HHS Public Access

Author manuscript

J Comput Graph Stat. Author manuscript; available in PMC 2016 April 01.

Published in final edited form as:

J Comput Graph Stat. 2015 April 1; 24(2): 357–378. doi:10.1080/10618600.2014.902764.

Markov Chain Monte Carlo from Lagrangian Dynamics

Shiwei Lan[*], **Vasileios Stathopoulos**[†], **Babak Shahbaba**[*], and **Mark Girolami**[†]

[*] Department of Statistics, University of California, Irvine, Irvine, CA 92697, USA

[†]Department of Statistical Science, University College London, London, WC1E 6BT, UK

Abstract

Hamiltonian Monte Carlo (HMC) improves the computational e ciency of the Metropolis-Hastings algorithm by reducing its random walk behavior. Riemannian HMC (RHMC) further improves the performance of HMC by exploiting the geometric properties of the parameter space. However, the geometric integrator used for RHMC involves implicit equations that require fixed-point iterations. In some cases, the computational overhead for solving implicit equations undermines RHMC's benefits. In an attempt to circumvent this problem, we propose an explicit integrator that replaces the momentum variable in RHMC by velocity. We show that the resulting transformation is equivalent to transforming Riemannian Hamiltonian dynamics to Lagrangian dynamics. Experimental results suggests that our method improves RHMC's overall computational e ciency in the cases considered. All computer programs and data sets are available online (http://www.ics.uci.edu/~babaks/Site/Codes.html) in order to allow replication of the results reported in this paper.

Keywords

Hamiltonian Monte Carlo; Riemannian Manifold; Lagrangian Dynamics; Explicit Integrator

1 Introduction

Hamiltonian Monte Carlo (HMC) (Duane et al., 1987) reduces the random walk behavior of the Metropolis-Hastings algorithm by proposing samples that are distant from the current state, but nevertheless have a high probability of acceptance. These distant proposals are found by numerically simulating Hamiltonian dynamics for some specified amount of fictitious time (Neal, 2010). Hamiltonian dynamics can be represented by a function, known as the *Hamiltonian*, of model parameters $\boldsymbol{\theta}$ and auxiliary momentum parameters $\mathbf{p} \sim N(0, \mathbf{M})$ (with the same dimension as $\boldsymbol{\theta}$) as follows:

$$H(\boldsymbol{\theta}, \mathbf{p}) = -\log p(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{M}^{-1}\mathbf{p} \quad (1)$$

where M is a symmetric, positive-definite *mass matrix*.

Hamilton's equations, which involve di erential equations derived from H, determine how $\boldsymbol{\theta}$ and $\mathbf{p}$ change over time. In practice, however, solving these equations exactly is di cult in general, so we need to approximate them by discretizing time, using some small step size ε. For this purpose, the *leapfrog* method is commonly used.

Hamiltonian dynamics is restricted by the smallest eigen-direction, requiring small step size to maintain the stability of the numerical discretization. Girolami and Calderhead (2011) propose a new method, called Riemannian HMC (RHMC), that exploits the geometric properties of the parameter space to improve the e ciency of standard HMC, especially in sampling distributions with complex structure (e.g., high correlation, non-Gaussian shape). Simulating the resulting dynamics, however, is computationally intensive since it involves solving two implicit equations, which require additional iterative numerical computation (e.g., fixed-point iteration).

In an attempt to increase the speed of RHMC, we propose a new integrator that is completely explicit: we propose to replace momentum with velocity in the definition of the Riemannian Hamiltonian dynamics. As we will see, this is equivalent to using Lagrangian dynamics as opposed to Hamiltonian dynamics. By doing so, we eliminate one of the implicit steps in RHMC. Next, we construct a time symmetric integrator to remove the remaining implicit step in RHMC. This leads to a valid sampling scheme (i.e., converges to the true target distribution) that involves only explicit equations. We refer to this algorithm as Lagrangian Monte Carlo (LMC).

In what follows, we begin with a brief review of RHMC and its geometric integrator. Section 3 introduces our proposed semi-explicit integrator based on defining Hamiltonian dynamics in terms of velocity as opposed to momentum. Next, in Section 4, we eliminate the remaining implicit equation and propose a fully explicit integrator. In Section 5, we use simulated and real data to evaluate our methods' performance. Finally, in Section 6, we discuss some possible future research directions.

2 Riemannian Hamiltonian Monte Carlo

As discussed above, although HMC explores the parameter space more e ciently than random walk Metropolis does, it does not fully exploit the geometric properties of parameter space defined by the density $p(\boldsymbol{\theta})$. Indeed, Girolami and Calderhead (2011) argue that dynamics over Euclidean space may not be appropriate to guide the exploration of parameter space. To address this issue, they propose a new method, called Riemannian HMC (RHMC), that exploits the Riemannian geometry of the parameter space (Amari and Nagaoka, 2006) to improve standard HMC's e ciency by automatically adapting to the local structure. They do this by using a position-specific mass matrix $\mathrm{M} = \mathrm{G}(\boldsymbol{\theta})$. More specifically, they set $\mathrm{G}(\boldsymbol{\theta})$ to the Fisher information matrix $\mathrm{E}\left[\nabla log\, p\,(\boldsymbol{\theta})\, \nabla log\, p(\boldsymbol{\theta})^{\mathsf{T}}\right]$. As a result, $\phi\,(\boldsymbol{\theta}) := -log\, p\,(\boldsymbol{\theta}) + \frac{1}{2}\, log\, det\, \mathbf{G}\,(\boldsymbol{\theta})$, and the Hamiltonian is defined as follows:

$$H\,(\boldsymbol{\theta}, \mathbf{p}) = -log\, p\,(\boldsymbol{\theta}) + \frac{1}{2} log\, det\, \mathbf{G}\,(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\, \mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p} = \phi\,(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^{\mathsf{T}}\, \mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p} \quad (2)$$

where $\phi\,(\boldsymbol{\theta}) := -log\, p\,(\boldsymbol{\theta}) + \frac{1}{2}\, log\, det\, \mathbf{G}\,(\boldsymbol{\theta})$. Based on this Hamiltonian, Girolami and Calderhead (2011) propose the following Hamiltonian dynamics on the Riemmanian manifold:

$$\begin{aligned}\dot{\boldsymbol{\theta}} &= \nabla_{\mathbf{p}} H(\boldsymbol{\theta}, \mathbf{p}) = \mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p} \\ \dot{\mathbf{p}} &= -\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}, \mathbf{p}) = -\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}) + \tfrac{1}{2}\boldsymbol{\nu}(\boldsymbol{\theta}, \mathbf{p})\end{aligned} \quad (3)$$

With the shorthand notation $\partial_i = \partial/\partial\theta_i$ for partial derivative, the ith element of the vector $\nu(\boldsymbol{\theta}, \mathrm{p})$ is

$$(\boldsymbol{\nu}(\boldsymbol{\theta}, \mathbf{p}))_i = -\mathbf{p}^{\mathsf{T}}\,\partial_i\left(\mathbf{G}(\boldsymbol{\theta})^{-1}\right)\mathbf{p} = \left(\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}\right)^{\mathsf{T}}\partial_i\mathbf{G}(\boldsymbol{\theta})\,\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}$$

The above equations of dynamics are non-separable (they contain products of $\boldsymbol{\theta}$ and p), and the resulting map $(\boldsymbol{\theta}, \mathrm{p}) \rightarrow (\boldsymbol{\theta}^*, \mathbf{p}^*)$ based on the standard leapfrog method is neither time-reversible nor symplectic. Therefore, we cannot use the standard leapfrog algorithm (Girolami and Calderhead, 2011). Instead, they use the Stömer-Verlet (Verlet, 1967) method as follows:

$$\mathbf{p}^{(n+1/2)} = \mathbf{p}^{(n)} - \frac{\varepsilon}{2}\left[\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}^{(n)}\right) - \frac{1}{2}\boldsymbol{\nu}\left(\boldsymbol{\theta}^{(\mathbf{n})}, \mathbf{p}^{(\mathbf{n+1/2})}\right)\right] \quad (4)$$

$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} + \frac{\varepsilon}{2}\left[\mathbf{G}^{-1}\left(\boldsymbol{\theta}^{(n)}\right) + \mathbf{G}^{-1}\left(\boldsymbol{\theta}^{(n+1)}\right)\right]\mathbf{p}^{(n+1/2)} \quad (5)$$

$$\mathbf{p}^{(n+1)} = \mathbf{p}^{(n+1/2)} - \frac{\varepsilon}{2}\left[\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}^{(n+1)}\right) - \frac{1}{2}\boldsymbol{\nu}\left(\boldsymbol{\theta}^{(\mathbf{n+1})}, \mathbf{p}^{(\mathbf{n+1/2})}\right)\right] \quad (6)$$

where ε is the size of time step. This is also known as generalized leapfrog, which can be derived by concatenating a symplectic Euler-B integrator of (3) with its adjoint symplectic Euler-A integrator (See more details in Leimkuhler and Reich, 2004). The above series of transformations are (i) deterministic (ii) reversible and (iii) volume-preserving. Therefore, the e ective proposal distribution is a delta function $\delta((\boldsymbol{\theta}^{(1)}, \mathbf{p}^{(1)}), (\boldsymbol{\theta}^{(L+1)}, \mathbf{p}^{(L+1)}))$ and the acceptance probability is simply:

$$exp\left(H\left(\boldsymbol{\theta}^{(1)}, \mathbf{p}^{(1)}\right) - H\left(\boldsymbol{\theta}^{(L+1)}, \mathbf{p}^{(L+1)}\right)\right) \quad (7)$$

Here, $(\boldsymbol{\theta}^{(1)}, \mathbf{p}^{(1)})$ is the current state, and $(\boldsymbol{\theta}^{(L+1)}, \mathbf{p}^{(L+1)})$ is the proposed sample after L leapfrog steps.

As an illustrative example, Figure 1 shows the sampling paths of random walk Metropolis (RWM), HMC, and RHMC for an artificially created banana-shaped distribution. (See Girolami and Calderhead, 2011, discussion by Luke Bornn and Julien Cornebise.) For this example, we fix the trajectory and choose the step sizes such that the acceptance probability for all three methods remains around 0.7. RWM moves slowly and spends most of iterations at the distribution's low-density tail, and HMC explores the parameter space in an indirect way, while RHMC moves directly to the high-density region and explores the distribution more e ciently.

One major drawback of this geometric integrator, which is both time-reversible and volume-preserving, is that it involves two implicit functions: Equations (4) and (5). These functions require extra numerical analysis (e.g. fixed-point iteration), which results in higher computational cost and simulation error. This is especially true when solving $\boldsymbol{\theta}^{(n+1)}$ because the fixed-point iteration for (5) repeatedly inverts matrix **G.** To address this problem, we propose an alternative approach that uses velocity instead of momentum in the equations of motion.

3 Moving from Momentum to Velocity

In the equations of Hamiltonian dynamics (3), the term $\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}$ appears several times. This motivates us to re-parameterize the dynamics in terms of velocity, $\mathbf{v} = \mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}$**.** Note that this in fact corresponds to the usual definition of velocity in physics, i.e., momentum divided by mass. The transformation $\mathbf{p} \rightarrow \mathbf{v}$ changes the Hamiltonian dynamics (3) to the following form (derivation in Appendix A):

$$\begin{aligned} \dot{\boldsymbol{\theta}} &= \mathbf{v} \\ \dot{\mathbf{v}} &= -\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{v}) - \mathbf{G}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}) \end{aligned} \quad (8)$$

where $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{v})$ is a vector whose kth element is $\sum_{i,j}\Gamma_{ij}^{k}(\boldsymbol{\theta})\,\mathbf{v}^{i}\mathbf{v}^{j}$. Here,

$\Gamma_{ij}^{k}(\boldsymbol{\theta}) := \frac{1}{2}\sum_{l}\mathrm{g}^{kl}\left(\partial_{i}\mathrm{g}_{lj} + \partial_{j}\mathrm{g}_{il} - \partial_{l}\mathrm{g}_{ij}\right)$ are the Christoffel symbols, where g_{ij} and g^{ij} denote (i, j)th element of $\mathbf{G}(\boldsymbol{\theta})$ and $\mathbf{G}(\boldsymbol{\theta})^{-1}$ respectively.

This transformation moves the complexity of the dynamics in the first equation for $\boldsymbol{\theta}$ to its second equation where it spends more time in finding a "good" direction **v**. By resolving the implicitness of updating $\boldsymbol{\theta}$ in the generalized leapfrog method, we reduce the associated computational cost. Concatenating the Euler-B integrator with its adjoint Euler-A integrator (Leimkuhler and Reich, 2004, see also Appendix B) leads to the following semi-explicit integrator:

$$\mathbf{v}^{(n+1/2)} = \mathbf{v}^{(n)} - \frac{\varepsilon}{2}\left[\boldsymbol{\eta}\left(\boldsymbol{\theta}^{(\mathbf{n})}, \mathbf{v}^{(\mathbf{n+1/2})}\right) + \mathbf{G}\left(\theta^{(\mathbf{n})}\right)^{-1}\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}^{(\mathbf{n})}\right)\right] \quad (9)$$

$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} + \varepsilon\mathbf{v}^{(n+1/2)} \quad (10)$$

$$\mathbf{v}^{(n+1)} = \mathbf{v}^{(n+1/2)} - \frac{\varepsilon}{2}\left[\boldsymbol{\eta}\left(\boldsymbol{\theta}^{(\mathbf{n+1})}, \mathbf{v}^{(\mathbf{n+1/2})}\right) + \mathbf{G}\left(\boldsymbol{\theta}^{(\mathbf{n+1})}\right)^{-1}\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}^{(\mathbf{n+1})}\right)\right] \quad (11)$$

Note that equation (9) updating **v** remains implicit (more details are available in Appendix B).

The introduction of velocity **v** in place of **p** is also advocated by Beskos et al. (2011) to avoid large variables **p** for the sake of numerical stability. They consider a constant mass so the resulting dynamics is still Hamiltonian. In general, however, the new dynamics (8)

cannot be recognized as Hamiltonian dynamics of ($\boldsymbol{\theta}$, $\mathbf{v}$); rather, it is equivalent to the following Euler-Lagrange equation of the second kind:

$$\frac{\partial \mathbf{L}}{\partial \boldsymbol{\theta}} = \frac{d}{dt}\frac{\partial \mathbf{L}}{\partial \dot{\boldsymbol{\theta}}}$$

which is the solution to variation of the *Lagrangian*: $\mathbf{L} = \frac{1}{2}\mathbf{v}^{\mathsf{T}}\,\mathbf{G}(\boldsymbol{\theta})\,\mathbf{v} - \phi(\boldsymbol{\theta})$. That is, in our case,

$$\ddot{\boldsymbol{\theta}} = -\boldsymbol{\eta}\left(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}\right) - \mathbf{G}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta})$$

Therefore, we refer to the new dynamics (8) as *Lagrangian* dynamics. Although the resulting dynamics is not Hamiltonian, it nevertheless remains a valid proposal-generating mechanism, which preserves the original Hamiltonian $H(\boldsymbol{\theta}, \mathbf{p} = \mathbf{G}(\boldsymbol{\theta})\mathbf{v})$ (proof in Appendix A); thus, the acceptance probability is only determined by the discretization error from the numerical integration as usual (to be discussed below).

Because $\mathbf{p}|\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}(\boldsymbol{\theta}))$, the distribution of $\mathbf{v}|\boldsymbol{\theta}$ is $\mathcal{N}\left(\mathbf{0}, \mathbf{G}(\boldsymbol{\theta})^{-1}\right)$. We define the *energy* function $\mathbf{E}(\boldsymbol{\theta}, \mathbf{v})$ as the sum of the potential energy, $U(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta})$ and kinetic energy $K(\boldsymbol{\theta}, \mathbf{v}) = -\log(p(\mathrm{v}|\boldsymbol{\theta}))$:

$$\mathbf{E}(\boldsymbol{\theta}, \mathbf{v}) = -\log p(\boldsymbol{\theta}) - \frac{1}{2}\log\det \mathbf{G}(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{v}^{\mathsf{T}}\,\mathbf{G}(\boldsymbol{\theta})\,\mathbf{v} \quad (12)$$

Note that the integrator (9)-(11) is (i) reversible and (ii) energy-preserving up to a global error with order $\mathcal{O}(\varepsilon)$, where ε is the step size. (See Proposition 1 in Appendix B and a proof of distribution invariance in Appendix B.1.) The resulting map, however, is not volume-preserving. Therefore, the acceptance probability based on $\mathbf{E}(\boldsymbol{\theta}, \mathbf{v})$ must be adjusted with the determinant of the transformation to ensure that a detailed balance condition is satisfied (Peter J. Green, 1995),

$$\alpha_{sLMC} = min\left\{1, exp\left(-\mathbf{E}\left(\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)}\right) + \mathbf{E}\left(\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)}\right)\right)|det \quad \mathbf{J}_{sLMC}|\right\}$$

where J_{sLMC} is the Jacobian matrix of $(\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)}) \to (\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)})$ according to (9)-(11) with the following determinant (see more in Appendix B.2):

$$det \quad \mathrm{J}_{sLMC} := \left|\frac{\partial\left(\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)}\right)}{\partial\left(\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)}\right)}\right| = \prod_{n=1}^{L}\frac{det\left(I - \varepsilon\Omega\left(\boldsymbol{\theta}^{(n+1)}, \mathbf{v}^{(n+1/2)}\right)\right)}{det\left(I + \varepsilon\Omega\left(\boldsymbol{\theta}^{(n)}, \mathbf{v}^{(n+1/2)}\right)\right)}$$

Here, $\boldsymbol{\Omega}(\boldsymbol{\theta}^{(n+1)}, \mathbf{v}^{(n+1/2)})$ is a matrix whose (i, j)th element is $\sum_k \mathbf{v}_k^{(n+1/2)}\Gamma_{kj}^{i}\left(\boldsymbol{\theta}^{(n+1)}\right)$.

The dynamics is now defined by the semi-explicit integrator (9)-(11). Algorithm 1 (see pseudocode in Appendix D) provides the corresponding steps. We refer to this approach as semi-explicit Lagrangian Monte Carlo (sLMC), which has a physical interpretation as exploring the parameter space along the path on a Riemannian manifold that minimizes the action (total Lagrangian). Contrast this to RHMC augmenting parameter space with momentum, sLMC augments parameter space with velocity. In Section 5, we use several experiments to show that switching from momentum to velocity may lead to improvements in computational e ciency in some cases.

4 Explicit Lagrangian Monte Carlo

To further resolve the the remaining implicit equation (9), we modify $\eta\theta^{(n)}, \mathbf{v}^{(n+1/2)})$, to be an asymmetric function of both $\mathbf{v}^{(n)}$ and $\mathbf{v}^{(n+1/2)}$ such that $\mathbf{v}^{(n+1/2)}$ can be solved explicitly (see details in Appendix C). We now propose a fully explicit integrator for Lagrangian dynamics (8) as follows:

$$\mathbf{v}^{(n+1/2)}=\left[\mathbf{I}+\frac{\varepsilon}{2}\Omega\left(\boldsymbol{\theta}^{(n)},\mathbf{v}^{(n)}\right)\right]^{-1}\left[\mathbf{v}^{(n)}-\frac{\varepsilon}{2}\mathbf{G}\left(\boldsymbol{\theta}^{(n)}\right)^{-1}\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}^{(n)}\right)\right] \quad (13)$$

$$\boldsymbol{\theta}^{(n+1)}=\boldsymbol{\theta}^{(n)}+\varepsilon\mathbf{v}^{(n+1/2)} \quad (14)$$

$$\mathbf{v}^{(n+1)}=\left[\mathbf{I}+\frac{\varepsilon}{2}\Omega\left(\boldsymbol{\theta}^{(n+1)},\mathbf{v}^{\left(n+\frac{1}{2}\right)}\right)\right]^{-1}\left[\mathbf{v}^{(n+1/2)}-\frac{\varepsilon}{2}\mathbf{G}\left(\boldsymbol{\theta}^{(n+1)}\right)^{-1}\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}^{(n+1)}\right)\right] \quad (15)$$

In Appendix C.1, we prove that the solution obtained from this integrator numerically converges to the true solution of the Lagrangian dynamics. This integrator is (i) reversible and (ii) energy-preserving up to a global error with order $\mathcal{O}(\varepsilon)$. The resulting map is not volume-preserving as the Jacobian determinant of $(\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)}) \rightarrow (\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)})$ by (13)-(15) is (see details in Appendix C.2):

$$det \quad \mathbf{J}_{LMC} := \prod_{n=1}^{L} \frac{det\left(\mathbf{G}\left(\boldsymbol{\theta}^{(n+1)}\right)-\varepsilon/2\tilde{\Omega}\left(\boldsymbol{\theta}^{(n+1)},\mathbf{v}^{(n+1)}\right)\right) det\left(\mathbf{G}\left(\boldsymbol{\theta}^{(n)}\right)-\varepsilon/2\tilde{\Omega}\left(\boldsymbol{\theta}^{(n)},\mathbf{v}^{(n+1/2)}\right)\right)}{det\left(\mathbf{G}\left(\boldsymbol{\theta}^{(n+1)}\right)+\varepsilon/2\tilde{\Omega}\left(\boldsymbol{\theta}^{(n+1)},\mathbf{v}^{(n+1/2)}\right)\right) det\left(\mathbf{G}\left(\boldsymbol{\theta}^{(n)}\right)+\varepsilon/2\tilde{\Omega}\left(\boldsymbol{\theta}^{(n)},\mathbf{v}^{(n)}\right)\right)}$$

Here, $\tilde{\Omega}(\boldsymbol{\theta},\mathbf{v})$ denotes $\mathbf{G}(\boldsymbol{\theta})\boldsymbol{\Omega}(\boldsymbol{\theta},\mathbf{v})$ whose (k, j)th element is equal to $\sum_i \mathbf{v}^i\tilde{\Gamma}_{ijk}(\boldsymbol{\theta})$, with $\tilde{\Gamma}_{ijk}(\boldsymbol{\theta})=g_{kl}\Gamma^l_{ij}(\boldsymbol{\theta})=\frac{1}{2}(\partial_i g_{kj}+\partial_i g_{ik}-\partial_k g_{ij})$. As a result, the acceptace probability must be adjusted as follows:

$$\alpha_{LMC}=min\left\{1, exp\left(-\mathrm{E}\left(\theta^{(L+1)},\mathrm{v}^{(L+1)}\right)+\mathrm{E}\left(\theta^{(1)},\mathrm{v}^{(1)}\right)\right)|det \quad \mathrm{J}_{LMC}|\right\}$$

We refer to this approach that involves a fully explicit integrator (13)-(15) as Lagrangian Monte Carlo (LMC). Algorithm 2 (see pseudocode in Appendix D) shows the corresponding steps for this method. In both algorithms 1 and 2, the position update is relatively simple

while the computational time is dominated by choosing the "right" direction (velocity) using the geometry of parameter space. In sLMC, solving $\boldsymbol{\theta}$ explicitly reduces computation cost by $(F-1)\,\mathcal{O}\left(D^{2.373}\right)$ where F is the number of fixed-point iterations, and D is the number of parameters. This is because for each fixed-point iteration, it takes $\mathcal{O}\left(D^{2.373}\right)$ elementary linear algebraic operations to invert $\mathbf{G}(\boldsymbol{\theta})$. The connection terms $\tilde{\Gamma}(\boldsymbol{\theta})$ in $\tilde{\Omega}$ do not add substantial computational cost since they are obtained from permuting three dimensions of the array $\partial\mathbf{G}(\boldsymbol{\theta})$, which is also computed in RHMC. The additional price of determinant adjustment is $O(D^{2.373})$.

LMC avoids the fixed-point iteration method. Therefore, it reduces computation by $(F-1)\,\mathcal{O}\left(D^{2}\right)$. Further, it resolves possible convergence issues associated with using the fixed-point iteration method. However, because it involves additional matrix inversions to update $\mathbf{v}$, its benefits could be undermined occasionally. This is evident from our experimental results presented in Section 5.

5 Experimental Results

In this section, we use both simulated and real data to evaluate our methods, sLMC and LMC, compared to standard HMC and RHMC. Following Girolami and Calderhead (2011), we use a time-normalized effective sample size (ESS) to compare these methods. For B posterior samples we calculate $ESS = B\left[1+2\sum_{k=1}^{K}\gamma(k)\right]^{-1}$ for each parameter, where $\sum_{k=1}^{K}\gamma(k)$ is the sum of K monotone sample autocorrelations (Geyer, 1992). Minimum, median, and maximum values of ESS over all parameters are provided for comparing different algorithms. More specifically, we use the minimum ESS normalized by CPU time (s), min(ESS)/s, as the measure of sampling e ciency. All computer programs and data sets discussed in this paper are available online at http://www.ics.uci.edu/~babaks/Site/Codes.html.

5.1 Simulating a banana-shaped distribution

The banana-shaped distribution, which we used above for illustration, can be constructed as the posterior distribution of $\theta = (\theta_1, \theta_2)|$ based on the following model:

$$\begin{aligned} y|\theta &\sim N\left(\theta_1+\theta_2^2, \sigma_y^2\right) \\ \theta &\sim N\left(0, \sigma_\theta^2\mathbf{I}_2\right) \end{aligned}$$

The data $\{y_i\}_{i=1}^{100}$ are generated with $\theta_1+\theta_2^2=1$, $\sigma_y = 2$, and $\sigma_\theta = 1$.

As we can see in Figure 2, similar to RHMC, sLMC and LMC explore the parameter space efficiently by adapting to its local geometry. Table 1 compares the performances of these algorithms based on 20000 MCMC iterations after 5000 burn-in. For this specific example, sLMC has the best performance followed by LMC. As discussed above, although LMC is fully explicit, its numerical benefits (obtained by removing implicit equations) can be

negated in certain examples since it involves additional matrix inversion operations to update v. The histograms of posterior samples shown in Figure 3 confirm that our algorithms converge to the true posterior distributions of θ_1 and θ_2, whose density functions are shown as red solid curves.

5.2 Logistic Regression Models

Next, we evaluate our methods based on five binary classification problems used in Girolami and Calderhead (2011). These are Australian Credit data, German Credit data, Heart data, Pima Indian data, and Ripley data. These data sets are publicly available from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml). For each problem, we use a logistic regression model,

$$\begin{aligned} p\left(y_i{=}1|\boldsymbol{x}_i,\boldsymbol{\beta}\right) &= \frac{exp\left(\boldsymbol{x}_i^{\mathrm{T}}\,\boldsymbol{\beta}\right)}{1{+}exp\left(\boldsymbol{x}_i^{\mathrm{T}}\,\boldsymbol{\beta}\right)}, \quad i{=}1,\ldots,n \\ \boldsymbol{\beta} &\sim N\left(\mathbf{0},\mathbf{100}\boldsymbol{I}\right) \end{aligned}$$

where y_i is a binary outcome for the ith observation, x_i is the corresponding vector of predictors (with the first element equal to 1), and β is the set of regression parameters.

We use standard HMC, RHMC, sLMC, and LCM to simulate 20000 posterior samples for β. We fix the trajectory length for different algorithms, and tune the step sizes so that they have comparable acceptance rates. Results (after discarding the initial 5000 iterations) are summarized in Table 2, and show that in general our methods improve the sampling efficiency measured in terms of minimum ESS per second compared to RHMC on these examples.

5.3 Results for Multivariate T-distributions

The computational complexity of standard HMC is $\mathcal{O}\left(D\right)$. This is substantially lower than $\mathcal{O}\left(D^{2.373}\right)$, which is the computational complexity of the three geometrically motivated methods discussed here (RHMC, sLMC, and LMC). On the other hand, these three methods could have substantially better mixing rates compared to standard HMC, whose mixing time is mainly determined by the *condition number* of the target distribution defined as the ratio of the maximum and minimum eigenvalues of its covariance matrix: $\lambda_{\max}/\lambda_{\min}$.

In this section, we illustrate how efficiency of these sampling algorithms changes as the condition number varies using multivariate t-distributions with the following density function:

$$p\left(x\right) = \frac{\Gamma\left(\left(\nu{+}D\right)/2\right)}{\Gamma\left(\nu/2\right)}\left(\pi\nu\right)^{-D/2}{}_{|\Sigma|^{-1/2}}\left(1{+}\frac{1}{\nu}x^{T}\Sigma^{-1}x\right)^{-(\nu+D)/2} \quad (16)$$

where ν is the degrees of freedom and D is the dimension. In our first simulation, we fix the dimension at $D = 20$ and vary the condition number of Σ from 10 to 10^5. As the condition number increases, one can expect HMC to be more restricted by the smallest eigen-direction, whereas RHMC, sLMC, and LMC would adapt to the local geometry. Results

presented Figure 4 (left panel) show that this is in fact the case: for higher conditional numbers, geometrically motivated methods perform substantially better than standard HMC. Note that our two proposed algorithms, sLMC and LMC, provide substantial improvements over RHMC.

For our second simulation, we fix the condition number at 10000 and let the dimension changes from 10 to 50. Our results (Figure 4, right panel) show that the gain by exploiting geometric properties of the target distribution could be undermined eventually as the dimension increases.

5.4 Finite Mixture of Gaussians

Finally we consider finite mixtures of univariate Gaussian components of the form

$$p\left(x_i|\boldsymbol{\theta}\right)=\sum_{k=1}^{K}\pi_k\mathcal{N}\left(x_i|\mu_k,\sigma_k^2\right) \quad (17)$$

where $\boldsymbol{\theta}$ is the vector of size $D = 3K - 1$ of all the parameters $g=p_k$, μ_k and σ_k^2 and $\mathcal{N}\left(\cdot|\mu,\sigma^2\right)$ is a Gaussian density with mean μ and variance σ^2. A common choice of prior takes the form

$$p\left(\boldsymbol{\theta}\right)=\mathcal{D}\left(\pi_1,\ldots,\pi_K|\lambda\right)\prod_{k=1}^{K}\mathcal{N}\left(\mu_k|m,\beta^{-1}\sigma_k^2\right)\mathcal{IG}\left(\sigma_k^2|b,c\right) \quad (18)$$

where $\mathcal{D}\left(\cdot|\lambda\right)$ is the symmetric Dirichlet distribution with parameter λ, and $\mathcal{IG}\left(\cdot|b,c\right)$ is the inverse Gamma distribution with shape parameter b and scale parameter c.

Although the posterior distribution associated with this model is formally explicit, it is computationally intractable, since it can be expressed as a sum of K^N terms corresponding to all possible allocations of observations x_i to mixture components (Marin et al., 2005, chap. 9). We want to use this model to test the efficiency of posterior sampling $\boldsymbol{\theta}$ using the four methods. A more extensive comparison of Riemannian Manifold MCMC and HMC, Gibbs sampling and standard Metropolis-Hastings for finite Gaussian mixture models can be found at Stathopoulos and Girolami (2011). Due to the non-analytic nature of the expected Fisher Information, $\boldsymbol{I}(\boldsymbol{\theta})$, we use the empirical Fisher information as metric tensor (McLachlan and Peel, 2000, chap. 2),

$$\boldsymbol{G}\left(\boldsymbol{\theta}\right)=\boldsymbol{S}^{\mathsf{T}}\,\boldsymbol{S}-\frac{1}{N}\boldsymbol{s}\boldsymbol{s}^{\mathsf{T}}$$

where $N \times D$ score matrix $\boldsymbol{S}$ has elements $S_{i,d}=\frac{\partial\, log\; p\left(x_i|\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}_d}$ and $\boldsymbol{s}=\sum_{i=1}^{N}\boldsymbol{S}_i^{\mathsf{T}}$.

We show five Gaussian mixtures in Table 3 and Figure 5 and compare sampling efficiency of HMC, RMHMC, sLMC and LMC using simulated datasets in Table 4. As before, our two algorithms outperform RHMC.

6 Conclusions and Discussion

Following the method of Girolami and Calderhead (2011) for more e cient exploration of parameter space, we have proposed new sampling schemes to reduce the computational cost associated with using a position-specific mass matrix. To this end, we have developed a semi-explicit (sLMC) integrator and a fully explicit (LMC) integrator for RHMC and demonstrated their advantage in improving computational efficiency over the generalized leapfrog (RHMC) method used by Girolami and Calderhead (2011). It is easy to show that if $\mathbf{G}(\boldsymbol{\theta}) \equiv \mathbf{M}$, our method reduces to standard HMC.

Compared to HMC, whose local and global errors are $\mathcal{O}\left(\varepsilon^3\right)$ and $\mathcal{O}\left(\varepsilon^2\right)$ respectively, LMC's local error is $\mathcal{O}\left(\varepsilon^2\right)$, and its global error is $\mathcal{O}(\varepsilon)$ (Proof in Appendix C.1). Although the numerical solutions converge to the true solutions of the corresponding dynamics at a slower rate for LMC compared to HMC, in general, the approximation remains adequate leading to reasonably high acceptance rates while providing a more computationally e cient sampling mechanism. Compared to RHMC, our LMC method has the additional advantage of being more stable by avoiding implicit updates relying on the fixed point iteration method: RHMC could occasionally give highly divergent solutions, especially for ill conditioned metrics, $\mathbf{G}(\boldsymbol{\theta})$.

Future directions could involve splitting Hamiltonian (Dullweber et al., 1987; Sexton and Weingarten, 1992; Neal, 2010; Shahbaba et al., 2013) to develop explicit geometric integrators. For example, one could split a non-separable Hamiltonian dynamic into several smaller dynamics some of which can be solved analytically. A similar idea has been explored by Chin (2009), where the Hamiltonian, instead of the dynamic, is split.

Because our methods involve costly matrix inversions, another possible research direction could be to approximate the mass matrix (and the Christo el symbols as well) to reduce computational cost. For many high-dimensional problems, the mass matrix could be appropriately approximated by a highly sparse or structured sparse (e.g., tridiagonal) matrix. This could further improve our method's computational efficiency.

Acknowledgements

VS and MAG are grateful to the UK Engineering and Physical Sciences Research Council (EPSRC) for supporting this research through project grant EP/K015664/1. MAG is grateful for support by an EPSRC Established Research Fellowship EP/J016934/1 and a Royal Society Wolfson Research Merit Award. SL and BS are supported by NSF grant IIS-1216045 and NIH grant R01-AI107034.

Appendices: Derivations and Proofs

In what follows, we show the detailed derivations of our methods. We adopt Einstein notation, so whenever the index appears twice in a mathematical expression, we sum over it: e.g., $a_i b^i := \sum_i a_i b^i, \Gamma_{ij}^k v^i v^j := \sum_{i,j} \Gamma_{ij}^k v^i v^j$. A lower index is used for the covariant tensor, whose components vary by the same transformation as the change of basis (e.g., gradient), whereas the upper index is reserved for the contravariant tensor, whose

components vary in the opposite way as the change of basis in order to compensate (e.g. velocity vector). Interested readers should refer to Bishop and Goldberg (1980).

A Transformation of Hamiltonian dynamic

To derive the dynamic (8) from the Hamiltonian dynamic (3), the first equation in (8) is directly obtained from the assumed transformation: $\boldsymbol{\theta}^k = \mathrm{g}^{kl}\mathbf{p}_l = \mathbf{v}^k$. For the second equation in (8), we have

$$\dot{\mathbf{p}}_l = \frac{d\left(\mathrm{g}_{lj}\left(\boldsymbol{\theta}\right)\mathbf{v}^j\right)}{dt} = \frac{\partial \mathrm{g}_{lj}}{\partial \boldsymbol{\theta}_i}\dot{\boldsymbol{\theta}}^i\mathbf{v}^j + \mathrm{g}_{lj}\dot{v}^j = \partial_i \mathrm{g}_{lj}\mathbf{v}^i\mathbf{v}^j + \mathrm{g}_{lj}\dot{\mathbf{v}}^j$$

Further, from Equation (3) we have

$$\begin{aligned}\dot{\mathbf{p}}_l = & -\partial_l\phi\left(\boldsymbol{\theta}\right) + \tfrac{1}{2}\mathbf{v}^{\mathsf{T}}\,\partial_l\mathbf{G}\left(\boldsymbol{\theta}\right)\mathbf{v} = -\partial_l\phi + \tfrac{1}{2}\mathrm{g}_{ij,l}\mathbf{v}^i\mathbf{v}^j \\ = & \partial_i \mathrm{g}_{lj}\mathbf{v}^i\mathbf{v}^j + \mathrm{g}_{lj}\dot{\mathbf{v}}^j\end{aligned}$$

which means

$$\mathrm{g}_{lj}\dot{\mathbf{v}}^j = -\left(\partial_i \mathrm{g}_{lj} - \frac{1}{2}\partial_l \mathrm{g}_{ij}\right)\mathbf{v}^i\mathbf{v}^j - \partial_l\phi$$

By multiplying $\mathbf{G}^{-1} = (\mathrm{g}^{kl})$ on both sides, we have

$$\dot{\mathbf{v}}^k = \delta^k_j\dot{\mathbf{v}}^j = -\mathrm{g}^{kl}\left(\partial_i \mathrm{g}_{lj} - \frac{1}{2}\partial_l \mathrm{g}_{ij}\right)\mathbf{v}^i\mathbf{v}^j - \mathrm{g}^{kl}\partial_l\phi \quad (19)$$

Since *i*, *j* are symmetric in the first summand, switching them gives the following equations:

$$\dot{\mathbf{v}}^k = -\mathrm{g}^{kl}\left(\partial_j \mathrm{g}_{li} - \frac{1}{2}\partial_l \mathrm{g}_{ji}\right)\mathbf{v}^i\mathbf{v}^j - \mathrm{g}^{kl}\partial_l\phi \quad (20)$$

which in turn gives the final form of Equation (8) after adding equations (19) and (20) and dividing the results by two:

$$\dot{\mathbf{v}}^k = -\Gamma^k_{ij}\left(\boldsymbol{\theta}\right)\mathbf{v}^i\mathbf{v}^j - \mathrm{g}^{kl}\left(\boldsymbol{\theta}\right)\partial_l\phi\left(\boldsymbol{\theta}\right)$$

Here, $\Gamma^k_{ij}\left(\boldsymbol{\theta}\right) := \frac{1}{2}\mathrm{g}^{kl}\left(\partial_i \mathbf{g}_{lj} + \partial_j \mathbf{g}_{il} - \partial_l \mathrm{g}_{ij}\right)$ are the Christoffel symbols of second kind.

Note that the new dynamics (8) still preserves the original Hamiltonian $H(\boldsymbol{\theta}, \mathbf{p} = \mathbf{G}(\boldsymbol{\theta})\mathbf{v})$. This is of course intuitive, but it also can be proven as follows:

$$\begin{aligned}\frac{d}{dt}H\left(\boldsymbol{\theta}.\mathbf{G}\left(\boldsymbol{\theta}\right)\mathbf{v}\right) &= \dot{\boldsymbol{\theta}}^{\mathsf{T}}\frac{\partial}{\partial\boldsymbol{\theta}}H\left(\boldsymbol{\theta},\mathbf{G}\left(\boldsymbol{\theta}\right)\mathbf{v}\right)+\dot{\mathbf{v}}^{\mathsf{T}}\frac{\partial}{\partial\mathbf{v}}H\left(\boldsymbol{\theta},\mathbf{G}\left(\boldsymbol{\theta}\right)\mathbf{v}\right)\\ &= \mathbf{v}^{\mathsf{T}}\left[\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}\right)+\frac{1}{2}\mathbf{v}^{\mathsf{T}}\partial\mathbf{G}\left(\boldsymbol{\theta}\right)\mathbf{v}\right]+\left[-\mathbf{v}^{\mathsf{T}}\Gamma\left(\boldsymbol{\theta}\right)\mathbf{v}-\mathbf{G}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}\right)\right]^{\mathsf{T}}\mathbf{G}\left(\boldsymbol{\theta}\right)\mathbf{v}\\ &= \mathbf{v}^{\mathsf{T}}\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}\right)-\left(\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}\right)\right)^{\mathsf{T}}\mathbf{v}+\frac{1}{2}\mathbf{v}^{\mathsf{T}}\left(\mathbf{v}^{\mathsf{T}}\partial\mathbf{G}\left(\boldsymbol{\theta}\right)\mathbf{v}\right)-\left(\mathbf{v}^{\mathsf{T}}\tilde{\Gamma}\left(\boldsymbol{\theta}\right)\mathbf{v}\right)^{\mathsf{T}}\mathbf{v}\\ &= 0+0=0\end{aligned}$$

where $\mathbf{v}^{\mathsf{T}}\Gamma\left(\boldsymbol{\theta}\right)\mathbf{v}$ is a vector whose *k*th element is $\Gamma_{ij}^{k}\left(\boldsymbol{\theta}\right)\mathbf{v}^{i}\mathbf{v}^{j}$. The second 0 is due to the triple form $\left(\mathbf{v}^{\mathsf{T}}\tilde{\Gamma}\left(\boldsymbol{\theta}\right)\mathbf{v}\right)^{\mathsf{T}}\mathbf{v}=\tilde{\Gamma}_{ijk}\mathbf{v}^{i}\mathbf{v}^{j}\mathbf{v}^{k}=\frac{1}{2}\partial_{k}\mathrm{g}_{ij}\mathbf{v}^{i}\mathbf{v}^{j}\mathbf{v}^{k}$, where $\tilde{\Gamma}$ is the Christoffel symbol of first kind with elements $\tilde{\Gamma}_{ijk}\left(\boldsymbol{\theta}\right):=\mathrm{g}_{kl}\Gamma_{ij}^{l}\left(\boldsymbol{\theta}\right)=\frac{1}{2}\left(\partial_{i}\mathrm{g}_{kj}+\partial_{j}\mathrm{g}_{ik}-\partial_{k}\mathrm{g}_{ij}\right)$.

B Derivation of semi-explicit Lagrangian Monte Carlo (sLMC)

To construct a time-reversible integrator for (8), we concatenate a half step of the following Euler-B integrator of (8) (Leimkuhler and Reich, 2004, chap 4):

$$\begin{aligned}\boldsymbol{\theta}^{(n+1/2)} &= \boldsymbol{\theta}^{(n)}+\frac{\varepsilon}{2}\mathbf{v}^{(n+1/2)}\\ \mathbf{v}^{(n+1/2)} &= \mathbf{v}^{(n)}-\frac{\varepsilon}{2}\left[\left(\mathbf{v}^{(n+1/2)^{\mathsf{T}}}\Gamma\left(\boldsymbol{\theta}^{(n)}\right)\mathbf{v}^{(n+1/2)}+\mathbf{G}\left(\boldsymbol{\theta}^{(n)}\right)^{-1}\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}^{(n)}\right)\right]\end{aligned}$$

with another half step of its adjoint Euler-A integrator:

$$\begin{aligned}\boldsymbol{\theta}^{(n+1)} &= \boldsymbol{\theta}^{(n+1/2)}+\frac{\varepsilon}{2}\mathbf{v}^{(n+1/2)}\\ \mathbf{v}^{(n+1)} &= \mathbf{v}^{(n+1/2)}-\frac{\varepsilon}{2}\left[\left(\mathbf{v}^{(n+1/2)^{\mathsf{T}}}\Gamma\left(\boldsymbol{\theta}^{(n+1)}\right)\mathbf{v}^{(n+1/2)}+\mathbf{G}\left(\boldsymbol{\theta}^{(n+1)}\right)^{-1}\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}^{(n+1)}\right)\right]\end{aligned}$$

Then we obtain the following semi-explicit integrator:

$$\begin{aligned}\mathbf{v}^{(n+1/2)} &= \mathbf{v}^{(n)}-\frac{\varepsilon}{2}\left[\left(\mathbf{v}^{(n+1/2)^{\mathsf{T}}}\Gamma\left(\boldsymbol{\theta}^{(n)}\right)\mathbf{v}^{(n+1/2)}+\mathbf{G}\left(\boldsymbol{\theta}^{(n)}\right)^{-1}\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}^{(n)}\right)\right]\\ \boldsymbol{\theta}^{(n+1/2)} &= \boldsymbol{\theta}^{(n)}+\varepsilon\mathbf{v}^{(n+1/2)}\\ \mathbf{v}^{(n+1)} &= \mathbf{v}^{(n+1/2)}-\frac{\varepsilon}{2}\left[\left(\mathbf{v}^{(n+1/2)^{\mathsf{T}}}\Gamma\left(\boldsymbol{\theta}^{(n+1)}\right)\mathbf{v}^{(n+1/2)}+\mathbf{G}\left(\boldsymbol{\theta}^{(n+1)}\right)^{-1}\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}^{(n+1)}\right)\right]\end{aligned}$$

The resulting integrator, however, is no longer volume-preserving (see subsection B.2). Nevertheless, based on proposition 1, we can still have detailed balance after determinant adjustment (Also see Peter J. Green, 1995).

Proposition 1 (Detailed Balance Condition with determinant adjustment). *Denote* $\mathbf{z}=(\boldsymbol{\theta},\mathbf{v})$, $\mathbf{z}'=T_{\widehat{L(\mathbf{z})}}$ for some time reversible integrator $\mathrm{T}_{\widehat{\mathrm{L}}}$ to the Lagrangian dynamic. If the acceptance probability is adjusted in the following way:

$$\tilde{\alpha}\left(\mathbf{z},\mathbf{z}'\right)=min\left\{1,\frac{exp\left(-H\left(\mathbf{z}'\right)\right)}{exp\left(-H\left(\mathbf{z}\right)\right)}|det\quad\hat{T}_{L}|\right\}\quad(21)$$

then the detailed balance condition still holds

$$\tilde{\alpha}\left(\mathbf{z},\mathbf{z}'\right)\mathbb{P}\left(d\mathbf{z}\right)=\tilde{\alpha}\left(\mathbf{z}',\mathbf{z}\right)\mathbb{P}\left(d\mathbf{z}'\right) \quad (22)$$

Proof.

$$\begin{aligned}\tilde{\alpha}\left(\mathbf{z},\mathbf{z}'\right)\mathbb{P}\left(d\mathbf{z}\right)= \quad & min\left\{1,\frac{exp\left(-H\left(\mathbf{z}'\right)\right)}{exp(-H(\mathbf{z}))}\left|\frac{d\mathbf{z}'}{d\mathbf{z}}\right|\right\}exp\left(-H\left(\mathbf{z}\right)\right)d\mathbf{z}\\ & \mathbf{z}=\hat{T}_{\frac{L}{=}}^{-1}\left(\mathbf{z}'\right)min\left\{exp\left(-H\left(\mathbf{z}\right)\right),exp\left(-H\left(\mathbf{z}'\right)\right)\left|\frac{d\mathbf{z}'}{d\mathbf{z}}\right|\right\}\left|\frac{d\mathbf{z}}{d\mathbf{z}'}\right|d\mathbf{z}'\\ & =min\left\{1,\frac{exp(-H(\mathbf{z}))}{exp\left(-H\left(\mathbf{z}'\right)\right)}\left|\frac{d\mathbf{z}}{d\mathbf{z}'}\right|\right\}exp\left(-H\left(\mathbf{z}'\right)\right)\mathbf{z}'=\tilde{\alpha}\left(\mathbf{z}',\mathbf{z}\right)\mathbb{P}\left(d\mathbf{z}'\right)\end{aligned}$$

Note, the acceptance probability should be calculated based on $\boldsymbol{H}(\boldsymbol{\theta}, \mathbf{G}(\boldsymbol{\theta})\mathbf{v})$. However, it could also be calculated as follows based on the energy function $\mathbf{E}(\boldsymbol{\theta}, \mathbf{v})$ defined in section 3.

To show their equivalence, we note that $\left|\frac{\partial\left(\boldsymbol{\theta}',\mathbf{p}'\right)}{\partial(\boldsymbol{\theta},\mathbf{p})}\right|=\frac{det\left(\mathbf{G}\left(\boldsymbol{\theta}'\right)\right)}{det(\mathbf{G}(\boldsymbol{\theta}))}\left|\frac{\partial\left(\boldsymbol{\theta}',\mathbf{v}'\right)}{\partial(\boldsymbol{\theta},\mathbf{v})}\right|$ and prove as follows:

$$\begin{aligned}\tilde{\alpha} \quad & =min\left\{1,\frac{exp\left(-H\left(\boldsymbol{\theta}',\mathbf{p}'\right)\right)}{exp(-H(\boldsymbol{\theta},\mathbf{p}))}\left|\frac{\partial\left(\boldsymbol{\theta}',\mathbf{p}'\right)}{\partial(\boldsymbol{\theta},\mathbf{p})}\right|\right\}=min\left\{1,\frac{exp\left(-H\left(\boldsymbol{\theta}',\mathbf{G}\left(\boldsymbol{\theta}'\right)\mathbf{v}'\right)\right)}{exp(-H(\boldsymbol{\theta},\mathbf{G}(\boldsymbol{\theta})\mathbf{v}))}\left|\frac{\partial\left(\boldsymbol{\theta}',\mathbf{p}'\right)}{\partial(\boldsymbol{\theta},\mathbf{p})}\right|\right\}\\ & =min\left\{1,\frac{epx\left\{-\left(log\,p\left(\boldsymbol{\theta}'\right)+\frac{1}{2}\,log\,det\,\mathbf{G}\left(\boldsymbol{\theta}'\right)+\frac{1}{2}\mathbf{v}'^{\mathsf{T}}\,\mathbf{G}\left(\boldsymbol{\theta}'\right)\mathbf{v}'\right)\right\}det\left(\mathbf{G}\left(\boldsymbol{\theta}'\right)\right)}{exp\left\{-\left(log\,p(\boldsymbol{\theta})+\frac{1}{2}\,log\,det\mathbf{G}(\boldsymbol{\theta})+\frac{1}{2}\mathbf{v}^{\mathsf{T}}\,\mathbf{G}(\boldsymbol{\theta})\mathbf{v}\right)\right\}det(\mathbf{G}(\boldsymbol{\theta}))}\left|\frac{\partial\left(\boldsymbol{\theta}',\mathbf{v}'\right)}{\partial(\boldsymbol{\theta},\mathbf{v})}\right|\right\}\\ & =min\left\{1,\frac{exp:\left\{-\left(log\,p\left(\boldsymbol{\theta}'\right)-\frac{1}{2}\,log\,det\,\mathbf{G}\left(\boldsymbol{\theta}'\right)+\frac{1}{2}\mathbf{v}'^{\mathsf{T}}\,\mathbf{G}\left(\boldsymbol{\theta}'\right)\mathbf{v}'\right)\right\}}{exp\left\{-\left(log\,p(\boldsymbol{\theta})-\frac{1}{2}\,log\,det\,\mathbf{G}(\boldsymbol{\theta})+\frac{1}{2}\mathbf{v}^{\mathsf{T}}\,\mathbf{G}(\boldsymbol{\theta})\mathbf{v}\right)\right\}}\left|\frac{\partial\left(\boldsymbol{\theta}',\mathbf{v}'\right)}{\partial(\boldsymbol{\theta},\mathbf{v})}\right|\right\}\\ & =min\left\{1,\frac{exp\left(-\mathbf{E}\left(\boldsymbol{\theta}',\mathbf{v}'\right)\right)}{exp(-\mathbf{E}(\boldsymbol{\theta},\mathbf{v}))}\left|\frac{\partial\left(\boldsymbol{\theta}',\mathbf{v}'\right)}{\partial(\boldsymbol{\theta},\mathbf{v})}\right|\right\}\end{aligned}$$

B.1 Distribution Invariance with Volume Correction

Now with proposition 1 we can prove that the Markov Chain derived by our reversible integrator with adjusted acceptance probability (21) converges to the true target density. One can also find the similar proof in (Liu, 2001, chap 9).

Starting from position $\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\boldsymbol{X})$ at time 0, we generate a velocity $\mathbf{v} \sim \mathcal{N}(0, \mathbf{G}(\boldsymbol{\theta}))$. Then evolve $(\boldsymbol{\theta}, \mathbf{v})$ according to our time reversible integrator $\hat{T}$ to reach reach a new state $(\boldsymbol{\theta}^*, \mathbf{v}^*)$ with $\boldsymbol{\theta}^* \sim f(\boldsymbol{\theta}^*)$. We want to prove that $f(\cdot) = p(\cdot|X)$, which can be done by showing $\mathrm{E}_f[h(\boldsymbol{\theta}^*)] = \mathrm{E}_{\theta|X}[h(\boldsymbol{\theta}^*)]$ for any square integrable function h. Denote $\mathbf{z} := (\boldsymbol{\theta}, \mathbf{v})$ and $\mathbb{P}(d\mathbf{z}) := exp(-\mathbf{E}(\mathbf{z}))\,d\mathbf{z}$.

Note that $\mathbf{z}^* = (\boldsymbol{\theta}^*, \mathbf{v}^*)$ can be reached from two scenarios: either the proposal is accepted or rejected. Therefore,

$$\begin{aligned}\mathrm{E}_f\left[h\left(\boldsymbol{\theta}^*\right)\right] = & \int h\left(\boldsymbol{\theta}^*\right)\left[\mathbb{P}\left(d\hat{T}^{-1}\left(\mathbf{z}^*\right)\right)\tilde{\alpha}\left(\hat{T}^{-1}\left(\mathbf{z}^*\right),\mathbf{z}^*\right)+\mathbb{P}\left(d\mathbf{z}^*\right)\left(1-\tilde{\alpha}\left(\mathbf{z}^*,\hat{T}\left(\mathbf{z}^*\right)\right)\right)\right] \\ = & \int h\left(\boldsymbol{\theta}^*\right)\mathbb{P}\left(d\mathbf{z}^*\right)+\int h\left(\boldsymbol{\theta}^*\right)\left[\mathbb{P}\left(d\hat{T}^{-1}\left(\mathbf{z}^*\right)\right)\tilde{\alpha}\left(\hat{T}^{-1}\left(\mathbf{z}^*\right),\mathbf{z}^*\right)-\mathbb{P}\left(d\mathbf{z}^*\right)\tilde{\alpha}\left(\mathbf{z}^*,\hat{T}\left(\mathbf{z}\right)\right)\right]\end{aligned}$$

So it suffices to prove

$$\int h\left(\boldsymbol{\theta}^*\right)\mathbb{P}\left(d\hat{T}^{-1}\left(\mathbf{z}^*\right)\right)\tilde{\alpha}\left(\hat{T}^{-1}\left(\mathbf{z}^*\right)\mathbf{z}^{ast}\right)=\int h\left(\boldsymbol{\theta}^*\right)\mathbb{P}\left(d\mathbf{z}^*\right)\tilde{\alpha}\left(\mathbf{z}^*,\hat{T}\left(\mathbf{z}^*\right)\right) \quad (23)$$

Denote the involution $\nu : (\boldsymbol{\theta}, v) \to (\boldsymbol{\theta}, -v)$. First, by time reversibility we have $\hat{T^{-1}}(\mathbf{z}^*) = \nu\hat{T}\nu(\mathbf{z}^*)$. Further, we claim $\tilde{\alpha}\left(\nu\left(\mathbf{z}\right),\mathbf{z}^{'}\right)=\tilde{\alpha}\left(\mathbf{z},\nu\left(\mathbf{z}^{'}\right)\right)$. This is true because: i) **E** is quadratic in v so $\mathbf{E}(\nu(\mathbf{z})) = \mathbf{E}(\mathbf{z})$; ii) $|\frac{d\mathbf{z}^{'}}{d\nu(\mathbf{z})}|=|\frac{d\nu\left(\mathbf{z}^{'}\right)}{d\nu(\nu(\mathbf{z}))}|=|\frac{d\nu\left(\mathbf{z}^{'}\right)}{d\mathbf{z}}|$. Then that follows from definition of the adjusted acceptance probability (21) and the equivalence discussed under proposition 1. Therefore

$$\begin{aligned}\int h\left(\boldsymbol{\theta}\right)\mathbb{P}\left(d\hat{T}^{-1}\left(\mathbf{z}^*\right)\right)\tilde{\alpha}\left(\hat{T}^{-1}\left(\mathbf{z}^*\right),\mathbf{z}^*\right) = & \int h\left(\boldsymbol{\theta}^*\right)\mathbb{P}\left(d\nu\hat{T}\nu\left(\mathbf{z}^*\right)\right)\tilde{\alpha}\left(\nu\hat{T}\nu\left(\mathbf{z}^*\right),\mathbf{z}^*\right) \\ = & \int h\left(\boldsymbol{\theta}^*\right)\mathbb{P}\left(d\hat{T}\nu\left(\mathbf{z}^*\right)\right)\tilde{\alpha}\left(\hat{T}\nu\left(\mathbf{z}^*\right),\nu\left(\mathbf{z}^*\right)\right)\end{aligned} \quad (24)$$

Next, applying the detailed balance condition (22) to $\nu(\mathbf{z}^*)$ we get

$$\mathbb{P}\left(d\hat{T}\nu\left(\mathbf{z}^*\right)\right)\tilde{\alpha}\left(\hat{T}\nu\left(\mathbf{z}^*\right),\nu\left(\mathbf{z}^*\right)\right)=\mathbb{P}\left(d\nu\left(\mathbf{z}^*\right)\right)\tilde{\alpha}\left(\nu\left(\mathbf{z}^*\right),\hat{T}\nu\left(\mathbf{z}^*\right)\right)$$

substitute it in (24) and continue,

$$\begin{aligned}&=\int h\left(\boldsymbol{\theta}^*\right)\mathbb{P}\left(d\nu\left(\mathbb{z}^*\right)\right)\tilde{\alpha}\left(\nu\left(\mathbf{z}^*\right),\hat{T}\nu\left(\mathbf{z}^*\right)\right) \\ &\overset{\nu(\mathbf{z}^*)\mapsto}{=} \int h\left(\boldsymbol{\theta}^*\right)\mathbb{P}\left(d\mathbf{z}^*\right)\tilde{\alpha}\left(\mathbf{z}^*,\hat{T}\left(\mathbf{z}^*\right)\right)\end{aligned}$$

Therefore, Equation (23) holds, and we complete the proof.

B.2 Volume Correction

To adjust volume, we must derive the Jacobian determinant, $\det \mathbf{J}:=|\frac{\partial\left(\boldsymbol{\theta}^{(L+1)},\mathbf{v}^{(L+1)}\right)}{\partial\left(\boldsymbol{\theta}^{(1)},\mathbf{v}^{(1)}\right)}|$, which can be calculated using wedge products.

Definition 1 (Differential Forms, Wedge Product). *The differential one-form* $\alpha{:}TM^D \to \mathbb{R}$ on a differential manifold $\mathrm{M^D}$ is a smooth mapping from tangent space $\mathrm{T\ M^D}$ to $\mathbb{R}$, which can be expressed as a linear combination of di erentials of local coordinates: $\alpha = f_i dx^i =: f \cdot dx$.

For example, if $f:\mathbb{R}^D \to \mathbb{R}$ is a smooth function, then its directional derivative along a vector $v \in \mathbb{R}^D$, *denoted by df(v) is given by*

$$df\,(v) = \frac{\partial f}{\partial z_i} v^i$$

then df(·) is a linear functional of v, called the differential of f at z and is an example of a differential one-form. In particular, $dz^i(v) = v^i$, *thus*

$$df\,(v) = \frac{\partial f}{\partial z_i} dz^i\,(v)\,, \quad then \quad df = \frac{\partial f}{\partial z_i} dz^i$$

The wedge Product of two one-forms α, β is a 2-form αΛβ, an anti-symmetric bilinear function on tangent space which has the following properties (α, β, γ one-forms, A be a square matrix of same dimension D):

- $\alpha \wedge \alpha = 0$
- $\alpha \wedge (\beta + \gamma) = \alpha \wedge \beta + \alpha \wedge \gamma$ (thus $\alpha \wedge \beta = -\beta \wedge \alpha$)
- $\alpha \wedge A\beta = A^T \alpha \wedge \beta$

The following proposition enables us to calculate the Jacobian determinant denoted as det **J.**

Proposition 2. *Let* $T_L : (\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)}) \to (\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)})$ *be evolution of a smooth flow, then*

$$d\boldsymbol{\theta}^{(L+1)} \wedge d\mathbf{v}^{(L+1)} = \frac{\partial\left(\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)}\right)}{\partial\left(\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)}\right)} d\boldsymbol{\theta}^{(1)} \wedge d\mathbf{v}^{(1)}$$

Note that the Jacobian determinant det **J** can also be regarded as a Radon-Nikodym derivative of two probability measures: $\mathbf{J} = \frac{\left(d\boldsymbol{\theta}^{(L+1)}, d\mathbf{v}^{(L+1)}\right)}{\left(d\boldsymbol{\theta}^{(1)}, d\mathbf{v}^{(1)}\right)}$, where $\mathbb{P}\,(d\boldsymbol{\theta}, \mathbf{dv}) = p\,(\boldsymbol{\theta}, \mathbf{v})\, d\boldsymbol{\theta}\mathbf{dv}$. We have

$$\begin{aligned} d\mathbf{v}^{(m+1/2)} &= d\mathbf{v}^{(n)} - \varepsilon\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}} \Gamma\left(\boldsymbol{\theta}^{(n)}\right) d\mathbf{v}^{(n+1/2)} + (**)\, d\boldsymbol{\theta}^{(n)} \\ d\boldsymbol{\theta}^{(n+1)} &= d\boldsymbol{\theta}^{(n)} + \varepsilon d\mathbf{v}^{(n+1/2)} \\ d\mathbf{v}^{(n+1)} &= d\mathbf{v}^{(n+1/2)} - \varepsilon\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}} \Gamma\left(\boldsymbol{\theta}^{(n+1)}\right) d\mathbf{v}^{(n+1/2)} + (**)\, d\boldsymbol{\theta}^{(n+1)} \end{aligned}$$

where $\mathbf{v}^{\mathsf{T}}\,\Gamma\,(\boldsymbol{\theta})$ is a matrix whose (*k*, *j*)th element is $\mathbf{v}^i \Gamma^k_{ij}\,(\theta)$. Therefore,

$$
\begin{aligned}
d\boldsymbol{\theta}^{(n+1)} \wedge d\mathbf{v}^{(n+1)} &= \left[\mathbf{I} - \varepsilon\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}} \Gamma\left(\boldsymbol{\theta}^{(n+1)}\right)\right]^{\mathsf{T}} d\boldsymbol{\theta}^{(n+1)} \wedge d\mathbf{v}^{(n+1/2)} \\
&= \left[\mathbf{I} - \varepsilon\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}} \Gamma\left(\boldsymbol{\theta}^{(n+1)}\right)\right]^{\mathsf{T}} d\boldsymbol{\theta}^{(n)} \wedge d\mathbf{v}^{(n+1/2)} \\
&= \left[\mathbf{I} - \varepsilon\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}} \Gamma\left(\boldsymbol{\theta}^{(n+1)}\right)\right]^{\mathsf{T}} \left[\mathbf{I} + \varepsilon\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}} \Gamma\left(\boldsymbol{\theta}^{(n)}\right)\right]^{\mathsf{T}} d\boldsymbol{\theta}^{(n)} \wedge d\mathbf{v}^{(n)}
\end{aligned}
$$

For volume adjustment, we must use the following Jacobian determinant accumulated along the integration steps:

$$
det \quad \mathbf{J}_{sLMC} := \left|\frac{\partial\left(\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)}\right)}{\partial\left(\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)}\right)}\right| = \prod_{n=1}^{L} \frac{det\left(I - \varepsilon\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}} \Gamma\left(\boldsymbol{\theta}^{(n+1)}\right)\right)}{det\left(I + \varepsilon\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}} \Gamma\left(\boldsymbol{\theta}^{(n)}\right)\right)} \tag{25}
$$

As a result, the acceptance probability becomes

$$
\alpha_{sLMC} = min\left\{1, exp\left(-\mathbf{E}\left(\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)}\right) + \mathbf{E}\left(\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)}\right)\right) |det \quad \mathbf{J}_{sLMC}|\right\}
$$

C Derivation of explicit Lagrangian Monte Carlo (LMC)

To resolve the remaining implicit equation (9), we now propose an additional modification motivated by the following relationship (notice the symmetry of lower indices in):

$$
\mathbf{v}^{\mathsf{T}} \Gamma \mathbf{u} = \frac{1}{2}\left[(\mathbf{v}+\mathbf{u})^{\mathsf{T}} \Gamma (\mathbf{v}+\mathbf{u}) - \mathbf{v}^{\mathsf{T}} \Gamma \mathbf{v} - \mathbf{u}^{\mathsf{T}} \Gamma \mathbf{u}\right]
$$

To keep time-reversibility, we make the modification to both (9) and (11) as shown below:

$$
\begin{aligned}
\mathbf{v}^{(n+1/2)} &= \mathbf{v}^{(n)} - \tfrac{\varepsilon}{2}\left[\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}} \Gamma\left(\boldsymbol{\theta}^{(n)}\right) \mathbf{v}^{(n+1/2)} + \mathbf{G}\left(\boldsymbol{\theta}^{(n)}\right)^{-1} \nabla_{\boldsymbol{\theta}} \phi\left(\boldsymbol{\theta}^{(n)}\right)\right] \\
&\Downarrow \\
\mathbf{v}^{(n+1/2)} &= \mathbf{v}^{(n)} - \tfrac{\varepsilon}{2}\left[\left(\mathbf{v}^{(n)}\right)^{\mathsf{T}} \Gamma\left(\boldsymbol{\theta}^{(n)}\right) \mathbf{v}^{(n+1/2)} + \mathbf{G}\left(\boldsymbol{\theta}^{(n)}\right)^{-1} \nabla_{\boldsymbol{\theta}} \phi\left(\boldsymbol{\theta}^{(n)}\right)\right]
\end{aligned} \tag{26}
$$

$$
\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} + \varepsilon \mathbf{v}^{(n+1/2)} \tag{27}
$$

$$
\begin{aligned}
\mathbf{v}^{(n+1)} &= \mathbf{v}^{(n+1/2)} - \tfrac{\varepsilon}{2}\left[\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}} \Gamma\left(\boldsymbol{\theta}^{(n+1)}\right) \mathbf{v}^{(n+1/2)} + \mathbf{G}\left(\boldsymbol{\theta}^{(n+1)}\right)^{-1} \nabla_{\boldsymbol{\theta}} \phi\left(\boldsymbol{\theta}^{(n+1)}\right)\right] \\
&\Downarrow \\
\mathbf{v}^{(n+1)} &= \mathbf{v}^{(n+1/2)} - \tfrac{\varepsilon}{2}\left[\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}} \Gamma\left(\boldsymbol{\theta}^{(n+1)}\right) \mathbf{v}^{(n+1)} + \mathbf{G}\left(\boldsymbol{\theta}^{(n+1)}\right)^{-1} \nabla_{\boldsymbol{\theta}} \phi\left(\boldsymbol{\theta}^{(n+1)}\right)\right]
\end{aligned} \tag{28}
$$

The time-reversibility of the integrator (26)-(28) can be shown by switching $(\boldsymbol{\theta}, \mathbf{v})^{(n+1)}$ and $(\boldsymbol{\theta}, \mathbf{v})^{(n)}$ and negating velocity. The resulting integrator is completely explicit since both

updates of velocity (26) and (28) can be solved by collecting terms containing $\mathbf{v}^{(n+1/2)}$ and $\mathbf{v}^{(n+1)}$ respectively:

$$\mathbf{v}^{(n+1/2)}=\left[\mathbf{I}+\frac{\varepsilon}{2}\left(\mathbf{v}^{(n)}\right)^{\mathsf{T}}\Gamma\left(\boldsymbol{\theta}^{(n)}\right)\right]^{-1}\left[\mathbf{v}^{(n)}-\frac{\varepsilon}{2}\mathbf{G}\left(\boldsymbol{\theta}^{(n)}\right)^{-1}\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}^{(n)}\right)\right] \quad (29)$$

$$\mathbf{v}^{(n+1)}=\left[\mathbf{I}+\frac{\varepsilon}{2}\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}}\Gamma\left(\boldsymbol{\theta}^{(n+1)}\right)\right]^{-1}\left[\mathbf{v}^{(n+1/2)}-\frac{\varepsilon}{2}\mathbf{G}\left(\boldsymbol{\theta}^{(n+1)}\right)^{-1}\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}^{(n+1)}\right)\right] \quad (30)$$

C.1 Convergence of Numerical Solution

We now show that the discretization error

$e_n=\|\mathbf{z}(t_n)-\mathbf{z}^{(n)}\|=\|(\boldsymbol{\theta}(\mathbf{t_n}),\mathbf{v}(\mathbf{t_n}))-\left(\boldsymbol{\theta}^{(n)},\mathbf{v}^{(n)}\right)\|$ (i.e. the difference between the true solution and the numerical solution) accumulated over final time interval [0, T] is bounded and goes to zero as the stepsize ε goes to zero. (See Leimkuhler and Reich (2004) for a similar proof for the generalized leapfrog method.) Here, we assume that

$\mathbf{f}(\boldsymbol{\theta},\mathbf{v}):=\mathbf{v}^{\mathsf{T}}\Gamma(\boldsymbol{\theta})\mathbf{v}+\mathbf{G}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta})$ is smooth; hence, f and its derivatives are uniformly bounded as ($\boldsymbol{\theta}$, $\mathbf{v}$) evolves within finite time duration T . We expand the true solution $\mathbf{z}(t_{n+1})$ at t_n:

$$\begin{aligned}\mathbf{z}(t_{n+1})&=\mathbf{z}(t_n)+\dot{\mathbf{z}}(t_n)\varepsilon+\tfrac{1}{2}\ddot{\mathbf{z}}(t_n)\varepsilon^2+o(\varepsilon^2)\\&=\begin{bmatrix}\boldsymbol{\theta}(\mathbf{t_n})\\\mathbf{v}(t_n)\end{bmatrix}+\begin{bmatrix}\mathbf{v}(t_n)\\-\mathbf{f}(\boldsymbol{\theta}(\mathbf{t_n}),\mathbf{v}(\mathbf{t_n}))\end{bmatrix}\varepsilon+\tfrac{1}{2}\begin{bmatrix}-\mathbf{f}(\boldsymbol{\theta}(\mathbf{t_n}),\mathbf{v}(\mathbf{t_n}))\\-\frac{\partial\mathbf{f}}{\partial\boldsymbol{\theta}}\mathbf{v}(t_n)+\frac{\partial\mathbf{f}}{\partial\mathbf{v}}\mathbf{f}(\boldsymbol{\theta}(\mathbf{t_n}),\mathbf{v}(\mathbf{t_n}))\end{bmatrix}\varepsilon^2+o(\varepsilon^2)\\&=\begin{bmatrix}\boldsymbol{\theta}(\mathbf{t_n})\\\mathbf{v}(t_n)\end{bmatrix}+\begin{bmatrix}\mathbf{v}(t_n)\\-\mathbf{f}(\boldsymbol{\theta}(\mathbf{t_n}),\mathbf{v}(\mathbf{t_n}))\end{bmatrix}\varepsilon+O(\varepsilon^2)\end{aligned}$$

Next, we simplify the expression of the numerical solutions $\mathbf{z}^{(n+1)}=\begin{bmatrix}\boldsymbol{\theta}^{(n+1)}\\\mathbf{v}^{(n+1)}\end{bmatrix}$ for the fully explicit integrator and compare it to the above true solutions. To this end, we rewrite equation (29) as follows:

$$\begin{aligned}\mathbf{v}^{(n+1/2)}&=\left[\mathbf{I}+\tfrac{\varepsilon}{2}\left(\mathbf{v}^{(n)}\right)^{\mathsf{T}}\Gamma\left(\boldsymbol{\theta}^{(n)}\right)\right]^{-1}\left[\mathbf{v}^{(n)}-\tfrac{\varepsilon}{2}\mathbf{G}\left(\boldsymbol{\theta}^{(n)}\right)^{-1}\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}^{(n)}\right)\right]\\&=\mathbf{v}^{(n)}-\left[\mathbf{I}+\tfrac{\varepsilon}{2}\left(\mathbf{v}^{(n)}\right)^{\mathsf{T}}\Gamma\left(\boldsymbol{\theta}^{(n)}\right)\right]^{-1}\left[\left(\mathbf{v}^{(n)}\right)^{\mathsf{T}}\Gamma\left(\boldsymbol{\theta}^{(n)}\right)\mathbf{v}^{(n)}+\tfrac{\varepsilon}{2}\mathbf{G}\left(\boldsymbol{\theta}^{(n)}\right)^{-1}\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}^{(n)}\right)\right]\\&=\mathbf{v}^{(n)}-\left[\mathbf{I}+\tfrac{\varepsilon}{2}\left(\mathbf{v}^{(n)}\right)^{\mathsf{T}}\Gamma\left(\boldsymbol{\theta}^{(n)}\right)\right]^{-1}\tfrac{\varepsilon}{2}\mathbf{f}\left(\boldsymbol{\theta}^{(n)},\mathbf{v}^{(n)}\right)\\&=\mathbf{b}^{(n)}-\tfrac{\varepsilon}{2}\mathbf{f}\left(\boldsymbol{\theta}^{(n)},\mathbf{v}^{(n)}\right)+\tfrac{\varepsilon^2}{4}\left[\mathbf{I}+\tfrac{\varepsilon}{2}\left(\mathbf{v}^{(n)}\right)^{\mathsf{T}}\Gamma\left(\boldsymbol{\theta}^{(n)}\right)\right]^{-1}\left[\left(\mathbf{v}^{(n)}\right)\right]^{-1}\left[\left(\mathbf{v}^{(n)}\right)^{\mathsf{T}}\Gamma\left(\boldsymbol{\theta}^{(n)}\right)\right]\mathbf{f}\left(\boldsymbol{\theta}^{(n)},\mathbf{v}^{(n)}\right)\\&=\mathbf{v}^{(n)}-\tfrac{\varepsilon}{2}\mathbf{f}\left(\boldsymbol{\theta}^{(n)},\mathbf{v}^{(n)}\right)+O(\varepsilon^2)\end{aligned}$$

Similarly, from equation (30) we have

$$\mathbf{v}^{(n+1)}=\mathbf{v}^{(n+1/2)}-\frac{\varepsilon}{2}\mathbf{f}\left(\boldsymbol{\theta}^{(n+1)},\mathbf{v}^{(n+1/2)}\right)+O\left(\varepsilon^2\right)$$

Substituting $\mathbf{v}^{(n+1/2)}$ in the above equation, we obtain $\mathbf{v}^{(n+1)}$ as follows:

$$\begin{aligned}\mathbf{v}^{(n+1)} &=\mathbf{v}^{(n)}-\tfrac{\varepsilon}{2}\mathbf{f}\left(\boldsymbol{\theta}^{(n)},\mathbf{v}^{(n)}\right)-\tfrac{\varepsilon}{2}\mathbf{f}\left(\boldsymbol{\theta}^{(n+1)},\mathbf{v}^{(n)}\right)+O\left(\varepsilon^2\right)\\ &=\mathbf{v}^{(n)}-\mathbf{f}\left(\boldsymbol{\theta}^{(n)},\mathbf{v}^{(n)}\right)\varepsilon+\tfrac{\varepsilon}{2}\left[\mathbf{f}\left(\boldsymbol{\theta}^{(n)},\mathbf{v}^{(n)}\right)-\mathbf{f}\left(\boldsymbol{\theta}^{(n)}+O\left(\varepsilon\right),\mathbf{v}^{(n)}\right]+O\left(\varepsilon^2\right)\\ &=\mathbf{v}^{(n)}-\mathbf{f}\left(\boldsymbol{\theta}^{(n)},\mathbf{v}^{(n)}\right)\varepsilon+O\left(\varepsilon^2\right)\end{aligned}$$

From (29), (27), and (30), we have the following numerical solution:

$$\mathbf{z}^{(n+1)}=\begin{bmatrix}\boldsymbol{\theta}^{(n+1)}\\ \mathbf{v}^{(n+1)}\end{bmatrix}=\begin{bmatrix}\boldsymbol{\theta}^{(n)}\\ \mathbf{v}^{(n)}\end{bmatrix}+\begin{bmatrix}\mathbf{v}^{(n)}\\ -\mathbf{f}\left(\boldsymbol{\theta}^{(n)},\mathbf{v}^{(n)}\right)\end{bmatrix}\varepsilon+O\left(\varepsilon^2\right)$$

Therefore, the local error is

$$\begin{aligned}\epsilon_{n+1}&=\|\mathbf{z}\left(t_{n+1}\right)\\ &\quad-\mathbf{z}^{(n+1)}\|\\ &=\left\|\begin{bmatrix}\boldsymbol{\theta}\left(\mathbf{t_n}\right)-\boldsymbol{\theta}^{(\mathbf{n})}\\ \mathbf{v}\left(t_n\right)-\mathbf{v}^{(n)}\end{bmatrix}+\begin{bmatrix}\mathbf{v}\left(t_n\right)-\mathbf{v}^{(n)}\\ -\left[\mathbf{f}\left(\boldsymbol{\theta}\left(\mathbf{t_n}\right),\mathbf{v}\left(\mathbf{t_n}\right)\right)-\mathbf{f}\left(\boldsymbol{\theta}^{(n)},\mathbf{v}^{(n)}\right)\right]\end{bmatrix}\varepsilon+O\left(\varepsilon^2\right)\right\|\leq\left(1+M\varepsilon\right)e_n+O\left(\varepsilon^2\right)\end{aligned}$$

where $M=c\quad sup_{i\in[0,T]}\|\nabla\mathbf{f}\left(\boldsymbol{\theta}\left(\mathbf{t}\right),\mathbf{v}\left(\mathbf{t}\right)\right)\|$ for some constant $c > 0$. Accumulating the local errors by iterating the above inequality for $L = T/\varepsilon$ steps provides the following global error:

$$\begin{aligned}e_{n+1} &\leq\left(1+M\varepsilon\right)e_n+O\left(\varepsilon^2\right)\leq\left(1+M\varepsilon\right)^2e_{n-1}+2O\left(\varepsilon^2\right)\leq\cdots\leq\left(1+M\varepsilon\right)^ne_1+nO\left(\varepsilon^2\right)\\ &\leq\left(1+M\varepsilon\right)^L\varepsilon+LO\left(\varepsilon^2\right)\leq\left(e^{MT}+T\right)\varepsilon\rightarrow 0,\quad as\quad\varepsilon\rightarrow 0\end{aligned}$$

C.2 Volume Correction

As before, using the wedge product on the system (29), (27), and (30), the Jacobian matrix is

$$\begin{aligned}\frac{\partial\left(\boldsymbol{\theta}^{(n+1)},\mathbf{v}^{(n+1)}\right)}{\partial\left(\boldsymbol{\theta}^{(n)},\mathbf{v}^{(n)}\right)}= &\left[\mathbf{I}+\tfrac{\varepsilon}{2}\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}}\Gamma\left(\boldsymbol{\theta}^{(n+1)}\right)\right]^{\mathsf{T}}\left[\mathbf{I}-\tfrac{\varepsilon}{2}\left(\mathbf{v}^{(n+1)}\right)^{\mathsf{T}}\Gamma\left(\boldsymbol{\theta}^{(n+1)}\right)\right]^{\mathsf{T}}\cdot\\ &\left[\mathbf{I}+\tfrac{\varepsilon}{2}\left(\mathbf{v}^{\mathsf{T}}\Gamma\left(\boldsymbol{\theta}^{(n)}\right)\right)\right]^{-\mathsf{T}}\left[\mathbf{I}-\tfrac{\varepsilon}{2}\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}}\Gamma\left(\boldsymbol{\theta}^{(n)}\right)\right]^{\mathsf{T}}\end{aligned}$$

As these new equations show, our derived integrator is not symplectic so the acceptance probability needs to be adjusted by the following Jacobian determinant, det **J**, in order to preserve the detailed balance condition:

$$
\begin{aligned}
det \quad \mathbf{J}_{LMC} &:= \left| \frac{\partial\left(\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)}\right)}{\partial\left(\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)}\right)} \right| \\
&= \prod_{n=1}^{L} \frac{det\left(\mathbf{I} - \varepsilon/2\left(\mathbf{v}^{(n+1)}\right)^{\mathsf{T}} \Gamma\left(\boldsymbol{\theta}^{(n+1)}\right)\right) det\left(\mathbf{I} - \varepsilon/2\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}} \Gamma\left(\boldsymbol{\theta}^{(n)}\right)\right)}{det\left(\mathbf{I} - \varepsilon/2\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}} \Gamma\left(\boldsymbol{\theta}^{(n+1)}\right)\right) det\left(\mathbf{I} - \varepsilon/2\left(\mathbf{v}^{(n)}\right)^{\mathsf{T}} \Gamma\left(\boldsymbol{\theta}^{(n)}\right)\right)} \\
&= \prod_{n=1}^{L} \frac{det\left(\mathbf{G}\left(\boldsymbol{\theta}^{(n+1)}\right) - \varepsilon/2\left(\mathbf{v}^{(n+1)}\right)^{\mathsf{T}} \tilde{\Gamma}\left(\boldsymbol{\theta}^{(n+1)}\right)\right) det\left(\mathbf{G}\left(\boldsymbol{\theta}^{(n)}\right) - \varepsilon/2\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}} \tilde{\Gamma}\left(\boldsymbol{\theta}^{(n)}\right)\right)}{det\left(\mathbf{G}\left(\boldsymbol{\theta}^{(n+1)}\right) + \varepsilon/2\left(\mathbf{v}^{(n+1/2)^{\mathsf{T}}} \tilde{\Gamma}\left(\boldsymbol{\theta}^{(n+1)}\right)\right) det\left(\boldsymbol{\theta}^{(n)}\right) + \varepsilon/2\left(\varepsilon^{(n)}\right)^{\mathsf{T}} \tilde{\Gamma}\left(\boldsymbol{\theta}^{(n)}\right)\right)}
\end{aligned} \tag{31}
$$

As a result, the acceptance probability is

$$
\alpha_{LMC} = min\left\{1, exp\left(-\mathbf{E}\left(\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(+1)}\right) + \mathbf{E}\left(\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)}\right)\right) |det \quad \mathbf{J}_{LMC}|\right\}
$$

D Pseudocodes

References

Amari, S.; Nagaoka, H. Methods of Information Geometry. Oxford University Press; 2000.

Beskos A, Pinski FJ, Sanz-Serna JM, Stuart AM. Hybrid Monte Carlo on Hilbert spaces. Stochastic Processes and Applications. 2011; 121(10):2201–2230.

Bishop, Richard L.; Goldberg, Samuel I. Tensor Analysis on Manifolds. Dover Publications, Inc.; 2000.

Chin SA. Explicit symplectic integrators for solving nonseparable Hamiltonians. Physical Review E. 2009; 80:037701.

Christensen OF, Roberts GO, Rosenthal JS. Scaling Limits for the Transient Phase of Local Metropolis-Hastings Algorithms. Journal of the Royal Statistical Society: Series B. Apr; 2005 67(2):253–268.

Duane S, Kennedy AD, Pendleton BJ, Roweth D. Hybrid monte carlo. Physics Letters B. 1987; 195(2):216 – 222.

Dullweber A, Leimkuhler B, McLachlan R. Split-Hamiltonian Methods for Rigid Body Molecular dynamic. Journal of Chemical Physics. 1997; 107(15):5840 – 5852.

Geyer CJ. Practical markov chain monte carlo. Statistical Science. 1992; 7(4):473–483.

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

Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination Biometrika. 1995; 82(4):711–732.

Hoffman MD, Gelman A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. 2011 arXiv:1111.4246v1.

Leimkuhler, B.; Reich, S. Simulating Hamiltonian dynamic. Cambridge University Press; 2004.

Liu, JS. Molecular dynamic and Hybrid Monte Carlo.. In: Liu; Jun, S., editors. Monte Carlo Strategies in Scientific Computing. Springer-Verlag; 2001.

Marin, JM.; Mengersen, KL.; Robert, C. Bayesian modeling and inference on mixtures of distributions.. In: Dey, D.; Rao, CR., editors. Handbook of Statistics Volume 25. Elsevier; 2005.

McLachlan, GJ.; Peel, D. Finite Mixture Models. John Wiley & Sons, Inc.; New York: 2000.

Neal, RM. MCMC using Hamiltonian dynamic.. In: Brooks, S.; Gelman, A.; Jones, G.; Meng, XL., editors. Handbook of Markov Chain Monte Carlo. Chapman and Hall/CRC.; 2010.

Sexton JC, Weingarten DH. Hamiltonian evolution for the hybrid Monte Carlo algorithm. Nuclear Physics B. 1992; 380(3):665–677.

Shahbaba B, Lan S, Johnson W, Neal R. Split hamiltonian monte carlo. Statistics and Computing. 20132013:1–11.

Stathopoulos, V.; Girolami, M. Manifold MCMC for Mixtures.. In: Mengersen, K.; Robert, CP.; Titteringhton, MD., editors. Mixture Estimation and Applications. John Wiley & Sons; 2011.

Verlet, Loup. Computer ”Experiments” on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. Physical Review. 1967; 159:98–103.

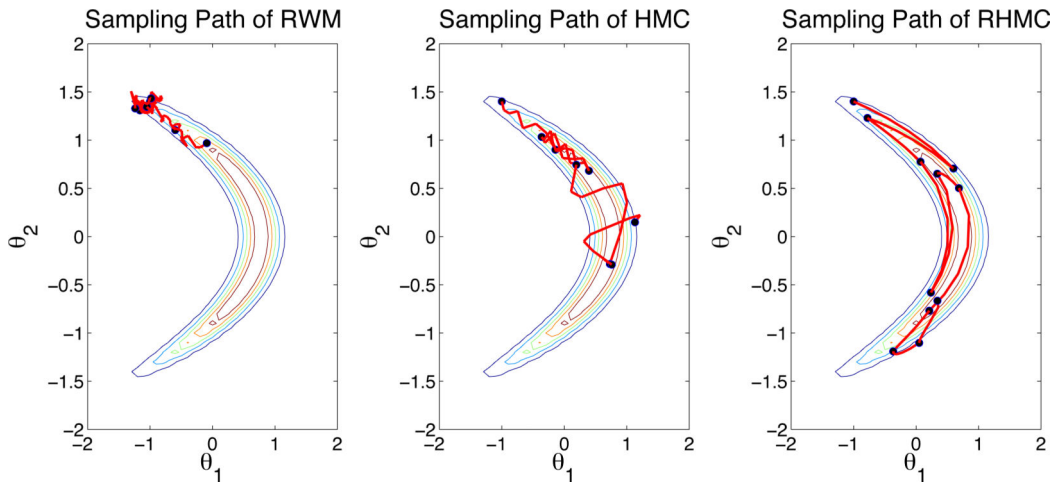

Figure 1.

The first 10 iterations in sampling from a banana shaped distribution with random walk Metropolis (RWM), Hamiltonian Monte Carlo (HMC), and Riemannian HMC (RHMC). For all three methods, the trajectory length (i.e., step size ε times number of integration steps L) is set to 1. For RWM, L=17, for HMC, L=7, and for RHMC, L=5. Solid red lines are the sampling paths, and black circles are the accepted proposals.

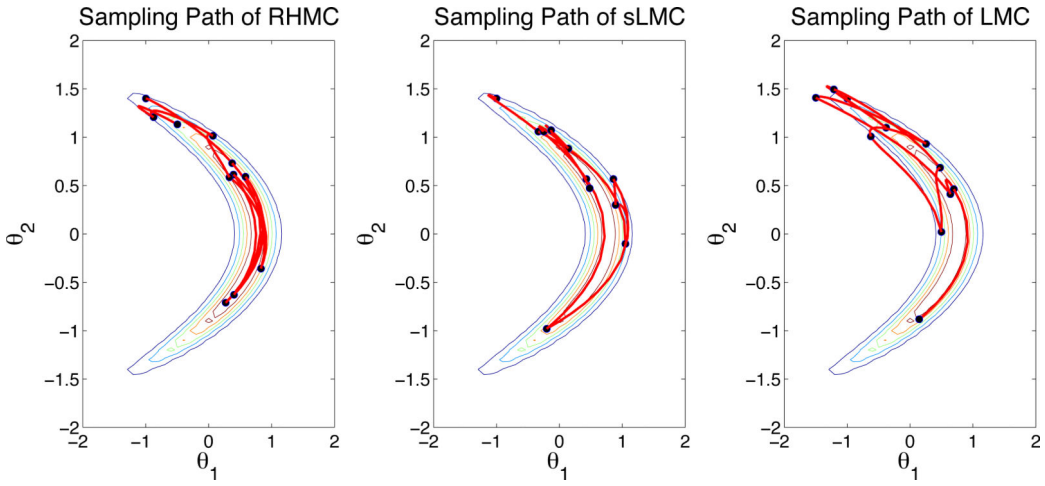

Figure 2.

The first 10 iterations in sampling from the banana-shaped distribution with Riemannian HMC (RHMC), semi-explicit Lagrange Monte Carlo (sLMC) and explicit LMC (LMC). For all three methods, the trajectory length (i.e., step size times number of integration steps) is set to 1.45 and number of integration steps is set to 10. Solid red lines show the sampling path, and each point represents an accepted proposal.

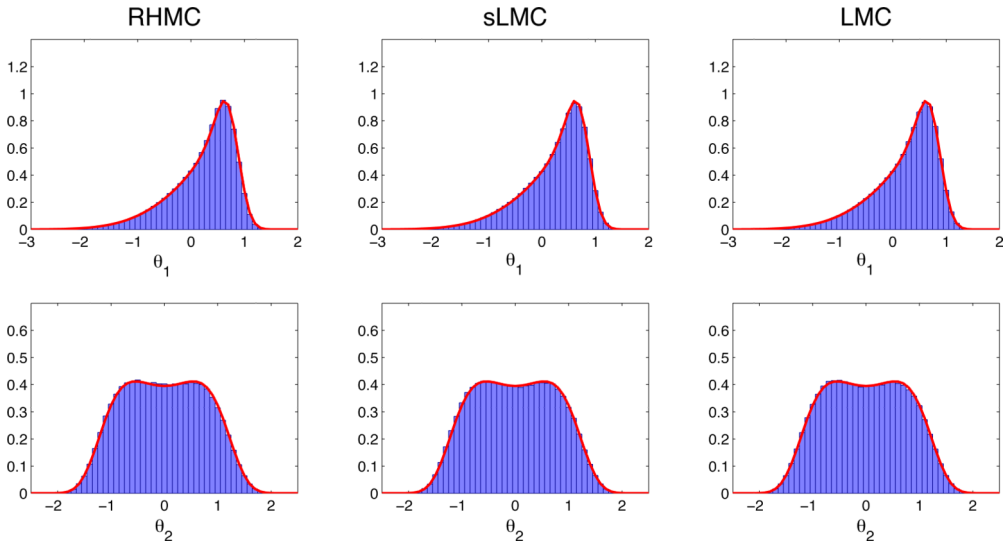

Figure 3.

Histograms of 1 million posterior samples of θ_1 and θ_2 for the banana-shaped distribution using RHMC (left), sLMC (middle) and LMC (right). Solid red curves are the true density functions.

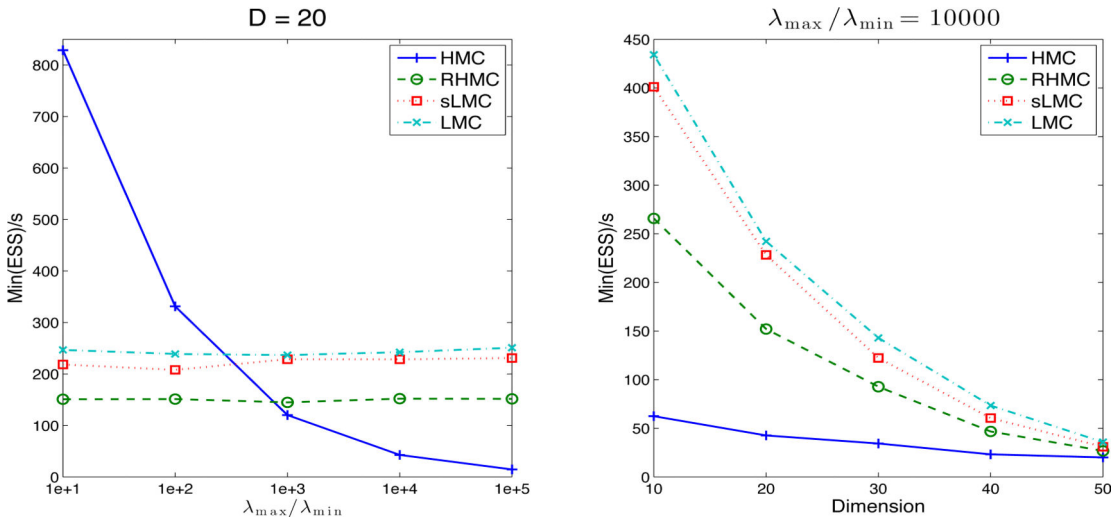

Figure 4.

Left: Sampling efficiency, Min(ESS)/s vs. the condition number for a fixed dimension (D = 20). Right: Sampling efficiency vs dimension for a fixed condition number ($\lambda_{max}/\lambda_{min}$ = 10000). Each algorithm is tuned to have an acceptance rate around 70%. Results are based on 5000 samples after discarding the initial 1000 samples.

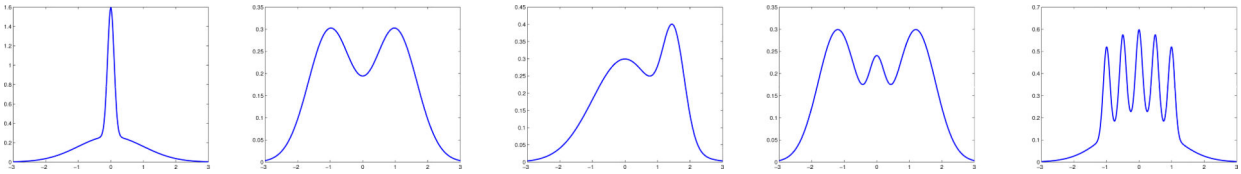

Figure 5.

Densities used to generate synthetic datasets. From left to right the densities are in the same order as in Table 3. The densities are taken from McLachlan and Peel (2000)

Table 1

Comparing alternative methods using a banana-shaped distribution. For each method, we provide the acceptance probability (AP), the CPU time (s) for each iteration, ESS (min., med., max.) and the time-normalized ESS.

Method	AP	s	ESS	min(ESS)/s
HMC	0.79	6.96e-04	(288,614,941)	20.65
RHMC	0.78	4.56e-03	(4514,5779,7044)	49.50
sLMC	0.84	7.90e-04	(2195,3476,4757)	138.98
LMC	0.73	7.27e-04	(1139,2409,3678)	78.32

Table 2

Comparing alternative methods using five binary classification problems discussed in Girolami and Calderhead (2011). For each dataset, the number of predictors, D, and the number of observations, *N*, are specified. For each method, we provide the acceptance probability (AP), the CPU time (s) for each iteration, ESS (min., med., max.) and the time-normalized ESS.

Data	method	AP	s	ESS	min(ESS)/s
Australian D=14,N=690	HMC	0.75	6.13E-03	(1225,3253,10691)	13.32
	RHMC	0.72	2.96E-02	(7825,9238,9797)	17.62
	sLMC	0.83	2.17E-02	(10184,13001,13735)	31.29
	LMC	0.75	1.60E-02	(9636,10443,11268)	40.17
German D=24,N=1000	HMC	0.74	1.31E-02	(766,4006,15000)	3.90
	RHMC	0.76	6.55E-02	(14886,15000,15000)	15.15
	sLMC	0.71	4.13E-02	(13395,15000,15000)	21.64
	LMC	0.70	3.74E-02	(13762,15000,15000)	24.54
Heart D=13,N=270	HMC	0.71	1.75E-03	(378,850,2624)	14.44
	RHMC	0.73	2.12E-02	(6263,7430,8191)	19.68
	sLMC	0.77	1.30E-02	(10318,11337,12409)	52.73
	LMC	0.76	1.15E-02	(10347,10724,11773)	59.80
Pima D=7,N=532	HMC	0.85	5.75E-03	(887,4566,12408)	10.28
	RHMC	0.81	1.64E-02	(4349,4693,5178)	17.65
	sLMC	0.81	8.98E-03	(4784,5437,5592)	35.50
	LMC	0.82	7.90E-03	(4839,5193,5539)	40.84
Ripley D=2,N=250	HMC	0.88	1.50E-03	(820,3077,15000)	36.39
	RHMC	0.74	1.09E-02	(12876,15000,15000)	78.83
	sLMC	0.80	6.79E-03	(15000,15000,15000)	147.38
	LMC	0.79	5.36E-03	(12611,15000,15000)	157.02

Table 3

Densities used for the generation of synthetic Mixture of Gaussian data sets.

Dataset name	Density function	Num. of parameters
Kurtotic	$\frac{2}{3}\mathrm{N}(x \mid 0, 1) + \frac{1}{3}\mathrm{N}\left(x \mid 0, \left(\frac{1}{10}\right)^2\right)$	5
Bimodal	$\frac{1}{2}\mathrm{N}\left(x \mid -1, \left(\frac{2}{3}\right)^2\right) + \frac{1}{2}\mathrm{N}\left(x \mid 1, \left(\frac{2}{3}\right)^2\right)$	5
Skewed	$\frac{3}{4}\mathrm{N}(x \mid 0, 1) + \frac{1}{4}\mathrm{N}\left(x \mid \frac{3}{2}, \left(\frac{1}{3}\right)^2\right)$	5
Trimodal	$\frac{9}{20}\mathrm{N}\left(x \mid -\frac{6}{5}, \left(\frac{3}{5}\right)^2\right) + \frac{9}{20}\mathrm{N}\left(x \mid \frac{6}{5}, \left(\frac{3}{5}\right)^2\right) + \frac{1}{10}\mathrm{N}\left(x \mid 0, \left(\frac{1}{4}\right)^2\right)$	8
Claw	$\frac{1}{2}\mathrm{N}(x \mid 0, 1) + \sum_{i=0}^{4} \frac{1}{10}\mathrm{N}\left(x \mid \frac{i}{2} - 1, \left(\frac{1}{10}\right)^2\right)$	17

Table 4

Acceptance probability (AP), seconds per iteration (s), ESS (min., med., max.) and time-normalized ESS for Gaussian mixture models. Results are calculated on a 5,000 sample chain with a 5,000 sample burn-in session. For HMC the burn-in session was 20,000 samples in order to ensure convergence.

Data	Method	AP	s	ESS	min(ESS)/s
claw	HMC	0.88	7.01E-01	(1916, 3761, 4970)	0.54
	RHMC	0.80	5.08E-01	(1524, 3474, 4586)	0.60
	sLMC	0.86	3.76E-01	(2531, 4332, 5000)	1.35
	LMC	0.82	2.92E-01	(2436, 3455, 4608)	1.67
trimodal	HMC	0.77	3.43E-01	(2244, 2945, 3159)	1.30
	RHMC	0.79	9.94E-02	(4701, 4928, 5000)	9.46
	sLMC	0.82	4.02E-02	(4978, 5000, 5000)	24.77
	LMC	0.80	4.84E-02	(4899, 4982, 5000)	20.21
skewed	HMC	0.83	1.78E-01	(2915, 3237, 3630)	3.27
	RHMC	0.85	5.10E-02	(5000, 5000, 5000)	19.63
	sLMC	0.82	2.26E-02	(4698, 4940, 5000)	41.68
	LMC	0.84	2.52E-02	(4935, 5000, 5000)	39.09
kurtotic	HMC	0.78	2.85E-01	(3013, 3331, 3617)	2.11
	RHMC	0.82	4.72E-02	(5000, 5000, 5000)	21.20
	sLMC	0.85	2.54E-02	(5000, 5000, 5000)	39.34
	LMC	0.81	2.70E-02	(5000, 5000, 5000)	36.90
bimodal	HMC	0.73	1.61E-01	(2923, 2991, 3091)	3.62
	RHMC	0.86	5.38E-02	(5000, 5000, 5000)	18.56
	sLMC	0.81	2.06E-02	(4935, 4996, 5000)	48.00
	LMC	0.85	2.06E-02	(5000, 5000, 5000)	46.43

Algorithm 1

Semi-explicit Lagrangian Monte Carlo (sLMC)

Initialize $\boldsymbol{\theta}^{(1)}$ = current $\boldsymbol{\theta}$

Sample new velocity $\mathbf{v}^{(1)} \sim \boldsymbol{N}(0, \mathbf{G}^{-1}(\boldsymbol{\theta}^{(1)}))$

Calculate current $\mathbf{E}(\boldsymbol{\theta}^{(1)}, \mathbf{v}^{(1)})$ according to equation (12)

for n = 1 to L (leapfrog steps) **do**

% Update the velocity with fixed point iterations

$$\hat{\mathbf{v}}^{(0)} = \mathbf{v}^{(n)}$$

for i = 1 to NumOfFixedPointSteps **do**

$$\hat{\mathbf{v}}^{(i)} = \mathbf{v}^{(n)} - \frac{\boldsymbol{\varepsilon}}{2}\mathbf{G}\left(\boldsymbol{\theta}^{(n)}\right)^{-1}\left[\left(\hat{\mathbf{v}}^{(i-1)}\right)^{\mathsf{T}}\tilde{\boldsymbol{\Gamma}}\left(\boldsymbol{\theta}^{(n)}\right)\hat{\mathbf{v}}^{(i-1)} + \nabla_{\boldsymbol{\theta}}\boldsymbol{\phi}\left(\boldsymbol{\theta}^{(n)}\right)\right]$$

end for

$$\mathbf{v}^{(n+1/2)} = \hat{\mathbf{v}}^{(last\ i)}$$

% *Update the position only with simple one step*

$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} + \boldsymbol{\varepsilon}\mathbf{v}^{(n+1/2)}$$

$$\boldsymbol{\Delta}\ \mathbf{log\ det}_{\boldsymbol{n}} = \mathbf{log\ det}\left(\mathbf{I} - \boldsymbol{\varepsilon\Omega}\left(\boldsymbol{\theta}^{(n+1)}, \mathbf{v}^{(n+1/2)}\right)\right) - \mathbf{log\ det}\left(\mathbf{I} + \boldsymbol{\varepsilon\Omega}\left(\boldsymbol{\theta}^{(n)}, \mathbf{v}^{(n+1/2)}\right)\right)$$

% Update the velocity exactly

$$\mathbf{v}^{(n+1)} = \mathbf{v}^{(n+1/2)} - \frac{\boldsymbol{\varepsilon}}{2}\mathrm{G}\left(\boldsymbol{\theta}^{(n+1)}\right)^{-1}\left[\left(\mathbf{v}^{(n+1/2)}\right)^{\mathsf{T}}\tilde{\boldsymbol{\Gamma}}\left(\boldsymbol{\theta}^{(n+1)}\right)\mathbf{v}^{(n+1/2)} + \nabla_{\boldsymbol{\theta}}\boldsymbol{\phi}\left(\boldsymbol{\theta}^{(n+1)}\right)\right]$$

end for

Calculate proposed $\mathbf{E}(\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)})$ according to equation (12)

$$\mathrm{logRatio} = -\mathrm{Proposed}H + \mathrm{Current}H + \sum_{n=1}^{N}\Delta\log\det_n$$

Accept or reject the proposal $(\boldsymbol{\theta}^{(L+1)}, \mathbf{v}^{(L+1)})$ according to logRatio

Algorithm 2

Explicit Lagrangian Monte Carlo (LMC)

Initialize $\boldsymbol{\theta}^{1)} = \text{current } \boldsymbol{\theta}$

Sample new velocity $\mathbf{v}^{(1)} \sim N(0, \mathbf{G}(\boldsymbol{\theta}^{1)})^{-1})$

Calculate current $\mathbf{E}(\boldsymbol{\theta}^{1)}, \mathbf{v}^{(1)})$ according to equation (12)

$\Delta \log\det = 0$

for $n = 1$ to L **do**

$$\Delta \log\det = \Delta \log\det - \log\det\left(\mathbf{G}\left(\boldsymbol{\theta}^{(n)}\right) + \varepsilon/2\tilde{\Omega}\left(\boldsymbol{\theta}^{(n)}, \mathbf{v}^{(n)}\right)\right)$$

% Update the velocity *explicitly* with a half step:

$$\mathbf{v}^{(n+1/2)} = \left[\mathbf{G}\left(\boldsymbol{\theta}^{(n)}\right) + \frac{\varepsilon}{2}\tilde{\Omega}\left(\boldsymbol{\theta}^{(n)}, \mathbf{v}^{(n)}\right)\right]^{-1}\left[\mathbf{G}\left(\boldsymbol{\theta}^{(n)}\right)\mathbf{v}^{(n)} - \frac{\varepsilon}{2}\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}^{(n)}\right)\right]$$

$$\mathbf{\Delta \log\det = \Delta \log\det + \log\det}\left(\mathbf{G}\left(\boldsymbol{\theta}^{(n)}\right) - \varepsilon/2\tilde{\Omega}\left(\boldsymbol{\theta}^{(n)}, \mathbf{v}^{(n+1/2)}\right)\right)$$

% Update the position with a full step:

$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} + \varepsilon\mathbf{v}^{\left(n+\frac{1}{2}\right)}$$

$$\mathbf{\Delta \log\det = \Delta \log\det + \log\det}\left(\mathbf{G}\left(\boldsymbol{\theta}^{(n+1)}\right) - \varepsilon/2\tilde{\Omega}\left(\boldsymbol{\theta}^{(n+1)}, \mathbf{v}^{(n+1/2)}\right)\right)$$

% Update the velocity *explicitly* with a half step:

$$\mathbf{v}^{(n+1)} = \left[\mathbf{G}\left(\boldsymbol{\theta}^{(n+1)}\right) + \frac{\varepsilon}{2}\tilde{\Omega}\left(\boldsymbol{\theta}^{(n+1)}, \mathbf{v}^{(n+1/2)}\right)\right]^{-1}\left[\mathbf{G}\left(\boldsymbol{\theta}^{(n+1)}\right)\mathbf{v}^{(n+1/2)} - \frac{\varepsilon}{2}\nabla_{\boldsymbol{\theta}}\phi\left(\boldsymbol{\theta}^{(n+1)}\right)\right]$$

$$\Delta \log\det = \Delta \log\det + \log\det\left(\mathbf{G}\left(\boldsymbol{\theta}^{(n+1)}\right) - \varepsilon/2\tilde{\Omega}\left(\boldsymbol{\theta}^{(n+1)}, \mathbf{v}^{(n+1)}\right)\right)$$

end for

Calculate proposed $\mathbf{E}(\boldsymbol{\theta}^{L+1)}, \mathbf{v}^{(L+1)})$ according to equation (12)

logRatio = -Proposed**E** + Current**E** + Alogdet

Accept or reject the proposal $(\boldsymbol{\theta}^{L+1)}, \mathbf{v}^{(L+1)})$ according to logRatio