
Metropolis Adjusted Microcanonical Hamiltonian Monte Carlo

Jakob Robnik
Physics Department,
University of California at Berkeley,
Berkeley, CA 94720, USA
jakob_robnik@berkeley.edu

Reuben Cohn-Gordon
Physics Department,
University of California at Berkeley,
Berkeley, CA 94720, USA
reubenharry@gmail.com

Uroš Seljak
Physics Department,
University of California at Berkeley
and Lawrence Berkeley National Laboratory, Berkeley,
Berkeley, CA 94720, USA
useljak@berkeley.edu

Abstract

Sampling from high dimensional distributions is a computational bottleneck in many scientific applications. Hamiltonian Monte Carlo (HMC), and in particular the No-U-Turn Sampler (NUTS), are widely used, yet they struggle on problems with a very large number of parameters or a complicated geometry. Microcanonical Langevin Monte Carlo (MCLMC) has been recently proposed as an alternative which shows striking gains in efficiency over NUTS, especially for high-dimensional problems. However, it produces biased samples, with a bias that is hard to control in general. We introduce the *Metropolis-Adjusted Microcanonical sampler* (MAMS), which relies on the same dynamics as MCLMC, but introduces a Metropolis-Hastings step and thus produces asymptotically unbiased samples. We develop an automated tuning scheme for the hyperparameters of the algorithm, making it applicable out of the box. We demonstrate that MAMS outperforms NUTS across the board on benchmark problems of varying complexity and dimensionality, achieving up to a factor of seven speedup.

1 Introduction

Drawing samples from a given probability density $p(\mathbf{x})$, for $\mathbf{x} \in \mathbb{R}^d$, has applications in a wide range of scientific disciplines, from Bayesian inference for statistics (Štrumbelj et al., 2024; Carpenter et al., 2017) to biology (Gelman and Rubin, 1996), statistical physics (Janke, 2008), quantum mechanics (Gattringer and Lang, 2010) and cosmology (Campagne et al., 2023). From the perspective of a practitioner in these fields, what is often desirable is a *black-box algorithm*, in the sense of taking as input an unnormalized density and returning samples from the corresponding distribution. An important special case is where the density is differentiable (either analytically or by automatic differentiation (Griewank and Walther, 2008)).

Markov Chain Monte Carlo (Metropolis et al., 1953; Hastings, 1970) is a broad class of methods suited to this task, which construct a Markov Chain $\{\mathbf{x}_i\}_{i=1}^n$ such that $\mathbf{x}_j$ are samples from the target distribution p . When the density’s gradient is available, Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 2011) is a leading method. In particular, the No-U-Turn sampler is a black-box version of HMC, where users do not need to manually select the hyperparameters. It has

been implemented in libraries like Stan (Carpenter et al., 2017) and Pyro (Bingham et al., 2019), serving a wide community of scientists.

NUTS is a powerful method, but there have been many attempts in the past decade to replace it with algorithms that can achieve the same accuracy at a lower computational cost. One such method is Microcanonical Langevin Monte Carlo (MCLMC) (Robnik et al., 2024); it replaces the HMC dynamics with a velocity-norm preserving dynamics, resulting in a method that is more stable to large gradients. Benchmarking in cosmology (Simon-Onfroy et al., 2025), Bayesian inference (Robnik et al., 2024; Sommer et al., 2024), and field theories (Robnik and Seljak, 2024) suggests MCLMC is a promising candidate to replace NUTS as a go-to gradient-based sampler.

The drawback of MCLMC is that it is biased; that is, the samples it produces do not correspond to samples from the true target distribution p , but rather to samples from a nearby distribution $\tilde{p}$. Although this bias is controllable, in many fields, asymptotically unbiased samplers are wanted, limiting the widespread utility of MCLMC. For HMC, this problem is resolved by the use of a Metropolis-Hastings (MH) step (Metropolis et al., 1953; Hastings, 1970), which accepts or rejects proposed moves $\mathbf{x} \rightarrow \mathbf{x}'$ of the Markov chain according to an acceptance $\min(1, e^{-W(\mathbf{x}', \mathbf{x})})$, where

$$e^{-W(\mathbf{x}', \mathbf{x})} = \frac{p(\mathbf{x}') q(\mathbf{x}|\mathbf{x}')}{p(\mathbf{x}) q(\mathbf{x}'|\mathbf{x})}. \quad (1)$$

But for MCLMC, the corresponding quantity W has not been previously derived. Moreover, an adaptation scheme for choosing the hyperparameters of the algorithm in the MH adjusted case is needed to make the algorithm usable out of the box, so that it can serve the same use cases as NUTS.

Contributions In this paper, we derive the acceptance probabilities for microcanonical dynamics with and without Langevin noise in Section 5 and Section 4, respectively. Notably, W turns out to be the energy error induced by discretization of the dynamics, as in HMC. We term the resulting sampler the *Metropolis-Adjusted Microcanonical Sampler* (MAMS), and develop an automatic adaptation scheme (Section 6) to make MAMS applicable without having to specify hyperparameters manually. We test MAMS on standard benchmarks in Section 7 and find that *it outperforms the state-of-the-art HMC with NUTS tuning by a factor of two at worst, and seven at best*. The algorithm is implemented in blackjax (Cabezas et al., 2024), applicable out-of-the-box, and is publicly available, together with documentation and tutorials¹. The code for reproducing numerical experiments is also available².

2 Related work

A wide variety of gradient-based samplers have been proposed, including Metropolis Adjusted Langevin trajectories (Riou-Durand and Vogrinc, 2022), generalized HMC (Horowitz, 1991a; Neal, 2020), the Metropolis Adjusted Langevin Algorithm (MALA; Grenander and Miller (1994)), Deterministic Langevin Monte Carlo (Grumitt et al., 2022), Nose-Hoover (Evans and Holian, 1985; Leimkuhler and Reich, 2009), Riemannian HMC (Girolami and Calderhead, 2011), Magnetic HMC (Tripuraneni et al., 2017) and the Barker proposal (Livingstone and Zanella, 2022). Some of these methods come with automatic tuning schemes that make them black-box, for example MALT (Riou-Durand and Vogrinc, 2022; Riou-Durand et al., 2023), generalized HMC (Hoffman et al., 2021) and HMC (Sountsov and Hoffman, 2022; Hoffman et al., 2021), but these schemes are designed for the many-short-chains MCMC regime (Sountsov et al., 2024; Margossian et al., 2024), which we do not consider here³. To our knowledge, NUTS remains the state-of-the-art black-box method for selecting the trajectory length in HMC-like algorithms for the single chain regime.

The dynamics described by Equation (3) have been independently proposed several times. In computational chemistry, they were derived by constraining Hamiltonian dynamics to have a fixed velocity norm (Tuckerman et al., 2001; Minary et al., 2003) and termed *isokinetic dynamics*. More recently, Steeg and Galstyan (2021) proposed them as a time-rescaling of Hamiltonian dynamics

¹<https://blackjax-devs.github.io/sampling-book/algorithms/mclmc.html>

²<https://github.com/reubenhenry/sampler-benchmarks>

³Many-short-chains approach is to run multiple short chains in parallel instead of a single long chain. This regime is interesting when parallel resources are available. However, it is often not applicable, either because one only has a single CPU, or because the parallel resources are needed elsewhere, for example, for parallelizing the model (Gattringer and Lang, 2010) or for performing multiple sampling tasks (Robnik et al., 2024).

with non-standard kinetic energy and no momentum resampling. Robnik et al. (2024) observed that while HMC aims to reach a stationary distribution known in statistical mechanics as the canonical distribution, it is also possible to target what is known as the *microcanonical* distribution, i.e. the delta function at some level set of the energy. The Hamiltonian H must then be chosen carefully to ensure that the position marginal of the microcanonical distribution is the desired target p , and one such choice is the Hamiltonian from Steeg and Galstyan (2021). They propose adding velocity resampling every n steps or Langevin noise every step as a method to obtain ergodicity.

In all of these instances, MCLMC has been proposed without Metropolis-Hastings, and as such, has been proposed as a biased sampler. This is the shortcoming that the present work resolves. We refer to our sampler as using *microcanonical* dynamics in reference to previous work on MCLMC, In contrast, we will refer to the dynamics in standard HMC as *canonical* dynamics.

3 Technical Preliminaries

Hamiltonian Monte Carlo Let $\mathcal{L}$ be the negative log likelihood of p up to a constant, i.e. $p(\mathbf{x}) = e^{-\mathcal{L}(\mathbf{x})}/Z$, where $Z = \int e^{-\mathcal{L}(\mathbf{x})} d\mathbf{x}$. If gradients $\nabla \mathcal{L}(\mathbf{x})$ are available and are sufficiently smooth, HMC is the gold standard proposal distribution. In HMC, each parameter x_i has an associated velocity u_i . Parameters and their velocities evolve by a set of differential equations

$$\dot{\mathbf{x}} = \mathbf{u} \quad \dot{\mathbf{u}} = \nabla \log p(\mathbf{x}), \quad (2)$$

which are designed to have $p_{HMC}(\mathbf{x}, \mathbf{u}) = p(\mathbf{x})\mathcal{N}(\mathbf{u})$ as their stationary distribution. Here $\mathcal{N}$ is the standard normal distribution. Note that the marginal distribution $\int p_{HMC}(\mathbf{x}, \mathbf{u}) d\mathbf{u}$ is equal to $p(\mathbf{x})$, the distribution we want to sample from. Thus, sampling from $p(\mathbf{x})$ reduces to solving Equation (2). In practice, the dynamics has to be simulated numerically, by iteratively solving for $\mathbf{x}$ at fixed $\mathbf{u}$ and vice versa, and updating the variables by a time step ϵ at each iteration. The discrepancy between this approximation and the true dynamics causes the stationary distribution to differ from the target distribution, but this can be corrected by MH; that is, we can use discretized Hamiltonian dynamics as a proposal q . Furthermore, to attain ergodicity, the velocities $\mathbf{u}$ must be resampled after every n steps.

The resulting algorithm has two hyperparameters: the discretization step size of the dynamics ϵ and the trajectory length between each resampling $L = n\epsilon$. Choosing good values for these two hyperparameters is crucial (Beskos et al., 2013; Neal, 2011; Betancourt, 2018), and so a practical sampler must also provide a robust *adaptation scheme* for choosing them.

Microcanonical dynamics An alternative to HMC is microcanonical dynamics (Robnik et al., 2024; Tuckerman et al., 2001; Minary et al., 2003; Steeg and Galstyan, 2021) defined by:

$$\dot{\mathbf{x}} = \mathbf{u} \quad \dot{\mathbf{u}} = (I - \mathbf{u}\mathbf{u}^T)\nabla \log p(\mathbf{x})/(d-1), \quad (3)$$

where $\mathbf{u}$ has unit norm which is preserved by the dynamics. The proposed benefit is that the normalization of the velocity makes the dynamics more stable to large gradients. When integrated exactly, these dynamics have $p_{MCLMC}(\mathbf{x}, \mathbf{u}) = p(\mathbf{x})\mathcal{U}_{S^{d-1}}(\mathbf{u})$ as a stationary distribution (see Appendix B.4), so that the marginal is still $p(\mathbf{x})$. Here $\mathcal{U}_{S^{d-1}}$ is the uniform distribution on the $d-1$ sphere. Robnik et al. (2024) propose using these dynamics without MH in order to *approximately* sample from $p(\mathbf{x})$. In this case, the velocity is partially resampled after every step, and the step size of the discretized dynamics is chosen small enough to limit deviation from the target distribution to acceptable levels.

While this algorithm works well in practice when the step size is properly tuned, the numerical integration error is not corrected, resulting in an asymptotic bias which is hard to control. In HMC this is solved by the MH step, which requires calculating W , as defined in Equation (1). In this case, W can be easily derived since the integrator is symplectic (volume preserving) and $q(\mathbf{x}|\mathbf{x}')/q(\mathbf{x}'|\mathbf{x}) = 1$. The integrator used for microcanonical dynamics is not symplectic, so it not immediately clear how to calculate W .

4 Metropolis adjustment for canonical and microcanonical dynamics

Both canonical and microcanonical dynamics can be numerically solved by separately solving the differential equation for the parameters $\mathbf{x}$, at fixed velocities $\mathbf{u}$ and vice versa. For a time interval ϵ ,

we refer to the position update as $A_\epsilon(\mathbf{x}, \mathbf{u})$ and the velocity update as $B_\epsilon(\mathbf{x}, \mathbf{u})$. The solution of the combined dynamics at time $t = n\epsilon$ is then constructed by a composition of these updates:

$$\varphi = \mathcal{T} \circ \underbrace{\Phi_{t/n} \circ \Phi_{t/n} \circ \dots \circ \Phi_{t/n}}_n, \quad (4)$$

$$\Phi_\epsilon = B_{\epsilon/2} \circ A_\epsilon \circ B_{\epsilon/2}. \quad (5)$$

This is known as the leapfrog (or velocity Verlet) scheme. A final time reversal map $\mathcal{T}(\mathbf{x}, \mathbf{u}) = (\mathbf{x}, -\mathbf{u})$, is inserted to ensure the map is an involution, i.e. $\varphi \circ \varphi = id$, where id is the identity map. This is useful in the Metropolis step but does not affect the dynamics in any way, because a full velocity refreshment is performed after the Metropolis step, erasing the effect of time reversal.

Both HMC and MCHMC possess a quantity, which we refer to as energy, which is conserved for exact dynamics, but only approximately conserved by the discrete update from Equation (5). The energy H is composed of two parts, a potential energy V and kinetic energy⁴ K . The position updates $(\mathbf{x}', \mathbf{u}) = A_\epsilon(\mathbf{x}, \mathbf{u})$ change the potential energy by

$$V(A_\epsilon(\mathbf{z})) - V(\mathbf{z}) = -\log \frac{p(\mathbf{x}')}{p(\mathbf{x})}, \quad (6)$$

while the velocity updates $(\mathbf{x}, \mathbf{u}') = B_\epsilon(\mathbf{x}, \mathbf{u})$ change the kinetic energy by

$$K(B_\epsilon(\mathbf{z})) - K(\mathbf{z}) = \frac{1}{2} \|\mathbf{u}'\|^2 - \frac{1}{2} \|\mathbf{u}\|^2 \quad (7)$$

for HMC and by

$$K(B_\epsilon(\mathbf{z})) - K(\mathbf{z}) = (d-1) \log \{ \cosh \delta + \mathbf{e} \cdot \mathbf{u} \sinh \delta \} \quad (8)$$

for MAMS. Here $\|\cdot\|$ is the Euclidean norm, $\mathbf{e} = -\nabla \mathcal{L}(\mathbf{x}) / \|\nabla \mathcal{L}(\mathbf{x})\|$ and $\delta = \epsilon \|\nabla \mathcal{L}(\mathbf{x})\| / (d-1)$.

To derive the MH ratio, the key is to realize that the A and B updates are deterministic

$$q(\mathbf{z}'|\mathbf{z}) = \delta(\varphi(\mathbf{z}) - \mathbf{z}'), \quad (9)$$

where the transition map is generated by a dynamical system (Fang et al., 2014) for $\mathbf{z} = (\mathbf{x}, \mathbf{u})$,

$$\dot{\mathbf{z}}(t) = F(\mathbf{z}(t)). \quad (10)$$

Here, $\mathbf{z}(0) = \mathbf{z}$ and $\mathbf{z}(T) = \varphi(\mathbf{z}) = \mathbf{z}'$. The drift vector field F in canonical and microcanonical dynamics can be read from Equations (2) and (3), respectively. For the former, it equals $F_A(\mathbf{x}, \mathbf{u}) = (\mathbf{u}, 0)$ during the A updates, and $F_B(\mathbf{x}, \mathbf{u}) = (0, -\nabla \mathcal{L}(\mathbf{x}))$ during the B updates. For the latter, it equals $F_A(\mathbf{x}, \mathbf{u}) = (\mathbf{u}, 0)$ during the A updates, and $F_B(\mathbf{x}, \mathbf{u}) = (0, -(1 - \mathbf{u}\mathbf{u}^T)\nabla \mathcal{L}(\mathbf{x})/(d-1))$ during the B updates. These fields can be used to explicitly solve for the A and B updates; the solutions are given in Appendix B.3.

Lemma 4.1. *For proposals which are deterministic involutions generated by a dynamical system of the form (10), W in the MH acceptance probability (1) equals*

$$W(\mathbf{z}', \mathbf{z}) = -\log \frac{p(\mathbf{z}')}{p(\mathbf{z})} - \int_0^T \nabla \cdot F(\mathbf{z}(s)) ds.$$

Proof. The first term comes from the first factor in Equation (1). For the second term, observe that the ratio of transition probabilities is

$$\frac{q(\mathbf{z}|\mathbf{z}')}{q(\mathbf{z}'|\mathbf{z})} = \frac{\delta(\varphi(\mathbf{z}') - \mathbf{z})}{\delta(\varphi(\mathbf{z}) - \mathbf{z}')} = \frac{\delta(\varphi(\mathbf{z}') - \mathbf{z})}{\delta(\mathbf{z} - \varphi(\mathbf{z}'))} \left| \frac{\partial \varphi}{\partial \mathbf{z}}(\mathbf{z}) \right| = \left| \frac{\partial \varphi}{\partial \mathbf{z}}(\mathbf{z}) \right|,$$

where in the second step we have used reversibility, as well as standard properties of the delta function⁵. This last expression is the Jacobian determinant of the transition map φ . Finally, the second term of W in Lemma 4.1 follows from Eq. (11) by the Abel–Jacobi–Liouville identity. $\square$

⁴Note that the MAMS dynamics of Equation (3) are not Hamiltonian, so for MAMS, K is not kinetic energy in the standard sense. In Appendix B.5, a relationship between MAMS dynamics and a Hamiltonian dynamics for which K actually is kinetic energy is given, justifying the name.

⁵Recall that $\delta(x-a)f(x) = \delta(x-a)f(a)$, and $\delta(f(x)) = \sum_i \delta(x-a_i) \left| \frac{df}{dx} a_i \right|^{-1}$, where a_i are the roots of f . In our case, $f(z) = \phi(z) - z'$, so that $\delta(\phi(z) - z) = \delta(z - \phi(z')) \left| \frac{\partial \phi}{\partial z}(\phi(z')) \right|^{-1} = \delta(z - \phi(z')) \left| \frac{\partial \phi}{\partial z}(z) \right|^{-1}$

Lemma 4.2. For a proposal of the form $\varphi = \mathcal{T} \circ B_{\epsilon_N} \circ A_{\eta_N} \circ \dots \circ B_{\epsilon_1} \circ A_{\eta_1}$, where $\epsilon_k \in \mathbb{R}$ and $\eta_k \in \mathbb{R}$ for all k , W in both HMC and MAMS equals the total energy change of the proposal, that is, the sum of all the energy changes:

$$W(\varphi(z), z) = \sum_{k=1}^N V(A_{\epsilon_k}(z_{k-1})) - V(z_{k-1}) + K(B_{\epsilon_k} \circ A_{\eta_k}(z_{k-1})) - K(A_{\eta_k}(z_{k-1}))$$

where $z_k = B_{\epsilon_k} \circ A_{\eta_k} \circ \dots \circ B_{\epsilon_1} \circ A_{\eta_1}(z)$ and energy changes are taken from Equations (6) and (7) for HMC, and (6) and (8) for MAMS.

Proof. The position update A in both HMC and MAMS has a vanishing divergence: $\nabla \cdot F_A = \frac{\partial u_i}{\partial x_i} = 0$, so during the position update, only the first term in Lemma 4.1 survives and $W(z', z) = -\log p(x', u')/p(x, u) = -\log p(x')/p(x) = \Delta V$ from Equation (6).

The velocity update in HMC has vanishing divergence $\nabla \cdot F_B = \frac{-\partial \nabla_i \mathcal{L}(x)}{\partial u_i} = 0$, so during the HMC velocity update, only the first term of W in Lemma 4.1 survives and $W(z', z) = -\log p(x', u')/p(x, u) = -\log p(u')/p(u) = \Delta K$ from Equation (8).

The velocity update in MAMS, on the other hand, has a non-zero divergence:

$$\begin{aligned} \nabla \cdot F_B &= -\|\nabla \mathcal{L}(x)\| u(t) \cdot e = \|\nabla \mathcal{L}(x)\| \frac{\sinh \delta + \cosh \delta (e \cdot u)}{\cosh \delta + \sinh \delta (e \cdot u)} \\ &= -(d-1) \frac{d}{dt} \log \{ \cosh \delta + \sinh \delta (e \cdot u) \}. \end{aligned}$$

In the first equality, we have used the divergence from Robnik and Seljak (2024), which we also re-derive in Appendix B.7. In the second equality, we used the explicit form of the velocity update from Appendix B.3, namely Equation (28). So we find that the velocity update for MAMS has

$$W(\varphi_{F_B}(z), z) = -\int_0^T \nabla \cdot F_B(z(s)) ds = (d-1) \log \{ \cosh \delta + e \cdot u \sinh \delta \} = \Delta K$$

from Equation (7). A more direct derivation of the above is provided in Appendix B.6 for the interested reader.

W of a composition of maps of the form in Lemma 4.1 is a sum of the individual W , i.e., $W(\varphi(\chi(z)), z) = W(\varphi(\chi(z)), \chi(z)) + W(\chi(z), z)$. This yields the formula in the statement of the theorem. $\square$

This result shows how to perform Metropolis adjustment in MAMS: analogously to HMC, for every B or A update that the algorithm performs in the proposal, the energy difference must be calculated and added to the cumulative energy difference W (B update energy difference is calculated by Equation (7), A update by Equation (6)). Final MH acceptance probability is $\min(1, e^{-W})$. This is a favorable result, because both the HMC and MAMS numerical integrators keep the energy error small, even over long trajectories (Leimkuhler and Matthews, 2015), implying that a high acceptance rate can be maintained. As an empirical illustration that the MH acceptance probability from Lemma 4.2 is correct, Section 4 shows a histogram of 2 million samples from a 100-dimensional Gaussian (1st dimension shown) using the MH-adjusted kernel (orange), given a step size of 20. The kernel without MH adjustment is also shown (blue) and exhibits asymptotic bias.

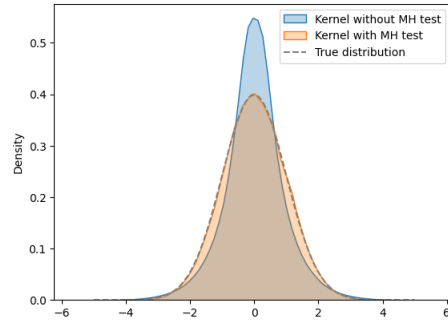


Figure 1: Histogram of MAMS samples with (orange) and without MH adjustment (blue) from 100-dim standard normal (1st dim shown). Step size is chosen very large ($\epsilon = 20$) to highlight the bias the MH step removes.

5 Sensitivity to hyperparameters

The performance of HMC is known to be very sensitive to the choice of the trajectory length, and the problem becomes even more pronounced for ill-conditioned targets, where different directions may require different trajectory lengths for optimal performance (Neal, 2011). This is further illustrated in Appendix D. Two solutions to this problem are randomizing the trajectory length (Bou-Rabee and Sanz-Serna, 2017) and replacing the full velocity refreshment with partial refreshment after every step, also known as the underdamped Langevin Monte Carlo (Horowitz, 1991b). Here, we will pursue both approaches with respect to MAMS.

Random integration length We randomize the integration length by taking $n_k = \lceil 2h_k L/\epsilon \rceil$ integration steps to construct the k -th MH proposal. Here h_k can either be random draws from the uniform distribution $\mathcal{U}(0, 1)$ or the k -th element of Halton’s sequence, as recommended in Owen (2017); Hoffman et al. (2021). Other distributions of the trajectory length were also explored in the literature (Sountsov and Hoffman, 2022) but with no gain in performance. The factor of two is inserted to make sure that we do L/ϵ steps on average⁶.

Partial refreshment Partially refreshing the velocity after every step also has the effect of randomizing the time before the velocity coherence is lost, and therefore has similar benefits to randomizing the integration length (Jiang, 2023). However, while the flipping of velocity, needed for the deterministic part of the update to be an involution, is made redundant by a full resampling of velocity, this is not the case for partial refreshment. This results in rejected trajectories backtracking some of the progress that was made in the previously accepted proposals (Riou-Durand and Vogrinc, 2022). Skipping the velocity flip is possible, but it results in a small bias in the stationary distribution (Akhmatskaya et al., 2009), and it is not clear that it has any advantages over full refreshment. Two popular solutions for LMC are to either use a non-reversible MH acceptance probability as in Neal (2020); Hoffman and Sountsov (2022) or to add a full velocity refreshment before the MH step as in MALT (Riou-Durand and Vogrinc, 2022). We will prove that both can be straightforwardly used with microcanonical dynamics, and then concentrate on the MALT strategy in the remainder of the paper.

We will generate the Langevin dynamics by the “OBABO” scheme (Leimkuhler and Matthews, 2015), where BAB is the deterministic φ_ϵ map from Equation (5) and O is the partial velocity refreshment. In LMC, $O_\epsilon(\mathbf{u}) = c_1 \mathbf{u} + c_2 \mathbf{Z}$, where $\mathbf{Z}$ is the standard normal distributed variable, $c_1 = e^{-\epsilon/L_{\text{partial}}}$ and $c_2 = \sqrt{1 - c_1^2}$. L_{partial} is a parameter that controls the partial refreshment’s strength and is comparable with HMC’s trajectory length L . With microcanonical dynamics, a similar expression that additionally normalizes the velocity has been proposed (Robnik and Seljak, 2024):

$$O_\epsilon(\mathbf{u}) = \frac{c_1 \mathbf{u} + c_2 \mathbf{Z} / \sqrt{d}}{\|c_1 \mathbf{u} + c_2 \mathbf{Z} / \sqrt{d}\|}. \quad (11)$$

Denote by $\Delta(\mathbf{z}', \mathbf{z})$ the energy error accumulated in the *deterministic* (φ_ϵ) part of the update. Note that for microcanonical Langevin dynamics, only the deterministic part of the update changes the energy, while in canonical Langevin dynamics, the O update also does but is not included in Δ .

Theorem 5.1. *The Metropolis-Hastings acceptance probability of the MAMS proposal $q(\mathbf{z}'|\mathbf{z})$, corresponding to $\mathcal{TOBABO}$ is $\min(1, e^{-\Delta(\mathbf{z}', \mathbf{z})})$.*

The generalized HMC strategy (Hoffman and Sountsov, 2022) only uses the one-step proposal, so Lemma 4.2 shows that it can be generalized to the microcanonical update, simply by using the microcanonical energy instead of the canonical energy. The MALT proposal, on the other hand, consists of n LMC (or in our case, microcanonical LMC) steps and a full refreshment of the velocity, as shown in Algorithm 1.

Theorem 5.2. *$\{\mathbf{x}_i\}_{i>0}$ defined in Alg 1 is a Markov chain whose stationary distribution is $p(\mathbf{z})$.*

Proofs of both theorems are in Appendix A.

⁶More precisely, $\int_0^1 \lceil 2uL/\epsilon \rceil du \neq L/\epsilon$, because of the ceiling function. In the implementation, we use the correct expression, which is $n_k = \lceil 2yh_k L/\epsilon \rceil$, where $y = \frac{Y(Y+1)}{Y+1-L/\epsilon}$ and $Y = \lfloor 2L/\epsilon - 1 \rfloor$ is the integer part of y . This follows from solving $L/\epsilon = \mathbb{E}[n_k] = \frac{1+2+\dots+Y+(y-Y)(Y+1)}{y} = \frac{(Y+1)(y-Y/2)}{y}$ for y .

Input:

negative log-density function $\mathcal{L} : \mathbb{R}^d \rightarrow \mathbb{R}$, initial condition $\mathbf{x}^0 \in \mathbb{R}^d$, number of samples $N > 0$, step size $\epsilon > 0$, steps per sample $L/\epsilon \in \mathbb{N}$, partial refreshment parameter L_{partial} . The last three parameters can be determined automatically as in Section 6.

Returns: samples $\{\mathbf{x}^n\}_{n=1}^N$ from $p(\mathbf{x}) \propto e^{-\mathcal{L}(\mathbf{x})}$.

for $I \leftarrow 0$ **to** N **do**

$\mathbf{u} \sim \mathcal{U}_{S^{d-1}}$

$\mathbf{z}^0 \leftarrow (\mathbf{x}^I, \mathbf{u})$

$\delta \leftarrow 0$

for $i \leftarrow 0$ **to** n **do**

$\mathbf{z} \leftarrow O_\epsilon(\mathbf{z}_i)$

$\mathbf{z}' \leftarrow \Phi_\epsilon(\mathbf{z})$

$\mathbf{z}^{i+1} \leftarrow O_\epsilon(\mathbf{z}')$

$\delta \leftarrow \delta + \Delta(\mathbf{z}', \mathbf{z})$

end for

 draw a random uniform variable $U \sim \mathcal{U}(0, 1)$

if $U < e^{-\delta}$ **then**

$\mathbf{x}^{I+1} \leftarrow \mathbf{z}^{n-1}[0]$

else

$\mathbf{x}^{I+1} \leftarrow \mathbf{z}^0[0]$

end if

end for

Algorithm 1: MAMS - Langevin

6 Automatic hyperparameter tuning

MAMS has two hyperparameters, stepsize ϵ and the trajectory length L , where L/ϵ is the (average) number of steps in a proposal's trajectory. The Langevin version of the algorithm has an additional hyperparameter L_{partial} that determines the partial refreshment strength during the proposal trajectories, i.e., the amount of Langevin noise. In addition, it is common to use a preconditioning matrix M to linearly transform the configuration space, in order to reduce the condition number of the covariance matrix. The algorithm's performance crucially depends on these hyperparameters, so we here develop an automatic tuning scheme. First, the stepsize is tuned, then the preconditioning matrix, and finally, the trajectory length. L_{partial} is directly set by the trajectory length, see Appendix D. We adopt a schedule similar to the one in (Robnik et al., 2024), where each of these three stages takes 10% of the total sampling time, so that tuning does not significantly increase the total sampling cost.

Stepsize We extend the argument from (Beskos et al., 2013; Neal, 2011) to MAMS in Appendix C, showing that the optimal acceptance rate in MAMS is the same as in HMC, namely 65%. For HMC, larger acceptance rates have been observed to perform better in practice (Phan et al., 2019a), with some theoretical justification (Betancourt et al., 2015). We set the acceptance rate to 90%. In the first stage of tuning, we use a stochastic optimization scheme, dual averaging (Nesterov, 2009) from (Hoffman and Gelman, 2014), to adapt the step size until a desired acceptance rate is achieved.

Preconditioning matrix In the second stage, we determine the preconditioning matrix. A simple choice of diagonal preconditioning matrix is obtained by estimating variance along each parameter.

Trajectory length Microcanonical and canonical dynamics are extremely efficient in exploring the configuration space, while staying on the typical set. Therefore, we do not wish to reduce them to a comparatively inefficient diffusion process by adding too much noise, i.e., having too low L . On the other hand, we want to prevent the dynamics from being caught in cycles or quasi-cycles to maintain efficient exploration. Heuristically, we should send the dynamics in a new direction at the time scale that the dynamics needs to move to a different part of the configuration space, producing a new effective sample (Robnik et al., 2024). This suggests two approaches for tuning L (Robnik et al., 2024).

The simpler is to estimate the size of the typical set by computing the average of the eigenvalues of the covariance matrix, which is equal to the mean of the variances in each dimension (Robnik et al.,

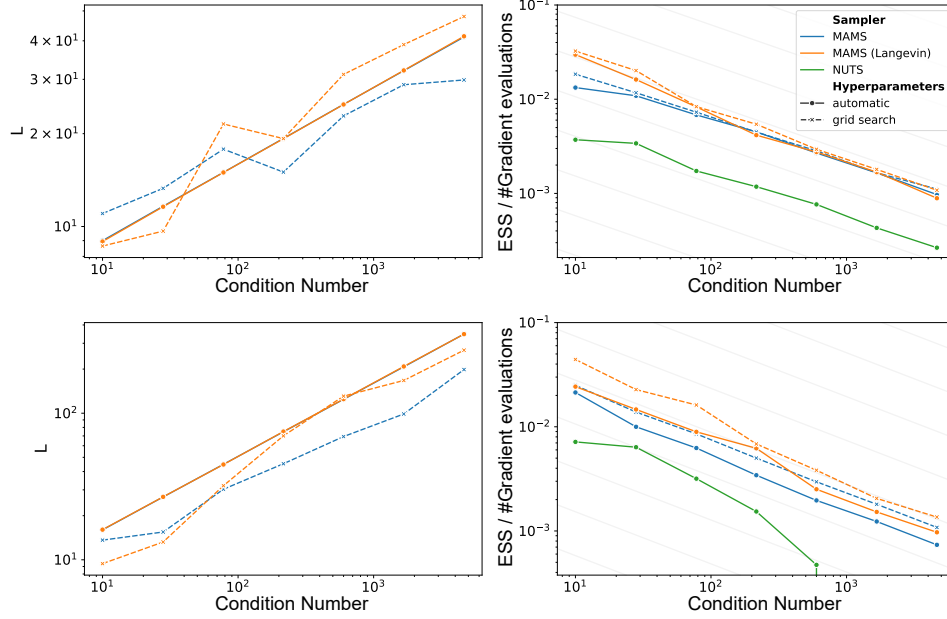


Figure 2: Tuning performance on Gaussians as a function of condition number. Gaussians are 100d with eigenvalues log-uniform distributed (top row) and outlier distributed (bottom). Left panels: the value of L from the automatic tuning algorithm is shown with a solid line, and the optimal L obtained by a grid search (dashed). As can be seen, automatic tuning achieves close to optimal values. Right panels: ESS (for the worst parameter) is shown as a function of condition number. MAMS tuning scheme (solid lines) achieves ESS, which is very close to the grid search results (dashed lines). MAMS scaling is similar to NUTS and goes as $ESS \propto \text{condition number}^{-1/2}$, which is shown as grey lines in the background. However, MAMS has around four times better proportionality constant.

2024). With a linearly preconditioned target, these variances are 1, and the estimate for the optimal L is $L = \sqrt{d}$. We will use this as an initial value. A more refined approach is to set L to be on the same scale as the time passed between effective samples:

$$L_{ALBA} \propto \mathbb{E}[\text{time between effective samples}] = \mathbb{E}[\text{time between samples}] \tau_{\text{int}} = L \tau_{\text{int}}, \quad (12)$$

The proportionality constant is of order one and will be determined numerically, based on Gaussian targets. Integrated autocorrelation time τ_{int} is the ratio between the total number of (correlated) samples in the chain and the number of effectively uncorrelated samples. It depends on the observable $f(\mathbf{x})$ that we are interested in and can be calculated as

$$\tau_{\text{int}}[f] = 1 + 2 \sum_{t=1}^{\infty} \rho_t[f], \quad (13)$$

where

$$\rho_t[f] = \frac{\mathbb{E}[(f(\mathbf{x}(s)) - \mathbb{E}[f])(f(\mathbf{x}(s+t)) - \mathbb{E}[f]))]}{\text{Var}[f]} \quad (14)$$

is the chain autocorrelation function in stationarity. We take $f(x_i) = x_i$ and harmonically average $\tau_{\text{int}}[x_i]$ over i . We determine the proportionality constant of Equation (12) in a way that L equals the optimal L , determined by a grid search, for the standard Gaussian. We find a proportionality constant of 0.3 for MAMS without Langevin noise and 0.23 with Langevin noise.

7 Experiments

We aim to compare MAMS with NUTS, which is the main competitive *black-box* sampler in the single chain setting, see Section 2. We also compare against MALA (i.e., HMC with one step per trajectory), which is also popular in some settings.

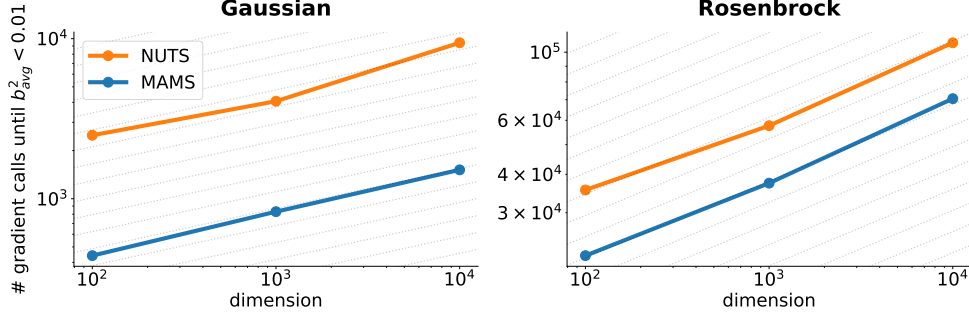


Figure 3: Sampling performance as a function of the dimensionality of the problem. On the left, the problem is a standard Gaussian, on the right, multiple independent copies of the Rosenbrock function (a banana-shaped target). Number of gradient calls to convergence scales as $d^{1/4}$ (grey lines in the background) for both MAMS and NUTS, but MAMS has a better proportionality constant.

Evaluation metric We follow Hoffman and Sountsov (2022) and define the squared error of the expectation value $\mathbb{E}[f(\mathbf{x})]$ as

$$b^2(f) = \frac{(\mathbb{E}_{\text{sampler}}[f] - \mathbb{E}[f])^2}{\text{Var}[f]}, \quad (15)$$

and consider the largest second-moment error across parameters, $b_{\max}^2 \equiv \max_{1 \leq i \leq d} b^2(x_i^2)$, because in our problems of interest, there is typically a parameter of particular interest that has a significantly higher error than the other parameters (for example, a hierarchical parameter in many Bayesian models). For distributions which are a product of independent low-dimensional distributions (standard Gaussian, Rosenbrock function, and Cauchy problems), we take an average instead of the maximum because all parameters should have the same error. For the Cauchy distribution, the second moment $\mathbb{E}[x^2]$ diverges, so we instead consider the expected value of $-\log p(x)$, i.e., the entropy of the distribution. b^2 can be interpreted as an accuracy equivalent of 100 effective samples (Hoffman and Sountsov, 2022).

In typical applications, computing the gradients $\nabla \log p(\mathbf{x})$ dominates the total sampling cost, so we take the number of gradient evaluations as a proxy of wall-clock time. For very simple models, gradient calls might not dominate the cost, and the exact implementation of the numerical integration becomes important. Our implementation is efficient; for example, 1000 samples with $L = 2$ and stepsize = 1 for stochastic volatility take 1.5 seconds with MAMS on a single CPU and 4.8 seconds with NUTS. We do not report these numbers since they are irrelevant for the more expensive models that the method is meant to be applied to in practice. As in (Hoffman and Sountsov, 2022), we measure a sampler’s performance as the number of gradient calls n needed to achieve low error, $b_{\max}^2 < 0.01$. We note that it is common to report effective sample size (ESS) per gradient evaluation instead, but both carry similar information, since n can be interpreted as $100/(\text{ESS}/\# \text{ gradient evaluations})$. Furthermore, we argue that the former is of primary interest and the latter is only used as a proxy for the former.

Scaling with the condition number and the dimensionality Figure 2 compares MAMS with NUTS on 100-dimensional Gaussians with varying condition number. Two distributions of covariance matrix eigenvalues are tested: uniform in log and outlier distributed. *Outlier distributed* means that two eigenvalues are κ while the other eigenvalues are 1. For both samplers, the number of gradients to low error scales with the condition number κ as $\kappa^{-\frac{1}{2}}$, but MAMS is faster by a factor of around 4. Figure 3 compares MAMS and NUTS scaling with the problem’s dimensionality. Both have the known $d^{\frac{1}{4}}$ scaling law (Neal, 2011), albeit MAMS has a better proportionality constant.

Benchmarks Table 1 compares MAMS with NUTS and MALA on a set of benchmark problems, mostly adapted from the Inference Gym (Sountsov et al., 2020). Problems vary in dimensionality (36–2429), are both synthetic and with real data, and include a distribution with a very long tail (Cauchy), a bimodal distribution (Bimodal), and many Bayesian inference problems. Problem details

	NUTS	MALA	MAMS	MAMS (Langevin)	MAMS (Grid Search)
Gaussian	19,652	11,010	3,249	3,172	3,121
Banana	95,519	140,524	14,078	14,818	15,288
Bimodal	210,758	$> 10^6$	139,418	136,770	123,295
Rosenbrock	161,359	$> 10^6$	94,184	103,545	93,782
Cauchy	171,373	824,429	110,404	155,963	87,900
Brownian	29,816	597,119	13,528	15,232	14,015
German Credit	88,975	$> 10^6$	55,748	49,979	52,265
ItemResp	76,043	249,470	45,371	56,902	45,640
StochVol	843,768	$> 10^6$	430,088	510,190	431,957
Funnel	$> 10^8$	$> 10^7$	2,346,899	1,765,311	1,013,048

Table 1: Number of gradients calls needed to get the squared error on the worst second moment below 0.01. Lower is better; number of gradients is roughly proportional to wall clock time.

are in Appendix E.1. We do not include Bayesian neural networks because they lack an established ground truth (Izmailov et al., 2021), would require the use of stochastic gradients, which MAMS is not directly amenable to and because there is evidence that suggests that higher accuracy samples are not necessary for the good performance (Wenzel et al., 2020), so unadjusted methods achieve superior performance (Sommer et al., 2024).

For all algorithms, we use an initial run to find a diagonal preconditioning matrix. For NUTS and MALA, the only remaining parameter to tune is step size, which is tuned by dual averaging, targeting 80% acceptance rate for NUTS and 57.4% for MALA (Neal, 2011). For MAMS, we further tune L using the scheme of Section 6. We take the adaptation steps as our burn-in, initializing the chain with the final state returned by the adaptation procedure. NUTS is run using the BlackJax (Cabezas et al., 2024) implementation, with the provided window adaptation scheme. Table 1 shows the number of gradient calls in the chain (excluding tuning) used to reach squared error of 0.01. To reduce variance in these results, we run at least 128 chains for each problem and take the median of the error across chains at each step.

Results In all cases, MAMS outperforms NUTS, typically by a factor of two at worst and seven at best. MALA drastically underperforms relative to MAMS and NUTS. Using Langevin noise instead of the trajectory length randomization has little effect in most cases, analogous to the situation for HMC (Jiang, 2023; Riou-Durand et al., 2023). For Neal’s Funnel, we need an acceptance rate of 0.99 for MAMS to converge. We were unable to obtain convergence for NUTS. This problem is a known NUTS failure mode, so it is of note that MAMS converges.

To assess how successful the tuning scheme from Section 6 is at finding the optimal value of L , we perform a grid search over L , by first performing a long NUTS run to obtain a diagonal of the covariance matrix and an initial L , and then for each new candidate value of L , tuning step size by dual averaging with a target acceptance rate of 0.9. In Table 1 we show the number of gradients to low error using this optimal L . Performance is very close to optimal on all benchmark problems.

8 Conclusions

Our core contribution is MAMS, an out-of-the-box gradient-based sampler applicable in the same settings as NUTS HMC and intended as a successor to it. Our experiments found substantial performance gains for MAMS over NUTS in terms of statistical efficiency. These experiments were on problems varying in dimension (up to 10^4) and included real datasets, multimodality, and long tails. This said, to reach the maturity of NUTS, the method needs to be battle-tested over many years on an even broader variety of problems.

We note that MAMS is simple to implement, with little code change from standard HMC. A promising future direction is the many-short-chains MCMC regime (Hoffman and Sountsov, 2022; Margossian et al., 2024), since MAMS is not as control-flow heavy as NUTS and since MAMS with Langevin noise can have a fixed number of steps per trajectory, making parallelization more efficient (Sountsov et al., 2024).

Acknowledgments and Disclosure of Funding

This material is based upon work supported in part by the Heising-Simons Foundation grant 2021-3282 and by NSF CSSI grant award number 2311559.

References

- Akhmatskaya, E., N. Bou-Rabee, and S. Reich (2009, April). A comparison of generalized hybrid Monte Carlo methods with and without momentum flip. *Journal of Computational Physics* 228(6), 2256–2265.
- Beskos, A., N. Pillai, G. Roberts, J.-M. Sanz-Serna, and A. Stuart (2013, November). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli* 19(5A), 1501–1534. Publisher: Bernoulli Society for Mathematical Statistics and Probability.
- Betancourt, M. (2018, July). A Conceptual Introduction to Hamiltonian Monte Carlo. arXiv:1701.02434 [stat].
- Betancourt, M. J., S. Byrne, and M. Girolami (2015, February). Optimizing The Integrator Step Size for Hamiltonian Monte Carlo. arXiv:1411.6669 [stat].
- Bingham, E., J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, and N. D. Goodman (2019). Pyro: Deep universal probabilistic programming. *Journal of machine learning research* 20(28), 1–6.
- Bou-Rabee, N. and J. M. Sanz-Serna (2017, August). Randomized Hamiltonian Monte Carlo. *The Annals of Applied Probability* 27(4). arXiv:1511.09382 [math].
- Cabezas, A., A. Corenflos, J. Lao, and R. Louf (2024). Blackjax: Composable bayesian inference in jax.
- Campagne, J.-E., F. Lanusse, J. Zuntz, A. Boucaud, S. Casas, M. Karamanis, D. Kirkby, D. Lanzieri, Y. Li, and A. Peel (2023, April). JAX-COSMO: An End-to-End Differentiable and GPU Accelerated Cosmology Library. *The Open Journal of Astrophysics* 6, 10.21105/astro.2302.05163. arXiv:2302.05163 [astro-ph].
- Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017, January). Stan: A Probabilistic Programming Language. *Journal of Statistical Software* 76, 1–32.
- Creutz, M. (1988, August). Global Monte Carlo algorithms for many-fermion systems. *Physical Review D* 38(4), 1228–1238. Publisher: American Physical Society.
- Crooks, G. E. (1999, September). Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E* 60(3), 2721–2726. Publisher: American Physical Society.
- Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987, September). Hybrid Monte Carlo. *Physics Letters B* 195(2), 216–222.
- Evans, D. J. and B. L. Holian (1985, October). The Nose-Hoover thermostat. *Journal of Chemical Physics* 83, 4069–4074. Publisher: AIP ADS Bibcode: 1985JChPh..83.4069E.
- Evans, D. J. and D. J. Searles (1994, August). Equilibrium microstates which generate second law violating steady states. *Physical Review E* 50(2), 1645–1648. Publisher: American Physical Society.
- Evans, D. J. and D. J. Searles (2002, November). The Fluctuation Theorem. *Advances in Physics* 51(7), 1529–1585. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/00018730210155133>.
- Fang, Y., J.-M. Sanz-Serna, and R. D. Skeel (2014, May). Compressible Generalized Hybrid Monte Carlo. *The Journal of Chemical Physics* 140(17), 174108. arXiv:1402.7107 [physics, stat].

- Gattringer, C. and C. B. Lang (2010). *Quantum Chromodynamics on the Lattice: An Introductory Presentation*, Volume 788 of *Lecture Notes in Physics*. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Gelman, A. and D. B. Rubin (1996). Markov chain monte carlo methods in biostatistics. *Statistical methods in medical research* 5(4), 339–355.
- Girolami, M. and B. Calderhead (2011, March). Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 73(2), 123–214.
- Grenander, U. and M. I. Miller (1994, November). Representations of Knowledge in Complex Systems. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 56(4), 549–581.
- Griewank, A. and A. Walther (2008, January). Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, Second Edition. Society for Industrial and Applied Mathematics. Edition: Second.
- Grumitt, R. D. P., B. Dai, and U. Seljak (2022, October). Deterministic Langevin Monte Carlo with Normalizing Flows for Bayesian Inference. arXiv:2205.14240 [cond-mat, physics:physics, stat].
- Hastings, W. K. (1970, April). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1), 97–109.
- Hoffman, M., A. Radul, and P. Sountsov (2021, March). An Adaptive-MCMC Scheme for Setting Trajectory Lengths in Hamiltonian Monte Carlo. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 3907–3915. PMLR. ISSN: 2640-3498.
- Hoffman, M. D. and A. Gelman (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* 15(47), 1593–1623.
- Hoffman, M. D. and P. Sountsov (2022, May). Tuning-Free Generalized Hamiltonian Monte Carlo. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 7799–7813. PMLR. ISSN: 2640-3498.
- Horowitz, A. M. (1991a, October). A generalized guided Monte Carlo algorithm. *Physics Letters B* 268(2), 247–252.
- Horowitz, A. M. (1991b). A generalized guided monte carlo algorithm. *Physics Letters B* 268(2), 247–252.
- Izmailov, P., S. Vikram, M. D. Hoffman, and A. G. G. Wilson (2021, July). What Are Bayesian Neural Network Posteriors Really Like? In *Proceedings of the 38th International Conference on Machine Learning*, pp. 4629–4640. PMLR. ISSN: 2640-3498.
- Janke, W. (2008). Monte carlo methods in classical statistical physics. In *Computational many-particle physics*, pp. 79–140. Springer.
- Jiang, Q. (2023, January). On the dissipation of ideal Hamiltonian Monte Carlo sampler. *Stat* 12(1), e629.
- Leimkuhler, B. and C. Matthews (2015). *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*, Volume 39 of *Interdisciplinary Applied Mathematics*. Cham: Springer International Publishing.
- Leimkuhler, B. and S. Reich (2004). *Simulating hamiltonian dynamics*. Number 14. Cambridge university press.
- Leimkuhler, B. and S. Reich (2009, July). A Metropolis adjusted Nosé-Hoover thermostat. *ESAIM: Mathematical Modelling and Numerical Analysis* 43(4), 743–755.
- Livingstone, S. and G. Zanella (2022, April). The Barker Proposal: Combining Robustness and Efficiency in Gradient-Based MCMC. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84(2), 496–523.

- Margossian, C. C., M. D. Hoffman, P. Sountsov, L. Riou-Durand, A. Vehtari, and A. Gelman (2024, January). Nested R[^]: Assessing the Convergence of Markov Chain Monte Carlo When Running Many Short Chains. *Bayesian Analysis* -1(-1), 1–28. Publisher: International Society for Bayesian Analysis.
- Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953, June). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* 21(6), 1087–1092.
- Minary, P., G. J. Martyna, and M. E. Tuckerman (2003, February). Algorithms and novel applications based on the isokinetic ensemble. II. *Ab initio* molecular dynamics. *The Journal of Chemical Physics* 118(6), 2527–2538.
- Neal, R. M. (2011, May). *MCMC using Hamiltonian dynamics*. arXiv:1206.1901 [physics, stat].
- Neal, R. M. (2020, January). Non-reversibly updating a uniform [0,1] value for Metropolis accept/reject decisions. arXiv:2001.11950 [stat].
- Nesterov, Y. (2009, August). Primal-dual subgradient methods for convex problems. *Mathematical Programming* 120(1), 221–259.
- Owen, A. B. (2017). A randomized halton algorithm in r. *arXiv preprint arXiv:1706.02808*.
- Phan, D., N. Pradhan, and M. Jankowiak (2019a, December). Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. arXiv:1912.11554 [stat].
- Phan, D., N. Pradhan, and M. Jankowiak (2019b). Composable effects for flexible and accelerated probabilistic programming in numpyro. In *Proceedings of the 2nd Symposium on Advances in Approximate Bayesian Inference*. arXiv preprint arXiv:1912.11554.
- Riou-Durand, L., P. Sountsov, J. Vogrinc, C. Margossian, and S. Power (2023, April). Adaptive Tuning for Metropolis Adjusted Langevin Trajectories. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pp. 8102–8116. PMLR. ISSN: 2640-3498.
- Riou-Durand, L. and J. Vogrinc (2022). Metropolis adjusted langevin trajectories: a robust alternative to hamiltonian monte carlo. *arXiv preprint arXiv:2202.13230*.
- Robnik, J., A. E. Bayer, M. Charisi, Z. Haiman, A. Lin, and U. Seljak (2024, October). Periodicity significance testing with null-signal templates: reassessment of PTF’s SMBH binary candidates. *Monthly Notices of the Royal Astronomical Society* 534(2), 1609–1620.
- Robnik, J., R. Cohn-Gordon, and U. Seljak (2024). Black-box unadjusted Hamiltonian Monte Carlo. *arXiv preprint arXiv:2412.08876*.
- Robnik, J., G. B. De Luca, E. Silverstein, and U. Seljak (2024, March). Microcanonical Hamiltonian Monte Carlo. *The Journal of Machine Learning Research* 24(1), 311:14696–311:14729.
- Robnik, J. and U. Seljak (2024, July). Fluctuation without dissipation: Microcanonical Langevin Monte Carlo. In *Proceedings of the 6th Symposium on Advances in Approximate Bayesian Inference*, pp. 111–126. PMLR. ISSN: 2640-3498.
- Sevick, E. M., R. Prabhakar, S. R. Williams, and D. J. Searles (2008, May). Fluctuation Theorems. *Annual Review of Physical Chemistry* 59(1), 603–633. arXiv:0709.3888 [cond-mat].
- Simon-Onfroy, H., F. Lanusse, and A. d. Mattia (2025, April). Benchmarking field-level cosmological inference from galaxy redshift surveys. arXiv:2504.20130 [astro-ph].
- Skeel, R. D. (2009). What makes molecular dynamics work? *SIAM Journal on Scientific Computing* 31(2), 1363–1378.
- Sommer, E., J. Robnik, G. Nozadze, U. Seljak, and D. Rügamer (2024, October). Microcanonical Langevin Ensembles: Advancing the Sampling of Bayesian Neural Networks.
- Sountsov, P., C. Carroll, and M. D. Hoffman (2024, November). Running Markov Chain Monte Carlo on Modern Hardware and Software. arXiv:2411.04260 [stat].

- Sountsov, P. and M. D. Hoffman (2022, May). Focusing on Difficult Directions for Learning HMC Trajectory Lengths. arXiv:2110.11576 [stat].
- Sountsov, P., A. Radul, and contributors (2020). Inference gym.
- Steeg, G. V. and A. Galstyan (2021, December). Hamiltonian Dynamics with Non-Newtonian Momentum for Rapid Sampling. arXiv:2111.02434 [physics].
- Tripuraneni, N., M. Rowland, Z. Ghahramani, and R. Turner (2017, July). Magnetic Hamiltonian Monte Carlo. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3453–3461. PMLR. ISSN: 2640-3498.
- Tuckerman, M. E. (2023). *Statistical mechanics: theory and molecular simulation*. Oxford university press.
- Tuckerman, M. E., Y. Liu, G. Ciccotti, and G. J. Martyna (2001). Non-hamiltonian molecular dynamics: Generalizing hamiltonian phase space principles to non-hamiltonian systems. *The Journal of Chemical Physics* 115(4), 1678–1702.
- Wenzel, F., K. Roth, B. S. Veeling, J. Świątkowski, L. Tran, S. Mandt, J. Snoek, T. Salimans, R. Jenatton, and S. Nowozin (2020, July). How Good is the Bayes Posterior in Deep Neural Networks Really? arXiv:2002.02405 [stat].
- Štrumbelj, E., A. Bouchard-Côté, J. Corander, A. Gelman, H. Rue, L. Murray, H. Pesonen, M. Plummer, and A. Vehtari (2024, February). Past, Present and Future of Software for Bayesian Inference. *Statistical Science* 39(1), 46–61. Publisher: Institute of Mathematical Statistics.

A Metropolis adjusted Microcanonical Langevin dynamics proofs

Denote by $o(\mathbf{z}'|\mathbf{z})$ the density corresponding to the O update and by $q(\mathbf{z}'|\mathbf{z})$, the density corresponding to the single step proposal $TOBABO$. We will use a shorthand notation for the time reversal: $\bar{\mathbf{z}} = \mathcal{T}(\mathbf{z})$ and denote by $\Delta(\mathbf{z}', \mathbf{z})$ the energy error accumulated in the *deterministic* part of the update.

A.1 Proof of Theorem 5.1

Proof. For the MH ratio we will need

$$\frac{q(\mathbf{z}|\bar{\mathbf{z}}')}{q(\bar{\mathbf{z}}'|\mathbf{z})} = \frac{\int o(\bar{\mathbf{z}}|\mathbf{Z}') \delta(\mathbf{Z}', \varphi(\mathbf{Z})) o(\mathbf{Z}|\bar{\mathbf{z}}') d\mathbf{Z} d\mathbf{Z}'}{\int o(\mathbf{z}'|\mathbf{Z}') \delta(\mathbf{Z}', \varphi(\mathbf{Z})) o(\mathbf{Z}|\mathbf{z}) d\mathbf{Z} d\mathbf{Z}'} = \frac{\int o(\bar{\mathbf{z}}|\varphi(\mathbf{Z})) o(\mathbf{Z}|\bar{\mathbf{z}}') d\mathbf{Z}}{\int o(\mathbf{z}'|\varphi(\mathbf{Z})) o(\mathbf{Z}|\mathbf{z}) d\mathbf{Z}}, \quad (16)$$

where we have used the delta function to evaluate the integral over $\mathbf{Z}'$.

We can further simplify the numerator

$$\begin{aligned} \int o(\bar{\mathbf{z}}|\varphi(\mathbf{Z})) o(\mathbf{Z}|\bar{\mathbf{z}}') d\mathbf{Z} &= \int o(\mathbf{z}|\overline{\varphi(\mathbf{Z})}) o(\bar{\mathbf{Z}}|\mathbf{z}') d\mathbf{Z} = \int o(\mathbf{z}|\overline{\varphi(\mathbf{Z})}) o(\mathbf{Z}|\mathbf{z}') d\mathbf{Z} \\ &= \int o(\mathbf{z}|\varphi^{-1}(\mathbf{Z})) o(\mathbf{Z}|\mathbf{z}') d\mathbf{Z} = \int o(\mathbf{z}|\mathbf{Z}) o(\varphi(\mathbf{Z})|\mathbf{z}') \left| \frac{\partial \varphi(\mathbf{Z})}{\partial \mathbf{Z}} \right| d\mathbf{Z} = \int o(\mathbf{z}'|\varphi(\mathbf{Z})) o(\mathbf{Z}|\mathbf{z}) \left| \frac{\partial \varphi(\mathbf{Z})}{\partial \mathbf{Z}} \right| d\mathbf{Z}. \end{aligned}$$

In the first step we have used that $o(\bar{\mathbf{x}}|\bar{\mathbf{y}}) = o(\mathbf{x}|\mathbf{y})$ and that time reversal is an involution. In the second step, we have performed a change of variables from $\bar{\mathbf{Z}}$ to $\mathbf{Z}$ (for which, the Jacobian determinant of the transformation is 1). In the third step we used that $\overline{\varphi(\mathbf{Z})} = \varphi^{-1}(\mathbf{Z})$. In the fourth step we change variables to $\varphi^{-1}(\mathbf{Z})$ instead of $\mathbf{Z}$. In the last step we use that $o(\mathbf{y}|\mathbf{x}) = o(\mathbf{x}|\mathbf{y})$.

Since o only connects states with the same $\mathbf{x}$ there is only one $\mathbf{Z}$ which makes the integral nonvanishing and we get

$$\frac{q(\bar{\mathbf{z}}'|\mathbf{z})}{q(\mathbf{z}|\bar{\mathbf{z}}')} = \left| \frac{\partial \varphi(\mathbf{Z})}{\partial \mathbf{Z}} \right|,$$

as if there were no O updates. The O updates also preserve the target density, so we see that the acceptance probability is only concerned with the BAB part of the update. In this case, the desired acceptance probability was already derived in Lemma 4.2. $\square$

A.2 Proof of Theorem 5.2

Proof. Following a similar structure of the proof as in (Riou-Durand and Vogrinc, 2022) we will work on the space of trajectories $\mathbf{z}_{0:L} = (\mathbf{z}_0, \dots, \mathbf{z}_L) \in \mathcal{M}^{L+1}$. We will define a kernel $\mathcal{Q}$ on the space of trajectories, with q as a marginal- $\mathbf{x}_0$ kernel. We will prove that $\mathcal{Q}$ is reversible with respect to the extended density

$$\mathcal{P}(\mathbf{z}_{0:L}) = \prod_{i=1}^L q(\mathbf{z}_i|\mathbf{z}_{i-1}) p(\mathbf{z}_0),$$

and use it to show that q is reversible with respect to the marginal $p(\mathbf{x}_0)$.

We define the Gibbs update, corresponding to the conditional distribution $\mathcal{P}(\cdot|\mathbf{x}_0)$:

$$\mathcal{G}(\mathbf{z}'_{0:L}|\mathbf{z}_{0:L}) = \delta(\mathbf{x}'_0 - \mathbf{x}_0) U_{S^{d-1}}(\mathbf{u}'_0) \prod_{i=1}^L q(\mathbf{z}'_i|\mathbf{z}'_{i-1}).$$

The Gibbs kernel $\mathcal{G}$ is reversible with respect to $\mathcal{P}$ by construction. Built upon a deterministic proposal of the backward trajectory

$$\bar{\mathbf{z}}_{0:L} = (\bar{\mathbf{z}}_L, \bar{\mathbf{z}}_{L-1}, \dots, \bar{\mathbf{z}}_0), \quad (17)$$

we introduce a Metropolis update:

$$M(\mathbf{z}'_{0:L}|\mathbf{z}_{0:L}) = P_{MH} \delta(\mathbf{z}'_{0:L} - \bar{\mathbf{z}}_{0:L}) + (1 - P_{MH}) \delta(\mathbf{z}'_{0:L} - \mathbf{z}_{0:L}),$$

where $P_{MH}(\mathbf{z}_{0:L}) = \min(1, e^{-\Delta(\mathbf{z}_{0:L})})$. For $\eta > 0$, the distribution $\mathcal{P}$ admits a density with respect to Lebesgue's measure. Therefore

$$e^{-\Delta(\mathbf{z}_{0:L})} = \frac{\mathcal{P}(\bar{\mathbf{z}}_{0:L})}{\mathcal{P}(\mathbf{z}_{0:L})} \left| \frac{\partial \bar{\mathbf{z}}_{0:L}}{\partial \mathbf{z}_{0:L}} \right| \quad (18)$$

ensures that the Metropolis kernel M is reversible with respect to $\mathcal{P}$.

Before proceeding with the proof, we express Equation (18) in a simple, easy-to-compute form. The Jacobian is $\frac{\partial \bar{\mathbf{z}}_{0:L}}{\partial \mathbf{z}_{0:L}} = \sigma \otimes \frac{\partial \bar{\mathbf{z}}}{\partial \mathbf{z}}$, where σ is the matrix of the permutation $\sigma(i) = L - i$ and $\frac{\partial \bar{\mathbf{z}}}{\partial \mathbf{z}} = I_{d \times d} \oplus -I_{d-1 \times d-1}$. Both of these matrices have determinant ± 1 , so the determinant of their Kronecker product is also ± 1 and its absolute value is 1.

We get

$$\begin{aligned} e^{-\Delta(\mathbf{z}_{0:L})} &= \frac{\mathcal{P}(\bar{\mathbf{z}}_{0:L})}{\mathcal{P}(\mathbf{z}_{0:L})} = \frac{p(\bar{\mathbf{z}}_L) \prod_{i=1}^L q(\bar{\mathbf{z}}_{i-1}|\bar{\mathbf{z}}_i)}{p(\mathbf{z}_0) \prod_{i=1}^L q(\mathbf{z}_i|\mathbf{z}_{i-1})} = \frac{\prod_{i=1}^L p(\bar{\mathbf{z}}_i) \prod_{i=1}^L q(\bar{\mathbf{z}}_{i-1}|\bar{\mathbf{z}}_i)}{\prod_{i=1}^L p(\mathbf{z}_{i-1}) \prod_{i=1}^L q(\mathbf{z}_i|\mathbf{z}_{i-1})} \\ &= \prod_{i=1}^L \frac{q(\bar{\mathbf{z}}_{i-1}|\bar{\mathbf{z}}_i)p(\bar{\mathbf{z}}_i)}{q(\mathbf{z}_i|\mathbf{z}_{i-1})p(\mathbf{z}_{i-1})} = e^{-\sum_{i=1}^L \Delta(\mathbf{z}_i, \mathbf{z}_{i-1})}, \end{aligned} \quad (19)$$

where $\Delta(\mathbf{z}_i, \mathbf{z}_{i-1})$ is the energy error in step i , by Theorem 5.1.

We are now in a position to define the trajectory-space kernel:

$$\mathcal{Q} = \mathcal{G}M\mathcal{G}. \quad (20)$$

The palindromic structure of $\mathcal{Q}$ ensures reversibility with respect to $\mathcal{P}$. Since the transition $\mathcal{G}(\cdot|\mathbf{z}_{0:L}) = \mathcal{G}(\cdot|\mathbf{x}_0)$ only depends on the starting position $\mathbf{x}_0 \in \mathbb{R}^d$ and $p(\mathbf{x})$ is the marginal of $\mathcal{P}$, we obtain that $q(\mathbf{x}'_0|\mathbf{x}_0) = \int \mathcal{Q}(\mathbf{z}'_{0:L}|\mathbf{z}_{0:L}) d\mathbf{u}_0 \prod_{i=1}^L d\mathbf{z}_i$ defines marginally a Markov kernel on $\mathbb{R}^d$, reversible with respect to p . In particular, the distribution of $\{\mathbf{x}_i\}_{i \geq 0}$ in Algorithm 1 coincides with the distribution of a Markov chain generated by q . $\square$

B Microcanonical dynamics

In this appendix, we establish a relationship between the microcanonical dynamics of Equation (3) and a Hamiltonian system with energy E from which it can be derived by a time-rescaling operation. As well as motivating the dynamics of Equation (3), this allows us to show that W in Lemma 4.2 for the dynamics of Equation (3) corresponds to the change in energy E of the Hamiltonian system. We also provide a complete derivation of the form of W for microcanonical dynamics. Familiarity with the basics of Hamiltonian mechanics is assumed throughout.

B.1 Sundman transformation

We begin by introducing a transformation to a Hamiltonian system known as a Sundman transform (Leimkuhler and Reich, 2004)

$$S(F)(\mathbf{z}(t)) = w(\mathbf{z}(t))F(\mathbf{z}(t)),$$

where w is any function $\mathbb{R}^{2d} \rightarrow \mathbb{R}$. Intuitively, this is a $\mathbf{z}$ -dependent time rescaling of the dynamics. Therefore it is not surprising that:

Lemma B.1. *The integral curves of $S(F)$ are the same as of F (Skeel, 2009)*

Proof. To see this, first use $\mathbf{z}_G$ to refer to the dynamics from a field G , and posit that $\mathbf{z}_{S(F)}(s) = \mathbf{z}_F(t(s))$, where $\frac{dt(s)}{ds} = w(\mathbf{z}(s))$. Then we see that

$$\frac{d\mathbf{z}_{S(F)}(s)}{ds} = \frac{d\mathbf{z}_F(t(s))}{ds} = \frac{d\mathbf{z}_F(t)}{dt} \frac{dt}{ds} = F(\mathbf{z}_F(s))w(\mathbf{z}_F(s)),$$

which shows that, indeed, $z_{S(F)} = z_F \circ s$, where s is a function $\mathbb{R} \rightarrow \mathbb{R}$, which amounts to what we set out to show. $\square$

However, note that the stationary distribution is not necessarily preserved, on account of the phase space dependence of the time-rescaling, which means that in a volume of phase space, different particles will move at different velocities.

B.2 Obtaining the dynamics of Equation (3)

Consider the Hamiltonian system⁷ given by $H = T + V$, with $T(\mathbf{\Pi}) = (d-1) \log \|\mathbf{\Pi}\|$ and $V(\mathbf{x}) = \mathcal{L}(\mathbf{x})$. Here, $\mathbf{\Pi}$ is the canonical momentum associated to position $\mathbf{x}$. Then the dynamics derived from Hamilton's equations of motion are:

$$\frac{d}{dt} \begin{bmatrix} \mathbf{x} \\ \mathbf{\Pi} \end{bmatrix} = \begin{bmatrix} \frac{\partial H}{\partial \mathbf{\Pi}} \\ -\frac{\partial H}{\partial \mathbf{x}} \end{bmatrix} = \begin{bmatrix} (d-1) \frac{\mathbf{\Pi}}{\|\mathbf{\Pi}\|^2} \\ -\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}) \end{bmatrix} := F(\mathbf{z}). \quad (21)$$

Any Hamiltonian dynamics has $p(\mathbf{z}) \propto \delta(H - C)$ as a stationary distribution, which can be sampled from by integrating the equations if ergodicity holds. As observed in Steeg and Galstyan (2021) and Robnik et al. (2024), the closely related Hamiltonian $d \log \|\mathbf{\Pi}\| + \mathcal{L}(\mathbf{x})$ has the property that the marginal of this stationary distribution is the desired target, namely $p(\mathbf{x}) \propto e^{-\mathcal{L}(\mathbf{x})}$. However, numerical integration of these equations is unstable due to the $\frac{1}{\|\mathbf{\Pi}\|^2}$ factor, and moreover, MH adjustment is not possible since numerical integration induces error in H , which would result in proposals always being rejected, due to the delta function.

Both problems can be addressed with a Sundman transform and a subsequent change of variables. To that end, we choose $w(\mathbf{z}) = \|\mathbf{\Pi}\|/(d-1)$ (which corresponds, up to a factor, to the weight r in Steeg and Galstyan (2021), and to w in Robnik et al. (2024)), we obtain:

$$\frac{d}{dt} \begin{bmatrix} \mathbf{x} \\ \mathbf{\Pi} \end{bmatrix} = \begin{bmatrix} \mathbf{\Pi} / \|\mathbf{\Pi}\| \\ -\nabla \mathcal{L}(\mathbf{x}) \|\mathbf{\Pi}\| / (d-1) \end{bmatrix}. \quad (22)$$

Changing variables to $\mathbf{u} = \mathbf{\Pi} / \|\mathbf{\Pi}\|$, we obtain precisely the microcanonical dynamics of Equation (3):

$$\frac{d}{dt} \begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{u} \\ -(I - \mathbf{u}\mathbf{u}^T) \nabla \mathcal{L}(\mathbf{x}) / (d-1) \end{bmatrix} := \begin{bmatrix} B_{\mathbf{x}} \\ B_{\mathbf{u}} \end{bmatrix},$$

where we have used that the Jacobian $\frac{d\mathbf{u}}{d\mathbf{\Pi}} = \frac{1}{\|\mathbf{\Pi}\|} (I - \frac{\mathbf{\Pi}\mathbf{\Pi}^T}{\|\mathbf{\Pi}\|^2})$. Note that $B_{\mathbf{x}} = S(F)_{\mathbf{x}}$, since this final change of variable only targets $\mathbf{\Pi}$.

B.3 Discrete updates

For completeness, we here state the position and velocity updates of the canonical and microcanonical dynamics, which are obtained by solving dynamics at fixed velocity for the position update and at fixed position for the velocity update. For canonical dynamics, this amounts to solving

$$\frac{d}{d\epsilon} A_{\epsilon} = \mathbf{u}(t) \quad \frac{d}{d\epsilon} B_{\epsilon} = -\nabla \mathcal{L}(\mathbf{x}(t)), \quad (23)$$

with initial condition $A_0 = \mathbf{x}(t)$ and $B_0 = \mathbf{u}(t)$. These solution is trivial:

$$A_{\epsilon} = \mathbf{x}(t) + \epsilon \mathbf{u}(t) \quad B_{\epsilon} = \mathbf{u}(t) - \epsilon \nabla \mathcal{L}(\mathbf{x}(t)). \quad (24)$$

⁷Here we follow Robnik et al. (2024) and Steeg and Galstyan (2021), but our Hamiltonian differs by a factor, to avoid the need for a weighting scheme used in those papers.

For microcanonical dynamics, one needs to solve

$$\frac{d}{d\epsilon}A_\epsilon = \mathbf{u}(t) \quad \frac{d}{d\epsilon}B_\epsilon = -(1 - \mathbf{u}(t)\mathbf{u}(t)^T)\nabla\mathcal{L}(\mathbf{x}(t))/(d-1), \quad (25)$$

with initial condition $A_0 = \mathbf{x}(t)$ and $B_0 = \mathbf{u}(t)$. The velocity equation is a vector version of the Riccati equation (Steeg and Galstyan, 2021). Denote $\mathbf{g} = -\nabla\mathcal{L}(\mathbf{x}(t))/(d-1)$ and replace the variable B_ϵ by $\mathbf{y}_\epsilon$, such that

$$B_\epsilon = \frac{\frac{d}{d\epsilon}\mathbf{y}_\epsilon}{\mathbf{g} \cdot \mathbf{y}_\epsilon}. \quad (26)$$

This is convenient, because the equation for B_ϵ is a nonlinear first-order differential equation, but the equation for $\mathbf{y}_\epsilon$ is a *linear* second-order differential equation

$$\frac{d^2}{d\epsilon^2}\mathbf{y}_\epsilon = (\mathbf{g}\mathbf{g}^T)\mathbf{y}_\epsilon, \quad (27)$$

which is easy to solve and yields the updates

$$A_\epsilon = \mathbf{x}(t) + \epsilon\mathbf{u}(t) \quad B_\epsilon = \frac{\mathbf{u}(t) + (\sinh \delta + \mathbf{e} \cdot \mathbf{u}(t)(\cosh \delta - 1))\mathbf{e}}{\cosh \delta + \mathbf{e} \cdot \mathbf{u}(t) \sinh \delta}, \quad (28)$$

where $\delta = \epsilon \|\nabla\mathcal{L}(\mathbf{x}(t))\|/(d-1)$ and $\mathbf{e} = -\nabla\mathcal{L}(\mathbf{x})/\|\nabla\mathcal{L}(\mathbf{x})\|$.

B.4 Obtaining the stationary distribution of Equation (3)

We can derive the stationary distribution of Equation (3) following the approach of Tuckerman (2023). There, it is shown that for a flow F , if there is a g such that $\frac{d}{dt} \log g = -\nabla \cdot F$, and Λ is the conserved quantity under the dynamics, then $p(\mathbf{z}) \propto g(\mathbf{z})f(\Lambda(\mathbf{z}))$, where f is any function.

We note that $\nabla \cdot F = \mathbf{u} \cdot \nabla\mathcal{L}(\mathbf{x}) = \frac{d}{dt}\mathcal{L}(\mathbf{x})$, using Appendix B.7 in the first step. Therefore $\log g = -\mathcal{L}(\mathbf{x})$. Further, $|\mathbf{u}|$ is preserved by the dynamics if we initialize with $|\mathbf{u}_0| = 1$, as can easily be seen: $\frac{d}{dt}(\mathbf{u} \cdot \mathbf{u}) = 2\mathbf{u} \cdot \dot{\mathbf{u}} = 2\mathbf{u} \cdot (I - \mathbf{u}\mathbf{u}^T)(-\nabla\mathcal{L}(\mathbf{x})/(d-1)) = 2(1 - \mathbf{u} \cdot \mathbf{u})(\mathbf{u} \cdot \nabla\mathcal{L}/(d-1)) = 0$. Thus a stationary distribution is:

$$p(\mathbf{x}, \mathbf{u}) \propto e^{-\mathcal{L}(\mathbf{x})}\delta(\|\mathbf{u}\| - 1). \quad (29)$$

Importantly, because even the discretized dynamics are norm preserving, the condition $\delta(\|\mathbf{u}\| - 1)$ is always satisfied, so that $\frac{p(\mathbf{z}')}{p(\mathbf{z})}$ is always well defined. This makes it possible to perform MH adjustment, in contrast to the original Hamiltonian dynamics as discussed in Appendix B.2.

B.5 W as energy change

In the non-equilibrium physics literature, W (termed the dissipation function) is interpreted as work done on the system and the second term in Lemma 4.1 is the dissipated heat (Evans and Searles, 1994, 2002; Sevcik et al., 2008). W plays a central role in fluctuation theorems, for example, Crook's relation (Crooks, 1999) states that the transitions $\mathbf{z} \rightarrow \mathbf{z}'$ are more probable than $\mathbf{z}' \rightarrow \mathbf{z}$ by a factor $e^{W(\mathbf{z}', \mathbf{z})}$. In statistics, this fact is used by the MH algorithm to obtain reversibility, or *detailed balance*, a sufficient condition for convergence to the target distribution.

Here we will justify why it can also be interpreted as an energy change in microcanonical dynamics.

Lemma B.2. *W , calculated for the microcanonical dynamics over a time interval $[0, T]$ is equal to ΔE of the Hamiltonian $(d-1) \log \|\mathbf{\Pi}\| + \mathcal{L}(\mathbf{x})$ for an interval $[s(0), s(T)]$, where s is the time rescaling arising from the Sundman transformation $w(\mathbf{z}) = \|\mathbf{\Pi}_F\|/(d-1)$.*

Proof. Recall that for a flow field F :

$$W(\mathbf{z}_F(T), \mathbf{z}_F(0)) = -\log \frac{p(\mathbf{z}_F(T))}{p(\mathbf{z}_F(0))} - \int_0^T \nabla \cdot F(\mathbf{z}_F(s))ds. \quad (30)$$

Given the form of the stationary distribution induced by B , derived in Appendix B.4, we see that the first term of the work, $\log \frac{P(\mathbf{z}_B(0))}{P(\mathbf{z}_B(T))} = \mathcal{L}(\mathbf{x}_B(T)) - \mathcal{L}(\mathbf{x}_B(0)) = \mathcal{L}(\mathbf{x}_{S(F)}(T)) - \mathcal{L}(\mathbf{x}_{S(F)}(0)) = \mathcal{L}(\mathbf{x}_F(s(T))) - \mathcal{L}(\mathbf{x}_F(s(0)))$ which is equal to ΔV for an interval of time $[s(0), s(T)]$.

As for the second term, observe that

$$\frac{dK(\boldsymbol{\Pi}(t(s)))}{ds} = \frac{\partial H}{\partial \boldsymbol{\Pi}} \cdot \frac{d\boldsymbol{\Pi}}{dt} \frac{dt}{ds} = \frac{d\mathbf{x}}{dt} \frac{dt}{ds} \cdot \frac{d\boldsymbol{\Pi}}{dt} = \mathbf{u} \cdot (-\nabla \mathcal{L}(\mathbf{x})) = -\nabla \cdot B,$$

which is precisely the integrand of the second term. □

This shows that $W = \Delta K + \Delta V = \Delta E$, where ΔE is the energy change of the original Hamiltonian, over the rescaled time interval $[s(0), s(T)]$. As we know, $\Delta E = 0$ for the exact Hamiltonian flow, and indeed $W = 0$ for the exact dynamics of Equation (3), which is to say that for the exact dynamics, no MH correction would be needed for an asymptotically unbiased sampler.

However, our practical interest is in the discretized dynamics arising from a Velocity Verlet numerical integrator. In this case, we wish to calculate W for $B_{\mathbf{u}}$ and $B_{\mathbf{x}}$ separately, and consider the sum, noting that W is an additive quantity with respect to the concatenation of two dynamics. Considering W with respect to only $B_{\mathbf{x}}$, we see that the first term of W remains ΔV , since the stationary distribution gives uniform weight to all values of $\mathbf{u}$ of unit norm, and the dynamics are norm preserving. The second term vanishes, because $\nabla_{\mathbf{x}} B_{\mathbf{x}} = \nabla_{\mathbf{x}} \mathbf{u} = 0$. As for $B_{\mathbf{u}}$, since the norm preserving change in $\mathbf{u}$ leaves the density unchanged, the first term of W vanishes. Meanwhile, the second term is ΔK , from the above derivation, since $\nabla \cdot B = \nabla_{\mathbf{x}} \cdot B_{\mathbf{x}} + \nabla_{\mathbf{u}} \cdot B_{\mathbf{u}} = \nabla_{\mathbf{u}} \cdot B_{\mathbf{u}}$. Thus, the full W is equal to $\Delta V + \Delta K = \Delta E$, as desired. For HMC, it is easily seen that W for $F_{\mathbf{x}}$ is ΔV , and for $F_{\mathbf{u}}$ is ΔT . Putting this together, we maintain the result of Lemma B.2, but now in a setting where W is not 0 so that MH adjustment is of use.

B.6 Direct calculation of velocity update W

We here provide a self-contained derivation of the MH ratio for the velocity update from Equation (28). The MH ratio is a scalar with respect to state space transformations, i.e. it is the same in all coordinate systems. We can therefore select convenient coordinates for its computation. We will choose spherical coordinates in which $\mathbf{e}$ is the north pole and

$$\mathbf{u} = \cos \vartheta \mathbf{e} + \sin \vartheta \mathbf{f}, \quad (31)$$

for some unit vector $\mathbf{f}$, orthogonal to $\mathbf{e}$. ϑ is then a coordinate on the S^{d-1} manifold. The velocity updating map from Equation (28),

$$\mathbf{u}' = \frac{1}{\cosh \delta + \cos \vartheta \sinh \delta} \mathbf{u} + \frac{\sinh \delta + \cos \vartheta (\cosh \delta - 1)}{\cosh \delta + \cos \vartheta \sinh \delta} \mathbf{e} \quad (32)$$

$$= \frac{\sinh \delta + \cos \vartheta \cosh \delta}{\cosh \delta + \cos \vartheta \sinh \delta} \mathbf{e} + \frac{\sin \vartheta}{\cosh \delta + \cos \vartheta \sinh \delta} \mathbf{f}, \quad (33)$$

can be expressed in terms of the ϑ variable:

$$\cos \vartheta' = \frac{\sinh \delta + \cos \vartheta \cosh \delta}{\cosh \delta + \cos \vartheta \sinh \delta} \quad \sin \vartheta' = \frac{\sin \vartheta}{\cosh \delta + \cos \vartheta \sinh \delta}. \quad (34)$$

The Jacobian of the $\vartheta \mapsto \vartheta'$ transformation is

$$\left| \frac{d\vartheta'}{d\vartheta} \right| = \left| \frac{d\vartheta'}{d \cos \vartheta'} \frac{d \cos \vartheta'}{d\vartheta} \right| = \frac{1}{|\cosh \delta + \cos \vartheta \sinh \delta|} \quad (35)$$

and the density ratio is

$$\frac{p(\vartheta')}{p(\vartheta)} = \sqrt{\frac{g(\vartheta')}{g(\vartheta)}} = \left(\frac{\sin \vartheta'}{\sin \vartheta} \right)^{d-2} = \frac{1}{(\cosh \delta + \cos \vartheta \sinh \delta)^{d-2}}, \quad (36)$$

where g is the metric determinant on a S^{d-1} sphere. Combining the two together yields

$$W = (d-1) \log (\cosh \delta + \cos \vartheta \sinh \delta), \quad (37)$$

which is the kinetic energy from Equation (8).

B.7 Direct calculation of the velocity update divergence

For completeness, we here derive the divergence of the microcanonical velocity update flow field F . We will use the divergence theorem, which states that the integral of the divergence of a vector field over some volume Ω equals the flux of this vector field over the boundary of Ω . Here, flux is $F \cdot n$ where n is the unit vector, normal to the boundary.

We will use the coordinate system defined in Equation (31) and pick as the volume Ω a thin spherical shell, centered around the north pole e and spanning the ϑ range $[\vartheta, \vartheta + \Delta\vartheta]$. The boundary of Ω are two spheres in $d - 2$ dimensions with radii $\sin \vartheta$ and $\sin(\vartheta + \Delta\vartheta)$. Note that F is normal to this boundary and flux is a constant on each shell. It is outflowing on the boundary which is closer to the north pole and inflowing on the other boundary.

Note that for $\Delta\vartheta \rightarrow 0$, we have that $\nabla \cdot F$ is a constant on Ω . The divergence theorem in this limit therefore implies

$$(\nabla \cdot F) V(S^{d-2})(\sin \vartheta)^{d-2} = -\frac{d}{d\vartheta}(|F|V(S^{d-2})(\sin \vartheta)^{d-2}), \quad (38)$$

where $V(S^{d-2})$ is the volume of the unit sphere in $d-2$ dimensions and we have used that the volume of a n -dimensional sphere with radius r is $V(S^n)r^n$. By rearranging we get:

$$\nabla \cdot F = -\frac{\frac{d}{d\vartheta}(|F|(\sin \vartheta)^{d-2})}{(\sin \vartheta)^{d-2}}. \quad (39)$$

We have

$$F = \frac{\|\nabla \mathcal{L}(\mathbf{x})\|}{d-1}(1 - \mathbf{u}\mathbf{u}^T)\mathbf{e}, \quad (40)$$

so

$$|F| = \frac{\|\nabla \mathcal{L}(\mathbf{x})\|}{d-1} \sqrt{e(1 - \mathbf{u}\mathbf{u}^T)\mathbf{e}} = \frac{\|\nabla \mathcal{L}(\mathbf{x})\|}{d-1} \sqrt{1 - (\mathbf{e} \cdot \mathbf{u})^2} = \frac{\|\nabla \mathcal{L}(\mathbf{x})\|}{d-1} \sin \vartheta. \quad (41)$$

Inserting $|F|$ in Equation (39) yields

$$\nabla \cdot F = -\frac{\|\nabla \mathcal{L}(\mathbf{x})\|}{d-1} \frac{(d-1)(\sin \vartheta)^{d-2} \cos \vartheta}{(\sin \vartheta)^{d-2}} = -\|\nabla \mathcal{L}\| \mathbf{e} \cdot \mathbf{u}. \quad (42)$$

C Optimal acceptance rate

The optimal acceptance rate argument for MAMS is analogous to the one in Neal (2011). We will use two general properties of the deterministic MH proposal:

1. The expected value of the MH ratio under the stationary distribution, $\mathbb{E}_{z \sim p}[e^{-W(z', z)}]$, is

$$\int p(\mathbf{z}) e^{-W(\mathbf{z}', \mathbf{z})} d\mathbf{z} = \int p(\mathbf{z}) \frac{q(\mathbf{z}|\mathbf{z}')p(\mathbf{z}')}{q(\mathbf{z}'|\mathbf{z})p(\mathbf{z})} d\mathbf{z} = \int p(\mathbf{z}') \left| \frac{\partial \varphi}{\partial \mathbf{z}}(\mathbf{z}) \right| d\mathbf{z} = \int p(\varphi(\mathbf{z})) d\varphi(\mathbf{z}) = 1.$$

This is the Jarzynski equality. In statistical literature it was used by Neal (2011); Creutz (1988) in the special case when φ is symplectic.

2. In equilibrium,

$$P(W > 0 | \text{accepted}) = P(W < 0 | \text{accepted}) = \frac{1}{2}$$

by the design of the MH algorithm (Neal, 2011). Since $P(\text{accepted} | W < 0) = 1$, we have that

$$\frac{1}{2} = P(w < 0 | \text{accepted}) = \frac{P(W < 0 | \text{accepted})P(W < 0)}{P(\text{accepted})} = \frac{P(W < 0)}{P(\text{accepted})},$$

so $P(\text{accepted}) = 2P(W < 0)$.

Let us approximate the stationary distribution over W as $\mathcal{N}(\mu, \sigma^2)$, as in Neal (2011). We then have by the Jarzynski equality:

$$1 = \int p(z) e^{-W(z', z)} dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(W-\mu)^2/2\sigma^2} e^{-W} dW = e^{\frac{\sigma^2}{2} - \mu}, \quad (43)$$

implying that $\sigma^2 = 2\mu$. By property (2) we then have

$$P(\text{accept}) = 2\Phi(-\mu/\sqrt{2\mu}), \quad (44)$$

where Φ is the Gaussian cumulative density function. Denote by K_{accepted} the number of accepted proposals needed for a new effective sample. This corresponds to moving a distance on the order of the size of the typical set

$$K_{\text{accepted}} N \epsilon \propto \sqrt{d}, \quad (45)$$

since $\sqrt{d}$ is the size of the standard Gaussian's typical set. The number of effective samples per gradient call is then

$$\text{ESS} / \# \text{ gradient evaluations} = \frac{1}{K_{\text{total}} N} = \frac{P(\text{accept})}{K_{\text{accepted}} N} \propto \frac{\epsilon P(\text{accept})}{\sqrt{d}}. \quad (46)$$

The error of the MCHMC Velocity Verlet integrator for an interval of fixed length is (Robnik et al., 2024)

$$\sigma^2/d \propto \epsilon^4/d^2, \quad (47)$$

implying that $\sigma^2 = 2\mu \propto \epsilon^4/d$. Therefore

$$\text{ESS} / \# \text{ gradient evaluations} \propto \mu^{1/4} \Phi(-\sqrt{\mu/2}) d^{-1/4}, \quad (48)$$

so we see that the efficiency drops as $d^{-1/4}$. ESS is maximal at $\mu = 0.41$, corresponding to $P(\text{accept}) = 65\%$. From Equation (47) we then see that the optimal stepsize grows as $\epsilon \propto d^{1/4}$ instead of the $d^{1/2}$ that would correspond to the unimpaired efficiency.

Note that this result is different if a higher-order integrator is used. For example, when using a fourth order integrator $\sigma^2/d \propto (\epsilon^2/d)^4$, the optimal setting is $\mu = 0.13$ and $P(\text{accept}) = 80\%$.

Empirically, we find that even for a second-order integrator, targeting a higher acceptance rate, of 90%, works well in practice; we use this in our experiments.

D Optimal rate of partial refreshments

We will here demonstrate the sensitivity of HMC to trajectory length for ill-conditioned Gaussian distributions and show how Langevin dynamics alleviates this issue. We will then obtain the optimal rate of partial refreshments in the limit of large condition number. For MALT, Riou-Durand and Vogrinc (2022) derive the ESS of the second moments in the continuous-time limit for the Gaussian targets $\mathcal{N}(0, \sigma)$:

$$\text{ESS}(\beta, T) = \frac{1 - \rho^2}{1 + \rho^2}, \quad (49)$$

where

$$\rho = e^{-\beta T} \left(\cos \omega T + \frac{\beta}{\omega} \sin \omega T \right) \quad \omega = \sqrt{\frac{1}{\sigma^2} - \beta^2}. \quad (50)$$

T is the trajectory length, β is the LMC damping parameter ($\gamma = 2\beta$ in Riou-Durand and Vogrinc (2022)).

Figure 4 shows ESS as a function of the trajectory length and the rate of partial refreshments, both for the standard Gaussian and for the Gaussian in the limit of very large condition number. For the standard Gaussian, HMC achieves better performance than MALT, but only if trajectory length is chosen well. For ill-conditioned Gaussian however, HMC drastically underperforms compared to MALT. Optimal MALT hyperparameter settings for the ill conditioned Gaussians are $\beta = 0.567$ and $T = 1.413$. Therefore the optimal ratio of the decoherence time scales of the partial and full refreshment are

$$\frac{t_{\text{partial}}}{t_{\text{full}}} = \frac{1/\beta}{T} = 1.25. \quad (51)$$

We will use the same setting for Langevin MAMS, so $L_{\text{partial}}/L = 1.25$.

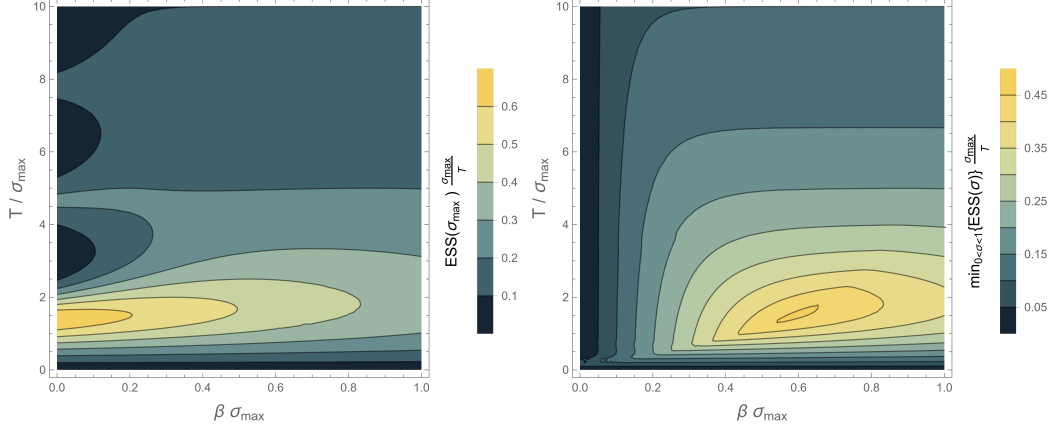


Figure 4: Effective sample size in continuous time for MALT LMC on Gaussian targets. x-axis is the LMC damping parameter, y-axis the trajectory length. $x = 0$ is the HMC line ($\beta = 0$ means there is no damping and no Langevin noise), $x = 1$ is LMC with critical damping $\beta = 1/\sigma_{\max}$. Left panel: isotropic Gaussian $\mathcal{N}(0, \sigma_{\max})$. Note that HMC achieves the optimal performance if properly tuned, the only reason to introduce Langevin noise would be to potentially make the tuning easier. Right panel: extremely ill-conditional Gaussian with all scales $(0, \sigma_{\max}]$. ESS along the worst direction is shown. HMC performs poorly as it cannot be tuned to all scales, damping of $\beta \sigma_{\max} = 0.57$ performs best. Note that these results do not imply MALT having non-zero ESS in the infinite condition number limit: we only study continuous time MALT here.

E Further experimental details

E.1 Benchmark Inference Models

We detail the inference models used in Section 7. For models adapted from the Inference gym (Sountsov et al., 2020) we give model’s inference gym name in the parenthesis.

- Gaussian is 100-dimensional with condition number 100 and eigenvalues uniformly spaced in log.
- Banana (Banana) is a two-dimensional, banana-shaped target.
- Bimodal: A mixture of two Gaussians in 50 dimensions, such that

$$p(\mathbf{x}) = (1 - a)\mathcal{N}(\mathbf{x}|0, I) + a\mathcal{N}(\mathbf{x}|\mu, \sigma^2 I),$$

where $a = 0.25$, $\mu = (4, 0, \dots, 0)$ and $\sigma = 0.6$.

- Rosenbrock is a banana-shaped target in 36 dimensions. It is 18 copies of the Rosenbrock functions with $Q = 0.1$, see ().
- Cauchy is a product of 100 1D standard Cauchy distributions.
- Brownian Motion (BrownianMotionUnknownScalesMissingMiddleObservations) is a 32-dimensional hierarchical problem, where Brownian motion with unknown innovation noise is fitted to the noisy and partially missing data.
- Sparse logistic regression (GermanCreditNumericSparseLogisticRegression) is a 51-dimensional Bayesian hierarchical model, where logistic regression is used to model the approval of the credit based on the information about the applicant.
- Item Response theory (SyntheticItemResponseTheory) is a 501-dimensional hierarchical problem where students’ ability is inferred, given the test results.
- Stochastic Volatility is a 2429-dimensional hierarchical non-Gaussian random walk fit to the S&P500 returns data, adapted from numpyro (Phan et al., 2019b).
- Neal’s funnel (Neal, 2011) is a funnel shaped target with a hierarchical parameter $z_1 \sim \mathcal{N}(0, 3)$ that controls the variance of the other parameters $z_i \sim \mathcal{N}(0, e^{z_1/2})$ for $i = 2, 3, \dots, d$. We take $d = 20$.

	MAMS	MAMS ($\times 10$)	MAMS ($/ 10$)
Gaussian	3249	3162	3169
Banana	14,078	13,098	13,469
Rosenbrock	94,184	91,620	94,533
Brownian	13,528	13,759	13,973
German Credit	55,748	59,777	57,118
ItemResp	45,371	43,840	450,40

Table 2: Number of gradients calls needed to get the squared error on the worst second moment below 0.01. MAMS is compared with MAMS where the initial step size has been multiplied (third column) or divided (fourth column) by a factor of 10. This demonstrates that the performance of the automatic hyperparameter adaptation is not sensitive to the initialization.

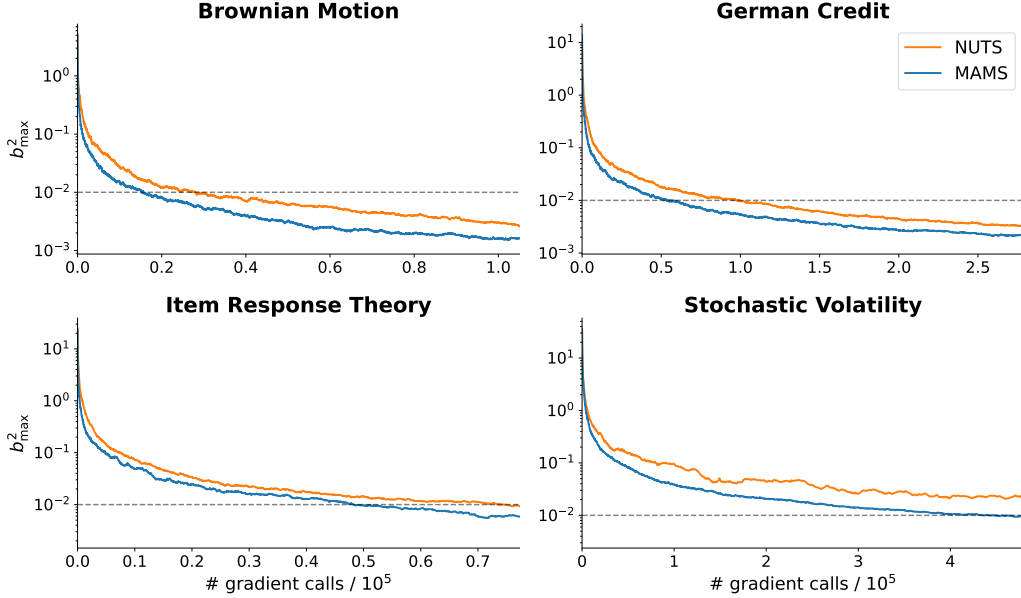


Figure 5: Sampling error, as measured by $b_{\max}^2$, as a function of computational cost, as measured by number of gradient calls. MAMS is compared with NUTS on several benchmark problems, demonstrating it consistently outperforms NUTS. Gradient calls to $b_{\max}^2 = 0.01$ dashed line is the number reported in Table 1.

Ground truth expectation values $\mathbb{E}[x^2]$ and $\text{Var}[x^2] = \mathbb{E}[(x^2 - \mathbb{E}[x^2])^2]$ are computed analytically for the Ill Conditioned Gaussian, by generating exact samples for Banana, Rosenbrock and Neal’s funnel and by very long NUTS runs for the other targets.

E.2 Sensitivity to the hyperparameter initialization

MAMS hyperparameters are automatically determined by the algorithm from Section 6. However, the algorithm is initialized with some stepsize $\epsilon_{\text{init}} = 0.2\sqrt{d}$ and trajectory length $L_{\text{init}} = \sqrt{d}$. MAMS performance could in principle be sensitive to this initialization. In Table 2 we show that this is not the case.

E.3 Convergence curves

Figure 5 shows the convergence of the second moments as a function of gradient calls used for NUTS and MAMS. $b_{\max}^2$ is computed as in Section 7. Sampling error for both samplers steadily decreases, but MAMS is consistently faster.

Model	Metric	MAMS	NUTS
Banana	b_{max}	0.42%	1.52%
Bimodal Gaussian	b_{max}	6.34%	2.07%
Brownian Motion	b_{max}	0.36%	0.22%
Cauchy	b_{avg}	5.37%	0.02%
German Credit	b_{max}	0.12%	0.09%
Item Response	b_{max}	1.05%	0.17%
Rosenbrock	b_{avg}	0.06%	0.02%
Standard Gaussian	b_{avg}	1.04%	0.13%
Stochastic Volatility	b_{max}	0.10%	0.09%

Table 3: Relative uncertainty associated with Table 1.

E.4 Uncertainty in Table 1

Table 3 shows the relative uncertainty in the results from Table 1. Uncertainty is calculated by bootstrap: for a given model, we produce a set of chains (usually 128), and calculate the bias b_{max} at each step of the chain. We then resample (with replacement) 100 times from this set, and compute our final metric (number of gradients to low bias) 100 times. We take the standard deviation of this list of length 100 to obtain an estimate of the error. This error, relative to the values in Table 1 is then reported in percent.

E.5 Computational architecture

The experiments were run on 128 CPU cores, where each core is a 2x AMD EPYC 7763 (Milan) CPU.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. The paper clearly introduces the Metropolis-Adjusted Microcanonical Sampler (MAMS), describing it as a novel approach that outperforms the No-U-Turn Sampler (NUTS) across multiple benchmarks. It further claims theoretical derivations for the Metropolis adjustment and automated tuning, both of which are thoroughly supported by the mathematical derivations and experimental results in the paper. The scope of the claims is aligned with the presented results, and the paper does not overstate its contributions.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Limitations of MAMS are very similar to that of NUTS as demonstrated in Experiments and discussed in the Conclusions.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: The paper provides a full set of assumptions and complete, rigorously derived proofs for all theoretical results, including the derivation of the Metropolis-Hastings acceptance probability for the microcanonical dynamics (Section 4) and the calculation of the Jacobian and divergence of the velocity update (Appendix). Each theorem and lemma is clearly numbered and referenced in the main text, and the proofs are presented in the appendix, ensuring clarity and completeness. Assumptions for each result are explicitly stated.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: The paper provides comprehensive details necessary to reproduce the main experimental results. It clearly outlines the algorithms (including pseudocode in Algorithm 1), the adaptation scheme, and the benchmarking methodology. The paper also specifies the hyperparameter settings and evaluation metrics. Additionally, the paper provides details about the benchmark problems, including dimensionality and problem structure (Section 7 and Appendix E). The implementation details, such as the use of BlackJax for NUTS and the code structure for MAMS, are also discussed, ensuring reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: MAMS is implemented in open source package BlackJax and comes with a tutorial on how to use it. The code for reproducing the experiments and implementations of NUTS and MALA are also provided as a GitHub repository.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: The paper specifies all the experimental details necessary to understand the results. It clearly describes the datasets used, including both synthetic and real-world benchmarks. Hyperparameters adaptation scheme is described in detail (Section 7: Experiments and Section 6: Adaptation). The choice of evaluation metrics is explained, and the setup for the baseline methods, including NUTS with BlackJax, is described. The experimental

protocol, including the number of chains, initialization, and the method for calculating performance metric, is also detailed. These specifications ensure that the experimental setting is fully transparent.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: we use a bootstrap method to calculate standard deviation of results, which are fully reported in Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: The paper provides information on the runtime of the algorithm on Stochastic volatility of the model in comparison to NUTS and states the computational architecture used. That being said, the computational architecture is not really relevant, as the algorithm here is tested on simple models with cheap to compute gradient evaluations, but it is intended to be used on more involved models. Number of gradient evaluations is therefore the relevant metric, not the runtime in our experiments, as we explain in the Experiments section 7.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research conducted in this paper fully conforms with the NeurIPS Code of Ethics. It adheres to the ethical principles, including transparency, integrity, and responsible use of computational resources. The paper does not involve any unethical data collection, misuse of computational power, or other ethical concerns.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: The paper is on foundational research and not tied to particular applications.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper is on foundational research and not tied to particular applications.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: The paper properly credits the creators and original owners of all assets used, including code, data, and models. The relevant assets are cited with appropriate references, and their licenses and terms of use are respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: MAMS implementation in BlackJax comes with a well-written documentation and a tutorial on how to use it.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.