

Non-asymptotic convergence bound for the Unadjusted Langevin Algorithm

Eric Moulines

Institut Mines-Télécom / Telecom ParisTech / Laboratoire Traitement et Communication de
l'Information

Joint work with: A. Durmus

1 Motivation

2 Framework

3 Langevin diffusions

4 Naïve Euler discretisation of the Langevin diffusion

5 Deviation inequalities

6 Conclusion

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 and probit regression

- **Likelihood:** Binary regression set-up in which the binary observations (responses) (Y_1, \dots, Y_n) are conditionally independent Bernoulli random variables with success probability $F(\beta^T X_i)$, where
 - 1 X_i is a d dimensional vector of known covariates,
 - 2 β is a d dimensional vector of unknown regression coefficient
 - 3 F is a distribution function.
- Two important special cases:
 - 1 **probit regression:** F is the standard normal distribution function,
 - 2 **logistic regression:** F is the standard logistic distribution function,
 $F(t) = e^t / (1 + e^t)$.

New challenges

■ Problems

- 1 the number of predictor variables d is **large** (10^4 and up). Inverting a linear system is problematic.
- 2 the number of predictor usually of predictor variables exceeds the number of observations... ("**short,fat**" data sets)
- 3 or... both the number of predictor and observations are large... (**Big Data**)

■ 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. For simplicity $\pi(\beta) = \mathcal{N}(0, \Sigma_\beta)$ but in high dimension prior elicitation is challenging.
- The posterior of β is given by Bayes' rule, up to a proportionality constant by

$$\pi(\beta | (Y, X)) \propto \exp \left(\sum_{i=1}^n Y_i \beta^T X_i - \log(1 + e^{\beta^T X_i}) \right) \pi(\beta)$$

- For probit and logistic regression, the posterior density is **intractable**:
 - expectations with respect to $\pi(\beta | (Y, X))$ **cannot be computed in closed form**.

Data Augmentation

- These difficulties have spurred the development of many Markov chain Monte Carlo Methods (MCMC)
- The most popular algorithms in Bayesian Binary regression are based on **data augmentation**:
 - 1 probit regression: Albert and Chib (1993).
 - 2 logistic regression: Polya-Gamma sampler, Polsson and Scott (2012)... !
- Bayesian lexicon: **Data Augmentation** instead on sampling $\pi(\beta|(Y, X))$ sample $\pi(\beta, W|(Y, X))$ and marginalize.
 - Typical application of the Gibbs sampling: sample $\pi(\beta|W, Y, X)$ and $\pi(W|\beta, X, Y)$
 - The choice of the data augmentation should these two steps reasonably easy...

Data Augmentation algorithms

- These two algorithms have been shown to be uniformly geometrically ergodic, **BUT**
 - The geometric rate of convergence is exponentially small with the dimension
 - do not allow to construct **honest** confidence intervals, credible regions
- The algorithms are very demanding in terms of computational resources...
 - applicable only when $d = 10$ to $d = 100$ and certainly not 10^4 in their present implementation.

A daunting problem ?

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

$$\pi_{\beta}(\beta | (\mathbf{y}_i)_{1 \leq i \leq p}) \propto \exp(-U(\beta)) .$$

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

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

- The potential $\beta \mapsto U(\beta)$ is **smooth, strongly convex**...
- The corresponding optimization problem is conceptually straightforward and relatively to solve in high-dimension...

Tales for the Machine Learning Community

- Focus on **first order method**...
 - second-order information is so-costly to compute that the improvements are generally offset by computational cost.
 - solving a $d \times d$ linear system of equations can be difficult when $d \ll 1$.
- Algorithms which are easy to **parallelize** or **distribute** are a must.

1 Motivation

2 **Framework**

3 Langevin diffusions

4 Naïve Euler discretisation of the Langevin diffusion

5 Deviation inequalities

6 Conclusion

Framework

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

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

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:

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

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

- Unique strong solution $(Y_t)_{t \geq 0}$ which is a Markov process.
- $\pi \propto e^{-U}$ is **reversible** w.r.t. the Markov semi-group and is therefore the unique **invariant probability** measure.
- The convergence to the stationary distribution takes place at **geometrical rate**.
 - Precise estimates of the convergence rate can be obtained (using **Poincaré** or **Log-Sobolev** inequalities)

Discretized Langevin diffusion

- **Idea:** Sample the stationary distribution π by sampling approximately 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. $N(0, I_d)$
 - $(\gamma_k)_{k \geq 1}$ 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 **gradient algorithm** which is the core of most machine learning algorithms...

Discretized Langevin diffusion: constant stepsize

- When $\gamma_k = \gamma$, then $(X_k)_{k \geq 1}$ is an **homogeneous Markov chain** with Markov kernel

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

- Under some appropriate conditions, this Markov chain is irreducible and positive recurrent, but **its stationary distribution π_γ is no longer equal to the target π** .
- Even if π is not the stationary distribution, it is sensible to guess that $\pi_\gamma \rightarrow_{\gamma \rightarrow 0} \pi$.
- **Beware!** of course, when $\gamma = 0$ the chain becomes singular...

$$R_\gamma(x, \cdot) \Rightarrow_{\gamma \rightarrow 0} \delta_x .$$

Metropolis-Adjusted Langevin Algorithm

- To correct the bias, a Metropolis-Hastings step can be included: every EM step is a proposal in a Metropolis-Hastings algorithm, leading to **Metropolis Adjusted Langevin Algorithm** (MALA).
 - **Key references** Roberts and Tweedie, 1996
- **Algorithm:**

- 1 Propose $Y_{k+1} \sim X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} Z_{k+1}$, $Z_{k+1} \sim N(0, I_d)$
- 2 Compute the acceptance ratio $\alpha_\gamma(X_k, Y_{k+1})$

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

- 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** ,

$$\lim_{\|x\| \rightarrow \infty} \int_{A_\gamma(x) \Delta \mathcal{I}_\gamma(x)} r_\gamma(x, y) dy = 0 .$$

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

$$A_\gamma(x) = \{y, \pi(x)r_\gamma(x, y) \leq \pi(y)r_\gamma(y, x)\}$$

- ... *The 'inward converging' property is often easy to evaluate...*
Roberts, Tweedie, 1996... 20 years later, are you still so sure Gareth ?

Discretized Langevin diffusion: decreasing stepsize

- If $\lim_{k \rightarrow \infty} \downarrow \gamma_k = 0$, then $(X_k)_{k \geq 0}$ is a **non-homogeneous Markov chain**
- **Questions:**
 - 1 **Convergence** to the stationary distribution (in total variation ? Wasserstein distance ?)
 - 2 **Rate of convergence** to the stationary distribution
 - 3 **Non-asymptotic control** for additive functionals
- **Available results** Weak convergence (Lamberton and Pagès, 2002-2003, Lemaire and co-authors 2010) of the weighted empirical measure + CLT

$$\sum_{k=1}^p \frac{\gamma_k}{\Gamma_1^p} \delta_{X_k} \xrightarrow{p \rightarrow \infty} \pi, \quad \Gamma_1^p = \sum_{k=1}^p \gamma_k.$$

- 1 Motivation
- 2 Framework
- 3 Langevin diffusions**
- 4 Naïve Euler discretisation of the Langevin diffusion
- 5 Deviation inequalities
- 6 Conclusion

Level-0: Ergodicity

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

$$dY_t = -\nabla U(Y_t)dt + \sqrt{2}dB_t$$

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.
- In addition, the semi-group is reversible w.r.t. to π , and hence admits π as its (unique) invariant measure.
- Hence, for all initial distribution,

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

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,

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

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

- 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).

- The case $\alpha = 1$ may be dealt with, but with slightly different techniques.

Pinsker inequalities

Lemma (Generalized Pinsker inequality)

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

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

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

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

where $d\mu/d\nu$ 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

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

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

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

Level-1: "Carré du champ" inequalities

Theorem

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

$$\int \psi(h) d\pi \leq C_\psi \int \psi''(h) \mathcal{G}(h, h) d\pi,$$

where \mathcal{G} is the *carré du champ* operator,

$$\mathcal{G}(f, g) = 2^{-1} \{ \mathcal{A}fg - f\mathcal{A}g - g\mathcal{A}f \} \quad (= \langle \nabla f, \nabla g \rangle \text{ Langevin}).$$

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

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

Poincaré inequality under Lyapunov condition

Theorem (after Barthe, Cattiaux, Guillin, 2009)

Assume that U is L -smooth and that $\mathcal{A}V \leq -\theta V + b\mathbb{1}_{B(0,R)}$. Then π satisfies a *Poincaré inequality* with constant

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

where

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

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

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

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 $\text{Tr}(\Sigma)$ (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 $h \in \mathcal{D}(\mathcal{A})$, $\text{Ent}_\pi(f) < \infty$,

$$\text{Ent}_\pi(h) \leq C_{\text{LS}} \int h^{-1} \mathcal{G}(h) d\pi \quad ,$$

then for all $t \geq 0$,

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

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 $d\mu_0/d\pi \in L^2(\pi)$, we have

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

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

Wasserstein distance convergence

Theorem

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

- (i) For all probability measures μ_0 and $\nu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and $t \geq 0$,

$$W_2(\mu_0 P_t, \nu_0 P_t) \leq e^{-mt} W_2(\mu_0, \nu_0)$$

- (ii) The stationary distribution π satisfies

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

In addition, for any $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$,

$$W_2(\mu_0 P_t, \pi) \leq e^{-mt} W_2(\mu_0, \pi) .$$

Elements of proof

$$\begin{cases} dY_t &= -\nabla U(Y_t)dt + \sqrt{2}dB_t, \\ d\tilde{Y}_t &= -\nabla U(\tilde{Y}_t)dt + \sqrt{2}dB_t, \end{cases}$$

where (Y_0, \tilde{Y}_0) is some coupling between μ and ν .

Elements of proof

$$\begin{cases} dY_t &= -\nabla U(Y_t)dt + \sqrt{2}dB_t, \\ d\tilde{Y}_t &= -\nabla U(\tilde{Y}_t)dt + \sqrt{2}dB_t, \end{cases}$$

where (Y_0, \tilde{Y}_0) is some coupling between μ and ν . Since μ and ν are in $\mathcal{P}_2(\mathbb{R}^d)$ and ∇U is Lipschitz, then this SDE has a unique strong solution $(Y_t, \tilde{Y}_t)_{t \geq 0}$ associated to $(B_t)_{t \geq 0}$.

Elements of proof

$$\begin{cases} dY_t &= -\nabla U(Y_t)dt + \sqrt{2}dB_t, \\ d\tilde{Y}_t &= -\nabla U(\tilde{Y}_t)dt + \sqrt{2}dB_t, \end{cases}$$

where (Y_0, \tilde{Y}_0) is some coupling between μ and ν . Since μ and ν are in $\mathcal{P}_2(\mathbb{R}^d)$ and ∇U is Lipschitz, then this SDE has a unique strong solution $(Y_t, \tilde{Y}_t)_{t \geq 0}$ associated to $(B_t)_{t \geq 0}$. Moreover

$$\|Y_t - \tilde{Y}_t\|^2 = \|Y_0 - \tilde{Y}_0\|^2 - 2 \int_0^t \langle (\nabla U(Y_s) - \nabla U(\tilde{Y}_s)), Y_s - \tilde{Y}_s \rangle ds,$$

which implies using Grönwall's inequality that

$$\|Y_t - \tilde{Y}_t\|^2 \leq \|Y_0 - \tilde{Y}_0\|^2 - 2m \int_0^t \|Y_s - \tilde{Y}_s\|^2 ds \leq \|Y_0 - \tilde{Y}_0\|^2 e^{-2mt}.$$

Elements of proof

$$\begin{cases} dY_t &= -\nabla U(Y_t)dt + \sqrt{2}dB_t, \\ d\tilde{Y}_t &= -\nabla U(\tilde{Y}_t)dt + \sqrt{2}dB_t, \end{cases}$$

where (Y_0, \tilde{Y}_0) is some coupling between μ and ν . Since μ and ν are in $\mathcal{P}_2(\mathbb{R}^d)$ and ∇U is Lipschitz, then this SDE has a unique strong solution $(Y_t, \tilde{Y}_t)_{t \geq 0}$ associated to $(B_t)_{t \geq 0}$. Moreover

$$\|Y_t - \tilde{Y}_t\|^2 = \|Y_0 - \tilde{Y}_0\|^2 - 2 \int_0^t \langle (\nabla U(Y_s) - \nabla U(\tilde{Y}_s)), Y_s - \tilde{Y}_s \rangle ds,$$

which implies using Grönwall's inequality that

$$\|Y_t - \tilde{Y}_t\|^2 \leq \|Y_0 - \tilde{Y}_0\|^2 - 2m \int_0^t \|Y_s - \tilde{Y}_s\|^2 ds \leq \|Y_0 - \tilde{Y}_0\|^2 e^{-2mt}.$$

The proof follows since for all $t \geq 0$, the law of (Y_t, \tilde{Y}_t) is a coupling between μP_t and νP_t .

- 1 Motivation
- 2 Framework
- 3 Langevin diffusions
- 4 Naïve Euler discretisation of the Langevin diffusion**
- 5 Deviation inequalities
- 6 Conclusion

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

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

- The 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. $N(0, I_d)$, independent of X_0 .

- Markov kernel R_γ and $x \in \mathbb{R}^d$ by

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

- The sequence $(X_n)_{n \geq 0}$ is a (possibly nonhomogeneous) Markov chain with respect to the sequence of Markov kernels $(R_{\gamma_n})_{n \geq 1}$.

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

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

where $V(x) = \exp(U(x)/2)$. In addition, with $Q_\gamma^{n,p} = R_{\gamma_n} \cdots R_{\gamma_p}$,

$$Q_\gamma^n V(x) \leq e^{-s\gamma_1 \Gamma_{1,n}} V(x) + b_{\gamma_1, \alpha, \rho} \sum_{i=1}^n \gamma_i e^{-s\gamma_1 \Gamma_{i+1,n}}$$

Error decomposition

The total variation distance of the target probability measure π with the sequence $(\mu_0 Q_\gamma^n)_{n \geq 0}$, where μ_0 is the initial distribution, is decomposed as follows: for all $n, p \geq 1$ and $n \leq p$,

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

- We have the right tools to control the second terms on the RHS of the previous equation.
- We need to find a way to compare the total variation distance between the diffusion and its discretization started at time Γ_n from the same distribution.

A trivial coupling

- For all $x \in \mathbb{R}^d$, denote by $\mu_{n,p}^x$ and $\bar{\mu}_{n,p}^x$ the laws on $C([\Gamma_n, \Gamma_p], \mathbb{R}^d)$ 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 $(Y_t, \bar{Y}_t)_{t \geq 0}$ with initial distribution equals to ζ_0 , and defined for $t \geq 0$ by

$$\begin{cases} dY_t = -\nabla U(Y_t)dt + \sqrt{2}dB_t \\ d\bar{Y}_t = -\bar{\nabla} \bar{U}(\bar{Y}_t)dt + \sqrt{2}dB_t \end{cases}$$

and

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

Change of measure

- The **Girsanov theorem** shows that $\mu_{n,p}^x \ll \bar{\mu}_{n,p}^x$ with density

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

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

$$\begin{aligned} \|\delta_x Q_{\gamma}^{n+1,p} - \delta_x P_{\Gamma_{n+1,p}}\|_{\text{TV}} &\leq 2^{-1} \left(\text{Ent}_{\bar{\mu}_{n,p}^x} \left(\frac{d\mu_{n,p}^x}{d\bar{\mu}_{n,p}^x} \right) \right)^{1/2} \\ &\leq 4^{-1} \left(\int_{\Gamma_n}^{\Gamma_p} \mathbb{E}_x \left[\|\nabla U(\bar{Y}_s) - \bar{\nabla} U(\bar{Y}_s)\|^2 \right] ds \right)^{1/2}. \end{aligned}$$

Change of measure

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

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

- If U is L -smooth,

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

Back to the decomposition of the error

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

- The first term goes to zero if $\sum_{k=n+1}^p \gamma_k^2 \rightarrow 0$ as $n, p \rightarrow \infty$.
- The second-term goes to zero using either the Poincaré or the Log-Sobolev inequalities provided that $\Gamma_{n+1,p} = \sum_{k=n+1}^p \gamma_k \rightarrow \infty \dots$
- **Main result:** For all $n, p \geq 1$, $n \leq p$, and $x \in \mathbb{R}^d$

$$\|\mu_0 Q_\gamma^p - \pi\|_{\text{TV}} \leq C_1 V(x) \lambda^{\Gamma_{n+1,p}} + \left(C_2 V(x) \sum_{k=n+1}^p \gamma_k^2 \right)^{1/2}$$

- **Same conditions** than for stochastic approximation, $\sum_k \gamma_k = \infty$ and $\sum_k \gamma_k^2 < \infty \dots$

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 !

$$\begin{cases} dY_t = -\nabla U(Y_t)dt + \sqrt{2}dB_t \\ d\bar{Y}_t = -\bar{\nabla} \bar{U}(\bar{Y}_t)dt + \sqrt{2}dB_t \end{cases}$$

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

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

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

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

1 Motivation

2 Framework

3 Langevin diffusions

4 Naïve Euler discretisation of the Langevin diffusion

5 Deviation inequalities

6 Conclusion

Bounds for functionals

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

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

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:

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

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

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

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

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

and

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

Gaussian Poincaré inequality

- If $Z = (Z_1, \dots, Z_d) \sim N(\mu, I_d)$, then

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

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

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

A martingale decomposition

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

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

where $(\mathcal{G}_k)_{k \geq 0}$ is the natural filtration of $(X_k)_{k \geq 0}$ (sorry Sean, no Poisson equation today...).

- **Variance:**

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

Martingale decomposition

- $\hat{\pi}_n^N(f)$ is a sum and $\mathbb{E}_x^{\mathcal{G}^{k+1}} [\hat{\pi}_n^N(f)] - \mathbb{E}_x^{\mathcal{G}^k} [\hat{\pi}_n^N(f)]$ 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: $\text{Var}_x \{ \hat{\pi}_n^N(f) \} = \sum_{k=1}^N V_k + W_N$ 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 $S_{n,k+1}^N$ is Lipschitz and use recursively, backward in time, the Gaussian Poincaré inequality;
- **Step 1:**

$$\begin{aligned} |S_{n,k+1}^N(y) - S_{n,k+1}^N(z)| &= \left| \omega_{k+1,n}^N \{f(y) - f(z)\} \right. \\ &\quad \left. + \sum_{i=k+2}^{N+n} \omega_{i,n}^N \{Q_\gamma^{k+2,i} f(y) - Q_\gamma^{k+2,i} f(z)\} \right|. \end{aligned}$$

- **Step 2:** (Monge-Kantorovitch duality)

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

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. $N(0, I_d)$. Consider the processes $(X_{n-1,k}^1, X_{n-1,k}^2)_{k \geq n-1}$ with initial distribution ζ_0 and defined for $k \geq n-1$ by

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

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. $N(0, I_d)$. Consider the processes $(X_{n-1,k}^1, X_{n-1,k}^2)_{k \geq n-1}$ with initial distribution ζ_0 and defined for $k \geq n-1$ by

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

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

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. $N(0, I_d)$. Consider the processes $(X_{n-1,k}^1, X_{n-1,k}^2)_{k \geq n-1}$ with initial distribution ζ_0 and defined for $k \geq n-1$ by

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

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

$$\begin{aligned} \|\Delta_{n-1,k+1}\|^2 &= \|\Delta_{n-1,k}\|^2 + \gamma_{k+1}^2 \|\nabla U(X_{n-1,k}^1) - \nabla U(X_{n-1,k}^2)\|^2 \\ &\quad - 2\gamma_{k+1} \langle \Delta_{n-1,k}, \nabla U(X_{n-1,k}^1) - \nabla U(X_{n-1,k}^2) \rangle \leq (1 - \kappa \gamma_{k+1}) \|\Delta_{n-1,k}\|^2 \end{aligned}$$

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we get

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

where

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

- 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

1 Motivation

2 Framework

3 Langevin diffusions

4 Naïve Euler discretisation of the Langevin diffusion

5 Deviation inequalities

6 Conclusion

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 ?)