# Compressible generalized hybrid Monte Carlo

Youhan Fang, J. M. Sanz-Serna, and Robert D. Skeel

Citation: The Journal of Chemical Physics **140**, 174108 (2014); doi: 10.1063/1.4874000
View online: http://dx.doi.org/10.1063/1.4874000
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/140/17?ver=pdfcov
Published by the AIP Publishing

# Compressible generalized hybrid Monte Carlo

Youhan Fang,[1,a)] J. M. Sanz-Serna,[2,b)] and Robert D. Skeel[1,c)]
[1]*Department of Computer Science, Purdue University, Indiana 47907-2107, USA*
[2]*Departamento de Matemática Aplicada, Universidad de Valladolid, Valladolid, Spain*

(Received 27 February 2014; accepted 17 April 2014; published online 6 May 2014)

One of the most demanding calculations is to generate random samples from a specified probability distribution (usually with an unknown normalizing prefactor) in a high-dimensional configuration space. One often has to resort to using a Markov chain Monte Carlo method, which converges only in the limit to the prescribed distribution. Such methods typically inch through configuration space step by step, with acceptance of a step based on a Metropolis(-Hastings) criterion. An acceptance rate of 100% is possible in principle by embedding configuration space in a higher dimensional phase space and using ordinary differential equations. In practice, numerical integrators must be used, lowering the acceptance rate. This is the essence of *hybrid Monte Carlo* methods. Presented is a general framework for constructing such methods under relaxed conditions: the only geometric property needed is (weakened) reversibility; volume preservation is not needed. The possibilities are illustrated by deriving a couple of explicit hybrid Monte Carlo methods, one based on barrier-lowering variable-metric dynamics and another based on isokinetic dynamics. © *2014 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4874000]

## I. SUMMARY

Generating random samples from a prescribed distribution is of immense importance in the simulation of materials and in Bayesian statistics and machine learning. For example, for a protein, just a single (unbiased) sample of its structure can, with high probability, reveal its native structure.

The computational task is to generate samples from a distribution, whose probability density function (p.d.f.),

$$\rho_x(x) \propto \exp(-V(x)), \quad x = [x_1, x_2, \ldots, x_N]^{\mathsf{T}},$$

is known up to a multiplicative constant. We assume that the state variables have been scaled and rendered dimensionless. If $V(x)$ denotes energy or enthalpy, its value is in units of $k_{\mathrm{B}}T$.

To generate unbiased samples generally requires a Markov chain Monte Carlo (MCMC) method of the type proposed in the 1953 landmark paper by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (MRRTT),[1] which uses conditional acceptance to enforce unbiased sampling. Unfortunately, such methods, if based on random walk proposals, generate highly correlated samples, so many steps are required to produce effectively independent samples.

Hybrid Monte Carlo (HMC) methods, introduced in 1987[2] and generalized in 1991,[3] offer the possibility of reducing correlation between successive samples. Such methods are based on the numerical integration of ordinary differential equations (ODEs). Key requirements in their theoretical justification[4,5] are that the dynamics and its discretization must be reversible and preserve phase-space volume. Yet, there is in the literature[6] a realization that the assumption of volume conservation can be relaxed by including a Jacobian in the probability ratio of the acceptance test. Here we fully develop this idea and also weaken the reversibility assumption, providing a general framework for designing novel HMC dynamics under relaxed conditions. Not all interesting MCMC propagators are included, e.g., Ref. 7 proposes a method having a more complicated structure than those considered here.

The convergence of an MCMC method relies on two properties: stationarity and ergodicity. Presented in this article is a significant weakening of sufficient conditions for stationarity: preservation of volume in phase space is not required and reversibility is required only in the form of a bijection rather than an involution. The omission of the volume preservation requirement has been noted already for special cases.[8,9] Weaker requirements expand possibilities for designing better proposals for MCMC moves. In particular, many "purely dynamical" samplers can be made unbiased, e.g., the non-Hamiltonian sampling dynamics presented in Ref. 10. The construction of a hybrid Monte Carlo method begins with the selection of a system of ODEs in a higher dimensional phase space having an invariant probability density whose marginalization to the given state variables $x$ is the target density $\rho_x(x)$. Three examples, two novel, are given for such a construction, all of which violate phase-space volume preservation. A HMC method requires a reversible numerical integrator, but it can be difficult to find one that is efficient—unless a change of variables is effected. Given in this article is a simple general formula for changing variables, a special case occurring in Ref. 9. It is worth noting that the analytical formulations and derivations presented here require only elementary calculus (and not differential geometry[10,11]).

A familiarity with basic MCMC methods is assumed so that the ideas can be presented in an abstract style that

a)yfang@purdue.edu. URL: https://www.cs.purdue.edu/people/graduate-students/yfang/.
b)sanzsern@mac.uva.es. URL: http://www.sanzserna.org.
c)rskeel@purdue.edu. URL: http://bionum.cs.purdue.edu.

 

highlights key concepts. The mathematics requires substantial use of the chain rule: clarity is enhanced (and mistakes avoided!) by liberal use of the function-composition operator $\circ$, e.g., $f \circ g(x)$ instead of $f(g(x))$. To avoid excessive parentheses, the raised dot is used for scalar and matrix multiplication.

### A. Discussion

The extent to which the relaxed requirements can aid in the creation of better MCMC propagators remains to be seen. Yet, there are already a couple of examples from the work of others. Also included are some simple numerical experiments with results that indicate potential benefit for the generalizations developed in this article.

For molecular simulation, one of the advantages of traditional MCMC over MD (or HMC) is the ability to use a set of *large* moves tailored to the characteristics of the problems, e.g., concerted rotations of dihedral angles that deform a polymer only locally.[12] It is conceivable that ODEs can be designed that make tailored moves.

### B. Outline of paper

Section II of this article is a review of basic ideas of MCMC and especially HMC, which also serves to introduce notation. Section III presents the generalized MCMC and HMC methods. Three examples follow in Sec. IV. Discussion of (modified) detailed balance is deferred until Sec. V. It is more straightforward, it seems, to verify stationarity directly than to use detailed balance as a stepping stone. The importance of detailed balance is its role in assuring self-adjointness of the operator that propagates probabilities.

## II. MCMC METHODS

This section reviews basic ideas.

### A. Markov chain methods

Consider a Markov chain in configuration space $\mathbf{x}^0 \to \mathbf{x}^1 \to \cdots \to \mathbf{x}^n$ that samples from a distribution with unnormalized density $\exp(-V(x))$, where boldface denotes a random variable. The aim is to estimate "observables" $\langle A(\mathbf{x})\rangle = \int A(x)\rho_{\mathrm{x}}(x)\mathrm{d}x$ by means of MCMC path averages $\bar{\mathbf{A}}_n = (1/n)\sum_{k=1}^{n} A(\mathbf{x}^k)$.

In molecular simulation, at least, there are two ways to generate such Markov chains. The first is to use discretized (stochastic) differential equations. Such an approach produces a systematic error due to discretization, and for this reason, it is shunned by the statistics community. The second way to generate a Markov chain is by using a MRRTT method,[1] based on conditional acceptance. This produces no systematic error (in the limit as $n \to \infty$). Not surprisingly, such an approach tends to be less efficient, and it does not scale as well with $N$. It is a research challenge, e.g., Ref. 8, to make such methods competitive with those that unconditionally accept each move.

Convergence for a Markov chain depends on two properties:

1. *stationarity*, which means that if $\mathbf{x}^k$ has probability density $\rho_{\mathrm{x}}(x)$, so will $\mathbf{x}^{k+1}$; and
2. *ergodicity*, which means that the chain almost surely visits every set of positive probability.

Ergodicity is typically ensured by inclusion of stochastic randomness.

A *nonstandard* introduction to MRRTT methods follows, which enables the seamless inclusion of hybrid Monte Carlo as a special case.

### B. Basic Monte Carlo methods

Two examples are given of the process of generating a sample $\mathbf{x}^{k+1}$ from $\mathbf{x}^k$.

#### 1. Random walk sampler

Given a sample $\mathbf{x}$ from the target p.d.f. $\rho_{\mathrm{x}}(x)$, the next sample is determined as follows:

(1) Generate auxiliary random variables $\mathbf{y}$ from a Gaussian distribution with mean $\mathbf{x}$ and covariance matrix $2\tau I$. Hence, $\rho(x, y) = \rho_{\mathrm{x}}(x)(4\pi\tau)^{-N/2}\exp\left(-(y-x)^2/(4\tau)\right)$ serves as a joint probability density for $(\mathbf{x}, \mathbf{y})$.

(2a) Construct a proposed move: $(\mathbf{x}', \mathbf{y}') = (\mathbf{y}, \mathbf{x})$. With probability

$$\min\left\{1, \frac{\rho(\mathbf{x}', \mathbf{y}')}{\rho(\mathbf{x}, \mathbf{y})}\right\}$$

choose the next sample to be $\mathbf{x}'$.

(2b) Otherwise, choose it to be $\mathbf{x}$ again.

Due to symmetry in the proposal, the probability ratio simplifies to $\rho_{\mathrm{x}}(\mathbf{x}')/\rho_{\mathrm{x}}(\mathbf{x})$.

#### 2. Brownian sampler

This scheme aims at a higher acceptance rate by generating $\mathbf{y}$ from a Gaussian distribution with mean $\mathbf{x} - \tau\nabla V(\mathbf{x})$ and covariance matrix $2\tau I$. The limit $\tau \to 0$ yields Brownian dynamics. The basic idea is similar to the force-bias MC method[13] and almost identical to smart Monte Carlo,[14] both proposed in 1978. Extension to general densities is given in Refs. 15 and 16 and is named MALA (Metropolis-adjusted Langevin algorithm) in the latter reference. The asymmetric proposal of this scheme is an example of the Hastings[17] generalization of the MRRTT acceptance-test ratio.

Drift-diffusion dynamics is a slow way to explore configuration space. Addition of inertia remedies this. Moreover, there is a very broad set of possibilities based on approximating the flow of a dynamical system having an invariant density $\rho(x, y)$ whose marginalization to $\mathbf{x}$ is $\rho_{\mathrm{x}}(x)$. Indeed, the name "hybrid" stems from the use of *dynamics* to construct proposals.

### C. Hybrid Monte Carlo methods

With a change of auxiliary variables $y = x + \sqrt{2\tau}\,p$ in the random walk sampler, the probability density becomes (see formula (10))

$$\rho(x,\,p) = \rho_{\mathrm{x}}(x)(2\pi)^{-N/2}\exp(-p^2/2) \propto \mathrm{e}^{-H(x,p)},$$

where $H(x,\,p) = V(x) + \frac{1}{2}p^{\mathsf{T}}p$. The value $p_i$ serves as the momentum for configuration variable $x_i$. This density is preserved by a Hamiltonian flow $\Phi_\tau$ of any duration $\tau$, where the flow $\Phi_t$, $0 \le t \le \tau$ is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t(z) = f(\Phi_t(z)), \quad \Phi_0(z) = z,$$

with

$$z = \begin{bmatrix} x \\ p \end{bmatrix}, \quad f(z) = \begin{bmatrix} \nabla_p H(x,\,p) \\ -\nabla_x H(x,\,p) \end{bmatrix}.$$

To define a Hamiltonian system, all you need is an even number of variables, which can always be arranged by introducing additional variables.

Hybrid Monte Carlo is introduced in a seminal paper by Duane, Kennedy, Pendleton, and Roweth in 1987.[2] The HMC propagator is $\Psi \approx \Phi_\tau$ where $\Psi$ is composed of $\nu$ steps of a numerical integrator $\Psi_{\Delta t} \approx \Phi_{\Delta t}$ with $\Delta t = \tau/\nu$. For example, the Störmer/Verlet/leapfrog scheme is given by

$$\Psi_{\Delta t} = \Phi^{\mathrm{B}}_{\Delta t/2} \circ \Phi^{\mathrm{A}}_{\Delta t} \circ \Phi^{\mathrm{B}}_{\Delta t/2}, \tag{1}$$

where

$$\Phi^{\mathrm{A}}_t(x,\,p) = \begin{bmatrix} x + tp \\ p \end{bmatrix}, \quad \Phi^{\mathrm{B}}_t(x,\,p) = \begin{bmatrix} x \\ p - t\nabla V(x) \end{bmatrix}. \tag{2}$$

are each the exact flow of a piece of the Hamiltonian system.

Assuming a joint density $\rho(x,\,y)$ has been chosen for which $\int \rho(x,\,y)\mathrm{d}y = \rho_{\mathrm{x}}(x)$, a general recipe for HMC is the following: Let $\mathbf{x}$ be a sample from the target density $\rho_{\mathrm{x}}(x)$.

(1) Generate $\mathbf{y}$ from the marginalized distribution with density $\rho(y|\mathbf{x}) = \rho(\mathbf{x},\,y)/\int \rho(\mathbf{x},\,y)\mathrm{d}y$ and set $\mathbf{z} = [\,\mathbf{x}^{\mathsf{T}}\ \mathbf{y}^{\mathsf{T}}\,]^{\mathsf{T}}$.

(2a) Calculate a proposed move: $\mathbf{z}' = \Psi(\mathbf{z})$. With probability

$$\min\left\{1,\,\frac{\rho(\mathbf{z}')}{\rho(\mathbf{z})}\right\},$$

choose the next sample to be $\mathbf{x}'$ where $[\,(\mathbf{x}')^{\mathsf{T}}\ (\mathbf{y}')^{\mathsf{T}}\,]^{\mathsf{T}} = \mathbf{z}'$.

(2b) Otherwise, choose it to be $\mathbf{x}$ again.

For HMC to satisfy stationarity, it is sufficient that[4]

1. $\Psi$ be reversible with respect to (w.r.t.) $\rho$, meaning that there is a linear mapping $R$ such that $R \circ R = \mathrm{id}$, $\Psi^{-1} = R \circ \Psi \circ R$, and $\rho \circ R = \rho$. For classical HMC, $R(z) = [\,x^{\mathsf{T}} - p^{\mathsf{T}}\,]^{\mathsf{T}}$.
2. $\Psi$ be volume-preserving, meaning that $|\mathcal{J}\Psi| = 1$ where $\mathcal{J}\Psi = \det \partial_z \Psi$ is the Jacobian (determinant of the Jacobian matrix).

Although the leapfrog method is only second order accurate, its accuracy is difficult to beat for larger step sizes. An alternative is proposed in Ref. 18, which tunes a second order splitting to a target density that is Gaussian. Another alternative is shadow Hamiltonian HMC,[19–21] which samples from a modified Hamiltonian that is better conserved by the leapfrog integrator, but it requires a slight reweighting of the samples. An important practical issue is that of choosing the step size. See Ref. 22 for a theoretical study with references to earlier work.

### D. Generalized MCMC

In 1991, Horowitz[3] generalizes HMC by doing MCMC in phase space,

$$\mathbf{z}^0 \to \mathbf{z}^1 \to \cdots \to \mathbf{z}^n,$$

with only a partial refresh of the auxilliary variables from one step to the next. This modification allows samples to be taken more frequently: one can use a smaller duration $\tau$ without losing momentum.

Given a sample $\mathbf{x}$, $\mathbf{y}$ from the chosen joint distribution, the next sample is determined as follows:

(1) Make a random change $\mathbf{y}'$ to $\mathbf{y}$ that preserves density $\rho(\mathbf{x},\,y)$.

(2a) Calculate a proposed move $\mathbf{z}'' = \Psi(\mathbf{z}')$. With probability

$$\min\left\{1,\,\frac{\rho(\mathbf{z}'')}{\rho(\mathbf{z}')}\right\},$$

choose the next sample to be $\mathbf{z}''$.

(2b) Otherwise, choose it to be $R(\mathbf{z}')$.

The reversal substep (2b) is, of course, an unfortunate outcome, because it means going backwards on the same trajectory and only slowly diffusing away from it.

There are a range of possibilities for substep (1). At one extreme, regular HMC generates a fresh $\mathbf{y}'$ in each step; at the other extreme with $\mathbf{y}' = \mathbf{y}$, the result is a random walk along a single discrete trajectory (which is almost certain not to be ergodic, and, even if it were, convergence would be extremely slow).

#### 1. Langevin sampler

A specific example is the method L2MC of Horowitz:[3] Given a sample $\mathbf{x}$, $\mathbf{p}$, it generates $\mathbf{p}'$ from a Gaussian distribution with mean $\sqrt{1-2\gamma\tau}\mathbf{p}$ and covariance matrix $2\gamma\tau I$. A configuration space proposal is obtained from $\mathbf{z}'' = \Psi_\tau(\mathbf{z}')$ where $\Psi_\tau$ is the leapfrog integrator defined by Eqs. (1) and (2). The limit $\tau \to 0$ gives Langevin dynamics.

### E. Variable-metric hybrid Monte Carlo

The integration step size $\Delta t$ is limited by the fastest time scale. Inserting a metric tensor into the "kinetic energy" term can compress the range of time scales, resulting in faster sampling within a basin of the potential energy surface.[23] The idea

is to integrate a Hamiltonian system with Hamiltonian

$$H(x,\,p) = V(x) + \frac{1}{2}p^{\mathsf{T}}M(x)^{-1}p + \frac{1}{2}\log\det M(x).$$

This has invariant density $\propto \exp(-H(x,p))$, and the marginal density for $x$ is $\propto \exp(-V(x))$. With $M_k = (\partial/\partial x_k)M$ and with $e_k$ denoting the unit vector for the $k$th coordinate direction, one obtains the equations

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}x &= M^{-1}p,\\ \frac{\mathrm{d}}{\mathrm{d}t}p &= -\nabla_x V - \frac{1}{2}\sum_k(\mathrm{tr}(M^{-1}M_k) - p^{\mathsf{T}}M^{-1}M_kM^{-1}p)e_k,\end{aligned} \tag{3}$$

where tr denotes the trace.

The idea of mass-tensor dynamics to improve sampling goes back as far as Bennett in 1975.[24] That paper mentions but does not develop the idea of nonconstant mass tensors for sampling. It is, of course, unaware of the importance of using a geometric integrator. The special case of a constant diagonal mass matrix is explored for molecular simulation in Ref. 25.

Recently, Girolami and Calderhead[11] propose the use of mass-tensor dynamics with a symplectic integrator for hybrid Monte Carlo for Bayesian statistics. For the mass matrix, they propose the use of the expected Fisher information matrix plus the negative Hessian of the log-prior. For Bayesian inference, the expected Fisher information matrix is normally positive definite and always semi-positive definite. A numerical experiment[26] for a banana-shaped distribution illustrates the potential effectiveness of this approach for Bayesian statistics.

## III. GENERAL THEORY

Presented here are the most general conditions known to be sufficient for stationarity. Previously stated results are special cases.

### A. General MRRTT methods

The target density $\rho_x(x)$ is given, but the phase-space density $\rho(x,\,y)$ and the proposal mapping $\Psi(x,\,y)$ must be chosen. It is not assumed that $y$ has the same dimension as $x$.

It is required that $\int\rho(x,\,y)\mathrm{d}y = \rho_x(x)$. As a practical matter, it must be easy to sample $\mathbf{y}$ from the conditional density $\rho(y|\mathbf{x}) \propto \rho(\mathbf{x},\,y)$.

In practice, $\Psi(x,\,y)$ may be formally undefined for some combinations of $y$ with $x$, typically because the resulting $x'$ is outside (the designated) configuration space. Let $\Omega$ be that part of phase space on which $\Psi$ is defined. The one requirement on $\Psi$ is that of *generalized reversibility*: $\Psi$ is said to be *reversible* w.r.t. $\rho$ if

$$\Psi^{-1} = R^{-1}\circ\Psi\circ R, \tag{4}$$

where $R$ is a bijection on phase space that preserves probability, i.e.,

$$\rho\circ R = \rho/|\mathcal{J}R|, \tag{5}$$

and is such that $R\circ\Psi$ is a bijection on $\Omega$. Note that $|\mathcal{J}R| = 1$ if both $R$ is linear and $R\circ R = \mathrm{id}$. (Recall that $\mathcal{J}R = \det\partial_z R$.)

The generalization to compressible proposal mappings is enabled by including the factor $|\mathcal{J}\Psi(\mathbf{z}')|$ in the probability ratio of the acceptance test:

ALGORITHM I. Compressible generalized Monte Carlo

Let $\mathbf{z} = [\mathbf{x}^{\mathsf{T}}\quad \mathbf{y}^{\mathsf{T}}]^{\mathsf{T}}$ be given. One step of the Markov chain is as follows:

(1) Make a random change $\mathbf{y}'$ to $\mathbf{y}$ that preserves density $\rho(x,\,y')$ and set $\mathbf{z}' = [\mathbf{x}^{\mathsf{T}}\quad (\mathbf{y}')^{\mathsf{T}}]^{\mathsf{T}}$.

(2a) If $\mathbf{z}'\in\Omega$, set $\mathbf{z}'' = \Psi(\mathbf{z}')$ and accept $\mathbf{z}''$ with probability

$$\min\left\{1,\,\frac{\rho(\mathbf{z}'')}{\rho(\mathbf{z}')}|\mathcal{J}\Psi(\mathbf{z}')|\right\}.$$

(2b) If $\mathbf{z}'\notin\Omega$ or $\mathbf{z}''$ is rejected, choose $R(\mathbf{z}')$.

*Notes.* (i) It is *not* required that $\mathbf{z}\in\Omega$, because substep (1) can produce a value $\mathbf{z}'\in\Omega$. (ii) The specified density may contain Dirac delta functions. These represent conservation laws, and the propagator must enforce them exactly.

It is shown in Appendix A that the second substep of phase space HMC satisfies stationarity if the propagator is reversible. A direct demonstration of stationarity is easier than proving detailed balance, which is the topic of Sec. V.

### B. Flows

The benefit of compressibility for HMC is that the underlying continuous flow $\Phi_\tau$ can be non-Hamiltonian. The dynamics must nonetheless preserve the chosen joint density, which we express as $\rho(x,\,y) = \rho(z) \propto \exp(-H(z))$, where $H(z)$ does not necessarily play the role of a Hamiltonian. For a flow $\Phi_t$ to have $\rho$ as an invariant density, it must hold that

$$\int_{\Phi_t(A)}\rho(z)\mathrm{d}z = \int_A\rho(z)\mathrm{d}z$$

for an arbitrary set $A$. Suppose the flow $\Phi_t$ can be defined in terms of a vector field $f(z)$ by an ODE system $(\mathrm{d}/\mathrm{d}t)\Phi_t = f\circ\Phi_t$, $\Phi_0 = \mathrm{id}$. It has invariant density $\rho$ if the continuity equation $\nabla\cdot(\rho f) = 0$ is satisfied.[27] This is conveniently expressed in terms of $H(z)$ by

$$f\cdot\nabla H = \nabla\cdot f. \tag{6}$$

From Eqs. (4) and (5), it follows that the flow $\Phi_t$ is reversible w.r.t. $\rho$ if and only if there is a bijection $R$ such that

$$f\circ R = -(\partial_z R)f \tag{7}$$

and

$$H\circ R = H + \log|\mathcal{J}R|. \tag{8}$$

To obtain Eq. (7), apply $\mathrm{d}/\mathrm{d}t$ to $\Phi_t\circ R = R\circ\Phi_{-t}$, which follows from Eq. (4), and set $t = 0$. If $R$ is linear, then $(\partial_z R)f$ simplifies to $R\circ f$.

Therefore, once $\rho(z)$, and therefore $H(z) = -\log\rho(z) + \mathrm{const}$, has been chosen, one needs to find $f(z)$ and $R(z)$ that satisfy continuity equation (6) and reversibility conditions (8)/(7). The aim is to design an ODE $(\mathrm{d}/\mathrm{d}t)z = f(z)$

which explores state space rapidly and accommodates an *explicit* reversible integrator with large step sizes $\Delta t$. This is a different perspective from that of Ref. 10, where the system of ODEs rather than the invariant density is assumed to be given.

### C. Integrators from splittings

We consider the use of splitting methods to approximate the dynamics. Splitting methods are easy to design both to be reversible and to be explicit with respect to force $F(x) = -\nabla V(x)$ and energy $V(x)$. The idea is to express $f$ as a sum of terms $f^\kappa$ such that each system $(\mathrm{d}/\mathrm{d}t)z = f^\kappa(z)$ can be integrated analytically (or numerically using a reversible discretization that is explicit with respect to force and energy). A symmetric splitting gives at least second order accuracy.

For the examples that follow, assume a Trotter-Strang splitting

$$f = \frac{1}{2}f^{\mathrm{B}} + f^{\mathrm{A}} + \frac{1}{2}f^{\mathrm{B}}$$

with the proposal calculated using

$$z_{1/4} = \Phi^{\mathrm{B}}_{\Delta t/2}(z_0), \quad z_{3/4} = \Phi^{\mathrm{A}}_{\Delta t}(z_{1/4}), \quad z_1 = \Phi^{\mathrm{B}}_{\Delta t/2}(z_{3/4}).$$

Then the acceptance test ratio would be a product of factors of the form

$$\frac{\rho(\Psi_{\Delta t}(z_0))}{\rho(z_0)}\mathcal{J}\Psi_{\Delta t}(z_0) = \exp(H(z_0) - H(z_1))\mathcal{J}\Phi^{\mathrm{B}}_{\Delta t/2}(z_0) \times \mathcal{J}\Phi^{\mathrm{A}}_{\Delta t}(z_{1/4})\mathcal{J}\Phi^{\mathrm{B}}_{\Delta t/2}(z_{3/4}).$$

If the density $\rho(z)$ contains Dirac delta function(s), the associated conservation law must, in practice, be satisfied by each part of the splitting. Alternatively, the conservation law might be used to eliminate a degree of freedom. An example of each possibility is given in Secs. IV B and IV C, respectively.

If $(\mathrm{d}/\mathrm{d}t)z = f^\kappa(z)$ is integrated analytically, the exact value of $\mathcal{J}\Phi^\kappa_t(z)$ is needed. Its formula can be obtained directly from the formula for $\Phi^\kappa_t$. Alternatively, for some vector fields $f^\kappa(z)$, one can find a *compressibility integral*[27] $K^\kappa(z)$ such that $f^\kappa \cdot \nabla K^\kappa = \nabla \cdot f^\kappa$, in which case one can use

$$\mathcal{J}\Phi^\kappa_t = \exp\left(K^\kappa \circ \Phi^\kappa_t - K^\kappa\right). \tag{9}$$

This is derived as follows: Omitting the superscript $\kappa$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t = f \circ \Phi_t \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}t}\partial_z\Phi_t = (\partial_z f) \circ \Phi_t \cdot \partial_z\Phi_t.$$

Applying the Abel-Jacobi-Liouville identity gives

$$\mathcal{J}\Phi_t = \det \partial_z\Phi_t = (\det \partial_z\Phi_0)\exp\left(\int_0^t \mathrm{tr}(\partial_z f) \circ \Phi_s \,\mathrm{d}s\right) = \exp\left(\int_0^t (\nabla \cdot f) \circ \Phi_s \,\mathrm{d}s\right).$$

The result follows from $(\nabla \cdot f) \circ \Phi_t = (f \cdot \nabla K) \circ \Phi_t = (\mathrm{d}/\mathrm{d}t)K \circ \Phi_t$.

### D. Change of variables

A change of variables can facilitate numerical integration. This is illustrated in Sec. IV A with an example from Eq. (8) of Ref. 9.

Assume Eqs. (6)–(7) hold. Let $z = \zeta(w)$ be a bijection from some set $\bar{\Omega}$ to $\Omega$. Then the following relationships hold: The random variable $\mathbf{w} = \zeta^{-1}(\mathbf{z})$ has p.d.f. $\bar{\rho}(w) \propto \exp(-\bar{H}(w))$, where

$$\bar{H} = H \circ \zeta - \log|\det \partial_w\zeta|. \tag{10}$$

The transformed map

$$\bar{\Phi}_t = \zeta^{-1} \circ \Phi_t \circ \zeta \quad \text{has invariant density} \quad \bar{\rho}(w). \tag{11}$$

The map $\bar{\Phi}_t$ satisfies $(\mathrm{d}/\mathrm{d}t)\bar{\Phi}_t = \bar{f} \circ \bar{\Phi}_t$ for the vector field $\bar{f}(w)$ given by

$$\bar{f} = (\partial_w\zeta)^{-1} f \circ \zeta. \tag{12}$$

The flow $\bar{\Phi}_\tau$ is reversible w.r.t. $\bar{\rho}$ for the bijection $\bar{R}(w)$ given by

$$\bar{R} = \zeta^{-1} \circ R \circ \zeta. \tag{13}$$

Derivations for Eqs. (10)–(13) are provided in Appendix B.

## IV. THREE EXAMPLES

For the first two of these examples, it seems that viable algorithms are possible only for compressible formulations of the dynamics.

### A. Explicit variable-metric HMC

The Hamiltonian variable-metric HMC method requires solving nonlinear equations involving a variable metric tensor $M(x)$ at each time step. Motivated by Lagrangian mechanics, it is shown in Ref. 9 that solving equations can be avoided using compressible HMC.

Applying the change of variables formula of Sec. III D for $(x, p) = \zeta(x, v) = (x, M(x)v)$, one obtains

$$\bar{H}(x, v) = V(x) + \frac{1}{2}v^{\mathsf{T}}M(x)v - \frac{1}{2}\log\det M(x).$$

From

$$\partial_w\zeta = \begin{bmatrix} I & 0 \\ \sum_k M_k v e_k^{\mathsf{T}} & M \end{bmatrix}$$

and Eq. (3), one obtains the equations

$$\frac{\mathrm{d}}{\mathrm{d}t}x = v,$$
$$M\frac{\mathrm{d}}{\mathrm{d}t}v = -\nabla_x V - \frac{1}{2}\sum_k (\mathrm{tr}(M^{-1}M_k) - v^{\mathsf{T}}M_k v)e_k - \sum_k v_k M_k v, \tag{14}$$

which agrees with Eq. (8) of Ref. 9.

With $(\mathrm{d}/\mathrm{d}t)x = 0$ in the equations for part B of the splitting, this leaves quadratic differential equations to solve for $v$. These are discretized[9] using a symmetric linearly implicit scheme of Kahan,[28,29] based on geometric averaging of terms

quadratic in $v$. This requires solving one linear system of equations per time step, instead of iteratively solving a nonlinear system at each time step.

### 1. Barrier-lowering dynamics

The effect of decreasing mass in a small region $X$ of configuration space is to increase speed across that region. Since the probability of being in $X$ is unchanged, the region $X$ must be crossed more often. This principle can be used to increase barrier crossings. Let $x^0$ be the approximate location of a saddle point and $m$, $m^{\mathsf{T}}m = 1$, the direction of negative curvature. Specifically, assume the Hessian of $V(x)$ at $x = x^0$ has a single negative eigenvalue with corresponding eigenvector $m$. Define

$$M(x) = \mu(x)^{N-1} mm^{\mathsf{T}} + \mu(x)^{-1}(I - mm^{\mathsf{T}}),$$

where

$$\mu(x) = 1 - (1 - \mu_0)\exp\left(-\frac{\|x - x^0\|^2}{2r^2}\right)$$

localizes the effect to a neighborhood of $x^0$. Here $r$ is perhaps the distance to an inflection point and $\log(1/\mu_0^2)$ is some fraction of the barrier height. The "tunnel metric" proposed independently in Ref. 30 is similar. However, it uses a switching function that is different in form and defined by two points rather than just $x^0$. And it lacks the normalization $\det M(x) = 1$, which helps to reduce integration error.

To handle several such barriers, an elegant approach is to use

$$M = \exp\left(\sum_{\kappa} \log M^{\kappa}\right),$$

which, in practice, would be replaced by a Trotter or Trotter-Strang splitting.

The method is tested on a high-dimensional mixture of two Gaussians. Specifically, the first dimension is the average of two Gaussians located at $\pm 2.5$, and the other 128 dimensions are all Gaussians. Hence, the components of $\mathbf{x}$ are independent. The standard deviation of each Gaussian in the first dimension is 1, and the standard deviations in other dimensions are uniformly distributed from 1 to 2. This is the simplest problem that features high dimensionality and multiple modes and is not especially sensitive to the duration $\tau$ of the dynamics run. (And it can be implemented very efficiently.) The observable is a sigmoid function of the first dimension, $A(x) = 1/(1 + \exp(-x_1))$, which is chosen to mimic the second eigenfunction shown in Fig. 1 of Ref. 31. The step size $\Delta t = 0.5$, the number of integration steps $\nu = 10$, and the number of samples $n = 10^6$. These many samples give an estimated standard deviation of 3 for the mean value of the auto-correlation time for HMC with a MC step of duration $\tau = 5$ and a number of integrator steps $\nu = 10$. The auto-correlation time, computed by *Acor*,[32] gives the ratio of the sample size to the *effective sample size*. Results are displayed in Figures 1 and 2 for $m$ in the direction of the $x_1$-axis, $x_0 = 0$, $r = 1$, and a range of values of the minimum mass $\mu_0^{N-1}$. For a fairly wide range of values of $\mu_0^{N-1}$, the barrier lowering variable mass method produces about 100% more effective samples.

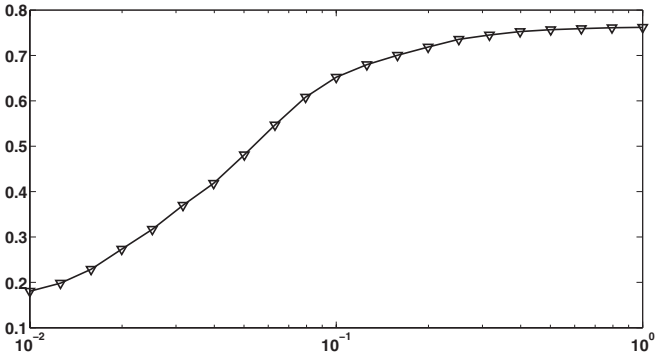


FIG. 1. x-axis is the minimum mass and y-axis is the acceptance probability.

## B. Isokinetic HMC

The equations for isokinetic dynamics are

$$\frac{\mathrm{d}}{\mathrm{d}t}x = \frac{N-1}{N}p, \quad \frac{\mathrm{d}}{\mathrm{d}t}p = \left(I - \frac{pp^{\mathsf{T}}}{p^{\mathsf{T}}p}\right)F,$$

where $F = -\nabla_x V$. With masses thus chosen, the density

$$\rho(x, p) \propto \exp(-NV(x)/(p^{\mathsf{T}}p))\delta(p^{\mathsf{T}}p - N)$$

is invariant and $\langle \mathbf{p}_i^2 \rangle = 1$. For the purpose of satisfying Eq. (6), $\delta$ may be taken to be a mollified delta function.[33] Isokinetic dynamics is intended for sampling from the canonical ensemble $\mathrm{e}^{-V(x)}$.[34] There is current interest in isokinetic dynamics because it is less sensitive to resonances induced by multiple time-stepping.[35,36]

A convenient splitting is the following:

$$f^{\mathrm{A}} = \begin{bmatrix} ((N-1)/N)p \\ 0 \end{bmatrix}, \quad f^{\mathrm{B}} = \begin{bmatrix} 0 \\ F - ((p^{\mathsf{T}}F)/(p^{\mathsf{T}}p))p \end{bmatrix}.$$

To satisfy stationarity, $\Psi_{\Delta t}$ must conserve kinetic energy exactly, which is true if the flow $\Phi_t^{\mathrm{B}}$ is obtained analytically.

To solve the equations $(\mathrm{d}/\mathrm{d}t)z = f^{\mathrm{B}}(z)$, introduce scalar quantities $\xi = (F^{\mathsf{T}}F)^{1/2}$, $\zeta = (p^{\mathsf{T}}p)^{1/2}$, and $\eta = F^{\mathsf{T}}p/(\xi\zeta)$. The equations then reduce to solving

$$\frac{\mathrm{d}}{\mathrm{d}t}\eta = (\xi/\zeta)(1 - \eta^2),$$

where $\xi$ and $\zeta$ are constant. The solution is

$$\eta(t) = \frac{\eta(0)\cosh(\xi t/\zeta) + \sinh(\xi t/\zeta)}{\cosh(\xi t/\zeta) + \eta(0)\sinh(\xi t/\zeta)},$$

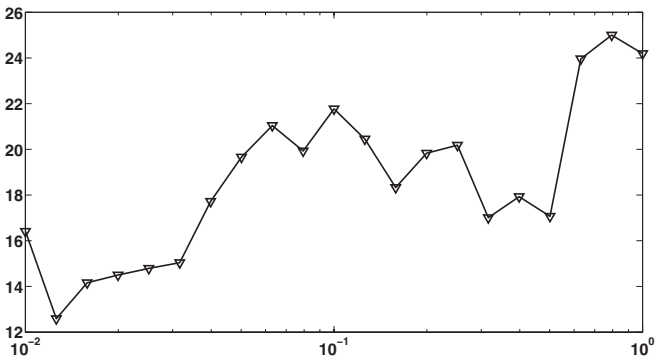


FIG. 2. x-axis is the minimum mass and y-axis is the auto-correlation time.

TABLE I. Effective number of samples per 1000 force evaluations for Hamiltonian HMC.

| | $\nu$ | | | |
|---|---|---|---|---|
| $\tau$ | 6 | 8 | 10 | 12 |
| 4 | 1.44 | 2.51 | 3.03 | 2.55 |
| 5 | 2.04 | 4.41 | 4.13 | 3.65 |
| 6 | ... | 1.52 | 3.42 | 3.19 |

TABLE II. Effective number of samples per 1000 force evaluations for isokinetic HMC.

| | $\nu$ | | | |
|---|---|---|---|---|
| $\tau$ | 6 | 8 | 10 | 12 |
| 4 | 3.16 | 3.52 | 2.81 | 3.20 |
| 5 | 3.83 | 4.52 | 4.91 | 4.84 |
| 6 | 0.72 | 4.11 | 3.60 | 3.29 |

which can be written

$$\frac{\xi}{\zeta}\eta(t) = \frac{1}{\sigma(t)}\frac{\mathrm{d}\sigma}{\mathrm{d}t}(t),$$
$$\text{where} \quad \sigma(t) = \cosh(\xi t/\zeta) + \eta(0)\sinh(\xi t/\zeta).$$

From this, it can be shown that

$$p(t) = \frac{1}{\sigma(t)}\left(p(0) + \frac{\xi^2}{\zeta^2}\left(\frac{\mathrm{d}\sigma}{\mathrm{d}t}(t) - \frac{\mathrm{d}\sigma}{\mathrm{d}t}(0)\right)F\right),$$

which is equivalent to the formula given in Sec. IV B of Ref. 35.

The compressibility $\nabla\cdot f^{\mathrm{A}} = 0$, but $\nabla\cdot f^{\mathrm{B}} = -(N-1)\,\xi\eta/\zeta$. The compressibility $\nabla\cdot f^{\mathrm{B}}$ has an integral $K^{\mathrm{B}} = \frac{1}{2}(N-1)\log(1-\eta^2)$; whence, using Eq. (9), we have $\mathcal{J}\Phi_t^{\mathrm{B}}(z(0)) = ((1-\eta(t)^2)/(1-\eta(0)^2))^{(N-1)/2} = (\sigma(0)/\sigma(t))^{N-1}$.

Isokinetic dynamics can be transformed to make it divergence-free, see Ref. 27. However, this involves placing the prefactor $\exp(-(N-1)V(x)/(p^{\mathsf{T}}p))$ into the equations of motion, for which there seems to be no reversible integrator explicit in force and energy.

#### 1. Numerical experiments

Using the same test problem as in Sec. IV A 1, Hamiltonian HMC is compared to isokinetic HMC, in both cases with unit masses. Given in Tables I and II are the estimated effective sample sizes per 1000 force evaluations for each, for varying values of the duration $\tau$ of a MC step and the number of integrator steps $\nu$ per MC step. A pair of parameter values that maximizes the number of effective samples per integrator step is in the center of each table. A dash indicates a failure of all proposed moves. The isokinetic method performs a little better than the Hamiltonian method and *is less sensitive to tuning parameters.*

### C. Nosé-Hoover thermostat

The Nosé-Hoover thermostat employs the extended system

$$\frac{\mathrm{d}}{\mathrm{d}t}x = p, \quad \frac{\mathrm{d}}{\mathrm{d}t}p = F - \frac{\pi}{Q}p, \quad \frac{\mathrm{d}}{\mathrm{d}t}s = \frac{\pi}{Q},$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\pi = p^{\mathsf{T}}p - N.$$

The invariant density is

$$\rho(z,s) \propto \mathrm{e}^{Ns}\delta(H(z) + Ns),$$

where $H(z) = \frac{1}{2}p^{\mathsf{T}}p + V(x) + \pi^2/(2Q) - E$, $Q$ is "thermal mass," and $E$ is some constant. The vector $z$ consists of $x$, $p$, and a scalar $\pi$. Again, time is normalized so that $\langle \mathbf{p}_i^2\rangle = 1$.

The quantity $H(z) + Ns$ is conserved. However, there seems to be no practical numerical integrator that conserves this quantity exactly. Therefore, we eliminate the equation for $(\mathrm{d}/\mathrm{d}t)s$, which is possible since no other variable depends on it, to get a system with invariant density $\rho(z) \propto \exp(-H(z))$.

Leimkuhler and Reich[8] develop a compressible generalized HMC version of this scheme, in which only the thermostat variable $\pi$ is partially refreshed, and they prove stationarity. Their aim is to show the feasibility of introducing conditional acceptance into stochastic sampling dynamics. For an integrator, they use a compact splitting of Ref. 37:

$$f^{\mathrm{A}} = \begin{bmatrix} p \\ 0 \\ p^{\mathsf{T}}p - N \end{bmatrix}, \quad f^{\mathrm{B}} = \begin{bmatrix} 0 \\ F - \pi p/Q \\ 0 \end{bmatrix}$$

(p. 746 of Ref. 8). The system $(\mathrm{d}/\mathrm{d}t)z = f^{\mathrm{B}}(z)$ has nonzero compressibility $\nabla\cdot f^{\mathrm{B}} = -N\pi$. Obtaining the Jacobian directly gives $\mathcal{J}\Phi_t^{\mathrm{B}} = \exp(-N\pi t/Q)$ (from which one can obtain the compressibility integral $K^{\mathrm{B}} = N\log\|F - \pi p/Q\|$).

There are a couple of ways of making the Nosé-Hoover scheme divergence-free, see Refs. 8, 10, and 27, but in either case we are left with a system having a conservation law whose exact enforcement seems impractical.

## V. MODIFIED DETAILED BALANCE

Many compressible MCMC schemes satisfy not only stationarity but also a modified detailed balance property. The benefit of this is that the operator that propagates probabilities is self-adjoint, making it easier to analyze and understand the behavior of the Markov chain. A Markov process satisfies *(modified) detailed balance* if

$$\rho(z'|R(z))\rho(z) = \rho(z|R(z'))\rho(z'), \tag{15}$$

where $\rho(z'|z)$ is the conditional p.d.f. for a step. (The definition given in Ref. 38 is recovered by replacing $z$ by $R(z)$ and making the assumptions $R\circ R = \mathrm{id}$ and $\rho\circ R = \rho$.)

It is easy to show that modified detailed balance implies stationarity: Replace $z$ by $R(z)$ and make use of Eq. (5) to get

$$\int \rho(z'|z)\rho(z)\mathrm{d}z = \int \rho(z'|R(z))\rho(z)\mathrm{d}z.$$

The result follows from applying Eq, (15) and using the fact that

$$\int \rho(z|R(z'))\mathrm{d}z = 1.$$

### A. The forward transfer operator and self-adjointness

The forward transfer operator (or propagator) $\mathcal{P}$ relates the *relative p.d.f.* $u_n(z) = \rho_n(z)/\rho(z)$ of a sample $\mathbf{z}_n$ to that of its predecessor $\mathbf{z}_{n-1}$:

$$\rho_n/\rho = u_n = \mathcal{P}u_{n-1} = \mathcal{P}(\rho_{n-1}/\rho).$$

Hence,

$$\mathcal{P}u(z) = \frac{1}{\rho(z)}\int \rho(z|z')u(z')\rho(z')\mathrm{d}z',$$

where $\rho(z|z')$ is the transition probability density.

It happens that the forward transfer operator $\mathcal{P}$ is *self-adjoint* with respect to the $\rho$-weighted inner product

$$(u, v) = \int u(z)v(z)\rho(z)\mathrm{d}z,$$

defined on the space of functions for which $u(R(z)) = u(z)$ if the Markov process satisfies modified detailed balance. We have

$$(\mathcal{P}u, v) = \iint \rho(z|z')u(z')\rho(z')v(z)\mathrm{d}z'\mathrm{d}z.$$

If we replace $z'$ by $R(z')$ and make use of Eq. (5), we get

$$(\mathcal{P}u, v) = \iint \rho(z|R(z'))u(z')\rho(z')v(z)\mathrm{d}z'\mathrm{d}z.$$

Similarly, we have

$$(\mathcal{P}v, u) = \iint \rho(z'|R(z))v(z)\rho(z)u(z')\mathrm{d}z'\mathrm{d}z,$$

where we have interchanged not only $u$ and $v$ but also the names of the dummy variables. Equality, $(\mathcal{P}u, v) = (u, \mathcal{P}v)$, follows from Eq. (15).

### B. Generalized HMC (GHMC) and detailed balance

*Generalized HMC does not formally satisfy detailed balance.* To illustrate why this is so, consider the case of a Langevin sampler for $V(x) = x^2/2$, $2\gamma\tau = 1$, $\tau = \pi/2$, but with an exact integrator:

Given scalars $x$ and $p$,

(1) replace $p$ with a sample $p'$ from a Gaussian distribution, and
(2) set $x'' = x + p'$ and $p'' = p' - x$.

The first substep is a random move in the direction of the p axis; the second is a 90° clockwise rotation in the x-p plane. For most pairs $(x'', p'')$ generated this way, it is *impossible* to get back to $(x, p)$ in a single MCMC step.

Generally, to ensure detailed balance for generalized HMC, there should be a third substep that is the *time reversal*[39] of the process given by the first substep. The reversal of a process with conditional p.d.f. $\rho(z'|z)$ has a conditional p.d.f. $\hat{\rho}(z'|z) = \rho(z|z')\rho(z')/\rho(z)$. Specifically, it is verified in Appendix C that a complete step of a 3-substep MCMC scheme satisfies modified detailed balance if

1. The substep1 process has a conditional p.d.f. $\rho_1(z'|z)$ that satisfies

$$\rho_1(R(z')|R(z)) = \rho_1(z'|z)/|\mathcal{J}R(z')|. \tag{16}$$

2. The middle substep satisfies detailed balance.
3. The substep3 process has a conditional p.d.f.

$$\rho_3(z'|z) = \rho_1(z|z')\rho(z')/\rho(z).$$

To ensure that a *GHMC proposal substep* satisfies detailed balance, make the additional assumption that $R$ is an involution:

$$R \circ R = \mathrm{id}.$$

Under this assumption, it is shown in Appendix D that the second substep of phase space HMC satisfies modified detailed balance.

The Langevin sampler can be made to formally satisfy detailed balance by splitting the initial substep into initial and final substeps, each of which multiplies the momenta by $(1 - 2\gamma\tau)^{1/4}$ and adds independent Gaussians of mean 0 and variance $(1 - 2\gamma\tau)^{1/2} - 1$. This is, of course, equivalent to the original method, showing that GHMC effectively satisfies modified detailed balance.


## ACKNOWLEDGMENTS

A precursor to this investigation was an exploration of the applicability of variable-metric HMC[11] to particle simulations. We would like to thank Mari Paz Calvo for her contribution in bringing that investigation to an end. The work of J.M.S.-S. is supported by Project MTM2010-18246-C03-01, Ministerio de Ciencia e Innovación, Spain.


## APPENDIX A: DEMONSTRATION OF STATIONARITY

Here the variables have one less prime than those of Algorithm 1 in Sec. III A. The conditional probability density for substep (2) is

$$\rho(z'|z) = \theta(z)\delta(z' - \Psi(z)) + (1 - \theta(z))\delta(z' - R(z)), \tag{A1}$$

where

$$\theta(z) = \min\left\{1, \frac{\rho(\Psi(z))}{\rho(z)}|\mathcal{J}\Psi(z)|\right\},$$

if $z \in \Omega$ and $\theta(z) = 0$ otherwise. It must be shown that $\int \rho(z'|z)\rho(z)\mathrm{d}z = \rho(z')$. We have

$$\begin{aligned}
&\int \rho(z'|z)\rho(z)\mathrm{d}z \\
&\quad = \int_\Omega \min\{\rho(z), \rho(\Psi(z))|\mathcal{J}\Psi(z)|\}\,\delta(z' - \Psi(z))\mathrm{d}z \\
&\qquad - \int_\Omega \min\{\rho(z), \rho(\Psi(z))|\mathcal{J}\Psi(z)|\}\,\delta(z' - R(z))\mathrm{d}z \\
&\qquad + \int \delta(z' - R(z))\rho(z)\mathrm{d}z.
\end{aligned}$$

Make the substitution $z = R \circ \Psi(\bar{z})$ in the first integral and drop the bars:

$$\begin{aligned}\text{1st term} &= \int_\Omega \min\{\rho\circ R\circ\Psi(z),\, \rho\circ R(z)|(\mathcal{J}\Psi)\circ R\circ\Psi(z)|\}\\ &\quad\times \delta(z' - R(z))|\mathcal{J}(R\circ\Psi)(z)|\mathrm{d}z\\ &= \int_\Omega \min\{\rho\circ R\circ\Psi(z)|\mathcal{J}(R\circ\Psi)(z)|,\\ &\quad \rho\circ R(z))|\mathcal{J}(\Psi\circ R\circ\Psi)(z)|\}\delta(z' - R(z))\mathrm{d}z.\end{aligned}$$

Using Eq. (5),

$$\begin{aligned}\text{1st term} &= \int_\Omega \min\{\rho\circ\Psi(z)|\mathcal{J}(R\circ\Psi)(z)/(\mathcal{J}R)\circ\Psi(z)|,\\ &\quad \rho(z)\}\delta(z' - R(z))\mathrm{d}z,\end{aligned}$$

which cancels the 2nd term. Also,

$$\begin{aligned}\text{3rd term} &= \int \delta(z' - z)\rho\circ R^{-1}(z)|\mathcal{J}(R^{-1})(z)|\mathrm{d}z\\ &= \rho(z')|(\mathcal{J}R)\circ R^{-1}(z')||\mathcal{J}(R^{-1})(z')| = \rho(z').\end{aligned}$$

## APPENDIX B: DERIVATION OF FORMULAS FOR CHANGE OF VARIABLES

Let $A$ be an arbitrary subset of $\bar{\Omega}$. Then

$$\begin{aligned}\int_A \bar{\rho} &= \Pr(\mathbf{w}\in A) = \Pr(\zeta^{-1}(\mathbf{z})\in A)\\ &= \int_{\zeta(A)}\rho = \int_A \rho\circ\zeta\,|\det\partial_w\zeta|,\end{aligned}$$

and Eq. (10) follows from equating integrands and taking their negative logarithms.

To see (11), observe that

$$\begin{aligned}\Pr(\mathbf{w}\in\bar{\Phi}_t(A)) &= \Pr(\zeta^{-1}(\mathbf{z})\in\bar{\Phi}_t(A)) = \Pr(\mathbf{z}\in\Phi_t(\zeta(A)))\\ &= \Pr(\mathbf{z}\in\zeta(A)) = \Pr(\zeta^{-1}(\mathbf{z})\in A)\\ &= \Pr(\mathbf{w}\in A).\end{aligned}$$

To verify Eq. (12), observe that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\bar{\Phi}_t &= \partial_z(\zeta^{-1})\circ\Phi_t\circ\zeta\,\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t\circ\zeta\\ &= (\partial_w\zeta\circ\zeta^{-1}\circ\Phi_t\circ\zeta)^{-1}\,f\circ\Phi_t\circ\zeta\\ &= (\partial_w\zeta\circ\bar{\Phi}_t)^{-1}\,f\circ\zeta\circ\bar{\Phi}_t.\end{aligned}$$

To establish formula (13), we need to verify Eqs. (8) and (7) for $\bar{H}$, $\bar{f}$, and $\bar{R}$. For Eq. (7),

$$\begin{aligned}RHS &= -\partial_w(\zeta^{-1}\circ R\circ\zeta)\cdot(\partial_w\zeta)^{-1}f\circ\zeta\\ &= -\partial_z(\zeta^{-1})\circ R\circ\zeta\cdot\partial_z R\circ\zeta\cdot f\circ\zeta\\ &= (\partial_w\zeta)^{-1}\circ\zeta^{-1}\circ R\circ\zeta\cdot f\circ R\circ\zeta = LHS.\end{aligned}$$

For Eq. (8), its left-hand side would be

$$\begin{aligned}LHS &= H\circ R\circ\zeta - \log|\mathcal{J}\zeta\circ\zeta^{-1}\circ R\circ\zeta|\\ &= H\circ\zeta + \log|\mathcal{J}R\circ\zeta| - \log|\mathcal{J}\zeta\circ\zeta^{-1}\circ R\circ\zeta|.\end{aligned}$$

The right-hand side would be

$$RHS = H\circ\zeta - \log|\mathcal{J}\zeta| + \log|\mathcal{J}(\zeta^{-1}\circ R\circ\zeta)|.$$

We have

$$\mathcal{J}(\zeta^{-1}\circ R\circ\zeta) = (\mathcal{J}\zeta\circ\zeta^{-1}\circ R\circ\zeta)^{-1}\mathcal{J}R\circ\zeta\cdot\mathcal{J}\zeta,$$

whence $LHS = RHS$.

## APPENDIX C: DEMONSTRATION OF MODIFIED DETAILED BALANCE FOR A 3-SUBSTEP SCHEME

We have

$$\rho(z'|z) = \iint \rho_3(z'|w')\rho_2(w'|w)\rho_1(w|z)\mathrm{d}w\mathrm{d}w',$$

where $\rho_3(z'|w') = \rho_1(w'|z')\rho(z')/\rho(w')$. It is enough to show

$$\frac{\rho(z'|R(z))}{\rho(z')} = \frac{\rho(z|R(z'))}{\rho(z)}.$$

The left-hand side

$$\begin{aligned}LHS &= \iint \rho_1(w'|z')\rho_1(w|R(z))\frac{\rho_2(w'|w)}{\rho(w')}\mathrm{d}w\mathrm{d}w'\\ &= \iint \rho_1(w'|z')\rho_1(R(w)|R(z))\frac{\rho_2(w'|R(w))}{\rho(w')}\\ &\quad\times|\mathcal{J}R(w)|\mathrm{d}w\mathrm{d}w'\\ &= \iint \rho_1(w'|z')\rho_1(w|z)\frac{\rho_2(w|R(w'))}{\rho(w)}\mathrm{d}w\mathrm{d}w',\end{aligned}$$

where the final step uses Eq. (16). The right-hand side $RHS$ is this same expression, but with $z$ and $z'$ interchanged. Equality, $RHS = LHS$, follows by interchanging $w$ and $w'$ in the expression for $RHS$ and applying modified detailed balance (Eq. (15)).

## APPENDIX D: DEMONSTRATION OF MODIFIED DETAILED BALANCE FOR A GHMC PROPOSAL

We show that the second substep of compressible HMC satisfies modified detailed balance if the propagator is reversible and the reversing map $R$ is an involution. First, consider the case where $z\notin\Omega$, $z'\notin\Omega$. We demonstrate detailed balance, Eq. (15) separately for each term of $\rho(z'|z)$ given by Eq. (A1). This means showing

$$\theta(R(z))\delta(z' - \Psi(R(z)))\rho(z) = \theta(R(z'))\delta(z - \Psi(R(z')))\rho(z') \tag{D1}$$

and

$$(1-\theta(R(z)))\delta(z'-z)\rho(z) = (1-\theta(R(z')))\delta(z-z')\rho(z').$$

The second of these obviously holds. To show the first of these, we write

$$\begin{aligned}\delta(z - \Psi(R(z'))) &= \delta(\Psi\circ R\circ\Psi\circ R(z) - \Psi\circ R(z'))\\ &= \delta(\partial_z(\Psi\circ R)(z')\cdot(\Psi\circ R(z) - z'))\\ &= |\mathcal{J}(\Psi\circ R)(z')|^{-1}\delta(z' - \Psi\circ R(z))\\ &= |\mathcal{J}(\Psi\circ R)\circ\Psi\circ R(z)|^{-1}\delta(z' - \Psi\circ R(z))\\ &= |\mathcal{J}(\Psi\circ R)|\delta(z' - \Psi\circ R(z)),\end{aligned}$$

where the last step is a consequence of $1 = \mathcal{J}(\Psi \circ R) \circ \Psi \circ R \cdot \mathcal{J}(\Psi \circ R)$. Therefore, to show Eq. (D1) reduces to showing

$$\theta \circ R \cdot \rho = \theta \circ R \circ \Psi \circ R \cdot \rho \circ \Psi \circ R \, |\mathcal{J}(\Psi \circ R)|,$$

where omitted arguments are all $z$. Replacing $z$ by $R(z)$ reduces this to showing

$$\theta \cdot \rho = \theta \circ R \circ \Psi \cdot \rho \circ \Psi \, |\mathcal{J}\Psi|. \tag{D2}$$

To simplify $\theta \circ R \circ \Psi$, note that

$$\rho \circ \Psi \circ R \circ \Psi = \rho \circ R = \rho/|\mathcal{J}R|,$$

$$\rho \circ R \circ \Psi = \rho \circ \Psi/|\mathcal{J}R \circ \Psi|,$$

$$\mathcal{J}\Psi \circ R \circ \Psi \cdot \mathcal{J}R \circ \Psi \cdot \mathcal{J}\Psi = \mathcal{J}R,$$

whence

$$\theta \circ R \circ \Psi = \min\left\{1, \frac{\rho}{\rho \circ \Psi |\mathcal{J}\Psi|}\right\},$$

from which Eq. (D2) and hence Eq. (D1) follows. Now, consider the case where either $R(z)$ or $R(z')$ is not in $\Omega$. Without loss of generality, suppose $R(z) \notin \Omega$. Suppose $R(z') \in \Omega$. Then $R \circ \Psi \circ R(z') \in \Omega$ and $R(z) \neq R \circ \Psi \circ R(z')$, implying $\delta(z - \Psi \circ R(z')) = 0$. Hence, $\rho(z|R(z')) = \delta(R(z') - R(z))$. This last equality also holds if $R(z') \notin \Omega$. Hence, modified detailed balance reduces to $\delta(z' - z)\rho(z) = \delta(z - z')\rho(z')$, which always holds.


[1] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, "Equation of state calculations by fast computing machines," J. Chem. Phys. **21**, 1087–1092 (1953).

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

[3] A. M. Horowitz, "A generalized guided Monte-Carlo algorithm," Phys. Lett. B **268**, 247–252 (1991).

[4] B. Mehlig, D. W. Heermann, and B. M. Forrest, "Hybrid Monte Carlo method for condensed-matter systems," Phys. Rev. B **45**, 679–685 (1992).

[5] J. M. Sanz-Serna, "Markov chain Monte Carlo and numerical differential equations," in *Current Challenges in Stability Issues for Numerical Differential Equations*, Lecture Notes in Mathematics Vol. 2082, edited by L. Dieci and N. Guglielmi (Springer, 2014), pp. 39–88.

[6] See Ref. 8, p. 126 of Ref. 11 and Ref. 9.

[7] J. Sohl-Dickstein, M. Mudigonda, and M. R. DeWeese, "Hamiltonian Monte Carlo without detailed balance," in *Proceedings of the 31st International Conference on Machine Learning, 2014* [J. Mach. Learn. Res. W&CP **32**, 719–726 (2014)].

[8] B. Leimkuhler and S. Reich, "A metropolis adjusted Nosé-Hoover thermostat," Math. Model. Numer. Anal. **43**, 743–755 (2009).

[9] S. Lan, V. Stathopoulos, B. Shahbaba, and M. Girolami, "Lagrangian dynamical Monte Carlo," J. Comput. Graph. Stat. (in press).

[10] M. Tuckerman, Y. Liu, G. Ciccotti, and G. Martyna, "Non-Hamiltonian molecular dynamics: Generalizing Hamiltonian phase space principles to non-Hamiltonian systems," J. Chem. Phys. **115**, 1678–1702 (2001).

[11] M. Girolami and B. Calderhead, "Riemann manifold Langevin and Hamiltonian Monte Carlo methods," J. R. Stat. Soc. B **73**(Pt 2), 123–214 (2011).

[12] J. Hu, A. Ma, and A. R. Dinner, "Monte Carlo simulations of biomolecules: The MC module in CHARMM," J. Comput. Chem. **27**, 203–216 (2006).

[13] C. S. Pangali, M. Rao, and B. J. Berne, "On a novel Monte Carlo scheme for simulating water and aqueous solutions," Chem. Phys. Lett. **55**, 413–417 (1978).

[14] P. J. Rossky, J. D. Doll, and H. L. Friedman, "Brownian dynamics as smart Monte Carlo simulation," J. Chem. Phys. **69**, 4628–4633 (1978).

[15] J. Besag, "Discussion of the paper by Grenander and Miller," J. R. Stat. Soc. Ser. B **56**, 591–592 (1994).

[16] G. O. Roberts and R. L. Tweedie, "Exponential convergence of Langevin diffusions and their discrete approximations," Bernoulli **2**, 341–363 (1996).

[17] W. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," Biometrika **57**, 97–109 (1970).

[18] S. Blanes, F. Casas, and J. Sanz-Serna, "Numerical integrators for the hybrid Monte Carlo method" (submitted).

[19] J. A. Izaguirre and S. S. Hampton, "Shadow hybrid Monte Carlo: An efficient propagator in phase space of macromolecules," J. Comput. Phys. **200**, 581–604 (2004).

[20] C. R. Sweet, S. S. Hampton, R. D. Skeel, and J. A. Izaguirre, "A separable shadow Hamiltonian hybrid Monte Carlo method," J. Chem. Phys. **131**, 174106 (2009).

[21] E. Akhmatskaya and S. Reich, "New hybrid Monte Carlo methods for efficient sampling: From physics to biology and statistics," in *The Joint International Conference of the 7th Supercomputing in Nuclear Application and the 3rd Monte Carlo (SNA+MC 2010)*, Progress in Nuclear Science and Technology Vol. 2 (Taylor & Francis, 2012), pp. 447–462.

[22] A. Beskos, N. Pillai, G. Roberts, J. M. Sanz-Serna, and A. Stuart, "Optimal tuning of the hybrid Monte-Carlo algorithm," Bernoulli J. **19**, 1501–1534 (2013).

[23] See Sec. 3.4 of A. Beskos, F. Pinski, J. Sanz-Serna, and A. Stuart, "Hybrid Monte-Carlo on Hilbert spaces," Stoch. Proc. Appl. **121**, 2201–2230 (2011).

[24] C. H. Bennett, "Mass tensor molecular dynamics," J. Comput. Phys. **19**, 267–279 (1975).

[25] I.-C. Lin and M. E. Tuckerman, "Enhanced conformational sampling of peptides via reduced side-chain and solvent masses," J. Phys. Chem. B **114**, 15935–15940 (2010).

[26] See Fig. 1 in Ref. 9.

[27] R. D. Skeel, "What makes molecular dynamics work?" SIAM J. Sci. Comput. **31**, 1363–1378 (2009).

[28] J. M. Sanz-Serna, "An unconventional symplectic integrator of W. Kahan," Appl. Numer. Math. **16**, 245–250 (1994).

[29] W. Kahan and R.-C. Li, "Unconventional schemes for a class of ordinary differential equations—With applications to the Korteweg-de Vries (KdV) equation," J. Comput. Phys. **134**, 316–331 (1997).

[30] S. Lan, J. Streets, and B. Shahbaba, "Wormhole Hamiltonian Monte Carlo," preprint arXiv:1306.0063v2.

[31] C. Schütte and W. Huisinga, "Biomolecular conformations as metastable sets of Markov chains," in *Proceedings of the 38th Annual Allerton Conference on Communication, Control, and Computing* (University of Illinois, 2000), pp. 1106–1115.

[32] J. Goodman, "Acor, statistical analysis of a time series," (2009), see http://www.math.nyu.edu/faculty/goodman/software/acor/.

[33] In fact, $\delta$ can be taken to be any differentiable function for which $\rho(x, p)$ is a valid p.d.f.

[34] D. J. Evans and G. Morriss, "The isothermal isobaric molecular dynamics ensemble," Phys. Lett. A **98**, 433–436 (1983).

[35] P. Minary, G. J. Martyna, and M. E. Tuckerman, "Algorithms and novel applications based on the isokinetic ensemble. I. Biophysical and path integral molecular dynamics," J. Chem. Phys. **118**, 2510 (2003).

[36] I. Omelyan and A. Kovalenko, "Generalised canonical-isokinetic ensemble: speeding up multiscale molecular dynamics and coupling with 3D molecular theory of solvation," Mol. Simul. **39**, 25–48 (2013).

[37] P. Klein, "Pressure and temperature control in molecular dynamics simulations: a unitary approach in discrete time," Model. Simul. Mater. Sci. Eng. **6**, 405–421 (1998).

[38] C. W. Gardiner, *Handbook of Stochastic Methods*, 3rd ed. (Springer-Verlag, Berlin, 2004) .

[39] G. E. Crooks, "Path-ensemble averages in systems driven far from equilibrium," Phys. Rev. E **61**, 2361–2366 (2000).