# A Family of MCMC Methods on Implicitly Defined Manifolds

**Marcus A Brubaker**[*,†] **Mathieu Salzmann**[†] **Raquel Urtasun**[†]
[*]University of Toronto
[†]Toyota Technological Institute at Chicago

## Abstract

Traditional MCMC methods are only applicable to distributions defined on $\mathbb{R}^n$. However, there exist many application domains where the distributions cannot easily be defined on a Euclidean space. To address this limitation, we propose a general constrained version of Hamiltonian Monte Carlo, and give conditions under which the Markov chain is convergent. Based on this general framework we define a family of MCMC methods which can be applied to sample from distributions on non-linear manifolds. We demonstrate the effectiveness of our approach on a variety of problems including sampling from the Bingham-von Mises-Fisher distribution, collaborative filtering and pose estimation.

## 1 Introduction

Markov Chain Monte Carlo (MCMC) is a popular approach to sampling complex distributions because it requires relatively little knowledge about the distribution of interest. Typically, one only needs to be able to evaluate the unnormalized probability density in order to sample from the target distribution. As a consequence, MCMC has been employed in a wide range of applications such as computer vision [14], machine learning [2], and computational biology [26].

Traditional MCMC methods have generally targeted distributions defined on $\mathbb{R}^n$. However distributions over non-Euclidean spaces arise in a number of problem domains, *e.g.*, protein conformation modelling with the Fisher-Bingham distribution [9], texture analysis using distributions over rotations [15], fixed-rank matrix factorization for collaborative filtering [27, 22].

Some rejection and Gibbs sampling approaches have been developed for distributions on specific manifolds, such as Gaussian-like distributions over unit length vectors and orthogonal matrices [3, 30, 31, 13, 11], but these methods are highly specific and not applicable to different target distributions or manifolds. In computational physics a number of methods have been developed for sampling from submanifolds [17] but these techniques have not been explored outside of that community. The line of research most closely related to ours is in molecular dynamics [4], where schemes to sample from constrained systems were proposed [10, 18].

In this paper, we construct a family of MCMC methods for distributions defined on manifolds and explore their performance. In particular, we derive the Constrained Hamiltonian Monte Carlo (CHMC) algorithm which generalizes several HMC methods [10, 18, 6]. Based on this we also define a novel Metropolis Monte Carlo sampler for sampling on constrained spaces without gradients of the target posterior. The methods provide a range of options for sampling from distributions on manifolds. We demonstrate the effectiveness of our algorithms on a variety of problems including sampling a Gaussian distribution under linear constraints, sampling unit vectors from the Bingham-von Mises-Fisher distribution, sampling orthonormal matrices for collaborative filtering, and sampling human poses under length constraints for 3D reconstruction from 2D data.

In the remainder of this paper, we first review the concepts of Hamiltonian dynamics that will be used in our derivations. We then describe CHMC, prove that it converges to the desired distribution, and describe some of its variants. We then present our experimental evaluation, and conclude with a discussion of previous and future work.

## 2 Hamiltonian Dynamics

Hamiltonian dynamics are a crucial component of traditional HMC methods [23]. Our approach also exploits these dynamics, but, in contrast to traditional HMC, the system of interest is subject to constraints. In the following, we briefly review concepts from the Lagrangian and Hamiltonian dynamics of constrained

Appearing in Proceedings of the $15^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX. 

systems. We refer the reader to [20] for more details.

Let $\mathcal{M} = \{q \in \mathbb{R}^n | c(q) = 0\}$ be a connected, differentiable submanifold of $\mathbb{R}^n$, where $C(q) = \frac{\partial c}{\partial q}$ is the Jacobian of the constraints, which is assumed to have full rank everywhere. The *tangent bundle* of $\mathcal{M}$ is defined as $\mathcal{TM} = \{(q, \dot{q}) | c(q) = 0 \text{ and } C(q)\dot{q} = 0\}$. The *tangent space* of $\mathcal{M}$ at a point $q \in \mathcal{M}$ is defined as $\mathcal{T}_q\mathcal{M} = \{\dot{q} | C(q)\dot{q} = 0\}$, with the constraint $C(q)\dot{q} = 0$ found by differentiating the constraint $c(q)$.

The *Lagrangian* $\mathcal{L} : \mathcal{TM} \to \mathbb{R}$ of a constrained mechanical system is defined as the difference between kinetic and potential energies, *i.e.*, $\mathcal{L} = T - U - \lambda^T c(q)$, where $\lambda$ is a vector of Lagrange multipliers. For the purpose of this paper we assume that the potential energy, $U(q)$, does not depend on the velocity, and that the kinetic energy has the form $T(q, \dot{q}) = \frac{1}{2}\dot{q}^T M(q)\dot{q}$, where $M(q)$ is a symmetric, positive definite matrix called the mass matrix. The equations of the dynamics of the system in terms of its Lagrangian are obtained as the solution to the Euler-Lagrange equation $\frac{\partial \mathcal{L}}{\partial q} = \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{q}}$ coupled with the constraint $c(q) = 0$.

The *Hamiltonian* $\mathcal{H} : \mathcal{T}^*\mathcal{M} \to \mathbb{R}$ can be constructed by first defining the *momentum* of the system $p = \frac{\partial \mathcal{L}}{\partial \dot{q}}$, and taking the *Legendre transformation* of $\mathcal{L}$, such that $\mathcal{H} = p^T\dot{q} - \mathcal{L}$, which, in this context, becomes

$$\mathcal{H}(p, q) = T(p, q) + U(q) + \lambda^T c(q) \,, \qquad (1)$$

where the kinetic energy $T(p, q) = \frac{1}{2}p^T M(q)^{-1} p$ is now defined in terms of the momentum. The space defined by the momentum and the state is called the *cotangent bundle*, $\mathcal{T}^*\mathcal{M} = \{(p, q) | c(q) = 0 \text{ and } C(q)\frac{\partial \mathcal{H}}{\partial p}(p, q) = 0\}$, and the space of momentum at $q$ is the *cotangent space* $\mathcal{T}^*_q\mathcal{M} = \{p | C(q)\frac{\partial \mathcal{H}}{\partial p}(p, q) = 0\}$. The dynamics of the system in terms of its Hamiltonian are given by

$$\begin{aligned} \dot{p} &= -\frac{\partial \mathcal{H}}{\partial q} \,, \\ \dot{q} &= \frac{\partial \mathcal{H}}{\partial p} \,, \\ c(q) &= 0 \,. \end{aligned}$$

The dynamics of Hamiltonian systems have several important properties. First, *symmetry*, *i.e.*, forward simulation can be inverted by reversing the direction of time. Second, *$\rho$-reversibility* with respect to the map $\rho(p, q) = (-p, q)$, meaning that inverting the momentum is equivalent to reversing the direction of time. Third, the value of the Hamiltonian is conserved over time, *i.e.*, $d\mathcal{H}/dt = 0$. Finally, Hamiltonian dynamics are symplectic. A mapping $f : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ is *symplectic* if $F(x)^T J F(x) = J$ where $F = \frac{\partial f}{\partial x}$ is the Jacobian of $f$, and $J = \begin{bmatrix} 0 & \mathbf{I}_{n\times n} \\ -\mathbf{I}_{n\times n} & 0 \end{bmatrix}$. Symplectic dynamics imply volume preservation of the cotangent bundle, *i.e.*, the volume of the $(p, q)$ space is preserved under forward simulation. This can be easily seen, since $F(x)^T J F(x) = J$ implies that $\det(F(x)) \det(J) \det(F(x)) = \det(J)$, thus $\det(F(x))^2 = 1$. We refer the reader to [8, 21, 20] for proofs of these properties and a more formal discussion on Hamiltonian dynamics and symplectic mappings. These properties of Hamiltonian dynamics are crucial for the proofs of detailed balance and ergodicity.

## 3 Constrained HMC

In this section, we introduce our Constrained Hamiltonian Monte Carlo algorithm, and prove that it converges to the desired distribution. Finally, we show how CHMC can be used to derive other constrained Monte Carlo algorithms.

Let $\pi : \mathbb{R}^n \to \mathbb{R}$ be a continuous probability measure on $\mathcal{M}$ such that $\int_{\mathcal{M}} \pi(q)dq = 1$ and $\pi(q) \geq 0$ for all $q \in \mathcal{M}$. Adopting the terminology of [5], we define two different Hamiltonians: the *acceptance Hamiltonian* $\mathcal{H}$ and the *guidance Hamiltonian* $\hat{\mathcal{H}}$. The acceptance Hamiltonian is used to compute the acceptance probability of the step, and is based on the target distribution $\pi(q)$. In contrast, the guidance Hamiltonian is used for simulation.

Following Eq. (1), both Hamiltonians are expressed as the sum of kinetic and potential energies plus a term due to the constraints. For the acceptance Hamiltonian, the kinetic energy is $T(p, q) = \frac{1}{2}p^T M(q)^{-1} p$ for some mass matrix $M(q)$, and the potential energy is $U(q) = \frac{1}{2}\log|M(q)| - \log\pi(q)$. This choice of kinetic and potential energies means that $\exp(-\mathcal{H}(p, q)) = \pi(q)\mathcal{N}(p|0, M(q))$, where $\mathcal{N}(\cdot|\mu, \Sigma)$ is a multivariate Gaussian probability density with mean $\mu$ and covariance $\Sigma$. The guidance Hamiltonian, $\hat{\mathcal{H}}$ uses the same kinetic energy but can vary in the choice of potential energy function, $\hat{U}(q)$.

A step of CHMC proceeds as follows. First, a new momentum is drawn from the distribution $\mathcal{N}(p_0|0, M(q_0))$ subject to the constraint that $C(q)\frac{\partial \mathcal{H}}{\partial p}(p_0, q_0) = 0$. This can be done by first sampling from the unconstrained Gaussian distribution, and then projecting the momentum onto $\mathcal{T}^*_{q_0}\mathcal{M}$. This step is, in essence, a Gibbs sampler of the momentum for the augmented distribution $\exp(-\mathcal{H}(p, q))$ on $\mathcal{T}^*\mathcal{M}$. Then, starting at $(p_0, q_0)$, the *guidance Hamiltonian* is simulated for $L$ steps with a step size of $h$ ending at $(p_L, q_L)$. The resulting state of the simulation, $q_L$, is then accepted or rejected using a Metropolis test, *i.e.*, $q_L$ becomes the next state with probability $\min(1, \exp\{\mathcal{H}(p_0, q_0) - \mathcal{H}(p_L, q_L)\})$, otherwise the next state is $q_0$. This is summarized in Algorithm 1.

**Algorithm 1** Constrained Hamiltonian Monte Carlo

**Input:** $q_0$, $M(q)$, $h$, $L$, $\mathcal{H}(p,q)$, $\hat{\mathcal{H}}(p,q)$
  $p_0 \sim \mathcal{N}(0, M(q_0)|C(q_0)M(q_0)^{-1}p_0 = 0)$
  **for** $i = 1, \ldots, L$ **do**
    $(p_i, q_i) \leftarrow \Phi_h^{\hat{\mathcal{H}}}(p_{i-1}, q_{i-1})$
  **end for**
  $u \sim U(0,1)$
  **if** $u \leq \min\left(1, \exp\left\{\mathcal{H}(p_0,q_0) - \mathcal{H}(p_L,q_L)\right\}\right)$ **then**
    **return** $q_L$ {accept the proposal}
  **else**
    **return** $q_0$ {reject the proposal}
  **end if**

### 3.1 Numerical Simulation

At the heart of CHMC is the simulation of Hamiltonian dynamics. In practice, simulation of general Hamiltonians requires the use of numerical approximations. Perhaps unexpectedly, many properties which hold for exact Hamiltonian dynamics can also hold for numerical approximations, so long as the numerical integration method satisfies certain properties. In the following, we describe the necessary properties to ensure convergence of the chain, and introduce the integrator used in our experiments.

We denote by $\Phi_h^{\mathcal{H}} : \mathcal{T}^*\mathcal{M} \to \mathcal{T}^*\mathcal{M}$ the numerical integrator which approximates the dynamics of the Hamiltonian $\mathcal{H}$ at a time $h$ into the future. Note that this implicitly requires that $\Phi_h^{\mathcal{H}}$ satisfies both the state constraints $c(q) = 0$ and the momentum constraints $C(q)\frac{\partial \mathcal{H}}{\partial p}(p_0, q_0) = 0$. We further require the integrator to be both symmetric and symplectic. An integrator is *symmetric* if $(p,q) = \Phi_{-h}^{\mathcal{H}}(\Phi_h^{\mathcal{H}}(p,q))$, and is *symplectic* if $\Phi_h^{\mathcal{H}}$ is a symplectic mapping when applied to a smooth Hamiltonian $\mathcal{H}$.

While a symplectic integrator implies volume preservation of the cotangent bundle, it also implies that there exists a *discrete Lagrangian*, $\mathcal{L}'_h$, which the numerical method is integrating (Theorem 2.1.1, [21]; Theorem 5.6, Section IX.5.2, [8]). That is, starting at $(p_0,\ q_0)$, under the integrator, $q_1$ is the solution to the equation $p_0 + \frac{\partial}{\partial q_0}\mathcal{L}'_h(q_0, q_1) = \lambda_0^T C(q_0)$.

The final requirement is that the integrator is consistent. A symplectic integrator of a continuous Hamiltonian $\mathcal{H}$ with corresponding continuous Lagrangian $\mathcal{L}$ is of order $r$ if, for some sufficiently small $h$, we have

$$\mathcal{L}'_h(q_0, q_1) = \int_0^h \mathcal{L}(q(t), \dot{q}(t))dt + h^r e_h(q_0, q_1)\ , \quad (2)$$

where $\mathcal{L}'_h : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ is the *discrete Lagrangian* of the integrator, $q(t)$ is the solution to the Euler-Lagrange equation applied to $\mathcal{L}$ with boundary conditions $q(0) = q_0$, $q(h) = q_1$, and $e_h$ is a bounded error function. An integrator is said to be *consistent* if $r \geq 1$. Note that one property not satisfied by the numerical integration scheme is exact conservation of the Hamiltonian. Any consistent method will, however, approximately conserve the Hamiltonian for sufficiently small $h$. In CHMC, the error in this approximation is corrected through the use of the Metropolis acceptance test.

One choice of numerical integration method which satisfies the aforementioned conditions is known as RATTLE [1, 25]. It is a generalization of the Leapfrog integrator (typically used with unconstrained HMC) to handle the manifold constraints and the more general form of Hamiltonian. A step of the generalized RATTLE algorithm consists of solving the system of non-linear equations

$$\begin{aligned}
p_{1/2} &= p_0 - \frac{h}{2}\left(\frac{\partial \hat{\mathcal{H}}(p_{1/2}, q_0)}{\partial q} + C(q_0)^T\lambda\right)\ ,\\
q_1 &= q_0 + \frac{h}{2}\left(\frac{\partial \hat{\mathcal{H}}(p_{1/2}, q_0)}{\partial p} + \frac{\partial \hat{\mathcal{H}}(p_{1/2}, q_1)}{\partial p}\right)\ ,\\
0 &= c(q_1)\ ,\\
p_1 &= p_{1/2} - \frac{h}{2}\left(\frac{\partial \hat{\mathcal{H}}(p_{1/2}, q_1)}{\partial q} + C(q_1)^T\mu\right)\ ,\\
0 &= C(q_1)\frac{\partial \hat{\mathcal{H}}(p_1, q_1)}{\partial p}\ ,
\end{aligned}$$

for the unknowns $p_{1/2}, q_1, p_1, \lambda, \mu$ where $\hat{\mathcal{H}}$ is the simulation Hamiltonian, and $\lambda$ and $\mu$ are the Lagrange multipliers associated with the state and momentum constraints at the end of the step, *i.e.*, $c(q_1) = 0$ and $C(q_1)\frac{\partial \hat{\mathcal{H}}(p_1,q_1)}{\partial p} = 0$. A solution to this system of non-linear equations can be obtained using Newton's method.[1] Note that this method is symplectic, symmetric, of order 2 (and therefore consistent), and respects the manifold constraints, ensuring that the solution lies in the cotangent space $\mathcal{T}^*\mathcal{M}$. Furthermore, as the integrator works with any continuous Hamiltonian, the method can naturally handle a state dependent mass matrix [6]. Note that, although we use this integrator for our experiments, other integrators which preserve the aforementioned properties can be used, *e.g.*, a partitioned Runge-Kutta method using the Lobatto IIIA-IIIB pair can be used to create higher order methods on a manifold [8, 12].

[1] Solving this system can be done more efficiently by noting that the first three equations are independent of $p_1, \mu$ and and the last two equations and that the last two equations are linear in $p_1, \mu$ for quadratic kinetic energies.

### 3.2 Convergence of CHMC

We now prove that CHMC converges to the target posterior $\pi(q)$ from any starting point $q \in \mathcal{M}$. We begin by stating the conditions under which CHMC satisfies detailed balance.

**Theorem 1** (Detailed Balance). *Let $\hat{\mathcal{H}}$ be $\mathcal{C}^2$-continuous, $\mathcal{M} = \{q \in \mathbb{R}^n | c(q) = 0\}$ be a connected, smooth and differentiable manifold with $\frac{\partial c}{\partial q}$ full-rank everywhere, $M(q)$ be positive definite on $\mathcal{M}$, and $\pi(q)$ be smooth. If $\Phi_h^{\hat{\mathcal{H}}}$ is symmetric and symplectic, then*

$$\int_{Q'} \int_{Q} \pi(q) T(q \to q') dq dq' = \int_{Q} \int_{Q'} \pi(q') T(q' \to q) dq' dq \,, \tag{3}$$

*where $Q, Q' \subset \mathcal{M}$ and $T$ is the transition kernel.*

**Proof:** In appendix. □

The proof proceeds by first showing that CHMC satisfies detailed balance with respect to the augmented distribution $\exp(-\mathcal{H}(p,q))$ on $\mathcal{T}^*\mathcal{M}$. Because the integration is volume preserving and symmetric, it is straightforward to compute the transition probability between two regions, regardless of the guidance Hamiltonian. After some algebraic manipulation of the acceptance probability, detailed balance follows. Then, since $\pi(q)$ is the marginal distribution of $\exp(-\mathcal{H}(p,q))$, it follows that the chain satisfies detailed balance with respect to $\pi(q)$ by simply ignoring the momentum.

**Theorem 2** (Accessibility). *Let $\hat{\mathcal{H}}$ be $\mathcal{C}^2$-continuous, $\mathcal{M} = \{q \in \mathbb{R}^n | c(q) = 0\}$ be a connected, smooth and differentiable manifold with $\frac{\partial c}{\partial q}$ full-rank everywhere, and $M(q)$ be positive definite on $\mathcal{M}$. If $\Phi_h^{\hat{\mathcal{H}}}$ is symmetric, symplectic and consistent, then for any $q_0, q_1 \in \mathcal{M}$ and $h$ sufficiently small, there exist finite $p_0 \in \mathcal{T}^*_{q_0}\mathcal{M}$, $p_1 \in \mathcal{T}^*_{q_1}\mathcal{M}$ and Lagrange multipliers $\lambda_0, \lambda_1$ such that $(p_1, q_1) = \Phi_h^{\hat{\mathcal{H}}}(p_0, q_0)$.*

**Proof:** In appendix. □

The proof follows from the connection between symplectic maps and Lagrangian dynamics noted above. The existence of a discrete Lagrangian, $\mathcal{L}'_h$, directly provides a formula for the momentum and Lagrange multipliers given $q_0$ and $q_1$. The final piece consists in proving that the derivatives of $\mathcal{L}'_h$ exist, which can be done using the fact that the integrator is consistent and the Hamiltonians are smooth.

While the above theorem shows that there exist momentum and Lagrange multipliers to move between any two points, it is possible that, for a given initial state $(p_0, q_0)$, there may be multiple choices of Lagrange multipliers. For instance, for a single step of the RATTLE integrator (discussed in Section 3.1) on a sphere, there will generally be two choices of $\lambda$ corresponding to points on opposite hemispheres.

To prove irreducibility we must then make additional assumptions about how the integrator selects between these two Lagrange multipliers given a particular set of initial conditions. Informally, we assume that for small regions around every point on the manifold, the integrator can uniquely choose the Lagrange multiplier which moves between points within the region.

**Theorem 3** (Irreducibility). *Let $\hat{\mathcal{H}}$ be $\mathcal{C}^2$-continuous, $\mathcal{M} = \{q \in \mathbb{R}^n | c(q) = 0\}$ be a connected, smooth and differentiable manifold with $\frac{\partial c}{\partial q}$ full-rank everywhere, $M(q)$ be positive definite on $\mathcal{M}$, $\Phi_h^{\hat{\mathcal{H}}}$ be symmetric, symplectic and consistent, and $\pi(q)$ be strictly positive on $\mathcal{M}$. Let $\mathcal{B}_\ell(q) = \{q' \in \mathcal{M} | d(q', q) \leq \ell\}$ be a ball defining all points on $\mathcal{M}$ with geodesic distance at most $\ell$ from $q$. If there is a constant $\ell > 0$ such that for every $q \in \mathcal{M}$, $q' \in \mathcal{B}_\ell(q)$ there is a unique choice of Lagrange multiplier and momentum $p \in \mathcal{T}^*_q\mathcal{M}$, $p' \in \mathcal{T}^*_{q'}\mathcal{M}$ for which $(p', q') = \Phi_h^{\hat{\mathcal{H}}}(p, q)$, and, if, given $(p, q)$, this is the Lagrange multiplier selected by the integrator, then, for $h$ sufficiently small and any $q_0, q_1 \in \mathcal{M}$, $\pi(q_1) > 0$, there exists an $n \in \mathbb{N}$ such that*

$$T^n(q_0 \to q_1) > 0 \,. \tag{4}$$

**Proof:** In appendix. □

The proof proceeds by noting that a composition of $L$ integration steps can, itself, be considered a single integration step. Symmetry, sympleticness and consistency are all preserved by the composition and, hence, Theorem 2 applies to the composite step. Then, because the manifold is connected, there is a path of some length between $q_0$ and $q_1$ which can be divided into chunks of length less than $\ell$. Using the additional assumption of uniqueness of the choice of Lagrange multipliers made by the integrator within neighborhoods of size $\ell$, and the fact that $\pi(q_1) > 0$, it follows that the acceptance probability is non-zero. In order to prove convergence, it is necessary to prove aperiodicity. This is proven in appendix as an almost direct consequence of irreducibility.

**Theorem 4** (Convergence). *Let $\hat{\mathcal{H}}$ be $\mathcal{C}^2$-continuous, $\mathcal{M} = \{q \in \mathbb{R}^n | c(q) = 0\}$ be a connected, smooth and differentiable manifold with $\frac{\partial c}{\partial q}$ full-rank everywhere, $M(q)$ be positive definite on $\mathcal{M}$, $\Phi_h^{\hat{\mathcal{H}}}$ be symmetric, symplectic and consistent, and $\pi(q)$ be smooth and strictly positive on $\mathcal{M}$. Let $\mathcal{B}_\ell(q) = \{q' \in \mathcal{M} | d(q', q) \leq \ell\}$ be a ball defining all points on $\mathcal{M}$ with geodesic distance at most $\ell$ from $q$. If there is a constant $\ell > 0$ such that for every $q \in \mathcal{M}$, $q' \in \mathcal{B}_\ell(q)$ there is a unique choice of Lagrange multiplier and momentum $p \in \mathcal{T}^*_q\mathcal{M}$, $p' \in \mathcal{T}^*_{q'}\mathcal{M}$ for which $(p', q') = \Phi_h^{\hat{\mathcal{H}}}(p, q)$,*

*and, if, given $(p, q)$, this is the Lagrange multiplier selected by the integrator, then for all $q_0 \in \mathcal{M}$*

$$\lim_{n\to\infty} \|T^n(q_0 \to \cdot) - \pi(\cdot)\| = 0 \,.$$

**Proof:** In appendix. □

### 3.3 Variations of CHMC

The above proofs give us flexibility in the choice of integrator $\Phi$, mass matrix $M(q)$, number of steps $L$ and guidance Hamiltonian $\hat{\mathcal{H}}$. Note that unconstrained HMC [5], and Riemann Manifold HMC [6] are both instances of CHMC when $\mathcal{M} = \mathbb{R}^n$. We now describe some other variations of CHMC.

**Constrained Langevin MC:** The Constrained Langevin Monte Carlo method is a special case of CHMC with $\hat{\mathcal{H}} = \mathcal{H}$ and for a single simulation step, *i.e.*, $L = 1$. It is sometimes preferred to CHMC because it requires fewer gradient evaluations and prevents wasting computation on long trajectories which may ultimately be rejected. However, it can also perform poorly in some cases, as the small number of steps might cause the random initial momentum to dominate, exhibiting a random walk-like behavior.

**Constrained Metropolis Monte Carlo:** The primary practical limitation of HMC is that it requires the gradient of the target posterior to be computed. This can be limiting as in some applications analytic gradients are not readily available, or are cumbersome and expensive to compute. In those cases, using Metropolis Monte Carlo can be advantageous. The unconstrained Metropolis algorithm with proposal covariance $\Sigma$ can be considered a special case of HMC with simulation Hamiltonian $\hat{\mathcal{H}}(p, q) = \frac{1}{2} p^T M^{-1} p$ and parameters $h = 1$, $L = 1$ and $M = \Sigma^{-1}$. Based on this observation, a Constrained Metropolis algorithm can be derived in the same fashion. This results in an MCMC method on a constrained space which does not require the gradient of the target posterior.

**Constrained Riemann Manifold HMC:** As in unconstrained HMC, proper tuning of the mass matrix is crucial for good performance. Furthermore, for some problems, there may not be a single mass matrix which is suitable everywhere. The recently proposed Riemann Manifold HMC method [6] addresses this issue by exploiting geometric information of the probability distribution to set and adapt the mass matrix. As a consequence, the mass matrix becomes dependent on the state $q$. A similar variable mass matrix is naturally handled in our formulation.

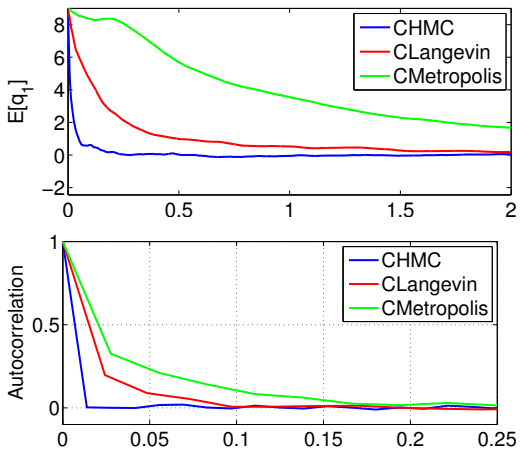


Figure 1: **Sampling from a Gaussian distribution with linear constraints.** (Top) Mean of the first coordinate of $q$ as a function of computation time. (Bottom) Autocorrelation of $-\log \pi(q)$ for the different methods after convergence, with the chains initialized at the mean $\mu$.

## 4 Experimental Evaluation

We demonstrate the effectiveness of our approach at sampling various target distributions on different manifolds, including sampling a multivariate Gaussian distribution under linear constraints, sampling unit length vectors from a Bingham-von Mises-Fisher distribution, sampling orthonormal, low-rank matrices for collaborative filtering, and sampling human poses under length constraints for 3D estimation from 2D observations.

### 4.1 Linearly Constrained Gaussian

Our first example consists of sampling from a multivariate Gaussian distribution subject to linear equality constraints. Because the constraint manifold is linear, sampling in this case is equivalent to sampling from a Gaussian in a subspace of the original space, allowing a clear assessment of the accuracy and effectiveness of our methods. Specifically, we define $\pi(q) \propto \mathcal{N}(q|\mu, \Sigma)$ subject to the constraints $c(q) = Aq - b$, where $A \in \mathbb{R}^{D\times n}$ and $b \in \mathbb{R}^D$. For this experiment, we used $\mu = (0, 0, 0, 0)^T$ and $\Sigma = \text{diag}(1, 1, 0.01, 0.01)$. Two constraints were applied, $A_1 = (1, 1, 1, 1)$, $A_2 = (1, 1, -1, 1)$ and $b = (0, 0)^T$.

In a first experiment, we compare the ability of the different algorithms presented in the previous section to find the mode of the Gaussian. To this end, we initialized all the chains to the same state, $q = (9, -9, 11, -11)^T$. The top row of Fig. 1 depicts the estimated mean of the first coordinate for each method. It can be seen that CHMC and CLangevin algorithms performed best, quickly converging to the correct value 0. In comparison, the CMetropolis algo-

| Method | $E[-\log\pi(q)]$ | ESS % | ESS/second |
|---|---|---|---|
| CHMC (L = 4) | -999.021 | 27.3 | 183.756 |
| CHMC (L = 3) | -998.759 | 25.4 | 217.427 |
| CHMC (L = 2) | -999.121 | 37.9 | 440.898 |
| CLangevin | -998.757 | 33.0 | 619.339 |
| CMetropolis | -998.82 | 3.8 | 90.1513 |
| Gibbs [11] | -998.742 | 50.8 | 160.722 |

Table 1: **Efficiency for sampling the Bingham-von Mises-Fisher Distribution.** We used parameters $d = (100, 0, 0, 0, 0, 0)$ and $A = diag(-1000, -600, -200, 200, 600, 1000)$, $M = 2000$. Results are average over 10 runs; for CHMC and CLangevin $h = 1$ and for CMetropolis $h = 0.4$.

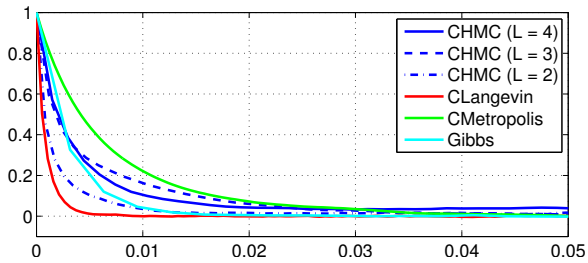


Figure 2: **Autocorrelation of $-\log\pi(q)$ on the Bingham-von Mises-Fisher Distribution.** Distribution parameters were set to $d = (100, 0, 0, 0, 0, 0)$ and $A = diag(-1000, -600, -200, 200, 600, 1000)$. Results were averaged over 10 independent runs.

rithm took much longer, both in number of steps and in computation time, to eventually find the mode.

We then tested the performance of the methods once the mode is found. We initialize each sampler at the mode of the distribution $q = (0, 0, 0, 0)^T$ and computed the autocorrelation of $-\log\pi(q)$ averaged over ten independent runs. The bottom row of Fig. 1 depicts the autocorrelation of these chains as a function of computation time, again showing that CHMC is more efficient than CMetropolis and CLangevin.

### 4.2 Bingham-von Mises-Fisher Distribution

Distributions over directions (*i.e.*, vectors of unit length) play a significant role in applications such as structural biology [9], geology [15], computer vision [24] and robotics [7]. A natural distribution over directions is the Bingham-von Mises-Fisher distribution, which can be derived as a Gaussian distribution in $\mathbb{R}^n$ restricted to the unit sphere $\mathbb{S}^{n-1}$. More formally, its density function is given by $\pi(q) \propto \exp\left(d^T q + q^T A q\right)$ restricted to the manifold $\mathbb{S}^{n-1} = \left\{q \in \mathbb{R}^n | q^T q = 1\right\}$. Note that if $d$ is the zero-vector, it reduces to the Bingham distribution, and if $A$ is the zero-matrix it is the von Mises-Fisher distribution. If the vector $d$ is non-zero, the distribution is antipodally asymmetric (*i.e.*, $\pi(q) \neq \pi(-q)$), and it has a bias towards values that point in the same direction as $d$ (*i.e.*, $d^T q > 0$). The matrix $A$ describes the spread and concentration of the distribution [15]. Note that $A$ is not uniquely defined as the matrices $A$ and $A + \alpha I$ describe the same distribution for any choice of scalar $\alpha$. This is due to the structure of the manifold as $\exp(q^T(A + \alpha I)q) = \exp(q^T A q + \alpha q^T q) \propto \exp(q^T A q)$.

Table 1 and Fig. 2 depict the effective sample size (ESS) and ESS/time, as well as the autocorrelation for our algorithms and for the state-of-the-art Gibbs sampler [11]. From the table we can see that, while each step of the Gibbs sampler is the more independent (*i.e.*, has the highest ESS), the cost of each one of its steps is higher, causing it to lose out to CHMC and CLangevin when the ESS is normalized by the computation time. This is particularly noticeable in Fig. 2 where the difference between the Gibbs sampler and the best performing of our methods (CLangevin) is very significant.

The computational performance of the Gibbs sampler is determined largely by the efficiency of a rejection sampler which was intractably slow for the choice of $A$ and $d$ explored here. Indeed, for some choices of $A$ and $d$, the rejection sampler is completely unable to move, *e.g.*, with acceptance probabilities of $10^{-100}$. In contrast, our methods were still able to perform well in such cases. To work around this problem, the rejection sampling step of the Gibbs sampler was replaced with 20 steps of an Independence Metropolis sampler with the same proposal distribution as used in the rejection sampler [11].

Note that the CLangevin sampler outperformed CHMC. This result is in contrast with the general belief in unconstrained MCMC that multiple steps perform better. One explanation for this behavior is that the ability to reach more distant points with multiple steps is less valuable with closed, compact spaces such as $\mathbb{S}^{n-1}$. Indeed, the decrease in ESS (*i.e.*, increase in autocorrelation) with $L > 2$ suggests that the simulation begins to oscillate with longer simulations.

### 4.3 Collaborative Filtering

A typical example of collaborative filtering arises in the context of user ratings of movies. One popular formulation of this problem is to find a low-rank decomposition of a partially observed matrix. Given a matrix $\mathbf{Y} \in \mathbb{R}^{N \times M}$ of observed ratings, we seek to find a low-rank decomposition of $\mathbf{Y} = f(\mathbf{U}^T \mathbf{S} \mathbf{V})$ with $\mathbf{U} \in \mathbb{R}^{r \times N}$, $\mathbf{V} \in \mathbb{R}^{r \times M}$, and $\mathbf{S}$ a diagonal matrix. The element-wise function $f()$ is often taken as the identity function. However, other choices, such as the logistic function, are possible. In practice, only a few entries in $\mathbf{Y}$ are known (*e.g.*, each user only rated a few movies), and the goal is to use the decomposition to estimate missing values.

| | 5 | 10 | 15 |
|---|---|---|---|
| HMC | 1.577 ± 0.39 | 2.001 ± 0.66 | 2.306 ± 0.25 |
| HMC-l | 0.909 ± 0.008 | 0.949 ± 0.01 | 0.99 ± 0.01 |
| CHMC | 0.893 ± 0.01 | 0.888 ± 0.01 | 0.889 ± 0.01 |
| CHMC-l | **0.888 ± 0.01** | **0.881 ± 0.01** | **0.881 ± 0.01** |

| | 5 | 10 | 15 |
|---|---|---|---|
| HMC | 0.435 ± 0.008 | 0.465 ± 0.016 | 0.503 ± 0.002 |
| HMC-l | **0.413 ± 0.002** | 0.429 ± 0.004 | 0.445 ± 0.007 |
| CHMC | 0.419 ± 0.003 | 0.418 ± 0.003 | 0.419 ± 0.004 |
| CHMC-l | 0.418 ± 0.004 | **0.415 ± 0.003** | **0.416 ± 0.002** |

Figure 3: **1M MovieLens:** Comparison of our approach (CHMC) with its unconstrained version (HMC) in the weak settings and for rank 5, 10 and 15. A *-l* after the method's name indicates the use of the logistic function in the predictor. (Left) RMSE. (Right) NMAE. Note that CHMC outperforms HMC in most cases.

In our experiments, we sample $\mathbf{U}$, $\mathbf{V}$, and the diagonal of $\mathbf{S}$ for a fixed rank $r$ under orthonormality constraints of the form $\mathbf{U}\mathbf{U}^T = \mathbf{I}_{r\times r}$ , and $\mathbf{V}\mathbf{V}^T = \mathbf{I}_{r\times r}$. The vector $q$ contains the concatenated vectorized forms of $\mathbf{U}$, $\mathbf{V}$, and $diag(\mathbf{S})$. Given a set $\mathcal{E}$ of pairs of indices corresponding to known entries in $\mathbf{Y}$, a density can be expressed as

$$\pi(q) \propto \prod_{(i,j)\in\mathcal{E}} \exp\left(-\left(f\left(\mathbf{U}_i^T\mathbf{S}\mathbf{V}^j\right) - \mathbf{Y}_{i,j}\right)^2/\sigma_p^2\right) ,$$

where $\mathbf{U}_i$ is the $i^{th}$ column of $\mathbf{U}$, $\mathbf{V}^j$ is the $j^{th}$ row of $\mathbf{V}$, and $\sigma_p$ is the expected prediction error. The quality of the decomposition is evaluated by computing the prediction error on the test entries of $\mathbf{Y}$.

We tested our approach on two popular data sets: 1M MovieLens and EachMovie. In both cases, we used the partitions of [19], and performed our experiments under the weak generalization setting, *i.e.*, a single rating per user is withheld for the test set. At test time, the prediction error was computed both as the RMSE and as the NMAE. In the latter case, following [16], we used 1.6 and 1.944 as normalization constant for MovieLens and EachMovie, respectively. Our results were obtained by computing the mean predictions over 2000 samples and comparing it with the test data.

As shown in Figs. 3 and 4 CHMC outperforms unconstrained HMC in most cases. The two exceptions occur only when error is evaluated in terms of NMAE. As both approaches optimize RMSE, we consider this measure the most indicative of performance, particularly since NMAE involves rounding, thus convoluting the results. Further note that, without constraints, sampling tends to overfit with larger ranks, thus performing poorly. This was confirmed by checking the prediction error on the training set, which was indeed much lower in the unconstrained case than in the constrained one. Note that, even though our approach was not specifically designed to address the collaborative filtering problem, our results are comparable to the state-of-the-art methods, whose NMAE typically range from 0.4342 [19] to 0.3916 [32] for MovieLens, and from 0.4520 [19] to 0.4109 [32] for EachMovie.

### 4.4 Human Pose Estimation

In the next experiment, we tackle the task of 3D human pose estimation from monocular 2D observations. The goal is to estimate the pose of a person, parameterized by the 3D locations of the joints of a stick figure. In this case, the manifold of interest is defined by constraints encoding the fixed length of the skeleton segments. More specifically, let $q$ be the vector of 3D coordinates of $N$ joints. The constraints can be formulated as

$$\left\|q^i - q^j\right\|_2^2 = l_{i,j}^2 \ , \ \forall (i,j) \in \mathcal{J} \ , \tag{5}$$

where $q^i$ encodes the 3D position of joint $i$, $l_{i,j}$ is the known length of the limb, and $\mathcal{J}$ is the set of limbs.

To perform reconstruction from monocular images, we assume that we are given the noisy image locations $x^i \in \mathbb{R}^2$ of the skeleton joints, as well as the matrix $\mathbf{A}$ of internal camera parameters. Furthermore, to regularize our reconstructions, we rely on a linear pose model obtained by performing PCA on a set of training 3D poses. This lets us write the density as

$$\begin{aligned}\pi(q) \propto & \textstyle\prod_{i=1}^N \exp\left(-\|\hat{x}^i(q^i) - x^i\|^2/\sigma_m^2\right) \cdot \\ & \textstyle\prod_{j=1}^{3N} \exp\left(-\left(\mathbf{P}_j^T(q-q_0)\right)^2/\sigma_j^2\right) ,\end{aligned} \tag{6}$$

where $\mathbf{P}_j$ is the column vector containing the $j^{th}$ eigenpose obtained by PCA, $\sigma_j^2$ is the corresponding eigenvalue, $q_0$ is the mean pose of the training data,

$$\hat{x}^i(q^i) = \begin{pmatrix} (\mathbf{A}^1 q^i)/(\mathbf{A}^3 q^i) \\ (\mathbf{A}^2 q^i)/(\mathbf{A}^3 q^i) \end{pmatrix} \tag{7}$$

is the projection of joint $i$, with $\mathbf{A}^k$ the $k^{th}$ row of $\mathbf{A}$, and $\sigma_m^2$ is the expected variance of the image measurements. The first part of the density encodes the reprojection error of the joints given their corresponding image measurements, and the second part defines the PCA-based regularizer. Note that, for simplicity, we assume that the pose is expressed in the camera referential, which is equivalent to assuming that the camera is fully calibrated.

We performed our experiments on the walking sequence of subject 1 in the HumanEva data set [28]. This data set depicts a person walking in circles in a room. We set aside one circle to learn the linear pose model, and tested our approach on the remaining circle. To create image measurements, we used a known camera to project the ground-truth 3D poses, and added noise to the projections with standard deviation ranging from 0 to 10 pixels. For each frame, we

| | 5 | 10 | 15 |
|---|---|---|---|
| HMC | 1.153 ± 0.002 | 1.161 ± 0.002 | 1.204 ± 0.018 |
| HMC-l | 1.155 ± 0.007 | 1.164 ± 0.001 | 1.184 ± 0.004 |
| CHMC | 1.144 ± 0.002 | 1.121 ± 0.001 | 1.116 ± 0.001 |
| CHMC-l | **1.137 ± 0.003** | **1.115 ± 0.002** | **1.11 ± 0.002** |

| | 5 | 10 | 15 |
|---|---|---|---|
| HMC | 0.44 ± 0.003 | 0.44 ± 0.003 | 0.448 ± 0.005 |
| HMC-l | **0.437 ± 0.003** | 0.436 ± 0.002 | 0.443 ± 0.0015 |
| CHMC | 0.444 ± 0.003 | 0.434 ± 0.003 | 0.432 ± 0.002 |
| CHMC-l | 0.44 ± 0.002 | **0.43 ± 0.002** | **0.428 ± 0.003** |

Figure 4: **EachMovie:** Comparison of our approach (CHMC) with its unconstrained version (HMC) in the weak settings and for rank 5, 10 and 15. A *-l* after the method's name indicates the use of the logistic function in the predictor. (Left) RMSE. (Right) NMAE. As for MovieLens, CHMC outperforms HMC in most cases.

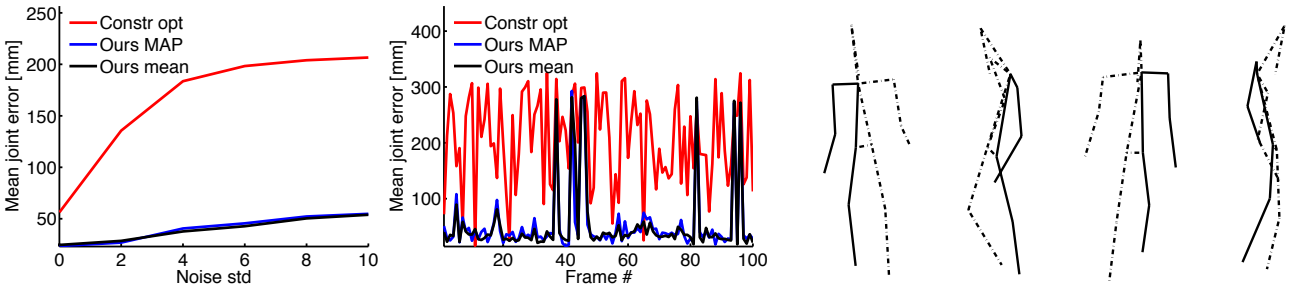


Figure 5: **Human pose estimation:** (a) Mean joint-to-joint error, averaged over 100 frames, as a function of the image measurement noise. (b) Similar error as a function of the frame number for a noise std 10. Note that constrained optimization yields bad local optima in most cases, whereas our approach almost always gives an accurate solution. (c)-(f) Examples of our mean reconstructions for a noise std 10.

initialized $q$ with a training pose chosen uniformly at random, and let our HMC simulation run for $L = 200$ steps. Furthermore, we used the Hessian of the negative log of the observations as a mass matrix. In this experiment, we allowed this matrix to vary with $q$, thus computing its gradient at each step. Fig. 5(left) depicts reconstruction error as a function of the measurement noise. This error is computed as the mean joint-to-joint distance between reconstruction and ground-truth, averaged over 100 frames of the sequence. We compare our MAP solution and the mean of our samples to reconstructions obtained with a constrained optimization method following a projected gradient descent approach. Note that we clearly outperform this baseline, which tends to get stuck in local optima due to the poor initialization. As illustrated in Fig. 5(right) where we plot the error for each frame for a noise std 10, our approach manages to get out of such local minima for most frames.

## 5 Discussion

In this paper we have presented a general framework for constructing Markov chains on manifolds defined by implicit constraints. Using this framework we constructed a family of different sampling methods, *i.e.*, Constrained Hamiltonian Monte Carlo, Constrained Metropolis and Constrained Langevin. Furthermore, we have also explicitly stated the conditions necessary for the Markov Chain to converge to the desired distribution, and have demonstrated the benefits of our algorithms on a variety of problem domains involving different distributions and manifolds.

Traditional HMC [5, 23] can be seen as a special case of our framework, where the manifold of interest is $\mathbb{R}^n$. Further, the generality of our framework also allows the use of state dependent mass matrices, thus extending the Riemann Manifold HMC (RMHMC) of [6] to handle constrained configuration spaces. Note that the manifold geometry exploited by RMHMC is independent of the constraint manifolds discussed here. The methods most strongly related to ours are that of [10, 18], where sampling schemes were proposed for constrained systems in molecular dynamics. Our work goes well beyond these methods by allowing the use of state dependent mass matrices and, more importantly, our method and proofs are applicable to any appropriate integration method applied to any guidance Hamiltonian. The greater generality of our framework also allowed the introduction of the novel Constrained Metropolis algorithm which can be used to construct Markov chains on manifolds without the need for gradients of the target distribution.

In the future, we intend to study the effect of other choices of guidance Hamiltonians and integration methods, such as the partitioned Runge-Kutta method of [12]. We believe that the significant body of research in structure-preserving numerical integration methods [8] could be exploited in conjunction with our framework to the benefit of HMC-type algorithms. *Finally, Matlab code implementing CHMC is available at* `http://www.cs.toronto.edu/~mbrubake/`.

**Acknowledgements** This work was partially supported by NSF ID 1017626.

## References


[1] H. C. Andersen. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *Journal of Computational Physics*, 52 (1):24 – 34, 1983. ISSN 0021-9991. doi: 10.1016/0021-9991(83)90014-1.

[2] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50(1):5–43, 2003.

[3] C. Bingham. An antipodally symmetric distribution on the sphere. *The Annals of Statistics*, 2(6): 1201–1225, Nov 1974.

[4] G. Ciccotti and J. P. Ryckaert. Molecular dynamics simulation of rigid molecules. *Computer Physics Report*, 4(6):346–392, 1986.

[5] S. Duane, A.D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid monte carlo. *Physics Letters B*, 195(2):216 – 222, 1987. ISSN 0370-2693.

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

[7] J. Glover, G. Bradsky, and R. Rusu. Monte carlo pose estimation with quaternion kernels and the bingham distribution. In *RSS*, 2011.

[8] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration*. Springer, 2nd edition, 2006.

[9] T. Hamelryck, J. T. Kent, and A. Krogh. Sampling realistic protein conformations using local structural bias. *PLoS Computational Biology*, 2 (9):e131, 09 2006.

[10] C. Hartmann. An ergodic sampling scheme for constrained hamiltonian systems with applications to molecular dynamics. *Journal of Statistical Physics*, 130:687–711, 2008.

[11] P. D. Hoff. Simulation of the matrix binghamvon mises-fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18:438–456, 2009.

[12] L. Jay. Symplectic partitioned runge-kutta methods for constrained hamiltonian systems. *SIAM Journal of Numerical Analysis*, 33(1):368–387, Feb 1996.

[13] J. T. Kent, P. D.L. Constable, and F. Er. Simulation for the complex Bingham distribution. *Statistics and Computing*, 14(1):53–57, 2004.

[14] Z. Khan, T. Balch, and F. Dellaert. An mcmc-based particle filter for tracking multiple interacting targets. In *ECCV 2004*, volume 3024, pages 279–290. Springer Berlin / Heidelberg, 2004.

[15] K. Kunze and H. Schaeben. The bingham distribution of quaternions and its spherical radon transform in texture analysis. *Mathematical Geology*, 36:917–943, 2004. ISSN 0882-8121.

[16] N. D. Lawrence and R. Urtasun. Non-linear matrix factorization with gaussian processes. In *Proceedings of ICML*, 2009.

[17] T. Lelièvre, M. Rousset, and G. Stoltz. *Free energy computations: A Mathematical Perspective*. Imperial College Press, 2010.

[18] T. Lelièvre, M. Rousset, and G. Stoltz. Langevin dynamics with constraints and computation of free energy differences. 2010. URL `http://arxiv.org/abs/1006.4914`.

[19] B. Marlin. Collaborative filtering: A machine learning perspective. Master's thesis, University of Toronto, 2004.

[20] J. E. Marsden and T. S. Ratiu. *Introduction to mechanics and symmetry*. Springer, 1999.

[21] J. E. Marsden and M. West. Discrete mechanics and variational integrators. *Acta Numerica*, pages 357–514, 2001.

[22] G. Meyer, S. Bonnabel, and Rodolphe Sepulchre. Linear regression under fixed-rank constrains: A riemannian approach. In *Proceedings of International Conference on Machine Learning*, 2011.

[23] R. M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC, 2011.

[24] G. Pons-Moll, A. Baak, J. Gall, L. Leal-Taixe, M. Mueller, H.-P. Seidel, and B. Rosenhahn. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In *ICCV*, 2011.

[25] S. Reich. Symplectic integration of constrained hamiltonian systems by runge-kutta methods. Technical Report 93-13, University of British Columbia, 1993.

[26] W. Rieping, M. Habeck, and M. Nilges. Inferential structure determination. *Science*, 309(5732): 303–306, 2005.

[27] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*, pages 880–887, 2008.

[28] L. Sigal and M. J. Black. Humaneva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion. Technical report, Department of Computer Science, Brown University, 2006.

[29] L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22 (4):1701–1728, 1994.

[30] A. T. A. Wood. The simulation of spherical distributions in the Fisher-Bingham family. *Communications in Statistics - Simulation and Computation*, 16(3):885–898, 1987.

[31] A. T.A. Wood. Simulation of the von Mises-Fisher distribution. *Communications in Statistics - Simulation and Computation*, 23(1):157–164, 1994.

[32] M. Zhou, C. Wang, M. Chen, J. Paisley, D. Dunson, and L. Carin. Nonparametric bayesian matrix completion. In *SAM*, 2010.

# A Appendix

Here, we detail the proofs of the theorems in the main paper.

**Theorem 1** (Detailed Balance)**.** *Let $\hat{\mathcal{H}}$ be $\mathcal{C}^2$-continuous, $\mathcal{M} = \{q \in \mathbb{R}^n | c(q) = 0\}$ be a connected, smooth and differentiable manifold with $\frac{\partial c}{\partial q}$ full-rank everywhere, $M(q)$ be positive definite on $\mathcal{M}$, and $\pi(q)$ be smooth. If $\Phi_h^{\hat{\mathcal{H}}}$ is symmetric and symplectic, then*

$$\int_{Q'}\int_{Q} \pi(q)T(q \to q')dqdq' = \int_{Q}\int_{Q'} \pi(q')T(q' \to q)dq'dq \;, \tag{8}$$

*where $Q, Q' \subset \mathcal{M}$ and $T$ is the transition kernel.*

*Proof.* We begin by showing that CHMC satisfies detailed balance with respect to $\bar{\pi}(p,q) = \exp(-\mathcal{H}(p,q))$ when viewed as a transition on the augmented state space $\mathcal{T}^*\mathcal{M}$. Let $R, R' \subset \mathcal{T}^*\mathcal{M}$ be sufficiently small regions such that the augmented distribution $\bar{\pi}$ is constant over them, with values $\bar{\pi}(R)$ and $\bar{\pi}(R')$ respectively and $R'$ is the image of $R$ under $T$. Let $T(r \to r')$ be the transition kernel utilized by CHMC to transition from $r \in R$ to $r' \in R'$. Since $\Phi_h^{\hat{\mathcal{H}}}$ is symmetric, if $R'$ is the image of $R$ under $T$ then the image of $R'$ under $T$ is $R$ with the sign of the momentum reversed. Furthermore, since the integration method is symplectic, the phase-space volume is preserved, *i.e.*, $\int dR = \int dR' = \delta V$. Thus,

$$\begin{aligned}
&\int_{R'}\int_{R} \frac{1}{Z_{\bar{\pi}}}\bar{\pi}(r)T(r \to r')drdr' \\
&= \frac{1}{Z_{\bar{\pi}}}\bar{\pi}(R)\delta V \min\left(1, \frac{\bar{\pi}(R')}{\bar{\pi}(R)}\right) \\
&= \frac{1}{Z_{\bar{\pi}}}\bar{\pi}(R')\delta V \min\left(1, \frac{\bar{\pi}(R)}{\bar{\pi}(R')}\right) \\
&= \int_{R}\int_{R'} \frac{1}{Z_{\bar{\pi}}}\bar{\pi}(r')T(r' \to r)dr'dr \;,
\end{aligned}$$

where $Z_{\bar{\pi}} = \int_{\mathcal{T}^*\mathcal{M}} \bar{\pi}(r)dr$ is the normalization constant for the augmented distribution. Note that the second equality above can be derived trivially by multiplying the previous equation by $\bar{\pi}(R')/\bar{\pi}(R')$.

$\pi(q)$ can be easily shown to be the marginal distribution of $\bar{\pi}(p,q)$ as follows

$$\begin{aligned}
&\int \bar{\pi}(p,q)dp \\
&= \int_{\mathcal{T}_q^*\mathcal{M}} \pi(q)\frac{1}{Z_{\mathcal{N}_{\mathcal{T}_q^*\mathcal{M}}}}\mathcal{N}(p|0, M(q))dp \\
&= \pi(q)\int_{\mathcal{T}_q^*\mathcal{M}} \frac{1}{Z_{\mathcal{N}_{\mathcal{T}_q^*\mathcal{M}}}}\mathcal{N}(p|0, M(q))dp \\
&= \pi(q) \;,
\end{aligned}$$

where $Z_{\mathcal{N}_{\mathcal{T}_q^*\mathcal{M}}} = \int_{\mathcal{T}_q^*\mathcal{M}} \mathcal{N}(p|0, M(q))dp$ is the normalizing constant of the Gaussian distribution restricted to the tangent space. Combining this with the detailed balance result on the augmented space gives

$$\begin{aligned}
&\int_{Q'}\int_{Q} \pi(q)T(q \to q')dqdq' \\
&= \int_{Q'}\int\int_{Q}\int \pi(q)\pi(p|q)T(p, q \to p', q')dpdqdp'dq' \\
&= \int_{R'}\int_{R} \bar{\pi}(r)T(r \to r')drdr' \\
&= \int_{R}\int_{R'} \bar{\pi}(r')T(r' \to r)dr'dr \\
&= \int_{Q}\int\int_{Q'}\int \pi(q')\pi(p'|q')T(p', q' \to p, q)dp'dq'dpdq \\
&= \int_{Q}\int_{Q'} \pi(q')T(q' \to q)dq'dq \;.
\end{aligned}$$

□

**Theorem 2** (Accessibility)**.** *Let $\hat{\mathcal{H}}$ be $\mathcal{C}^2$-continuous, $\mathcal{M} = \{q \in \mathbb{R}^n | c(q) = 0\}$ be a connected, smooth and differentiable manifold with $\frac{\partial c}{\partial q}$ full-rank everywhere, and $M(q)$ be positive definite on $\mathcal{M}$. If $\Phi_h^{\hat{\mathcal{H}}}$ is symmetric, symplectic and consistent, then for any $q_0, q_1 \in \mathcal{M}$ and $h$ sufficiently small, there exist finite $p_0 \in \mathcal{T}_{q_0}^*\mathcal{M}$, $p_1 \in \mathcal{T}_{q_1}^*\mathcal{M}$ and Lagrange multipliers $\lambda_0, \lambda_1$ such that $(p_1, q_1) = \Phi_h^{\hat{\mathcal{H}}}(p_0, q_0)$.*

*Proof.* If the integrator is symplectic then there exists a corresponding discrete Lagrangian $\mathcal{L}'_h(q_0, q_1)$ for which the result of the integrator satisfies the discrete Euler-Lagrange equations $p_0 + \frac{\partial_h \mathcal{L}'(q_0,q_1)}{\partial q_0} = \lambda_0^T C(q_0)$. See Theorem 2.1.1 of [21]; Theorem 5.6, Section IX.5.2 of [8] for a formal proof. For any given $q_0$ and $q_1$, $p_0$ can be computed as $p_0 = -\frac{\partial \mathcal{L}'_h(q_0,q_1)}{\partial q_0} + \lambda_0^T C(q_0)$ with $\lambda_0$ chosen so that $p_0 \in \mathcal{T}_{q_0}^*\mathcal{M}$. By symmetry of the integrator, we also have $p_1 = -\frac{\partial \mathcal{L}'_{-h}(q_1,q_0)}{\partial q_1} + \lambda_1^T C(q_1)$ with $\lambda_1$ chosen so that $p_1 \in \mathcal{T}_{q_1}^*\mathcal{M}$. Such choices of $\lambda_0$ and $\lambda_1$ exist so long as $C(q_0)$, $M(q_0)$, $C(q_1)$ and $M(q_1)$ all have full rank.

Since $\Phi_h^{\hat{\mathcal{H}}}$ is consistent we have

$$\mathcal{L}'_h(q_0, q_1) = \int_0^h \mathcal{L}(q(t), \dot{q}(t))dt + h^r e_h(q_0, q_1) \ , \quad (9)$$

where $r \geq 1$, $e_h$ is a smooth and bounded error function and $q(t) : [0, h] \to \mathcal{M}$ is the solution to the continuous Euler-Lagrange equations on $\mathcal{M}$ with $q(0) = q_0$ and $q(h) = q_1$. This solution exists, and thus the above holds, so long as $\mathcal{M}$ is connected and $\mathcal{L}$ is twice differentiable. This holds since $\mathcal{L} = p^T q - \hat{\mathcal{H}}$, and $\hat{\mathcal{H}}$ is twice differentiable by assumption. Then,

$$\begin{aligned}
&\frac{\partial \mathcal{L}'_h}{\partial q_0}(q_0, q_1) \\
&= \int_0^h \frac{\partial}{\partial q_0}\mathcal{L}(q(t), \dot{q}(t))dt + h^r \frac{\partial e_h}{\partial q_0}(q_0, q_1) \\
&= \int_0^h \left( \frac{\partial \mathcal{L}}{\partial q}\frac{\partial q}{\partial q_0} + \frac{\partial \mathcal{L}}{\partial \dot{q}}\frac{\partial \dot{q}}{\partial q_0} \right) dt + h^r \frac{\partial e_h}{\partial q_0}(q_0, q_1) \\
&= \left. \frac{\partial \mathcal{L}}{\partial \dot{q}}\frac{\partial q}{\partial q_0} \right|_0^h + \int_0^h \left( \frac{\partial \mathcal{L}}{\partial q} - \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{q}} \right) \frac{\partial q}{\partial q_0} dt + h^r \frac{\partial e_h}{\partial q_0}(q_0, q_1) \\
&= -\frac{\partial \mathcal{L}}{\partial \dot{q}}(q_0, \dot{q}_0) + h^r \frac{\partial e_h}{\partial q_0}(q_0, q_1) \ .
\end{aligned}$$

The third equality can be derived by doing partial integration, while the last step comes from the properties of the Euler-Lagrange equations. Similarly, one can derive that $\frac{\partial \mathcal{L}'_h}{\partial q_1}(q_0, q_1) = \frac{\partial \mathcal{L}}{\partial \dot{q}}(q_1, \dot{q}_1) + h^r \frac{\partial e_h}{\partial q_1}(q_0, q_1)$. Thus, since $\mathcal{L}$ is differentiable and $e_h$ is smooth and bounded, the gradients $\partial \mathcal{L}'$, and therefore the momenta $p_0, p_1$, exist and are finite. □

**Theorem 3.** *Let $\hat{\mathcal{H}}$ be $\mathcal{C}^2$-continuous, $\mathcal{M} = \{q \in \mathbb{R}^n | c(q) = 0\}$ be a connected, smooth and differentiable manifold with $\frac{\partial c}{\partial q}$ full-rank everywhere, $M(q)$ be positive definite on $\mathcal{M}$, $\Phi_h^{\hat{\mathcal{H}}}$ be symmetric, symplectic and consistent, and $\pi(q)$ be strictly positive on $\mathcal{M}$. Let $\mathcal{B}_\ell(q) = \{q' \in \mathcal{M} | d(q', q) \leq \ell\}$ be a ball defining all points on $\mathcal{M}$ with geodesic distance at most $\ell$ from $q$. If there is a constant $\ell > 0$ such that for every $q \in \mathcal{M}$, $q' \in \mathcal{B}_\ell(q)$ there is a unique choice of Lagrange multiplier and momentum $p \in \mathcal{T}_q^*\mathcal{M}$, $p' \in \mathcal{T}_{q'}^*\mathcal{M}$ for which $(p', q') = \Phi_h^{\hat{\mathcal{H}}}(p, q)$, and, if, given $(p, q)$, this is the Lagrange multiplier selected by the integrator, then, for $h$ sufficiently small and any $q_0, q_1 \in \mathcal{M}$, $\pi(q_1) > 0$, there exists an $n \in \mathbb{N}$ such that*

$$T^n(q_0 \to q_1) > 0 \ . \quad (10)$$

*Proof.* If $\mathcal{M}$ is connected, then there exists a geodesic curve between any two points $q$ and $q'$ on the manifold. Let $d(q, q')$ be the geodesic distance between $q$ and $q'$. Let $q_0, \ldots, q_n$ be an ordered sequence of points on the geodesic between $q$ and $q'$ with $q_0 = q$, $q_n = q'$, and $n = \lceil \frac{d(q,q')}{\ell} \rceil$, such that $d(q_{i-1}, q_i) \leq \ell$. By assumption, for each $1 \leq i \leq n$, there exists $p_{i-1}$ such that $(p'_i, q_i) = \Phi_h^{\mathcal{H}}(p_{i-1}, q_{i-1})$. The probability density of making this sequence of transitions is $T^n(q \to q') = \prod_{i=1}^n T(q_{i-1} \to q_i)$ which is non-zero so long as all the individual transitions have non-zero probability. This holds by Theorem 2 and because $\pi(q)$ is strictly positive. □

**Lemma 1.** *Let $\hat{\mathcal{H}}$ be $\mathcal{C}^2$-continuous, $\mathcal{M} = \{q \in \mathbb{R}^n | c(q) = 0\}$ be a connected, smooth and differentiable manifold with $\frac{\partial c}{\partial q}$ full-rank everywhere, $M(q)$ be positive definite everywhere, $\Phi_h^{\hat{\mathcal{H}}}$ be symmetric, symplectic and consistent, and $\pi(q)$ be strictly positive on $\mathcal{M}$. Let $\mathcal{B}_\ell(q) = \{q' \in \mathcal{M} | d(q', q) \leq \ell\}$ be a ball defining all points on $\mathcal{M}$ with geodesic distance at most $\ell$ from $q$. If there is a constant $\ell > 0$ such that for every $q \in \mathcal{M}$, $q' \in \mathcal{B}_\ell(q)$ there is a unique choice of Lagrange multiplier and momentum $p \in \mathcal{T}_q^*\mathcal{M}$, $p' \in \mathcal{T}_{q'}^*\mathcal{M}$ for which $(p', q') = \Phi_h^{\hat{\mathcal{H}}}(p, q)$, and, if, given $(p, q)$, this is the Lagrange multiplier selected by the integrator, then CHMC is aperiodic. That is, there is no period $p$ and disjoint subsets $Q_0, \ldots, Q_{p-1} \subset \mathcal{M}$ such that for $i = 0, \ldots, p-1$ and $T(q \to Q_{(i+1)\%p}) = 1$ for all $q \in Q_i$.*

*Proof.* Suppose by contradiction, that CHMC is periodic, and that there exists a period $p$ and a sequence of nonempty, disjoint subsets $Q_0, \ldots, Q_{p-1} \subset \mathcal{M}$ such that for $i = 0, \ldots, n-1$ $T(q \to Q_{(i+1)\%p}) = 1$ for all $q \in Q_i$. This implies that $T^{1+\alpha p}(q \to Q') = 1$ for all $\alpha \in \mathbb{N}$ and $Q' \subset \mathcal{M} \setminus Q_{(i+1)\%n}$. Let $Q' \subset \mathcal{M}$ be any nonempty subset of $\mathcal{M}$ which is at least partially distinct from each $Q_i$, *i.e.*, $\pi(Q' \setminus Q_i) > 0$ for all $i$. Then, by Theorem 3, for all $i$ and $q \in Q_i$ $T^{1+\alpha p}(q \to Q' \setminus Q_{(i+1)\%p}) > 0$ where $\alpha$ is such that $1 + \alpha p \geq n$ with $n$ from Theorem 3. This contradicts the original assumption and, hence, CHMC must be aperiodic. □

**Theorem 4.** *Let $\hat{\mathcal{H}}$ be $\mathcal{C}^2$-continuous, $\mathcal{M} = \{q \in \mathbb{R}^n | c(q) = 0\}$ be a connected, smooth and differentiable manifold with $\frac{\partial c}{\partial q}$ full-rank everywhere, $M(q)$ be positive definite on $\mathcal{M}$, $\Phi_h^{\hat{\mathcal{H}}}$ be symmetric, symplectic and consistent, and $\pi(q)$ be smooth and strictly positive on $\mathcal{M}$. Let $\mathcal{B}_\ell(q) = \{q' \in \mathcal{M} | d(q', q) \leq \ell\}$ be a ball defining all points on $\mathcal{M}$ with geodesic distance at most $\ell$ from $q$. If there is a constant $\ell > 0$ such that for every $q \in \mathcal{M}$, $q' \in \mathcal{B}_\ell(q)$ there is a unique choice of Lagrange multiplier and momentum $p \in \mathcal{T}_q^*\mathcal{M}$, $p' \in \mathcal{T}_{q'}^*\mathcal{M}$ for which $(p', q') = \Phi_h^{\hat{\mathcal{H}}}(p, q)$, and, if, given $(p, q)$, this is the Lagrange multiplier selected by the integrator, then for all $q_0 \in \mathcal{M}$*

$$\lim_{n \to \infty} \| T^n(q_0 \to \cdot) - \pi(\cdot) \| = 0 \ .$$

*Proof.* Since CHMC satisfies detailed balance with respect to $\pi$ (Theorem 1), is $\pi$-irreducible (Theorem 3)

and is aperiodic (Lemma 1), then the desired result follows by Theorem 1 of [29]. □