

GIST: Gibbs self-tuning for locally adaptive Hamiltonian Monte Carlo

Nawaf Bou-Rabee* Bob Carpenter† Milo Marsden‡

April 24, 2024

Abstract

We present a novel and flexible framework for localized tuning of Hamiltonian Monte Carlo samplers by sampling the algorithm’s tuning parameters conditionally based on the position and momentum at each step. For adaptively sampling path lengths, we show that randomized Hamiltonian Monte Carlo, the No-U-Turn Sampler, and the Apogee-to-Apogee Path Sampler all fit within this unified framework as special cases. The framework is illustrated with a simple alternative to the No-U-Turn Sampler for locally adapting path lengths.

1 Introduction

Tuning the parameters of Markov chain Monte Carlo (MCMC) algorithms is critical to their performance, but notoriously difficult in practice. This challenge is particularly acute for Hamiltonian Monte Carlo (HMC), where the tuning of path length (integration time) [31, 15, 4, 34], step length (time discretization parameter) [2, 5, 6], and pre-conditioner (mass matrix) [25, 33, 52] frequently presents a complex trade-off between computational cost and mixing. The successful self-tuning of path length provided by the No-U-Turn Sampler (NUTS) has led to its widespread adoption as the default sampler in probabilistic programming languages [18, 48, 22, 24, 45].

In NUTS, the path length is adaptively sampled according to U-turn avoiding conditions, which roughly speaking stop the underlying Hamiltonian trajectory whenever it doubles back [31, 4]. The main idea is to numerically integrate Hamilton’s equations forward and backward in time until a U-turn occurs.¹ A point along this path is then randomly selected such that: (i) the resulting sampler has the correct invariant distribution; and (ii) points far from the starting point are more likely to be selected.

Can the algorithm’s other tuning parameters be similarly self-tuned? Motivated by the benefits of NUTS, this paper addresses this question by presenting a new framework for adaptively sampling

*Department of Mathematical Sciences, Rutgers University, nawaf.bourabee@rutgers.edu

†Center for Computational Mathematics, Flatiron Institute, bcarpenter@flatironinstitute.org

‡Department of Mathematics, Stanford University, mmarsden@stanford.edu

¹A U-turn must occur in a proper density, but in practice, a cap such as 1024 iterations is imposed.

HMC tuning parameters such as path length, step length, pre-conditioner, etc.² This framework is termed Gibbs self-tuning HMC or GIST for short. In the context of adaptively sampling path lengths, we demonstrate that GIST includes as special cases almost all locally adaptive HMC samplers in current use including randomized HMC [15, 9], NUTS [31, 4], and the Apogee-to-Apogee Path Sampler [20]. Owing to its simplicity and generality, GIST can be naturally extended to adapting the algorithm’s remaining tuning parameters including step size and mass matrix.

To be sure, the primary contribution of this paper is the GIST sampler, which admits a relatively simple proof of correctness. Moreover, we demonstrate the utility of the GIST sampling framework as a theoretical tool by unifying the proofs of correctness for existing locally adaptive HMC samplers and as a practical tool by developing a novel alternative for locally adapting path lengths.

The remainder of the paper is organized as follows. Section 2 presents the GIST sampling framework for self-tuning HMC samplers. In the context of adaptively sampling path lengths, Section 3 considers some fundamental special cases of GIST samplers including randomized HMC (Section 3.1), NUTS (Section 3.4), and the Apogee-to-Apogee Path Sampler (Section 3.5). Section 4 compares GIST samplers for path-length adaptation in the continuous-time context applied to a truncation of an infinite-dimensional Gaussian target measure. Section 5 considers path-length adaptation in the discrete-time context, and presents several concrete proposals for sampling the number of leapfrog steps based on the current position and momentum given a fixed step size and mass matrix. Section 6 describes a novel sampler for adapting path lengths along with experimental evaluation and an empirical comparison to NUTS.

2 A framework for self-tuning Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) is a class of MCMC methods for sampling absolutely continuous probability target distributions of the form

$$\mu(d\theta, d\rho) \propto e^{-H(\theta, \rho)} m(d\theta d\rho), \quad (1)$$

which have a density relative to Lebesgue measure m on phase space \mathbb{R}^{2d} , with position $\theta \in \mathbb{R}^d$ and momentum $\rho \in \mathbb{R}^d$. For simplicity, we assume the case of a separable Hamiltonian function with unit mass matrix, i.e.,

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

for a continuously differentiable potential energy function $U : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $\int e^{-U(\theta)} d\theta < \infty$. The non-normalized target density of interest is $e^{-U(\theta)}$ for model parameters θ .

A defining ingredient of any HMC algorithm is a reversible, volume-preserving map $F(\alpha) : \mathbb{R}^{2d} \rightarrow \mathbb{R}^{2d}$ where α encapsulates the algorithm tuning parameters. This map is typically obtained from discretizing Hamilton’s equations for the Hamiltonian function $H(\theta, \rho)$, i.e.,

$$\frac{d}{dt} \theta_t = \rho_t, \quad \frac{d}{dt} \rho_t = -\nabla U(\theta_t).$$

²The approach presented here could also be applied to Metropolis samplers other than HMC.

GIST(θ) 0

$\theta \in \mathbb{R}^d$ position 2
 $\rho \in \mathbb{R}^d$ momentum 2
 $\alpha \in \mathcal{A}$ algorithm tuning parameter 2

(INITIALIZE) 0
 $\theta_0 = \theta$ 12

(GIBBS) 0
 $\rho_0 \sim \text{normal}(0, \mathbf{I}_{d \times d})$ 0 (complete momentum refreshment) 13
 $\alpha \sim p(\cdot \mid \theta_0, \rho_0)$ 17 (sample tuning parameter) 14

(METROPOLIS-WITHIN-GIBBS) 0
 $(\theta^*, \rho^*) = F(\alpha)(\theta_0, \rho_0)$ 12 (compute proposal) 15
 $u \sim \text{uniform}([0, 1])$ 184
if $u < e^{-\Delta H(\theta_0, \rho_0)} p(\pi(\theta_0, \rho_0)(\alpha) \mid \mathcal{S} \circ F(\alpha)(\theta_0, \rho_0))$ 17
 $p(\alpha \mid \theta_0, \rho_0)$ 29
return θ^* 58 (accept) 0
else 18
return θ_0 58 (reject) 0

Figure 1: **GIST Sampling**. The GIST sampler differs from standard HMC in the sampling of the tuning parameter and subsequent adjustment of the acceptance probability. Here we use the shorthand $\Delta H(\theta, \rho) := H \circ F(\alpha)(\theta, \rho) - H(\theta, \rho)$ and $\mathbf{I}_{d \times d}$ denotes the $d \times d$ identity matrix. 170

Reversibility and volume-preservation ensure that the map $F(\alpha)$ is Metropolis adjustable [10]. Note that $\mathcal{S} \circ F(\alpha)$ is a volume-preserving involution where $\mathcal{S} : \mathbb{R}^{2d} \rightarrow \mathbb{R}^{2d}$ is the momentum flip map defined by $\mathcal{S}(\theta, \rho) = (\theta, -\rho)$ for all $(\theta, \rho) \in \mathbb{R}^{2d}$ [29]. We suppose that the tuning parameter α takes values in a set \mathcal{A} where $(\mathcal{A}, \mathcal{B}, \eta)$ is a measure space with σ -algebra \mathcal{B} and background measure η . 41

To “self tune” the parameter α , the state space \mathbb{R}^{2d} is enlarged to a product space $\mathbb{R}^{2d} \times \mathcal{A}$. On this enlarged space, an enlarged target measure is defined by specifying a conditional distribution of the tuning parameter α given the position and momentum (θ, ρ) , i.e., 25

$$\mu_e(d\theta, d\rho, d\alpha) \propto e^{-H(\theta, \rho)} p(\alpha \mid \theta, \rho) (m \otimes \eta)(d\theta d\rho d\alpha). \quad (2) \quad 0$$

Note that the (θ, ρ) -marginal of μ_e is the desired target measure μ . Explicit forms of the conditional distribution $p(\alpha \mid \theta, \rho)$ are specified in subsequent sections for the case of path length adaptivity. 197

For any position and momentum $(\theta, \rho) \in \mathbb{R}^{2d}$, let $\pi(\theta, \rho) : \mathcal{A} \rightarrow \mathcal{A}$ be a measurable map acting on the tuning parameter $\alpha \in \mathcal{A}$ such that 26

$$G : (\theta, \rho, \alpha) \mapsto (S \circ F(\alpha)(\theta, \rho), \pi(\theta, \rho)(\alpha)) \quad (3) \quad 38$$

is a $(m \otimes \eta)$ -preserving involution on $\mathbb{R}^{2d} \times \mathcal{A}$. This clearly holds when the map $\pi(\theta, \rho)$ is taken to be the identity map on \mathcal{A} .

In terms of this notation, Figure 1 describes a transition step of the “GIST” (Gibbs self-tuning HMC) sampler. The following theorem indicates that the GIST sampler is correct.

Theorem 1. *The GIST sampler in Figure 1 generates a Markov Chain $\theta_0, \theta_1, \dots$ on \mathbb{R}^d that is reversible with respect to $e^{-U(\theta)} m(d\theta)$.*

A detailed, self-contained proof of Theorem 1 is provided in Appendix A. By interpreting the GIST sampler as a Gibbs sampler, the proof shows that each update in Figure 1 leaves invariant the enlarged target (2). In particular, the accept/reject step is interpreted as a Metropolis-within-Gibbs step on the enlarged space $\mathbb{R}^{2d} \times \mathcal{A}$ with target measure defined in (2), proposal given by G defined in (3), and corresponding acceptance probability

$$a_e(\theta_0, \rho_0, \alpha) = 1 \wedge \left(\frac{e^{-\Delta H(\theta_0, \rho_0)} p(\pi(\theta_0, \rho_0)(\alpha) | \mathcal{S} \circ F(\alpha)(\theta_0, \rho_0))}{p(\alpha | \theta_0, \rho_0)} \right) \quad (4)$$

where we used the shorthand $\Delta H(\theta, \rho) := H \circ F(\alpha)(\theta, \rho) - H(\theta, \rho)$ for $(\theta, \rho, \alpha) \in \mathbb{R}^{2d} \times \mathcal{A}$. Note, this Metropolis-within-Gibbs step is correct because we assume that the map G on $\mathbb{R}^{2d} \times \mathcal{A}$ is a $(m \otimes \eta)$ -preserving involution; for a detailed justification see Lemma 2 in Appendix A.

Remark 1. Just as the Markov chain for standard HMC can be formulated over phase space variables $(\theta, \rho) \in \mathbb{R}^{2d}$, the Markov chain for GIST can also be formulated over triples $(\theta, \rho, \alpha) \in \mathbb{R}^{2d} \times \mathcal{A}$. However, in the case treated here of full momentum and tuning parameter refreshment, the sequence of θ values by itself forms a Markov chain and including the other variables is superfluous.

Remark 2. In addition to adaptively sampling HMC tuning parameters, the GIST sampler can also be applied to randomize the time integrator for the Hamiltonian flow, as in [13, 12]. In this case, the tuning parameter would specify a particular time integrator within a parametric family of time integrators that are each reversible and volume-preserving. In certain representative models, randomized time integrators have provably better complexity for Hamiltonian MCMC than the frequently used leapfrog integrator [49, 30, 17, 13, 16, 14].

3 Adaptively sampling path lengths

To make the GIST sampling framework more concrete and demonstrate its breadth, this section shows that several fundamental HMC samplers that locally adapt path lengths fit within this unified framework.

3.1 Sampling path lengths in randomized HMC

In randomized forms of HMC, the path length or the step length is generated randomly at each iteration independently of the current position and momentum [37]; using, e.g., an exponential distribution over path length [15, 9], an empirically learned distribution over path length [53], or a uniform distribution over path and step lengths [37, 43]. These can all be analyzed as instances of the GIST sampler given in Figure 1.

As an example, consider $F(\alpha)(\theta, \rho) = \varphi_\alpha(\theta, \rho)$ where we have introduced the exact Hamiltonian flow $\varphi_\alpha : \mathbb{R}^{2d} \rightarrow \mathbb{R}^{2d}$ at time $\alpha \geq 0$. Take $\mathcal{A} = [0, \infty)$ and η to be the Lebesgue measure on \mathbb{R} . Define $p(\alpha | \theta, \rho) = \lambda e^{-\lambda\alpha}$ where $\lambda > 0$; that is, α is an exponential random variable with parameter λ . In this case, the map G in (3) is given explicitly by $G : (\theta, \rho, \alpha) \mapsto (\mathcal{S} \circ F(\alpha)(\theta, \rho), \alpha)$, for which $a_e \equiv 1$ (the proposal is always accepted), and Figure 1 reduces to a draw from randomized HMC at the first jump time [15, 9].

The other randomized forms of HMC that generate randomly the tuning parameter independently of the current position and/or momentum can be analyzed in exactly the same way. In Section 4, in the setting of a truncation of an infinite-dimensional Gaussian measure, the performance of randomized HMC is compared with a GIST sampler that we describe next.

3.2 Adapting path lengths in exact HMC

Consider again $F(\alpha) = \varphi_\alpha$ where $\varphi_\alpha : \mathbb{R}^{2d} \rightarrow \mathbb{R}^{2d}$ is the exact Hamiltonian flow map at time $\alpha \geq 0$. Take $\mathcal{A} = [0, \infty)$ and η to be the Lebesgue measure on \mathbb{R} . Let $\tau(\theta, \rho) : \mathbb{R}^{2d} \rightarrow (0, \infty)$ be any measurable function. Define

$$p(\alpha | \theta, \rho) = \frac{1}{\tau(\theta, \rho)} \mathbb{1}_{[0, \tau(\theta, \rho)]}(\alpha).$$

That is, conditioned on the position and momentum (θ, ρ) , α is a uniform random variable over the interval $[0, \tau(\theta, \rho)]$. Here $\mathbb{1}_A$ denotes the standard indicator function of a set A . As a shorthand, let $\tau_1 = \tau(\theta_0, \rho_0)$ and $\tau_2 = \tau(\mathcal{S} \circ \varphi_\alpha(\theta_0, \rho_0))$, i.e., the function τ evaluated at the current state of the chain (θ_0, ρ_0) and the proposed state but with momentum reversed $\mathcal{S} \circ \varphi_\alpha(\theta_0, \rho_0)$. In this case, the map G in (3) is given explicitly by

$$G : (\theta, \rho, \alpha) \mapsto (\mathcal{S} \circ F(\alpha)(\theta, \rho), \alpha)$$

and the corresponding acceptance probability in (4) reduces to

$$a_e(\theta_0, \rho_0, \alpha) = 1 \wedge \left(\frac{\tau_1}{\tau_2} \mathbb{1}_{\tau_2 \geq \alpha} \right)$$

because $\Delta H(\theta_0, \rho_0) = 0$ for the exact Hamiltonian flow. The indicator in this Metropolis ratio indicates that $p(\alpha | \mathcal{S} \circ \varphi_\alpha(\theta_0, \rho_0)) \neq 0$.

Conditions avoiding U-turns in the exact Hamiltonian flow can be used to specify the function $\tau(\theta, \rho)$. There are several ways to characterize such U-turn conditions. For example, here is a condition based on when the angle between the initial velocity $\rho_0 \in \mathbb{R}^d$ and the velocity $\rho_t \in \mathbb{R}^d$ at time $t \geq 0$ first exceeds $\pi/2$

$$\tau(\theta_0, \rho_0) := \inf\{t > 0 : \rho_0 \cdot \rho_t \leq 0\}, \quad (5)$$

where we have introduced $(\theta_t, \rho_t) := \varphi_t(\theta_0, \rho_0)$ for $t \geq 0$. Another condition is based on when the squared distance $\Gamma(t) := |\theta_0 - \theta_t|^2$ between the initial configuration $\theta_0 \in \mathbb{R}^d$ and the configuration $\theta_t \in \mathbb{R}^d$ at time $t \geq 0$ first decreases

$$\tau(\theta_0, \rho_0) := \inf\{t > 0 : \Gamma'(t) < 0\}. \quad (6)$$

These continuous-time U-turn conditions have discrete-time analogs, which are discussed in the next example and Section 5.

3.3 Adapting path lengths in numerical HMC 163

Fix $h > 0$. Let $\Phi_h : \mathbb{R}^{2d} \rightarrow \mathbb{R}^{2d}$ denote one step of the leapfrog integrator with step length h . Consider $F(\alpha) = \Phi_h^\alpha$ where α is the number of leapfrog steps. Take $\mathcal{A} = \mathbb{N}$ and let η be the counting measure. Similarly to the above, let $\tau : \mathbb{R}^{2d} \rightarrow \mathbb{N}$ be a measurable function and define

$$p(\alpha \mid \theta, \rho) = \frac{1}{\tau(\theta, \rho)} \mathbb{1}_{\{0, \dots, \tau(\theta, \rho) - 1\}}(\alpha).$$

Similarly to the above, the map G in (3) is given explicitly by

$$G : (\theta, \rho, \alpha) \mapsto (\mathcal{S} \circ F(\alpha)(\theta, \rho), \alpha),$$

and the corresponding acceptance probability in (4) simplifies to

$$a_e(\theta_0, \rho_0, \alpha) = 1 \wedge \left(e^{-\Delta H(\theta_0, \rho_0)} \frac{\tau_1}{\tau_2} \mathbb{1}_{\{\tau_2 \geq \alpha\}} \right)$$

where $\tau_1 = \tau(\theta_0, \rho_0)$ and $\tau_2 = \tau(\mathcal{S} \circ \Phi_h^\alpha(\theta_0, \rho_0))$.

3.4 Adapting path lengths in NUTS 89

Here we show that NUTS is a special case of the GIST sampler presented in Figure 1.

In order to write NUTS as a GIST sampler, we first explain NUTS more precisely following the terminology and framework introduced in [4, 23]. Given the current position and momentum $(\theta_0, \rho_0) \in \mathbb{R}^{2d}$, the idea behind NUTS is to first randomly generate a finite subset of integers $J \subset \mathbb{Z}$, and in turn, sample the next state of the chain from among the leapfrog iterates $\{\Phi_h^i(\theta_0, \rho_0)\}_{i \in J}$. Following the notation of [23], let $P_h(J \mid \theta_0, \rho_0)$ be the probability of selecting J and let $Q_h(i \mid J, \theta_0, \rho_0)$ be the probability of selecting the i -th leapfrog iterate $\Phi_h^i(\theta_0, \rho_0)$. In [23], these are referred to as the “orbit” and “index” selection kernels, respectively.

More precisely, the random procedure to generate the set J can be described by induction. First, take $J_0 = \{0\}$. Then, given J_k , we first check whether the leapfrog iterates $\{\Phi_h^i(\theta_0, \rho_0)\}_{i \in J_k}$ satisfy the U-turn condition:

$$\rho_+ \cdot (\theta_+ - \theta_-) < 0 \quad \text{and} \quad \rho_- \cdot (\theta_- - \theta_+) < 0$$

where $(\theta_+, \rho_+) = \Phi_h^{\max J_k}(\theta_0, \rho_0)$ and $(\theta_-, \rho_-) = \Phi_h^{\min J_k}(\theta_0, \rho_0)$. If this condition is met, we output $J = J_k$. If instead no U-turn is encountered, then sample I_{k+1} uniformly from $\{J_k + |J_k|, J_k - |J_k|\}$. This set is the next proposed extension to the set J_k .

This proposed extension will either be accepted, yielding $J_{k+1} = J_k \cup I_{k+1}$, or rejected, returning $J = J_k$. To ensure the correctness and efficiency of the algorithm the following symmetry condition is imposed: $P_h(J \mid \theta_0, \rho_0) = P_h(J - i \mid \Phi_h^i(\theta_0, \rho_0))$ for all $i \in J$. This condition ensures that the probability of obtaining a particular set of leapfrog iterates $\{\Phi_h^i(\theta_0, \rho_0)\}_{i \in J}$ is uniform with respect to the starting point $\Phi_h^i(\theta_0, \rho_0)$ within this set for all $i \in J$.

Imposing this constraint will require that some proposed extensions I_{k+1} are rejected. For every $i \in J_k$, corresponds an alternate sequence of extensions I'_1, \dots, I'_k which would have produced the set $J_k - i$ had the orbit generating procedure instead been started at $\Phi_h^i(\theta_0, \rho_0)$. To enforce

$P_h(J \mid \theta_0, \rho_0) = P_h(J - i \mid \Phi_h^i(\theta_0, \rho_0))$, it is sufficient to require that J_k satisfies the No-Intermediate-U-Turn property. This property states that at no stage in this alternate sequence of extensions I'_1, \dots, I'_k would the U-turn condition have been triggered.

To ensure that every J_k satisfies the No-Intermediate-U-Turn property, we simply need to reject any proposal I_{k+1} where $J_{k+1} = J_k \cup I_{k+1}$ does not satisfy the No-Intermediate-U-Turn property. While naively this appears to involve checking over exponentially many sequences of extensions, this step can be implemented efficiently. Indeed, over different choices of $i \in J_k$ many of the alternate sequences will refer to the same set of leapfrog iterates. Since the U-turn condition depends only on the leapfrog iterates, we therefore need only check that all possible intermediate sets of leapfrog iterates have no U-turns. As the number of possible intermediate sets of leapfrog iterates is relatively small, this observation greatly reduces the complexity of checking the No-Intermediate-U-Turn property. See Remark 3 for more details about the computational complexity of the orbit generation procedure.

Having generated J using this randomized orbit selection procedure, NUTS then generates $i \in J$ according to the index selection kernel $Q_h(i \mid \theta_0, \rho_0, J)$ and returns $\Phi_h^i(\theta_0, \rho_0)$ as the next state. There are two index selection kernels commonly used in practice. The first is given by the Boltzmann weight of the corresponding leapfrog iterate, i.e., $Q_h(i \mid \theta_0, \rho_0, J) \propto e^{-H(\Phi_h^i(\theta_0, \rho_0))}$. The second is biased progressive sampling, as detailed in [4].

We are now in position to describe NUTS as a GIST sampler. Let $\mathcal{A} = \mathcal{P}_F \times \mathbb{Z}$ where $\mathcal{P}_F = \{J \subset \mathbb{Z} \mid |J| < \infty\}$, let η be the counting measure on $\mathcal{P}_F \times \mathbb{Z}$, and write the tuning parameter in components as $\alpha = (J, i) \in \mathcal{A}$. For all $(\theta, \rho, \alpha) \in \mathbb{R}^{2d} \times \mathcal{A}$, define

$$p(J, i \mid \theta, \rho) = P_h(J \mid \theta, \rho) Q_h(i \mid \theta, \rho, J).$$

As in Section 3.3, let $F(\alpha)(\theta, \rho) = \Phi_h^i(\theta, \rho)$. The map G given in (3) can be written explicitly as

$$G(\theta, \rho, J, i) = (S \circ \Phi_h^i(\theta, \rho), -(J - i), i).$$

The resulting GIST sampler in Figure 1 corresponds to NUTS. Notably, the Metropolis-within-Gibbs step for NUTS is always accepted. Indeed, since $\{\Phi_h^i(S(\theta, \rho))\}_{i \in -J} = \{\Phi_h^{-i}(S(\theta, \rho))\}_{i \in J} = \{S(\Phi_h^i(\theta, \rho))\}_{i \in J}$ a short coupling construction shows $P_h(-J \mid S(\theta, \rho)) = P_h(J \mid \theta, \rho)$. Combining this with $P_h(J - i \mid \Phi_h^i(\theta, \rho)) = P_h(J \mid \theta, \rho)$ gives $P_h(-(J - i) \mid S \circ \Phi_h^i(\theta, \rho)) = P_h(J \mid \theta, \rho)$. Since the index selection kernels described above satisfy the following detailed balance condition

$$e^{-H(\Phi_h^i(\theta_0, \rho_0))} Q_h(i \mid S \circ \Phi_h^i(\theta_0, \rho_0), -(J - i)) = e^{-H(\theta_0, \rho_0)} Q_h(i \mid \theta_0, \rho_0, J),$$

it follows that the acceptance probability in Figure 1 is always equal to 1. For the index selection kernel based on the Boltzmann weights this detailed balance condition can be seen by a direct computation, and for biased progressive sampling this property follows from Proposition 6 in [23].

Remark 3 (Computational Cost of Orbit Selection). As noted above, one can efficiently check whether J_{k+1} satisfies the No-Intermediate-U-Turn condition by checking whether all possible intermediate sets of leapfrog iterates are free of U-turns. If we define random variables $\beta_1, \dots, \beta_{k+1}$ as $\beta_l = \mathbb{1}_{I_l = J_l - |J_l|}$ and $B = \sum_{l=1}^{k+1} 2^{l-1} \beta_l$ then one can prove by induction that $J_{k+1} = \{-B, \dots, 2^k - 1 - B\} = [2^k] - (B + 1)$. The possible sets of intermediate leapfrog iterates corresponding to

J_{k+1} then correspond exactly to the sets $\{\Phi_h^i(\theta, \rho)\}_{i \in [m \cdot 2^l : (m+1)2^l - 1] - (B+1)}$, for $l = 1, \dots, k$ and $m = 0, 1, \dots, 2^{k+1-l} - 1$, where $[m : n] = \{m, m+1, \dots, n-1, n\}$. 55

Since the U-turn condition involves only the endpoints of the set of leapfrog iterates, the cost of checking whether or not a given set of leapfrog iterates satisfies the U-turn condition is constant in the size of the set. Therefore, when generating an orbit starting from (θ_0, ρ_0) the total cost of checking whether $J_{k+1} = I_{k+1} \cup J_k$ satisfies the No-Intermediate-U-Turn condition is upper bounded by the number of sets we need to check. 56

Since each choice of l above corresponds to 2^{k-l} sets the total cost for checking that J_{k+1} satisfies the No-Intermediate-U-Turn property is upper bounded by $\sum_{l=1}^k 2^{k-l} \leq 2^{k+1}$. Consequently, returning a set J with $|J| = M = 2^k$ has computational cost $O(M)$. In practice, by induction we need only check those intermediate sets of leapfrog iterates which correspond to subsets of I_{k+1} which gives further speedup. 57

Compare this to the simpler strategy of extending J_k one point at a time by sampling I_{k+1} uniformly from $\{\min(J_k) - 1, \max(J_k) + 1\}$. For such a strategy, the cost of appending I_{k+1} is again of order $|J_k|$ and therefore to return a set J with $|J| = M$ has computational cost $O(M^2)$. Reducing the cost of producing large orbits motivates the strategy of having the orbits J_k double in size with each iteration. For more details on the computational cost of the orbit selection procedure in NUTS as well as other aspects of the implementation, see Appendix A in [4]. 58

3.5 Adapting path lengths in the Apogee-to-Apogee Path Sampler 5

The Apogee-to-Apogee Path Sampler (AAPS) introduced in [20] is a special case of the GIST sampler presented in Figure 1. The idea behind AAPS is to replace the No-U-Turn condition for adapting the number of leapfrog steps with an alternative condition based on partitioning the discrete path $\{\Phi_h^i(\theta, \rho)\}_{i \in \mathbb{Z}}$ into ‘segments’ $\{S_i(\theta, \rho)\}_{i \in \mathbb{Z}}$. The i -th segment $S_i(\theta, \rho)$ consists of points along the discrete path originating from the current point (θ, ρ) that lie between consecutive local maxima (termed “apogees”) of the underlying potential energy $U(x)$. 60

In order to write AAPS as a GIST sampler, we first explain AAPS more precisely starting with the partitioning mentioned above. Our description follows [20]. Proceeding both forwards and backwards in time from (θ, ρ) via leapfrog steps, consecutive outputs belong to the same segment if the potential U is either instantaneously increasing at both points or instantaneously decreasing at both points. Otherwise, an apogee occurs between this pair of points, and hence, a new segment is defined. Iterating this procedure produces a two-sided sequence of segments: 61

$$\dots, S_{-1}(\theta, \rho), S_0(\theta, \rho), S_1(\theta, \rho), \dots \quad 176$$

where $S_0(\theta, \rho)$ is the initial segment. Following the notation of [20], let $S_{\#}((\theta, \rho), (\theta', \rho'))$ be the index of the segment started at (θ, ρ) containing the leapfrog iterate (θ', ρ') . For instance, $S_{\#}((\theta, \rho), (\theta, \rho)) = 0$ since the starting point is always in the initial segment $S_0(\theta, \rho)$ and $S_{\#}((\theta, \rho), (\theta', \rho')) = k$ for every $(\theta', \rho') \in S_k(\theta, \rho)$ and $k \in \mathbb{Z}$. Additionally, let $S_{a:b}(\theta, \rho) = \bigcup_{i=a, \dots, b} S_i(\theta, \rho)$ for $a, b \in \mathbb{Z}$ such that $a \leq b$. 62

A positive integer K and a user-defined weight function $w : \mathbb{R}^{4d} \rightarrow [0, \infty)$ are required in AAPS. The latter assigns weights to points within selected sets of segments. From the current point (θ, ρ) , 63

AAPS uniformly randomly chooses among the $K + 1$ sets of segments: $S_{0:K}, S_{-1:K-1}, \dots, S_{-K:0}$. These sets of segments are all possible unions of $K + 1$ consecutive segments whose union contains the current point. Having selected such a set of segments $S_{-c:K-c}$, AAPS then randomly selects a leapfrog iterate $(\theta', \rho') \in S_{-c:K-c}$ as a proposal with weight $w((\theta, \rho), (\theta', \rho'))$. Finally, this proposal is Metropolized with Metropolis-Hastings acceptance probability

$$1 \wedge \frac{e^{-\Delta H(\theta, \rho)} w((\theta', \rho'), (\theta, \rho)) \sum_{(\tilde{\theta}, \tilde{\rho})} w((\theta, \rho), (\tilde{\theta}, \tilde{\rho}))}{w((\theta, \rho), (\theta', \rho')) \sum_{(\tilde{\theta}, \tilde{\rho})} w((\theta', \rho'), (\tilde{\theta}, \tilde{\rho}))} \quad (7)$$

The weight function can be chosen to preferentially select proposals according to various desiderata. For instance, one might take $w((\theta, \rho), (\theta', \rho')) = e^{-H(\theta', \rho')}$ to select points which will always be accepted under the Metropolis-Hastings step, or $w((\theta, \rho), (\theta', \rho')) = e^{-H(\theta', \rho')} \|\theta - \theta'\|^2$ to bias toward proposals which are farther from the current point. In addition to these, several other choices of weight functions are suggested in [20].

We are now in position to write AAPS as a GIST sampler. Let $\mathcal{A} = \mathbb{Z}^2$, let η be the counting measure on \mathcal{A} , and write the algorithm tuning parameter in components as $\alpha = (c, i) \in \mathcal{A}$. The component c represents the choice of set of segments $S_{-c:K-c}$ while the component i represents the choice of leapfrog iterate $\Phi_h^i(\theta, \rho)$ within this set of segments. Define

$$p(c, i \mid \theta, \rho) = \frac{1}{K+1} \frac{w((\theta, \rho), \Phi_h^i(\theta, \rho)) \mathbb{1}_{S_{-c:K-c}(\theta, \rho)}(\Phi_h^i(\theta, \rho))}{\sum_j w((\theta, \rho), \Phi_h^j(\theta, \rho)) \mathbb{1}_{S_{-c:K-c}(\theta, \rho)}(\Phi_h^j(\theta, \rho))}$$

where $c \in \{0, \dots, K\}$. The map G in (3) is given explicitly by

$$G : (\theta, \rho, c, i) \mapsto (\mathcal{S} \circ \Phi_h^i(\theta, \rho), c + S_{\#}((\theta, \rho), \mathcal{S} \circ \Phi_h^i(\theta, \rho)), i),$$

and the corresponding acceptance probability in (4) simplifies to (7). Thus, the GIST sampler in Figure 1 with the above specification of the conditional distribution of the algorithm tuning parameter $\alpha = (c, i)$ given the state (θ, ρ) correspond to AAPS.

4 Truncation of an infinite-dimensional Gaussian

This section demonstrates that the exact self-tuned HMC algorithm described in Section 3.2 can achieve a similar mean-squared jump distance (MSJD) as randomized HMC from Section 3.1 on a target distribution that can be interpreted as a truncation of an infinite-dimensional Gaussian measure [3, 11]. This is a worst-case example since the corresponding Hamiltonian dynamics is highly oscillatory [44]. More precisely, the target distribution is a d -dimensional centered normal distribution with covariance matrix given by

$$\Sigma = \text{diag}(\sigma_1, \dots, \sigma_d), \quad \text{where} \quad \sigma_i = \begin{cases} 1 & \text{for } i \in \{1, \dots, d\} \end{cases} \quad (8)$$

This example has traditionally been used to illustrate the importance of path length randomization to avoid slow mixing due to periodicities or near-periodicities in the underlying Hamiltonian dynamics [43, 15, 9]. Here we use this worst-case example to demonstrate the efficacy of the self-tuned HMC algorithms based on the U-turn conditions defined by (5) and (6).

For this model, the exact solution to the corresponding Hamiltonian dynamics (θ_t, ρ_t) at time $t \geq 0$ from initial condition (θ_0, ρ_0) is given in components by

$$\theta_t^i = \cos\left(\frac{t}{\sigma_i}\right) \theta_0^i + \sigma_i \sin\left(\frac{t}{\sigma_i}\right) \rho_0^i \quad \text{and} \quad \rho_t^i = -\frac{1}{\sigma_i} \sin\left(\frac{t}{\sigma_i}\right) \theta_0^i + \cos\left(\frac{t}{\sigma_i}\right) \rho_0^i$$

for $i \in \{1, \dots, d\}$. Using these solutions, the evaluation of the U-turn path lengths in (5) or (6) reduces to finding the first positive root of a scalar function of time, which itself crucially relies on a good initialization.

To be precise, we provide a complete description of the initialization procedure for the case of the U-turn path length given in (5); a similar initialization is used for (6) and therefore omitted. Our goal is to find the first positive root of f , i.e.,

$$\tau = \inf\{t \geq 0 : f(t) = 0\}.$$

where we have defined the following function of time

$$f(t) = \sum_{i=1}^d f_i(t),$$

where

$$f_i(t) = -\frac{1}{\sigma_i} \sin\left(\frac{t}{\sigma_i}\right) \theta_0^i \rho_0^i + \cos\left(\frac{t}{\sigma_i}\right) (\rho_0^i)^2.$$

The first positive root of f_i can be computed analytically, i.e.,

$$\tau_i = \arctan\left(\frac{\rho_0^i}{\theta_0^i}\right) + k^* \pi$$

where

$$k^* = \min \left\{ k \in \mathbb{Z} : \arctan\left(\frac{\rho_0^i}{\theta_0^i}\right) + k\pi \geq 0 \right\}$$

The mean of $\{\tau_i\}_{i=1}^d$ is used to seed the root solver for finding τ . Using the above root finding procedure to evaluate τ per transition step, the numerically estimated mean-acceptance probability (mean a_e), mean-squared jump distance (MSJD), and mean path length (mean τ) for $d = 1000$ and using 10^5 transition steps is summarized in the following table.

	mean a_e	MSD	mean τ
randomized HMC	100%	429.81	1.00
exact self-tuned HMC (5)	97.4%	174.85	0.44
exact self-tuned HMC (6)	94.4%	573.15	1.16

Randomized HMC is operated using the optimal choice of mean path length for maximizing expected square jump distance in this example, which corresponds to the standard deviation of the least constrained coordinate [15, Section 4]. The self-tuned HMC algorithm based on (6) slightly outperformed exact randomized HMC in terms of MSD, while the one based on (5) performed worse. Remarkably, neither degenerated in high dimension. This high-dimensional example demonstrates both the efficacy of the U-turn conditions and the leniency of the corresponding acceptance probability in the self-tuned HMC algorithm.

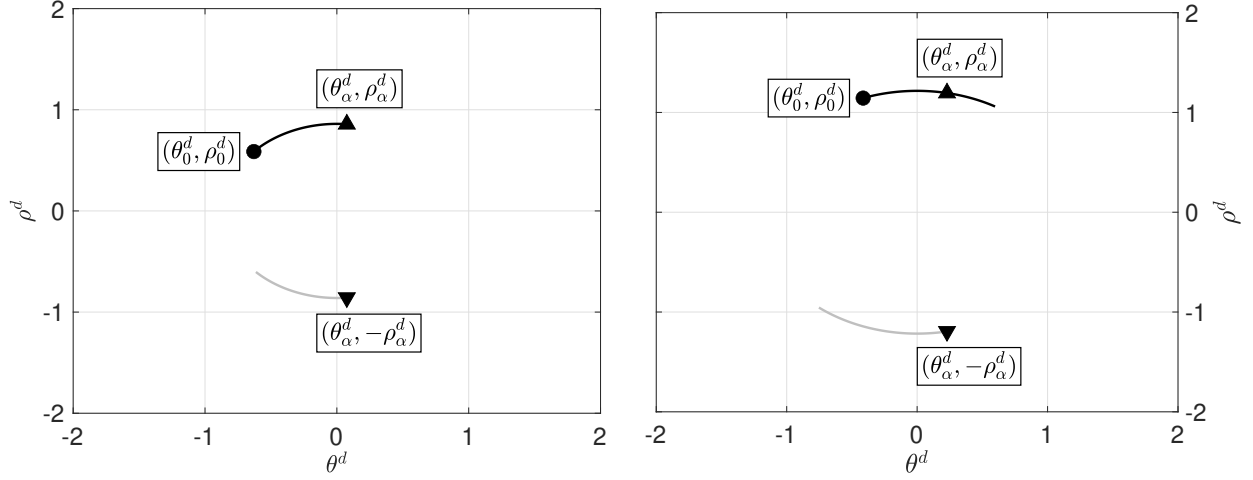


Figure 2: **Self-Tuning Path Lengths for Exact HMC.** This figure plots forward and backward trajectories of the GIST sampler using (5) in the phase space of the least constrained coordinate of the d -dimensional Gaussian measure with covariance matrix given in (8) with $d = 10^3$. The left panel illustrates a rejected proposal (due to $\alpha > \tau_2$) and the right panel illustrates an accepted proposal.

Remark 4. These numerical findings are of independent interest, since they motivate using GIST samplers based on U-turn avoiding conditions to sample from perturbed Gaussian measures on Hilbert spaces. This class of target measures is relevant to several important applications including Path Integral Molecular Dynamics [19, 39, 28, 36], Transition Path Sampling [47, 46, 7, 40], and Bayesian statistical inverse problems for Hilbert spaces [32, 50, 21, 8]. Since the corresponding Hamiltonian dynamics is potentially highly oscillatory in high modes [44], in actual numerical implementations, preconditioning [10] or strongly stable integrators [35] are necessary to be able to choose the step size independently of the dimension.

5 Path length sampling to avoid U-turns

We now turn to an extended example of a novel U-turn avoiding sampler based on Gibbs self tuning. We will focus on locally adapting the number of steps. Thus for each HMC step, the self-tuned HMC algorithm will generate the number of steps probabilistically according to a distribution $p(L \mid \theta, \rho, \epsilon, \Sigma)$. We consider several such distributions based on the no U-turn condition in the next section and evaluate them in the following sections.

Figure 3 provides pseudocode for a general algorithm adapting the number of leapfrog steps. The ratio of the proposal joint density to the starting point joint density can be factored as

$$\frac{p(\theta^*, \rho^*, L^* \mid \epsilon, \Sigma)}{p(\theta^{(0)}, \rho^{(0)}, L^* \mid \epsilon, \Sigma)} = \frac{p(\theta^*)}{p(\theta^{(0)})} \cdot \frac{p(\rho^* \mid \Sigma)}{p(\rho^{(0)} \mid \Sigma)} \cdot \frac{p(L^* \mid \theta^*, \rho^*, \epsilon, \Sigma)}{p(L^* \mid \theta^{(0)}, \rho^{(0)}, \epsilon, \Sigma)},$$

where $p(\theta) = p(\theta \mid y)$ is the target probability density function, $p(\rho \mid \Sigma) = \text{normal}(0, \Sigma)$ is the momentum probability density function, and $p(L \mid \theta, \rho, \epsilon, \Sigma)$ is the conditional step size probability mass function.

GIST ($\theta, \rho, \epsilon, \Sigma$)	0	
$\theta \in \mathbb{R}^d$	initial position	105
$\rho \in \mathbb{R}^d$	initial momentum (unused—discarded in Gibbs update)	106
$\epsilon \in (0, \infty)$	step size	107
$\Sigma \in \mathbb{R}^{d \times d}$	symmetric, positive definite mass matrix	108
$p(\theta)$	target density (log density evaluation & gradient)	109
$p(L \mid \theta, \rho, \epsilon, \Sigma)$	conditional steps distribution (sampler & log density evaluation)	109
<hr/>		
(INITIALIZE)	0	
$\theta^{(0)} = \theta$	0	
<hr/>		
(GIBBS)	0	
$\rho^{(0)} \sim \text{normal}(0, \Sigma)$	0	(Gibbs sample momentum) 111
$L \sim p(L \mid \theta^{(0)}, \rho^{(0)}, \epsilon, \Sigma)$	17	(Gibbs sample number of steps) 112
<hr/>		
(METROPOLIS-WITHIN-GIBBS)	0	
for ℓ from 0 to $L - 1$ (inclusive)	113	(L leapfrog steps) 113
$\rho^{(\ell+1/2)} = \rho^{(\ell)} + \frac{\epsilon}{2} \cdot \nabla \log p(\theta^{(\ell)})$	21	(half step momentum) 3
$\theta^{(\ell+1)} = \theta^{(\ell)} + \epsilon \cdot \Sigma^{-1} \cdot \rho^{(\ell+1/2)}$	21	(full step position) 3
$\rho^{(\ell+1)} = \rho^{(\ell+1/2)} + \frac{\epsilon}{2} \cdot \nabla \log p(\theta^{(\ell+1)})$	21	(half step momentum) 3
$\theta^*, \rho^* = \theta^{(L)}, -\rho^{(L)}$	12	(proposal flips momentum) 114
$u \sim \text{uniform}([0, 1])$	184	(sample acceptance probability) 115
if $u < \frac{p(\theta^*, \rho^*, L \mid \epsilon, \Sigma)}{p(\theta^{(0)}, \rho^{(0)}, L \mid \epsilon, \Sigma)}$	16	(Metropolis accept condition) 0
return θ^*, ρ^*, L	58	(accept) 0
else	18	
return $\theta^{(0)}, \rho^{(0)}, L$	58	(reject) 0

Figure 3: *GIST sampling for path-length self-tuning.* The GIST sampler for path-length self-tuning differs from standard HMC in sampling the number of steps each iteration and then adjusting the acceptance probability to ensure detailed balance. Note, the Gibbs steps for refreshment of both momentum and number of steps are exact draws from the corresponding conditional distributions.

5.1 Step distributions avoiding U-turns 39

To make our sampler concrete, a specific distribution over the number of steps must be defined. We evaluate a few related choices, all of which are motivated by the observation driving the No U-Turn Sampler (NUTS) [31], which is that it's wasteful to evaluate the Hamiltonian dynamics beyond the point at which the trajectory has made a U-turn and is heading back toward the starting position. Figure 4 illustrates the U-turn condition, which is made precise in Equation (9).

Let $U(\theta, \rho, \epsilon, \Sigma)$ be the maximum number of leapfrog steps that can be taken before a U-turn, starting from (θ, ρ) and using step size ϵ and mass matrix Σ . The point just before a U-turn can be

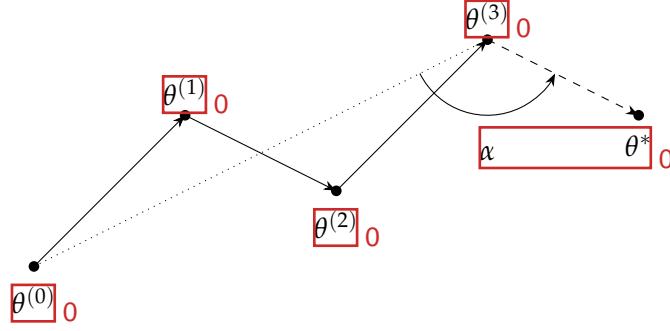


Figure 4: **U-turn condition.** A Hamiltonian trajectory of positions (not momenta) in two dimensions, consisting of three leapfrog steps plus a potential fourth step. The dotted line connects the initial position $\theta^{(0)}$ to the current position $\theta^{(3)}$. The dashed line connects the current position to the next potential position θ^* and runs in the direction of the current momentum $\rho^{(3)}$. The trajectory is extended one step to θ^* if the next step moves away from the initial position, which requires the absolute value of the angle α between the dotted line and the dashed line to be greater than 90° , which arises when $(\theta^{(3)} - \theta^{(0)})^\top \cdot (\theta^* - \theta^{(3)}) > 0$, or equivalently, when $(\theta^{(3)} - \theta^{(0)}) \cdot \rho^{(3)} > 0$.

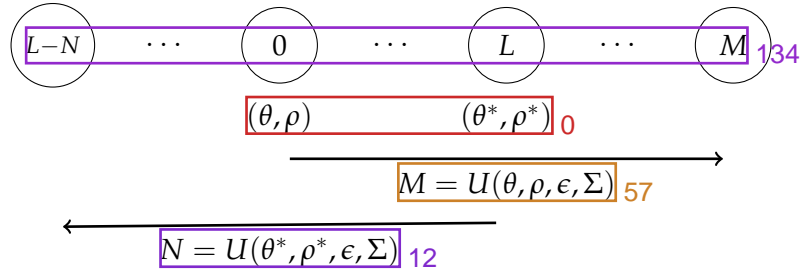


Figure 5: **Self-tuned steps proposal.** Starting from the initial position and momentum (θ, ρ) , the algorithm makes M forward leapfrog steps until a U-turn. It then Gibbs samples a number of steps L between 0 and M for which the leapfrog integrator produces the proposal (θ^*, ρ^*) . Then N backward leapfrog steps are taken until a U-turn. If $L - N > 0$, the proposal is rejected. The algorithm takes $M + N - L$ unique leapfrog steps.

defined for the discrete steps of the leapfrog integrator by

$$U(\theta^{(0)}, \rho^{(0)}, \epsilon, \Sigma) = \arg \min_{n \in \mathbb{N}} \left(\left(\theta^{(n)} - \theta^{(0)} \right)^\top \cdot \rho^{(n)} < 0 \right) \quad (9)$$

where $\theta^{(n)}, \rho^{(n)}$ is the location in phase space after n leapfrog steps from $\theta^{(0)}, \rho^{(0)}$ with step size ϵ and positive-definite mass matrix Σ .

Figure 5 shows a single step of the algorithm. The number of steps is sampled conditionally based on the number of steps possible before a U-turn. To preserve detailed balance, the trajectory from the selected point backward in time is evaluated until it makes a U-turn and the probability of selecting the initial state (i.e., the same number of steps) is used to balance the selection. This requires sampling zero or more steps backward in time before the initial position.

5.2 Conditional distribution of steps 123

In this section, we consider a few closely related approaches to generating the number of steps given the number of steps to a U-turn. 130

5.2.1 Steps generated uniformly 34

The most obvious choice for a conditional distribution over the number of steps is uniform between 1 and the number of steps before a U-turn. 131

$$p(L \mid \theta, \rho, \epsilon, \Sigma) = \text{uniform}(L \mid 1, U(\theta, \rho, \epsilon, \Sigma)), \quad (10) \quad 132 \quad 0$$

where the bounds are read inclusively. 133

5.2.2 Steps generated uniformly from later states 46

One of the techniques NUTS uses to take such long jumps on average is to bias the selection of a candidate toward the end of the Hamiltonian trajectory [31]. To create a simple approximation to this, the uniform distribution can be restricted to the latter part of the trajectory by changing the lower bound from 1 to something greater. Any number between 0 and $M = U(\theta, \rho, \epsilon, \Sigma)$ is valid as a lower bound. The choice of 1 corresponds to the uniform distribution of the previous section. We will also evaluate lower bounds of $\lfloor \frac{M}{2} \rfloor$ and $\lfloor \frac{3}{4} M \rfloor$, where $\lfloor x \rfloor$ is the floor of x , which is largest integer less than or equal to x . With smaller intervals, the proposed trajectory lengths will be longer, but the reversibility balance condition will reduce the acceptance probability. 134

5.2.3 Binomial step generation 49

Non-uniform distributions may also be used. For example, a binomial distribution for the number of steps could be defined for a fixed $\psi \in (0, 1)$ as 135

$$p(L \mid \theta, \rho, \epsilon, \Sigma) = \text{binomial}(L \mid \psi, U(\theta, \rho, \epsilon, \Sigma)). \quad (11) \quad 17 \quad 0$$

The expected value of L starting from (θ, ρ) is thus $\psi \cdot U(\theta, \rho, \epsilon, \Sigma)$. We evaluated the binomial method and it wound up generating proposals that were too concentrated and thus hard to balance and accept, so we do not include results here. We did not evaluate a more dispersed beta-binomial distribution due to the cost of the beta and gamma functions required for normalization. 137

6 Empirical evaluation 7

In this section, we evaluate the performance of our various proposals for step size adaptation and compare their performance to the state-of-the-art No U-Turn Sampler as coded in Stan [31, 4]. 138

6.1 Models evaluated 129

The models considered are multivariate standard normal, ill-conditioned multivariate normal, correlated multivariate normal, eight schools meta-analysis, item-response theory 2 parameter 139

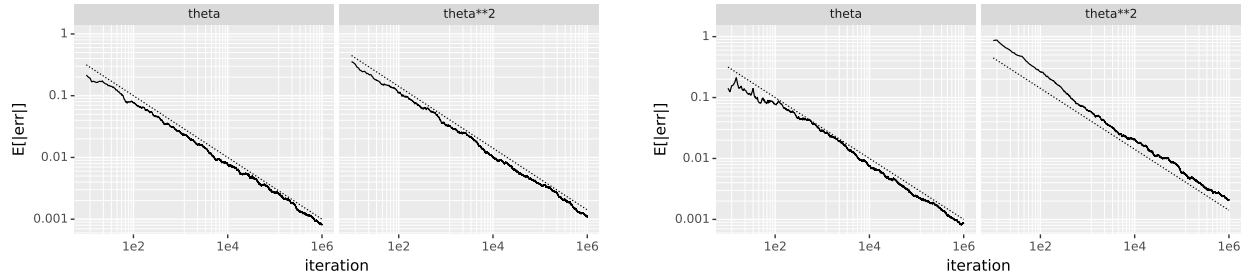


Figure 6: **Learning curve validation.** The absolute error in parameter and parameter squared estimates for i.i.d. draws (left) and the uniform self-tuning algorithm (right). In each plot, the left half shows error for parameter estimates and the right half shows error for squared parameter estimates (averaged over the 100 identical dimensions). For reference, the dotted line is the standard error derived from independent draws, which at iteration n is $1/\sqrt{n}$ for parameter estimates and $\sqrt{2}/\sqrt{n}$ for squared parameter estimates.

logistic model, mixed effects Poisson regression, normal mixture, hidden Markov model, one-compartment pharmacokinetic / pharmacodynamic model (PK/PD), Lotka-Volterra population dynamics, autoregressive time series (AR), autoregressive moving average time series (ARMA), and generalized autoregressive conditional heteroskedasticity time series (GARCH). All but the normal models are drawn from the posterior package [38], but reparameterized according to Stan best practices so that NUTS can sample them.

6.2 Learning curve

As a simple validation that the sampler is targeting the correct distribution, Figure 6 plots a learning curve (expected absolute error) versus iteration for the uniform sampler with the full path for the eight schools model. The plot shows that error decreases as expected at a rate of $1/\sqrt{n}$ and the efficiency is greater than that of independent draws for the parameters and slightly worse for the parameters squared in this simple case.

6.3 Effect of step size and path fraction

In Figure 7, the performance of the uniform self-tuning sampler is shown for two step sizes, 0.36 (blue lines) and 0.18 (red lines), across a range of lower bound fractions (x axis) for uniform sampling. The step size 0.36 is what NUTS adapted for an 80% average Metropolis acceptance probability (Stan's default); 0.18 is the step size for roughly 95% Metropolis acceptance. Halving the step size roughly doubles the number of leapfrog steps taken, as shown in the upper left of the plot. The remaining plots show that performance is better with a smaller step size.

Mean square jump distance (MSJD) is also shown in Figure 7; it is defined for M steps of sampling by

$$\text{MSJD} = \frac{1}{M} \sum_{m=1}^M \left\| \theta^{(m+1)} - \theta^{(m)} \right\|_2^2 \quad (12)$$

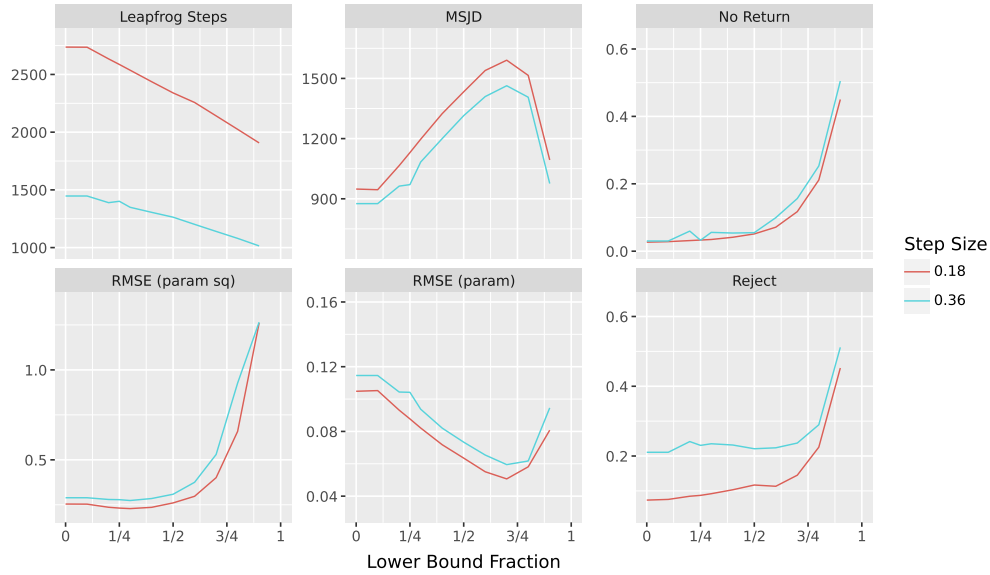


Figure 7: *Performance versus step size and lower bound fraction.* The plots represent averages (over 200 repetitions) of results for 100 iterations of uniform self-tuning HMC starting from a draw from the 500-dimensional standard normal target distribution. The x-axis represents the fraction of the number of steps to a U-turn to use as the lower bound in uniformly drawing a number of leapfrog steps. Line color indicates step sizes 0.36 (blue) and 0.18 (red). The titles of the subplots describe the values on the y-axis. The label "No Return" (top right) is for the fraction of non-reversible proposals due to a U-turn before returning to the starting point, and the proportion of rejections (lower right) includes those due to non-reversibility.

The MSJD is a reasonable estimate of the expected squared jump distance,

$$\text{ESJD} = \mathbb{E} \left[\left\| \theta^{(t+1)} - \theta^{(t)} \right\|^2 \right] \quad (13)$$

As the path fraction goes up, the MSJD goes up until the no-return rejection rate overtakes it and it begins to decrease.

The rejection rate is broken down into total rejection rate, and then the number of rejections due to not being able to return to the origin before hitting a U-turn (see Figure 5). Smaller step sizes do a better job at preserving the Hamiltonian and thus have lower rejection rates.

Root mean square error (RMSE) is defined by

$$\text{RMSE} = \sqrt{\frac{1}{M} \sum_{m=1}^M \left\| \theta^{(m)} - \theta \right\|^2} \quad (14)$$

where θ is the value of the parameter and the $\theta^{(m)}$ are MCMC draws. If the draws were independent, the expected RMSE for parameters is 0.1 (the standard deviation, 1, divided by the square root of the number of i.i.d. draws, $\sqrt{100}$). The expected RMSE for parameters squared is $\sqrt{2}/10$ because standard normal variates squared follow a chi-squared distribution with 1 degree of freedom, the standard deviation of which is $\sqrt{2}$. The RMSE for parameters is below that of independent

draws, going down to roughly half at lower bound fraction 0.6. Not surprisingly, because MSJD tracks inverse lag-one autocorrelation, the best RMSE for parameters is achieved where the MSJD is maximized. The RMSE for parameters squared is maximized at around a 0.35 lower-bound fraction, though values are very close between 0 and 1/2. With a single lower bound fraction, a trade off must be made, as it must be for HMC and NUTS. Here a fraction of 0.6 appears reasonable. The RMSE is lower for the smaller step sizes, but doubling the number of iterations with the larger step size would reduce expected RMSE by a factor of $1/\sqrt{2}$ (about 70%).

6.4 Evaluations for multiple models

We run all of our reference models for 10,000 warmup and 40,000 iterations of NUTS using Stan's defaults and use the sampling draws to define reference means for parameters and parameters squared. The results for the uniform sampler are shown in Figure 8. There are four plots, which show RMSE for parameters, RMSE for squared parameters, MSJD, and leapfrog steps.

We evaluate NUTS (in red) and three settings of uniform sampling, S(hort), which uses $\text{uniform}(1, L)$ sampling, where L is the number of steps to a U-turn, M(edium), which uses $\text{uniform}(\lfloor 0.3 \cdot L \rfloor, L)$, and L(ong), which uses $\text{uniform}(\lfloor 0.5 \cdot L \rfloor, L)$, and XL (extra long), which uses $\text{uniform}(\lfloor 0.7 \cdot L \rfloor, L)$ sampling. Two step sizes are considered, (1) the step size adapted by NUTS with its default acceptance target (80%), and (1/2), which is half of that step size, which is approximately what NUTS would adapt for a target of 95% step size, which is recommended for the numerical stability of the leapfrog integrator in more difficult problems.

Starting with leapfrog steps (the bottom plot), it is clear that the shorter expected paths require more steps for self-tuning HMC, because they are likely to have to go further past the origin backward in time before hitting a U-turn. There is never more than a factor of two difference among systems in terms of leapfrog steps. The steps required for NUTS and ST-HMC are similar; for some models, ST-HMC requires fewer steps than NUTS and for others it requires more steps. For the three normal cases, ST-HMC requires fewer steps for the standard normal and correlated normal, but about the same number for the ill-conditioned normal.

Smaller step sizes lead to roughly twice as many leapfrog steps, but not quite that in complicated models (or models that run up against the 1024 maximum step limit imposed on both NUTS and ST-HMC).

The second plot from the bottom shows the MSJD for all samplers, which is inversely related to autocorrelation at lag 1. The clear trend is that larger steps lead to lower error, but this is not always borne out. For instance, in the standard normal, the ST-HMC sampler takes longer steps and improves the estimate of parameters, but NUTS has lower error estimating parameters squared.

The first two plots show RMSE for parameters and parameters squared. All plots have a y -axis starting at 0 so that relative performance may be assessed visually. Many of the more complicated models are neck and neck in RMSE for parameters and parameters squared, with NUTS consistently slightly outperforming ST-HMC.

In conclusion, NUTS slightly outperforms ST-HMC overall. This is highly encouraging in that this is a relatively simple first attempt at self-tuned HMC, whereas NUTS is represented by Stan's

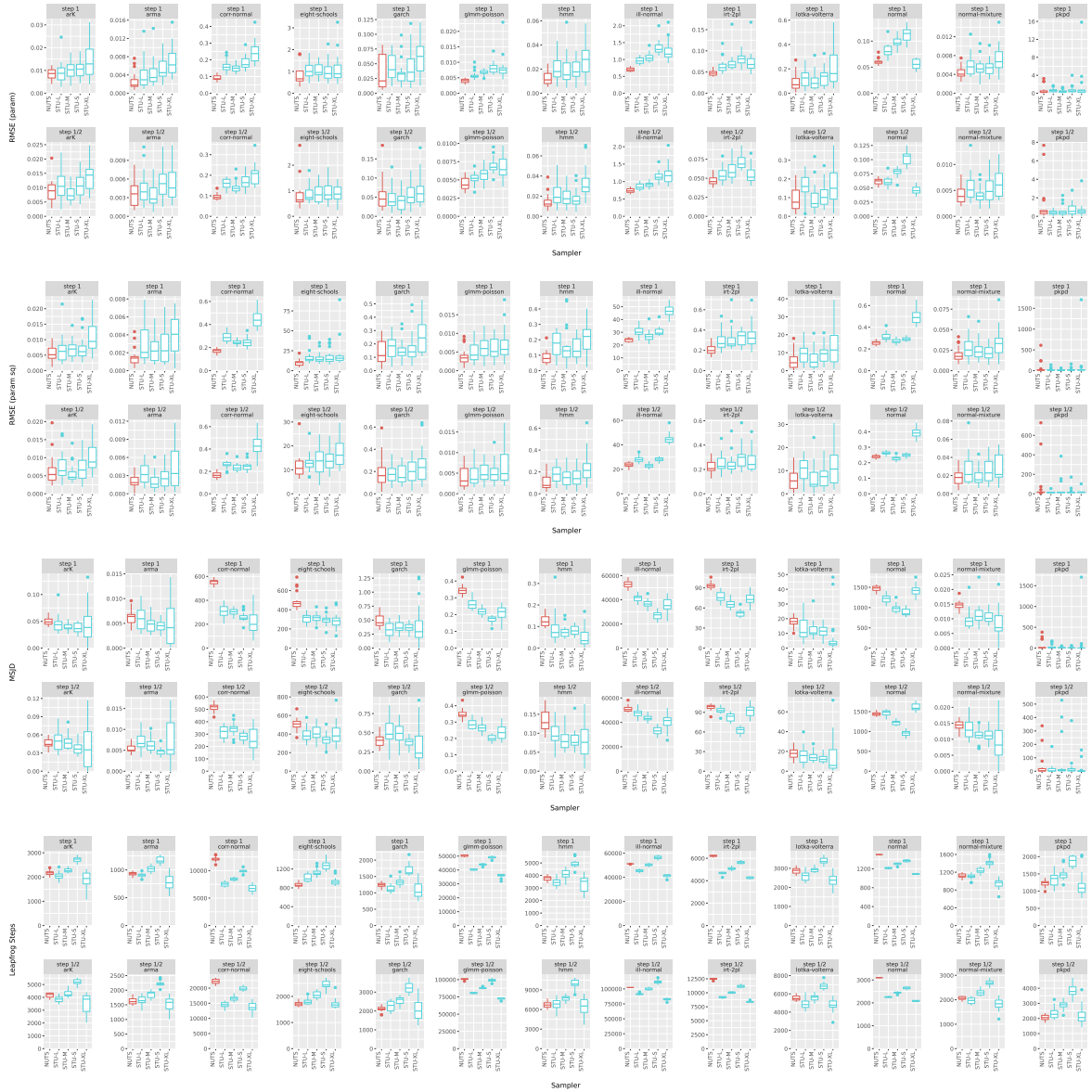


Figure 8: *Empirical Evaluations vs. NUTS*. From top to bottom: root mean square error of parameters, 151
 root mean square error of parameters squared (lower is better), mean square jump distances (higher is better), 153
 and number of leapfrog steps (smaller is more efficient). Each column is for a different model and the rows 156
 are for runs at NUTS default adapted step size (targeting 0.8 average Metropolis acceptance) and half of 144
 that rate (which targets approximately a 0.95 Metropolis acceptance, which increases stability of the leapfrog 156
 integrator). The red result is NUTS and the blue results are for the self-tuning uniform sampler over intervals 155
 with 0.0 (S), 0.3 (M), 0.5 (L), and 0.7 (XL) path fraction. 156

current implementation, which has been implemented robustly and improved considerably over 161
 the first NUTS paper (and first version in Stan) [4, 31]. 161

6.5 Open-source code and reproducibility ²

The results and plots can be reproduced from our code distribution, which is available from GitHub under a permissive open-source license.³

7 Other Related Prior work ⁰

The GIST sampler may be viewed as a dynamic version of the empirical HMC sampler [53]. This empirical HMC sampler learns an empirical distribution of the underlying path lengths to U-turns during a warmup phase and then fixes this empirical distribution, which is then used to sample path lengths at each step during sampling. Like randomized HMC in Section 3.1, the empirical HMC sampler is an instance of the GIST sampler where the conditional distribution of path lengths does not depend on the current position and/or momentum.

Both NUTS and the Apogee-to-Apogee Path Sampler reviewed in Section 3 can be formulated within the dynamic HMC framework introduced in [23]. Dynamic HMC can be viewed as an instance of a generalized Hit-and-Run sampler in the sense of [1]. On the other hand, as discussed in Section 2, the GIST sampler is a Gibbs sampler (with a Metropolis-within-Gibbs step) aimed at the enlarged target measure (2). A key benefit of incorporating a Metropolis-within-Gibbs step in GIST is that it relaxes the restrictive symmetry conditions required for correctness of dynamic HMC in practice, and hence, offers significant flexibility to adapt other algorithm tuning parameters.

The autoMALA sampler introduced a self-tuning version of the Metropolis-adjusted Langevin (MALA) sampler [33, 6]. It uses a forward and reverse non-deterministic scheme to choose adaptation parameters in a way that satisfies detailed balance. MALA is equivalent to one-step Hamiltonian Monte Carlo, but is simpler in not needing to evolve momentum. The GIST sampler can be viewed as a probabilistic generalization of the autoMALA adaptation selection criteria.

For step size and mass matrix tuning, two approaches have been popular in practice. In the first, an adaptation phase is used to estimate algorithm tuning parameters. These tuning parameters are then fixed so that the resulting chain is Markovian. This is the strategy used by NUTS for step size and mass matrix adaptation for HMC [31]. In the second approach, the adaptation phase is never turned off, but the amount of adaptation is decreased so that asymptotically the results are valid. This is the strategy used by delayed rejection Metropolis (DRAM) [27].

Another approach for step size adaptivity is the recently introduced delayed rejection HMC sampler of [42]. The delayed rejection algorithm [41, 26] is a generalization of Metropolis-Hastings to a sequence of proposal moves. This proposal sequence can be tuned to start with larger scale moves and then scaled down for subsequent proposals [27]. In the same vein, the delayed rejection HMC sampler automatically tries smaller step sizes if proposals with larger step sizes are rejected, allowing it to sample from target densities with varying scale. Like other delayed rejection methods, it generates “ghost points” using reversed proposals, which are then used as part of the acceptance probability to ensure detailed balance [26].

³The code is distributed under the MIT License at <https://github.com/bob-carpenter/adaptive-hmc>.

Acknowledgements₀

We would like to thank Edward Roualdes, Tore Kleppe, Andreas Eberle, Sam Livingstone, and Chirag Modi for feedback on the Gibbs self-tuning idea. ₁₆₈

N. Bou-Rabee has been partially supported by NSF Grant No. DMS-2111224. ₁₆₉

References₀

- [1] H. Andersen and P. Diaconis. “Hit and run as a unifying device”. In: *J. Soc. Fr. Stat. & Rev. Stat. Appl.* 148 (Jan. 2007) (cit. on p. 19). ₂ ₆
- [2] A. Beskos, N. S. Pillai, G. O. Roberts, J. M. Sanz-Serna, and A. M. Stuart. “Optimal tuning of the Hybrid Monte Carlo algorithm”. In: *Bernoulli* 19 (2013), pp. 1501–1534 (cit. on p. 1). ₃ ₂
- [3] A. Beskos, F. J. Pinski, J. M. Sanz-Serna, and A. M. Stuart. “Hybrid Monte Carlo on Hilbert spaces”. In: *Stochastic Processes and their Applications* 121.10 (2011), pp. 2201–2230 (cit. on p. 9). ₂ ₁₃₈
- [4] M. Betancourt. “A conceptual introduction to Hamiltonian Monte Carlo”. In: *arXiv preprint 1701.02434* (2017) (cit. on pp. 1, 2, 6–8, 14, 18). ₂ ₁₆₅
- [5] M. Betancourt, S. Byrne, and M. Girolami. “Optimizing the integrator step size for Hamiltonian Monte Carlo”. In: *arXiv preprint 1411.6669* (2014) (cit. on p. 1). ₂ ₁₄₄
- [6] M. Biron-Lattes, N. Surjanovic, S. Syed, T. Campbell, and A. Bouchard-Côté. “autoMALA: Locally adaptive Metropolis-adjusted Langevin algorithm”. In: *27th International Conference on Artificial Intelligence and Statistics*. Vol. PMLR 238. 2024 (cit. on pp. 1, 19). ₆₆ ₂ ₁₆₅
- [7] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler. “Transition path sampling: Throwing ropes over rough mountain passes, in the dark”. In: *Annual Review of Physical Chemistry* 53.1 (2002), pp. 291–318. ISSN: 0066-426X (cit. on p. 11). ₂ ₉₇ ₈₂
- [8] J. Borggaard, N. Glatt-Holtz, and J. Krometis. “A Bayesian approach to estimating background flows from a passive scalar”. In: *arXiv preprint 1808.01084* (2018) (cit. on p. 11). ₇₈ ₁₀₀
- [9] N. Bou-Rabee and A. Eberle. “Couplings for Andersen dynamics in high dimension”. In: *Ann. Inst. H. Poincaré Probab. Statist* 58.2 (2022), pp. 916–944 (cit. on pp. 2, 4, 5, 9). ₂ ₁₀₀
- [10] N. Bou-Rabee and J. M. Sanz-Serna. “Geometric integrators and the Hamiltonian Monte Carlo Method”. In: *Acta Numerica* 27 (2018), pp. 113–206 (cit. on pp. 3, 11). ₂ ₂
- [11] N. Bou-Rabee and A. Eberle. “Two-scale coupling for preconditioned Hamiltonian Monte Carlo in infinite dimensions”. In: *Stochastics and Partial Differential Equations: Analysis and Computations* 9.1 (2021), pp. 207–242 (cit. on p. 9). ₂ ₃ ₃
- [12] N. Bou-Rabee and T. S. Kleppe. “Randomized Runge-Kutta-Nyström”. In: *arXiv preprint 2310.07399* (2023) (cit. on p. 4). ₂ ₂
- [13] N. Bou-Rabee and M. Marsden. “Unadjusted Hamiltonian MCMC with stratified Monte Carlo time integration”. In: *arXiv preprint 2211.11003* (2022) (cit. on p. 4). ₂ ₃₂
- [14] N. Bou-Rabee and S. Oberdörster. “Mixing of Metropolis-Adjusted Markov Chains via couplings: The high acceptance regime”. In: *arXiv preprint 2308.04634* (2023) (cit. on p. 4). ₉₉ ₆₆
- [15] N. Bou-Rabee and J. M. Sanz-Serna. “Randomized Hamiltonian Monte Carlo”. In: *Ann. Appl. Probab.* 27.4 (2017), pp. 2159–2194 (cit. on pp. 1, 2, 4, 5, 9, 10). ₂ ₂
- [16] N. Bou-Rabee and K. Schuh. “Nonlinear Hamiltonian Monte Carlo and its particle approximation”. In: *arXiv preprint 2308.11491* (2023) (cit. on p. 4). ₂ ₂

- [17] Y. Cao, J. Lu, and L. Wang. "Complexity of randomized algorithms for underdamped Langevin dynamics". In: *Communications in Mathematical Sciences* 19.7 (2021), pp. 1827–1853 (cit. on p. 4).₂^{155 165}
- [18] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. "Stan: A probabilistic programming language". In: *Journal of Statistical Software* 20 (2016), pp. 1–37 (cit. on p. 1).₂^{0 3}
- [19] D. Chandler and P. G. Wolynes. "Exploiting the isomorphism between quantum theory and classical statistical mechanics of polyatomic fluids". In: *The Journal of Chemical Physics* 74.7 (1981), pp. 4078–4095. ISSN: 0021-9606 (cit. on p. 11).₂^{137 119}
- [20] S. U. Chris Sherlock and M. Ludkin. "The apogee to apogee path sampler". In: *Journal of Computational and Graphical Statistics* 32.4 (2023), pp. 1436–1446 (cit. on pp. 2, 8, 9).₇₈¹⁶³
- [21] M. Dashti and A. M. Stuart. "The Bayesian approach to inverse problems". In: *Handbook of Uncertainty Quantification* (2017), pp. 311–428 (cit. on p. 11).₂¹⁰⁰
- [22] P. de Valpine, D. Turek, C. Paciorek, C. Anderson-Bergman, D. Temple Lang, and R. Bodik. "Programming with models: writing statistical algorithms for general model structures with NIMBLE". In: *Journal of Computational and Graphical Statistics* 26 (2017), pp. 403–417 (cit. on p. 1).₀^{2 119 72}
- [23] A. Durmus, S. Gruffaz, M. Kailas, E. Saksman, and M. Vihola. "On the convergence of dynamic implementations of Hamiltonian Monte Carlo and no U-turn samplers". In: *arXiv preprint 2307.03460* (2023) (cit. on pp. 6, 7, 19).₂⁴⁷
- [24] H. Ge, K. Xu, and Z. Ghahramani. "Turing: A language for flexible probabilistic inference". In: *International Conference on Artificial Intelligence and Statistics, (AISTATS)*. 2018, pp. 1682–1690 (cit. on p. 1).₂^{155 66}
- [25] M. Girolami and B. Calderhead. "Riemann manifold Langevin and Hamiltonian Monte Carlo methods". In: *J R Statist Soc B* 73 (2011), pp. 123–214 (cit. on p. 1).₇₄²
- [26] P. J. Green and A. Mira. "Delayed rejection in reversible jump Metropolis–Hastings". In: *Biometrika* 88.4 (2001), pp. 1035–1053 (cit. on p. 19).₂¹⁰⁰
- [27] H. Haario, M. Laine, A. Mira, and E. Saksman. "DRAM: efficient adaptive MCMC". In: *Statistics and computing* 16 (2006), pp. 339–354 (cit. on p. 19).₇₈⁰
- [28] S. Habershon, D. E. Manolopoulos, T. E. Markland, and T. F. Miller. "Ring-polymer molecular dynamics: Quantum effects in chemical dynamics from classical trajectories in an extended phase space". In: *Annual Review of Physical Chemistry* 64.1 (2013), pp. 387–413. ISSN: 0066-426X (cit. on p. 11).₂^{36 163}
- [29] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration*. en. 2nd ed. Springer Series in Computational Mathematics. Berlin, Germany: Springer, Feb. 2010 (cit. on p. 3).₉₉¹⁰⁰
- [30] Y. He, K. Balasubramanian, and M. A. Erdogdu. "On the ergodicity, bias and asymptotic normality of randomized midpoint sampling method". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. 2020, pp. 7366–7376 (cit. on p. 4).₂^{137 36 212}
- [31] M. D. Hoffman and A. Gelman. "The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo". In: *Journal of Machine Learning Research* 15.1 (2014), pp. 1593–1623 (cit. on pp. 1, 2, 12, 14, 18, 19).₂^{89 3}

- [32] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Vol. 160. Applied Mathematical Sciences. Springer Science & Business Media, 2005 (cit. on p. 11). 201
- [33] T. S. Kleppe. “Adaptive step size selection for Hessian-based manifold Langevin samplers”. In: *Scandinavian Journal of Statistics* 43.3 (2016), pp. 788–805 (cit. on pp. 1, 19). 166 102
- [34] T. S. Kleppe. “Connecting the dots: Numerical randomized Hamiltonian Monte Carlo with state-dependent event rates”. In: *Journal of Computational and Graphical Statistics* 31.4 (2022), pp. 1238–1253 (cit. on p. 1). 2 61
- [35] R. Korol, J. L. Rosa-Raíces, N. Bou-Rabee, and T. F. Miller. “Dimension-free path-integral molecular dynamics without preconditioning”. In: *The Journal of Chemical Physics* 152.10 (2020), p. 104102 (cit. on p. 11). 2 66
- [36] J. Lu, Y. Lu, and Z. Zhou. “Continuum limit and preconditioned Langevin sampling of the path integral molecular dynamics”. In: *Journal of Computational Physics* 423 (2020), p. 109788 (cit. on p. 11). 2 7 89
- [37] P. B. Mackenzie. “An improved hybrid Monte Carlo method”. In: *Physics Letters B* 226.3 (1989), pp. 369–371 (cit. on p. 4). 2
- [38] M. Magnusson, P. Bürkner, and A. Vehtari. *posterioradb: a set of posteriors for Bayesian inference and probabilistic programming*. <https://github.com/stan-dev/posterioradb>. Version 0.5. 2023 (cit. on p. 15). 2 58 3
- [39] T. F. Miller and D. E. Manolopoulos. “Quantum diffusion in liquid para-hydrogen from ring-polymer molecular dynamics”. In: *The Journal of Chemical Physics* 122.18 (2005), p. 184503. ISSN: 0021-9606 (cit. on p. 11). 2 61 66
- [40] T. F. Miller III and C. Predescu. “Sampling diffusive transition paths”. In: *The Journal of Chemical Physics* 126.14 (2007), p. 144102. ISSN: 0021-9606 (cit. on p. 11). 2 82
- [41] A. Mira. “On Metropolis-Hastings algorithms with delayed rejection”. In: *Metron* 59.3-4 (2001), pp. 231–241 (cit. on p. 19). 2 167
- [42] C. Modi, A. Barnett, and B. Carpenter. “Delayed rejection Hamiltonian Monte Carlo for sampling multiscale distributions”. In: *Bayesian Analysis* 1.1 (2023), pp. 1–28 (cit. on p. 19). 19 73
- [43] R. M. Neal. “MCMC using Hamiltonian dynamics”. In: *Handbook of Markov Chain Monte Carlo* 2 (2011), pp. 113–162 (cit. on pp. 4, 9). 2 3
- [44] L. R. Petzold, L. O. Jay, and J. Yen. “Numerical solution of highly oscillatory ordinary differential equations”. In: *Acta Numer.* 6 (1997), pp. 437–483. ISSN: 0962-4929 (cit. on pp. 9, 11). 0 85 2
- [45] D. Phan, N. Pradhan, and M. Jankowiak. “Composable effects for flexible and accelerated probabilistic programming in NumPyro”. In: *arXiv preprint 1912.11554* (2019) (cit. on p. 1). 3 155
- [46] F. J. Pinski and A. M. Stuart. “Transition paths in molecules at finite temperature”. In: *The Journal of Chemical Physics* 132.18 (2010), p. 184104 (cit. on p. 11). 102 61
- [47] M. G. Reznikoff and E. Vanden-Eijnden. “Invariant measures of stochastic partial differential equations and conditioned diffusions”. In: *Comptes Rendus Mathématique* 340.4 (2005), pp. 305–308 (cit. on p. 11). 2 66 4
- [48] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. “Probabilistic programming in Python using PyMC3”. In: *PeerJ Computer Science* 2 (2016), e55 (cit. on p. 1). 2 100
- [49] R. Shen and Y. T. Lee. “The randomized midpoint method for log-concave sampling”. In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on p. 4). 99 155

- [50] A. M. Stuart. "Inverse problems: a Bayesian perspective". In: *Acta Numerica* 19 (2010), pp. 451–559 (cit. on p. 11).₂
- [51] L. Tierney. "A note on Metropolis-Hastings kernels for general state spaces". In: *The Annals of Applied Probability* 8.1 (1998), pp. 1–9 (cit. on p. 23).₁₀₂
- [52] P. A. Whalley, D. Paulin, and B. Leimkuhler. "Randomized time Riemannian manifold Hamiltonian Monte Carlo". In: *Statistics and Computing* 34.1 (2024), p. 48 (cit. on p. 1).₃
- [53] C. Wu, J. Stoehr, and C. P. Robert. "Faster Hamiltonian Monte Carlo by learning leapfrog scale". In: *arXiv preprint 1810.04449* (2018) (cit. on pp. 4, 19).₂

A Proof of correctness₆

The correctness of the GIST sampler given in Figure 1 follows from the observation that the GIST sampler is a Gibbs sampler which interleaves Gibbs resampling of the velocity variable ρ and tuning parameter α with a Metropolis-within-Gibbs step that uses a proposal given by a measure-preserving involution. In fact, this idea is also the essence of the standard HMC sampler, and the GIST sampler is a natural extension of this idea to the selection of tuning parameters as well.₁₇₀

The proof of correctness relies on the following general lemma. This lemma is not new; see, e.g., special case 2 of Theorem 2 in [51]. However, a complete proof is given for the reader's convenience.₁₇₁

Lemma 2. Suppose that $(S, \mathcal{F}, \lambda)$ is a measure space, ν is a probability measure with strictly positive density γ with respect to λ , and $F : S \rightarrow S$ is a measurable involution which preserves λ . Given a state $Z \in S$, define a Markov transition kernel by the following procedure:₁₇₂

- Draw $U \sim \text{uniform}([0, 1])$.₁₇₃
- Set $Z' = F(Z)$ if $U \leq \min\left(1, \frac{\gamma(F(Z))}{\gamma(Z)}\right)$ and $Z' = Z$ otherwise.₁₇₃

Then, this transition step is reversible with respect to ν .₂₈

Proof of Lemma 2. Let $Z \sim \nu$. The proof shows that₁₇₄

$$\mathbb{P}(Z \in A, Z' \in B) = \mathbb{P}(Z' \in A, Z \in B), \quad 185$$

for all measurable sets $A, B \in \mathcal{F}$. Let $\beta(x, F(x)) = \min\left(1, \frac{\gamma(F(x))}{\gamma(x)}\right)$ for all $x \in S$. For all $x \in S$ and for any measurable set $A \in \mathcal{F}$, the transition kernel associated to the transition step described in Lemma 2 is given by₅₀

$$p(x, A) = \beta(x, F(x))\delta_{F(x)}(A) + (1 - \beta(x, F(x)))\delta_x(A). \quad 175$$

Hence,₈₁

$$\begin{aligned} \mathbb{P}(Z \in A, Z' \in B) &= \int_A p(x, B) \nu(dx) \\ &= \int_A [\beta(x, F(x))\mathbb{1}_B(F(x)) + (1 - \beta(x, F(x)))\mathbb{1}_B(x)] \nu(dx) \\ &= \int_S [\beta(x, F(x))\mathbb{1}_B(F(x))\mathbb{1}_A(x) + (1 - \beta(x, F(x)))\mathbb{1}_A(x)\mathbb{1}_B(x)] \nu(dx). \end{aligned} \quad (15)$$

The second term in (16) is already symmetric in the sets A, B — so we turn our attention to the first term. For this term, we use the elementary identity $a \min(1, \frac{a}{b}) = b \min(1, \frac{b}{a})$ valid for all $a, b \neq 0$, as follows

$$\begin{aligned} & \int_S \beta(x, F(x)) \mathbb{1}_B(F(x)) \mathbb{1}_A(x) \nu(dx) \\ &= \int_S \beta(x, F(x)) \mathbb{1}_B(F(x)) \mathbb{1}_A(x) \gamma(x) \lambda(dx) \\ &= \int_S \mathbb{1}_B(F(x)) \mathbb{1}_A(x) \gamma(x) \min\left(1, \frac{\gamma(F(x))}{\gamma(x)}\right) \lambda(dx) \\ &= \int_S \mathbb{1}_B(F(x)) \mathbb{1}_A(x) \gamma(F(x)) \min\left(1, \frac{\gamma(x)}{\gamma(F(x))}\right) \lambda(dx) \end{aligned}$$

But, since the map F preserves λ and is an involution we have, by change of variables under F ,

$$\begin{aligned} & \int_S \mathbb{1}_B(F(x)) \mathbb{1}_A(x) \gamma(F(x)) \min\left(1, \frac{\gamma(x)}{\gamma(F(x))}\right) \lambda(dx) \\ &= \int_S \mathbb{1}_B(F(F(x))) \mathbb{1}_A(F(x)) \gamma(F(F(x))) \min\left(1, \frac{\gamma(F(x))}{\gamma(F(F(x)))}\right) \lambda(dx) \\ &= \int_S \mathbb{1}_B(x) \mathbb{1}_A(F(x)) \gamma(x) \min\left(1, \frac{\gamma(F(x))}{\gamma(x)}\right) \lambda(dx) \\ &= \int_S \beta(x, F(x)) \mathbb{1}_B(x) \mathbb{1}_A(F(x)) \nu(dx) = \int_B \beta(x, F(x)) \mathbb{1}_A(F(x)) \nu(dx) . \end{aligned} \quad (17)$$

Inserting (17) into (16), and comparing with (15), we observe that

$$\mathbb{P}(Z \in A, Z' \in B) = \mathbb{P}(Z \in B, Z' \in A)$$

as required. \square

With Lemma 2 in hand, we are now in position to prove correctness of the GIST sampler.

Proof of Theorem 1. It is clear that the iterates $\theta_0, \theta_1, \dots$ generated by the GIST sampler in Figure 1 form a Markov chain, so we need only establish reversibility. In the notation of Figure 1, suppose $\theta_0 = \theta \sim \mu_\theta$. After selecting $\rho_0 \sim \text{normal}(0, I_{d \times d})$ and $\alpha \sim p(\cdot \mid \theta_0, \rho_0)$, then $(\theta_0, \rho_0, \alpha) \sim \mu_e$.

Let $u \sim \text{uniform}([0, 1])$ and set

$$(\theta^*, \rho^*, \alpha^*) = \begin{cases} G(\theta_0, \rho_0, \alpha) & \text{if } u \leq e^{-\Delta H(\theta_0, \rho_0)} \frac{p(\pi(\theta_0, \rho_0)(\alpha) \mid S \circ F(\alpha)(\theta_0, \rho_0))}{p(\alpha \mid \theta_0, \rho_0)} \\ G(\theta_0, \rho_0, \alpha) & \text{else} \end{cases}$$

Since G is a measure-preserving involution by assumption, by Lemma 2 we have

$$\mathbb{P}((\theta_0, \rho_0, \alpha) \in A, (\theta^*, \rho^*, \alpha^*) \in B) = \mathbb{P}((\theta^*, \rho^*, \alpha^*) \in A, (\theta_0, \rho_0, \alpha) \in B) \quad (18)$$

for any measurable $A, B \subset \mathbb{R}^{2d} \times \mathcal{A}$. In particular, for Borel sets $\tilde{A}, \tilde{B} \subset \mathbb{R}^d$,

$$\mathbb{P}(\theta \in \tilde{A}, \theta^* \in \tilde{B}) = \mathbb{P}(\theta^* \in \tilde{A}, \theta \in \tilde{B})$$

by taking $A = \tilde{A} \times \mathbb{R}^d \times \mathcal{A}$ and $B = \tilde{B} \times \mathbb{R}^d \times \mathcal{A}$ in (18). Hence, the GIST sampler produces a Markov Chain that is reversible with respect to $e^{-U(x)}m(d\theta)$. \square

As a corollary, Theorem 1 implies correctness of the GIST samplers presented in Sections 3.1 and 3.2, as well as the GIST sampler in Section 3.3. \square

Corollary 3. *If $\pi(\theta, \rho)(\alpha) = \alpha$ for all $(\theta, \rho, \alpha) \in \mathbb{R}^{2d} \times \mathcal{A}$, the corresponding GIST sampler is correct.*

Proof. We need only verify the map $G : \mathbb{R}^{2d} \times \mathcal{A} \rightarrow \mathbb{R}^{2d} \times \mathcal{A}$ defined by $G(\theta, \rho, \alpha) = (\mathcal{S} \circ F(\alpha)(\theta, \rho), \alpha)$ is a measure-preserving involution. As $\mathcal{S} \circ F(\alpha)$ is an involution G is automatically an involution on $\mathbb{R}^{2d} \times \mathcal{A}$. Additionally $\mathcal{S} \circ F(\alpha)$ preserves Lebesgue measure m on \mathbb{R}^{2d} for every fixed α . Hence, by Fubini's theorem, G preserves $(m \otimes \eta)$ on $\mathbb{R}^{2d} \times \mathcal{A}$.

Thus, iterating the transition step in Figure 1 produces a Markov chain reversible with respect to μ_θ by Theorem 1. \square

The next two corollaries of Theorem 1 show the correctness of the GIST samplers in Section 3.4 and Section 3.5.

Corollary 4. *The No-U-Turn Sampler presented in Section 3.4 is correct.*

Proof. In this case, the map G is of the form $G(\theta, \rho, J, i) = (\mathcal{S} \circ \Phi_h^i(\theta, \rho), -(J - i), i)$. This is clearly an involution. Moreover, for θ, ρ fixed the map $(J, i) \rightarrow (-(J - i), i)$ is a bijection and thus preserves the counting measure on $\mathcal{P}_F \times \mathbb{Z}$. By Fubini's theorem G then preserves $(m \otimes \eta)$ on $\mathbb{R}^{2d} \times \mathcal{P}_F \times \mathbb{Z}$.

Thus, by Theorem 1, we get correctness of NUTS as presented in Section 3.4. \square

Corollary 5. *The Apogee-to-Apogee Path Sampler presented in Section 3.5 is correct.*

Proof. Here, we take $G(\theta, \rho, c, i) = (\mathcal{S} \circ \Phi_h^i(\theta, \rho), c + S_\#((\theta, \rho), \mathcal{S} \circ \Phi_h^i(\theta, \rho)), i)$. Since by definition $S_\#((\theta, \rho), (\theta', \rho')) = -S_\#((\theta', \rho'), (\theta, \rho))$, we observe by direct computation that G is an involution.

For fixed (θ, ρ, i) , the map $c \mapsto c + S_\#((\theta, \rho), \mathcal{S} \circ \Phi_h^i(\theta, \rho))$ is a bijection and hence preserves counting measure. Additionally, for i fixed, $\theta, \rho \mapsto \mathcal{S} \circ \Phi_h^i(\theta, \rho)$ preserves Lebesgue measure.

Applying Fubini's theorem then implies that the map G is measure preserving. Hence, by Theorem 1, the Apogee-to-Apogee Path Sampler presented in Section 3.5 is correct. \square

Remark 5. In the description above, the proposal $G(\theta_0, \rho_0, \alpha)$ has phase space state given by $\mathcal{S} \circ F(\alpha)(\theta_0, \rho_0)$ whereas the proposal in the original algorithm is $F(\alpha)(\theta_0, \rho_0)$. This difference is inconsequential since the acceptance probabilities are the same and the map \mathcal{S} affects only the velocity variable. Since this variable is immediately discarded in the next step of the chain, we disregard this difference. Equivalently, one could adapt the description above so that \mathcal{S} is applied to the phase space coordinates immediately after Metropolizing the proposal. Since \mathcal{S} preserves the extended target distribution μ_e , the whole procedure will still preserve the extended target.

A.1 Reduction to standard Metropolis-Hastings 200

It is worth noting that the above yields the original Metropolis-Hastings algorithm as a special case. Let $(X, \mathcal{F}, \lambda)$ be a measure space with σ -finite measure λ . Suppose $\mu(dx) = p_0(x)\lambda(dx)$ is an absolutely continuous target probability measure and $P(x, dy) = p(x, y)\lambda(dy)$ is an absolutely continuous transition kernel with strictly positive densities. Analogously to (2), define the following enlarged target measure on $X \times X$:

$$\mu_e(dx dy) = \gamma(x, y)\lambda(dx dy) = p_0(x)p(x, y)(\lambda \otimes \lambda)(dx dy).$$

Note that the marginal of μ_e in the first component gives the desired target.

On $X \times X$, and from a state with first component $x \in X$, first generate a sample from the second component y conditional on the first, i.e.,

$$y \sim p(x, \cdot).$$

Then, apply the Metropolis procedure to the involutive proposal $\phi : (x, y) \mapsto (y, x)$ which clearly preserves $\lambda \otimes \lambda$. According to Lemma 2, the corresponding acceptance probability is given by

$$a_e(x, y) = 1 \wedge \left(\frac{\gamma(\phi(x, y))}{\gamma(x, y)} \right) = 1 \wedge \left(\frac{p_0(y)p(y, x)}{p_0(x)p(x, y)} \right)$$

which we recognize as the standard Metropolis-Hastings acceptance probability.

B Test models 139

In this section, we briefly describe the models used for evaluations. Greek letters are used for parameters, roman letters for constants, predictors and modeled data, and italics for indexes. Where not otherwise specified, parameters have weakly informative priors concentrated on their rough scale.

Standard normal A 500-dimensional standard normal, with $\alpha \sim \text{normal}(0, I)$, where I is a 500×500 identity matrix.

Correlated normal A 250-dimensional correlated normal $\alpha \sim \text{normal}(0, S)$, where the covariance is that of a unit scale first-order random walk, $S_{i,j} = r^{|i-j|}$, with correlation $r = 0.9$.

Ill-conditioned normal A 250-dimensional ill-conditioned normal, with $\alpha \sim \text{normal}(0, \text{diag-matrix}(\sigma))$, where $\sigma = \begin{bmatrix} 1 & 2 & \cdots & 250 \\ 250 & 250 & \cdots & 250 \\ \vdots & \vdots & \ddots & \vdots \\ 250 & 250 & \cdots & 250 \end{bmatrix}^\top$.

Poisson generalized linear mixed model The expected count for observation i is $\lambda_i = \exp(\alpha + \beta_1 \cdot t_i + \beta_2 \cdot t_i^2 + \beta_3 \cdot t_i^3 + \varepsilon_i)$, with a hierarchical prior $\varepsilon_i \sim \text{normal}(0, \sigma)$, and a likelihood $y_i \sim \text{Poisson}(\lambda_i)$, for $i < 40$.

Eight-schools meta-analysis Observations of mean effects y_j and standard errors σ_j in school j are used to estimate a school effect θ_j with likelihood $y_j \sim \text{normal}(\theta_j, \sigma_j)$ and a non-centered parameterization of the prior $\theta_j \sim \text{normal}(\mu, \tau)$.

Order K autoregressive A K -th order autoregressive model with likelihood $y_t \sim \text{normal}(\alpha + [y_{t-1} y_{t-2} \cdots y_{t-K}] \beta, \sigma)$. The prior does not restrict coefficient vectors β to ensure stationary. A history of $K = 5$ and $T = 200$ time points is used in the evaluations.

- Autoregressive moving average (order 1, 1)** An autoregressive time series with first element $y_1 \sim \text{normal}(\mu + \phi \cdot \mu, \sigma)$, and subsequent elements $y_t \sim \text{normal}(\mu + \phi \cdot y_{t-1} + \theta \cdot \epsilon_{t-1}, \sigma)$ for $t > 1$. 207
- Generalized autoregressive conditional heteroskedasticity** An autoregressive time series model with stochastic volatility, with $y_t \sim \text{normal}(\mu, \sigma_t)$, where $\sigma_1 = s$ is given as data and $\sigma_t = \sqrt{\alpha_0 + \alpha_1 \cdot (y_{t-1} - \mu)^2 + \beta_1 \cdot \sigma_{t-1}^2}$. 208
- Hidden Markov model** A hidden Markov model (HMM) with normal observations. The data generating process is Markovian in hidden state, $z_t \sim \text{categorical}(\phi_{z_{t-1}})$, and then normal in observation, $y_t \sim \text{normal}(\mu_{z_t}, \sigma_{z_t})$. In the implementation, the forward algorithm is used to marginalize the z_t to calculate the likelihood. The vector μ is constrained to ascending order for identifiability. 209
- Hierarchical item-response theory 2 parameter logistic** For test item i , a student j 's correctness is generated by $y_{i,j} \sim \text{bernoulli}(\text{logit}^{-1}(\alpha_i \cdot (\theta_j - \beta_i)))$, for ability θ , difficulty β , and discrimination α . For identifiability, β has a standard hierarchical prior, $\beta_i \sim \text{normal}(\mu^\beta, \sigma^\beta)$, whereas the prior for α is centered, $\alpha_i \sim \text{normal}(0, \sigma^\alpha)$, and the prior for θ is fixed to unit scale, $\theta \sim \text{normal}(0, 1)$. 210
- Normal mixture** A normal mixture model with $z_n \sim \text{bernoulli}(\theta)$, and $y_n \sim \text{normal}(\mu_{z_n}, \sigma_{z_n})$, with μ constrained to ascending order for identifiability. The responsibility parameter z_n is marginalized out. 211
- Lotka-Volterra population dynamics** A lognormal model of population time series for prey ($y_{t,1}$) and predator ($y_{t,2}$). The population dynamics is modeled by a system of ordinary differential equations, $\frac{d}{dt}u(t) = (\alpha - \beta \cdot v(t)) \cdot u(t)$ and $\frac{d}{dt}v(t) = (-\gamma + \delta \cdot u(t)) \cdot v(t)$ with unknown starting point $(u(0), v(0))$ and discrete observations modeled by $y_{t,1} \sim \text{lognormal}(\log u(t), \sigma_1)$ and $y_{t,2} \sim \text{lognormal}(\log v(t), \sigma_2)$, where $u(t)$ and $v(t)$ are solutions to the ODE. 212
- Pharmacometric/pharmacodynamic model** A one-compartment PK/PD elimination model with Michaelis-Menten dynamics defined by a differential equation for concentration with dosing (D) of $\frac{d}{dt}C = \frac{V}{u+C(t)} - \exp(-\delta \cdot t) \cdot D \cdot \frac{\partial}{\partial V}$, with lognormal likelihood for discrete observations $y_t \sim \text{lognormal}(\log C(t), \sigma)$, with all parameters constrained to be positive. 213