The Langevin MCMC: Theory and Methods - Course 2

Alain Durmus, Eric Moulines

ENS Paris-Saclay and Ecole Polytechnique

August 9, 2019

Framework

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

$$x \mapsto \pi(x) \stackrel{\text{def}}{=} \mathrm{e}^{-U(x)} / \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)} \mathrm{d}y \ ,$$

Implicitly, $d \gg 1$.

- Assumption: U is L-smooth : twice continuously differentiable and there exists a constant L such that for all $x, y \in \mathbb{R}^d$,

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

(Overdamped) Langevin diffusion

- Langevin SDE:

$$\mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \ ,$$

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

- Notation: $(P_t)_{t \geq 0}$ 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.

Ergodicity

- Key property 1: For all $x \in \mathbb{R}^d$,

$$\lim_{t \to +\infty} \|\delta_x P_t - \pi\|_{\mathrm{TV}} = 0 \,.$$

- Key property 2: for "nice" functions

$$\frac{1}{T}\int_0^T f(X_t)\mathrm{d}t \overset{\mathbb{P}_x - \text{a.s.}}{\longrightarrow} \pi(f) = \int \pi(\mathrm{d}x) f(x)$$

$$\frac{1}{\sqrt{T}}\int_0^T \{f(X_t) - \pi(f)\}\mathrm{d}t \overset{\mathbb{P}_x}{\Longrightarrow} \mathcal{N}(0, \sigma^2(\pi, f)) \,.$$

- The Langevin diffusion provides a mean to sample any smooth distribution... Of course, this is a highly theoretical solution...

Discretized Langevin diffusion

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

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

  where
  - $(Z_k)_{k \geq 1}$ is i.i.d. $\mathcal{N}(0, \mathrm{I}_d)$
  - $(\gamma_k)_{k \geq 1}$ 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.

Discretized Langevin diffusion: 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 $\rightsquigarrow$ unique invariant distribution π_γ which does not coincide with the target distribution π.
- Questions:
  - For a given precision $\epsilon > 0$, how should I choose the stepsize $\gamma > 0$ and the number of iterations n so that : $\|\delta_x R_\gamma^n - \pi\|_{\mathrm{TV}} \leq \epsilon$
  - Is there a way to choose the starting point x cleverly ?
  - Auxiliary question: quantify the distance between π_γ and π.

Discretized Langevin diffusion: 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_γ^p is the composition of Markov kernels

$$Q_\gamma^p = R_{\gamma_1} R_{\gamma_2} \dots R_{\gamma_p}$$

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 $\|\delta_x Q_\gamma^p - \pi\|_{\mathrm{TV}} \to 0$ 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: $\|\delta_x Q_\gamma^p - \pi\|_{\mathrm{TV}} \leq \epsilon$ starting from a given point x
  - Should we use fixed or decreasing step sizes ?

Strongly convex potential

- Assumption: U is L-smooth and m-strongly convex

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

- Outline of the proof
  1. Control in W_2 the distance of the laws of the Langevin diffusion and its discretized version.
  2. Relate W_2 control to total variation.
- Key technique: (Synchronous and Reflection) coupling !
- Reference: Durmus and Moulines (2018), forthcoming paper in Bernoulli.

Coupling of probability measures

Definition

- A coupling of two probability measures $(\xi, \xi') \in \mathbb{M}_1(\mathcal{X}) \times \mathbb{M}_1(\mathcal{X})$ is a probability measure γ on the product space $(\mathsf{X} \times \mathsf{X}, \mathcal{X} \otimes \mathcal{X})$ whose marginals are ξ and ξ', *i.e.* $\gamma(A \times \mathsf{X}) = \xi(A)$ and $\gamma(\mathsf{X} \times A) = \xi'(A)$ for all $A \in \mathcal{X}$.
- The set of all couplings of ξ and ξ' is denoted by $\mathcal{C}(\xi, \xi')$.
- A coupling $\gamma \in \mathcal{C}(\xi, \xi')$ is said to be optimal for the Hamming distance if $\gamma(\Delta^c) = d_{\mathrm{TV}}(\xi, \xi')$ where $\Delta = \left\{(x, x') \in \mathsf{X}^2 \,:\, x = x'\right\}$ is the diagonal of $\mathsf{X} \times \mathsf{X}$

Wasserstein distance

Definition

For $p \geq 1$ and $\xi, \xi' \in \mathbb{M}_1(\mathcal{X})$, the Wasserstein distance of order p between ξ and ξ' denoted by $\mathbf{W}_{\mathrm{d},p}(\xi,\xi')$, is defined by

$$\mathbf{W}^p_{\mathrm{d},p}(\xi,\xi') = \inf_{\gamma \in \mathcal{C}(\xi,\xi')} \int_{\mathsf{X}\times\mathsf{X}} \mathrm{d}^p(x,x')\gamma(\mathrm{d}x\mathrm{d}x') ,$$

where $\mathcal{C}(\xi,\xi')$ is the set of coupling of ξ and ξ'. For $p=1$, we simply write $\mathbf{W}_{\mathrm{d}}$.

Properties of the Wasserstein distance

- The Wasserstein distance can be expressed in terms of random variables as:

$$\mathbf{W}_{\mathrm{d},p}\left(\xi,\xi'\right) = \inf_{(X,X')\in\mathcal{C}(\xi,\xi')} \left\{\mathbb{E}[\mathrm{d}^p(X,X')]\right\}^{1/p} ,$$

  where $(X,X') \in \mathcal{C}(\xi,\xi')$ that the distribution of the pair of random elements (X,X') is a coupling of ξ and ξ'.
- Any particular coupling therefore provides an upper bound of the Wasserstein distance.
- By Hölder's inequality, it obviously holds that if $p \leq q$, then for all $\xi,\xi' \in \mathbb{M}_1(\mathcal{X})$,

$$\mathbf{W}_{\mathrm{d},p}\left(\xi,\xi'\right) \leq \mathbf{W}_{\mathrm{d},q}\left(\xi,\xi'\right) .$$

Wasserstein distance convergence

Theorem

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

$$\mathbf{W}_2(\delta_x P_t, \delta_y P_t) \leq \mathrm{e}^{-mt} \|x - y\|$$

The contraction depends only on the strong convexity constant.

Synchronous Coupling

$$\begin{cases} \mathrm{d}Y_t &= -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \;, \\ \mathrm{d}\tilde{Y}_t &= -\nabla U(\tilde{Y}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \;, \end{cases} \quad \text{where } (Y_0, \tilde{Y}_0) = (x, y).$$

This SDE has a unique strong solution $(Y_t, \tilde{Y}_t)_{t \geq 0}$. Since

$$\mathrm{d}\{Y_t - \tilde{Y}_t\} = -\left\{\nabla U(Y_t) - \nabla U(\tilde{Y}_t)\right\}\mathrm{d}t$$

The product rule for semimartingales imply

$$\mathrm{d}\left\|Y_t - \tilde{Y}_t\right\|^2 = -2\left\langle \nabla U(Y_t) - \nabla U(\tilde{Y}_t), Y_t - \tilde{Y}_t \right\rangle \mathrm{d}t \;.$$

Synchronous Coupling

$$\left\|Y_t - \tilde{Y}_t\right\|^2 = \left\|Y_0 - \tilde{Y}_0\right\|^2 - 2\int_0^t \left\langle (\nabla U(Y_s) - \nabla U(\tilde{Y}_s)), Y_s - \tilde{Y}_s \right\rangle \mathrm{d}s \;,$$

Since U is strongly convex $\langle \nabla U(y) - \nabla U(y'), y - y' \rangle \geq m \left\| y - y' \right\|^2$ which implies

$$\left\|Y_t - \tilde{Y}_t\right\|^2 \leq \left\|Y_0 - \tilde{Y}_0\right\|^2 - 2m\int_0^t \left\|Y_s - \tilde{Y}_s\right\|^2 \mathrm{d}s \;.$$

Grömwall inequality:

$$\left\|Y_t - \tilde{Y}_t\right\|^2 \leq \left\|Y_0 - \tilde{Y}_0\right\|^2 \mathrm{e}^{-2mt}$$

Theorem

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

$$\mathbb{E}_x\left[\|Y_t - x^\star\|^2\right] \leq \|x - x^\star\|^2 \mathrm{e}^{-2mt} + \frac{d}{m}(1 - \mathrm{e}^{-2mt}) \; .$$

where

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

The stationary distribution π satisfies

$$\int_{\mathbb{R}^d} \|x - x^\star\|^2 \pi(\mathrm{d}x) \leq d/m.$$

The constant depends only linearly in the dimension d.

Elements of proof

- The generator $\mathscr{A}$ associated with $(P_t)_{t\geq 0}$ is given, for all $f \in C^2(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$ by:

$$\mathscr{A} f(x) = -\langle \nabla U(x), \nabla f(x) \rangle + \Delta f(x) \; .$$

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

$$\mathscr{A} V(x) = 2\left(-\langle \nabla U(x) - \nabla U(x^\star), x - x^\star \rangle + d\right) \leq 2\left(-mV(x) + d\right) \; .$$

Elements of proof

Key relation

$$\mathscr{A}V(x) \leq 2\left(-mV(x) + d\right) .$$

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

$$v(t,x) = P_t V(x) = \mathbb{E}_x\left[\|Y_t - x^\star\|^2\right]$$

We have

$$\frac{\partial v(t,x)}{\partial t} = P_t \mathscr{A} V(x) \leq -2mP_t V(x) + 2d = -2mv(t,x) + 2d ,$$

Grönwall inequality

$$v(t,x) = \mathbb{E}_x\left[\|Y_t - x^\star\|^2\right] \leq \|x - x^\star\|^2 \mathrm{e}^{-2mt} + \frac{d}{m}(1 - \mathrm{e}^{-2mt}) .$$

Elements of proof

Set $V(x) = \|x - x^\star\|^2$. By Jensen's inequality and for all $c > 0$ and $t > 0$, we get

$$\begin{aligned}\pi(V \wedge c) = \pi P_t(V \wedge c) \leq \pi(P_t V \wedge c) \\ = \int \pi(\mathrm{d}x)\, c \wedge \left\{ \|x - x^*\|^2 \mathrm{e}^{-2mt} + \frac{d}{m}(1 - \mathrm{e}^{-2mt}) \right\} \\ \leq \pi(V \wedge c)\mathrm{e}^{-2mt} + (1 - \mathrm{e}^{-2mt}) d/m \,.\end{aligned}$$

Taking the limit as $t \to +\infty$, we get $\pi(V \wedge c) \leq d/m$.

Contraction property of the discretization

Theorem

Assume that U is L-smooth and m-strongly convex. Then,

(i) *Let $(\gamma_k)_{k\geq 1}$ be a nonincreasing sequence with $\gamma_1 \leq 2/(m+L)$. For all $x, y \in \mathbb{R}^d$ and $\ell \geq n \geq 1$,*

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

where $\kappa = 2mL/(m+L)$.

(ii) *For any $\gamma \in (0, 2/(m+L))$, for all $x \in \mathbb{R}^d$ and $n \geq 1$,*

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

A coupling proof (I)

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

$$\mathbf{W}_2\left(\delta_x Q_\gamma^n, \pi P_{\Gamma_n}\right)$$

  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 semi group of the diffusion
- Idea ! synchronous coupling between the diffusion and the interpolation of the Euler discretization.

A coupling proof (II)

For all $n \geq 0$ and $t \in [\Gamma_n, \Gamma_{n+1})$ by

$$\begin{cases} Y_t = Y_{\Gamma_n} - \int_{\Gamma_n}^t \nabla U(Y_s) \mathrm{d}s + \sqrt{2}(B_t - B_{\Gamma_n}) \\ \bar{Y}_t = \bar{Y}_{\Gamma_n} - \int_{\Gamma_n}^t \nabla U(\bar{Y}_{\Gamma_n}) \mathrm{d}s + \sqrt{2}(B_t - B_{\Gamma_n}) \end{cases},$$

with $Y_0 \sim \pi$ and $\bar{Y}_0 = x$

For all $n \geq 0$,

$$\mathbf{W}_2^2 \left(\delta_x P_{\Gamma_n}, \pi Q_\gamma^n \right) \leq \mathbb{E}[\|Y_{\Gamma_n} - \bar{Y}_{\Gamma_n}\|^2] \,,$$

Explicit bound in Wasserstein distance

Theorem

Assume that U is m-strongly convex and L-smooth. Let $(\gamma_k)_{k\geq 1}$ be a nonincreasing sequence with $\gamma_1 \leq 1/(m+L)$. Then

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

where $u_n^{(1)}(\gamma) = 2 \prod_{k=1}^{n} (1 - \kappa\gamma_k)$ *with* $\kappa = mL/(m+L)$ *and*

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

Can be sharpened if U is three times continuously differentiable and there exists $\tilde{L}$ such that for all $x, y \in \mathbb{R}^d$, $\left\|\nabla^2 U(x) - \nabla^2 U(y)\right\| \leq \tilde{L} \|x - y\|$.

Results

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

$$\mathbf{W}_2\left(\delta_{x_*} R_\gamma^p, \pi\right) \leq \epsilon \quad \text{in } p = \mathcal{O}(\sqrt{d}\epsilon^{-1}) \text{ iterations}$$

where x_* is the unique maximum of π

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

$$\mathbf{W}_2\left(\delta_{x_*} Q_\gamma^n, \pi\right) = \sqrt{d}\mathcal{O}(n^{-\alpha}) \ .$$

- These results are tight (check with $U(x) = 1/2\|x\|^2$).

Total Variation

Definition

For μ, ν two probabilities measure on $\mathbb{R}^d$, define

$$d_{\mathrm{TV}}(\mu,\nu) = \frac{1}{2} \sup_{\|f\|_\infty \leq 1} |\mu(f) - \nu(f)| = \inf_{(X,Y) \in \mathcal{C}(\mu,\nu)} \mathbb{P}(X \neq Y),$$

where $(X,Y) \in \mathcal{C}(\mu,\nu)$ if $X \sim \mu$ and $Y \sim \nu$.

$$\begin{aligned} |\mu(f) - \nu(f)| &= \mathbb{E}[f(X) - f(Y)] \\ &= \mathbb{E}[\{f(X) - f(Y)\}\mathbb{1}_{\{X \neq Y\}}] \leq \mathrm{osc}(f)\mathbb{P}(X \neq Y) \ . \end{aligned}$$

From the Wasserstein distance to the TV

Theorem

If U is strongly convex, then for all $x, y \in \mathbb{R}^d$,

$$\|P_t(x,\cdot) - P_t(y,\cdot)\|_{\mathrm{TV}} \leq 1 - 2\boldsymbol{\Phi}\left\{-\frac{\|x-y\|}{\sqrt{(4/m)(\mathrm{e}^{2mt}-1)}}\right\}$$

Use reflection coupling (Lindvall and Rogers, 1986)

Hints of Proof I

$$\begin{cases} \mathrm{d}\mathbf{X}_t &= -\nabla U(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t^d \\ \mathrm{d}\mathbf{Y}_t &= -\nabla U(\mathbf{Y}_t)\mathrm{d}t + \sqrt{2}(\mathrm{Id} - 2\mathrm{e}_t\mathrm{e}_t^T)\mathrm{d}B_t^d \end{cases}, \qquad \text{where } \mathrm{e}_t = \mathrm{e}(\mathbf{X}_t - \mathbf{Y}_t)$$

with $\mathbf{X}_0 = x$, $\mathbf{Y}_0 = y$, $\mathrm{e}(z) = z/\|z\|$ for $z \neq 0$ and $\mathrm{e}(0) = 0$ otherwise. Define the coupling time $T_c = \inf\{s \geq 0 \mid \mathbf{X}_s \neq \mathbf{Y}_s\}$. By construction $\mathbf{X}_t = \mathbf{Y}_t$ for $t \geq T_c$.

$$\tilde{B}_t^d = \int_0^t (\mathrm{Id} - 2\mathrm{e}_s\mathrm{e}_s^T)\mathrm{d}B_s^d$$

is a d-dimensional Brownian motion, therefore $(\mathbf{X}_t)_{t\geq 0}$ and $(\mathbf{Y}_t)_{t\geq 0}$ are weak solutions to Langevin diffusions started at x and y, respectively. Then by Lindvall's inequality, for all $t > 0$ we have

$$\|P_t(x,\cdot) - P_t(y,\cdot)\|_{\mathrm{TV}} \leq \mathbb{P}(\mathbf{X}_t \neq \mathbf{Y}_t) \ .$$

Hints of Proof II

For $t < T_c$ (before the coupling time)

$$\mathrm{d}\{\mathbf{X}_t - \mathbf{Y}_t\} = -\{\nabla U(\mathbf{X}_t) - \nabla U(\mathbf{Y}_t)\}\,\mathrm{d}t + 2\sqrt{2}e_t \mathrm{dB}_t^1 \ .$$

Using Itô's formula

$$\begin{aligned}\|\mathbf{X}_t - \mathbf{Y}_t\| &= \|x-y\| - \int_0^t \langle \nabla U(\mathbf{X}_s) - \nabla U(\mathbf{Y}_s), e_s\rangle\,\mathrm{d}s + 2\sqrt{2}\mathrm{B}_t^1 \\ &\leq \|x-y\| - m\int_0^t \|\mathbf{X}_s - \mathbf{Y}_s\|\,\mathrm{d}s + 2\sqrt{2}\mathrm{B}_t^1 \ .\end{aligned}$$

and Grönwall's inequality implies

$$\|\mathbf{X}_t - \mathbf{Y}_t\| \leq \mathrm{e}^{-mt}\|x-y\| + 2\sqrt{2}\mathrm{B}_t^1 - m2\sqrt{2}\int_0^t \mathrm{B}_s^1 \mathrm{e}^{-m(t-s)}\mathrm{d}s \ .$$

Hint of Proof III

Therefore by integration by part, $\|\mathbf{X}_t - \mathbf{Y}_t\| \leq \mathsf{U}_t$ where $(\mathsf{U}_t)_{t \in (0,T_c)}$ is the one-dimensional Ornstein-Uhlenbeck process defined by

$$\mathsf{U}_t = \mathrm{e}^{-mt} \|x - y\| + 2\sqrt{2} \int_0^t \mathrm{e}^{m(s-t)} \mathrm{d}\mathsf{B}_s^1 = \mathrm{e}^{-mt} \|x - y\| + \int_0^{8t} \mathrm{e}^{m(s-t)} \mathrm{d}\tilde{B}_s^1$$

Therefore, for all $x, y \in \mathbb{R}^d$ and $t \geq 0$, we get

$$\mathbb{P}(T_c > t) \leq \mathbb{P}\left(\min_{0 \leq s \leq t} \mathsf{U}_t > 0\right) .$$

Finally the proof follows from the tail of the hitting time of (one-dimensional) OU (see Borodin and Salminen,2002).

From the Wasserstein distance to the TV (II)

$$\|P_t(x,\cdot) - P_t(y,\cdot)\|_{\mathrm{TV}} \leq \frac{\|x-y\|}{\sqrt{(2\pi/m)(\mathrm{e}^{2mt}-1)}}$$

Consequences:

1. $(P_t)_{t\geq 0}$ converges exponentially fast to π in total variation at a rate e^{-mt}.
2. For all $f : \mathbb{R}^d \to \mathbb{R}$, measurable and $\sup|f| \leq 1$, then the function $x \mapsto P_t f(x)$ is Lipschitz with Lipshitz constant smaller than

$$1/\sqrt{(2\pi/m)(\mathrm{e}^{2mt}-1)} \; .$$

Explicit bound in total variation

Theorem

- *Assume* U *is* L*-smooth and strongly convex. Let* $(\gamma_k)_{k\geq 1}$ *be a nonincreasing sequence with* $\gamma_1 \leq 1/(m+L)$.
- *(Optional assumption)* $U \in C^3(\mathbb{R}^d)$ *and there exists* $\tilde{L}$ *such that for all* $x, y \in \mathbb{R}^d$: $\left\|\nabla^2 U(x) - \nabla^2 U(y)\right\| \leq \tilde{L}\,\|x-y\|$.

Then there exist sequences $\{\tilde{u}_n^{(1)}(\gamma), n \in \mathbb{N}\}$ *and* $\{\tilde{u}_n^{(1)}(\gamma), n \in \mathbb{N}\}$ *such that for all* $x \in \mathbb{R}^d$ *and* $n \geq 1$,

$$\|\delta_x Q_\gamma^n - \pi\|_{\mathrm{TV}} \leq \tilde{u}_n^{(1)}(\gamma)\left\{\|x-x^\star\|^2 + d/m\right\} + \tilde{u}_n^{(2)}(\gamma)\,.$$

Constant step sizes

- For any $\epsilon > 0$, the minimal number of iterations to achieve $\|\delta_x Q_\gamma^p - \pi\|_{\mathrm{TV}} \leq \epsilon$ is

$$p = \mathcal{O}(\sqrt{d}\log(d)\epsilon^{-1}\left|\log(\epsilon)\right|) \ .$$

- For a given stepsize γ, letting $p \to +\infty$, we get:

$$\|\pi_\gamma - \pi\|_{\mathrm{TV}} \leq C\gamma\left|\log(\gamma)\right| \ .$$

Super-exponential density

- Super-exponential condition If there exist $\alpha > 1$, $\rho > 0$ and $M_\rho \geq 0$ such that for all $y \in \mathbb{R}^d$, $\|y\| \geq M_\rho$:

$$\langle \nabla U(y), y \rangle \geq \rho \|y\|^\alpha \ .$$

- If U is super-exponential, then $V(x) = \exp(U(x)/2)$ is a Lyapunov function.
- A function $V \in C^2(\mathbb{R}^d)$ is a Lyapunov function if $V \geq 1$ and if there exists $\theta > 0$, $b \geq 0$ and $R > 0$ such that,

$$\mathscr{A} V \leq -\theta V + b \mathbb{1}_{\mathrm{B}(0,R)} \ ,$$

where $\mathscr{A} f = -\langle \nabla U, \nabla f \rangle + \Delta f$ is the generator of the diffusion

Geometric convergence of the Euler discretization

- Let $(\gamma_k)_{k\geq 1}$ be a sequence of positive and non-increasing step sizes
- Euler discretization:

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

where $(Z_k)_{k\geq 1}$ is i.i.d. $\mathcal{N}(0, \mathrm{I}_d)$, independent of X_0.
- Markov kernel R_γ and $x \in \mathbb{R}^d$ by

$$R_\gamma(x, A) = \int_A \frac{1}{(4\pi\gamma)^{d/2}} \exp\left(-(4\gamma)^{-1} \|y - x + \gamma\nabla U(x)\|^2\right) \mathrm{d}y \ .$$

- The sequence $(X_n)_{n\geq 0}$ is a (possibly) time-nonhomogeneous Markov chain whose distribution is specified by the Markov kernels $(R_{\gamma_n})_{n\geq 1}$.

Level-0 results

- The Markov kernel R_γ is strongly Feller, irreducible, and hence all the compact sets are therefore small.
- Typically, the R_γ satisfies a Foster-Lyapunov drift condition: there exists $\kappa \in [0,1)$, $b > 0$ such that for all $\gamma > 0$

$$R_\gamma V \leq \kappa^\gamma V + \gamma b \; .$$

- R_γ admits a unique stationary distribution π_γ and is V-uniformly geometrically ergodic.

A drift condition for R_γ

Theorem

Assume U *is* L*-smooth and there exist* $\rho > 0$, $\alpha > 1$ *and* $M_\rho \geq 0$ *such that :*

$$\langle \nabla U(y), y \rangle \geq \rho \|y\|^\alpha \;, \quad \text{for all } y \in \mathbb{R}^d, \; \|y\| \geq M_\rho$$

Then for all $\bar{\gamma} \in \left(0, L^{-1}\right)$, *there exists* $b \geq 0$ *and* $s > 0$ *such that*

$$R_\gamma V(x) \leq \kappa^\gamma V(x) + \gamma b \;, \quad \text{for all } \gamma \in (0, \bar{\gamma}] \text{ and } x \in \mathbb{R}^d,$$

where

$$V(x) = \exp(U(x)/2).$$

Control of moments

- By a straightforward induction, we get for all $n \geq 0$ and $x \in \mathbb{R}^d$,

$$Q_\gamma^n V \leq \kappa^{\Gamma_{1,n}} V + b \sum_{i=1}^{n} \gamma_i \kappa^{\Gamma_{i+1,n}} .$$

  where for

$$n \leq p$$

  we have set $Q_\gamma^{n,p} = R_{\gamma_n} \cdots R_{\gamma_p}$.

- Note that for all $n \geq 1$, we have

$$\sum_{i=1}^{n} \gamma_i \kappa^{\Gamma_{i+1,n}} \leq \gamma_1 (1 - \kappa^{\Gamma_{1,n}})/(1 - \kappa^{\gamma_1}) .$$

Error decomposition

- Error decomposition

$$\|\mu_0 Q_\gamma^p - \pi\|_{\mathrm{TV}} \leq \|\mu_0 Q_\gamma^n Q_\gamma^{n+1,p} - \mu_0 Q_\gamma^n P_{\Gamma_{n+1,p}}\|_{\mathrm{TV}} + \|\mu_0 Q_\gamma^n P_{\Gamma_{n+1,p}} - \pi\|_{\mathrm{TV}} .$$

where

$$\Gamma_{n,p} \stackrel{\text{def}}{=} \sum_{k=n}^{p} \gamma_k , \qquad \Gamma_n = \Gamma_{1,n} .$$

- Second term on the RHS: contraction of the Markov semi-group of the diffusion (which is exponential)
- Problem: Find a way to compare the total variation distance between the diffusion and its discretization started at time Γ_n from the same distribution.

Coupling

- For all $x \in \mathbb{R}^d$, denote by $\mu^x_{n,p}$ and $\bar{\mu}^x_{n,p}$ the distributions on $\mathrm{C}([\Gamma_n, \Gamma_p], \mathbb{R}^d)$ of the Langevin diffusion $(Y_t)_{\Gamma_n \leq t \leq \Gamma_p}$ and of the Euler discretisation $(\bar{Y}_t)_{\Gamma_n \leq t \leq \Gamma_p}$ both started at x at time Γ_n.
- For any $\zeta_0 \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$, consider the diffusion $(Y_t, \overline{Y}_t)_{t \geq 0}$ with initial distribution equals to ζ_0, and defined for $t \geq 0$ by

$$\begin{cases} \mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \\ \mathrm{d}\bar{Y}_t = -\overline{\nabla U}(\bar{Y}, t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \end{cases}$$

and

$$\overline{\nabla U}(y, t) = \sum_{k=0}^{\infty} \nabla U(y_{\Gamma_k}) \mathbb{1}_{[\Gamma_k, \Gamma_{k+1})}(t)$$

Change of measure

- Let $(\xi_t)_{t\geq 0}$ and $(\eta_t)_{t\geq 0}$ be two diffusion type processes with

$$\mathrm{d}\xi_t = a_t(\xi)\mathrm{d}t + \sigma \mathrm{d}B_t, \qquad \text{for } t > 0,$$

 and

$$\mathrm{d}\eta_t = b_t(\eta)\mathrm{d}t + \sigma \mathrm{d}B_t \qquad \text{for } t > 0,$$

 where $\xi_0 = \eta_0$ is an $\mathcal{F}_0$ measurable random variable and σ is a positive constant.

- Suppose that the nonanticipative functionals $(a_t)_{t\geq 0}$ and $(b_t)_{t\geq 0}$ are such that a unique (continuous) strong solution exist for these equations.
- Suppose in addition that for any fixed $T > 0$,

$$\int_0^T [|a_s(\xi)|^2 + |b_s(\xi)|^2]\mathrm{d}s < \infty \text{ (a.s.) and } \int_0^T [|a_s(\eta)|^2 + |b_s(\eta)|^2]\mathrm{d}s < \infty \text{ (a.s.)},$$

Change of measure

Proposition

Under the stated assumptions, $\mu_\xi^T = \mathcal{L}(\xi_{[0,T]}) \backsim \mu_\eta^T = \mathcal{L}(\eta_{[0,T]})$ *and the densities are given by*

$$\frac{\mathrm{d}\mu_\eta^T}{\mathrm{d}\mu_\xi^T}(\xi) = \exp\left(-\sigma^{-2}\int_0^T \langle a_s(\xi) - b_s(\xi), \mathrm{d}\xi_s\rangle + \frac{1}{2\sigma^2}\int_0^T [|a_s(\xi)|^2 - |b_s(\xi)|^2]\mathrm{d}s\right)$$

and

$$\frac{\mathrm{d}\mu_\xi^T}{\mathrm{d}\mu_\eta^T}(\eta) = \exp\left(\sigma^{-2}\int_0^T \langle a_s(\eta) - b_s(\eta), \mathrm{d}\eta_s\rangle - \frac{1}{2\sigma^2}\int_0^T [|a_s(\eta)|^2 - |b_s(\eta)|^2]\mathrm{d}s\right).$$

Finally, the Kullback-Leibler divergence is given by

$$\mathrm{KL}(\mu_\xi^T, \mu_\eta^T) = \frac{1}{2}\mathbb{E}\left[\int_0^T |a_s(\xi) - b_s(\xi)|^2\mathrm{d}s\right].$$

Change of measure

- The Girsanov theorem for diffusion-like processes show that $\mu^x_{n,p} \sim \bar{\mu}^x_{n,p}$ with density

$$\frac{\mathrm{d}\mu^x_{n,p}}{\mathrm{d}\bar{\mu}^x_{n,p}}(\bar{Y}_s) = \exp\left(\frac{1}{2}\int_{\Gamma_n}^{\Gamma_p} \langle \nabla U(\bar{Y}_s) - \overline{\nabla U}(\bar{Y}_s, s), \mathrm{d}\bar{Y}_s\rangle \right.$$
$$\left. - \frac{1}{4}\int_{\Gamma_n}^{\Gamma_p} \left\{ \left\|\nabla U(\bar{Y}_s)\right\|^2 - \left\|\overline{\nabla U}(\bar{Y}, s)\right\|^2 \right\} \mathrm{d}s \right) .$$

- The Pinsker inequality implies that for all $x \in \mathbb{R}^d$

$$\|\delta_x Q_\gamma^{n+1,p} - \delta_x P_{\Gamma_{n+1,p}}\|_{\mathrm{TV}} \le 2^{-1}\left(\mathrm{Ent}_{\bar{\mu}^x_{n,p}}\left(\frac{\mathrm{d}\mu^x_{n,p}}{\mathrm{d}\bar{\mu}^x_{n,p}}\right)\right)^{1/2}$$
$$\le 4^{-1}\left(\int_{\Gamma_n}^{\Gamma_p} \mathbb{E}_x\left[\left\|\nabla U(\bar{Y}_s) - \overline{\nabla U}(\bar{Y}_s, s)\right\|^2\right] \mathrm{d}s\right)^{1/2} .$$

Change of measure

- Pinsker inequality: for all $x \in \mathbb{R}^d$

$$\|\delta_x Q_\gamma^{n+1,p} - \delta_x P_{\Gamma_{n+1,p}}\|_{\mathrm{TV}} \leq 4^{-1} \left(\int_{\Gamma_n}^{\Gamma_p} \mathbb{E}_x \left[\left\| \nabla U(\bar{Y}_s) - \overline{\nabla U}(\bar{Y}_s, s) \right\|^2 \right] \mathrm{d}s \right)^{1/2} .$$

- If U is L-smooth,

$$\|\delta_x Q_\gamma^{n+1,p} - \delta_x P_{\Gamma_{n+1,p}}\|_{\mathrm{TV}} \leq 4^{-1} L \left(\sum_{k=n+1}^{p} \left\{ (\gamma_k^3/3) \mathbb{E}_x \left[\|\nabla U(X_k)\|^2 \right] + d\gamma_k^2 \right\} \right)^{1/2} .$$

Convergence of the Euler discretization

Assumption

- There exist $\alpha > 1$, $\rho > 0$ and $M_\rho \geq 0$ such that for all $y \in \mathbb{R}^d$, $\|y\| \geq M_\rho$:

$$\langle \nabla U(y), y \rangle \geq \rho \|y\|^\alpha \ .$$

- U is convex.

Results Durmus and Moulines (2017).

- If $\lim_{\gamma_k \to +\infty} \gamma_k = 0$, and $\sum_k \gamma_k = +\infty$ then

$$\lim_{p \to +\infty} \|\delta_x Q_\gamma^p - \pi\|_{\mathrm{TV}} = 0 \ .$$

- $\|\pi_\gamma - \pi\|_{\mathrm{TV}} \leq C\sqrt{\gamma}$ (instead of γ)

Target precision ϵ: the convex case

- Setting U is convex. Constant stepsize
- Optimal stepsize γ and number of iterations p to achieve ϵ-accuracy in TV:

$$\|\delta_x Q_\gamma^p - \pi\|_{\mathrm{TV}} \leq \epsilon \; .$$

-

	d	ε	L
γ	$\mathcal{O}(d^{-3})$	$\mathcal{O}(\varepsilon^2/\log(\varepsilon^{-1}))$	$\mathcal{O}(L^{-2})$
p	$\mathcal{O}(d^5)$	$\mathcal{O}(\varepsilon^{-2}\log^2(\varepsilon^{-1}))$	$\mathcal{O}(L^2)$

- In the strongly convex case, $\sqrt{d}$!

Strongly convex outside a ball potential: Durmus and Moulines (2017)

- U is convex everywhere and there exist $r \geq 0$ and $m > 0$, such that for all $x, y \in \mathbb{R}^d$, $\|x - y\| \geq r$,

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

- Optimal stepsize γ and number of iterations p to achieve ϵ-accuracy in TV (starting point $x^\star \in \arg\min_{\mathbb{R}^d} U$):

$$\|\delta_{x^\star} R_\gamma^p - \pi\|_{\mathrm{TV}} \leq \epsilon \ .$$

	d	ε	L	m	r
γ	$\mathcal{O}(d^{-1})$	$\mathcal{O}(\varepsilon^2 / \log(\varepsilon^{-1}))$	$\mathcal{O}(L^{-2})$	$\mathcal{O}(m)$	$\mathcal{O}(r^{-4})$
p	$\mathcal{O}(d \log(d))$	$\mathcal{O}(\varepsilon^{-2} \log^2(\varepsilon^{-1}))$	$\mathcal{O}(L^2)$	$\mathcal{O}(m^{-2})$	$\mathcal{O}(r^8)$

How it works ?

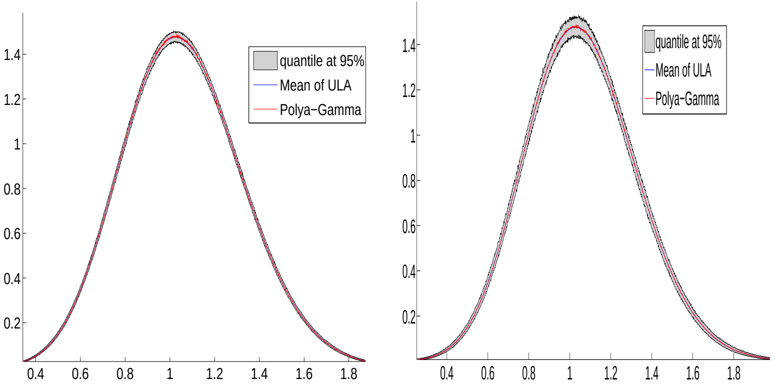

Figure: Empirical distribution comparison between the Polya-Gamma Gibbs Sampler and ULA. Left panel: constant step size $\gamma_k = \gamma_1$ for all $k \geq 1$; right panel: decreasing step size $\gamma_k = \gamma_1 k^{-1/2}$ for all $k \geq 1$

Data set	Observations p	Covariates d
German credit	1000	25
Heart disease	270	14
Australian credit	690	35
Musk	476	167

Table: Dimension of the data sets

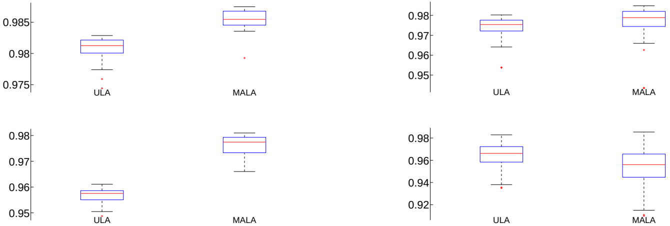

Figure: Marginal accuracy across all the dimensions. Upper left: German credit data set. Upper right: Australian credit data set. Lower left: Heart disease data set. Lower right: Musk data set

Conclusion

- Our goal is to avoid a Metropolis-Hastings accept-reject step We explore the efficiency and applicability of DMCMC to high-dimensional problems arising in a Bayesian framework, without performing the Metropolis-Hastings correction step.
- When classical (or adaptive) MCMC fails (for example, due to computational time restrictions or inability to select good proposals), we show that diffusion MCMC is a viable alternative which requires little input from the user and can be computationally more efficient.

References I

Durmus, A. and E. Moulines (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab. 27*(3), 1551–1587.

Durmus, A. and E. Moulines (2018, May). High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *ArXiv e-prints, Forthcoming in Bernoulli*.