Multiscale MCMC sampling
with delayed rejection generalized HMC

Bob Carpenter

Center for Computational Mathematics
Flatiron Institute

bcarpenter@flatironinstitute.org

September 2023 University of Michigan SciML Webinar

FLATIRON INSTITUTE

The problem and the solution

- **Goal**: Bayesian posterior inference for multiscale posteriors
- **Measure of curvature**: Spectrum (eigenvalues) of Hessian (2nd derivative matrix) of log posterior density
- **Multiscale**: Spectrum varies with parameters
  - **Examples**: hierarchical prior for varying effects, stochastic volatility models, ODEs of varying stiffness w.r.t. parameters, etc.
- **Problem**:
  - **0th order** (Gibbs, RWM) and **1st order** (MALA, HMC, NUTS) methods fail.
  - **2nd order** (Riemannian HMC) too expensive in high dimensions.
- **Solution**: multiscale integrator (generalized HMC with delayed rejection)

Bayesian quantities of interest are expectations

- **Posterior** $p(\theta \mid y) \propto p(y \mid \theta) \cdot p(\theta)$ with **data** y and **parameters** $\theta \in \mathbb{R}^D$.
- **Parameter estimate** minimizing expected square error:

$$\widehat{\theta} = \mathbb{E}[\theta \mid y] = \int_{\mathbb{R}^D} \theta \cdot p(\theta \mid y) \, \mathrm{d}\theta$$

- **Event probability** for event $A \subseteq \mathbb{R}^D$:

$$\Pr[A \mid y] = \mathbb{E}[\mathrm{I}(\theta \in A) \mid y] = \int_{\mathbb{R}^D} \mathrm{I}(\theta \in A) \cdot p(\theta \mid y) \, \mathrm{d}\theta$$

- **Posterior predictive density** for new data $\tilde{y}$:

$$p(\tilde{y} \mid y) = \mathbb{E}\big[p(\tilde{y} \mid \theta) \mid y\big] = \int_{\mathbb{R}^D} p(\tilde{y} \mid \theta) \cdot p(\theta \mid y) \, \mathrm{d}\theta$$

Monte Carlo method (Fermi, Ulam 1930s–1940s)

- Given a Bayesian **posterior density** $p(\theta \mid y)$, with support for **parameters** $\theta \in \mathbb{R}^D$ and **data** y, draw a **sample**

$$\theta^{(1)}, \ldots \theta^{(M)} \sim p(\theta \mid y)$$

- to evaluate **posterior expectations** of functions f

$$\begin{aligned}
\mathbb{E}[f(\theta) \mid y] &= \int_{\mathbb{R}^D} f(\theta) \cdot p(\theta \mid y) \, \mathrm{d}\theta \\
&= \lim_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} f\left(\theta^{(m)}\right) \\
&\approx \frac{1}{M} \sum_{m=1}^{M} f\left(\theta^{(m)}\right)
\end{aligned}$$

Markov chain Monte Carlo (Metropolis et al. 1950)

- Usually impossible to draw an independent sample from a target density.
- Instead, set up a **Markov chain** where the **stationary distribution** is the target distribution.
- Same **plug-in estimator** still works with correlated draws.
- **MCMC central limit theorem** says estimation standard error is $\frac{\mathrm{sd}}{\sqrt{\mathrm{ESS}}}$, where
  - sd is the **posterior standard deviation** of the estimand,
  - and ESS is the **effective sample size** of the sample (as measured in independent draws).
  - With HMC, effective sample size can exceed sample size

Hessians are second derivatives

- Given a posterior density $p(\theta \mid y)$, its **Hessian** is the matrix of **second (partial) derivatives**,

$$\mathrm{H}(\theta) = \nabla_\theta \, \nabla_\theta^\top \, p(\theta \mid y).$$

  with entries

$$\mathrm{H}_{i,j}(\theta) = \frac{\partial^2}{\partial \theta_i \, \partial \theta_j} p(\theta \mid y).$$

- If $p(\theta \mid y) = \mathrm{normal}(\theta \mid \mu, \Sigma)$ with **positive definite covariance** Σ, then the Hessian is the negative inverse covariance (i.e., negative precision),

$$\mathrm{H}(\theta) = -\Sigma^{-1}.$$

- $\Sigma = \mathrm{diag}([\sigma_1^2 \cdots \sigma_D^2])$ is **diagonal**, then its Hessian is $\mathrm{diag}([\sigma_1^{-2} \cdots \sigma_D^{-2}])$

The spectrum of eigenvalues

- If A is a $D \times D$ matrix, its **eigendecomposition** is

$$A = Q \cdot \mathrm{diag}(\lambda) \cdot Q^{-1}$$

  λ a D-vector of **eigenvalues**, Q a $D \times D$ orthonormal matrix of **eigenvectors**

- Eigenvalues are **inverse squared scales** in the direction of the eigenvalues

Positive definiteness and log concavity

- A matrix is **positive definite** if the eigenvalues are all positive
- A density is **log concave** at a point if its Hessian is positive definite.
- A multivariate normal with **diagonal covariance** $\Sigma = \mathrm{diag}([\sigma_1^2 \cdots \sigma_D^2])$ has
  - axis-aligned eigenvectors, $Q = \mathrm{I}$ (I is identity)
  - eigenvalues $\lambda = \sigma_1^{-2}, \ldots, \sigma_D^{-2}$
- Eigenvalues are **rotation invariant**.
- For non-diagonal covariance, just **rotate to diagonal**.

Condition numbers and iterative algorithms

- The **condition** of a positive definite matrix A is the ratio of largest to smallest eigenvalue,

$$c = \frac{\max(\lambda)}{\min(\lambda)}.$$

- To move a "unit," gradient-based algorithms take **steps proportional to smallest scale** and a **number of steps equal to the condition**.

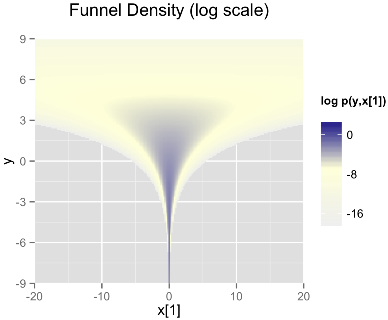
- A posterior $p(\theta \mid y)$ has
  - **varying curvature** if its Hessian changes for different θ, and
  - **varying scale** if its smallest scale changes for different θ.

- Thus **varying scales require varying step sizes** (for gradient-based algo).

Neal's funnel as a proxy for hierarchical priors

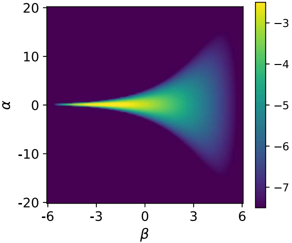
- Neal's funnel for log scale (times two) $y \in \mathbb{R}$ and varying effects $x \in \mathbb{R}^N$ is

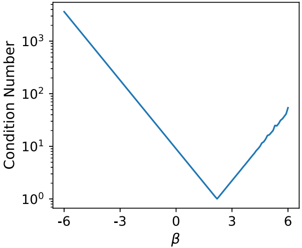
$$p(x, y) = \text{normal}(y \mid 0, 3) \cdot \prod_{n=1}^{N} \text{normal}(x_n \mid 0, \exp(y/2)).$$



Neal's funnel has varying curvature and scale

- Here's a plot of the (rotated) funnel and its condition number vs. scale β
- central 95% interval for constant scale β—condition worsens in tails
- Eigenvectors change orientation (biggest along β in neck, along α in mouth)





Hamiltonian dynamics

- **Potential energy** at $\theta \in \mathbb{R}^D$ is negative log density $U(\theta) = -\log\left(p(\theta \mid y)\right)$.
- **Kinetic energy** for momentum $\rho \in \mathbb{R}^D$ is $V(\rho) = -\log\left(\text{normal}(\rho \mid 0, 1)\right)$.
- **Hamiltonian** is sum $H(\theta) = U(\theta) + V(\theta)$
- **Leapfrog step** for Hamiltonian dynamics w. discretization time $\epsilon > 0$

$$
\begin{aligned}
\rho_{t+1/2} &= \rho_t - \frac{\epsilon}{2} \cdot \nabla U(\theta) \\
\theta_{t+1} &= \theta_t - \epsilon \cdot \nabla V(\theta) \\
\rho_{t+1} &= \rho_{t+1/2} - \frac{\epsilon}{2} \cdot \nabla U(\theta)
\end{aligned}
$$

- **Precondition** with pos. def. metric Σ by $V(\rho) = -\log\left(\text{normal}(\rho \mid 0, Sigma)\right)$.

Hamiltonian Monte Carlo

(Duane et al. 1987)

- **Input**: initial position $\theta^{(0)}$, step size ϵ, steps L, metric Σ, sample size M
- For each iteration $m \in 1, \ldots, M$
  - (Gibbs) Resample momentum $\rho \sim \mathrm{normal}(0, \Sigma)$
  - (Metropolis) Run leapfrog algorithm L steps from $(\theta^{(m-1)}, \underbrace{-\rho}_{\text{flip}})$ to (θ^*, ρ^*)
  - $\text{accept} = \text{uniform}(0,1) < \min\left(1, \dfrac{\exp(-H(\theta^*, \rho^*))}{\exp(-H(\theta^{(m-1)}, \rho))}\right)$
  - $(\theta^{(m)}, \rho^{(m)}) = (\theta^*, \underbrace{-\rho^*}_{\text{flip}})$ if accept else $(\theta^{(m-1)}, \rho^{(m-1)})$.
- **Output**: sample $\theta^{(1)}, \ldots, \theta^{(M)}$

Generalized Hamiltonian Monte Carlo (Horowitz 1991)

- **Generalized HMC**: Partially resample momentum each iteration

$$\rho \sim \mathrm{normal}\left(\sqrt{1-\lambda} \cdot \rho^{(m-1)},\ \lambda \cdot \Sigma\right).$$

- Still **(exact) Gibbs** sampling
  - if $\rho^{(m-1)} \sim \mathrm{normal}(0, \Sigma)$, then $\sqrt{1-\lambda} \cdot \rho \sim \mathrm{normal}(0, (1-\lambda) \cdot \Sigma)$ and

$$\rho \sim \mathrm{normal}(0, \Sigma)$$

  - weights balance variance (sqrt converts to scale)
- Usually take just **one leapfrog iteration**
  - one step of HMC is identical to **Metropolis-adjusted Langevin** (MALA)
  - but it operates on position and momentum vector

HMC works, but generalized HMC fails

- **HMC scales in dimension** by making **directed progress** per iteration
- **Hamiltonian flow** keeps trajectory in region of high probability
- Leapfrog integrator is **symplectic**
  - preserves Hamiltonian well, so **high Metropolis accept** rate
  - it's *not* an accurate ODE solver (but that's OK)
- **G-HMC reverts to random walk** because of the **flipped momentum**
  - G-HMC usually configured to use **one leapfrog step** (like MALA)
  - required to preserve stationarity (cf. 100% refreshed in standard HMC)
  - **reverses momentum** on failure, so need **sequences of acceptances**
  - need large step size to move, small step sizes for acceptance

Non-uniform acceptance fixes G-HMC (Neal 2020)

- Neal (2020) replaced the i.i.d. $u^{(m)} \sim \mathrm{uniform}(0,1)$ variate in Metropolis,

  $$\mathrm{accept} = \mathrm{uniform}(0,1) < \min(1, \cdots)$$

  with an identically distributed but not independent variate carving out a **sawtooth pattern**

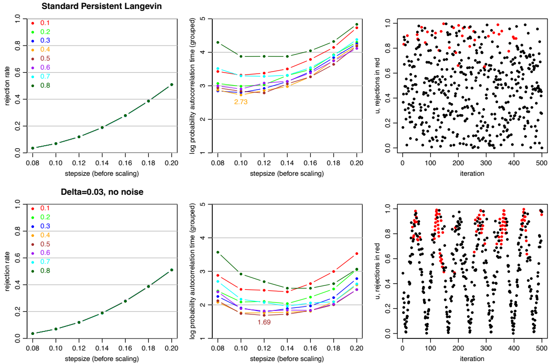
  $$u^{(m)} = u^{(m-1)} + \delta + \mathrm{uniform}(0, \sigma^{\mathrm{jitter}})$$

  and if $u^{(m)} \notin (0,1)$ add or subtract 2 until it is.

- Jitter is for **ergodicity** so that $u^{(m)} \sim \mathrm{uniform}(0,1)$ marginally (correlated)
- **Acceptances cluster** at sequences of small values of $u^{(m)}$.
- Adds **tuning parameters** $\delta, \sigma^{\mathrm{jitter}} \in (0, \infty)$.

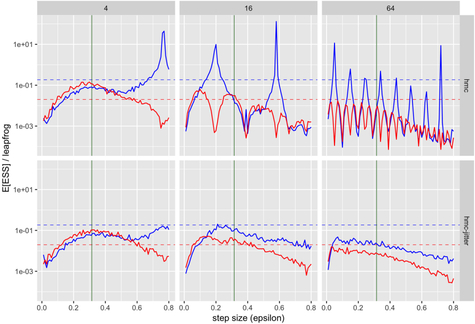
Neal’s evaluation of non-reversible u for G-HMC

- for Bayesian neural network, 1.25 times **faster than HMC**!
- 16 pairs of normal variables with unit variance and 0.99 correlation, α color coded:



(Neal 2020)

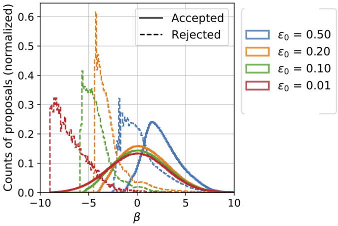
HMC sensitive to integration time (steps $\times$ num steps)

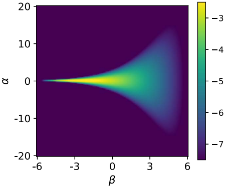


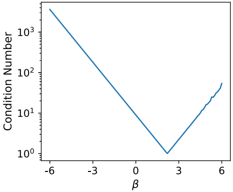
- Standard normal, **1000 dimensions**; vertical axis ESS (log scale); horizontal axis step size (ϵ); columns (4, 16, 64) steps (L); top row HMC, bottom row uniformly steps-jittered HMC; blue mean estimate, red variance; dashed line is NUTS (Stan)

HMC & MALA fail on the funnel

- **Fixed step size** leads to **truncated sampling** with HMC (and NUTS), either
  - **Neck**: step size too big, **Hamiltonian diverges** and we **reject**.
  - **Mouth**: step size too small, **diffusion too slow**. explore.
  - Result is **biased estimation** of quantities of interest.
- Vertical dashed lines show the left truncation (color = step size)



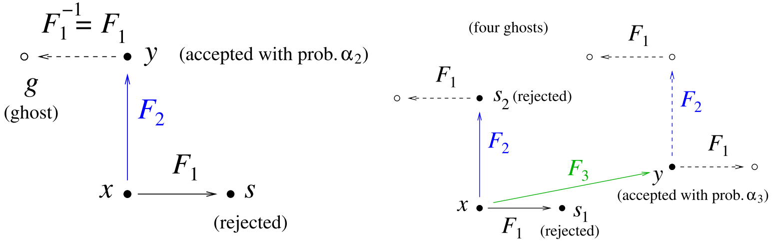




Delayed rejection Metropolis **(Mira and Green 2001)**

- Within a single iteration, **try again if proposal rejected**.
- Require **Hastings adjustment** for detailed balance for trying again.
- Assume first level **proposal** F_1 and second-level F_2, and so on
- **First level**: accept $s = F_1(x)$ $\alpha_1(x, s) = \min\left(1, \frac{p(s)}{p(x)}\right)$.
- **Second level**: accept $x \mapsto z$: $\alpha_2(x, y) = \min\left(1, \frac{p(y)}{p(x)} \frac{1 - \alpha_1(y, g)}{1 - \alpha_1(x, s)}\right)$.
  - where $g = F_1(v)$ is a first level **"ghost proposal"**
- **Third level (and beyond)**: next page figure (paper for **general recursion**)

Picture of delayed rejection



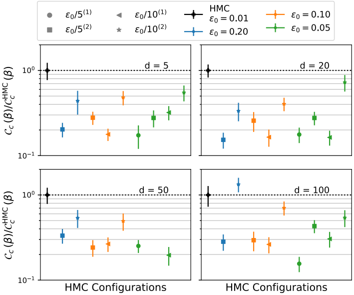
Delayed rejection HMC

(Modi, Barnett, Carpenter 2022)

- For HMC, key is to try again with **reduced step size**.
  - earlier attempts tried to save computation by extending rejected proposal (Sohl-Dickstein et al. 2014, Campos and Sanz-Serna 2014)
- We evaluated up to 3 levels of retries,
  - with step sizes $\epsilon, \epsilon \cdot \lambda, \epsilon \cdot \lambda^2$ for $\lambda = \frac{1}{2}, \frac{1}{3}, \frac{1}{5}$

Evaluation of DR-HMC

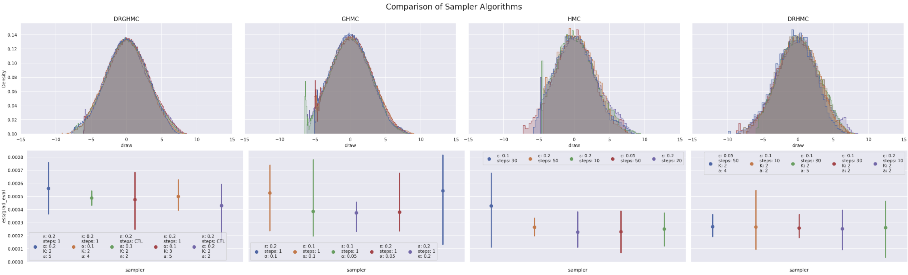
- **Neal's funnel** various dims, step sizes, step reduction ratios
- vertical axis (log scale) is **cost in gradients vs. ground truth**
- **DR-HMC works** and is **also cheaper** (HMC isn't convergent here)



Delayed rejection, generalized HMC (Turok et al. 2023+)

- **Two great tastes** that go great together.
- **Swaps** delayed rejection for Neal's non-reversible uniform accept probs
- Two benefits:
  - **high acceptance rate** needed for mixing in G-HMC
  - works for **multiscale distributions**
- DR-G-HMC **mixes faster** than DR-HMC per gradient
  - DR-HMC mixes as fast or faster than HMC but also handled varying scales
- **Gilad Turok** was an (undergrad) intern this summer with Chirag Modi.
- **Edward Roualdes** is working on adaptation (led to BridgeStan package!).

DR-G-HMC evaluation



- HMC and G-HMC fail; **DR-G-HMC outperforms DR-HMC** (as in Neal's evaluations)
- Results similar with **constant integration time** on retries (multiplying steps)
- **Paper** in progress as is **code for Bayes-Kit** (Python).

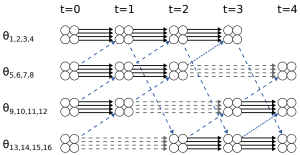
MEADS: Adaptation for G-HMC (Hoffman, Sountsov 2022)

- **Starting point** is Neal's non-reversible acceptance G-HMC
- **Less wasteful** than HMC/NUTS (cf. Nicholas Chopin's "waste-free" SMC)
  - vs. HMC: doesn't **reject long chain** of leapfrog steps
  - vs. NUTS: doesn't go **forward and backward in time** and choose **non-final point** on trajectory
- **Easier to deploy** than HMC/NUTS
  - much easier to **parallelize** than NUTS recursion
  - easier to **adaptively tune** (steps more granular)

MEADS (cont.)

(Hoffman, Sountsov 2022)

- **Ensemble** of chains for **complementary chain adaptation**
  - cf. Goodman-Weare affine-invariant, ter Braak differential evolution



- Heuristic **eigenvalue estimator for step size**
  - $\epsilon = \dfrac{1}{2 \cdot \sqrt{\lambda^{\max}(-\overline{H})}}$, where $\lambda^{\max}$ is max eigenvalue operator
  - $\overline{H} = \mathbb{E}[\mathrm{H}(\Theta) \mid y] = \mathbb{E}[\nabla\nabla^\top \log p(\Theta \mid y)]$, estimated with empirical average

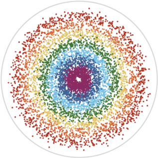
Summary and Conclusions

- **delayed rejection HMC** enables multiscale sampling (Modi et al.)
- **one-step generalized HMC** can be tuned to be as efficient as HMC with non-reversible acceptance (Neal)
- **delayed rejection** works as well as non-reversible acceptance *and* enables multiscale sampling (Turok et al.)
- **ensemble methods** and **eigenvalue step size estimate** allow automatic tuning of one-step G-HMC (Hoffman and Sountsov)
- **same adaptation** works for DR-G-HMC (Roualdes et al.)

Dramatis Personae

Gilad Turok
gil2rok

Chirag Modi
modichirag



Edward A. Roualdes
roualdes

Alex Barnett
ahbarnett

Radford Neal
radfordneal

Matt Hoffman
matthewdhoffman

Pavel Sountsov
SiegeLordEx

References

- Turok, G., Roualdes, E., Modi, C., and Carpenter, B. In preparation. **Delayed rejection generalized Hamiltonian Monte Carlo**.
- Modi, C., Barnett, A. and Carpenter, B., 2023. **Delayed rejection Hamiltonian Monte Carlo for sampling multiscale distributions**. *Bayesian Analysis*.
- Hoffman, M.D. and Sountsov, P., 2022. **Tuning-free generalized Hamiltonian Monte Carlo**. *AISTATS*.
- Neal, R.M., 2020. **Non-reversibly updating a uniform [0, 1] value for Metropolis accept/reject decisions**. *arXiv*.