Hybrid Monte-Carlo on Hilbert Spaces

A. Beskos[a,*], F.J. Pinski[b], J. M. Sanz-Serna[c], A.M. Stuart[d]

[a] *Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK*
[b] *Physics Department, University of Cincinnati, Geology-Physics Building, P. O. Box 210011, Cincinnati, OH 45221, USA*
[c] *Departamento de Matemática Aplicada, Universidad de Valladolid, Spain*
[d] *Mathematics Institute, University of Warwick, Coventry, CV4 7AL, UK*

Abstract

The Hybrid Monte-Carlo (HMC) algorithm provides a framework for sampling from complex, high-dimensional target distributions. In contrast with standard Markov chain Monte-Carlo (MCMC) algorithms, it generates nonlocal, non-symmetric moves in the state space, alleviating random walk type behaviour for the simulated trajectories. However, similarly to algorithms based on random walk or Langevin proposals, the number of steps required to explore the target distribution typically grows with the dimension of the state space. We define a generalized HMC algorithm which overcomes this problem for target measures arising as finite-dimensional approximations of measures π which have density with respect to a Gaussian measure on an infinite-dimensional Hilbert space. The key idea is to construct an MCMC method which is well-defined on the Hilbert space itself.

We successively address the following issues in the infinite-dimensional setting of a Hilbert space: (i) construction of a probability measure Π in an enlarged phase space having the target π as a marginal, together with a Hamiltonian flow that preserves Π; (ii) development of a suitable geometric numerical integrator for the Hamiltonian flow; and (iii) derivation of an accept/reject rule to ensure preservation of Π when using the above numerical integrator instead of the actual Hamiltonian flow. Experiments are reported that compare the new algorithm with standard HMC and with a version of the Langevin MCMC method defined on a Hilbert space.

1. Introduction

Several applications of current interest give rise to the problem of sampling a probability measure π on a separable Hilbert space $\left(\mathcal{H}, \langle\cdot,\cdot\rangle, |\cdot|\right)$ defined via

*Corresponding Author
Email addresses: alex@stats.ucl.ac.uk (A. Beskos), frank.pinski@uc.edu (F.J. Pinski), sanzsern@mac.uva.es (J. M. Sanz-Serna), a.m.stuart@warwick.ac.uk (A.M. Stuart)

its density with respect to a Gaussian measure π_0:

$$\frac{d\pi}{d\pi_0}(q) \propto \exp\bigl(-\Phi(q)\bigr) \ . \tag{1}$$

Measures with this form arise, for example, in the study of conditioned diffusions [13] and the Bayesian approach to inverse problems [27]. The aim of this paper is to develop a generalization of the Hybrid Monte-Carlo (HMC) method which is tailored to the sampling of measures π defined as in (1).

Any algorithm designed to sample π will in practice be implemented on a finite-dimensional space of dimension N; key to the efficiency of the algorithm will be its cost as a function of N. Mathematical analysis in the simplified scenario of product targets [22, 23, 1], generalizations to the non-product case in [2] and practical experience together show that the MCMC methods studied in these references require $\mathcal{O}(N^a)$ steps to explore the approximate target in $\mathbb{R}^N$, for some $a > 0$. Indeed for specific targets and proposals it is proven that for the standard Random-Walk Metropolis (RWM), Metropolis-adjusted Langevin algorithm (MALA) and HMC methods $a = 1, 1/3$ and $1/4$ respectively. The growth of the required number of steps with N occurs for one or both of the following two reasons: either because the algorithms are not defined in the limit $N = \infty$ [1] or because the proposals at $N = \infty$ are distributed according to measures which are not absolutely continuous with respect to the target measure π. Finite-dimensional approximations then require, as N increases, smaller and smaller moves to control these shortcomings. On the other hand, when π_0 is Gaussian it is now understood that both the MALA and RWM algorithms can be generalized to obtain methods which require $\mathcal{O}(1)$ steps to explore the approximate target in $\mathbb{R}^N$ [3, 4, 5]. This is achieved by designing algorithms where the the method is well-defined even on an infinite-dimensional Hilbert space $\mathcal{H}$. In this paper we show that similar ideas can be developed for the HMC method.

The standard HMC algorithm was introduced in [9]. It is based on the observation that the exponential of a separable Hamiltonian, with potential energy given by the negative logarithm of the target density, is invariant under the Hamiltonian flow. In contrast with standard Markov chain Monte-Carlo (MCMC) methods such as RWM and MALA, the HMC algorithm generates nonlocal moves in the state space, offering the potential to overcome undesirable mixing properties associated with random walk behaviour; see [19] for an overview. It is thus highly desirable to generalize HMC to the infinite-dimensional setting required by the need to sample measures of the form (1).

The paper proceeds as follows. In Section 2 we review those aspects of the standard HMC method that are helpful to motivate later developments. The new Hilbert space algorithm is presented in Section 3. Therein, we successively address the following issues in the infinite-dimensional setting of a Hilbert space:

[1] The restriction is then analogous to a Courant stability restriction in the numerical approximation of partial differential equations

(i) construction of a probability measure Π in an enlarged phase space having the target π as a marginal, together with a Hamiltonian flow that preserves Π; (ii) development of a suitable geometric numerical integrator for the Hamiltonian flow; and (iii) derivation of an accept/reject rule to ensure preservation of Π when using the above numerical integrator instead of the actual Hamiltonian flow. All required proofs have been collected in Section 4. Section 5 contains numerical experiments illustrating the advantages of our generalized HMC method over both the standard HMC method [9] and the modified MALA algorithm which is defined in a Hilbert space [3]. We make some concluding remarks in Section 6 and, in the Appendix, we gather some results from the Hamiltonian formalism.

Our presentation in the paper is based on constructing an HMC algorithm on an infinite dimensional Hilbert space; an alternative presentation of the same material could be based on constructing an HMC algorithm which behaves uniformly on a sequence of finite dimensional target distributions in $\mathbb{R}^N$ in which the Gaussian part of the target density has ratio of smallest to largest variances that approaches 0 as $N \to \infty$. Indeed the proofs in section 4 are all based on finite dimensional approximation, and passage to the limit. Theorem 4.1 is a central underpinning result in this context showing that both the numerical integrator and the acceptance probability can be approximated in finite dimensions and, importantly, that the acceptance probabilities do not degenerate to zero as $N \to \infty$ while keeping a fixed step-size in the integrator. It is this key algorithmic fact that makes the methodology proposed in this paper practical and useful. We choose the infinite dimensional perspective to present the material as it allows for concise statement of key ideas that are not based on any particular finite dimensional approximation scheme. The proofs use a spectral approximation. Such spectral methods have been widely used in the PDE literature to prove results concerning measure preservation for semilinear Hamiltonian PDEs [15, 7], but other methods such as finite differences or finite elements can be used in practical implementations. Indeed the numerical results of section 5 employ a finite difference method and, in the context of random walk algorithms on Hilbert space, both finite difference and spectral approximations are analyzed in [2]; of course, similar finite-difference based approximations could also be analyzed for HMC methods.

2. Standard HMC on $\mathbb{R}^N$

In order to facilitate the presentation of the new Hilbert-space-valued algorithm to be introduced in Section 3 below, it is convenient to first review the standard HMC method defined in [9] from a perspective that is related to our ultimate goal of using similar ideas to sample the infinite dimensional measure given by (1). For broader perspectives on the HMC method the reader is referred to the articles of Neal [19, 20].

The aim of MCMC methods is to sample from a probability density function π in $\mathbb{R}^N$. In order to link to our infinite dimensional setting in later sections we

write this density function in the form

$$\pi(q) \propto \exp\Big(-\tfrac{1}{2}\langle q, Lq\rangle - \Phi(q)\Big) \ , \tag{2}$$

where L is a symmetric, positive semi-definite matrix. At this stage the choice $L = 0$ is not excluded. When L is non-zero, (2) clearly displays the Gaussian and non-Gaussian components of the probability density, a decomposition that will be helpful in the last subsection. HMC is based on the combination of three elements: (i) a Hamiltonian flow, (ii) a numerical integrator and (iii) an accept/reject rule. Each of these is discussed in a separate subsection. The final subsection examines the choice of the mass matrix required to apply the method, specially in the case where L in (2) is 'large'. Also discussed in that subsection are the reasons, based on scaling, that suggest to implement the algorithm using the velocity in lieu of the momentum as an auxiliary variable.

The Hamiltonian formalism and the numerical integration of Hamiltonian differential equations are of course topics widely studied in applied mathematics and physics. We employ ideas and terminology from these fields and refer the reader to the texts [11, 25] for background and further details.

2.1. Hamiltonian Flow

Consider a Hamiltonian function ('energy') in $\mathbb{R}^{2N}$ associated with the target density (2):

$$H(q,p) = \tfrac{1}{2}\langle p, M^{-1}p\rangle + \tfrac{1}{2}\langle q, Lq\rangle + \Phi(q) \ . \tag{3}$$

Here p is an auxiliary variable ('momentum') and M a user-specified, symmetric positive definite 'mass' matrix. Denoting by

$$f = -\nabla\Phi$$

the 'force' stemming from the 'potential' Φ, the corresponding canonical Hamiltonian differential equations read as (see the Appendix)

$$\frac{dq}{dt} = \frac{\partial H}{\partial p} = M^{-1}p \ , \quad \frac{dp}{dt} = -\frac{\partial H}{\partial q} = -Lq + f(q) \ . \tag{4}$$

HMC is based on the fact that, for any fixed t, the t-flow Ξ^t of (4), i.e. the map $\Xi^t : \mathbb{R}^{2N} \to \mathbb{R}^{2N}$ such that

$$(q(t), p(t)) = \Xi^t(q(0), p(0)) \ ,$$

preserves both the volume element $dq\,dp$ and the value of H. As a result, Ξ^t also preserves the measure in the phase space $\mathbb{R}^{2N}$ with density

$$\Pi(q,p) \propto \exp(-H(q,p)) = \exp\big(-\tfrac{1}{2}\langle p, M^{-1}p\rangle\big)\ \exp\big(-\tfrac{1}{2}\langle q, Lq\rangle - \Phi(q)\big) \ , \tag{5}$$

whose q and p marginals are respectively the target (2) and a centred Gaussian with M as a covariance matrix. It follows that if we assume that the initial value $q(0)$ is distributed according to (2) and we draw $p(0) \sim N(0, M)$, then

$q(t)$ will also follow the law (2). This shows that the implied Markov transition kernel $q(0) \mapsto q(T)$, with a user-defined fixed $T > 0$, defines a Markov chain in $\mathbb{R}^N$ that has (2) as an invariant density and makes nonlocal moves in that state space (see [19, 26]). Furthermore, the chain is reversible, in view of the symmetry of the distribution $N(0, M)$ and of the time-reversibility ([25]) of the dynamics of (4): that is, if $\Xi^T(q, p) = (q', p')$, then $\Xi^T(q', -p') = (q, -p)$.

2.2. Numerical Integrator

In general the analytic expression of the flow Ξ^t is not available and it is necessary to resort to numerical approximations to compute the transitions. The integrator of choice, the Verlet/leap-frog method, is best presented as a splitting algorithm, see e.g. [25]. The Hamiltonian (3) is written in the form

$$H = H_1 + H_2 \ , \quad H_1 = \tfrac{1}{2}\langle q, Lq\rangle + \Phi(q) \ , \quad H_2 = \tfrac{1}{2}\langle p, M^{-1}p\rangle \ ,$$

where the key point is that the flows Ξ_1^t, Ξ_2^t of the split Hamiltonian systems

$$\frac{dq}{dt} = \frac{\partial H_1}{\partial p} = 0 \ , \quad \frac{dp}{dt} = -\frac{\partial H_1}{\partial q} = -Lq + f(q)$$

and

$$\frac{dq}{dt} = \frac{\partial H_2}{\partial p} = M^{-1}p \ , \quad \frac{dp}{dt} = -\frac{\partial H_2}{\partial q} = 0$$

may be explicitly computed:

$$\Xi_1^t(q, p) = \big(q, p - tLq + tf(q)\big) \ , \quad \Xi_2^t(q, p) = \big(q + tM^{-1}p, p\big) \ . \tag{6}$$

Then a time-step of length $h > 0$ of the Verlet algorithm is, by definition, carried out by composing three substeps:

$$\Psi_h = \Xi_1^{h/2} \circ \Xi_2^h \circ \Xi_1^{h/2} \ ; \tag{7}$$

and the exact flow Ξ^T of (4) is approximated by the transformation $\Psi_h^{(T)}$ obtained by concatenating $\lfloor \frac{T}{h} \rfloor$ Verlet steps:

$$\Psi_h^{(T)} = \Psi_h^{\lfloor \frac{T}{h} \rfloor} \ . \tag{8}$$

Since the mappings Ξ_1^t and Ξ_2^t are exact flows of Hamiltonian systems, the transformation $\Psi_h^{(T)}$ itself is symplectic and preserves the volume element $dq\,dp$ (see the Appendix). Also the symmetry in the right hand-side of (7) (Strang's splitting) results in $\Psi_h^{(T)}$ being *time-reversible:*

$$\Psi_h^{(T)}(q, p) = (q', p') \quad \Leftrightarrow \quad \Psi_h^{(T)}(q', -p') = (q, -p) \ .$$

The map $\Psi_h^{(T)}$ is an example of a geometric integrator [11]: it preserves various geometric properties of the flow Ξ^T and in particular the symplectic and time-reversible nature of the underlying flow. However $\Psi_h^{(T)}$ does not preserve the value of H: it makes an $\mathcal{O}(h^2)$ error and, accordingly, it does not exactly preserve the measure with density (5).

2.3. Accept/Reject Rule

The invariance of (5) in the presence of integration errors is ensured through an accept/reject Metropolis-Hastings mechanism; the right recipe is given in steps (iii) and (iv) of Table 1, that summarizes the standard HMC algorithm [9, 19].

HMC on $\mathbb{R}^N$:

(i) *Pick $q^{(0)} \in \mathbb{R}^N$ and set $n = 0$.*

(ii) *Given $q^{(n)}$, compute*

$$(q^\star, p^\star) = \Psi_h^{(T)}(q^{(n)}, p^{(n)})$$

where $p^{(n)} \sim N(0, M)$ and propose $q^\star$.

(iii) *Calculate*

$$a = \min\Big(1, \exp\big(H(q^{(n)}, p^{(n)}) - H(q^*, p^*)\big)\Big) \ .$$

(iv) *Set $q^{(n+1)} = q^\star$ with probability a; otherwise set $q^{(n+1)} = q^{(n)}$.*

(v) *Set $n \to n + 1$ and go to (ii).*

Table 1: Standard HMC algorithm on $\mathbb{R}^N$. It generates a Markov chain $q^{(0)} \mapsto q^{(1)} \mapsto \dots$ reversible with respect to the target probability density function (2). The numerical integrator $\Psi_h^{(T)}$ is defined by (6)–(8).

2.4. Choice of Mass Matrix

As pointed out above, the mass-matrix M is a 'parameter' to be selected by the user; the particular choice of M will have great impact on the efficiency of the algorithm ([10]). A rule of thumb ([17]) is that directions where the target (2) possesses larger variance should be given smaller mass so that the Hamiltonian flow can make faster progress along them. This rule of thumb is used to select the mass matrix to study a polymer chain in [14].

In order to gain understanding concerning the role of M and motivate the material in Section 3, we consider in the remainder of this section the case where in (2) the matrix L is positive-definite and $\Phi(q)$ is small with respect to $\langle q, Lq\rangle$, i.e. the case where the target is a perturbation of the distribution $N(0, L^{-1})$. In agreement with the rule of thumb above, we set $M = L$ so that (4) reads

$$\frac{dq}{dt} = L^{-1}p \ , \quad \frac{dp}{dt} = -Lq + f(q) \ . \tag{9}$$

Let us now examine the limit situation where the perturbation vanishes, i.e. $\Phi \equiv 0$. From (5), at stationarity, $q \sim N(0, L^{-1})$, $p \sim N(0, L)$. Furthermore in (9), $f \equiv 0$ so that, after eliminating p,

$$\frac{d^2q}{dt^2} = -q \ .$$

Thus, $q(t)$ undergoes oscillations with angular frequency 1 *regardless of the size of the eigenvalues of L/(co)-variances of the target.* From a probabilistic point of view, this implies that, if we think of q as decomposed in independent Gaussian scalar components, the algorithm (as intended with the choice of mass matrix $M = L$) automatically adjusts itself to the fact that different components may possess widely different variances. From a numerical analysis point of view, we see that the Verlet algorithm will operate in a setting where it will not be necessary to reduce the value of h to avoid *stability* problems originating from the presence of fast frequencies.[2]

Remark 1. Let us still keep the choice $M = L$ but drop the assumption $\Phi \equiv 0$ and suppose that L has some very large eigenvalues (i.e. the target distribution presents components of very small variance). As we have just discussed, we do not expect such large eigenvalues to negatively affect the dynamics of q. However we see from the second equation in (9) that p (which, recall, is only an auxiliary variable in HMC) will in general be large. In order to avoid variables of large size, it is natural to rewrite the algorithm in Table 1 using throughout the scaled variable

$$v = L^{-1}p$$

rather than p. Since $v = M^{-1}p = dq/dt$, the scaled variable possesses a clear meaning: it is the 'velocity' of q.

In terms of v the system (9) that provides the required flow reads

$$\frac{dq}{dt} = v \ , \quad \frac{dv}{dt} = -q + L^{-1}f(q) \ ; \tag{10}$$

the value of the Hamiltonian (3) to be used in the accept/reject step is given by

$$H = \tfrac{1}{2}\langle v, Lv\rangle + \tfrac{1}{2}\langle q, Lq\rangle + \Phi(q) \ , \tag{11}$$

and the invariant density (5) in $\mathbb{R}^{2N}$ becomes

$$\Pi(q, v) \propto \exp\big(-\tfrac{1}{2}\langle v, Lv\rangle\big) \exp\big(-\tfrac{1}{2}\langle q, Lq\rangle - \Phi(q)\big) \ . \tag{12}$$

[2]Note that the Verlet algorithm becomes unstable whenever $h\omega \geq 2$, where ω is any of the angular frequencies present in the dynamics. While the choice of mass matrix $M = L$ precludes the occurrence of *stability* problems in the integration, the standard HMC algorithm in the present setting ($M = L$, $\Phi \equiv 0$) still suffers from the restriction $h = \mathcal{O}(N^{1/4})$ discussed in [1] (the restriction stems from *accuracy* —rather than stability— limitations in the Verlet integrator). The new integrator to be introduced later in the paper is *exact* in the setting $\Phi \equiv 0$, and hence eliminates this problem.

Note that the marginal for v is

$$v \sim N(0, L^{-1}) \; ; \tag{13}$$

the initial value $v^{(n)}$ at step (ii) of Table 1 should be drawn accordingly. Note that this formulation has the desirable attribute that, when $\Phi \equiv 0$, the position and velocity are independent draws from the same distribution.

We finish the section with two comments concerning introduction of the variable v in place of p. The algorithm expressed in terms of v may be found either by first replacing p by v in the differential equations (9) to get (10) and then applying the Verlet algorithm; or by first applying the Verlet algorithm to the system (9) and then replacing p by v in the equations of the integrator: the Verlet discretization commutes with the scaling $p \mapsto v = M^{-1}p$. In addition, it is important to note that (10) is also a Hamiltonian system, albeit of a non-canonical form, see the Appendix.

3. The Algorithm

In this section we define the new algorithm on a Hilbert space $\big(\mathcal{H}, \langle\cdot,\cdot\rangle, |\cdot|\big)$, and outline the main mathematical properties of the algorithm. After introducing the required assumptions on the distribution to be sampled, we discuss successively the flow, the numerical integrator and the accept/reject strategy.

3.1. Assumptions on π_0 and Φ

Throughout we assume that π_0 in (1) is a non-degenerate (non-Dirac) centred Gaussian measure with covariance operator $\mathcal{C}$. Thus, $\mathcal{C}$ is a positive, self-adjoint, nuclear operator (i.e. its eigenvalues are summable) whose eigenfunctions span $\mathcal{H}$. For details on properties of Gaussian measures on a Hilbert space see section 2.3 of [8], and for the Banach space setting see [6, 16].

Let $\{\phi_j\}_{j\geq 1}$ be the (normalised) eigenfunctions of $\mathcal{C}$ and λ_j^2 its eigenvalues, so that

$$\mathcal{C}\phi_j = \lambda_j^2 \, \phi_j \; , \quad j \geq 1 \; .$$

The expansion

$$q = \sum_{j=1}^{\infty} q_j \phi_j \tag{14}$$

establishes an isomorphism between $\mathcal{H}$ and the space

$$\ell_2 = \Big\{ \{q_j\}_{j=1}^{\infty} \in \mathbb{R}^{\infty} : \sum q_j^2 < \infty \Big\} \tag{15}$$

that maps each element q into the corresponding sequence of coefficients $\{q_j\}_{j\geq 1}$. This isomorphism gives rise to subspaces ($s > 0$) and superspaces ($s < 0$) of $\mathcal{H}$:

$$\mathcal{H}^s := \Big\{ \{q_j\}_{j=1}^{\infty} \in \mathbb{R}^{\infty} : |q|_s < \infty \Big\} \; ,$$

where $|\cdot|_s$ denotes the following Sobolev-like norm:

$$|q|_s := \Big(\sum_{j=1}^{\infty} j^{2s} q_j^2\Big)^{1/2} , \quad s \in \mathbb{R} .$$

Note that $\mathcal{H}^0 = \mathcal{H}$. For an introduction to Sobolev spaces defined this way see Appendix A of [24].

If $q \sim N(0, \mathcal{C})$, then

$$q_j \sim N(0, \lambda_j^2) \tag{16}$$

independently over j. Thus, λ_j is the standard deviation, under the reference measure π_0, of the j^{th} coordinate. We shall impose the following condition, with the bound $\kappa > 1/2$ being required to ensure that $\mathcal{C}$ is nuclear, so that the Gaussian distribution is well-defined:

Condition 3.1. *The standard deviations $\{\lambda_j\}_{j\geq 1}$ decay at a polynomial rate $\kappa > 1/2$, that is*

$$\lambda_j = \Theta(j^{-\kappa})$$

i.e. $\liminf_{j\to\infty} j^{\kappa}\lambda_j > 0$ *and* $\limsup_{j\to\infty} j^{\kappa}\lambda_j < \infty$.

From this condition and (16), a direct computation shows that $\mathbb{E}|q|_s^2 < \infty$ for $s \in [0, \kappa - 1/2)$ and hence that

$$|q|_s < \infty \ , \ \pi_0 - \text{a.s.} \ , \ \text{for any} \ \ s \in [0, \kappa - 1/2) \ .$$

Therefore, we have the following.

Proposition 3.1. *Under Condition 3.1, the probability measure π_0 is supported on $\mathcal{H}^s$ for any $s < \kappa - 1/2$.*

Let us now turn to the hypotheses on the real-valued map ('potential') Φ. In the applications which motivate us (see [13, 27]) Φ is typically defined on a dense subspace of $\mathcal{H}$. To be concrete we will assume throughout this paper that the domain of Φ is $\mathcal{H}^\ell$ for some fixed $\ell \geq 0$. Then the Fréchet derivative $D\Phi(q)$ of Φ is, for each $q \in \mathcal{H}^\ell$, a linear map from $\mathcal{H}^\ell$ into $\mathbb{R}$ and therefore we may identify it with an element of the dual space $\mathcal{H}^{-\ell}$. We use the notation $f(q) = -D\Phi(q)$ and, from the preceding discussion, view f ('the force') as a function $f : \mathcal{H}^\ell \to \mathcal{H}^{-\ell}$. The first condition concerns properties of f.

Condition 3.2. *There exists $\ell \in [0, \kappa - \frac{1}{2})$, where κ is as in Condition 3.1, such that $\Phi : \mathcal{H}^\ell \to \mathbb{R}$ is continuous and $f = -D\Phi : \mathcal{H}^\ell \to \mathcal{H}^{-\ell}$ is globally Lipschitz continuous, i.e. there exists a constant $K > 0$ such that, for all $q, q' \in \mathcal{H}^\ell$,*

$$|f(q) - f(q')|_{-\ell} \leq K\, |q - q'|_\ell \ .$$

The next condition is a bound from below on Φ, characterizing the idea that the change of measure is dominated by the Gaussian reference measure.

Condition 3.3. *Fix ℓ as given in Condition 3.2. Then, for any $\epsilon > 0$ there exists $M = M(\epsilon) > 0$ such that, for all $q \in \mathcal{H}^\ell$*

$$\Phi(q) \geq M - \epsilon |q|_\ell^2 \ .$$

Under Conditions 3.1, 3.2 and 3.3 , (1) defines π as a probability measure absolutely continuous with respect to π_0; Proposition 3.1 ensures that π is supported in $\mathcal{H}^s$ for any $s < \kappa - 1/2$ and in particular that $\pi(\mathcal{H}^\ell) = 1$; Condition 3.3 guarantees, via the Fernique Theorem (2.6 in [8]), the integrability of $\exp(-\Phi(q))$ with respect to π_0; the Lipschitz condition on f in Condition 3.2 ensures continuity properties of the measure with respect to perturbations of various types. The reader is referred to [27] for further details. The conditions above summarize the frequently occurring situation where a Gaussian measure π_0 dominates the target measure π. This means intuitively that the random variable $\langle u, \phi_j \rangle$ behaves, for large j, almost the same under $u \sim \pi$ and under $u \sim \pi_0$. It then possible to construct effective algorithms to sample from π by using knowledge of π_0.

We remark that our global Lipschitz condition could be replaced by a local Lipschitz assumption at the expense of a more involved analysis; indeed we will give numerical results in Subsection 5.2 for a measure arising from conditioned diffusions where f is only locally Lipschitz.

We shall always assume hereafter that Conditions 3.1–3.3 are satisfied and use the symbols κ and ℓ to refer to the two fixed constants that arise from them.

3.2. Flow

There is a clear analogy between the problem of sampling from π given by (1) in $\mathcal{H}$ and the problem, considered in Subsection 2.4, of sampling from the density (2) in $\mathbb{R}^N$ with L positive definite and $\Phi(q)$ small with respect to $\langle q, Lq \rangle$. In this analogy, $\pi_0(dq)$ corresponds to the measure $\exp(-(1/2)\langle q, Lq \rangle)dq$ and therefore the covariance operator $\mathcal{C}$ corresponds to the matrix L^{-1}: L is the precision operator. Many of the considerations that follow are built on this parallelism.

The key idea in HMC methods is to double the size of the state space by adding an auxiliary variable related to the 'position' q. We saw in Remark 1 in Subsection 2.4, that, in the setting considered there, large eigenvalues of L lead to large values of the momentum p but do not affect the size of v. In the Hilbert space setting, the role of L is played by $\mathcal{C}^{-1}$ which has eigenvalues $1/\lambda_j^2$ of arbitrarily large size. This suggests working with the velocity $v = dq/dt$ as an auxiliary variable and not the momentum. Equation (13) prompts us to use π_0 as the marginal distribution of v and introduce the following Gaussian measure Π_0 on $\mathcal{H} \times \mathcal{H}$

$$\Pi_0(dq, dv) = \pi_0(dq) \otimes \pi_0(dv) \ .$$

We define accordingly (cf. (12)):

$$\frac{d\Pi}{d\Pi_0}(q, v) \propto \exp\bigl(-\Phi(q)\bigr) \ , \tag{17}$$

so that the marginal on q of Π is simply the target distribution π. Furthermore (10) suggests to chose

$$\frac{dq}{dt} = v \ , \quad \frac{dv}{dt} = -q + \mathcal{C}\, f(q) \tag{18}$$

as the equations to determine the underlying dynamics that will provide (when solved numerically) proposals for the HMC algorithm with target distribution π.

Our first result shows that (18) defines a well-behaved flow Ξ^t in the subspace $\mathcal{H}^\ell \times \mathcal{H}^\ell$ of $\mathcal{H} \times \mathcal{H}$ which, according to Proposition 3.1, has full Π_0 (or Π) measure. The space $\mathcal{H}^\ell \times \mathcal{H}^\ell$ is assumed to have the product topology of the factor spaces $\big(\mathcal{H}^\ell, |\cdot|_\ell\big)$. We state precisely the dependence of the Lipschitz constant for comparison with the situation arising in the next section where we approximate in $N \gg 1$ dimensions and with time-step h, but Lipschitz constants are independent of N and h and exhibit the same dependence as in this section.

Proposition 3.2.

(i) For any initial condition $\big(q(0), v(0)\big) \in \mathcal{H}^\ell \times \mathcal{H}^\ell$ and any $T > 0$ there exists a unique solution of (18) in the space $C^1([-T, T], \mathcal{H}^\ell \times \mathcal{H}^\ell)$.

(ii) Let $\Xi^t : \mathcal{H}^\ell \times \mathcal{H}^\ell \to \mathcal{H}^\ell \times \mathcal{H}^\ell$, $t \in \mathbb{R}$ denote the group flow of (18), so that

$$\big(q(t), v(t)\big) = \Xi^t\big(q(0), v(0)\big) \ .$$

The map Ξ^t is globally Lipschitz with a Lipschitz constant of the form $\exp(K|t|)$, where K depends only on $\mathcal{C}$ and Φ.

(iii) Accordingly, for each $T > 0$, there exists constant $C(T) > 0$ such that, for $0 \le t \le T$,

$$|q(t)|_\ell + |v(t)|_\ell \le C(T)\big(1 + |q(0)|_\ell + |v(0)|_\ell\big) \ .$$

Our choices of measure (17) and dynamics (18) have been coordinated to ensure that Ξ^t preserves Π:

Theorem 3.1. *For any $t \in \mathbb{R}$, the flow Ξ^t preserves the probability measure Π given by* (17).

The theorem implies that π will be an invariant measure for the Markov chain for q defined through the transitions $q^{(n)} \mapsto q^{(n+1)}$ determined by

$$(q^{(n+1)}, v^{(n+1)}) = \Xi^T(q^{(n)}, v^{(n)}) \ , \quad v^{(n)} \sim \pi_0 \ , \tag{19}$$

where the $v^{(n)}$ form an independent sequence. This chain is actually reversible:

Theorem 3.2. *For any $t \in \mathbb{R}$, the Markov chain defined by (19) is reversible under the distribution $\pi(q)$ in (1).*

We conclude by examining whether the dynamics of (18) preserve a suitable Hamiltonian function. The Hilbert-space counterpart of (11) is given by

$$\mathsf{H}(q, v) = \frac{1}{2}\,\langle v, \mathcal{C}^{-1} v\rangle + \frac{1}{2}\,\langle q, \mathcal{C}^{-1} q\rangle + \Phi(q) \tag{20}$$

and it is in fact trivial to check that H and therefore $\exp(-\mathsf{H})$ are *formal* invariants of (18). However the terms $\langle q, \mathcal{C}^{-1} q\rangle$ and $\langle v, \mathcal{C}^{-1} v\rangle$ are almost surely infinite in an infinite-dimensional context. This may be seen from the fact that $|\mathcal{C}^{-\frac{1}{2}} \cdot|$ is the Cameron-Martin norm for π_0, see e.g. [6, 8], or directly from a zero-one law applied to a series representation of the inner-product. For further discussion on the Hamiltonian nature of (18) see the Appendix.

3.3. Numerical Integrator

Our next task is to study how to numerically approximate the flow Ξ^T. As in the derivation of the Verlet algorithm in Subsection 3.3 we resort to the idea of splitting; however the splitting that we choose is different, dictated by a desire to ensure that the resulting MCMC method is well-defined on Hilbert space. The system (18) is decomposed as (see the Appendix)

$$\frac{dq}{dt} = 0 \ , \quad \frac{dv}{dt} = \mathcal{C} f(q) \tag{21}$$

and

$$\frac{dq}{dt} = v \ , \quad \frac{dv}{dt} = -q \tag{22}$$

with the explicitly computable flows

$$\Xi_1^t(q, v) = (q, \ v + t\, \mathcal{C} f(q)) \ , \tag{23}$$

and

$$\Xi_2^t(q, v) = \Big(\cos(t)\, q + \sin(t)\, v, \ -\sin(t)\, q + \cos(t)\, v \Big) \ . \tag{24}$$

This splitting has also been recently suggested in the review [20], although without the high-dimensional motivation of relevance here.

A time-step of length $h > 0$ of the integrator is carried out by the symmetric composition (Strang's splitting)

$$\Psi_h = \Xi_1^{h/2} \circ \Xi_2^h \circ \Xi_1^{h/2} \tag{25}$$

and the exact flow Ξ^T, $T > 0$, of (18) is approximated by the map $\Psi_h^{(T)}$ obtained by concatenating $\lfloor \frac{T}{h} \rfloor$ steps:

$$\Psi_h^{(T)} = \Psi_h^{\lfloor \frac{T}{h} \rfloor}. \tag{26}$$

This integrator is time-reversible —due to the symmetric pattern in the Strang splitting and the time-reversibility of Ξ_1^t and Ξ_2^t— and if applied in a finite-dimensional setting would also preserve the volume element $dq\, dv$. In the case where $\Phi \equiv 0$, the integrator coincides with the rotation Ξ_2^t; it is therefore exact and preserves exactly the measure Π_0. However, in general, $\Psi_h^{(T)}$ does not preserve formally the Hamiltonian (20), a fact that renders necessary the introduction of an accept/reject criterion, as we will describe in the following subsection.

The next result is analogous to Proposition 3.2:

Proposition 3.3.

(i) For any $(q,v) \in \mathcal{H}^\ell \times \mathcal{H}^\ell$ *we have* $\Psi_h(q,v) \in \mathcal{H}^\ell \times \mathcal{H}^\ell$ *and therefore* $\Psi_h^{(T)}(q,v) \in \mathcal{H}^\ell \times \mathcal{H}^\ell$.

(ii) Ψ_h*, and therefore* $\Psi_h^{(T)}$*, preserves absolute continuity with respect to* Π_0 *and* Π.

(iii) $\Psi_h^{(T)}$ *is globally Lipschitz as a map from* $\mathcal{H}^\ell \times \mathcal{H}^\ell$ *onto itself with a Lipschitz constant of the form* $\exp(KT)$ *with* K *depending only on* $\mathcal{C}$ *and* Φ.

(iv) Accordingly, for each $T > 0$ *there exists* $C(T) > 0$ *such that, for all* $0 \le ih \le T$,

$$|q_i|_\ell + |v_i|_\ell \le C(T)\big(1 + |q_0|_\ell + |v_0|_\ell\big) \ ,$$

where

$$(q_i, v_i) = \Psi_h^i(q_0, v_0) \ . \tag{27}$$

3.4. Accept/Reject Rule

The analogy with the standard HMC would suggest the use of

$$1 \wedge \exp\Big(\mathsf{H}(q^{(n)}, v^{(n)}) - \mathsf{H}(\Psi_h^{(T)}(q^{(n)}, v^{(n)})) \Big)$$

to define the acceptance probability. Unfortunately and as pointed out above, H is almost surely infinite in our setting. We will bypass this difficulty by deriving a well behaved expression for the energy difference

$$\Delta\mathsf{H}(q,v) = \mathsf{H}(\Psi_h^{(T)}(q,v)) - \mathsf{H}(q,v)$$

in which the two infinities cancel.

A straightforward calculation using the definition of $\Psi_h(q,v)$ gives, for one time-step $(q',v') = \Psi_h(q,v)$:

$$\begin{aligned}\mathsf{H}(q',v') - \Phi(q') = \mathsf{H}(q,v) - \Phi(q) + \frac{h^2}{8}\Big(|\mathcal{C}^{\frac12} f(q)|^2 - |\mathcal{C}^{\frac12} f(q')|^2\Big) \\ + \frac{h}{2}\Big(\langle f(q), v\rangle + \langle f(q'), v'\rangle\Big) \ .\end{aligned}$$

Using this result iteratively, we obtain for $I = \lfloor T/h \rfloor$ steps (subindices refer to time-levels along the numerical integration):

$$\begin{aligned}\Delta\mathsf{H}(q_0,v_0) = \Phi(q_I) - \Phi(q_0) + \frac{h^2}{8}\Big(|\mathcal{C}^{\frac12} f(q_0)|^2 - |\mathcal{C}^{\frac12} f(q_I)|^2\Big) \\ + h\sum_{i=1}^{I-1}\langle f(q_i), v_i\rangle + \frac{h}{2}\left(\langle f(q_0), v_0\rangle + \langle f(q_I), v_I\rangle\right) \ . \end{aligned}\tag{28}$$

(We note in passing that in the continuum $h \to 0$ limit, (28) gives formally:

$$\mathsf{H}\big(q(T), v(T)\big) - \mathsf{H}\big(q(0), v(0)\big) = \Phi\big(q(T)\big) - \Phi\big(q(0)\big) + \int_0^T \langle f\big(q(t)\big), v(t)\rangle \, dt \ ,$$

HMC on $\mathcal{H}^\ell$:

(i) Pick $q^{(0)} \sim \Pi_0$ and set $n = 0$.

(ii) Given $q^{(n)}$, compute

$$(q^\star, v^*) = \Psi_h^{(T)}(q^{(n)}, v^{(n)})$$

where $v^{(n)} \sim N(0, \mathcal{C})$ and propose q^*.

(iii) Using (28),(29), define

$$a = a(q^{(n)}, v^{(n)}) \ .$$

(iv) Set $q^{(n+1)} = q^\star$ with probability a; otherwise set $q^{(n+1)} = q^{(n)}$.

(v) Set $n \to n+1$ and go to (ii).

Table 2: The HMC algorithm on a Hilbert space, for sampling from π in (1). The numerical integrator $\Psi_h^{(T)}$ is defined by (23)–(26).

with the right hand side here being identically 0: the gain in potential energy equals the power of the applied force. This is a reflection of the formal energy conservation by the flow (18) pointed out before.) Condition 3.2 and parts (ii) and (iv) of Lemma 4.1 in Section 4 now guarantee that $\Delta\mathsf{H}(q, v)$, as defined in (28), is a Π-a.s. finite random variable; in fact $\Delta\mathsf{H} : \mathcal{H}^\ell \times \mathcal{H}^\ell \to \mathbb{R}$ is continuous according to parts (iii) and (v) of that Lemma. We may therefore define the *acceptance probability* by

$$a(q, v) = \min\Big(1, \exp\big(-\Delta\mathsf{H}(q, v)\big)\Big) \ . \tag{29}$$

We are finally ready to present an HMC algorithm on $\mathcal{H}^\ell$ aiming at simulating from $\pi(q)$ in equilibrium. The pseudo-code is given in Table 2. Our main result asserts that the algorithm we have defined achieves its goal:

Theorem 3.3. *For any choice of $T > 0$, the algorithm in Table 2 defines a Markov chain which is reversible under the distribution $\pi(q)$ in (1).*

The practical application of the algorithm requires of course to replace $\mathcal{H}$, π_0 and Φ by finite-dimensional approximations. Once these have been chosen, it is a trivial matter to write the corresponding versions of the differential system (18), of the integrator and of the accept/reject rule. The case where the discretization is performed by a spectral method is presented and used for the purposes of analysis in Subsection 4.2. However many alternative possibilities exist: for example in Subsection 5.2 we present numerical results based on finite-dimensionalization using finite differences. For any finite-dimensional approximation of the state

space, the fact that the algorithm is defined in the infinite-dimensional limit imparts robustness under refinement of finite-dimensional approximation.

Of course in any finite dimensional implementation used in practice the value of H will be finite, but large, and so it would be possible to evaluate the energy difference by subtracting two evaluations of H. However this would necessitate the subtraction of two large numbers, something well known to be undesirable in floating-point arithmetic. In contrast the formula we derive has removed this subtraction of two infinities, and is hence suitable for floating-point use.

4. Proofs and Finite Dimensional Approximation

This section contains some auxiliary results and the proofs of the theorems and propositions presented in Section 3. The method of proof is to consider finite dimensional approximation in $\mathbb{R}^N$, and then pass to the limit as $N \to \infty$. In so doing we also prove some useful results concerning the behaviour of finite dimensional implementations of our new algorithm. In particular Theorem 4.1 shows that the acceptance probability does not degenerate to 0 as N increases, for fixed timestep h in the integrator. This is in contrast to the standard HMC method where the choice $h = \mathcal{O}(N^{-\frac{1}{4}})$ is required to ensure $\mathcal{O}(1)$ acceptance probabilities. We discuss such issues further in section 5.

4.1. Preliminaries

Recall the fixed values of κ and ℓ defined by Conditions 3.1 and 3.2 respectively. The bounds for $f = -D\Phi$ provided in the following lemma will be used repeatedly. The proof relies on the important observation that Condition 3.1 implies that

$$|\mathcal{C}^{-s/2\kappa} \cdot | \asymp | \cdot |_s, \tag{30}$$

where we use the symbol $\asymp$ to denote an equivalence relation between two norms.

Lemma 4.1. *There exists a constant $K > 0$ such that*

(i) for all $q, q' \in \mathcal{H}^\ell$,

$$|\mathcal{C}f(q) - \mathcal{C}f(q')|_\ell \leq K|q - q'|_\ell \ ;$$

(ii) for all $q, v \in \mathcal{H}^\ell$,

$$|\langle f(q), v\rangle| \leq K\big(1 + |q|_\ell\big)|v|_\ell \ ;$$

(iii) for all $q, q', v, v' \in \mathcal{H}^\ell$,

$$|\langle f(q), v\rangle - \langle f(q'), v'\rangle| \leq K|v|_\ell|q - q'|_\ell + K\big(1 + |q'|_\ell\big)|v - v'|_\ell \ ;$$

(iv) for all $q \in \mathcal{H}^\ell$,

$$|\mathcal{C}^{\frac{1}{2}} f(q)| \leq K\big(1 + |q|_\ell\big) \ ;$$

(v) for all $q, q' \in \mathcal{H}^\ell$,

$$|\mathcal{C}^{\frac{1}{2}} f(q) - \mathcal{C}^{\frac{1}{2}} f(q')| \leq K|q - q'|_\ell \ .$$

Proof. From (30)

$$|\mathcal{C}\cdot|_\ell \asymp |\mathcal{C}^{1-\frac{\ell}{2\kappa}}\cdot| \ , \quad |\cdot|_{-\ell} \asymp |\mathcal{C}^{\frac{\ell}{2\kappa}}\cdot| \ ,$$

and, since $\ell < \kappa - 1/2$, we have that $1 - \ell/(2\kappa) > \ell/(2\kappa)$. Thus, there is a constant K such that

$$|\mathcal{C}\cdot|_\ell \le K|\cdot|_{-\ell} \ .$$

Item (i) now follows from Condition 3.2. For item (ii) note that, by (30) and Condition 3.2, we have

$$\begin{aligned}|\langle f(q), v\rangle| &= |\langle \mathcal{C}^{\ell/2\kappa} f(q), \mathcal{C}^{-\ell/2\kappa} v\rangle| \\ &\le |\mathcal{C}^{\ell/2\kappa} f(q)||\mathcal{C}^{-\ell/2\kappa} v| \le K|f(q)|_{-\ell}|v|_\ell \le K\big(1+|q|_\ell\big)|v|_\ell \ .\end{aligned}$$

The proof of item (iii) is similar. For (iv) we write, by (30) and since $0 < \ell < \kappa$,

$$|\mathcal{C}^{\frac{1}{2}} f(q)| \le K|f(q)|_{-\kappa} \le K|f(q)|_{-\ell} \ .$$

Item (v) is proved in an analogous way. □

Proof of Proposition 3.2. Lemma 4.1 shows that $\mathcal{C}f$ is a globally Lipschitz mapping from $\mathcal{H}^\ell$ into itself. Therefore (18) is an ordinary differential equation in $\mathcal{H}^\ell \times \mathcal{H}^\ell$ with globally Lipschitz right hand-side which proves directly the statement. □

Proof of Proposition 3.3. Part (i) is a consequence of (i) in Lemma 4.1. For part (ii) it is clearly sufficient to address the case of the Gaussian law Π_0. From the definition of Ψ_h as a composition, it is enough to show that Ξ_1^t and Ξ_2^t defined in (23) and (24) preserve absolute continuity with respect to Π_0. The rotation Ξ_2^t preserves Π_0 exactly. The transformation Ξ_1^t leaves q invariant and thus it suffices to establish that for, any fixed $q \in \mathcal{H}^\ell$, the mapping $v \mapsto v + t\,\mathcal{C}\,f(q)$ preserves absolute continuity with respect to $N(0, \mathcal{C})$. Writing $\mathcal{C}f(q) = \mathcal{C}^{1/2}\{\mathcal{C}^{1/2} f(q)\}$, we see from Lemma 4.1(iv) that $\mathcal{C}^{1/2}\, f(q)$ is an element of $\mathcal{H}$; then, the second application of $\mathcal{C}^{1/2}$ projects $\mathcal{H}$ onto the Cameron-Martin space of the Gaussian measure $N(0, \mathcal{C})$. It is well known (see e.g. Theorem 2.23 in [8]) that translations by elements of the Cameron-Martin space preserve absolute continuity of the Gaussian measure.

Parts (iii) and (iv) are simple consequences of the fact that both Ξ_1^t and Ξ_2^t are globally Lipschitz continuous with constants of the form $1 + \mathcal{O}(|t|)$ as $t \to 0$. □

4.2. Finite-Dimensional Approximations

The proofs of the the main Theorems 3.1, 3.2 and 3.3, to be presented in the next subsection, and involve demonstration of invariance or reversibility properties of the algorithmic dynamics, rely on the use of finite-dimensional approximations.

Taking into account the spectral decomposition (14), we introduce the subspaces ($N \in \mathbb{N}$):

$$\mathcal{H}_N = \{q \in \mathcal{H} : q = \sum_{j=1}^{N} q_j \, \phi_j, \; q_j \in \mathbb{R}\} \; ,$$

and denote by $\mathrm{proj}_{\mathcal{H}_N}$ the projection of $\mathcal{H}$ onto $\mathcal{H}_N$. For $q, v \in \mathcal{H}$, we also employ the notations:

$$q^N = \mathrm{proj}_{\mathcal{H}_N}(q) \; , \quad v^N = \mathrm{proj}_{\mathcal{H}_N}(v) \; .$$

We will make use of the standard isomorphism $\mathcal{H}_N \leftrightarrow \mathbb{R}^N$ and will sometimes treat a map $\mathcal{H}_N \to \mathcal{H}_N$ as one $\mathbb{R}^N \to \mathbb{R}^N$; this should not create any confusion. If we think of elements of $\mathcal{H}$ as functions of 'spatial' variables, then the process of replacing $\mathcal{H}$ by $\mathcal{H}_N$ corresponds to space discretization by means of a spectral method.

We introduce the distributions in $\mathcal{H}_N$ (equivalently, $\mathbb{R}^N$) given by

$$\pi_{0,N} = N(0, C_N) \; , \qquad \pi_N(q) \propto \exp\{-\tfrac{1}{2}\langle q, C_N^{-1} q\rangle - \Phi_N(q)\} \; , \tag{31}$$

where C_N is the $N \times N$ diagonal matrix

$$C_N = \mathrm{diag}\{\lambda_1^2, \lambda_2^2, \ldots, \lambda_N^2\}$$

and Φ_N is the restriction of Φ to $\mathcal{H}_N$, i.e.

$$\Phi_N(q) = \Phi(q) \; , \quad \text{for } q \in \mathcal{H}_N \; .$$

To sample from π_N we reformulate the algorithm in Table 2 in the present finite-dimensional setting. Once more we discuss the flow, the integrator and the accept/reject rule, now for the finite dimensional approximation.

Since, for $q \in \mathcal{H}_N$, $D\Phi_N(q) \equiv \mathrm{proj}_{\mathcal{H}_N} D\Phi(q)$, instead of the system (18) we now consider:

$$\frac{dq}{dt} = v \; , \quad \frac{dv}{dt} = -q + \mathcal{C}\, \mathrm{proj}_{\mathcal{H}_N} f(q) \tag{32}$$

(for convenience we have written $\mathcal{C}$ here instead of C_N; both coincide in $\mathcal{H}_N$). The following result, similar to Proposition 3.2 holds:

Proposition 4.1.

(i) For any initial condition $\big(q(0), v(0)\big) \in \mathcal{H}_N \times \mathcal{H}_N$ and any $T > 0$ there exists a unique solution of (32) in the space $C^1([-T, T], \mathcal{H}_N \times \mathcal{H}_N)$.

(ii) Let $\Xi_N^t : \mathcal{H}_N \times \mathcal{H}_N \to \mathcal{H}_N \times \mathcal{H}_N$, $t \in \mathbb{R}$ denote the group flow of (32). The map Ξ_N^t is globally Lipschitz with respect to the norm induced by $\mathcal{H}^\ell \times \mathcal{H}^\ell$, with Lipschitz constant of the form $\exp(K|t|)$ where K is independent of N and depends only on $\mathcal{C}$ and Φ.

(iii) For each $T > 0$, there exists $C(T) > 0$ independent of N such that for $0 \le t \le T$ and $q(0), v(0) \in \mathcal{H}^\ell$, if we set

$$\big(q^N(t), v^N(t)\big) = \Xi_N^t\big(\mathrm{proj}_{\mathcal{H}_N} q(0), \mathrm{proj}_{\mathcal{H}_N} v(0)\big) \; ,$$

then

$$|q^N(t)|_\ell + |v^N(t)|_\ell \le C(T)\big(1 + |q^N(0)|_\ell + |v^N(0)|_\ell\big) \le C(T)\big(1 + |q(0)|_\ell + |v(0)|_\ell\big)\ , \tag{33}$$

and, for any $s \in (\ell, \kappa - 1/2)$,

$$|q^N(t) - q(t)|_\ell + |v^N(t) - v(t)|_\ell \le C(T)\Big(\frac{1}{N^{s-\ell}}\big(|q(0)|_s + |v(0)|_s\big) + \frac{1}{N}\big(1 + |q(0)|_\ell + |v(0)|_\ell\big)\Big)\ , \tag{34}$$

where $(q(t), v(t)) = \Xi^t(q(0), v(0))$ *is as specified in Proposition 3.2.*

Proof. We only derive the approximation result (34); the other statements are standard. We begin with the chain of inequalities (K will denote a constant independent of N whose value may vary from one occurrence to the next):

$$\begin{aligned}
|\mathcal{C}\, f\big(q^N(t)\big) - \mathcal{C}\,\mathrm{proj}_{\mathcal{H}_N} f\big(q^N(t)\big)|_\ell^2 &\le K|(I - \mathrm{proj}_{\mathcal{H}_N}) f\big(q^N(t)\big)|_{\ell-2\kappa}^2 \\
&\le \frac{K}{N^{4(\kappa-\ell)}}|f\big(q^N(t)\big)|_{-\ell}^2 \\
&\le \frac{K}{N^{4(\kappa-\ell)}}\big(1 + |q^N(t)|_\ell\big)^2 \\
&\le \frac{K}{N^2}\big(1 + |q^N(t)|_\ell\big)^2 \\
&\le \frac{K}{N^2}\big(1 + |q(0)|_\ell + |v(0)|_\ell\big)^2,
\end{aligned}$$

where we have used successively (30) with $s = -2\kappa$, the basic approximation inequality (35) below, the facts that (recalling Condition 3.2) $|D\Phi\big(q^N(t)\big)|_{-\ell} \le K(1 + |q^N(t)|_\ell)$ and $2(\kappa - \ell) > 1$, and finally (33). The basic approximation inequality is the fact that, for all $u \in \mathcal{H}^b$ and all $a < b$,

$$|(I - \mathrm{proj}_{\mathcal{H}_N})u|_a^2 \le \frac{1}{N^{2(b-a)}}|u|_b^2. \tag{35}$$

This may be proved by representing u in the basis $\{\phi_j\}_{j\ge1}$ and employing the definitions of the projection and norms.

Using triangle inequality and Lemma 4.1(i) we may now write

$$|\mathcal{C}f(q(t)) - \mathcal{C}\,\mathrm{proj}_{\mathcal{H}_N} f(q^N(t))|_\ell \le K\Big(|q(t) - q^N(t)|_\ell + \frac{1}{N}\big(1 + |q(0)|_\ell + |v(0)|_\ell\big)\Big)\ .$$

Subtracting the differential equations satisfied by $(q(t), v(t))$ and $(q^N(t), v^N(t))$, a standard Gronwall argument (see [24] for example) leads to

$$\begin{aligned}
&|q(t) - q^N(t)|_\ell + |v(t) - v^N(t)|_\ell \\
&\quad\le C(T)\Big(|q(0) - q^N(0)|_\ell + |v(0) - v^N(0)|_\ell + \frac{1}{N}\big(1 + |q(0)|_\ell + |v(0)|_\ell\big)\Big)
\end{aligned}$$

and (34) follows from (35) with $a = \ell$ and $b = s$. □

Clearly (32) is the Hamiltonian system associated with the following Hamiltonian function in $\mathbb{R}^{2N}$

$$H_N(q,v) = \Phi_N(q) + \frac{1}{2}\,\langle q, C_N^{-1} q\rangle + \frac{1}{2}\,\langle v, C_N^{-1}\ v\rangle$$

thus, we immediately have the following result:

Proposition 4.2. *For any $t \in \mathbb{R}$, the flow Ξ_N^t preserves the probability measure $\pi_N(dq)\pi_{0,N}(dv) = \exp\big(-H_N(q,v)\big)dqdv$.*

Note also that H_N is the restriction to $\mathcal{H}_N \times \mathcal{H}_N$ of the Hamiltonian H in (20).

We will also need to deal with the integrator of (32) and the relevant acceptance probability. By splitting (32) as we did in the case of (18), we construct mappings similar to Ψ_h and $\Psi_h^{(T)}$ in (25) and (26) respectively. The following definitions will be useful in this context.

Definition 4.1.

i) Let $\Psi_{h,N} : \mathcal{H}_N \times \mathcal{H}_N \to \mathcal{H}_N \times \mathcal{H}_N$ be as Ψ_h in (25) with the only difference that the former has $\mathcal{C}\,\mathrm{proj}_{\mathcal{H}_N} f(q)$ wherever the latter has $\mathcal{C}f(q)$ (in (23)). Also, let

$$\Psi_{h,N}^{(T)} = \Psi_{h,N}^{\lfloor \frac{T}{h} \rfloor}\ .$$

ii) Let $a_N : \mathcal{H}_N \times \mathcal{H}_N \to [0,1]$ be defined as a in (29), (28) but with $f(\cdot)$ replaced by $\mathrm{proj}_{\mathcal{H}_N} f(\cdot)$ in the latter formula (and with the q_i's, v_i's appearing in (28) now derived by applying iteratively the integrator $\Psi_{h,N}$.

The bounds in the following proposition are the discrete-time counterparts of those in Proposition 4.1:

Proposition 4.3.

(i) $\Psi_{h,N}^{(T)}$ is a globally Lipschitz map in $\mathcal{H}_N \times \mathcal{H}_N$ with respect to the norm induced by $\mathcal{H}^\ell \times \mathcal{H}^\ell$ with Lipschitz constant of the form $\exp(KT)$, where K is independent of N and depends only on $\mathcal{C}$ and Φ.

(ii) For each $T > 0$, there exists $C(T) > 0$ independent of N such that for $0 \le i \le \lfloor T/h \rfloor$, and q_0, v_0 in $\mathcal{H}^\ell$, if we set

$$(q_i^N, v_i^N) = \Psi_{h,N}^i(\mathrm{proj}_{\mathcal{H}_N} q_0, \mathrm{proj}_{\mathcal{H}_N} v_0) \tag{36}$$

then

$$|q_i^N|_\ell + |v_i^N|_\ell \le C(T)\big(1 + |q_0^N|_\ell + |v_0^N|_\ell\big) \le C(T)\big(1 + |q_0|_\ell + |v_0|_\ell\big)\ , \tag{37}$$

and, for $s \in (\ell, \kappa - \frac{1}{2})$,

$$|q_i^N - q_i|_\ell + |v_i^N - v_i|_\ell \le C(T)\Big(\frac{1}{N^{s-\ell}}\big(|q_0|_s + |v_0|_s\big) + \frac{1}{N}\big(1 + |q_0|_\ell + |v_0|_\ell\big)\Big)\ . \tag{38}$$

Proof. The convergence bound (38) is established by an argument similar to that used for (34). The role played there by Gronwall's lemma is now played by the stability of the numerical scheme, i.e. by the property in item (i). □

The integrator $\Psi_{h,N}^{(T)}$ is time-reversible and, as a composition of Hamiltonian flows, symplectic. As a consequence, it also preserves volume in $\mathbb{R}^{2N}$. In total, $\Psi_{h,N}^{(T)}$ and the acceptance probability a_N can be brought together to formulate an HMC sampling algorithm in $\mathbb{R}^N$ similar to that in Table 2.

Proposition 4.4. *The algorithm in $\mathbb{R}^N$ with proposal $(q^*, p^*) = \Psi_{h,N}^{(T)}(q^{(n)}, v^{(n)})$, where $v^{(n)} \sim N(0, C_N)$, and acceptance probability $a_N(q^{(n)}, v^{(n)})$ gives rise to a Markov chain reversible under the distribution π_N in (31).*

Proof. In view of the reversibility and volume preservation of the integrator, this algorithm corresponds to a standard HMC algorithm on Euclidean space, so the required result follows directly from known general properties of the HMC algorithm, see e.g. [9]. □

4.3. Invariance of Measures

This subsection contains the proofs of Theorems 3.1 and 3.3. The proof of Theorem 3.2 is similar and will be omitted.

Proof of Theorem 3.1. We wish to show that, for any bounded continuous function $g : \mathcal{H}^\ell \times \mathcal{H}^\ell \to \mathbb{R}$ and for any $t \in \mathbb{R}$,

$$\int_{\mathcal{H}^\ell \times \mathcal{H}^\ell} g\big(\Xi^t(q,v)\big)\Pi(dq,dv) = \int_{\mathcal{H}^\ell \times \mathcal{H}^\ell} g\big(q,v\big)\Pi(dq,dv) \ .$$

or, equivalently, that:

$$\int_{\mathcal{H}^\ell \times \mathcal{H}^\ell} g\big(\Xi^t(q,v)\big)e^{-\Phi(q)}\Pi_0(dq,dv) = \int_{\mathcal{H}^\ell \times \mathcal{H}^\ell} g\big(q,v\big)e^{-\Phi(q)}\Pi_0(dq,dv) \ . \quad (39)$$

First observe that it suffices to prove that

$$\int_{\mathcal{H}^\ell \times \mathcal{H}^\ell} g\big(\Xi_N^t(q^N,v^N)\big)e^{-\Phi(q^N)}\Pi_0(dq,dv) = \\ = \int_{\mathcal{H}^\ell \times \mathcal{H}^\ell} g\big(q^N,v^N\big)e^{-\Phi(q^N)}\Pi_0(dq,dv) \ . \quad (40)$$

This follows by dominated convergence: analytically, from Condition 3.3 the integrands $g(\Xi_N^t(q^N,v^N))\exp(-\Phi(q^N))$ and $g(q^N,v^N)\exp(-\Phi(q^N))$ are both dominated by $K\exp(\epsilon\,|q|_\ell^2)$ for some $K = K(\epsilon)$, which is integrable with respect to Π_0 by Fernique's Theorem (2.6 [8]); they also converge pointwise Π_0-a.s. to their counterparts in (39) by virtue of (34), continuity of g and continuity of Φ from Condition 3.2 (to see this, recall that Π_0 is supported in $\mathcal{H}^s \times \mathcal{H}^s$ for any $s \in (\ell, \kappa - \frac{1}{2})$, cf. Proposition 3.1).

Thus it remains to establish (40). This identity may be rewritten as

$$\int_{\mathbb{R}^N \times \mathbb{R}^N} g\big(\Xi_N^t(q,v)\big)\pi_N(dq)\pi_{0,N}(dv) = \int_{\mathbb{R}^N \times \mathbb{R}^N} g\big(q,v\big)\pi_N(dq)\pi_{0,N}(dv) \ .$$

But $\pi_N(dq)\pi_{0,N}(dv) = \exp\big(-H_N(q,v)\big)dq\,dv$ and, from Proposition 4.2, this measure is preserved by Ξ_N^t. □

We now turn to Theorem 3.3. The dynamics of the HMC Markov chain on $\mathcal{H}^\ell$ described in Table 2 correspond to the following one-step transitions:

$$q_1 = \mathbb{I}\,[\,U \le a\,]\; P_q \Psi_h^{(T)}(q_0, v_0) + \mathbb{I}\,[\,U > a\,]\, q_0\ , \tag{41}$$

where $U \sim U[0,1]$, $v_0 \sim N(0, \mathcal{C})$ and $a = 1 \wedge \exp(-\Delta\mathsf{H}(q_0, v_0))$; also P_q denotes the projection $P_q(q,v) = q$. Here $\Delta\mathsf{H}$ is given by (28) and U and v_0 are independent. Our last goal is to prove that this Markov chain is reversible under $\pi(q)$, that is:

$$\pi(dq)\, P(q, dq') = \pi(dq')\, P(dq', q) \tag{42}$$

with $P(\cdot,\cdot)$ being the transition kernel corresponding to the Markov dynamics (41). We begin with the following standard result:

Lemma 4.2. *The detailed balance equation (42) is satisfied if and only:*

$$I(g) = I(g^\top)$$

for any continuous bounded function $g : \mathcal{H}^\ell \times \mathcal{H}^\ell \to \mathbb{R}$, where:

$$I(g) = \int_{\mathcal{H}^\ell \times \mathcal{H}^\ell} g(q, P_q\, \Psi_h^{(T)}(q,v))\, a(q,v)\, e^{-\Phi(q)}\, \pi_0(dq)\pi_0(dv) \tag{43}$$

and $g^\top(q,q') := g(q',q)$.

Proof. The two probability measures in (42) are equal if and only if (see e.g. [21]):

$$\int_{\mathcal{H}^\ell \times \mathcal{H}^\ell} g(q,q')\pi(dq)\, P(q, dq') = \int_{\mathcal{H}^\ell \times \mathcal{H}^\ell} g(q,q')\pi(dq')\, P(dq', q)$$

for any continuous bounded g. In terms of expectations, this can be equivalently re-written as:

$$\mathrm{E}\,[\,g(q_0, q_1)\,] = \mathrm{E}\,[\,g^\top(q_0, q_1)\,]\ ,$$

with $q_0 \sim \pi$ and $q_1 \mid q_0$ determined via (41). Integrating out U, we get that:

$$\mathrm{E}\,[\,g(q_0, q_1)\,] = \mathrm{E}\,[\,g(q_0, P_q\, \Psi_h^{(T)}(q_0, v_0))\, a(q_0, v_0)\,] + \mathrm{E}\,[\,g(q_0, q_0)\,(1 - a(q_0, v_0))\,]\ .$$

The desired result follows from the fact that the second expectation on the right hand side will not change if we replace $g \leftrightarrow g^\top$. □

We first apply this lemma in the discretized finite-dimensional setting. Recall the definition of a_N in Definition 4.1. Taking into account the reversibility of the discretized chain with respect to π_N (Proposition 4.4) we have that, for any continuous and bounded function $\hat{g} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$:

$$\int_{\mathbb{R}^{2N}} \hat{g}(q, P_q(\Psi_{h,N}^{(T)}(q,v)))a_N(q,v)\pi_N(dq)\pi_{N,0}(dv) = \int_{\mathbb{R}^{2N}} \hat{g}^\top(q, P_q(\Psi_{h,N}^{(T)}(q,v)))a_N(q,v)\pi_N(dq)\pi_{N,0}(dv)$$

and after selecting

$$\hat{g}(q,v) = g\Big(\sum_{j=1}^{N} q_j\phi_j, \sum_{i=1}^{N} v_j\phi_j\Big)$$

we reach the conclusion

$$I_N(g) = I_N(g^\top) \tag{44}$$

where

$$I_N(g) = \int_{\mathcal{H}^\ell\times\mathcal{H}^\ell} g(q^N, P_q\,\Psi^{(T)}_{h,N}(q^N,v^N))a_N(q^N,v^N)\,e^{-\Phi(q^N)}\,\pi_0(dq)\pi_0(dv)\ . \tag{45}$$

The idea now is to conclude the proof by taking the limit $N\to\infty$ in (44) to show that $I(g) = I(g^\top)$.

Theorem 4.1. *As $N\to\infty$ then*

$$P_q\,\Psi^{(T)}_{h,N}(q^N,v^N) \to P_q\,\Psi^{(T)}_{h}(q,v)\ , \quad a_N(q^N,v^N)\to a(q,v)\ ,$$

Π_0*-almost surely.*

Proof. The first result follows directly from the bound (38), since Π_0 is concentrated in $\mathcal{H}^s$ for any $s<\kappa-1/2$. We proceed to the second result. We define q_i^N, v_i^N as in (36) and q_i, v_i as in (27), where (for both cases) now the starting positions are q and v (instead of q_0, v_0 appearing in the definitions). As a direct consequence of the definition of $a(q,v)$ and $a_N(q^N,v^N)$, to prove the required result it suffices to show the following statements are true Π_0-a.s.:

$$\Phi(q_i^N)-\Phi(q_i)\to 0\ ;$$
$$|\mathcal{C}^{\frac12}\mathrm{proj}_{\mathcal{H}_N} f(q_i^N)|^2 \to |\mathcal{C}^{\frac12} f(q_i)|^2\ ;$$
$$\langle \mathrm{proj}_{\mathcal{H}_N} f(q_i^N), v_i^N\rangle \to \langle f(q_i), v_i\rangle\ .$$

The first of these results follows directly from the continuity of Φ in Condition 3.2 and (38). For the other two limits, we observe that from the continuity properties of the involved functions in Lemma 4.1(iii) and (iv), it suffices to prove the following:

$$|\mathcal{C}^{\frac12}\mathrm{proj}_{\mathcal{H}_N} f(q_i^N)|^2 - |\mathcal{C}^{\frac12} f(q_i^N)|^2\to 0\ ;$$
$$\langle \mathrm{proj}_{\mathcal{H}_N} f(q_i^N), v_i^N\rangle - \langle f(q_i^N), v_i^N\rangle \to 0\ .$$

For the first of these, note that:

$$|\mathcal{C}^{\frac12}\mathrm{proj}_{\mathcal{H}_N} f(q_i^N)| + |\mathcal{C}^{\frac12} f(q_i^N)| \le 2|\mathcal{C}^{\frac12} f(q_i^N)| \le K\big(1+|q_0|_\ell+|v_0|_\ell\big)$$

by Lemma 4.1(iv) and (37). Now, using in succession Condition 3.1, standard approximation theory, and Condition 3.2 with (37), we obtain:

$$\big|\mathcal{C}^{\frac12}(I-\mathrm{proj}_{\mathcal{H}_N})f(q_i^N)\big| \le K\,\big|(I-\mathrm{proj}_{\mathcal{H}_N})f(q_i^N)\big|_{-\kappa} \le \frac{K}{N^{\kappa-\ell}}|f(q_i^N)|_{-\ell} \le \frac{K}{N^{\kappa-\ell}}\big(1+|q_0|_\ell+|v_0|_\ell\big)\ .$$

Since $\kappa - \ell > \frac{1}{2}$ the desired convergence follows. For the remaining limit, note that the difference can be bounded by $|(I - \mathrm{proj}_{\mathcal{H}_N})f(q_i^N)|_{-\ell}|v_i^N|_\ell$. Since $|v_i^N|_\ell$ is bounded independently of N we see that it is sufficient to show that $|(I - \mathrm{proj}_{\mathcal{H}_N})f(q_i^N)|_{-\ell} \to 0$. We note that

$$\begin{aligned}|(I - \mathrm{proj}_{\mathcal{H}_N})f(q_i^N)|_{-\ell} &\le |(I - \mathrm{proj}_{\mathcal{H}_N})f(q_i)|_{-\ell} + |f(q_i^N) - f(q_i)|_{-\ell} \\ &\le |(I - \mathrm{proj}_{\mathcal{H}_N})f(q_i)|_{-\ell} + K|q_i^N - q_i|_\ell.\end{aligned}$$

The first term goes to zero because, since $|q_i|_\ell$ is finite, Condition 3.2 shows that hence $|f(q_i)|_{-\ell}$ is finite. The second term goes to zero by (38). □

Proof of Theorem 3.3. Using Theorem 4.1 and the continuity of Φ from Condition 3.2, the integrand in $I_N(g)$ (see (45)) converges Π_0-a.s. to the integrand of $I(g)$ (see (43)). Also, for every $\epsilon > 0$, there is $K = K(\epsilon)$ such that the former integrand is dominated by $K\exp(\epsilon|q|_\ell^2)$, by Condition 3.3. Since $\pi_0(\mathcal{H}^\ell) = 1$, Fernique's theorem enables us to employ dominated convergence to deduce that

$$I_N(g) \to I(g)\ .$$

Equation (44) and Lemma 4.2 now prove the theorem. □

5. Numerical Illustrations

We present two sets of numerical experiments which illustrate the performance of the function space HMC algorithm suggested in this paper. In Subsection 5.1 we compare the new algorithm with the standard HMC method. This experiment illustrates that use of the new algorithm on high N-dimensional problems removes the undesirable N dependence in the acceptance probability that arises for the standard method. This reflects the fact that the new algorithm is well-defined on infinite dimensional Hilbert space, in contrast to the standard method. The experiment in Subsection 5.2 compares the new HMC method on Hilbert space with a Hilbert space Langevin MCMC algorithm introduced in [3]. Neither of these Hilbert space algorithms exhibit N dependence in the required number of steps, precisely because they are both defined in the limit $N \to \infty$; however the experiments show the clear advantage of the HMC method in alleviating the random walk behaviour of algorithms, such as those using Langevin proposals, which are based on local moves.

5.1. Comparison with Standard HMC

Consider the target distribution π in the space ℓ^2 of square integrable sequences (see (15)):

$$\frac{d\pi}{d\pi_0}(q) \propto \exp\{-\tfrac{1}{2}\langle q, \mathcal{C}^{-\alpha/2}q\rangle\} \tag{46}$$

with the reference Gaussian measure given by:

$$\pi_0 = N(0, \mathcal{C})\,; \quad \mathcal{C} = \mathrm{diag}\{j^{-2\kappa};\, j \ge 1\}\ .$$

Since π is itself Gaussian, with independent co-ordinates, it may be easily sampled using standard approaches. However, it provides a useful test case on which to illustrate the differences between standard HMC and our new HMC method.

We start by discussing Conditions 3.1, 3.2 and 3.3 for this problem. Intuitively they encode the idea that the reference measure π_0 dominates the change of measure and hence will involve a restriction on the size of α. Clearly, $\{j^{-2\kappa}\}$ are precisely the eigenvalues of $\mathcal{C}$ and Condition 3.1, which is necessary and sufficient for $\mathcal{C}$ to be a trace-class operator on ℓ^2, becomes

$$\kappa > 1/2 \; ; \tag{47}$$

recall also that π_0 will be concentrated in $\mathcal{H}^s$ for any $s < \kappa - 1/2$. Notice that, by (30), $\Phi(q) \asymp |q|^2_{\alpha\kappa/2}$. Thus we choose $\ell = \alpha\kappa/2$ and restrict α to satisfy $\kappa\,\alpha/2 < \kappa - 1/2$, ie.

$$\alpha < 2 - \tfrac{1}{\kappa} \; , \tag{48}$$

to ensure that $\Phi(q) < \infty$, π_0-a.s. With regard to Condition 3.2, we note that clearly Φ is continuous on $\mathcal{H}^\ell$ as a norm on this space; similarly, since

$$f(q) = -D\Phi(q) = \mathcal{C}^{-\alpha/2} q$$

it follows that $f : \mathcal{H}^\ell \to \mathcal{H}^{-\ell}$ is Lipschitz, using (30) twice. Condition 3.3 is trivially satisfied since Φ here is lower bounded. In total, specification of κ and α under the restrictions (47) and (48) places the target (46) in the general setting presented in previous sections, with all relevant conditions satisfied.

The problem is discretized by the spectral technique which we introduced in Subsection 4.2 for theoretical purposes. Because of the product structure of the target, the resulting sampling methods then correspond to applying either the standard or the new HMC as described in Tables 1 or 2 to sample from the marginal distribution of the first N co-ordinates of π:

$$\pi_N(q) \propto \exp\{-\tfrac{1}{2}\langle q, C_N^{-1} q\rangle - \tfrac{1}{2}\langle q, C_N^{-\alpha/2} q\rangle\} \tag{49}$$

where $C_N = \mathrm{diag}\{j^{-2\kappa}; j = 1, 2, \ldots, N\}$.

We applied the algorithms in Tables 1 (with mass matrix $M = C_N^{-1}$) and 2 to sample, for various choices of N, from the target distribution π_N in (49) with $\kappa = 1$, $\alpha = 1/2$. We have chosen the following algorithmic parameters: length of integration of Hamiltonian dynamics $T = 1$; discretisation increment $h = 0.2$; number of MCMC iterations $n = 5,000$. Fig. 5.1 shows empirical average acceptance probabilities from applications of the MCMC algorithms for increasing $N = 2^{10}, 2^{11}, \ldots 2^{20}$. Execution times for the Hilbert-space algorithm were about $2.5 - 3$ times greater than for the standard HMC.

For $N = 2^{10}$ the standard HMC algorithm gives an average acceptance probability of 0.89, whereas the Hilbert-space HMC gives 0.965. Thus the new integrator, with $h = 0.2$, appears to be marginally more accurate than the standard Verlet integrator. Critically, as the dimension increases, the average acceptance probability deteriorates for the standard HMC method until it eventually becomes 0. In contrast, the Hilbert-space algorithm is well-defined even

in the limit $N = \infty$ and the average acceptance probability approaches a non-zero limit as N grows. Indeed we have proved in Theorem 4.1 that the limit of the acceptance probability as $N \to \infty$ exists for the new HMC method, and this limiting behaviour is apparent in Figure 5.1. In practice, when applying the standard HMC method a user would have to use smaller h for increasing N (with $h \to 0$ as $N \to \infty$) to attain similar decorrelation to that given by the new HMC method with a fixed step-size h. The result is that the new method has smaller asymptotic variance, for given computational work; and this disparity increases with dimension.

The degeneration of the acceptance probability in standard HMC can be alleviated by applying the scaling $h = \mathcal{O}(N^{-\frac{1}{4}})$ (see [1, 20]). Heuristically this results in the need for $\mathcal{O}(h^{-1}) = \mathcal{O}(N^{\frac{1}{4}})$ steps to explore the target measure; in contrast the new HMC method requires no restriction on h in terms of N and, heuristically, explores the state space in $\mathcal{O}(1)$ steps.

We conclude this discussion with some remarks concerning choice of the mass matrix for the standard HMC method. We have given the algorithm the benefit of the choice $M = C_N^{-1}$, based on our discussion in Remark 1. This equalizes the frequencies of the Hamiltonian oscillator underlying the HMC method to one, when applied to sample the Gaussian reference measure π_0. If we had made the choice $M = I$ then the frequencies of this oscillator would have been $\{1, 2, \cdots, N\}$ (for $\kappa = 1$) resulting in the need to scale $h \propto N^{-5/4} = N^{-1} \times N^{-1/4}$ to obtain an order one acceptance probability. Intuitively the factor N^{-1} comes from a *stability* requirement[3] required to control integration of fast frequencies, whilst the factor of $N^{-1/4}$ comes (as for the choice $M = C_N^{-1}$) from an *accuracy* requirement related to controlling deviations in the energy in the large N limit. The choice $M = C_N^{-1}$ may be viewed as a *preconditioner* and the choice $M = I$ the unpreconditioned case. This viewpoint is discussed for MALA algorithms in [3] where, for the same π_0 used here, the preconditioned method requires the relevant proposal timestep Δt to be chosen as $\Delta t \propto N^{-1/3}$ whilst the unpreconditioned method requires $\Delta t \propto N^{-7/3}$.

5.2. Comparison with a MALA Algorithm

This subsection is devoted to a comparison of the new method with a MALA (Langevin based) MCMC method, also defined on the infinite dimensional space, as introduced in [3]. As the methods are both defined on Hilbert space no $N-$dependence is expected for either of them. We illustrate the fact that the HMC method breaks the random walk type behaviour resulting from the local moves used in the Langevin algorithm and can consequently be far more efficient, in terms of asymptotic variance per unit of computational cost. It is pertinent in this regard to note that the cost of implementing one step of the new HMC Markov chain from Table 2 is roughly equivalent (in fact slightly less than) the cost of T/h Langevin steps with time-step $\Delta t \propto h^2$. This is because the HMC

[3] Analogous to a Courant-Friedrichs-Lewy condition in the numerical approximation of PDEs

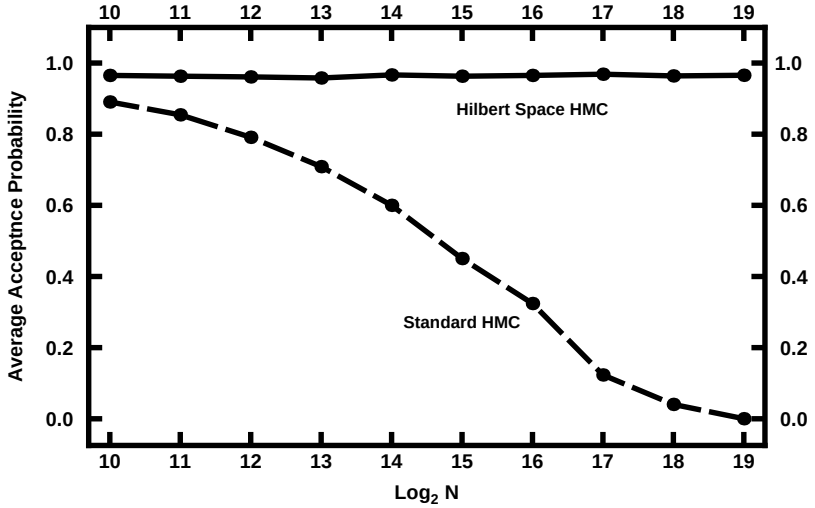

Figure 1: Empirical average acceptance probabilities corresponding to implementations of the standard and Hilbert-space HMC (with $h = 0.2$ and $T = 1$, and $n = 5,000$ iterations) with target distribution π_N in (49) (with $\kappa = 1$, $\alpha = 1/2$), for $N = 2^{10}, 2^{11}, \ldots 2^{20}$.

algorithm can be implemented by a straightforward adaptation of the Langevin code, as noted for the standard HMC and Langevin methods in [18]. This follows from the fact that using the HMC method (standard or Hilbert space versions) with one step of integration ($T = h$) is equivalent to use of the MALA method with a time-step $\Delta t \propto h^2$.

The target measure that we study is defined via a bridge diffusion. Consider the stochastic differential equation

$$dq(\tau) = -V'(q(\tau))dt + \sqrt{10}\, dW(\tau) \tag{50}$$

subject to the end-point conditions $q(0) = q(20) = 0$ and with $V(u) = \big(u^2 - 1\big)^2$. Use of the Girsanov formula, together with an integration by parts using the Itô formula, shows [3] that the resulting probability measure for $u \in L^2((0,20);\mathbb{R})$ may be written in the form (1) with π_0 Brownian bridge measure on $(0,20)$ and

$$\Phi(q) = \int_0^{20} \frac{1}{2}\Big(|V'\big(q(\tau)\big)|^2 - 10\, V''\big(q(\tau)\big)\Big)\, d\tau \ .$$

The precision operator for Brownian bridge is the second order differential operator $L = -d^2/d\tau^2$ with domain $H_0^1(I) \cap H^2(I)$ and $I = (0,20)$. The reference measure $\pi_0 = N(0, L^{-1})$ hence satisfies Condition 3.1 with $\kappa = 1$. Using the polynomial properties of Φ and Sobolev embedding it is then possible to show that Conditions 3.2 and 3.3 are satisfied for suitably chosen ℓ, and with the proviso that the Lipschitz property of the force f is only local. Because of the

symmetry of V and π_0 about the origin it is clear that π is also symmetric about the origin. Thus we may use the HMC and Langevin Markov chains to compute, via the ergodic theorem, an approximation to the mean function under π, knowing that the true mean is zero. The empirical mean we denote by $\hat{q}(\tau)$ We also compute the (signed) empirical standard deviation functions by first computing the empirical mean of the function $\left(q(\tau) - \hat{q}(\tau)\right)^2$ and then taking both positive and negative square roots.

To implement the comparison between MALA and HMC the target measure is approximated by a finite difference method employing 10^5 points in $[0, 20]$ and a value of $\Delta\tau$ given by $\Delta\tau = 2.0 \times 10^{-4}$. The HMC algorithm from Table 2 is run with a value of $h = 8.944272 \times 10^{-3}$ (leading to an acceptance rate of more than 90% for both values of T used below), and the Langevin algorithm from [3] with a value of $\Delta t = 8 \times 10^{-5}$ (leading to an acceptance rate of 78%.) For the results displayed we have used both the values $T = 3.13 \approx \pi$ and $T = 1.001$ for the HMC algorithm. Note that for $\Phi = 0$ the choice $T = \pi$ gives anti-correlated samples from the Gaussian reference measure and is natural for this reason; however the results now discussed show that choosing $T = 1$ is equally effective. We run both the MALA and Langevin algorithms for a number of steps determined so that the computational work for each is almost identical. Figures 2, 3 and 4 show the empirical mean function, together with error bar functions, computed by adding/subtracting the empirical standard deviation to these means. Comparison clearly shows the advantage of the HMC method over the Langevin method, in terms of asymptotic variance for fixed computational cost. This is primarily manifest in the empirical mean which is much closer to the true mean function 0 for both runs of the HMC algorithm than for Langevin. Furthermore, the empirical standard deviations are much closer to their true values for the HMC algorithm, than for the Langevin algorithm. The true values are shown in Figure 5 (in fact these are computed by running the HMC algorithm until it the ergodic averages have converged, around 10 times as many steps as for Figure 2).

We now examine computational efficiency in more detail. To this end we define N_t as the number of time integration steps employed before accept/reject (this is always 1 for MALA), and N_{MC} as the number of accept/reject tests. Thus the product $N_t \times N_{MC}$ is representative of the computational effort. In Figure 6 we plot E, defined as

$$E = \int_0^{20} |\hat{q}(\tau)| \, d\tau,$$

as a function of the product $N_t \times N_{MC}$. We see that the Langevin method requires considerably more computations than the HMC method to achieve convergence. Furthermore, using either $T \approx \pi$ or $T \approx 1$ in the HMC method results in comparable computational efficiency.

As with most computations by means of MCMC, caution should be applied. In particular: (i) our "true" standard deviations are themselves only computed by an MCMC method; (ii) we have chosen to study a particular test statistic

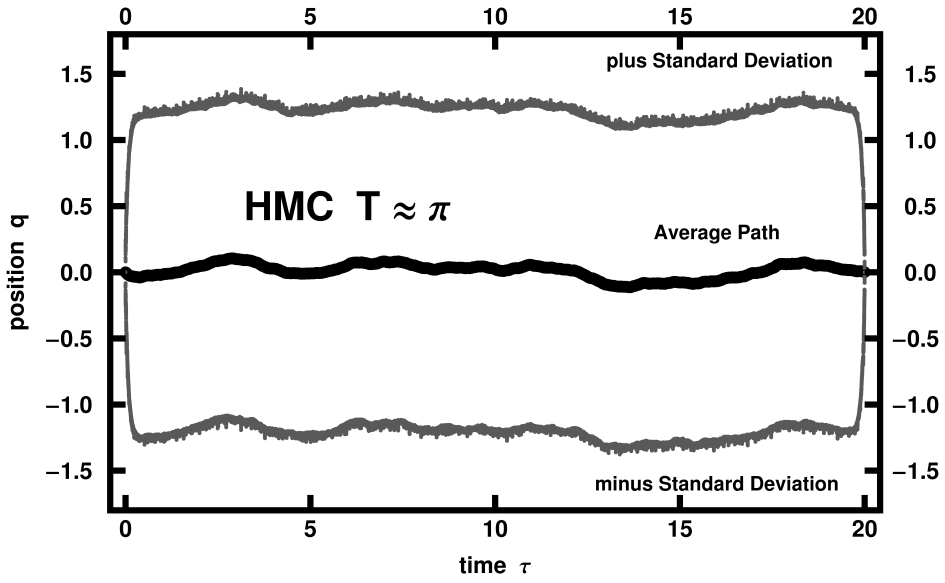

Figure 2: Empirical mean function, and empirical standard deviation functions, for Hilbert space-valued HMC algorithm, with $T \approx \pi$.

(the mean function) which possesses a high degree of symmetry so that conclusions could differ if a different experiment were chosen. However our experience with problems of this sort leads us to believe that the preliminary numerical indications do indeed show the favorable properties of the function space HMC method over the function space MALA method.

6. Conclusions

We have suggested (see Table 2) and analyzed a generalized HMC algorithm that may be applied to sample from Hilbert-space probability distributions π defined by a density with respect to a Gaussian measure π_0 as in (1). In practice the algorithm has to be applied to a discretized N-dimensional version π_N of π, but the fact that the algorithm is well defined in the limit case $N = \infty$ ensures that its performance when applied to sample from π_N does not deteriorate as N increases. In this way, and as shown experimentally in Section 5, the new algorithm eliminates a shortcoming of the standard HMC when used for large values of N. On the other hand, we have also illustrated numerically how the algorithm suggested here benefits from the rationale behind all HMC methods: the capability of taking nonlocal steps when generating proposals alleviates the random-walk behaviour of other MCMC algorithms; more precisely we have shown that the Hilbert-space HMC method clearly improves on a Hilbert-space MALA counterpart in an example involving conditioned diffusions.

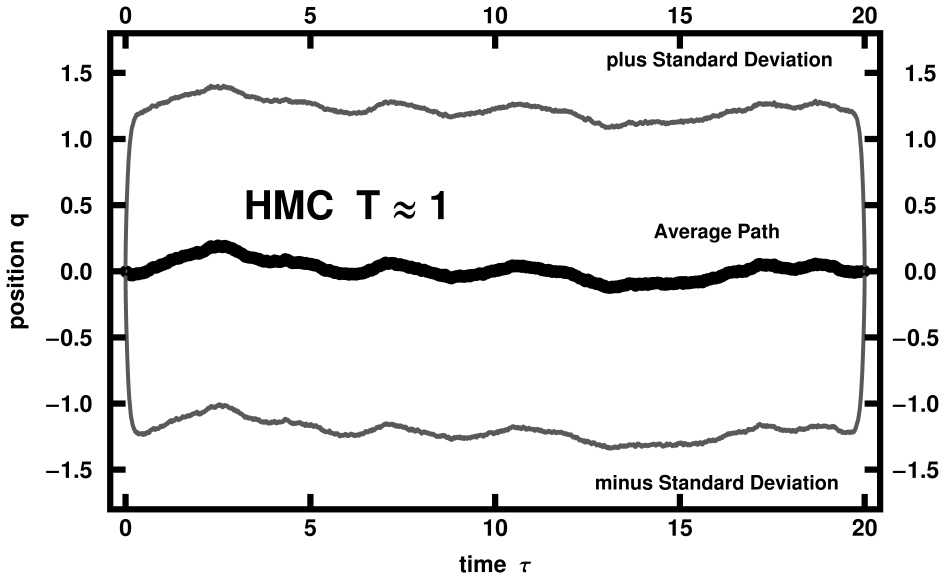

Figure 3: Empirical mean function, and empirical standard deviation functions, for Hilbert space-valued HMC algorithm, with $T \approx 1$.

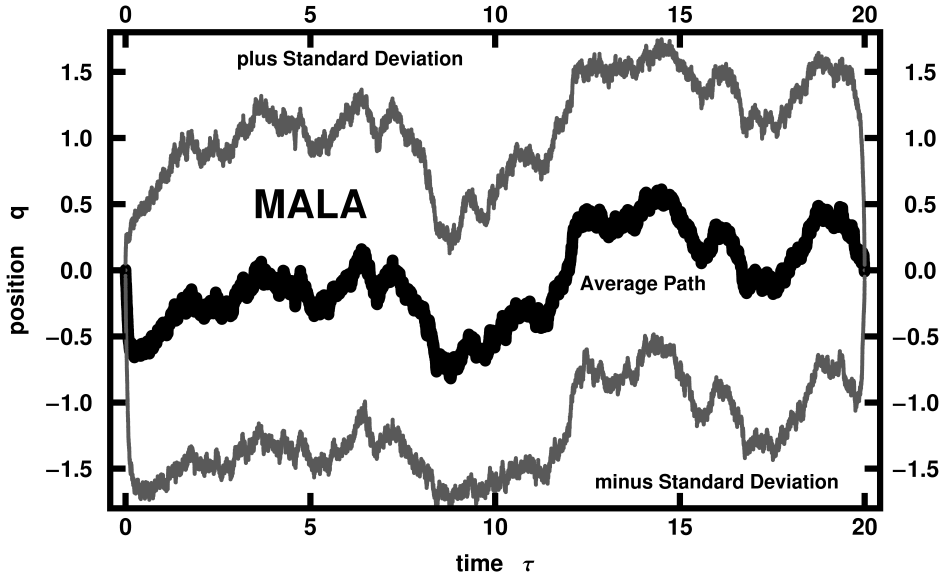

Figure 4: Empirical mean function, and empirical standard deviation functions, for Hilbert space-valued Langevin algorithm.

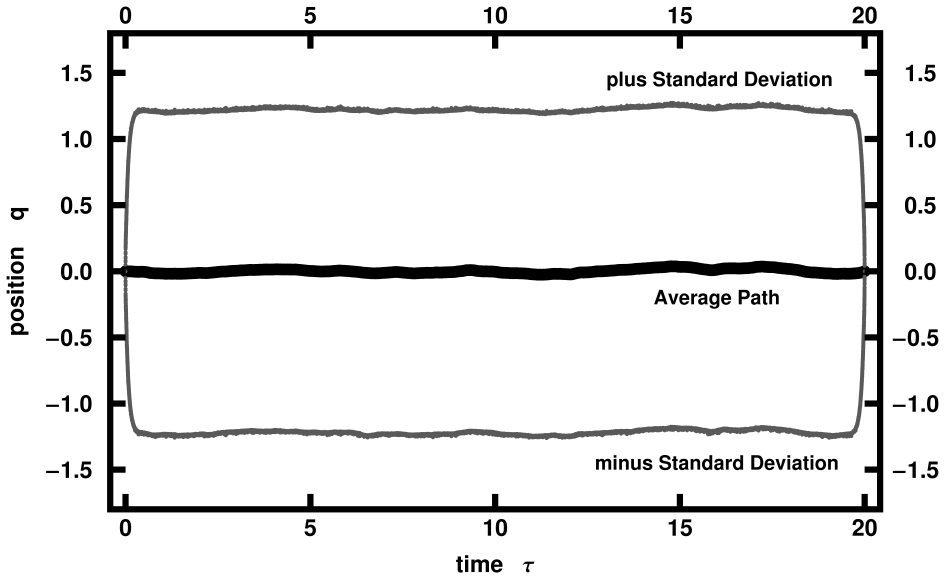

Figure 5: Mean function and standard deviation functions under the target measure.

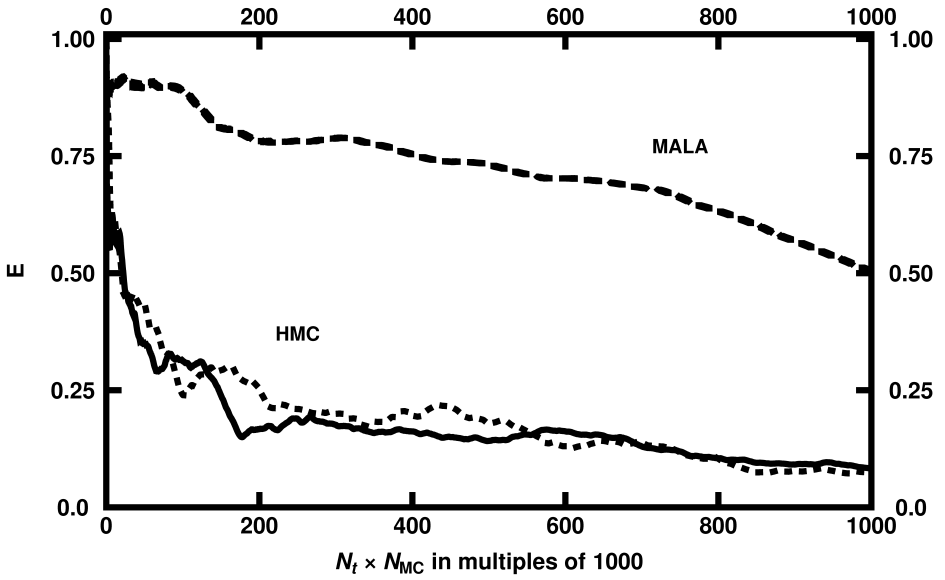

Figure 6: Comparison of the computational efficiencies of the two methods.

In order to define the algorithm we have successively addressed three issues:

- The definition of a suitable enlarged phase space for the variables q and $v = dq/dt$ and of corresponding measures Π_0 and Π having π_0 and π, respectively, as marginals on q. The probability measure Π is invariant with respect to the flow of an appropriate Hamiltonian system. Since the Hamiltonian itself is almost surely infinite under Π this result is proved by finite dimensional approximation and passage to the limit.

- A geometric numerical integrator to simulate the flow. This integrator is reversible and symplectic and, when applied in finite-dimensions, also volume-preserving. It preserves the measure Π exactly in the particular case $\Pi = \Pi_0$ and approximately in general. The integrator is built on the idea of Strang's splitting, see e.g. [25], [11]; more sophisticated splittings are now commonplace and it would be interesting to consider them as possible alternatives to the method used here.

- We have provided an accept/reject strategy that results in an algorithm that generates a chain reversible with respect to π. Here we note that straightforward generalizations of the formulae employed to accept/reject in finite dimensions are not appropriate in the Hilbert space context, as the Hamiltonian (energy) is almost surely infinite. However for our particular splitting method the energy *difference* along the trajectory is finite almost surely, enabling the algorithm to be defined in the Hilbert space setting.

There are many interesting avenues for future research opened up by the work contained herein. On the theoretical side a major open research program concerns proving ergodicity for MCMC algorithms applied to measures given by (1), and the related question of establishing convergence to equilibrium, at $N-$independent rates, for finite dimensional approximations. Addressing these questions is open for the HMC algorithm introduced in this paper, and for the Hilbert space MALA algorithms introduced in [3]. The primary theoretical obstacle to such results is that straightforward minorization conditions can be difficult to establish in the absence of a smoothing component in the proposal, due to a lack of absolute continuity of Gaussian measures with mean shifts outside the Cameron-Martin space. This issue was addressed in continuous time for the preconditioned Langevin SPDE in the paper [12], by use of the concept of "asymptotic strong Feller"; it is likely that similar ideas could be used in the discrete time setting of MCMC. We also remark that our theoretical analyses have been confined to covariance operators with algebraically decaying spectrum. It is very likely that this assumption may be relaxed to cover super-algebraic decay of the spectrum and this provides an interesting direction for further study.

There are also two natural directions for research on the applied side of this work. First we intend to use the new HMC method to study a variety of applications with the structure given in (1), such as molecular dynamics and inverse problems in partial differential equations, with applications to fluid mechanics

and subsurface geophysics. Secondly we intend to explore further enhancements of the new HMC method, for example by means of more sophisticated time-integration methods.

Appendix: Hamiltonian formalism

In this appendix we have collected a number of well-known facts from the Hamiltonian formalism (see e.g. [25]) that, while being relevant to the paper, are not essential for the definition of the Hilbert-space algorithm.

To each real-valued function $H(z)$ on the Euclidean space $\mathbb{R}^{2N}$ there correspond a *canonical* Hamiltonian system of differential equations :

$$\frac{dz}{dt} = J^{-1}\nabla_z H(z) \ ,$$

where J is the skew-symmetric matrix

$$J = \begin{pmatrix} 0_N & -I_N \\ I_N & 0_N \end{pmatrix} \ .$$

This system conserves the value of H, i.e. $H(z(t))$ remains constant along solutions of the system. Of more importance is the fact that the flow of the canonical equations preserves the *standard or canonical symplectic structure in* $\mathbb{R}^{2N}$, defined by the matrix J, or, in the language of differential forms, preserves the associated canonical differential form Ω. As a consequence the exterior powers Ω^n, $n = 2, \ldots, N$ are also preserved (Poincaré integral invariants). The conservation of the N-th power corresponds to conservation of the volume element dz. For the Hamiltonian function (3), with $z = (q, p)$, the canonical system is given by (4).

There are many *non-canonical* symplectic structures in $\mathbb{R}^{2N}$. For instance, the matrix J may be replaced by

$$\widehat{J} = \begin{pmatrix} 0_N & -L \\ L & 0_N \end{pmatrix} \ ,$$

with L an invertible symmetric $N \times N$ real matrix. Then the Hamiltonian system corresponding to the Hamiltonian function H is given by

$$\frac{dz}{dt} = \widehat{J}^{-1}\nabla_z H(z) \ .$$

Again H is an invariant of the system and there is a differential form $\widehat{\Omega}$ that is preserved along with its exterior powers $\widehat{\Omega}^n$, $n = 2, \ldots, N$. The N-th power is a constant multiple of the volume element[4] dz and therefore the standard volume is also preserved. With this terminology, the system (10) for the unknown

[4] In fact, any two non-zero $2N$-forms in $\mathbb{R}^{2N}$ differ only by a constant factor.

$z = (q, v)$, that was introduced through a change of variables in the canonical Hamiltonian system (9), is a (non-canonical) Hamiltonian system on its own right for the Hamiltonian function (11).

These considerations may be extended to a Hilbert space setting in an obvious way. Thus (18) is the Hamiltonian system in $\mathcal{H} \times \mathcal{H}$ arising from the Hamiltonian function H in (20) and the structure operator matrix

$$\widehat{J} = \begin{pmatrix} 0 & -\mathcal{C}^{-1} \\ \mathcal{C}^{-1} & 0 \end{pmatrix} .$$

However both H and the bilinear symplectic form defined by $\widehat{J}$, though densely defined in $\mathcal{H} \times \mathcal{H}$, are almost surely infinite in our context, as they only make sense in the Cameron-Martin space.

The splitting of (18) into (21) and (22) used to construct the Hilbert space integrator corresponds to the splitting

$$\mathsf{H} = \mathsf{H}_1 + \mathsf{H}_2, \quad \mathsf{H}_1(q, v) = \Phi(q), \quad \mathsf{H}_2(q, v) = \frac{1}{2} \langle v, \mathcal{C}^{-1} v \rangle + \frac{1}{2} \langle q, \mathcal{C}^{-1} q \rangle$$

of the Hamiltonian function and therefore the flows Ξ_1^t and Ξ_2^t in (23) and (24) are symplectic. The integrator Ψ_h is then symplectic as composition of symplectic mappings.

Acknowledgements The work of Sanz-Serna is supported by MTM2010-18246-C03-01 (Ministerio de Ciencia e Innovacion). The work of Stuart is suppported by the EPSRC and the ERC. The authors are grateful to the referees and editor for careful reading of the manuscript, and for many helpful suggestions for improvement.

References

[1] A. Beskos, N.S. Pillai, G.O. Roberts, J.M. Sanz-Serna, A.M. Stuart, Optimal Tuning of Hybrid Monte-Carlo, Technical Report, 2010. Submitted.

[2] A. Beskos, G.O. Roberts, A.M. Stuart, Optimal scalings for local Metropolis-Hastings chains on non-product targets in high dimensions, Ann. Appl. Probab. 19 (2009) 863–898.

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

[4] A. Beskos, A.M. Stuart, MCMC methods for sampling function space, in: R. Jeltsch, G. Wanner (Eds.), Invited Lectures, Sixth International Congress on Industrial and Applied Mathematics, ICIAM07, European Mathematical Society, 2009, pp. 337–364.

[5] A. Beskos, A.M. Stuart, Computational complexity of Metropolis-Hastings methods in high dimensions, in: P. L'Ecuyer, A.B. Owen (Eds.), Monte Carlo and Quasi-Monte Carlo Methods 2008, Springer-Verlag, 2010.

[6] V. Bogachev, Gaussian Measures, volume 62 of *Mathematical Surveys and Monographs*, American Mathematical Society, 1998.

[7] J. Bourgain, Periodic nonlinear Schrödinger equation and invariant measures, Commun. Math. Phys. 166 (1994) 1–26.

[8] G. Da Prato, J. Zabczyk, Stochastic equations in infinite dimensions, volume 44 of *Encyclopedia of Mathematics and its Applications*, Cambridge University Press, Cambridge, 1992.

[9] S. Duane, A.D. Kennedy, B. Pendleton, D. Roweth, Hybrid Monte Carlo, Phys. Lett. B 195 (1987) 216–222.

[10] M. Girolami, B. Calderhead, Riemann manifold Langevin and Hamiltonian Monte Carlo methods, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73 (2011) 123–214.

[11] E. Hairer, C. Lubich, G. Wanner, Geometric numerical integration, volume 31 of *Springer Series in Computational Mathematics*, Springer-Verlag, Berlin, second edition, 2006. Structure-preserving algorithms for ordinary differential equations.

[12] M. Hairer, A.M. Stuart, J. Voss, Analysis of spdes arising in path sampling part ii: the nonlinear case, Ann. App. Prob. 17 (2010) 1657–1706.

[13] M. Hairer, A.M. Stuart, J. Voss, Signal processing problems on function space: Bayesian formulation, stochastic pdes and effective MCMC methods, in: D. Crisan, B. Rozovsky (Eds.), Oxford Handbook of Nonlinear Filtering.

[14] A. Irback, Hybrid monte carlo simulation of polymer chains, The Journal of chemical physics 101 (1994) 1661–1667.

[15] J.L. Lebowitz, H.A. Rose, E.R. Speer, Statistical mechanics of the nonlinear schrödinger equation, J. Stat. Phys. 50 (1988) 657–687.

[16] M.A. Lifshits, Gaussian Random Functions, volume 322 of *Mathematics and its Applications*, Kluwer, 1995.

[17] J.S. Liu, Monte Carlo strategies in scientific computing, Springer Series in Statistics, Springer, New York, 2008.

[18] B. Mehlig, D.W. Heermann, B.M. Forrest, Exaxct langevin algorithms, Molecular Physics 8 (1992) 1347–1357.

[19] R.M. Neal, Probabilistic Inference Using Markov chain Monte Carlo methods, Technical Report, Department of Computer Science, University of Toronto, 1993.

[20] R.M. Neal, MCMC using Hamiltonian dynamics, in: S. Brooks, A. Gelman, G. Jones, X.L. Meng (Eds.), Handbook of Markov Chain Monte Carlo.

[21] K.R. Parthasarathy, Probability measures on metric spaces, Probability and Mathematical Statistics, No. 3, Academic Press Inc., New York, 1967.

[22] G.O. Roberts, A. Gelman, W.R. Gilks, Weak convergence and optimal scaling of random walk Metropolis algorithms, Ann. Appl. Probab. 7 (1997) 110–120.

[23] G.O. Roberts, J.S. Rosenthal, Optimal scaling of discrete approximations to Langevin diffusions, J. R. Stat. Soc. Ser. B Stat. Methodol. 60 (1998) 255–268.

[24] J.C. Robinson, Infinite Dimensional Dynamical Systems, Cambridge Texts in Applied Mathematics, Cambridge University Press, 2001.

[25] J.M. Sanz-Serna, M.P. Calvo, Numerical Hamiltonian problems, volume 7 of *Applied Mathematics and Mathematical Computation*, Chapman & Hall, London, 1994.

[26] C. Schütte, Conformational dynamics: Modelling, theory, algorithm, and application to biomolecules (1998). Habilitation Thesis, Dept. of Mathematics and Computer Science, Free University Berlin, Available at http://proteomics-berlin.de/89/.

[27] A.M. Stuart, Inverse problems: a Bayesian approach, Acta Numerica 19 (2010).