Bayesian Theory and Computation

Lecture 12: MCMC Theory

Cheng Zhang

School of Mathematical Sciences, Peking University

April 08, 2024

MCMC Theory

- So far, we have introduced many MCMC algorithms.
- Although these algorithms have been empirically shown to converge to the target distribution with good speed, in practice, we may want to know more precisely about the convergence behavior and assess the approximation error for a given computation budget.
- In this lecture, we discuss some theoretical results on the convergence of MCMC methods, with an emphasis on Langevin diffusion.

Ergodicity Recap

- Total Variation Distance: the total variation distance between two probability measures μ and ν on $\mathcal{X}$ is

$$d_{\mathrm{TV}}(\mu, \nu) = \sup_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)|.$$

- Ergodicity: if a Markov chain on a state space $\mathcal{X}$ is both ϕ-irreducible and aperiodic, and has a stationary distribution π, then for π-a.e. $x \in \mathcal{X}$,

$$\lim_{n \to \infty} d_{\mathrm{TV}}(\delta_x P^n, \pi) = 0.$$

In particular,

$$\lim_{n \to \infty} P^n(x, A) = \pi(A)$$

for all measurable $A \subseteq \mathcal{X}$.

Uniform Ergodicity

- For Markov chains that are irreducible and aperiodic, we have stronger convergence properties given certain conditions.
- Uniform Ergodicity: a Markov chain with invariant probability measure π and Markov transition kernel P is uniformly ergodic if

  $$d_{\mathrm{TV}}(\delta_x P^n, \pi) \leq M\rho^n, \quad \forall x \in \mathcal{X}$$

  for some constant M and $\rho < 1$.
- This means the total variation distance decreases geometrically fast, with ρ governing the rate, and the bound is independent of x.

Minorization Condition

- Minorization Condition:

$$P^m(x, A) \geq \epsilon\nu(A), \quad \forall x \in \mathcal{X}, A \subseteq \mathcal{X},$$

  for some $m \in \mathbb{N}, \epsilon > 0$ and probability measure ν.

- Loosely speaking, minorimzation condition guarantees that $\delta_x P^m$ and $\delta_y P^m$ have some degree of overlap, $\forall x, y \in \mathcal{X}$.

Theorem

Suppose the above minorization condition holds, then

$$d_{\mathrm{TV}}(\delta_x P^n, \pi) \leq (1-\epsilon)^{\lfloor \frac{n}{m} \rfloor}, \quad \forall x \in \mathcal{X}.$$

The Coupling Inequality

- Coupling: we say ζ is a coupling of two probability measure μ, ν if it is a probability measure on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d))$ such that

$$\zeta(A, \mathbb{R}^d) = \mu(A),\ \zeta(\mathbb{R}^d, A) = \nu(A),\ \forall A \in \mathcal{B}(\mathbb{R}^d).$$

- The coupling inequality:

$$d_{\mathrm{TV}}(\mu, \nu) \leq p(X \neq Y),$$

for any coupling (X, Y) of μ and ν.

Proof. $$\begin{aligned} d_{\mathrm{TV}}(\mu, \nu) &= \sup_A |\mu(A) - \nu(A)| \\ &= \sup_A |p(X \in A, Y \notin A) - p(X \notin A, Y \in A)| \\ &\leq p(X \neq Y). \end{aligned}$$

A Coupling Proof

- For simplicity, we consider the case when $m = 1$. That is

$$P(x, A) \geq \epsilon \nu(A).$$

- We define a residual Markov kernel

$$R(x, A) = \frac{P(x, A) - \epsilon \nu(A)}{1 - \epsilon}, \quad x \in \mathcal{X}, A \subseteq \mathcal{X},$$

and observe that $\delta_x P = \epsilon \nu + (1 - \epsilon) \delta_x R$.

- We will show an explicit coupling (Doeblin 1938) such that

$$p(X_n \neq Y_n) \leq (1 - \epsilon)^n$$

where $X_n \sim \delta_x P^n$ and $Y_n \sim \pi$.

A Coupling Proof

- Let $X_0 = x$ and $Y_0 \sim \pi$.
- Now follow the procedure for each time $n \geq 1$:
  1. If $X_{n-1} = Y_{n-1}$, sample $Z_n \sim P(X_{n-1}, \cdot)$, set $X_n = Y_n = Z_n$.
  2. Otherwise, with probability ϵ, sample $Z_n \sim \nu$ and set $X_n = Y_n = Z_n$.
  3. Otherwise, sample $X_n \sim R(X_{n-1}, \cdot)$ and $Y_n \sim R(Y_{n-1}, \cdot)$ independently.
- Note that we have not changed the marginal distributions of X_n or Y_n, so $X_n \sim \delta_x P^n$ and $Y_n \sim \pi$.
- We also observe that

$$p(X_n \neq Y_n) \leq (1 - \epsilon)^n.$$

Geometric Ergodicity

- The minorization condition allows us to successfully couple Markov chains with probability ϵ at each time, which is too strong in practice.
- A weaker condition is geometric ergodicity.
- Geometric Ergodicity: a Markov chain with stationary distribution π is geometrically ergodic if

$$d_{\mathrm{TV}}(\delta_x P^n, \pi) \leq M(x)\rho^n, \quad x \in \mathcal{X},$$

  for some $\rho < 1$, where $M(x) < \infty$ for π-a.e. $x \in \mathcal{X}$.
- Instead of a constant, the bound M now depends on x.

Small Sets and Drift Condition

- Small sets: a set $C \subseteq \mathcal{X}$ is small if

$$P^m(x, A) \geq \epsilon\nu(A), \quad x \in C,\ A \subseteq \mathcal{X},$$

  for some $m \in \mathbb{N}, \epsilon > 0$ and probability measure ν.

- Drift condition: there is a function $V : \mathcal{X} \mapsto [1, \infty]$ with $V(x) < \infty$ for at least one $x \in \mathcal{X}$, such that

$$\int_{\mathcal{X}} V(y)P(x, dy) \leq \lambda V(x) + b\mathbf{1}_C(x),$$

  where C is a small set, $\lambda \in (0, 1)$ and $b < \infty$.

- The drift condition guarantees that

$$\sup_{x\in C} \mathbb{E}_x \kappa^{\tau_C} < \infty, \qquad \text{for some } \kappa > 1,$$

  where $\tau_A = \inf\{n \geq 1 : X_n \in A\}$.

Geometric Ergodicity

Theorem (Meyn and Tweedie, 1993)

A Markov chain is geometrically ergodic if and only if it admits a small set and satisfies the drift condition.

Geometric ergodicity for various sampling algorithms.

	Generalized Gaussian distribution, $\pi(x) \propto \exp(-\|x\|^{\beta})$				
Sampling	$\beta \in (0,1)$	$\beta = 1$	$\beta \in (1,2)$	$\beta = 2$	$\beta > 2$
methods	Thick tails	Exponential		Gaussian	Light Tails
MALA (1D)	No	Yes	Yes	Yes	No
RWM	No			Yes	Yes
HMC	No	Yes	Yes	Yes	No

Non-asymptotic Analysis: Problem Setup

- Denote by π a target density w.r.t the Lebesgue measure on $\mathbb{R}^d$, known up to a normalizing constant

$$\pi(x) = \frac{\exp(-U(x))}{\int_{\mathbb{R}^d} \exp(-U(y))dy}$$

Here, $d \gg 1$.

- Assumption 1: U is L-smooth: twice continously differentiable, $\forall x, y \in \mathbb{R}^d$,

$$\|\nabla U(x) - \nabla U(y)\| \leq L\|x - y\|.$$

- Assumption 2: U is m-strongly convex, $\forall x, y \in \mathbb{R}^d$,

$$U(y) \geq U(x) + \nabla U(x)^T(y - x) + \frac{m}{2}\|y - x\|^2.$$

Overdamped Langevin Diffusion

- Langevin SDE:

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

  where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian Motion.

- Notation: $(P_t)_{t\geq 0}$ is the Markov semigroup associated to the Langevin diffusion:

$$P_t(x, A) = \mathbb{P}(X_t \in A | X_0 = x), \quad x \in \mathbb{R}^d,\ A \in \mathcal{B}(\mathbb{R}^d).$$

- $\pi(x) \propto \exp(-U(x))$ is the unique invariant probability measure,

$$\pi = \pi P_t, \quad \forall t \geq 0.$$

Discretized Langevin Diffusion

- Idea: Sample the diffusion path, using the Euler-Maruyama (EM) scheme:

  $$X_{k+1} = X_k - \gamma_{k+1}\nabla U(X_k) + \sqrt{2\gamma_{k+1}}\eta_{k+1}$$

  where

  - $(\eta_k)_{k\geq 1}$ is i.i.d $\mathcal{N}(0, I_d)$.
  - $(\gamma_k)_{k\geq 1}$ is a sequence of stepsizes, which can either be held constant or be chosen to decrease to 0 at a certain rate.
- Closely related to the (stochastic) gradient descent algorithm.
- Note that this is just MALA without MH correction. Hence, this is referred to as unadjusted Langevin algorithm.

Constant Stepsize

- When the stepsize is held constant, i.e. $\gamma_k = \gamma$, then $(X_k)_{k\geq 1}$ is an homogeneous Markov chain with Markov kernel R_γ.
- Under some appropriate conditions, this Markov chain is irreducible, positive recurrent. Hence, it has an unique invariant distribution π_γ which does not coincide with the target distribution π.
- Questions:
  - For a given precision $\epsilon > 0$, how could we choose the stepsize $\gamma > 0$ and the number of iterations n so that

    $$D(\delta_x R_\gamma^n, \pi) \leq \epsilon$$

    where D is some distance measure between distributions.
  - Is there a way to choose the starting point x cleverly?
  - How to quantify the distance between π_γ and π?

Decreasing Stepsize

- When $(\gamma_k)_{k\geq 1}$ is nonincreasing and non constant, $(X_k)_{k\geq 1}$ is an inhomogeneous Markov chain associated with the kernels $(R_{\gamma_k})_{k\geq 1}$.
- Notation: $Q_\gamma^{n,p}, n \leq p$ is the composition of Markov kernels

$$Q_\gamma^{n,p} = R_{\gamma_n} R_{\gamma_{n+1}} \cdots R_{\gamma_p}, \quad Q_\gamma^p = Q_\gamma^{1,p}$$

  with this notation, $\mathbb{E}_x[f(X_p)] = \delta_x Q_\gamma^p f$.
- Questions:
  - Convergence: is there a way to choose the step sizes so that $D(\delta_x Q_\gamma^p, \pi) \to 0$ and if yes, what is the optimal way of choosing the stepsizes?
  - Optimal choice of simulation parameters: what is the number of iterations required to reach a neighborhood of the target: $D(\delta_x Q_\gamma^p, \pi) \leq \epsilon$ starting form a given point x?
  - Should we use fixed or decreasing step sizes?

Wasserstein Distance

- Coupling: we say ζ is a coupling of two probability measure μ, ν if it is a probability measure on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d))$ such that

$$\zeta(A, \mathbb{R}^d) = \mu(A),\ \zeta(\mathbb{R}^d, A) = \nu(A),\ \forall A \in \mathcal{B}(\mathbb{R}^d).$$

- Wasserstein Distance: for two probability measure μ, ν on $\mathbb{R}^d$, define Wasserstein distance of order $p \geq 1$ as

$$W_p(\mu, \nu) = \inf_{(X,Y) \in \prod(\mu,\nu)} \mathbb{E}^{\frac{1}{p}} \left[\|X - Y\|^p \right],$$

  where $\prod(\mu, \nu)$ is the set of couplings of μ, ν.

- Let $\mathcal{P}_p(\mathbb{R}^d) = \{\mu : \int_{\mathbb{R}^d} \|x\|^p d\mu(x) < \infty\}$. $\mathcal{P}_p(\mathbb{R}^d)$ equipped with W_p is a complete separable metric space. In what follows, we use the case $p = 2$.

Wasserstein Distance Convergence

Theorem (Durums and Moulines, 2016)

Assume that U is L-smooth and m-strongly convex. Then $\forall x, y \in \mathbb{R}^d$ and $t \geq 0$,

$$W_2(\delta_x P_t, \delta_y P_t) \leq \exp(-mt)\|x - y\|$$

- The contraction depends only on the strong convexity constant.
- Key idea: synchronous coupling!

Synchronous Coupling

$$\begin{cases} dX_t &= -\nabla U(X_t)dt + \sqrt{2}dB_t \\ dY_t &= -\nabla U(Y_t)dt + \sqrt{2}dB_t \end{cases}, \quad \text{where } (X_0, Y_0) = (x, y).$$

- This SDE has a unique strong solution $(X_t, Y_t)_{t \geq 0}$. As $(B_t)_{t \geq 0}$ is shared, we have

$$dX_t - dY_t = -\left(\nabla U(X_t) - \nabla U(Y_t)\right) dt$$

- The product rule for semimartingales imply

$$\begin{aligned} d\|X_t - Y_t\|^2 &= 2\langle dX_t - dY_t, X_t - Y_t\rangle \\ &= -2\langle \nabla U(X_t) - \nabla U(Y_t), X_t - Y_t\rangle dt \end{aligned}$$

Synchronous Coupling

- Since U is m-strongly convex, $\forall x, y \in \mathbb{R}^d$

$$\langle \nabla U(x) - \nabla U(y), x - y \rangle \geq m\|x - y\|^2.$$

- This implies

$$d\|X_t - Y_t\|^2 \leq -2m\|X_t - Y_t\|^2 dt.$$

- By Grönwall inequality:

$$\|X_t - Y_t\|^2 \leq \exp(-2mt)\|X_0 - Y_0\|^2 = \exp(-2mt)\|x - y\|^2$$

- Therefore,

$$W_2(\delta_x P_t, \delta_y P_t) \leq \mathbb{E}^{\frac{1}{2}}\|X_t - Y_t\|^2 \leq \exp(-mt)\|x - y\|.$$

Wasserstein Distance Convergence

- Assume that U is L-smooth and m-strongly convex. Then, $\forall x \in \mathbb{R}^d$ and $t \geq 0$

  $$\mathbb{E}_x\|X_t - x^*\|^2 \leq \|x - x^*\|^2 \exp(-2mt) + \frac{d}{m}(1 - \exp(-2mt)),$$

  where

  $$x^* = \arg\min_{x \in \mathbb{R}^d} U(x).$$

- The stationary distribution π satisfies

  $$\int_{\mathbb{R}^d} \|x - x^*\|^2 \pi(dx) \leq \frac{d}{m}.$$

- $\forall x \in \mathbb{R}^d$ and $t > 0$,

  $$W_2(\delta_x P_t, \pi) \leq \exp(-mt)\left(\|x - x^*\| + \sqrt{\frac{d}{m}}\right).$$

Elements of Proof

- The generator $\mathcal{A}$ associated $(P_t)_{t\geq 0}$ is defined as

$$\begin{aligned}\mathcal{A}f(x) &= \lim_{t\downarrow 0}\frac{\mathbb{E}_x f(X_t) - f(x)}{t}\\ &= -\langle \nabla U(x), \nabla f(x)\rangle + \Delta f(x),\end{aligned}$$

  $\forall f \in C^2(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$.

- Set $V(x) = \|x - x^*\|^2$. Since $\nabla U(x^*) = 0$ and using the strong convexity,

$$\begin{aligned}\mathcal{A}V(x) &= 2\left(-\langle \nabla U(x) - \nabla U(x^*), x - x^*\rangle + d\right)\\ &\leq 2(-mV(x) + d).\end{aligned}$$

Elements of Proof

- Denote for all $t \geq 0$ and $x \in \mathbb{R}^d$ by

$$v(t,x) = P_t V(x) = \mathbb{E}_x \|X_t - x^*\|^2.$$

- We have

$$\frac{\partial v(t,x)}{\partial t} = P_t \mathcal{A} V(x) \leq -2m P_t V(x) + 2d = -2m v(t,x) + 2d.$$

- Grönwall inequality

$$\begin{aligned}\mathbb{E}_x \|X_t - x^*\|^2 &= v(t,x) \\ &\leq \|x - x^*\|^2 \exp(-2mt) + \frac{d}{m}(1 - \exp(-2mt)).\end{aligned}$$

Elements of Proof

- Using triangle inequality

$$\begin{aligned} W_2(\delta_x P_t, \pi) &\le W_2(\delta_x P_t, \delta_{x^*} P_t) + W_2(\delta_{x^*} P_t, \pi) \\ &\le W_2(\delta_x P_t, \delta_{x^*} P_t) + W_2(\delta_{x^*} P_t, \pi P_t) \\ &\le \exp(-mt)\|x - x^*\| + W_2(\delta_{x^*} P_t, \pi P_t). \end{aligned}$$

- Using a similar synchronous coupling strategy as before, with $X_0 = x^*, Y_0 \sim \pi$,

$$W_2(\delta_{x^*} P_t, \pi P_t) \le \exp(-mt)\mathbb{E}^{\frac{1}{2}}\|X_0 - Y_0\|^2 \le \exp(-mt)\sqrt{\frac{d}{m}}.$$

- This concludes the proof

$$W_2(\delta_x P_t, \pi) \le \exp(-mt)\left(\|x - x^*\| + \sqrt{\frac{d}{m}}\right).$$

Contraction Property of The Discretization

- Assume that U is L-smooth and m-strongly convex. Let $(\gamma_k)_{k\geq 1}$ be a nonincreasing sequence with $\gamma_1 \leq \frac{2}{m+L}$. Then, $\forall x, y \in \mathbb{R}^d$ and $\ell \geq n \geq 1$,

$$W_2(\delta_x Q_\gamma^{n,\ell}, \delta_y Q_\gamma^{n,\ell}) \leq \sqrt{\prod_{k=n}^{\ell}(1-\kappa\gamma_k)\|x-y\|^2}$$

where

$$\kappa = \frac{2mL}{m+L}.$$

- For any $\gamma \in (0, \frac{2}{m+L})$, $\forall x \in \mathbb{R}^d$ and $n \geq 1$,

$$W_2(\delta_x R_\gamma^n, \pi_\gamma) \leq (1-\kappa\gamma)^{n/2}\left(\|x-x^*\| + \sqrt{2d\kappa^{-1}}\right).$$

Elements of Proof

- Synchronous Coupling:

$$
\begin{aligned}
X_{k+1} &= X_k - \gamma_{k+1}\nabla U(X_k) + \sqrt{2\gamma_{k+1}}\eta_{k+1} \\
Y_{k+1} &= Y_k - \gamma_{k+1}\nabla U(Y_k) + \sqrt{2\gamma_{k+1}}\eta_{k+1}
\end{aligned}
$$

  where $(\eta_k)_{k\geq n}$ is i.i.d $\mathcal{N}(0, I_d)$ and $X_{n-1} = x, Y_{n-1} = y$.

- Cancel out η_{k+1} gives

$$
X_{k+1} - Y_{k+1} = X_k - Y_k - \gamma_{k+1}\left(\nabla U(X_k) - \nabla U(Y_k)\right)
$$

  which implies

$$
\begin{aligned}
\|X_{k+1} - Y_{k+1}\|^2 = {} & \|X_k - Y_k\|^2 + \gamma_{k+1}^2\|\nabla U(X_k) - \nabla U(Y_k)\|^2 \\
& - 2\gamma_{k+1}\langle \nabla U(X_k) - \nabla U(Y_k), X_k - Y_k\rangle
\end{aligned}
$$

Elements of Proof

- A stronger inequality for U (Nesterov 2004)

$$\langle \nabla U(x) - \nabla U(y), x - y \rangle \geq \frac{\kappa}{2}\|x-y\|^2 + \frac{1}{m+L}\|\nabla U(x) - \nabla U(y)\|^2.$$

- Using this inequality, we have

$$\|X_{k+1} - Y_{k+1}\|^2 \leq (1-\kappa\gamma_{k+1})\|X_k - Y_k\|^2$$

- By induction

$$\|X_\ell - Y_\ell\|^2 \leq \prod_{k=n}^{\ell}(1-\kappa\gamma_k)\|X_{n-1} - Y_{n-1}\|^2 = \prod_{k=n}^{\ell}(1-\kappa\gamma_k)\|x-y\|^2.$$

- Therefore,

$$W_2(\delta_x Q_\gamma^{n,\ell}, \delta_y Q_\gamma^{n,\ell}) \leq \mathbb{E}^{\frac{1}{2}}\|X_\ell - Y_\ell\|^2 \leq \sqrt{\prod_{k=n}^{\ell}(1-\kappa\gamma_k)\|x-y\|^2}.$$

Wasserstein Distance Convergence: Discrete Case

- Let x^* be the unique minimize of U. Then $\forall x \in \mathbb{R}^d$ and $\ell \geq n \geq 1$, given $X_{n-1} = x$,

$$\mathbb{E}_x \|X_\ell - x^*\|^2 \leq \rho_{n,\ell}(x)$$

where

$$\rho_{n,\ell}(x) = \prod_{k=n}^{\ell} (1 - \kappa\gamma_k) \|x - x^*\|^2 + 2d\kappa^{-1} \left(1 - \prod_{k=n}^{\ell} (1 - \kappa\gamma_k) \right)$$

- For any $\gamma \in (0, \frac{2}{m+L})$, R_γ has a unique stationary distribution π_γ and

$$\int_{\mathbb{R}^d} \|x - x^*\|^2 \pi_\gamma(dx) \leq 2d\kappa^{-1}.$$

Elements of Proof

- For any $\gamma \in (0, \frac{2}{m+L})$, we have $\forall x \in \mathbb{R}^d$

$$\begin{aligned}\int_{\mathbb{R}^d} \|y - x^*\|^2 R_\gamma(x, dy) &= \|x - \gamma \nabla U(x) - x^*\|^2 + 2\gamma d \\ &= \|x - x^* - \gamma(\nabla U(x) - \nabla U(x^*))\|^2 + 2\gamma d \\ &\leq (1 - \kappa\gamma)\|x - x^*\|^2 + 2\gamma d\end{aligned}$$

- By induction

$$\begin{aligned}\mathbb{E}_x\|X_\ell - x^*\|^2 &\leq (1 - \kappa\gamma_\ell)\mathbb{E}_x\|X_{\ell-1} - x^*\|^2 + 2\gamma_\ell d \\ &\leq \rho_{n,\ell}(x).\end{aligned}$$

- Similarly

$$\begin{aligned}W_2(\delta_x R_\gamma^n, \pi_\gamma) &\leq W_2(\delta_x R_\gamma^n, \delta_{x^*} R_\gamma^n) + W_2(\delta_{x^*} R_\gamma^n, \pi_\gamma R_\gamma^n) \\ &\leq (1 - \kappa\gamma)^{n/2} \left(\|x - x^*\| + \sqrt{2d\kappa^{-1}}\right).\end{aligned}$$

Convergence to The Target Distribution

- Objective: compute bound for $W_2(\delta_x Q_\gamma^n, \pi)$.
- Since $\pi P_t = \pi, \forall t \geq 0$, it suffices to get bounds of the Wasserstein distance

$$W_2(\delta_x Q_\gamma^n, \pi P_{\Gamma_n})$$

where

$$\Gamma_n = \sum_{k=1}^{n} \gamma_k.$$

  - $\delta_x Q_\gamma^n$: law of the discretized diffusion
  - $\pi P_{\gamma_n} = \pi$, where $(P_t)_{t \geq 0}$ is the semigroup of the diffusion.

Explicit Bound for The Euler Discretization

Theorem (Durums and Moulines, 2016)

Let $(\gamma_k)_{k\geq 1}$ be a non-increasing sequence with $\gamma_1 \leq \frac{1}{m+L}$. Then $\forall x \in \mathbb{R}^d$ and $n \geq 1$,

$$W_2^2(\delta_x Q_\gamma^n, \pi) \leq u_n^{(1)}(\gamma)\left(\|x - x^*\|^2 + \frac{d}{m}\right) + u_n^{(2)}(\gamma),$$

where

$$u_n^{(1)}(\gamma) = 2\prod_{k=1}^{n}(1 - \kappa\gamma_k/2)$$

$$u_n^{(2)}(\gamma) = L^2 d \sum_{i=1}^{n}\left[\gamma_i^2(\kappa^{-1} + \gamma_i)\left(2 + \frac{L^2\gamma_i}{m} + \frac{L^2\gamma_i^2}{6}\right)\prod_{k=i+1}^{n}(1 - \kappa\gamma_k/2)\right]$$

Idea: synchronous coupling between the diffusion and the interpolation of the Euler discretization!

A Coupling Proof

- $\forall n \geq 0$ and $t \in [\Gamma_n, \Gamma_{n+1})$, define

$$\begin{cases} X_t = X_{\Gamma_n} - \int_{\Gamma_n}^t \nabla U(X_s) ds + \sqrt{2}(B_t - B_{\Gamma_n}) \\ \bar{X}_t = \bar{X}_{\Gamma_n} - \int_{\Gamma_n}^t \nabla U(\bar{X}_{\Gamma_n}) ds + \sqrt{2}(B_t - B_{\Gamma_n}) \end{cases}$$

  with $X_0 \sim \pi$ and $\bar{X}_0 = x$.

- $\forall n \geq 0$,

$$W_2^2(\delta_x Q_\gamma^n, \pi P_{\Gamma_n}) \leq \mathbb{E}\|X_{\Gamma_n} - \bar{X}_{\Gamma_n}\|^2.$$

- Cancel out noise terms

$$\|X_{\Gamma_{n+1}} - \bar{X}_{\Gamma_{n+1}}\|^2 = \|X_{\Gamma_n} - \bar{X}_{\Gamma_n} - \int_{\Gamma_n}^{\Gamma_{n+1}} \nabla U(X_s) - \nabla U(\bar{X}_{\Gamma_n}) ds\|^2$$

$$= \|X_{\Gamma_n} - \bar{X}_{\Gamma_n}\|^2 + \|\int_{\Gamma_n}^{\Gamma_{n+1}} \nabla U(X_s) - \nabla U(\bar{X}_{\Gamma_n}) ds\|^2$$

$$-2\int_{\Gamma_n}^{\Gamma_{n+1}} \langle X_{\Gamma_n} - \bar{X}_{\Gamma_n}, \nabla U(X_s) - \nabla U(\bar{X}_{\Gamma_n})\rangle ds$$

A Coupling Proof

- Young's inequality and Jensen's inequality

$$\|\int_{\Gamma_n}^{\Gamma_{n+1}} \nabla U(X_s) - \nabla U(\bar{X}_{\Gamma_n})ds\|^2 \leq 2\gamma_{n+1}^2\|\nabla U(X_{\Gamma_n}) - \nabla U(\bar{X}_{\Gamma_n})\|^2 + 2\gamma_{n+1}\int_{\Gamma_n}^{\Gamma_{n+1}} \|\nabla U(X_s) - \nabla U(X_{\Gamma_n})\|^2 ds$$

- Since $\gamma_1 \leq 1/(m+L)$ and $(\gamma_k)_{k\geq 1}$ is non-increasing

$$\begin{aligned}\|X_{\Gamma_{n+1}} - \bar{X}_{\Gamma_{n+1}}\|^2 \leq &(1-\kappa\gamma_{n+1})\|X_{\Gamma_n} - \bar{X}_{\Gamma_n}\|^2 \\ &+ 2\gamma_{n+1}\int_{\Gamma_n}^{\Gamma_{n+1}} \|\nabla U(X_s) - \nabla U(X_{\Gamma_n})\|^2 ds \\ &- 2\int_{\Gamma_n}^{\Gamma_{n+1}} \langle X_{\Gamma_n} - \bar{X}_{\Gamma_n}, \nabla U(X_s) - \nabla U(X_{\Gamma_n})\rangle ds\end{aligned}$$

A Coupling Proof

- Using the Cauchy-Schwartz inequality, $\forall \epsilon > 0$

$$\|X_{\Gamma_{n+1}} - \bar{X}_{\Gamma_{n+1}}\|^2 \leq (1 - (\kappa - 2\epsilon)\gamma_{n+1})\|X_{\Gamma_n} - \bar{X}_{\Gamma_n}\|^2 + (2\gamma_{n+1} + (2\epsilon)^{-1}) \int_{\Gamma_n}^{\Gamma_{n+1}} \|\nabla U(X_s) - \nabla U(X_{\Gamma_n})\|^2 ds$$

Lemma

Let $(X_t)_{t\geq 0}$ be the solution of Langevin SDE with $X_0 = x$.
Then $\forall t \geq 0$ and $x \in \mathbb{R}^d$,

$$\mathbb{E}_x\|X_t - x\|^2 \leq dt\left(2 + \frac{L^2t^2}{3}\right) + \frac{3}{2}t^2L^2\|x - x^*\|^2.$$

A Coupling Proof

- Using the Lemma and L-smoothness

$$\mathbb{E}^{\mathcal{F}'_{\Gamma_n}}\|X_{\Gamma_{n+1}} - \bar{X}_{\Gamma_{n+1}}\|^2 \le (1-(\kappa-2\epsilon)\gamma_{n+1})\|X_{\Gamma_n} - \bar{X}_{\Gamma_n}\|^2$$
$$+L^2\gamma_{n+1}^2(\gamma_{n+1}+(4\epsilon)^{-1})\left(2d + L^2\gamma_{n+1}\|X_{\Gamma_n} - x^*\|^2 + dL^2\gamma_{n+1}^2/6\right)$$

- Note that $X_{\Gamma_n} \sim \pi$

$$\mathbb{E}\|X_{\Gamma_n} - x^*\|^2 \le \frac{d}{m}$$

- Let $\epsilon = \kappa/4$. By induction

$$\begin{aligned}\mathbb{E}\|X_{\Gamma_n} - \bar{X}_{\Gamma_n}\|^2 &\le \prod_{k=1}^{n}(1-\kappa\gamma_k/2)\mathbb{E}\|X_0 - x\|^2 + u_n^{(2)}(\gamma)\\ &\le u_n^{(1)}(\gamma)\mathbb{E}(\|x - x^*\|^2 + \|X_0 - x^*\|^2) + u_n^{(2)}(\gamma)\\ &\le u_n^{(1)}(\gamma)\left(\|x - x^*\|^2 + \frac{d}{m}\right) + u_n^{(2)}(\gamma).\end{aligned}$$

Results

- Fixed step size: $\forall \epsilon > 0$, one may choose γ so that

$$W_2(\delta_{x^*} R_\gamma^n, \pi) \leq \epsilon \quad \text{in } n = \mathcal{O}(d\epsilon^{-2}) \text{ iterations}$$

  where x^* is the unique maximum of π.

- Decreasing step size: with $\gamma_k = \gamma_1 k^{-\alpha},\ \alpha \in (0,1)$

$$W_2(\delta_{x^*} Q_\gamma^n, \pi) = \sqrt{d}\mathcal{O}(n^{-\alpha}).$$

- These results are tight (check with $U(x) = \frac{1}{2}\|x\|^2$).
- Similar results hold for total variation distance, see Durums and Moulines, 2016 for more details.

Underdamped Langevin Diffusion

- Underdamped Langevin SDE

$$
\begin{aligned}
dv_t &= -\gamma v_t dt - u\nabla U(x_t)dt + \sqrt{2\gamma u} dB_t \\
dx_t &= v_t dt
\end{aligned}
$$

- Notation: $(x_0, v_0) \sim p_0$ for some distribution p_0 on $\mathbb{R}^{2d}$. Then $(x_t, v_t) \sim p_t$. Let Φ_t denote the operator that maps from p_0 to p_t:

$$\Phi_t p_0 = p_t.$$

- $p^*(x, v) \propto \exp(-U(x) + \frac{1}{2u}\|v\|^2)$ is the unique invariant probability measure.

Discretized Underdamped Langevin Diffusion

- One step of the discrete underdamped Langevin diffusion is defined by the SDE

$$\begin{aligned} d\bar{v}_t &= -\gamma \bar{v}_t dt - u\nabla U(\bar{x}_0)dt + \sqrt{2\gamma u} dB_t \\ d\bar{x}_t &= \bar{v}_t dt \end{aligned} \tag{1}$$

  with an initial condition $(\bar{x}_0, \bar{v}_0) \sim \bar{p}_0$. Similarly, $(\bar{x}_t, \bar{v}_t) \sim \bar{p}_t$ and $\bar{\Phi}_t \bar{p}_0 = \bar{p}_t$.

- The above update has an analytical solution.

$$(\bar{x}_t, \bar{v}_t) \sim \mathcal{N}(\mu_t(\bar{x}_0, \bar{v}_0), \Sigma_t)$$

- With a small $t = \delta$, this can be used as a single step of discrete underdamped Langevin MCMC

$$(\bar{x}_{k+1}, \bar{v}_{k+1}) \sim \mathcal{N}(\mu_\delta(\bar{x}_k, \bar{v}_k), \Sigma_\delta) \tag{2}$$

Main Result

Theorem (Cheng et al., 2018)

Let $p^{(n)}$ be the distribution of $(\bar{x}_n, \bar{v}_n)$ after n iterations starting with $p^{(0)}(\bar{x}, \bar{v}) = 1_{x=x_0} \cdot 1_{v=0}$. Let the initial distance to optimum satisfy $\|x_0 - x^*\|^2 \leq \mathcal{D}^2$. If we set the step size to be

$$\delta = \frac{\epsilon}{104\kappa}\sqrt{\frac{1}{d/m + \mathcal{D}^2}},$$

and run update (2) for n iterations with

$$n \geq \frac{52\kappa^2}{\epsilon} \cdot \left(\sqrt{\frac{d}{m} + \mathcal{D}^2}\right) \cdot \log\left(\frac{24(\frac{d}{m} + \mathcal{D}^2)}{\epsilon}\right),$$

where $\kappa = L/m$, then we have the guarantee that

$$W_2(p^{(n)}, p^*) \leq \epsilon.$$

Remarks

- To converge to within ϵ of the target measure p^* in W_2 distance, underdamped Langevin diffusion requires $\mathcal{O}(\frac{\sqrt{d}}{\epsilon})$ iterations, which is a significant improvement over $\mathcal{O}(\frac{d}{\epsilon^2})$ of overdamped Langevin diffusion.
- The $\log(\frac{24(\frac{d}{m}+\mathcal{D}^2)}{\epsilon})$ factor can be removed by using a time-varying step size (Chen et al., 2018).
- Similar result holds for stochastic gradient underdamped Langevin diffusion, if the variance is made small enough (Chen et al., 2018).

Elements of Proof

Theorem (Cheng et al., 2018)

Let $u = \frac{1}{L}$ and $\gamma = 2$. $\forall t > 0$, there exists a coupling $\zeta_t(x_0, v_0, y_0, w_0) \in \prod(\Phi_t \delta_{x_0, v_0}, \Phi_t \delta_{y_0, w_0})$ such that

$$\begin{aligned} &\mathbb{E}_{(x_t, v_t, y_t, w_t) \sim \zeta_t(x_0, v_0, y_0, w_0)} \left(\|x_t - y_t\|^2 + \|(x_t + v_t) - (y_t + w_t)\|^2 \right) \\ &\qquad \leq e^{-\frac{t}{\kappa}} \left(\|x_0 - y_0\|^2 + \|(x_0 + v_0) - (y_0 + w_0)\|^2 \right) \end{aligned}$$

Corollary (Cheng et al., 2018)

Let p_0 be arbitrary distribution with $(x_0, v_0) \sim p_0$. Let q_0 and $\Phi_t q_0$ be the distributions of $(x_0, x_0 + v_0)$ and $(x_t, x_t + v_t)$, respectively. Then

$$W_2(\Phi_t q_0, q^*) \leq e^{-\frac{t}{2\kappa}} W_2(q_0, q^*)$$

where q^* is the distribution of $(x, x + v)$ when $(x, v) \sim p^*$.

Elements of Proof

- Sandwich Inequality

$$\frac{1}{2}W_2(p_t, p^*) \le W_2(q_t, q^*) \le 2W_2(p_t, p^*).$$

- Thus we also get convergence of $\Phi_t p_0$ to p^*

$$W_2(\Phi_t p_0, p^*) \le 4e^{-\frac{t}{2\kappa}} W_2(p_0, p^*).$$

- Bound the Discretization. Let $\delta \le 1$, $\forall p_0$

$$W_2(\Phi_\delta p_0, \bar{\Phi}_\delta p_0) \le \delta^2 \sqrt{\frac{2\mathcal{E}_K}{5}}$$

where $\mathcal{E}_K = 26(\frac{d}{m} + \mathcal{D}^2)$ is an upper bound of the kinetic energy

$$\mathbb{E}_{p_t}\|v\|^2 \le \mathcal{E}_K, \quad \forall t \in [0, \delta]$$

Elements of Proof

- Triangle inequality

$$\begin{aligned} W_2(q^{(i+1)}, q^*) &= W_2(\bar{\Phi}_\delta q^{(i)}, q^*) \\ &\le W_2(\Phi_\delta q^{(i)}, \bar{\Phi}_\delta q^{(i)}) + W_2(\Phi_\delta q^{(i)}, q^*) \\ &\le 2\delta^2 \sqrt{\frac{2\mathcal{E}_K}{5}} + e^{-\delta/2\kappa} W_2(q^{(i)}, q^*). \end{aligned}$$

- Let $\eta = e^{-\delta/2\kappa}$. By induction

$$\begin{aligned} W_2(q^{(n)}, q^*) &\le \eta^n W_2(q^{(0)}, q^*) + (1 + \eta + \ldots + \eta^{n-1}) 2\delta^2 \sqrt{\frac{2\mathcal{E}_K}{5}} \\ &\le 2\eta^n W_2(p^{(0)}, p^*) + \frac{2}{1-\eta}\delta^2 \sqrt{\frac{2\mathcal{E}_K}{5}}. \end{aligned}$$

Elements of Proof

- Using the sandwich inequality again

$$W_2(p^{(n)}, p^*) \leq 4\eta^n W_2(p^{(0)}, p^*) + \frac{4}{1-\eta}\delta^2\sqrt{\frac{2\mathcal{E}_K}{5}}. \qquad (3)$$

- Note that

$$\begin{aligned} W_2^2(p^{(0)}, p^*) &= \mathbb{E}_{(x,v)\sim p^*}\left(\|x_0 - x\|^2 + \|v\|^2\right) \\ &\leq 2\mathbb{E}_{x\sim p^*(x)}\|x - x^*\|^2 + 2\|x_0 - x^*\|^2 + \mathbb{E}_{v\sim p^*(v)}\|v\|^2 \\ &\leq \frac{2d}{m} + 2\mathcal{D}^2 + \frac{d}{L}. \end{aligned}$$

- Choosing δ and n to bound the right hand side of (3) completes the proof.

Conclusion

- We have shown some theoretical results on the convergence of MCMC algorithms.
- For the non-asymptotic analysis, similar results hold for other distances (e.g., total variation distance, KL divergence)
- These results have also been generalized to some non-convex settings (Cheng et al., 2018)
- These non-asymptotic results show that diffusion MCMC is a viable alternative to classic MCMC which requires little input from the user and can be computationally more efficient.

References

- W. Doeblin (1938), Exposé de la théorie des châines simples constantes de Markov á un nombre fini d'ètats. Revue Mathematique de l'Union Interbalkanique 2, 77–105.
- Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. Probability Surveys 1 20–71.
- Meyn, S.P. and Tweedie, R.L. (1993a) Markov Chains and Stochastic Stability. London: Springer-Verlag.
- Roberts, G. O. and Tweedie, R. L. (1996a). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. Biometrika 83 95–110.

References

- Roberts, G. O. and Tweedie, R. L. (1996b). Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli 341–363.
- S. Livingstone, M. Betancourt, S. Byrne, and M. Girolami. On the geometric ergodicity of Hamiltonian Monte Carlo. Bernoulli, 25(4A):3109–3138, 2019.
- Y. Nesterov. Introductory Lectures on Convex Optimization: A Basic Course. Applied Optimization. Springer, 2004.
- Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(3): 651–676, 2017a.

References

- Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In Proceedings of Algorithmic Learning Theory, volume 83 of Proceedings of Machine Learning Research, pages 186–211. PMLR, 07–09 Apr 2018.
- Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In Proceedings of the 2018 Conference on Learning Theory, volume 75 of Proceedings of Machine Learning Research, Stockholm, Sweden, 06–09 Jul 2018. PMLR.
- Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. Bernoulli, 25(4A):2854–2882, 2019.

References

- Alain Durmus, Szymon Majewski, and Blazej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. J. Mach. Learn. Res., 20:73–1, 2019.
- Xiang Cheng, Niladri S Chatterji, Yasin Abbasi-Yadkori, Peter L Bartlett, and Michael I Jordan. Sharp convergence rates for Langevin dynamics in the nonconvex setting. arXiv preprint arXiv:1805.01648, 2018.