

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

Lionel Riou-Durand  
University of Warwick



Statistics Seminar, UBC  
May 3, 2022

Joint work with



Jure Vogrinc  
University of Warwick

# Outline

## Tuning for Hamiltonian Monte-Carlo

- Tuning the time step: optimal scaling of the acceptance rate

- Tuning the integration time: control of the worst ACF

- Randomized HMC: a robust solution based on random integration times

## Langevin diffusion

- Robustness: positive damping enables control of the worst ACF

- Connections to Randomized HMC

- Quantitative exponential mixing rates

## Metropolis Adjusted Langevin Trajectories

- MALT: a robust alternative to HMC

- Tuning the time step: optimal scaling for any damping

- Numerical illustrations

# Notations

- Euclidean norm:  $|\mathbf{x}| \triangleq (\mathbf{x}^\top \mathbf{x})^{1/2}, \mathbf{x} \in \mathbb{R}^d$ .
- Supremum norm:  $\|f\|_\infty \triangleq \sup_{\mathbf{x} \in \mathbb{R}^d} |f(\mathbf{x})|$ .
- $\mathbb{L}^2(\pi_*)$  norm:  $\|f\| \triangleq (\int f^2 d\pi_*)^{1/2}$ .
- 2-Wasserstein metric:  $W_2(\nu, \mu) \triangleq \inf\{\mathbb{E}[|\mathbf{X} - \mathbf{Y}|^2]^{1/2}, \mathbf{X} \sim \nu, \mathbf{Y} \sim \mu\}$

# Hamiltonian dynamics

- Goal: approximate sampling from a target with density

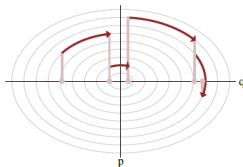
$$\pi(\mathbf{x}) \propto \exp\{-\Phi(\mathbf{x})\}, \quad \mathbf{x} \in \mathbb{R}^d.$$

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

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

- Hamiltonian dynamics for  $t \geq 0$ :

$$d \begin{bmatrix} \mathbf{X}_t \\ \mathbf{V}_t \end{bmatrix} = \begin{bmatrix} \mathbf{V}_t \\ -\nabla\Phi(\mathbf{X}_t) \end{bmatrix} dt.$$



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

$$\pi_*(\mathbf{x}, \mathbf{v}) \propto \exp\{-\Phi(\mathbf{x}) - |\mathbf{v}|^2/2\}, \quad (\mathbf{x}, \mathbf{v}) \in \mathbb{R}^{2d}.$$

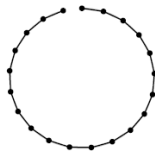
# Leapfrog integrator

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

$$\mathbf{v}_{h/2} = \mathbf{v}_0 - (h/2)\nabla\Phi(\mathbf{x}_0)$$

$$\mathbf{x}_h = \mathbf{x}_0 + h\mathbf{v}_{h/2}$$

$$\mathbf{v}_h = \mathbf{v}_{h/2} - (h/2)\nabla\Phi(\mathbf{x}_h).$$



- Each trajectory is composed of  $L$  leapfrog steps:  $\theta_h^L = \theta_h \circ \dots \circ \theta_h$ .

# Hamiltonian Monte Carlo

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

# Generalized Hamiltonian Monte Carlo

- Horowitz 1991
- GHMC for  $h > 0$ ,  $T > 0$ , and persistence parameter  $\alpha \in [0, 1)$ .
  - set  $L = \lceil T/h \rceil$
  - refresh the momentum  $\mathbf{V}' \leftarrow \alpha \mathbf{V}_0 + \sqrt{1 - \alpha^2} \boldsymbol{\xi} \sim \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ .
  - propose a trajectory  $(\mathbf{X}_L, \mathbf{V}_L) = \boldsymbol{\theta}_h^L(\mathbf{X}_0, \mathbf{V}')$
  - accept with probability  $\pi_*(\mathbf{X}_L, \mathbf{V}_L) / \pi_*(\mathbf{X}_0, \mathbf{V}')$
  - if rejected, flip the momentum  $(\mathbf{X}_L, \mathbf{V}_L) \leftarrow (\mathbf{X}_0, -\mathbf{V}')$
- 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  $\Phi(\mathbf{x}) = \sum_{i=1}^d \phi(x_i)$  where  $\phi \in C^4(\mathbb{R})$

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

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

## HMC: tuning the integration time

- Adaptively setting the path length  $T > 0$ .

$$d|\mathbf{X}_t - \mathbf{X}_0|^2 = 2(\mathbf{X}_t - \mathbf{X}_0)^\top \mathbf{V}_t dt.$$

- Idea: following Hamiltonian dynamics until  $(\mathbf{X}_T - \mathbf{X}_0)^\top \mathbf{V}_T < 0$ .
- Hoffman and Gelman 2014: No-U Turn Sampler.
- A black-box for tuning  $T > 0$ , aiming to maximize the jump distance.
- Euclidean metric... what mixing guarantees for a given component?

## HMC: tuning the integration time

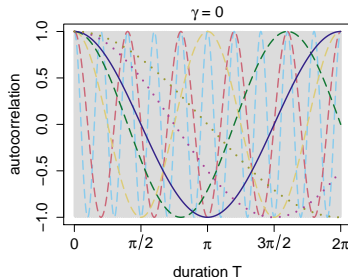
- Auto-Correlation Functions:  $\rho_i(T) \triangleq \text{Corr}(X_i(T), X_i(0))$ ,  $i = 1, \dots, d$ .

- Heterogeneity of scales, Gaussian

$$\Phi(\mathbf{x}) = \sum_{i=1}^d x_i^2 / (2\sigma_i^2).$$

- Periodic ACFs

$$\rho_i(T) = \cos(T/\sigma_i).$$



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

# Summary

## Tuning for Hamiltonian Monte-Carlo

Tuning the time step: optimal scaling of the acceptance rate

Tuning the integration time: control of the worst ACF

Randomized HMC: a robust solution based on random integration times

## Langevin diffusion

Robustness: positive damping enables control of the worst ACF

Connections to Randomized HMC

Quantitative exponential mixing rates

## Metropolis Adjusted Langevin Trajectories

MALT: a robust alternative to HMC

Tuning the time step: optimal scaling for any damping

Numerical illustrations

# Langevin diffusion

- Damping parameter  $\gamma \geq 0$ , a.k.a friction.

- Langevin SDE for  $t \geq 0$ :

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

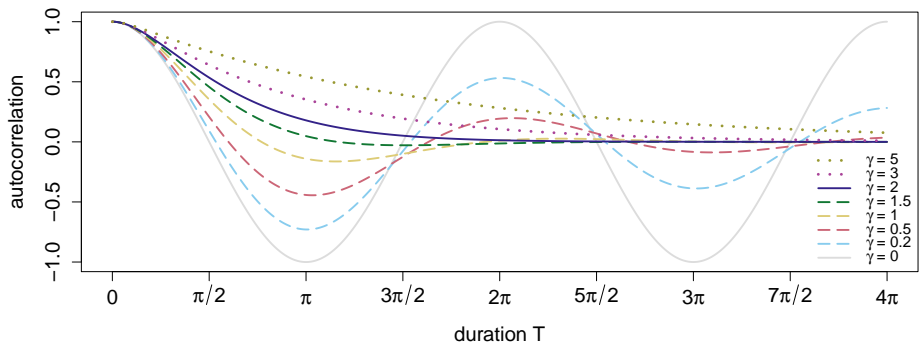
- Langevin dynamics = Hamiltonian dynamics with momentum refreshment continuously induced by a Brownian Motion  $(\mathbf{W}_t)_{t \geq 0}$ .
- Same invariant measure:  $\pi \otimes \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$ .

## Control of the worst ACF

- Running example, Gaussian with heterogeneous scales:

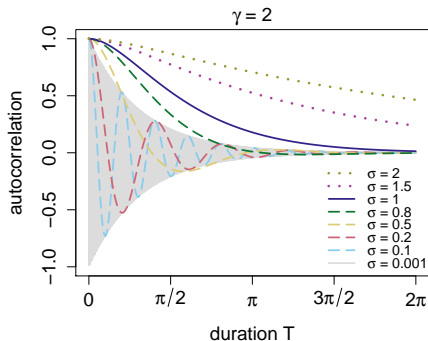
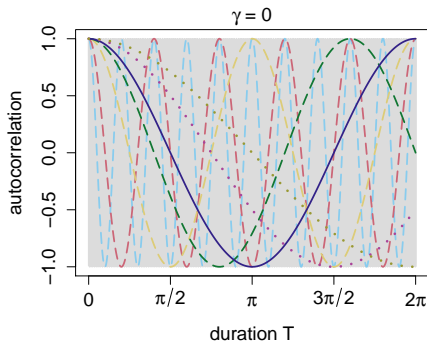
$$\Phi(\mathbf{x}) = \sum_{i=1}^d x_i^2 / (2\sigma_i^2).$$

- ACF for the Langevin diffusion: various frictions  $\gamma \geq 0$  for fixed  $\sigma_i = 1$ .



## 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

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

## Connection with Randomized HMC

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

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

- Randomized HMC and Langevin diffusion generators, for  $f \in C_c^\infty(\mathbb{R}^{2d})$ .

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

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

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

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

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

- **Proposition:** If  $\lambda = \frac{2\gamma}{1-\alpha^2}$  then  $\|\mathcal{L}_{\lambda, \alpha}^{\text{RH}} f - \mathcal{L}_{\gamma}^{\text{LD}} f\|_{\infty} \rightarrow 0$  as  $\alpha \rightarrow 1$ .
- The Langevin Diffusion is a limit of Randomized HMC, obtained as refreshments become infinitesimally partial and frequent.



## Quantitative mixing rates

- Denote  $\mathbf{P}^t$  the transition kernel of RHMC with parameters  $(\lambda, \alpha)$
- **A3:** the potential  $\Phi \in C^2(\mathbb{R}^d)$ , such that for some  $M \geq m > 0$

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

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

$$W_2((\nu \mathbf{P}^t)_{\mathbf{x}}, \pi) \leq C e^{-rt} W_2(\nu_{\mathbf{x}}, \pi), \quad \nu = \nu_{\mathbf{x}} \otimes \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$$

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

where

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

## Quantitative mixing rates

- The Langevin diffusion is a limit of Randomized HMC that achieves the fastest exponential mixing rate for strongly log-concave targets.
- An interpolation between results from Deligiannidis et al. 2018 and Dalalyan and R-D 2020.
- Motivates the construction of a sampling algorithm directly built upon Langevin trajectories.

# Summary

## Tuning for Hamiltonian Monte-Carlo

Tuning the time step: optimal scaling of the acceptance rate

Tuning the integration time: control of the worst ACF

Randomized HMC: a robust solution based on random integration times

## Langevin diffusion

Robustness: positive damping enables control of the worst ACF

Connections to Randomized HMC

Quantitative exponential mixing rates

## Metropolis Adjusted Langevin Trajectories

MALT: a robust alternative to HMC

Tuning the time step: optimal scaling for any damping

Numerical illustrations

## A discretization for Langevin Trajectories

- A standard integrator for Langevin dynamics (**A1**  $\Rightarrow$  strong accuracy):
- Set  $\alpha = e^{-\gamma h/2}$ , let  $(\mathbf{x}_h, \mathbf{v}_h) \sim \mathbf{Q}_{h,\gamma}((\mathbf{x}_0, \mathbf{v}_0), \cdot)$  such that

$$\mathbf{v}'_0 = \alpha \mathbf{v}_0 + \sqrt{1 - \alpha^2} \boldsymbol{\xi}$$

$$\mathbf{v}_{h/2} = \mathbf{v}'_0 - (h/2) \nabla \Phi(\mathbf{x}_0)$$

$$\mathbf{x}_h = \mathbf{x}_0 + h \mathbf{v}_{h/2}$$

$$\mathbf{v}'_h = \mathbf{v}_{h/2} - (h/2) \nabla \Phi(\mathbf{x}_h)$$

$$\mathbf{v}_h = \alpha \mathbf{v}'_h + \sqrt{1 - \alpha^2} \boldsymbol{\xi}'.$$

- Langevin Trajectory for  $L$  steps:  $(\mathbf{x}_{Lh}, \mathbf{v}_{Lh}) \sim \mathbf{Q}_{h,\gamma}^L((\mathbf{x}_0, \mathbf{v}_0), \cdot)$ .

# Metropolis Adjusted Langevin Trajectories

- MALT for friction  $\gamma \geq 0$ , time step  $h > 0$ , integration time  $T > 0$ .

- set  $L = \lceil T/h \rceil$
- refresh the momentum  $\mathbf{V}_0 \leftarrow \boldsymbol{\xi} \sim \mathcal{N}_d(\mathbf{0}_d, \mathbf{I}_d)$
- propose  $(\mathbf{X}_L, \mathbf{V}_L) \sim \mathbf{Q}_{h,\gamma}^L((\mathbf{X}_0, \mathbf{V}_0), \cdot)$
- accept with probability

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

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

- Remark: full refreshments erase the momentum flips.

# 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: we establish  $d^{1/4}$  scaling for MALT for any choice of friction, without additional assumptions.

## Optimal scaling: an extension to positive friction

- Choosing  $h$  for a given  $T$  and friction  $\gamma \geq 0$ ?

- **A2:** The potential writes  $\Phi(\mathbf{x}) = \sum_{i=1}^d \phi(x_i)$  where  $\phi \in C^4(\mathbb{R})$

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

- **Theorem:** optimal scaling of the acceptance rate, as  $d \rightarrow \infty$ .
- 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 illustrations

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

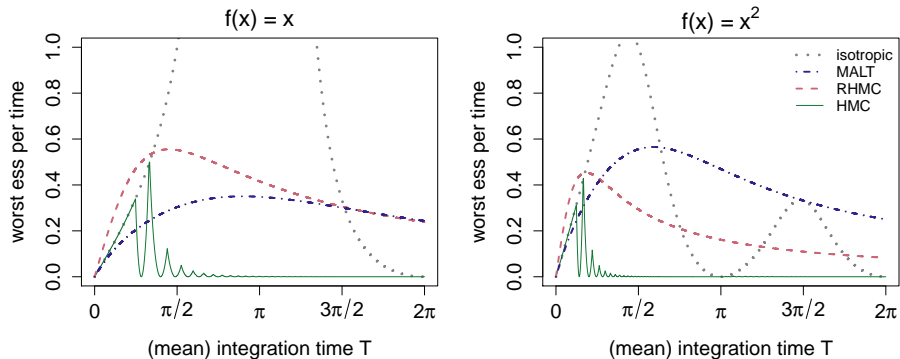


Figure: Gaussian  $d=50$ . Worst ESS per time for estimating the mean and variance.



## Numerical illustrations

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

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

	odd				even			
$f(x)$	$x$	$x^3$	$\text{sgn}(x)$	$\sin(x)$	$x^2$	$x^4$	$e^{- x }$	$\cos(x)$
MALT: $L = 8$	0.25	0.31	0.31	0.27	<b>0.40</b>	<b>0.42</b>	<b>0.43</b>	<b>0.40</b>
RHMC: $L = 5$	<b>0.40</b>	<b>0.43</b>	<b>0.45</b>	<b>0.41</b>	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

## Tuning for Hamiltonian Monte-Carlo

Tuning the time step: optimal scaling of the acceptance rate

Tuning the integration time: control of the worst ACF

Randomized HMC: a robust solution based on random integration times

## Langevin diffusion

Robustness: positive damping enables control of the worst ACF

Connections to Randomized HMC

Quantitative exponential mixing rates

## Metropolis Adjusted Langevin Trajectories

MALT: a robust alternative to HMC

Tuning the time step: optimal scaling for any damping

Numerical illustrations

## 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 MALT, without additional assumptions.

# Perspectives

- Adaptive tuning for  $h$ ,  $T$  and  $\gamma$ , software implementations.
- Further comparisons with HMC, GHMC, Randomized HMC.

# 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

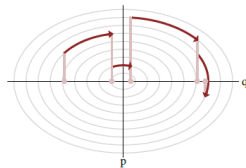
- [DR20]     Arnak S Dalalyan and Lionel R-D. “On sampling from a log-concave density using kinetic Langevin diffusions”. In: *Bernoulli* 26.3 (2020), pp. 1956–1988.
- [Dua+87]   Simon Duane et al. “Hybrid monte carlo”. In: *Physics letters B* 195.2 (1987), pp. 216–222.
- [HG14]     Matthew D Hoffman and Andrew Gelman. “The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.”. In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 1593–1623.
- [Hor91]     Alan M Horowitz. “A generalized guided Monte Carlo algorithm”. In: *Physics Letters B* 268.2 (1991), pp. 247–252.

## References III

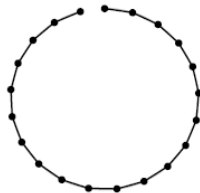
- [Nea+11] Radford M Neal et al. “MCMC using Hamiltonian dynamics”. In: *Handbook of markov chain monte carlo* 2.11 (2011), p. 2.

# Pictures

- Hamiltonian dynamics: Betancourt 2017



- Leapfrog integrator: Neal et al. 2011





Thank you !