A Stable Manifold MCMC Method for High Dimensions

Alexandros Beskos

Dept of Statistics and Applied Probability, National University of Singapore

Abstract

We combine two important recent advancements of MCMC algorithms: first, methods utilizing the intrinsic manifold structure of the parameter space; then, algorithms effective for targets in infinite-dimensions with the critical property that their mixing time is robust to mesh refinement.

Keywords: Manifold MCMC, Metropolis-adjusted Langevin algorithm, Cameron-Martin space, infinite dimensions.

1. Introduction

Manifold MCMC methods were introduced in Girolami and Calderhead (2011) and were shown to be effective for challenging target distributions with complex local-correlation structures. Furthermore, MCMC methods robust in high dimensions have been recently developed (see e.g. Beskos et al. (2008)) for important statistical models giving rise to targets defined as change of measures from Gaussian laws in infinite dimensions. This work aims to develop new MCMC algorithms with the objective of joining strengths from the two directions of recent methodological progress. The new methodology will be illustrated on a model structure involving a diffusion process observed with (small) error. This simple example has been selected to clearly showcase the effect of the new method, as the infinite-dimensional aspect of the MCMC algorithm will deal with the high-dimensionality of the diffusion path, while its manifold aspect will provide a principled mechanism for driving proposed diffusion paths close to the data.

We will focus on the Manifold Metropolis-adjusted Langevin algorithm (MMALA) of Girolami and Calderhead (2011). On the infinite-dimensional side, the methods in, e.g., Beskos et al. (2008); Cotter et al. (2013) are

relevant for targets Π defined as change of measures from Gaussian laws, i.e.:

$$\frac{d\Pi}{d\tilde{\Pi}}(x) = \exp\{-\Phi(x)\} \ , \quad x \in \mathcal{H} \ , \tag{1}$$

where $\mathcal{H}$ is a separable Hilbert space equipped with the inner product $\langle \cdot, \cdot \rangle$, $\tilde{\Pi} = \mathcal{N}(\mu, \mathcal{C})$ a Gaussian distribution on $\mathcal{H}$ of mean μ and covariance $\mathcal{C}$, and $\Phi : \mathcal{H} \mapsto \mathbb{R}$. See e.g. Da Prato and Zabczyk (1992) for a treatment of Gaussian laws on general Hilbert spaces. The approaches in Beskos et al. (2008) are effective in the case when $\mathcal{H}$ is infinite-dimensional since - upon finite-dimensional projection and selection of some relevant mesh size - they provide algorithms of *mesh-free* mixing time, separating themselves from standard MCMC algorithms for which mixing time deteriorates with increasing dimension (see e.g. Roberts and Rosenthal (2001)). The contribution of this paper will be to combine the manifold and infinite-dimensional methods, thus deriving new algorithms that could unify the positive computational effects of these two approaches. The main algorithm developed in this paper will be assigned the label ∞-MMALA (the label ∞-MALA will be used for the infinite-dimensional MALA of Beskos et al. (2008); recall that MMALA refers to the manifold method).

The structure of the paper is as folllows. In Section 2 we develop ∞-MMALA and state conditions under which it is well-defined in infinite dimensions. In Section 3 we show numerical results from applying ∞-MMALA on a diffusion-driven model. In Section 4 we discuss further directions.

2. ∞-MMALA: Manifold MALA in Infinite-Dimensions

We focus mainly on the practicalities of the derivation of the algorithm and avoid technicalities. Therefore, we will not discuss here the notion of differentiation (denoted by ∇) in general Hilbert spaces or the well-posedness of the Langevin stochastic differential equation (SDE) below or its manifold version on arbitrary Hilbert spaces. The reader could simply assume that the development of the algorithm happens on some N-dimensional projection of the infinite-dimensional target Π, for some large $N \geq 1$, so that the state-space is the Euclidean $\mathbb{R}^N$. Mathematical rigor will be applied when defining the final algorithm (involving the easier to handle time-discretised SDE dynamics). So, Π is used interchangeably below to denote both the infinite-dimensional target and the N-dimensional projection; a similar convention

is applied for other related notions, e.g. for $\tilde{\Pi}$. Also, from the definition of Π in (1) we have (in a formal sense, in the case of general Hilbert spaces):

$$\Pi(x) \propto \exp\{\ell(x)\} = \exp\left\{ -\Phi(x) - \tfrac{1}{2}\langle x-\mu, L(x-\mu)\rangle \right\} ,$$

where we have set $L = \mathcal{C}^{-1}$.

2.1. MMALA and ∞-MALA Algorithms

MMALA utilizes the dynamics of the Langevin SDE on the manifold space generated by a chosen metric tensor $G(x)$. Its expression is as follows:

$$dx = \tfrac{1}{2}\,\tilde{\nabla}\ell(x)dt + d\tilde{b} \,, \tag{2}$$

with $\tilde{\nabla} = G^{-1}(x)\nabla$ corresponding to differentiation along the manifold and $d\tilde{b}$ denoting infinitesimal increments of a Brownian motion on the manifold space. In agreement with the comments above, for all practical purposes one can assume that the state-space is $x \in \mathbb{R}^N$, so that $G^{-1}(x)$ is assumed to be a symmetric positive-definite matrix in $\mathbb{R}^{N\times N}$. Using the analytical expressions from Girolami and Calderhead (2011) we can equivalently re-express (2) in terms of the following more familiar SDE on Euclidean space:

$$dx = G(x)^{-1}\left\{\tfrac{1}{2}\,\nabla\ell(x) + \tfrac{1}{2}\nabla\log|G(x)| + \nabla\right\}dt + G(x)^{-1/2}\,db \,, \tag{3}$$

with db now denoting increments of standard Brownian motion and $|G(x)|$ is the determinant of $G(x)$. The Langevin SDE (3) will now be time-discretised to provide a proposal in a Metropolis-Hastings framework. The two algorithms, MMALA and ∞-MALA are now specified as follows:

- MMALA: it applies the standard Euler finite-difference scheme to time-discretise the dynamics (3). For current position x, this will provide a proposed transition to, say, x', accepted or rejected according to the related Metropolis-Hastings ratio. This algorithm will typically not be well-defined in infinite-dimensions.

- ∞-MALA: in this case $G(x) \equiv L$, so the drift function in (3) becomes:

$$G^{-1}(x)\,\tfrac{1}{2}\nabla\ell(x) \equiv \tfrac{1}{2}\left\{ -\mathcal{C}\,\nabla\Phi(x) - (x-\mu)\right\} .$$

  ∞-MALA employs a semi-implicit discretisation scheme, where the linear term x in the drift is replaced by with $\frac{x'+x}{2}$ when applying finite

differences. Solving w.r.t. x' delivers a proposal of positive acceptance probability even in infinite dimensions (under conditions). This is because due to the semi-implicit scheme, and under weak conditions on Φ, $\tilde{\Pi}$, the distributions of (x, x') and (x', x) are absolutely continuous w.r.t. each other, thus allowing for a non-zero Metropolis-Hastings ratio. We will discuss this also in the sequel, as ∞-MMALA will require some of the tools used for the development ∞-MALA.

2.2. Time-Discretisation of Langevin Dynamics for ∞-MMALA

We will in fact opt for a simplified version of the dynamics in (3), with the objective of combining designated moves with computational efficiency. As already noted in Girolami and Calderhead (2011) most of the strength of the manifold method is captured by the dynamics that only involve the term $G^{-1}(x)\,\frac{1}{2}\,\ell(x)$ in the drift function. Calculation of the removed Christofell symbols is typically expensive (of the order of $\mathcal{O}(N^3)$) and could eradicate in the balance their effect on improved mixing.

Guided by the semi-implicit idea behind ∞-MALA, we introduce a time-discretisation scheme which possesses the critical property of giving rise to a well-posed algorithm in infinite-dimensions. The scheme is as follows (we add/subtract $G(x)x$ in the drift and apply an implicit scheme in one term):

$$x' - x = \tfrac{1}{2}\,G(x)^{-1}\left\{ -G(x)\,\tfrac{x'+x}{2} + G(x)\,x + \nabla\ell(x)\right\} h + \sqrt{h}\,\mathcal{N}(0, G(x)^{-1})\ , \tag{4}$$

for a step-size $h > 0$, which can equivalently be written as:

$$x' = \tfrac{1-h/4}{1+h/4}\,x + \tfrac{h/2}{1+h/4}\,S(x) + \tfrac{\sqrt{h}}{1+h/4}\,\mathcal{N}(0, G(x)^{-1})\ , \tag{5}$$

where we have defined:

$$S(x) = -G(x)^{-1}\{\,\nabla\Phi(x) - (G(x) - L)x - L\mu\,\}\ .$$

Notice that the choice $G(x) = L$ will deliver ∞-MALA. Equation (5) provides the proposal for ∞-MMALA upon which we will apply the Metropolis-Hasting acceptance rule. In Section 2.5 below we give the expression for the acceptance probability of proposal (5) on the infinite-dimensional space $\mathcal{H}$.

2.3. Choice of Metric Tensor $G(x)$

Following Girolami and Calderhead (2011), an often effective approach is to choose the metric tensor as the expected Fisher information (we write $\Phi(x) = \Phi(x; y)$ to emphasize dependence of Φ on some data $Y = y$):

$$\begin{aligned} -\mathrm{E}_{Y|x}\nabla^2\ell(x) &= \mathrm{E}_{Y|x}\nabla_x^2\Phi(x;Y) + L \\ &= \mathrm{E}_{Y|x}\big[\,\nabla_x\Phi(x;Y)\,\{\nabla_x\Phi(x;Y)\}^\top\,\big] + L\ . \end{aligned} \tag{6}$$

Used in the context of high-dimensional $x \in \mathbb{R}^N$, this can sometimes lead to prohibitive computations as an order of N. Thus, for a given class of target distributions one could try to balance improved mixing due to a good choice of $G(x)$ with computational considerations, and maybe opt for a convenient proxy of the expected Fisher information (or the observed Fisher information, or other tensor understood to be appropriate in a given scenario).

2.4. Diversion on Metropolis-Hastings in General State-Spaces

Our objective is to show that the acceptance ratio is non-trivial when working on an infinite-dimensional Hilbert space $\mathcal{H}$. We denote by $Q(x, dx')$ the transition probability measure corresponding to ∞-MMALA proposal (5). As shown in Beskos et al. (2008) for ∞-MALA, a deviation from standard proofs of invariance for Metropolis-Hastings kernels is that there is typically no common dominating measure for the probability measures $Q(x, dx')$ over all $x \in \mathcal{H}$. So, one has to resort to a generalised definition of the Metropolis-Hastings ratio in Tierney (1998). Following Tierney (1998), one has to seek for conditions so that - for $x \sim \Pi(dx)$, in stationarity - the laws of (x, x') and (x', x) are absolutely continuous w.r.t. each other (we use the symbol ' $\simeq'$ to denote such a relation between probability laws); then, their Radon-Nikodym derivative provides the Metropolis-Hastings acceptance ratio.

More analytically, we define the bivariate probability measure on $\mathcal{H} \times \mathcal{H}$:

$$\mu(dx, dx') = \Pi(dx)Q(x, dx')\ .$$

and the corresponding symmetric measure $\mu^\top(dx, dx') := \mu(dx', dx)$. Following Tierney (1998), if $\mu \simeq \mu^\top$ then the Metropolis-Hastings acceptance probability is non-trivial and equal to:

$$1 \wedge \frac{d\mu^\top}{d\mu}(x, x')\ . \tag{7}$$

Thus, we will specify conditions under which $\mu \simeq \mu^\top$ and find $(d\mu^\top/d\mu)(x, x')$.

2.5. Proof of Well-Posedness of ∞-MMALA

The derivation below has connections with the one in Beskos et al. (2008) for ∞-MALA. To demonstrate well-posedness of ∞-MMALA in infinite dimensions we need some assumptions.

Assumption 1. *$\tilde{\Pi}$-a.s. in x, we have $\mathcal{N}(0, G(x)^{-1}) \simeq \mathcal{N}(0, L^{-1})$, with:*

$$\kappa(v; x) := \frac{d\mathcal{N}(0, G(x)^{-1})}{d\mathcal{N}(0, L^{-1})}(v) \ , \quad x, v \in \mathcal{H} \ .$$

Assumption 2. *$\tilde{\Pi}$-a.s. in x, quantity $S(x)$ is an element of the Cameron-Martin space of $\mathcal{N}(0, \mathcal{C})$, that is $S(x) \in \operatorname{Im} \mathcal{C}^{1/2}$.*

$\operatorname{Im} \mathcal{C}^{1/2}$ denotes the image space of $\mathcal{C}^{1/2}$. The Cameron-Martin space of $\mathcal{N}(0, \mathcal{C})$ is comprised of all elements of $\mathcal{H}$ that preserve absolute continuity of $\mathcal{N}(0, \mathcal{C})$ when translating it. In particular, we have the following result.

Proposition 1. *Consider $\mathcal{N}(0, \Sigma)$ on $\mathcal{H}$. If $R(x) = x + \Sigma^{1/2} x_0$ for a constant $x_0 \in \mathcal{H}$, then $\mathcal{N}(0, \Sigma) \simeq N(0, \Sigma) \circ R^{-1}$ with density:*

$$\frac{d\mathcal{N}(0, \Sigma) \circ R^{-1}}{d\mathcal{N}(0, \Sigma)}(x) = \exp \left\{ \langle x_0, \Sigma^{-1/2} x \rangle - \tfrac{1}{2} |x_0|^2 \right\} \ .$$

This is Theorem 2.21 of Da Prato and Zabczyk (1992). Notice that we can rewrite the proposal (5) as:

$$Q(x, dx') \ : \quad x' = \tfrac{1 - h/4}{1 + h/4} \, x + \tfrac{\sqrt{h}}{1 + h/4} \, \mathcal{N}(\tfrac{\sqrt{h}}{2} S(x), G(x)^{-1}) \ . \tag{8}$$

Let $\tilde{Q}(x, dx')$ denote the transition probability law for the update:

$$\tilde{Q}(x, dx') \ : \quad x' = \tfrac{1 - h/4}{1 + h/4} \, x + \tfrac{\sqrt{h}}{1 + h/4} \, \mathcal{N}(0, L^{-1}) \ . \tag{9}$$

We define the reference bivariate Gaussian measure:

$$\tilde{\mu}(dx, dx') = \tilde{\Pi}(dx) \tilde{Q}(x, dx') \ .$$

It is easy to check that $\tilde{\mu}(dx, dx')$ is symmetric (see Beskos et al. (2008) for details; this is because the sum of the squares of the scalars in front of x and $\mathcal{N}(0, L^{-1})$ in (9) is unit), so that $\tilde{\mu}(dx, dx') = \tilde{\mu}^{\top}(dx, dx') := \tilde{\mu}(dx', dx)$.

Proposition 2. *Under Assumptions 1 and 2 above, $\tilde{\Pi}$-a.s. in x we have that $\mathcal{N}(\frac{\sqrt{h}}{2}S(x), G^{-1}(x)) \simeq \mathcal{N}(0, L^{-1})$ with density:*

$$\lambda(v;x) = \frac{d\mathcal{N}(\frac{\sqrt{h}}{2}S(x), G^{-1}(x))}{d\mathcal{N}(0,L^{-1})}(v) =$$

$$\exp\left\{ \tfrac{\sqrt{h}}{2} \left\langle G^{1/2}(x)S(x), G(x)^{1/2}\, v \right\rangle - \tfrac{h}{8} \left| G(x)^{1/2}S(x) \right|^2 \right\} \times \kappa(v;x) \ .$$

Proof. We use the chain rule:

$$\tfrac{d\mathcal{N}(\frac{\sqrt{h}}{2}S(x), G^{-1}(x))}{d\mathcal{N}(0,L^{-1})}(v) = \tfrac{d\mathcal{N}(\frac{\sqrt{h}}{2}S(x), G^{-1}(x))}{d\mathcal{N}(0,G(x)^{-1})}(v) \times \tfrac{d\mathcal{N}(0,G(x)^{-1})}{d\mathcal{N}(0,L^{-1})}(v) \ . \tag{10}$$

The last density is found via Assumption 1. For the other, we use the fact that Assumption 1 implies that operators L^{-1}, $G(x)^{-1}$ have the same Cameron-Martin space (see Feldman-Hajek theorem in Da Prato and Zabczyk (1992)), so that applying Proposition 1 for $\Sigma \equiv G(x)^{-1/2}$, $x_0 \equiv (\sqrt{h}/2)\, G(x)^{1/2}S(x)$ (guaranteed to be a proper element of $\mathcal{H}$ due to having $S(x) \in \operatorname{Im} G(x)^{-1/2}$) will give that the first density on the RHS of (10) is equal to:

$$\exp\left\{ \tfrac{\sqrt{h}}{2} \left\langle G^{1/2}(x)S(x), G(x)^{1/2}\, v \right\rangle - \tfrac{h}{8} \left| G(x)^{1/2}S(x) \right|^2 \right\} .$$

The proof is now complete. □

From Proposition 2, eqs (8), (9) imply directly that $\tilde{\Pi}$-a.s. in x we have $Q(x, dx') \simeq \tilde{Q}(x, dx')$, thus also $\mu(dx, dx') \simeq \tilde{\mu}(dx, dx')$. This essentially completes the well-posedness of ∞-MMALA as - due to the symmetricity of $\tilde{\mu}(dx, dx')$ - it implies that $\mu(dx, dx') \simeq \mu(dx, dx')$. Indeed, we can find the density required in the acceptance probability (7) as follows. First, we find:

$$\frac{d\mu}{d\tilde{\mu}}(x, x') = \frac{d\Pi}{d\tilde{\Pi}}(x)\frac{dQ}{d\tilde{Q}}(x, x') = \exp\{-\Phi(x)\}\, \lambda(\rho^{-1}(x';x);x) \ ,$$

where we denote by $\rho^{-1}(\cdot;x)$ the inverse of the 1-1 mapping:

$$v \mapsto \rho(v;x) = \tfrac{1-h/4}{1+h/4}\, x + \tfrac{\sqrt{h}}{1+h/4}\, v \ .$$

From the definition of symmetric measures, we have that $(d\mu^\top/d\tilde{\mu}^\top)(x, x') = (d\mu/d\tilde{\mu})(x', x)$, thus we finally have that:

$$\frac{d\mu^\top}{d\mu}(x, x') = \frac{(d\mu/d\tilde{\mu})(x', x)}{(d\mu/d\tilde{\mu})(x, x')} = \frac{\exp\{-\Phi(x')\}\, \lambda(\rho^{-1}(x;x');x')}{\exp\{-\Phi(x)\}\, \lambda(\rho^{-1}(x';x);x)} \ . \tag{11}$$

∞-*MMALA:*

(i) Start with an initial value $x^{(0)}$ *from* $\mathcal{N}(\mu, L^{-1})$*, or another Gaussian law absolutely continuous w.r.t. the target* Π *and set* $k = 0$*.*

(ii) Given current $x = x^{(k)}$*, propose the transition:*

$$x' = \tfrac{1-h/4}{1+h/4}\, x + \tfrac{h/2}{1+h/4}\, S(x) + \tfrac{\sqrt{h}}{1+h/4}\, \mathcal{N}(0, G(x)^{-1}) \ ,$$

Set $x^{(k+1)} = x'$ *with probability* $1 \wedge (d\mu^\top/d\mu)(x, x')$ *for* $(d\mu^\top/d\mu)(x, x')$ *as specified in (11), otherwise set* $x^{(k+1)} = x$*.*

(iii) Set $k \to k + 1$ *and go to (ii).*

Table 1: Definition of ∞-MMALA.

Thus, we have proven that $\mu(dx, dx') \simeq \mu^\top(dx, dx')$ with the above density. The complete ∞-MMALA algorithm can now be summarized in Table 1.

Remark 1. *We note that earlier works (e.g. the recent Cotter et al. (2013)) have looked at* ∞*-MALA for a* constant *metric tensor* $G(x)$*; Cotter et al. (2013) use* ∞*-MALA corresponding to the choice* $G(x) = L$ *for the algorithm described here. The extension to a non-constant metric tensor is non-trivial and involved: i) the development of the discretisation scheme in (4) which is not an apparent generalisation of the scheme for a constant metric tensor of the earlier works; ii) the analytical justification of the well-posedness in infinite-dimensions of the new algorithm.*

3. Illustrative Application

We consider an SDE observed with small error. That is, we have:

$$dx_t = a(x_t)dt + dw_t \ , \quad x_0 = x^* \in \mathbb{R} \ , \tag{12}$$

where w denotes standard Brownian motion on $\mathbb{R}$, with data points:

$$y_i = f(x_{t_i}) + \mathcal{N}(0, \sigma^2) \ , \quad 1 \le i \le n \ ,$$

for times $t_1 < \cdots < t_n = T$, a drift $a : \mathbb{R} \mapsto \mathbb{R}$ and a mapping $f : \mathbb{R} \mapsto \mathbb{R}$. The target here is the posterior law $\Pi(dx)$ of the path $x = \{x_t;\, t \in [0, T]\}$

given $y = \{y_1, \ldots, y_n\}$, which is of the general form in (1) with:

$$\Phi(x) = \Phi_\sigma(x) - \Phi_b(x) \; ; \quad \Phi_\sigma(x) := \sum_{i=1}^{n} \frac{(y_i - f(x_{t_i}))^2}{2\sigma^2} \; ;$$
$$\Phi_a(x) := -\int_0^T a(x_s)dx_s + \tfrac{1}{2}\int_0^T a^2(x_s)ds \; . \tag{13}$$

$\tilde{\Pi}$ here is the law of a Brownian motion on $[0, T]$ started at x^*. Also, here $\mathcal{H} = L^2([0, T], \mathbb{R})$, i.e. the space of squared-integrable paths, equipped with the corresponding inner product. There are two main challenges for MCMC algorithms attempting to sample from Π.

(a) High-Dimensionality of state space. In theory, the state space is infinite dimensional. In practice, one will typically select a large $N \geq 1$ and apply finite-differences to obtain a vector $(x_1, x_2, \ldots, x_N) \in \mathbb{R}^N$ corresponding to times $\delta, 2\delta, \ldots, N\delta$ for mesh-size $\delta = T/N$. ∞-MALA will be stable under mesh-refinement: the computational cost per step will increase as $\mathcal{O}(N)$, but the mixing time will be $\mathcal{O}(1)$. MMALA of Girolami and Calderhead (2011) will deteriorate with decreasing mesh-size δ. The work e.g. in Roberts and Rosenthal (2001) suggests that one has to choose $h = \mathcal{O}(N^{-1/3})$ to control the acceptance probability, thus giving a mixing time of $\mathcal{O}(N^{1/3})$.

(b) Complex a-posteriori covariance structure. ∞-MALA will deteriorate for decreasing $\sigma > 0$ as target Π gets distanced from $\tilde{\Pi}$, but the proposal generation mechanism still uses a covariance matrix $G^{-1}(x) \equiv \mathcal{C}$ that does not adjust to the complex covariance structure of the posterior characterised by small marginal variances at the times of the data and larger ones further from those. In contrast to MALA, MMALA accommodates for the complex a-posteriori covariance structure of the x-path. The newly developed ∞-MMALA turns out to be robust both in increasing N and decreasing σ.

3.1. Algorithmic Specification and Numerical Results

As we are interested mainly in the effect of σ at the properties of the algorithm, we will obtain the metric tensor $G(x)$ by applying the expected information idea in (6) only upon Φ_σ in (13) - this also guarantees positive-definiteness for the induced $G(x)$. Thus, we have:

$$G(x) = \mathrm{diag}_N\Big\{ \sum_{i=1}^{n} \mathbb{I}_{t_i = j\delta} \, \tfrac{\{f'(x_j)\}^2}{\sigma^2} \, , \; 1 \leq j \leq N \Big\} + L \; ,$$

for the $N \times N$ tridiagonal covariance matrix of the Brownian motion:

$$L = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots \\ -1 & 2 & -1 & 0 & \cdots \\ 0 & -1 & 2 & -1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix} /\delta \ .$$

Critically for the computational properties of ∞-MMALA, due to $G(x)$ being tri-diagonal the cost per step is $\mathcal{O}(N)$ (the same holds for MMALA).

Remark 2. *Regarding Assumptions 1, 2 here, note that $\mathcal{N}(0, G(x)^{-1})$ corresponds to a Brownian motion $\mathcal{N}(0, L^{-1})$ given finite observations with error, thus clearly $\mathcal{N}(0, G(x)^{-1}) \simeq \mathcal{N}(0, L^{-1})$. Then, $\mathrm{Im}\, \mathcal{C}^{1/2}$ consists of paths with $x_0 = 0$ whose 1^{st} derivative (in a weak sense) is in $L^2([0,T], \mathbb{R})$, see e.g. Beskos et al. (2013). We do not present a proof here that $S(x) \in \mathrm{Im}\, \mathcal{C}^{1/2}$, but we mention that in our runs the paths $S(x)$ over the various x's appear to be everywhere differentiable apart from the instances of the data where they are only continuous, thus everywhere differentiable in the weak sense.*

We applied ∞-MMALA in the following scenario:

$$a(x) = 4 - x \ , \ \ x^* = 2 \ , \ \ t_i = i, \ \ n = 100 \ , \ \ f(x) = x^{3/2} \ , \ \ \sigma^2 = 0.1 \ .$$

We used the standard Euler scheme to discretise x with mesh-size $\delta = 10^{-2}$, so that $N = n\delta^{-1} = 10^4$. The algorithm was initiated at a path with $x_{t_i} = 2$, for $1 \leq i \leq n$, with Brownian bridges connecting these points; this position is very far from the center of the target (notice that the $n = 100$ 'true' x_{t_i}'s will be scattered around 4 due to the choice of drift, and $\sigma^2 = 0.1$).

We used step-size $h = 1.0$, giving average acceptance probability of 82% (this changed to 80% when trying $\delta' = \delta/2$, in an empirical manifestation of the mesh-free mixing time of ∞-MMALA). Fig.1 shows two traceplots for ∞-MMALA over $2,000$ iterations, corresponding to the position of the path x at times $t = 37$, $t = 36.5$. Notice that the y-axes are on the same scale, so the algorithm adjusts automatically to the different sizes of the marginal posteriors (at $t = 37$ there is an observation, thus a lot of information about x_t). Even if started far from stationarity, the algorithm converges almost instantaneously to the target distribution.

For comparison, we applied ∞-MALA, MMALA in the same setting. ∞-MALA was extremely poor, as we had to use $h = 10^{-5}$ to get acceptance

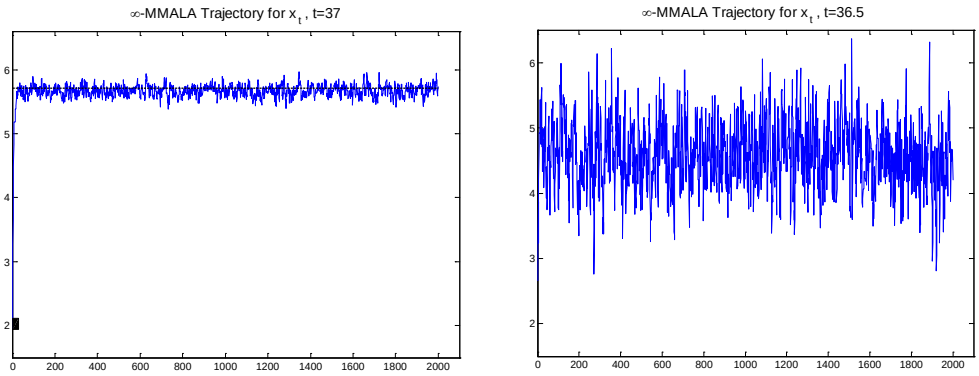

Figure 1: Traceplots over 2,000 ∞-MMALA steps. The left panel corresponds to x_t for $t = 37$, the right panel to $t = 36.5$. The horizontal line on the left panel indicates the value of $f^{-1}(y_{37})$; the black rectangle at position 2 of the y-axis highlights the initial position.

ratio of 53%, as the algorithm does not adjust to the different (a-posteriori) scales in the target, and needed a very small h to stay close to the data. For MMALA, we had to use a step-size of $h = 0.1$ to get average acceptance probability of 69% (reduced to 55% when trying $\delta' = \delta/2$), thus it is much less efficient that ∞-MMALA (execution times for both MMALA and ∞-MMALA were about 40s using Matlab on a standard PC). Fig.2 highlights a consequence of the fundamental structural difference between ∞-MMALA and MMALA: we took an initial path pinned at the data, so $x_{t_i} = f^{-1}(y_i)$ for $1 \leq i \leq n$ with Brownian bridges in-between, and run 1,000 iterations of ∞-MMALA and MMALA with $h = 1.0$. The plots in Fig.2 show the estimated Quadratic Variation (QVe) of all 1,000 *proposed* paths for both algorithms. Recall here that $QVe = \sum_{j=1}^{N}(x_j - x_{j-1})^2$, and this quantity will converge to $T = 100$ as $\delta \to 0$. As ∞-MMALA is well-defined in infinite-dimensions, the estimated QV of the path is very close to the limiting one for $N = \infty$ (the acceptance ratio was 81%). In contrast, MMALA gave QVe's wide off the mark; not surprisingly, all 1,000 proposed paths were rejected.

4. Conclusions and Further Directions

We presented a first attempt at merging in a principled manner recently developed manifold and infinite-dimensional MCMC algorithms. A simple SDE-model served well as an example where the new method indeed combines the benefits of the two directions. We aim to further develop this line of research and clarify the potential of new algorithms in important classes of applications. Some further investigations are summarised below.

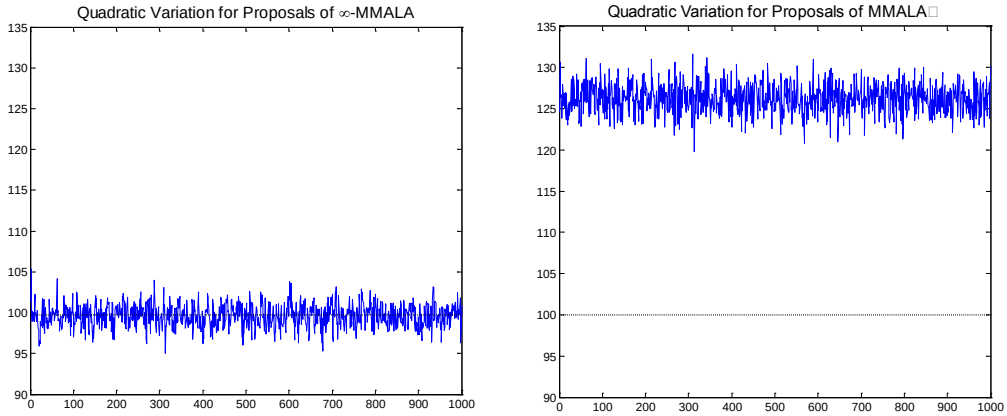

Figure 2: Estimated QV of the proposed x's over 1,000 iterations of ∞-MMALA (left) and MMALA (right) both using step-size $h = 1.0$. The horizontal lines highlight the limiting QV (equal to $T = 100$) in infinite dimensions. MMALA proposes x's with very wrong QV - so all of them got rejected; ∞-MMALA is well-defined in infinite-dimensions, thus the estimated QV of the proposed x's is very close to the limiting one (acceptance rate 81%).

Algorithmic Development: We aim to develop a Hybrid Monte-Carlo (HMC) version of ∞-MMALA. Also, other approaches for merging infinite-dimensional algorithms with manifold ones could be more appropriate in particular settings. In data assimilation, recent works have set-up a Bayesian framework in infinite-dimensions (see e.g. Stuart (2010) and the references therein) where information for important parameters of interest, such as the permeability field for sub-surface flow models, or the initial condition for fluid velocity dynamics, is expressed in the form of a posterior distribution defined as a change of measure from a Gaussian prior. ∞-MALA (or a random-walk version of it) has turned out to be overly costly in this context (see e.g. Stuart (2010)) as the posterior could be characterised by far more complex correlation structure than the prior. It seems very natural to develop an ∞-MMALA in this context, by applying MMALA at the low frequency components of the unknown parameters that are mostly informed by the data, and ∞-MALA for the high-frequency ones that are mainly determined by the Gaussian prior. A similar algorithmic construction could give critical computational advantages in the class of models of SDEs driven by fractional Brownian motion (fBM), where recent attempts (see e.g. Dureau et al.

(2013) and the references therein) to apply MCMC have to deal with the existence of complex correlation structures among model parameters together with a high-dimensional latent driving fBM. Such ideas can be relevant also for Bayesian non-parametric density estimation, e.g. the logistic Gaussian process prior (see e.g. Tokdar and Ghosh (2007); Cotter et al. (2013)).

Weakening Assumptions: Assumptions 1, 2 seem stronger than needed for the well-posedness of ∞-MMALA. In our example model for instance we indeed have $\mathcal{N}(0, G(x)^{-1}) \simeq \mathcal{N}(0, L^{-1})$ for any $\sigma > 0$, but not in the limit when $\sigma = 0$, this maybe suggesting (falsely, from our experiments) that the algorithm may deteriorate as $\sigma \to 0$. The resolution is that our assumptions and proof of well-posedness could involve some more 'appropriate' Gaussian measure instead of $\mathcal{N}(0, L^{-1})$, and in particular one for which its density w.r.t. $\mathcal{N}(0, G(x)^{-1})$ will not be trivial as $\sigma \to 0$. Such considerations are relevant beyond our example model.

Acknowledgments

I thank the anonymous referee and the Associate Editor for useful suggestions that have improved the content of the paper.

References

Beskos, A., Kalogeropoulos, K., Pazos, E., 2013. Advanced MCMC methods for sampling on diffusion pathspace. Stoch. Proc. Appl. 123 (4), 1415–1453.

Beskos, A., Roberts, G., Stuart, A., Voss, J., 2008. MCMC methods for diffusion bridges. Stoch. Dyn. 8 (3), 319–350.

Cotter, S., Roberts, G., Stuart, A., White, D., 2013. MCMC methods for functions: Modifying old algorithms to make them faster. Stat. Sci. 28 (3), 424–446.

Da Prato, G., Zabczyk, J., 1992. Stochastic equations in infinite dimensions. Vol. 44 of Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge.

Dureau, J., Beskos, A., Kalogeropoulos, K., 2013. Bayesian inference for partially observed SDEs driven by fractional Brownian motion, submitted.

Girolami, M., Calderhead, B., 2011. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (2), 123–214, with discussion and a reply by the authors.

Roberts, G., Rosenthal, J., 2001. Optimal scaling for various Metropolis-Hastings algorithms. Statist. Sci. 16 (4), 351–367.

Stuart, A., 2010. Inverse problems: a Bayesian perspective. Acta Numer. 19, 451–559.

Tierney, L., 1998. A note on Metropolis-Hastings kernels for general state spaces. Ann. Appl. Probab. 8 (1), 1–9.

Tokdar, S., Ghosh, J., 2007. Posterior consistency of logistic Gaussian process priors in density estimation. J. Statist. Plann. Inference 137 (1), 34–42.