# Supplementary Material for Magnetic Hamiltonian Monte Carlo

## 1 Section 3 and 4 Proofs

Here we provide proofs for results discussed in Section 3 of the main text regarding non-canonical dynamics.

**Lemma 1.** *The map $\mathbf{\Phi}_{\tau,H}^{\mathbf{A}}(\theta, \mathbf{p})$ defined by integrating the non-canonical Hamiltonian system*

$$\frac{\mathrm{d}}{\mathrm{dt}} \begin{bmatrix} \theta(t) \\ \mathbf{p}(t) \end{bmatrix} = \mathbf{A}\nabla_{\theta,\mathbf{p}} H(\theta(t), \mathbf{p}(t)) \tag{1}$$

*with initial conditions $(\theta, \mathbf{p})$ for time $\tau$, where $\mathbf{A} \in \mathcal{M}_{2n \times 2n}$ is any invertible, antisymmetric matrix induces a flow on the coordinates $(\theta, \mathbf{p})$ that is still energy-conserving $(\partial_\tau H(\mathbf{\Phi}_{\tau,H}^{\mathbf{A}}(\theta, \mathbf{p})) = 0)$ and symplectic with respect to $\mathbf{A}$ $([\nabla_{\theta,\mathbf{p}}\mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p})]^\top \mathbf{A}^{-1}[\nabla_{\theta,\mathbf{p}}\mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p})] = \mathbf{A}^{-1})$ which also implies volume-preservation of the flow.*

*Proof.* The proofs of both results simply uses the antisymmetry of $\mathbf{A}$.

 **Energy-Conservation** – Simply, we have that:

$$\partial_\tau H(\mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p})) = \nabla_{\theta,\mathbf{p}} H(\mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p}))\partial_\tau \mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p}) = \tag{2}$$

$$\nabla_{\theta,\mathbf{p}} H(\mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p}))^\top \mathbf{A}\nabla_{\theta,\mathbf{p}} H(\mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p})) = 0 \tag{3}$$

using the antisymmetry of $\mathbf{A}$ and symmetry of $\nabla_{\theta,\mathbf{p}} H(\mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p}))\nabla_{\theta,\mathbf{p}} H(\mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p}))^\top$.

 **Symplecticness** – We must show that the Jacobian of the flow generated by the dynamics preserves the non-canonical structure matrix $\mathbf{A}$, which amounts to showing:

$$\underbrace{[\nabla_{\theta,\mathbf{p}}\mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p})]^\top}_{F(\tau)^\top} \mathbf{A}^{-1} \underbrace{[\nabla_{\theta,\mathbf{p}}\mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p})]}_{F(\tau)} = \mathbf{A}^{-1} \tag{4}$$

where we define $F(\tau) = \nabla_{\theta,\mathbf{p}}\mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p})$ as the time-evolving Jacobian of the flow. First, note that $F(\tau)$ can be equivalently described as the solution to the differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}\tau} F(\tau) = \mathbf{A}\nabla_{\theta,\mathbf{p}}\nabla_{\theta,\mathbf{p}} H(\mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p}))F(\tau) \tag{5}$$

with the initial condition $F(0) = \mathbf{I}_{2d}$ (the Jacobian for the identity map at $t = 0$). Trivially, we have:

$$F(0)\mathbf{A}^{-1}F(0) = \mathbf{A}^{-1} \tag{6}$$

Then note that:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}(F(\tau)^\top \mathbf{A}^{-1}F(\tau)) =$$

$$F(\tau)^\top \mathbf{A}^{-1}\mathbf{A}\nabla_{\theta,\mathbf{p}}\nabla_{\theta,\mathbf{p}} H(\mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p}))F(\tau) + F(\tau)^\top \nabla_{\theta,\mathbf{p}}\nabla_{\theta,\mathbf{p}} H(\mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p}))\mathbf{A}^\top \mathbf{A}^{-1}F(\tau)$$

$$= F(\tau)^\top \nabla_{\theta,\mathbf{p}}\nabla_{\theta,\mathbf{p}} H(\mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p}))F(\tau) - F(\tau)^\top \nabla_{\theta,\mathbf{p}}\nabla_{\theta,\mathbf{p}} H(\mathbf{\Phi}_{\tau,H}(\theta, \mathbf{p}))F(\tau) = \mathbf{0}$$

as desired by simply using $\mathbf{A}^\top = -\mathbf{A}$. □

 **Time-Reversibility** – However, crucially it is *not* the case that the Hamilton equations are time-reversible in the traditional sense of canonical Hamiltonian dynamics.

**Lemma 2.** *If $(\theta(t), \mathbf{p}(t))$ is a solution to the non-canonical dynamics:*

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\theta(t)\\\mathbf{p}(t)\end{bmatrix} = \underbrace{\begin{bmatrix}\mathbf{E} & \mathbf{F}\\-\mathbf{F}^\top & \mathbf{G}\end{bmatrix}}_{\mathbf{A}}\begin{bmatrix}\nabla_\theta H(\theta(t),\mathbf{p}(t))\\\nabla_\mathbf{p} H(\theta(t),\mathbf{p}(t))\end{bmatrix} \tag{7}$$

*then $(\widetilde{\theta}(t), \widetilde{\mathbf{p}}(t)) = (\theta(-t), -\mathbf{p}(-t))$ is a solution to the modified non-canonical dynamics:*

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\widetilde{\theta}(t)\\\widetilde{\mathbf{p}}(t)\end{bmatrix} = \underbrace{\begin{bmatrix}-\mathbf{E} & \mathbf{F}\\-\mathbf{F}^\top & -\mathbf{G}\end{bmatrix}}_{\widetilde{\mathbf{A}}}\begin{bmatrix}\nabla_{\widetilde{\theta}} H(\widetilde{\theta}(t),\mathbf{p}(t))\\\nabla_{\widetilde{\mathbf{p}}} H(\widetilde{\theta}(t),\widetilde{\mathbf{p}}(t))\end{bmatrix} \tag{8}$$

*if $H(\theta, \mathbf{p}) = H(\theta, -\mathbf{p})$. In particular if $\mathbf{E} = \mathbf{G} = 0$ then $\mathbf{A} = \widetilde{\mathbf{A}}$, which reduces to the traditional time-reversal symmetry of canonical Hamiltonian dynamics.*

*Proof.* A direct calculation yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\tilde{\theta}(t)\\\tilde{\mathbf{p}}(t)\end{bmatrix} = \begin{bmatrix}-\frac{\mathrm{d}}{\mathrm{d}t}\theta(-t)\\\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{p}(-t)\end{bmatrix} = \begin{bmatrix}-\mathbf{E}\nabla_\theta H(\theta(-t)) - \mathbf{F}\nabla_\mathbf{p} H(\theta(-t))\\-\mathbf{F}^\top\nabla_\theta H(\theta(-t)) + \mathbf{G}\nabla_\mathbf{p} H(\theta(-t))\end{bmatrix}$$

$$= \begin{bmatrix}-\mathbf{E}\nabla_{\tilde{\theta}} H(\tilde{\theta}(t)) + \mathbf{F}\nabla_{\tilde{\mathbf{p}}} H(\tilde{\theta}(t))\\-\mathbf{F}^\top\nabla_{\tilde{\theta}} H(\tilde{\theta}(t)) - \mathbf{G}\nabla_{\tilde{\mathbf{p}}} H(\tilde{\theta}(t))\end{bmatrix} = \underbrace{\begin{bmatrix}-\mathbf{E} & \mathbf{F}\\-\mathbf{F}^\top & -\mathbf{G}\end{bmatrix}}_{\tilde{\mathbf{A}}}\begin{bmatrix}\nabla_{\tilde{\theta}} H(\tilde{\theta}(t))\\\nabla_{\tilde{\mathbf{p}}} H(\tilde{\theta}(t))\end{bmatrix}$$

$\square$

## 1.1 Non-Canonical Dynamics Variable Augmentation

As remarked in the main text it is necessary to flip the $\mathbf{E}$ and $\mathbf{G}$ matrices at the end of a deterministic simulation of the Hamiltonian dynamics in order to render the proposal time-reversible which is in turn necessary to satisfy detailed balance. This is crucial for the correctness of the algorithm especially when an approximate simulation of the dynamics is used (as is always often the case).

In particular, say that we wish to use $\mathbf{\Phi}^\mathbf{A}_{\tau, H}(\theta, \mathbf{p})$ as a transition kernel with fixed, non-zero values of $\mathbf{E} = \mathbf{E}_0$ and $\mathbf{G} = \mathbf{G}_0$. We first augment the state-space by placing a symmetric, binary distribution independently over $\mathbf{E}$ and $\mathbf{G}$ such that $p(\mathbf{E} = \mathbf{E}_0) = p(\mathbf{E} = -\mathbf{E}_0) = 1/2$ and $p(\mathbf{G} = \mathbf{G}_0) = p(\mathbf{G} = -\mathbf{G}_0) = 1/2$, independently of $\theta, \mathbf{p}$:

$$\rho(\theta, \mathbf{p}, \mathbf{E}, \mathbf{G}) \propto e^{-H(\theta, \mathbf{p})} p(\mathbf{E}) p(\mathbf{G}). \tag{9}$$

Importantly, this augmentation leaves the distribution over $\theta, \mathbf{p}$ intact when $\mathbf{E}$ and $\mathbf{G}$ are marginalized out. Just as applying the momentum flip operator, $\mathbf{\Phi}_\mathbf{p} : (\theta, \mathbf{p}, \mathbf{E}, \mathbf{G}) \to (\theta, -\mathbf{p}, \mathbf{E}, \mathbf{G})$, is a deterministic, energy-preserving, volume-preserving transformation, the $\mathbf{E}, \mathbf{G}$ flip operators, $\mathbf{\Phi}_\mathbf{E} : (\theta, \mathbf{p}, \mathbf{E}, \mathbf{G}) \to (\theta, \mathbf{p}, -\mathbf{E}, \mathbf{G})$ and $\mathbf{\Phi}_\mathbf{G} : (\theta, \mathbf{p}, \mathbf{E}, \mathbf{G}) \to (\theta, \mathbf{p}, \mathbf{E}, -\mathbf{G})$, are also deterministic, energy-preserving, volume-preserving transformations that leave (9) invariant for this particular augmentation with $p(\mathbf{E})$ and $p(\mathbf{G})$. We can now build a self-inverse operator $\widetilde{\mathbf{\Phi}}^\mathbf{A}_{\tau, H}(\theta, \mathbf{p})$, composed of simulating the Hamiltonian flow as $\mathbf{\Phi}^\mathbf{A}_{\tau, H}(\theta, \mathbf{p})$ plus $\mathbf{\Phi}_\mathbf{E} \circ \mathbf{\Phi}_\mathbf{G} \circ \mathbf{\Phi}_\mathbf{p}$, a flip of $\mathbf{p}, \mathbf{E}, \mathbf{G}$, as:

$$\widetilde{\mathbf{\Phi}}^\mathbf{A}_{\tau, H}(\theta, \mathbf{p}) = \mathbf{\Phi}_\mathbf{E} \circ \mathbf{\Phi}_\mathbf{G} \circ \mathbf{\Phi}_\mathbf{p} \circ \mathbf{\Phi}^\mathbf{A}_{\tau, H}(\theta, \mathbf{p}) \tag{10}$$

Now we have constructed a deterministic, self-inverse map $\widetilde{\mathbf{\Phi}}^\mathbf{A}_{\tau, H}(\theta, \mathbf{p})$. $\widetilde{\mathbf{\Phi}}^\mathbf{A}_{\tau, H}(\theta, \mathbf{p})$ can now be used as the proposal for a reversible MCMC algorithm.

An important point to note is that our choice of variable augmentation strategy, namely augmenting with binary distribution, is certainly not unique. However, it is perhaps the most natural and simplest choice which avoids the repetitive computation of matrix exponentials/diagonalizations since the approximate flow detailed in Section 2.2 will only need to compute matrix exponentials once upfront for $\pm\mathbf{G}$.

## 1.2 Mass Preconditioning Proofs

A common variation on standard HMC dynamics is to set the kinetic energy term in the Hamiltonian $H(\theta, \mathbf{p})$ to $\frac{1}{2}\mathbf{p}^\top \mathbf{M}^{-1}\mathbf{p}$ for some symmetric positive-definite matrix $\mathbf{M}$, and sample the initial momentum variable $\mathbf{p}$ from the corresponding distribution $\mathcal{N}(0, \mathbf{M})$. However, we can contextualize preconditioning using a non-canonical $\mathbf{A}$ matrix in the following manner:

**Lemma 3.** *i) Preconditioned HMC with momentum variable* $\mathbf{p} \sim \mathcal{N}(0, \mathbf{M})$ *in the* $(\theta, \mathbf{p})$ *coordinates, is exactly equivalent to simulating non-canonical HMC with* $\mathbf{p}' = \mathbf{M}^{-1/2}\mathbf{p} \sim \mathcal{N}(0, \mathbf{I})$ *and the non-canonical matrix:*

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{M}^{1/2} \\ -(\mathbf{M}^{1/2})^\top & \mathbf{0} \end{bmatrix}$$

*and then transforming back to* $(\theta, \mathbf{p})$ *coordinates using* $\mathbf{p} = \mathbf{M}^{1/2}\mathbf{p}'$. *Here* $\mathbf{M}^{1/2}$ *is a Cholesky factor for* $\mathbf{M}$.

*ii) On the other hand if we apply a change of basis (via an invertible matrix* $\mathbf{F}$) *to our coordinates* $\theta' = \mathbf{F}^{-1}\theta$, *simulate HMC in the* $(\theta', \mathbf{p})$ *coordinates, and transform back to the original basis using* $\mathbf{F}$, *this is exactly equivalent to non-canonical HMC with*

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{F} \\ -\mathbf{F}^\top & \mathbf{0} \end{bmatrix}$$

*Proof.* We first prove the equivalence regarding the change of basis in momentum space. Under the $\mathbf{M}$ mass matrix variant of HMC, $\mathbf{p}$ is drawn from a $\mathcal{N}(\mathbf{0}, \mathbf{M})$ distribution, and the dynamics of $\theta, \mathbf{p}$ are then given by

$$\frac{\mathrm{d}}{\mathrm{dt}} \begin{bmatrix} \theta \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{M}^{-1}\mathbf{p} \\ -\nabla_\theta U(\theta) \end{bmatrix}$$

Denoting the upper-triangular Cholesky factor of $\mathbf{M}^{-1}$ by $\mathbf{M}^{-1/2}$, and introducing a new variable $\mathbf{p}' = \mathbf{M}^{-1/2}\mathbf{p}$, we obtain the following dynamics for the joint variable $(\theta, \mathbf{p}')$:

$$\frac{\mathrm{d}}{\mathrm{dt}} \begin{bmatrix} \theta \\ \mathbf{p}' \end{bmatrix} = \frac{\mathrm{d}}{\mathrm{dt}} \begin{bmatrix} \theta \\ \mathbf{M}^{-1/2}\mathbf{p} \end{bmatrix} = \begin{bmatrix} (\mathbf{M}^{-1/2})^\top \mathbf{p}' \\ -\mathbf{M}^{-1/2}\nabla_\theta U(\theta) \end{bmatrix} = \begin{bmatrix} \mathbf{0} & (\mathbf{M}^{-1/2})^\top \\ -\mathbf{M}^{-1/2} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \nabla_\theta U(\theta) \\ \mathbf{p}' \end{bmatrix}$$

Further, note that if the marginal distribution of $\mathbf{p}$ is $\mathcal{N}(0, \mathbf{M})$, then under this change of variables $\mathbf{p}'$ has the marginal distribution $\mathcal{N}(0, \mathbf{I})$. Thus, simulating canonical HMC with a non-identity mass matrix is equivalent to making a change of basis in momentum space, simulating non-canonical HMC with a particular choice of non-canonical $\mathbf{A}$ matrix, and finally reverting back to the original basis.

We now prove the equivalence regarding the change of basis in $\theta$ space, which follows similarly. Consider non-canonical HMC on the state-momentum pair $(\theta, \mathbf{p})$, with the antisymmetric matrix $\mathbf{A}$ taking the particular form

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{F} \\ -\mathbf{F}^\top & \mathbf{0} \end{pmatrix}$$

The states $\theta, \mathbf{p}$ obtained from this algorithm are equal to those obtained by first changing basis to $\theta' = \mathbf{F}^{-1}\theta$, then simulating standard HMC dynamics for the pair $(\theta', \mathbf{p})$ with respect to the Hamiltonian

$$H'(\theta', \mathbf{p}) = U'(\theta') + \frac{1}{2}\mathbf{p}^\top \mathbf{p}$$

$$= U(\mathbf{F}\theta) + \frac{1}{2}\mathbf{p}^\top \mathbf{p}$$

and then reverting to the original basis as $\theta = \mathbf{F}\theta'$. To see this, first note that if we denote the distribution on the coordinates $\theta$ corresponding to the potential $U$ by $\pi(\theta) = e^{-U(\theta)}$, then the corresponding distribution on the coordinates $\theta'$ is given by $\pi'$, where

$$\pi'(\theta') = \det(\mathbf{F})\pi(\mathbf{F}\theta')$$

The corresponding potential $U'$ is therefore given by

$$U'(\theta') = U(\mathbf{F}\theta')$$

Running canonical HMC dynamics targeting the Hamiltonian $H'$ yields the dynamics:

$$\frac{\mathrm{d}}{\mathrm{dt}} \begin{bmatrix} \theta' \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{p} \\ -\nabla_{\theta'} U(\theta') \end{bmatrix}$$

But note then that the dynamics of the original coordinates are then given by:

$$\frac{\mathrm{d}}{\mathrm{dt}} \begin{bmatrix} \theta \\ \mathbf{p} \end{bmatrix} = \frac{\mathrm{d}}{\mathrm{dt}} \begin{bmatrix} \mathbf{F}\theta' \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{F}\mathbf{p} \\ -\nabla_{\theta'} U(\theta') \end{bmatrix} = \begin{bmatrix} \mathbf{F}\mathbf{p} \\ -\nabla_{\theta'} U(\mathbf{F}\theta) \end{bmatrix} = \begin{bmatrix} \mathbf{F}\mathbf{p} \\ -(\mathbf{F}^\top)\nabla_\theta U(\theta) \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{0} & \mathbf{F} \\ -\mathbf{F}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \nabla_\theta U(\theta) \\ \mathbf{p} \end{bmatrix}$$

which are exactly a special case of non-canonical HMC dynamics described above. $\qquad\square$

# 2 Magnetic HMC (MHMC)

Here we provide proofs related to the dynamics of magnetic HMC and it's symplectic integration scheme.

## 2.1 Non-Canonical Dynamics and Magnetism

We first establish the connection between the particular subcase of non-canonical Hamiltonian dynamics where

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{G} \end{bmatrix}$$

and Newton's law for a charged particle coupled to a magnetic field.

**Lemma 4.** *In 3-dimensions the non-canonical Hamiltonian dynamics, with Hamiltonian $H(\theta, \mathbf{p}) = U(\theta) + \frac{1}{2}\mathbf{p}^\top \mathbf{p}$, correspond to simulating the differential equations:*

$$\frac{\mathrm{d}}{\mathrm{dt}} \begin{bmatrix} \theta \\ \mathbf{p} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{G} \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} \nabla_\theta H \\ \nabla_\mathbf{p} H \end{bmatrix} \equiv \underbrace{\begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{G} \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} \nabla_\theta U(\theta) \\ \mathbf{p} \end{bmatrix} \qquad (11)$$

*where*

$$\mathbf{G} = \begin{bmatrix} 0 & -b_3 & b_2 \\ b_3 & 0 & -b_1 \\ -b_2 & b_1 & 0 \end{bmatrix}$$

*are equivalent to the Newtonian mechanics of a charged particle (with unit mass and charge) coupled to a magnetic field $\vec{\mathbf{B}} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$ which take the form:*

$$\frac{\mathrm{d}^2\theta}{\mathrm{dt}^2} = -\nabla_\theta U(\theta) + \frac{\mathrm{d}\theta}{\mathrm{dt}} \times \vec{\mathbf{B}} \qquad (12)$$

*where $\theta$ is simply a 3-dimensional vector and $\times$ the cross-product.*

*Proof.* If we let $\theta$ and $\mathbf{p}$ denote our position and momentum coordinates in 3 dimensions then Newton's law for a charged particle in a magnetic field (with $m = q = 1$) is:

$$\frac{\mathrm{d}^2\theta}{\mathrm{dt}^2} = -\nabla_\theta U(\theta) + \frac{\mathrm{d}\theta}{\mathrm{dt}} \times \vec{\mathbf{B}} \qquad (13)$$

Defining momentum canonically as $\frac{\mathrm{d}\theta}{\mathrm{dt}} = \mathbf{p}$ we have:

$$\frac{\mathrm{d}}{\mathrm{dt}} \begin{bmatrix} \theta \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{p} \\ -\nabla_\theta U(\theta) + \frac{\mathrm{d}\theta}{\mathrm{dt}} \times \vec{\mathbf{B}} \end{bmatrix} = \begin{bmatrix} \mathbf{p} \\ -\nabla_\theta U(\theta) + \mathbf{G}\mathbf{p} \end{bmatrix} \equiv \underbrace{\begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{G} \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} \nabla_\theta U(\theta) \\ \mathbf{p} \end{bmatrix} \qquad (14)$$

$\qquad\square$

We now show that the dynamics used in magnetic HMC cannot be reproduced by simply choosing a different smooth Hamiltonian, $H'(\theta, \mathbf{p})$ and using the canonical $\mathbf{A}$ matrix:

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{bmatrix}$$

to generate the dynamics.

**Lemma 5.** *The non-canonical Hamiltonian dynamics with magnetic $\mathbf{A}$ and Hamiltonian $H(\theta, \mathbf{p}) = U(\theta) + \frac{1}{2}\mathbf{p}^\top \mathbf{p}$ cannot be obtained using canonical Hamiltonian dynamics for any choice of smooth Hamiltonian.*

*Proof.* Consider the ODEs corresponding to non-canonical dynamics with magnetic $\mathbf{A}$ and $H(\theta, \mathbf{p}) = U(\theta) + \frac{1}{2}\mathbf{p}^\top \mathbf{p}$:

$$\frac{\mathrm{d}}{\mathrm{dt}} \begin{bmatrix} \theta \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{G} \end{bmatrix} \begin{bmatrix} \nabla_\theta U(\theta) \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{p} \\ -\nabla_\theta U(\theta) + \mathbf{G}\mathbf{p} \end{bmatrix}. \tag{15}$$

Assume, to obtain a contradiction, that these canonical Hamiltonian dynamics can be reproduced for some choice of smooth $H'(\theta, \mathbf{p})$ and canonical $\mathbf{A}$ matrix (i.e. $\mathbf{E} = \mathbf{G} = \mathbf{0}$, $\mathbf{F} = \mathbf{I}$):

$$\frac{\mathrm{d}}{\mathrm{dt}} \begin{bmatrix} \theta \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \nabla_\theta H'(\theta, \mathbf{p}) \\ \nabla_{\mathbf{p}} H'(\theta, \mathbf{p}) \end{bmatrix} = \begin{bmatrix} \mathbf{p} \\ -\nabla_\theta U(\theta) + \mathbf{G}\mathbf{p} \end{bmatrix}. \tag{16}$$

This implies:

$$\begin{bmatrix} \nabla_{\mathbf{p}} H'(\theta, \mathbf{p}) \\ \nabla_\theta H'(\theta, \mathbf{p}) \end{bmatrix} = \begin{bmatrix} \mathbf{p} \\ \nabla_\theta U(\theta) - \mathbf{G}\mathbf{p} \end{bmatrix} \implies \begin{bmatrix} \nabla_\theta \nabla_{\mathbf{p}} H'(\theta, \mathbf{p}) \\ \nabla_{\mathbf{p}} \nabla_\theta H'(\theta, \mathbf{p}) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ -\mathbf{G} \end{bmatrix}. \tag{17}$$

However, as long as the 2nd-order mixed partial derivatives are continuous they must be equal; so the conclusion follows. $\qquad\square$

## 2.2 Symplectic Integrator for Magnetic Dynamics

We begin by considering the symmetric splitting:

$$H(\theta, \mathbf{p}) = \underbrace{U(\theta)/2}_{H_1(\theta)} + \underbrace{\mathbf{p}^T \mathbf{p}/2}_{H_2(\mathbf{p})} + \underbrace{U(\theta)/2}_{H_1(\theta)} \tag{18}$$

The corresponding non-canonical dynamics for the sub-Hamiltonians $H_1(\theta)$ and $H_2(\mathbf{p})$ are:

$$\frac{\mathrm{d}}{\mathrm{dt}} \begin{bmatrix} \theta \\ \mathbf{p} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{E} & \mathbf{F} \\ -\mathbf{F}^\top & \mathbf{G} \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} \nabla_\theta U(\theta)/2 \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{E}\nabla_\theta U(\theta)/2 \\ -\mathbf{F}^\top \nabla_\theta U(\theta)/2 \end{bmatrix} \tag{19}$$

and

$$\frac{\mathrm{d}}{\mathrm{dt}} \begin{bmatrix} \theta \\ \mathbf{p} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{E} & \mathbf{F} \\ -\mathbf{F}^\top & \mathbf{G} \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} \mathbf{0} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{F}\mathbf{p} \\ \mathbf{G}\mathbf{p} \end{bmatrix}. \tag{20}$$

We denote the corresponding flows by $\mathbf{\Phi}^{\mathbf{A}}_{\epsilon, H_1(\theta)}$ and $\mathbf{\Phi}^{\mathbf{A}}_{\epsilon, H_2(\mathbf{p})}$ respectively. The flow in (19) is generally not explicitly tractable unless we take $\mathbf{E} = \mathbf{0}$ – in which case it is solved by an Euler translation as for standard Hamiltonian dynamics.

Crucially, the flow in (20) is a *linear* differential equation and hence analytically integrable. Without loss of generality, we restrict ourselves to the case $\mathbf{F} = \mathbf{I}$ (the case for general $\mathbf{F}$ follows similarly). The dynamics associated with the flow $H_2(\mathbf{p})$ introduced in Lemma 4 become

$$\frac{\mathrm{d}}{\mathrm{dt}} \begin{bmatrix} \theta(t) \\ \mathbf{p}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{p}(t) \\ \mathbf{G}\mathbf{p}(t) \end{bmatrix}$$

with initial condition $(\theta_0, \mathbf{p}_0)$. Using the power series representation of the matrix exponential, the second differential equation for $\mathbf{p}$ may be integrated analytically to yield the following flow in $\mathbf{p}$-space:

$$\mathbf{p}(t) = \exp(\mathbf{G}t)\mathbf{p}_0$$

Substituting this result into the differential equation for $\theta$ yields

$$\frac{d\theta}{dt} = \exp(\mathbf{G}t)\mathbf{p}_0$$

If $\mathbf{G}$ is invertible then once again using the power series representation of the matrix exponential and rearranging yields the solution

$$\mathbf{\Phi}_{\epsilon, H_2(\mathbf{p})} \begin{bmatrix} \theta \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \theta + \mathbf{G}^{-1}(\exp(\mathbf{G}\epsilon) - \mathbf{I})\mathbf{p} \\ \exp(\mathbf{G}\epsilon)\mathbf{p} \end{bmatrix}$$

If $\mathbf{G}$ is not invertible, then slightly more care must be taken to first diagonalize $\mathbf{G}$ and separate its invertible/singular components. Since $\mathbf{G}$ is strictly antisymmetric it can be written as $i\mathbf{H}$ where $\mathbf{H}$ is a Hermitian matrix. Thus it can be diagonalized over $\mathbb{C}$ as:

$$\mathbf{G} = \begin{bmatrix} U_{\mathbf{\Lambda}} & U_0 \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} U_{\mathbf{\Lambda}}^{\top} \\ U_0^{\dagger} \end{bmatrix}$$

where $\mathbf{\Lambda}$ is a diagonal submatrix consisting of the nonzero eigenvalues of $\mathbf{G}$. $\begin{bmatrix} U_{\mathbf{\Lambda}} & U_0 \end{bmatrix}$ and $\begin{bmatrix} U_{\mathbf{\Lambda}}^{\top} \\ U_0^{\dagger} \end{bmatrix}$ are unitary matrices where the columns of $U_{\mathbf{\Lambda}}$ are the eigenvectors of $\mathbf{G}$ corresponding to its nonzero eigenvalues while the columns of $U_0$ are the eigenvectors of $\mathbf{G}$ corresponding to its zero eigenvalues. Even if $\mathbf{G}$ is not invertible we still have:

$$\mathbf{p}(t) = \exp(\mathbf{G}t)\mathbf{p}_0$$

However it is more convenient to represent the matrix exponential as:

$$\exp(\mathbf{G}t) = \begin{bmatrix} U_{\mathbf{\Lambda}} & U_0 \end{bmatrix} \begin{bmatrix} \exp(\mathbf{\Lambda}t) & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} U_{\mathbf{\Lambda}}^{\top} \\ U_0^{\dagger} \end{bmatrix}$$

Substituting this result into the differential equation for $\theta$, this representation of $\exp(\mathbf{G}t)$ implies the non-identity block can be handled as in the invertible case while the $\mathbf{I}$ block follows trivially to give:

$$\theta(t) = \theta_0 + \begin{bmatrix} U_{\mathbf{\Lambda}} & U_0 \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda}^{-1}(\exp(\mathbf{\Lambda}t) - \mathbf{I}) & \mathbf{0} \\ \mathbf{0} & t\mathbf{I} \end{bmatrix} \begin{bmatrix} U_{\mathbf{\Lambda}}^{\top} \\ U_0^{\dagger} \end{bmatrix} \mathbf{p}_0$$

Note that if $\mathbf{G} = \mathbf{0}$ then the flow map will simply reduce to an Euler translation as in ordinary HMC. We can also combine the ideas of Section 1.2 to obtain a preconditioned, magnetic HMC algorithm corresponding to a general $\mathbf{A}$-matrix of the form

$$\begin{pmatrix} \mathbf{0} & \mathbf{F} \\ -\mathbf{F}^{\top} & \mathbf{G} \end{pmatrix}$$

Dealing with a non-zero $\mathbf{E}$ becomes more subtle, since the corresponding sub-Hamiltonian is no longer exactly integrable under the splitting construction. In order to exactly integrate this sub-block a more costly implicit integrator is needed.

## 2.3  Integration Error of Magnetic HMC

Since we are using a symmetric, leapfrog splitting scheme for magnetic HMC that *exactly* integrates each sub-Hamiltonian we obtain identical error scaling to the leapfrog integrator applied to canonical HMC. Indeed, symplectic integrators are well-known to have many nice error properties in general and so perhaps this result is not so surprising [3].

**Lemma 6.** *The symplectic leapfrog-like integrator for magnetic HMC will have the same local ($\sim \mathcal{O}(\epsilon^3)$) and global ($\sim \mathcal{O}(\epsilon^2)$) error scaling (over $\tau \sim \frac{L}{\epsilon}$ steps), as the canonical leapfrog integrator of standard HMC if the Hamiltonian is separable.*

*Proof.* Note that for the parametrization of **A** corresponding to magnetic HMC the Hamiltonian vector field $\vec{\mathbf{H}} = \nabla_{\mathbf{p}} H \nabla_\theta + (-\nabla_\theta + \mathbf{G} \nabla_{\mathbf{p}} H) \nabla_{\mathbf{p}} \equiv \vec{\mathbf{A}} + \vec{\mathbf{B}}$ will generate the *exact* flow corresponding to exactly simulating the dynamics. We obtain an $\mathcal{O}(\epsilon^3)$ local error by simply exploiting the separability of the Hamiltonian. The leapfrog integration scheme splits the Hamiltonian as: $H(\theta, \mathbf{p}) = H_1(\theta) + H_2(\mathbf{p}) + H_1(\theta)$ and *exactly* integrates each sub-Hamiltonian so:

$$\mathbf{\Phi}_{\epsilon,H}^{\text{frog}} = \mathbf{\Phi}_{\epsilon,H_1(\theta)} \circ \mathbf{\Phi}_{\epsilon,H_2(\mathbf{p})} \circ \mathbf{\Phi}_{\epsilon,H_1(\theta)} = \exp\left(\frac{\epsilon}{2}\vec{\mathbf{B}}\right) \circ \exp\left(\epsilon\vec{\mathbf{A}}\right) \circ \exp\left(\frac{\epsilon}{2}\vec{\mathbf{B}}\right) \tag{21}$$

$$\tag{22}$$

Via repeated applications of the Baker-Campbell-Hausdorff formula [3] obtain:

$$\exp\left(\frac{\epsilon}{2}\vec{\mathbf{B}}\right) \circ \exp\left(\epsilon\vec{\mathbf{A}} + \frac{\epsilon}{2}\vec{\mathbf{B}} + \frac{\epsilon^2}{2}[\vec{\mathbf{A}}, \vec{\mathbf{B}}]\right) + \mathcal{O}(\epsilon^3) = \tag{23}$$

$$\exp\left(\frac{\epsilon}{2}\vec{\mathbf{B}} + \epsilon\vec{\mathbf{A}} + \frac{\epsilon}{2}\vec{\mathbf{B}} + \frac{\epsilon^2}{2}[\vec{\mathbf{A}}, \vec{\mathbf{B}}] + \frac{1}{2}[\frac{\epsilon}{2}\vec{\mathbf{B}}, \epsilon\vec{\mathbf{A}} + \frac{\epsilon}{2}\vec{\mathbf{B}} + \frac{\epsilon^2}{2}[\vec{\mathbf{A}}, \vec{\mathbf{B}}]]\right) + \mathcal{O}(\epsilon^3) = \tag{24}$$

$$\exp\left(\epsilon\vec{\mathbf{H}} + \frac{\epsilon^2}{4}[\vec{\mathbf{A}}, \vec{\mathbf{B}}] + \frac{\epsilon^2}{4}[\vec{\mathbf{B}}, \vec{\mathbf{A}}] + \frac{\epsilon^2}{8}[\vec{\mathbf{B}}, \vec{\mathbf{B}}]\right) + \mathcal{O}(\epsilon^3) = \exp\left(\epsilon\vec{\mathbf{H}}\right) + \mathcal{O}(\epsilon^3) \tag{25}$$

where we have used the antisymmetry of the commutator. The global error scaling, for an integration time of $\tau = \frac{L}{\epsilon}$ follows straightforwardly:

$$\mathbf{\Phi}_{\tau,H}^{\text{frog}} = \left(\exp\left(\frac{\epsilon}{2}\vec{\mathbf{B}}\right) \circ \exp\left(\epsilon\vec{\mathbf{A}}\right) \circ \exp\left(\frac{\epsilon}{2}\vec{\mathbf{B}}\right)\right)^L \tag{26}$$

$$= \left(\exp\left(\epsilon\vec{\mathbf{H}}\right) + \mathcal{O}(\epsilon^3)\right)^L \tag{27}$$

$$= \exp\left(\epsilon L \vec{\mathbf{H}}\right) + L\epsilon\mathcal{O}(\epsilon^2) \tag{28}$$

$$= \exp\left(\tau \vec{\mathbf{H}}\right) + \tau\mathcal{O}(\epsilon^2) \tag{29}$$

$$= \exp\left(\tau \vec{\mathbf{H}}\right) + \mathcal{O}(\epsilon^2) \tag{30}$$

as desired. □

# 3  Section 6 Experimental Details

Here we provide relevant experimental details for some of the Experiments presented in the main text.

## 3.1  Gaussians

In both experiments the reported autocorrelation measures are averaged over all coordinates as well as over 100 independent runs of the HMC/MHMC chains.

### 3.1.1  2D Gaussian

For the uncorrelated, ill-conditioned 2D Gaussian experiment presented in the main text the magnetic **G** component only has one non-zero parameter which was set to $g = .2$.

### 3.1.2 10D Gaussian

For the uncorrelated, ill-conditioned 10D Gaussian experiment presented in the main text, the **G** matrix was set to encourage the flow of momentum between the directions of large marginal variance with covariance eigenvalues $10^6$ and the remaining 8 directions of directions of small marginal variance with covariance eigenvalues of 1. We denote the directions of large marginal variance as $x_1$, $x_2$, and the other 8 directions of directions of small marginal variance as $x_i$. **G** was set such that $\mathbf{G}_{1i} = \mathbf{G}_{2i} = g$, $\mathbf{G}_{i1} = \mathbf{G}_{i2} = -g$ and $\mathbf{G}_{12} = \mathbf{G}_{21} = 0$ for $g = .2$.

## 3.2 Mixture of Gaussians

The superior mixing of MHMC relative to HMC in this example holds true for a wide range of $(\epsilon, L)$ settings as we can see by looking at the maximum mean discrepancy as a function of the number of samples in both Figures 1 and 2.
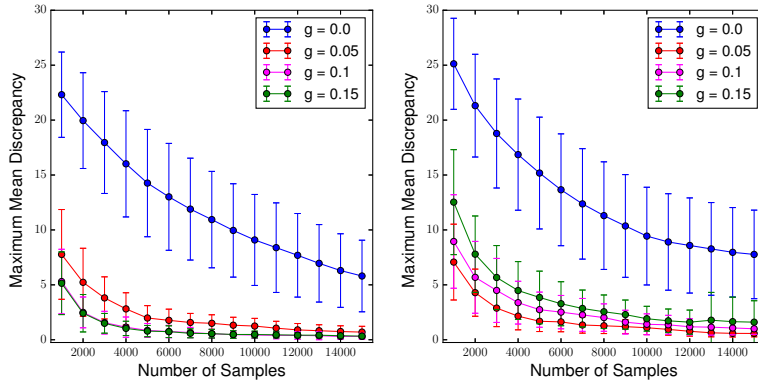


Figure 1: Left: MMD vs. Number of Samples for $(\epsilon = 1.5, L = 33)$. The acceptance rate was $\sim .74$ for all $g$. Right: MMD vs. Number of Samples for $(\epsilon = 1.9, L = 40)$. The acceptance rate was $\sim .43$ for all $g = 0$ and $\sim .34$ for all non-zero $g$. In both diagrams $g$ denotes the non-zero component of the magnetic field.
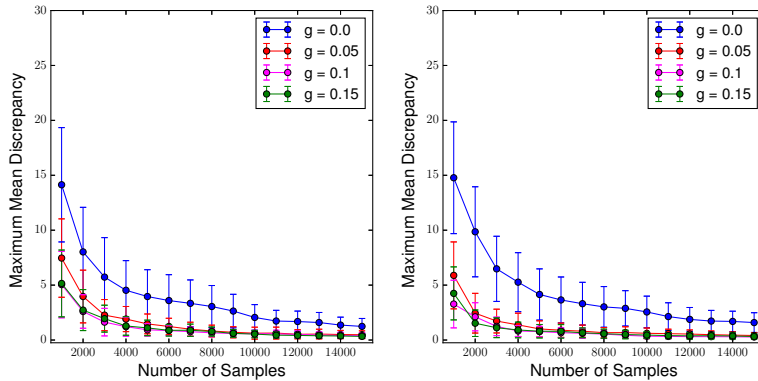


Figure 2: Left: MMD vs. Number of Samples for $(\epsilon = 1.0, L = 50)$. The acceptance rate was $\sim .87$ for all $g$. Right: MMD vs. Number of Samples for $(\epsilon = 0.5, L = 110)$. The acceptance rate was $\sim .95$ for all $g$. In both diagrams $g$ denotes the non-zero component of the magnetic field.

We found that tuning the parameters $(\epsilon, L)$ via Bayesian optimization often resulted in worse performance for ordinary HMC since the values found for $(\epsilon, L)$ were too conservative to encourage exploration between both modes. Moreover, more aggressive choices for $(\epsilon, L)$ for ordinary HMC led to a sharp drop in acceptance rate and significantly worse performance.

## 3.3 Gaussian Funnel

In this additional experiment, we consider the Gaussian funnel of [4] with density

$$p(\mathbf{x}, v) = \Pi_{i=1}^n \mathcal{N}(x_i|0, e^{-v})\mathcal{N}(v|0, 3^2)$$

in 10+1 dimensions (i.e. $n = 10$). This density illustrates the pathological correlation present in many hierarchical models between $\mathbf{x}$, a vector of low-level parameters, and $v$, a hyperparameter controlling their variability. As noted in [1, 5], Riemannian HMC methods, which incorporate local curvature information of the target, are well-suited to this problem as they help the dynamics traverse the energy surface which rapidly changes as a $v$ varies. HMC (as well as MHMC) do not exploit curvature information and will have more difficulty exploring the $v$ direction due to the rapid variation in density – see [1] for a detailed discussion. Despite this difficulty, we might intuitively expect that introducing a "curl" term into the entries of $\mathbf{G}$ which couple each $x_i$ and $v$ could increase exploration of the dynamics since these variables are nonlinearly correlated. In order to encourage the periodic flow of momentum between the marginal direction $v$ and the coordinates $x_i$ the $\mathbf{G}$ matrix was set such that $\mathbf{G}_{vi} = g$, $\mathbf{G}_{iv} = -g$, $\mathbf{G}_{ij} = 0$ with $g = .2$. To investigate this, we generated 10000 samples from both HMC and MHMC, discarding 1000 burn-in samples and computed the minimum effective sample size across $\mathbf{x}$ and $v$ and bias in the moments of the $v$ parameter similar to the set-up in [5] for various $\epsilon, L$ (see Table 1). We report results averaged over 100 different runs of the Markov chains.

Table 1: Comparison of HMC and MHMC targeting the Gaussian funnel for a variety of leapfrog steps

| algorithm | settings | time (s) | min ESS($\mathbf{x}$, $v$) | min ESS($\mathbf{x}$, $v$)/s | MSE($\mathbb{E}[v]$, $\mathbb{E}[v^2]$) |
|---|---|---|---|---|---|
| HMC | $\epsilon = 0.05, L = 100$ | 225 | 414, 85 | 1.84, 0.38 | .59, 1.57 |
| MHMC | $\epsilon = 0.05, L = 100$ | 270 | 463, 97 | 1.71, 0.36 | .29, 1.17 |
| HMC | $\epsilon = 0.05, L = 300$ | 705 | 1342, 118 | 1.90, 0.17 | .35, 1.15 |
| MHMC | $\epsilon = 0.05, L = 300$ | 837 | 1554, 122 | 1.86, 0.15 | .15, 1.05 |

We find that adding the magnetic field component decreases the bias in the moments and marginally increases the ESS, although both samplers struggle to explore the full target density, as the relatively low ESS figures indicate. Further details and experiments are provided in the Appendix.

Recall the density of the Gaussian funnel $p(\mathbf{x}, v) = \Pi_{i=1}^n \mathcal{N}(x_i|0, e^{-v})\mathcal{N}(v|0, 3^2)$. In order to encourage the periodic flow of momentum between the marginal direction $v$ and the coordinates $x_i$ the $\mathbf{G}$ matrix was set such that $\mathbf{G}_{vi} = g$, $\mathbf{G}_{iv} = -g$, $\mathbf{G}_{ij} = 0$ with $g = .2$. Moreover the reported results were averaged over 100 different runs of the Markov chains.

# 4   MHMC Proposals and Dynamics

In this section, we provide illustrations of the proposal distributions of MHMC in simple low-dimensional settings, to aid intuition and demonstrate the divergence of its behaviour from standard HMC.

## 4.1   Gaussian Densities

We first consider the case of an isotropic Gaussian target, and illustrate the proposal distribution of standard HMC, as well as MHMC with a variety of settings for the skew-symmetric matrix $\mathbf{A} = \begin{pmatrix} \mathbf{E} & \mathbf{F} \\ \mathbf{F} & \mathbf{G} \end{pmatrix}$ - see Figure 3. As in previous sections, we denote the off-diagonal element of the $\mathbf{G}$ matrix by $g$. We also provide proposal plots for an anisotropic Gaussian target distribution - see Figure 4.

## 4.2   Banana Density

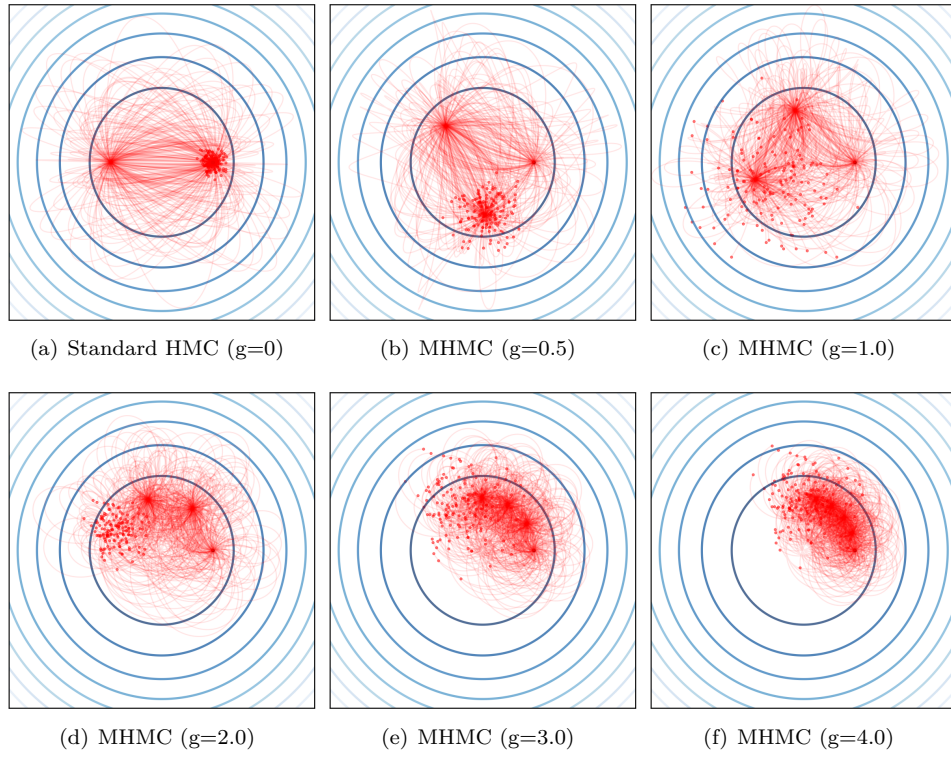We also provide proposal illustrations for the banana density of [2], as shown in Figure 5.

(a) Standard HMC (g=0)  (b) MHMC (g=0.5)  (c) MHMC (g=1.0)

(d) MHMC (g=2.0)  (e) MHMC (g=3.0)  (f) MHMC (g=4.0)

Figure 3: HMC and MHMC proposals for an isotropic Gaussian target.



(a) Standard HMC (g=0)  (b) MHMC (g=0.5)  (c) MHMC (g=1.0)

(d) MHMC (g=2.0)  (e) MHMC (g=3.0)  (f) MHMC (g=4.0)
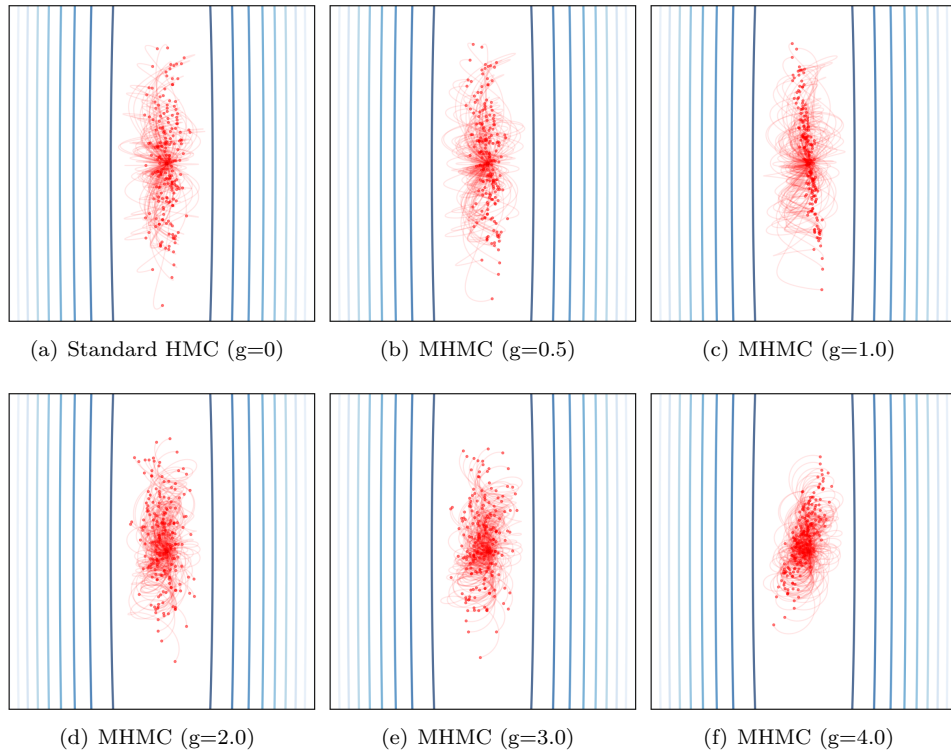
Figure 4: HMC and MHMC proposals for an anisotropic Gaussian target.

(a) Standard HMC (g=0)  (b) MHMC (g=0.1)  (c) MHMC (g=0.2)
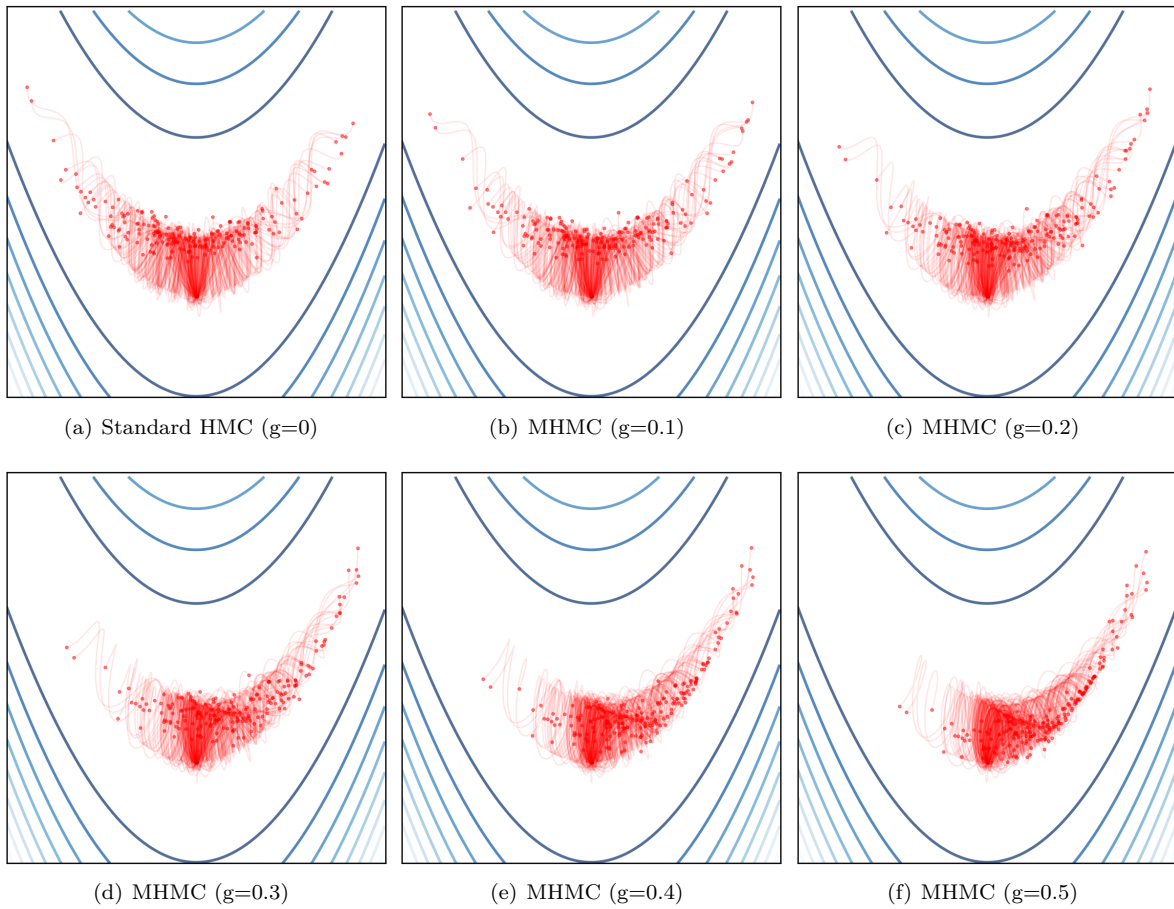
(d) MHMC (g=0.3)  (e) MHMC (g=0.4)  (f) MHMC (g=0.5)

Figure 5: HMC and MHMC proposals for the banana density target.

# References

[1] Michael Betancourt and Mark Girolami. *Hamiltonian Monte Carlo for Hierarchical Models*. CRC Press, Boca Raton, FL, 2015.

[2] Heikki Haario, Eero Saksman, and Johanna Tamminen. Adaptive proposal distribution for random walk metropolis algorithm. *Computational Statistics*, 14(3), 1999.

[3] Ernst Hairer, Marlis Hochbruck, Arieh Iserles, and Christian Lubich. Geometric Numerical Integration. *Oberwolfach Reports*, pages 805–882, 2006.

[4] Radford M. Neal. Slice Sampling: Rejoinder, 2003.

[5] Yichuan Zhang and Charles Sutton. Semi-separable Hamiltonian Monte Carlo for Inference in Bayesian Hierarchical Models. In *Advances in Neural Information Processing Systems*, pages 10–18, 2014.