Metropolis Adjusted Langevin Trajectories: a robust alternative to Hamiltonian Monte-Carlo.

Lionel Riou-Durand

WARWICK
THE UNIVERSITY OF WARWICK

50th anniversary - Parallel Sessions

September 9, 2022

Joint work with

Jure Vogrinc

Hamiltonian dynamics

- Goal: approximate sampling from a target with density

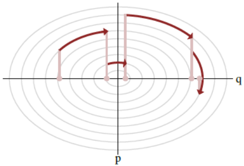
$$\Pi(\boldsymbol{x}) \propto \exp\{-\Phi(\boldsymbol{x})\}, \qquad \boldsymbol{x} \in \mathbb{R}^d.$$

- **A1:** The potential $\Phi \in C^1(\mathbb{R}^d)$ has a Lipschitz gradient

$$\exists M > 0, \qquad |\nabla\Phi(\boldsymbol{x}) - \nabla\Phi(\boldsymbol{y})| \leq M|\boldsymbol{x} - \boldsymbol{y}|, \qquad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

- Hamiltonian dynamics for $t \geq 0$:

$$\mathrm{d}\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{V}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{V}_t \\ -\nabla\Phi(\boldsymbol{X}_t) \end{bmatrix} \mathrm{d}t.$$



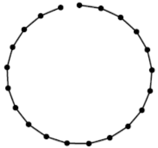
- Invariant measure: $\Pi \otimes \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ with density

$$\Pi_*(\boldsymbol{x}, \boldsymbol{v}) \propto \exp\{-\Phi(\boldsymbol{x}) - |\boldsymbol{v}|^2/2\}, \qquad (\boldsymbol{x}, \boldsymbol{v}) \in \mathbb{R}^{2d}.$$

Leapfrog integrator

- Leapfrog: a standard integrator for Hamiltonian dynamics.
- For a time step $h > 0$, define $\boldsymbol{\theta}_h : (\boldsymbol{x}_0, \boldsymbol{v}_0) \mapsto (\boldsymbol{x}_h, \boldsymbol{v}_h)$ as

$$
\begin{aligned}
\boldsymbol{v}_{h/2} &= \boldsymbol{v}_0 - (h/2)\nabla\Phi(\boldsymbol{x}_0) \\
\boldsymbol{x}_h &= \boldsymbol{x}_0 + h\boldsymbol{v}_{h/2} \\
\boldsymbol{v}_h &= \boldsymbol{v}_{h/2} - (h/2)\nabla\Phi(\boldsymbol{x}_h).
\end{aligned}
$$



- A trajectory is composed of $L = \lceil T/h \rceil$ steps: $\boldsymbol{\theta}_h^L = \boldsymbol{\theta}_h \circ \cdots \circ \boldsymbol{\theta}_h$.

Hamiltonian Monte Carlo

- Duane et al. 1987
- HMC for $h > 0$, $T > 0$. Set $L = \lceil T/h \rceil$.
  - refresh the momentum $\boldsymbol{V}' \leftarrow \boldsymbol{\xi} \sim \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$
  - propose a trajectory $(\boldsymbol{X}_L, \boldsymbol{V}_L) = \boldsymbol{\theta}_h^L(\boldsymbol{X}_0, \boldsymbol{V}')$
  - accept with probability $\pi_*(\boldsymbol{X}_L, \boldsymbol{V}_L)/\pi_*(\boldsymbol{X}_0, \boldsymbol{V}')$
  - if rejected, flip the momentum $(\boldsymbol{X}_L, \boldsymbol{V}_L) \leftarrow (\boldsymbol{X}_0, -\boldsymbol{V}')$
- Remark: full refreshments erase the momentum flips.

Generalized Hamiltonian Monte Carlo

- Horowitz 1991
- GHMC for $h > 0$, $T > 0$, and persistence $\alpha \in [0, 1)$. Set $L = \lceil T/h \rceil$.
  - refresh the momentum $\boldsymbol{V}' \leftarrow \alpha \boldsymbol{V}_0 + \sqrt{1-\alpha^2}\boldsymbol{\xi} \sim \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$.
  - propose a trajectory $(\boldsymbol{X}_L, \boldsymbol{V}_L) = \boldsymbol{\theta}_h^L(\boldsymbol{X}_0, \boldsymbol{V}')$
  - accept with probability $\pi_*(\boldsymbol{X}_L, \boldsymbol{V}_L)/\pi_*(\boldsymbol{X}_0, \boldsymbol{V}')$
  - if rejected, flip the momentum $(\boldsymbol{X}_L, \boldsymbol{V}_L) \leftarrow (\boldsymbol{X}_0, -\boldsymbol{V}')$
- Remark: momentum flips are only partially erased.

HMC: tuning the time step

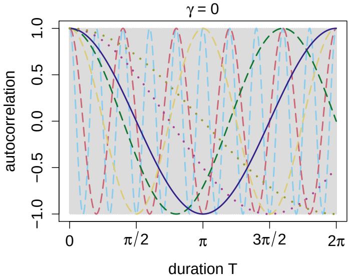
- Choosing h for a given T, when $\alpha = 0$ (full refreshments).
- **A2:** The potential writes $\Phi(\boldsymbol{x}) = \sum_{i=1}^{d} \phi(x_i)$ where $\phi \in C^4(\mathbb{R})$

$$\int_{\mathbb{R}} x^8 \exp\{\phi(x)\} \mathrm{d}x < \infty, \qquad \|\phi^{(k)}\|_\infty < \infty, \qquad k = 2, 3, 4.$$

- Beskos et al. 2013: optimal scaling of the acceptance rate, as $d \to \infty$.
- Choose $h = \ell_T d^{-1/4}$ to get an asym. acceptance rate $a(\ell_T) \approx 65\%$.

HMC: tuning the integration time

- Auto-Correlation Functions: $\rho_i(T) \triangleq \text{Corr}(X_i(T), X_i(0))$, $i = 1, ..., d$.
  - Heterogeneity of scales, Gaussian
  $$\Phi(\boldsymbol{x}) = \sum_{i=1}^{d} x_i^2/(2\sigma_i^2).$$
  - Periodic ACFs
  $$\rho_i(T) = \cos(T/\sigma_i).$$



- The worst ACF $\max_{i \in [\![1,d]\!]} |\rho_i(T)|$ can be arbitrarily erratic and close to 1.
- Bou-Rabee and Sanz-Serna 2017: $T \sim \mathcal{E}xp(\lambda)$, Randomized HMC.
- Smoothing effect: $\mathbb{E}[\rho_i(T)] = \frac{\sigma_i^2}{\sigma_i^2 + \lambda^{-2}} \leq \frac{\sigma_{\max}^2}{\sigma_{\max}^2 + \lambda^{-2}} \Rightarrow$ monotonic.

Langevin diffusion

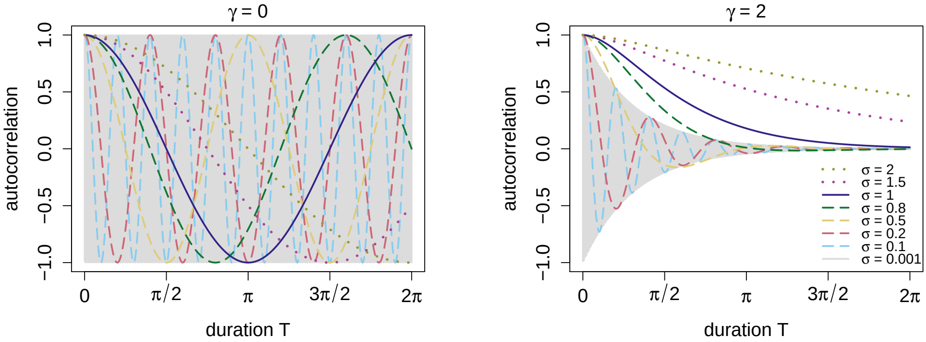
- Damping parameter $\gamma \geq 0$, a.k.a friction.
- Langevin SDE for $t \geq 0$:

$$\mathrm{d}\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{V}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{V}_t \\ -\nabla\Phi(\boldsymbol{X}_t) \end{bmatrix} \mathrm{d}t + \begin{bmatrix} \mathbf{0}_d \\ -\gamma \boldsymbol{V}_t \, \mathrm{d}t + \sqrt{2\gamma}\, \mathrm{d}\boldsymbol{W}_t \end{bmatrix}.$$

- Langevin dynamics = Hamiltonian dynamics with momentum refreshment continuously induced by a Brownian Motion $(\boldsymbol{W}_t)_{t\geq 0}$.
- Same invariant measure: $\Pi \otimes \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$.

Control of the worst ACF

- ACF for HMC and the Langevin diffusion ($\gamma = 2$), for various $\sigma_i > 0$.



- Positive damping enables a uniform control of the correlations

$$\gamma = 2/\sigma_{\max} \Rightarrow \max_{i \in [\![1,d]\!]} |\rho_{i,\gamma}(T)| \leq e^{-T/\sigma_{\max}} \left(1 + T/\sigma_{\max}\right).$$

Quantitative mixing rates

- Randomized HMC with parameters (λ, α), a jump-type SDE for $t \geq 0$:

$$\mathrm{d}\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{V}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{V}_t \\ -\nabla\Phi(\boldsymbol{X}_t) \end{bmatrix} \mathrm{d}t + \begin{bmatrix} \mathbf{0}_d \\ \left(\alpha \boldsymbol{V}_{t-} + \sqrt{1-\alpha^2}\boldsymbol{\xi}_{\boldsymbol{N}_{t-}} - \boldsymbol{V}_{t-}\right) \mathrm{d}\boldsymbol{N}_t \end{bmatrix}.$$

- **A3**: the potential $\Phi \in C^2(\mathbb{R}^d)$, such that for some $M \geq m > 0$

$$m\mathbf{I}_d \preceq \nabla^2\Phi(\boldsymbol{x}) \preceq M\mathbf{I}_d, \qquad \boldsymbol{x} \in \mathbb{R}^d.$$

- **Theorem:** Let $\lambda = \frac{2\sqrt{M+m}}{1-\alpha^2}$, then for any $\alpha \in [0,1)$ we have

$$W_2((\nu\mathbf{P}^t)_{\boldsymbol{x}}, \Pi) \leq Ce^{-rt}W_2(\nu_{\boldsymbol{x}}, \Pi), \qquad \nu = \nu_{\boldsymbol{x}} \otimes \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$$

$$\|\mathbf{P}^t f\| \leq C' e^{-rt}\|f\|, \qquad f \in \mathbb{L}_0^2(\Pi)$$

where

$$r = \frac{1+\alpha}{2}\left(\frac{m}{\sqrt{M+m}}\right), \qquad C, C' \leq 1.56$$

Quantitative mixing rates

- Interpolation of Deligiannidis et al. 2018 and Dalalyan and R-D 2020.
- Randomized HMC and Langevin diffusion generators, for $f \in C_c^\infty(\mathbb{R}^{2d})$.

$$\mathcal{L}^{\mathrm{RH}}_{\lambda,\alpha} \triangleq \mathcal{L}^{\mathrm{H}} + \lambda \mathcal{R}^{\mathrm{PP}}_{\alpha}$$

$$\mathcal{L}^{\mathrm{LD}}_{\gamma} \triangleq \mathcal{L}^{\mathrm{H}} + \gamma \mathcal{R}^{\mathrm{BM}}$$

$$\mathcal{L}^{\mathrm{H}} f(\boldsymbol{x},\boldsymbol{v}) \triangleq \boldsymbol{v}^\top \nabla_{\boldsymbol{x}} f(\boldsymbol{x},\boldsymbol{v}) - \nabla\Phi(\boldsymbol{x})^\top \nabla_{\boldsymbol{v}} f(\boldsymbol{x},\boldsymbol{v})$$

$$\mathcal{R}^{\mathrm{PP}}_{\alpha} f(\boldsymbol{x},\boldsymbol{v}) \triangleq \mathbb{E}\left[f(\boldsymbol{x}, \alpha\boldsymbol{v} + \sqrt{1-\alpha^2}\boldsymbol{\xi})\right] - f(\boldsymbol{x},\boldsymbol{v})$$

$$\mathcal{R}^{\mathrm{BM}} f(\boldsymbol{x},\boldsymbol{v}) \triangleq -\boldsymbol{v}^\top \nabla_{\boldsymbol{v}} f(\boldsymbol{x},\boldsymbol{v}) + \Delta_{\boldsymbol{v}} f(\boldsymbol{x},\boldsymbol{v}).$$

- **Proposition:** If $\lambda = \frac{2\gamma}{1-\alpha^2}$ then $\|\mathcal{L}^{\mathrm{RH}}_{\lambda,\alpha} f - \mathcal{L}^{\mathrm{LD}}_{\gamma} f\|_\infty \to 0$ as $\alpha \to 1$.
- The Langevin diffusion is a limit of Randomized HMC that achieves the fastest exponential mixing rate for strongly log-concave targets.
- Motivates the construction of a sampler drawing Langevin trajectories.

A discretization for Langevin Trajectories

- A standard integrator for Langevin dynamics (**A1** $\Rightarrow$ strong accuracy):

- Set $\alpha = e^{-\gamma h/2}$, let $(\boldsymbol{x}_h, \boldsymbol{v}_h) \sim \mathbf{Q}_{h,\gamma}((\boldsymbol{x}_0, \boldsymbol{v}_0), .)$ such that

$$
\begin{aligned}
\boldsymbol{v}_0' &= \alpha \boldsymbol{v}_0 + \sqrt{1-\alpha^2}\boldsymbol{\xi} \\
\boldsymbol{v}_{h/2} &= \boldsymbol{v}_0' - (h/2)\nabla\Phi(\boldsymbol{x}_0) \\
\boldsymbol{x}_h &= \boldsymbol{x}_0 + h\boldsymbol{v}_{h/2} \\
\boldsymbol{v}_h' &= \boldsymbol{v}_{h/2} - (h/2)\nabla\Phi(\boldsymbol{x}_h) \\
\boldsymbol{v}_h &= \alpha \boldsymbol{v}_h' + \sqrt{1-\alpha^2}\boldsymbol{\xi}'.
\end{aligned}
$$

- Langevin trajectory, $L = \lceil T/h \rceil$ steps: $(\boldsymbol{X}_L, \boldsymbol{V}_L) \sim \mathbf{Q}_{h,\gamma}^L((\boldsymbol{x}_0, \boldsymbol{v}_0), .)$

Metropolis Adjusted Langevin Trajectories

- MALT for $h > 0$, $T > 0$, and damping $\gamma \geq 0$. Set $L = \lceil T/h \rceil$.
  - refresh the momentum $\boldsymbol{V}_0 \leftarrow \boldsymbol{\xi} \sim \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$
  - propose $(\boldsymbol{X}_L, \boldsymbol{V}_L) \sim \mathbf{Q}^L_{h,\gamma}((\boldsymbol{X}_0, \boldsymbol{V}_0), .)$
  - accept with probability

$$\frac{\pi_*(\boldsymbol{X}_L, \boldsymbol{V}_L)}{\pi_*(\boldsymbol{X}_0, \boldsymbol{V}_0)} \times \prod_{i=1}^{L} \frac{q_{h,\gamma}((\boldsymbol{X}_i, -\boldsymbol{V}_i), (\boldsymbol{X}_{i-1}, -\boldsymbol{V}_{i-1}))}{q_{h,\gamma}((\boldsymbol{X}_{i-1}, \boldsymbol{V}_{i-1}), (\boldsymbol{X}_i, \boldsymbol{V}_i))}$$

  - if rejected, flip the momentum $(\boldsymbol{X}_L, \boldsymbol{V}_L) \leftarrow (\boldsymbol{X}_0, -\boldsymbol{V}_0)$

- Remark: full refreshments erase the momentum flips.

Algorithm 1: MALT (h, T, γ), set $L = \lfloor T/h \rfloor$ and $\alpha = \exp\{-\gamma h\}$

for $n \leftarrow 1$ *to* N **do**

  draw fresh momentum start $\boldsymbol{V}' \sim \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$

  set $(\boldsymbol{x}_0, \boldsymbol{v}_0) \leftarrow (\boldsymbol{X}^{n-1}, \boldsymbol{V}')$ and $\Delta \leftarrow 0$

  for $i \leftarrow 1$ *to* L **do**

    draw $\boldsymbol{\xi} \sim \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ and refresh $\boldsymbol{v}'_{i-1} = \alpha \boldsymbol{v}_{i-1} + \sqrt{1-\alpha^2}\boldsymbol{\xi}$

    perform a Leapfrog step $(\boldsymbol{x}_i, \boldsymbol{v}_i) = \boldsymbol{\theta}_h(\boldsymbol{x}_{i-1}, \boldsymbol{v}'_{i-1})$

    update $\Delta \leftarrow \Delta + (|\boldsymbol{v}_i|^2 - |\boldsymbol{v}'_{i-1}|^2)/2$

  end

  set $(\boldsymbol{X}^n, \boldsymbol{V}^n) \leftarrow (\boldsymbol{x}_L, \boldsymbol{v}_L)$ and $\Delta \leftarrow \Delta + \Phi(\boldsymbol{x}_L) - \Phi(\boldsymbol{x}_0)$

  Propose a Langevin trajectory.

  draw a uniform random variable U on $(0, 1)$

  if $U > \exp\{-\Delta\}$ **then**

    reject $\boldsymbol{X}^n \leftarrow \boldsymbol{X}^{n-1}$

  end

end

return $\boldsymbol{X}^1, \cdots, \boldsymbol{X}^N$.

Metropolis Adjusted Langevin Trajectories

- A neat Metropolis adjustment for the Langevin diffusion.
- The length of the trajectories can be chosen by the user.
- Momentum flips can be erased by full refreshments.
- For $\gamma > 0$ the trajectories are ergodic $\Rightarrow$ no U-turns.
- Positive damping enables control of the worst ACF.
- A robust extension to HMC: what about tuning & scaling?

Optimal scaling: an extension to positive friction

- Choosing h for a given T and friction $\gamma \geq 0$?
- **A2**: The potential writes $\Phi(\boldsymbol{x}) = \sum_{i=1}^{d} \phi(x_i)$ where $\phi \in C^4(\mathbb{R})$

$$\int_{\mathbb{R}} x^8 \exp\{\phi(x)\} \mathrm{d}x < \infty, \qquad \|\phi^{(k)}\|_\infty < \infty, \qquad k = 2, 3, 4.$$

- **Theorem**: optimal scaling of the acceptance rate, as $d \to \infty$.
- Choose $h = \ell_T d^{-1/4}$ to get an asym. acceptance rate $a(\ell_T) \approx 65\%$.
- An extension of Beskos et al. 2013 to any friction $\gamma \geq 0$.

Numerical illustration

- Gaussian: $\Phi(\boldsymbol{x}) = \sum_{i=1}^{d} x_i^2/(2\sigma_i^2)$. Heterogeneous scales: $\sigma_i^2 = i/d$.

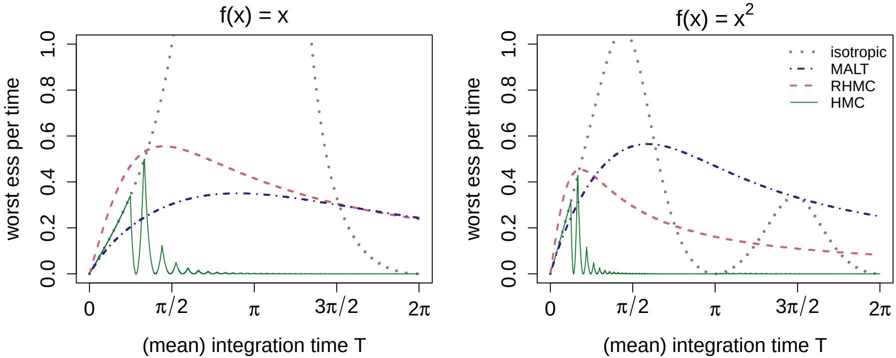


Figure: Gaussian d=50. *Worst ESS per time for estimating the mean and variance.*

Numerical illustration

- Setting: $h > 0$ is fixed to obtain acceptance rates close to 65%.
- Objective: tuning L to obtain good efficiency for every function.

Table: Gaussian d=50. *Worst ESS per gradient evaluation for various functions.*

	odd				even			
$f(x)$	x	x^3	$\mathrm{sgn}(x)$	$\sin(x)$	x^2	x^4	$e^{-\lvert x\rvert}$	$\cos(x)$
MALT: $L = 8$	0.25	0.31	0.31	0.27	**0.40**	**0.42**	**0.43**	**0.40**
RHMC: $L = 5$	**0.40**	**0.43**	**0.45**	**0.41**	0.29	0.31	0.31	0.29
HMC: $L = 3$	0.19	0.25	0.26	0.21	0.00	0.00	0.00	0.00
MALA $(L = 1)$	0.06	0.08	0.09	0.07	0.12	0.12	0.16	0.13

Summary of contributions

- Langevin diffusion is a limit of Randomized HMC that achieves the fastest exponential mixing rate for strongly log-concave targets.
- Positive damping enables control of the worst ACF.
- MALT, a neat Metropolis correction for Langevin trajectories:
  - the length of the trajectories can be chosen by the user
  - momentum flips can be erased by full refreshments
- Optimal scaling, an extension of Beskos et al. 2013: we establish $d^{1/4}$ scaling for any damping, without additional assumptions.

References I

[Bes+13] Alexandros Beskos et al. "Optimal tuning of the hybrid Monte Carlo algorithm". In: *Bernoulli* 19.5A (2013), pp. 1501–1534.

[Bet17] Michael Betancourt. "A conceptual introduction to Hamiltonian Monte Carlo". In: *arXiv preprint arXiv:1701.02434* (2017).

[BS17] Nawaf Bou-Rabee and Jesús Maria Sanz-Serna. "Randomized hamiltonian monte carlo". In: *The Annals of Applied Probability* 27.4 (2017), pp. 2159–2194.

[Del+18] George Deligiannidis et al. "Randomized Hamiltonian Monte Carlo as scaling limit of the bouncy particle sampler and dimension-free convergence rates". In: *arXiv preprint arXiv:1808.04299* (2018).

References II

[DR20] Arnak S Dalalyan and Lionel R-D. “On sampling from a log-concave density using kinetic Langevin diffusions”. In: *Bernoulli* 26.3 (2020), pp. 1956–1988.

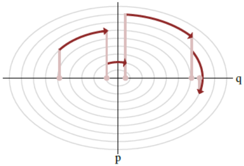
[Dua+87] Simon Duane et al. “Hybrid monte carlo”. In: *Physics letters B* 195.2 (1987), pp. 216–222.

[Hor91] Alan M Horowitz. “A generalized guided Monte Carlo algorithm”. In: *Physics Letters B* 268.2 (1991), pp. 247–252.

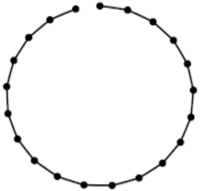
[Nea+11] Radford M Neal et al. “MCMC using Hamiltonian dynamics”. In: *Handbook of markov chain monte carlo* 2.11 (2011), p. 2.

Pictures

- Hamiltonian dynamics: Betancourt 2017



- Leapfrog integrator: Neal et al. 2011



Thank you !