Bayesian Analysis (2024) **19**, Number 3, pp. 815–842

Delayed rejection Hamiltonian Monte Carlo for sampling multiscale distributions

Chirag Modi[*,†], Alex Barnett[*], and Bob Carpenter[*]

Abstract. The efficiency of Hamiltonian Monte Carlo (HMC) can suffer when sampling a distribution with a wide range of length scales, because the small step sizes needed for stability in high-curvature regions are inefficient elsewhere. To address this we present a *delayed rejection* (DR) variant: if an initial HMC trajectory is rejected, we make one or more subsequent proposals each using a step size geometrically smaller than the last. To reduce the cost of DR approaches, we extend the standard delayed rejection to a probabilistic framework wherein we do not make multiple proposals at every rejection, but allow the probability of a retry to depend on the probability of accepting the previous proposal. We test the scheme in several sampling tasks, including statistical applications and multiscale model distributions such as Neal's funnel. Delayed rejection enables sampling multiscale distributions for which standard approaches such as HMC fail to explore the tails, and improves performance five-fold over optimally-tuned HMC as measured by effective sample size per gradient evaluation. Even for simpler distributions, delayed rejection provides increased robustness to step size misspecification.

Keywords: delayed rejection, Hamiltonian Monte Carlo, detailed balance, multiscale distributions.

1 Introduction

Hamiltonian Monte Carlo (HMC), including auto-tuned extensions like the no U-turn sampler (NUTS), have become the de facto standard for high performance sampling of high-dimensional, differentiable distributions (Duane et al., 1987; Neal, 2011; Hoffman and Gelman, 2011). One reason for this is that HMC scales much better with dimension than other Markov chain Monte Carlo (MCMC) methods such as random-walk Metropolis or Gibbs sampling. HMC's scalability derives from its ability to move large distances by approximating the Hamiltonian flow defined by the gradient of a distribution's log density function (Betancourt, 2017). As a result, with optimal tuning of parameters, HMC requires $\mathcal{O}(d^{5/4})$ iterations to generate an independent sample in d dimensions as compared to the $\mathcal{O}(d^2)$ iterations for random-walk Metropolis or Gibbs sampling (Neal, 2011; Livingstone and Zanella, 2019). The actual efficiency also depends strongly on geometric features of the density being sampled, particularly issues of high correlation between coordinates (leading to *stiffness*, i.e., ill-conditioning of the local curvature Hessian), and of spatially varying curvature (which defeats the use of global preconditioning to counteract stiffness).

arXiv: 2110.00610
[*]Center for Computational Mathematics, Flatiron Institute, New York, cmodi@flatironinstitute.org
[†]Center for Computational Astrophysics, Flatiron Institute, New York

 https://doi.org/10.1214/23-BA1360

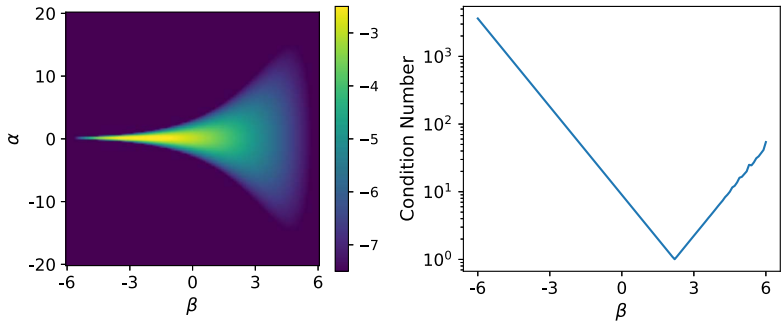


Figure 1: Neal's funnel. (Left) Natural log density of the two-dimensional funnel (Equation 27). The neck of the funnel, where $\beta \ll 0$, has low volume but high density; the mouth, where $\beta \gg 0$, has high volume but low density. (Right) The condition number of the inverse Hessian as a function of the log scale parameter β, with $\alpha = 0$ fixed. This plot covers two standard deviations of β, or about 95% of the probability mass, and the conditioning continues to get worse in the tails.

One of the most common pathologies plaguing these algorithms is the multiscale geometry of the posterior distributions (Betancourt and Girolami, 2015; Pourzanjani and Petzold, 2019). When the curvature of the log density varies spatially over a large dynamic range, small HMC time steps are needed for numerical stability in the high-curvature regions, preventing the use of the larger time steps needed for sampling in smoother regions. This geometry arises naturally in hierarchical models that provide a population model for a group of effects in order to support regularization and partial pooling. However since all the contributions at the bottom of the hierarchy depend on the common global parameter, a small change in these high level parameters can induce large changes in the conditional density of the effects. Consequently, when the data are sparse and inference is sensitive to the priors on these parameters, the posterior density of these models looks like a funnel with a neck of high density but low volume smoothly widening to a mouth of low density and high volume as the hierarchical scale increases. We show a two-dimensional example of this distribution in Figure 1. In the right panel of the same figure, we show the dramatic variations in condition number as the log scale parameter varies. Sampling this distribution is challenging because the mouth and the neck of the funnel contain equal probability mass and so any sampling algorithm needs to be able to sample both regions.

One option for managing varying curvature is to generalize HMC from Euclidean to Riemannian geometry. This allows the use of local curvature information the form of the Fisher information (Girolami and Calderhead, 2011) or conditioned Hessians (Betancourt, 2013). However, constructing and differentiating through higher-dimensional positive definite matrices does not scale well to the higher dimension problems that are common in applied statistics.

An alternative way to deal with high curvature is to approximate the Hamiltonian flow with an implicit symplectic integrator, which is able to naturally adjust stepping in

different regions of phase space (Pourzanjani and Petzold, 2019; Brofos and Lederman, 2021). Even simple implicit integration schemes like implicit midpoint are costly and present an algorithmic challenge for efficient and stable line search. Ultimately, we believe it will be necessary to combine implicit integration and delayed rejection to achieve greater robustness in the face of even more challenging posterior sampling problems.

In this work, we develop an alternate approach inspired by the use of *delayed rejection* (DR) methods to sample multiscale posterior distributions. Recall that a high rejection rate increases autocorrelation of the Markov chain, reducing sampling efficiency. Whenever a rejection would occur in the Metropolis algorithm, DR methods do additional work, which can even exploit knowledge of the first rejection, to make a new proposal with a higher chance of acceptance (Haario et al., 2006). Since such new (possibly expensive) proposals are mostly made only when the standard proposal is poor, efficiency can be increased. Although well studied in the context of random-walk Metropolis sampling (Mira, 1998; Tierney and Mira, 1999; Green and Mira, 2001; Bédard et al., 2014; Haario et al., 2006), there has been relatively little work done with these approaches for Hamiltonian Monte Carlo samplers (Sohl-Dickstein et al., 2014; Campos and Sanz-Serna, 2015). Delayed rejection has also been applied to deterministic volume-preserving proposals in related settings (Sherlock and Thiery, 2017; Vanetti et al., 2017; Andrieu et al., 2020).

Previous approaches employing delayed rejection with HMC extend the same trajectory upon rejection so as to balance the additional cost by making larger jumps in the state space (Sohl-Dickstein et al., 2014; Campos and Sanz-Serna, 2015). However, this approach will fail if for chains that are stuck in a region of high curvature where instability causes a high rejection rate. As with delayed rejection in random-walk Metropolis (Green and Mira, 2001; Haario et al., 2006), we show that chains can escape regions of high curvature if we instead reduce the step size in order to increase the probability of accepting a proposal. In this work, we use this idea, building upon the original idea of delayed rejection (Tierney and Mira, 1999; Green and Mira, 2001) to develop delayed rejection HMC (DR-HMC). Upon a rejection, DR-HMC makes one or more subsequent proposals with smaller step sizes, with the aim that these are more likely to give stable leapfrog integration than their rejected predecessors. The result is a form of step size adaptation, which has been very successful for numerical integration more generally.

One of the major concerns for delayed rejection approaches with random-walk Metropolis is that for a kth order scheme, the number of proposals (and hence the cost) per iteration before final rejection increases is $\mathcal{O}(2^k)$. However we show that a properly tuned delayed rejection HMC implementation is only a constant factor more expensive for each iteration than an HMC proposal *tuned locally*, which violates the fixed step size assumption of standard HMC. In the general delayed rejection method, the second and subsequent proposals may depend on the earlier proposals (Green and Mira, 2001). We exploit this property in proposing a probabilistic form of delayed rejection where we only retry if the original proposal had a low acceptance probability.

In the rest of this paper, we begin with background on Metropolis-Hastings, delayed rejection for random-walk Metropolis, and HMC in Section 2. We present DR-HMC for one or more proposals in Section 3. We show that DR-HMC obeys detailed balance,

discuss the cost of delayed proposals, and outline probabilistic delayed rejection to reduce the cost of retries. In Section 4, we consider some simple example target densities as well as actual Bayesian posteriors and show that DR-HMC can provide significant speedups compared to traditional HMC in sampling multiscale distributions. We also show that in cases with no such pathologies, probabilistic DR-HMC is no more expensive than HMC, which suggests that we can use DR-HMC as a robust alternative to traditional HMC. We conclude with discussion in Section 5. A short appendix reviews a proof useful for understanding HMC and DR-HMC.

2 Metropolis-Hastings, delayed rejection, and Hamiltonian Monte Carlo

In this section we recap necessary background material and set up notation. We refer the reader to Geyer (2011) and Neal (2011) on MCMC, to Andrieu et al. (2020) for a more rigorous framework, and to MacKay (1998), Sohl-Dickstein et al. (2014), and Betancourt (2017) for more conceptual overviews. For measure theory, see Stein and Shakarchi (2005), Hunter and Nachtergaele (2001), and Billingsley (2012).

We use $x \in S$ to denote the state in a continuous state space $S = \mathbb{R}^n$. For HMC sampling of a target density over $\mathbb{R}^d$, we will have $n = 2d$ after coupling momentum with position. Our goal is to sample from a target measure π over S, which we assume has strictly positive density with respect to the Lebesgue measure dx. As is common, we also use π to denote this probability density function (pdf). Unless indicated, all integrals are over S. All subsets of S (i.e., events) considered will be assumed to be Borel measurable. We assume the normalization $\int \pi(x)\, dx = 1$, although all MCMC methods we will discuss only require unnormalized densities.

2.1 Metropolis-Hastings

Given π, the random-walk Metropolis-Hastings algorithm constructs a Markov chain with kernel K such that $\pi K = \pi$, i.e., the target distribution π is invariant. Recall that, in general, the action of a kernel K on a measure π is $(\pi K)(A) := \int \pi(dx)K(x, A)$, for any (measurable) subset $A \subset S$. If the detailed balance condition

$$\int_A \pi(dx)K(x, B) = \int_B \pi(dx)K(x, A), \qquad \text{for all subsets } A, B \subset S \tag{1}$$

holds, then it is easy to establish (by choosing $A = S$) that π is invariant. The converse need not hold. Detailed balance is also known as reversibility, because it is equivalent to K being self-adjoint with respect to $\mathrm{L}^2_\pi(S)$. Detailed balance is equivalent to the $x \leftrightarrow y$ symmetry of the product measure over $S \times S$,

$$\pi(dx)K(x, dy) = \pi(dy)K(y, dx). \tag{2}$$

Given the current state x, one samples y from a proposal distribution $Q(x, \cdot)$. The proposal is then accepted with some probability $\alpha(x, y)$, in which case the next state is

y, otherwise the next state remains x. The resulting Markov chain transition kernel is

$$K(x, dy) = \alpha(x, y)Q(x, dy) + r(x)\delta_x(dy), \tag{3}$$

where $r(x) := 1 - \int \alpha(x, y)Q(x, dy)$ is the probability of rejection at x. Here δ_x is the unit Dirac measure at x, defined by $\delta_x(A) = 1$ if $x \in A$ or 0 otherwise.[1]

Since the second term in (3) is $x \leftrightarrow y$ symmetric whatever the form of $r(x)$, Metropolis-Hastings thus obeys detailed balance if the remaining condition

$$\pi(dx)Q(x, dy)\alpha(x, y) = \pi(dy)Q(y, dx)\alpha(y, x) \tag{4}$$

holds. In the classical random-walk Metropolis case where $Q(x, \cdot)$ is absolutely continuous with respect to Lebesgue measure, for each $x \in S$, then $Q(x, dy) = q(x, y)dy$ where $q(x, y)$ is some transition function. In that case, the condition is simply that the function $\pi(x)q(x, y)\alpha(x, y)$ is $x \leftrightarrow y$ symmetric. If $q(x, y)$ is everywhere positive, this condition is satisfied by the standard Metropolis-Hastings acceptance formula for $x, y \in S$,

$$\alpha(x, y) = \min\left(\frac{\pi(y)\, q(y, x)}{\pi(x)\, q(x, y)}, 1\right), \tag{5}$$

where the denominator is never zero given the above assumptions on π and q. For each $x, y \in S$, either $\alpha(x, y)$ or $\alpha(y, x)$ is 1. There are other formulae for α obeying (4), but with lower acceptance rates; by the Peskun ordering, they lead to higher asymptotic variance for estimated expectations (Tierney and Mira, 1999, Thm. 4), which is undesirable.

2.2 Delayed rejection for Metropolis-Hastings

Here we summarize standard delayed rejection as introduced by Mira (1998), Tierney and Mira (1999), and Green and Mira (2001). The idea is that if the first proposal distribution $Q_1(x, \cdot)$ from x to s is rejected, one tries a second proposal with distribution $Q_2(x, s, \cdot)$ to y; see Figure 3(a), which is analogous to Figure 1 of Green and Mira (2001). Note that Q_2 is a measure parameterized by both the current state x and the rejected state s. The transition kernel analogous to (3) must account for three possible ways to end up at state y: i) acceptance of $Q_1(x, dy)$, for which one uses the usual Metropolis-Hastings probability $\alpha_1(x, y)$ obeying detailed balance (4); ii) acceptance of the second proposal, which occurs with some new probability $\alpha_2(x, s, y)$; and iii) rejection of this second proposal with no change in state. For cases (ii) and (iii) one must marginalize over all possible rejected first tries s.

Thus the transition kernel from x to y is

$$K(x, dy) \;=\; Q_1(x, dy)\alpha_1(x, y)$$

[1] This measure theory notation $\delta_x(dy)$ is equivalent to $\delta_x(y)dy$ or $\delta(x-y)dy$ in the standard notation for Schwartz distributions in applied mathematics (Hunter and Nachtergaele, 2001, Ch. 11) (Lieb and Loss, 2001, Ch. 6).

$$+\int_{s\in S} Q_1(x,ds)[1-\alpha_1(x,s)][Q_2(x,s,dy)\alpha_2(x,s,y)+r_2(x,s)\delta_x(dy)], \quad (6)$$

where r_2 is the probability of rejection of the second proposal.[2] Since $1-\alpha_1(x,s)$ is the probability of rejection of the first proposal, the kernel $Q_1(x,ds)[1-\alpha_1(x,s)]$ in the integrand is the distribution of rejected s values. The goal is then to choose $\alpha_2(x,s,y)$ such that detailed balance is satisfied for the kernel $K(x,dy)$ given by (6). We have already established that this holds for the first term and the r_2 term, which leaves only the middle term involving α_2 itself. Substituting this middle term into the detailed balance condition (4) gives

$$\int_{s\in S} \pi(dx)Q_1(x,ds)[1-\alpha_1(x,s)]Q_2(x,s,dy)\alpha_2(x,s,y) \quad (7)$$

$$= \int_{s'\in S} \pi(dy)Q_1(y,ds')[1-\alpha_1(y,s')]Q_2(y,s',dx)\alpha_2(y,s',x), \quad (8)$$

where s and s' may be viewed as (unrelated) dummy variables labeling rejected states. Recall that the above is to hold in the sense of product measures in (x,y). As explained in (Mira, 1998, Sec. 5.2), one (but not the only) way to enforce this condition is simply to set $s'=s$ and equate the integrands.[3] This gives

$$\begin{aligned}&\pi(dx)Q_1(x,ds)[1-\alpha_1(x,s)]Q_2(x,s,dy)\alpha_2(x,s,y)\\ &\quad= \pi(dy)Q_1(y,ds)[1-\alpha_1(y,s)]Q_2(y,s,dx)\alpha_2(y,s,x),\end{aligned} \quad (9)$$

in the sense of product measures over $(x,y,s)\in S\times S\times S$. In the special case of absolutely continuous and everywhere-positive proposal kernels, q_j denoting the density function for kernel Q_j, $j=1,2$, the acceptance probability for the second proposal that obeys (9) with the least rejection is that of Tierney and Mira (1999),

$$\alpha_2(x,s,y)=\min\left(\frac{\pi(y)q_2(y,s,x)q_1(y,s)[1-\alpha_1(y,s)]}{\pi(x)q_2(x,s,y)q_1(x,s)[1-\alpha_1(x,s)]},1\right). \quad (10)$$

Compared to (5), this has an extra factor $\frac{q_1(y,s)[1-\alpha_1(y,s)]}{q_1(x,s)[1-\alpha_1(x,s)]}$, which is the ratio of probabilities of the first proposal being rejected from y and x respectively.

2.3 Classical Hamiltonian Monte Carlo

In this section, we outline the Hamiltonian Monte Carlo (HMC) sampling algorithm. We overload q following standard notation for Hamiltonian dynamics, so that $q\in\mathbb{R}^d$ also denotes the parameter of interest that is to be sampled.[4] The target density, which we now write as $\tilde{\pi}$, is assumed to be continuous, differentiable, and everywhere positive.

[2] As before with r, its form will be irrelevant because the measure $\delta_x(dy)$ is symmetric, so will not affect detailed balance.

[3] Alternatively, one can assume that there exists a differentiable and invertible mapping from (x,s,y) to (y,s',x), and apply a change of variables to identify a more generic acceptance equation that is not constrained to follow the same path via s from y to x, as in (Green and Mira, 2001).

[4] Confusion should not arise between uses of q as a parameter and a proposal density.

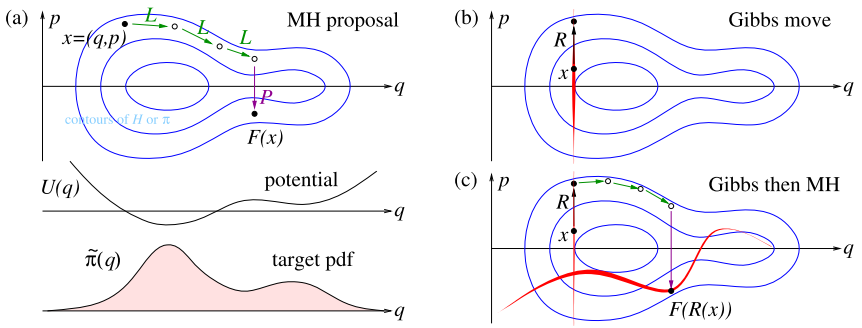


Figure 2: Overview of HMC, sketched in $d = 1$ dimensions. (a) shows the target density $\tilde{\pi}(q)$ (bottom), the associated potential $U(q)$ (middle), and the resulting contours (i.e., level sets) in the two-dimensional phase space (q, p) of the Hamiltonian H given by (11). Each leapfrog step L moves approximately along such a contour. For step 2 of HMC, the proposal move $F = L_\varepsilon^n P$ is sketched, for $n = 3$, where P is the momentum flip. (b) shows the p randomization (Gibbs move) in step 1 of HMC (red shows density of the kernel living on the d-dimensional slice $q =$ constant). (c) shows the composition of steps 1 and 2, comprising one HMC iteration (again red shows the resulting Markov kernel density, which lives on the union of a curved d-dimensional manifold and a constant-q slice).

To draw samples q from $\tilde{\pi}(q)$, HMC reinterprets the parameters of interest as a position vector with associated potential energy function $U(q) = -\log \tilde{\pi}(q)$. One introduces an auxiliary momentum vector $p \in \mathbb{R}^d$, which contributes a kinetic energy term $K(p) = \frac{1}{2} p^T M^{-1} p$, where M is some symmetric positive definite mass matrix that we take as fixed. Then the Hamiltonian $H : \mathbb{R}^{2d} \to \mathbb{R}$ gives the total energy for the state $x := (q, p)$,

$$H(x) = H(q, p) = U(q) + \frac{1}{2} p^T M^{-1} p. \tag{11}$$

The state space $S = \mathbb{R}^{2d}$ is called *phase space*; see Figure 2(a) for an illustration. HMC uses a Markov chain to generate samples x from the Gibbs density (also known as the Boltzmann or canonical distribution in statistical mechanics) defined by H, namely

$$\pi(x) := Z^{-1} e^{-H(x)} = Z^{-1} e^{-U(q)} e^{-\frac{1}{2} p^T M^{-1} p} = Z^{-1} \tilde{\pi}(q) e^{-\frac{1}{2} p^T M^{-1} p}, \tag{12}$$

where $Z = \int_{\mathbb{R}^{2d}} e^{-H(x)} \, dx = (2\pi)^{d/2} \sqrt{\det M}$ is the normalizing constant. Since H is the sum of potential and kinetic terms, in the Gibbs density q and p are independent, with the q-marginal of $\pi(x)$ being the target density $\tilde{\pi}(q)$. Thus, given samples $x^{(i)}$ from π, by extracting their first d coordinates one obtains samples from $\tilde{\pi}$.

HMC constructs a Markov chain to generate samples from this distribution. Given a current state $x^{(i)} := (q^{(i)}, p^{(i)})$, the Markov update comprises two steps:

Step 1. Gibbs sampling: Resample the momentum $p^{(i)}$ from its Gaussian marginal

distribution $p \sim \mathcal{N}(0, M)$, without changing $q^{(i)}$.[5] This randomization step is needed to mix efficiently between different H values (energy level-sets). It is shown as R in Figure 2(b). There are variants using partial randomization that we will not explore here (Horowitz, 1991; Neal, 2011; Sohl-Dickstein et al., 2014).

Step 2. Metropolis update: Given the momentum $p^{(i)}$, generate a new Metropolis-Hastings proposal via a deterministic map $y = F(x)$ which approximates Hamiltonian flow for (11) over a certain length of time T, with initial point $x = (q^{(i)}, p^{(i)})$, followed by negation of the final momentum. This is sketched in Figure 2(a). This proposal is then accepted with some probability $\alpha(x, y)$. The transition kernel for Step 2 is thus a mixture two Dirac masses,

$$K(x, dy) = \alpha(x, y)\delta_{F(x)}(dy) + r(x)\delta_x(dy), \tag{13}$$

where $r(x) = 1 - \alpha(x, F(x))$ is the rejection probability.

The key properties needed for the map F are that it be *volume-preserving* (i.e., the Lebesgue measures of A and $F(A)$ are equal, for all subsets $A \subset \mathbb{R}^{2d}$) and an *involution* (meaning that $F^{-1} = F$, i.e., it is *time reversible*). In this case, the kernel (13) of Step 2 preserves detailed balance when

$$\alpha(x, y) = \min\left(\frac{\pi(y)}{\pi(x)}, 1\right), \tag{14}$$

so that π is invariant under Step 2. This result was derived heuristically by Duane et al. (1987) and Neal (2011) and rigorously by Andrieu et al. (2020) (e.g., in Example 14). For convenience, we provide a simple proof in the Appendix (Modi et al., 2023). Since π is also invariant under Step 1, π is also invariant under their composition (i.e., under each HMC update).[6] Note that failure to approximate well the Hamiltonian flow *does not* impact detailed balance, although it can drastically reduce the mixing of the Markov chain, and hence the efficiency.

The most commonly used dynamics for the map F is the leapfrog (Verlet) integrator. Each leapfrog step, written $(q', p') = L_\varepsilon(q, p)$, comprises three substeps:

$$\begin{aligned} \bar{p} &\leftarrow p - \frac{\varepsilon}{2}\nabla U(q)\ , \\ q' &\leftarrow q + \varepsilon M^{-1}\bar{p}\ , \\ p' &\leftarrow \bar{p} - \frac{\varepsilon}{2}\nabla U(q'). \end{aligned} \tag{15}$$

The composition of $n = T/\varepsilon$ such steps, L_ε^n, is a $\mathcal{O}(\varepsilon^2)$-accurate approximation to Hamiltonian dynamics for time T (Leimkuhler and Reich, 2004). With the momentum-flip operator $P(q, p) := (q, -p)$, then $F = L_\varepsilon^n P$ (recalling that operators act leftwards)

[5] We use $\mathcal{N}(\mu, \Sigma)$ for normal distributions with mean vector μ and covariance matrix Σ, and in the univariate case, $\mathcal{N}(\mu, \sigma^2)$ where σ^2 is the variance.

[6] Detailed balance is not preserved under composition and does not hold in the full space (q, p), but in the case of complete momentum randomization it does obey detailed balance when viewed as a Markov chain in q alone, as shown by Mehlig et al. (1992) and in Proposition 1 of Andrieu et al. (2020).

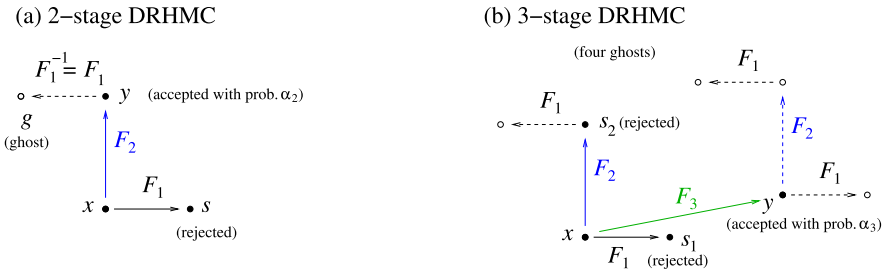


Figure 3: Sketch of states (in the augmented state space $\mathbb{R}^{2d}$) involved in delayed rejection HMC. In this sketch, locations are chosen purely to aid visualisation. (a) Basic 2-stage scheme with proposal maps F_1 and F_2, each of which is a certain number of leapfrog steps followed by a momentum-flip, hence an involution (see Section 3). The first map F_1 generates a proposal $s = F_1(x)$, which is accepted with probability $\alpha_1(x, s)$. If the first proposal is rejected, the second proposal $y = F_2(x)$ is accepted with probability $\alpha_2(x, s, y)$. The target density at the "ghost" $g = F_1(F_2(x))$ is also needed. (b) 3-stage scheme, involving a third proposal map F_3 (see Section 3.1). The target density must be evaluated at $2^3 = 8$ states, four of which are ghosts (shown by open circles).

is volume-preserving because each of the three substeps is a shear transformation,[7] and F is an involution because $L_\varepsilon(q', -p') = (q, -p)$, which can be verified by reversing the order of the substeps, so $L_\varepsilon P L_\varepsilon = P$ and $(L_\varepsilon^n P)^2 = I$. In practice, ε too large leads to instability (large H changes, hence low acceptance rates), whereas small ε is stable but inefficient since n is large. The main point of our DR-HMC proposal will be to locally adapt ε in a somewhat automated manner.

It is worth mentioning that while leapfrog integration (L) of Hamilton's equations is the most commonly used proposal in HMC, it is not the only choice. The leapfrog integrator itself can be extended to higher orders (Creutz and Gocksch, 1989; Yoshida, 1990). Neal points out that even the $\mathcal{O}(\varepsilon)$-accurate modified Euler step is valid (Neal, 2011), and recent works have proposed using other maps, such as implicit integrators (Pourzanjani and Petzold, 2019; Brofos and Lederman, 2021) for multiscale distributions or generalizing HMC with neural networks (Levy et al., 2017).

3 Delayed rejection for HMC

In this section, we introduce delayed rejection Hamiltonian Monte Carlo (DR-HMC). As with standard HMC, we work with the extended state $x = (q, p)$ to sample from the desired distribution $\tilde{\pi}(q)$, which is the marginal of $\pi(x)$, the resulting Gibbs pdf (12) over the extended state. As in Section 2.2, we use s to represent intermediate proposals

[7] A *shear* is a map of the form $(q, p) \mapsto (q + G(p), p)$ or $(q, p + G(q))$, and it is easy to check that the determinant of the $2d \times 2d$ Jacobian derivative matrix is 1.

that have been rejected, and y will always represent the most recent proposal (i.e., the proposal made in the current delayed rejection stage).

We keep the Gibbs step unchanged and apply DR only to the Metropolis step. Consider $F_1 = L_\varepsilon^n P$, a deterministic proposal map for some time step ε and number of leapfrogs n. The first acceptance probability remains the same as in standard HMC, $\alpha_1(x,s) = \min\left[\pi(s)/\pi(x), 1\right]$. If this first proposal with kernel $Q_1(x, ds) = \delta_{F_1(x)}(ds)$ is rejected, it suggests $\pi(s)/\pi(x)$ is much smaller than 1, indicating very poor approximate energy conservation so that ε was too large for stable integration. This motivates a second proposal via a mapping F_2 using a smaller ε; the second kernel, in terms of the final state y, is thus $Q_2(x, s, dy) = \delta_{F_2(x)}(dy)$.

We now derive the detailed balance condition for α_2. We assume only that the maps F_1 and F_2 are volume preserving involutions, which is true for the HMC steps discussed in the previous section. Substituting the above Q_1 and Q_2 kernels into the detailed balance condition (8) for the second proposal in delayed rejection gives

$$\begin{aligned}&\int_{s\in S} \pi(x)dx\,\delta_{F_1(x)}(ds)\,[1-\alpha_1(x,s)]\delta_{F_2(x)}(dy)\,\alpha_2(x,s,y)\\ &\quad= \int_{s'\in S} \pi(y)dy\,\delta_{F_1(y)}(ds')\,[1-\alpha_1(y,s')]\delta_{F_2(y)}(dx)\,\alpha_2(y,s',x)\end{aligned}$$

as product measures on (x, y). We apply the rule for Dirac measures for the s and s' integrals, then note that $dx\,\delta_{F_2(x)}(dy) = dy\,\delta_{F_2(y)}(dx)$ since F_2 is a volume preserving involution.[8] Fixing $y = F_2(x)$, thus the detailed balance condition on α_2 is

$$\pi(x)\,[1-\alpha_1(x,F_1(x))]\,\alpha_2(x,F_1(x),y) \;=\; \pi(y)\,[1-\alpha_1(y,F_1(y))]\,\alpha_2(y,F_1(y),x). \quad (16)$$

Note that the method of setting integrands equal, as done by Mira and Tierney to get (9), would fail here since the left-side Dirac sets $s = F_1(x)$ while the right side sets $s' = F_1(y)$, but F_1 is injective so $s = s'$ could only hold if $x = y$. To maximize the acceptance rate while obeying (16) we choose

$$\alpha_2(x,F_1(x),y) \;=\; \min\left(\frac{\pi(y)}{\pi(x)}\frac{1-\alpha_1\big(y,F_1(y)\big)}{1-\alpha_1\big(x,F_1(x)\big)},1\right), \qquad \text{for } \; y=F_2(x). \quad (17)$$

Since all proposals are deterministic in DR-HMC, it is now useful to simplify notation by folding the known image points into the acceptance probabilities,

$$\widetilde{\alpha}_1(x) \;:=\; \alpha_1(x,F_1(x)), \quad (18)$$

$$\widetilde{\alpha}_2(x) \;:=\; \alpha_2(x,F_1(x),F_2(x)). \quad (19)$$

This allows us to write the second acceptance probability obeying detailed balance as

$$\widetilde{\alpha}_2(x) \;=\; \min\left(\frac{\pi(F_2(x))}{\pi(x)}\,\frac{1-\widetilde{\alpha}_1(F_2(x))}{1-\widetilde{\alpha}_1(x)},1\right). \quad (20)$$

[8] A simple proof is that $\int_{x\in A}\int_{y\in B} dx\,\delta_{F_2(x)}(dy) = \mathrm{vol}(A\cap F_2^{-1}(B)) = \mathrm{vol}(F_2(A)\cap B) = \mathrm{vol}(F_2^{-1}(A)\cap B) = \int_{x\in A}\int_{y\in B} dy\,\delta_{F_2(y)}(dx)$, which holds for all (measurable) subsets $A, B \subset S$.

Note that this is reminiscent of the acceptance relation (10) from plain DR, but setting all the proposal densities q_j to unity. However, we emphasize that its derivation is quite different, requiring care with deterministic maps, and relying on them being volume-preserving involutions. In addition to the initial point of the trajectory x and the two proposals $F_1(x)$ and $F_2(x)$, this rule demands, via $\widetilde{\alpha}_1(F_2(x))$, evaluation of the pdf at a fourth state $g = F_1(F_2(x))$, even though it is never actually proposed. This is the first proposal that would have been made in a hypothetical chain, had we started the chain in the reverse direction, i.e., starting from y to go to x. Hence we call it a *ghost preimage* of the second proposal. See Figure 3(a) for a visual aid.

Finally, we describe the specific maps tested in this paper. We consider delayed rejections which reduce the step size of the leapfrog integrator by a constant *adaptivity factor* $a > 1$ but maintain the same trajectory length (time of integration T). Hence we will propose $F_2 = L^{an}_{\varepsilon/a} P$. In Section 4 we will show that this allows us to explore regions in the phase space that otherwise face persistent rejections with classical HMC. This completes the simplest form of DR-HMC; however, we find that the higher-order proposals described next can also help.

3.1 Higher order proposals

The previous section focused on making a second proposal when the first proposal in HMC gets rejected. The same formalism can be extended to allow a third proposal upon rejection of the second, a fourth upon rejection of the third, and so on. In this section, we explicitly derive the acceptance probability for the third proposal in DR-HMC and give a general recursive relation for the kth proposal. Mira (1998) presents similar acceptance probabilities for higher-order delayed proposals in the Metropolis-Hastings case; see also Andrieu et al. (2020). We also discuss the growth of the cost with number of proposals, since this determines the computational trade-off for the increased acceptance rate of DR-HMC.

Starting from a state $x \in S$, if we reject the first two proposals, denoted now by $s_1 = F_1(x)$ and $s_2 = F_2(x)$, we make a third proposal via a map F_3. The resulting proposal kernel, in terms of y, is $Q_3(x, s_1, s_2, dy) = \delta_{F_3(x)}(dy)$.

In this case, the transition kernel analogous to (6) must account for four possible ways to end up at a state y: i) accepting the first proposal, ii) accepting the second proposal, iii) accepting the third proposal, or iv) rejecting all three proposals and maintaining the current state. We have established the acceptance probabilities of case (i) and (ii) in the previous section. For cases (iii) and (iv), the transition kernel must marginalize over the rejected first and second proposals s_1, s_2,

$$\begin{aligned} K(x, dy) \;=\; & Q_1(x, dy)\alpha_1(x, y) + \int_{s_1 \in S} Q_1(x, ds_1)[1 - \alpha_1(x, s_1)]Q_2(x, s_1, dy)\alpha_2(x, s_1, y) \\ & + \int_{s_1 \in S} \int_{s_2 \in S} Q_1(x, ds_1)Q_2(x, s_1, ds_2)\,[1 - \alpha_1(x, s_1)]\,[1 - \alpha_2(x, s_1, s_2)] \\ & \qquad \times [Q_3(x, s_1, s_2, dy)\alpha_3(x, s_1, s_2, y) + r_3(x)\delta_x(dy)]\,, \end{aligned}$$

where $r_3(x)$ is the probability of rejecting the third proposal from state x.

As we saw in the previous section, since the first and second proposals are independent of the third proposal, their acceptance probabilities α_1 and α_2 are given by (4) and (16) respectively, so the first two terms of (21) maintain detailed balance. As with r_2, since r_3 lies on the diagonal, it will also maintain detailed balance regardless of its form. Thus to maintain detailed balance for the third proposal, a similar derivation as above gives, using the notation $\widetilde{\alpha}_3(x) := \alpha_3(x, F_1(x), F_2(x), F_3(x))$, a third acceptance probability

$$\widetilde{\alpha}_3(x) = \min\left(\frac{\pi(y)[1-\widetilde{\alpha}_1(y)][1-\widetilde{\alpha}_2(y)]}{\pi(x)[1-\widetilde{\alpha}_1(x)][1-\widetilde{\alpha}_2(x)]}, 1\right), \qquad \text{where } y = F_3(x). \tag{21}$$

Continuing in this way, one can write down a recursive relation for the acceptance probability of the kth proposal obeying detailed balance,

$$\widetilde{\alpha}_k(x) = \min\left(\frac{\pi(y)\prod_{i=1}^{k-1}[1-\widetilde{\alpha}_i(y)]}{\pi(x)\prod_{i=1}^{k-1}[1-\widetilde{\alpha}_i(x)]}, 1\right), \qquad \text{where } y = F_k(x), \tag{22}$$

for which we use the notational shorthand $\widetilde{\alpha}_k(x) := \alpha_k(x, F_1(x), \ldots, F_k(x))$.

Growth in cost with respect to the number of proposals

It can be tempting to keep making successively higher order proposals to increase the acceptance rate; however, it is important to be mindful of their increasing cost. With this aim, we explicitly write down the full form of $\widetilde{\alpha}_1$ and $\widetilde{\alpha}_2$ in the acceptance probability for the third proposal, to see the various ghost preimages that need to be evaluated. Recall that (20) gives

$$\widetilde{\alpha}_2(x) = \min\left(\frac{\pi(F_2(x))[1-\widetilde{\alpha}_1(F_2(x))]}{\pi(x)[1-\widetilde{\alpha}_1(x)]}, 1\right),$$

where for any $x \in S$ mapping to $y = F_1(x)$, the standard HMC acceptance probability is

$$\widetilde{\alpha}_1(x) = \min\left(\frac{\pi(F_1(x))}{\pi(x)}, 1\right).$$

Substituting these forms in (21), we see that the denominator involves estimating the density at points x, $F_1(x)$, $F_2(x)$ and $F_1(F_2(x))$, while the numerator requires the density at $F_3(x)$, $F_1(F_3(x))$, $F_2(F_3(x))$, and $F_1(F_2(F_3(x)))$. Of these $2^3 - 1 = 7$ points, only $F_1(x)$, $F_2(x)$ and $F_3(x)$ are proposals made in DR-HMC; there are an additional four ghost preimages that need to be evaluated to maintain detailed balance. This is sketched in Figure 3(b). Evaluating the acceptance probability for the kth proposal requires 2^k density evaluations.

For both HMC (and DR-HMC), computational cost is dominated by the gradient evaluations in leapfrog steps which are considerably more in number than the density

evaluations for calculating acceptance probabilities.[9] Despite this exponential growth in cost as stepsize decreases, we expect the actual cost of DR-HMC to be only a *constant* factor larger than classical HMC when a globally stable step size is chosen for the latter..[10] Depending on the distribution, it is possible that some regions with high curvature might require smaller step size than the optimal one to avoid divergences, for e.g. in the neck of Neal's funnel. In such cases other regions of the distribution and hence globally the algorithm can have more than $\sim 68\%$ acceptance rate. To see this, consider a target distribution wherein ε_0 is the largest step size that gives stable leapfrog integration at all points in the domain. To sample this distribution with DR-HMC, consider a setup where the first proposal is F_1 with n leapfrog steps of size ε, i.e., a trajectory time of $T = n\varepsilon$, and a sequence of step sizes $\varepsilon/a, \varepsilon/a^2, \ldots$ for subsequent higher order proposals upon rejection. For maximal gains, we will choose $\varepsilon > \varepsilon_0$. Furthermore, for a *properly tuned* implementation, we would like to stop making any higher order DR-HMC proposals once the step size falls below the stable stepsize for HMC, i.e., we make k proposals with k being the smallest integer such that $\varepsilon_0 > \varepsilon/a^{k-1}$.

Then for this stable step size ε_0 and same trajectory time T, standard HMC would need between $a^{k-2}n$ and $a^{k-1}n$ leapfrog steps per proposal. On the other hand for DR-HMC, in the worst-case that we are forced to make all the k iterations at every proposal, the total number of leapfrog steps needed is $2^{k-1}n + 2^{k-2}an + \cdots + a^{k-1}n$, where the first term is for the F_1 maps, the second for the F_2 maps, etc. (See Figure 3(b) for the $k = 3$ case.) This sum is nka^{k-1} for $a = 2$ and is order $\mathcal{O}(a^{k-1}n)$ for $a > 2$. Thus compared to the lower bound of $a^{k-2}n$ steps of standard HMC, a properly tuned DR-HMC algorithm cost is only $\mathcal{O}(ak)$ more for $a = 2$ and $\mathcal{O}(a)$ (independent of k) for $a > 2$. In practice, the cost of DR-HMC is expected to be lower, because not every proposal will require a maximum number of retries. If DR-HMC is not properly tuned and we continue to make further proposals after the step size $\varepsilon/a^k \ll \varepsilon_0$, then DR-HMC can be arbitrarily more expensive than HMC. We summarize this discussion as follows.

Remark 1. *For a kth order DR-HMC algorithm that is properly tuned in the sense that the last proposal has step size comparable to the stable step size of classical HMC, the cost per proposal as measured in terms of gradient evaluations is only a constant factor more expensive than HMC. This is true even though the number of ghost preimages grows exponentially in k.*

3.2 Probabilistic delayed rejection

One way to reduce the average cost per iteration for DR-HMC is to retry only when the first proposal had a high chance of rejection, which we can formulate within the general framework of delayed rejection as a retry that is dependent on the first try. To motivate

[9] With automatic differentiation, the log density evaluations come for free with the gradient calculations (Griewank, 1988).

[10] By stable step size here we do not mean what is sometimes referred to as optimal step size for HMC, i.e., the optimal 68% global acceptance rate under one set of conditions (Beskos et al., 2013) or the higher rate required for hierarchical models (Betancourt and Girolami, 2015).

how this can be helpful, consider a case when the cost of a secondary proposal is much higher than that of the first proposal, and, even though the first proposal function is well tuned for most of the state space, there are certain hard regions which can only be sampled by the second proposal. In this scenario, while we need DR to correctly sample the full distribution, we do not need it throughout phase space. Every time we make a secondary proposal upon getting a rejection in the easy regions, we might not be trading excess cost with higher acceptance rate effectively. Thus instead of making the second (and subsequent higher order) proposal automatically upon a rejection, we would like to make them probabilistic such that we make a second proposal with some probability $p_2(x,s) < 1$. This modifies the second proposal kernel to

$$Q_2(x,s,dy) = p_2(x,s)\,\delta_{F_2(x)}(dy) + (1-p_2(x,s))\,\delta_x(dy).$$

As was the case for the second proposal map F_2, this probability can also be informed by the previously rejected proposals in the same trajectory. Amending the proof of detailed balance in the previous section shows that the acceptance probability becomes

$$\widetilde{\alpha}_2(x) = \min\left(\frac{\pi(y)\,[1-\widetilde{\alpha}_1(y)]\,p_2\big(y,F_1(y)\big)}{\pi(x)\,[1-\widetilde{\alpha}_1(x)]\,p_2\big(x,F_1(x)\big)},1\right), \qquad \text{where } y = F_2(x). \tag{23}$$

Returning to the scenario outlined above, we see that one way to avoid secondary proposals in easy regions is to construct a proposal probability that makes it less likely for a secondary proposal if the first proposal was rejected on random chance despite having high acceptance probability. On the other hand, if the first proposal was rejected strongly, which might indicate that we are in a hard region of the state space for the first proposal, we make it more likely to make a subsequent proposal with a new function. A simple heuristic proposal probability to achieve this is

$$p_{j+1}(x) = 1 - \tilde{\alpha}_j(x), \tag{24}$$

recalling the notation of (18)-(19). Detailed balance is maintained by including this factor p_{j+1} in the acceptance condition α_{j+1} for the $(j+1)$th proposal along the lines of (23). In the next section, we show how probabilistic delayed rejection preserves the efficiency of standard HMC for simple distributions where standard HMC is effective.

4 Experiments

In this section, we compare the performance of delayed rejection HMC (DR-HMC) to that of standard HMC.

4.1 Setup

Given a current state x, HMC makes a proposal $y = F_1(x)$ where $F_1 = L_\varepsilon^n P$ is the deterministic mapping that integrates Hamiltonian dynamics with leapfrog integration for n steps and step size ε. In DR-HMC, we consider the first proposal to be the same as in HMC. Upon rejection of the first proposal, we make $k-1$ subsequent proposals. For

each of these, we reduce the step size by a fixed factor $a > 1$ while increasing the number of steps in proportion, to maintain a constant integration time. This corresponds to a deterministic mapping,

$$F_k = L^{na^{k-1}}_{\varepsilon a^{-(k-1)}} P.$$

For every experiment and configuration, we run 50 chains with 1000 iterations to reduce initialization bias (i.e., burn-in). This is followed by 20,000 sampling iterations except in Neal's funnel experiment where we sample 50,000 points.

Choice of parameters

HMC has three tuning parameters, the step size ε, the number of leapfrog steps n, and the mass matrix M. The total integration time is $T = n\varepsilon$.

We use Stan (Stan Development Team, 2022) to tune the reference values of these parameters using the following two steps.

1. We use the no-U-turn sampler (NUTS) (Hoffman and Gelman, 2011) to select the integration time T. NUTS is an adaptive algorithm that automatically stops every leapfrog trajectory when it starts to double back and retrace its steps and biases draws along the trajectory to later in the trajectory in an attempt to maximize expected squared jump distance. Therefore, NUTS does not require tuning for T during the warm-up phase. Following (Wu et al., 2018), we choose time of integration T to be the 90th percentile of the trajectories followed by NUTS.[11]

2. After fixing T, we re-run Stan with HMC to estimate the optimal step size ε_{f} and a diagonal mass metric, M.

In addition to the HMC tuning parameters for integration time and step size, DR-HMC has tuning parameters k for the total number of subsequent proposals made and a for the divisor by which step size is reduced for every subsequent proposal. To develop an understanding of how these parameters impact the performance of DR-HMC, we report results for the grid of configurations with $k \in \{2, 3, 4\}$ and $a \in \{2, 5, 10\}$.

With its ability to reduce step sizes in subsequent proposals, DR-HMC is more robust to the initial tuning of step size. To demonstrate this, we evaluate HMC and DR-HMC with fixed and initial step sizes at, above and below the adapted step size, $\varepsilon_0 = 0.5\varepsilon_{\mathrm{f}}, \varepsilon_{\mathrm{f}}, 2\,\varepsilon_{\mathrm{f}}, 5\,\varepsilon_{\mathrm{f}}$.

Metric of comparison

To measure sampling performance, we report the number of log density and gradient evaluations required per effective draw, that is,

$$\mathcal{C} = \frac{N_{\mathrm{eval}}}{\mathrm{ESS}}, \tag{25}$$

[11] Unlike (Wu et al., 2018), we do not jitter the number of leapfrog steps.

where N_{eval} is the total number of log density and gradient evaluations in the Markov chain, and ESS is the estimated effective sample size for an expectation being estimated (i.e., a parameter estimate). Log density and gradient evaluations dominate the cost of HMC, allowing us to ignore other implementation details in comparisons. Thus $\mathcal{C}$ is the inverse of efficiency; smaller $\mathcal{C}$ is better. Its value will depend on the expectation being evaluated, so we report results for posterior means of parameters θ and their squares θ^2, the latter of which measures performance in estimating variance.

If $\rho_t \in (-1, 1)$ is the autocorrelation of a quantity in the Markov chain at lag t, the effective sample size is

$$\text{ESS} = \frac{N}{1 + 2\sum_{t=1}^{\infty} \rho_t}, \tag{26}$$

where N is the total number of iterations (Geyer, 2011). Standard errors for estimating parameters are then derived from the MCMC central limit theorem as

$$\text{se} = \frac{\text{sd}}{\sqrt{\text{ESS}}},$$

where sd is posterior standard deviation. The MCMC central limit theorem states that as effective sample size grows, errors approach a normal distribution,

$$\widehat{\theta} - \theta \sim \mathcal{N}(0, \text{se}^2).$$

This is usually a reasonable approximation even for modest effective sample sizes. Alternatively, if we know the true posterior mean value θ, we can run independent Markov chains and calculate errors $\widehat{\theta} - \theta$. The sample standard deviation of the errors can be used to estimate se, from which we can back out effective sample size as

$$\text{ESS} = \left(\frac{\text{sd}}{\text{se}}\right)^2.$$

In the following experiments, depending on whether we have access to the true parameter distributions, we will show results in terms of cost per effective sample calculated by autocorrelation length ($\mathcal{C}_r$) or estimated through errors in cases where posterior means and variances are known ($\mathcal{C}_c$). In experiments with more than one parameter being sampled, we will show the cost for the parameter that mixes the slowest in the sense of having the lowest effective sample size. We run multiple Markov chains and measure per-chain variation in cost by applying the bootstrap technique across chains.

4.2 Neal's funnel

We begin our experiments with the problem of sampling the funnel density defined by Neal (2003), which is defined for a variance σ^2, log scale β, and coefficients $\alpha_2, \ldots, \alpha_d$ by

$$\beta \sim \mathcal{N}(0, \sigma^2),$$

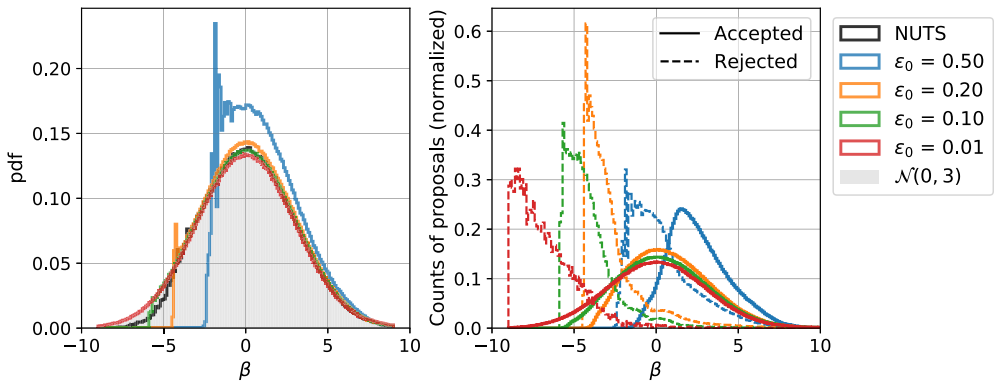


Figure 4: Histograms of draws of the log scale β for Neal's funnel in $d = 20$ dimensions. (Left) The marginal for β, which is $\mathcal{N}(0, 3^2)$ as shown in gray, when the funnel is sampled with NUTS (default settings and HMC for different step sizes. (Right) Normalized (to unit area) histogram of the number of accepted and rejected proposals for HMC as a function of β for different step sizes. All runs are done for 50,000 samples and histograms are generated with bin width of 0.1.

$$\alpha_i \sim \mathcal{N}(0, e^{\beta}), \quad i = 2, 3, \ldots, d, \tag{27}$$

where $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2. Following Neal (2003), we set $\sigma = 3$, so that the resulting target probability density function is

$$\pi(\beta, \alpha_2, \ldots, \alpha_d) = \mathcal{N}\big(\beta \mid 0, \sigma^2\big) \prod_{i=2}^{d} \mathcal{N}\big(\alpha_i \mid 0, e^{\beta}\big)\,.$$

This distribution has equal probability mass in the regions $\beta < 0$ and $\beta > 0$. However the distribution has a wide range of length scales due to the curvature changing as β ranges from large to small values (see Figure 1). This makes it challenging to sample the funnel efficiently with a constant step size. We can illustrate this with the help of Figure 4 which shows the empirical marginal of parameter β when sampling the funnel in $d = 20$ dimensions with NUTS and HMC for different settings. The correct marginal for β is $\mathcal{N}(0, 3^2)$ which means that about $\sim 5\%$ of samples should lie at $\beta < -5$. However even for $\varepsilon_0 = 0.2$, there are no samples in this regime. For NUTS to push to $\beta < -5$, the step size had to be reduced such that 99% of all proposals are accepted, as compared to 80% default value of Stan and 65% fraction considered optimal for normal distributions in HMC (Beskos et al., 2013). To explore beyond the 3σ region, as required to get the right results for modest tail statistics, we need to reduce the step size further to $\varepsilon = 0.01$.

As the step size is reduced, it takes proportionally more iterations to sample the mouth of the funnel, to the point where it becomes impractical because of the poor conditioning. As β grows, the marginals $p(\alpha_k)$, for $k = 2, \ldots d$, approach a lognormal

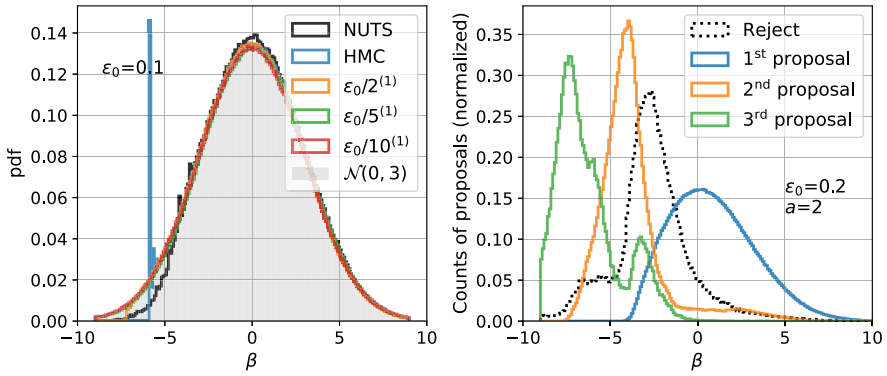


Figure 5: DR-HMC for Neal's funnel in 20 dimensions. (Left) Histograms of draws for the log scale parameter β sampled by NUTS, HMC and DR-HMC. With DR-HMC, an initial step size $\varepsilon_0 = 0.2$ with a single retry reducing step by factors of $a = 2, 5, 10$. (Right) Normalized (to unit area) histogram of the number of accepted proposals at different stages of DR-HMC and rejections (after the third proposals) as a function of β. Results are for DR-HMC configuration with the first step size $\varepsilon_0 = 0.2$ and step sizes reduced by a factor of $a = 2$ at every stage. All runs are done for 50,000 draws and histograms are generated with a bin width of 0.1.

distribution with $\sigma = 3$, and thus have long tails. The expected value of α^2 is on the order of 10^2, whereas the expectation of α^4 is on the order of 10^8. Due to the scale of the mouth of the funnel, the optimal step size is much larger than the $\varepsilon = 0.01$ required to sample the neck of the funnel.

Figure 4 provides an illustration of how well HMC can cover the mouth and neck of the funnel based on step size (left panel). In the right panel, we show the marginal densities of accepted and rejected proposals for various step sizes. All of the step sizes are able to sample the mouth of the funnel, however inefficiently, but the neck of the funnel is hard to explore; to sample in the tail, where $\beta < 3 \cdot \text{sd}[\beta]$, we need to reduce step size even further below $\varepsilon = 0.01$. Comparing the solid and dashed lines, one can immediately make three important observations: i) for every step size, the ratio of rejections to acceptances increases significantly as we go into the neck of the funnel, ii) there is a sharp cutoff below which HMC is even unable to explore the distribution, and iii) this threshold is pushed deeper into the neck of the funnel with decreasing step size.

These observations motivate us to construct a DR-HMC scheme that combines efficient exploration of the mouth of the funnel using large step sizes, while decreasing the step size upon rejections (which are primarily in the neck) to push deeper in the neck. Figure 5 shows how this mitigates the sharp cutoff in the neck of the funnel. The left plot shows the marginal density $p(\beta)$ sampled with NUTS and HMC for step size $\varepsilon_0 = 0.1$, as well as one-retry DR-HMC with different step size reduction factors, a. When the second proposal step size is reduced by factor of 10, DR-HMC is able to

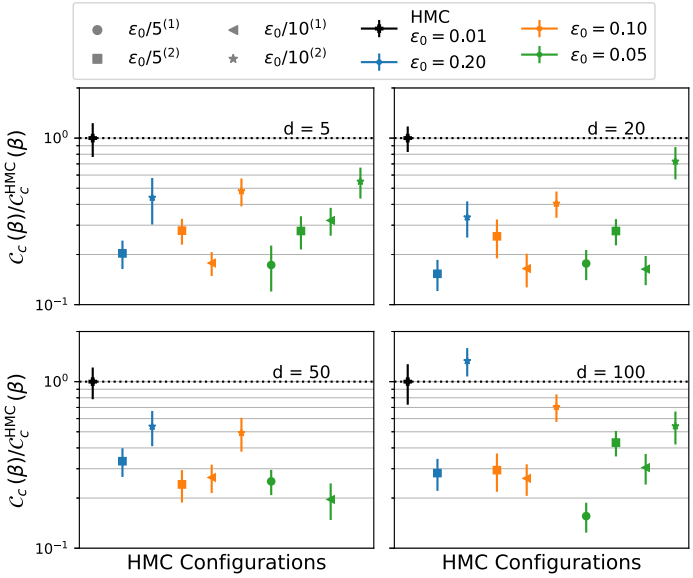


Figure 6: Efficiency of DR-HMC for Neal's funnel. These plots show the ratio of cost per effective sample of the log scale β for DR-HMC and HMC when sampling Neal's funnel for different dimensions ($d = 5, 20, 50, 100$). ESS for the cost is estimated using standard error with reference samples (26). Black points and horizontal dotted black lines show the reference ratio (=1 for HMC with $\varepsilon = 0.01$). Different colors show the cost for DR-HMC with different first step sizes ε_0; smaller is better. Different shapes correspond to different configurations (number of proposals k and reduction factor a). We show only configurations with $\varepsilon_{\min} \leq 0.01$. Error-bars are estimated by bootstrapping over chains.

sample β to $\pm 3\sigma$. The right panel explores a DR-HMC setup that starts with a larger step size ($e_0 = 0.2$), and allows multiple retries with a reduction of $a = 2$. This setup is also able to sample β to $\pm 3\sigma$. We show the density (distribution) of acceptances for the first, second, and third proposals, as well as rejections admitted after the three proposals. As we would have expected based on the right panel of Figure 4, the largest step size is sufficient to explore the mouth of the funnel and successive proposals are rarely made. However to explore deeper in the neck, larger step sizes are rarely accepted and we need second and third proposals with smaller step size for acceptance. In this simple example, we can use these cut-offs to guide the choices of hyperparameters ε_0, k, and a.

Figure 6 illustrates the efficiency gain of DR-HMC over HMC for Neal's funnel. We show the cost per effective sample of β for funnels of dimensions $d = 5, 10, 20, 100$. ESS_r can be biased to the high side because it only depends on the autocorrelation length of the chain and not that it is sampling the correct stationary distribution. Thus we use square errors to estimate effective sample size (ESS_c). This requires reference samples from the distribution, which are simple to generate independently using a non-centered parameterization of the funnel (Betancourt and Girolami, 2015).

Figure 6 shows that DR-HMC is consistently a factor of 4 more efficient than HMC in terms of log density and gradient evaluations required for a given effective sample size; in some configurations the advantage is as much as a factor of 8. We restrict attention to configurations for which HMC is able to sample β to plus or minus three standard deviations (i.e., $\varepsilon \leq 0.1$). For all configurations, we used simulation-based calibration tests (Gelman et al., 2020) at different quantiles of the distribution to ensure we are sampling both, the body and the tails of the distribution correctly.

4.3 Eight schools model

One of the first applications of Bayesian hierarchical modeling was a meta-analysis of the effects of a test preparation intervention on students in eight schools (Rubin, 1981). The data consists of the differences in pre-test and post-test scores, which are reported as an average y_n and standard deviation σ_n for each school n. The hierarchical model uses parameters θ_n for the efficacy in each school and assigns them a hierarchical normal prior with unknown location μ and scale τ. Gelman et al. (2013) provide the generative model[12]

$$\begin{aligned} \mu &\sim \mathcal{N}(0, 5^2) & \qquad \tau &\sim \text{Cauchy}_+(0, 5) \\ \theta_n &\sim \mathcal{N}(\mu, \tau^2) & \qquad y_n &\sim \mathcal{N}(\theta_n, \sigma_n^2). \end{aligned}$$

The hyperparameter μ represents the average treatment effect across schools and τ the scale of variation of effects among schools. As $\tau \to \infty$, the model approaches no pooling (i.e., each of the school treatment effects is estimated independently). As $\tau \to 0$, the model approaches complete pooling (i.e., all of the school treatment effects are the same). For small values of τ, the school-level effects θ_n are squeezed together; for large values, they are allowed to vary widely. This yields a multiscale, funnel-like geometry in the τ and θ parameters, where we would expect delayed rejection to improve the performance of baseline HMC.

Figure 7 evaluates several configurations of the DR-HMC algorithm as applied to the eight schools problem (see 4.1), plotting the cost of each configuration using the standard error method ($\mathcal{C}_c$) for the slowest mixing parameter. We use the reference samples provided with `posteriordb` package[13] to estimate the mean and variance of the parameters as needed to calculate error-based effective sample size (ESS_c). The best DR-HMC configuration improves over the best HMC configuration by a factor of three for estimating the parameter mean. Different configurations for HMC perform the best for the first and second moment, with the cost of second moment estimation by DR-HMC being on par with that of HMC.

4.4 Gull's lighthouse

Challenging posterior geometries arise even in simple two dimensional problems if the data is not very informative. For example, consider estimating the direction of flashes

[12] The positive half-Cauchy distribution uses a location-scale parameterization whereas the normal uses a location-variance parameterization.

[13] https://github.com/stan-dev/posteriordb.

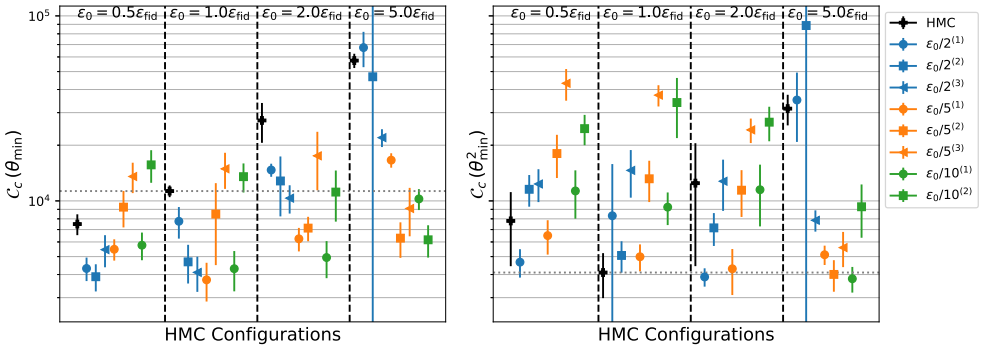


Figure 7: Cost per effective sample for HMC and DR-HMC for the eight schools model. The two panels show the cost for the slowest dimension for the first (θ) and second (θ^2) moments respectively. The cost for HMC is shown in black points, as estimated by using the standard error method for ESS. Different configurations for DR-HMC are shown in different colors (reduction factor, a) and shapes (number of proposals, k). Different step sizes are separated by vertical dashed black lines. Dotted horizontal black line shows the cost for the reference configuration (as fit by Stan) of HMC.

emanating from a coastal lighthouse (Gull, 1988, p. 59). Assume the lighthouse is at position x_0 along a straight coast at distance y into the sea. Its light is spinning and emits a series of collimated flashes at random intervals which are then detected, each at a single point on the coastline. Given N flashes recorded at positions x_i for $i = 1, \ldots, N$, we perform a Bayesian estimation of the position of the lighthouse (x_0, y).

The lighthouse flashes in a random direction θ, relative to vertical, drawn from a uniform distribution on $(-\pi/2, \pi/2)$. Such a flash will be observed at location x_i on the coast, where $\theta = \arctan((x_i - x_0)/y)$. Applying a change of variables, the likelihood of observing a flash is

$$\begin{aligned} p(x_i \mid x_0, y) &= \frac{y}{\pi(y^2 + (x_i - x_0)^2)} \\ &= \text{Cauchy}(x_i \mid x_0, y). \end{aligned}$$

With improper uniform priors on x_0 and y, and the assumption that the flashes are independent, the posterior is proportional to the product of observation likelihoods,

$$p(x_0, y \mid \{x_i\}) \propto \prod_i \text{Cauchy}(x_i \mid x_0, y).$$

In Figure 8, we show the cost $\mathcal{C}_r$ for the case with $N = 3$ flashes observed at $x_i = 0.9, 1.2, 1.21$, for both the parameters x_0 and y. We estimate ESS by measuring autocorrelation length of the chains since there are no reference samples available for this model. For estimating y, whose effective sample size is an order of magnitude lower than that of x_0 with HMC, DR-HMC is a factor of five more efficient; there are no gains in sampling the parameter x_0 that already mixes well with HMC.

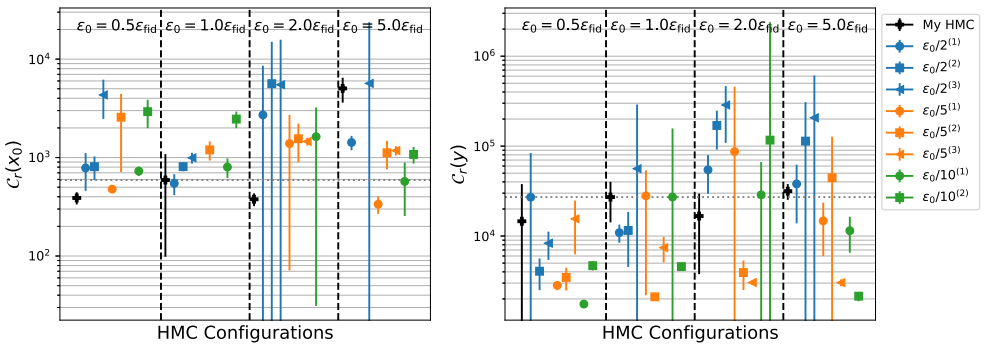


Figure 8: Cost per effective sample for HMC and DR-HMC for Gull's lighthouse model. The two panels show the cost for the two parameters of the model. The figure legend is the same as that of Figure 7. ESS is estimated by measuring autocorrelation length of the chains.

4.5 Gaussian mixture model

Mixture models present problems for samplers with fixed step sizes when the mixture components are of different scales. Multimodal distributions whose components have varying geometries also defeat global tuning for HMC. HMC relies on tuning these parameters before sampling (e.g., Stan first runs a warmup phase that performs adaptation before sampling begins). As in other intrinsically multiscale problems, DR-HMC has the potential to outperform baseline HMC by using different proposal scales in different regions of the state space.

To simulate the situation arising with multivariate posteriors, we consider a univariate Gaussian mixture with equal mixing weights on the components. We take weakly separated locations μ_i that still allow Markov chains to transition between components. The scales σ_i then vary by an order of magnitude. The model pdf is

$$p(\theta) = \sum_{i=1}^{2} \phi_i \cdot \mathcal{N}(\theta \mid \mu_i, \sigma_i^2), \tag{28}$$

where we fix $\phi_1 = 0.5$, $\phi_2 = 0.5$, $\mu_1 = 0$, $\mu_2 = 3$, $\sigma_1 = 0.1$, and $\sigma_2 = 1$. Our goal is to sample the parameter $\theta \in \mathbb{R}$. We choose this simple problem for illustration because sampling mixtures only becomes more challenging in higher dimensions with differently conditioned components, in situations where the modes are either more widely separated or more highly overlapping, or when the weights of the components are highly skewed. The optimal step size for the components is directly proportional to the component's scale, which varies by an order of magnitude.

Figure 9 shows that the best DR-HMC configurations can be twice as efficient as HMC. This gap can be made arbitrarily wide by increasing the number of dimensions and the difference in scales between the modes.

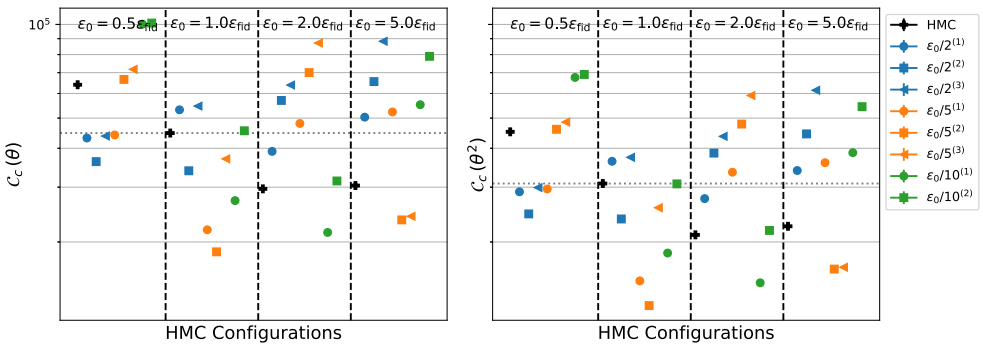


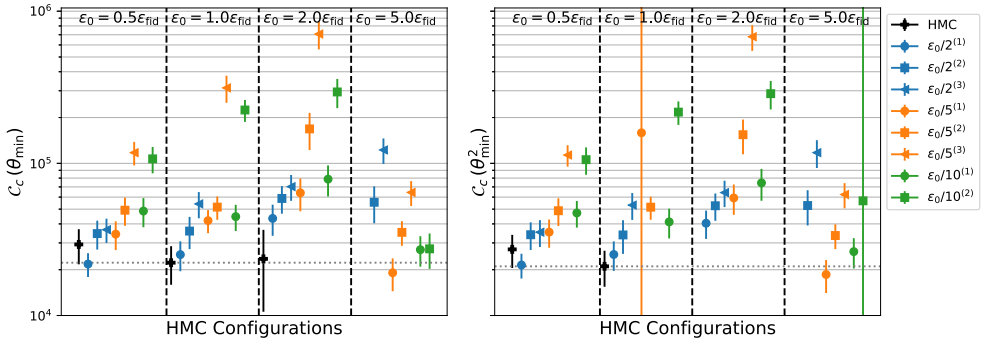
Figure 9: Cost per effective sample for HMC and DR-HMC for the Gaussian mixture model with components varying in scale by a factor of ten. The two panels show the cost for estimating the mean of θ and θ^2, the second of which determines variance. Symbol legends are the same as in Figure 7. ESS is estimated with the standard error method after generating reference samples from the Gaussian mixture model.

4.6 Stochastic volatility model

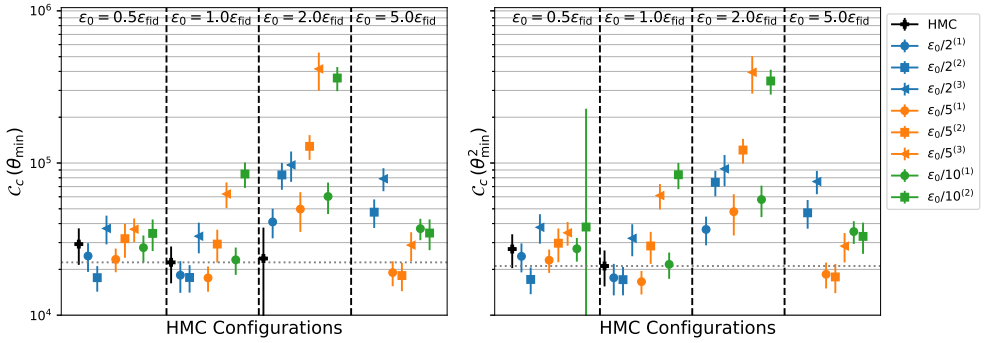
Finally we consider an example that does not suffer from the pathology of multiscale distributions, but is still challenging due to high dimensionality and correlated parameters. Stochastic volatility models seek to model the volatility (i.e., variance) of a return on a financial asset, such as an option to buy a security (Kim et al., 1998). This changing volatility is modeled as a latent stochastic process in discrete time. Given the mean corrected returns y_t on an underlying asset at T equally spaced time points as input data, we are interested in estimating the latent parameter h_t for the log volatility, mean μ and variance σ of log volatility, as well as the persistence of the volatility ϕ. Thus the parameter vector is $q = \{\mu, \sigma, \phi, h_{t=1,\ldots,T}\}$ and the generative model is as follows.

$$\begin{aligned} &\phi \sim \text{uniform}(-1, 1); \quad \sigma \sim \text{Cauchy}(0, 5); \quad \mu \sim \text{Cauchy}(0, 10) \\ &h_1 \sim \mathcal{N}\Big(\mu, \frac{\sigma^2}{1-\phi^2}\Big); \quad h_t \sim \mathcal{N}\big(\mu + \phi(h_{t-1} - \mu), \sigma^2\big), \quad t = 2, 3, \ldots, T \\ &\qquad\qquad y_t \sim \mathcal{N}\big(0, \mathrm{e}^{h_t}\big), \quad t = 1, 2, \ldots, T. \end{aligned}$$

Figure 10(a) shows that the additional computational cost of DR-HMC ends up making it more costly per effective sample than HMC. The cost of DR-HMC is higher because the original step size is optimal, so that retrying with a lower step size only doubles computational costs. This situation tests whether the probabilistic version of DR-HMC described in Section 3.2, in which a proposal is only retried with a probability given by (24), is competitive with standard HMC with a fixed step size. Figure 10(b) shows how retrying with a probability equal to the original failure chance avoids needless step size reduction, allowing DR-HMC to slightly exceed the efficiency of HMC.



(a) Delayed Rejection



(b) Probabilistic Delayed Rejection

Figure 10: Cost per effective sample for HMC and (a, top) DR-HMC and (b, bottom) Probabilisitic DR-HMC for Stochastic Volatility model. The two panels show the cost for the slowest dimension for first (θ) and second (θ^2) moments respectively. Symbol legends are the same as in Figure 7.

5 Discussion

We introduced a novel application of delayed rejection to Hamiltonian Monte Carlo (HMC) sampling in which subsequent proposals are made for the same integration time at a reduced step size. We derived acceptance probabilities that maintain detailed balance for an arbitrary number of delayed proposals. We showed that in multiscale posteriors such as mixture models or hierarchical models, delayed rejection can boost performance by a factor of five or more. We also showed that if the initial step size of a properly tuned DR-HMC is too large, it introduces at most a constant factor additional cost over choosing the stable baseline step size for HMC.

In cases where the target density is not multiscale, we introduced a novel form of delayed rejection where retries are only attempted when the previous proposal had a high chance of failure. Unlike the case for HMC, which will fail with potentially hard-to-diagnose biases, DR-HMC with probabilistic retries is robust to the initial choice of

step size, thus reducing overall costs when tuning step size is expensive.

In realistic problems one often does not know if the target distribution suffers from multiscale or varying geometry pathologies as in the cases we considered. For example, varying amounts of data and varying noise ratios in the data can dramatically change the posterior geometry, changing approximately normal posteriors into funnels or vice-versa, depending on the model parameterization (Papaspiliopoulos et al., 2007; Betancourt and Girolami, 2015). In these cases, we propose that DR-HMC is a better choice than standard HMC since it is more robust to varying curvature of the target distributions. Especially, the probabilistic variant of DR-HMC avoids making extra proposals needlessly at every iteration and hence can often enjoy robustness at the same cost as HMC.

Reducing the step size is not the only way of constructing delayed proposals. Another approach would be to replace the leapfrog integrator altogether for retries, for example with an implicit symplectic integrator (Pourzanjani and Petzold, 2019). Such integrators may additionally be able to deal with stiffness arising from high correlation. Delayed rejection HMC could also be combined with other improvements to HMC, such as ensemble conditioning (Matthews et al., 2016; Hoffman and Sountsov, 2022), Riemannian HMC (Girolami and Calderhead, 2011; Betancourt, 2013), and manifold-based HMC (Au et al., 2020; Betancourt et al., 2017).

Supplementary Material

Appendix for Delayed rejection Hamiltonian Monte Carlo for sampling multiscale distributions (DOI: 10.1214/23-BA1360SUPP; .pdf). We provide a simple proof of detailed balance for deterministic, volume preserving involution maps.

References

Andrieu, C., Lee, A., and Livingstone, S. (2020). "A general perspective on the Metropolis-Hastings kernel." *arXiv e-prints*, arXiv:2012.14881. 817, 818, 822, 825

Au, K. X., Graham, M. M., and Thiery, A. H. (2020). "Manifold lifting: scaling MCMC to the vanishing noise regime." *arXiv preprint arXiv:2003.03950*. 839

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M., and Stuart, A. (2013). "Optimal tuning of the hybrid Monte Carlo algorithm." *Bernoulli*, 19(5A): 1501–1534. MR3129023. doi: https://doi.org/10.3150/12-BEJ414. 827, 831

Betancourt, M. (2013). "A general metric for Riemannian manifold Hamiltonian Monte Carlo." In *International Conference on Geometric Science of Information*, 327–334. Springer. MR3126061. doi: https://doi.org/10.1007/978-3-642-40020-9_35. 816, 839

Betancourt, M. (2017). "A conceptual introduction to Hamiltonian Monte Carlo." *arXiv preprint arXiv:1701.02434*. MR1699395. doi: https://doi.org/10.1017/CBO9780511470813.003. 815, 818

Betancourt, M., Byrne, S., Livingstone, S., and Girolami, M. (2017). "The geometric foundations of Hamiltonian Monte Carlo." *Bernoulli*, 23(4A): 2257–2298. MR3648031. doi: https://doi.org/10.3150/16-BEJ810. 839

Betancourt, M. and Girolami, M. (2015). "Hamiltonian Monte Carlo for hierarchical models." *Current trends in Bayesian methodology with applications*, 79(30): 2–4. MR3644666. 816, 827, 833, 839

Billingsley, P. (2012). *Probability and Measure*. John Wiley and Sons, anniversary edition. MR2893652. 818

Brofos, J. and Lederman, R. R. (2021). "Evaluating the implicit midpoint integrator for Riemannian Hamiltonian Monte Carlo." In *International Conference on Machine Learning*, 1072–1081. PMLR. 817, 823

Bédard, M., Douc, R., and Moulines, E. (2014). "Scaling analysis of delayed rejection MCMC methods." *Methodology and Computing in Applied Probability*, 16(4): 811–838. MR3270597. doi: https://doi.org/10.1007/s11009-013-9326-y. 817

Campos, C. M. and Sanz-Serna, J. M. (2015). "Extra chance Generalized Hybrid Monte Carlo." *Journal of Computational Physics*, 281: 365–374. MR3281978. doi: https://doi.org/10.1016/j.jcp.2014.09.037. 817

Creutz, M. and Gocksch, A. (1989). "Higher-order hybrid Monte Carlo algorithms." *Physical Review Letters*, 63(1): 9. MR1001905. doi: https://doi.org/10.1103/PhysRevLett.63.9. 823

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). "Hybrid Monte Carlo." *Phys. Lett. B*, 195: 216–222. MR3960671. doi: https://doi.org/10.1016/0370-2693(87)91197-x. 815, 822

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman Hall/CRC, third edition. MR3235677. 834

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). "Bayesian Workflow." *arXiv e-prints*, arXiv:2011.01808. 834

Geyer, C. J. (2011). "Introduction to Markov chain Monte Carlo." In Brooks, S., Gelman, A., Jones, G. L., and Meng, X.-L. (eds.), *Handbook of Markov chain Monte Carlo*, chapter 1. Boca Raton, FL: Chapman and Hall/CRC. MR2858443. 818, 830

Girolami, M. and Calderhead, B. (2011). "Riemann manifold Langevin and Hamiltonian Monte Carlo methods." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2): 123–214. MR2814492. doi: https://doi.org/10.1111/j.1467-9868.2010.00765.x. 816, 839

Green, P. J. and Mira, A. (2001). "Delayed rejection in reversible jump Metropolis-Hastings." *Biometrika*, 88(4): 1035–1053. MR1872218. doi: https://doi.org/10.1093/biomet/88.4.1035. 817, 819, 820

Griewank, A. (1988). "On automatic differentiation." Preprint ANL/MCS-P10-1088, Argonne National Laboratory. MR1114312. 827

Gull, S. F. (1988). "Bayesian inductive inference and maximum entropy." In *Maximum-entropy and Bayesian methods in Science and Engineering*, 53–74. Springer. MR0970800. doi: https://doi.org/10.1007/978-94-009-3049-0_4. 835

Haario, H., Laine, M., Mira, A., and Saksman, E. (2006). "DRAM: efficient adaptive MCMC." *Statistics and Computing*, 16(4): 339–354. MR2297535. doi: https://doi.org/10.1007/s11222-006-9438-0. 817

Hoffman, M. D. and Gelman, A. (2011). "The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo." *arXiv e-prints*, arXiv:1111.4246. MR3214779. 815, 829

Hoffman, M. D. and Sountsov, P. (2022). "Tuning-Free generalized Hamiltonian Monte Carlo." In *International Conference on Artificial Intelligence and Statistics*, 7799–7813. 839

Horowitz, A. M. (1991). "A generalized guided Monte Carlo algorithm." *Physics Letters B*, 268(2): 247–252. 822

Hunter, J. and Nachtergaele, B. (2001). *Applied Analysis*. World Scientific. MR1829589. doi: https://doi.org/10.1142/4319. 818, 819

Kim, S., Shephard, N., and Chib, S. (1998). "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models." *The Review of Economic Studies*, 65(3): 361–393. 837

Leimkuhler, B. and Reich, S. (2004). *Simulating Hamiltonian Dynamics*. Number 14 in Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press. MR2132573. 822

Levy, D., Hoffman, M. D., and Sohl-Dickstein, J. (2017). "Generalizing Hamiltonian Monte Carlo with neural networks." *arXiv preprint arXiv:1711.09268*. 823

Lieb, E. H. and Loss, M. (2001). *Analysis*, volume 14 of *Graduate Studies in Mathematics*. AMS, Providence, RI, second edition. MR1817225. doi: https://doi.org/10.1090/gsm/014. 819

Livingstone, S. and Zanella, G. (2019). "The Barker proposal: combining robustness and efficiency in gradient-based MCMC." *arXiv e-prints*, arXiv:1908.11812. MR4412995. doi: https://doi.org/10.1111/rssb.12482. 815

MacKay, D. J. C. (1998). "Introduction to Monte Carlo Methods." In Jordan, M. I. (ed.), *Learning in Graphical Models*, NATO Science Series, 175–204. Kluwer Academic Press. 818

Matthews, C., Weare, J., and Leimkuhler, B. (2016). "Ensemble preconditioning for Markov chain Monte Carlo simulation." *arXiv preprint arXiv:1607.03954*. MR3747563. doi: https://doi.org/10.1007/s11222-017-9730-1. 839

Mehlig, B., Heermann, D. W., and Forrest, B. M. (1992). "Hybrid Monte Carlo method for condensed-matter systems." *Phys. Rev. B*, 45(2): 679–685. 822

Mira, A. (1998). "Ordering, Slicing and Splitting Monte Carlo Markov chains." Ph.D. thesis, University of Minnesota. MR2698214. 817, 819, 820, 825

Modi, C., Barnett, A., and Carpenter, B. (2023). "Appendix for Delayed rejection Hamiltonian Monte Carlo for sampling multiscale distributions." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1360SUPP. 822

Neal, R. M. (2003). "Slice sampling (with discussion)." *Ann. Stat.*, 31: 705–767. MR1994729. doi: https://doi.org/10.1214/aos/1056562461. 830, 831

Neal, R. M. (2011). "MCMC using Hamiltonian dynamics." In Brooks, S., Gelman, A., Jones, G. L., and Meng, X.-L. (eds.), *Handbook of Markov chain Monte Carlo*, chapter 5. Boca Raton, FL: Chapman and Hall/CRC. MR2858447. 815, 818, 822, 823

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007). "A general framework for the parametrization of hierarchical models." *Statistical Science*, 59–73. MR2408661. doi: https://doi.org/10.1214/088342307000000014. 839

Pourzanjani, A. A. and Petzold, L. R. (2019). "Implicit Hamiltonian Monte Carlo for sampling multiscale distributions." *arXiv e-prints*, arXiv:1911.05754. 816, 817, 823, 839

Rubin, D. B. (1981). "Estimation in parallel randomized experiments." *Journal of Educational Statistics*, 6(4): 377–401. 834

Sherlock, C. and Thiery, A. H. (2017). "A discrete bouncy particle sampler." *arXiv e-prints*, arXiv:1707.05200. MR4430961. doi: https://doi.org/10.1093/biomet/asab013. 817

Sohl-Dickstein, J., Mudigonda, M., and DeWeese, M. (2014). "Hamiltonian Monte Carlo without detailed balance." In *International Conference on Machine Learning*, 719–726. PMLR. 817, 818, 822

Stan Development Team (2022). *Stan Language User's Guide, Version 2.30*. Stan Project. 829

Stein, E. M. and Shakarchi, R. (2005). *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*. Number 3 in Princeton Lectures in Analysis. Princeton University Press. MR2129625. 818

Tierney, L. and Mira, A. (1999). "Some adaptive Monte Carlo methods for Bayesian inference." *Statist. Med.*, 18: 2507–2515. 817, 819, 820

Vanetti, P., Bouchard-Côté, A., Deligiannidis, G., and Doucet, A. (2017). "Piecewise-deterministic Markov chain Monte Carlo." *arXiv e-prints*, arXiv:1707.05296. 817

Wu, C., Stoehr, J., and Robert, C. P. (2018). "Faster Hamiltonian Monte Carlo by learning leapfrog scale." *arXiv e-prints*, arXiv:1810.04449. 829

Yoshida, H. (1990). "Construction of higher order symplectic integrators." *Physics letters A*, 150(5-7): 262–268. MR1078768. doi: https://doi.org/10.1016/0375-9601(90)90092-3. 823