Non-asymptotic convergence bound for the Unadjusted Langevin Algorithm

Alain Durmus, Eric Moulines

Telecom ParisTech et Ecole Polytechnique, CMAP

Many thanks to Andreas Eberle and Arnaud Guillin for sharing their invaluable insights and comments and introducing us to their new results. Thanks to Marcelo Pereyra for running the numerical experiments -not presented here- and pushing us to sharpen our results

Introduction

- Sampling distribution over high-dimensional state-space has recently attracted a lot of research efforts in computational statistics and machine learning community...
- Applications (non-exhaustive)
  1. Bayesian inference for high-dimensional models and Bayesian non parametrics
  2. Bayesian linear inverse problems (typically function space problems converted to high-dimensional problem by Galerkin method)
  3. Aggregation of estimators and experts
- Most of the sampling techniques known so far do not scale to high-dimension... Challenges are numerous in this area...

Logistic regression

- Likelihood: Binary regression set-up in which the binary observations (responses) $(Y_1, \ldots, Y_n)$ are conditionally independent Bernoulli random variables with success probability $F(\boldsymbol{\beta}^T X_i)$, where
  1. X_i is a d dimensional vector of known covariates,
  2. $\boldsymbol{\beta}$ is a d dimensional vector of unknown regression coefficient
  3. F is a distribution function.
- logistic regression: F is the standard logistic distribution function,

$$F(t) = \mathrm{e}^t / (1 + \mathrm{e}^t)$$

.

New challenges

Problem the number of predictor variables d is large (10^4 and up).
Examples

- text categorization,
- genomics and proteomics (gene expression analysis), ,
- other data mining tasks (recommendations, longitudinal clinical trials, ..).

Bayes 101

- Bayesian analysis requires a prior distribution for the unknown regression parameter

$$\pi(\boldsymbol{\beta}) = \mathrm{N}(0, \Sigma_{\boldsymbol{\beta}})$$

.

- The posterior of $\boldsymbol{\beta}$ is up to a proportionality constant given by

$$\pi(\boldsymbol{\beta}|(Y,X)) \propto \prod_{i=1}^{n} F^{Y_i}(\beta' X_i)(1 - F(\beta' X_i))^{1-Y_i} \pi(\boldsymbol{\beta})$$

- For probit and logistic link, the posterior density is intractable.

Data Augmentation

- The most popular algorithms for Bayesian inference in binary regression models are based on data augmentation:
  1. probit link: Albert and Chib (1993).
  2. logistic link: Polya-Gamma sampler, Polsson and Scott (2012)... !
- These two algorithms have been shown to be uniformly geometrically ergodic, BUT
  - The geometric rate of convergence is exponentially small with the dimension (show that the state space is a small set)
  - do not allow to construct honest confidence intervals, credible regions
- The algorithms are very demanding in terms of computational ressources...
  - applicable only when is d small 10 to moderate 100 but certainly not when d is large (10^4 or more).
  - convergence time prohibitive as soon as $d \geq 10^2$.

A daunting problem ?

- The posterior density distribution of β is given by Bayes' rule, up to a proportionality constant by

$$\pi(\boldsymbol{\beta}|(Y,X)) \propto \exp(-U(\boldsymbol{\beta}))\ .$$

where the potential $U(\boldsymbol{\beta})$ is given by

$$U(\boldsymbol{\beta}) = -\sum_{i=1}^{p}\{Y_i \log F(\boldsymbol{\beta}^T X_i) + (1-Y_i)\log(1-F(\boldsymbol{\beta}^T X_i))\} + (1/2)\boldsymbol{\beta}^T \Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}$$

- The potential $\boldsymbol{\beta} \mapsto U(\boldsymbol{\beta})$ is smooth, strongly convex...

Framework

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

$$x \mapsto \mathrm{e}^{-U(x)} / \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)} \mathrm{d}y \;,$$

  Implicitly, $d \gg 1$.

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

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

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.
- $\pi \propto \mathrm{e}^{-U}$ is reversible $\rightsquigarrow$ the unique invariant probability measure.
- The convergence to the stationary distribution takes place at geometrical rate.
  - Precise estimates of the convergence rate (TV, relative entropy) can be obtained using:
    - Functional inequalities: Poincaré or Log-Sobolev inequalities
    - Coupling techniques: synchronous or reflection coupling, depending upon the assumptions (Eberle, 2015)

Discretized Langevin diffusion

- Idea: Sample the diffusion paths, using for example 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 gradient algorithm.

Discretized Langevin diffusion: constant stepzize

- When $\gamma_k = \gamma$, 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.
- Problem: $\pi_\gamma \neq \pi$.

Metropolis-Adjusted Langevin Algorithm

- To correct the target distribution, a Metropolis-Hastings step can be included $\rightsquigarrow$ Metropolis Adjusted Langevin Agorithm (MALA).
  - Key references Roberts and Tweedie, 1996
- Algorithm:
  1. Propose $Y_{k+1} \sim X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} Z_{k+1}$, $Z_{k+1} \sim \mathcal{N}(0, \mathrm{I}_d)$
  2. Compute the acceptance ratio $\alpha_\gamma(X_k, Y_{k+1})$

$$\alpha_\gamma(x, y) = 1 \wedge \frac{\pi(y) r_\gamma(y, x)}{\pi(x) r_\gamma(x, y)} \;, r_\gamma(x, y) \propto \mathrm{e}^{-\|y - x - \gamma \nabla U(x)\|^2 / (4\gamma)}$$

  3. Accept / Reject the proposal.

MALA: pros and cons

- Require to compute 2 gradients at each iteration and to evaluate two times the objective function
- Geometric convergence is established under the condition that in the tail the acceptance region is inwards in q,

$$\lim_{\|x\| \to \infty} \int_{\mathcal{A}_\gamma(x) \Delta \mathcal{I}(x)} r_\gamma(x, y) \mathrm{d}y = 0 \; .$$

where $\mathcal{I}(x) = \{y, \|y\| \leq \|x\|\}$ and $A_\gamma(x)$ is the acceptance region

$$\mathcal{A}_\gamma(x) = \{y, \pi(x) r_\gamma(x, y) \leq \pi(y) r_\gamma(y, x)\}$$

Level-0: Ergodicity

- If the initial distribution μ_0 satisfies $\int \|x\|^2 \mu_0(\mathrm{d}x) < \infty$ then there exists a unique strong solution $(Y_t)_{t \geq 0}$ to

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

with Y_0 distributed according to μ_0.

- The semi-group $(P_t)_{t \geq 0}$ is
  - aperiodic, strong Feller (all compact sets are small).
  - reversible w.r.t. to π (admits π as its unique invariant distribution).
- For all initial distribution,

$$\lim_{t \to +\infty} \|\mu_0 P_t - \pi\|_{\mathrm{TV}} = 0 .$$

Foster-Lyapunov condition

- 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

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

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

then $V(x) = \exp(U(x)/2)$ is a Lyapunov function (constants are quantitative but may blow exponentially fast with the dimension of the state-space).

Geometric convergence

- If there exists a Lyapunov function then convergence to the stationary distribution can be shown to occur at an exponential rate.
- More precisely, there exists $\kappa \in [0, 1)$ such that for any initial distribution μ_0 and $t > 0$,

$$\|\mu_0 P_t - \pi\|_{\mathrm{TV}} \leq C(\mu_0)\kappa^t \ ,$$

  for some explicit function of the initial probability $C(\mu_0)$.
- Explicit expressions of the constant (the way dimension impacts theses constants critically depends on the assumptions on the potential U)

- 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

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

- 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 of a particular form, *i.e.* there exists $\kappa \in [0,1)$, $b > 0$ such that for all $\gamma > 0$

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

- It is well-known that under such assumption, R_γ admits a unique stationary distribution π_γ and that the Markov kernel is V-uniformly geometrically ergodic, in the sense that, for some constant $C < \infty$ and $\kappa \in [0,1)$, such that for all $x \in \mathbb{R}^d$,

$$\left\| R_\gamma^k(x, \cdot) - \pi_\gamma \right\|_V \leq C(\gamma) \kappa^{\gamma k} V(x) \, .$$

Example: 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 $x \in \mathbb{R}^d$,

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

- 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

- For $n \leq p$ set $Q_\gamma^{n,p} = R_{\gamma_n} \cdots R_{\gamma_p}$,
- 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.
- 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 laws 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

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

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

- The Girsanov theorem shows that $\mu^x_{n,p} \sim \bar{\mu}^x_{n,p}$ with density

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

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

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

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,

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

Back to the decomposition of the error

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

- Main result: For all $n, p \geq 1$, $n \leq p$, and $x \in \mathbb{R}^d$

$$\|\mu_0 Q_\gamma^p - \pi\|_{\mathrm{TV}} \leq C(\mu_0 Q_\gamma^n) V(x) \lambda^{\Gamma_{n+1,p}} + \left(D(d,\gamma) V(x) \sum_{k=n+1}^{p} \gamma_k^2 \right)^{1/2}$$

- If $\sum_k \gamma_k = \infty$, then

$$\|\mu_0 Q_\gamma^p - \pi\|_{\mathrm{TV}} \to 0 , \quad p \to \infty .$$

Controlling π_γ

- How far π_γ is from π ?
- Under the stated conditions, there exists an explicit constant $C(d)$ such that for all $\gamma \in [0, \bar{\gamma})$,

$$\|\pi - \pi_\gamma\|_V \leq C(d)\gamma^{1/2} .$$

Pinsker inequalities

Lemma (Generalized Pinsker inequality)

Let $\psi : \mathbb{R}^+ \to \mathbb{R}$ *be a* C^2 *convex function such that*

1. ψ *is uniformly convex on all bounded intervals,*
2. $\psi(1) = 0$ *and* $\lim_{u\to\infty} \psi(u)/u = +\infty$.

Then, for all (μ, ν) *on* $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ *such that* $\mu \ll \nu$,

$$\|\mu - \nu\|_{\mathrm{TV}} \leq c_\psi I_\psi^{1/2}(\mu|\nu) \,, \quad \text{where} \quad I_\psi(\mu,\nu) = \int \psi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu \,,$$

where $\mathrm{d}\mu/\mathrm{d}\nu$ *is the Radom-Nykodim derivative and* c_ψ *is a universal constant.*

Poincaré and Log-Sobolev inequalities

- Poincaré inequality: If $\psi(u) = (u-1)^2$, then $I_\psi(\mu,\nu)$ is the chi-square distance, $c_\psi = 1$ and

$$\|\mu - \nu\|_{\mathrm{TV}} \leq \mathrm{Var}_\nu^{1/2}\left\{\mathrm{d}\mu/\mathrm{d}\nu\right\} .$$

- Log-Sobolev inequality: If $\psi(u) = u\ln(u)$, then $I_\psi(\mu,\nu)$ is the Kullback-Leibler divergence and $c_\psi = 2$ and

$$\|\mu - \nu\|_{\mathrm{TV}} \leq \left(2\,\mathrm{KL}(\mu|\nu)\right)^{1/2} ,$$

"Carré du champ" inequalities

Theorem

Assume that there exists a constant C_ψ *such that for any density function* $h \in \mathcal{D}(\mathscr{A})$ *satisfying* $\int \psi(h) \mathrm{d}\pi < \infty$,

$$\int \psi\left(h\right) \mathrm{d}\pi \leq C_\psi \int \psi''\left(h\right) \|\nabla h\|^2 \mathrm{d}\pi \;.$$

Then, for all $t \geq 0$, *and any initial distribution* μ_0 *such that* $\mu_0 \ll \pi$,

$$\|\mu_0 P_t - \pi\|_{\mathrm{TV}} \leq c_\psi \mathrm{e}^{-t/C_\psi} I_\psi^{1/2}\left(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi} \cdot \pi, \pi\right) \;.$$

Poincaré inequality under Lyapunov condition

Theorem (after Barthe, Cattiaux, Guillin, 2009)

Assume that U is L-smooth and that

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

Then π satisfies a Poincaré inequality with constant

$$C_{\mathrm{lyap}} = -\theta^{-1}\left\{1 + b4R^2/\pi^2\, \mathrm{e}^{\mathrm{osc}_R(U)}\right\}$$

where

$$\mathrm{osc}_R(U) = \sup_{\mathrm{B}(0,R)}(U) - \inf_{\mathrm{B}(0,R)}(U) \;.$$

Poincaré inequality under convexity

Theorem (Bobkov, 1999)

Assume that U is L-smooth and convex. Then, π satisfies a Poincaré inequality with constant C_P given by

$$C_{\mathrm{cvx}} = 432 \int_{\mathbb{R}^d} \left\{ x - \int_{\mathbb{R}^d} y \mathrm{d}\pi(y) \right\}^2 \mathrm{d}\pi(x) \; .$$

If $\pi(x) = (2\pi)^{-d/2} \exp(-(1/2) x^T \Sigma^{-1} x)$ where Σ is a definite positive matrix, then C_{cvx} is proportional to $\mathrm{Tr}(\Sigma)$ (which typically scales linearly with the dimension).

Log-Sobolev inequalities

- If we apply the "carré du champ" inequality with $\psi(u) = u \ln(u)$, we obtain the log-Sobolev inequality.
- If there exists some constant C_{LS} such that, for any density $h \in \mathrm{D}(\mathscr{A})$ satisfying $\mathrm{Ent}_\pi(h) < \infty$ we have

$$\mathrm{Ent}_\pi(h) \leq C_{\mathrm{LS}} \int h^{-1} \|\nabla h\|^2 \mathrm{d}\pi \quad ,$$

then for all $t \geq 0$,

$$\|\mu_0 P_t - \pi\|_{\mathrm{TV}} \leq \exp(-t/C_{\mathrm{LS}}) \left(2\mathrm{Ent}_\pi(\mathrm{d}\mu_0/\mathrm{d}\pi)\right)^{1/2} \ .$$

Strong convexity

- Strong convexity There exists $m > 0$ such that for all $x, y \in \mathbb{R}^d$,

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

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

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

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

where

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

Log-Sobolev inequalities

Theorem

Assume that U is twice continuously differentiable, L-smooth and strongly convex. Then, for all probability measure $\mu_0 \ll \pi$ such that $(\mathrm{d}\mu_0/\mathrm{d}\pi)\log(\mathrm{d}\mu_0/\mathrm{d}\pi) \in \mathrm{L}^1(\pi)$, we have

$$\|\mu_0 P_t - \pi\|_{\mathrm{TV}} \le \mathrm{e}^{-mt}\left(2\mathrm{Ent}_\pi\left(\frac{\mathrm{d}\mu_0}{\mathrm{d}\pi}\right)\right)^{1/2} .$$

In such case, the ergodicity constant does not depend on the dimension.

The strongly convex case

- In the strongly convex case, a direct proof (with more explicit constants) can be obtained in Wasserstein distance...
- Idea: Bound with explicit constants, the Wasserstein distance between the diffusion and its discretized version by constructing a coupling between these two probabilities measures.
- Obvious candidate: synchronous coupling !

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

Theorem

Assume U is L-smooth and strongly convex. Let $(\gamma_k)_{k \geq 1}$ be a nonincreasing sequence with $\gamma_1 \leq 1/(m+L)$. Then for all $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and $n \geq 1$,

$$W_2^2(\mu_0 Q_\gamma^n, \pi) \leq u_n^{(1)}(\gamma) W_2^2(\mu_0, \pi) + u_n^{(2)}(\gamma) \;,$$

where

$$u_n^{(1)}(\gamma) \overset{\text{def}}{=} \prod_{k=1}^{n} (1 - \kappa\gamma_k/2) \quad \kappa = 2mL/(m+L)$$

and

$$u_n^{(2)}(\gamma) \overset{\text{def}}{=} L^2 \sum_{i=1}^{n} \gamma_i^2 \left\{\kappa^{-1} + \gamma_i\right\} (2d + dL^2\gamma_i/m + dL^2\gamma_i^2/6) \prod_{k=i+1}^{n} (1 - \kappa\gamma_k/2) \;,$$

Bounds for functionals

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Lipshitz function and $(X_k)_{k\geq 0}$ the Euler discretization of the Langevin diffusion. We approximate $\int_{\mathbb{R}^d} f(x)\pi(\mathrm{d}x)$ by the weighted average estimator

$$\hat{\pi}_n^N(f) = \sum_{k=N+1}^{N+n} \omega_{k,n}^N f(X_k) \; , \quad \omega_{k,n}^N = \gamma_{k+1}\Gamma_{N+2,N+n+1}^{-1} \; .$$

  where $N \geq 0$ is the length of the burn-in period, $n \geq 1$ is the number of effective samples.

- Objective: compute an explicit bounds for the Mean Square Error (MSE) of this estimator defined by:

$$\mathrm{MSE}_f(N,n) = \mathbb{E}_x\left[\left|\hat{\pi}_n^N(f) - \pi(f)\right|^2\right] \; .$$

- The MSE can be decomposed into the sum of the squared bias and the variance

$$\mathrm{MSE}_f(N,n) = \left\{\mathbb{E}_x[\hat{\pi}_n^N(f)] - \pi(f)\right\}^2 + \mathrm{Var}_x\left\{\hat{\pi}_n^N(f)\right\} ,$$

- Denote by ξ_k the optimal transference plan between $\delta_x Q_\gamma^k$ and π for W_2. Then by the Jensen inequality,

$$\begin{aligned}\mathrm{Bias}^2 &= \left(\sum_{k=N+1}^{N+n} \omega_{k,n}^N \int_{\mathbb{R}^d\times\mathbb{R}^d} \{f(z) - f(y)\}\xi_k(\mathrm{d}z,\mathrm{d}y)\right)^2 \\ &\leq \|f\|_{\mathrm{Lip}}^2 \sum_{k=N+1}^{N+n} \omega_{k,n}^N W_2^2(\delta_x Q_\gamma^k, \pi) .\end{aligned}$$

and

$$W_2^2(\delta_x Q_\gamma^k, \pi) \leq 2(\|x - x^\star\|^2 + d/m) u_k^{(1)}(\gamma) + u_k^{(2)}(\gamma) .$$

Gaussian Poincaré inequality

- If $Z = (Z_1, \ldots, Z_d) \sim \mathcal{N}(\mu, \mathrm{I}_d)$, then

$$\mathrm{Var}\left\{g(Z)\right\} \leq \|g\|_{\mathrm{Lip}}^2 \ .$$

- Idea: Apply to R_γ !... For any Lipshitz function $g : \mathbb{R}^d \to \mathbb{R}$, $\gamma > 0$ and $y \in \mathbb{R}^d$, we get

$$\begin{aligned} 0 &\leq R_\gamma \left\{g(\cdot) - R_\gamma g(y)\right\}^2 (y) \\ &= \int R_\gamma(y, \mathrm{d}z) \left\{g(z) - R_\gamma g(y)\right\}^2 \leq 2\gamma \|g\|_{\mathrm{Lip}}^2 \ . \end{aligned}$$

A martingale decomposition

- Idea Decompose $\hat{\pi}_n^N(f) - \mathbb{E}_x[\hat{\pi}_n^N(f)]$ as the sum of martingale increments,

$$\hat{\pi}_n^N(f) - \mathbb{E}_x[\hat{\pi}_n^N(f)] = \sum_{k=N}^{N+n-1} \left\{ \mathbb{E}_x^{\mathcal{G}_{k+1}}\left[\hat{\pi}_n^N(f)\right] - \mathbb{E}_x^{\mathcal{G}_k}\left[\hat{\pi}_n^N(f)\right] \right\} + \mathbb{E}_x^{\mathcal{G}_N}\left[\hat{\pi}_n^N(f)\right] - \mathbb{E}_x[\hat{\pi}_n^N(f)] ,$$

where $(\mathcal{G}_k)_{k\geq 0}$ is the natural filtration of $(X_k)_{k\geq 0}$.

- Variance:

$$\mathrm{Var}_x\left\{\hat{\pi}_n^N(f)\right\} = \sum_{k=N}^{N+n-1} \mathbb{E}_x\left[\left(\mathbb{E}_x^{\mathcal{G}_{k+1}}\left[\hat{\pi}_n^N(f)\right] - \mathbb{E}_x^{\mathcal{G}_k}\left[\hat{\pi}_n^N(f)\right]\right)^2\right] + \mathbb{E}_x\left[\left(\mathbb{E}_x^{\mathcal{G}_N}\left[\hat{\pi}_n^N(f)\right] - \mathbb{E}_x[\hat{\pi}_n^N(f)]\right)^2\right] .$$

Martingale decomposition

- Since $\hat{\pi}_n^N(f)$ is a sum, computing $\mathbb{E}_x^{\mathcal{G}_{k+1}}\left[\hat{\pi}_n^N(f)\right] - \mathbb{E}_x^{\mathcal{G}_k}\left[\hat{\pi}_n^N(f)\right]$ is easy.
- Set $S_{n,N+n}^N(x_{N+n}) = \omega_{N+n,n}^N f(x_{N+n})$ and define backward in time

$$S_{n,k}^N : x_k \mapsto \omega_{k,n}^N f(x_k) + R_{\gamma_{k+1}} S_{n,k+1}^N(x_k) \ .$$

- Variance: $\mathrm{Var}_x\left\{\hat{\pi}_n^N(f)\right\} = \sum_{k=1}^N V_k + W_N$ where

$$V_k = \mathbb{E}_x\left[R_{\gamma_{k+1}}\left\{S_{n,k+1}^N(\cdot) - R_{\gamma_{k+1}} S_{n,k+1}^N(X_k)\right\}^2(X_k)\right]$$

Bound of the incremental variance

- Idea: Prove that $S^N_{n,k+1}$ is Lipshitz and use recursively, backward in time, the Gaussian Poincaré inequality;
- Step 1:

$$\left|S^N_{n,k+1}(y) - S^N_{n,k+1}(z)\right| = \left|\omega^N_{k+1,n}\left\{f(y) - f(z)\right\} + \sum_{i=k+2}^{N+n} \omega^N_{i,n}\left\{Q^{k+2,i}_\gamma f(y) - Q^{k+2,i}_\gamma f(z)\right\}\right| .$$

- Step 2: (Monge-Kantorovitch duality)

$$W_1(\delta_y Q^{n,p}_\gamma, \delta_z Q^{n,p}_\gamma) \le \prod_{k=n}^{p} (1 - \kappa\gamma_k)^{1/2} \|y - z\| \ ;$$

Elements of proof

- Let ζ_0 be an OT plan of μ_0 and ν_0 and $(Z_k)_{k \geq n-1}$ be i.i.d. $\mathcal{N}(0, \mathrm{I}_d)$. Consider the processes $(X^1_{n-1,k}, X^2_{n-1,k})_{k \geq n-1}$ with initial distribution ζ_0 and defined for $k \geq n-1$ by

$$X^j_{n-1,k+1} = X^j_{n-1,k} - \gamma_{k+1} \nabla U(X^j_{n-1,k}) + \sqrt{2}\gamma_{k+1} Z_{k+1} \quad j = 1, 2 \; .$$

Elements of proof

- Let ζ_0 be an OT plan of μ_0 and ν_0 and $(Z_k)_{k\geq n-1}$ be i.i.d. $\mathcal{N}(0, \mathrm{I}_d)$. Consider the processes $(X^1_{n-1,k}, X^2_{n-1,k})_{k\geq n-1}$ with initial distribution ζ_0 and defined for $k \geq n-1$ by

$$X^j_{n-1,k+1} = X^j_{n-1,k} - \gamma_{k+1}\nabla U(X^j_{n-1,k}) + \sqrt{2}\gamma_{k+1}Z_{k+1} \quad j = 1, 2 \,.$$

- For any $p \geq n \geq 0$, $W_2^2(\mu_0 Q_\gamma^{n,p}, \nu_0 Q_\gamma^{n,p}) \leq \mathbb{E}\left[\left\|\Delta_{n-1,p}\right\|^2\right]$ with $\Delta_{n-1,k} = X^1_{n-1,k} - X^2_{n-1,k}$.

Elements of proof

- Let ζ_0 be an OT plan of μ_0 and ν_0 and $(Z_k)_{k \geq n-1}$ be i.i.d. $\mathcal{N}(0, \mathrm{I}_d)$. Consider the processes $(X^1_{n-1,k}, X^2_{n-1,k})_{k \geq n-1}$ with initial distribution ζ_0 and defined for $k \geq n-1$ by

$$X^j_{n-1,k+1} = X^j_{n-1,k} - \gamma_{k+1} \nabla U(X^j_{n-1,k}) + \sqrt{2}\gamma_{k+1} Z_{k+1} \qquad j = 1, 2 \, .$$

- For any $p \geq n \geq 0$, $W_2^2(\mu_0 Q^{n,p}_\gamma, \nu_0 Q^{n,p}_\gamma) \leq \mathbb{E}\left[\left\|\Delta_{n-1,p}\right\|^2\right]$ with $\Delta_{n-1,k} = X^1_{n-1,k} - X^2_{n-1,k}$.
- The strong convexity implies for $k \geq n-1$,

$$\left\|\Delta_{n-1,k+1}\right\|^2 = \left\|\Delta_{n-1,k}\right\|^2 + \gamma_{k+1}^2 \left\|\nabla U(X^1_{n-1,k}) - \nabla U(X^2_{n-1,k})\right\|^2$$
$$-2\gamma_{k+1} \left\langle \Delta_{n-1,k}, \nabla U(X^1_{n-1,k}) - \nabla U(X^2_{n-1,k}) \right\rangle \leq (1 - \kappa\gamma_{k+1}) \left\|\Delta_{n-1,k}\right\|^2$$

MSE

Theorem

Assume that U is L-smooth and strongly convex. Let $(\gamma_k)_{k\geq 1}$ be a nonincreasing sequence with $\gamma_1 \leq 2/(m+L)$. Then for all $N \geq 0$, $n \geq 1$ and Lipschitz functions $f : \mathbb{R}^d \to \mathbb{R}$, we get

$$\mathrm{Var}_x \left\{ \hat{\pi}_n^N(f) \right\} \leq 8\kappa^{-2} \|f\|_{\mathrm{Lip}}^2 \Gamma_{N+2,N+n+1}^{-1} u_{N,n}^{(3)}(\gamma)$$

where

$$u_{N,n}^{(3)}(\gamma) \overset{\text{def}}{=} \left\{ 1 + \Gamma_{N+2,N+n+1}^{-1}(\kappa^{-1} + 2/(m+L)) \right\} .$$

- The upper bound is independent of the dimension and allow to construct honest confidence bounds.
- The optimal rate for the variance is obtained for fixed stepsizes

MSE

	Bound for the MSE
$\alpha = 0$	$\gamma_1 + (\gamma_1 n)^{-1} \exp(-\kappa\gamma_1 N/2)$
$\alpha \in (0, 1/2)$	$\gamma_1 n^{-\alpha} + (\gamma_1 n^{1-\alpha})^{-1} \exp(-\kappa\gamma_1 N^{1-\alpha}/(2(1-\alpha)))$
$\alpha = 1/2$	$\gamma_1 \log(n) n^{-1/2} + (\gamma_1 n^{1/2})^{-1} \exp(-\kappa\gamma_1 N^{1/2}/4)$
$\alpha \in (1/2, 1)$	$n^{\alpha-1} \left\{\gamma_1 + \gamma_1^{-1} \exp(-\kappa\gamma_1 N^{1-\alpha}/(2(1-\alpha)))\right\}$
$\alpha = 1$	$\log(n)^{-1} \left\{\gamma_1 + \gamma_1^{-1} N^{-\gamma_1\kappa/2}\right\}$

Table: Bound for the MSE for $\gamma_k = \gamma_1 k^{-\alpha}$ as a function of γ_1, n and N

What's next ?

- A simple algorithm which scale easily in the dimension of the problem
- Computable bounds for convergence in TV, MSE, and deviation inequalities with constants which make sense !
- Future works
  - partial updates (coordinate descent)
  - sparsity inducing priors
  - detailed comparison with MALA
  - bias reduction ("exact estimation" à la Glynn and Rhee ?)