Adjustable Randomized Time Integrators for Hamiltonian Monte Carlo

Nawaf Bou-Rabee

Rutgers University Camden

joint with Milo Marsden @ Stanford

1. MCMC

Aims to approximately sample from a target probability measure μ by simulating a Markov chain

$$X_0, X_1, \ldots \quad \text{s.t.} \quad \mathrm{Law}(X_n) \approx \mu .$$

$$\boxed{\text{How many MCMC steps till } \mathrm{Law}(X_n) \approx \mu \, ?}$$

2. Hamiltonian MCMC

Hamiltonian Monte Carlo (HMC) is a class of MCMC methods for sampling

possibly rough.

$$\mu(dx) \propto \exp(-U(x))\, \lambda^d(dx)$$

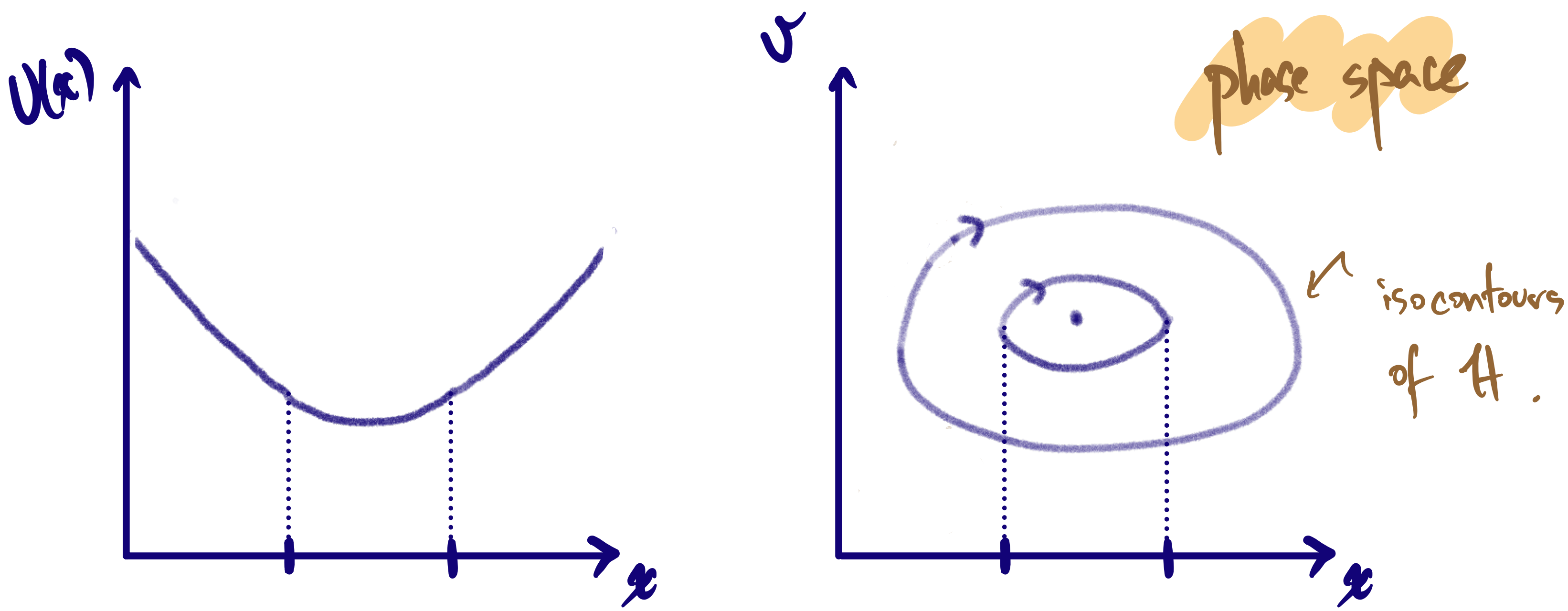
where $U: \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable.

> Motivated in part by success of momentum-based algorithms for optimization: Polyak and Nesterov's methods.

HMC constructs a Markov chain "aimed" at

$$\mu_{BG}(dx\, dv) \propto e^{-\mathcal{H}(x,v)} \lambda^d(dx)\, \lambda^d(dv)$$

where $\mathcal{H}(x,v) = \frac{1}{2}|v|^2 + U(x)$.



HMC uses a fictitious Hamiltonian dynamics

$$\dot{q}_t = v_t \qquad \dot{v}_t = -\nabla U(q_t)$$

Let $\varphi_t : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ be the corresponding flow.

$$(1) \quad H \circ \varphi_t \overset{\text{energy conservation}}{=} H \qquad (2) \quad \lambda^{2d} \circ \varphi_t \overset{\text{volume conservation}}{=} \lambda^{2d}$$

> However, Hamiltonian dynamics, by itself, preserves infinitely many invariant measures.

Exact HMC Algorithm

Input : duration hyperparameter $T > 0$

Output : $X_0, X_1, \ldots$

(Step 1) Draw $\xi_k \sim \mathcal{N}(0, I_d)$. ← velocity randomization

(Step 2) $X_{k+1} \leftarrow q_T(X_k, \xi_k)$.

Transition Kernel $\Pi_{ex}(x, \cdot) = \mathbb{E}\left[\delta_{q_T(x,\xi)}\right]$ $\quad \xi \sim \mathcal{N}(0, I_d)$

Theorem (Chen and Vempala 2019)

$\leftarrow$ builds on Mangoubi/Smith 2017

Suppose $U : \mathbb{R}^d \to \mathbb{R}$ satisfies

L-gradient Lipschitz

(A1) $\quad |\nabla U(x) - \nabla U(y)| \leq L\,|x-y|$,

K-strong convexity

(A2) $\quad (x-y)\cdot(\nabla U(x) - \nabla U(y)) \geq K\,|x-y|^2$.

Suppose $T>0$ satisfies $LT^2 \leq 1/8$. $\leftarrow$ NO-U-TURN condition

For all distributions $\nu, \eta \in \mathcal{P}(\mathbb{R}^{2d})$,

$$\boxed{\mathcal{W}^p(\nu \Pi_{ex}, \eta \Pi_{ex}) \leq e^{-c}\, \mathcal{W}^p(\nu, \eta)}$$

L^p - Wasserstein Contractivity

where $c = KT^2/6$.

$\boxed{\text{How many xHMC steps till } \mathrm{Law}(X_n) \approx \mu \text{ ?}}$

Given accuracy $\varepsilon > 0$. Choose $T \propto L^{-1/2}$.

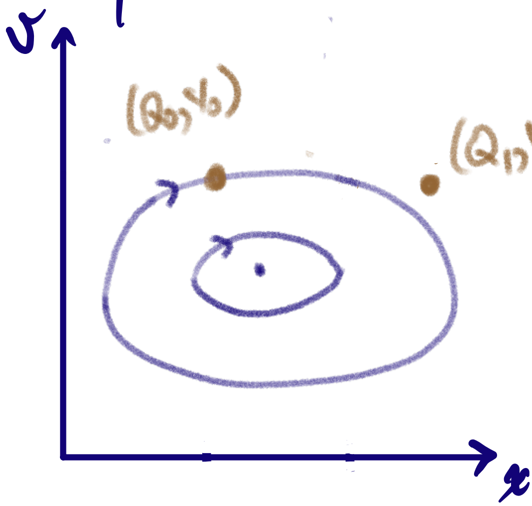
As long as $m \geq \frac{L}{K} \log\left(\frac{\mathcal{W}^p(\mu, \nu)}{\varepsilon}\right)^{+}$,

$$\mathcal{W}^p(\nu \Pi_{ex}^m, \mu) \leq \varepsilon .$$

L/K is the condition # of U.

3. Unadjusted HMC with time integrator randomization

In practice, exact Hamiltonian flow is unavailable.



$$Q_1 = Q_0 + h V_0 + \frac{h^2}{2} F_0$$

$$V_1 = V_0 + h F_0$$

$$F_0 = -\nabla U\left(Q_0 + \frac{h}{2} V_0\right)$$

Idea: evaluate force at a random midpoint

$$F_0 = -\nabla U(Q_0 + \mathcal{U} V_0) \quad \text{where } \mathcal{U} \sim \text{Uniform}(0, h)$$

Cf: Bou-Rabee, Schuh 2020 & Bou-Rabee, Eberle 2021

Unadjusted HMC replaces exact Hamilton flow w/ flow of randomized midpoint method

$$\dot{Q}_t = V_t \qquad \dot{V}_t = -\nabla U\left(Q_{\lfloor t \rfloor} + \gamma_t V_{\lfloor t \rfloor}\right)$$

where $\gamma_t := \mathcal{U}_{\lfloor t \rfloor / h}$ and $\{\mathcal{U}_k\} \overset{i.i.d}{\sim} \text{Uniform}(0,h)$

Transition Kernel $\Pi_u(x,\cdot) = E\left[\delta_{Q_T(x,\xi)}\right] \quad \xi \sim \mathcal{N}(0, I_d)$

Cf. Shen, Lee 2019 "RMM for Logconcave Sampling"
Cao, Lu, Wang 2021 "Complexity of Randomized Algorithms.."

Theorem (Bou-Rabee and Marsden 2022)

Suppose $U: \mathbb{R}^d \to \mathbb{R}$ satisfies
(A0) $U(0)=0$ and $\nabla U(0)=0$
(A1) L-gradient Lipschitz
(A2) K-strongly convex

Suppose $T>0$ and $h>0$ satisfy $LT^2 \le 1/8$ and $T/h \in \mathbb{Z}$.

For all distributions $\nu, \eta \in \mathcal{P}(\mathbb{R}^d)$,

$$\boxed{\mathcal{W}^2(\nu \Pi_u, \eta \Pi_u) \le e^{-c}\, \mathcal{W}^2(\nu, \eta)}$$

L^2-Wasserstein Contractivity

where $c = KT^2/6$.

N.b. $\exists!\ \tilde{\mu} = \tilde{\mu}\Pi_u$ but $\tilde{\mu} \approx \mu$.

Idea of Proof

The proof is based on synchronously coupling two copies of Π_u and applying

$$\boxed{|Q_T(x,v) - Q_T(y,v)|^2 \leq \left(1 - \frac{KT^2}{3}\right)|x-y|^2}$$

which holds for all $x, y, v \in \mathbb{R}^d$ almost surely.

Cf.: Bou-Rabee, Schuh 2020 "Convergence of uHMC .."

Theorem (Bou-Rabee and Marsden 2022)

Suppose $U : \mathbb{R}^d \to \mathbb{R}$ satisfies
(A0) $U(0) = 0$ and $\nabla U(0) = 0$
(A1) L-gradient Lipschitz
(A2) K-strongly convex

Suppose $T > 0$ and $h > 0$ satisfy $LT^2 \le 1/8$ and $T/h \in \mathbb{Z}$.

It holds that

$$\boxed{\mathcal{W}^2(\tilde{\mu}, \mu) \le 142\, d^{1/2} c^{-1} \left(\frac{L}{K}\right)^{1/2} L^{1/4} h^{3/2}}$$

L^2-Wasserstein Asymptotic Bias

Cf. Durmus, Eberle 2021 "Asymptotic Bias of Inexact MCMC..."

Idea of Proof

The proof is based on L^2-accuracy of the underlying randomized midpoint flow

$$\left(E\left[\,|Q_{t_k}(x,v) - q_{t_k}(x,v)|^2\right]\right)^{1/2} \leq 71 \cdot \left(|v| + \sqrt{L}\,|x|\right) L^{1/4}\, h^{3/2}$$

which holds for all $x, v \in \mathbb{R}^d$.

The $3/2$-order L^2-accuracy is a consequence of cancellations due to independence of the random midpoints.

See Fundamental Thm. for L^2-Convergence Milstein and Tretyakov 2021.

$$\boxed{\text{How many uHMC steps till } \mathrm{Law}(X_n) \approx \mu \text{ ?}}$$

$$\mathcal{W}^2(\mu, \nu \pi_u^m) \le \underbrace{\mathcal{W}^2(\mu, \tilde{\mu})}_{\text{asymp. bias}} + \underbrace{\mathcal{W}^2(\tilde{\mu}, \nu \pi_u^m)}_{\text{contractivity}}$$

Given accuracy $\varepsilon > 0$. Choose $T \propto L^{-1/2}$.

Choose $h \propto \varepsilon^{2/3} d^{-1/3} c^{2/3} (L/K)^{-1/3} L^{-1/6}$.

Complexity : $\underbrace{\frac{T}{h}}_{\substack{\text{\# of gradient evaluations} \\ \text{per uHMC step}}} \cdot m \propto \varepsilon^{-2/3} \left(\frac{d}{K}\right)^{1/3} \left(\frac{L}{K}\right)^{5/3}$

vs. $\varepsilon^{-1} \left(\frac{d}{K}\right)^{1/2} \left(\frac{L}{K}\right)^{2}$

4. Adjustable Randomized Time Integrators

A Metropolis adjustment step can be used to obtain a Markov chain w/out asymptotic bias.

$$\Pi_{ex} = \Pi \, \Gamma$$

$$\Gamma((q,v),\cdot) = \delta_q \otimes N(0, I_d)$$

$$\Pi((q,v),\cdot) = \delta_{Q_T(q,v)}$$

Since Γ leaves the v-marginal of μ_{BG} invariant, we focus on Π.

A key property of the exact Hamiltonian flow is

reversibility w.r.t. $S : (q, v) \mapsto (q, -v)$

$$\varphi_{-t} = S \circ \varphi_t \circ S .$$

Consequently, the corresponding transition kernel Π satisfies

generalized detailed balance w.r.t. μ_{BG}

$$\mu_{BG}(dq\, dv)\, \Pi((q,v), dq'\, dv') = \mu_{BG}(dq'\, dv')\, \Pi(S(q',v'), S(dq\, dv))$$

Defn. A randomized time integrator is adjustable if the corresponding transition kernel $\tilde{\Pi}$ satisfies

$$\mu_{BG}(dq\,dv)\,\tilde{\Pi}(S(q,v),\,S(dq'\,dv')) = \varrho((q,v),(q',v'))\,\mu_{BG}(dq'\,dv')\,\tilde{\Pi}((q',v'),\,dq\,dv)$$

Adjusted Kernel

$$\tilde{\Pi}_a((q,v),\,dq'\,dv') = \alpha((q,v),(q',v'))\,\tilde{\Pi}((q,v),\,dq'\,dv') + r(q,v)\,\delta_{S(q,v)}(dq'\,dv')$$

where $\alpha((q,v),(q',v')) := \min(1,\varrho((q',v'),(q,v)))$.

Adjustable Randomized Time Integrators

Let $\{\theta_i\}_{i\in I}$ be an indexed family of time integrators

such that for each $i \in I$

S-reversible

(1) $\lambda^{2d} \circ \theta_i = \lambda^{2d}$ (2) $\theta_i = S \circ \theta_i^{-1} \circ S$

vol. preserving

Let ϱ be a probability distribution over I.

Let $\theta_{i_N \cdots i_1} := \theta_{i_N} \circ \cdots \circ \theta_{i_2} \circ \theta_{i_1}$.

Then $$\tilde{\Pi}((q,v),\cdot) = \int_{I^N} \delta_{\theta_{h_N \cdots h_1}(q,v)} \prod_{i=1}^{N} \varrho(dh_i)$$

is adjustable.

Example: Randomized 2-Stage Palindromic Integrator

Let $I=[0,1]$, $h>0$, and $b\in I$.

$$\Theta_b := \varphi^{(A)}_{bh}\circ\varphi^{(B)}_{h/2}\circ\varphi^{(A)}_{(1-2b)h}\circ\varphi^{(B)}_{h/2}\circ\varphi^{(A)}_{bh}$$

where $\varphi_t^{(A)}(q,v) := (q+tv,\, v)$ and $\varphi_t^{(B)}(q,v) := (q,\, v+tF(q))$.

More explicitly,

$b=1/2$ corresponds to p. Verlet

$$(q,v)\mapsto\left(q+hv+(1-b)\frac{h^2}{2}F_+ + \frac{bh^2}{2}F_-\,,\; v+\frac{h}{2}\left[F_+ + F_-\right]\right)$$

where $F_+ := F(x+bhv)$ and $F_- := F\left(x+(1-b)hv+(1-2b)\frac{h^2}{2}F_+\right)$

Adjusted HMC Algorithm

Input: # of integration steps N, time step size $h>0$, and $\varrho \in \mathcal{P}(I)$.

Output: (X_0, V_0), (X_1, V_1), ...

(Step 1) Independently draw $\xi_k \sim \mathcal{N}(0, I_d)$ and $\{\mathcal{U}_i^k\}_{i=1}^N \overset{i.i.d.}{\sim} \varrho$.

(Step 2) $(\tilde{X}_{k+1}, \tilde{V}_{k+1}) \leftarrow \theta_{\mathcal{U}_N^k} \cdots \mathcal{U}_1^k (X_k, \xi_k)$

(Step 3) Draw $\beta^k \sim \mathrm{Bern}\big(\alpha((X_k, \xi_k), (\tilde{X}_{k+1}, \tilde{V}_{k+1}))\big)$.

(Step 4) $(X_{k+1}, V_{k+1}) \leftarrow \beta^k \cdot \underset{\text{Accept}}{(\tilde{X}_{k+1}, \tilde{V}_{k+1})} + (1-\beta^k) \cdot \underset{\text{Reject}}{S(\tilde{X}_k, \xi_k)}$.

- Summary -

- Hamiltonian MCMC provides a flexible framework for constructing implementable exact / in-exact Markov chains.
- Time integrator randomization relaxes regularity requirements on the target measure.
- Stochastic notions of integrator accuracy fit Hamiltonian MCMC and are more lenient.

– Questions –

- Can one similarly relax the Hessian Lipschitz condition for TV-convergence bounds?

  Cf. Bou-Rabee, Eberle 2021 "Mixing Time Guarantees for uHMC"

- Or, the regularity conditions for the discrete version of Monmarché's Modified Entropy Method?

  Cf. Camrud, Durmus, Monmarché, Stoltz 2023

- Can these results be extended to uHMC with partial velocity refreshment?

Cf. Monmarché 2022

Leimkuhler, Paulin, Whalley 2023

Appendix of Bou-Rabee, Marsden 2022

$$\text{Complexity} \in \mathcal{O}\left(\max\left(\left(\frac{d}{K}\right)^{1/4} \cdot \left(\frac{L}{K}\right)^{3/2} \cdot \varepsilon^{-1/2}, \left(\frac{d}{K}\right)^{1/3} \cdot \left(\frac{L}{K}\right)^{4/3} \cdot \varepsilon^{-2/3}\right)\right)$$