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

Lionel Riou-Durand & Jure Vogrinc, University of Warwick

SCAN ME

Langevin diffusion

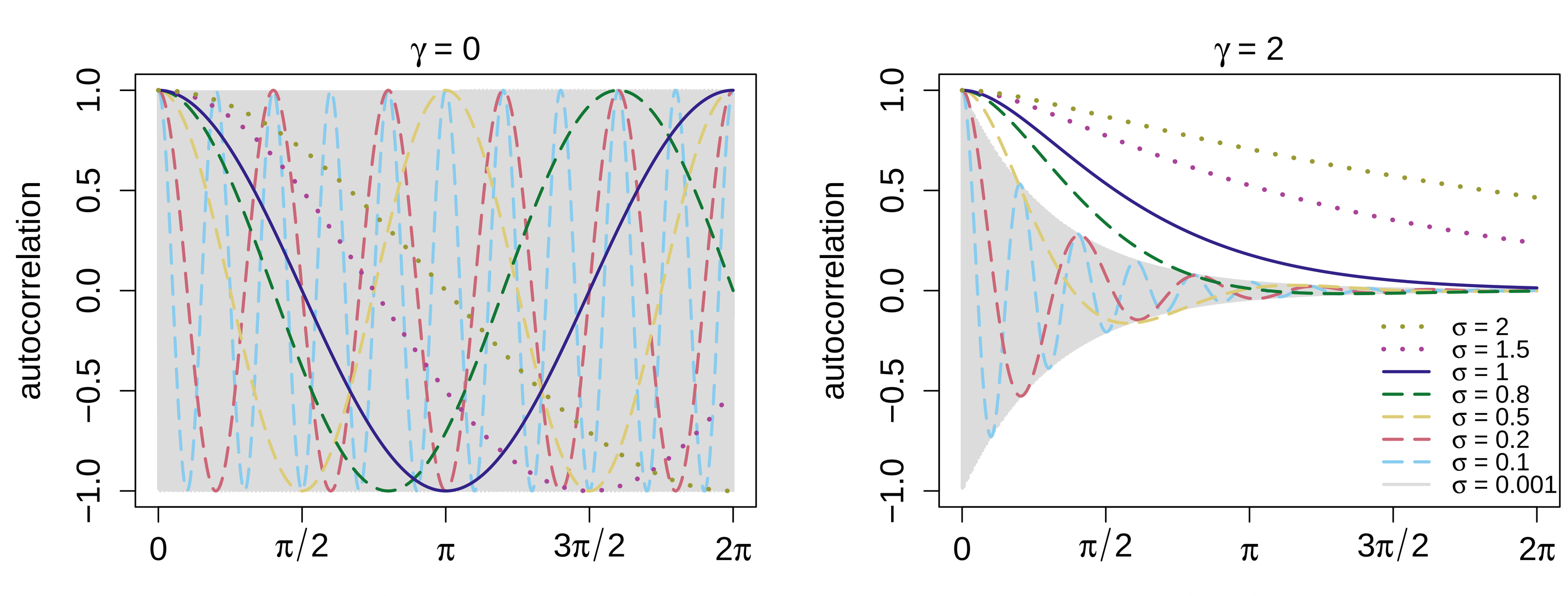
- Goal: sampling from $\pi(\boldsymbol{x}) \propto \exp\{-\Phi(\boldsymbol{x})\}$ for $\boldsymbol{x} \in \mathbb{R}^d$.
- Langevin SDE for $t \geq 0$ and damping $\gamma \geq 0$:
$$\mathrm{d}\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{V}_t\end{bmatrix} = \begin{bmatrix}\boldsymbol{V}_t \\ -\nabla\Phi(\boldsymbol{X}_t)\end{bmatrix}\mathrm{d}t + \begin{bmatrix}\mathbf{0}_d \\ -\gamma \boldsymbol{V}_t\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}\boldsymbol{W}_t\end{bmatrix}.$$
- Langevin dynamics = Hamiltonian dynamics with momentum refreshment continuously induced by a Brownian Motion $(\boldsymbol{W}_t)_{t\geq 0}$.
- Invariant measure: $\pi \otimes \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ with density
$$\pi_*(\boldsymbol{x}, \boldsymbol{v}) \propto \exp\{-\Phi(\boldsymbol{x}) - |\boldsymbol{v}|^2/2\}, \qquad (\boldsymbol{x}, \boldsymbol{v}) \in \mathbb{R}^{2d}.$$

Generalized Hamiltonian Monte Carlo

- 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 if $\alpha > 0$.

Tuning the integration time

- Anisotropic Gaussian: $\Phi(\boldsymbol{x}) = \sum_{i=1}^d x_i^2/(2\sigma_i^2)$, various scales $\sigma_i > 0$.
- Auto-Correlation Functions: $\rho_i(T) \triangleq \mathrm{Corr}(X_i(T), X_i(0))$, $i = 1, ..., d$.



- Randomized HMC: draw T at random to average & smooth the ACFs
$T \sim \mathcal{E}xp(\lambda) \Rightarrow \mathbb{E}[\rho_i(T)] = \frac{\sigma_i^2}{\sigma_i^2+\lambda^{-2}} \leq \frac{\sigma_{\max}^2}{\sigma_{\max}^2+\lambda^{-2}} \Rightarrow$ monotonic.
- Positive damping: an alternative to control the worst ACF
$\gamma = 2/\sigma_{\max} \Rightarrow \max_{i \in 1,d} |\rho_{i,\gamma}(T)| \leq e^{-T/\sigma_{\max}}(1 + T/\sigma_{\max})$.

Quantitative mixing rates

- Randomized HMC with parameters (λ, α), a jump-type SDE for $t \geq 0$:
$$\mathrm{d}\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{V}_t\end{bmatrix} = \begin{bmatrix}\boldsymbol{V}_t \\ -\nabla\Phi(\boldsymbol{X}_t)\end{bmatrix}\mathrm{d}t + \begin{bmatrix}\mathbf{0}_d \\ (\alpha \boldsymbol{V}_{t-} + \sqrt{1-\alpha^2}\boldsymbol{\xi}_{\boldsymbol{N}_{t-}} - \boldsymbol{V}_{t-})\,\mathrm{d}\boldsymbol{N}_t\end{bmatrix}.$$
- **A1**: 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
$$W_2((\nu \mathbf{P}^t)_{\boldsymbol{x}}, \pi) \leq C e^{-rt} W_2(\nu_{\boldsymbol{x}}, \pi), \qquad \nu = \nu_{\boldsymbol{x}} \otimes \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$$
$$\|\mathbf{P}^t f\| \leq C' e^{-rt}\|f\|, \qquad f \in \mathbb{L}_0^2(\pi)$$
where
$$r = \frac{1+\alpha}{2}\left(\frac{m}{\sqrt{M+m}}\right), \qquad C, C' \leq 1.56$$
- Randomized HMC and Langevin diffusion generators, for $f \in C_c^\infty(\mathbb{R}^{2d})$.
$$\mathcal{L}_{\lambda,\alpha}^{\mathrm{RH}} \triangleq \mathcal{L}^{\mathrm{H}} + \lambda \mathcal{R}_\alpha^{\mathrm{PP}}, \qquad \mathcal{L}_\gamma^{\mathrm{LD}} \triangleq \mathcal{L}^{\mathrm{H}} + \gamma \mathcal{R}^{\mathrm{BM}}$$
- **Proposition:** If $\lambda = \frac{2\gamma}{1-\alpha^2}$ then $\|\mathcal{L}_{\lambda,\alpha}^{\mathrm{RH}} f - \mathcal{L}_\gamma^{\mathrm{LD}} f\|_\infty \to 0$ as $\alpha \to 1$.

Time discretization

- 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
$$\begin{aligned}
&\boldsymbol{v}_0' = \alpha \boldsymbol{v}_0 + \sqrt{1-\alpha^2}\boldsymbol{\xi} \\
\text{Leapfrog}\quad \boldsymbol{\theta}_h: &\begin{cases} \boldsymbol{v}_{h/2} = \boldsymbol{v}_0' - (h/2)\nabla f(\boldsymbol{x}_0) \\ \boldsymbol{x}_h = \boldsymbol{x}_0 + h\boldsymbol{v}_{h/2} \\ \boldsymbol{v}_h' = \boldsymbol{v}_0 - (h/2)\nabla f(\boldsymbol{x}_h) \end{cases} \\
&\boldsymbol{v}_h = \alpha \boldsymbol{v}_h' + \sqrt{1-\alpha^2}\boldsymbol{\xi}'
\end{aligned}$$
- Langevin Trajectory for L steps: $(\boldsymbol{x}_{Lh}, \boldsymbol{v}_{Lh}) \sim \mathbf{Q}_{h,\gamma}^L((\boldsymbol{x}_0, \boldsymbol{v}_0), .)$.

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 $(\boldsymbol{X}_L, \boldsymbol{V}_L) \sim \mathbf{Q}_{h,\gamma}^L((\boldsymbol{X}_0, \boldsymbol{V}_0), .)$
  - 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.

Numerical illustration

- Gaussian with scales: $\sigma_i^2 = i/d$. Worst ESS among $d = 50$ components.

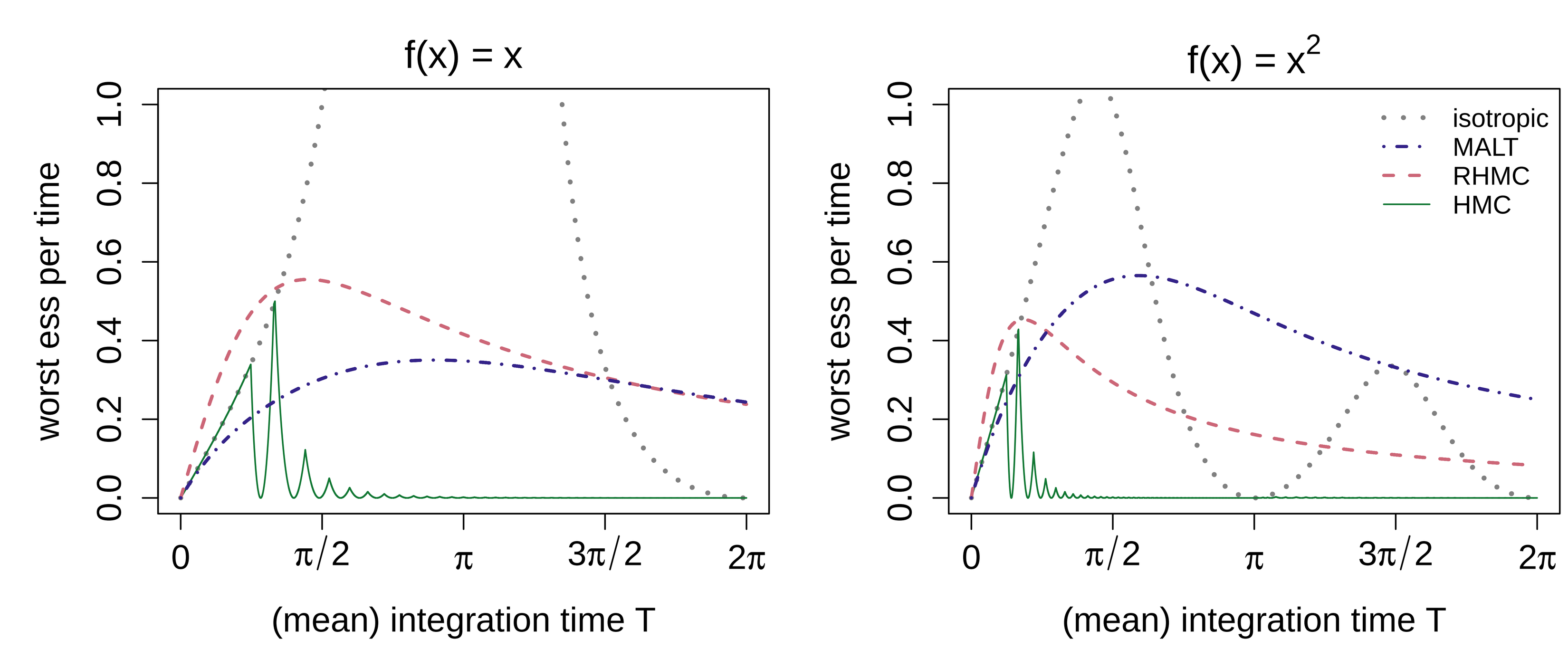


Figure: *Worst ESS per time - mean & variance estimation.*

Optimal scaling

- Optimal choice of 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 HMC's scaling limits to any friction $\gamma \geq 0$.

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: ergodic trajectories enable 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
- A robust extension to HMC: we establish $d^{1/4}$ scaling for MALT for any choice of friction, without additional assumptions.