DRHMC : Delayed Rejection Hamiltonian Monte-Carlo

Chirag Modi
Center for Computational Astrophysics (CCA)
Center for Computational Mathematics (CCM)
Flatiron Institute

StanCon 2023

w/ Alex Barnett, Bob Carpenter
arXiv:2110.00610

Hierarchical models

Q : Researcher is interested in learning the mean SAT score (**μ**).
Data : Collect students' scores for SAT exams from 5 schools.

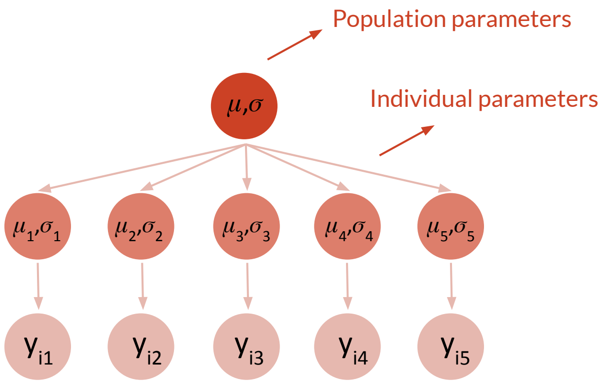
Hierarchical models

Q : Researcher is interested in learning the mean SAT score (**μ**).
Data : Collect students' scores for SAT exams from 5 schools.

Hierarchical model with partial pooling

$$\mu, \sigma \sim \boldsymbol{\pi}(\mu, \sigma); \quad \mu_j \sim \mathbb{N}(\mu, \sigma);$$

$$Y_{ij} \sim \mathbb{N}(\mu_j, \sigma_j)$$



Hierarchical models

Q : Researcher is interested in learning the mean SAT score (**μ**).
Data : Collect students' scores for SAT exams from 5 schools.

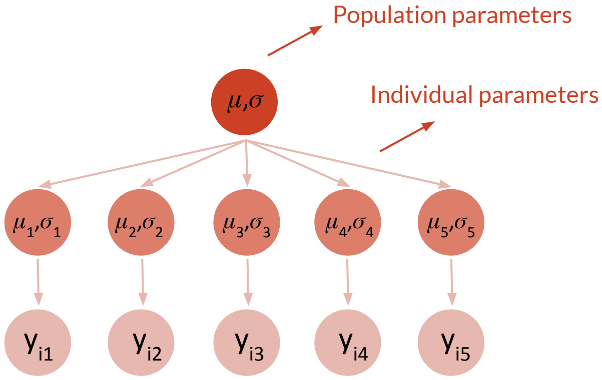
Hierarchical model with partial pooling

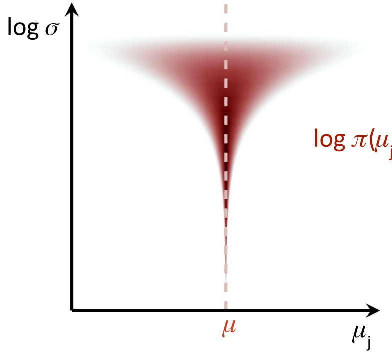
$$\mu, \sigma \sim \pi(\mu, \sigma); \quad \mu_j \sim \mathbb{N}(\mu, \sigma);$$

$$Y_{ij} \sim \mathbb{N}(\mu_j, \sigma_j)$$

Degeneracies of hierarchical model

- small σ results in μ_j concentrating around μ
- large σ results in μ_j varying wider range of values





Hierarchical models

Q : Researcher is interested in learning the mean SAT score (**μ**).
Data : Collect students' scores for SAT exams from 5 schools.

Hierarchical model with partial pooling

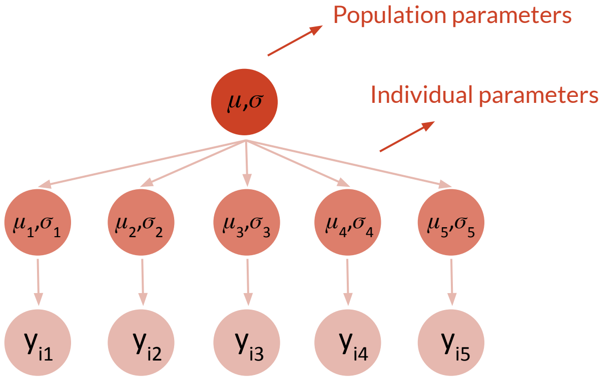
$$\mu, \sigma \sim \boldsymbol{\pi}(\mu, \sigma); \quad \mu_j \sim \mathbb{N}(\mu, \sigma);$$

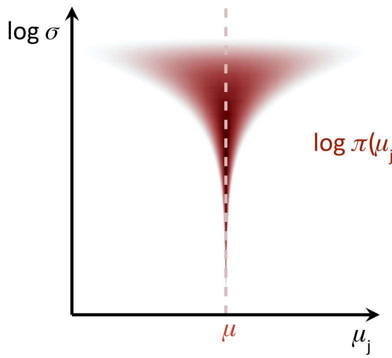
$$Y_{ij} \sim \mathbb{N}(\mu_j, \sigma_j)$$

Degeneracies of hierarchical model

- small σ results in μ_j concentrating around μ
- large σ results in μ_j varying wider range of values

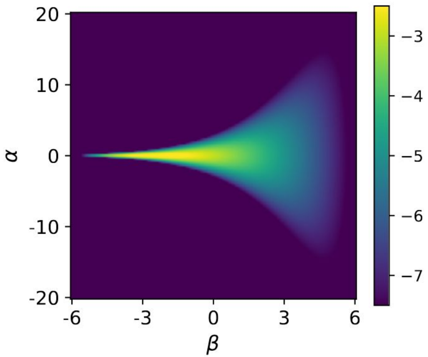
Strong coupling of the individual parameters (μ_j) to the population parameters (μ,σ)
→ **Funnel degeneracy**





Neal's funnel

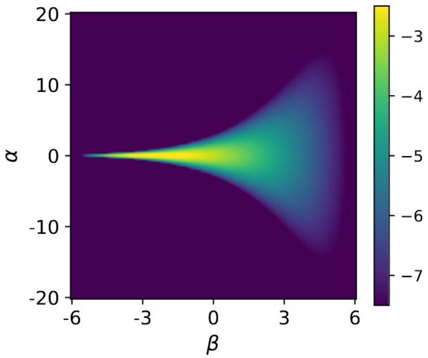
$$\beta \sim \mathcal{N}(0, \sigma^2)$$
$$\alpha_i \sim \mathcal{N}(0, e^{\beta}),$$

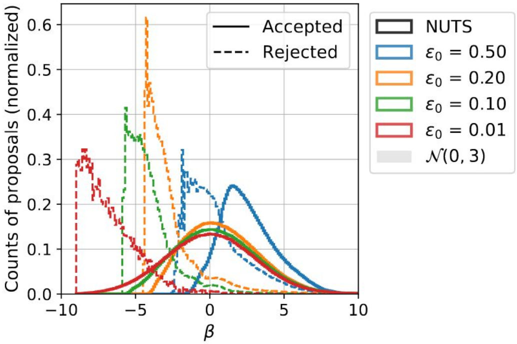


Neal’s funnel

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

- Need very different step-sizes in different regions
  - Small steps to probe the neck

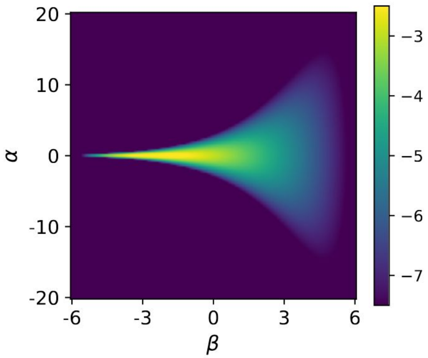


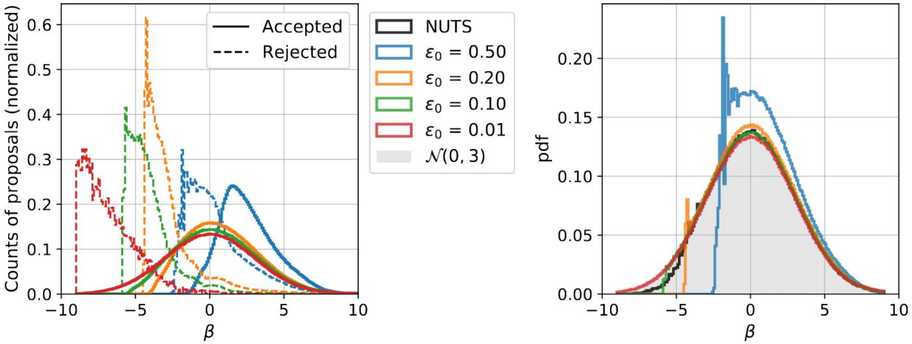


Neal’s funnel

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

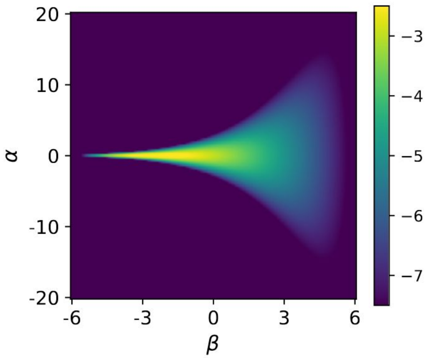
- Need very different step-sizes in different regions
    - Small steps to probe the neck



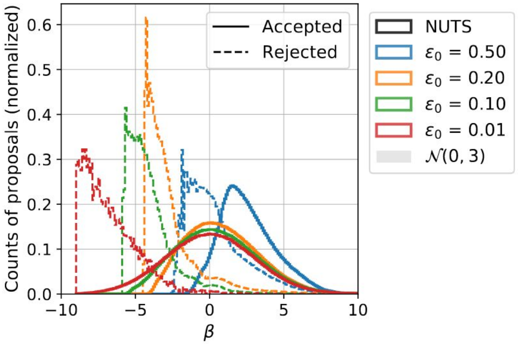


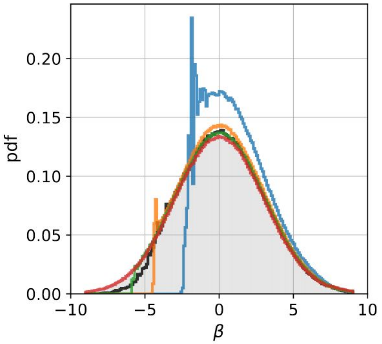
Neal's funnel

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



- Need very different step-sizes in different regions
  - Small steps to probe the neck
  - Large steps to probe the mouth



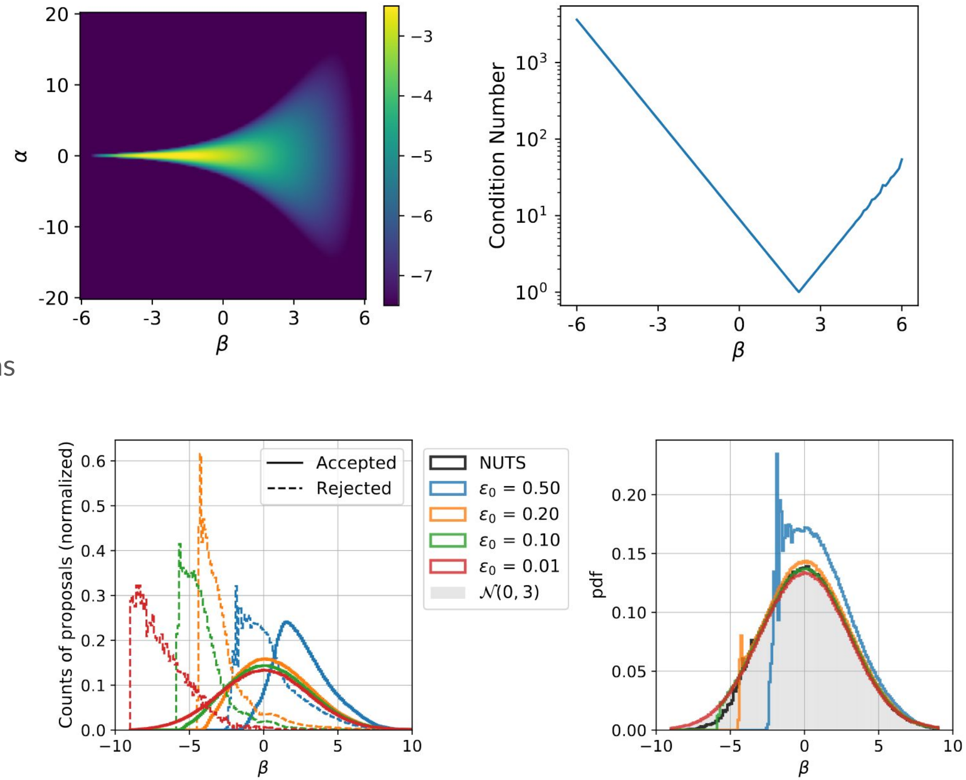


Neal's funnel

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

- Need very different step-sizes in different regions
  - Small steps to probe the neck
  - Large steps to probe the mouth

- Constant mass matrix is insufficient
  - **Multiscale!** condition number changes
  - Very badly conditioned away from origin



Motivation for DRHMC

For the distributions that do not have globally optimal configurations for the transition kernel, can we still benefit from different locally optimized transition kernels.*

(*when continuous adaptation is not feasible)

HMC

$$x \bullet \xrightarrow{F_1} \bullet\ y$$

Delayed Rejection HMC

Standard HMC

- Transition kernel $\mathbf{F_1}$: `n' leapfrog steps with step size `ε'

Random-walk Metropolis: Mira 1998, Mira & Tierney 99, Greene & Mira 2001

Delayed Rejection HMC

When faced with a rejection, make another proposal with a **different transition kernel** that has a better chance of getting accepted

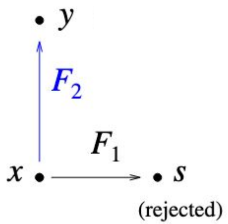
Standard HMC

- Transition kernel $\mathbf{F_1}$: `n' leapfrog steps with step size `ε'

2-stage DRHMC

HMC

(a) 2–stage DRHMC



Random-walk Metropolis: Mira 1998, Mira & Tierney 99, Greene & Mira 2001

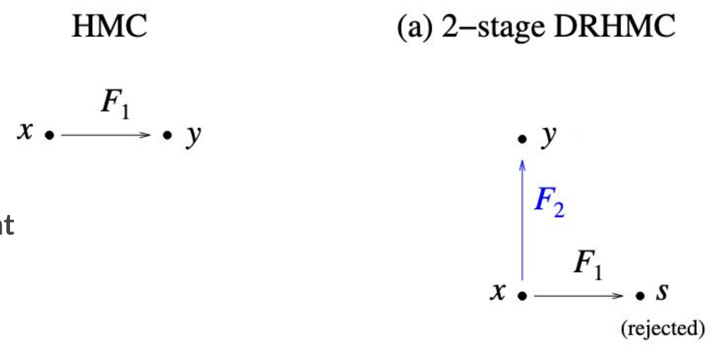
Delayed Rejection HMC

When faced with a rejection, make another proposal with a **different transition kernel** that has a better chance of getting accepted

Standard HMC

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

2-stage DRHMC



Detailed balance:

the probability of transitioning from ***x*** to ***y*** is the same as the probability of transitioning from ***y*** to ***x***

Random-walk Metropolis: Mira 1998, Mira & Tierney 99, Greene & Mira 2001

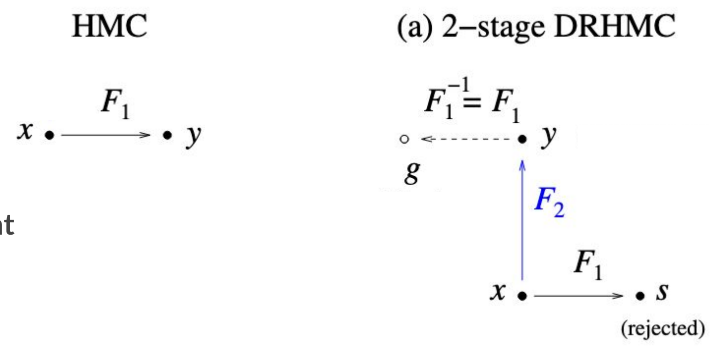
Delayed Rejection HMC

When faced with a rejection, make another proposal with a **different transition kernel** that has a better chance of getting accepted

Standard HMC

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

2-stage DRHMC



Detailed balance:

the probability of transitioning from ***x*** to ***y*** is the same as the probability of transitioning from ***y*** to ***x***

Random-walk Metropolis: Mira 1998, Mira & Tierney 99, Greene & Mira 2001

Delayed Rejection HMC

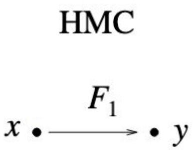
When faced with a rejection, make another proposal with a **different transition kernel** that has a better chance of getting accepted

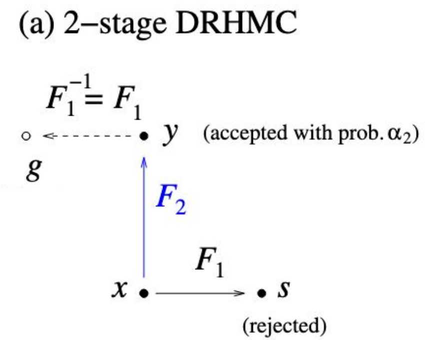
Standard HMC

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

2-stage DRHMC

$$\alpha_2(x, F_1(x), y) \;=\; \min\left(1, \frac{\pi(y)}{\pi(x)} \frac{1 - \alpha_1\big(y, F_1(y)\big)}{1 - \alpha_1\big(x, F_1(x)\big)}\right)$$





Detailed balance:

the probability of transitioning from ***x*** to ***y*** is the same as the probability of transitioning from ***y*** to ***x***

Random-walk Metropolis: Mira 1998, Mira & Tierney 99, Greene & Mira 2001

Delayed Rejection HMC

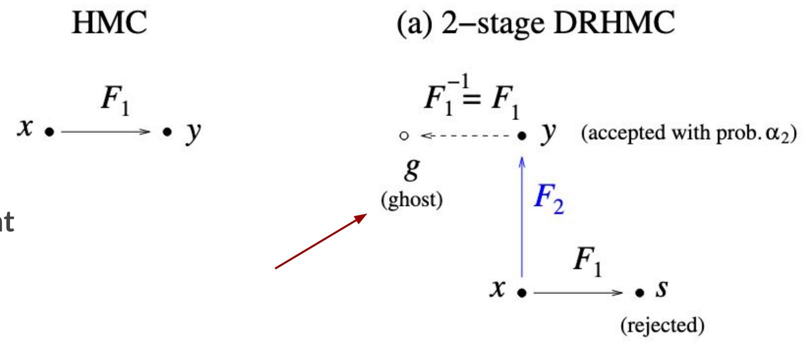
When faced with a rejection, make another proposal with a **different transition kernel** that has a better chance of getting accepted

Standard HMC

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

2-stage DRHMC

$$\alpha_2(x, F_1(x), y) \;=\; \min\left(1, \frac{\pi(y)}{\pi(x)} \frac{1-\alpha_1(y, F_1(y))}{1-\alpha_1(x, F_1(x))}\right)$$



Detailed balance:

the probability of transitioning from ***x*** to ***y*** is the same as the probability of transitioning from ***y*** to ***x***

Random-walk Metropolis: Mira 1998, Mira & Tierney 99, Greene & Mira 2001

Delayed Rejection HMC

When faced with a rejection, make another proposal with a **different transition kernel** that has a better chance of getting accepted

Standard HMC

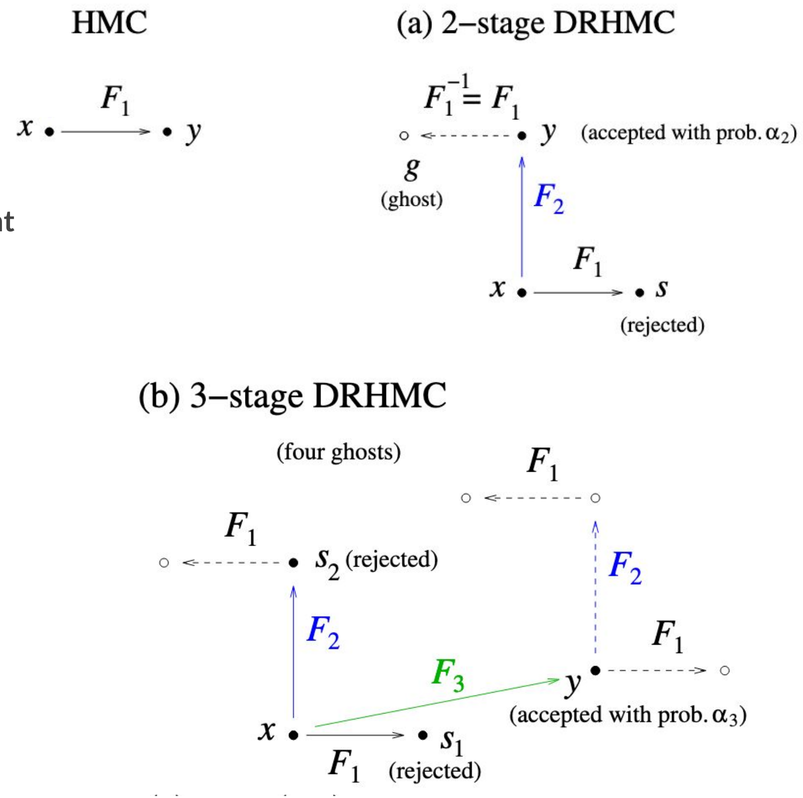
$$\alpha(x,y) = \min\left(\frac{\pi(y)}{\pi(x)}, 1\right)$$

2-stage DRHMC

$$\alpha_2(x, F_1(x), y) \;=\; \min\left(1, \frac{\pi(y)}{\pi(x)} \frac{1-\alpha_1(y, F_1(y))}{1-\alpha_1(x, F_1(x))}\right)$$

3-stage DRHMC

$$\widetilde{\alpha}_3(x) = \min\left[\frac{\pi(y)[1-\widetilde{\alpha}_1(y)][1-\widetilde{\alpha}_2(y)]}{\pi(x)[1-\widetilde{\alpha}_1(x)][1-\widetilde{\alpha}_2(x)]}, 1\right]$$



Random-walk Metropolis: Mira 1998, Mira & Tierney 99, Greene & Mira 2001

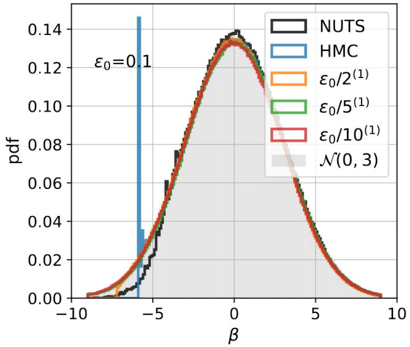
DRHMC for Neal's funnel

Delayed proposals with reduced step size:

- Starting step size: ε_0
- Step size decreases with factor 'a'
  $\varepsilon_0, \varepsilon_0/a, \varepsilon_0/a^2, ..., \varepsilon_0/a^k....$

Two hyperparameters:

- a: factor of reduction
- k: number of delayed rejections



DRHMC for Neal's funnel

Delayed proposals with reduced step size:

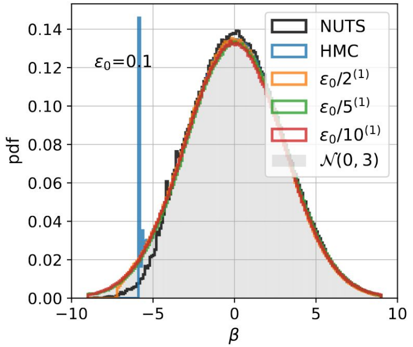
- Starting step size: ε_0
- Step size decreases with factor 'a'
  $\varepsilon_0, \varepsilon_0/a, \varepsilon_0/a^2, ..., \varepsilon_0/a^k ....$

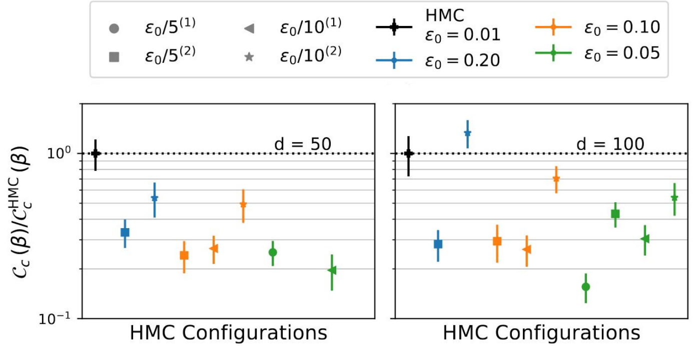
Two hyperparameters:

- a: factor of reduction
- k: number of delayed rejections

DRHMC requires ~5-10x less gradient evaluations vs standard HMC

- Similar gains for other hard problems





Lower is better

DRHMC for Neal's funnel

Delayed proposals with reduced step size:

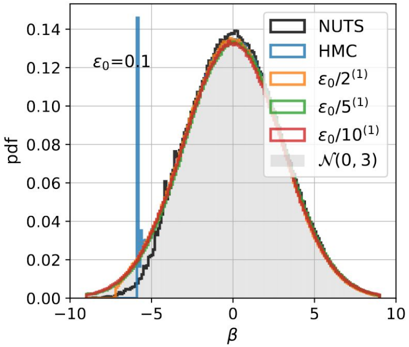
- Starting step size: ε_0
- Step size decreases with factor 'a'
  $\varepsilon_0, \varepsilon_0/a, \varepsilon_0/a^2, ..., \varepsilon_0/a^k ....$

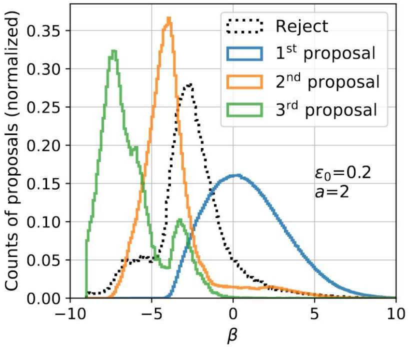
Two hyperparameters:

- a: factor of reduction
- k: number of delayed rejections

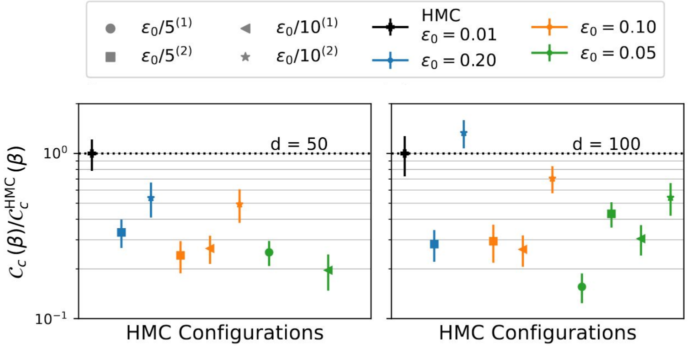
DRHMC requires ~5-10x less gradient evaluations vs standard HMC

- Similar gains for other hard problems





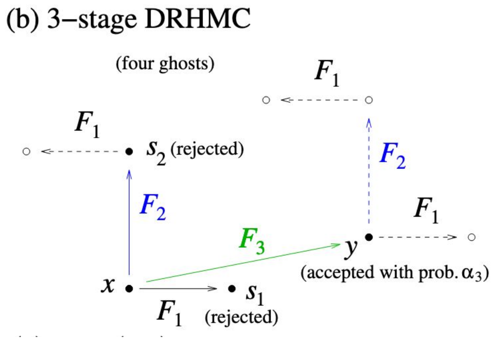
Delayed proposals made as we move into the neck of the funnel.



Lower is better

Cost of DRHMC

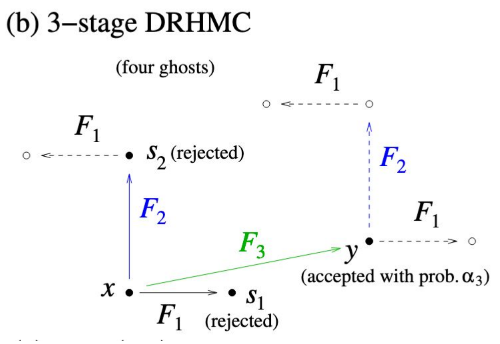
Number of proposals (density evaluations) for k^{th} order grows as 2^{k-1}



Cost of DRHMC

Number of proposals (density evaluations) for k^{th} order grows as 2^{k-1}

But, the dominant cost in HMC is number of gradient evaluations, leapfrog steps!



Cost of DRHMC

Number of proposals (density evaluations) for k^{th} order grows as 2^{k-1}

But, the dominant cost in HMC is number of gradient evaluations, leapfrog steps!

DRHMC

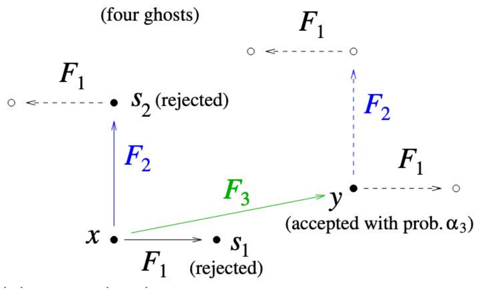
- Starting step size: ε
- Step size decreases with factor '*a*'
- $\varepsilon, \varepsilon / a, \varepsilon / a^2, \ldots, \varepsilon / a^k$

HMC

- Largest ***stable*** step size: ε_0

  $\varepsilon_0 \sim \varepsilon / a^{k-1}$

(b) 3–stage DRHMC



Cost of DRHMC

Number of proposals (density evaluations) for k^{th} order grows as 2^{k-1}

But, the dominant cost in HMC is number of gradient evaluations, leapfrog steps!

DRHMC

- Starting step size: ε
- Step size decreases with factor 'a'
- $\varepsilon, \varepsilon / a, \varepsilon / a^2, \ldots, \varepsilon / a^k$

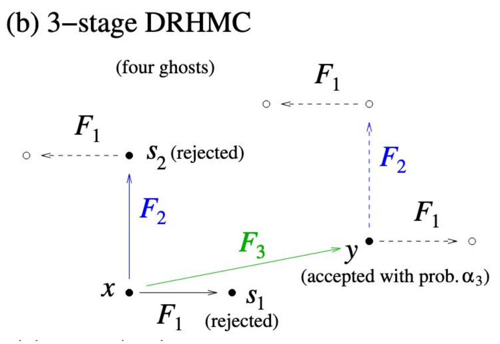
HMC

- Largest ***stable*** step size: ε_0

$\varepsilon_0 \sim \varepsilon / a^{k-1}$

Integration time for every proposal: $\boldsymbol{T = n\varepsilon}$

- Total number of leapfrog steps for HMC $\sim T/\varepsilon_0 \quad \sim na^{k-2}$



Cost of DRHMC

Number of proposals (density evaluations) for k^{th} order grows as 2^{k-1}

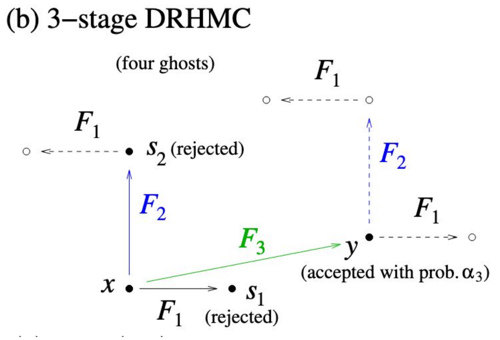
But, the dominant cost in HMC is number of gradient evaluations, leapfrog steps!

DRHMC

- Starting step size: ε
- Step size decreases with factor 'a'
- $\varepsilon, \varepsilon / a, \varepsilon / a^2, \ldots, \varepsilon / a^k$

HMC

- Largest ***stable*** step size: ε_0
  $\varepsilon_0 \sim \varepsilon / a^{k-1}$



Integration time for every proposal: $\boldsymbol{T = n\varepsilon}$

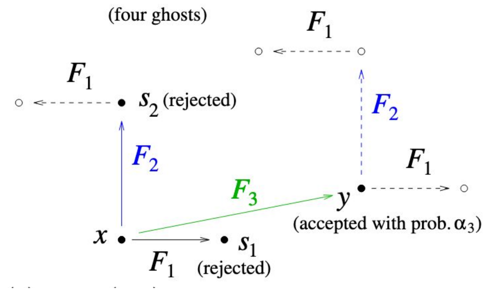
- Total number of leapfrog steps for HMC $\sim T/\varepsilon_0$ $\sim na^{k-2}$
- Total number of leapfrog steps for DRHMC* : $n \times 2^{k-1} + na \times 2^{k-2} + \ldots + na^{k-1} \times 1$ (GP)
  (*in the worst case) $\quad(\varepsilon) \quad (\varepsilon/a) \quad (\varepsilon/a^k)$

Cost of DRHMC

Number of proposals (density evaluations) for k^{th} order grows as 2^{k-1}

But, the dominant cost in HMC is number of gradient evaluations, leapfrog steps!

(b) 3–stage DRHMC



DRHMC

- Starting step size: ε
- Step size decreases with factor '*a*'
- $\varepsilon, \varepsilon / a, \varepsilon / a^2, \ldots, \varepsilon / a^k$

HMC

- Largest ***stable*** step size: ε_0
  $\varepsilon_0 \sim \varepsilon / a^{k-1}$

Integration time for every proposal: $T = n\varepsilon$

- Total number of leapfrog steps for HMC $\sim T/\varepsilon_0$ $\quad \sim na^{k-2}$
- Total number of leapfrog steps for DRHMC* $\quad : n \times 2^{k-1} + na \times 2^{k-2} + \ldots + na^{k-1} \times 1$ (GP)
  (*in the worst case)
  $= nka^{k-1}$ if $a = 2$ → **O(ak) more expensive than HMC**
  $= O(a^{k-1}n)$ if $a > 2$ → **O(a) more expensive than HMC,**
  Independent of k!

Are delayed proposals always beneficial?

Say you reject a proposal when α = 0.9
Should you make a delayed proposal?

Are delayed proposals always beneficial?

Say you reject a proposal when α = 0.9
Should you make a delayed proposal?
Not all rejections are bad

Are delayed proposals always beneficial?

Say you reject a proposal when α = 0.9
Should you make a delayed proposal?
Not all rejections are bad

Probabilistic DRHMC: Make the next proposal with probability

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

$$\widetilde{\alpha}_2(x) = \min\left[\frac{\pi(y)[1-\widetilde{\alpha}_1(y)]p_2\big(y, F_1(y)\big)}{\pi(x)[1-\widetilde{\alpha}_1(x)]\underline{p_2\big(x, F_1(x)\big)}}, 1\right]$$

Are delayed proposals always beneficial?

Say you reject a proposal when α = 0.9
Should you make a delayed proposal?
Not all rejections are bad

Probabilistic DRHMC: Make the next proposal with probability

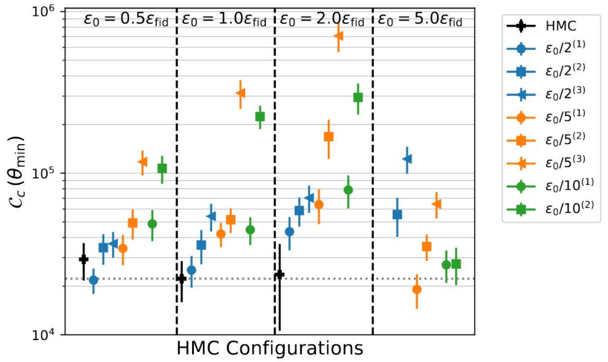
$$p_{j+1}(x) = 1 - \tilde{\alpha}_j(x)$$

$$\tilde{\alpha}_2(x) = \min\left[\frac{\pi(y)[1-\tilde{\alpha}_1(y)]p_2(y, F_1(y))}{\pi(x)[1-\tilde{\alpha}_1(x)]p_2(x, F_1(x))}, 1\right]$$

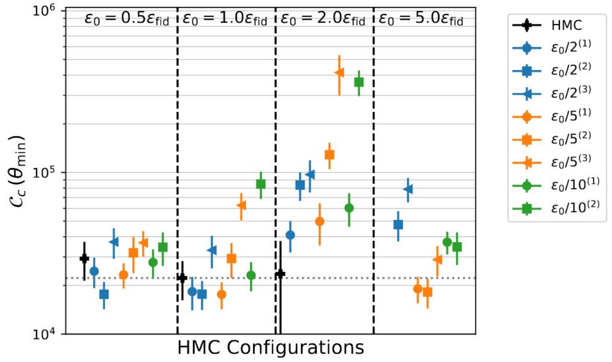
- Reduces the cost of DRHMC by only making delayed proposals only when needed

- Increases robustness to fitting step-size!

Stochastic Volatility



(a) Delayed Rejection



(b) Probabilistic Delayed Rejection

Variants of DRHMC

Does not necessarily need to reduce step size

- Different integrators (higher order leapfrog, implicit midpoint)
- Different kinetic energy
- Different mass matrix

Variants of DRHMC

Does not necessarily need to reduce step size

- Different integrators (higher order leapfrog, implicit midpoint)
- Different kinetic energy
- Different mass matrix

Extensions for DRHMC

WIP with Gilad Turok and Bob Carpenter

- Auto-tuning DRHMC hyper-parameters
- Continuous adaptation: Combine DR + Generalized HMC (partial momentum refresh)

Takeaways

- Delayed rejection HMC for pathological distributions (multiscale distributions like funnel)
  - benefit from multiple, locally optimized transition kernels

- Unlike DR for Metropolis Hastings, cost of a well-tuned DRHMC is a constant factor more than a stable HMC
  - if adapting step size

- Probabilistic DRHMC makes proposals probabilistically, and reduces the cost of DRHMC

- More stable to tuning parameters of HMC (for e.g. step size)

- Any valid transition kernels can be combined into delayed proposals

Takeaways

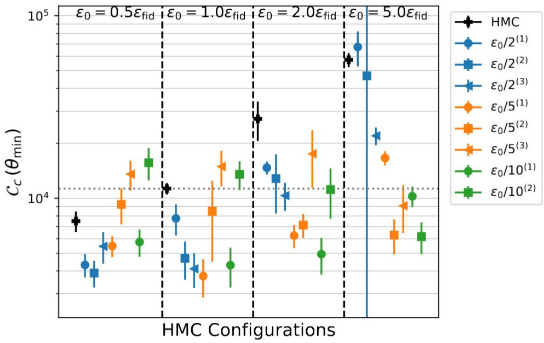
- Delayed rejection HMC for pathological distributions (multiscale distributions like funnel)
  - benefit from multiple, locally optimized transition kernels
- Unlike DR for Metropolis Hastings, cost of a well-tuned DRHMC is a constant factor more than a stable HMC
  - if adapting step size
- Probabilistic DRHMC makes proposals probabilistically, and reduces the cost of DRHMC
- More stable to tuning parameters of HMC (for e.g. step size)
- Any valid transition kernels can be combined into delayed proposals

Thank you

Extra slides

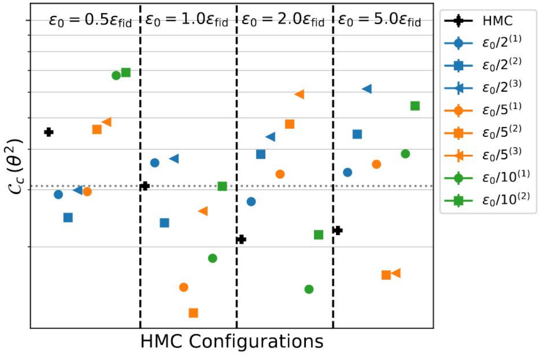
Other experiments

No ideal global proposal!



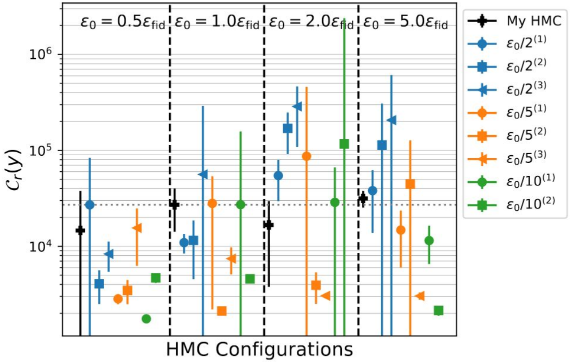
Eight school model- hierarchical, mildly multi-scale

~3x gains



Mixture of 2 Gaussians with different scales (σ = 0.1, 1)

~2x gains



Gull's lighthouse: poor data, ill-defined prior,Cauchy posterior

~5x gains

DRHMC for Neal's funnel

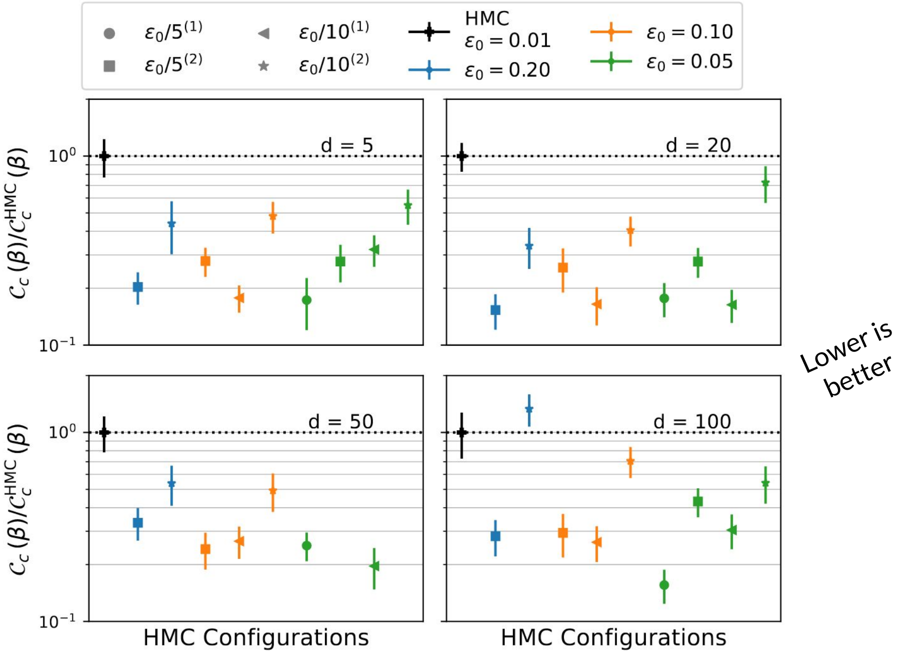
Delayed proposals with reduced step size:

- Starting step size: ε_0
- Step size decreases with factor 'a'
  $\varepsilon_0, \varepsilon_0/a, \varepsilon_0/a^2, ..., \varepsilon_0/a^k....$

Two hyperparameters:

- a: factor of reduction
- k: number of delayed rejections

Similar gains for other hard problems



Comparing cost of DRHMC vs standard HMC
~5-10x gains

Delayed Rejections (DR)

Delayed Rejection : when faced with a rejection, delay it.
Try to make more proposals which might get accepted

Well studied in the context of random-walk Metropolis sampling
Mira 1998, Mira & Tierney 99, Greene & Mira 2001

$$K(x, dy) = \alpha(x, y)Q(x, dy) + r(x)\delta_x(dy)$$

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

$$\begin{aligned} K(x, dy) \;&=\; Q_1(x, dy)\alpha_1(x, y) \\ &+ \int_{s \in S} \underline{Q_1(x, ds)[1 - \alpha_1(x, s)]}[Q_2(x, s, dy)\alpha_2(x, s, y) + r_2(x, s)\delta_x(dy)], \end{aligned}$$

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

Other DR methods in the literature

Extra Chance HMC : when turned down, keep moving on
Sohl-Dickstein et al., 2014; Campos and Sanz-Serna, 2015