OPTIMAL PROPOSALS FOR MCMC METHODS

Alexandre Thiéry
Under the supervision of G.O. Roberts and A.M. Stuart

1. Introduction

As A. Sokal states it at the beginning of its lectures on Monte Carlo methods [Sok96],

> "Monte Carlo is an extremely bad method; it should be used only when all alternative methods are worse."

Indeed, Monte Carlo methods are based on the Central Limit theorem so that a statistical error of order

$$\text{error} = \mathcal{O}\Big(\frac{\sigma}{\sqrt{\text{computational budget}}} \Big)$$

is unavoidable: at best, the constant σ can be improved; in many situations, especially for high dimensional problems, there is no better alternative. Any introductory text on Monte-Carlo methods describes dozens of area of science where these ideas are at work, and where there is no known way to tackle these problems more efficiently: see [Dia09] for an introduction and [Liu08, RC04] for book-length treatments.

Suppose that we are given a probability distribution π on a state space S and want to compute the expected value $\mathbb{E}_\pi[\varphi]$ of an observable $\varphi : S \to \mathbb{R}$. The basic Monte Carlo approach consists in sampling independent copies $X_1, \ldots, X_N$ from π and then to take the average. Nevertheless, except in very simple cases, the distribution π is impossible to sample directly from and is generally only known up to a normalization constant. To remedy to this problem, the Markov Chain Monte Carlo (MCMC) algorithm uses a Markov chain $x = \big\{x(k)\big\}_{k \geqslant 0}$ on S that has π as invariant measure: the ergodic theorem for Markov chains shows that under mild assumptions [MTH93] the quantity

$$I(\varphi, N) = \frac{1}{N}\Big(\varphi(x(1)) + \ldots + \varphi(x(N)) \Big)$$

is a consistent estimator of $\mathbb{E}_\pi[\varphi]$. In the 1990s, Tierney [Tie94] is one of the first paper to carefully laid out the assumptions needed to analyse MCMC algorithms and their properties, in particular, convergence of ergodic averages and central limit theorems: the more recent reference [MTH93] is dedicated to these topics. Maybe surprisingly, the construction of Markov chains that let π invariant is generally not difficult. They are even so many different ways of building such Markov chains that choosing an efficient candidate is often a delicate exercise: this is one of the main themes of this report.

1.1. **Notations.** In this text, sequence of Markov chains living in spaces of different dimensions are considered. A Markov chain $x^d = \{x^d(k)\}_{k\geqslant 0}$ evolves in a space of dimension d (typically $\mathbb{R}^d$) and its coordinates are written as

$$x^d(k) = \big(x_1^d(k), x_2^d(k), \dots, x_d^d(k)\big) \in \mathbb{R}^d.$$

The notations $\alpha_n \lesssim \beta_n$ and $\alpha_n \asymp \beta_n$ indicate that there exists a universal constant $K > 0$ satisfying $\alpha_n \leqslant K\beta_n$ and $K^{-1}\beta_n \leqslant \alpha_n \leqslant K\beta_n$ respectively.

2. Metropolis Hastings algorithm, a brief review

The probability distributions $\pi(x) = \frac{1}{Z}\tilde{\pi}(x)$ that we are interested in are typically known up to a normalization constant Z so that only the unormalized density $\tilde{\pi}$ is generally available:

- in statistical physics, an Hamiltonian $H : S \to \mathbb{R}$ is given and the associated Bolztman distribution $\pi_\beta(x) \propto \exp\{-\beta H(x)\}$ is used to describe the system at inverse temperature β,
- in Bayesian statistics, the posterior distribution $\mathbb{P}(\theta|y) \propto \mathbb{P}(y|\theta)\,\pi_0(\theta)$ is again defined up to a normalization constant.

A Markov kernel $\tilde{T}(x, y)$ is reversible with respect to the measure π if it satisfies the detailed balanced equations $\pi(x)\tilde{T}(x, y) = \pi(y)\tilde{T}(y, x)$: this implies in particular that π is an invariant distribution of the Markov kernel. The idea of the Metropolis-Hastings algorithm in order to construct a Markov chain $\big\{x(k)\big\}_{k\geqslant 0}$ that let π invariant is to take an arbitrary Markov chain with transition kernel T and to modify it in such a way that the detailed balanced equations (for the target distribution π) are enforced: the transition kernel of the Markov chain $\big\{x(k)\big\}_{k\geqslant 0}$ is defined by $\tilde{T}(x, y) = T(x, y)\alpha(x, y)$ where $\alpha(x, y) \in [0, 1]$ for $x \neq y$ is an appropriate holding probability. The detailed balance equations are satisfied if $\pi(x)\tilde{T}(x, y) = \pi(y)\tilde{T}(y, x)$ for any $x \neq y$, which also reads $\alpha(x, y) = \frac{\pi(y)T(y,x)}{\pi(x)T(x,y)}\alpha(y, x)$. Because $0 \leqslant \alpha(y, x) \leqslant 1$, this implies that $\alpha(x, y) \leqslant 1 \wedge \frac{\pi(y)T(y,x)}{\pi(x)T(x,y)}$; it can be checked that the choice of holding probabilities

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)T(y, x)}{\pi(x)T(x, y)} \qquad \text{for} \qquad x \neq y \tag{2.1}$$

satisfies the detailed balance equations: the quantity $\alpha(x, y)$ is usually called **Metropolis-Hastings ratio** or **acceptance probability**. The MCMC algorithm, discovered by Metropolis and co-authors in 1953 and described in their celebrated paper [MRR$^+$53] hence proceeds as follows:

(1) if the current position is x, propose a move y according to $T(x, y)$
(2) compute the acceptance probability $\alpha(x, y) := 1 \wedge \frac{\pi(y)T(y,x)}{\pi(x)T(x,y)}$
(3) with probability $\alpha(x, y)$ move from x to y; otherwise stay still
(4) go back to 1.

This mechanism clearly defines a Markov chain that is π-reversible: stability, convergence of ergodic averages and central limit theorems are available [MTH93]. There are thus many different ways of constructing a Markov chain that let π invariant: any reasonable transition kernel T can potentially be a candidate to build a Markov chain $\big\{x(k)\big\}_{k\geqslant 0}$ with transition kernel $\tilde{T}$ that let π invariant.

3. Optimal proposals

In the founding paper [MRR+53], the issue of optimal proposals was already mentioned: proposals of the form $y \overset{\mathcal{D}}{\sim} \text{Uniform}[x-\alpha, x+\alpha]$ were considered, and it was noted that

> "it may be mentioned in this connection that the maximum displacement α must be chosen with some care; if too large, most moves will be forbidden, and if too small, the configuration will not change enough. In either case it will then take longer to come to equilibrium."

The proposal density $T(x, y)$ is crucial to the success of a MCMC algorithm. The most common case involves a symmetric random-walk Metropolis algorithm (RMW) in which the proposal value is given by $y_n = x_n + Z_n$ where the increments Z_n are i.i.d. samples from an easy to simulate symmetric distribution (e.g., $Z_n \overset{\mathcal{D}}{\sim} \text{N}(0, \sigma^2 I_d)$). In this case, the question is how to scale the proposals (e.g., how to choose σ): too small a variance and the chain moves too slowly; too large and the proposals are rejected too often, leading also to poor performances.

Before the 90's, the tuning of the proposals was almost invariably performed by trial and errors. Several rules of thumb [BG93, BGHM95] to select the value of the local variance of the proposals were described, advocating an acceptance rate as high as 70%. It came as a surprise when it was proved in [RGG97] that, under certain assumptions described in the next section, it is optimal to accept a proportion of only 23% of the proposed moves. This means that for optimality the chain must stay still 77% of the time, which might seem counter-intuitive when the goal is to obtain a chain converging fast to its stationary distribution. The seminal result described in [RGG97] is only valid for random walk proposals with a very specific target distribution, and described the behaviour of the algorithm at stationarity: different authors have subsequently tried to relax the different assumptions needed to the application of the 23% rule; [CRR05] describes the transient phase of local Metropolis-Hastings algorithm, Langevin proposals are studied in [RR01, RR98, RT96], tuning of the Hybrid Monte-Carlo algorithm is considered in [BRS09], more general target probabilities are studied in [BR00, NRY07, BR08] and [MPS10] considers Random walk metropolis algorithms in an infinite dimensional framework.

3.1. Optimality criteria.

In this section we describe several efficiency criteria for the comparison of different Markov chains and focus on the so called Expected Squared Jumping Distance (ESJD) that is the benchmark that will be considered in the remaining of this text.

As described in the introduction, the MCMC algorithm is mainly used to compute an expectation $\mathbb{E}_\pi[g]$ through the estimator $I_N(g) = \frac{1}{N}\sum_{k=1}^N g(x(k))$: at stationarity, the variance of this estimator is equal to $\text{Var}[I_N(g)] = \frac{1}{N}\left\{1 + 2\sum_{k=1}^{N-1}(1-\frac{k}{N})\rho_k^{(g)}\right\}\text{Var}[g]$ where $\rho_k = \text{Corr}_\pi(g(x(t))g(x(t+k)))$ is the auto-correlation function of the stationary series $\{g(x(k))\}_k$. If $\sum_k \rho_k^{(g)}$ is summable it comes that

$$\lim_N \; N \cdot \text{Var}[I_N(g)] = \left\{1 + 2\sum_{k=1}^{\infty}\rho_k^{(g)}\right\}\text{Var}[g] = \tau^{(g)}\text{Var}[g],$$

where the quantity $\tau^{(g)} = 1 + 2\sum_{k=1}^{\infty} \rho_k^{(g)}$, usually called the **integrated autocorrelation time**, quantifies how much worse the MCMC estimator is with respect to the case where the $x(k)$ are i.i.d. samples from π: the bigger $\tau^{(g)}$, the larger the variance of the MCMC estimator.

The integrated autocorrelation time $\tau^{(g)}$ depends on the function g so that this is not a very useful efficiency criterion, in general: this is possible to construct two different Markov chains X and Y on the state space S such that $\tau_X^{(g)} < \tau_Y^{(g)}$ and $\tau_X^{(f)} > \tau_Y^{(f)}$ where $f, g : S \to \mathbb{R}$ are two different test functions. The use of the autocorrelation time as an efficiency criterion is thus very limited for the comparison of different MCMC algorithms. Instead, for real valued Markov chains, the first autocorrelation coefficient $\rho_1 = \mathrm{Corr}[x(k), x(k+1)]$ is a simple way of quantifying the dependence structure. Moreover, because for real valued Markov chains we have $\mathbb{E}\big[|x(k+1) - x(k)|^2\big] = 2\mathrm{Var}[x(k)](1 - \rho_1)$, minimizing the first autocorrelation coefficient ρ_1 is equivalent to maximizing the expected squared jumping distance

$$\mathrm{ESJD} = \mathbb{E}[\|x(k+1) - x(k)\|^2].$$

The ESJD is defined the same way for multidimensional Markov chains; this is the usual measure of efficiency that has been studied in the literature [BRS09, RR01, PG09]: indeed, the analytical tractability of the ESJD is one of the main reasons for its use. More sophisticated efficiency criteria such as spectral gaps, mixing time [LPW09] have been considered in the literature, but are usually much more difficult to analyse than the ESJD. It should be noted that a notion of distance on the state space is necessary to define the ESJD, which rules out the use of the ESJD in many important cases where MCMC methods are used. Before continuing, one might wonder why the expected squared jumping distance is considered while we could instead investigate the properties of $\mathbb{E}\big[\|x(k+1) - x(k)\|^p\big]$ for any power $p > 0$: as we will see, in many of the circumstances we are interested in, the MCMC algorithm can asymptotically be approached by a diffusion process. The squared jumping distance is thus directly related to the fact that the p-variation of a diffusion process is only interesting for $p = 2$. It is worth mentioning that in many situation where it is known that there is a diffusion limit, as this is described in the next section, maximization of the asymptotic integrated autocorrelation time and ESJD are equivalent: see [RR01] for a discussion.

3.2. **High Dimensional asymptotic: diffusion limits.** Quantifying computational complexity of an MCMC method is most naturally undertaken by studying the behaviour of the method on a family of probability distributions indexed by a parameter, and studying the cost of the algorithm as a function of that parameter. For example, the behaviour of the Gibbs sampler for the Ising model, as a function of the temperature, has been extensively studied [LPW09].

In this section we describe the costs of different MCMC algorithms as a function of the dimension of the target density. The seminal paper [RGG97] falls into this setting: this paper studied the behaviour of the Random Walk Metropolis algorithm (RWM) on target distributions with density [1]

$$\pi^d(x_1, \ldots, x_d) = \prod_{i=1}^{d} f(x_i),$$

[1]For ease of notation we do not distinguish between a measure and its density: all the densities are with respect to the usual Lebesgue measure on $\mathbb{R}^d$

where $f : \mathbb{R} \to \mathbb{R}^+$ is a fixed probability density. The RWM proposals are of the form $y = x + \sqrt{\ell d^{-\gamma}} Z$ where $Z \overset{\mathcal{D}}{\sim} \mathrm{N}(0, I_d)$ and $\ell, \gamma > 0$ are two parameters to be optimized: the larger γ, the smaller the variance of the proposed moves. It was shown that the ESJD is asymptotically maximized for $\gamma = \gamma_c = 1$, for any reasonable choice of function f. This means that for any fixed value of ℓ and γ,

$$\mathrm{ESJD}(d, \ell, \gamma) \leqslant \mathrm{ESJD}(d, \ell, \gamma_c), \qquad \gamma_c = 1$$

for d large enough where $\mathrm{ESJD}(d, \ell, \gamma)$ is the expected squared jumping distance for the algorithm with parameter ℓ and γ and target distribution π^d. Furthermore, the asymptotic optimal value of the parameter ℓ can be described as the one that leads to an asymptotic mean acceptance rate equal to 23%, independently of the function f.

These ideas have been generalized in many directions and are now better understood: for more complex target distributions π_d, where d stands for the dimension of the problem, and the proposal distribution $T^d_{\ell,\gamma}(x, y)$ has an asymptotic variance of order $d^{-\gamma}$, the choice of the scaling exponent γ is a delicate exercise. In a recent series of articles [BRS09, BS09, BRSJ] it was shown that for a variety of proposals $T^d_{\ell,\gamma}(x, y)$ there exists a critical exponent $\gamma_c > 0$ such that choice of $\gamma < \gamma_c$ leads to average acceptance probabilities which are smaller than any inverse power of d while for exponent $\gamma > \gamma_c$ the average acceptance probabilities tends to 1. This shows that in high dimensions, value of γ strictly smaller than γ_c lead to very poor mixing because of the negligible acceptance probability. However, it turns out that at the critical value γ_c the average acceptance probability is bounded away from 0 so that the algorithm does not degenerate. For optimality the Metropolis Markov chains $x^d = \{x^d(k)\}_{k \geqslant 0}$ must evolves in $\mathbb{R}^d$ with jumps of variance of order $d^{-\gamma_c}$. Because the variance of the jump $y - x$ is of order $d^{-\gamma_c}$, we need to speed up time by a factor $d^{-\gamma_c}$ to observe a non trivial limit: defining the continuous process z^d as a speeded up version of x^d by

$$z^d(k d^{-\gamma_c}) = x^d(k), \qquad k = 1, 2, \ldots$$

and linear interpolation in between, it is shown in many instances that z^d converges in a suitable way (see below) to a diffusion process. Care has to be taken in general since the different processes z^d live on different spaces: z^d is a d-dimensional process. A few examples include:

- the original paper [RGG97] shows that the first coordinate process $z^d_1 = \left\{z^d_1(t)\right\}_{t \in [0;T]}$ converges weakly in $\mathcal{C}([0, T]; \mathbb{R})$ to the scalar diffusion

$$dz = h(\ell)(\log f)'(z)\, dt + \sqrt{2h(\ell)}\, dW_t, \tag{3.1}$$

  where $h(\ell)$ is a computable constant depending on ℓ and $f(\cdot)$. Observe that the invariant distribution of the Langevin diffusion (3.1) is $f(x)\, dx$ and the constant $h(\ell)$ describes the speed of this diffusion. Moreover, it is shown in [RGG97] that

$$\lim_{d \to \infty} \mathrm{ESJD}(\ell, \gamma_c, d) = h(\ell), \qquad \gamma_c = 1$$

  so that in the asymptotic limit and for the critical choice of exponent $\gamma = \gamma_c$, maximizing the speed function over ℓ is equivalent to maximizing the ESJD. Maybe more importantly, the optimal value $\ell^\star$ of the parameter ℓ that leads to an asymptotic average acceptance probability is equal to 23%, independently of the target distribution f.

- similar results have been obtained in [RR98] for the MALA algorithm that proposes moves according to $y = \ell d^{-\gamma} \nabla(\log \pi^d)(x) + \sqrt{2\ell d^{-\gamma}}\, Z$ where again $Z \overset{\mathcal{D}}{\sim} \mathrm{N}(0, I_d)$ and π^d has density equal to $\prod_{i=1}^d f(x_i)$. The critical exponent is in this case equal to $\gamma_c = \frac{1}{3}$ and the limiting diffusion has exactly the same form: after speeding up the Metropolis Markov chain by a factor d^{γ_c}, the first coordinate process converges in laws to the diffusion 3.1. Nevertheless, the maximization of the speed function $h(\ell)$ leads to another universal value: independently of f, the optimal value of ℓ leads to an average acceptance probability equal to 57%.

These types of results have two immediate important messages:

- if the Metropolis chain x^d is speeded up by a factor d^{γ_c}, it can be approximated by a Langevin diffusion: this suggests that the MCMC algorithm takes $\mathcal{O}(d^{\gamma_c})$ to explore the d-dimensional target distribution π^d. Indeed, care has to be taken with this kind of heuristic, especially if it is only known that the first coordinate process converges to a diffusion. This is discussed in section 3.3.
- for practitioners, the universal constants 23% for the RWM algorithm and 57% for the MALA algorithm give very clear and straightforward ways of optimizing MCMC algorithms, at least in the setting were the results described above apply. It suffices to tune the variance of the proposals so that these optimal acceptance probabilities are achieved.

3.3. **Whole process scaling.** In this section we try to motivate the usefulness of the main theorem 4.3 of this report by arguing that the first coordinate scaling result of [RGG97] might not be satisfying to answer very simple questions. Indeed, if what we are interested in is a functional of the first coordinate, the first coordinate scaling is satisfying to give precise estimates of the rate of convergence of the RWM, say. One might even say that this is often the case that when dealing with high-dimensional problems we are in fact very often only interested in low dimensional statistics: these are the situation where the first coordinate scaling gives the correct answer. Nevertheless, this does not say how fast the whole process mixes.

As a first very simple example, consider a Gibbs sampler with target distribution density

$$\pi^d(x_1, \ldots, x_d) = \prod_{i=1}^d f(x_i) dx_i.$$

At each step, one coordinate $i \in \{1, 2, \ldots, d\}$ is chosen uniformly at random and x_i is updated by $x_i^\star \overset{\mathcal{D}}{\sim} f(x)\, dx$. It is obvious that once all the coordinates have been chosen at least one time the Markov chain has reached stationarity. Moreover, it takes $\tau \overset{\mathcal{D}}{\sim} \mathrm{Geom}(\frac{1}{d})$ steps before the first coordinate $i = 1$ is updated: this means that after $[\alpha d]$ steps the first coordinate $i = 1$ has a chance $\approx 1 - e^{-\alpha}$ of having reached stationarity. In other words, the first coordinate mixes on a time scale of order d. Nevertheless, this is not true that the Markov chain mixes on a time scale of order d since this is well known (coupon collector problem) that it takes $\mathcal{O}\big(d \ln(d)\big)$ to reach stationarity: there is a correlation cost between the different coordinates that has to be taken into account.

The next example is more realistic: consider the product form target density $\pi^d = \mathrm{N}(0, \frac{1}{d} I_d)$ *i.e.* the coordinates are independent and distributed as a centred Gaussian random

variables with variance $\sigma^2 = \frac{1}{d}$. An optimal RWM algorithm proposes moves of the form $y = x + \sqrt{\frac{\ell}{d^2}}\,\xi$ where $\xi \overset{\mathcal{D}}{\sim} \mathrm{N}(0, I_d)$: indeed, our situation is just a rescaled version (rescaled by a factor $\frac{1}{\sqrt{d}}$) of the one treated in [RGG97] with $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Because π^d is highly concentrated on the unit sphere of $\mathbb{R}^d$, one can argue that the MCMC algorithm behaves like a Brownian motion on $\mathbb{S}_{d-1}$ in the sense that the trajectory of x^d between 0 and k can be approximated by a Brownian trajectory W_t between time 0 and $h(\ell)\frac{k}{d^2}$, where $h(\ell) > 0$ is a computable constant. It is well known that a Brownian motion on $\mathbb{S}_{d-1}$ takes $\mathcal{O}(\frac{\ln(d)}{d})$ to mix, which shows that the RWM algorithm takes $\mathcal{O}(d\,\ln(d))$ to mix, and not $\mathcal{O}(d)$ as it could be believed from the first coordinate scaling result. As a simple illustration, if the Markov chain is started at $x^d(0) = (1, 0, 0, \ldots, 0) \in \mathbb{S}_{d-1}$ it takes order $\mathcal{O}(d\,\ln(d))$ for the first coordinate x_1 to be of order $\frac{1}{\sqrt{N}}$. Ongoing works try to make this heuristic more rigorous and adapt these ideas to more general target distributions.

4. S(P)DE limits for Langevin proposals

In this section we describe a variant of the main theorem of [RGG97]. This is the main result obtained during this first year of PhD and is joint work with Andrew Stuart and Natesh Pillai: this is a follow-up of the article [MPS10]. The generalization is twofold:

- the target density is not supposed to have a product form: the different coordinates are neither independent, nor identically distributed. A weak form of correlation is allowed, motivated by applications [Stu10a, BS09]. The target distribution π^d is a finite dimensional approximation of a limiting probability distribution π living on a separable Hilbert space $\mathcal{H}$: the probability π has a density with respect to a Gaussian measure $\pi_0 \overset{\mathcal{D}}{\sim} \mathrm{N}(0, C)$ on $\mathcal{H}$ which is supported by a linear subspace $\mathcal{H}^s \subset \mathcal{H}$,
$$\frac{d\pi}{d\pi_0}(x) \propto \exp\Big(-\Psi(x)\Big)$$
where $\Psi : \mathcal{H}^s \to \mathcal{H}^s$ is a reasonably well-behaved functional: precise assumptions on Ψ are given in section 4.2.
- in this setup where the coordinates are not supposed to be independent, the first coordinate process is not expected to converge to a Markovian diffusion. The distribution π^d is a finite dimensional approximation of π, living in $\mathbb{R}^d$. Nevertheless, looking at $\mathbb{R}^d$ as a finite dimensional subspace X^d of $\mathcal{H}$, the Metropolis Markov chain x^d that has π^d as invariant distribution can also be viewed as a Markov chain evolving in $X^d \subset \mathcal{H}$. All the Markov chains x^d living in the same Hilbert space $\mathcal{H}$, this allows to describe the limiting behaviour of the full Markov chains, not only the first coordinate process: we proved theorem 4.3 that shows that the appropriately rescaled Markov chains x^d converge weakly in $\mathcal{C}([0,T], \mathcal{H}^s)$ to the $\mathcal{H}^s$-valued diffusion
$$dz_t = -\big(z_t + C\nabla\Psi(z_t)\big)\,dt + dW_t \tag{4.1}$$
where W is a $\mathcal{H}^s$-valued Brownian motion.

The rigorous arguments leading to the main theorem being slightly too long to be presented in this report, we only give a detail sketch of the proof; the full argumentation can be found in the forthcoming paper [PST11].

4.1. **Preliminaries.** Let $\big(\mathcal{H}, \langle\cdot,\cdot\rangle\big)$ be a separable Hilbert space and C be a self-adjoint positive, trace class operator on $\mathcal{H}$. Let $\{\phi_j, \lambda_j^2\}$ be the eigenfunctions and eigenvalues of C respectively, so that

$$C\phi_j = \lambda_j^2\,\phi_j,\ j \in \mathbb{N}.$$

We assume a normalization under which $\{\phi_j\}$ forms a complete orthonormal basis in $\mathcal{H}$. We also assume that the eigenvalues are arranged in decreasing order and are all strictly positive: the expansion in this basis of a vector $x \in \mathcal{H}$ is $x = \sum_i x_i\phi_i$ where $x_i = \langle x, \phi_i\rangle$. Using this expansion, we define the Sobolev-like space $\mathcal{H}^s, s \in \mathbb{R}$, with norm defined by

$$\|x\|_s^2 \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} j^{2s}\, x_j^2. \tag{4.2}$$

Let π_0 denote a mean zero Gaussian measure on $\mathcal{H}$ with covariance operator C, *i.e.,* $\pi_0 \stackrel{\text{def}}{=} \mathrm{N}(0, C)$. If $x \stackrel{\mathcal{D}}{\sim} \pi_0$, then the $x_j = \langle x, \phi_j\rangle$ are independent $N(0, \lambda_j^2)$ Gaussian random variables and we may write (Karhunen-Loève expansion) $x = \sum_{j=1}^{\infty} \lambda_j\,\xi_j\,\phi_j$ where $\xi_j \stackrel{\mathcal{D}}{\sim} \mathrm{N}(0,1)$ are i.i.d. standard Gaussian random variables. We assume in the sequel that there exists $\kappa > 0$ such that $\lambda_j \asymp j^{-\kappa}$ and we fix $s \in [0; \kappa - \frac{1}{2})$. Observe that $\mathbb{E}_{\pi^0}\,\|x\|_s^2 < \infty$ so that π_0-almost every $x \in \mathcal{H}$ belong to $\mathcal{H}^s$: this is the motivation for the introduction of the Sobolev-like space $\mathcal{H}^s$.

Our goal is to sample from a measure π on $\mathcal{H}$, given by

$$\frac{d\pi}{d\pi_0} = M_\Psi\,\exp\Big(-\Psi(x)\Big)$$

where M_Ψ is a normalization constant and $\Psi : \mathcal{H}^s \to \mathbb{R}$ is a functional that only needs to be defined on $\mathcal{H}^s$. To this purpose, we first approximate π and π_0 by probability measures π^d and π_0^d living in a finite dimensional subspace $X^d \subset \mathcal{H}$ of dimension d. Consider the orthogonal projection $P^d : \mathcal{H} \to X^d \subset \mathcal{H}$ on

$$X^d = \mathrm{span}\big\{\phi_1, \ldots, \phi_d\big\}.$$

The approximate distribution π^d is a d-dimensional approximation of π defined as follows:

- the approximate prior probability measure π_0^d is defined as the image of π_0 under P^d: this is a Gaussian measure living in $X^d \cong \mathbb{R}^d$, with covariance $C_d = P^d \circ C \circ P^d$ and density with respect to the Lebesgue measure on $X^d \cong \mathbb{R}^d$ proportional to $\exp\Big(-\frac{1}{2}\langle x, C_d^{-1}x\rangle\Big)$.
- the function $\Psi : \mathcal{H}^s \to \mathcal{H}^s$ is approximated by $\Psi^d(x) = \Psi\big(P^d x\big)$: this can be viewed as a function on X^d.
- finally, the approximate probability distribution π^d is defined as a change of measure with respect to π_0^d,

$$\frac{d\pi^d}{d\pi_0^d}(x) = M_{\Psi^d}\,\exp\Big(-\Psi^d(x)\Big).$$

The probability π^d has a support equal to $X^d \cong \mathbb{R}^d$ and has density in $\mathbb{R}^d$ proportional to $\exp\Big(-\frac{1}{2}\langle x, C_d^{-1}x\rangle - \Psi^d(x)\Big)$, where $C_d = P^d \circ C \circ P^s$ is the restriction of C to X^d.

The precise assumptions on the functional Ψ and the covariance operator C are described in the next section.

4.2. **Assumptions on π_0 and Ψ.** In order for the main theorem 4.3 to be valid, the posterior distribution π must not be very different from the Gaussian prior π_0: this is translated by growth conditions imposed on the functional Ψ. The regularity of the functional Ψ is needed to control the behaviour of its gradient $\nabla\Psi$ which is used in the proposals of the different Markov chains (see section 4.3). Also, we impose a rate of decay of the eigenvalues λ_i^2 of C: this is not fundamental but simplifies the proof to a certain extent.

To avoid technicalities we assume that $\Psi(x)$ is quadratically bounded and lower bounded in $\mathcal{H}^s$, with first derivative linearly bounded and second derivative globally bounded. Weaker assumptions could be dealt with by use of stopping time arguments.

Assumptions 4.1. *The operator C and functional Ψ satisfy the following:*

(1) ***Decay of Eigenvalues λ_i^2 of C:*** *there is an exponent $\kappa > \frac{1}{2}$ such that*

$$\lambda_j \asymp j^{-\kappa}$$

(2) ***Assumptions on Ψ:*** *There exist $s \in [0, \kappa - 1/2)$ such that for all $x \in \mathcal{H}^s$ we have*

$$\begin{aligned} 1 \lesssim \Psi(x) &\lesssim \Big(1 + \|x\|_s^2\Big) \\ \|\nabla\Psi(x)\|_{-s} &\lesssim \Big(1 + \|x\|_s\Big) \\ \|\partial^2\Psi(x)\|_{L(\mathcal{H}^s, \mathcal{H}^{-s})} &\lesssim 1. \end{aligned}$$

Remark 4.2. *the condition $\kappa > \frac{1}{2}$ ensures that C is a trace class operator. Also, the $\mathcal{H}^s$ norm of $x \overset{\mathcal{D}}{\sim} N(0, C)$ is almost surely finite because $\mathbb{E}\Big(\|x\|_s^2\Big) < \infty$ for $x \overset{\mathcal{D}}{\sim} N(0, C)$. A simple example of a function Ψ satisfying the above assumptions is $\Psi(x) = \|x\|_s^2$.*

The above regularity assumptions on Ψ imply in particular that the functional $\mu(x) = -\big(x + C\nabla\Psi(x)\big)$ satisfies

$$\|\mu(x) - \mu(y)\|_s \leqslant K \cdot \|x - y\|_s$$

for a certain constant $K > 0$. This is important since μ is the drift of the Langevin diffusion (4.1). This is used to establish the continuity of the Itô map $\Theta : \mathcal{H}^s \times \mathcal{C}([0,T], \mathcal{H}^s)$ that maps a couple $(x_0, w) \in \mathcal{H}^s \times \mathcal{C}([0,T], \mathcal{H}^s)$ to the unique solution of the integral equation $x_t = x_0 + \int_0^t \mu(x_s)\, ds + w_t$.

4.3. **Algorithm description and main theorem.** The goal is now to sample from π^d: we use the Metropolis Adjusted Langevin Algorithm (MALA) [RT96, RS02] that we describe now. It is readily checked that the $\mathbb{R}^d$-valued diffusion $dX = A\nabla D(X)\, dt + \sqrt{2A}\, dW_t$, where A is any positive definite matrix and $D : \mathbb{R}^d \to \mathbb{R}$ a smooth function that defines a probability distribution $e^{-D(x)}\, dx$, has precisely $e^{-D(x)}\, dx$ as invariant distribution: the matrix C is usually called a preconditioning matrix and the process is called "Langevin diffusion" in the literature. Since this diffusion has $e^{-D(x)}\, dx$ as invariant diffusion, this is tempting to use Euler-Maruyama discretizations of this diffusion as MCMC proposals in order to study the distribution $e^{-D(x)}\, dx$. In other words, proposals of the MALA are defined by $y = x + A\nabla D(x)\, \Delta t + \sqrt{2A\Delta t}\, \mathrm{N}(0, I_d)$ and then accepted or rejected according to the usual

Metropolis-Hastings rule.

The probability π^d has density proportional to $\exp\Big(-\frac{1}{2}\langle x, C_d^{-1}x\rangle - \Psi^d(x)\Big)$ on $X^d \cong \mathbb{R}^d$ so that the MALA proposals with preconditioning matrix equal to C_d read

$$y = x - \ell\Delta t\Big(x + C_d\nabla\Psi^d(x)\Big) + \sqrt{2\ell\Delta t\, C_d}\,\xi^d, \qquad \xi^d = \sum_{i=1}^{d}\xi_i\phi_i \tag{4.3}$$

where ℓ and Δt are parameters to be optimized and $\xi_i \overset{\mathcal{D}}{\sim} \mathrm{N}(0,1)$ are i.i.d. Gaussian random variables. It is known [RR98] that the critical exponent for the MALA algorithm is $\gamma_c = \frac{1}{3}$ so that Δt is chosen equal to

$$\Delta t = d^{-\frac{1}{3}}.$$

To be more precise, the Markov chain $x^d = \{x^d(k)\}_{k\geqslant 0}$ evolving in $\mathcal{H}$ and with target distribution π^d is defined as follows: if the current position is $x = x^d(k)$, a proposal $y^d(k)$ distributed according to (4.3) is considered and accepted with probability $\alpha^d(x^d(k), y^d(k))$ equal to the Metropolis-Hastings ratio of the move from $x^d(k)$ to $y^d(k)$. A Bernoulli random variable $\gamma^d(k)$ which can be defined as $1_{\{U<\alpha^d(x(k),y(k))\}}$ is introduced, where $U \overset{\mathcal{D}}{\sim} \mathrm{Uniform}(0,1)$ is independent of any other source of randomness : the next position $x^d(k+1)$ of the Metropolis Markov chain is

$$x^d(k+1) = \big(1-\gamma^d(k)\big)x^d(k) + \gamma^d(k)y^d(k).$$

To observe a diffusion limit, the Markov chain x^d has to be speeded up by a factor $d^{\frac{1}{3}}$: the continuous interpolate z^d of x^d is thus defined by

$$z^d(t) = \big(t/\Delta t - k\big)\, x^d(k+1) + \big(k+1-t/\Delta t\big)\, x^d(k), \qquad k\Delta t \leqslant t < (k+1)\Delta t. \tag{4.4}$$

We show that the asymptotic mean acceptance probability of this algorithm is given by $\alpha(\ell) = \mathbb{E}\big[1\wedge\exp(Z_\ell)\big]$ where $Z_\ell \overset{\mathcal{D}}{\sim} \mathrm{N}(-\frac{\ell^3}{4}, \frac{\ell^3}{2})$.

Theorem 4.3. *Suppose that the assumptions 4.1 are satisfied and that the MALA Markov chain x^d is started at stationarity,*

$$x^d(0) \overset{\mathcal{D}}{\sim} \pi^d.$$

Then the sequence of rescaled continuous interpolates z^d defined by (4.4) converges weakly in $\mathcal{C}([0,T];\mathcal{H}^s)$ to the $\mathcal{H}^s$-valued diffusion process z

$$\begin{cases} dz & = -h(\ell)\left(z + C\nabla\Psi(z)\right)dt + \sqrt{2h(\ell)}\,dW_t \\ z(0) & \overset{\mathcal{D}}{\sim} \pi \end{cases} \tag{4.5}$$

where $\{W_t\}_{t\geqslant 0}$ is a $\mathcal{H}$-valued Brownian motion with covariance operator C and the speed function $h(\ell)$ is given by

$$h(\ell) = \ell\,\alpha(\ell).$$

Remark 4.4. *Observe that the $\mathcal{H}$-valued Brownian motion with covariance C takes value in $\mathcal{H}^s$ since π^0-almost every element of $\mathcal{H}$ are in $\mathcal{H}^s$.*

This theorem thus describes the asymptotic behaviour of the whole Markov chain. It is interesting to notice that the speed function $h(\ell)$ has exactly the same expression as the analogous theorem describing diffusion limit of the first coordinate process [RR98]: it is maximized for $\ell^\star$ satisfying $\alpha(\ell^\star) = 0.574$ to three decimal places. Remarkably, the optimal

acceptance probability identified in [RR98] for product measures, is also optimal for the non-product measures studied in this paper.

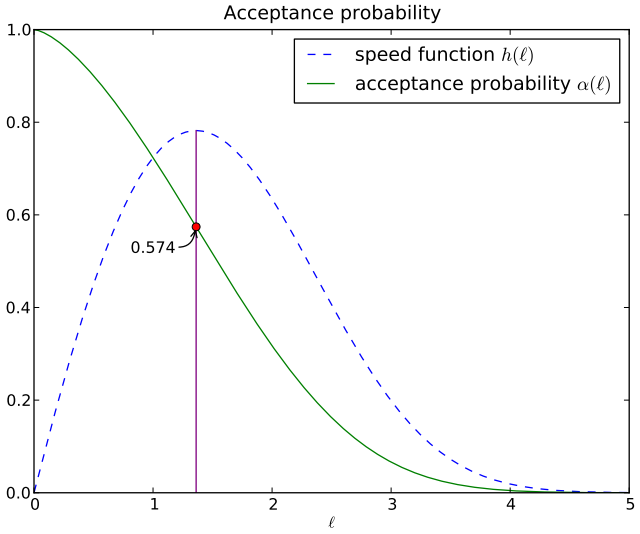

FIGURE 1. Optimal acceptance probability = 0.574

4.4. **Sketch of the proof.** This section describes the main ideas behind the proof of theorem 4.3; the full proof can be found in [PST11]. The majority of the MCMC diffusion limits present in the literature [RGG97, RR98, BR00, Béd07] is based on the generator approach. The generalization of this technique to the Hilbert space setting is difficult: instead, following [MPS10], we adopt a pedestrian approach which has the advantages of keeping the technicalities at their minimum and of offering a better understanding of the diffusion limit.

Consider a scalar Lipschitz function $\mu : \mathbb{R} \to \mathbb{R}$ and constants $\ell, C > 0$: the usual theory of diffusion approximation for Markov processes [EK86] shows that the sequence $x^d = \left\{x^d(k)\right\}_k$ of scalar Markov chains

$$x^d(k+1) - x^d(k) = \mu(x^d(k))\ell\Delta t + \sqrt{2\ell\Delta t}C^{\frac{1}{2}}\xi(k),$$

with $\Delta t \to 0$ and $\xi^k \overset{\mathcal{D}}{\sim} \mathrm{N}(0,1)$ converges in any reasonable sense, when speeded up by a factor $(\Delta t)^{-1}$, to the scalar diffusion $dz_t = \ell\mu(z_t)\,dt + \sqrt{2\ell}\,dW_t$, where W is a Brownian motion with variance $\mathrm{Var}(W_t) = Ct$. Also, if $\{\gamma(k)\}_k$ is an i.i.d. sequence of Bernoulli random variables with success rate $\alpha(\ell)$, independent from the Markov chain x^d, then it can be proved that the sequence

$$x^d(k+1) - x^d(k) = \gamma(k)\,\mu(x^d(k))\ell\Delta t + \sqrt{2\ell\Delta t}\gamma(k)\,C^{\frac{1}{2}}\xi(k),$$

converges, after being speeded up by a factor $(\Delta t)^{-1}$, to the diffusion $dz_t = h(\ell)\mu(z_t)\,dt + \sqrt{2h(\ell)}\,dW_t$ where $h(\ell) = \ell\alpha(\ell)$. Hence, the Bernoulli random variables γ^k have slowed down the original Markov chain by a factor $\alpha(\ell)$, as expected.

The proof of theorem 4.3 is an application of this idea in a slightly more general setting:

- instead of working for scalar diffusions, the result holds for a Hilbert space valued diffusions. The difference is small but the correlation structure between the different coordinates has to be taken into account.
- instead of working with a single function μ, a sequence of approximations $\mu^d(x)$ has to be taken into account.
- the Bernoulli random variables $\gamma(k)$ are not i.i.d. and have an autocorrelation structure. Moreover, the Bernoulli random variables $\gamma(k)$ are not independent from the Markov chain x^d. This is the main difficulty in the proof.
- it should be emphasized that the main theorem crucially uses the fact that the MALA Markov chain is started at stationarity: this in particular implies that $x^d(k) \overset{\mathcal{D}}{\sim} \pi^N$ for any $k \geqslant 0$. Indeed, this is known that in many instances, if the Markov chain is started "far" from stationarity, a fluid limit is first observed [CRR05].

The main ingredient of the proof is a Gaussian approximation. We now give the main steps of the proof:

(1) **Gaussian approximation**: it can be shown by simple algebraic manipulations that the Metropolis-Hastings ratio $\alpha^d(x,y) = 1 \wedge e^{Q^d(x,\xi^d)}$ introduced in section 4.3 satisfies

$$Q^d(x,\xi^d) = Z^d(x,\xi^d) + \mathbf{e}^d(x,\xi^d) \tag{4.6}$$

where $Z^d(x,\xi^d) = -\frac{\ell^3}{4} - \frac{\ell^{\frac{3}{2}}}{\sqrt{2}} d^{-\frac{1}{2}} \sum_{1 \leqslant j \leqslant d} \frac{\xi_i x_i}{\lambda_j}$ and $\mathbf{e}^d(x,\xi^d)$ is an error term. For any fixed value of x the random variable $Z^d(x,\xi)$ is Gaussian with mean $-\frac{\ell^3}{4}$ and variance $\frac{\ell^3}{2} \frac{1}{d} \sum_{i=1}^d \frac{x_i^2}{\lambda_i^2}$: at stationarity, as this is the case in theorem 4.3, x is distributed according to π^d and usual Gaussian concentration shows that the quantity $\frac{1}{d} \sum_{i=1}^d \frac{x_i^2}{\lambda_i^2}$ is approximately equal to 1 with overwhelming probability. In [PST11] we precisely quantify the error in the approximation

$$Q^d(x,y) \approx Z_\ell$$

where $Z_\ell \overset{\mathcal{D}}{\sim} \mathrm{N}(-\frac{\ell^3}{4}, \frac{\ell^3}{2})$ is a Gaussian random variables independent from the local position x and the noise term ξ: this is the main part of the proof. This in particular implies that the asymptotic acceptance rate is equal to $\alpha(\ell) = \mathbb{E}\big[1 \wedge e^{Z_\ell}\big]$.

(2) **Drift-Martingale decomposition**: the increment $x^d(k+1) - x^d(k)$ is decomposed as

$$x^d(k+1) - x^d(k) = \mu^d\big(x^d(k)\big)\,\Delta t + \sqrt{2\Delta t}\,\Gamma^d(k)$$

where $\mu^d : \mathcal{H} \to \mathcal{H}$ is a deterministic function and $\Gamma^d(k)$ defines a martingale $M^d(k) = \sum_{j \leqslant k-1} \Gamma^d(j)$ adapted to the natural filtration of x^d. In other words, $\mu^d(x) = (\Delta t)^{-1}\,\mathbb{E}\big[x^d(k+1) - x^d(k)\,|x^d(k) = x\big]$. It can be proved that

$$\lim_d \mathbb{E}^{\pi^d}\big[\,\big\|\mu^d(x) - h(\ell)\mu(x)\big\|\,\big] = 0$$

where $\mu(x) = -\Big(x + C\nabla\Psi(x)\Big)$.

(3) **Invariance principle**: an invariance principle for Hilbert space valued martingales [Ber86] and the Gaussian approximation (4.6) are used to show that the continuous rescaled sequence of martingales $w^d(k\Delta t) = M^d(k)$ converges weakly in $\mathcal{C}([0,T];\mathcal{H}^s)$ to a Brownian motion W with covariance $h(\ell)C$.

In other words, the decomposition $x^d(k+1) - x^d(k) = \mu^d\big(x^d(k)\big)\,\Delta t + \sqrt{2\Delta t}\,\Gamma^d(k)$ 'resembles' the Euler-Maruyama discretization of the SPDE

$$dz = -h(\ell)\,\big(z + C\nabla\Psi(z)\big)\,dt + \sqrt{2h(\ell)}\,dW_t.$$

(4) **continuity of the Itô map**: the Itô map $\Theta : \mathcal{H} \times \mathcal{C}([0,T];\mathcal{H}^s) \to \mathcal{C}([0,T];\mathcal{H}^s)$ that sends a couple $(z_0, w) \in \mathcal{H}^s \times \mathcal{C}([0,T];\mathcal{H}^s)$ to the unique solution $z \in \mathcal{C}([0,T],\mathcal{H}^s)$ of the integral equation

$$z(t) = z_0 + \int_0^t \mu\big(z(u)\big)\,du + w(t), \qquad \forall t \in [0,T]$$

is continuous: the usual Picard's iteration proof of the Cauchy uniqueness theorem for ODEs works exactly the same. This would not be true if the noise in the SPDE (4.5) was not additive *i.e* the volatility function were not constant.

The continuous interpolate z^d defined by (4.4) satisfies

$$z^d = \Theta(x^d(0), w^d) + (\text{error})$$

where it can be proved that the error term converges in probability to the null function in $\mathcal{C}([0,T],\mathcal{H}^s)$ and is thus asymptotically negligible. The end of the proof follows from the continuous mapping theorem: the law of the diffusion (4.5) is equal to the law of $\Theta(z(0), W)$, with $z(0) \overset{\mathcal{D}}{\sim} \pi$ and it can be proved that $\big(x^d(0), w^d\big)$ converges weakly in $\mathcal{H}^s \times \mathcal{C}([0,T];\mathcal{H}^s)$ to $(z(0), W)$.

Acknowledgement: I would like to thank my supervisors Gareth Roberts and Andrew Stuart for the patient guidance and invaluable advices they are giving me during this PhD.

References

[Béd07] M. Bédard. Weak convergence of Metropolis algorithms for non-iid target distributions. *Annals of Applied Probability*, 17(4):1222–1244, 2007.

[Ber86] E. Berger. Asymptotic behaviour of a class of stochastic approximation procedures. *Probab. Theory Relat. Fields*, 71(4):517–552, 1986.

[BG93] J. Besag and P.J. Green. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(1):25–37, 1993.

[BGHM95] J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 10(1):3–41, 1995.

[BR00] LA Breyer and GO Roberts. From Metropolis to diffusions: Gibbs states and optimal scaling. *Stochastic Processes and their Applications*, 90(2):181–206, 2000.

[BR08] 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.

[BRS09] A. Beskos, G. Roberts, and A. Stuart. Optimal scalings for local Metropolis-Hastings chains on non-product targets in high dimensions. *Ann. Appl. Prob*, 19:863–898, 2009.

[BRSJ] A. Beskos, G. Roberts, A. Stuart, and V. Jochen. MCMC methods for diffusion bridges.

[BS09] A. Beskos and A. Stuart. MCMC methods for sampling function space. In *ICIAM 07: 6th International Conference on Industrial and Applied Mathematics, Zurich, Switzerland, 16-20 July 2007: invited lectures*, page 337. Amer Mathematical Society, 2009.

[CRR05] O.F. Christensen, G.O. Roberts, and J.S. Rosenthal. Scaling limits for the transient phase of local Metropolis–Hastings algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):253–268, 2005.

[Dia09] P. Diaconis. The markov chain monte carlo revolution. *AMERICAN MATHEMATICAL SOCIETY*, 46(2):179–205, 2009.

[EK86] S.N. Ethier and T.G. Kurtz. *Markov processes: Characterization and convergence*, volume 6. Wiley New York, 1986.

[GCC09] M. Girolami, B. Calderhead, and S.A. Chin. Riemannian manifold hamiltonian monte carlo. *Arxiv preprint arXiv:0907.1100*, 2009.

[Liu08] J.S. Liu. *Monte Carlo strategies in scientific computing*. Springer Verlag, 2008.

[LPW09] D.A. Levin, Y. Peres, and E.L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2009.

[MPS10] J.C. Mattingly, N.S. Pillai, and A.M. Stuart. Diffusion Limits of the Random Walk Metropolis Algorithm in High Dimensions. *Arxiv preprint arXiv:1003.4306*, 2010.

[MRR+53] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, et al. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087, 1953.

[MTH93] S.P. Meyn, R.L. Tweedie, and JL Hibey. *Markov chains and stochastic stability*. Springer London et al., 1993.

[NRY07] P. Neal, G. Roberts, and J. Yuen. Optimal scaling of random walk Metropolis algorithms with discontinuous target densities. 2007.

[PG09] C. Pasarica and A. Gelman. Adaptively scaling the Metropolis algorithm using expected squared jumped distance. *Statistica Sinica*, 2009.

[PS08] G.A. Pavliotis and AM Stuart. *Multiscale methods: averaging and homogenization*. Springer Verlag, 2008.

[PST11] N.S. Pillai, A.M. Stuart, and A.H. Thiéry. Langevin Algorithm in High Dimensions. *To appear*, 2011.

[RC04] C.P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Verlag, 2004.

[RGG97] G.O. Roberts, A. Gelman, and W.R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, 1997.

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

[RR01] G.O. Roberts and J.S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

[RS02] GO Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and Computing in Applied Probability*, 4(4):337–357, 2002.

[RT96] G.O. Roberts and R.L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[Sok96] A.D. Sokal. *Monte Carlo methods in statistical mechanics: foundations and new algorithms*. 1996.

[Stu10] A.M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica, To appear*, 2010.

[Tie94] L. Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, 22(4):1701–1728, 1994.