Weak Convergence and Optimal Proposals for Monte Carlo Algorithms

Ali Zaidi

Under the supervision of Professor J.S. Rosenthal

Abstract

Metropolis-Hastings Markov chain Monte Carlo (MCMC) algorithms and other popular MCMC algorithms induce a Markov chain which has the target distribution as its stationary distribution. Optimal scaling refers to the need to tune the parameters of the proposal kernel in order to ensure the Markov chain obtained from the algorithm converges as fast as possible to stationarity. Theoretical results for optimal scaling are obtained by approximating high dimensional algorithm trajectories by diffusions, and then maximizing the speed measure of the corresponding diffusion with respect to a transformation of the scale parameter. These notes summarize most of the literature regarding the optimal tuning of Markov chain Monte Carlo algorithms, and discuss the applications of these findings in practice. We have added an extensive appendix which summarizes some needed background to understand the proofs and concepts presented in the various papers.

Contents

1 Introduction and First Results

1.1 The “Goldilocks” Principle

Achieving the optimal convergence speed for a Markov chain Monte Carlo (MCMC) algorithm is important and often a crucial matter in order to ensure efficiency in computation. Due to the flexibility of the Metropolis-Hastings algorithm and other popular Monte Carlo algorithms, it is usually not difficult to set up a Markov chain that *eventually* converges to the stationary distribution of interest in many problems. However, the fundamental question that must be answered after one is able to construct such a Markov chain is the following: how many steps is sufficient in order for the chain to be close to the stationary distribution of interest? The amount of steps necessary before we have “convergence” is called the “mixing time” of the chain/algorithm. One way of reaching some level of optimality in the performance for Metropolis-Hastings algorithms is through the selection of the scaling parameter of the proposal distribution that leads to the fastest mixing time. This leads to the topic of optimal scaling. The selection of the optimal scaling parameter is an interesting application of probability theory to MCMC, as the results are often proven using the theory of Markov processes and weak convergence to diffusions, although some recent extensions have instead relied on the results of functional analysis and infinite dimensional analysis of probability spaces. As much as possible, we have summarized these results in the appendix (see A).

If we want to run a Metropolis-Hastings algorithm for a continuous d-dimensional density $\pi_u : \mathbb{R}^d \mapsto [0, \infty)$ with a large dimension d, then the optimal scaling problem is concerned with the question of how we should choose the proposal distribution for this algorithm. If we consider the specific examples such as the random walk Metropolis algorithm (RWM) with Gaussian proposals or the Metropolis-Adjusted Langevin algorithm (MALA), then the specific concern is the choice of the scaling parameter σ^2 in the proposal distribution, as was mentioned earlier. These two types of Metropolis algorithms are explained in Section 1.2.1.

If σ^2 is chosen to be too small, then nearly all proposed moves will be accepted, and the Markov chain induced by the Metropolis algorithm will converge very slowly since each increment is very small. Thus, the problem with such an algorithm is that this prevents sufficient "mixing", in that the chain will move slowly and thus lead to a poor performance. On the other hand, if σ^2 is chosen to be too large, then the chain will almost never accept new proposals, and thus remain idle in a few states for large periods of time. Extremal variances lead to an algorithm which explores its state space in a lazy fashion, and thus converges slowly to the target stationary distribution. Hence, practitioners are eager to find a value of σ^2 that is optimal, which allows sufficient mixing and exploration of the space in consideration (i.e., the one that satisfies the Goldilocks principle). Similar to the plots shown in Section 2 of [Ros10], Figure 1 shows three trace plots of the Markov chain induced by a RWM algorithm applied to a multivariate normal target distribution. The first two plots show poor mixing due to extremal choices of the proposal variance. The third plot shows a well-chosen proposal variance, which provides sufficient mixing of the Markov chain and a good estimate of the distribution of interest.

Indeed, from the onset of the Metropolis algorithms, practitioners and theoreticians were already aware of the issue of optimal proposals. In the founding paper on Metropolis algorithms, [MRR+53], the authors considered proposals of the form $y \sim \text{Uniform}\,[x - \alpha, x + \alpha]$, and they observed (page 1089):

> *"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."*

An alternative way of looking at this dilemma is by examining the source of correlation in the induced Markov chain. An ideal proposal distribution for the Metropolis algorithm would be one that reduces the correlation between subsequence realizations of Markov chain. However, note that there are two sources which produce the correlation within the Markov chain: the first one is due to the fact that there is correlation between the current state and the subsequent proposed value through the proposal, and secondly, there is correlation introduced whenever a proposal is rejected and thus the value of the Markov chain is retained. In order to eliminate these two sources we would ideally want to a proposal distribution that allows for fast changes/mixing in the induced Markov chain, and also a high probability of acceptance. Unfortunately, these are two competing goals, as discussed above, reducing one source of correlation increases the other source.

The first two papers to obtain theoretical results on optimal scaling for Metropolis algorithms are discussed in subsections 1.3 and 1.4. Remarkably, despite starting with fairly stringent conditions on the form of the target density, the results obtained in these two papers have shown to apply in fairly general circumstances. The papers discussed in the later sections very frequently reinforce the results found in these two papers.

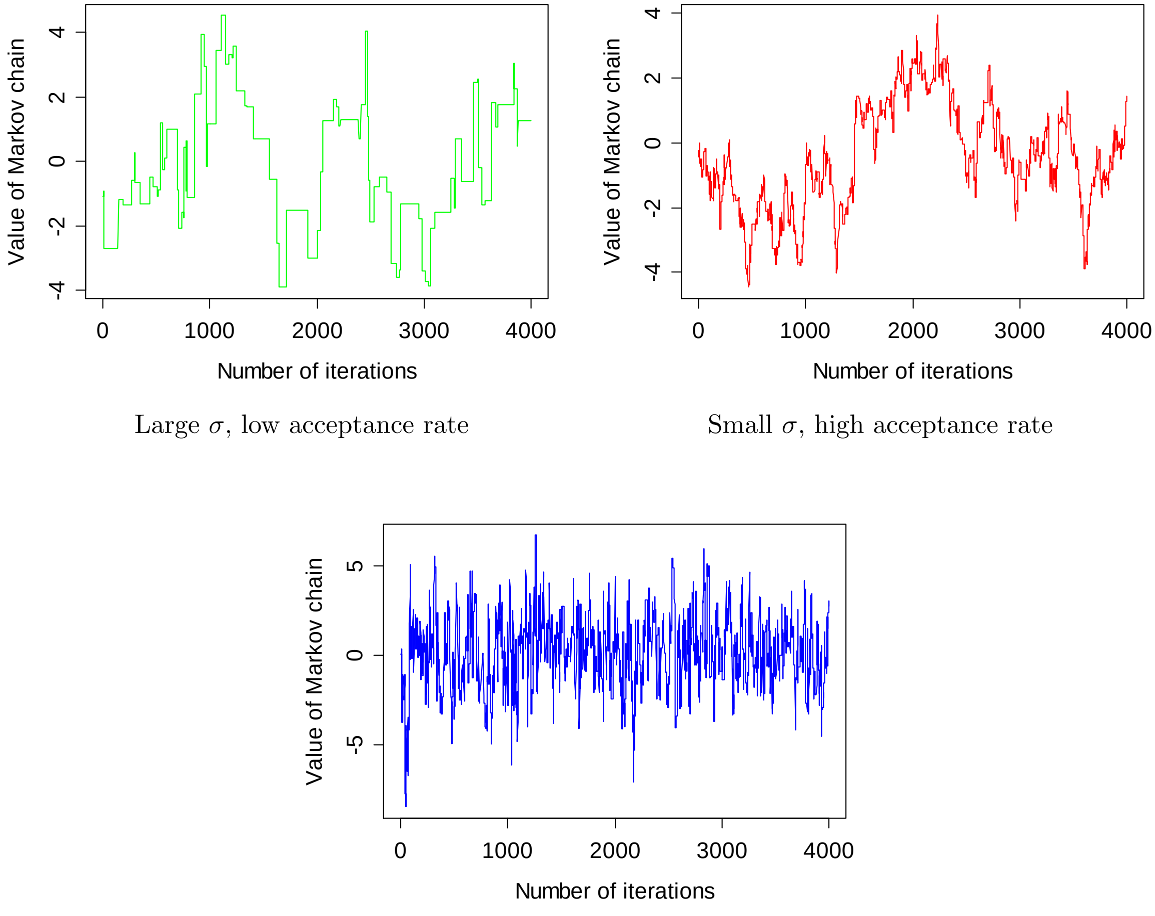
1.2 Some Preliminaries and Notation

1.2.1 Metropolis Algorithms

Before we begin to provide theoretical results concerning optimality of the Metropolis algorithm, we first need to describe the algorithm and explain why it works. We will be necessarily brief in this subsection, omitting several key topics, which can be found in any treatise on Markov chain Monte Carlo methods (for instance, a thorough review of the Metropolis-Hastings algorithm can be found in Chapter 7 of [RC04]).

The basic idea of the Metropolis algorithm as formulated in Metropolis, *et al.* [MRR+53] is that each new state of the Markov chain is chosen by considering a small perturbation to the current state, and accepting or rejecting this change based on how the probability of the altered state compares to that of the current state. For the random

Figure 1: Trace plots of RWM applied to a MVN Target



Large σ, low acceptance rate

Small σ, high acceptance rate

Medium σ, medium acceptance rate.

walk Metropolis algorithm, one proposes a Gaussian density,

$$q\left(\boldsymbol{x}^{(d)},\boldsymbol{y}^{(d)}\right)=\frac{1}{\left(2\pi\sigma^{2}\right)^{d/2}}\exp\left\{-\frac{1}{2\sigma^{2}}\left|\boldsymbol{y}^{(d)}-\boldsymbol{x}^{(d)}\right|^{2}\right\}. \tag{1.1}$$

We will denote this proposal increment $q_{\sigma}\left(\boldsymbol{x}^{(d)}\right)d\boldsymbol{x}\sim\mathcal{N}\left(0,I_{d}\sigma^{2}\right)$.

The Metropolis adjusted Langevin algorithm (MALA) is obtained by discretization of the Langevin diffusion,

$$dV_{t}=\sigma dB_{t}+\frac{\sigma^{2}}{2}\nabla\log\pi_{d}\left(V_{t}\right)dt, \tag{1.2}$$

where B_t is a standard d-dimensional Brownian motion. The natural discrete approximation can be given by

$$V_{t+1}=V_{t}+\sigma Z_{t+1}+\frac{\sigma^{2}}{2}\nabla\log\pi\left(V_{t}\right), \tag{1.3}$$

where the random variables Z_t are i.i.d. standard normal. The MALA thus has proposal increments $q_{\sigma}\left(\boldsymbol{x}^{(d)}\right)d\boldsymbol{x}\sim\mathcal{N}\left(\frac{\sigma^{2}}{2}\nabla\log\pi\left(X_{t}\right),\sigma^{2}I_{d}\right)$. Analogous to equation (1.1), we have for the proposal distribution for the MALA given by

$$q\left(\boldsymbol{x}^{(d)},\boldsymbol{y}^{(d)}\right)=\frac{1}{\left(2\pi\sigma^{2}\right)^{d/2}}\exp\left(-\frac{1}{2\sigma^{2}}\left|\boldsymbol{y}-\boldsymbol{x}-\frac{\sigma^{2}}{2}\nabla\log\pi\left(\boldsymbol{x}\right)\right|^{2}\right).$$

As mentioned above, the MALA algorithm is obtained by considering a discrete approximation of the Langevin diffusion. Unfortunately, discrete approximations like the one (1.3) can have different behaviors than the diffusion process they attempt to approximate, such as (1.2). In particular, it is possible that the approximation is transient, regardless of how small the step variance σ_d^2 is chosen to be (see [RT96]). Therefore, it is necessary to further introduce a Metropolis accept and reject step to ensure that π is a stationarity distribution for the process.

For both algorithms, we shall denote the acceptance probability

$$\alpha\left(\boldsymbol{X}_{t},\boldsymbol{Y}_{t+1}\right)=\min\left\{\frac{\pi\left(\boldsymbol{Y}_{t+1}\right)q\left(\boldsymbol{Y}_{t+1},\boldsymbol{X}_{t}\right)}{\pi\left(\boldsymbol{X}_{t}\right)q\left(\boldsymbol{X}_{t},\boldsymbol{Y}_{t+1}\right)},1\right\}.$$

Due to the fact the Gaussian distribution proposal for the RWM is symmetric, we have the simplified acceptance probability given by

$$\alpha\left(\boldsymbol{X}_{t},\boldsymbol{Y}_{t+1}\right)=\min\left\{\frac{\pi\left(\boldsymbol{Y}_{t+1}\right)}{\pi\left(\boldsymbol{X}_{t}\right)},1\right\}.$$

In case of rejection, which happens with probability $1-\alpha\left(\boldsymbol{X}_{t},\boldsymbol{Y}_{t+1}\right)$, set $\boldsymbol{X}_{t+1}=\boldsymbol{X}_{t}$.

1.2.2 Efficiency

To compare different implementations of MCMC, it is necessary to obtain some notion of efficiency of Markov chains for comparison criteria. Suppose $\{\boldsymbol{X}_t\}_{t\geq1}$ is the Markov chain of the algorithm in consideration, and suppose that the chain has stationary distribution $\pi\left(\cdot\right)$. The Markov chain is said to be *geometrically ergodic* with stationary distribution $\pi\left(\cdot\right)$ if

$$\left\|P^{n}\left(\boldsymbol{x},\cdot\right)-\pi\left(\cdot\right)\right\|_{1}\leq M\left(\boldsymbol{x}\right)r^{n},$$

for some $r<1$ and $M\left(\cdot\right)\geq0$, and where the metric above is the total variation distance between measures (see [MT09] for a detailed explanation of stationarity, ergodicity, etc.). Efficiency of a geometrically ergodic algorithm is measured by the geometric rate of convergence, r. Over a large number of iterations, the geometric rate of convergence is well approximated by the second largest eigenvalue of the transition kernel (the largest eigenvalue being 1, corresponding to the stationary distribution $\pi\left(\cdot\right)$) .

What is crucial for practical MCMC is that any geometrically ergodic reversible Markov chain satisfies a central limit theorem for all functions with finite second moment with respect to $\pi\left(\cdot\right)$. Hence, there exists $\sigma_f^2<\infty$ such that for any arbitrary square-integrable function g,

$$n^{\frac{1}{2}}\left(\hat{g}_{n}-\mathbb{E}_{\pi}\left[g\left(\boldsymbol{X}\right)\right]\right)\Rightarrow\mathcal{N}\left(0,\sigma_{g}^{2}\right), \tag{1.4}$$

where $\hat{g}_n$ is the sample average of the generated Markov chain $\boldsymbol{X}$,

$$\hat{g}_{n}=\frac{1}{n}\sum_{i=1}^{n}g\left(\boldsymbol{X}\left(i\right)\right).$$

The central limit theorem above not only guarantees convergence of the Markov chain, but also provides its standard error, which decreases as $n^{-1/2}$. Thus, for all $k \geq 0$ and $i \geq 0$, we have

$$\operatorname{Corr}\left[g\left(\boldsymbol{X}_k\right), g\left(\boldsymbol{X}_{k+i}\right)\right] = \operatorname{Corr}\left[g\left(\boldsymbol{X}_0\right), g\left(\boldsymbol{X}_i\right)\right].$$

This is simply the autocorrelation at lag i. Using (1.4), we have

$$\begin{aligned} \sigma_g^2 &= \lim_{n\to\infty} n \operatorname{Var}\left[\hat{g}_n\right] \\ &= \operatorname{Var}\left[g\left(\boldsymbol{X}_0\right)\right] + 2\sum_{i=1}^{\infty} \operatorname{Corr}\left[g\left(\boldsymbol{X}_0\right), g\left(\boldsymbol{X}_i\right)\right]. \end{aligned}$$

If one observes that for a stationary chain with independent elements, σ_g^2 is simply $\operatorname{Var}\left[g\left(\boldsymbol{X}_0\right)\right]$. Hence, a measure of efficiency of the Monte Carlo estimate relative to the perfect i.i.d. sample is

$$\frac{\sigma_g^2}{\operatorname{Var}\left[g\left(\boldsymbol{X}_0\right)\right]} = \tau_g = 1 + 2\sum_{i=1}^{\infty} \operatorname{Corr}\left(g\left(\boldsymbol{X}_0\right), g\left(\boldsymbol{X}_i\right)\right),$$

which is often termed as the *integration autocorrelation factor.* This quantity represents the effective number of dependent samples that is equivalent to a single independent sample. Furthermore, one could denote $n^* = n/\tau_g$ as the effective equivalent sample size, had the elements of chain been independent.

If a central limit theorem for $\boldsymbol{X}$ and g exists, then the variance of the estimator $\sum_1^n g\left(\boldsymbol{X}_i\right)/n$ for estimating $\mathbb{E}_\pi\left(g\left(\boldsymbol{X}\right)\right)$ is approximately $\operatorname{Var}_\pi\left(g\left(\boldsymbol{X}\right)\right) \cdot \frac{\tau_g}{n}$. This suggests that the efficiency of Markov chains can be compared by the reciprocal of their integrated autocorrelation times, i.e.,

$$\begin{aligned} e_g\left(\sigma\right) &= \left(\operatorname{Var}_\pi\left(g\left(\boldsymbol{X}\right)\right)\tau_g\right)^{-1} \\ &= \left(\lim_{n\to\infty} n\operatorname{Var}\left(\frac{\sum_1^n g\left(\boldsymbol{X}_i\right)}{n}\right)\right)^{-1}. \end{aligned} \tag{1.5}$$

However, this measure of efficiency is highly dependent on the choice of the function g, and also, estimating the higher order autocorrelations can be very difficult as these involve estimation of an integral with respect to Dirac measure (see Section 1.2 of [PG10]). Hence, for two different Markov chains, different functions g could order their efficiency differently, making this criteria not very robust, unless the specific interest is in a particular function. When one is interested in a whole collection of functionals of the target distribution (i.e., the cdf of a component of interest), then the use of this criterion is problematic.

Furthermore, in the high-dimensional limit, when these algorithms behave like diffusion processes, all efficiency measures are virtually equivalent. The most sensible measure of efficiency for a diffusion process is the speed measure of the given diffusion. Indeed, it turns out that a tractable quantity related to the speed measure is the algorithm's *acceptance rate*, given by α, which is the probability in stationarity that the chain's proposed move is accepted. By the ergodic theorem, the acceptance rate equals the long-term proportion of proposed moves accepted by the chain,

$$\begin{aligned} a &= \int \alpha\left(\boldsymbol{x}, \boldsymbol{y}\right)\pi\left(\boldsymbol{x}\right) q\left(\boldsymbol{x}, \boldsymbol{y}\right) d\boldsymbol{x} d\boldsymbol{y} \\ &= \lim_{n\to\infty} n^{-1} \#\left\{\text{accepted moves}\right\}. \end{aligned}$$

Roberts and Rosenthal [RR01], provides theoretical justification for a commonly used strategy of implementing the multivariate RWM algorithm. The strategy involves estimating the correlation structure of the target distribution (either empirically or perhaps numerically from curvature applications on the target density) and using a proposal distribution for the random walk algorithm to be a scalar multiple of the estimated correlation matrix. In the Gaussian case, if the correlation structure is accurately estimated, then scaling the proposal in each direction proportional to the target scaling can be shown to optimize the algorithm's efficiency.

Consider the continuous time version of (1.5):

$$e_g^{(\infty)}\left(\ell\right) = \left(\lim_{n\to\infty} n\operatorname{Var}\left(\frac{\int_{t=0}^n g\left(\boldsymbol{X}_t\right)}{n}\right)\right)^{-1}.$$

Using (1.5), we have

$$
\begin{aligned}
e_g^{(d)}\left(\ell d^{-\frac{1}{2}}\right) &= \left(\lim_{n\to\infty} \lfloor nd \rfloor \operatorname{Var}\left(\frac{\sum_{i=1}^{\lfloor nd \rfloor} g\left(X_i^{(1)}\right)}{\lfloor nd \rfloor}\right)\right)^{-1} \\
&= \left(\lim_{n\to\infty} \lfloor nd \rfloor \operatorname{Var}\left(\frac{\sum_{i=1}^{\lfloor nd \rfloor} g\left(Z_{i/d}^{d}\right)}{\lfloor nd \rfloor}\right)\right)^{-1} \\
&\approx \left(\lim_{n\to\infty} \lfloor nd \rfloor \operatorname{Var}\left(\frac{\int_0^n g\left(Z_s\right) ds}{n}\right)\right)^{-1} \\
&\approx d^{-1} e_g^{(\infty)}(\ell),
\end{aligned}
$$

so we have

$$\lim_{d\to\infty} d e_g^{(d)}\left(\ell d^{-\frac{1}{2}}\right) = e_g^{(\infty)}(\ell) = h(\ell)\, e_g^L .$$

Another approach to measure the efficiency in Markov chains is with the estimator which is called the first order efficiency, which is the average squared jumping distance for the article (see Section 3 in [RR98]). The first order efficiency of a multi-dimensional Markov chain $\boldsymbol{X}$ with first component $X^{(1)}$ is given by

$$
\begin{aligned}
\text{first order efficiency} &= \mathbb{E}\left[\left\|\left(X_{t+1}^{(1)} - X_t^{(1)}\right)\right\|^2\right] && (1.6)\\
&= 2\left(1 - \operatorname{Corr}\left(X_1^{(1)}, X_0^{(1)}\right)\right) \cdot \operatorname{Var}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right). && (1.7)
\end{aligned}
$$

Note further that for a single component of the target with variance $\sigma_1^2 = \operatorname{Var}\left(X_t^{(1)}\right) = \operatorname{Var}\left(X_{t+1}^{(1)}\right)$, and furthermore, $\mathbb{E}\left[X_{t+1}^{(1)} - X_t^{(1)}\right] = 0$, so

$$
\begin{aligned}
\mathbb{E}\left[\left(X_{t+1}^{(1)} - X_t^{(1)}\right)^2\right] &= \operatorname{Var}\left[X_{t+1}^{(1)} - X_t^{(1)}\right] \\
&= 2\sigma_t^2\left(1 - \operatorname{Corr}\left[X_t^{(1)}, X_{t+1}^{(1)}\right]\right).
\end{aligned}
$$

Thus, maximizing (1.6) amounts to minimizing the lag-1 autocorrelations in the Markov chain $\{\boldsymbol{X}_t\}_{t\geq 1}$.

1.3 Roberts, Gelman, and Gilks (1997): The Paper that Started it All

1.3.1 Introduction to the optimal scaling problem for i.i.d. targets

The line of research concerning weak convergence of Metropolis algorithms to tractable diffusions was initiated by Roberts, Gelman, and Gilks in 1997. Their paper ([RGG97]) is the first paper to discuss theoretical results about the optimal scaling for Metropolis algorithms with Gaussian proposals. The paper establishes that for high-dimensional target distributions formed of i.i.d. components (1.8), the acceptance rate optimizing the efficiency of the process approaches 0.234 as $d \to \infty$. The i.i.d. components assumption for the d-dimensional target density amounts to assuming that the density has the form

$$\pi\left(\boldsymbol{x}^{(d)}\right) = \prod_{i=1}^{d} f\left(x_i\right), \tag{1.8}$$

where $f(\cdot)$ is positive, continuous over $\mathbb{R}$ and also belongs to C^2 (continuous functions with continuous second derivatives), and $\boldsymbol{x}^{(d)} \in \mathbb{R}^d$. As the interest here is in the infinite-dimension process obtained in the limit as $d \to \infty$, one needs need to introduce a time-scaling factor in order to prevent the algorithm from stopping as the size of the dimension increases. By considering suitably regular sequences of canonical target densities and rescaling the proposal density by a factor of d^{-1}, the authors manage to obtain a weak convergence (convergence of distributions in the Skorokhod topology) result for the sequence of algorithms restricted to a fixed finite set of components to a Langevin diffusion. A Langevin diffusion for a multivariate probability density function π, is the unique (up to a speed measure) non-explosive diffusion which is reversible with respect to π.

In particular, the authors consider a proposal scaling of the form $\sigma^2(d) = \ell^2/d$, where ℓ is a scaling parameter to be chosen by the practitioners. This amounts to updating the algorithm every d^{-1} time units, allowing the algorithm to update more quickly, and allows one to consider a non-trivial limiting process when $d \to \infty$. In the limit, the jump sizes become increasingly smaller and more frequent, causing the limiting process of the algorithm to be characterized by a continuous sample path, and ultimately, a diffusion. In order to preserve the Markov property for the limiting diffusion, it is necessary to use the memoryless property of the exponential distribution. More specifically, one considers the new sped up version of the RWM algorithm where the jumps occur in accordance to a Poisson process of rate d. The main result of the paper is that for each fixed one-dimensional component of the Markov chain $\{\boldsymbol{X}_0, \boldsymbol{X}_1, \dots\}$, where the chain is updated using RWM algorithm, the one-dimensional process converges weakly (in the Skorokhod topology) to the appropriate Langevin diffusion.

Furthermore, it was believed that the result would often hold in various regular and general circumstances. The authors are able to transform the conclusion achieved in the weak convergence of the limiting process into a statement about the efficiency as a function of the acceptance rate. It is found that the optimal choice of the scaling parameter σ (as a function of parameter ℓ and dimension d) is that which maintains an acceptance rate of 0.234. A very striking, and indeed counter-intuitive feature that will be discussed more below, is that low-dimensional algorithms have higher optimal acceptance rates than higher-dimensional ones. An acceptance rate as small as 23% indicates that high-dimensional algorithms should remain idle at 77% of the time in order to obtain the fastest mixing chain, which is indeed an unexpected result for many practitioners.

1.3.2 Gaussian proposals and the Metropolis algorithm

We recall, the Gaussian proposal density for the RWM on $\mathbb{R}^d$ is given by

$$q\left(\boldsymbol{x}^{(d)}, \boldsymbol{y}^{(d)}\right) = \frac{1}{(2\pi\sigma^2 d)^{d/2}} \exp\left\{-\frac{1}{2\sigma^2(d)} \left|\boldsymbol{y}^{(d)} - \boldsymbol{x}^{(d)}\right|^2\right\},$$

which produces a Markov chain $\boldsymbol{X}^{(d)} = \{X_0, X_1, X_2, \dots\}$, where each $\boldsymbol{X}^{(d)}(t)$ is chosen according to a rule. The rule is as follows: for given $\boldsymbol{X}^{(d)}(t)$, consider a new proposal $\boldsymbol{Y}^{(d)}(t+1) \sim q\left(\boldsymbol{X}^{(d)}(t), \boldsymbol{Y}^{(d)}(t)\right)$, which is accepted with probability

$$\alpha\left(\boldsymbol{X}^{(d)}(t), \boldsymbol{Y}^{(d)}(t+1)\right) = \min\left(1, \frac{\pi\left(\boldsymbol{Y}^{(d)}(t+1)\right)}{\pi\left(\boldsymbol{X}^{(d)}(t)\right)}\right), \tag{1.9}$$

and in the case of rejection, $\boldsymbol{X}^{(d)}(t+1) = \boldsymbol{X}^{(d)}(t)$. The Markov chain $\left\{\boldsymbol{X}^{(d)}(t)\right\}_{t\geq 0}$ is reversible with respect to π, and is π-irreducible, and hence ergodic. Thus, it indeed converges to the target density π.

Remark 1.1. Observe that although the arguments of the acceptance function (1.9) are multidimensional, the values of the function are scalars. Hence, at any given time, the algorithm either accepts all the proposed moves for the d components, or rejects them. Therefore, as d increases, and thus the number of proposed moves increases, which are independent for each component, if one were to use the constant proposal scaling σ^2, then it becomes increasingly probable to obtain a rejection for at least one of the components (and thus reject entirely the proposed move). It is for this reason that the authors choose to consider a proposal scaling that is inversely proportional to the dimension. Algorithms which consider proposals for a fixed set of coordinates at a time are an interesting variation of the Metropolis algorithm, examples of which are the Metropolis-within-Gibbs, and also the Gibbs algorithm. These are considered in Section 2.4.

1.3.3 Conditions on target distribution f

Along with the assumption of the symmetric product form of the target density (1.8), the authors also fix certain conditions on f. These conditions are more stringent than necessary, and a discussion of relaxation of assumptions is discussed later. First, assume that f'/f is Lipschitz continuous, and we have the moment conditions

$$\mathbb{E}_f\left[\left(\frac{f'(X)}{f(X)}\right)^8\right] \equiv M < \infty \qquad \text{and} \qquad \mathbb{E}_f\left[\left(\frac{f''(X)}{f(X)}\right)^4\right] < \infty. \tag{1.10}$$

We let C^2 denote the space of $\mathbb{R}$-valued functions with continuous second derivatives, and let $\sigma^2(d) = \ell^2(d-1)^{-1}$, for $d \in \mathbb{Z}$. Since the target distribution is composed of identical and independent components, one can study any particular component of interest without loss of generality. For ease, the authors choose to study the first component by defining a new process $U_t^{(d)}$ which consists entirely of the first component of the Metropolis induced Markov chain $\boldsymbol{X}^{(d)}$ sped up by a factor of d, i.e., $U_t^{(d)} = X_{[td]}^{(1)}$. It has been parametrized to make jumps of size $\mathcal{O}\left(\sigma\left(d\right)\right) = \mathcal{O}\left(d^{-1/2}\right)$ every d^{-1} time units. In finite dimensions, the RWM algorithm contains d components, which are constantly interacting with one another. Therefore, there is no reason a priori to believe that in finite dimensions, any individual component, or even subset of components, should behave like a Markov chain. However, when $d \to \infty$, each individual component does behave like a Markov chain, and due to the independent structure of all the components, one can study the asymptotic behavior of a finite collection of components as well.

1.3.4 Theorems and Corollaries

The main result is given in the following theorem.

Theorem 1.2. (Theorem 1.1 in [RGG97]) *Suppose the conditions (1.10) hold, and consider a RWM algorithm with proposal distribution $\boldsymbol{Y}^{(d)} \sim \mathcal{N}\left(\boldsymbol{x}^{(d)}, \frac{\ell^2}{d} I_d\right)$ applied to the target density as in (1.8). Consider the process $\left\{U_t^{(d)}\right\}_{t\geq 0}$ and let $\boldsymbol{X}^{(d)}\left(0\right)$ be distributed according to the target density π. Then,*

$$U_t^{(d)} \Rightarrow U_t \qquad \text{as } d \to \infty,$$

where U_t satisfies the Langevin SDE

$$dU_t = \left(h\left(\ell\right)\right)^{\frac{1}{2}} dB_t + h\left(\ell\right) \frac{f'\left(U_t\right)}{2f\left(U_t\right)} dt, \tag{1.11}$$

where the speed measure $h\left(\ell\right)$ follows

$$h\left(\ell\right) = 2\ell^2 \Phi\left(-\frac{\ell\sqrt{I}}{2}\right),$$

(Φ being the standard normal cdf), and

$$I \equiv \mathbb{E}_f\left[\left(\frac{f'\left(X\right)}{f\left(X\right)}\right)^2\right].$$

Equivalently, by unit speed measure, we can write $U_t = V_{h(\ell)t}$ where V is the Langevin diffusion:

$$dV_t = dB_t = \frac{f'\left(V_t\right)}{2f\left(V_t\right)} dt.$$

It is thus clear that the most "efficient" asymptotic diffusion has the largest speed measure, as this corresponds to the fastest convergence of our Markov chain $X_{[dt],1}^n$. Furthermore, since U^d is produced by speeding up time by a factor of d, the complexity of the algorithm is d.

Remark 1.3. Since $h\left(\ell\right) = \ell^2 \times 2\Phi\left(-\frac{\sqrt{I}\ell}{2}\right)$ and $I = \mathbb{E}_f\left[\left(\left(\log f\left(X\right)\right)'\right)^2\right]$, it follows that

$$h\left(\ell\right) = \ell^2 \times A\left(\ell\right),$$

where $A\left(\ell\right)$ is the limiting overall acceptance rate of the algorithm, i.e., the proportion of proposed Metropolis moves ultimately accepted. This leads to the fact that the formulation of the maximization problem of the limiting diffusion's speed measure can be written entirely in terms of the algorithm's acceptance rate:

$$h\left(\ell\right) = \frac{4}{I}\left(\Phi^{-1}\left(A\left(\ell\right)\right)\right)^2 A\left(\ell\right).$$

Corollary 1.4. (Corollary 1.2 in [RGG97]) *Denote the average acceptance rate of the RWM in d dimensions as follows,*

$$\begin{aligned} a\left(d,\ell\right) &= \mathbb{E}\left[\min\left(1, \frac{\pi\left(\boldsymbol{Y}^{(d)}\right)}{\pi\left(\boldsymbol{X}^{(d)}\right)}\right)\right] \\ &= \int\int \pi\left(\boldsymbol{x}^{(d)}\right) \alpha\left(\boldsymbol{x}^{(d)}, \boldsymbol{y}^{(d)}\right) q\left(\boldsymbol{x}^{(d)}, \boldsymbol{y}^{(d)}\right) d\boldsymbol{x}^{(d)} d\boldsymbol{y}^{(d)}. \end{aligned}$$

Then,

$$\lim_{d\to\infty} a(d,\ell) = a(\ell),$$

and $h(\ell)$ is maximized by

$$\ell = \hat{\ell} = 2.38 I^{-\frac{1}{2}}.$$

Furthermore,

$$h\left(\hat{\ell}\right) = 1.3 I^{-1}.$$

Note 1.5. The optimal value of ℓ is scaled by $I^{-1/2}$, rather than the standard deviation of the target density as is often suggested. However, in situations where f is Gaussian, it can be easily verified that $I \equiv \sigma_f^{-2}$. Hence, I is in fact the variance of the derivative of the log density of f. In general, I is a measure of the roughness of the distribution f, and thus, it is negatively related with the optimal value $\hat{\ell}$.

Note 1.6. Although the Theorem in the paper is only proven for univariate components, the proof provides a stronger statement which sufficiently shows for the case for integers $c > 1$, the process consisting of the first c components of $\boldsymbol{X}_{[dt]}$ converges to a collection of c independent processes, each distributed according to (1.2).

1.3.5 Lemmas and Proofs

In order to prove the above theorem, the paper provides a uniform convergence result of the generator G_n to G for a suitably large class of real valued functions V, where G is the generator of the limiting one-dimensional Langevin diffusion, where

$$GV(x) = h(\ell)\left[\frac{1}{2}V''(x) + \frac{1}{2}\frac{d}{dx}(\log f)(x)V'(x)\right],$$

and the discrete time generator of $\boldsymbol{X}^n$,

$$G_d V\left(\boldsymbol{x}^{(d)}\right) = d\mathbb{E}_{\boldsymbol{Y}^{(d)}}\left[\left(V\left(\boldsymbol{Y}^{(d)}\right) - V\left(\boldsymbol{x}^{(d)}\right)\right)\left(1 \wedge \frac{\pi\left(\boldsymbol{Y}^{(d)}\right)}{\pi\left(\boldsymbol{x}^{(d)}\right)}\right)\right].$$

The choice of the class of functions, for which V is a member is important, as it should be general enough to encompass the necessary cases, but it also must make sense in the definition of the discrete time generator of $\boldsymbol{X}^{(d)}$. By virtue of working in the Skorokhod topology (see sectionB.2 for a brief introduction about the Skorokhod topology), it is acceptable to treat G_d as a continuous time generator if we consider it as a process with jumps at times of a Poisson process at rate d.

Then, by noting that the putative diffusion limit is Markov, and the sequence of approximations $\{U^{(d)}\}_{d\geq 1}$, while not Markov, are embedded in the sequence of Markov processes $\left\{\boldsymbol{Z}^{(d)}\right\}_{d\geq 1}$, with

$$\boldsymbol{Z}_t^{(d)} = \left(X_{[dt],1}^{(d)}, \dots, X_{[dt],d}^{(d)}\right),$$

so $U^{(d)}$ is the first component of $\boldsymbol{Z}^{(d)}$.

The following sequence of sets is then defined: $\left\{F_d \subset \mathbb{R}^d\right\}_{d>1}$ with

$$F_d = \left\{\left|R_d(x_2,\dots,x_d) - I\right| < d^{-\frac{1}{8}}\right\} \cap \left\{\left|S_d(x_2,\dots,x_d) - I\right| < d^{-\frac{1}{8}}\right\},$$

where

$$R_d(x_2,\dots,x_d) = \frac{1}{d-1}\sum_{i=2}^{d}\left[(\log f(x_i))'\right]^2, \quad \text{and} \quad S_d(x_2,\dots,x_d) = -\frac{1}{d-1}\sum_{i=2}^{d}\left[(\log f(x_i))''\right].$$

With the above sequences defined, the following Lemma is obtained.

Lemma 1.7. (Lemma 2.2 in [RGG97]**)** *For fixed t,*

$$\mathbb{P}\left[\boldsymbol{Z}_s^{(d)} \in F_d,\, 0 \leq s \leq t\right] \overset{d\to\infty}{\longrightarrow} 1.$$

Proof. Since the sequence starts in stationarity, with π being stationary, we have $\boldsymbol{Z}_0^{(d)} \sim \pi$ and $\boldsymbol{Z}_s^{(d)} \sim \pi$. Therefore,

$$\mathbb{P}\left[\boldsymbol{Z}_s^{(d)} \notin F_d, \text{ for some } 0 \leq s \leq t\right] \leq t d \mathbb{P}_\pi \left[\boldsymbol{Z} \notin F_d\right].$$

Furthermore, since $\mathbb{E}\left[R_d\left(X_2, \ldots, X_d\right)\right] = 1$ according to distribution π, we have by the weak law of large numbers, for all $\epsilon > 0$

$$\mathbb{P}_\pi \left\{\left|R\left(\boldsymbol{Z}\right) - I\right| > \epsilon\right\} \xrightarrow{d\to\infty} 0.$$

Hence, by Markov's inequality and the first assumption (1.10), we have

$$\mathbb{P}_\pi \left[\boldsymbol{Z} \notin F_d\right] \leq \mathbb{E}_\pi \left[\left(R_d\left(\boldsymbol{Z}\right) - I\right)^4\right] d^{1/2} \leq \frac{3M}{\left(d-1\right)^{3/2}}.$$

Therefore,

$$\mathbb{P}\left[\boldsymbol{Z}_s^{(d)} \in \left\{\left|R_d - I\right| < d^{\frac{1}{8}}\right\}, 0 \leq s \leq t\right] \to 1.$$

The proof of

$$\mathbb{P}\left[\boldsymbol{Z}_s^{(d)} \in \left\{\left|S_d\left(x_2, \ldots, x_d\right) - I\right| < d^{-\frac{1}{8}}\right\}, 0 \leq s \leq t\right] \to 1$$

follows from the second assumption in (1.10). □

Proposition 1.8. (Proposition 2.3 in [RGG97]) *The function $g\left(x\right) = \min\left(1, e^x\right)$ is Lipschitz with coefficient* 1*, that is,*

$$\left|g\left(x\right) - g\left(y\right)\right| \leq \left|x - y\right|,$$

for all $x, y \in \mathbb{R}$.

Lemma 1.9. (Proposition 2.4 in [RGG97]) *Let*

$$W_d\left(x_i\right) = \sum_{i=2}^{d}\left[\frac{\left(\log f\left(x_i\right)\right)''}{2}\left(Y_i - x_i\right)^2 + \frac{\ell^2}{2\left(d-1\right)}\left(\log f\left(x_i\right)\right)'^2\right],$$

where $Y_i \sim \mathcal{N}\left(x_i, \ell^2/d\right)$ are i.i.d. for all $i = 2, \ldots, d$. Then

$$\sup_{\boldsymbol{x}^{(d)} \in F_d} \mathbb{E}\left[\left|W_d\right|\right] \xrightarrow{d\to\infty} 0.$$

Proof. Observe,

$$\begin{aligned}\mathbb{E}\left[\left|W_d\right|\right]^2 &\leq \mathbb{E}\left[W_d^2\right] \\ &= \frac{1}{4\left(d-1\right)^2}\left(\sum_{i=2}^{d}\left(\log f\left(x_i\right)\right)'' + \left(\left(\log f\left(x_i\right)\right)'\right)^2\right)^2 + \frac{2}{4\left(d-1\right)^2}\sum_{i=2}^{d}\left(\log f\right)''^2\left(x_i\right).\end{aligned}$$

For $\boldsymbol{x}^{(d)} \in F_d$,

$$\left|\sum_{i=1}^{d}\frac{\left(\log f\left(x_i\right)\right)'' + \left(\left(\log f\left(x_i\right)\right)'\right)^2}{2\left(d-1\right)}\right| \leq d^{-1/8}.$$

Since $\left(\log f\right)''$ is bounded, $\mathbb{E}\left[\left|W_d\right|\right]^2 \to 0$ uniformly for $\boldsymbol{x}^{(d)} \in F_d$. □

Proposition 1.10. (Proposition 2.5 in [RGG97]) *If $A \sim \mathcal{N}\left(\mu, \sigma^2\right)$, then*

$$\mathbb{E}\left[\min\left(1, e^A\right)\right] = \Phi\left(\frac{\mu}{\sigma}\right) + \exp\left(\mu + \sigma^2/2\right)\Phi\left(-\sigma - \frac{\mu}{\sigma}\right),$$

where $\Phi\left(\cdot\right)$ is the standard normal cumulative distribution.

Lemma 1.11. (Lemma 2.6 in [RGG97]) *For $V \in C_c^\infty$,*

$$\limsup_{d\to\infty}\sup_{x_1\in\mathbb{R}} d\left|\mathbb{E}_{Y_1}\left[V\left(Y_1\right) - V\left(x_1\right)\right]\right| < \infty,$$

with $Y_1 \sim \mathcal{N}\left(x_1, \sigma_d^2\right)$.

Proof.

$$V\left(Y_1\right)-V\left(x_1\right)=V'\left(x_1\right)\left(Y_1-x_1\right)+\frac{1}{2}V''\left(Z_1\right)\left(Y_1-x_1\right)^2,$$

for some $Z_1 \in (x_1, Y_1)$ or (Y_1, x_1). Thus,

$$d\mathbb{E}_{Y_1}\left[V\left(Y_1\right)-V\left(x_1\right)\right]\leq\frac{Kd\ell^2}{d-1},$$

where K is an upper bound for V''. □

Lemma 1.12. (Lemma 2.7 in [RGG97]) *Suppose $V \in C_c^\infty$ is a function of the first component of $\boldsymbol{Z}^d$. Then*

$$\sup_{\boldsymbol{x}^{(d)}\in F_d}\left|G_dV\left(\boldsymbol{x}^{(d)}\right)-GV\left(x_1\right)\right|\xrightarrow{d\to\infty}0.$$

Proof. After decomposing the proposal $\mathbf{Y}^{(d)}$ into $\left(Y_1, \mathbf{Y}^{(d)-}\right)$ as

$$G_dV\left(\boldsymbol{x}^{(d)}\right)=d\mathbb{E}_{Y_1}\left[\left(V\left(\mathbf{Y}^{(d)}\right)-V\left(\boldsymbol{x}^{(d)}\right)\right)\mathbb{E}_{\mathbf{Y}^{(d)-}}\left[\min\left(1,\frac{\pi\left(\mathbf{Y}^{(d)}\right)}{\pi\left(\boldsymbol{x}^{(d)}\right)}\right)\right]\right],$$

one can focus on the inner expectation of this term. Let the inner expectation be denoted by $E\left(Y_1\right)$, so denoting with $\varepsilon\left(Y_1\right)=\log\left(\frac{f(Y_1)}{f(x_1)}\right)$, we have

$$\begin{aligned}E\left(Y_1\right) &= \mathbb{E}\left[\min\left(1,\exp\left\{\varepsilon\left(Y_1\right)+\sum_{i=2}^{d}\left(\log f\left(Y_i\right)-\log f\left(x_i\right)\right)\right\}\right)\right],\\ &= \mathbb{E}\left[\min\left(1,\exp\left\{\varepsilon\left(Y_1\right)+\sum_{i=2}^{d}\left(\log f\left(x_i\right)\right)'\left(Y_i-x_i\right)\right.\right.\right.\\ &\quad\left.\left.\left.+\frac{1}{2}\left(\log f\left(x_i\right)\right)''\left(Y_i-x_i\right)^2+\frac{1}{6}\left(\log f\left(Z_i\right)\right)'''\left(Y_i-x_i\right)^3\right]\right\}\right],\end{aligned}$$

for some $Z_i \in (x_i, Y_i)$ or (Y_i, x_i). By Proposition 1.8,

$$\left|E\left(Y_1\right)-\mathbb{E}\left[\min\left(1,\exp\left\{\varepsilon\left(Y_1\right)+\sum_{i=2}^{d}\left[\left(\log f\left(x_i\right)\right)'\left(Y_i-x_i\right)-\frac{\ell^2}{2\left(d-1\right)}\left(\left(\log f\left(x_i\right)\right)'\right)^2\right]\right\}\right)\right]\right|$$

$$\leq\mathbb{E}\left[\left|W_d\right|\right]+\sup_{z\in\mathbb{R}}\left|\left(\log f\left(z\right)\right)'''\right|\frac{1}{6\left(d-1\right)^{\frac{1}{2}}}\frac{4\ell^3}{\left(2\pi\right)^{\frac{1}{2}}},$$

where W_d is as defined in Lemma 1.9. Furthermore, the term

$$\sup_{\boldsymbol{x}^{(d)}\in F_d}\left|E\left(Y_1\right)-\mathbb{E}\left[\min\left(1,\exp\left\{\varepsilon\left(Y_1\right)+\sum_{i=2}^{d}\left[\left(\log f\left(x_i\right)\right)'\left(Y_i-x_i\right)-\frac{\ell^2}{2\left(d-1\right)}\left(\left(\log f\left(x_i\right)\right)'\right)^2\right]\right\}\right)\right]\right|.$$

Denote this term by $\varphi\left(d\right)$, which converges to zero as $d\to\infty$.

However,

$$\varepsilon\left(Y_1\right)+\sum_{i=2}^{d}\left[\left(\log f\left(x_i\right)\right)'\left(Y_i-x_i\right)-\frac{\ell^2}{2\left(d-1\right)}\left(\left(\log f\left(x_i\right)\right)'\right)^2\right],$$

is distributed according $\mathcal{N}\left(\varepsilon\left(Y_1\right)-\ell^2R_d/2,\ell^2R_d\right)$, so that by Proposition 1.10,

$$\left|E-\mathbb{E}\left[\min\left(1,\exp\left\{\varepsilon\left(Y_1\right)+\sum_{i=2}^{d}\left[\left(\log f\left(x_i\right)\right)'\left(Y_i-x_i\right)-\frac{\ell^2}{2\left(d-1\right)}\left(\left(\log f\left(x_i\right)\right)'\right)^2\right]\right\}\right)\right]\right|$$

$$\leq\mathbb{E}\left[\left|W_d\right|\right]+\sup_{z\in\mathbb{R}}\left|\left(\log f\left(z\right)\right)''\right|\frac{1}{6\left(d-1\right)^{\frac{1}{2}}}\frac{4\ell^3}{\left(2\pi\right)^{\frac{1}{2}}}.$$

As was the case two equations ago, the above converges to zero as $d \to \infty$. However,

$$\varepsilon\left(Y_1\right)+\sum_{i=2}^{d}\left[\left(\log f\left(x_i\right)\right)'\left(Y_i-x_i\right)-\frac{\ell^2}{2\left(d-1\right)}\left(\left(\log f\left(x_i\right)\right)'\right)^2\right]$$

is distributed according to $\mathcal{N}\left(\varepsilon\left(Y_1\right)-\ell^2 R_d/2, \ell^2 R_d\right)$, so that Proposition 1.10,

$$\mathbb{E}\left[\min\left(1, \exp\left\{\varepsilon\left(Y_1\right)+\sum_{i=2}^{d}\left[\left(\log f\left(x_i\right)\right)'\left(Y_i-x_i\right)-\frac{\ell^2}{2\left(d-1\right)}\left(\left(\log f\left(x_i\right)\right)'\right)^2\right]\right\}\right)\right]$$

$$=\Phi\left(R_d^{-\frac{1}{2}}\left(\ell^{-1}\varepsilon\left(Y_1\right)-\frac{\ell R_d}{2}\right)\right)+e^{\varepsilon\left(Y_1\right)}\Phi\left(-\frac{\ell R_d^{\frac{1}{2}}}{2}-\varepsilon\left(Y_1\right)R_d^{-\frac{1}{2}}\ell^{-1}\right).$$

Denote the last equation by $M\left(\varepsilon\right)$.

Therefore,

$$\sup_{\boldsymbol{x}^{(d)}\in F_d}\left|G_d V-d\mathbb{E}\left[\left(V\left(Y_1\right)-V\left(x_1\right)\right)M\left(\log\left(\frac{f\left(Y_1\right)}{f\left(x_1\right)}\right)\right)\right]\right|\le\varphi\left(d\right)d\mathbb{E}\left[\left|V\left(Y_1\right)-V\left(x_1\right)\right|\right],$$

and the right hand side converges to zero as $d \to \infty$.

Finally, consider the term

$$d\mathbb{E}\left[\left(V\left(Y_1\right)-V\left(x_1\right)\right)M\left(\log\frac{f\left(Y_1\right)}{f\left(x_1\right)}\right)\right].$$

A Taylor expansion of the integrand about x_1 yields

$$\begin{aligned}\left(V\left(Y_1\right)-V\left(x_1\right)\right)M\left(\log\frac{f\left(Y_1\right)}{f\left(x_1\right)}\right) &= \left(V'\left(x_1\right)\left(Y_1-x_1\right)+\frac{1}{2}V''\left(x_1\right)\left(Y_1-x_1\right)^2+\frac{V'''\left(Z_1\right)}{6}\left(Y_1-x_1\right)^3\right)\\ &\quad\times\left[M\left(0\right)+\left(Y_1-x_1\right)M'\left(0\right)\left(\log f\left(x_i\right)\right)'+\frac{1}{2}\left(Y_1-x_1\right)^2T\left(x_1,W_1\right)\right],\end{aligned}$$

where

$$T\left(x_1,W_1\right)=M''\left(\log\frac{f\left(W_1\right)}{f\left(x_1\right)}\right)\left(\left(\log f\left(W_1\right)\right)'\right)^2+\left(\log f\left(W_1\right)\right)''M'\left(\log\frac{f\left(W_1\right)}{f\left(x_1\right)}\right),$$

and where $Z_1, W_1 \in [x_1, Y_1]$ or $[Y_1, x_1]$. Since V has compact support, which we denote by S, there exists $K<\infty$ such that $\left|\left(\log f\right)^{(i)}\left(x\right)\right|, \left|V^{(i)}\left(x\right)\right|\le K$, for $x\in S$, $i=1,2,3$, and one can show that M' and M'' are bounded, say by K again. Thus,

$$M\left(0\right)=2M'\left(0\right)=2\Phi\left(-\frac{\ell R_d^{\frac{1}{2}}}{2}\right),$$

so that

$$\begin{aligned}\mathbb{E}\left[d\left(V\left(Y_1\right)-V\left(x_1\right)\right)M\left(\log\frac{f\left(Y_1\right)}{f\left(x_1\right)}\right)\right] &= 2n\Phi\left(-\frac{R_d^{\frac{1}{2}}\ell}{2}\right)\left[\left(\frac{1}{2}V''\left(x_1\right)\right.\right.\\ &\quad\left.\left.+\frac{1}{2}\left(\log f\left(x_i\right)\right)'V'\left(x_1\right)\right]\mathbb{E}\left[\left(Y_1-x_1\right)^2\right]+\mathbb{E}\left[B\left(x_1,Y_1,d\right)\right]\end{aligned}$$

where

$$\mathbb{E}\left[\left|B\left(x_1,Y_1,d\right)\right|\right]\le a_1\left(K\right)d\mathbb{E}\left[\left|Y_1-x\right|^3\right]+a_2\left(K\right)d\mathbb{E}\left[\left|Y_1-x_1\right|^4\right]+a_3\left(K\right)d\mathbb{E}\left[\left|Y_1-x_1\right|^5\right],$$

and a_i are polynomials in K for $i=1,2,3$. Hence, $\mathbb{E}\left[B\left(x_1,Y_1,d\right)\right]$ is uniformly $\mathcal{O}\left(n^{-\frac{1}{2}}\right)$, and thus

$$\sup_{\boldsymbol{x}^{(d)}\in F_d}\left|G_d V\left(\boldsymbol{x}\right)-GV\left(\boldsymbol{x}\right)\right|\xrightarrow{d\to\infty}0.$$

□

1.4 Roberts and Rosenthal (1998): The MALA Algorithm

Results similar to the paper discussed in the previous section can be obtained for other Metropolis-Hastings algorithms. Following the approach of [RGG97], Roberts and Rosenthal (1998) [RR98] demonstrate that the Metropolis-Adjusted Langevin Algorithm (MALA, see 1.2.1), applied to target distributions of the form (1.8), is optimized by fixing the variance of the proposal distribution to be $\mathcal{O}\left(d^{-1/3}\right)$, which yields an overall asymptotically optimal acceptance rate close to 0.574 and requires $\mathcal{O}\left(d^{1/3}\right)$ steps to converge. This compares favorably to the RWM algorithm, which require $\mathcal{O}\left(d\right)$ steps to explore the same class of target measures. The trade off being the need to calculate the gradient of the Markov chain at every iteration, adding to the computational burden.

1.4.1 Target distribution and assumptions

Consider Metropolis-adjusted discrete approximations $\{\boldsymbol{X}_t\}$ to the Langevin diffusion for π as above, with

$$\pi\left(\boldsymbol{x}^{(d)}, d\right) = \prod_{i=1}^{d} f\left(x_i\right) = \prod_{i=1}^{d} \exp\left\{g\left(x_i\right)\right\}. \tag{1.12}$$

Assume $\boldsymbol{X}_0 \sim \pi$, i.e., the chain starts out in stationarity. Furthermore, we assume that g is a C^8 function with derivatives $g^{(i)}$ satisfying

$$\left|g\left(x\right)\right|, \left|g^{(i)}\left(x\right)\right| \leq M_0\left(x\right),$$

for $i \in [1, 8]$ for some polynomial $M_0\left(\cdot\right)$, and that

$$\int_{\mathbb{R}} x^k f\left(x\right) dx < \infty,$$

for $k = 1, 2, 3, \ldots$. To be able apply results stemming from SDE theory, we assume that g' is a Lipschitz function.

To compare the discrete approximations to limiting continuous-time processes, it is helpful to define the discrete approximations as jump processes with exponential holding times. In particular, let $\{J_t\}$ be a Poisson process with rate $d^{1/3}$, and let $\boldsymbol{\Gamma}^{(d)} = \left\{\boldsymbol{\Gamma}_t^{(d)}\right\}_{t\geq 0}$ be the d-dimensional jump process defined by $\boldsymbol{\Gamma}_t^{(d)} = \boldsymbol{X}_{J_t}$, where we set $\sigma_d^2 = \ell^2 n d^{-1/3}$ in the definitions above, with ℓ an arbitrary positive constant. It is assumed throughout that $\{\boldsymbol{X}_t\}$ is non-explosive. Let

$$\alpha\left(\ell, d\right) = \int\int \pi\left(\boldsymbol{x}\right) q\left(\boldsymbol{x}, \boldsymbol{y}\right) \alpha\left(\boldsymbol{x}, \boldsymbol{y}, d\right) d\boldsymbol{x} d\boldsymbol{y} = \mathbb{E}\left(\min\left\{\frac{\pi\left(\boldsymbol{Y}\right) q\left(\boldsymbol{Y}, \boldsymbol{X}\right)}{\pi\left(\boldsymbol{X}\right) q\left(\boldsymbol{X}, \boldsymbol{Y}\right)}, 1\right\}\right), \tag{1.13}$$

denote the π-average acceptance rate of the algorithm generating $\boldsymbol{\Gamma}$.

The first primary result of the paper gives an analytic formula for the asymptotic acceptance probability of the algorithm.

1.4.2 Main Results

Theorem 1.13. *The sequence (1.13) converges as $d \to \infty$,*

$$\lim_{d\to\infty} \alpha\left(\ell, d\right) = \alpha\left(\ell\right),$$

with $\alpha\left(\ell\right) = 2\Phi\left(-K\ell^3/2\right)$, where $\Phi\left(x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$, and

$$K = \sqrt{\mathbb{E}\left(\frac{5g'''\left(X^2\right) - 3g''\left(X\right)^3}{48}\right)} > 0,$$

where the expectation is taken over X having density $f = \exp\left(g\right)$.

The second result of the paper shows that for the target density π as given above and large values of d, a corresponding MALA algorithm for the target π with proposal variances given by $\sigma\left(d\right)$, and $\ell = \sigma_d d^{1/6}$, the speed of the process is given by $h\left(\ell\right) = 2\ell^2\Phi\left(-K\ell^3/2\right)$. Furthermore, the optimal value $\hat{\ell}$ of ℓ which maximizes this speed function is such that the asymptotic acceptance probability $\alpha\left(\hat{\ell}\right) \approx 0.574$.

Theorem 1.14. *Let $\left\{U_t^{(d)}\right\}_{t\geq 0}$ be the process corresponding to the first component of $\mathbf{\Gamma}^{(d)}$. With $d\to\infty$, the process $U_t^{(d)}$ converges weakly (in the Skorokhod topology) to the Langevin diffusion $\boldsymbol{U}_t$, which follows the dynamics of the SDE*

$$d\boldsymbol{U}_t = \left(h\left(\ell\right)\right)^{1/2} dB_t + \frac{1}{2} h\left(\ell\right) \frac{d}{dx} \log \pi_1\left(U_t\right) dt,$$

with $h\left(\ell\right) = 2\ell^2 \Phi\left(-K\ell^3/2\right)$ denoting the speed of the limiting diffusion. Furthermore, $h\left(\ell\right)$ is maximized at the unique value of ℓ for which $\alpha\left(\ell\right) = 0.574$.

Remark 1.15. The theorem further implies that the proposal variances $\sigma^2\left(d\right)$ should be tuned to be approximately $\hat{\ell}^2 d^{-1/3}$, which is the value for which the acceptance probability becomes approximately 0.574.

1.4.3 Simulation Studies

In order to measure efficiency of the algorithm, the authors use the first order efficiency as given by (1.6). We had observed in section 1.2.2 that maximizing this quantity is equivalent to minimizing the lag-1 autocorrelations of the Markov chain. The first simulation studies done by the authors show that the results hold well for symmetric functions f. When f is not symmetric, then the asymptotic picture is more opaque, and results are usually considered on a case by case basic.

1.4.4 Proofs

To prove the main results discussed above, we define generators of the discrete approximation process $\mathbf{\Gamma}^{(d)}$ and of the (first-component) Langevin diffusion process with speed $h\left(\ell\right)$,

$$G_d V\left(\boldsymbol{x}^{(d)}\right) = d^{1/3}\mathbb{E}\left(\left(V\left(\boldsymbol{Y}\right) - V\left(\boldsymbol{x}^{(d)}\right)\right)\left(\min\left\{\frac{\pi\left(\boldsymbol{Y}\right) q\left(\boldsymbol{Y}, \boldsymbol{x}^{(d)}\right)}{\pi\left(\boldsymbol{x}^{(d)}\right) q\left(\boldsymbol{x}^{(d)}, \boldsymbol{Y}\right)}, 1\right\}\right)\right),$$

where the expectation is taken over $\boldsymbol{Y} \sim q\left(\boldsymbol{x}^{(d)}, \cdot\right)$, and

$$GV\left(\boldsymbol{x}^{(d)}\right) = h\left(\ell\right)\left(\frac{1}{2}V''\left(x_1\right) + \frac{1}{2}g'\left(x_1\right)V'\left(x_1\right)\right),$$

where $g\left(x_1\right) = \log f\left(x_1\right)$.

1.5 Discussion

The papers we have so far discussed provided very useful analytical framework for the optimal scaling for two oft-used Metropolis-Hastings algorithms. However, the papers we discussed assumed a fairly stringent assumption on the target density by assuming that it is comprised of i.i.d. components. Thus, we are begged to consider further generalizations of this target density in order to see if the results found earlier are still valid. In their comprehensive survey on MCMC algorithms ([RR98]), Roberts and Rosenthal pose the open question:

> *"Determine the extent to which the above optimal scaling results continue to apply, when when assumption (1.8) is violated."*

The aim of the following two sections is to discuss how optimal scaling results can be obtained for more general forms of the target distribution, and the papers therein partially answer the aforementioned open question. Many of the papers corroborate with the findings in this section however, indicating that the acceptance rate of 0.234 for RWM and 0.574 for MALA are considerably robust to different forms of the the target distribution.

2 Extensions to More General Target Distributions

2.1 Breyers and Roberts (2000): Finite Range Gibbs Random-Field Targets

Following the publication of the results described in the previous section, many researchers attempted to relax the i.i.d. assumption for the target distribution. It was shown by Breyers and Roberts [BR00] that the acceptance rate of 0.234 still holds when applied to suitably behaved target densities with partial correlations of finite range, i..e, when the local correlations are sufficiently small so that no phase transitions occur. The Gibbs measure used as the target distribution for this model is such that the field's correlations decrease exponentially fast as a function of distance. In particular, they consider a Gibbs measure on $\mathbb{R}^{\mathbb{Z}^r}$ which has a density with respect to $\prod_{k\in\mathbb{Z}^r}\mu\left(dx_k\right)$ given by

$$\exp\left\{-\sum_{k\in\mathbb{Z}^r}U_k\left(\boldsymbol{x}\right)\right\},$$

where U_k is a finite-range potential, depending only on a finite number of neighboring terms of k. Markov random fields with significant phase-transition behavior will require algorithms that use proposals with large jumps in order to provide sufficient mixing. This will often yield degenerate algorithms, as the acceptance rate will converge to 0, making the algorithm practically useless.

2.2 Roberts and Rosenthal (2001): & Inhomogeneous Targets

Another validation of the asymptotic optimal acceptance rate equaling 0.234 for the RWM was proven by Roberts and Rosenthal (2001) [RR01] in the case of inhomogeneous target distributions. Specifically, they considered inhomogeneous target distributions of the form

$$\pi\left(\boldsymbol{x}\right)=\prod_{i=1}^{d}C_if\left(C_ix_i\right), \tag{2.1}$$

where the $\{C_i\}$ are themselves i.i.d. for some fixed distribution, and f is again a fixed one-dimensional density satisfying the regularity conditions discussed in (1.10). The components in this target are uncorrelated, but heterogeneously scaled. The target (1.8) used in [RGG97] corresponds to the case where the $\{C_i\}$ are all constant in (2.1). In order to obtain general results for the limiting theory of the RWM algorithm applied to (2.1), it is assumed that C_i have mean 1 and variance $b<\infty$.

Theorem 2.1. (Theorem 5 in [RR01]) *Consider RWM with target density (2.1), and with proposal distribution $\mathcal{N}\left(0,I_d\left(\ell^2/d\right)\right)$, for some $\ell>0$. Consider the time-rescaled process consisting of the first component of the induced Markov chain, i.e., $W_t^{(d)}=C_1X_{[td]}^{(1)}$. As $d\to\infty$, $W_t^d\Rightarrow W_t$, where W_t satisfies the Langevin SDE*

$$dW_t=\frac{1}{2}g'\left(W_t\right)\left(C_1s\right)^2dt+\left(C_1s\right)dB_t,$$

with B_t being a standard Brownian motion and where

$$\begin{aligned}s^2&=2\ell^2\Phi\left(-\ell b^{1/2}I^{1/2}/2\right)\\&=\frac{1}{b}\cdot2\left(\ell^2b\right)\Phi\left(-\left(\ell^2b\right)^{1/2}I^{1/2}/2\right),\end{aligned}$$

with $I=\mathbb{E}_f\left[\left(g'\left(X\right)\right)^2\right]$.

Remark 2.2. When considering a functional of the first coordinate of the algorithm, the efficiency of the algorithm as a function of acceptance rate is the same as that found in [RGG97], but multiplied with a global "inhomogeneous" factor C_1^2/b. For a fixed function of f, the optimal asymptotic efficiency is proportional to C_1^2/bd. In particular, the efficiency of the algorithm is now slowed down by a factor of $b=\mathbb{E}\left(C_i^2\right)/\mathbb{E}\left(C_i\right)^2$, which by Jensen's inequality is greater than or equal to 1, with equality coming only when the factors C_i are all constant, which is equivalent to the form of (1.8).

Remark 2.3. When the values of C_i are known, then instead of using the proposal distribution $\mathcal{N}\left(0,\sigma^2\left(d\right)I_d\right)$, one should use an inhomogeneous proposal which is scaled in proportion to C_i for each component, i..e, we should consider proposals of the form $\mathcal{N}\left(0,\bar{\sigma}^2\left(d\right)\operatorname{diag}\left(C_1,\ldots,C_d\right)\right)$. This would then yield a RWM applied to the form of the target (1.8), since the scaling in the proposal and in the target would cancel each other out. This would alleviate the inefficiency factor of b as discussed above. See Theorem 6 and the subsequent discussion in [RR01] for a more detailed assessment of inhomogeneous proposals for such target distributions.

The results above provide some theoretical justification for a frequently used strategy as suggested by Tierney in [Tie94], which is to use a proposal distribution that is a scalar multiple of the correlation matrix obtained by estimating the correlation structure of the target distribution, perhaps through an empirical estimate based on a sample MCMC run, or numerically through curvature calculations on the target distribution itself. Using inhomogeneous proposals by first obtaining a preliminary estimate of the C_i can thus lead to more efficient Metropolis algorithms.

2.3 Bédard: Various Scaling Terms for Components of the Target

2.3.1 Target distribution with varying scaling terms

In a series of papers ([Béd06, Béd07, Béd08a, Béd08b, Béd10, BR08]), Bédard relaxes the i.i.d. assumption imposed on the target distribution, and considers the case for d-dimensional target distributions with independent components, where each component possesses its own scaling term which may be a function of the dimension of the target distribution. Specifically, she analyzes the following d-dimensional target density:

$$\pi\left(d, \boldsymbol{x}^{(d)}\right) = \prod_{j=1}^{d} \theta_j\left(d\right) f\left(\theta_j\left(d\right) x_j\right). \tag{2.2}$$

This target density does not consist of d identical components, as was the case in the sections above, but rather the scaling term of each component (which is independent of the others) in the target is allowed to depend on the dimension of the target distribution (hence, they are no longer identically distributed). The target distributions considered in the papers above, (1.8) and (2.1) are special cases of (2.2) where the scaling terms are equal in the first case, and identically distributed but independent of the dimension d in the second case.

The paper [Béd07] provides a necessary and sufficient condition on the scaling terms under which the algorithm admits the same limiting process as discussed previously, and hence the asymptotically optimal acceptance rate (AOAR) is again the seemingly ubiquitous 0.234. The condition is that the scaling terms converge at roughly the same rate. This is achieved in the first place by determining the appropriate form for the proposal scaling as a function of d, which is now different from the i.i.d. case. This is an important result for the class of targets (2.2), where an explicit condition is provided [Béd07, Béd08a] for which one can calculate the exact form the optimal scaling for the proposal variance. By verifying L^1 convergence of the generators, she proves that the sequence of stochastic processes formed by the i^{th} component of each Markov chain (rescaled appropriately) converges to a Langevin diffusion process with a certain speed measure. One is then able to obtain the AOAR by optimizing the speed measure of this limiting diffusion.

Although the regularity conditions imposed in [Béd07] are weaker than those imposed in [RGG97], they are still stronger than necessary. All that is required is that f be a C^2 density, $(\log f)'$ be Lipschitz continuous, and $\mathbb{E}\left[\left|f''\left(X\right)/f\left(X\right)\right|^{1+\epsilon}\right] < \infty$ for some $\epsilon > 0$. When obtaining a limiting version of the Metropolis algorithm, one will obtain an infinite-dimensional version of the target density (2.2), and thus a family of scaling vectors $\Theta^2\left(d\right)$ as $d \to \infty$. Although the d target components are independent, they are not identically distributed. As $d \to \infty$, there will be some scaling terms that will be repeated an infinite number of times, whereas other components that will be non-replicated. Assume that there are $0 < m < \infty$ terms that appear infinitely often in the limit, and $n < \infty$ terms that are distinct from the rest. Furthermore, let the first $n+m$ components of the vector $\Theta^2\left(d\right)$ consist of the n non-replicated scaling terms, followed by the m different scaling terms that are replicated. Lastly, assume that within these two groups, distinct and repeating families of scaling terms, the components are arranged according to an asymptotically increasing order.

For the first n components, the j^{th} scaling term will be denoted by K_j/d^{λ_j}, and then denote the following m terms by $\frac{K_{n+i}}{d^{\gamma_i}}$for $i = 1, \ldots, m$, where γ_i and λ_j reside in $\mathbb{R}$, and K_j resides in $\mathbb{R}_+$. While the components in the m group of replicated terms appear infinitely often in the limit, for any fixed and finite value of d they may not each be replicated the same amount of times. In order to determine the proportion of $\Theta^2\left(d\right)$ occupied by any particular term in the replicated group, it is convenient to define the cardinality of the sets $\mathcal{J}\left(i, d\right)$:

$$c\left(\mathcal{J}\left(i, d\right)\right) = \#\left\{j \in \{1, \ldots, d\} : \theta_j^2 = \frac{K_{n+i}}{d^{\gamma_i}}\right\}, \quad \text{for} \quad i = 1, \ldots, m.$$

To accommodate the feature that each of these scaling terms appear infinitely often in the limit, it follows that $c\left(\mathcal{J}\left(i, d\right)\right)$ is some polynomial function of d satisfying $\lim_{d\to\infty} c\left(\mathcal{J}\left(i, d\right)\right) = \infty$.

In order to study the limiting distribution of each component forming the d-dimensional Markov process, one must first set the scaling term of the target component of interest, to 1, i.e., $\theta_{i^*}\left(d\right) = 1$. This can be done by applying a linear transformation to the target distribution.

2.3.2 Proposal scaling and time-acceleration factor

In this present scenario, the proposal variance needs to be selected so that it converges to zero faster than any of the individual scaling components, $\theta_i^2\left(d\right)$. It is shown in [Béd07] that the optimal form for the proposal scaling takes the form $\sigma^2(d) = \ell^2/d^\alpha$, where α is the smallest number satisfying

$$\lim_{d\to\infty} \frac{d^{\lambda_1}}{d^\alpha} < \infty \qquad \text{and} \qquad \lim_{d\to\infty} \frac{d^{\gamma_i} c(\mathcal{J}(i,d))}{d^\alpha} < \infty, \tag{2.3}$$

for $i = 1, \ldots, m$. For this to hold, it follows that at least one of the $m+1$ limits above must converge to some positive constant, and the rest converge to zero. Since the scaling term of the component of interest is taken to be equal to 1, the largest possible form for the proposal scaling is $\sigma^2 = \ell^2$, and thus it will not diverge as $d \to \infty$.

Note 2.4. When we proceed to prove the convergence of the limiting process of the Metropolis algorithm to a Langevin diffusion, it will be useful to consider the reciprocal of equation (2.3). This equivalent condition is

$$\begin{aligned} &\lim_{d\to\infty} \frac{K_1}{d^{\lambda_1}} \left(\frac{d^{\lambda_1}}{K_1} + \cdots + \frac{d^{\lambda_n}}{K_n} + \cdots \right. \\ &\left. c(\mathcal{J}(i,d)) \frac{d^{\gamma_1}}{K_{n+1}} + \cdots + c(\mathcal{J}(m,d)) \frac{d^{\gamma_m}}{K_{n+m}} \right) \quad = \quad \infty. \end{aligned} \tag{2.4}$$

Since all the distinct terms in the first n positions appear only a finite number of times, it follows that for (2.4) to be satisfied, there must exist at least one $i \in \{1, \ldots, m\}$ such that $\lim_{d\to\infty} \frac{K_1}{d^{\lambda_1}} c(\mathcal{J}(i,d)) \frac{d^{\gamma_i}}{K_{n+i}} = \infty$. Furthermore, this stipulates that $\lim_{d\to\infty} \frac{d^{\lambda_1}}{d^\alpha} = 0$ and hence the scaling term $\theta_1^{-2}(d)$ is immaterial for the choice of α.

As was the case in the papers above, a time-accelerating adjustment is necessary in order to ensure the limiting process of the Metropolis algorithm is non-trivial. For this purpose, a new sped-up continuous time version of the original algorithm is introduced, paramterized to make jumps every $d^{-\alpha}$ time units, i.e., $\boldsymbol{Z}^d(t) = \left(X_1^{(d)}([d^\alpha t]), \ldots, X_d^{(d)}([d^\alpha t]) \right)$, where $[\cdot]$ is the integer function. As was the case earlier, the only way to preserve the Markov property of the discrete-time process of the Metropolis algorithm while making it a continuous process is to resort to the memoryless property of the exponential distribution. By allowing the process to jump according to a Poisson process with rate d^α, the new sped up process $\boldsymbol{Z}^d(t)$ moves about d^α times in every time unit, and furthermore, behaves asymptotically like a continuous Markov process.

2.3.3 Optimal proposal and scaling values

As has been noted in the preceding paragraphs, the major difference between the target (2.2) and the targets considered earlier is the inclusion of the various scaling terms. Since the preceding sections proved results for targets where the scaling terms are all identical or identically distributed, the question that is of interest is how big a discrepancy between these various scaling terms is required in order to effect the limiting behavior of the algorithm. The results following show that if none of the first n scaling terms are significantly different than the remaining scaling terms, then the limiting behavior is unaffected.

Theorem 2.5. (Theorem 1 in [Béd07]) *Consider a RWM algorithm with proposals $\boldsymbol{Y}^{(d)}(t+1) \sim \mathcal{N}(0, \ell^2 I_d/d^\alpha)$, with α satisfying (2.3), applied to a target distribution of the form (2.2), which satisfies the specified conditions on f. Consider the i^{*th} component of the process $\left\{\boldsymbol{Z}^{(d)}\right\}_{t\geq 0}$, i.e., $\left\{Z_{i^*}^{(d)}(t)\right\}_{t\geq 0} = \left\{X_{i^*}^{(d)}([d^\alpha t])\right\}_{t\geq 0}$, and let the algorithm start in stationarity (i.e., $\boldsymbol{X}^{(d)}(0) \sim \pi$).*

Then $\left\{Z_{i^*}^{(d)}(t)\right\}_{t\geq 0} \Rightarrow \{Z(t)\}_{t\geq 0}$, where $Z(0)$ is distributed according to the density f and $\{Z(t)\}_{t\geq 0}$ satisfies the Langevin SDE

$$dZ(t) = v(\ell)^{\frac{1}{2}} dB_t + \frac{h(\ell) \nabla \log f(Z(t))}{2} dt,$$

if and only if

$$\lim_{d\to\infty} \frac{d^{\lambda_1}}{\sum_{j=1}^n d^{\lambda_j} + \sum_{i=1}^m c(\mathcal{J}(i,d)) d^{\gamma_i}} = 0. \tag{2.5}$$

Here $h(\ell) = 2\ell^2 \Phi\left(-\ell\sqrt{E_R}/2\right)$, and

$$E_R = \lim_{d\to\infty} \sum_{i=1}^m \frac{c(\mathcal{J}(i,d))}{d^\alpha} \frac{d^{\gamma_i}}{K_{n+1}} \mathbb{E}\left[\left(\frac{f'(X)}{f(X)}\right)^2\right].$$

Furthermore, we have $\lim_{d\to\infty} a(d,\ell) = 2\Phi\left(-\ell\sqrt{E_R}/2\right) \equiv a(\ell)$. The speed measure $v(\ell)$ is maximized at the unique value $\hat{\ell} = 2.38/\sqrt{E_R}$, where $a\left(\hat{\ell}\right) = 0.234$.

Note 2.6. Theorem 2.5 provides a condition that when the asymptotically smallest scaling term of the distinct component is not significantly smaller than the remaining scaling components, than the limiting process is the same as the one found in [RGG97]. However, it is possible that $\theta_1^{-2}(d)$ is not the smallest scaling term, it could be that $\theta_{n+1}^{-2}(d)$ is the smallest scaling term. This situation is verified in the theorem above with $c(\mathcal{J}(i,d))\,\theta_{n+1}^{2}(d)$ in the denominator.

Remark 2.7. The function E_R has an inverse effect on the optimal value of the scaling parameter $\hat{\ell}$. As was the case in the section above, for smoother values of f, the smaller is E_R and thus the larger is $\hat{\ell}$. The added component of the scaling vector $\Theta^2(d)$ effects the optimal value of $\hat{\ell}$ through E_R by the following: for larger values of K_{n+i}, and thus the variance of the component, the larger is the value of $\hat{\ell}$.

Proof. (of Theorem 2.5). As is shown in the appendix (give citation to "core" section), to prove the weak convergence result of the rescaled RWM to the Langevin diffusion, it suffices to show that for a test function $h \in C_c^\infty$, the following limit holds:

$$\lim_{d\to\infty} \mathbb{E}\left[\left|Gh\left(d,\boldsymbol{X}^{(d)}\right) - G_L h\left(X_{i^*}\right)\right|\right] = 0,$$

where

$$Gh\left(d,\boldsymbol{X}^{(d)}\right) = d^{\alpha}\mathbb{E}_{\boldsymbol{Y}^{(d)}}\left[\left(h\left(Y_{i^*}\right) - h\left(X_{i^*}\right)\right)\left(\min\left\{1, \frac{\pi\left(d,\boldsymbol{Y}^{(d)}\right)}{\pi\left(d,\boldsymbol{X}^{(d)}\right)}\right\}\right)\right]$$

denotes the discrete-time generator of the rescaled Metropolis algorithm, and

$$G_L h\left(X_{i^*}\right) = v(\ell)\left[\frac{1}{2}h''\left(X_{i^*}\right) + \frac{1}{2}h'\left(X_{i^*}\right)\left(\log f\left(X_{i^*}\right)\right)'\right]$$

is the generator of a Langevin diffusion process with speed measure $v(\ell)$ as in Theorem 2.5. Instead of proving the convergence of these generators, one can define a generator $\tilde{G}h\left(d,\boldsymbol{X}^{(d)}\right)$ that is equivalent to $Gh\left(d,\boldsymbol{X}^{(d)}\right)$ in the sense that

$$\lim_{d\to\infty} \mathbb{E}\left[\left|Gh\left(d,\boldsymbol{X}^{(d)}\right) - \tilde{G}h\left(d,\boldsymbol{X}^{(d)}\right)\right|\right] = 0.$$

The result thus follows if one is able to show the $\mathcal{L}^1$ convergence of this equivalent generator $\tilde{G}h\left(d,\boldsymbol{X}^{(d)}\right)$ to the generator of a Langevin diffusion.

The equivalent generator is defined as

$$\begin{aligned}\tilde{G}h\left(d,\boldsymbol{X}^{(d)}\right) \quad &= \quad \frac{1}{2}\ell^2 h''\left(X_{i^*}\right)\mathbb{E}_{\boldsymbol{Y}^{(d)-}} \min\left\{1, \exp\left(\sum_{j=1,j\neq i^*}^{d} \varepsilon\left(d,X_j,Y_j\right)\right)\right\} + \\ &\quad +\ell^2 h'\left(X_{i^*}\right)\left(\log f\left(X_{i^*}\right)\right)'\mathbb{E}_{\boldsymbol{Y}^{(d)-}}\left(\exp\left(\sum_{j=1,j\neq i^*}^{d} \varepsilon\left(d,X_j,Y_j\right)\Big|_{\sum_{j=1,j\neq i^*}^{d}\varepsilon(d,X_j,Y_j)<0}\right)\right),\end{aligned}$$

where

$$\varepsilon\left(d,X_j,Y_j\right) = \log\frac{f\left(\theta_j(d)\,Y_j\right)}{f\left(\theta_j(d)\,X_j\right)}.$$

First we prove the $\mathcal{L}^1$ convergence of the generator $\tilde{G}h\left(d,\boldsymbol{X}^{(d)}\right)$ to the generator of the Langevin diffusion. We compute the $\mathcal{L}^1$ distance

$$\begin{aligned}\mathbb{E}\left[\left|\tilde{G}h\left(d,\boldsymbol{X}^{(d)}\right)-G_Lh\left(X_{i^*}\right)\right|\right] \leq\ & \ell^2\mathbb{E}_{\boldsymbol{X}^{(d)-}}\left[\left|\frac{1}{2}\mathbb{E}_{\boldsymbol{Y}^{(d)-}}\left[\min\left(1,\exp\left(\sum_{\substack{j=1\\ j\neq i^*}}^{d}\varepsilon\left(d,X_j,Y_j\right)\right)\right)\right]\right.\right.\\ & \left.\left.-\Phi\left(-\frac{\ell\sqrt{E_R}}{2}\right)\right|\right]\mathbb{E}\left[\left|h''\left(X_{i^*}\right)\right|\right]\\ & +\ell^2\mathbb{E}_{\boldsymbol{X}^{(d)-}}\left[\left|\mathbb{E}_{\boldsymbol{Y}^{(d)-}}\left[\exp\left(\sum_{\substack{j=1\\ j\neq i^*}}^{d}\varepsilon\left(d,X_j,Y_j\right)\right):\sum_{\substack{j=1\\ j\neq i^*}}^{d}\varepsilon\left(d,X_j,Y_j\right)<0\right]\right.\right.\\ & \left.\left.-\Phi\left(-\frac{\ell\sqrt{E_R}}{2}\right)\right|\right]\mathbb{E}\left[\left|h'\left(X_{i^*}\right)\left(\log f\left(X_{i^*}\right)\right)'\right|\right].\end{aligned}$$

Since h has compact support, it is absolutely bounded and so are its derivatives. Therefore, the expectations $\mathbb{E}\left[\left|h''\left(X_{i^*}\right)\right|\right]$ and $\mathbb{E}\left[\left|h'\left(X_{i^*}\right)\left(\log f\left(X_{i^*}\right)\right)'\right|\right]$ are also bounded. Therefore, by Lemmas (add 8,9) the first expectation on the right-hand side of the expression above converges to 0 as $d\to\infty$. The same result holds for the remaining expectations by use of Lemmas (add 10,11). □

Theorem 2.5 yields that each component of the time-accelerated Metropolis algorithm follows the dynamics of a Langevin diffusion process. Although the different components of $\boldsymbol{X}^{(d)}(t)$ depend on one another in finite dimensions, they become independent in the asymptotic limit, $d\to\infty$.

2.3.4 Inhomogeneous proposal scaling and extensions

In order for one to obtain a stochastic process from the rescaled Metropolis algorithm $\left\{\boldsymbol{Z}^{(d)}(t)\right\}_{t\geq 0}$, it is necessary that each of its components are sped up by the same factor. Otherwise, some components would move more frequently than the others in the same time unit, and thus the acceptance probability for a proposed move would be ill-defined. However, it is still possible to consider a scheme that adjusts the proposal scaling of each component as a function of d separately from the rest. In particular, Bédard considers the scheme where the proposal scaling of the last $d-n$ components are personalized, whereas the first n components are scaled as previously. One then treats each of the m groups of scaling terms appearing infinitely often as a different portion of the scaling vector, and then determine the appropriate α (from 2.3) for each group separately.

Consider the following scaling form for the proposals for each X_j: let $\sigma_j^2=\ell^2/d^{\alpha_j}$, where $\alpha_j=\alpha$ for $j=1,\ldots,n$, and let α_j be the smallest number such that $\lim_{d\to\infty}c\left(\mathcal{J}(i,d)\right)d^{\gamma_i}/d^{\alpha_j}<\infty$ for $j=n+1,\ldots,d$, $j\in\mathcal{J}(i,d)$. For component of interest X_{i^*}, it is still necessary to assume that $\theta_{i^*}(d)=1$, but now let $\boldsymbol{Z}^{(d)}(t)=\boldsymbol{X}^{(d)}\left(\left[d^{\alpha_i^*}t\right]\right)$. A new theorem can be presented as an analogue to Theorem 2.5, but now with the new inhomogeneous proposal scaling adopted.

Theorem 2.8. (Theorem 3 in [Béd07]) *in the setting of Theorem 2.5 but with the inhomogeneous proposal scaling adopted, the conclusions of Theorem 2.5 are preserved, but with*

$$E_R=\lim_{d\to\infty}\sum_{i=1}^{m}\frac{c\left(\mathcal{J}(i,d)\right)}{d^{\alpha_n+i}}\frac{d^{\gamma_i}}{K_{n+i}}\mathbb{E}\left[\left(\frac{f'(X)}{f(X)}\right)^2\right].$$

Remark 2.9. The proposal considered in this theorem has been personalized for the terms which appear Since the proposal scaling has been personalized for each distinct group of components, each constant term $K_{n+1},\ldots,K_{n+m}$ has an impact on the limiting process, yielding a larger value for E_R. Therefore, the optimal value of $\hat{\ell}$ will be smaller in the current situation than before.

Note 2.10. While simulations showed ([RR01]) that the asymptotically optimal acceptance rate of 0.234 proved to be fairly accurate in small dimensions (i.e., $d\geq 10$) for targets of the form (2.1), this is not the case in the context of targets of the form (2.2). For instance, if there exists a finite number of scaling terms λ_j that are close to the exponent α in (2.3), then the optimal acceptance rate converges at a very slow rate to 0.234.

2.3.5 Generalizing the K_j parameters

The conclusions of 2.8 can be extended to allow the terms $\{K_j, j \in \mathcal{J}(i,d)\}$ for $i = 1, \ldots, m$ to be i.i.d. random variables with some distribution with $\mathbb{E}\left[K_j^{-2}\right] < \infty$. Denote the scaling vector now as

$$\Theta^{-2}(d) = \left(\frac{K_1}{d^{\lambda_1}}, \ldots, \frac{K_n}{d^{\lambda_n}}, \ldots, \frac{K_{n+c(\mathcal{J}(i,d))}}{d^{\gamma_1}}, \ldots, \frac{K_{n+\sum_1^{m-1} c(\mathcal{J}(i,d))+1}}{d^{\gamma_m}}, \ldots \frac{K_d}{d^{\gamma_m}} \right). \tag{2.6}$$

In order to provide meaningful results, it is helpful (and is costless in respect to the generality afforded by the subsequent theorems) to denote $\mathbb{E}\left[K_j^{-1}\right] = b_i$ for $j \in \mathcal{J}(i,d)$. Since the scaling term of interest is not dependent on dimension, it follows that $\theta_{i^*}^{-2}(d) = K_{i^*}$. To keep consistent with the ordering adopted above, we denote $-\infty < \gamma_m < \gamma_{m-1} < \cdots < \gamma_1 < \infty$. Furthermore, to prevent redundancy in the specification of scaling terms, all the terms λ_j are assumed to be distinct from γ_i, for $j = 1, \ldots, n$ and $i = 1, \ldots, m$. Hence, all the scaling terms appearing infinitely often in the limiting process are grouped together in one of the m groups.

Theorem 2.11. (Theorem 4 in [Béd07]) *In the setting of Theorem 2.5, with scaling vector (2.6), consider the parameter of interest with scaling parameter $\theta_{i^*} = K_{i^*}^{-1/2}$ (i.e., normalized so $d^{\lambda_i} = 1$). We then have the weak convergence of $\left\{Z_{i^*}^{(d)}(t)\right\}_{t\geq 0} \Rightarrow \{Z(t)\}_{t\geq 0}$, where $Z(0)$ is distributed according to the density $\theta_{i^*} f(\theta_{i^*} x)$ and $\{Z(t)\}_{t\geq 0}$ satisfies the Langevin SDE*

$$dZ(t) = (v(\ell))^{1/2} dB(t) + \frac{1}{2} v(\ell) \left(\log f(\theta_{i^*} Z(t))\right)' dt,$$

if and only if (2.5) is satisfied. The speed measure is as in Theorem 2.5 , with

$$E_R = \lim_{d\to\infty} \sum_{i=1}^{m} \frac{c(\mathcal{J}(i,d))\, d^{\gamma_i}}{d^{\alpha}} b_i \mathbb{E}\left[\left(\frac{f'(X)}{f(X)}\right)^2\right],$$

with $c(\mathcal{J}(i,d)) = \#\left\{j \in [n+1, d] : \theta_j(d) \in \mathcal{O}\left(d^{\gamma_i/2}\right)\right\}$.

There is a final extension that is given in [Béd07], allowing the terms $\left\{\theta_j^{-2}(d), j \in \mathcal{J}(i,d)\right\}$ to vary within each of the m subgroups, provided that they are at least of the same order. The interested reader can read of this extension in the paper; the particular generalization itself should not be too difficult to deduce following the discussion above.

2.3.6 Optimal scaling when there exists scaling terms significantly smaller than the rest

In the subsections above, the target distribution under consideration with such that its scaling terms were not significantly different from the rest. This condition was stipulated by equation (2.5). Suppose this condition is violated, i.e., we have the complement of equation (2.5) given by

$$\lim_{d\to\infty} \frac{\theta_1^2(d)}{\sum_{j=1}^{d} \theta_j^2(d)} > 0, \tag{2.7}$$

which states that there exists at least one scaling term that is significantly smaller than the rest. By the manner with which we formulated the scaling terms, it is clear that the asymptotically smallest scaling term would have to be $\theta_1^{-2}(d)$ or $\theta_{n+1}^{-2}(d)$. However, if condition (2.7) is satisfied, then K_1/d^{λ_1} must be the smallest for large values of d, as its inverse dominates the sum of the remaining components. This differs from the case considered above, since before either $\theta_1^{-2}(d)$ or $\theta_{n+1}^{-2}(d)$ could have been the asymptotically smallest term. If there is more than one target component possessing a scaling term of $\mathcal{O}\left(d^{-\lambda_1}\right)$, then let

$$b = \max_{j\in\{1,\ldots,n\}} (\lambda_j = \lambda_1)$$

denote the number of such components.

For the optimal scaling problem in the current context, $\theta_1^{-2}(d)$ will be of paramount importance, since the choice of proposal scaling must be such that the algorithm does not propose jumps too large for this component. It is shown by Bédard in [Béd08a] that the optimal scaling for the target still is given by $\sigma^2(d) = \ell^2/d^\alpha$, where α is the smallest number satisfying (2.3). If equation (2.7) holds as well, then its reciprocal has to satisfy

$$\lim_{d\to\infty}\sum_{j=1}^{d}\theta_1^{-2}(d)\,\theta_j^2(d)$$

$$= \lim_{d\to\infty}\frac{K_1}{d^{\lambda_1}}\left(\frac{d^{\lambda_1}}{K_1}+\cdots+\frac{d^{\lambda_n}}{K_n}+c\left(\mathcal{J}(1,d)\right)\frac{d^{\gamma_1}}{K_{n+1}}+\cdots+c\left(\mathcal{J}(m,d)\right)\frac{d^{\gamma_m}}{K_{n+m}}\right) \quad < \quad \infty.$$

Theorem 1 in [Béd08a] stipulates that in order for a diffusion limit to exist in the case where (2.7) holds, it is necessary that there exists at least one $i \in \{1,\dots,m\}$ such that

$$\lim_{d\to\infty}\frac{c\left(\mathcal{J}(i,d)\right)d^{\gamma_i}}{d^{\lambda_1}} > 0. \tag{2.8}$$

Furthermore, it is proven there that the optimal value of ℓ maximizes

$$\begin{aligned} v(\ell) &= 2\ell^2\mathbb{E}_{\boldsymbol{Y}^{(b)},\boldsymbol{X}^{(b)}}\left[\Phi\left(\frac{\sum_{j=1}^{b}\log\left(f\theta_jY_j\right)/f\left(\theta_jX_j\right)-\ell^2E_R/2}{\sqrt{\ell^2E_R}}\right)\right] \\ &= \ell^2a(\ell). \end{aligned} \tag{2.9}$$

Example 2.12. Consider the distribution (2.2), for $d = 10$, where $f_i \sim \mathcal{N}\left(1,\theta_i^2(d)\right)$, where $\theta_i^2(d) = \frac{1}{d}$ for the first five components, and $d = 25$ for the remaining components. In this case, $\alpha = 1$ in (2.3), so we consider proposal variances of the form $\sigma^2(d) = \ell^2/d$. In this situation, the first five components yield a finite limit in (2.3), and they thus play an active role in the determination of the AOAR. Here we can use Theorem 1 in [Béd08a], and find the value of $\hat{\ell}$ that maximizes (2.9). Doing so yields

$$\lim_{d\to\infty}\frac{d}{5d+(d-5)/2} \to 0.1984 > 0$$

as $d\to\infty$, and $\hat{\ell} = 1.95$ with AOAR$= 0.177$.

The code in Section C.1 allows us to obtain empirical estimates as well.

Example 2.13. In this example we will consider a target distribution of the form of (2.2), with

$$\pi\left(\boldsymbol{x}^{(d)}\right) \propto x_1^4\exp\left(-x_1\right)x_2^4\exp\left(-x_2\right)\prod_{i=3}^{d}x_i^4\exp\left(-x_i/5\sqrt{d}\right),$$

for $x > 0$ and $d = 100$. In the increasing ordering notation as defined in section 2.3, we take the scaling vector $\Theta^{-2}(d) = (1,1,25d,25d,25d,\dots)$. The value of α satisfying (2.3) is $\alpha = 0$, so we consider proposals of the form $\sigma^2(d) = \ell^2$. In this situation, we observe that the first two scaling terms are significantly smaller than the remaining $d-2$ scaling terms. In particular, it is clear that both conditions (2.7) and (2.8) hold, with

$$\lim_{d\to\infty}\left(2+\frac{d-2}{25d}\right)^{-1} = \frac{25}{51},$$

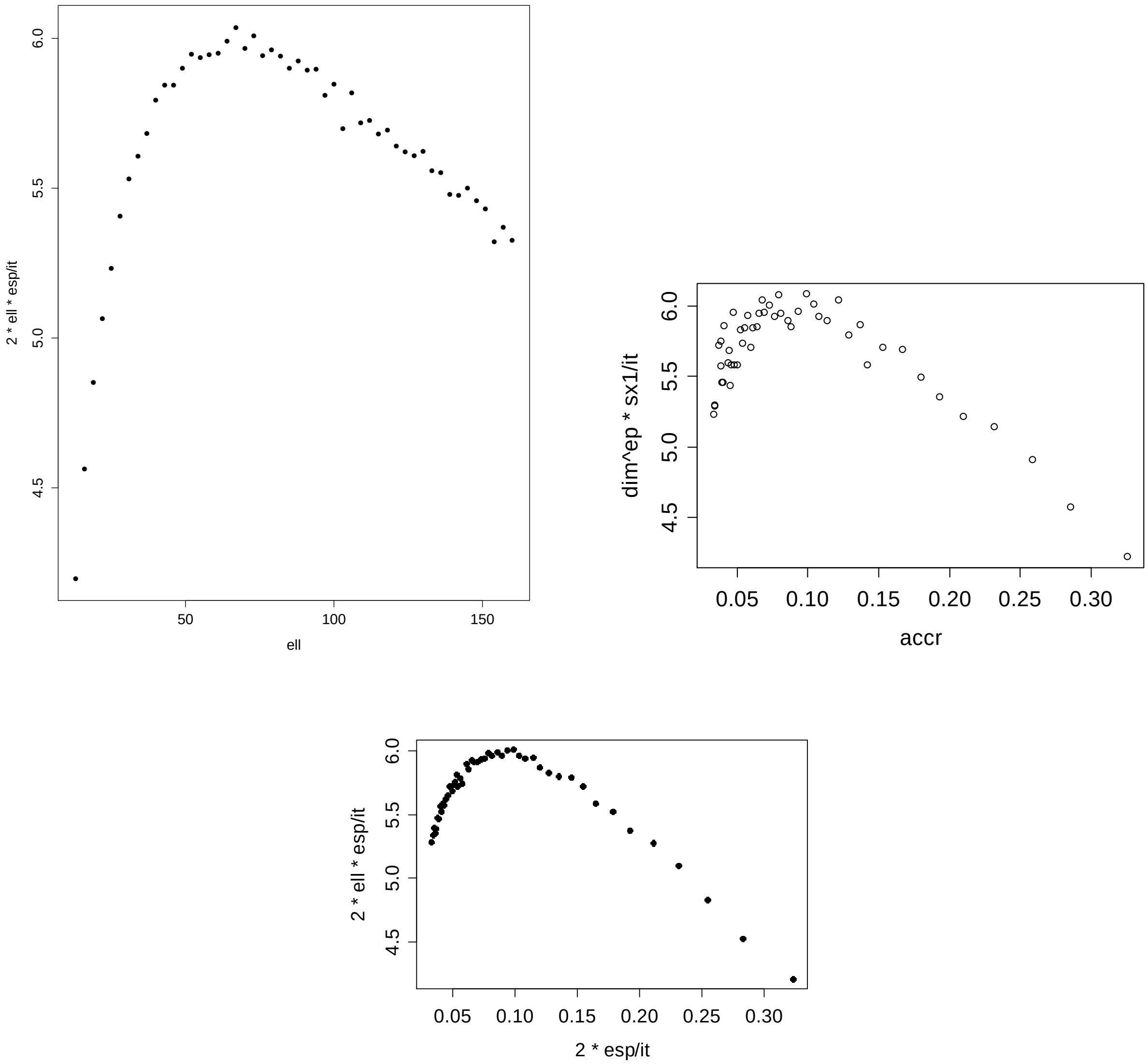
and

$$\lim_{d\to\infty}\frac{c(1,d)\,\theta_3^2(d)}{d^0} = \lim_{d\to\infty}\frac{d-2}{25d} = \frac{1}{25}.$$

The graph below implements the code in Section C.2, where Monte Carlo simulation has been preformed first on the easier to simulate pairs X_1, X_2 (since they are defined on a narrower interval than the rest), and the expectation in (2.9) is subsequently replaced with $E_R = \frac{1}{75}$. Integer values of ℓ^2 in the vector $[13,160]$ are all applied in the proposal variance to obtain a vector of values for the speed measure $v(\ell)$.

A scan of the speed function $h(\ell)$ yields a critical value for the scaling parameter as $\hat{\ell} \approx 61$, and an acceptance rate $a\left(\hat{\ell}\right) = 0.01213$. Using (2.9), we are also able to obtain theoretical values for the maximization of the diffusion's speed measure. This is in the third panel of the Figure 2.

Figure 2: Example 2.13



2.4 Metropolis-within-Gibbs algorithms

In this section we move our consideration from Metropolis algorithms that make updates for all components simultaneously to partially updating algorithms. This means that the updates are chosen to be lower dimensional than the target density itself. Optimal scaling in this context requires two choices, one for the scaling of the proposal density, and the second for the dimensionality of the proposal. The paper by Neal and Roberts, [NR05], proves optimality results for the Metropolis-within-Gibbs algorithm with random walk and Langevin updates separately. In the first instance, they find that full-dimensional random walk Metropolis updates are not better asymptotically than smaller dimensional updating schemes, and thus, with the added factor of computational cost for higher dimensional updates (computational overhead at every iteration is a non-decreasing function of the proportion of the density being updated), it is always optimal to use lower dimensional updates when possible. On the other hand, they find that full dimensional Langevin algorithms are worthwhile in general, although one needs to consider the computational overhead associated with their implementation when deciding on optimality.

For RWM/MALA-within-Gibbs algorithms, the implementation of the MCMC algorithm requires choosing $d \cdot c_d$ components, where c_d denotes the fixed value of updating proportion, chosen at random at each iteration, and then attempt to update them jointly according to the RWM/MALA mechanism, respectively. Hence, the two algorithms (RWM and MALA, respectively) propose new values for coordinate i according to

$$\begin{aligned} Y_i^{(d)} &= x_i^{(d)} + \chi_i^{(d)} \sigma_{d,c_d} Z_i, \\ Y_i^{(d)} &= x_i^{(d)} + \chi_i^{(d)} \left\{ \sigma_{d,c_d} Z_i + \frac{\sigma_{d,c_d}^2}{2} \frac{\partial}{\partial x_i} \log \pi_d \left(\boldsymbol{x}^{(d)} \right) \right\}, \end{aligned}$$

where $1 \leq i \leq d$. The random variables $\{Z_i\}_{i\in[1,d]}$ are i.i.d. with distributions $Z \sim \mathcal{N}(0,1)$, and each of the characteristic functions $\chi_i^{(d)}$ are chosen independently of the Z_i's. At each iteration, a new random subset A is chosen from $\{1, 2, \ldots, d\}$, and $\chi_i^d = 1$ if $i \in A$, and is zero otherwise. The Metropolis-Hastings acceptance rule for the two above setup when applied to the i.i.d. components target (1.8) yields the acceptance probability

$$\alpha_d^{c_d} \left(\boldsymbol{x}^{(d)}, \boldsymbol{Y}^{(d)} \right) = \min \left\{ 1, \frac{\pi \left(\boldsymbol{Y}^{(d)} \right) q \left(\boldsymbol{Y}^{(d)}, \boldsymbol{x}^{(d)} \right)}{\pi \left(\boldsymbol{x}^{(d)} \right) q \left(\boldsymbol{x}^{(d)}, \boldsymbol{Y}^{(d)} \right)} \right\},$$

where $q(\cdot,\cdot)$ is the chosen proposal density. In case of rejection, we choose $\boldsymbol{X}_m^{(d)} = \boldsymbol{X}_{m-1}^{(d)}$. As we had done for the cases considered previously, we can use the notion of average acceptance rate as means for measuring efficiency and optimality of the algorithm. We denote

$$\begin{aligned} a_d^{c_d}(\ell) &= \mathbb{E}_\pi \left[\alpha_d^{c_d} \left(\boldsymbol{X}^d, \boldsymbol{Y}^d \right) \right] \\ &= \mathbb{E}_\pi \left[\min \left\{ 1, \frac{\pi_d \left(\boldsymbol{Y}^d \right) q \left(\boldsymbol{Y}^d, \boldsymbol{x}^d \right)}{\pi_d \left(\boldsymbol{x}^d \right) q \left(\boldsymbol{x}^d, \boldsymbol{Y}^d \right)} \right\} \right], \end{aligned}$$

where $\sigma_{d,c_d}^2 = \ell^2 d^{-s}$. Furthermore, as always, it is assumed throughout that the algorithm starts at stationarity. The scaling exponent s is given by $s = 1$ for RWM and $s = \frac{1}{3}$ for MALA.

2.4.1 RWM-within-Gibbs

Consider first the RWM-within-Gibbs algorithm applied to the i.i.d. target density. As was done in [RR98], it is helpful to express the density in the exponential target form

$$\pi \left(\boldsymbol{x}^{(d)} \right) = \prod_{i=1}^{d} f \left(x_i^{(d)} \right) = \prod_{i=1}^{d} \exp \left\{ g \left(x_i^{(d)} \right) \right\},$$

The proposal standard deviation is denoted by $\sigma(d) = \frac{\ell}{\sqrt{d-1}}$, for some $\ell > 0$. Since the target is again formed of i.i.d. components, all individual components of the target density are equivalent. For convenience, it will be again useful to consider the first component of the resulting algorithm, with a time-increasing factor of d, which is denoted by $U_{t,1}^d$, obtained from

$$\boldsymbol{U}_t^{(d)} = \left(X_{[dt],1}, X_{[dt],2}, \ldots X_{[dt],d} \right).$$

Theorem 2.14. (Theorem 3.1 in [NR05]) *Suppose f is positive, is of class C^3, satisfies (1.10), and $(\log f)' = g'$ is a Lipschitz function. Suppose $c_d \to c$ as $d \to \infty$, for some $0 < c \leq 1$. Let $\boldsymbol{X}_0^{(\infty)} = \left(X_{0,1}^{(1)}, X_{0,2}^{(2)}, \dots\right)$ be such that all of its components are distributed according to f and assume further that $X_{0,i}^{(j)} = X_{0,i}^{(i)}$. Then $U_t \Rightarrow U$ as $d \to \infty$, where U_0 is distributed according to f and U satisfies the Langevin SDE*

$$dU_t = \left(h_c\left(\ell\right)\right)^{1/2} dB_t + \frac{1}{2} h_c\left(\ell\right) g'\left(U_t\right) dt$$

and

$$h_c\left(\ell\right) = 2c\ell^2 \Phi\left(-\frac{\ell\sqrt{cI}}{2}\right),$$

and

$$I = \mathbb{E}_f\left[\left(\frac{f'\left(X\right)}{f\left(X\right)}\right)^2\right] = \mathbb{E}_g\left[g'\left(X\right)^2\right].$$

Corollary 2.15. (Corollary 3.2 in [NR05]) *Suppose $c_d \to c$ as $d \to \infty$, for some $0 < c \leq 1$. Then,*

$$\lim_{d\to\infty} a_d^{c_d}\left(\ell\right) = a^c\left(\ell\right) = 2\Phi\left(-\frac{\ell\sqrt{cI}}{2}\right),$$

and the unique value of ℓ , denoted by $\hat{\ell}$, which maximizes the speed function for $c = 1$, is given by $\hat{\ell} = c^{1/2}\hat{\ell}_c$, where $\hat{\ell}_c$ is the unique value of ℓ which maximizes $h_c\left(\ell\right)$ for all other values of $c \in (0, 1]$. Furthermore, for all $c \in (0, 1]$, the optimal acceptance rate is given by $a^c\left(\hat{\ell}_c\right) = 0.234$.

Remark 2.16. As mentioned in the introductory paragraphs for this section, since the optimal acceptance rate is independent of the value of c, there is no added efficiency to be gained by using larger block updates in the RWM setup.

2.4.2 MALA-within-Gibbs

The above subsection showed that there is no theoretical justification for using high-dimensional updates for the RWM algorithm. The results of this section will show that the same conclusion does not hold for MALA. Along with the assumptions in 1.4.1, assume further that g' is Lipschitz. A Poisson process $\{J_t\}$ with rate $d^{1/3}$ is again needed for the theorem of this section, along with a d-dimensional jump-process $\Gamma_t^{(d)} = \boldsymbol{X}_{J_t}^{(d)}$, and suppose the proposal form is given by $\sigma^2\left(d\right) = \ell^2 d^{\frac{1}{3}}$.

Theorem 2.17. (Theorems 4.1 and 4.2 in [NR05]) *Suppose that $c_d \to c$ for $c \in (0, 1]$, as $d \to \infty$. Then*

$$\lim_{d\to\infty} \{a_d^{c_d}\left(\ell\right)\} = a^c\left(\ell\right) = 2\Phi\left(-\frac{\sqrt{c}K\ell^3}{2}\right),$$

where

$$K^2 = \mathbb{E}\left[\frac{5g'''\left(X^2\right) - 3g''\left(X\right)^3}{48}\right] > 0.$$

Let $\left\{U_t^{(d)}\right\}_{t\geq 0}$ be the process corresponding to the first component of $\Gamma^{(d)}$. Then, $U_t^{(d)} \Rightarrow U_t$, where U_t satisfies the Langevin SDE

$$dU_t = h_c\left(\ell\right)^{1/2} dB_t + \frac{1}{2} h_c\left(\ell\right) g'\left(U_t\right) dt,$$

and $h_c\left(\ell\right) = 2c\ell^2\Phi\left(-\frac{\sqrt{c}\ell^3 K}{2}\right)$ denotes the speed measure of the Langevin diffusion.

Corollary 2.18. (Corollary 4.3 in [NR05]) *The unique value $\hat{\ell}$ of ℓ which maximizes $h_1\left(\ell\right) = 2\ell^2\Phi\left(-\frac{\ell^3 K}{2}\right)$ on $\mathbb{R}_+$ is given by $h_1\left(\hat{\ell}\right) = h_c\left(\hat{\ell}_c\right) \cdot c^{-2/3}$, where $\hat{\ell}_c = c^{-1/6}\hat{\ell}$, is the unique value of ℓ which maximizes $h_c\left(\ell\right)$ on $\mathbb{R}_+$. Furthermore, for all $c \in (0, 1]$, the optimal acceptance rate $a^c\left(\hat{\ell}_c\right) = 0.574$.*

2.5 Optimal Scaling of RWM with Discontinuous Target Densities

2.5.1 Inefficiencies of full-dimensional updates using RWM applied to discontinuous targets

All the results proven thus far have been for continuous target densities. In [NRY05], the authors provide a weak convergence result for a high-dimensional RWM algorithm applied to a target distribution which has a discontinuous probability density function. In particular, they show that when the proposal variance is scaled by d^{-2}, the sequence of stochastic processes formed by the first component of each Markov chain converges to a Langevin diffusion process whose speed measure is optimized when the algorithm is tuned to yield an acceptance rate of $e^{-2} \approx 0.1353$. Hence, the mixing time for the RWM algorithm when applied to discontinuous densities is $\mathcal{O}\left(d^2\right)$. This work serves as a validation of the result of [NR05], where it was found that lower dimensional updates are more efficient in the RWM setting due to the lower computational cost. In fact, [NR05] showed that the Metropolis-within-Gibbs algorithms mix with rate $\mathcal{O}\left(d\right)$, which compares favorably to the mixing rate of $\mathcal{O}\left(d^2\right)$ required in the current context for RWM algorithms applied to discontinuous target densities.

The target density in consideration is given by

$$\pi\left(\boldsymbol{x}^{(d)}, d\right) = \prod_{i=1}^{d} f\left(x_i\right),$$

where

$$f\left(x\right) \propto \exp\left(g\left(x\right)\right),$$

with $0 < x < 1$, $g\left(\cdot\right)$ twice differentiable on $[0,1]$, and $f\left(x\right) = 0$ otherwise.

2.5.2 Algorithm and Main Results

For $d \geq 1$, consider a RWM algorithm on the d-dimensional hypercube with target density $\pi\left(\boldsymbol{x}^{(d)}, d\right)$ as given above. As always, suppose the algorithm starts in stationarity, so $\boldsymbol{X}_0^{(d)} \sim \pi\left(\cdot\right)$, and for $t \geq 0$ and $i \geq 1$, denote by $Z_{t,i}$ elements of a sequence of i.i.d. uniform random variables with $Z \sim \mathcal{U}\left[-1, 1\right]$, with $\boldsymbol{Z}_t^{(d)} = \left(Z_{t,1}, Z_{t,2}, \ldots, Z_{t,d}\right)$. For $d \geq 1$, $t \geq 0$ and $\ell > 0$, consider the proposals given by

$$\boldsymbol{Y}_{t+1}^{(d)} = \boldsymbol{X}_t^{(d)} + \sigma\left(d\right) \boldsymbol{Z}_t^{(d)},$$

where $\sigma\left(d\right) = \ell/d$. Accept these proposals with probability $\min\left(1, \frac{\pi\left(\boldsymbol{Y}_{t+1}^{(d)}\right)}{\pi\left(\boldsymbol{X}_t^{(d)}\right)}\right)$. Thus, for $d \geq 1$, let

$$h\left(\boldsymbol{z}^{(d)}, d\right) = \begin{cases} 2^{-d}, & \boldsymbol{z}^{(d)} \in \left(-1, 1\right)^d \\ 0, & \text{otherwise.} \end{cases}$$

Given the current state of the process $\boldsymbol{x}^{(d)}$, denote by $J\left(\boldsymbol{x}^{(d)}, d\right)$ the probability of accepting a proposal,

$$J\left(\boldsymbol{x}^{(d)}, d\right) = \int h\left(\boldsymbol{z}^{(d)}, d\right) \left\{\min\left(1, \frac{\pi\left(\boldsymbol{x}^{(d)} + \sigma\left(\boldsymbol{z}^{(d)}, d\right)\right)}{\pi\left(\boldsymbol{x}^{(d)}\right)}\right)\right\} d\boldsymbol{z}^{(d)}.$$

It is now necessary to define a pseudo-RWM process, which is identical to the regular RWM process except that it moves at every iteration. For $d \geq 1$, let $\hat{\boldsymbol{X}}_0^{(d)}, \hat{\boldsymbol{X}}_1^{(d)}, \ldots$ denote the successive states of the pseudo-RWM process, with $\hat{\boldsymbol{X}}_0^{(d)} \sim \pi\left(\cdot\right)$. The pseudo-RWM process is a Markov process, where for $t \geq 0$ $\hat{\boldsymbol{X}}_{t+1}^{(d)} = \hat{\boldsymbol{X}}_t^{(d)} + \sigma\left(d\right) \hat{\boldsymbol{Z}}_t^{(d)}$, and given that $\hat{\boldsymbol{X}}_t^{(d)} = \boldsymbol{x}^{(d)}$, the pdf of $\hat{\boldsymbol{Z}}_t^{(d)}$ is given by

$$\zeta\left(\boldsymbol{z}^{(d)} | \boldsymbol{x}^{(d)}\right) = h\left(\boldsymbol{z}^{(d)}, d\right) \min\left\{1, \frac{\pi\left(\boldsymbol{x}^{(d)} + \sigma\left(d\right) \boldsymbol{z}^{(d)}\right)}{\pi\left(\boldsymbol{x}^{(d)}\right)}\right\} \Big/ J\left(\boldsymbol{x}^{(d)}, d\right),$$

with $\boldsymbol{z}^{(d)} \in \mathbb{R}^d$.

3 Nonproduct Target Laws and Infinite Dimensional Target Measures

3.1 Infinite-Dimensional Framework

3.1.1 Problem specific context and Radon-Nikodym derivative for target measure

It has been suggested by a group of authors to consider targets defined on an infinite dimensional Hilbert space $\mathcal{H}$. The target measure π, which lives in the function space defined by $\mathcal{H}$, has density with respect to a Gaussian reference measure, meaning that it is absolutely continuous with respect to this Gaussian reference measure (the target distribution is thus defined as a change of measure from a product law). In this context, the target distribution π can be expressed using its Radon-Nikodym derivative, which has the form:

$$\frac{d\pi}{d\pi_0} = M_\Psi \exp\left(-\Psi\left(x\right)\right), \tag{3.1}$$

where π_0 is the Gaussian reference measure, $\Psi\left(\cdot; y\right)$ is some potential function and M_Ψ is the inverse of a normalization constant. Often, in specific examples, it is further assumed that the Gaussian reference measure π_0 has known mean and covariance operator, i.e., the eigenpairs (eigenvalues and eigenfunctions) of the covariance operator are known. Such a framework has been used to tackle problems in various fields, not limited to: molecular dynamics, inverse problems in Bayesian a framework where a product prior is combined with the likelihood (for instance, in the setting of Navier-Stokes equations, continuous-time signal processing problems, etc.), and conditioned diffusions ([BS09] provides a good survey for a few of these applications, [HSV10] highlights an array of signal processing problems which are tackled in the Bayesian framework, where the posterior measure arising in these problems all have the common structure discussed in this section, and [BRSV08] looks solely at the problem of sampling for diffusion bridges). In the context of optimal scaling, one reason for considering infinite dimensional space for the target measure is that if the target distribution is not of product form, then the interactions of the various coordinates with each other and thus could yield a limiting diffusion which take values in an infinite dimensional space.

3.1.2 Generality of non-product target measure

At first glance the framework developed in this section appears to be wholly general, as in finite dimensions it is a trivial problem to find a reference measure for which the target measure is absolutely continuous with respect to (i.e., the Lebesgue measure). However, in the infinite-dimensional context, this is not so clear, as no clear analog of the Lebesgue measure exists in the infinite dimensional context (see the discussion in A.5). Furthermore, provided the necessary reference measure exists, the assumption that the eigenpairs of the covariance is known is a fairly stringent assumption. For some applications, such as for conditioned diffusions, this may be a reasonable assumption; whereas for other cases, it may not be, and there it would be more natural to consider the isotropic proposal covariance instead. (See [MPS11], page 5) We will not consider isotropic proposals in this paper.

Since the framework used here is drastically different than the one adopted in the previous sections, it is no surprise that the proofs will use rather different methods. In particular, the papers considered here do not use the theory of generators of the associated Markov processes as was done in the papers considered above and that is given attention in [EK86], but instead rely on infinite dimensional analysis and functional analysis. For the benefit of the reader not familiar with these aspects, the necessary background has been summarized in the Appendix.

As mentioned previously, this non-product target form arises naturally in many applications. We discuss the specific application of sampling nonlinear diffusion bridges as motivation for this type of target measure in the following section.

3.2 Likelihood Estimation for Discretely Observed Diffusions

Diffusion processes are often used for modeling continuous-time phenomena, and are ubiquitous in finance and economics. A generic diffusion process V is defined as the solution of the stochastic differential equation (SDE) of the type:

$$dV_s = b\left(V_s; \theta\right) ds + \sigma\left(V_s; \theta\right) dB_s, \tag{3.2}$$

where B is a scalar Brownian motion. The functionals $b\left(\cdot; \theta\right)$ and $\sigma\left(\cdot; \theta\right)$ are called the drift and diffusion coefficient respectively, and are allowed to depend on some parameters $\theta \in \Theta$. Provided that they satisfy certain regularity conditions (see chapter 4 of [KP99]), it is guaranteed that there exists a weakly unique, global solution of the SDE above.

If one considers a small time increment dt, $V_{t+dt} - V_t$ is approximately Gaussian with mean and variance given by the Euler-Maruyama approximation

$$V_{t+dt} \approx V_t + b\left(V_t;\theta\right)dt + \sigma\left(V_t;\theta\right)\sqrt{dt}\cdot Y, \qquad Y \sim \mathcal{N}\left(0,1\right).$$

Higher order approximations are also available. The exact dynamics of the diffusion process are govern by its transition density:

$$p_t\left(v,w;\theta\right) = \mathbb{P}\left[V_t \in dw | V_0 = v;\theta\right]/dw, \quad t>0, w,v\in\mathbb{R}. \tag{3.3}$$

Suppose that there exists a finite set of observations for the diffusion process above, i.e., one has the collection of time instances $0 < t_0 < t_1 < \cdots < t_n$ with corresponding observations given at $\boldsymbol{v} = \{V_{t_0}, V_{t_1}, \cdots, V_{t_n}\}$. Denote the time increment between consecutive observations by $\Delta t_i = t_i - t_{i-1}$ for $1 \le i \le n$.

Then the log-likelihood of the data set $\boldsymbol{v}$ is given by

$$\ell\left(\theta|\boldsymbol{v}\right) = \sum_{i=1}^{n} \ell_i\left(\theta\right), \qquad \ell_i\left(\theta\right) := \log p_{\Delta t_i}\left(V_{t_{i-1}}, V_{t_i};\theta\right).$$

Unfortunately, other than a few special cases, (e.g., when the diffusion follows a geometric Brownian motion), this transition density is rarely available in an analytically tractable form. Hence, deriving maximum likelihood estimates for generic discretely observed diffusions is a challenging problem. One approach started by Aït-Sahalia is to expand the transition density using Hermite polynomials. It is shown that this closed-form approximation is an unbiased estimator of the true (unknown) transition density. A second approach is to use simulation to obtain the MLE estimates. Exact simulation of diffusion sample paths is a recent approach that is feasible for a wide class of diffusion processes, although not all diffusions can be handled by this method. The Exact Algorithm was introduced by Beskos [BPR06], and it uses technique called *retrospective sampling.*

First it is necessary to apply a variance-stabilization transformation to the diffusion process (3.2). Consider the Lamperti transform, $V_s \mapsto \eta\left(V_s;\theta\right) = X_s$,

$$\eta\left(u;\theta\right) = \int^u \frac{1}{\sigma\left(z;\theta\right)}dz.$$

If $\sigma\left(\cdot;\theta\right)$ is continuously differentiable, then we can apply Itô's rule to obtained the transformed diffusion's SDE:

$$dX_s = \alpha\left(X_s;\theta\right)ds + dB_s,$$

where $X_0 = x = \eta\left(V_0;\theta\right)$ for $s \in [0,t]$, and

$$\alpha\left(u;\theta\right) = \frac{b\left\{\eta^{-1}\left(u;\theta\right);\theta\right\}}{\sigma\left\{\eta^{-1}\left(u;\theta\right)\theta;\theta\right\}} - \sigma'\left\{\eta^{-1}\left(u;\theta\right);\theta\right\}/2, \quad u\in\mathbb{R};$$

η^{-1} denotes the inverse of the Lamperti transformation, and σ' denotes the derivative w.r.t. the space variable. Assume that $\alpha\left(\cdot;\theta\right)$ is continuously differentiable, $\left(\alpha^2+\alpha'\right)\left(\cdot;\theta\right)$ is bounded below, and the Girsanov formula for X, given by

$$\exp\left\{\int_0^t \alpha\left(\omega_s;\theta\right)d\omega_s - \frac{1}{2}\int_0^t \alpha^2\left(\omega_s;\theta\right)ds\right\},$$

is a martingale w.r.t. Wiener measure. Define

$$A\left(u;\theta\right) = \int^u \alpha\left(z;\theta\right)dz$$

to be any anti-derivative of α. Then the transition density of X is defined as

$$\tilde{p}_t\left(x,y;\theta\right) = \mathbb{P}\left[X_t \in dy | X_0 = x;\theta\right]/dy, \quad t>0, x,y\in\mathbb{R}.$$

Let $\mathbb{Q}_\theta^{(t,x,y)}$ denote the distribution of the process X conditioned to start at $X_0 = x$ and finish at $X_t = y$ for some fixed x, y, and $\mathbb{W}^{(t,x,y)}$ be the probability measure for the corresponding Brownian bridge. The objective of the EA is to obtain a rejection sampling algorithm to obtain draws from $\mathbb{Q}_\theta^{(t,x,y)}$. The following Proposition provides the necessary methodology for this. Let $\mathcal{N}_u\left(0,t\right)$ denote the density of a normal random variable with mean zero and variance t evaluated at $u\in\mathbb{R}$.

Proposition 3.1. *Under the conditions above, $\mathbb{Q}_\theta^{(t,x,y)}$ is absolutely continuous with respect to $\mathbb{W}^{(t,x,y)}$ with density:*

$$\frac{d\mathbb{Q}_\theta^{(t,x,y)}}{d\mathbb{W}^{(t,x,y)}} = \frac{\mathcal{N}_{y-x}\left(0,t\right)}{\tilde{p}_t\left(x,y;\theta\right)}\exp\left\{A\left(y;\theta\right) - A\left(x;\theta\right) - \int_0^t \frac{1}{2}\left(\alpha^2+\alpha'\right)\left(\omega_s;\theta\right)ds\right\}.$$

3.3 Optimal scaling for the RWM algorithm when applied to non-product target measures

3.3.1 MCMC methods for function space

The topic of optimal scaling when implementing infinite-dimensional analogs of the RWM and MALA algorithm for targets of the form (3.1) was started in [BRS09], and was made rigorous in [MPS11] for the RWM algorithm and in [PST11] for MALA. This section will look at the method and the main results found for the RWM algorithm, while the next section will look at the similar proof for MALA. While the target measures are defined on an infinite-dimensional Hilbert space, the MCMC methods necessary for sampling these probability measures will require a discretization of the function space leading to a high-dimensional measure $\mathbb{R}^d$, $d \gg 1$. This will be done using a spectral truncation method called the Karhunen-Loève expansion.

3.3.2 Karhunen-Loève expansion and invariant SPDEs for the target measure

As is shown in appendix A.2, Gaussian measures are the natural generalization of measures from $\mathbb{R}^N$ to infinite dimensional Hilbert spaces. In the current context, the Hilbert space in question, $(\mathcal{H}, \langle\cdot,\cdot\rangle)$, is assumed to be real and separable, with full-measure under π_0. Furthermore, it is assumed that the mean and covariance operator of the reference Gaussian measure π_0 are known, and will be denoted by m and $\mathcal{C}$ respectively. In order for the Gaussian law by π_0 to be well defined it is necessary for its covariance operator to be of trace class, which simply means that its eigenvalues are summable. With these conditions in place, it is possible to construct of an orthonormal basis of eigenfunctions $\{\varphi_i\}_{i=1}^{\infty}$ and corresponding eigenvalues $\left\{\lambda_i^2\right\}_{i=1}^{\infty}$ satisfying $\mathcal{C}\varphi_i = \lambda_i^2 e_i$ with $\sum_i \lambda_i^2 < \infty$. Since the eigenfunctions form a orthonormal system in $\mathcal{H}$, any function $x \in \mathcal{H}$ can be represented using the expansion

$$x = \sum_{i=1}^{\infty} x_i \varphi, \qquad x_i = m_i + \lambda_i \xi_i, \tag{3.4}$$

where the sequence $\{\xi_i\}_{i=1}^{\infty}$ is i.i.d. with $\xi_1 \sim \mathcal{N}(0,1)$ and $m_i = \langle m, \varphi_i\rangle$. This expansion is precisely the Karhunen-Loève expansion. Since $\mathcal{H}$ has full measure under π_0, the above statement is equivalent to saying that any realization $x \overset{\mathcal{D}}{\sim} \pi_0$ (meaning x is distributed according to π_0) can be expressed by allowing all the x_i to be independent Gaussian random variables, $x_j \sim \mathcal{N}\left(0, \lambda_j^2\right)$. By the form of Radon-Nikodym derivative (3.1), the measure π is absolutely continuous with respect to π_0, so any almost sure properties under π_0 also hold for π. In particular, by the law of large numbers, we have that almost surely with respect to π_0 (and thus π),

$$\frac{1}{N}\sum_{i=1}^{N} \frac{x_i^2}{\lambda_i^2} \overset{N\to\infty}{\longrightarrow} 1.$$

Hence, a typical draw from the target measure π must behave similarly to a typical draw from π_0 in large j coordinates. This structure of absolute continuity between measures allows us to exploit the product structure underlying Gaussian measure, when represented in the Karhunen-Loève coordinate system. Hence, in the coordinates $\{x_i\}_{i\geq 1}$, the prior has product structure of the form

$$\pi_0^N(x) = \prod_{i=1}^{n} \lambda_i^{-1} f\left(\lambda_i^{-1} x_i\right). \tag{3.5}$$

Since it is assumed the eigenpairs are known, sampling from the reference measure is thus straightforward. Furthermore, using the tensor product $\otimes$, it is also verified that

$$\mathbb{E}\left[(X-m)\otimes(X-m)\right] = \sum_{i=1}^{\infty} \lambda_i^2 \left(\varphi_i \otimes \varphi_i\right) = \mathcal{C}.$$

Using (3.4), it is clear that the isomorphism $X \mapsto \{x_i\}$ allows one to view $\mathcal{N}(m, \mathcal{C})$ as a product measure on the ℓ_2, the space of square summable sequences.

Since the target measure resides in an infinite dimensional Hilbert space, the limiting diffusion (if it exists) of the Metropolis induced Markov chain will necessarily be an infinite dimensional Hilbert space valued SDE/SPDE. However, without a link between the target (3.1) and the targets discussed earlier, it is hard to predict what this limiting process will be. It is precisely due to the presence of the potential function $\Psi(\cdot; y)$ that we cannot use the results from the previous sections, as $\Psi(\cdot;y)$ provides interactions between the components, and so individual components of the Markov chain cannot be expected, a priori, to convergence to a scalar SDE. Fortunately, it was proven in a series of papers [HSVW05, HSV07] that the target measure (3.1) is in fact invariant for Hilbert-space valued SPDEs with the form

$$\frac{dz}{dt} = -h(\ell)\left(z + \mathcal{C}\nabla\Psi(z)\right) + \sqrt{2h(\ell)}\frac{dW}{dt}, \quad z(0) = z^0, \tag{3.6}$$

where W is a Brownian motion on $\mathcal{H}$ with covariance operator $\mathcal{C}$. Hence, (3.6) will now play the role that was occupied by the Langevin diffusion process in the previous sections, and optimization of the RWM algorithm in the current context simply amounts to maximizing the speed measure of (3.6).

3.3.3 Sampling algorithm in finite dimensions

For sampling in the finite-dimensional context, the authors define $\Psi^N(\cdot) = \Psi\left(P^N\right)$, where P^N denotes projection in $\mathcal{H}$onto the first N eigenfunctions of $\mathcal{C}$ (more precisely, this is a projection on a dense subset of $\mathcal{H}$ where Ψ is defined) . One can then consider the finite-dimensional analog of (3.1), with the form

$$\frac{d\pi^N}{d\pi_0}(x) \propto \exp\left(-\Psi^N(x)\right).$$

Finite dimensional versions of other operators can also be obtained by taking suitable transformations with respect to P^N. For instance, $\mathcal{C}^N = P^N\mathcal{C}P^N$. This measure above coincides with π_0 on $\mathcal{H}\backslash P^N\mathcal{H}$ and has a density with respect to Lebesgue measure on $P^N\mathcal{H}$, in the coordinates $\{x_i\}_{i=1}^N$, as shown below.

It is now possible to implement a random walk method on $\mathbb{R}^N$ in the coordinate system $\{x_i\}_{i=1}^N$, enabling us to sample from π^N in $P^N\mathcal{H}$. Consider proposal distributions for the RWM which exploit the covariance structure of π_0 and can be expressed in $\mathcal{H}$ as:

$$y = x + \sqrt{\frac{2\ell^2}{N}}\mathcal{C}^{1/2}\xi, \qquad \text{where} \qquad \xi = \sum_{j=1}^N \xi_j\varphi_j \qquad \text{with} \qquad \xi_j \sim \mathcal{N}(0,1).$$

Observe that the proposal variance scales as $N^{-\gamma}$ with $\gamma = 1$.

Suppose further that the tails of the potential are sufficiently light so that for some $\mathcal{H}^r \subset \mathcal{H}$ we have $\pi_0(\mathcal{H}^r) = 1$. The subspace is defined by the inner-products described below. Now, if one denotes by X^N the span of the eigenfunctions of $\mathcal{C}$, after they have been transformed by an operator B_r with diagonal entries i^{2r}, then on X^N, π^N has Lebesgue density

$$\begin{aligned}\pi^N(x) &= M_{\Psi^N}\exp\left(-\Psi^N(x) - \frac{1}{2}\left\langle P^N x, \mathcal{C}^{-1}\left(P^N x\right)\right\rangle\right), \\ &= M_{\Psi^N}\exp\left(-\Psi^N(x) - \frac{1}{2}\left\langle x, \left(\mathcal{C}^N\right)^{-1}x\right\rangle\right),\end{aligned}$$

for $x \in X^N$. The acceptance probability corresponding to this RWM is given by

$$\alpha(x,\xi) = \min(1, Q(x,\xi)),$$

where

$$Q(x,\xi) = \frac{1}{2}\left\|\mathcal{C}^{-\frac{1}{2}}\left(P^N x\right)\right\|^2 - \frac{1}{2}\left\|\mathcal{C}^{-\frac{1}{2}}\left(P^N y\right)\right\|^2 + \Psi^N(x) - \Psi^N(y). \tag{3.7}$$

The induced Markov chain of this algorithm is given by $\left\{x^k\right\}$, for $k \geq 0$ with

$$\begin{aligned} x^{k+1} &= \gamma^{k+1}y^{k+1} + \left(1-\gamma^{k+1}\right)x^k, \qquad (3.8)\\ y^{k+1} &= x^k + \sqrt{\frac{2\ell^2}{N}}\mathcal{C}^{\frac{1}{2}}\xi^{k+1}, \\ \gamma^{k+1} &= \gamma\left(x^k, \xi^{k+1}\right) \overset{\mathcal{D}}{\sim} \text{Bernoulli}\left(\alpha\left(x^k,\xi^{k+1}\right)\right) \\ \xi^{k+1} &= \sum_{i=1}^N \xi_i^{k+1}\varphi_i. \end{aligned}$$

The main result of this paper will look at the continuous interpolant of the Markov chain given by

$$z^N(t) = (Nt - k)\,x^{k+1} + (k + 1 - Nt)\,x^k, \tag{3.9}$$

for $k \leq Nt < k + 1$.

3.3.4 Assumptions and Main Theorem

We won't provide the proof here, but the paper shows that with the correct choice of scaling, the one step transition for the RWM Markov chain behaves nearly like an Euler scheme applied to (3.6). Noise enters (3.6) additively, so the induced Itô map which takes Wiener trajectories into solutions is continuous in the supremum-in-time topology. The authors then proceed to show that the martingale increments converge weakly to a Hilbert-space valued Wiener process using a martingale central limit theorem. Since weak convergence is preserved under a continuous map, the fact that the Itô map is continuous implies the RWM Markov chain converges to (3.6). Such a limit is proven to exist for the RWM algorithm with proposal variance $\frac{2\ell^2}{N}C$, and the time constant $h(\ell)$ is maximized with average acceptance probability of 0.234, as was shown by [RGG97] for the product case. Furthermore, from the form of the interpolant (3.9), it is clear that the $\mathcal{O}(N)$ steps are required for the Markov chain to explore the invariant measure (3.6).

In order to prove the theorem, it is necessary to impose certain conditions on the functional Ψ acting as the potential in (3.1). For this, consider Sobolev spaces (inner product spaces with weak derivatives) $\mathcal{H}^r$, $r \in \mathbb{R}$, with inner-products and norms given by

$$\langle x, y\rangle_r = \sum_{i=1}^{\infty} i^{2r} x_i y_i, \quad \text{and} \quad \|x\|_r^2 = \sum_{i=1}^{\infty} i^{2r} x_i^2.$$

Assumption 3.2. (Assumption 3.1 in [MPS11]) *The operator $\mathcal{C}$ and potential Ψ satisfy the following conditions:*

1. ***Decay of Eigenvalues λ_i^2 of $\mathcal{C}$:*** *There exists $M_1, M_2 \in (0, \infty)$ and $\kappa > \frac{1}{2}$ such that*

$$M_1 \leq i^{\kappa}\lambda_i \leq M_2,$$

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

2. ***Assumptions on Ψ:*** *There exists constants $M_i \in \mathbb{R}$, $3 \leq i \leq 6$ and $s \in \left[0, \kappa - \frac{1}{2}\right)$ such that*

$$\begin{aligned} M_3 &\leq \Psi(x) \leq M_4\left(1 + \|x\|_s^2\right), \\ \|\nabla\Psi(x)\|_{-s} &\leq M_5\left(1 + \|x\|_s\right) \\ \left\|\partial^2\Psi(x)\right\|_{\mathcal{L}(\mathcal{H}^s, \mathcal{H}^{-s})} &\leq M_6, \end{aligned}$$

for all $x \in \mathcal{H}^s$.

Suppose further that the functions $\Psi^N : X^N \mapsto \mathbb{R}$ satisfy the same conditions as above, with the same constants uniformly in N.

Remark 3.3. The condition $\kappa > \frac{1}{2}$ ensures that the covariance operator $\mathcal{C}$ is trace-class as required.

Theorem 3.4. (Theorem 3.6 in [MPS11]) *Let Assumption 3.2 hold. Suppose the initial condition x^0 of the RWM algorithm is such that $x^0 \overset{\mathcal{D}}{\sim} \pi^N$ and let $z^N(t)$ be a piecewise linear, continuous interpolant (3.9) of the RWM algorithm (3.8). Then $z^N(t)$ converges weakly in $C([0,T], \mathcal{H}^s)$ to the diffusion process $z(t)$ given by (3.6) with $z(0) \overset{\mathcal{D}}{\sim} \pi$.*

3.4 Optimal scaling for MALA when applied to non-product target measures

Following the results of 3.3, [PST11] is a follow up to [MPS11]and studies the efficiency of the Metropolis-adjusted Langevin algorithm (MALA) when applied to the same framework of non-product target laws. The authors show that when started at stationarity, a suitably chosen interpolated and scaled version of the Markov chain corresponding to the MALA algorithm applied to an N-dimensional approximation of the target will take $\mathcal{O}\left(N^{\frac{1}{3}}\right)$ steps to explore the invariant measure, which is again provided by the infinite dimensional Hilbert space valued diffusion

(3.6), rather than a scalar diffusion . Furthermore, as a by-product of the diffusion limit, it follows that the MALA algorithm is optimized at an average acceptance probability of 0.574.

The Langevin proposal for the RWM obtained from the discretization of the Langevin diffusion is

$$y = x^{k,N} + \delta \nabla \log \left(\pi^N \left(x^{k,N}\right)\right) + \sqrt{2\delta} Z^N, \quad Z^n \sim \mathcal{N}\left(0, I_N\right),$$

where 2δ is the proposal variance, which in the discretized case is a small parameter representing the discrete time increment. In order to minimize the computational complexity of the algorithm, optimal scaling in the RWM environment consists of determining the "optimal" size of the increment δ as a function of dimension N. In this regard, the authors define a continuous interpolant of the Markov chain

$$z^N(t) = \frac{1}{\Delta t}\left(t - t_k\right) x^{k+1,N} + \frac{1}{\Delta t}\left(t_{k+1} - t\right) x^{k,N} \quad \text{for} \quad t \in [t_k, t_{k+1}), \tag{3.10}$$

where $t_k \overset{\text{def}}{=} k\Delta t$. The proposal variance is chosen in the form $2\ell\Delta t$, where $\Delta t = N^{-\gamma}$, so the scaling is in terms of the dimension of the parameter ℓ is a tuning parameter. If the interpolated chain z^N converges weakly to a suitable stationary diffusion process, the number of Markov chain steps required in stationarity is inversely proportional to the proposal variance, i.e., N^γ.

3.4.1 Sampling Algorithm

The difference in the two papers discussed in this section and the immediately preceding section is only in the form of the algorithm being used. The discussion in subsection 3.3.2 applies in the current context as well, so we do not repeat it here. Since the MALA algorithm incorporates the gradient of the previous iteration to make "smarter" moves, the authors of the current paper have to incorporate this added complexity to the sampling algorithm. For this purpose define $\mu(x) := -\left(x + C\nabla\Psi(x)\right)$ and its N-dimensional counterpart $\mu^N(x) = -\left(P^N x + C^N \nabla \Psi^N(x)\right)$.

From state $x \in X^N$, consider proposal $y \in X^N$ given by

$$y - x = \delta\mu^N(x) = \sqrt{2\delta}\left(C^N\right)^{\frac{1}{2}} \xi^N, \qquad \delta = \ell N^{-\frac{1}{3}}, \tag{3.11}$$

where $\xi^N = \sum_1^N \xi_i \varphi_i$, and $\xi_i \sim \mathcal{N}(0,1)$. As above, $\Delta t := \ell^{-1}\delta = N^{-\frac{1}{3}}$. Now study the Markov chain $\left\{x^{k,N}\right\}_{k\geq 0}$ obtained by "Metropolizing" this proposal. As always, it is assumed the algorithm starts in stationarity. Therefore, the initial position $x^{0,N} \sim \pi^N$, so $x^{0,N} \in X^N$. The acceptance probability only depends on first N coordinates of x and y, so

$$\begin{aligned} \alpha^N\left(x, \xi^N\right) &= \max\left\{1, \frac{\pi^N(y) T^N(y,x)}{\pi^N(x) T^N(x,y)}\right\} \\ &= \max\left\{1, \exp\left(Q_N\left(x, \xi^N\right)\right)\right\}, \end{aligned}$$

where the transition kernel in the Langevin environment is given by the density

$$T^N(x,y) \propto \exp\left\{-\frac{1}{4\delta}\left\|y - x - \delta\mu^N(x)\right\|_{C^N}^2\right\}.$$

The local mean acceptance probability is given by

$$\alpha^N(x) = \mathbb{E}_x\left[\alpha^N\left(x, \xi^N\right)\right].$$

Therefore, the induced Markov chain for $x^N = \left\{x^{k,N}\right\}_{k\geq 0}$ can also be expressed as

$$\begin{cases} y^{k,N} = x^{k,N} + \delta\mu^N\left(x^{k,N}\right) + \sqrt{2\delta}\left(C^N\right)^{\frac{1}{2}} \xi^{k,N} \\ x^{k+1,N} = \gamma^{k,N} y^{k,N} + \left(1 - \gamma^{k,N}\right) x^{k,N}, \end{cases}$$

where $\xi^{k,N} \sim \xi^N$ and $\gamma^{k,N} = \gamma^N\left(x^{k,N}, \xi^{k,N}\right) \sim \text{Bernoulli}\left(\alpha^N\left(x^{k,N}, \xi^{k,N}\right)\right)$. The exponent in acceptance probability can be written as

$$\begin{aligned} Q^N\left(x, \xi^N\right) &= -\frac{1}{2}\left(\|y\|_{C^N}^2 - \|x\|_{C^N}^2\right) - \left(\Psi^N(y) - \Psi^N(x)\right) \\ &= -\frac{1}{4\delta}\left\{\left\|x - y - \delta\mu^N(y)\right\|_{C^N}^2 - \left\|y - x - \delta\mu^N(x)\right\|_{C^N}^2\right\}. \end{aligned} \tag{3.12}$$

Observe that this exponent differs from (3.7) in the incorporation of the new gradient terms. Once projected onto X^N, the Markov chain in $\mathcal{H}^s$ is equivalent to the standard MALA on $\mathbb{R}^N$ for π^N.

3.4.2 Main Theorem and Outline of Proof

Consider $\alpha(\ell) = \mathbb{E}\left[\max\left(1, \exp\left\{Z_\ell\right\}\right)\right]$, with $Z_\ell \sim \mathcal{N}\left(-\frac{\ell^3}{4}, \frac{\ell^3}{2}\right)$, and $h(\ell) = \ell\alpha(\ell)$, where $\alpha(\ell)$ denotes the limiting expected acceptance probability of the MALA algorithm, and $h(\ell)$ is the asymptotic speed function of the limiting diffusion.

The normalizing constant $h(\ell)$ defined above ensures the drift function d^N (3.13) and the martingale difference array (3.14) defined below are asymptotically independent from the parameter ℓ.

Theorem 3.5. (Theorem 5.2 in [PST11]) *Suppose the conditions 3.2 hold, and that the initial condition $x^{0,N}$ of the MALA algorithm is such that $x^{0,N} \sim \pi^N$, and let $z^N(t)$ be a piecewise linear, continuous interpolant of the MALA algorithm with $\Delta t = N^{-\frac{1}{3}}$. Then, for any $T > 0$, $z^N(t) \overset{C([0,T];\mathcal{H}^s)}{\Rightarrow} z(t)$ given by (3.6), with $z(0) \sim \pi$.*

Proof Strategy:

- The main result is that the quantity Q^N defined in (3.12) can be approximated by Gaussian random variables $Q^N\left(x, \xi^N\right) \approx Z_\ell$, with $Z_\ell \sim \mathcal{N}\left(-\frac{\ell^3}{4}, \frac{\ell^3}{2}\right)$.
- Hence, the aforementioned Bernoulli random variables $\gamma^N\left(x, \xi^N\right)$ can be approximated by a single Bernoulli random variable, independent of x, which has probability of success given by $\alpha(\ell) = \mathbb{E}\left[\max\left(1, e^{Z_\ell}\right)\right]$.
- The limiting acceptance probability of the MALA is thus given by $\alpha(\ell)$ in the prior line.
- For $\Delta t = N^{-\frac{1}{3}}$, consider the interpolant yielding the drift function, which acts like an operator on $\mathcal{H}^s$, given by
$$d^N(x) = \left(h(\ell)\Delta t\right)^{-1} \mathbb{E}\left[x^{1,N} - x^{0,N} | x^{0,N} = x\right], \tag{3.13}$$
and the diffusion term, given by the martingale difference array $\left\{\Gamma^{k,N}\right\}_{k\geq 0}$, for $\Gamma^{k,N} = \Gamma^N\left(x^{k,N}, \xi^{k,N}\right)$ with
$$\Gamma^{k,N} = \left(2h(\ell)\Delta t\right)^{-1}\left(x^{k+1,N} - x^{k,N} - h(\ell)\Delta t d^N\left(x^{k,N}\right)\right). \tag{3.14}$$
- Thus, the drift-martingale decomposition of the Markov chain $\left\{x^{k,N}\right\}_{k\geq 0}$ is given by
$$x^{k+1,N} - x^{k,N} = h(\ell)\Delta t d^N\left(x^{k,N}\right) + \sqrt{2h(\ell)\Delta t}\Gamma^{k,N}.$$
- The Gaussian behavior of $Q^N\left(x, \xi^N\right)$ given in (3.12) yields the approximations
$$d^N(x) \approx \mu(x) \qquad \text{and} \qquad \Gamma^{k,N} \approx \mathcal{N}(0, C).$$
- The sequence of continuous processes $\left\{W^N\right\}_{N\geq 1}$, where each continuous process $W^N \in C\left([0,T]; \mathcal{H}^s\right)$ is defined as
$$W^N(t) = \sqrt{\Delta t}\sum_{j=0}^{k}\Gamma^{j,N} + \frac{t - k\Delta t}{\sqrt{\Delta t}}\Gamma^{k+1,N}, \quad \text{for} \quad t \in [k\Delta t, (k+1)\Delta t),$$
converges weakly in $C\left([0;T], \mathcal{H}^s\right)$ to a Brownian motion W in $\mathcal{H}^s$ with covariance operator C_s.
- Main theorem follows from a continuous mapping argument, using the Itô map
$$\Theta : \mathcal{H}^s \times C\left([0,T]; \mathcal{H}^s\right) \to C\left([0,T]; \mathcal{H}^s\right).$$
- The Itô map is continuous, and $z = \Theta\left(z^0, W\right)$ solves the SPDE in (3.6).
- Moreover, the piecewise constant interpolant $\bar{z}^N$ of x^N is defined by $\bar{z}^n(t) = x^k$ for $t \in [t_k, t_{k+1})$, with $t_k = k\Delta t$, and the continuous piecewise linear interpolant z^N from (3.10) satisfies
$$z^N(t) = x^{0,N} + h(\ell)\int_0^t d^N\left(\bar{z}^N(u)\right)du + \sqrt{2h(\ell)}W^N(t), \quad \text{for all } t \in [0,T].$$

- By using closeness of drift function (3.13) with $\mu(x)$, and z^N with $\bar{z}^N$, it follows there exists a process $\hat{W}^N \Rightarrow W$ as $N \to \infty$ such that

$$z^N(t) = x^{0,N} + h(\ell) \int_0^t \mu\left(z^N(u)\right) du + \sqrt{2h(\ell)}\hat{W}^N(t),$$

  so $z^N = \Theta\left(x^{0,N}, \hat{W}^N\right)$.

  - Since Θ is continuous, it follows $z^N = \Theta\left(x^{0,N}, \hat{W}^N\right) \Rightarrow \Theta\left(z^0, W\right) = z$ as $N \to \infty$.

A Infinite Dimensional Analysis

A.1 Introduction

There is no natural analogue of the Lebesgue measure on an infinite dimensional Hilbert space. A natural substitute is provided by Gaussian measures. The theory expounded in these notes define Gaussian measures on a finite dimensional space, and then through an infinite product of measures, on the infinite dimensional Hilbert space $\mathcal{H}$, assumed to be separable.

Later we talk about the Cameron-Martin formula, which is an important tool discussing absolute continuity and singularity of a Gaussian measure and its translates.

We consider first Gaussian measure on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$, where $\mathcal{H}$ is a real separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$, and $\mathcal{B}(\mathcal{H})$ is the Borel sigma algebra on $\mathcal{H}$. Let $\mathcal{L}(\mathcal{H})$ denote the Banach algebra of all continuous linear operators on $\mathcal{H}$, $\mathcal{L}^+(\mathcal{H})$ denote the set of all $T \in \mathcal{L}(H)$ that are symmetric ($\langle Tx, y\rangle = \langle x, Ty\rangle, \forall x, y \in \mathcal{H}$) and positive ($\langle Tx, x\rangle \geq 0, \forall x \in \mathcal{H}$), and finally $\mathcal{L}_1^+(\mathcal{H})$ the set of all operators $Q \in \mathcal{L}^+(\mathcal{H})$ of trace class, i.e., operators Q such that $\operatorname{trace}(Q) := \sum_{k\geq 1} \langle Qe_k, e_k\rangle < \infty$ for all complete orthonormal systems $\{e_k\}_{k\geq 1} \in \mathcal{H}$.

A.1.1 Infinitely Divisible Laws

The law of a random variable X is the probability measure $X_{\#}\mathbb{P}$ on $(E, \mathcal{B}(E))$ defined as

$$X_{\#}\mathbb{P}(I) = \mathbb{P}\left(X^{-1}(I)\right) = \mathbb{P}(X \in I), \quad I \in \mathcal{B}(E).$$

The convolution $\nu_1 \star \nu_2$ of two finite Borel measures ν_1 and ν_2 on $\mathbb{R}^N$ is given by

$$\nu_1 \star \nu_2(\Gamma) = \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \mathbb{I}_\Gamma(\boldsymbol{x} + \boldsymbol{y})\, \nu_1(d\boldsymbol{x})\,(d\boldsymbol{y}), \qquad \Gamma \in \mathcal{B}_{\mathbb{R}^N},$$

and the distribution of the sum of two independent random variables is the convolution of their distributions. We want to describe the probability measure μ that, for each $n \geq 1$, can be written as the n-fold convolution power $\mu_{\frac{1}{n}}^{\star n}$ of some probability measure $\mu_{\frac{1}{n}}$.

Recall that the Fourier transform takes convolution into ordinary multiplication, the Fourier formulation for infinite divisibility involves describing those Borel probability measures on $\mathbb{R}^N$ whose Fourier transform $\hat{\mu}$ has, for each $n \in \mathbb{Z}^+$, an n^{th} root which is again the Fourier transform of a Borel probability measure on $\mathbb{R}^N$.

A.2 Gaussian Measures in Hilbert Spaces

A.2.1 One-dimensional Hilbert spaces

Let us first consider the well-known case where $\mathcal{H}$ is a one-dimensional Hilbert space. Consider $(a, \lambda) \in \mathbb{R}^2$ with $a \in \mathbb{R}$ and $\lambda \in \mathbb{R}^+$. We define a probability measure $\mathcal{N}_{a,\lambda}$ in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ as follows. If $\lambda = 0$, we set

$$\mathcal{N}_{a,0} = \delta_a,$$

where δ_a is the Dirac measure at a,

$$\delta_a(B) = \mathbb{I}_{\{a \in B\}},$$

for $B \in \mathcal{B}(\mathbb{R})$. This is a degenerate Gaussian measure, and is not of much use to us for our applications.

Alternatively, for $\lambda > 0$, we set

$$\mathcal{N}_{a,\lambda}(B) = \frac{1}{\sqrt{2\pi\lambda}} \int_B \exp\left(-\frac{(x-a)^2}{2\lambda}\right) dx, \qquad B \in \mathcal{B}(\mathbb{R}).$$

We are used to using $\lambda = \sigma^2$ in standard notation, but we have not adopted this notation here in accordance to the definition in larger dimensions. Indeed, one can think of the standard deviation in one-dimension as an operator as well. For example, in the case $\sigma^2 = 1$, we can think of it as the identity operator. Note that $\mathcal{N}_{a,\lambda}$ is a probability measure in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ since

$$\begin{aligned} \mathcal{N}_{a,\lambda}(\mathbb{R}) &= \frac{1}{\sqrt{2\pi\lambda}} \int_{\mathbb{R}} \exp\left(-\frac{(x-a)^2}{2\lambda}\right) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-x^2/2\right) dx = 1. \end{aligned}$$

Note that if $\lambda > 0$, then $\mathcal{N}_{a,\lambda}$ is absolutely continuous with respect to the Lebesgue measure, and thus we may write

$$\mathcal{N}_{a,\lambda}(dx) = \frac{1}{\sqrt{2\pi\lambda}} \exp\left(-\frac{(x-a)^2}{2\lambda}\right) dx.$$

The case where $a = 0$ corresponds to centered Gaussian measures, and we denote them by $\mathcal{N}_\lambda$ for short.

We may easily calculate the mean, variance and Fourier transform (characteristic function) of the Gaussian measure $\mathcal{N}_{a,\lambda}$:

$$\begin{aligned} \int_{\mathbb{R}} x\mathcal{N}_{a,\lambda} dx &= \int_{\mathbb{R}} \frac{x}{\sqrt{2\pi\lambda}} \exp\left(-\frac{(x-a)^2}{2\lambda}\right) dx = a \\ \int_{\mathbb{R}} (x-a)^2 \mathcal{N}_{a,\lambda}(dx) &= \int_{\mathbb{R}} \frac{(x-a)^2}{\sqrt{2\pi\lambda}} \exp\left(-\frac{(x-a)^2}{2\lambda}\right) dx = \lambda \\ \widehat{\mathcal{N}_{a,\lambda}}(h) &= \int_{\mathbb{R}} e^{ihx} N_{a,\lambda}(dx) = e^{iah - \frac{1}{2}\lambda h^2}, \quad h \in \mathbb{R}. \end{aligned}$$

Also observe that the normal (standard Gaussian) distribution function Φ is defined by the relation

$$\Phi(t) = \int_{-\infty}^{t} \mathcal{N}_{0,1}(dx).$$

Its inverse function, Φ^{-1}, is defined on $(0,1)$, and by convention we employ the rules $\Phi^{-1}(0) = -\infty$ and $\Phi^{-1}(1) = \infty$. Thus, we arrive at the following lemma which gives bounds for $\Phi(t)$.

Lemma A.1. *For any $t > 0$, the following inequalities hold*

$$\frac{1}{\sqrt{2\pi}} \left(\frac{1}{t} - \frac{1}{t^3}\right) e^{-t^2/2} \le 1 - \Phi(t) \le \frac{1}{\sqrt{2\pi}} \frac{1}{t} e^{-t^2/2}.$$

Proof. By the integration of parts formula, we have

$$\begin{aligned} \int_t^\infty s e^{-s^2/2} \frac{1}{s} ds &= \frac{e^{-t^2/2}}{t} - \int_t^\infty \frac{1}{s^2} e^{-s^2/2} ds \\ &\le \frac{e^{-t^2/2}}{t}. \end{aligned}$$

The lower bound is by an identical argument. □

In order to motivate some of the theory, we look at the following classical result.

Theorem A.2. *Let ξ_n be a sequence of independent, centered Gaussian random variables with variances λ_n. Then the following conditions are equivalent:*

1. *The series $\sum_1^\infty \xi_n$ converges almost everywhere.*

2. *The series in (1) converges in probability.*

3. *The series in (1) converges in L^2.*

4. $\sum_1^\infty \lambda_n < \infty$.

Proof. By the dominated convergence theorem and the independence condition, any of the first three conditions implies convergence of the product

$$\prod_{n=1}^{\infty} \exp\left(i\xi_n\right) dP = \prod_{n=1}^{\infty} \exp\left(-\lambda_n/2\right),$$

which yields the last condition. That the last condition implies the third condition is true by virtue of independence and symmetry of the random variables in question, i.e., we may write

$$\int \left(\xi_k + \cdots + \xi_{k+m}\right)^2 dP = \int \sum_{j=0}^{m} \xi_{k+j}^2 dP,$$

so we have that the second condition is verified. It remains to show that the last condition implies the first. Towards this end, note that the conditional expectation of the square-integrable random variable $\xi = \sum_1^n \xi_n$ with respect to the σ-field $\mathcal{F}_m$, generated by $\xi_1, \ldots, \xi_m$, coincides with $\sum_1^m \xi_n$, since the random variables ξ_n are centered and independent. Furthermore, the sequence of partial sums of the series defining ξ is a martingale with respect to $\{\mathcal{F}_n\}$. Almost sure convergence thus follows from Doob's inequality. □

A.2.2 Finite Dimensional Gaussian measures

We now suppose $\mathcal{H}$ is a finite-dimensional Hilbert space, and proceed to define a Gaussian measure $\mathcal{N}_{a,Q}$ for any $a \in \mathcal{H}$ and any $Q \in \mathcal{L}^+(\mathcal{H})$. Suppose we have a trace class operator, $Q \in \mathcal{L}^+(\mathcal{H})$ and let $(e_1, \ldots, e_d)$ denote an orthonormal basis on $\mathcal{H}$ such that $Qe_k = \lambda_k e_k$, for $k = 1, \ldots, d$, for $\lambda_k \geq 0$. Next we set

$$x_k = \langle x, e_k \rangle, \quad x \in \mathcal{H},\ k = 1, \ldots, d,$$

and identify $\mathcal{H}$ with $\mathbb{R}^d$ through the isomorphism γ,

$$\gamma : \mathcal{H} \to \mathbb{R},\ x \mapsto \gamma(x) = (x_1, \ldots, x_d).$$

We now define a probability measure $\mathcal{N}_{a,Q}$ on $\left(\mathbb{R}^d, \mathcal{B}\left(\mathbb{R}^d\right)\right)$ by setting

$$\mathcal{N}_{a,Q} = \prod_{k=1}^{d} \mathcal{N}_{a_k,\lambda_k}.$$

(Note the product here is defined on the product σ-algebra of the d Gaussian measures). As we had written in the one-dimensional case, when $a = 0$, i.e., the centered case, we shall denote the measure simply by $\mathcal{N}_Q$.

Proposition A.3. *Let $a \in \mathcal{H}$, $Q \in \mathcal{L}^+(\mathcal{H})$ and $\mu = N_{a,Q}$. Then we have*

$$\int_{\mathcal{H}} x\mathcal{N}_{a,Q}(dx) = a,$$

and

$$\int_{\mathcal{H}} \langle y, x-a \rangle \langle z, x-a \rangle \mathcal{N}_{a,Q}(dx) = \langle Qy, z \rangle, \quad y, z \in \mathcal{H}.$$

Furthermore, the Fourier transform of $\mathcal{N}_{a,Q}$ is given by

$$\begin{aligned} \widehat{\mathcal{N}_{a,Q}}(h) &= \int_{\mathcal{H}} e^{i\langle h,x \rangle} \mathcal{N}_{a,Q}(dx) \\ &= e^{i\langle a,h \rangle - \frac{1}{2}\langle Qh,h \rangle}, \end{aligned}$$

for $h \in \mathcal{H}$. Lastly, if the determinant of Q is positive, then $\mathcal{N}_{a,Q}$ is absolutely continuous with respect to Lebesgue measure in $\mathbb{R}^d$ and we have

$$N_{a,Q}(dx) = \frac{1}{\sqrt{(2\pi)^d \det Q}} \exp\left(-\frac{1}{2}\left\langle Q^{-1}(x-a), x-a\right\rangle\right) dx.$$

We call a the mean and Q the covariance operator of $N_{a,Q}$.

By uniqueness of the Fourier transform, we obtain the following result.

Corollary A.4. *Let $\mathcal{H}$ be a finite dimensional Hilbert space, and suppose $a \in \mathcal{H}$, $Q \in \mathcal{L}^+(\mathcal{H})$. Let μ be a probability measure on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$ such that*

$$\int_{\mathcal{H}} e^{i\langle h,x\rangle}\mu(dx) = e^{i\langle a,h\rangle - \frac{1}{2}\langle Qh,h\rangle}, \quad h \in \mathcal{H}.$$

Then we have $\mu = N_{a,Q}$.

We may define Gaussian measures on $\mathbb{R}^n$, or any finite-dimensional $\mathcal{H}$, as the image of standard Gaussian measure on $\mathbb{R}^n$, i.e., as the product of n copies of the standard Gaussian measure on $\mathbb{R}^1$, under affine mappings $x \mapsto \sqrt{Qx} + a$. On the linear subspace $\sqrt{K}(\mathbb{R}^n)$ we may define the inner product

$$(u,v)_{\mathcal{N}_{a,Q}} = \left(\sqrt{Q}^{-1}u, \sqrt{Q}^{-1}v\right).$$

A.2.3 Measures in Hilbert spaces

Consider an infinite dimensional separable Hilbert space $\mathcal{H}$ and suppose $\{e_k\}_{k\geq 1}$ denotes a complete orthonormal system in $\mathcal{H}$. For any $n \in \mathbb{N}$, consider the projection mapping $P_n : \mathcal{H} \to P_n(\mathcal{H})$ defined as

$$P_n x = \sum_{k=1}^{n} \langle x, e_k\rangle e_k, \quad x \in \mathcal{H}. \tag{A.1}$$

Clearly, $\lim_{n\to\infty} P_n x = x$ for all $x \in \mathcal{H}$.

Proposition A.5. *Suppose μ, ν are (Borel) probability measures such that*

$$\int_{\mathcal{H}} \varphi(x)\mu(dx) = \int_{\mathcal{H}} \varphi(x)\nu(x)\,dx,$$

for all continuous and bounded functions $\varphi : \mathcal{H} \mapsto \mathbb{R}$ in $\mathcal{H}$. Then $\mu = \nu$.

Proof. See Prop 1.5 in [Da06]. □

Proposition A.6. *Let μ and ν be probability measures on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$. If $(P_n)_{\#}\mu = (P_n)_{\#}\nu$ for any $n \in \mathbb{N}$, then $\mu = \nu$.*

Proposition A.7. *Let μ and ν be probability measures on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$. If $\hat{\mu}(h) = \hat{\nu}(h)$ for all $h \in \mathcal{H}$, then we have $\mu = \nu$.*

A.2.4 Mean and Covariance of Probability Measures

Consider a fixed probability measure μ on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$. Assume that

$$\int_{\mathcal{H}} |x|\,\mu(dx) < \infty.$$

Then, for any $h \in \mathcal{H}$, the linear functional $F : \mathcal{H} \to \mathbb{R}$ defined as

$$F(h) = \int_{\mathcal{H}} \langle x, h\rangle \mu(dx), \quad h \in \mathcal{H},$$

is continuous, since

$$|F(h)| \leq \int_{\mathcal{H}} |x| \, \mu(dx) \, |h|, \quad h \in \mathcal{H}.$$

It follows from the Riesz representation theorem there exists $m \in \mathcal{H}$ such that

$$\langle m, h \rangle = \int_{\mathcal{H}} \langle x, h \rangle \, \mu(dx), \quad h \in \mathcal{H}.$$

We call m the mean of the probability measure μ. Furthermore, we shall write

$$m = \int_{\mathcal{H}} x\mu(dx).$$

If we further assume that

$$\int_{\mathcal{H}} |x|^2 \, \mu(dx) < \infty,$$

then we may consider a bilinear form $G : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ defined as

$$G(h,k) = \int_{\mathcal{H}} \langle h, x-m \rangle \langle k, x-m \rangle \, \mu(dx), \quad h,k \in \mathcal{H}.$$

Furthermore, another application of the Riesz representation theorem guarantees the existence of a unique linear bounded operator $Q \in \mathcal{L}(\mathcal{H})$ such that

$$\langle Qh, k \rangle = \int_{\mathcal{H}} \langle h, x-m \rangle \langle k, x-m \rangle \, \mu(dx), \qquad h,k \in \mathcal{H}.$$

The operator Q is called the covariance of the probability measure μ.

Proposition A.8. *Let μ be a probability measure on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$ with mean m and covariance operator Q. Then $Q \in \mathcal{L}_1^+(\mathcal{H})$, i.e., Q is symmetric, positive, and of trace class.*

Proof. That Q is symmetric and positive is clear from its construction. To prove that it is of trace class, observe

$$\langle Qe_k, e_k \rangle = \int_{\mathcal{H}} |\langle x-m, e_k \rangle|^2 \, \mu(dx), \quad k \in \mathbb{N}.$$

By the monotone convergence theorem and an application of Parseval's identity, we have

$$\text{trace}Q = \sum_{k=1}^{\infty} \int_{\mathcal{H}} \left| \langle x-m, e_k \rangle^2 \right| \mu(dx) = \int_{\mathcal{H}} |x-m|^2 \, \mu(dx) < \infty.$$

□

A.3 Gaussian Measures

For $a \in \mathcal{H}$ and $Q \in \mathcal{L}_1^+(\mathcal{H})$, we call the measure $\mu := \mathcal{N}_{a,Q}$ on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$ a Gaussian measure μ with mean a and covariance operator Q. Its Fourier transform is given by

$$\widehat{\mathcal{N}_{a,Q}}(h) = \exp\left\{ i \langle a, h \rangle - \frac{1}{2} \langle Qh, h \rangle \right\}, \quad h \in \mathcal{H}.$$

The Gaussian measure $\mathcal{N}_{a,Q}$ is said to be non-degenerate if $\text{Ker}(Q) = \{x \in \mathcal{H} : Qx = 0\} = \{0\}$.

A.3.1 Existence and Uniqueness

We can show that for arbitrary $a \in \mathcal{H}$ and $Q \in \mathcal{L}_1^+(\mathcal{H})$, there exists a unique Gaussian measure $\mu = \mathcal{N}_{a,Q}$ in $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$. Observe that since $Q \in \mathcal{L}_1^+(\mathcal{H})$, there exists a complete orthonormal basis $\{e_k\}_{k \geq 1}$ on $\mathcal{H}$ and a sequence of non-negative numbers $\{\lambda_k\}_{k \geq 1}$ such that

$$Qe_k = \lambda e_k, \quad k \in \mathbb{N}.$$

Thus, for any $x \in \mathcal{H}$, we may set $x_k = \langle x, e_k \rangle$, $k \in \mathbb{N}$. Let us now consider the natural isomorphism γ between $\mathcal{H}$ and the Hilbert space ℓ^2 of all sequences $\{x_k\}_{k \geq 1}$ of real numbers such that $\sum_{k=1}^{\infty} |x_k|^2 < \infty$,defined by

$$\mathcal{H} \to \ell^2, \, x \mapsto \gamma(x) = \{x_k\}.$$

We thus may identify $\mathcal{H}$ with ℓ^2. We shall later consider the product measure

$$\mu := \prod_{k=1}^{\infty} N_{a_k, \lambda_k}.$$

Although μ is defined on the space $\mathbb{R}^{\infty} = \prod_{k=1}^{\infty} \mathbb{R}$ rather than ℓ^2, we shall show that it is concentrated in ℓ^2. Furthermore, we shall prove that μ is a Gaussian measure $N_{a,Q}$.

A.3.2 Preliminaries on Countable Product of Measures

Let (μ_k) be a sequence of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We want to define a product measure on the space $\mathbb{R}^{\infty} = \prod_{k=1}^{\infty} \mathbb{R}$, consisting of all the sequences $x = (x_k)$ of real numbers. Let us endow $\mathbb{R}^{\infty}$ with the metric

$$d(x, y) = \sum_{n=1}^{\infty} 2^{-n} \frac{\max\{|x_k - y_k| : 1 \leq k \leq n\}}{1 + \max\{|x_k - y_k| : 1 \leq k \leq n\}}.$$

With this metric, $\mathbb{R}^{\infty}$ is a complete metric space and the topology corresponding to this metric is in fact the product topology.

We then define $\mu = \prod_{k=1}^{\infty} \mu_k$ on the family $\mathcal{C}$ of all cylindrical subsets $I_{n,A}$ of $\mathbb{R}^{\infty}$, where $n \in \mathbb{N}$ and $A \in \mathcal{B}(\mathbb{R}^n)$, and

$$I_{n,A} = \{x = (x_k) \in \mathbb{R}^{\infty} : (x_1, \ldots, x_n) \in A\}.$$

Observe that

$$I_{n,A} = I_{n+k, A \times X_{n+1} \times \cdots \times X_{n+k}}, \quad k, n \in \mathbb{N}.$$

Hence, with this identity, it is easy to observe that $\mathcal{C}$ is an algebra. In particular, if $I_{n,A}$ and $I_{m,B}$ are two cylindrical sets, we then have

$$\begin{aligned} I_{n,A} \cup I_{m,B} &= I_{m+n, A \times X_{n+1} \times \cdots \times X_{n+m}} \cup I_{m+n, B \times X_{m+1} \times \cdots \times X_{m+n}} \\ &= I_{m+n, A \times X_{n+1} \times \cdots \times X_{n+m} \cup B \times X_{m+1} \times \cdots \times X_{m+n}}, \end{aligned}$$

and $I_{n,A}^c = I_{n,A^c}$. More crucially, the σ-algebra generated by $\mathcal{C}$ coincides with $\mathcal{B}(\mathbb{R}^{\infty})$ since any ball (with respect to the metric of $\mathbb{R}^{\infty}$) is a countable intersection of cylindrical sets.

Let us now consider the product measure

$$\mu(I_{n,A}) = (\mu_1 \times \cdots \times \mu_n)(A), \quad I_{n,A} \in \mathcal{C}.$$

The previous relation shows that μ is additive. An application of Caratheodory's extension theorem shows that μ is σ-additive on $\mathcal{C}$, and so μ can be uniquely extended to a probability measure on the product σ-algebra $\mathcal{B}(\mathbb{R}^{\infty})$.

Theorem A.9. *The measure μ is aσ-additive on $\mathcal{C}$ and it possesses a unique extension to a probability measure $(\mathbb{R}^{\infty}, \mathcal{B}(\mathbb{R}^{\infty}))$.*

Proof. See Theorem 1.9 in [Da06]. □

A.3.3 Definition of Gaussian measures

We now show that

$$\mu = \prod_{k=1}^{\infty} \mathcal{N}_{a_k, \lambda_k}$$

is a Gaussian measure on $\mathcal{H} = \ell^2$ with mean a and covariance Q. First observe that ℓ^2 is Borel subset of $\mathbb{R}^{\infty}$. Furthermore, we have $\mu(\ell^2) = 1$. The second statement follows from the monotone convergence theorem,

$$\int_{\mathbb{R}^{\infty}} \sum_{k=1}^{\infty} x_k^2 \mu(dx) \quad = \quad \sum_{k=1}^{\infty} \int_{\mathbb{R}} x_k^2 N_{a_k \lambda_k} d(x_k) = \sum_{k=1}^{\infty} (\lambda_k + a_k^2). \tag{A.2}$$

Thus,

$$\mu\left(\left\{x \in \mathbb{R}^{\infty} : |x_k|_{\ell^2}^2 < \infty\right\}\right) = 1.$$

Theorem A.10. *There exists a unique probability measure μ on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$ with mean a, covariance Q and Fourier transform*

$$\hat{\mu}(h) = \exp\left\{ i \langle a, h \rangle - \frac{1}{2} \langle Qh, h \rangle \right\}, \quad h \in \mathcal{H}. \tag{A.3}$$

We denote μ by $N_{a,Q}$, and if $a = 0$ we simply write N_Q.

Proof. We check that the restriction of ℓ^2 of the product measure μ, defined by (A.2) fulfills the necessary properties.

First note that by (A.2) we have

$$\int_{\mathcal{H}} \left| x^2 \right| \mu(dx) = \text{trace} Q + |a|^2. \tag{A.4}$$

Assume for simplicity, that $\text{Ker}(Q) = \{0\}$, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq \cdots$. Let (P_n) be the sequence of projectors defined by (A.1), and let $h \in \mathcal{H}$. Since $|\langle x, h \rangle| \leq |x| |h|$, and $\int_{\mathcal{H}} |x| \mu(dx)$ is finite by (A.4), the dominated convergence theorem thus implies

$$\int_{\mathcal{H}} \langle x, h \rangle \mu(dx) = \lim_{n \to \infty} \int_{\mathcal{H}} \langle P_n x, h \rangle \mu(dx).$$

However,

$$\begin{aligned} \int_{\mathcal{H}} \langle P_n x, h \rangle \mu(dx) &= \sum_{k=1}^{n} \int_{\mathcal{H}} x_k h_k \mu(dx) \\ &= \sum_{k=1}^{n} h_k \int_{\mathbb{R}} x_k N_{a_k, \lambda_k}(dx_k) \\ &= \sum_{k=1}^{n} h_k a_k = \langle P_n a, h \rangle \to \langle a, h \rangle, \end{aligned}$$

as $n \to \infty$. Therefore, the mean of μ is a. Likewise, in order to determine the covariance of μ, we fix $y, z \in \mathcal{H}$ and write

$$\int_{\mathcal{H}} \langle x - a, y \rangle \langle x - a, z \rangle \mu(dx) = \lim_{n \to \infty} \int_{\mathcal{H}} \langle P_n(x - a), y \rangle \langle P_n(x - a), z \rangle \mu(dx).$$

Furthermore,

$$\begin{aligned} \int_{\mathcal{H}} \langle P_n(x - a), y \rangle \langle P_n(x - a), z \rangle \mu(dx) &= \sum_{k=1}^{n} \int_{\mathcal{H}} (x_k - a_k)^2 y_k z_k \mu(dx) \\ &= \sum_{k=1}^{n} y_k z_k \int_{\mathbb{R}} (x_k - a_k)^2 N_{a_k, \lambda_k}(dx_k) \\ &= \sum_{k=1}^{n} y_k z_k \lambda_k = \langle P_n Q y, z \rangle \to \langle Qy, z \rangle, \end{aligned}$$

as $n \to \infty$. Hence, the covariance of μ is indeed Q.

Finally, for any $h \in \mathcal{H}$,

$$\begin{aligned} \int_{\mathcal{H}} e^{i\langle x, h \rangle} \mu(dx) &= \lim_{n \to \infty} \int_{\mathcal{H}} e^{i\langle P_n x, h \rangle} \mu(dx) \\ &= \lim_{n \to \infty} \prod_{k=1}^{n} \int_{\mathbb{R}} e^{i x_k h_k} N_{a_k, \lambda_k}(dx_k) \\ &= \lim_{n \to \infty} \prod_{k=1}^{n} e^{i a_k h_k - \frac{1}{2} \lambda_k h_k^2} \\ &= \lim_{n \to \infty} e^{i\langle P_n a, h \rangle} e^{-\frac{1}{2} \langle P_n Q h, h \rangle} \\ &= e^{i\langle a, h \rangle} e^{-\frac{1}{2} \langle Qh, h \rangle}. \end{aligned}$$

This verifies that the Fourier transform of μ is given by (A.3). □

Note A.11. For any $\varepsilon < \frac{1}{\lambda_1}$, the linear operator $1 - \varepsilon Q$ is invertible, and $(1 - \varepsilon Q)^{-1}$ is bounded. Furthermore,

$$(1 - \varepsilon Q)^{-1} x = \sum_{k=1}^{\infty} \frac{1}{1 - \varepsilon \lambda_k} \langle x, e_k \rangle e_k, \quad x \in \mathcal{H}.$$

In this case, we define the determinant of $(1 - \varepsilon Q)$ by setting

$$\det (1 - \varepsilon Q) := \lim_{n \to \infty} \prod_{k=1}^{n} (1 - \varepsilon \lambda_k) := \prod_{k=1}^{\infty} (1 - \varepsilon \lambda_k).$$

By assuming $\sum_{k \geq 1} \lambda_k < \infty$, the infinite product above is finite and positive.

A.4 Brownian Motion and Wiener Measure

Brownian motion is a real-valued stochastic process. They are interesting to study since an extremely large set of interesting class of processes contain Brownian motion, including, Gaussian processes, Markov processes, diffusions, Lévy processes, and more.

A.4.1 Infinite Dimensional Wiener Process

In the sections previous, we have discussed the standard finite dimensional Wiener process $W_t = \left(w_t^{(1)}, \dots, w_t^{(d)} \right)$. We now consider an infinite sequence $W_t = \left\{ w_t^{(j)} \right\}_{j \geq 1}$ of independent Wiener processes. Treating W_t as an infinite sequence at each t is the point of view of the theory of cylindrical Wiener process. We focus on the view where W_t is a random element of a state space in which the whole stochastic process $\{W_t\}_{t \geq 0}$ could be realized.

A.4.2 Wiener Measure

The distribution of Brownian motion is called Wiener measure in honor of the first mathematician to construct it, Norbert Wiener. Wiener constructed this measure by looking at a Gaussian measure on an infinite dimensional space, and we follow his method of construction in these notes. We are thus dealing with measures on a real Banach space E that are centered Gaussian in the sense that for each $x^* \in E^*$, $x \in E \mapsto \langle x, x^* \rangle \in \mathbb{R}$ is a centered Gaussian random variable. We thus call the measure this measure in consideration a centered Gaussian measure at x.

Note that $C\left(\mathbb{R}^N\right)$ is a separable, complete metric space. We would like to consider a separable Banach space. With this in mind, we define $\Theta\left(\mathbb{R}^N\right)$ to be the space of continuous paths $\boldsymbol{\theta} : [0, \infty) \to \mathbb{R}^N$ with the properties that $\boldsymbol{\theta}(0) = 0$ and $\lim_{t \to \infty} t^{-1} |\boldsymbol{\theta}(t)| = 0$.

Lemma A.12. *The map*

$$\boldsymbol{\psi} \in C\left(\mathbb{R}^N\right) \mapsto \|\boldsymbol{\psi}\|_{\Theta(\mathbb{R}^N)} \equiv \sup_{t \geq 0} \frac{|\boldsymbol{\psi}(t)|}{1 + t} \in [0, \infty]$$

is lower semicontinuous, and the pair $\left(\Theta\left(\mathbb{R}^N\right), \|\cdot\|_{\Theta(\mathbb{R}^N)}\right)$ *is a separable Banach space that is continuously embedded as a Borel measurable subset of* $C\left(\mathbb{R}^N\right)$*. Furthermore,* $\mathcal{B}_{\Theta(\mathbb{R}^N)}$ *coincides with* $\mathcal{B}_{C(\mathbb{R}^N)}\left[\Theta\left(\mathbb{R}^N\right)\right] = \left\{A \cap \Theta\left(\mathbb{R}^N\right) : A \in \mathcal{B}_{C(\mathbb{R}^N)}\right\}$*. Moreover, the dual space* $\Theta\left(\mathbb{R}^N\right)^*$ *of* $\Theta\left(\mathbb{R}^N\right)$ *can be identified with the space of* $\mathbb{R}^N$*-valued, Borel* $\boldsymbol{\lambda}$ *on* $[0, \infty)$ *with the properties that* $\boldsymbol{\lambda}(\{\mathbf{0}\}) = 0$ *and*

$$\|\boldsymbol{\lambda}\|_{\Theta(\mathbb{R}^N)^*} = \int_{[0,\infty)} (1 + t) |\boldsymbol{\lambda}| (dt) < \infty,$$

when the duality relation is given by

$$\langle \boldsymbol{\theta}, \boldsymbol{\lambda} \rangle = \int_{[0,\infty)} \boldsymbol{\theta}(t) \cdot \boldsymbol{\lambda}(dt).$$

If $(\boldsymbol{B}(t), \mathcal{F}_t, \mathbb{P})$ *is a* $\mathbb{R}^N$*-valued Brownian motion, then* $\boldsymbol{B} \in \Theta\left(\mathbb{R}^N\right)$ $\mathbb{P}$*-almost surely and*

$$\mathbb{E}^{\mathbb{P}}\left[\|\boldsymbol{B}\|_{\Theta(\mathbb{R}^N)}^2\right] \leq 32N.$$

Thus, $\mathbb{R}^N$-valued Brownian motion induces a Borel measure $\mathcal{W}^{(N)}$ on the separable Banach space $\Theta\left(\mathbb{R}^N\right)$, and throughout we refer to this measure as the classical Wiener measure.

Lemma A.13. *Let E with norm $\|\cdot\|_E$ be a separable, real Banach space, and denote*

$$(x, x^*) \in E \times E^* \mapsto \langle x, x^* \rangle \in \mathbb{R}$$

the duality relation between E and its dual space E^*. The Borel field $\mathcal{B}_E$ coincides with the σ-algebra generated by the maps $x \in E \mapsto \langle x, x^* \rangle$ as x^* runs over E^*. For any Borel probability measure μ on E, define its Fourier transform $\hat{\mu} : E^* \to \mathbb{C}$ by

$$\hat{\mu}\left(x^*\right) = \int_E \exp\left[\sqrt{-1}\left\langle x, x^*\right\rangle\right] \mu\left(dx\right), \qquad x^* \in E^*,$$

and then $\hat{\mu}$ i s continuous function of weak convergence on Θ^*, and $\hat{\mu}$ uniquely determines in the sense that if ν is a second element of $\boldsymbol{M}_1\left(\Theta\right)$ and $\hat{\mu} = \hat{\nu}$, then $\mu = \nu$.*

A.4.3 Cameron-Martin Space

To proceed from Wiener measure on a Banach space to Wiener on a Hilbert space, we use the results from Cameron-Martin. Note, that in finite dimensions, every centered, Gaussian measure with non-degenerate covariance can be thought of as the canonical/standard Gaussian measure on a Hilbert space. If $\mu_{\mathbf{0},\boldsymbol{C}}$ is the Gaussian measure on $\mathbb{R}^N$ with mean $\mathbf{0}$ and non-degenerate covariance $\boldsymbol{Q}$, consider $\mathbb{R}^N$ as a Hilbert space $\mathcal{H}$ with inner product $\left(\boldsymbol{g}, \boldsymbol{h}\right)_{\mathcal{H}} = \left(\boldsymbol{g}, \boldsymbol{Q}\boldsymbol{h}\right)_{\mathbb{R}^N}$, and take $\lambda_{\mathcal{H}}$ to be the natural Lebesgue measure, the one which assigns measure 1 to a unit cube in $\mathcal{H}$, or equivalently, the one which is obtained by pushing the usual Lebesgue measure $\lambda_{\mathbb{R}^N}$ forward under the linear transformation $\boldsymbol{C}^{\frac{1}{2}}$.We may thus write

$$\mu_{\mathbf{0},\boldsymbol{Q}}\left(d\boldsymbol{h}\right) = \frac{1}{\left(2\pi\right)^{\frac{N}{2}}} \exp\left(-\frac{1}{2}\left\|\boldsymbol{h}\right\|_{\mathcal{H}}^2\right) \lambda_{\mathcal{H}}\left(d\boldsymbol{h}\right)$$

and

$$\widehat{\mu_{\mathbf{0},\boldsymbol{Q}}}\left(\boldsymbol{h}\right) = \exp\left(-\frac{1}{2}\left\|\boldsymbol{h}\right\|_{\mathcal{H}}^2\right).$$

One question would be the equivalence and singularity of the measures $\mu_{\mathbf{0},\boldsymbol{Q}}$ and $\mu_{a,\boldsymbol{Q}}$ for some $a \in \mathcal{H}$.

Recall that for two measures on $\left(\Omega, \mathcal{F}\right)$, μ and ν, we say that μ is absolutely continuous with respect to ν, which we will denote by $\mu \ll \nu$, if for each null set of ν, $\nu\left(A\right) = 0$, we have $\mu\left(A\right) = 0$. An important result regarding absolute continuity of measures is the Radon-Nikodym derivative (see [SS05]), which states that if $\mu \ll \nu$, then there exists a unique function $f \in L^1\left(\Omega, \mathcal{F}, \nu\right)$ such that

$$\mu\left(A\right) = \int_A f d\nu,$$

for all $A \in \mathcal{F}$. If $\mu \ll \nu$ and $\nu \ll \mu$, then we say μ and ν are equivalent. Clearly, μ and ν are both absolutely continuous with respect to $\zeta = \frac{1}{2}(\mu + \nu)$ on $(\Omega, \mathcal{F})$. This fact gives rise to the Hellinger integral of μ and ν, defined by

$$H\left(\mu, \nu\right) = \int_\Omega \sqrt{\frac{d\mu}{d\zeta}\frac{d\nu}{d\zeta}} d\zeta.$$

If $H\left(\mu, \nu\right) = 0$, say the pair μ and ν are singular.

Theorem A.14. ***(Cameron Martin Theorem, see Theorem 2.8 in [Da06])*** *Let $a \in \mathcal{H}$, and consider two measures $\mu = \mathcal{N}_{\mathbf{0},\boldsymbol{Q}}$ and $\nu = \mathcal{N}_{a,\boldsymbol{Q}}$, with $Q \in L_1^+\left(\mathcal{H}\right)$.*

1. *If $a \notin Q^{1/2}\left(\mathcal{H}\right)$, then μ and ν are singular.*

2. *If $a \in Q^{1/2}\left(\mathcal{H}\right)$, then μ and ν are equivalent.*

3. *If μ and ν are equivalent, then the density $\frac{d\nu}{d\mu}$ is given by*

$$\frac{d\nu}{d\mu}\left(x\right) = \exp\left\{-\frac{1}{2}\left|Q^{-1/2}a\right|^2 + W_{Q^{-1/2}a}\left(x\right)\right\}, \qquad x \in \mathcal{H}.$$

A.5 Lack of a Canonical Gaussian Distribution in Infinite Dimensions

The construction above the Gaussian measure in an infinite-dimensional Hilbert space $\mathcal{H}$, requires that the covariance operator of the Gaussian measure is of trace class, i.e., the trace is finite. Amongst other issues, this requirement rules out the canonical $\mathcal{N}(0, I_\infty)$ Gaussian distribution.

Consider the projection of this distribution onto finite subspaces of the Hilbert space. We would like this distribution to be Gaussian as well. Furthermore, we would want every ϵ-ball in the Hilbert space to have non-zero probability measure, otherwise we are dealing with the trivial measure (in that we assign measure zero to every bounded set; one can further generalize by the method of cylindrical measure, which is a way of setting a countable union of bounded sets to have measure one) which is of not much use. However, one can easily observe that the measure of the projection of any particular ϵ-ball onto a finite subspace is greater than the measure of the original ϵ-ball, while the measure of the projection converges to zero as the dimension increases.

Probabilistically, we can use the following argument to explain why no such distribution can exist. Suppose $\mathcal{H}$ is an infinite-dimensional Hilbert space equipped with a Gaussian Borel measure μ. Furthermore, suppose that the covariance matrix of μ is the identity. Let $\{e_1, e_2, \dots\}$ be an orthonormal sequence. If we define $X_i : \mathcal{H} \mapsto \mathbb{R}$ as the continuous linear functional $X_i(x) = \langle x, e_i \rangle$, then X_i is a centered Gaussian random variable on the probability space $(\mathcal{H}, \mu)$ with unit variance. Since the X_i are jointly Gaussian distributed and are uncorrelated (since the e_i were orthogonal), they are thus independent. Hence, $\{X_i\}$ forms an i.i.d. $\mathcal{N}(0,1)$ sequence. However,

$$\sum_i |X_i(x)|^2 = \sum_i |\langle x, e_i \rangle|^2 \leq \|x\|^2 < \infty,$$

by Bessel's inequality (see page 166 in [SS05]). In particular, we have that $X_i \to 0$ surely, which is an absurd implication for an i.i.d. sequence.

If we take a deeper look at the sequence of the finite canonical Gaussians, we my intuitively expect them to converge to a point mass (Dirac δ-distribution) at infinity in the Hilbert space, or we might expect them to lead to a distribution in the surface of an infinite dimensional hypersphere with infinite radius (these are two different implications however, due to the strong law of large numbers).

If we have a vector of i.i.d. Gaussian random variables, $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, then the squares of the vector components are χ^2 random variables with one degree of freedom. Therefore, as $d \to \infty$, we have

$$\frac{1}{d}\sum_{i=1}^{d} X_i^2 = 1, \text{ almost surely.}$$

That is, if $X = (X_1, \dots, X_d)$, then the norm of X,

$$\|X\| \to \sqrt{d}, \text{ almost surely as } d \to \infty.$$

Furthermore, by the strong law of large numbers we have that the mean vector is the zero vector, so whatever the infinite-dimensional distribution may be, it is also centered at the origin. Furthermore, if X and X' are any two such vectors, then

$$\left\langle \frac{X}{\|X\|}, \frac{X'}{\|X'\|} \right\rangle \xrightarrow{d\to\infty} 0.$$

Since X is the zero vector with probability zero, it follows that almost surely each pair of vectors is orthogonal in the limit. This leads to the implication that points end up mutually orthogonal in the surface of a hypersphere about the origin.

B Markov Processes

In this section we provide some results and proofs about the class of continuous parameter stochastic processes that satisfy the Markov property, i.e., those processes that are known as Markov processes. The aim here is to discuss briefly the weak convergence of these processes, which will require a foray into the infinitesimal generator of the process and a look at operator semigroups. The presentation here is necessarily concise, only looking at those elements which are of immediate concern to the topic of this paper. For a comprehensive survey of convergence of Markov processes and characterization methods of such processes, the reader is referred to the classic text of Etheir and Kurtz [EK86], from which the majority of these notes are based. A more modern treatment of the theory expounded here can be found in Chapter VI of Jean Jacod's and Albert Shiryaev's text [JS03], Section III of Cosma Shalizi's manuscript entitled Almost None of the Theory Stochastic Processes and Olav Kallenberg's Probability textbook (Chapters 7, 14 and 17 in particular), [SK11, Kal97].

It is worth mentioning that the theory expounded here is not the unique theory for which convergence of Markov processes has developed. In fact, there are three different strands of development that have been used to deal with the convergence of Markov processes. The method used used, using operator semigroups convergence theorems, is only one method of proving Markov process convergence. This method was originally spearheaded by H.F. Trotter using diffusion approximations. A second approach was developed by Daniel Stroock and S.R.S. Varadhan also using convergence of generators, but instead of using operator semigroups, they authors developed weak compactness arguments and the martingale characterization of the limit. The last method used requires representing the process as a solution to a stochastic equation, and then by use of classical analysis, one tries to show that convergence of the equations implies convergence of their solutions.

B.1 Introduction to the Problem

When proofs of weak convergence were given for the Metropolis algorithm (properly rescaled) to the Langevin SDE in this paper, it was supposed that this convergence was under the Skorokhod topology, and this convergence relied on some general conditions that needs to be elucidated. The main result that provides the necessary characterization of convergence is the fact that Markov processes can be expressed according to a corresponding operator semigroup, which then allows one to characterize the process by the generator of this semigroup. This is the result of Theorem 8.2 of Chapter 4 in [EK86]. It is through convergence of generators (in the appropriate sense) that implies convergence of semigroups, which in turn provides convergence of Markov processes. Furthermore, weak convergence is characterized by convergence in the Skorokhod topology, which is the topology that equips the space of right continuous functions with left limits in such a manner that the resulting space is complete and separable (i.e., the space is Polish).

B.2 The Skorokhod Topology

The space $\mathbb{D}\left(\mathbb{R}^d\right)$ is defined to be the space of cádlág functions on $[0,\infty)$, i.e., those functions which are right continuous and possess left limits. Such functions are also said to *possess discontinuities of the first kind*, i.e., for each $t \geq 0$ and each function $x \in \mathbb{D}\left(\mathbb{R}^d\right)$, the limits $\lim_{x\to t^+} x(s) = x(t)$ and $\lim_{x\to t^-} x(s) = x(t)$ exist. It would be a great advantage if for a complete and separable metric space E, there is a metric ρ that makes $\mathbb{D}\left(\mathbb{R}^d\right)$ complete and separable as well. This metric induces the aforementioned Skorokhod topology. The main usefulness of this topology for the purposes used in the topic of optimal scaling is that it allows for a small deformation of the time scale. Indeed, for a probability space μ, the Skorokhod topology is the coarsest topology for which the mappings: $\mu \rightsquigarrow \mu(f)$ are continuous for all bounded continuous functions f on $\mathbb{D}\left(\mathbb{R}^d\right)$.

B.3 Operator Semigroups

As mentioned in the introduction to this section, operator semigroups are a useful tool for developing the characterization and convergence of Markov processes. Consider a real Banach space (complete normed vector space) L with norm $\|\cdot\|$.

Definition B.1. (Page 6, [EK86]) A one-parameter family $\{T(t)\}_{t\geq 0}$ of bounded linear operators on L is called a semigroups if

1. $T(0) = I$, where I denotes the identity matrix.
2. $T(s+t) = T(t)T(s)$ for all $s, t \geq 0$.

A semigroup is thus an operator varying with t.

Definition B.2. (Page 6, [EK86]) A semigroup $\{T(t)\}$ on L is said to be strongly continuous if for all $h \in L$

$$\lim_{t\to 0} T(t)h = h.$$

Definition B.3. (Page 6, [EK86]) A contraction semigroup is such that $\|T(t)\| \leq 1$ for all $t \geq 0$, where

$$\|T(t)\| = \sup_{\substack{h\in L\\ h\neq 0}} \frac{\|T(t)h\|}{\|h\|}$$

denotes the operator norm.

The following definition of a generator of a semigroup can be used to characterize a semigroup on L.

Definition B.4. (Page 8, [EK86]) The infinitesimal generator of a semigroup $\{T(t)\}$ on L is the linear operator A defined by

$$Ah = \lim_{t \to 0} \left(\frac{T(t)h - h}{t} \right).$$

The domain $\mathcal{D}(A)$ of A is the subspace of all $h \in L$ for which this limit exists.

The following proposition says that strongly continuous contraction semigroups are uniquely determined by their corresponding generator.

Proposition B.5. *Let $\{T(t)\}$ and $\{S(t)\}$ be strongly continuous contraction semigroups on L with generators A and B, respectively. If $A = B$, then $T(t) = S(t)$ for all $t \geq 0$.*

Proof. See page 15 in [EK86]. □

B.4 Markov Processes and Semigroups

For a given metric space (E, r), denote the collection of all real-valued, Borel measurable functions on E by $\mathcal{M}(E)$. The Banach space of bounded functions with supremum norm, $B(E)$ is a subset of $\mathcal{M}(E)$, i.e., any bounded function with

$$\|h\| = \sup_{x \in E} |h(x)|,$$

is Borel measurable . Finally denote the space of all bounded continuous functions on E (which are clearly Borel measurable on E) by $\bar{C}(E)$.

Let $\{X(t)\}_{t \geq 0}$ be a stochastic process defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in E, and denote the filtration up to t by $\mathcal{F}^X(t) = \sigma(X(s) : s \leq t)$. The process X is said to be Markov if

$$\mathbb{P}\left(X(t+s) \in \Gamma | \mathcal{F}^X(t)\right) = \mathbb{P}\left(X(t+s) \in \Gamma | X(t)\right),$$

for all $s, t \geq 0$ and $\Gamma \in \mathcal{B}(E)$, where $\mathcal{B}(E)$ is the Borel sigma algebra corresponding to the metric space E Therefore, for any bounded function h we have

$$\mathbb{E}\left[h(X(t+s)) | \mathcal{F}^X(t)\right] = \mathbb{E}\left[h(X(t+s)) | X(t)\right],$$

for all $s, t \geq 0$.

Definition B.6. An operator semigroup may be defined as

$$T(t)h(x) = \int h(y) \mathbb{P}(t, x, dy).$$

Remark B.7. Observe that

$$\begin{aligned}
\|T(t)\| &= \sup_{\substack{h \text{ bounded on } E \\ h \neq 0}} \frac{\|T(t)h\|}{\|h\|} \\
&= \sup_{\substack{h \text{ bounded on } E \\ h \neq 0}} \left(\frac{\sup_{x \in \mathbb{R}} |T(t)h(x)|}{\sup_{x \in \mathbb{R}} |h(x)|} \right) \\
&= \sup_{\substack{h \text{ bounded on } E \\ h \neq 0}} \left(\frac{\sup_{x \in \mathbb{R}} \left|\int h(y) \mathbb{P}(t, x, dy)\right|}{\sup_{x \in \mathbb{R}} \left|\int h(y) \delta_x(dy)\right|} \right) \leq 1.
\end{aligned}$$

Therefore, $\{T(t)\}$ is a contraction semigroup by Definition B.3.

C R Codes for Simulations

Listed below are various functions with annotations used for the simulations displayed in this paper.

C.1 Code for Example

This code corresponds to Example 2.12.

```
> library(mvtnorm)
> # it is the number of iterations for each Monte Carlo
> #   Simulation
> it = 50000
> # number of dimensions for target
> dim = 10
> # n is number of scaling parameters used
> n = 50
> # vector of scaling terms
> ell = (1:n)/10
> # exponent satisfying condtn
> # for iid case, RWM ep = 1, MALA ep = 1/3, HMC = 1/4.
> ep = 1
> targmean = c(rep(1, dim))
> targSigma = diag(c(rep(1/dim, 5), rep(25, dim - 5)))
> logpi = function(X) {
+     dmvnorm(X, targmean, targSigma, log = TRUE)
+ }
> # initialise acceptances
> accr = rep(0, n)
> # initial efficiencies
> vlist = rep(0, n)
> for (j in 1:n) {
+     # set overdispersed starting distribution
+     X = runif(dim, -5, 5)
+     # initialize Markov chain values
+     x1list = rep(0, it)
+     # proposal variance vector as fn of scaling and dimension
+     propSigma = diag(ell[j]/dim^ep, dim)
+     for (i in 1:it) {
+         # proposal value
+         Y = X + rmvnorm(1, sigma = propSigma)
+         # for accept/reject U = runif(1)
+         # We will use logarithmic scale to avoid zeroes in
+         #   accept/reject step
+         if (log(U) < logpi(Y) - logpi(X)) {
+             X = Y
+             # accept proposal
+             accr[j] = accr[j] + 1
+         }
+         x1list[i] = X[1]
+     }
+     varfact = function(xxx) {
+         2 * sum(acf(xxx, plot = FALSE, lag.max = 250)$acf) -
+             1
+     }
+     vfact = varfact(x1list)
+     vlist[j] = vfact
+ }
> accr = accr/it
> vvlist = vlist^-1
> plot(accr, vvlist)
```

C.2 Code for Example 2.13

```
> # This function is based on the M. Bedard's code for Figure
> #   3 in Bed08b.
> GamS <- function(it = 5e+05, dim = 100, ep = 0, n = 50) {
+     # it is the number of iterations for each Monte Carlo sim
+     # ep is the smallest expnt stsfng fntnss of scaling trms
+     #   (condition 3.5)
+     # dim is the dimension of the space
+     # in this case, ep = 0, meaning sigma(d)=ell
+     # n is the number of scaling parameters used
+     # ell is vector for values of the squared scaling parameter
+     ell <- 10 + (1:n) * 3
+     # vsd is proposal vsd as fn of ell, dim and ep
+     vsd = ell/dim^ep
+     # a is scale parameter for gamma distribution
+     a <- 5
+     # lamb is the scaling vector
+     lamb <- c(rep(1/sqrt(dim)/5, dim - 2), 1, 1)
+     # ll is number of observations for target & proposal
+     ll <- n * dim
+     # now form the target distribution
+     x <- rgamma(ll, a, rep(lamb, n))
+     xp1 <- x[((0:(n - 1)) * dim) + 1]
+     sx1 <- 0
+     # accr is the vector of accptnce rates
+     accr <- rep(0, n)
+     for (i in 1:(it - 1)) {
+         # initialize the proposals
+         y <- rnorm(ll, mean = x, sd = rep(vsd^0.5, each = dim))
+         # set negative values to zero
+         y[y < 0] <- rep(0, ll)[y < 0]
+         # calculate number of zeroes
+         vv <- cumsum((y == 0) * 1)[(1:n) * dim]
+         # take first order lags
+         vv <- diff(c(0, vv))
+         y <- (rep(vv, each = dim) == 0) * y
+         yl <- y
+         yl[yl == 0] <- x[yl == 0]
+         sum1 <- cumsum(log(yl) - log(x))[(1:n) * dim]
+         sum1 <- diff(c(0, sum1))
+         sumxy <- cumsum(rep(lamb, n) * (yl - x))[(1:n) * dim]
+         sumxy <- diff(c(0, sumxy))
+         alphal <- pmin((a - 1) * sum1 - sumxy, 0) + (vv != 0) *
+             (-10^7)
+         rrl <- log(runif(n))
+         x[rep(rrl, each = dim) < rep(alphal, each = dim)] <- y[rep(rrl,
+             each = dim) < rep(alphal, each = dim)]
+         sx1 <- sx1 + (x[((0:(n - 1)) * dim) + 1] - xp1)^2
+         xp1 <- x[((0:(n - 1)) * dim) + 1]
+         accr <- accr + (rrl < alphal)
+     }
+     aacr <- accr/it
+     plot(accr, dim^ep * sx1/it)
+     # plot(ell, dim^ep*sx1/it)
+     result = list(ell, accr, dim^ep * sx1/it)
+     return(result)
+ }
```

Subsequently, one can scan through the returned vectors easily using the following code, which also yields theoretical curves for the same example:

```
> ScanGam = function(it = 5e+05, dim = 50, ep = 0, n = 50) {
+     ell = 10 + (1:n) * 3
+     ER = 1/75
+     vsd = ell/dim^ep
+     a = 5
+     # initialize the speed measure
+     esp = 0
+     for (i in (1:it)) {
+         x = rgamma(2 * n, a, 1)
+         y = rnorm(2 * n, x, rep(vsd^0.5, each = 2))
+         eps = matrix(log(dgamma(y, a, 1)/dgamma(x, a, 1)), 2,
+             n)
+         # calculate the ratio
+         eps = (eps[1, ] + eps[2, ] - ell * ER/2)/sqrt(ell * ER)
+         esp = esp + pnorm(eps)
+     }
+     # scan the vectors
+     ell[which.max(2 * ell * esp/it)]
+     OAR = 2 * esp[which.max(2 * ell * esp/it)]/it
+     plot(2 * esp/it, 2 * ell * esp/it, type = "l")
+     result = list(ell, 2 * esp/it, 2 * ell * esp/it, which.max(2 *
+         ell * esp/it), ellhat, OAR)
+     return(result)
+ }
```

Acknowledgement. Words cannot express the gratitude I have for my supervisor, Professor Jeffrey S. Rosenthal. He graciously lent me this time and introduced me to the exciting field of optimal scaling of Monte Carlo algorithms, providing positive encouragement throughout the process. Without his guidance, none of this would have been possible.

References

[Béd06] Bédard, M. (2006). *On the Robustness of Optimal Scaling for Random Walk Metropolis Algorithms.* Ph.D. dissertation, Department of Statistics, University of Toronto.

[Béd07] Bédard, M. (2007). Weak Convergence of Metropolis Algorithms for Non-*iid* Target Distributions. *Ann. Appl. Probab.* **17**, 1222-44.

[Béd08a] Bédard, M. (2008). Optimal Acceptance Rates for Metropolis Algorithms: Moving Beyond 0.234. *Stochastic Process. Appl.* **118**, 2198-222.

[Béd08b] Bédard, M. (2008). Efficient Sampling using Metropolis Algorithms: Applications of Optimal Scaling Results. *J. Comput. Graph. Statist.* **17**, 312-32.

[Béd10] Bédard, M. (2010). On the Optimal Scaling Problem of Metropolis Algorithms for Hierarchical Target Distributions, Unpublished.

[BR08] Bédard, M., Rosenthal, J.S. (2008). Optimal scaling of Metropolis algorithms: Heading toward general target distributions. *Canadian Journal of Statistics*, **36**(4), 483-503.

[BPR06] Beskos, A., Papaspiliopoulos, O. and Roberts, G.O. (2006). Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli*, **12**, 1077-1098.

[BPR08] Beskos, A., Papaspiliopoulos, O. and Roberts, G.O. (2008). A factorisation of diffusion measure and finite sample path constructions. *Methodoloogy and Computing in Applied Probability*, **10**, 85-104.

[BPR09] Beskos, A., Papaspiliopoulos, O. and Roberts, G.O. (2009). Monte Carlo maximum likelihood estimation for discretely observed diffusion porcesses. *The Annals of Statistics*, **37**, 223-245.

[BPRF06] Beskos, A., Papaspiliopoulos, O. Roberts, G.O., and Fearnhead, P. (2006) Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *Journal of Royal Statistical Society, Series B: Statistical Methodology*, **68**, 333-382.

[BRS09] Beskos, A., Roberts, G.O., and Stuart, A.M. (2009). Optimal scalings for local Metropolis-Hastings on non-product targets in high dimensions. *Annals of Applied Probability*, **19**(3), 863-898.

[BRSV08] Beskos, A., Roberts, G.O., Stuart, A.M., and Voss, J. (2008) MCMC methods for diffusion bridges. *Stochastic Dynamics*, **8**, 319-350.

[BS09] Beskos, A. and Stuart, A.M. (2009). MCMC methods for sampling function space. In *Invited Lectures, Sixth International Congres on Industrial and Applied Mathematics, ICIAM07, Editiors Rolf Jeltsch and Gerhard Wanner, European Mathematical Society*, 337-364.

[BR00] Breyer, L.A., and Roberts, G.O. (2000). From Metropolis to diffusions: Gibbs states and optimal scaling. *Stochastic Processes and their Applications*, **90**, 181-206.

[Da06] Da Prato, G. (2006). *An Introduction to Infinite-Dimensional Analysis.* Springer, New York.

[EK86] Ethier, S.N. and Kurtz, T.G. (1986). *Markov processes, characterization and convergence*, Wiley, New York.

[GRG96] Gelman, A., Roberts, G.O., and Gilks, W.R. (1996). Efficient Metropolis jumping rules. *Bayesian Statistics* **5**, 599-608.

[HSV07] Hairer, M., Stuart, A.M., and Voss, J. (2007). Analysis of SPDES Arising in Path Sampling. Part II: The Nonlinear Case. *Annals of Applied Probability*, **17**(5-6), 1657-1706.

[HSV10] Hairer, M., Stuart, A.M., and Voss, J. (2010). Signal Processing Problems on Function Space: Bayesian Formulation, Stochastic PDEs and Effective MCMC Methods. *The Oxford Handbook of Nonlinear Filtering*, D. Crisan and B. Rozovsky (editors).

[HSVW05] Hairer, M., Stuart, A.M., Voss, J., and Wiberg, P. (2005). Analysis of SPDES Arising in Path Sampling. Part I: The Gaussian Case. *Communications in Mathematical Sciences*, **3**, 587-603.

[JS03] Jean, J. and A.N. Shiryaev. (2003). *Limit Theorems for Stochastic Processes, 2nd ed.* Springer-Verlag, New York.

[Kal97] Kallenberg, O. (1997). *Foundations of Modern Probability.* Springer-Verlag, New York.

[KP99] Kloeden, P.E., and Platen, E. (1999). *Numerical Solution of Stochastic Differential Equations*, Applied Mathematics, **23**, Third corrected printing, Springer, New York.

[MPS11] Mattingly, J.C., Pillai, N.S., and Stuart, A.M. (2011). Diffusion Limits of Random Walk Metropolis Algorithm in High Dimensions. *Annals of Applied Probability*, to appear.

[MRR+53] Metropolis, N, Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6): 1087.

[MT09] Meyn, S., and Tweedie., R.L. (2009). *Markov Chains and Stochastic Stability, Second Edition.* Cambridge University Press, NY.

[NR05] Neal, P., and Roberts, G.O. (2006). Optimal Scaling for Partially Updating MCMC Algorithms*. Annals of Applied Probability.* **16(2)**, 475-515.

[NRY05] Neal, P., Roberts, G.O., and Yuen, W.K. (2007). Optimal Scaling of Random Walk Metropolis Algorithms with Discontinuous Target Densities. *Under revision for Annals of Applied Probability.*

[PST11] Pillai, N.S., Stuart A.M., and Thiéry (2011). Optimal Scaling and Diffusion Limits for the Langevin Algorithm in High Dimensions. *Submitted to the Annals of Statistics.*

[PG10] Pasarica, C. and Gelman, A. (2010). Adaptively Scaling the Metropolis Algorithm Using Expected Squared Jumping Distance. *Statistica Sinica*, **20**, 343-364.

[RC04] Robert, C. and Casella, G. (2004) *Monte Carlo Statistical Methods, Second Edition.* Springer, NY.

[RGG97] Roberts, G.O., Gelman, A., Gilks, W.R. (1997). Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. *Ann. Appl. Probab.* **7**, 110-20.

[RR98] Roberts, G.O., Rosenthal, J.S. (1998). Optimal Scaling of Discrete Approximations to Langevin Diffusions. *J.R. Stat. Soc. Ser. B Stat. Methodol.* **60**, 255-268.

[RR01] Roberts, G.O., Rosenthal, J.S. (2001). Optimal Scaling for various Metropolis-Hastings algorithms. *Statis. Sci.* **16**, 351-67.

[RR04] Roberts, G.O., Rosenthal, J.S. (2004). General State Space Markov Chains and MCMC Algorithms. *Probab. Surveys* **1**, 20-71.

[RT96] Roberts, G.O. and Tweedie, R.L. (1996). Exponential Convergence of Langevin Distributions and Their Discrete Approximations. *Bernoull*, **2(4)**, 341-363.

[Ros08a] Rosenthal, J.S. (2008). Optimising Monte Carlo Search Strategies for Automated Pattern Detection.

[Ros08b] Rosenthal, J.S. (2008). Optimal Proposal Distributions and Adaptive MCMC. Prepared for MCMC Handbook.

[Ros10] Rosenthal, J.S. (2010). Optimal Proposal Distributions and Adaptive MCMC. *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones, and X.-L. Meng (editors), Chapman & Hall / CRC Press.

[SS05] Shakarchi, R. and Stein, E.M. (2005). *Real Analysis - Measure Theory, Integration and Hilbert Spaces*, Princeton University Press, Princeton.

[SK11] Shalizi, C.R. with Aryeh Kontorovich. (2011) *Almost None of the Theory of Stochastic Processes*, Unpublished Manuscript. Available online http://www.stat.cmu.edu/~cshalizi/almost-none/v0.1.1/almost-none.pdf

[Stro11] Stroock, D. (2011). *Probability Theory - An Analytic View, Second Edition.* Cambridge University Press, New York.

[Tie94] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics.* **22**, 1701-1762.