

When does Metropolized Hamiltonian Monte Carlo provably outperform Metropolis-adjusted Langevin algorithm?

UQSay seminars

Yuansi Chen

joint work with Khashayar Gatmiry
and Minhui Jiang

ETH Zürich



Markov chain Monte Carlo setup

Given $\mu(x) \propto e^{-f(x)}$ on \mathbb{R}^d , we want to sample from μ .

Bayesian inference as computing an integral w.r.t. μ :

$$\int g(\theta) d\mu(\theta) \approx \frac{1}{N} \sum_{i=1}^N g(\theta_i), \quad \theta_i \sim \mu$$

Markov chain Monte Carlo setup

Given $\mu(x) \propto e^{-f(x)}$ on \mathbb{R}^d , we want to sample from μ .

Bayesian inference as computing an integral w.r.t. μ :

$$\int g(\theta) d\mu(\theta) \approx \frac{1}{N} \sum_{i=1}^N g(\theta_i), \quad \theta_i \sim \mu$$

MCMC approach

- Construct a Markov chain with transition kernel $P(x, dy)$
- Run k steps to get $\theta_k \sim \mu_0 P^k$
- Hope for large k , $\mu_0 P^k \approx \mu$

Hamiltonian Monte Carlo (HMC) is default for continuous measure in many packages



Algorithms > MCMC Sampling **Stan**

MCMC Sampling

This chapter presents the two Markov chain Monte Carlo (MCMC) algorithms used in Stan, the Hamiltonian Monte Carlo (HMC) algorithm and its adaptive variant the no-U-turn sampler (NUTS), along with details of their implementation and configuration.

- No-U-Turn Sampler (NUTS) is the default for continuous measure
- NUTS is a specialized form of Hamiltonian Monte Carlo (HMC) that automatically tunes its parameters

Why studying HMC?

- HMC originates from physics literature [Alder & Wainwright '59], [Duane, Kennedy, Pendleton & Roweth '87]
- Introduced to statistics in [Neal '94]
- known to scale well in high dimensions empirically

Our goal:

- Quantify the non-asymptotic convergence rate of HMC: how large k should be to guarantee $d_{\text{TV}}(\mu_0 P^k, \mu) \leq \epsilon$?
- Enable rigorous comparison for MCMC algorithms

Simplified setting for rigorous comparison

Given access to $\mu \propto e^{-f}$, want to generate samples

Two settings:

1. Smooth and strongly log-concave target:

$$m\mathbb{I}_d \preceq \nabla^2 f(x) \preceq L\mathbb{I}_d, \quad \forall x \in \mathbb{R}^d$$

2. Smooth with isoperimetry

- $\|\nabla^2 f(x)\|_2 \leq L, \quad \forall x \in \mathbb{R}^d$
- μ satisfies isoperimetric inequality with constant ψ : for any partition S_1, S_2, S_3 of \mathbb{R}^d

$$\mu(S_3) \geq \psi \cdot d(S_1, S_2) \cdot \min\{\mu(S_1), \mu(S_2)\}$$

- may add higher-order smoothness

Metropolis-adjusted Langevin algorithm (MALA)

Langevin diffusion

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t$$

MALA

- Euler-Maruyama discretization

$$y = x - h\nabla f(x) + \sqrt{2h}\xi, \quad \xi \sim \mathcal{N}(0, I_d)$$

- Metropolis-Hastings accept/reject step, returning y with probability $\alpha(x, y)$

$$\begin{aligned}\alpha(x, y) &= \min \left\{ 1, \frac{\mu(y)p(y, x)}{\mu(x)p(x, y)} \right\} \\ &= \min \left\{ 1, e^{f(x)-f(y)+\frac{1}{4h}(\|y-x+h\nabla f(x)\|^2-\|x-y+h\nabla f(y)\|^2)} \right\}\end{aligned}$$

mainly because we desire high-accuracy samplers

- Metropolized algorithms are high-accuracy samplers
[Dwivedi, C., Wainwright & Yu '19]: mixing time $\propto \text{polylog}(1/\epsilon)$
- Unadjusted algorithms, such as unadjusted Langevin algorithm (ULA) has mixing time $\propto 1/\epsilon^2$ [Dalalyan '17]

Hamilton's equations

Start with a Hamiltonian system with potential energy $f(q)$ and kinetic energy $\frac{1}{2} \|p\|_2^2$

$$\mathcal{H}(q, p) = f(q) + \frac{1}{2} \|p\|_2^2.$$

Hamilton's equations express the dynamics of the system:

$$\begin{aligned}\frac{dq_t}{dt} &= \frac{\partial \mathcal{H}}{\partial p} \\ \frac{dp_t}{dt} &= -\frac{\partial \mathcal{H}}{\partial q}\end{aligned}$$

Hamilton's equations

Start with a Hamiltonian system with potential energy $f(q)$ and kinetic energy $\frac{1}{2} \|p\|_2^2$

$$\mathcal{H}(q, p) = f(q) + \frac{1}{2} \|p\|_2^2.$$

Hamilton's equations express the dynamics of the system:

$$\begin{aligned}\frac{dq_t}{dt} &= p_t \\ \frac{dp_t}{dt} &= -\nabla f(q_t)\end{aligned}$$

In words,

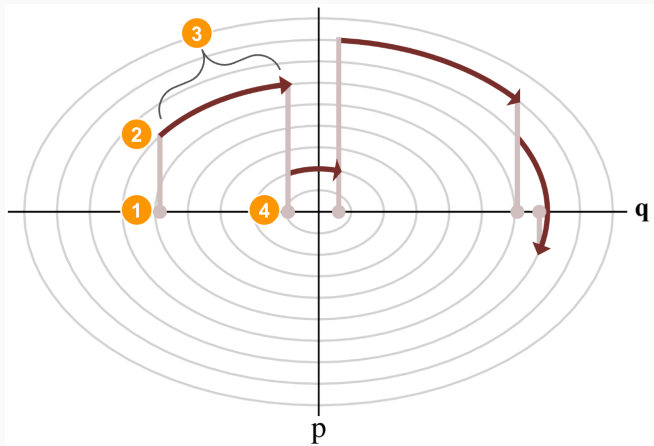
- rate of change of position q_t is given by momentum p_t
- rate of change of momentum p_t is given by negative gradient of potential energy

Starting from current position \mathbf{q}_0 , perform:

- Sample initial momentum $\mathbf{p}_0 \sim \mathcal{N}(0, I_d)$
- Evolve according to Hamilton's equations for time T to get proposal $(\mathbf{q}_T, \mathbf{p}_T)$
- Return \mathbf{q}_T as the next sample

It preserves the joint distribution $\mu \times \mathcal{N}(0, \mathbb{I}_d) \propto e^{-f(q) - \frac{1}{2}\|p\|_2^2}$

Illustration



HMC + leapfrog integrator + Metropolis-Hastings

Integer $K \geq 1$ and step-size $\eta > 0$. Iteratively do, from current state (q_0, p_0) :

- **Leapfrog integrator**: for $k = 0, 1, \dots, K - 1$

$$p_{k+\frac{1}{2}} = p_k - \frac{\eta}{2} \nabla f(q_k)$$

$$q_{k+1} = q_k + \eta p_{k+\frac{1}{2}}$$

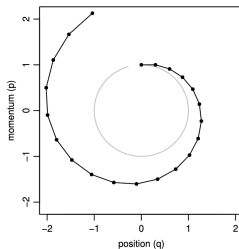
$$p_{k+1} = p_{k+\frac{1}{2}} - \frac{\eta}{2} \nabla f(q_{k+1})$$

- Metropolis-Hastings accept/reject step, returning q_K with probability $\alpha((q_0, p_0), (q_K, p_K))$

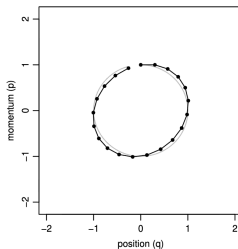
$$\alpha((q_0, p_0), (q_K, p_K)) = \min \left\{ 1, e^{-\mathcal{H}(q_K, p_K) + \mathcal{H}(q_0, p_0)} \right\}$$

Illustration of leapfrog integrator

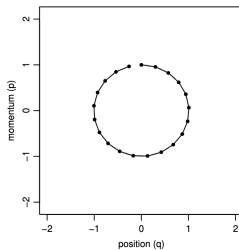
(a) Euler's Method, stepsize 0.3



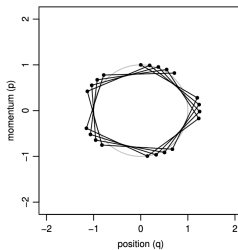
(b) Modified Euler's Method, stepsize 0.3

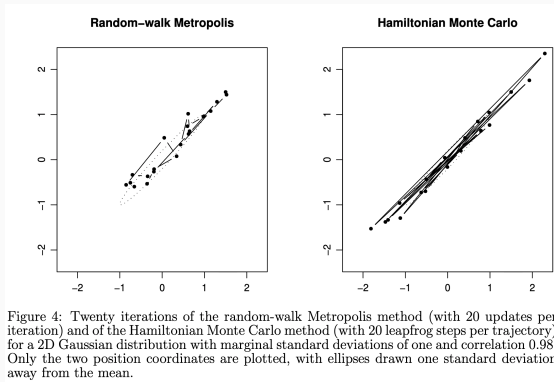


(c) Leapfrog Method, stepsize 0.3



(d) Leapfrog Method, stepsize 1.2





[Neal '94]

MALA as a special case of leapfrog HMC

MALA can be viewed as a special case of leapfrog HMC with $K = 1$ and step-size $h = \frac{\eta^2}{2}$:

$$\begin{aligned}q_1 &= q_0 + \eta p_{\frac{1}{2}} \\ &= q_0 + \eta \left(p_0 - \frac{\eta}{2} \nabla f(q_0) \right) \\ &= q_0 - \frac{\eta^2}{2} \nabla f(q_0) + \eta p_0\end{aligned}$$

Mixing time in total variation distance

$$\tau_{\text{mix}}(\epsilon, \mu_0) = \min \left\{ k \in \mathbb{N} \mid d_{\text{TV}}(\mu_0 P^k, \mu) \leq \epsilon \right\}$$

To enable fair comparison between MALA and HMC,
number of gradient evaluations is used

$$\# \text{grad evals} = \tau_{\text{mix}} * \# \text{grad evals per step}$$

Existing results on MALA and HMC

- **Optimal scaling:** $d^{-\frac{1}{3}}$ asymptotic step-size to achieve constant acceptance rate for smooth product measure
[Roberts & Rosenthal '98] \implies relaxation time $\lesssim d^{\frac{1}{3}}$ for smooth product measure (achieved by Gaussian)
- Other results with higher-order smoothness assumption
[Bou-Rabee et al. '10], [Eberle '14], [Hairer et al. '14] ...

Nonasymptotic results are motivated by [Dalalyan '14]:

“... a striking fact is that the convergence properties of optimisation algorithms are much better understood than those of the approximate sampling algorithms”

- Huge literature on ULA and underdamped Langevin, but not the focus of our talk

Previous nonasymptotic results on MALA (2)

For L -smooth and m -strongly log-concave sampling ($\kappa = L/m$):

- **Without warm-start:** $\tau_{\text{mix}} \lesssim \kappa d$ [Dwivedi, C., Wainwright & Yu '18]; [C. Dwivedi, Wainwright & Yu '19], [Lee, Shen, Tian '20]
 - Shown to be tight (\approx) in [Lee, Shen, Tian '21]
- **With warm-start ($\frac{\mu_0}{\mu} \leq M$),** up to poly-log(M/ϵ)
 - $\tau_{\text{mix}} \lesssim \kappa^{\frac{3}{2}} d^{\frac{1}{2}}$ [Chewi, Lu, Ahn, Cheng, Le Gouic, Rigollet '21]
 - $\tau_{\text{mix}} \approx \kappa d^{\frac{1}{2}}$ [Wu, Schmidler, C. '22], also tight

Previous nonasymptotic results on MALA (2)

For L -smooth and m -strongly log-concave sampling ($\kappa = L/m$):

- **Without warm-start:** $\tau_{\text{mix}} \lesssim \kappa d$ [Dwivedi, C, Wainwright & Yu '18]; [C. Dwivedi, Wainwright & Yu '19], [Lee, Shen, Tian '20]
 - Shown to be tight (\approx) in [Lee, Shen, Tian '21]
- **With warm-start ($\frac{\mu_0}{\mu} \leq M$), up to poly-log(M/ϵ)**
- $\tau_{\text{mix}} \lesssim \kappa^{\frac{3}{2}} d^{\frac{1}{2}}$ [Chewi, Lu, Ahn, Cheng, Le Gouic, Rigollet '21]
- $\tau_{\text{mix}} \approx \kappa d^{\frac{1}{2}}$ [Wu, Schmidler, C. '22], also tight

Extended to isoperimetric setting: $\|\nabla^2 f\|_2 \leq L$, isoperimetric constant ψ and sub-exponential tail [C., Gatmiry '23],

$\Upsilon = \sup_x \text{Tr}(\nabla^2 f_x)$:

$$\tau_{\text{mix}} \lesssim \frac{(L\Upsilon)^{\frac{1}{2}}}{\psi^2}$$

For L -smooth and m -strongly log-concave sampling ($\kappa = L/m$):

$$\min_{h>0} \max_{\mu} \max_{\mu_0 \text{ warm}} \tau_{\text{mix}}(\epsilon, \mu_0) \propto \kappa d^{\frac{1}{2}} \text{poly-log}(M/\epsilon).$$

For L -smooth and m -strongly log-concave sampling ($\kappa = L/m$):

$$\min_{h>0} \max_{\mu} \max_{\mu_0 \text{ warm}} \tau_{\text{mix}}(\epsilon, \mu_0) \propto \kappa d^{\frac{1}{2}} \text{poly-log}(M/\epsilon).$$

Worst-case construction: perturbed Gaussian (inspired by [Lee, Shen, Tian '20], [Chewi, Lu, Ahn, Cheng, Le Gouic, Rigollet '21])

$$f_{\delta}(x) = \frac{L}{2} \sum_{i=1}^{d-1} x_{[i]}^2 - \frac{1}{2d^{\frac{1}{2}-2\delta}} \sum_{i=1}^{d-1} \cos\left(d^{\frac{1}{4}-\delta} L^{\frac{1}{2}} x_{[i]}\right) + \frac{m}{2} x_{[d]}^2$$

Can warm-start be constructed for MALA?

For log-concave sampling: **yes! rigorously!**

- $\mathcal{N}(0, \frac{1}{L}\mathbb{I}_d)$ is only κ^d -warm
 - \implies additional d loss due to $\log M$ (not good enough!)
- Unadjusted underdamped Langevin algorithm serves as a warm-start with only $O(d^{\frac{1}{2}})$ cost [Chewi & Altschuler '23]

Can warm-start be constructed for MALA?

For log-concave sampling: **yes! rigorously!**

- $\mathcal{N}(0, \frac{1}{L}\mathbb{I}_d)$ is only κ^d -warm
 \implies additional d loss due to $\log M$ (not good enough!)
- Unadjusted underdamped Langevin algorithm serves as a warm-start with only $O(d^{\frac{1}{2}})$ cost [Chewi & Altschuler '23]

Overall, $\kappa d^{\frac{1}{2}}$ is tight for MALA for log-concave sampling, $d^{\frac{1}{3}}$ for Gaussian

Q: When does Metropolized HMC with leapfrog integrator provably outperform MALA?

Can we quantify these:

- Non-asymptotic convergence rate of HMC?
- When is it beneficial to choose $K > 1$ instead of $K = 1$?
- How does the answer depend on the target measure?

Intuition from numerical analysis viewpoint

- Euler-Maruyama method is a **first-order** integrator
- Leapfrog method is a **second-order** integrator

With enough smoothness on the target measure, we expect better acceptance rate and thus better convergence rate for HMC

- **Optimal scaling:** $d^{-\frac{1}{4}}$ asymptotic step-size to achieve constant acceptance rate for smooth product measure (with high-order derivatives) [Beskos, Pillai, Roberts, Sanz-Serna & Stuart '13] \implies relaxation time $\lesssim d^{\frac{1}{4}}$ (achieved by Gaussian)

Previous nonasymptotic results on Metropolized HMC (1)

For L -smooth and m -strongly log-concave sampling ($\kappa = L/m$):

- Without warm-start (lower-bound construction):
grad evals $\gtrsim \kappa d^{\frac{1}{2}}$ [Lee, Shen, Tian '20]
- With warm-start:
 - $d^{1/4}$ for Gaussian
 - We can pick $K = 1$, then HMC=MALA, so $d^{1/2}$
 - Can we do better with $K > 1$? Yes. $\kappa d^{1/4}$ with third-order smoothness [C., Gtmiry & Jiang '26]

Assume

- L -smooth, $\|\nabla^2 f\|_2 \leq L$
- isoperimetric constant ψ
- sub-exponential tail
- third-order smoothness: $\gamma L^{\frac{3}{2}}$ -Lipschitz Hessian ($\gamma = O(1)$)

$$\|\nabla^2 f_x - \nabla^2 f_y\|_2 \leq \gamma L^{\frac{3}{2}} \|x - y\|_2, \quad \forall x, y \in \mathbb{R}^d$$

then from a warm start,

$$\tau_{\text{mix}}^{\text{HMC}} \lesssim \frac{1}{K^2 \eta^2 \psi^2}$$

for $K\eta + K\eta^3 d^{\frac{1}{2}} + K\eta^5 d + K\eta^7 d^{\frac{3}{2}} = O(1)$

Assume

- $m\mathbb{I}_d \preceq \nabla^2 f(x) \preceq L\mathbb{I}_d$, (then isoperimetry $\psi \gtrsim m^{\frac{1}{2}}$)
- $\gamma L^{\frac{3}{2}}$ -Lipschitz Hessian ($\gamma = O(1)$)

then from a warm start, for choice $K\eta \approx 1, \eta \approx \frac{1}{d^{\frac{1}{4}} L^{\frac{1}{2}}}$,

$$\tau_{\text{mix}}^{\text{HMC}} \lesssim \kappa d^{\frac{1}{4}}$$

Proof sketch

Conductance based mixing time proof

- First developed for finite-state Markov chains (e.g. [Levin, Peres & Wilmer '09])
- Popularized in continuous-state by [Lovasz & Simonovits '93]
- [Lovasz & Simonovits '93] reduces mixing time analysis to two separate parts:
 - Global geometry (governed by isoperimetry ψ)
 - local transition overlap: for $\|x - y\|_2$ small, can we bound

$$d_{\text{TV}}(P_x, P_y) \leq ?$$

Conductance based proof - step 1

Let $P_x = P(x, \cdot)$ be the transition kernel at x . **Conductance:**

$$\Phi := \inf_{\mu(A) \leq \frac{1}{2}} \frac{\int_A P_x(A^c) d\mu(x)}{\mu(A)}$$

[Lov'asz & Simonovits '93]

Conductance lower bound \implies mixing time upper bound from warm start:

$$\tau_{\text{mix}} \lesssim \frac{1}{\Phi^2} \log \frac{M}{\epsilon}$$

[Lovász '99] provides a way to lower bound conductance through local calculation of transition overlap (local-to-global scheme):

[Lovász '99]

If $\|x - y\|_2 \leq r \implies d_{TV}(P_x, P_y) \leq 0.99$, then

$$\Phi \gtrsim r\psi$$

Transition overlap

To control the transition overlap $d_{TV}(P_x, P_y)$, we bound

- Proposal overlap $d_{TV}(Q_x, Q_y)$, where Q_x is before accept/reject step
- Acceptance rate

Transition overlap

To control the transition overlap $d_{\text{TV}}(P_x, P_y)$, we bound

- Proposal overlap $d_{\text{TV}}(Q_x, Q_y)$, where Q_x is before accept/reject step
- Acceptance rate

For **HMC**

- Q_x is the push-forward of $\mathcal{N}(0, \mathbb{I}_d)$ through the leapfrog integrator K times
- Acceptance rate is basically discretization error in Hamiltonian \mathcal{H}
 - Well-studied in low-dim [\[Hairer, Lubich & Wanner '06\]](#)
 - High-dim concentration and careful integration by parts is needed for high-dim analysis

Tuning MALA:

- For L -log-smooth sampling, step-size $h \approx \frac{1}{Ld^{\frac{1}{2}}}$ is needed to maintain good acceptance rate from a warm-start
- OK to pick $h \approx \frac{1}{Ld^{\frac{1}{3}}}$ if Gaussian-like target
- For non-warm-start
 - It is not good to tune h based on acceptance rate, since it forces h to be too small (like $\frac{1}{Ld}$)
 - Better to use ULA/underdamped Langevin as a warm-start

Tuning HMC:

- We don't really have a lower bound yet, so be skeptical
- third-order smoothness is needed for $K > 1$ to achieve $d^{\frac{1}{4}}$
- From a warm start, to achieve $d^{\frac{1}{4}}$ # grad evals, take integration time $K\eta \approx 1/L^{\frac{1}{2}}$ and $K \approx d^{\frac{1}{4}}$

Future directions and open problems

For Gaussian, HMC with **longer and random** integration time $K\eta \approx 1/m^{\frac{1}{2}}$ achieves $\sqrt{\kappa}d^{\frac{1}{4}}$ (**Nesterov-type acceleration or hypocoercivity!**), we only did short integration time analysis $K\eta \approx 1/L^{\frac{1}{2}}$

1. [Eberle & Lörler '24] did L^2 -analysis for continuous HMC to achieve $\sqrt{\kappa}$ via lifting and space-time Poincaré inequality
Can we do it for the actual HMC algorithm?
2. No-U-Turn Sampler (NUTS) is designed for acceleration, but it does not always accelerate [Oberdörster '25]
Can we design better adaptive tuning?
3. **Other high-order numerical schemes?**

The logo for PolytopeWalk features the word "POLYTOPEWALK" in a light blue, sans-serif font. The letter "P" is a solid dark blue. The text is centered within a white horizontal bar that has a dark blue background on either side. On the left side, a red diagonal line cuts through the white bar. On the right side, a light blue diagonal line cuts through the white bar, forming a chevron shape pointing to the right.

POLYTOPEWALK

Supports

- Uniform distribution truncated on a polytope (working on log-concave now)
- preprocessing of the constraint set via facial reduction [Borwein & Wolkowicz '80]
- sparse specification of the constraint set
- feasible initialization

<https://github.com/ethz-randomwalk/polytopewalk>

[Sun & C. '24]

Thank you!