Non-asymptotic analysis of HMC and Langevin diffusion for MCMC

Pierre Monmarché

Université d'Angers

Développements récents autour des méthodes numériques probabilistes pour le machine learning

Kinetic Markov chain Monte Carlo

Markov process $(X_t)_{t\geqslant 0}$ on $\mathbb{R}^d$, ergodic w.r.t. $\mu \propto e^{-U(x)}\mathrm{d}x$:

$$a.s. \qquad \frac{1}{t}\int_0^t f(X_s)\mathrm{d}s \underset{t\to+\infty}{\longrightarrow} \int_{\mathbb{R}^d} f(x)\mu(\mathrm{d}x)\,.$$

Kinetic Markov chain Monte Carlo

Markov process $(X_t)_{t\geqslant 0}$ on $\mathbb{R}^d$, ergodic w.r.t. $\mu \propto e^{-U(x)}\mathrm{d}x$:

$$a.s. \qquad \frac{1}{t}\int_0^t f(X_s)\mathrm{d}s \underset{t\to+\infty}{\longrightarrow} \int_{\mathbb{R}^d} f(x)\mu(\mathrm{d}x)\,.$$

Kinetic MCMC : $(X_t, V_t)_{t\geqslant 0}$ process on $\mathbb{R}^{2d}$, $\mathrm{d}X_t = V_t\mathrm{d}t$,

$$\frac{1}{t}\int_0^t f(X_s, V_s)\mathrm{d}s \underset{t\to+\infty}{\longrightarrow} \int_{\mathbb{R}^d} f(x,v)\pi(\mathrm{d}x\mathrm{d}v)$$

où $\pi(x,v) \propto e^{-U(x)}e^{-|v|^2/2} = e^{-H(x,v)}$,

$$H(x,v) = U(x) + |v|^2/2\,.$$

Langevin and HMC

- Langevin diffusion :

$$\begin{cases} \mathrm{d}X_t &= V_t \mathrm{d}t \\ \mathrm{d}V_t &= -\nabla U(X_t)\mathrm{d}t - \gamma V_t \mathrm{d}t + \sqrt{2\gamma}\mathrm{d}B_t, \qquad \gamma > 0 \end{cases}$$

- Hamiltonian Monte Carlo (HMC) :

$$\mathrm{d}X_t = V_t \mathrm{d}t \qquad \mathrm{d}V_t = -\nabla U(X_t)\mathrm{d}t \qquad t \in [0, T),$$

then $X_T = X_{T-}$ and

$$V_T = \eta V_{T-} + \sqrt{1-\eta^2}G, \qquad \eta \in [0,1), \; G \sim \mathcal{N}(0, I_d)$$

Splitting scheme

Based on $e^{t(L_1+L_2)} = e^{t/2L_2}e^{tL_1}e^{t/2L_2} + o(t^2)$.

- For the Hamiltonian dynamics,

$$L_1 = v \cdot \nabla_x \qquad L_2 = -\nabla U(x) \cdot \nabla_v \,,$$

which gives, with a stepsize $\delta > 0$, the Verlet integrator

$$\begin{cases} v & \leftarrow & v - \delta/2\nabla U(x) \\ x & \leftarrow & x + \delta v \\ v & \leftarrow & v - \delta/2\nabla U(x) \end{cases}$$

Splitting scheme

Based on $e^{t(L_1+L_2)} = e^{t/2L_2} e^{tL_1} e^{t/2L_2} + o(t^2)$.

- For the Hamiltonian dynamics,

$$L_1 = v \cdot \nabla_x \qquad L_2 = -\nabla U(x) \cdot \nabla_v \,,$$

which gives, with a stepsize $\delta > 0$, the Verlet integrator

$$\left\{ \begin{array}{lcl} v & \leftarrow & v - \delta/2 \nabla U(x) \\ x & \leftarrow & x + \delta v \\ v & \leftarrow & v - \delta/2 \nabla U(x) \end{array} \right.$$

- For the Langevin diffusion, same L_1, L_2 and add

$$L_3 = -\gamma v \cdot \nabla_v + \gamma \Delta_v \,,$$

whose transition is given by

$$V_t \;=\; e^{-\gamma t} V_0 + \sqrt{1 - e^{-2\gamma t}} G \,, \qquad G \sim \mathcal{N}(0, I_d) \,.$$

Splitting schemes

Given parameters

- time step $\delta > 0$
- Number of Verlet steps $K \in \mathbb{N}_*$
- Damping/friction parameter $\eta \in [0, 1)$,

consider the Markov chain $z = (x, v) \in \mathbb{R}^{2d}$ with a transition given by

$$\begin{array}{rcl} v & \leftarrow & \eta v + \sqrt{1-\eta^2}G \\ K \text{ times} \left\{ \begin{array}{l} v \\ x \\ v \end{array} \right. & \begin{array}{c} \leftarrow \\ \leftarrow \\ \leftarrow \end{array} & \begin{array}{l} v - \delta/2\nabla U(x) \\ x + \delta v \\ v - \delta/2\nabla U(x) \end{array} \\ v & \leftarrow & \eta v + \sqrt{1-\eta^2}G'. \end{array}$$

- Langevin diffusion : $K = 1$, $\eta = e^{-\gamma\delta/2} = 1 - \gamma\delta/2 + o(\delta)$.
- (position) HMC : $K = T/\delta$, $\eta = 0$.

Splitting schemes

Given parameters

- time step $\delta > 0$
- Number of Verlet steps $K \in \mathbb{N}_*$
- Damping/friction parameter $\eta \in [0, 1)$,

consider the Markov chain $z = (x, v) \in \mathbb{R}^{2d}$ with a transition given by

$$\begin{array}{rcl} v & \leftarrow & \eta v + \sqrt{1-\eta^2}\, G \\ K \text{ times} \left\{ \begin{array}{l} v \\ x \\ v \end{array} \right. & \begin{array}{l} \leftarrow \\ \leftarrow \\ \leftarrow \end{array} & \begin{array}{l} v - \delta/2 \nabla U(x) \\ x + \delta v \\ v - \delta/2 \nabla U(x) \end{array} \\ v & \leftarrow & \eta v + \sqrt{1-\eta^2}\, G'. \end{array}$$

- Langevin diffusion : $K = 1$, $\eta = e^{-\gamma\delta/2} = 1 - \gamma\delta/2 + o(\delta)$.
- (position) HMC : $K = T/\delta$, $\eta = 0$.

Remark : here, unadjusted algorithms (no Metropolis accept/reject step).

Log-concave target measures

In the following, we assume that there exist $m, L > 0$ such that

$$\forall x \in \mathbb{R}^d \,, \qquad 0 < m \leqslant \nabla^2 U(x) \leqslant L \,.$$

Restrictive but standard to get explicit bounds (in the dimension d in particular) : Cheng et al 2018, Durmus, Moulines 2019, Dwivedi et al 2019, Chen, Vempala 2019, Dalalyan, Riou-Durand 2020, Chen et al 2020, Bou-Rabee Schuh 2020, Deligianidis et al 2021, Sanz-Serna, Zygalakis 2021, Mangoubi, Smith 2021...

Parallel coupling

For M symmetric definite positive matrix, $\|z\|_M^2 = z \cdot Mz$.

Theorem (M, 2021)

For $K = 1$, $\eta = e^{-\gamma\delta/2}$, let $(Z_k, Z'_k)_{k\in\mathbb{N}}$ be the parallel coupling (i.e. same Gaussian variables at each step) of two chains. Assume that $\gamma \geqslant 2\sqrt{L}$ and $\delta \leqslant m/(33\gamma^3)$. Then

$$\|Z_{k+1} - Z'_{k+1}\|_M^2 \leqslant (1-\kappa\delta)\|Z_k - Z'_k\|_M^2$$

with

$$\kappa = \frac{m}{3\gamma}, \qquad \frac{1}{2}\left(\frac{1}{L}|x|^2 + |v|^2\right) \leqslant \|z\|_M^2 \leqslant \frac{3}{2}\left(\frac{1}{L}|x|^2 + |v|^2\right).$$

First consequence

For $p \geqslant 1$,

$$\mathcal{W}_{p,M}\left(\mu,\nu\right) = \inf_{\xi \in \mathcal{C}(\nu,\mu)} \left(\mathbb{E}_\xi\left(\|Z - Z'\|_M^p\right)\right)^{1/p} .$$

Corollary

Let Q be the transition operator of the chain. Then, under the previous conditions, for all $n \in \mathbb{N}$ and all initial distributions μ, ν,

$$\mathcal{W}_{p,M}\left(\mu Q^n, \nu Q^n\right) \leqslant (1 - \kappa\delta)^{n/2} \mathcal{W}_{p,M}\left(\mu,\nu\right) .$$

There exists a unique invariant measure π_δ.

Friction should be high enough

The condition $\gamma \geqslant 2\sqrt{L}$ is not far from optimal (to get a contraction) :

Proposition (M', 2020)

Let $(P_t)_{t\geqslant 0}$ be the semi-group of the continuous-time Langevin diffusion.

1. If $\gamma(\sqrt{m}+\sqrt{L}) > L-m$ and $U \in \mathcal{C}^2(\mathbb{R}^n)$ is such that

$$\forall x \in \mathbb{R}^n\,, \qquad m \leqslant \nabla^2 U(x) \leqslant L\,,$$

   then there exist some M and $\rho > 0$ such that, for all $p \geqslant 1$,

$$\mathcal{W}_{p,M}(\nu P_t, \mu P_t) \leqslant e^{-\rho t}\mathcal{W}_{p,M}(\nu,\mu) \tag{1}$$

2. If $\gamma(\sqrt{m}+\sqrt{L}) \leqslant L-m$ and $U \in \mathcal{C}^2(\mathbb{R}^n)$ is such that

$$\exists x, x' \in \mathbb{R}^n, \qquad \nabla^2 U(x) \;=\; mI_n\,, \qquad \nabla^2 U(x') \;=\; LI_n\,,$$

   then, (1) cannot hold for any M, $\rho > 0$, $p \geqslant 1$.

Corollaries

- Concentration inequalities for ergodic averages (confidence intervals)
- Dimension-free convergence rate in total variation by regularization
- Non-asymptotic efficiency bounds (contraction + numerical analysis) :

Proposition

Assumption	$\mathcal{W}_1(\pi, \pi_\delta)$	$\|\pi - \pi_\delta\|_{TV}$
$m \leqslant \nabla^2 U \leqslant L$	$\delta\sqrt{d}$	$\delta d(1 + \lvert\ln(\delta^3 d)\rvert)$
$+\lvert\nabla^3 U\rvert \leqslant L_2$	$\delta^2 d$	$\delta^2 d^{3/2}(1 + \lvert\ln(\delta^3 d)\rvert)$
+i.i.d.	$\delta^2\sqrt{d}$	$\delta^2 d(1 + \lvert\ln(\delta^3 d)\rvert)$

Remark : $\delta^2\sqrt{d}$ is sharp in the Gaussian case.

Extension

Proposition (M, 2022)

In the general case $K \geqslant 1$, $\eta \in [0,1)$, there exist explicit c, c' depending on m and L such that, if

$$K\delta \leqslant c \qquad \text{et} \qquad 1-\eta^2 \geqslant c' K\delta\,,$$

then there exists an explicit matrix M such that, with a parallel coupling,

$$\|Z_{k+1} - Z'_{k+1}\|_M^2 \leqslant (1-\kappa)\|Z_k - Z'_k\|_M^2\,, \qquad \kappa = \frac{(\delta K)^2 m}{54(1-\eta^2)}\,.$$

- The expressions of c, c' are not sharp but the conditions are necessary.
- This unifies previous results (also, rate for the discrete-time chain).
- In the two regimes $K\delta \to T > 0$, η constant and $K\delta \to 0$, $\eta = 1 - \gamma K\delta + o(K\delta)$, the convergence rate $\ln(\kappa)/K$ is of order δ.

The question

Goal : full optimization in $\mathfrak{p} = (\delta, K, \eta)$ in the Gaussian case $U(x) = x \cdot Sx/2$, uniformly ($\neq$ specific Gaussian target) over

$$\mathcal{M}_s(m, L) = \{\text{symmetric } s,\ \ m \leqslant S \leqslant L\}.$$

More precisely, at a given accuracy (fair comparison)

$$\varepsilon(\mathfrak{p}) := \sup_{S \in \mathcal{M}_s(m,L)} \mathcal{W}_2\left(\pi_S, \pi_{S,\mathfrak{p}}\right)$$

find $\mathfrak{p}$ which maximizes

$$\rho(\mathfrak{p}) := \inf_{S \in \mathcal{M}_s(m,L)} \lim_{n \to +\infty} -\frac{1}{Kn} \ln\left(\sup_{\nu \in \mathcal{P}_2 \setminus \{\pi_{S,\mathfrak{p}}\}} \frac{\mathcal{W}_2(\nu Q^n_{S,\mathfrak{p}}, \pi_{S,\mathfrak{p}})}{\mathcal{W}_2(\nu, \pi_{S,\mathfrak{p}})}\right).$$

First remarks

- Unique invariant measure if $\delta^2 L \leqslant 4$, which depends only on δ, and

$$\varepsilon(\mathfrak{p}) = \sqrt{d}\frac{1 - \sqrt{1 - \delta^2 L/4}}{\sqrt{L}} .$$

- At a fixed δ, due to scaling properties of $\mathcal{W}_2$, the optimal K, η should be functions of the rescaled time-step $\delta\sqrt{L}$ and the condition number L/m.

Proposition

For $\mathfrak{p} = (\delta, K, \eta)$ with $\delta^2 L < 4$,

$$\rho(\mathfrak{p}) = \frac{-\ln g\left(h(K,\delta),\eta\right)}{K}$$

where, writing $\varphi_\lambda = \arccos(1 - \delta^2\lambda/2)$ for $\lambda \in [m, L]$,

$$\begin{aligned} h(K,\delta) &= \sup\{|\cos(K\varphi_\lambda)|,\ \lambda \in [m,L]\} \\ g(c,\eta) &= \eta \vee \frac{(1+\eta^2)c}{2} + \sqrt{\left(\left(\frac{(1+\eta^2)c}{2}\right)^2 - \eta^2\right)_+}. \end{aligned}$$

Proposition

For $\mathfrak{p} = (\delta, K, \eta)$ with $\delta^2 L < 4$,

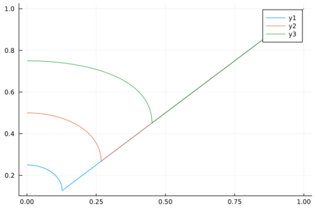
$$\rho(\mathfrak{p}) = \frac{-\ln g\left(h(K,\delta),\eta\right)}{K}$$

where, writing $\varphi_\lambda = \arccos(1-\delta^2\lambda/2)$ for $\lambda \in [m, L]$,

$$\begin{aligned} h(K,\delta) &= \sup\{|\cos(K\varphi_\lambda)|,\ \lambda \in [m,L]\} \\ g(c,\eta) &= \eta \vee \frac{(1+\eta^2)c}{2} + \sqrt{\left(\left(\frac{(1+\eta^2)c}{2}\right)^2 - \eta^2\right)_+}. \end{aligned}$$

For $c \in [0,1]$, $\eta \mapsto g(c,\delta)$ minimal at

$$\eta_*(c) = \eta_*(c) = (1-\sqrt{1-c^2})/c$$



Scaling limits

Proposition

Let $(\mathfrak{p}_n)_{n\in\mathbb{N}} = (\delta_n, K_n, \eta_n)_{n\in\mathbb{N}}$ with $\delta_n \to 0$ as $n \to +\infty$. Up to extracting a subsequence, we are necessarily in one of the three following cases :

1. There exists $T > 0$ and $\eta \in [0,1)$ such that $K_n\delta_n \to T$ and $\eta_n \to \eta$;

$$\rho(\mathfrak{p}_n) \underset{n\to+\infty}{\simeq} \frac{\delta_n |\ln g(h_*(T),\eta)|}{T} := \delta_n\sqrt{L}\bar{\rho}_{HMC}(T,\eta)$$

with $h_*(T) = \sup\{|\cos(x)|,\ x \in [T\sqrt{m}, T\sqrt{L}]\}$.

2. $K_n\delta_n \to 0$ and $\eta_n = 1 - \gamma K_n\delta_n + o(K_n\delta_n)$ for some $\gamma > 0$;

$$\rho(\mathfrak{p}_n) \underset{n\to+\infty}{\simeq} \left(\gamma - \sqrt{(\gamma^2 - m)_+}\right)\delta_n := \delta_n\sqrt{L}\bar{\rho}_{Lang}(\gamma) .$$

3. $\rho(\mathfrak{p}_n) = o(\delta_n)$ as $n \to +\infty$.

Optimization

- Langevin scaling : optimize

$$\gamma \mapsto \bar{\rho}_{Lang}(\gamma) \qquad \text{or} \qquad \eta \mapsto \rho(\delta, 1, \eta).$$

- HMC (or general) scaling : optimize

$$(T, \eta) \mapsto \bar{\rho}_{HMC}(T, \eta) \qquad \text{or} \qquad (K, \eta) \mapsto \rho(\delta, K, \eta).$$

- position HMC : optimize

$$T \mapsto \bar{\rho}_{HMC}(T, 0) \qquad \text{or} \qquad K \mapsto \rho(\delta, K, 0).$$

Conclusion

With $\kappa = L/m$, $\varepsilon' = \varepsilon\sqrt{L/d}$, $\delta' = \delta\sqrt{L}$,

1. With $K = 1$ (Langevin), the optimal rate is

$$\rho \underset{\varepsilon'\to 0}{\simeq} \frac{\delta'}{\sqrt{\kappa}}\,.$$

2. Optimal rate with $K = T_*/\delta$ (HMC), $T_* = \pi/(\sqrt{L}+\sqrt{m})$,

$$\rho \underset{\varepsilon'\to 0}{\simeq} \frac{\delta\sqrt{L}\,(1+1/\sqrt{\kappa})}{\pi} \ln\left(\frac{\cos\left(\pi/(1+\sqrt{\kappa})\right)}{1-\sin\left(\pi/(1+\sqrt{\kappa})\right)}\right) \underset{\kappa\to+\infty}{\simeq} \frac{\delta'}{\sqrt{\kappa}}\,.$$

3. If $K\delta\sqrt{L} \geqslant \pi$, $\rho = 0$ (very sensitive ! If $\sqrt{L_{true}} \geqslant \sqrt{L}+\sqrt{m}\ldots$)
4. Optimal position HMC ($\eta = 0$) for $K = T_*/\delta$,

$$\rho \underset{\varepsilon'\to 0}{\simeq} -\frac{\delta\sqrt{L}\ln\left(\cos\left(\pi/(1+\sqrt{\kappa})\right)\right)}{\pi/(1+\sqrt{1/\kappa})} \underset{\kappa\to+\infty}{\simeq} \frac{\pi\delta'}{\kappa}$$

Conclusion

- Unadjusted HMC and Langevin united (uniform non-asymptotic dimension-free efficiency bounds).
- In the Gaussian case, results in favor of $\eta > 0$ (partial refreshments of the velocity), contrary to the standard practice in the Proba/Stat community. Langevin competitive for badly conditionned problems (and less sensitive).
- Non-convex case : work in progress (hypocoercive modified entropy + numerical error)

Thanks for your attention !