,N}^{-1} - 1\right| \mathbb{1}_{W_{x,N} \in (0,\gamma)}\right] \leq \left(\sup_{x \in \mathcal{X}} \mathbb{E}\left[W_{x,N_0}^{-p}\right]\right)^{\frac{1}{p}} \left(\sup_{x \in \mathcal{X}} \mathbb{P}\left[W_{x,N} < \gamma\right]\right)^{\frac{p-1}{p}}.$$

Since $\gamma < 1$ and by W1

$$\lim_{N \to \infty} \sup_{x \in \mathcal{X}} \mathbb{P}\left[W_{x,N} < \gamma\right] = 0,$$

implying

$$\lim_{N \to \infty} \sup_{x \in \mathcal{X}} \mathbb{E}\left[\left|W_{x,N}^{-1} - 1\right| \mathbb{1}_{W_{x,N} \in (0,\gamma)}\right] = 0. \tag{4.3}$$

Now, for fixed $\gamma \in (0,1)$ the function $g(x) = x^{-1}$ is bounded and continuous on $[\gamma, \infty)$, implying by part (iv) of Lemma 4.1

$$\lim_{N \to \infty} \sup_{x \in \mathcal{X}} \mathbb{E}\left[\left|W_{x,N}^{-1} - 1\right| \mathbb{1}_{W_{x,N} \in [\gamma,\infty)}\right] = 0. \tag{4.4}$$

Finally, using (4.3) and (4.4)

$$\lim_{N \to \infty} \sup_{x \in \mathcal{X}} \mathbb{E}\left[\left|W_{x,N}^{-1} - 1\right|\right] = 0,$$

and by the triangle inequality

$$\sup_{x \in \mathcal{X}} \mathbb{E}\left[\left|W_{x,N}^{-1} - 1\right|\right] \geq \sup_{x \in \mathcal{X}} \mathbb{E}\left[W_{x,N}^{-1}\right] - 1,$$

the result follows. □

4.1.1 Remarks on some examples

Equipped with the previous results, we return to the examples in Section 3.1.2 and Section 3.1.3. Even though the noisy chain can be transient in these examples, the behaviour is quite different when considering weights that are arithmetic averages of the form in (2.24). Since in both examples the weights are uniformly bounded by the constant b, they immediately satisfy W1. Additionally, by Proposition 4.1, condition W2 is satisfied for the example in Section 3.1.2. This is not the case for example in Section 3.1.3, but condition P1* is satisfied by taking $V = \pi^{-\frac{1}{2}}$. Therefore, applying Theorem 3.1 and Theorem 3.2 to examples in Section 3.1.2 and in Section 3.1.3 respectively, as N increases the corresponding chains will go from being transient to geometrically ergodic.

Despite conditions W1 and W2 guaranteeing the inheritance of geometric ergodicity for the noisy chain, they are not necessary. Consider a modification of the example in Section 3.1.2, where the weights are given by

$$W_{m,1} = (b_m - \varepsilon_m)Ber(s_m) + \varepsilon_m,$$
$$\text{where } b_m > 1 \text{ and } \varepsilon_m \in (0,1] \text{ for all } m \geq 1.$$

Again, there exists a relationship between the variables b_m, ε_m and s_m for ensuring the expectation of the weights is equal to one. Let $Bin\,(N,s)$ denote a binomial distribution of parameters $N \in \mathbb{N}$ and $s \in (0,1)$. Then, in the arithmetic average context, $W_{m,N}$ becomes

$$W_{m,N} = \frac{(b_m - \varepsilon_m)}{N} Bin\,(N, s_m) + \varepsilon_m, \qquad \text{where } b_m > 1 \text{ and } \varepsilon_m \in (0,1] \text{ for all } m \geq 1. \tag{4.5}$$

For particular choices of the sequences $\{b_m\}_{m\in\mathbb{N}}$ and $\{\varepsilon_m\}_{m\in\mathbb{N}}$, the resulting noisy chain can be geometrically ergodic for all $N \geq 1$, even though neither W1 nor W2 hold.

Proposition 4.3. *Consider a geometric target density as in* (3.6) *and a proposal density as in* (3.7)*. In addition, let the weights be as in* (4.5) *with* $b_m \to \infty$, $\varepsilon_m \to 0$ *as* $m \to \infty$ *and*

$$\lim_{m\to\infty} \frac{\varepsilon_{m-1}}{\varepsilon_m} = l, \quad \text{where } l \in \mathbb{R}^+ \cup \{+\infty\}.$$

Then, the chain generated by the noisy kernel $\tilde{P}_N$ is geometrically ergodic for any $N \in \mathbb{N}$.

Proof. First notice that if $l < \infty$ then $l \geq 1$. To see this, define

$$a_m := \frac{\varepsilon_{m-1}}{\varepsilon_m},$$

then for fixed $\delta > 0$, there exists $M \in \mathbb{N}$ such that for $m \geq M$

$$a_m < l + \delta.$$

Then, for $m \geq M$

$$\frac{\varepsilon_1}{\varepsilon_m} = \prod_{j=2}^{m} a_j < (l+\delta)^{m-M} \frac{\varepsilon_1}{\varepsilon_M},$$

and because $\varepsilon_m \to 0$, it is clear that $(l+\delta)^m \to \infty$ as $m \to \infty$. Therefore, $l + \delta > 1$ and since δ can be taken arbitrarily small, it is true that $l \geq 1$.

Now, for weights as in (4.5) and using a simple random walk proposal, the noisy acceptance probability can be expressed as

$$\begin{aligned} \tilde{\alpha}_N(m, m-1) = \sum_{j=0}^{N} \sum_{k=0}^{N} \min\left\{1, \frac{2\theta}{1-\theta} \frac{b_{m-1} j + (N-j)\,\varepsilon_{m-1}}{b_m k + (N-k)\,\varepsilon_m}\right\} \binom{N}{j} \\ \times \binom{N}{k} (s_{m-1})^j (s_m)^k (1-s_{m-1})^{N-j} (1-s_m)^{N-k} \end{aligned} \tag{4.6}$$

and

$$\begin{aligned} \tilde{\alpha}_N(m, m+1) = \sum_{j=0}^{N} \sum_{k=0}^{N} \min\left\{1, \frac{1-\theta}{2\theta} \frac{b_{m+1} j + (N-j)\,\varepsilon_{m+1}}{b_m k + (N-k)\,\varepsilon_m}\right\} \binom{N}{j} \\ \times \binom{N}{k} (s_{m+1})^j (s_m)^k (1-s_{m+1})^{N-j} (1-s_m)^{N-k}. \end{aligned} \tag{4.7}$$

Since $b_m \to \infty$, then $s_m \to 0$ as $m \to \infty$; therefore, any term in (4.6) and (4.7), for which $j + k \neq 0$, tends to zero as $m \to \infty$. Hence,

$$\tilde{\alpha}_N(m, m-1) = \min\left\{1, \frac{2\theta}{1-\theta} \times \frac{\varepsilon_{m-1}}{\varepsilon_m}\right\} (1-s_{m-1})^N (1-s_m)^N + o(1),$$

and

$$\tilde{\alpha}_N(m, m+1) = \min\left\{1, \frac{1-\theta}{2\theta} \times \frac{\varepsilon_{m+1}}{\varepsilon_m}\right\} (1 - s_{m+1})^N (1 - s_m)^N + o(1),$$

implying

$$\lim_{m\to\infty} \frac{\tilde{P}_N(m, \{m-1\})}{\tilde{P}_N(m, \{m+1\})} = \frac{(1-\theta) \lim_{m\to\infty} \min\left\{1, \frac{2\theta}{1-\theta} \times \frac{\varepsilon_{m-1}}{\varepsilon_m}\right\}}{\theta \lim_{m\to\infty} \min\left\{1, \frac{1-\theta}{2\theta} \times \frac{\varepsilon_{m+1}}{\varepsilon_m}\right\}}. \tag{4.8}$$

If $l = +\infty$, (4.8) tends to $+\infty$, whereas if $l < \infty$

$$\lim_{m\to\infty} \frac{\tilde{P}_N(m, \{m-1\})}{\tilde{P}_N(m, \{m+1\})} = 2l \frac{\min\{1-\theta, 2\theta l\}}{\min\{2\theta l, 1-\theta\}} \geq 2.$$

In any case, this implies

$$\lim_{m\to\infty} \tilde{P}_N(m, \{m-1\}) \geq 2 \lim_{m\to\infty} \tilde{P}_N(m, \{m+1\}),$$

and since

$$\lim_{m\to\infty} \tilde{P}_N(m, \{m-1\}) = \min\{1-\theta, 2\theta l\} > 0,$$

the noisy chain is geometrically ergodic according to Proposition 3.5. □

Finally, in many of the previous examples, increasing the value of N seems to improve the ergodic properties of the noisy chain. However, the geometric ergodicity property is not always inherited, no matter how large N is taken. The following proposition shows an example rather similar to Proposition 4.3, but in which the ratio $\varepsilon_{m-1}/\varepsilon_m$ does not converge as $m \to \infty$.

Proposition 4.4. *Consider a geometric target density as in* (3.6) *and a proposal density as in* (3.7)*. In addition, let the weights be as in* (4.5) *with* $b_m = m$ *and*

$$\varepsilon_m = m^{-(3-(m \pmod 3))}.$$

Then, the chain generated by the noisy kernel $\tilde{P}_N$ *is transient for any* $N \in \mathbb{N}$.

Proof. Noting that

$$\frac{\varepsilon_{m-1}}{\varepsilon_m} \in \begin{cases} O\left(m^2\right) & \\ O\left(m^{-1}\right) & \end{cases} \text{if} \begin{array}{c} m \ (\text{mod } 3) = 0, \\ m \ (\text{mod } 3) \in \{1,2\}, \end{array}$$

$$\text{and} \quad \frac{\varepsilon_{m+1}}{\varepsilon_m} \in \begin{cases} O\left(m^{-2}\right) & \\ O\left(m\right) & \end{cases} \text{if} \begin{array}{c} m \ (\text{mod } 3) = 2, \\ m \ (\text{mod } 3) \in \{0,1\}, \end{array}$$

expressions in (4.6) and (4.7) become

$$\tilde{\alpha}_N(m, m-1) = (1-s_{m-1})^N (1-s_m)^N \mathbb{1}\left(m \ (\text{mod } 3) = 0\right) + O\left(m^{-1}\right),$$

and

$$\tilde{\alpha}_N(m, m+1) = (1-s_{m+1})^N (1-s_m)^N \mathbb{1}\left(m \ (\text{mod } 3) = 0,1\right) + O\left(m^{-1}\right).$$

Therefore,

$$\begin{aligned} \frac{\tilde{P}_N\left(m, \{m-1\}\right)}{\tilde{P}_N\left(m, \{m+1\}\right)} = \left(\frac{1-\theta}{\theta}\right) \frac{(1-s_{m-1})^N + O(m^{-1})}{(1-s_{m+1})^N + O(m^{-1})} \mathbb{1}\left(m \ (\text{mod } 3) = 0\right) \\ + O\left(m^{-1}\right) \mathbb{1}\left(m \ (\text{mod } 3) = 1\right) \\ + O(1) \mathbb{1}\left(m \ (\text{mod } 3) = 2\right), \end{aligned}$$

implying there exists $C \in \mathbb{R}^+$ such that for $j = 0, 2$

$$\lim_{k\to\infty} \frac{\tilde{P}_N\left(3k+j, \{3k+j-1\}\right)}{\tilde{P}_N\left(3k+j, \{3k+j+1\}\right)} \leq C,$$

and

$$\lim_{k\to\infty} \frac{\tilde{P}_N\left(3k+1, \{3k\}\right)}{\tilde{P}_N\left(3k+1, \{3k+2\}\right)} = 0.$$

Then, for fixed $\delta > 0$ there exists $k_0 \in \mathbb{N}$ such that whenever $k \geq k_0$

$$\frac{\tilde{P}_N\left(3k+j, \{3k+j-1\}\right)}{\tilde{P}_N\left(3k+j, \{3k+j+1\}\right)} < C + \delta, \quad \text{for } j = 0, 2$$

and

$$\frac{\tilde{P}_N\left(3k+1, \{3k\}\right)}{\tilde{P}_N\left(3k+1, \{3k+2\}\right)} < \delta.$$

Let

$$c_m := \frac{\tilde{P}_N(m, \{m-1\})}{\tilde{P}_N(m, \{m+1\})},$$

then for $k \geq k_0 + 1$

$$\prod_{j=2}^{3k+1} c_j = \prod_{j=1}^{k} c_{3j-1}c_{3j}c_{3j+1} \leq \left((C+\delta)^2 \delta\right)^{k-k_0} \prod_{j=1}^{k_0} c_{3j-1}c_{3j}c_{3j+1}.$$

Take δ small enough, such that $(C+\delta)^2 \delta < 1$, hence

$$\begin{aligned}
\sum_{k=1}^{\infty} \prod_{j=2}^{3k+1} c_j &= \sum_{k=1}^{k_0} \prod_{j=2}^{3k+1} c_j + \sum_{k=k_0}^{\infty} \prod_{j=2}^{3k+1} c_j \\
&\leq \sum_{k=1}^{k_0} \prod_{j=2}^{3k+1} c_j + \prod_{j=1}^{k_0} c_{3j-1}c_{3j}c_{3j+1} \sum_{k=k_0}^{\infty} \left((C+\delta)^2 \delta\right)^{k-k_0} \\
&= \sum_{k=1}^{k_0} \prod_{j=2}^{3k+1} c_j + \frac{\prod_{j=1}^{k_0} c_{3j-1}c_{3j}c_{3j+1}}{1-(C+\delta)^2 \delta} < \infty.
\end{aligned}$$

Similarly, it can be proved that

$$\sum_{k=0}^{\infty} \prod_{j=2}^{3k+2} c_j < \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \prod_{j=2}^{3k} c_j < \infty,$$

thus

$$\sum_{m=2}^{\infty} \prod_{j=2}^{m} c_j < \infty,$$

implying the noisy chain is transient according to Proposition 3.5. □

4.1.2 Rate of convergence

We conclude the arithmetic average setting with a result providing a rate of convergence for the total variation between $\tilde{\pi}_N$ and π. In order to obtain an explicit expression for $r(N)$, appearing in Theorem 3.4 and such that

$$\sup_{x \in \mathcal{X}} \|\tilde{P}_N(x, \cdot) - P(x, \cdot)\|_{TV} \leq \frac{1}{r(N)},$$

we must impose a slightly stronger assumption than W3.

Assumption (W3*). *There exists $k > 0$, such that the weights $\{W_x\}_x$ satisfy*

$$\sup_{x \in \mathcal{X}} \mathbb{E}\left[W_x^{1+k}\right] < \infty.$$

Proposition 4.5. *Assume P1, P2, W4 and W3*. Alternatively, assume P1*, P2 and W3*. Then, there exists $D_k > 0$ and $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$,*

$$\|\tilde{\pi}_N(\cdot) - \pi(\cdot)\|_{TV} \leq D_k \frac{\log(N)}{N^{\frac{\tau}{2+k}}},$$

where $\tau = k$ if $k \in (0,1)$ and $\tau = \frac{1+k}{2}$ if $k \geq 1$. If in addition W3 holds for all $k > 0$, then for any $\varepsilon \in (0, 1/6)$ there will exist $D_\varepsilon > 0$ and $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$,*

$$\|\tilde{\pi}_N(\cdot) - \pi(\cdot)\|_{TV} \leq D_\varepsilon \frac{\log(N)}{N^{\frac{1}{2}-\varepsilon}}.$$

Proof. From (3.11) and taking $\delta < \frac{1}{2}$, $\eta = \frac{\delta}{1-\delta}$

$$\sup_{x \in \mathcal{X}} \|\tilde{P}_N(x,\cdot) - P(x,\cdot)\|_{TV} \leq 3\delta + 4 \sup_{x \in \mathcal{X}} \mathbb{P}\left[\left|W_{x,N} - 1\right| \geq \frac{\delta}{2}\right],$$

implying by Markov's inequality

$$\begin{aligned}\sup_{x \in \mathcal{X}} \|\tilde{P}_N(x,\cdot) - P(x,\cdot)\|_{TV} &\leq 3\delta + 4 \sup_{x \in \mathcal{X}} \mathbb{P}\left[\left|W_{x,N} - 1\right|^{1+k} \geq \left(\frac{\delta}{2}\right)^{1+k}\right] \\ &\leq 3\delta + \frac{2^{3+k}}{\delta^{1+k}} \sup_{x \in \mathcal{X}} \mathbb{E}\left[\left|W_{x,N} - 1\right|^{1+k}\right].\end{aligned}$$

By the Marcinkiewicz-Zygmund inequality for i.i.d random variables, see e.g. Gut (2013, Chapter 3, Corollary 8.2), there exists $B_k < \infty$ such that

$$\mathbb{E}\left[\left|W_{x,N} - 1\right|^{1+k}\right] \leq B_k \mathbb{E}\left[\left|W_x - 1\right|^{1+k}\right] N^{-\tau},$$

where

$$\tau = \begin{cases} k & \text{if } k \in (0,1) \\ \frac{1+k}{2} & \text{if } k \geq 1. \end{cases}$$

Therefore,

$$\sup_{x\in\mathcal{X}} \|\tilde{P}_N(x,\cdot) - P(x,\cdot)\|_{TV} \leq 3\delta + \frac{2^{3+k}B_k}{\delta^{1+k}N^{\tau}} \sup_{x\in\mathcal{X}} \mathbb{E}\left[\left|W_x - 1\right|^{1+k}\right].$$

Now, let

$$C_k = B_k \sup_{x\in\mathcal{X}} \mathbb{E}\left[\left|W_x - 1\right|^{1+k}\right],$$

then the convex function $f : \mathbb{R}^+ \to \mathbb{R}^+$ where

$$f(s) = 3s + \frac{2^{3+k}C_k}{s^{1+k}N^{\tau}},$$

is minimised at

$$s_* = \left(\frac{(1+k)2^{3+k}C_k}{3N^{\tau}}\right)^{\frac{1}{2+k}} = O\left(N^{-\frac{\tau}{2+k}}\right).$$

Then,

$$\begin{aligned}\sup_{x\in\mathcal{X}} \|\tilde{P}_N(x,\cdot) - P(x,\cdot)\|_{TV} \leq f(s_*) &= O\left(N^{-\frac{\tau}{2+k}}\right) + O\left(N^{-\tau+\frac{\tau(1+k)}{2+k}}\right) \\ &= O\left(N^{-\frac{\tau}{2+k}}\right).\end{aligned}$$

Applying Theorem 3.4 by taking $r(N) \propto N^{\frac{\tau}{2+k}}$, the result is obtained.

For the second claim, take $k_\varepsilon \geq (2\varepsilon)^{-1} - 2 \geq 1$ and apply the first part. □

4.2 Conditions for noisy particle MCMC

More complex examples arise when using particle MCMC methods, for which noisy versions can also be performed. They may prove to be useful in some inference problems as seen in the following example. Recall that in this setting the weights $\{W_{\theta,N}\}$ are given by (2.25), i.e.

$$W_{\theta,N} = \frac{\gamma^N_{\theta,T}(1_T)}{\gamma_{\theta,T}(1_T)},$$

where $\gamma_{\theta,T}(1_T) = l(\theta; y_{1:T})$ is the likelihood function for θ in the HMM from Figure 2.1 on page 16, and $\gamma^N_{\theta,T}(1_T)$ is the unbiased estimator based on N particles as a result of Algorithm 2.2.

Example 4.1. Consider the hidden Markov model given by Figure 2.1 on page 16, where $X_0 = x_0$ is fixed, and let

$$m_\theta(x_{n-1}, \cdot) = \mathcal{N}\left(ax_{n-1}, \sigma_X^2\right) \quad \text{and} \quad g_\theta(\cdot\,|x_n) = \mathcal{N}\left(x_n, \sigma_Y^2\right).$$

In this case, the set of parameters θ is

$$\theta = \{x_0, a, \log(\sigma_X), \log(\sigma_Y)\},$$

and once a prior distribution for θ is specified, p_0 say, the aim is to conduct Bayesian inference on the posterior distribution

$$\pi(\theta\,|y_{1:T}) \propto p_0(\theta)\, l(\theta; y_{1:T}).$$

As noted in Chapter 2, for this particular setting the posterior distribution is tractable. This will allows us to compare the results obtained from the marginal, pseudo-marginal and noisy chains, the latter two relying on the SMC estimator $\gamma_{\theta,T}^N(1_T)$ of the likelihood, defined in (2.19) and obtained using Algorithm 2.2. Using uniform improper priors over $\mathbb{R}$ for the parameters and a random walk proposal, Figure 4.1 shows the run and autocorrelation function (ACF) for the autoregressive parameter a of the marginal chain.

Similarly, Figure 4.2 shows the corresponding run and ACF for both the pseudo-marginal and the noisy chain when $N = 250$. It is noticeable how the pseudo-marginal gets "stuck", resulting in a lower acceptance than the marginal and noisy chains. In addition, the ACF of the noisy chain seems to decay faster than that of the pseudo-marginal chain.

Finally, Figure 4.3 and Figure 4.4 show the estimated posterior densities for the parameters when $N = 250$ and $N = 750$, respectively. There, the trade-off between the pseudo-marginal and the noisy algorithm is noticeable. For lower values of N, the pseudo-marginal will require more iterations due to the slow mixing, whereas the noisy converges faster towards an unknown noisy invariant distribution. By increasing N, the mixing in the pseudo-marginal improves and the noisy invariant approaches the true posterior.

4.2.1 Bounding the relative variance

In the noisy particle MCMC context, obtaining specific conditions on the weights $\{W_{\theta,N}\}$ is not as straightforward as for arithmetic averages. The main reason is because expressions for the moments of the estimated likelihood $\gamma_{\theta,T}^N(1_T)$ are usually

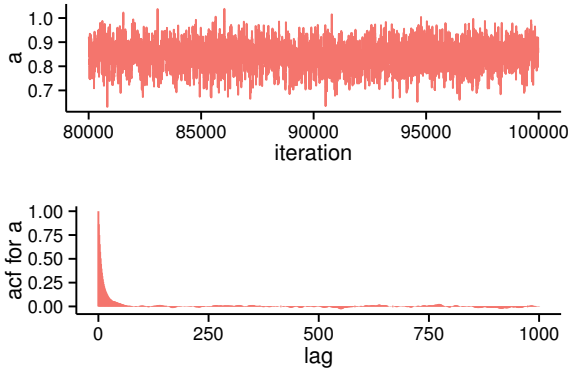


Figure 4.1: Last 20,000 iterations of the marginal algorithm for the autoregressive parameter a (top). Estimated ACF of the corresponding marginal chain (bottom). The mean acceptance probability was 0.256

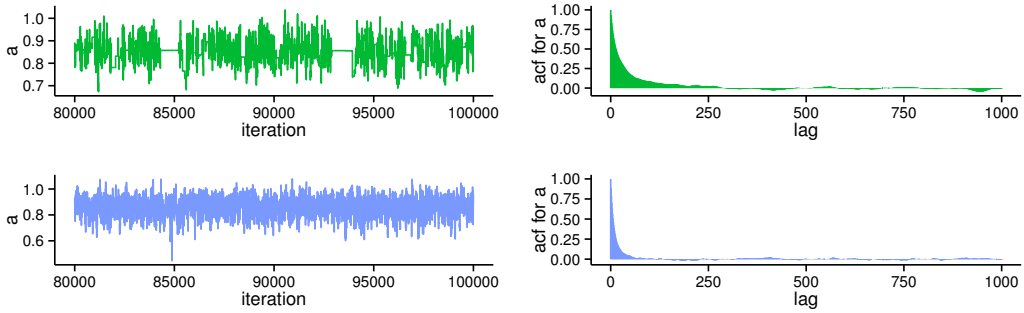


Figure 4.2: Last 20,000 iterations of the pseudo-marginal (top left) and noisy (bottom left) algorithms, for the autoregressive parameter a when $N = 250$. Estimated ACFs of the corresponding pseudo-marginal (top right) and noisy (bottom right) chains. The mean acceptance probabilities were 0.104 for the pseudo-marginal and 0.283 for the noisy chain

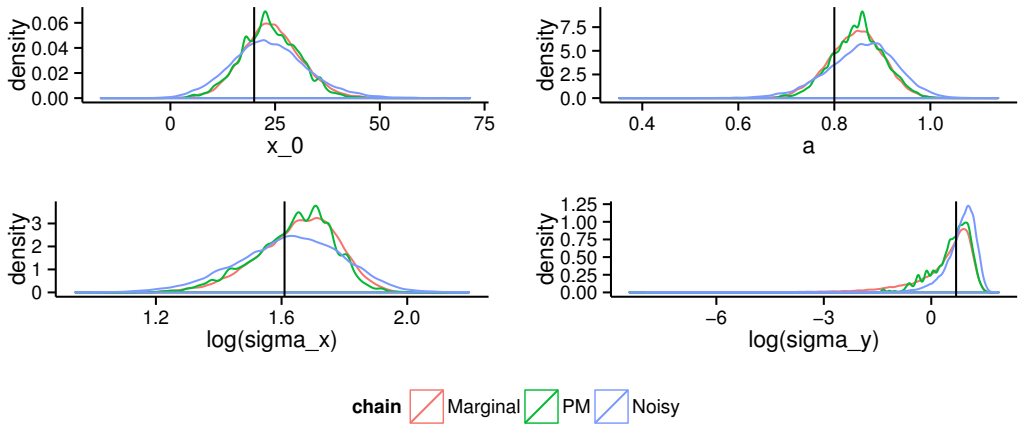


Figure 4.3: Estimated densities using the marginal, pseudo-marginal and noisy chains for the 4 parameters when $N = 250$. Vertical lines indicate the real values.

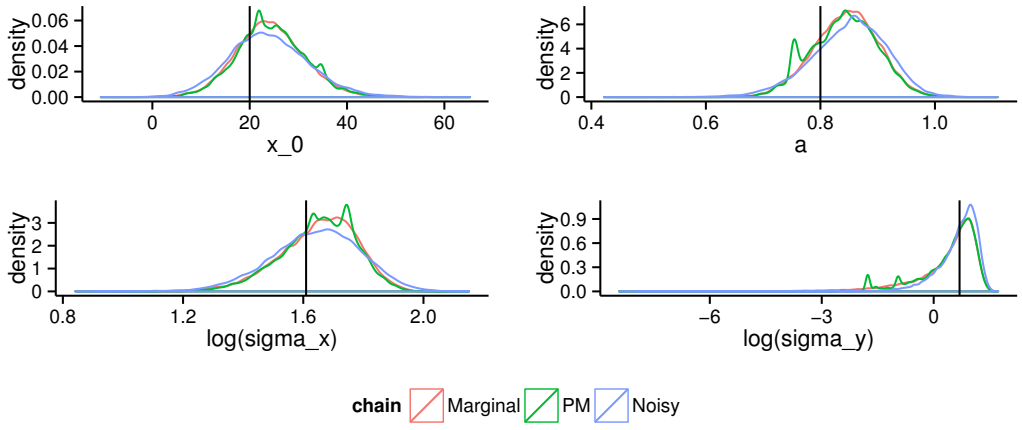


Figure 4.4: Estimated densities using the marginal, pseudo-marginal and noisy chains for the 4 parameters, when $N = 750$. Vertical lines indicate the real values.

unavailable. Nevertheless, the relative variance of $\gamma_{\theta,T}^{N}(1_T)$ has been object of major study, which will allow us to explore a sufficient conditions leading to W1.

Consider a vector $b = (b_1, \ldots, b_{T+1}) \in \{0,1\}^{T+1}$, from Lee and Whiteley (2015, Lemma 2) the second moment of $\gamma_{\theta,T}^{N}(1_T)$ can be expressed by

$$\mathbb{E}\left[\gamma_{\theta,T}^{N}(1_T)^2\right] = \sum_{b\in\{0,1\}^{T+1}} \prod_{n=1}^{T+1} \left\{ \left(\frac{1}{N}\right)^{b_n} \left(1-\frac{1}{N}\right)^{1-b_n} \right\} \mu_{\theta,T}(b), \tag{4.9}$$

where

$$\mu_{\theta,T}(b) := \mathbb{E}\left[\prod_{n=1}^{T} g_\theta(y_n|X_n)\, g_\theta\left(y_n\middle|X_n'\right)\right], \tag{4.10}$$

and $(X_n, X_n')_{1\leq n\leq T}$ is a joint Markov chain on $\mathcal{X}^2$ distributed as follows:

- If $b_1 = 0$, then $X_1 \sim m_\theta(x_0,\cdot)$ and $X_1' \sim m_\theta(x_0,\cdot)$ independently; whereas if $b_1 = 1$ then $X_1' = X_1 \sim m_\theta(x_0,\cdot)$.
- For $n \in \{2,\ldots,T\}$, if $b_n = 0$ then $X_n \sim m_\theta(X_{n-1},\cdot)$ and $X_n' \sim m_\theta\left(X_{n-1}',\cdot\right)$ independently; if $b_n = 1$ then $X_n' = X_n \sim m_\theta(x_{n-1},\cdot)$.

An equivalent expression for the aforementioned second moment was first proved in Cérou et al. (2011, Proposition 3.4), however for our purposes (4.9) will be more useful. The following assumption will imply W1, as proved in the proposition stated immediately after.

Assumption (W5)**.** *For fixed $T \in \mathbb{N}$ and for any $b \in \{0,1\}^{T+1}$*

$$\sup_{\theta\in\Theta} \frac{\mu_{\theta,T}(b)}{\gamma_{\theta,T}(1_T)^2} < \infty.$$

Proposition 4.6. *For weights as in* (2.25)*, W5 implies W1.*

Proof. By Markov's inequality, it is enough to show that

$$\lim_{N\to\infty} \sup_{\theta} \mathbb{V}[W_{\theta,N}] = 0. \tag{4.11}$$

Notice that for the zero vector, denoted by 0_{T+1}, $\mu_{\theta,T}(0_{T+1}) = \gamma_{\theta,T}(1_T)^2$. Hence,

using (4.9),

$$\begin{aligned}
\mathbb{V}\left[W_{\theta,N}\right] &= \frac{\mathbb{E}\left[\gamma_{\theta,T}^{N}\left(1_T\right)^2\right]}{\gamma_{\theta,T}\left(1_T\right)^2} - 1 = \sum_{j=1}^{T+1} \left(\begin{array}{c} T+1 \\ j \end{array} \right) (-1)^j N^{-j} \\
&+ \sum_{b \neq 0_{T+1}} \prod_{n=1}^{T+1} \left\{ \left(\frac{1}{N}\right)^{b_n} \left(1 - \frac{1}{N}\right)^{1-b_n} \right\} \frac{\mu_{\theta,T}\left(b\right)}{\gamma_{\theta,T}\left(1_T\right)^2} \\
&= O\left(N^{-1}\right) + \left(1 - \frac{1}{N}\right)^{T+1} \sum_{b \neq 0_{n+1}} \frac{1}{(N-1)^{\sum_{n=1}^{T+1} b_n}} \left(\frac{\mu_{\theta,T}\left(b\right)}{\gamma_{\theta,T}\left(1_T\right)^2} \right) \\
&= O\left(N^{-1}\right) + O\left(N^{-1}\right) \sum_{b \neq 0_{n+1}} \frac{\mu_{\theta,T}\left(b\right)}{\gamma_{\theta,T}\left(1_T\right)^2}.
\end{aligned}$$

The result follows by taking the supremum over θ and then $N \to \infty$. □

Similarly to Corollary 4.1, the following corollary is obtained as an immediate consequence of the above proposition and Theorem 3.2.

Corollary 4.2. *Let the weights be as in* (2.25)*. Assume P1*, P2 and W5. Then, there exists* $N_0 \in \mathbb{N}$ *such that for all* $N \geq N_0$*, the noisy chain with transition kernel* $\tilde{P}_N$ *is geometrically ergodic.*

4.2.2 Simpler but stronger condition

Although Assumption W5 may be a direct condition for inheriting geometric ergodicity, it is probably difficult to verify. Instead, one could rely on the following condition on the conditional density g_θ.

Assumption (W5*)**.** *For each* $n \in \{1, \ldots, T\}$ *there exists* $C_n \in (0, \infty)$ *such that the conditional density* g_θ *satisfies*

$$C_n := \sup_{x, x', \theta} \frac{g_\theta(y_n|x)}{g_\theta(y_n|x')} < \infty.$$

Proposition 4.7. *W5* implies W5.*

Proof. Consider two independent copies of the chain $X_{1:T}$, say $\bar{X}_{1:T}^{(1)}$ and $\bar{X}_{1:T}^{(2)}$. Then,

using Jensen's inequality,

$$\frac{\mu_{\theta,T}(b)}{\gamma_{\theta,T}(1_T)^2} = \frac{\mathbb{E}\left[\prod_{n=1}^T g_\theta(y_n|X_n)\, g_\theta(y_n|X'_n)\right]}{\mathbb{E}\left[\prod_{n=1}^T g_\theta\left(y_n\middle|\bar{X}_n^{(1)}\right)\right]\mathbb{E}\left[\prod_{n=1}^T g_\theta\left(y_n\middle|\bar{X}_n^{(2)}\right)\right]}$$
$$\leq \mathbb{E}\left[\prod_{n=1}^T \frac{g_\theta(y_n|X_n)\, g_\theta(y_n|X'_n)}{g_\theta\left(y_n\middle|\bar{X}_n^{(1)}\right) g_\theta\left(y_n\middle|\bar{X}_n^{(2)}\right)}\right] \leq \prod_{n=1}^T C_n^2 < \infty,$$

as required. □

However, W5* turns out to be quite strong since it is rarely satisfied in practice, not even for the simple and tractable model in Example 4.1. This assumption will mainly hold if both, the state space of the hidden chain $\mathcal{X}$ and the parameter space Θ are compact. As stated below, if W5* holds then the weights are uniformly bounded, which will also imply W2.

Proposition 4.8. *Under W5*, the weights are uniformly positive and uniformly bounded in θ. Hence, W2 is satisfied.*

Proof. Let $\bar{X}_{1:T}$ be an independent copy of the chain $X_{1:T}$. On one side, if W5* holds and using Jensen's inequality

$$W_{\theta,N} = \frac{\gamma_{\theta,T}^N(1_T)}{\gamma_{\theta,T}(1_T)} = \frac{\prod_{n=1}^T \left\{\frac{1}{N}\sum_{j=1}^N g_\theta\left(y_n\middle|X_n^{(j)}\right)\right\}}{\mathbb{E}\left[\prod_{n=1}^T g_\theta\left(y_n|\bar{X}_n\right)\right]}$$
$$\leq \mathbb{E}\left[\prod_{n=1}^T \frac{1}{N}\sum_{j=1}^N \frac{g_\theta\left(y_n\middle|X_m^{(j)}\right)}{g_\theta\left(y_n|\bar{X}_n\right)}\middle|\left\{X_{1:T}^{(j)}\right\}_{j=1}^N\right] \leq \prod_{n=1}^T C_n < \infty.$$

Similarly, using properties of the arithmetic and harmonic averages,

$$W_{\theta,N}^{-1} = \mathbb{E}\left[\prod_{n=1}^T \frac{g_\theta\left(y_n|\bar{X}_n\right)}{\frac{1}{N}\sum_{j=1}^N g_\theta\left(y_n\middle|X_n^{(j)}\right)}\middle|\left\{X_{1:T}^{(j)}\right\}_{j=1}^N\right]$$
$$\leq \mathbb{E}\left[\prod_{n=1}^T \frac{1}{N}\sum_{j=1}^N \frac{g_\theta\left(y_n|\bar{X}_n\right)}{g_\theta\left(y_n\middle|X_n^{(j)}\right)}\middle|\left\{X_{1:T}^{(j)}\right\}_{j=1}^N\right] \leq \prod_{n=1}^T C_n < \infty.$$

Hence, by the Cauchy-Schwartz inequality,

$$\mathbb{E}\left[\left|W_{\theta,N}^{-1}-1\right|\right]=\mathbb{E}\left[\frac{|W_{\theta,N}-1|}{W_{\theta,N}}\right]\leq\left(\mathbb{V}\left[W_{\theta,N}\right]\right)^{1/2}\left(\mathbb{E}\left[W_{\theta,N}^{-2}\right]\right)^{1/2}$$
$$=\left(\mathbb{V}\left[W_{\theta,N}\right]\right)^{1/2}\prod_{n=1}^{T}C_n,$$

implying

$$\sup_{\theta}\mathbb{E}\left[\left|W_{\theta,N}^{-1}-1\right|\right]\leq\left(\sup_{\theta}\mathbb{V}\left[W_{\theta,N}\right]\right)^{1/2}\prod_{n=1}^{T}C_n.$$

The result follows from (4.11). □

Finally, throughout this section the number of observations T has been considered fixed, focusing only on increasing N irrespective of T. If instead, we allow T to increase, the variance of $W_{\theta,N}$ could in principle be controlled using the bound in (2.20). However, stronger conditions than W5* are usually required, see e.g. Cérou et al. (2011, Theorem 5.1).

4.3 Towards a more general method

Until now, we have studied a particular perturbation of the marginal chain with transition kernel P, involving a randomised acceptance ratio leading to the following noisy acceptance

$$\tilde{\alpha}_N\left(x,y\right):=\mathbb{E}\left[\min\left\{1,r\left(x,y\right)\frac{W_{y,N}}{W_{x,N}}\right\}\right],$$

where $W_{y,N}$ and $W_{x,N}$ are independent positive random variables with expectation 1. However, the above assumptions on the weights $W_{y,N}$ and $W_{x,N}$ can be relaxed, allowing us to encompass more general settings as done in Alquier et al. (2014).

Looking back at Theorem 3.2, the result is still valid even when $W_{y,N}$ and $W_{x,N}$ are dependent or when they have expectation different to 1, as long as W1 holds. In this section, we rely on a similar condition that can provide stronger results. This is Assumption W0 (stated below), which controls the difference between $\tilde{\alpha}_N$ and α and the dissimilarity between the kernels $\tilde{P}_N$ and P, as stated in the lemma immediately after.

Assumption (W0). *The weights satisfy*

$$\lim_{N\to\infty}\sup_{x\in\mathcal{X}}\mathbb{E}\left[\left|\log\left(W_{x,N}\right)\right|\right]=0.$$

Lemma 4.2. *Suppose* $\sup_{x\in\mathcal{X}}\mathbb{E}\left[\left|\log\left(W_{x,N}\right)\right|\right]<\infty$*, then*

$$\sup_{(x,y)\in\mathcal{X}^2}\left|\tilde{\alpha}_N\left(x,y\right)-\alpha\left(x,y\right)\right|\leq 2\sup_{x\in\mathcal{X}}\mathbb{E}\left[\left|\log\left(W_{x,N}\right)\right|\right]$$

and

$$\sup_{x\in\mathcal{X}}\left\|\tilde{P}_N\left(x,\cdot\right)-P\left(x,\cdot\right)\right\|_{TV}\leq 4\sup_{x\in\mathcal{X}}\mathbb{E}\left[\left|\log\left(W_{x,N}\right)\right|\right].$$

Proof. First, by Jensen's inequality

$$\begin{aligned}\left|\tilde{\alpha}_N\left(x,y\right)-\alpha\left(x,y\right)\right| &= \left|\left[\mathbb{E}\min\left\{1,r\left(x,y\right)\frac{W_{y,N}}{W_{x,N}}\right\}\right]-\min\left\{1,r\left(x,y\right)\right\}\right| \\ &\leq \mathbb{E}\left[\left|\min\left\{1,r\left(x,y\right)\frac{W_{y,N}}{W_{x,N}}\right\}-\min\left\{1,r\left(x,y\right)\right\}\right|\right].\end{aligned}$$

The function $g\left(x\right)=\min\left\{1,\exp\left\{x\right\}\right\}$ is Lipschitz with coefficient 1, hence

$$\begin{aligned}&\left|\tilde{\alpha}_N\left(x,y\right)-\alpha\left(x,y\right)\right|\leq\mathbb{E}\left[\left|\min\left\{1,r\left(x,y\right)\frac{W_{y,N}}{W_{x,N}}\right\}-\min\left\{1,r\left(x,y\right)\right\}\right|\right] \\ &\leq\mathbb{E}\left[\left|\log\left(r\left(x,y\right)\frac{W_{y,N}}{W_{x,N}}\right)-\log\left(r\left(x,y\right)\right)\right|\right]\leq\mathbb{E}\left[\left|\log\left(W_{y,N}\right)\right|\right]+\mathbb{E}\left[\left|\log\left(W_{x,N}\right)\right|\right] \\ &\leq 2\sup_{x}\mathbb{E}\left[\left|\log\left(W_{x,N}\right)\right|\right],\end{aligned}$$

and the first result follows.

For the second claim, take $A\in\mathcal{B}\left(\mathcal{X}\right)$ and apply the triangle and Jensen's inequalities to obtain

$$\begin{aligned}&\left|\tilde{P}_N\left(x,A\right)-P\left(x,A\right)\right|\leq\left|\int_A\left[\tilde{\alpha}_N\left(x,y\right)-\alpha\left(x,y\right)\right]q\left(x,dy\right)\right| \\ &+\left|\int_{\mathcal{X}}\left[\tilde{\alpha}_N\left(x,y\right)-\alpha\left(x,y\right)\right]q\left(x,dy\right)\right|\leq 2\int_{\mathcal{X}}\left|\tilde{\alpha}_N\left(x,y\right)-\alpha\left(x,y\right)\right|q\left(x,dy\right) \\ &\leq 2\sup_{x,y}\left|\tilde{\alpha}_N\left(x,y\right)-\alpha\left(x,y\right)\right|.\end{aligned}$$

□

Similarly to the results in Chapter 3, under Assumption W0 and using the previous lemma it is possible to show (provided P1 and P2 hold and N is sufficiently

large) the existence of a noisy invariant $\tilde{\pi}_N$ for $\tilde{P}_N$, and also

$$\lim_{N\to\infty} \|\tilde{\pi}_N(\cdot) - \pi(\cdot)\|_{TV} = 0.$$

This is not pursued here, instead we focus on an interesting consequence of Lemma 4.2 when the weights are given by an arithmetic average as in (2.24). For such case a bound on

$$\sup_{x\in\mathcal{X}} \mathbb{E}\left[\left|\log\left(W_{x,N}\right)\right|\right]$$

in terms of N is available, which in turn provides a sharper rate of convergence as $N \to \infty$ for

$$\left\|\tilde{\pi}_N\left(\cdot\right) - \pi\left(\cdot\right)\right\|_{TV}.$$

This is discussed next and accompanied with a neat example. In the following chapter, the idea of relaxing assumptions on the weights $W_{y,N}$ and $W_{x,N}$ is developed in more depth. There, motivated by a correction to the noisy algorithm, a novel approximate method is analysed.

4.3.1 Arithmetic averages revisited

We introduce the following assumption on the single variables W_x, that guarantees W0 holds.

Assumption (W0*)**.** *The weights* $\{W_x\}_x$ *satisfy*

$$\sup_{x\in\mathcal{X}} \mathbb{V}\left[W_x\right] < \infty \qquad \textit{and} \qquad \sup_{x\in\mathcal{X}} \mathbb{E}\left[W_x^{-1}\right] < \infty.$$

Remark 4.1. Notice that W0* implies Assumptions W1 and W2. In addition and combined with part (*iii*) of Lemma 4.1, Assumption W0* implies

$$\sup_{x,N} \mathbb{E}\left[W_{x,N}^{-1}\right] < \infty.$$

The convergence result for the kernels $\tilde{P}_N$ and P is now presented. As a consequence, an improved rate of convergence for $\tilde{\pi}_N$ and π is achieved in comparison to Proposition 4.5. In contrast to such result, notice (apart from the ε improvement)

the much weaker requirement of

$$\sup_x \mathbb{V}\left[W_x\right] < \infty,$$

as opposed to W3* holding for all $k > 0$.

Proposition 4.9. *For weights as in* (2.24) *and under W0**

$$\sup_{x \in \mathcal{X}} \left\| \tilde{P}_N\left(x, \cdot\right) - P\left(x, \cdot\right) \right\|_{TV} \leq O\left(N^{-1/2}\right).$$

Additionally, if P1 and P2 hold then

$$\left\| \tilde{\pi}_N\left(\cdot\right) - \pi\left(\cdot\right) \right\|_{TV} \leq O\left(N^{-1/2} \log\left(N\right)\right).$$

Proof. First, notice that the logarithmic function satisfies for any $x > 0$

$$\left|\log\left(x\right)\right| \leq \frac{\left|x - 1\right|}{\sqrt{x}}.$$

Hence

$$\sup_x \mathbb{E}\left[\left|\log\left(W_{x,N}\right)\right|\right] \leq \sup_x \mathbb{E}\left[\left|\frac{W_{x,N} - 1}{\sqrt{W_{x,N}}}\right|\right],$$

and using the Cauchy-Schwartz inequality

$$\begin{aligned}\sup_x \mathbb{E}\left[\left|\log\left(W_{x,N}\right)\right|\right] &\leq \left(\sup_x \mathbb{E}\left[W_{x,N}^{-1}\right]\right)^{1/2} \left(\sup_x \mathbb{V}\left[W_{x,N}\right]\right)^{1/2} \\ &= O\left(N^{-1/2}\right).\end{aligned}$$

Therefore, by Lemma 4.2

$$\sup_x \left\| \tilde{P}_N\left(x, \cdot\right) - P\left(x, \cdot\right) \right\|_{TV} \leq O\left(N^{-1/2}\right).$$

The second claim follows directly from Theorem 3.4 by applying the the first part. □

We now look at an example which shows the rate of convergence for the kernels $\tilde{P}_N$ and P in Theorem 4.9 is tight. It also shows that better rates may be possible if the marginal acceptance ratio $r\left(x, y\right)$ lies outside a neighbourhood around 1.

Proposition 4.10. *Consider a geometric target distribution π on the non-negative integers with density*

$$\pi(x) \propto \left(\frac{1}{2}\right)^x \mathbb{1}(x \in \mathbb{N}_0).$$

In addition, take a random walk proposal on the integers where $q(x, x+1) = \theta = 1 - q(x, x-1)$ and let $W_x \sim Exp(1)$. Then, the noisy and marginal kernels satisfy

$$\sup_x \left\| \tilde{P}_N(x, \cdot) - P(x, \cdot) \right\|_{TV} = \begin{cases} O\left(N^{-1}\right) & \text{if } \theta \neq \frac{1}{3} \\ O\left(N^{-1/2}\right) & \text{if } \theta = \frac{1}{3}. \end{cases}$$

Proof. The arithmetic average assumption implies

$$\frac{W_{y,N}}{W_{x.N}} \stackrel{d}{=} \frac{G_1}{G_2} \stackrel{d}{=} \frac{V}{1-V},$$

where $G_i \overset{i.i.d}{\sim} Gamma(N, 1)$ and $V \sim Beta(N, N)$. Let $B(x, y)$ denote the beta function, then for $x \in \mathbb{N}_0$

$$\begin{aligned} \tilde{P}_N(x, x+1) &= \theta \mathbb{E}\left[\min\left\{1, \left(\frac{1-\theta}{2\theta}\right)\frac{V}{1-V}\right\}\right] \\ &= \theta\left[\left(\frac{1-\theta}{2\theta}\right)\frac{B(N+1, N-1)}{B(N,N)} I_{\frac{2\theta}{1+\theta}}(N+1, N-1) + 1 - I_{\frac{2\theta}{1+\theta}}(N,N)\right] \\ &= \theta\left[\left(\frac{1-\theta}{2\theta}\right)\frac{N}{N-1} I_{\frac{2\theta}{1+\theta}}(N+1, N-1) + 1 - I_{\frac{2\theta}{1+\theta}}(N,N)\right], \end{aligned}$$

where $I_z(x, y)$ is the regularised incomplete beta function given by

$$I_z(x,y) = \frac{\int_0^z v^{x-1}(1-v)^{y-1} dv}{B(x,y)}, \quad \text{for } z \in [0,1].$$

Similarly, for $x \in \mathbb{N}$

$$\begin{aligned} \tilde{P}_N(x, x-1) &= (1-\theta)\mathbb{E}\left[\min\left\{1, \left(\frac{2\theta}{1-\theta}\right)\frac{V}{1-V}\right\}\right] \\ &= (1-\theta)\left[\left(\frac{2\theta}{1-\theta}\right)\frac{N}{N-1} I_{\frac{1-\theta}{1+\theta}}(N+1, N-1) + 1 - I_{\frac{1-\theta}{1+\theta}}(N,N)\right]. \end{aligned}$$

Then, if $\theta < \frac{1}{3}$, and using Proposition B.2 in Appendix B

$$\tilde{P}_N(x, x+1) = \theta + o\left(N^{-1}\right) \quad \text{and} \quad \tilde{P}_N(x, x-1) = 2\theta + O\left(N^{-1}\right),$$

whereas if $\theta > \frac{1}{3}$

$$\tilde{P}_N(x, x+1) = \frac{1}{2}(1-\theta) + O\left(N^{-1}\right) \quad \text{and} \quad \tilde{P}_N(x, x-1) = 1 - \theta + o\left(N^{-1}\right).$$

Therefore, if $\theta \neq \frac{1}{3}$

$$\sup_x \left\| \tilde{P}_N(x, \cdot) - P(x, \cdot) \right\|_{TV} = O\left(N^{-1}\right).$$

If $\theta = \frac{1}{3}$

$$\begin{aligned} \tilde{P}_N(x, x+1) &= \frac{1}{3}\left[\frac{N}{N-1} I_{\frac{1}{2}}(N+1, N-1) + \frac{1}{2}\right] \\ &= \frac{1}{3}\left[\frac{N}{N-1}\left(\frac{1}{2} - \frac{1}{2^{2N-1} N B(N, N)}\right) + \frac{1}{2}\right], \end{aligned}$$

and applying Proposition B.1 in Appendix B

$$\tilde{P}_N(x, x+1) = \frac{1}{3} + O\left(N^{-1/2}\right).$$

Similarly,

$$\tilde{P}_N(x, x-1) = \frac{2}{3} + O\left(N^{-1/2}\right),$$

which implies

$$\sup_x \left\| \tilde{P}_N(x, \cdot) - P(x, \cdot) \right\|_{TV} = O\left(N^{-1/2}\right)$$

when $\theta = \frac{1}{3}$. □

As a final remark, notice that the bound for $\|\tilde{\pi}_N(\cdot) - \pi(\cdot)\|_{TV}$ in Proposition 4.9 is not tight since for the above example and if $\theta \neq \frac{1}{3}$

$$\|\tilde{\pi}_N(\cdot) - \pi(\cdot)\|_{TV} = O\left(N^{-1}\right).$$

In addition, if $\theta = \frac{1}{3}$ the noisy invariant $\tilde{\pi}_N$ is in fact equal to π since for $x \neq 0$

$$\tilde{\pi}_N(x) \propto \left(\frac{\theta \mathbb{E}\left[\min\left\{1, \left(\frac{1-\theta}{2\theta}\right)\frac{V}{1-V}\right\}\right]}{(1-\theta)\,\mathbb{E}\left[\min\left\{1, \left(\frac{2\theta}{1-\theta}\right)\frac{V}{1-V}\right\}\right]} \right)^x = \left(\frac{1}{2}\right)^x.$$

Chapter 5

Approximate MCMC via Geometric Averages

The noisy MH algorithm may prove useful when dealing with an intractable distribution π on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. For its implementation, one relies on the availability of unbiased estimates of the point-wise evaluation of the target density. These are

$$\hat{\pi}_x := \pi(x) W_x \tag{5.1}$$

where $W_x \sim Q_x(\cdot)$ taking values in $\mathcal{W} \subseteq \mathbb{R}^+$ and satisfying $\mathbb{E}[W_x] \equiv 1$ for all $x \in \mathcal{X}$. With such estimates, the usual acceptance ratio $r(x, y)$ given by

$$r(x, y) = \frac{\pi(y) q(y, x)}{\pi(x) q(x, y)},$$

is estimated using

$$R^{noi}_{x,y} := \frac{\hat{\pi}_y q(y, x)}{\hat{\pi}_x q(x, y)} = r(x, y) \frac{W_y}{W_x}. \tag{5.2}$$

Usually, as discussed in Sections 4.1 and 4.2, the variables $\{W_x\}_{x \in \mathcal{X}}$ may be also indexed by $N \in \mathbb{N}$ such that $W_{x,N} \overset{a.s.}{\to} 1$ as $N \to \infty$. Even though the chain produced by the noisy algorithm is not invariant under π, for some cases and under some conditions discussed throughout Chapter 3, a noisy invariant distribution $\tilde{\pi}_N$ exists for the noisy chain. Additionally, as $N \to \infty$, $\tilde{\pi}_N$ can converge (in an appropriate sense) to the desired target π. In some cases, however, $\tilde{\pi}_N$ may not provide a good approximation of the target unless N is very large. Hence, with N fixed, is it possible to modify the noisy algorithm in order to obtain a better approximation of π?

In this chapter, we study a plausible approach for (approximately) correcting

chains generated by acceptance probabilities of the form

$$\min\{1, R_{x,y}\}, \tag{5.3}$$

where $R_{x,y}$ is a random approximation for the usual ratio $r(x,y)$. This correction is motivated by a class of algorithms with randomised acceptance as in (5.3), that produce reversible chains with respect to the target π. Hence, the idea is to modify $R_{x,y}$ in such a way that the resulting chain is π-reversible or "close" to being so. From this point onwards, the subscript N is dropped since it is merely a distraction for this chapter.

Consider the following generic form for $R_{x,y}$, serving as proxy for $r(x,y)$,

$$R_{x,y} := r(x,y)\, V_{x,y}, \tag{5.4}$$

where $V_{x,y} \sim \nu_{x,y}(\cdot)$ having support $\mathcal{V} \subseteq \mathbb{R}^+$. Notice that the ratio $R^{noi}_{x,y}$ used in the noisy MH algorithm is a special case of (5.4), where $V_{x,y} = W_y/W_x$. A Markov chain generated by the randomised acceptance rule in (5.3) will satisfy detailed balance in (2.2) if the following condition (similar to (2.12)) holds

$$r(y,x)\,\frac{\mathbb{E}\left[\min\{1, R_{x,y}\}\right]}{\mathbb{E}\left[\min\{1, R_{y,x}\}\right]} = 1. \tag{5.5}$$

Notice that

$$\begin{aligned}
\mathbb{E}\left[\min\{1, R_{x,y}\}\right] &= \mathbb{E}\int_0^\infty \mathbb{1}\left(u \le \min\{1, R_{x,y}\}\right) du \\
&= \int_0^\infty \mathbb{1}\left(u \le 1\right) \mathbb{P}\left[u \le R_{x,y}\right] du = \int_0^1 \mathbb{P}\left[V_{x,y} \ge \frac{u}{r(x,y)}\right] du \\
&= r(x,y) \int_0^{1/r(x,y)} \mathbb{P}\left[V_{x,y} \ge u\right] du,
\end{aligned}$$

implying that the left hand side of (5.5) can be expressed as

$$\begin{aligned}
r(y,x)\,\frac{\mathbb{E}\left[\min\{1, R_{x,y}\}\right]}{\mathbb{E}\left[\min\{1, R_{y,x}\}\right]} &= r(y,x)\,\frac{r(x,y)\int_0^{1/r(x,y)} \mathbb{P}\left[V_{x,y} \ge u\right] du}{r(y,x)\int_0^{1/r(y,x)} \mathbb{P}\left[V_{y,x} \ge u\right] du} \\
&= \frac{r(x,y)\int_0^{1/r(x,y)} \mathbb{P}\left[V_{x,y} \ge u\right] du}{\int_0^{r(x,y)} \mathbb{P}\left[V_{y,x} \ge u\right] du}.
\end{aligned} \tag{5.6}$$

The expression above is close to $\mathbb{E}\left[V_{x,y}\right]$ (provided this value is finite) if $r(x,y)$ is small, whereas for large values of $r(x,y)$ the expression is close to $\left(\mathbb{E}\left[V_{y,x}\right]\right)^{-1}$.

Therefore, assuming $\mathbb{E}[V_{x,y}] < \infty$ for all $(x, y) \in \mathcal{X}^2$, one possible modification to the acceptance ratio $R_{x,y}$ is to consider instead

$$\tilde{R}_{x,y} := \frac{R_{x,y}}{\mathbb{E}[V_{x,y}]}. \tag{5.7}$$

In O'Neill et al. (2000) and Beaumont (2003), the authors suggest using the following modified acceptance instead of $R^{noi}_{x,y}$ in (5.2)

$$R^{noi}_{x,y}\frac{r(x,y)}{\mathbb{E}\left[R^{noi}_{x,y}\right]} = R^{noi}_{x,y}\frac{1}{\mathbb{E}\left[W_y W_x^{-1}\right]} = R^{noi}_{x,y}\frac{1}{\mathbb{E}\left[W_x^{-1}\right]},$$

which in fact coincides with (5.7) when $V_{x,y} = W_y/W_x$ and W_y is independent of W_x. However, this correction is not applicable in general since $\mathbb{E}[V_{x,y}] = \mathbb{E}\left[W_x^{-1}\right]$ is usually unknown. Additionally, it is not clear if for moderate values of $r(x, y)$ the correction is of some use.

The rest of this chapter is organised as follows. In Section 5.1, we take a look at a class of algorithms defined in terms of randomised acceptance ratios that are exact, in the sense that the resulting chain is invariant under the target π. Also there, we will see that for specific cases the corrected ratio in (5.7) produces an exact algorithm. In Section 5.2, a new approximate MCMC algorithm (Algorithm 5.1) is presented and motivated by the penalty method, the latter belonging to the aforementioned class of methods with randomised ratios. Section 5.3 contains some examples for which Algorithm 5.1 is applicable. Even though some of these examples are quite simple or artificial, we will be able to identify the existing trade-off when implementing and tuning Algorithm 5.1. Finally, in Section 5.4, we present some theoretical results dealing with a diffusion limit of the approximate chain and the convergence of the approximate stationary distribution towards π.

5.1 Exact algorithms with randomised acceptance

Recall the generic form of a randomised acceptance ratio given by (5.4). Since the resulting chain may not be reversible with respect to π, would it be possible to transform $V_{x,y}$ in such way that the new chain is reversible? Consider a function $h_{x,y} : \mathcal{V} \to [0, \infty)$ for fixed $(x, y) \in \mathcal{X}^2$, detailed balance condition in (5.5) can be

expressed by

$$\pi(x)\, q(x,y) \int_{\mathcal{V}} \nu_{x,y}(u) \min\{1, r(x,y)\, h_{x,y}(u)\}\, du \tag{5.8}$$
$$= \pi(y)\, q(y,x) \int_{\mathcal{V}} \nu_{y,x}(\bar{u}) \min\{1, r(y,x)\, h_{y,x}(\bar{u})\}\, d\bar{u}.$$

Taking $\bar{u} = g_{x,y}(u)$ for some invertible and continuously differentiable function $g_{x,y} : \mathcal{V} \to \mathcal{V}$, a sufficient condition for satisfying (5.8) is

$$h_{x,y}(u) = \frac{\nu_{y,x}(\bar{u})}{\nu_{x,y}(u)} \left|\frac{\partial \bar{u}}{\partial u}\right| \quad \text{and} \quad h_{y,x}(\bar{u}) = \frac{\nu_{x,y}(u)}{\nu_{y,x}(\bar{u})} \left|\frac{\partial \bar{u}}{\partial u}\right|^{-1},$$

since this implies

$$\pi(x) q(x,y)\, \nu_{x,y}(u) \min\{1, r(x,y)\, h_{x,y}(u)\}$$
$$= \pi(y)\, q(y,x)\, \nu_{y,x}(g_{x,y}(u)) \min\{1, r(y,x)\, h_{y,x}(g_{x,y}(u))\} \left|\frac{\partial}{\partial u} g_{x,y}(u)\right|. \tag{5.9}$$

By the inverse function theorem

$$h_{y,x}(\bar{u}) = \frac{\nu_{x,y}\left(g_{x,y}^{-1}(\bar{u})\right)}{\nu_{y,x}(\bar{u})} \left|\frac{\partial}{\partial \bar{z}} g_{x,y}^{-1}(\bar{u})\right|,$$

which imposes an involution-type condition, in this case $g_{x,y}^{-1} = g_{y,x}$. Additionally, notice that the above arguments remain valid for more general domains $\mathcal{V} \subseteq \mathbb{R}$.

Theorem 5.1. *Let $V_{x,y} \sim \nu_{x,y}(\cdot)$ defined on $\mathcal{V} \subseteq \mathbb{R}$ and suppose $g_{x,y} : \mathcal{V} \to \mathcal{V}$ is a continuously differentiable function satisfying $g_{x,y}^{-1} = g_{y,x}$, the Markov chain with proposal q generated by the acceptance ratio*

$$\bar{R}_{x,y} := r(x,y) \frac{\nu_{y,x}(g_{x,y}(V_{x,y}))}{\nu_{x,y}(V_{x,y})} \left|\frac{\partial}{\partial u} g_{x,y}(u)\Big|_{u=V_{x,y}}\right|, \tag{5.10}$$

is invariant under π.

This type of chain was first studied in Nicholls et al. (2012) where the 'very detailed balance' condition in (5.9) was established. In Maire et al. (2014) this idea is studied in more generality in the context of inhomogeneous Markov chains. We now discuss two particular cases for which $\bar{R}_{x,y}$ provides an explicit expression for the acceptance probability

$$\mathbb{E}\left[\min\left\{1, \bar{R}_{x,y}\right\}\right].$$

5.1.1 Barker's acceptance

The chain generated using Barker's acceptance

$$\alpha_B(x,y) = \frac{r(x,y)}{1+r(x,y)}$$

can be seen as a special case of the randomised acceptance setting. Barker's algorithm dates back to Barker (1965) and, as discussed in Section 2.1.3, it is less efficient in terms of asymptotic variance when compared to a MH algorithm using the same proposal q. Nevertheless, the authors in Łatuszyński and Roberts (2013) have shown that the asymptotic variance of Barker's algorithm is at most, roughly speaking, two times that of MH. Additionally, Barker's acceptance is everywhere differentiable (in terms of $r \equiv r(x,y)$), which may be more appealing in situations where derivatives of the acceptance $\alpha(r)$ are required.

We do not discuss Barker's acceptance further on since it is not the main focus of this chapter. However, up to our knowledge, this is the first time the acceptance probability in Barker's algorithm is regarded as an expectation of a randomised MH acceptance. The following proposition states such result.

Proposition 5.1. *Let $V \sim \nu(\cdot)$ have a density with respect to μ^{Leb} given by*

$$\nu(u) = 2(1+u)^{-3}\mathbb{1}(u>0),$$

and take the involution on $\mathbb{R}^+$ given by $g(u) = u^{-1}$. Then, using $\bar{R}_{x,y}$ defined in (5.10),

$$\mathbb{E}\left[\min\left\{1, \bar{R}_{x,y}\right\}\right] = \alpha_B(x,y).$$

Proof. Set $r = r(x,y)$, then

$$\begin{aligned}
\mathbb{E}\left[\min\left\{1, \bar{R}_{x,y}\right\}\right] &= \mathbb{E}\left[\min\left\{1, r\frac{\left(1+V^{-1}\right)^{-3}}{V^2(1+V)^{-3}}\right\}\right] \\
&= \mathbb{E}\left[\min\left\{1, rV\right\}\right] = \int_0^{r^{-1}} 2ru(1+u)^{-3}\,du + \int_{r^{-1}}^{\infty} 2(1+u)^{-3}\,du \\
&= \frac{r}{(1+r)^2} + \frac{r^2}{(1+r)^2} = \frac{r}{1+r}.
\end{aligned}$$

□

5.1.2 The penalty method

The second example, which is our main focus, appears when log-unbiased estimates of the acceptance ratio $r(x,y)$ are available. As noted in Nicholls et al. (2012), if

$$\log\left(R_{x,y}\right) \sim \mathcal{N}\left(\log\left(r\left(x,y\right)\right), \tau_{x,y}^2\right), \tag{5.11}$$

it is possible to obtain an exact algorithm by making a correction to the randomised acceptance. This idea was first introduced in Ceperley and Dewing (1999) under the name of the penalty method, with an increasing relevance in recent work, see e.g. Deligiannidis et al. (2015) or Yıldırım (2016).

Proposition 5.2. *Consider the randomised acceptance in* (5.4)*, where* $R_{x,y}$ *satisfies* (5.11) *and* $\tau_{x,y}^2 = \tau_{y,x}^2$ *for all* $(x,y) \in \mathcal{X}^2$*. Therefore, the acceptance ratio of the penalty method given by*

$$\bar{R}_{pen}\left(x,y\right) := R_{x,y} \exp\left\{-\frac{1}{2}\tau_{x,y}^2\right\},$$

defines an exact algorithm.

Proof. Let $Z_{x,y} \sim \mathcal{N}\left(0, \tau_{x,y}^2\right)$ and consider the involution on $\mathbb{R}$ given by $g_{x,y}\left(z\right) = \tau_{x,y}^2 - z$, applying Theorem 5.1 the acceptance ratio

$$\bar{R}_{x,y} = r\left(x,y\right) \exp\left\{Z_{x,y}\right\} \exp\left\{-\frac{1}{2}\tau_{x,y}^2\right\}$$

defines an exact algorithm. Since $R_{x,y}$ satisfies (5.11) it can be expressed as

$$R_{x,y} = r\left(x,y\right) V_{x,y} = r\left(x,y\right) \exp\left\{Z_{x,y}\right\},$$

which completes the proof. □

In order to analyse in more depth the properties of the penalty chain, let $P : \mathcal{X} \times \mathcal{B}\left(\mathcal{X}\right) \to [0,1]$ and $\bar{P}_{pen} : \mathcal{X} \times \mathcal{B}\left(\mathcal{X}\right) \to [0,1]$ denote the kernels associated to the marginal chain and the penalty method, respectively. The corresponding expressions are given by

$$P\left(x,dy\right) := q\left(x,dy\right)\alpha\left(x,y\right) + \delta_x\left(dy\right)\left[1 - \int_{\mathcal{X}} q\left(x,dy\right)\alpha\left(x,y\right)\right]$$

$$\text{and} \quad \bar{P}_{pen}\left(x,dy\right) := q\left(x,dy\right)\bar{\alpha}_{pen}\left(x,y\right) + \delta_x\left(dy\right)\left[1 - \int_{\mathcal{X}} q\left(x,dy\right)\bar{\alpha}_{pen}\left(x,y\right)\right],$$

where

$$\alpha(x,y) := \min\{1, r(x,y)\}$$
$$\text{and} \quad \bar{\alpha}_{pen}(x,y) := \mathbb{E}\left[\min\left\{1, \bar{R}_{pen}(x,y)\right\}\right].$$

Remark 5.1. Notice that

$$\begin{aligned}\bar{\alpha}_{pen}(x,y) &= \mathbb{E}\left[\min\left\{1, e^{\log\left(\bar{R}_{pen}(x,y)\right)}\right\}\right] \\ &= \Phi\left(\frac{\log(r(x,y)) - \frac{1}{2}\tau_{x,y}^2}{\tau_{x,y}}\right) + r(x,y)\,\Phi^c\left(\frac{\log(r(x,y)) + \frac{1}{2}\tau_{x,y}^2}{\tau_{x,y}}\right),\end{aligned} \tag{5.12}$$

where Φ denotes the cumulative distribution function of the standard Normal distribution and $\Phi^c := 1 - \Phi$.

We now look at a relationship between the penalty acceptance $\bar{\alpha}_{pen}$ and the marginal acceptance α. The acceptance probability of the penalty method can be upper and lower bounded in terms of the marginal acceptance.

Lemma 5.1. *The acceptance probabilities $\bar{\alpha}_{pen}$ and α satisfy*

$$2\Phi\left(-\frac{\tau_{x,y}}{2}\right)\alpha(x,y) \le \bar{\alpha}_{pen}(x,y) \le \alpha(x,y).$$

Proof. Let $Z_{x,y} \sim \mathcal{N}\left(0, \tau_{x,y}^2\right)$, then

$$\begin{aligned}\bar{\alpha}_{pen}(x,y) &= \mathbb{E}\left[\min\left\{1, r(x,y)\exp\left\{Z_{x,y} - \frac{1}{2}\tau_{x,y}^2\right\}\right\}\right] \\ &\ge \mathbb{E}\left[\min\left\{1, \exp\left\{Z_{x,y} - \frac{1}{2}\tau_{x,y}^2\right\}\right\}\min\{1, r(x,y)\}\right].\end{aligned}$$

Using Remark 5.1 with $r(x,y) \equiv 1$

$$\mathbb{E}\left[\min\left\{1, \exp\left\{Z_{x,y} - \frac{1}{2}\tau_{x,y}^2\right\}\right\}\right] = 2\Phi\left(-\frac{1}{2}\tau_{x,y}\right),$$

which implies the first inequality. The second part is proved by Jensen's inequality,

$$\bar{\alpha}_{pen}(x,y) \le \min\left\{1, r(x,y)\,\mathbb{E}\left[\exp\left\{Z_{x,y} - \frac{1}{2}\tau_{x,y}^2\right\}\right]\right\} = \alpha(x,y). \qquad \square$$

Using Peskun's off-diagonal ordering, the above result directly implies the penalty method is less efficient, in terms of asymptotic variance, than a MH chain with the same target and proposal. Additionally, one can show the penalty method

chain is geometrically ergodic whenever the marginal chain is geometric and the variance $\tau_{x,y}^2$ is uniformly bounded. This is shown in the following theorem using Dirichlet forms for reversible Markov chains (see Appendix A).

Theorem 5.2. *The chain generated by $\bar{P}_{pen}$ inherits geometric ergodicity from the marginal chain if*

$$\sup_{(x,y)\in\mathcal{X}^2} \tau_{x,y}^2 < \infty.$$

Proof. Using Lemma 5.1, the Dirichlet form of $\bar{P}_{pen}$ (see Appendix A) satisfies

$$\begin{aligned}
\mathcal{E}_{\bar{P}_{pen}}(g) &= \frac{1}{2}\int \pi(dx)\, q(x,dy)\, \bar{\alpha}_{pen}(x,y)\, [g(x)-g(y)]^2 \\
&\geq \int \pi(dx)\, q(x,dy)\, \alpha(x,y)\, \Phi\left(-\frac{1}{2}\tau_{x,y}\right) [g(x)-g(y)]^2 \\
&\geq 2 \inf_{x,y} \Phi\left(-\frac{1}{2}\tau_{x,y}\right) \mathcal{E}_P(g),
\end{aligned}$$

and

$$\mathcal{E}_{\bar{P}_{pen}}(g) \leq \frac{1}{2}\int \pi(dx)\, q(x,dy)\, \alpha(x,y)\, [g(x)-g(y)]^2 = \mathcal{E}_P(g).$$

Therefore,

$$\mathrm{Gap}_R\left(\bar{P}_{pen}\right) = \inf_{g:\pi(g)=0,\pi(g^2)=1} \mathcal{E}_{\bar{P}_{pen}}(g) \geq 2 \inf_{x,y} \Phi\left(-\frac{1}{2}\tau_{x,y}\right) \mathrm{Gap}_R(P)$$

and

$$\mathrm{Gap}_L\left(\bar{P}_{pen}\right) = \inf_{g:\pi(g)=0,\pi(g^2)=1} \left(2 - \mathcal{E}_{\bar{P}_{pen}}(g)\right) \geq \mathrm{Gap}_L(P).$$

The result follows by the existence of an absolute spectral gap for P (Theorem A.1), and the fact that

$$\inf_{x,y} \Phi\left(-\frac{1}{2}\tau_{x,y}\right) > 0$$

whenever $\sup_{x,y} \tau_{x,y}^2 < \infty$. □

In practice, log-normal estimates for the acceptance ratio as in (5.11) are not always available. Moreover, an expression for the variance of such estimates

$\tau^2_{x,y}$ may be unknown, making the implementation of the penalty method difficult. An approximate version of this method has been presented in Nicholls et al. (2012, Algorithm 5), which relies on an estimate for $\tau^2_{x,y}$ but still requires log-normality. In the following section we explore a similar approximate algorithm using geometric averages of estimators, where the log-normal assumption is not essential and in some cases can produce good results, as seen in the examples of Section 5.3.

5.2 Correcting the bias with geometric averages

The penalty method described in the previous section is applicable to the noisy MH algorithm, with ratio $R^{noi}_{x,y}$ given by (5.2), whenever $\log(W_x) \sim \mathcal{N}\left(-\frac{1}{2}\sigma^2_x, \sigma^2_x\right)$ and if σ^2_x is available as a function of $x \in \mathcal{X}$. For such case, assuming W_x and W_y are independent,

$$\begin{aligned}\log\left(R^{noi}_{x,y}\right) &= \log(r(x,y)) + \log(W_y) - \log(W_x) \\ &\sim \mathcal{N}\left(\log(r(x,y)) - \frac{1}{2}\left(\sigma^2_y - \sigma^2_x\right), \sigma^2_y + \sigma^2_x\right).\end{aligned}$$

However, neither the independence assumption between W_x and W_y nor the unbiasedness condition $\mathbb{E}[W_x] \equiv 1$ are essential for obtaining an exact method. In this section we propose an approach that attempts mimicking the penalty method; this is based on geometric averages of estimators $\left\{R^{(i)}_{x,y}\right\}$ satisfying (5.4) and the introduction of an approximate penalty term.

Consider again the estimator $R_{x,y}$ from (5.4) and define for all $(x,y) \in \mathcal{X}^2$

$$m_{x,y} := \mathbb{E}[\log(V_{x,y})] \qquad \text{and} \qquad \tau^2_{x,y} := \mathbb{V}[\log(V_{x,y})], \tag{5.13}$$

where $m_{x,y}$ and $\tau^2_{x,y}$ satisfy

$$m_{x,y} = -m_{y,x} \quad \text{and} \quad \tau^2_{x,y} = \tau^2_{y,x}.$$

From Proposition 5.2, and provided $\tau^2_{x,y} < \infty$, we can correct $R_{x,y}$ using $m_{y,x}$ as stated in the following corollary.

Assumption 1. $\tau^2_{x,y} = \mathbb{V}[\log(V_{x,y})] < \infty$ *for any* $(x,y) \in \mathcal{X}^2$.

Corollary 5.1. *Consider the randomised acceptance ratio* $R_{x,y}$ *in* (5.4). *Under Assumption 1 and if* $R_{x,y}$ *is log-normally distributed for any* $(x,y) \in \mathcal{X}^2$, *the acceptance*

ratio

$$\bar{R}_{x,y} := R_{x,y} \exp\left\{ m_{y,x} - \frac{1}{2}\tau_{x,y}^2 \right\}$$

defines an exact algorithm.

Remark 5.2. The correction given by $\tilde{R}_{x,y}$ in (5.7) is identical to $\bar{R}_{x,y}$ above when the variables $\{W_x\}$ are log-normally distributed.

Notice that if $\tau_{x,y}^2$ is large, the probability of accepting a move may be small even for large values of the acceptance ratio $r(x,y)$. Thus, although exact, the algorithm might mix poorly requiring an infeasible number of iterations for being of any use. To alleviate this, we consider the geometric average of M independent log-normal estimators $\left\{R_{x,y}^{(i)}\right\}_{i=1}^{M}$ as follows

$$\begin{aligned}\bar{R}_M(x,y) &:= \left(\prod_{i=1}^{M} R_{x,y}^{(i)}\right)^{1/M} \exp\left\{ m_{y,x} - \frac{1}{2M}\tau_{x,y}^2 \right\} \\ &= r(x,y) \exp\left\{ \frac{\tau_{x,y}}{\sqrt{M}}\xi - \frac{1}{2M}\tau_{x,y}^2 \right\}, \end{aligned} \tag{5.14}$$

where $\xi \sim \mathcal{N}(0,1)$.

Corollary 5.2. *Consider M independent random variables $\left\{R_{x,y}^{(i)}\right\}_{i=1}^{M}$, where each $R_{x,y}^{(i)}$ satisfies* (5.4) *and is log-normally distributed for any $(x,y) \in \mathcal{X}^2$. Under Assumption 1, the acceptance ratio $\bar{R}_M$ defines an exact algorithm for any $M \in \mathbb{N}$.*

As commented before, the estimator $R_{x,y}$ may not be log-normally distributed and even in such case the expressions for $m_{x,y}$ and $\tau_{x,y}^2$ may be unknown. Despite this, the log-normality assumption can be relaxed with a large enough sample $\left\{R_{x,y}^{(i)}\right\}_i$ and relying on the Central Limit Theorem (CLT). Additionally, in some cases $m_{x,y}$ and $\tau_{x,y}^2$ can be approximated using the variables $\left\{R_{x,y}^{(i)}\right\}_i$, we will see examples of this later on. Let $\hat{m}_{x,y,M}$ and $\hat{\tau}_{x,y,M}^2$ be such approximations; one could then consider the following approximate acceptance ratio

$$\tilde{R}_M(x,y) := \left(\prod_{i=1}^{M} R_{x,y}^{(i)}\right)^{1/M} \exp\left\{ \hat{m}_{y,x,M} - \frac{1}{2M}\hat{\tau}_{x,y,M}^2 \right\}, \tag{5.15}$$

which is implemented in Algorithm 5.1 and is a variant of Nicholls et al. (2012, Algorithm 5).

Algorithm 5.1 Approximate Penalty Method (APM)

Assumptions: Suitable approximations for $m_{x,y}$ and $\tau_{x,y}^2$ in (5.13), say $\hat{m}_{x,y,M}$ and $\hat{\tau}_{x,y,M}^2$ respectively, are available.
Given the current state $X_i = x$, simulate X_{i+1} as follows:

1. Sample $Y = y \sim q(x, \cdot)$.
2. Draw $\left\{R_{x,y}^{(i)}\right\}_{i=1}^{M}$ satisfying (5.4) and Assumption 1.
3. Compute $\hat{m}_{y,x,M}$ and $\hat{\tau}_{x,y,M}^2$.
4. With probability $\min\left\{1, \tilde{R}_M(x,y)\right\}$, where $\tilde{R}_M$ is given by (5.15):
   set $X_{i+1} = Y$;
   otherwise:
   set $X_{i+1} = x$.

Since $\mathbb{V}[\log(V_{x,y})] = \mathbb{V}[\log(R_{x,y})]$, a natural unbiased and consistent estimator for $\tau_{x,y}^2$ is obtained via the sample variance

$$\hat{\tau}_{x,y,M}^2 := \frac{1}{M-1}\sum_{i=1}^{M}\left(L_{x,y}^{(i)} - \bar{L}_{x,y,M}\right)^2, \tag{5.16}$$

where $L_{x,y}^{(i)} := \log\left(R_{x,y}^{(i)}\right)$ for each $i \in \{1, \ldots, M\}$ and

$$\bar{L}_{x,y,M} := \frac{1}{M}\sum_{j=1}^{M} L_{x,y}^{(j)}.$$

In contrast, estimating $m_{x,y}$ is not always as straightforward as estimating $\tau_{x,y}^2$. In the following section some examples for which $m_{x,y}$ is known or suitably approximated are explored. For our purposes, suitably approximated implies that the mean absolute error of $\hat{m}_{x,y,M}$ decays polynomially as a function of M^{-1} and the distance $\|y - x\|$, as stated below.

Assumption 2. *The estimator $\hat{m}_{x,y,M}$ satisfies for some $k \in \mathbb{N}$*

$$\mathbb{E}\left[|\hat{m}_{x,y,M} - m_{x,y}|\right] \lesssim \mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, \|y - x\|\right),$$

where $\mathscr{P}_k^{\beta,\gamma}(a,b) = \sum_{j=1}^{k} a^{\beta_j} b^{\gamma_j}$, $\beta_j, \gamma_j \in \mathbb{R}_0^+$ with $\beta_j + \gamma_j > 0$ for all $j \in \{1, \ldots, k\}$,

and

$$\lim_{a,b\to 0} b^{-2}\mathscr{P}_k^{\beta,\gamma}(a,b) = 0.$$

The above assumption will be important for addressing some aspects of the theoretical results presented in Section 5.4. In the following section, we explore some examples for which Assumption 2 can be verified.

5.3 Examples

We consider four examples for which the Approximate Penalty Method (APM) in Algorithm 5.1 is implemented. The first two, presented in Sections 5.3.1 and 5.3.2, deal with special cases of noisy MH relying on estimates of the density of π, as in (5.1). For the first one we will see that the log-normal assumption is not essential for still obtaining accurate results. The second one assumes log-normality, but provides a way for estimating $m_{x,y}$ for the implementation of the APM algorithm.

The last two examples, appearing in Section 5.3.3, deal with intractability arising from large data sets, where sub-sampling approaches are considered. This idea is not new and has been considered in the past, see e.g. Bardenet et al. (2015) for a good overview. We only focus in two possible expressions for $R_{x,y}$ under this scenario, briefly discussing the trade-off arising from the choice of the sub-sample size and the value of M (the number of estimators involved).

5.3.1 Homogeneous noise

When the noise of the estimates $\hat{\pi}_x$ in (5.1) is homogeneous with respect to the state space, i.e. when $W_x \equiv W \sim Q(\cdot)$ for all $x \in \mathcal{X}$, then

$$m_{x,y} = \mathbb{E}\left[\log(W_y) - \log(W_x)\right] \equiv 0$$

for all $(x,y) \in \mathcal{X}^2$.

Example 5.1. Consider a Gaussian target π on $\mathbb{R}$ and let

$$Q(dw) \propto (1+w)^{-(1+\beta)}\mathbb{1}(w>0)\,\mu^{Leb}(dw),$$

where $\beta > 0$.

In Figure 5.1, we can see that the APM may perform much better than the corresponding pseudo-marginal. One interesting aspect of this example is that the

APM provides a good approximation of the target density for small values of M, even when the noise variables W are not log-normally distributed and have infinite variance. However, the homogeneous assumption on the weights $\{W_x\}$ is difficult to attain in practice.

5.3.2 Log-normal noise

If the distribution of W_x is concentrated around 1, the delta method may be a suitable approach for estimating $\mathbb{E}\left[\log\left(W_x\right)\right]$. Using a second order Taylor expansion,

$$\mathbb{E}\left[\log\left(W_x\right)\right] \approx -\frac{1}{2}\mathbb{E}\left[\left(1-W_x\right)^2\right] \approx -\frac{1}{2}\mathbb{V}\left[W_x\right].$$

Similarly, with a first order approximation $\mathbb{V}\left[\log\left(W_x\right)\right] \approx \mathbb{V}\left[W_x\right]$, implying

$$\mathbb{E}\left[\log\left(W_x\right)\right] \approx -\frac{1}{2}\mathbb{V}\left[\log\left(W_x\right)\right].$$

In fact, when W_x has expectation 1 and is log-normally distributed the above expression becomes an equality.

Define $L_x := \log\left(\hat{\pi}_x\right)$ and let $s^2_{x,M}$ denote the sample variance of $\left\{L_x^{(i)}\right\}_{i=1}^M$, which follows a similar expression to (5.16). Although $-\frac{1}{2}\left(s^2_{y,M} - s^2_{x,M}\right)$ is an unbiased and consistent estimator of $m_{x,y} = -\frac{1}{2}\left(\sigma_y^2 - \sigma_x^2\right)$, its variance may be large affecting drastically the ergodic properties of the resulting chain. Instead, a better approach would be to estimate the difference $\sigma_y^2 - \sigma_x^2$ directly. Using Taylor expansions,

$$\sigma_y^2 - \sigma_x^2 \approx (y-x)\cdot\nabla\sigma_x^2 \quad \text{and} \quad \sigma_x^2 - \sigma_y^2 \approx (x-y)\cdot\nabla\sigma_y^2,$$

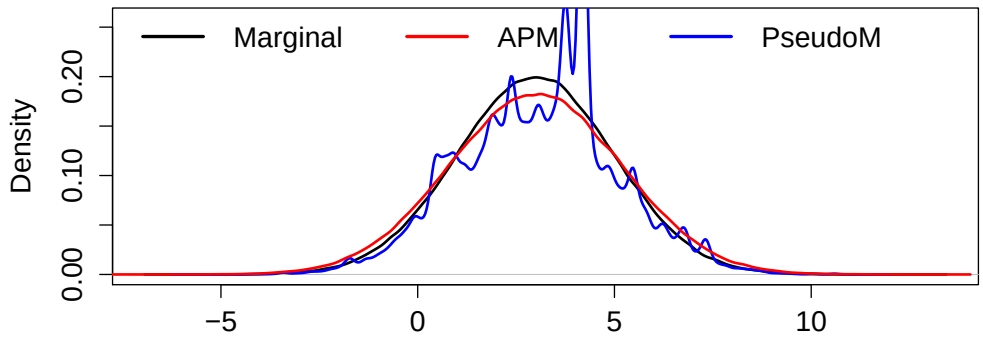
leading to

$$\sigma_y^2 - \sigma_x^2 \approx \frac{1}{2}(y-x)\cdot\left(\nabla\sigma_y^2 + \nabla\sigma_x^2\right).$$

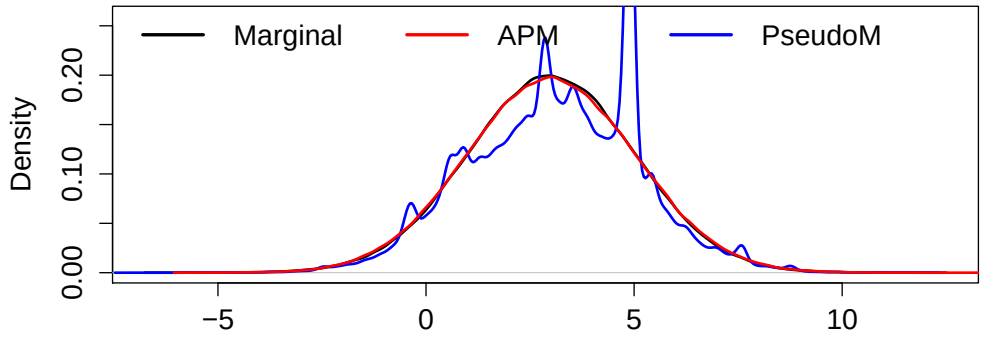
The aim is now to appropriately estimate the gradient of σ_x^2.

Proposition 5.3. *Let $L_x = \log\left(\hat{\pi}_x\right)$ be normally distributed and suppose an expression for $\nabla_x L_x$ is available. For a sample $\left\{L_x^{(i)}\right\}_{i=1}^M$ define*

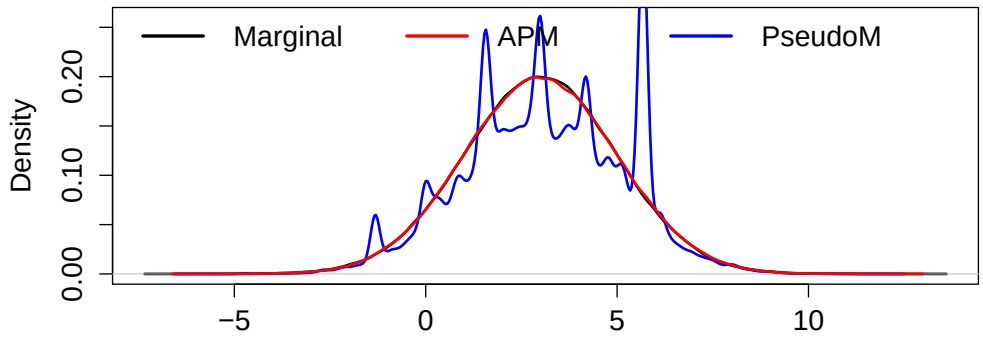
$$ds^2_{x,M} := \nabla_x s^2_{x,M} = \frac{2}{M-1}\sum_{i=1}^{M}\left[\left(L_x^{(i)} - \bar{L}_{x,M}\right)\nabla_x L_x^{(i)}\right],$$



(a) $M = 2$.



(b) $M = 5$.



(c) $M = 10$.

Figure 5.1: Estimated densities of Example 5.1 where $\beta = 1.1$, using 1 million iterations for the marginal (black), APM (red) and pseudo-marginal (blue) chains.

where

$$\bar{L}_{x,M} := \frac{1}{M}\sum_{j=1}^{M} L_x^{(j)}.$$

Then

$$ds^2_{x,M} \sim \frac{\chi^2_{(M-1)}}{M-1}\nabla\sigma_x^2,$$

where χ^2_d denotes a Chi-square distribution with d degrees of freedom.

Proof. Due to normality, L_x can be expressed as

$$L_x = \log\left(\pi\left(x\right)\right) - \frac{1}{2}\sigma_x^2 + \sigma_x\xi,$$

where $\xi \sim \mathcal{N}\left(0,1\right)$. Then,

$$\begin{aligned} ds^2_{x,M} &= \frac{2\sigma_x}{M-1}\sum_{i=1}^{M}\left[\left(\xi^{(i)} - \bar{\xi}\right)\left(\nabla\log\left(\pi\left(x\right)\right) - \frac{1}{2}\nabla\sigma_x^2 + \xi^{(i)}\nabla\sigma_x\right)\right] \\ &= \frac{2\sigma_x\nabla\sigma_x}{M-1}\sum_{i=1}^{M}\left[\left(\xi^{(i)} - \bar{\xi}\right)\xi^{(i)}\right] = \frac{\nabla\sigma_x^2}{M-1}\left[\sum_{i=1}^{M}\left(\xi^{(i)}\right)^2 - M\bar{\xi}^2\right] \\ &\sim \frac{\nabla\sigma_x^2}{M-1}\chi^2_{(M-1)}. \end{aligned}$$

□

Remark 5.3. In order for this approach to be useful, one relies on the availability of $\nabla_x L_x$. This is the case when $\hat{\pi}_x$ is obtained via importance sampling, see e.g. Andrieu and Roberts (2009).

One can then consider the following estimate

$$\hat{m}_{x,y,M} = -\frac{1}{4}\left(y - x\right)\cdot\left(ds^2_{y,M} + ds^2_{x,M}\right),$$

which satisfies Assumption 2, provided $\nabla\sigma_x^2$ is bounded and the Hessian $\nabla^2\sigma_x^2$ is

(component-wise) Lipschitz, since

$$\begin{aligned}
\mathbb{E}\left[|\hat{m}_{x,y,M} - m_{x,y}|\right] &\leq \mathbb{E}\left[|\hat{m}_{x,y,M} - \mathbb{E}\hat{m}_{x,y,M}|\right] + |[\mathbb{E}\hat{m}_{x,y,M}] - m_{x,y}| \\
&\leq \sqrt{\mathbb{V}\left[\hat{m}_{x,y,M}\right]} + \frac{1}{2}\left|\left(\sigma_y^2 - \sigma_x^2\right) - \frac{1}{2}(y-x)\cdot\left(\nabla\sigma_y^2 + \nabla\sigma_x^2\right)\right| \\
&\leq \sqrt{\frac{\|y-x\|^2}{8(M-1)}\left(\left\|\nabla\sigma_y^2\right\|^2 + \left\|\nabla\sigma_x^2\right\|^2\right) + \frac{\|y-x\|^2}{8}\left\|\nabla^2\sigma_{\rho_1}^2 - \nabla^2\sigma_{\rho_2}^2\right\|} \\
&\lesssim \frac{\|y-x\|}{\sqrt{M}} + \|y-x\|^2\|\rho_1 - \rho_2\| \lesssim \frac{\|y-x\|}{\sqrt{M}} + \|y-x\|^3,
\end{aligned}$$

where ρ_1 and ρ_2 lie on the segment joining x and y.

Example 5.2. Consider a Gaussian target π on $\mathbb{R}$ and let $\log(W_x) \sim \mathcal{N}\left(-\frac{1}{2}\sigma_x^2, \sigma_x^2\right)$ with

$$\sigma_x^2 = \sigma^2\left(1 + a\sin(x)\right),$$

where $|a| < 1$. In this case we want to control the distance between the proposed move y and the current state x. For doing this, we use a Gaussian random-walk proposal q with variance δ^2.

Table 5.1 presents estimated effective sample sizes (ESS) expressed as percentage, for different values of δ and M. The values coloured in red correspond to the APM, whereas the values for the pseudo-marginal appear on the extreme left column in blue. The estimated ESS for the marginal chain is presented on the top-left cell of the table. Clearly, an increase of M results in an increase of the estimated ESS, whereas decreasing the proposal standard deviation δ has the opposite effect.

12.66	$M\backslash\delta$	$\delta_0 = 2$	$\delta_0/2$	$\delta_0/2^2$	$\delta_0/2^3$	$\delta_0/2^4$
0.13	$M_0 = 3$	2.63	0.88	0.25	0.08	0.04
0.15	$2M_0$	4.30	1.57	0.46	0.13	0.05
0.22	2^2M_0	5.79	2.29	0.74	0.19	0.07
0.38	2^3M_0	6.59	2.98	0.87	0.24	0.08
0.43	2^4M_0	7.64	3.41	1.06	0.30	0.09

Table 5.1: Estimated effective sample size (ESS) of Example 5.2 (expressed as percentage) for the marginal (black), APM (red) and pseudo-marginal (blue) chains. The proposal is $\delta_0 = 2$ for the marginal and pseudo-marginal chains.

However, there is a trade-off present when choosing δ and M, which is noticeable from looking at Tables 5.2 and 5.3. These contain estimated values of the mean squared errors (MSE) of the expected value and the 90th percentile, respec-

tively, for the target distribution π. We observe how the APM may provide good estimates in this setting for some values of δ and M. For instance, possible good choices for δ are 0.5 or 0.25, together when M is 12 or 24 (italic bold numbers on the tables). Here, the trade-off between the number of iterations and the choice of δ is important. Taking smaller values of δ can improve the estimates but with the cost of increasing the length of the chain.

0.01	$M \backslash \delta$	$\delta_0 = 2$	$\delta_0/2$	$\delta_0/2^2$	$\delta_0/2^3$	$\delta_0/2^4$
0.70	$M_0 = 3$	26.64	25.92	16.50	14.85	7.31
1.88	$2M_0$	24.36	26.09	7.43	3.45	0.88
3.26	$2^2 M_0$	16.84	23.62	***2.15***	***0.97***	3.23
0.25	$2^3 M_0$	14.16	24.25	***1.82***	***0.30***	2.40
0.38	$2^4 M_0$	10.47	23.11	2.60	0.86	2.84

Table 5.2: Estimated mean squared error (MSE) of Example 5.2 (expressed in thousands) of the expected value of π for the marginal (black), APM (red) and pseudo-marginal (blue) chains. The proposal is $\delta_0 = 2$ for the marginal and pseudo-marginal chains.

0.01	$M \backslash \delta$	$\delta_0 = 2$	$\delta_0/2$	$\delta_0/2^2$	$\delta_0/2^3$	$\delta_0/2^4$
9.36	$M_0 = 3$	58.38	44.80	59.13	67.95	62.60
0.64	$2M_0$	28.66	8.64	5.31	2.75	4.03
0.58	$2^2 M_0$	21.28	3.73	***0.20***	***2.48***	5.40
0.37	$2^3 M_0$	17.12	2.23	***0.17***	***0.44***	1.43
0.57	$2^4 M_0$	13.12	1.52	0.17	0.45	4.29

Table 5.3: Estimated mean squared error (MSE) of Example 5.2 (expressed in thousands) of the 90th percentile of π for the marginal (black), APM (red) and pseudo-marginal (blue) chains. The proposal is $\delta_0 = 2$ for the marginal and pseudo-marginal chains.

Finally in Figure 5.2, estimated densities for some values of δ and M are shown. Specifically, we fix $M = 12$ and consider

$$\delta \in \{\delta_0 = 2, \quad \delta_2 = \delta_0/2 = 0.5, \quad \delta_4 = \delta_0/4 = 0.125\}\,.$$

The lines of the estimated densities contain numbers which correspond to the value of δ used in the simulation. Notice that the chains for δ_2 and δ_4 provide more accurate estimates than the pseudo-marginal chain, the latter shown in blue colour.

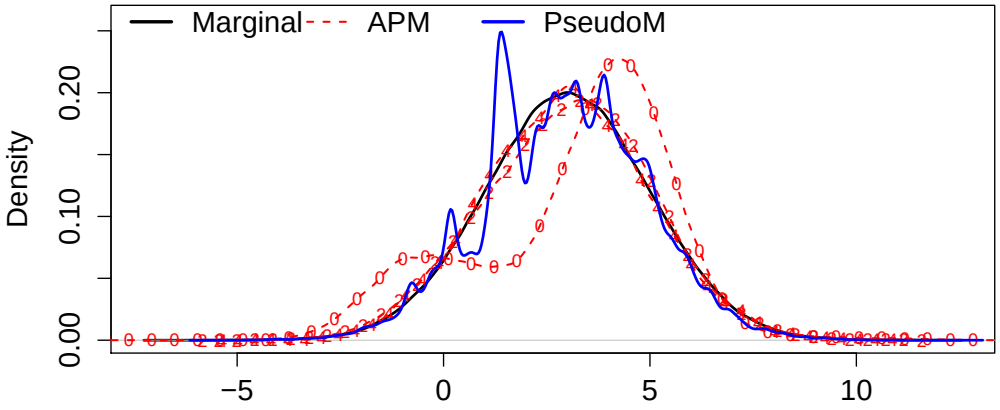


Figure 5.2: Estimated densities of Example 5.2 where $M = 12$, using 1 million iterations for the marginal (black), APM (red) and pseudo-marginal (blue) chains. Numbers indicate the subscript in δ_k, where $\delta_0 = 2$, $\delta_2 = \delta_0/2^2$ and $\delta_4 = \delta_0/2^4$.

5.3.3 Sub-sampling

A different type of intractability arises when dealing with large data sets, where sub-sampling approaches may prove useful. Suppose we have i.i.d. observations $\boldsymbol{D} = \{D_i\}_{i=1}^n$ coming from $f(\cdot\,|x)$; in a Bayesian framework the aim is to sample from the posterior

$$\pi(x) = \pi(x\,|\boldsymbol{D}) \propto p_0(x) \exp\{l(x\,|\boldsymbol{D})\},$$

where p_0 is the prior for the parameter x, and l is the log-likelihood function given by

$$l(x\,|\boldsymbol{D}) := \sum_{i=1}^{n} l_i(x) := \sum_{i=1}^{n} \log(f(D_i\,|x)).$$

Evaluating the above posterior may be difficult when n is very large; nevertheless, a possible (approximate) solution is to take a random sub-sample $\{D_{i_j}\}_{j=1}^k$ at each iteration of the MH algorithm. This idea is not new and has been studied extensively in the past, see e.g. Korattikara et al. (2014), Quiroz et al. (2014) or Bardenet et al. (2015) for a more complete review. For this case and considering a symmetric proposal, the Metropolis-Hastings ratio can be estimated with the following naive

estimator

$$R_{x,y}^{(I)} = \frac{p_0(y)}{p_0(x)} \exp\left\{ \frac{n}{k} \sum_{j=1}^{k} \left[l_{i_j}(y) - l_{i_j}(x) \right] \right\},$$

which is a log-unbiased estimator of acceptance ratio $r(x,y)$ since

$$\begin{aligned}
\mathbb{E}\left[\log\left(R_{x,y}^{(I)} \right) \right] &= \log\left(\frac{p_0(y)}{p_0(x)} \right) + \frac{n}{k} \mathbb{E}\left[\sum_{j=1}^{k} \left(l_{i_j}(y) - l_{i_j}(x) \right) \right] \\
&= \log\left(\frac{p_0(y)}{p_0(x)} \right) + \frac{n}{k} \sum_{i=1}^{n} \left[l_i(y) - l_i(x) \right] \mathbb{E}\left[\mathbb{1}\left(i \in \{i_1, \ldots, i_k\} \right) \right] \\
&= \log\left(r(x,y) \right).
\end{aligned}$$

Therefore, for this case $m_{x,y} \equiv 0$ and the implementation of the APM algorithm is straightforward. We now present a very simple Gaussian example.

Remark 5.4. Throughout this section we only consider sub-sampling without replacement. However, for implementational purposes, one may prefer using sub-samples with replacement.

Example 5.3. Consider observations $\{D_i\}_{i=1}^{n}$, where $D_i \overset{i.i.d}{\sim} \mathcal{N}\left(\mu, \sigma_0^2\right)$ with σ_0^2 known and take a uniform prior for the parameter μ. The Metropolis ratio given by

$$r(\mu, \mu^*) = \exp\left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} \left[(D_i - \mu^*)^2 - (D_i - \mu)^2 \right] \right\},$$

is estimated using a sub-sample $\left\{D_{i_j}\right\}_{j=1}^{k}$ via

$$R_{\mu,\mu^*}^{(I)} = \exp\left\{ -\frac{n}{2k\sigma_0^2} \sum_{j=1}^{k} \left[\left(D_{i_j} - \mu^*\right)^2 - \left(D_{i_j} - \mu\right)^2 \right] \right\}.$$

Provided n is large and considering $\|\mu - \mu^*\| = O\left(n^{-1/2}\right)$, which allows control over the marginal acceptance, the variance $\tau_{\mu,\mu^*}^2 = \mathbb{V}\left[\log\left(R_{\mu,\mu^*}^{(I)} \right) \right]$ satisfies for large n

$$\tau_{\mu,\mu^*}^2 \approx \|\mu - \mu^*\|^2 \, n \frac{(n-k)}{k} = O\left(\frac{n-k}{k} \right).$$

This suggests taking $kM \propto n$ for controlling the acceptance rate of the penalty method as n increases. Figure 5.3 shows the estimated densities when the product kM is a fixed proportion of n. Notice there that the accuracy of the estimation will

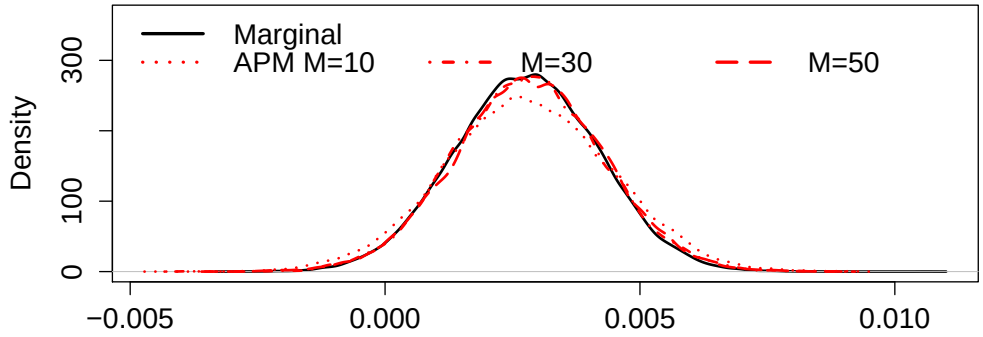


Figure 5.3: Estimated densities of Example 5.3 using $n = 10^6$ observations for the marginal (black) and APM (red) chains. The values of k and M are such that $kM = n/10$.

depend on how M and k are individually chosen.

It is possible to reduce the variability of $R^{(I)}$ using a control variate approach, as done e.g. in Bardenet et al. (2015, Section 7.2), but sacrificing the log-unbiased condition. By a Taylor expansion argument

$$l\left(y\,|\boldsymbol{D}\right) - l\left(x\,|\boldsymbol{D}\right) \approx \left(y - x\right) \cdot \frac{\nabla l\left(y\,|\boldsymbol{D}\right) + \nabla l\left(x\,|\boldsymbol{D}\right)}{2},$$

and using a sub-sample $\left\{D_{i_j}\right\}_{j=1}^{k}$, the gradient terms $\nabla l\left(\cdot\,|\boldsymbol{D}\right)$ are estimated unbiasedly using

$$\nabla l\left(\hat{x}\,|\boldsymbol{D}\right) + \frac{n}{k}\sum_{j=1}^{k}\left[\nabla l_{i_j}\left(\cdot\right) - \nabla l_{i_j}\left(\hat{x}\right)\right],$$

where $\hat{x}$ is some reference point, e.g. the maximum likelihood estimator (MLE). Hence, the acceptance ratio $r\left(x, y\right)$ can be estimated using

$$R_{x,y}^{(II)} = \frac{p_0\left(y\right)}{p_0\left(x\right)} \times \exp\left\{\left(y - x\right) \cdot \left(\nabla l\left(\hat{x}\,|\boldsymbol{D}\right) + \frac{n}{k}\sum_{j=1}^{k}\left[\frac{\nabla l_{i_j}\left(y\right) + \nabla l_{i_j}\left(x\right)}{2} - \nabla l_{i_j}\left(\hat{x}\right)\right]\right)\right\}.$$

For this case and assuming the Hessian $\nabla^2 l(x\,|\boldsymbol{D})$ is (component-wise) Lipschitz,

$$\begin{aligned}
|m_{x,y}| &= \left|\mathbb{E}\left[\log\left(R^{(II)}_{x,y}\right)\right] - \log\left(r\left(x,y\right)\right)\right| \\
&= \left|(y-x)\cdot\frac{\nabla l\left(y\,|\boldsymbol{D}\right)+\nabla l\left(x\,|\boldsymbol{D}\right)}{2} - \left(l\left(y\,|\boldsymbol{D}\right)-l\left(x\,|\boldsymbol{D}\right)\right)\right| \\
&= \frac{1}{4}\left|(y-x)^T\left[\nabla^2 l\left(\rho_1\,|\boldsymbol{D}\right)-\nabla^2 l\left(\rho_2\,|\boldsymbol{D}\right)\right](y-x)\right| \\
&\lesssim \|y-x\|^2\left\|\nabla^2 l\left(\rho_1\,|\boldsymbol{D}\right)-\nabla^2 l\left(\rho_2\,|\boldsymbol{D}\right)\right\| \lesssim \|y-x\|^3,
\end{aligned}$$

where ρ_1 and ρ_2 lie on the segment joining x and y. Therefore, Assumption 2 is immediately satisfied by considering $\hat{m}_{x,y,M}\equiv 0$.

Notice that for Example 5.3, $R^{(II)}_{\mu,\mu^*}$ is in fact not random. This will not be true in general, however $R^{(II)}$ can perform much better than $R^{(I)}$ in more complex scenarios, provided a reference point $\hat{x}$ is available. A logistic regression example is now presented.

Example 5.4. Consider a set of observations $\{x_i, y_i\}_{i=1}^n$, where $x_i\in\mathbb{R}$, $y_i\in\{0,1\}$ and for $(\beta_0,\beta_1)\in\mathbb{R}^2$

$$\theta_i\left(\boldsymbol{\beta}\right) := \mathbb{P}\left[Y_i=1|\,x_i,\beta_0,\beta_1\right] = \frac{1}{1+\exp\left\{-\left(\beta_0+\beta_1 x_i\right)\right\}}.$$

Using uniform prior distributions for β_0 and β_1, the usual acceptance ratio is estimated with a sub-sample $\left\{x_{i_j}, y_{i_j}\right\}_{j=1}^k$ by

$$\begin{aligned}
&R^{(II)}_{\boldsymbol{\beta},\boldsymbol{\beta}^*} \\
&\quad= \exp\left\{(\boldsymbol{\beta}^*-\boldsymbol{\beta})\cdot\left(\nabla l\left(\hat{\boldsymbol{\beta}}\,|\boldsymbol{y},\boldsymbol{x}\right)+\frac{n}{k}\sum_{j=1}^{k}\left[\frac{\nabla l_{i_j}\left(\boldsymbol{\beta}^*\right)+\nabla l_{i_j}\left(\boldsymbol{\beta}\right)}{2}-\nabla l_{i_j}\left(\hat{\boldsymbol{\beta}}\right)\right]\right)\right\},
\end{aligned}$$

where

$$l\left(\cdot\,|\boldsymbol{y},\boldsymbol{x}\right)=\sum_{i=1}^{n} l_i\left(\cdot\right)=\sum_{i=1}^{n}\left[y_i\log\left(\theta_i\left(\cdot\right)\right)+\left(1-y_i\right)\log\left(1-\theta_i\left(\cdot\right)\right)\right].$$

Figure 5.4 shows the estimated densities for the parameter β_1 and for different values of k and M using using $R^{(II)}_{\boldsymbol{\beta},\boldsymbol{\beta}^*}$. Even though $kM\ll n$, the method can perform well with a much lower computational cost than a standard MH algorithm, the latter requiring $n=10^6$ evaluations at every iteration. In this example, the reference point $\hat{x}$ considered is the MLE, which was computed numerically in advance but with the advantage of performing this step only once. Once more, notice from the plots the

existing trade-off between k and M, where possibly larger values of M should always be preferred in order for the CLT to kick in. More analysis on this trade-off is clearly required.

5.4 Theoretical results

This section contains some theoretical results under fairly strong conditions on the target π and the noise $V_{x,y}$ from the estimator $R_{x,y}$ in 5.4, contrasting with the empirical findings of the examples in the previous section. We investigate the behaviour and ergodic properties of the APM chain as the proposal variance $\delta^2 \to 0$ and the number of estimators $M \to \infty$.

For our analysis, let $\bar{P}_M : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ and $\tilde{P}_M : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ be the kernels associated to the chains generated by the penalty method and the APM, with acceptance ratios $\bar{R}_M$ and $\tilde{R}_M$ given by (5.15) and (5.14), respectively. These are

$$\begin{aligned} \bar{P}_M(x, dy) &:= q(x, dy)\,\bar{\alpha}_M(x, y) + \delta_x(dy)\left[1 - \int_{\mathcal{X}} q(x, dy)\,\bar{\alpha}_M(x, y)\right] \\ \text{and} \quad \tilde{P}_M(x, dy) &:= q(x, dy)\,\tilde{\alpha}_M(x, y) + \delta_x(dy)\left[1 - \int_{\mathcal{X}} q(x, dy)\,\tilde{\alpha}_M(x, y)\right], \end{aligned}$$

where

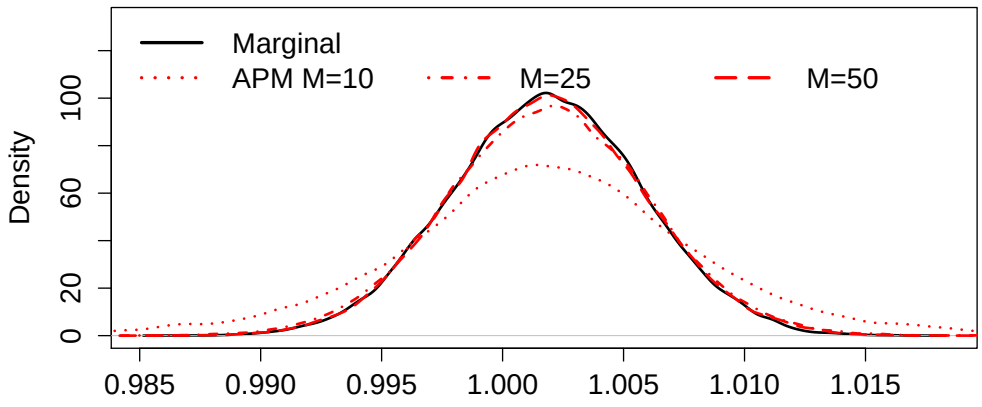
$$\begin{aligned} \bar{\alpha}_M(x, y) &:= \mathbb{E}\left[\min\left\{1, \bar{R}_M(x, y)\right\}\right] \\ \text{and} \quad \tilde{\alpha}_M(x, y) &:= \mathbb{E}\left[\min\left\{1, \tilde{R}_M(x, y)\right\}\right]. \end{aligned}$$

Assumption 2 involves the distance between the proposed moved y and the current state x. Therefore, random-walk proposals are a natural choice for controlling the estimator $\hat{m}_{x,y,M}$. Consider a family of random-walk proposals $\{q_n(x, \cdot)\}_{(x,n)\in\mathcal{X}\times\mathbb{N}}$, and for simplicity assume they are Gaussian. Here, $Y_n \sim q_n(x, \cdot)$ implies

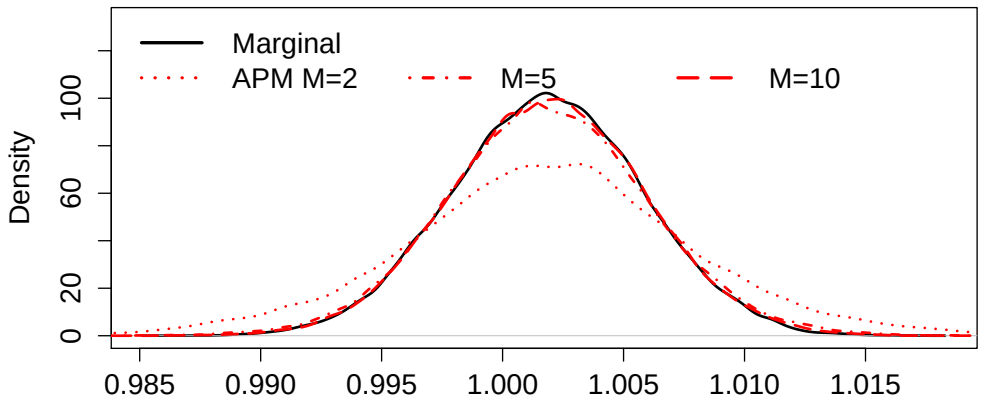
$$Y_n = x + \delta_n \xi, \tag{5.17}$$

where $\{\delta_n\}_n$ is a decreasing sequence of positive terms converging to 0 as $n \to \infty$ and ξ is an independent Gaussian random variable with mean $\mathbb{E}[\xi] = 0$ and variance $\mathbb{E}\left[\xi^2\right] = 1$.

The first main result of this section concerns with the weak convergence towards a diffusion process of the sped-up of chain generated by the penalty method



(a) k=1.



(b) k=5.

Figure 5.4: Estimated densities for β_1 in Example 5.4 using $n = 10^6$ observations for the marginal (black) and APM (red) chains. The different values for the product kM are comparable among the two plots.

kernel $\bar{P}_{M,n}$, for fixed $M \in \mathbb{N}$ and as $n \to \infty$. Such diffusion depends on the expression for $\tau_{x,x}$, but when $\tau_{x,x} \equiv \tau$ it reduces to the usual Langevin diffusion satisfying

$$dW(t) = \frac{1}{2}\gamma \nabla \log(\pi(W(t)))\, dt + \sqrt{\gamma} dB(t), \tag{5.18}$$

for some $\gamma > 0$ and where $B(t)$ denotes standard Brownian motion. The analysis of limiting diffusions for MCMC chains has been exploited before in the context of scaling limits as the dimension of the state space $\mathcal{X}$ increases. This was initially done in Roberts et al. (1997) which has led to several related publications, for example Roberts and Rosenthal (2001), Bédard and Rosenthal (2008), Mattingly et al. (2012) and Beskos et al. (2015) to name a few. In fact, the resulting diffusion in the following theorem is also discussed in Beskos et al. (2015), where the authors address the behaviour of the random-walk Metropolis algorithm when π concentrates on a sub-manifold. Here we tackle a much simpler problem, assuming the state space $\mathcal{X} = \mathbb{R}$ and some conditions on the target π and noise $V_{x,y}$ from the random ratio $R_{x,y}$ in (5.4). Recall from Section 5.2 the following three important variables

$$\begin{aligned} L_{x,y} &= \log(R_{x,y}) = \log(r(x,y)) + \log(V_{x,y}), \\ m_{x,y} = \mathbb{E}[\log(V_{x,y})] \qquad &\text{and} \qquad \tau^2_{x,y} = \mathbb{V}[\log(V_{x,y})] = \mathbb{V}[L_{x,y}]. \end{aligned}$$

Assumption 3. $\sup_{x,y} \mathbb{E}\left[(L_{x,y} - \mathbb{E}[L_{x,y}])^4\right] < \infty$ *and* $\inf_{x,y} \tau^2_{x,y} > 0$.

Assumption 4. *The function* $\tau^2_{x,y} : \mathcal{X}^2 \to \mathbb{R}^+$ *and target density* $\pi(x)$ *on* $\mathbb{R}$ *satisfy:*

1. $\tau^2_{x,y} \in C^2(\mathbb{R}^2)$ *with bounded 1st and 2nd order partial derivatives;*
2. $\log(\pi(x)) \in C^2(\mathbb{R})$ *with bounded 1st and 2nd derivatives.*

Theorem 5.3. *Set* $T > 0$, $M \in \mathbb{N}$ *and let* $\{\bar{X}_{M,n}(k)\}_{k \geq 0}$ *be the discrete-time Markov chain generated by* $\bar{P}_{M,n}$ *with proposal distribution* q_n. *If* $\bar{X}_{M,n}(0) \sim \mu(\cdot)$ *and under Assumptions 3 and 4, the continuous-time process*

$$\bar{\mathfrak{X}}_{M,n}(t) := \bar{X}_{M,n}\left(\left\lfloor t\delta_n^{-2} \right\rfloor\right)$$

converges weakly as $n \to \infty$, *in the Skorokhod topology on* $\mathbb{D}([0,T], \mathbb{R})$, *to the diffusion process* $\{W_M(t)\}_{t \in [0,T]}$ *satisfying*

$$\begin{cases} dW_M(t) = \frac{1}{2} b_M(W_M(t))\, dt + \sqrt{a_M(W_M(t))} dB(t) \\ W_M(0) \sim \mu(\cdot), \end{cases}$$

where

$$a_M(x) = 2\Phi\left(-\frac{\tau_{x,x}}{2\sqrt{M}}\right)$$
$$\text{and} \quad b_M(x) = \nabla a_M(x) + a_M(x)\nabla \log(\pi(x)).$$

See Appendix C for a brief review of the Skorokhod topology on $\mathbb{D}([0,T],\mathbb{R})$. Nevertheless, the accuracy of the APM relies on increasing M. The following result shows that the sped-up chain generated by the APM with kernel $\tilde{P}_{M,n}$ also converges to a diffusion limit, as both $M, n \to \infty$ in a suitable way. Looking back at the previous result and taking $M \to \infty$, a correct guess for such limit is the Langevin diffusion satisfying (5.18) with $\gamma = 1$.

Theorem 5.4. *Set $T > 0$ and let $\left\{\tilde{X}_n(k)\right\}_{k\geq 0}$ be the discrete-time Markov chain generated by $\tilde{P}_{M,n}$ with proposal q_n and where $M = M_n$ is such that $M_n\delta_n^4 \to \infty$ as $n \to \infty$. If $\tilde{X}_n(0) \sim \mu(\cdot)$ and under Assumptions 2, 3 and 4, the continuous-time process*

$$\tilde{\mathfrak{X}}_n(t) := \tilde{X}_n\left(\left\lfloor t\delta_n^{-2}\right\rfloor\right)$$

converges weakly as $n \to \infty$, in the Skorokhod topology on $\mathbb{D}([0,T],\mathbb{R})$, to the diffusion process $\{W(t)\}_{t\in[0,T]}$ satisfying

$$\begin{cases} dW(t) = \frac{1}{2}\nabla \log(\pi(W(t)))\,dt + dB(t) \\ W(0) \sim \mu(\cdot). \end{cases}$$

The above result provides some information on the relationship between M and δ_n, although it appears to be quite restrictive for practical purposes. For instance, halving the value of the proposal variance δ^2 not only suggests doubling the length of the chain, but also taking a new M greater than 4 times the original value, as the relationship $M_n\delta_n^2 \to \infty$ indicates.

Even though the corresponding limiting diffusions of the APM chain and marginal chain converge for a fixed time horizon T, this does not imply the APM chain is even positive recurrent. We now address the existence of an invariant distribution for the kernel $\tilde{P}_{M,n}$, this is done via geometric ergodicity. From Theorem 5.2, we know that the chain generated by the penalty method kernel $\bar{P}_{M,n}$ inherits geometric ergodicity from the marginal chain if $\sup_{x,y}\tau_{x,y}^2 < \infty$. This implies, by Theorem 2.3, that there exists a function $V_{M,n} \geq 1$, $\lambda_{M,n} < 1$, $b_{M,n} < \infty$ and a

small set $S_{M,n} \subseteq \mathcal{X}$ such that the following drift condition holds

$$\bar{P}_{M,n} V_{M,n}(x) \leq \lambda_{M,n} V_{M,n}(x) + b_{M,n} \mathbb{1}(x \in S_{M,n}), \qquad \text{for all } x \in \mathcal{X}.$$

The idea is to prove a similar condition for $\tilde{P}_{M,n}$ for large enough n and M. However, this becomes a difficult task unless we know $V_{M,n}$, $\lambda_{M,n}$, $b_{M,n}$ and $S_{M,n}$ explicitly, or unless we have uniform control over them. Under some conditions, the latter approach is feasible allowing us to prove the inheritance of geometric ergodicity from the marginal chain with kernel P_n, provided M and n are large.

Assumption 5. *The target density $\pi(x)$ is continuous, positive, symmetric and β-log-concave in the tails. The latter meaning there are $x^* > 0$ and $\beta > 0$ such that for all $y \geq x \geq x^*$ and $y \leq x \leq -x^*$*

$$\log(\pi(x)) - \log(\pi(y)) \geq \beta |x - y|.$$

Theorem 5.5. *Suppose Assumptions 2, 3 and 5 hold. For n sufficiently large, there exists $M_n \in \mathbb{N}$ such that the class of kernels $\left\{P_n, \tilde{P}_{M_n,n}, \tilde{P}_{M_n+1,n}, \dots\right\}$ is simultaneously geometrically ergodic.*

Of course, geometric ergodicity guarantees the existence of a noisy invariant $\tilde{\pi}_{M,n}$ if M and n are large enough. As done in Section 3.3.2 for obtaining (3.14), the simultaneous geometric ergodicity property from the above theorem implies the existence of $R_n < \infty$ and $\tau_n < 1$ such that for any $k \in \mathbb{N}$

$$\|\tilde{\pi}_{M,n}(\cdot) - \pi(\cdot)\|_{TV} \leq R_n \tau_n^k + k \sup_{x \in \mathcal{X}} \|\tilde{P}_{M,n}(x, \cdot) - P_n(x, \cdot)\|_{TV}. \tag{5.19}$$

In this case, one can show

$$\sup_{x \in \mathcal{X}} \|\tilde{P}_{M,n}(x, \cdot) - P_n(x, \cdot)\|_{TV} \lesssim \mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, \delta_n\right) + M^{-1/2},$$

and under some conditions for R_n and τ_n the invariants will converge in total variation.

Theorem 5.6. *Suppose Assumptions 2, 3 and 5 hold. If n is sufficiently large, there are $M_n \in \mathbb{N}$, $R_n < \infty$ and $\tau_n < 1$ such that if $M \geq M_n$ the invariants $\tilde{\pi}_{M,n}$ and π satisfy*

$$\|\tilde{\pi}_{M,n}(\cdot) - \pi(\cdot)\|_{TV} \leq \frac{1 + \log\left(r_n(M) \log\left(\tau_n^{-1}\right) R_n \tau_n^{-1}\right)}{r_n(M) \log\left(\tau_n^{-1}\right)},$$

where $r_n(M)^{-1} \propto \mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, \delta_n\right) + M^{-1/2}$. *Additionally, if*

$$\lim_{M,n\to\infty} \frac{\log(R_n)}{r_n(M)\log\left(\tau_n^{-1}\right)} \to 0, \tag{5.20}$$

then

$$\lim_{M,n\to\infty} \|\tilde{\pi}_{M,n}(\cdot) - \pi(\cdot)\|_{TV} = 0.$$

In practice, showing (5.20) holds may be challenging. Under some additional conditions, one could possibly obtain explicit expressions for R_n and τ_n as done in Rosenthal (1995). However this is not attempted here.

Finally, notice that Assumptions 3, 4 and 5 are fairly strong, but they were chosen in this way for clarity and simplicity in the proofs of the previous theorems. We envisage that many of the aforementioned assumptions could be relaxed, allowing us to obtain similar results for more complicated targets on $\mathbb{R}^d$. We finish this chapter with a set of technical results, leading to the proofs of the above theorems.

5.4.1 Bounds on the acceptance probabilities

This section provides some bounds for the acceptance probability $\tilde{\alpha}_M$ in terms of the penalty method acceptance $\bar{\alpha}_M$ and the marginal acceptance α. These bounds will be useful in the following sections since they provide enough control on the dissimilarity between the various chains considered. In order to do so, let $\check{P}_M : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ be the kernel of the chain generated using the acceptance ratio $\bar{R}_M$ in (5.14), but without the log-normality assumption. To avoid confusion, let $\check{R}_M$ denote such acceptance ratio, i.e.

$$\check{R}_M(x,y) = \left(\prod_{i=1}^{M} R_{x,y}^{(i)}\right)^{1/M} \exp\left\{m_{y,x} - \frac{1}{2M}\tau_{x,y}^2\right\},$$

where $\left\{R_{x,y}^{(i)}\right\}_i$ are not necessarily log-normally distributed. Therefore,

$$\check{P}_M(x,dy) := q(x,dy)\,\check{\alpha}_M(x,y) + \delta_x(dy)\left[1 - \int_{\mathcal{X}} q(x,dy)\,\check{\alpha}_M(x,y)\right],$$

where

$$\check{\alpha}_M(x,y) := \mathbb{E}\left[\min\left\{1, \check{R}_M(x,y)\right\}\right].$$

The following result bounds $\bar{\alpha}_M$ in terms of the marginal acceptance α.

Corollary 5.3. *The acceptance probability $\bar{\alpha}_M$ satisfies*

$$\bar{\alpha}_M(x,y) = \Phi\left(\frac{\log\left(r(x,y)\right) - \frac{1}{2M}\tau_{x,y}^2}{\tau_{x,y}/\sqrt{M}}\right) + r(x,y)\,\Phi^c\left(\frac{\log\left(r(x,y)\right) + \frac{1}{2M}\tau_{x,y}^2}{\tau_{x,y}/\sqrt{M}}\right),$$

with partial derivative

$$\begin{aligned}\frac{\partial}{\partial y}\bar{\alpha}_M(x,y) = r(x,y)\,\nabla\log\pi(y)\,\Phi^c\left(\frac{\log\left(r(x,y)\right) + \frac{1}{2M}\tau_{x,y}^2}{\tau_{x,y}/\sqrt{M}}\right)\\ - \frac{1}{\sqrt{M}}\frac{\partial}{\partial y}(\tau_{x,y})\,\phi\left(\frac{\log\left(r(x,y)\right) - \frac{1}{2M}\tau_{x,y}^2}{\tau_{x,y}/\sqrt{M}}\right).\end{aligned}$$

Additionally, the acceptance probabilities $\bar{\alpha}_M$ and α satisfy

$$2\Phi\left(-\frac{\tau_{x,y}}{2\sqrt{M}}\right)\alpha(x,y) \leq \bar{\alpha}_M(x,y) \leq \alpha(x,y).$$

Proof. The expression for $\bar{\alpha}_M$ and the bounds in terms of α follow from Remark 5.1 and Lemma 5.1 respectively, by considering $\tau_{x,y}/\sqrt{M}$ instead of $\tau_{x,y}$. The expression for the partial derivative follows from direct calculations. □

The next two lemmas provide bounds for the difference between $\check{\alpha}_M$ and $\bar{\alpha}_M$, and between $\tilde{\alpha}_M$ and $\check{\alpha}_M$. These results will be useful later on for addressing the proofs of Theorems 5.4, 5.5 and 5.6.

Lemma 5.2. *Under Assumption 3, the acceptance probabilities $\check{\alpha}_M$ and $\bar{\alpha}_M$ satisfy*

$$\sup_{(x,y)\in\mathcal{X}^2}\left|\check{\alpha}_M(x,y) - \bar{\alpha}_M(x,y)\right| \lesssim M^{-1/2}.$$

Proof. Using Jensen's inequality,

$$\begin{aligned}&\left|\check{\alpha}_M(x,y) - \bar{\alpha}_M(x,y)\right|\\ &\qquad = \left|\mathbb{E}\left[\min\left\{1, \check{R}_M(x,y)\right\}\right] - \mathbb{E}\left[\min\left\{1, \bar{R}_M(x,y)\right\}\right]\right|\\ &\qquad = \left|\int_0^1 \left(\mathbb{P}\left[u < \check{R}_M(x,y)\right] - \mathbb{P}\left[u < \bar{R}_M(x,y)\right]\right) du\right|\\ &\qquad \leq \int_0^1 \left|\mathbb{P}\left[u < \check{R}_M(x,y)\right] - \mathbb{P}\left[u < \bar{R}_M(x,y)\right]\right| du.\end{aligned}$$

From the expressions of $\check{R}_M$ and $\bar{R}_M$

$$\begin{aligned}\mathbb{P}\left[u < \check{R}_M(x,y)\right] &= \mathbb{P}\left[\log(u) + \frac{1}{2M}\tau_{x,y}^2 < \bar{L}_{x,y,M} - m_{x,y}\right] \\ &= \mathbb{P}\left[\log(u) + \frac{1}{2M}\tau_{x,y}^2 - \log(r(x,y)) < \bar{L}_{x,y,M} - \mathbb{E}\left[L_{x,y}\right]\right] \\ &= \mathbb{P}\left[\frac{\bar{L}_{x,y,M} - \mathbb{E}\left[L_{x,y}\right]}{\tau_{x,y}/\sqrt{M}} > \frac{\log(u) + \frac{1}{2M}\tau_{x,y}^2 - \log(r(x,y))}{\tau_{x,y}/\sqrt{M}}\right],\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}\left[u < \bar{R}_M(x,y)\right] &= \mathbb{P}\left[\log(u) + \frac{1}{2M}\tau_{x,y}^2 < \frac{\tau_{x,y}}{\sqrt{M}}\xi + \log(r(x,y))\right] \\ &= \Phi^c\left(\frac{\log(u) + \frac{1}{2M}\tau_{x,y}^2 - \log(r(x,y))}{\tau_{x,y}/\sqrt{M}}\right).\end{aligned}$$

Hence, by the Berry-Esseen Theorem

$$\begin{aligned}&\left|\mathbb{P}\left[u < \check{R}_M(x,y)\right] - \mathbb{P}\left[u < \bar{R}_M(x,y)\right]\right| \\ &\qquad \leq \sup_{w\in\mathbb{R}}\left|\mathbb{P}\left[\frac{\bar{L}_{x,y,M} - \mathbb{E}\left[L_{x,y}\right]}{\tau_{x,y}/\sqrt{M}} \leq w\right] - \Phi(w)\right| \\ &\qquad \lesssim \frac{\mathbb{E}\left[\left|L_{x,y} - \mathbb{E}\left[L_{x,y}\right]\right|^3\right]}{\tau_{x,y}^3\sqrt{M}},\end{aligned}$$

implying

$$\sup_{x,y}\left|\check{\alpha}_M(x,y) - \bar{\alpha}_M(x,y)\right| \lesssim M^{-1/2}\frac{\sup_{x,y}\mathbb{E}\left[\left|L_{x,y} - \mathbb{E}\left[L_{x,y}\right]\right|^3\right]}{\inf_{x,y}\tau_{x,y}^3}. \qquad \square$$

Lemma 5.3. *Under Assumptions 2 and 3 the acceptance probabilities $\tilde{\alpha}_M$ and $\check{\alpha}_M$ satisfy*

$$\left|\tilde{\alpha}_M(x,y) - \check{\alpha}_M(x,y)\right| \lesssim \mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, |y-x|\right) + M^{-3/2}.$$

Proof. Since the function $g(x) = \min\{1, \exp\{x\}\}$ is Lipschitz with coefficient equal to 1,

$$
\begin{aligned}
\left|\tilde{\alpha}_M(x,y) - \check{\alpha}_M(x,y)\right| &\leq \mathbb{E}\left[\left|\min\left\{1, \tilde{R}_M(x,y)\right\} - \min\left\{1, \check{R}_M(x,y)\right\}\right|\right] \\
&\leq \mathbb{E}\left[\left|\log\left(\tilde{R}_M(x,y)\right) - \log\left(\check{R}_M(x,y)\right)\right|\right] \\
&\leq \mathbb{E}\left[\left|\hat{m}_{x,y,M} - m_{x,y}\right|\right] + \frac{1}{2M}\mathbb{E}\left[\left|\hat{\tau}^2_{x,y,M} - \tau^2_{x,y}\right|\right].
\end{aligned}
$$

By Assumption 2

$$
\mathbb{E}\left[\left|\hat{m}_{x,y,M} - m_{x,y}\right|\right] \lesssim \mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, |y-x|\right),
$$

and using Assumption 3

$$
\begin{aligned}
\mathbb{E}\left[\left|\hat{\tau}^2_{x,y,M} - \tau^2_{x,y}\right|\right] &\leq \left(\mathbb{V}\left[\hat{\tau}^2_{x,y,M}\right]\right)^{1/2} \leq M^{-1/2} \sup_{x,y} \mathbb{E}\left[\left(L_{x,y} - \mathbb{E}\left[L_{x,y}\right]\right)^4\right] \\
&\lesssim M^{-1/2}.
\end{aligned}
$$

Therefore

$$
\left|\tilde{\alpha}_M(x,y) - \check{\alpha}_M(x,y)\right| \lesssim \mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, |y-x|\right) + M^{-3/2},
$$

as required. □

The lemma appearing below provides a relationship between the approximate acceptance $\tilde{\alpha}_M$ and the marginal acceptance α. This result will be useful for showing compact sets are small for the APM chain.

Lemma 5.4. *Under Assumptions 2 and 3 there exists $C > 0$ such that the acceptance probabilities $\tilde{\alpha}_M$ and α satisfy*

$$
\tilde{\alpha}_M(x,y) \gtrsim \alpha(x,y) \exp\left\{-C\mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, |y-x|\right)\right\}.
$$

Proof. Applying the inequality $\min\{1, ab\} \geq \min\{1, a\}\min\{1, b\}$ and Jensen's in-

equality in that order,

$$\begin{aligned}\frac{\tilde{\alpha}_M(x,y)}{\alpha(x,y)} &\geq \mathbb{E}\left[\min\left\{1, \exp\left\{\frac{1}{M}\sum_{i=1}^{M}\log\left(V_{x,y}^{(i)}\right) + \hat{m}_{y,x,M} - \frac{1}{2M}\hat{\tau}_{x,y,M}^2\right\}\right\}\right] \\ &\geq \mathbb{E}\left[\exp\left\{-\frac{1}{M}\sum_{i=1}^{M}\left|\log\left(V_{x,y}^{(i)}\right) - m_{x,y}\right| - |m_{x,y} - \hat{m}_{x,y,M}| - \frac{1}{2M}\hat{\tau}_{x,y,M}^2\right\}\right] \\ &\geq \exp\left\{-\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left|\log\left(V_{x,y}^{(i)}\right) - m_{x,y}\right|\right] - \frac{1}{2M}\mathbb{E}\left[\hat{\tau}_{x,y,M}^2\right]\right\} \\ &\quad \times \exp\left\{-\mathbb{E}\left[|m_{x,y} - \hat{m}_{x,y,M}|\right]\right\}.\end{aligned}$$

By Assumption 3

$$\mathbb{E}\left[\left|\log\left(V_{x,y}\right) - m_{x,y}\right|\right] \leq \sup_{x,y}\mathbb{E}\left[\left|L_{x,y} - \mathbb{E}\left[L_{x,y}\right]\right|\right] < \infty,$$

and

$$\mathbb{E}\left[\hat{\tau}_{x,y,M}^2\right] = \tau_{x,y}^2 \leq \sup_{x,y}\mathbb{E}\left[\left(L_{x,y} - \mathbb{E}\left[L_{x,y}\right]\right)^2\right] < \infty.$$

Therefore, using Assumption 2, there exists $C > 0$ such that

$$\begin{aligned}\frac{\tilde{\alpha}_M(x,y)}{\alpha(x,y)} &\gtrsim \exp\left\{-\mathbb{E}\left[|m_{x,y} - \hat{m}_{x,y,M}|\right]\right\} \\ &\geq \exp\left\{-C\mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, |y-x|\right)\right\}.\end{aligned}$$

□

5.4.2 Proofs of Theorems 5.3 and 5.4

Proof of Theorem 5.3. In order to prove weak convergence, we need to show that the convergence of the finite-dimensional distributions and relative compactness hold. This is done by showing convergence of the corresponding generators in an appropriate sense. For the first task, we rely on Ethier and Kurtz (2008, Chapter 4, Theorem 8.2), where equations (8.8)-(8.11) must be satisfied.

Equation (8.8). We need to prove

$$\sup_n \sup_{t\leq T}\mathbb{E}\left|\varsigma_{M,n}(t)\right| < \infty, \tag{5.21}$$

where

$$\varsigma_{M,n}(t) := \delta_n^{-2}\int_0^{\delta_n}\mathbb{E}\left[\psi\left(\bar{\mathfrak{X}}_{M,n}(t+s)\right)\left|\bar{\mathfrak{X}}_{M,n}(t)\right.\right]ds,$$

and ψ belonging to the core of the generator of W_M. From Ethier and Kurtz (2008, Chapter 8, Theorem 2.1) and under Assumptions 3 and 4, the set of smooth and compactly supported functions, denoted by C_c^∞, is a suitable core; hence (5.21) is immediately satisfied.

Equation (8.11). As noted in Ethier and Kurtz (2008, Chapter 4, Remark 8.3 (a)), it suffices to show

$$\lim_{n\to\infty} \mathbb{E}\left|G_{M,n}\psi\left(\bar{\mathfrak{X}}_{M,n}(t)\right) - \mathcal{G}_M\psi\left(\bar{\mathfrak{X}}_{M,n}(t)\right)\right| = 0, \tag{5.22}$$

where $G_{M,n}$ and $\mathcal{G}_M$ are the generators of $\bar{\mathfrak{X}}_{M,n}$ and W_M, respectively, given by

$$\begin{aligned} G_{M,n}\psi(x) &:= \delta_n^{-2}\mathbb{E}\left[\psi\left(\bar{\mathfrak{X}}_{M,n}\left(t+\delta_n^2\right)\right) - \psi\left(\bar{\mathfrak{X}}_{M,n}(t)\right)\left|\bar{\mathfrak{X}}_{M,n}(t) = x\right.\right], \\ \mathcal{G}_M\psi(x) &:= \frac{1}{2}\left[b_M(x)\,\psi'(x) + a_M(x)\,\psi''(x)\right]. \end{aligned}$$

Set $k = \left\lfloor t\delta_n^{-2}\right\rfloor$ and notice that $\bar{\mathfrak{X}}_{M,n}(t) = \bar{X}_{M,n}(k)$ and $\bar{\mathfrak{X}}_{M,n}\left(t+\delta_n^2\right) = \bar{X}_{M,n}(k+1)$, which implies

$$\begin{aligned} &G_{M,n}\psi\left(\bar{\mathfrak{X}}_{M,n}(t)\right) = G_{M,n}\psi\left(\bar{X}_{M,n}(k)\right) \\ &\quad = \delta_n^{-2}\mathbb{E}\left[\left(\psi\left(Y_n^{(k+1)}\right) - \psi\left(\bar{X}_{M,n}(k)\right)\right)\bar{\alpha}_M\left(\bar{X}_{M,n}(k), Y_n^{(k+1)}\right)\left|\bar{X}_{M,n}(k)\right.\right], \end{aligned}$$

where $Y_n^{(k+1)} \sim q_n\left(\bar{X}_{M,n}(k), \cdot\right)$. Take $x \in supp(\psi)$ and for simplicity let $\alpha_{x,M}(y) := \bar{\alpha}_M(x,y)$, then using a Taylor expansion for ψ and $\alpha_{x,M}(y)$

$$\begin{aligned} &(\psi(y) - \psi(x))\,\alpha_{x,M}(y) \\ &\quad = \psi'(x)(y-x)\left(\alpha_{x,M}(x) + \alpha'_{x,M}(x)(y-x)\right) + \frac{1}{2}\psi''(x)\,\alpha_{x,M}(x)(y-x)^2 \\ &\quad + (y-x)^3\left(\frac{1}{6}\psi'''\left(\rho^{(1)}\right)\alpha_{x,M}(y) + \frac{1}{2}\psi''(x)\,\alpha'_{x,M}\left(\rho^{(2)}\right) + \frac{1}{2}\psi'(x)\,\alpha''_{x,M}\left(\rho^{(3)}\right)\right), \end{aligned}$$

where $\rho^{(i)} \in (x,y)\cup(y,x)$ for all $i \in \{1,2,3\}$. For Y_n as in (5.17),

$$\begin{aligned} G_{M,n}\psi(x) &= \delta_n^{-2}\mathbb{E}\left[(\psi(Y_n) - \psi(x))\,\alpha_{x,M}(Y_n)\right] \\ &= \alpha'_{x,M}(x)\,\psi'(x) + \frac{1}{2}\alpha_{x,M}(x)\,\psi''(x) + \frac{1}{6}\delta_n\mathbb{E}\left[\xi^3\psi'''\left(\rho_n^{(1)}\right)\alpha_{x,M}(Y_n)\right] \\ &\quad + \frac{1}{2}\delta_n\psi''(x)\,\mathbb{E}\left[\xi^3\alpha'_{x,M}\left(\rho_n^{(2)}\right)\right] + \frac{1}{2}\delta_n\psi'(x)\,\mathbb{E}\left[\xi^3\alpha''_{x,M}\left(\rho_n^{(3)}\right)\right]. \end{aligned} \tag{5.23}$$

Now, from Corollary 5.3 $\alpha_{x,M}(x) = a_M(x)$, $\alpha'_{x,M}(x) = \frac{1}{2}b_M(x)$ and under Assump-

tions 3 and 4 it is not difficult to show

$$\sup_{x\in supp(\psi)}\left|\alpha'_{x,M}\left(\rho_n^{(2)}\right)\right|\lesssim 1 \quad \text{and} \quad \sup_{x\in supp(\psi)}\left|\alpha''_{x,M}\left(\rho_n^{(3)}\right)\right|\lesssim M^{1/2}\left(1+\delta_n\left|\xi\right|\right),$$

implying

$$\sup_{x\in supp(\psi)}\left|G_{M,n}\psi\left(x\right)-\mathcal{G}_M\psi\left(x\right)\right|\lesssim \delta_n M^{1/2}.$$

Additionally, if $x\notin supp\left(\psi\right)$ then $\mathcal{G}_M\psi\left(x\right)\equiv 0$ and

$$\left(\psi\left(y\right)-\psi\left(x\right)\right)\alpha_{x,M}\left(y\right)=\frac{1}{6}\psi'''\left(\rho\right)\left(y-x\right)^3\alpha_{x,M}\left(y\right),$$

which directly implies

$$\sup_{x\notin supp(\psi)}\left|G_n\psi\left(x\right)-\mathcal{G}\psi\left(x\right)\right|\lesssim \delta_n.$$

Therefore,

$$\left|G_{M,n}\psi\left(\bar{\mathfrak{X}}_{M,n}\left(t\right)\right)-\mathcal{G}_M\psi\left(\bar{\mathfrak{X}}_{M,n}\left(t\right)\right)\right|\lesssim \delta_n M^{1/2} \tag{5.24}$$

and (5.22) follows since M is fixed.

Equation (8.9). We must show

$$\sup_n\sup_{t\leq T}\mathbb{E}\left|G_{M,n}\psi\left(\bar{\mathfrak{X}}_{M,n}\left(t\right)\right)\right|<\infty. \tag{5.25}$$

From (5.24) and using the triangle inequality

$$\begin{aligned}\left|G_{M,n}\psi\left(\bar{\mathfrak{X}}_{M,n}\left(t\right)\right)\right| &\leq \left|G_{M,n}\psi\left(\bar{\mathfrak{X}}_{M,n}\left(t\right)\right)-\mathcal{G}_M\psi\left(\bar{\mathfrak{X}}_{M,n}\left(t\right)\right)\right|+\left|\mathcal{G}_M\psi\left(\bar{\mathfrak{X}}_{M,n}\left(t\right)\right)\right| \\ &\lesssim \delta_n M^{1/2}+\left|\mathcal{G}_M\psi\left(\bar{\mathfrak{X}}_{M,n}\left(t\right)\right)\right|.\end{aligned}$$

Since $\mathcal{G}_M\psi\left(x\right)\equiv 0$ for $x\notin supp\left(\psi\right)$, then

$$\left|\mathcal{G}_M\psi\left(x\right)\right|\lesssim\left(\sup_{x\in supp(\psi)}\left|b_M\left(x\right)\right|+1\right),$$

implying

$$\sup_x\left|G_{M,n}\psi\left(x\right)\right|\lesssim M^{1/2}. \tag{5.26}$$

Condition in (5.25) follows immediately since M is fixed.

Equation (8.10). Also from Ethier and Kurtz (2008, Chapter 4, Remark 8.3 (a)), it suffices to show

$$\lim_{n\to\infty} \mathbb{E}\left|\varsigma_{M,n}(t) - \psi\left(\bar{\mathfrak{X}}_{M,n}(t)\right)\right| = 0. \tag{5.27}$$

First notice that $\bar{\mathfrak{X}}_{M,n}(t+s) = \bar{\mathfrak{X}}_{M,n}(t)$ for $s \in \left[0, \delta_n^2\right)$, hence for any $s \in \left[0, \delta_n^2\right]$

$$\begin{aligned}
&\left|\mathbb{E}\left[\psi\left(\bar{\mathfrak{X}}_{M,n}(t+s)\right) - \psi\left(\bar{\mathfrak{X}}_{M,n}(t)\right)\left|\bar{\mathfrak{X}}_{M,n}(t)\right.\right]\right| \\
&\quad \leq \left|\mathbb{E}\left[\psi\left(\bar{\mathfrak{X}}_{M,n}\left(t+\delta_n^2\right)\right) - \psi\left(\bar{\mathfrak{X}}_{M,n}(t)\right)\left|\bar{\mathfrak{X}}_{M,n}(t)\right.\right]\right| = \delta_n^2\left|G_{M,n}\psi\left(\bar{\mathfrak{X}}_{M,n}(t)\right)\right|.
\end{aligned}$$

Thus,

$$\begin{aligned}
\left|\varsigma_{M,n}(t) - \psi\left(\bar{\mathfrak{X}}_{M,n}(t)\right)\right| &= \left|\delta_n^{-2}\int_0^{\delta_n^2} \mathbb{E}\left[\psi\left(\bar{\mathfrak{X}}_{M,n}(t+s)\right) - \psi\left(\bar{\mathfrak{X}}_{M,n}(t)\right)\left|\bar{\mathfrak{X}}_{M,n}(t)\right.\right] ds\right| \\
&\leq \delta_n^{-2}\int_0^{\delta_n^2} \left|\mathbb{E}\left[\psi\left(\bar{\mathfrak{X}}_{M,n}(t+s)\right) - \psi\left(\bar{\mathfrak{X}}_{M,n}(t)\right)\left|\bar{\mathfrak{X}}_{M,n}(t)\right.\right]\right| ds \\
&\leq \delta_n^2\left|G_{M,n}\psi\left(\bar{\mathfrak{X}}_{M,n}(t)\right)\right|.
\end{aligned}$$

Finally, from (5.26)

$$\left|\varsigma_{M,n}(t) - \psi\left(\bar{\mathfrak{X}}_{M,n}(t)\right)\right| \lesssim \delta_n^2 M^{1/2}, \tag{5.28}$$

and (5.27) follows since M is fixed.

To prove relative compactness we use Ethier and Kurtz (2008, Chapter 4, Corollary 8.6), where equations (8.33) and (8.34) must be satisfied.

Equation (8.33). We must show

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{t\leq T}\left|\varsigma_{M,n}(t) - \psi\left(\bar{\mathfrak{X}}_{M,n}(t)\right)\right|\right] = 0, \tag{5.29}$$

which is immediate since (5.28) implies

$$\sup_{t\leq T}\left|\varsigma_{M,n}(t) - \psi\left(\bar{\mathfrak{X}}_{M,n}(t)\right)\right| \lesssim \delta_n^2 M^{1/2}.$$

Equation (8.34). We need to prove for some $p > 1$

$$\sup_n \mathbb{E}\left[\left(\int_0^T \left|G_{M,n}\psi\left(\bar{\mathfrak{X}}_{M,n}(t)\right)\right|^p dt\right)^{1/p}\right] < \infty, \tag{5.30}$$

which follows directly from (5.26). □

Remark 5.5. From the previous proof, notice that the result is obtained if the following sufficient condition holds

$$\lim_{n\to\infty} \sup_x \left|G_{M,n}\psi(x) - \mathcal{G}_M\psi(x)\right| = 0.$$

If instead M increases as δ_n decreases, the corresponding limiting process is the usual Langevin diffusion satisfying (5.18) with $\gamma = 1$. The following lemma states such result.

Lemma 5.5. *Set $T > 0$ and let $\left\{\bar{X}_n(k)\right\}_{k\geq 0}$ be the discrete-time Markov chain generated by $\bar{P}_{M,n}$ with proposal q_n and where $M = M_n$ is such that $M_n\delta_n^2 \to \infty$ as $n \to \infty$. If $\bar{X}_{M,n}(0) \sim \mu(\cdot)$ and under Assumptions 3 and 4, the continuous-time process*

$$\bar{\mathfrak{X}}_n(t) := \bar{X}_n\left(\left\lfloor t\delta_n^{-2}\right\rfloor\right)$$

converges weakly as $n \to \infty$, in the Skorokhod topology on $\mathbb{D}([0,T],\mathbb{R})$, to the diffusion process $\{W(t)\}_{t\in[0,T]}$ satisfying

$$\begin{cases} dW(t) = \frac{1}{2}\nabla \log\left(\pi\left(W(t)\right)\right)dt + dB(t) \\ W(0) \sim \mu(\cdot). \end{cases}$$

Proof. The proof requires showing similar expressions of equations (5.21)-(5.30), although now M needs to be considered as a function of n. Using Remark 5.5, it suffices to show

$$\lim_{n\to\infty} \sup_x \left|G_n\psi(x) - \mathcal{G}\psi(x)\right| = 0,$$

where G_n and $\mathcal{G}$ are given by

$$\begin{aligned} G_n\psi(x) &:= \delta_n^{-2}\mathbb{E}\left[\psi\left(\bar{\mathfrak{X}}_n\left(t+\delta_n^2\right)\right) - \psi\left(\bar{\mathfrak{X}}_n(t)\right)\middle|\bar{\mathfrak{X}}_n(t) = x\right], \\ \mathcal{G}\psi(x) &:= \frac{1}{2}\left[\nabla \log\left(\pi(x)\right)\psi'(x) + \psi''(x)\right]. \end{aligned} \tag{5.31}$$

Similarly to (5.23), a Taylor series expansion for ψ and $\alpha_{x,M}(y)$ gives

$$\begin{aligned} G_n\psi(x) &= \delta_n^{-2}\mathbb{E}\left[\left(\psi(Y_n)-\psi(x)\right)\alpha_{x,M}(Y_n)\right] \\ &= \mathbb{E}\left[\xi^2\mathbb{E}\left[\alpha'_{x,M}(x+\delta_n\xi U)\middle|\xi\right]\right]\psi'(x)+\frac{1}{2}\alpha_{x,M}(x)\psi''(x) \\ &\quad+\frac{1}{6}\delta_n\alpha_{x,M}(x)\mathbb{E}\left[\xi^3\psi'''\left(\rho_n^{(1)}\right)\right]+\frac{1}{2}\delta_n\mathbb{E}\left[\xi^3\psi''\left(\rho_n^{(3)}\right)\alpha'_{x,M}\left(\rho_n^{(2)}\right)\right], \end{aligned}$$

where U is an independent uniform random variable coming from the integral form of the remainder. As before, $\alpha_{x,M}(x)=a_M(x)$ and

$$\sup_{x\in supp(\psi)}\left|\alpha'_{x,M}\left(\rho_n^{(2)}\right)\right|\lesssim 1,$$

which implies

$$\begin{aligned} &\sup_{x\in supp(\psi)}\left|G_n\psi(x)-\mathcal{G}\psi(x)\right| \\ &\quad\lesssim \sup_{x\in supp(\psi)}\left|\mathbb{E}\left[\xi^2\alpha'_{x,M}(x+\delta_n\xi U)\right]-\tfrac{1}{2}\nabla\log(\pi(x))\right| \\ &\quad+\sup_{x\in supp(\psi)}\left|a_M(x)-1\right|+\delta_n, \end{aligned} \tag{5.32}$$

Additionally, if $x\notin supp(\psi)$ then $\mathcal{G}\psi(x)\equiv 0$ and

$$\left(\psi(y)-\psi(x)\right)\alpha_{x,M}(y)=\frac{1}{6}\psi'''(\rho)(y-x)^3\alpha_{x,M}(y),$$

which directly implies

$$\sup_{x\notin supp(\psi)}\left|G_n\psi(x)-\mathcal{G}\psi(x)\right|\lesssim\delta_n.$$

The proof reduces to show the right hand side of the inequality (5.32) goes to zero. Using a Taylor expansion for $a_M(x)$ in terms of $M^{-1/2}$

$$\sup_{x\in supp(\psi)}|a_M(x)-1|=\sup_{x\in supp(\psi)}|a_M(x)-a_\infty(x)|\lesssim M^{-1/2}.$$

Respecting the other term, looking back at the analytic expression for $\alpha'_{x,M}(t)$ in Corollary 5.3 and under the assumptions for π, it is enough to show

$$\lim_{n\to\infty}\sup_{x\in supp(\psi)}\left|\nabla\log\pi(x)\left(\mathbb{E}\left[\xi^2\Phi^c\left(\frac{\log(r(x,T_n))+\frac{1}{2M}\tau^2_{x,T_n}}{\tau_{x,T_n}/\sqrt{M}}\right)\right]-\frac{1}{2}\right)\right|=0, \tag{5.33}$$

where $T_n=x+\delta_n\xi U$ with ξ and U independent.

Take $\varepsilon > 0$ and define $A_\varepsilon := \{x \in supp(\psi) : |\nabla \log \pi(x)| > \varepsilon\}$, then

$$\sup_{x \in supp(\psi) \setminus A_\varepsilon} |\nabla \log \pi(x)| \leq \varepsilon. \tag{5.34}$$

Since $supp(\psi)$ is compact and $\nabla \log \pi$ is continuous, there exists $\gamma > 0$ (independent of x) such that for any $x \in A_\varepsilon$ and any z such that$|z - x| < \gamma$

$$\nabla \log \pi(z) \geq \frac{\varepsilon}{2} \qquad \text{or} \qquad \nabla \log \pi(z) \leq -\frac{\varepsilon}{2}.$$

We only deal with the latter case, the proof of the former is similar. If $Y_n \in (x, x+\gamma)$ there exists $Z_n \in (x, T_n) \subseteq (x, Y_n)$ such that

$$\begin{aligned}\Phi^c\left(\frac{\log(r(x,T_n)) + \frac{1}{2M}\tau_{x,T_n}^2}{\tau_{x,T_n}/\sqrt{M}}\right) &= \Phi^c\left(\frac{\nabla \log \pi(Z_n)(T_n - x) + \frac{1}{2M}\tau_{x,T_n}^2}{\tau_{x,T_n}/\sqrt{M}}\right) \\ &= \Phi\left(\frac{|\nabla \log \pi(Z_n)|\,\delta_n \xi U - \frac{1}{2M}\tau_{x,T_n}^2}{\tau_{x,T_n}/\sqrt{M}}\right) \geq \Phi\left(\sqrt{M}\delta_n \frac{\varepsilon \xi U}{2\bar{\tau}} - \frac{1}{2\sqrt{M}}\bar{\tau}\right),\end{aligned}$$

where $\bar{\tau} := \sup_{x,y} \tau_{x,y}$. If instead $Y_n \in (x-\gamma, x)$, then $Z_n \in (T_n, x) \subseteq (Y_n, x)$ and

$$\begin{aligned}\Phi^c\left(\frac{\log(r(x,T_n)) + \frac{1}{2M}\tau_{x,T_n}^2}{\tau_{x,T_n}/\sqrt{M}}\right) &= \Phi\left(\frac{|\nabla \log \pi(Z_n)|\,\delta_n \xi U - \frac{1}{2M}\tau_{x,T_n}^2}{\tau_{x,T_n}/\sqrt{M}}\right) \\ &= \Phi^c\left(\frac{|\nabla \log \pi(Z_n)|\,\delta_n |\xi| U + \frac{1}{2M}\tau_{x,T_n}^2}{\tau_{x,T_n}/\sqrt{M}}\right) \leq \Phi^c\left(\sqrt{M}\delta_n \frac{\varepsilon |\xi| U}{2\bar{\tau}}\right).\end{aligned}$$

Consequently, recalling that $Y_n = x + \delta_n \xi$,

$$\begin{aligned}\sup_{x \in A_\varepsilon}\left|\mathbb{E}\left[\xi^2 \Phi^c\left(\frac{\log(r(x,T_n)) + \frac{1}{2M}\tau_{x,T_n}^2}{\tau_{x,T_n}/\sqrt{M}}\right)\right] - \frac{1}{2}\right| &\leq \mathbb{E}\left[\xi^2 \mathbb{1}\left(|\xi| \geq \frac{\gamma}{\delta_n}\right)\right] \\ &+ \mathbb{E}\left[\xi^2 \Phi^c\left(\sqrt{M}\delta_n \frac{\varepsilon|\xi|U}{2\bar{\tau}}\right)\mathbb{1}\left(-\frac{\gamma}{\delta_n} < \xi < 0\right)\right] \\ &+ \left(\frac{1}{2} - \mathbb{E}\left[\xi^2 \Phi\left(\sqrt{M}\delta_n \frac{\varepsilon \xi U}{2\bar{\tau}} - \frac{1}{2\sqrt{M}}\bar{\tau}\right)\mathbb{1}\left(0 < \xi < \frac{\gamma}{\delta_n}\right)\right]\right),\end{aligned}$$

which implies by the Dominated Convergence Theorem, assuming $M_n \delta_n^2 \to \infty$,

$$\lim_{n\to\infty} \sup_{x \in A_\varepsilon}\left|\mathbb{E}\left[\xi^2 \Phi^c\left(\frac{\log(r(x,T_n)) + \frac{1}{2M}\tau_{x,T_n}^2}{\tau_{x,T_n}/\sqrt{M}}\right)\right] - \frac{1}{2}\right| = 0. \tag{5.35}$$

Therefore, from (5.34) and (5.35)

$$\begin{aligned}\lim_{n\to\infty}\sup_{x\in supp(\psi)}&\left|\nabla\log\pi(x)\left(\mathbb{E}\left[\xi^2\Phi^c\left(\frac{\log\left(r\left(x,T_n\right)\right)+\frac{1}{2M}\tau_{x,T_n}^2}{\tau_{x,T_n}/\sqrt{M}}\right)\right]-\frac{1}{2}\right)\right|\\ &\lesssim\lim_{n\to\infty}\sup_{x\in A_\varepsilon}\left|\mathbb{E}\left[\xi^2\Phi^c\left(\frac{\log\left(r\left(x,T_n\right)\right)+\frac{1}{2M}\tau_{x,T_n}^2}{\tau_{x,T_n}/\sqrt{M}}\right)\right]-\frac{1}{2}\right|\\ &\quad+\lim_{n\to\infty}\sup_{x\in supp(\psi)\setminus A_\varepsilon}\left|\nabla\log\pi(x)\right|\lesssim\varepsilon,\end{aligned}$$

and (5.33) holds since ε can be taken arbitrarily small.

Thus,

$$\lim_{n\to\infty}\sup_x\left|G_n\psi(x)-\mathcal{G}\psi(x)\right|=0 \tag{5.36}$$

as required. □

Finally, we are in shape for proving Theorem 5.4.

Proof of Theorem 5.4. As before, it suffices to show

$$\lim_{n\to\infty}\sup_x\left|\tilde{G}_n\psi(x)-\mathcal{G}\psi(x)\right|=0,$$

where $\tilde{G}_n$ and $\mathcal{G}$ are given by

$$\begin{aligned}\tilde{G}_n\psi(x)&:=\delta_n^{-2}\mathbb{E}\left[\psi\left(\tilde{\mathfrak{X}}_n\left(t+\delta_n^2\right)\right)-\psi\left(\tilde{\mathfrak{X}}_n(t)\right)\middle|\tilde{\mathfrak{X}}_n(t)=x\right],\\ \mathcal{G}\psi(x)&:=\frac{1}{2}\left[\nabla\log\left(\pi(x)\right)\psi'(x)+\psi''(x)\right].\end{aligned}$$

Consider G_n as in (5.31), then

$$\begin{aligned}&\left|\tilde{G}_n\psi(x)-G_n\psi(x)\right|\\ &\quad=\delta_n^{-2}\left|\mathbb{E}\left[\left(\psi\left(Y_n\right)-\psi(x)\right)\left(\tilde{\alpha}_M\left(x,Y_n\right)-\bar{\alpha}_M\left(x,Y_n\right)\right)\right]\right|\\ &\quad\lesssim\delta_n^{-2}\mathbb{E}\left|\tilde{\alpha}_M\left(x,Y_n\right)-\bar{\alpha}_M\left(x,Y_n\right)\right|\\ &\quad\lesssim\delta_n^{-2}\left(\mathbb{E}\left|\tilde{\alpha}_M\left(x,Y_n\right)-\check{\alpha}_M\left(x,Y_n\right)\right|+\mathbb{E}\left|\check{\alpha}_M\left(x,Y_n\right)-\bar{\alpha}_M\left(x,Y_n\right)\right|\right),\end{aligned}$$

where Y_n is given by (5.17). By Lemmas 5.2 and 5.3

$$|\tilde{\alpha}_M(x,y) - \check{\alpha}_M(x,y)| + |\check{\alpha}_M(x,y) - \bar{\alpha}_M(x,y)| \lesssim \mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, |y-x|\right) + M^{-1/2},$$

leading to

$$\sup_x \left|\tilde{G}_n\psi(x) - G_n\psi(x)\right| \lesssim \delta_n^{-2}\left(\mathbb{E}\left[\mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, \delta_n|\xi|\right)\right] + M^{-1/2}\right) \lesssim \delta_n^{-2}\mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, \delta_n\right) + \delta_n^{-2}M^{-1/2}.$$

Therefore, by the triangle inequality and using (5.36)

$$\begin{aligned}\lim_{n\to\infty}\sup_x \left|\tilde{G}_n\psi(x) - \mathcal{G}\psi(x)\right| &\leq \lim_{n\to\infty}\sup_x \left|\tilde{G}_n\psi(x) - G_n\psi(x)\right| \\ &+ \lim_{n\to\infty}\sup_x |G_n\psi(x) - \mathcal{G}\psi(x)| \lesssim \lim_{n\to\infty}\delta_n^{-2}\mathscr{P}_k^{\beta,\gamma}\left(M_n^{-1}, \delta_n\right) \\ &+ \lim_{n\to\infty}\delta_n^{-2}M_n^{-1/2} = 0,\end{aligned}$$

as required. □

5.4.3 Proofs of Theorems 5.5 and 5.6

We first look at the marginal kernel P_n which, under some assumptions, satisfies a geometric drift condition as in (2.7) with uniform V, b and S. The term λ_n will converge to 1 as $n \to \infty$, but an explicit rate of convergence can be obtained that will prove useful. Before such result, a technical lemma is proved involving the standard Gaussian cumulative distribution function Φ.

Proposition 5.4. *Consider $x \in [0,1]$, then*

$$\left|\exp\left\{\frac{1}{2}x^2\right\}\Phi^c(x) - \left(\frac{1}{2} - \frac{x}{\sqrt{2\pi}} + \frac{x^2}{4}\right)\right| \leq x^3.$$

Proof. Set

$$f(x) := \exp\left\{\frac{1}{2}x^2\right\}\Phi^c(x),$$

a Taylor expansion around 0 gives

$$f(x) = \frac{1}{2} - \frac{x}{\sqrt{2\pi}} + \frac{x^2}{4} + \frac{x^3}{6}f'''(\xi),$$

where $\xi \in (0, x)$. The result is obtained by noting that for $x \in [0, 1]$

$$\begin{aligned}\left|f'''(\xi)\right| &= \left|f(\xi)\left(3\xi + \xi^3\right) - \frac{2+\xi^2}{\sqrt{2\pi}}\right| \leq \frac{1}{2}\exp\left\{\frac{1}{2}x^2\right\}\left(3x + x^3\right) + \frac{2+x^2}{\sqrt{2\pi}} \\ &\leq 6.\end{aligned}$$

□

Lemma 5.6. *Suppose Assumption 5 holds, then for any $n \in \mathbb{N}$ there are $\lambda_n > 0$, $b < \infty$ and $x^* > 0$, such that the kernel P_n satisfies the following condition*

$$P_n V(x) \leq \lambda_n V(x) + b\mathbb{1}(x \in S),$$

where $S = \{x : |x| \leq x^*\}$ and $V(x) = \exp\left\{\frac{\beta}{2}|x|\right\}$. Moreover, as $\delta_n \to 0$

$$\lambda_n = 1 - \frac{\beta^2\delta_n^2}{8} + O\left(\delta_n^3\right).$$

Proof. Let ϕ_n denote the density of a Gaussian distribution with mean 0 and variance δ_n^2. From the proof of Mengersen and Tweedie (1996, Theorem 3.2) there are $x^* > 0$ and $C^* < \infty$, independent of n, such that if $x > x^*$

$$\begin{aligned}\lambda_{x,n} &:= \frac{P_n V(x)}{V(x)} \\ &\leq 1 - \int_0^x \left(1 - \exp\left\{-\frac{\beta}{2}z\right\}\right)^2 \phi_n(dz) + 2\Phi^c\left(\frac{x}{\delta_n}\right) \\ &\leq 1 - \int_0^\infty \left(1 - \exp\left\{-\frac{\beta}{2}z\right\}\right)^2 \phi_n(dz) + 3\Phi^c\left(\frac{x^*}{\delta_n}\right) =: \lambda_n,\end{aligned}$$

and

$$\lambda_{x,n} \leq C^*$$

whenever $0 \leq x \leq x^*$. Hence, due to symmetry,

$$\begin{aligned}P_n V(x) &= \lambda_{x,n} V(x) \\ &\leq V(x)\left(\lambda_n \mathbb{1}(|x| > x^*) + C^* \mathbb{1}(|x| \leq x^*)\right) \\ &\leq \lambda_n V(x) + C^* V(x)\,\mathbb{1}(|x| \leq x^*) \\ &\leq \lambda_n V(x) + C^* V(x^*)\,\mathbb{1}(|x| \leq x^*).\end{aligned}$$

Now, notice that

$$\int_0^\infty \exp\{-sz\}\,\phi_n(dz) = \exp\left\{\frac{1}{2}s^2\delta_n^2\right\}\Phi^c(s\delta_n),$$

implying

$$\lambda_n \leq 1 - \left(\frac{1}{2} - 2\exp\left\{\frac{1}{2}\left(\frac{\beta^2\delta_n^2}{4}\right)\right\}\Phi^c\left(\frac{\beta\delta_n}{2}\right) + \exp\left\{\frac{1}{2}\beta^2\delta_n^2\right\}\Phi^c(\beta\delta_n)\right) + o\left(\delta_n^3\right).$$

Thus, by Proposition 5.4 and taking $\delta_n \leq \beta^{-1}$

$$\begin{aligned}\lambda_n &\leq 1 - \left(\frac{(\beta\delta_n)^2}{4} - \frac{(\beta\delta_n)^2}{8}\right) + O\left(\delta_n^3\right) \\ &= 1 - \frac{\beta^2\delta_n^2}{8} + O\left(\delta_n^3\right).\end{aligned}$$

□

The penalty method kernel $\bar{P}_{M,n}$ will satisfy a similar geometric drift condition as P_n, if M is large enough. In fact, the drift condition is satisfied using the same V, b and S from the previous result.

Lemma 5.7. *Assume P_n satisfies for all $n \in \mathbb{N}$*

$$P_nV(x) \leq \lambda_nV(x) + b\mathbb{1}(x \in S),$$

where $\lambda_n > 0$, $b < \infty$, S is a small set and $V \geq 1$. If $\sup_{x,y}\tau_{x,y}^2 < \infty$, then there exists $C \in (0,\infty)$ such that

$$\bar{P}_{M,n}V(x) \leq \left(\lambda_n + CM^{-1/2}\right)V(x) + b\mathbb{1}(x \in S).$$

Proof. Let $\bar{\tau} := \sup_{x,y}\tau_{x,y}$, applying Corollary 5.3

$$\begin{aligned}\bar{P}_{M,n}V(x) &= \int_{\mathcal{X}} V(y)\,\bar{\alpha}_M(x,y)\,q_n(x,dy) + V(x)\left[1 - \int_{\mathcal{X}}\bar{\alpha}_M(x,y)\,q_n(x,dy)\right] \\ &\leq \int_{\mathcal{X}} V(y)\,\alpha(x,y)\,q_n(x,dy) + V(x)\left[1 - 2\Phi\left(-\frac{\bar{\tau}}{2\sqrt{M}}\right)\int_{\mathcal{X}}\alpha(x,y)\,q_n(x,dy)\right] \\ &\leq P_nV(x) + V(x)\left[1 - 2\Phi^c\left(\frac{\bar{\tau}}{2\sqrt{M}}\right)\right]\int_{\mathcal{X}}\alpha(x,y)\,q_n(x,dy) \\ &\leq \lambda_nV(x) + b\mathbb{1}(x \in S) + V(x)\left[1 - 2\Phi^c\left(\frac{\bar{\tau}}{2\sqrt{M}}\right)\right] \\ &= \left[\lambda_n + 1 - 2\Phi^c\left(\frac{\bar{\tau}}{2\sqrt{M}}\right)\right]V(x) + b\mathbb{1}\{x \in S\}.\end{aligned}$$

The result follows since

$$0 \leq 1 - 2\Phi^c\left(\frac{\bar{\tau}}{2\sqrt{M}}\right) \leq \frac{\bar{\tau}}{\sqrt{2\pi M}}. \qquad \square$$

Using the above lemmas, we can now present the proofs of Theorems 5.5 and 5.6.

Proof of Theorem 5.5. First we show that the kernel $\tilde{P}_{M,n}$ satisfies a similar drift to the one satisfied by P_n in Lemma 5.6. Notice that V is log-Lipschitz, therefore from Lemma 5.2

$$\begin{aligned}
\frac{\left|\check{P}_{M,n}V(x) - \bar{P}_{M,n}V(x)\right|}{V(x)} &= \left|\mathbb{E}\left[\left(\frac{V(Y_n)}{V(x)} - 1\right)\left(\check{\alpha}_M(x, Y_n) - \bar{\alpha}_M(x, Y_n)\right)\right]\right| \\
&\lesssim M^{-1/2}\mathbb{E}\left|\exp\left\{\log\left(\frac{V(Y_n)}{V(x)}\right)\right\} - 1\right| \lesssim M^{-1/2}\mathbb{E}\left[\exp\left\{C\left|Y_n - x\right|\right\} - 1\right] \\
&\lesssim M^{-1/2}\mathbb{E}\exp\left\{C\delta_1\left|\xi\right|\right\} \lesssim M^{-1/2}.
\end{aligned}$$

Now, by Lemma 5.3

$$\begin{aligned}
\frac{\left|\tilde{P}_{M,n}V(x) - \check{P}_{M,n}V(x)\right|}{V(x)} &= \left|\mathbb{E}\left[\left(\frac{V(Y_n)}{V(x)} - 1\right)\left(\check{\alpha}_M(x, Y_n) - \bar{\alpha}_M(x, Y_n)\right)\right]\right| \\
&\lesssim \mathbb{E}\left|\left(\exp\left\{\log\left(\frac{V(Y_n)}{V(x)}\right)\right\} - 1\right)\left(\mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, \left|Y_n - x\right|\right) + M^{-3/2}\right)\right| \\
&\lesssim \mathbb{E}\left[\exp\left\{C\delta_1\left|\xi\right|\right\}\left(\mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, \delta_n\left|\xi\right|\right) + M^{-3/2}\right)\right] \\
&\lesssim \mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, \delta_n\right) + M^{-3/2}.
\end{aligned}$$

Thus, by the triangle inequality and Lemmas 5.6 and 5.7, there exists $C > 0$ such that

$$\begin{aligned}
\tilde{P}_{M,n}V(x) &\leq \check{P}_{M,n}V(x) + C\left(\mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, \delta_n\right) + M^{-3/2}\right)V(x) \\
&\leq \bar{P}_{M,n}V(x) + C\left(\mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, \delta_n\right) + M^{-1/2}\right)V(x) \\
&\leq \left(\lambda_n + C\left(\mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, \delta_n\right) + M^{-1/2}\right)\right)V(x) + b\mathbb{1}\left\{x \in S\right\} \\
&\leq \left(1 - \frac{\beta^2\delta_n^2}{8} + C\left(\delta_n^3 + \mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, \delta_n\right) + M^{-1/2}\right)\right)V(x) + b\mathbb{1}\left\{x \in S\right\}.
\end{aligned}$$

By Assumption 2 and taking δ_n sufficiently small, there exists $M_1(n) \in \mathbb{N}$ such that

$$\sup_{M \geq M_1} \tilde{\lambda}_{M,n} := 1 - \frac{\beta^2 \delta_n^2}{8} + C\left(\delta_n^3 + \mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, \delta_n\right) + M^{-1/2}\right) < 1.$$

What is left to show is the smallness of the set $S = \{x : |x| \leq x^*\}$, which appears in Lemma 5.6, for the class of kernels $\left\{P_n, \tilde{P}_{M_1,n}, \tilde{P}_{M_1+1,n}, \dots\right\}$. Take $A \subseteq S$, by Lemma 5.4 there exists $C > 0$ such that

$$\begin{aligned}
\tilde{P}_{M,n}(x, A) &\geq \int_A \tilde{\alpha}_M(x, y)\, q_n(x, dy) \\
&\gtrsim \int_A \alpha(x, y) \exp\left\{-C\mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, |y - x|\right)\right\} q_n(x, dy) \\
&\geq \exp\left\{-C \sup_{x,y \in S} \mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, |y - x|\right)\right\} \int_A \alpha(x, y)\, q_n(x, dy) \\
&\geq \exp\left\{-C\mathscr{P}_k^{\beta,\gamma}(1, 2x^*)\right\} \int_A \alpha(x, y)\, q_n(x, dy) \\
&\gtrsim \int_A \alpha(x, y)\, q_n(x, dy).
\end{aligned}$$

Finally, by Theorem 2.4, the conditions for π and q_n imply S is small for P_n, and from Remark 2.3, the minorisation condition is attained by the sub-kernel

$$\alpha(x, y)\, q_n(x, dy),$$

which completes the proof. □

Proof of Theorem 5.6. First notice

$$\begin{aligned}
\|\tilde{P}_{M,\delta_n}(x, \cdot) - P_{\delta_n}(x, \cdot)\|_{TV} &= \sup_{A \in \mathcal{B}(\mathcal{X})} \left|\tilde{P}_{M,\delta_n}(x, A) - P_{\delta_n}(x, A)\right| \\
&\leq \sup_{A \in \mathcal{B}(\mathcal{X})} \left|\int_A \left(\tilde{\alpha}_M(x, y) - \alpha(x, y)\right) q_n(x, dy)\right| \\
&\quad + \sup_{A \in \mathcal{B}(\mathcal{X})} \mathbb{1}(x \in A) \left|\int_{\mathcal{X}} \left(\alpha(x, y) - \tilde{\alpha}_M(x, y)\right) q_n(x, dy)\right| \\
&\leq 2 \int_{\mathcal{X}} \left|\tilde{\alpha}_M(x, y) - \alpha(x, y)\right| q_n(x, dy) \\
&\lesssim \int_{\mathcal{X}} \left|\tilde{\alpha}_M(x, y) - \bar{\alpha}_M(x, y)\right| q_n(x, dy) + \int_{\mathcal{X}} \left|\bar{\alpha}_M(x, y) - \alpha(x, y)\right| q_n(x, dy).
\end{aligned}$$

Then, by Lemmas 5.2 and 5.3

$$\|\tilde{P}_{M,\delta_n}(x,\cdot) - P_{\delta_n}(x,\cdot)\|_{TV} \lesssim \mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, \int_{\mathcal{X}} |y-x|\, q_n(x,dy)\right) + M^{-3/2}$$
$$+ M^{-1/2} + 1 - 2\Phi\left(-\frac{\tau_{x,y}}{2\sqrt{M}}\right) \lesssim \mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, \delta_n\right) + M^{-1/2}.$$

Consider $r_n(M)^{-1} \propto \mathscr{P}_k^{\beta,\gamma}\left(M^{-1}, \delta_n\right) + M^{-1/2}$, the right hand side of (5.19) is minimised either at $\left\lfloor k^*_{n,M} \right\rfloor$ or $\left\lceil k^*_{n,M} \right\rceil$, where

$$k^*_{n,M} := \frac{\log\left(r_n(M)\, R_n \log\left(\tau_n^{-1}\right)\right)}{\log\left(\tau_n^{-1}\right)}.$$

This implies,

$$\begin{aligned}\|\tilde{\pi}_{M,n}(\cdot) - \pi(\cdot)\|_{TV} &\leq R_n \tau_n^{k^*_{n,M}+1} + \frac{k^*_{n,M}+1}{r_n(M)} \\ &\leq \frac{1 + \log\left(r_n(M)\log\left(\tau_n^{-1}\right) R_n \tau_n^{-1}\right)}{r_n(M)\log\left(\tau_n^{-1}\right)}.\end{aligned}$$

Notice that the variables $\{R_n, R_{n+1}, \dots\}$ can be taken in such way that

$$\inf_n \log(R_n) > 0,$$

hence the result follows using (5.20). □

Chapter 6

Final Discussion

In this thesis some fundamental stability properties and examples of approximate MCMC algorithms were explored. The approximate Markov kernels considered are perturbed MH kernels defined by a collection of state-dependent distributions. General results for the noisy MH algorithm, involving random weights with expectation 1, were covered in Chapter 3. There, we do not assume a specific form for these weights, which can be simple arithmetic averages or more complex random variables as discussed in Chapter 4. The former may arise when unbiased importance sampling estimates of a target density are used (Section 4.1), while the latter may arise when such densities are estimated unbiasedly using a particle filter (Section 4.2). Additionally, in an attempt of correcting the noisy MH algorithm, a novel method was studied in Chapter 5. This is presented in Algorithm 5.1 and is based on the penalty method, the latter belonging to a class of exact algorithms with randomised acceptance probabilities (Section 5.1).

In Chapter 3, two different sets of sufficient conditions were provided under which the noisy MH chain inherits geometric ergodicity from the marginal chain. The first pair of conditions, W1 and W2, involve a stronger version of the Law of Large Numbers for the weights and uniform convergence of the first negative moment, respectively. For the second set, W1 is still required but W2 can be replaced with P1*, which imposes a condition on the proposal distribution. These conditions also imply simultaneous geometric ergodicity of a sequence of noisy Markov kernels together with the marginal Markov kernel, which then ensures that the noisy invariant $\tilde{\pi}_N$ converges to π in total variation as N increases (Theorem 3.3). Moreover, an explicit bound for the rate of convergence between $\tilde{\pi}_N$ and π is possible whenever an explicit bound is available for the convergence between $\tilde{P}_N(x, \cdot)$ and $P(x, \cdot)$ (Theorem 3.4).

Specific conditions for inheriting geometric ergodicity from the corresponding marginal chain were given in Chapter 4 for the cases when the weights are arithmetic averages as in (2.24), or when they arise from a particle filter as in (2.25). In the arithmetic averages context, the uniform integrability condition in W3 ensures that W1 is satisfied, whereas W4 is essential for satisfying W2. Regarding the noisy invariant distribution $\tilde{\pi}_N$, W0* (which implies W1 and W2) leads to an explicit bound on the rate of convergence of the difference between $\tilde{\pi}_N$ and π (Proposition 4.9). For the particle filter case, the bounded condition in W5 is enough for satisfying W1, although it may be difficult to verify in practice. A simpler but fairly stronger condition is given in W5*, essentially requiring the parameter and hidden state spaces to be compact.

The Approximate Penalty Method (APM) was studied in Chapter 5, providing some applications (Section 5.3) and stability properties (Section 5.4). The method is based on geometric averages of estimators $R_{x,y}$ of the usual MCMC acceptance ratio $r(x, y)$, and introduces a correction in an attempt to mimic the exact penalty method. The examples illustrate the possible potential of the algorithm in some scenarios, including the noisy MH setting or when dealing with sub-samples for large data sets. The theoretical results mainly deal with convergence of the APM chain towards a Langevin diffusion and simultaneous geometric ergodicity in tandem with the marginal chain.

6.1 Open Questions

- The noisy MH algorithm remains undefined when the weights have positive probability of being zero. If both weights were zero one could accept the move, reject the move or keep sampling new weights until one of them is not zero. Each of these lead to different behaviour.

- As seen in the examples of Section 4.1.1, the behaviour of the ratio of the weights (at least in the tails of the target) plays an important role in the ergodic properties of the noisy chain. It seems plausible to obtain geometric noisy chains, even when the marginal is not, if the ratio of the weights decays sufficiently fast to zero in the tails.

- As noted also in Nicholls et al. (2012), the further exploration of exact algorithms with randomised acceptance could lead to novel and useful approaches. The penalty method and Barker's algorithm are only two possibilities.

- The first two examples of Section 5.3 implement APM in the noisy MH context,

when the weights are homogeneous or log-normally distributed. It remains an open question how to fulfil these conditions in more realistic, practical problems. In this respect, a connection to the work done in Bérard et al. (2014) may be investigated. Additionally, for the sub-sampling setting, more analysis on the trade-off between M (the number of estimators used in the geometric average) and k (the size of the sub-sample) is required.

- The theoretical results of Section 5.4 are only proved for the one dimensional case, and under fairly strong assumptions on the target π and the noise coming from the estimators $R_{x,y}$. It seems plausible to relax the stated assumptions in order to obtain similar results for more general targets on $\mathbb{R}^d$.

Appendix A

Variance Bounding, Geometric Ergodicity and Spectral Gaps

Definition. A Markov chain with transition kernel P and unique stationary distribution π is variance bounding if

$$\sup_{g:\pi(g)=0,\pi(g^2)=1} \sigma_{g,P}^2 < \infty,$$

where $\sigma_{g,P}^2$ denotes the asymptotic variance given by (2.5).

For π-reversible Markov chains, variance bounding is the weakest property that guarantees the existence of a Central Limit Theorem for ergodic averages of any function in $L^2(\pi) := \left\{f : \pi\left(f^2\right) < \infty\right\}$ (see e.g. Roberts and Rosenthal, 2008, Theorem 7). Under reversibility, variance bounding and geometric ergodicity follow a close relationship, this is easily seen when looking at Markov kernels as operators on function spaces.

A π-reversible Markov kernel $P : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ can be regarded as a self-adjoint linear operator acting on $L^2(\pi)$. The typical inner-product in $L^2(\pi)$ is given by

$$\langle f, g\rangle = \int_{\mathcal{X}} f(x)\, g(x)\, \pi(dx),$$

which induces the norm $\|f\|_{L^2}^2 = \langle f, f\rangle$. Define the Dirichlet form of P, for $g \in L^2(\pi)$, as follows

$$\mathcal{E}_P(g) := \langle g, (I - P)\, g\rangle = \frac{1}{2}\int \pi(dx)\, P(x, dy)\, [g(x) - g(y)]^2.$$

The right spectral gap of P is given by

$$\mathrm{Gap}_R(P) := \inf_{g:\pi(g)=0,\pi(g^2)=1} \mathcal{E}_P(g),$$

which can be useful for assessing the variance bounding property of reversible Markov chains, as stated below.

Theorem (Roberts and Rosenthal, 2008, Theorem 14)**.** *A π-reversible Markov chain with transition kernel P is variance bounding if and only if $Gap_R(P) > 0$.*

Additionally, the left spectral gap of P is given by

$$\mathrm{Gap}_L(P) := \inf_{g:\pi(g)=0,\pi(g^2)=1} (2 - \mathcal{E}_P(g)),$$

allowing us to define the absolute spectral gap of P as follows

$$\mathrm{Gap}(P) := \min\{\mathrm{Gap}_R(P), \mathrm{Gap}_L(P)\}.$$

The following result relates geometric ergodicity with the existence of an absolute spectral gap for reversible Markov chains, see e.g. Roberts and Tweedie (2001, Theorem 2) together with Roberts and Rosenthal (1997, Theorem 2.1) or Kontoyiannis and Meyn (2012, Proposition 1.2) for a proof.

Theorem A.1. *A π-reversible, φ-irreducible and aperiodic Markov chain with transition kernel P is π-a.e. geometrically ergodic if and only if $Gap(P) > 0$.*

Appendix B

Some Properties of the Beta Function

Proposition B.1. *Let $B(x,y)$ denote the beta function, then for $N \in \mathbb{N}$*

$$B(N,N) = O\left(\frac{1}{2^{2N}\sqrt{N}}\right).$$

Proof. Stirling's approximation for the factorial implies

$$N! = O\left(N^{N+1/2}e^{-N}\right).$$

Hence, the beta function satisfies

$$\begin{aligned} B(N,N) &= \frac{2(N!)^2}{N(2N)!} = O\left(\frac{N^{2N+1}e^{-2N}}{N(2N)^{2N+1/2}e^{-2N}}\right) \\ &= O\left(\frac{1}{2^{2N}\sqrt{N}}\right). \end{aligned}$$

□

Proposition B.2. *Let $B(x,y)$ denote the beta function, and let $I_z(x,y)$ be the regularised incomplete beta function which is given by*

$$I_z(x,y) = \frac{\int_0^z v^{x-1}(1-v)^{y-1}\,dv}{B(x,y)}, \quad \text{for } z \in [0,1].$$

For $N \in \mathbb{N}$ and $z < \frac{1}{2}$,

$$\left.\begin{array}{r} I_z(N,N) \\ I_z(N+1,N-1) \\ I_z(N-1,N+1) \end{array}\right\} = o\left(N^{-1}\right).$$

Proof. Notice that for $z < \frac{1}{2}$

$$\int_0^z v^{N-1}(1-v)^{N-1}\,dv \leq z^N(1-z)^{N-1}$$

and

$$B(N,N) \geq \left(\frac{1}{2}-z\right)\left(\frac{1+2z}{4}\right)^{N-1}\left(\frac{3-2z}{4}\right)^{N-1},$$

implying

$$\begin{aligned} I_z(N,N) &\leq \frac{z^N(1-z)^{N-1}}{\left(\frac{1}{2}-z\right)\left(\frac{1+2z}{4}\right)^{N-1}\left(\frac{3-2z}{4}\right)^{N-1}} \\ &\leq O\left(\left(\frac{16z(1-z)}{(1+2z)(3-2z)}\right)^N\right). \end{aligned}$$

Since

$$\rho_z := \frac{16z(1-z)}{(1+2z)(3-2z)} < 1$$

if $z \neq \frac{1}{2}$, then

$$I_z(N,N) \leq O\left(e^{N\log(\rho_z)}\right).$$

Now, using integration by parts

$$\begin{aligned} I_z(N,N) &= I_z(N+1,N-1) + \frac{z^N(1-z)^{N-1}}{NB(N,N)} \\ &= I_z(N-1,N+1) - \frac{z^{N-1}(1-z)^N}{NB(N,N)}. \end{aligned}$$

Therefore, what is left to prove is

$$\frac{z^N(1-z)^N}{NB(N,N)} = o\left(N^{-1}\right),$$

which is immediate using Proposition B.1 and the fact that $4z^2 - 4z + 1 > 0$ if $z \neq \frac{1}{2}$. □

Appendix C

Skorokhod topology on $\mathbb{D}\left([0,T],\mathbb{R}\right)$

Let $\mathbb{D}\left([0,T],\mathbb{R}\right)$ be the space of càdlàg (right continuous with limits on the left) functions from the closed interval $[0,T]$ into $\mathbb{R}$. This space of functions with "jumps" is convenient for the study of various phenomena and its probabilistic importance dates back to Skorokhod (1956). An appropriate metric on this space should be able to deal with the difficulties arising when considering discontinuous functions. Consider $x \in \mathbb{D}\left([0,T],\mathbb{R}\right)$, the uniform metric defined in terms of the uniform norm

$$\|x\|_T := \sup_{0\leq t\leq T}\left\{|x(t)|\right\},$$

works well on the subspace of continuous functions. However, when discontinuities are present we should not insist on jumps occurring exactly at the same time in order for two functions to be close.

Among the different Skorokhod's topologies introduced in Skorokhod (1956), J_1 is the most appropriate for studying $\mathbb{D}\left([0,T],\mathbb{R}\right)$ since it allows small perturbations of time when addressing the closeness of two functions.

Definition. Let Λ denote the class of strictly increasing homeomorphisms on $[0,T]$. We say that a sequence of functions $\{x_n\}$ on $\mathbb{D}\left([0,T],\mathbb{R}\right)$ converges to $x \in \mathbb{D}\left([0,T],\mathbb{R}\right)$ as $n\to\infty$, in the Skorokhod (J_1) topology, if

$$d_{J_1}(x_n,x) := \inf_{\lambda\in\Lambda}\max\left\{\|\lambda - I\|_T, \|x_n\circ\lambda - x\|_T\right\} \to 0,$$

where I is the identity map.

Remark. The space $\mathbb{D}\left([0,T],\mathbb{R}\right)$ is separable but not complete under the metric

d_{J_1}. In Billingsley (1968), an equivalent metric was introduced that induces the Skorokhod topology and makes $\mathbb{D}([0,T],\mathbb{R})$ a Polish space.

Bibliography

P. Alquier, N. Friel, R. Everitt, and A. Boland. Noisy Monte Carlo: convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, pages 1–19, 2014.

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2):697–725, 04 2009.

C. Andrieu and M. Vihola. Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Probab.*, 25(2):1030–1077, 04 2015.

C. Andrieu and M. Vihola. Establishing some order amongst exact approximations of MCMCs. *Ann. Appl. Probab.*, 26(5):2661–2696, 10 2016.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 72(3):pp. 269–342, 2010.

R. Bardenet, A. Doucet, and C. Holmes. On Markov chain Monte Carlo methods for tall data. *ArXiv e-prints*, May 2015. URL `http://arxiv.org/abs/1505.02827v1`.

A. A. Barker. Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics*, 18:119, Apr. 1965.

M. A. Beaumont. Estimation of Population Growth or Decline in Genetically Monitored Populations. *Genetics*, 164(3):1139–1160, 2003.

M. Bédard and J. S. Rosenthal. Optimal scaling of Metropolis algorithms: Heading toward general target distributions. *Canadian Journal of Statistics*, 36(4):483–503, 2008.

J. Bérard, P. Del Moral, and A. Doucet. A lognormal central limit theorem for particle approximations of normalizing constants. *Electron. J. Probab.*, 19:28 pp., 2014.

A. Beskos, G. Roberts, A. Thiery, and N. Pillai. Asymptotic Analysis of the Random-Walk Metropolis Algorithm on Ridged Densities. *ArXiv e-prints*, Oct. 2015. URL `http://arxiv.org/abs/1510.02577v1`.

P. Billingsley. *Convergence of probability measures*. Wiley series in probability and mathematical statistics. Wiley, New York, 1968.

L. Breyer, G. O. Roberts, and J. S. Rosenthal. A note on geometric ergodicity and floating-point roundoff error. *Statistics & Probability Letters*, 53(2):123 – 127, 2001.

H. Callaert and J. Keilson. On exponential ergodicity and spectral structure for birth-death processes, II. *Stochastic Processes and their Applications*, 1(3):217 – 235, 1973.

O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer, 2005.

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

F. Cérou, P. Del Moral, and A. Guyader. A nonasymptotic theorem for unnormalized Feynman-Kac particle models. *Ann. Inst. H. Poincaré Probab. Statist.*, 47(3): 629–649, 08 2011.

K. S. Chan and C. J. Geyer. Discussion: Markov Chains for Exploring Posterior Distributions. *Ann. Statist.*, 22(4):1747–1758, 12 1994.

T. K. Chandra. Uniform Integrability in the Cesàro Sense and the Weak Law of Large Numbers. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 51(3):pp. 309–317, 1989.

N. Chopin. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Ann. Statist.*, 32(6):2385–2411, 12 2004.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages n/a–n/a, 2016.

P. Del Moral. *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Probability and Its Applications. Springer, 2004.

G. Deligiannidis and A. Lee. Which ergodic averages have finite asymptotic variance? *ArXiv e-prints*, June 2016. URL `http://arxiv.org/abs/1606.08373v2`.

G. Deligiannidis, A. Doucet, and M. K. Pitt. The Correlated Pseudo-Marginal Method. *ArXiv e-prints*, Nov. 2015. URL `http://arxiv.org/abs/1511.04992v3`.

R. Douc and O. Cappé. Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pages 64–69, Sept 2005.

R. Douc, E. Moulines, and J. Olsson. Long-term stability of sequential Monte Carlo methods under verifiable conditions. *Ann. Appl. Probab.*, 24(5):1767–1802, 10 2014.

A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009.

A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Information Science and Statistics. Springer New York, 2001.

A. Doucet, M. K. Pitt, G. Deligiannidis, and R. Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313, 2015.

A. Durmus and E. Moulines. Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm. *ArXiv e-prints*, July 2015. URL `http://arxiv.org/abs/1507.05021v3`.

S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 2008.

J. Fernández-Villaverde and J. F. Rubio-Ramírez. Estimating Macroeconomic Models: A Likelihood Approach. *The Review of Economic Studies*, 74(4):1059–1087, 2007.

D. Ferré, L. Hervé, and J. Ledoux. Regular perturbation of V-geometrically ergodic Markov chains. *J. Appl. Probab.*, 50(1):184–194, 03 2013.

N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140: 107–113(6), April 1993.

A. Gut. *Probability: A Graduate Course*. Springer Texts in Statistics. Springer New York, 2013.

M. Hairer and J. C. Mattingly. Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, pages 109–117. Springer, 2011.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

J. E. Johndrow, J. C. Mattingly, S. Mukherjee, and D. Dunson. Approximations of Markov Chains and Bayesian Inference. *ArXiv e-prints*, Aug. 2015. URL `http://arxiv.org/abs/1508.03387v2`.

R. C. Jun S. Liu. Sequential Monte Carlo Methods for Dynamic Systems. *Journal of the American Statistical Association*, 93(443):1032–1044, 1998.

R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

A. Khuri and G. Casella. The Existence of the First Negative Moment Revisited. *The American Statistician*, 56(1):pp. 44–47, 2002.

I. Kontoyiannis and S. P. Meyn. Geometric ergodicity and the spectral gap of non-reversible Markov chains. *Probability Theory and Related Fields*, 154(1):327–339, 2012.

A. Korattikara, Y. Chen, and M. Welling. Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 181–189. JMLR Workshop and Conference Proceedings, 2014. URL `http://jmlr.org/proceedings/papers/v32/korattikara14.pdf`.

K. Łatuszyński and G. O. Roberts. CLTs and Asymptotic Variance of Time-Sampled Markov Chains. *Methodology and Computing in Applied Probability*, 15(1):237–247, 2013.

A. Lee and K. Łatuszyński. Variance bounding and geometric ergodicity of Markov chain Monte Carlo kernels for approximate Bayesian computation. *Biometrika*, 2014.

A. Lee and N. Whiteley. Variance estimation in the particle filter. *ArXiv e-prints*, Sept. 2015. URL `http://arxiv.org/abs/1509.00394v2`.

T. Li, M. Bolic, and P. M. Djuric. Resampling Methods for Particle Filtering: Classification, implementation, and strategies. *IEEE Signal Processing Magazine*, 32(3):70–86, May 2015.

T. Lindvall. *Lectures on the Coupling Method*. Dover Books on Mathematics Series. Dover Publications, Incorporated, 2002.

A.-M. Lyne, M. Girolami, Y. Atchadé, H. Strathmann, and D. Simpson. On Russian Roulette Estimates for Bayesian Inference with Doubly-Intractable Likelihoods. *Statist. Sci.*, 30(4):443–467, 11 2015.

F. Maire, R. Douc, and J. Olsson. Comparison of asymptotic variances of inhomogeneous Markov chains with application to Markov chain Monte Carlo methods. *Ann. Statist.*, 42(4):1483–1510, 08 2014.

F. Maire, N. Friel, and P. Alquier. Light and Widely Applicable MCMC: Approximate Bayesian Inference for Large Datasets. *ArXiv e-prints*, Mar. 2015. URL `http://arxiv.org/abs/1503.04178v2`.

J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

J. C. Mattingly, N. S. Pillai, and A. M. Stuart. Diffusion limits of the random walk Metropolis algorithm in high dimensions. *Ann. Appl. Probab.*, 22(3):881–930, 06 2012.

T. J. McKinley, J. V. Ross, R. Deardon, and A. R. Cook. Simulation-based Bayesian inference for epidemic models. *Computational Statistics & Data Analysis*, 71(0): 434 – 447, 2014.

F. J. Medina-Aguayo, A. Lee, and G. O. Roberts. Stability of noisy Metropolis–Hastings. *Statistics and Computing*, 26(6):1187–1211, 2016.

K. L. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24(1):101–121, 02 1996.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

S. P. Meyn and R. L. Tweedie. Computable Bounds for Geometric Convergence Rates of Markov Chains. *Ann. Appl. Probab.*, 4(4):981–1011, 11 1994.

A. Y. Mitrophanov. Sensitivity and convergence of uniformly ergodic Markov chains. *J. Appl. Probab.*, 42(4):1003–1014, 12 2005.

G. K. Nicholls, C. Fox, and A. Muir Watt. Coupled MCMC with a randomized acceptance probability. *ArXiv e-prints*, May 2012. URL `http://arxiv.org/abs/1205.6857v1`.

J. Norris. *Markov Chains*. Number no. 2008 in Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 1999.

P. D. O'Neill, D. J. Balding, N. G. Becker, M. Eerola, and D. Mollison. Analyses of Infectious Disease Data from Household Outbreaks by Markov chain Monte Carlo Methods. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 49(4):pp. 517–542, 2000.

P. H. Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973.

W. W. Piegorsch and G. Casella. The Existence of the First Negative Moment. *The American Statistician*, 39(1):pp. 60–62, 1985.

N. S. Pillai and A. Smith. Ergodicity of Approximate MCMC Chains with Applications to Large Data Sets. *ArXiv e-prints*, May 2014. URL `http://arxiv.org/abs/1405.0182v2`.

M. K. Pitt, R. dos Santos Silva, P. Giordani, and R. Kohn. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134 – 151, 2012. Bayesian Models, Methods and Applications.

M. Quiroz, M. Villani, and R. Kohn. Speeding Up MCMC by Efficient Data Subsampling. *ArXiv e-prints*, Apr. 2014. URL `http://arxiv.org/abs/1404.4178v4`.

C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer New York, 2013.

G. Roberts and J. Rosenthal. Geometric Ergodicity and Hybrid Markov Chains. *Electron. Commun. Probab.*, 2:no. 2, 13–25, 1997.

G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, 16(4):351–367, 11 2001.

G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

G. O. Roberts and J. S. Rosenthal. Variance bounding Markov chains. *Ann. Appl. Probab.*, 18(3):1201–1214, 06 2008.

G. O. Roberts and R. L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996a.

G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 12 1996b.

G. O. Roberts and R. L. Tweedie. Geometric L^2 and L^1 convergence are equivalent for reversible Markov chains. *J. Appl. Probab.*, 38A:37–41, 02 2001.

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120, 02 1997.

G. O. Roberts, J. S. Rosenthal, and P. O. Schwartz. Convergence properties of perturbed Markov chains. *J. Appl. Probab.*, 35(1):1–11, 03 1998.

J. S. Rosenthal. Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):pp. 558–566, 1995.

D. Rudolf and N. Schweizer. Perturbation theory for Markov chains via Wasserstein distance. *ArXiv e-prints*, Mar. 2015. URL `http://arxiv.org/abs/1503.04123v2`.

T. Shardlow and A. M. Stuart. A Perturbation Theory for Ergodic Markov Chains and Application to Numerical Approximations. *SIAM Journal on Numerical Analysis*, 37(4):1120–1137, 2000.

C. Sherlock, A. H. Thiery, G. O. Roberts, and J. S. Rosenthal. On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.*, 43(1): 238–275, 02 2015.

S. Singh, M. Wick, and A. McCallum. Monte Carlo MCMC: efficient inference by sampling factors. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 111–115. Association for Computational Linguistics, 2012. URL `http://www.aclweb.org/anthology/W12-3021`.

A. V. Skorokhod. Limit Theorems for Stochastic Processes. *Theory of Probability & Its Applications*, 1(3):261–290, 1956.

Y. W. Teh, A. H. Thiery, and S. J. Vollmer. Consistency and Fluctuations For Stochastic Gradient Langevin Dynamics. *Journal of Machine Learning Research*, 17(7):1–33, 2016.

H. Thorisson. *Coupling, Stationarity, and Regeneration*. Probability and Its Applications. Springer New York, 2013.

L. Tierney. Markov Chains for Exploring Posterior Distributions. *Ann. Statist.*, 22 (4):1701–1728, 12 1994.

N. Whiteley. Stability properties of some particle filters. *Ann. Appl. Probab.*, 23(6): 2500–2537, 12 2013.

S. Yıldırım. On the Use of Penalty MCMC for Differential Privacy. *ArXiv e-prints*, Apr. 2016. URL `http://arxiv.org/abs/1604.07177v1`.