On Langevin Monte Carlo methods
CIRM
Advances in Computational Statistical Physics

Alain Durmus[1]
Joint work with: Nicolas Brosse[2], Eric Moulines[2],
Szymon Majewski[3], Błażej Miasojedow[3]

[1]ENS Paris-Saclay

[2]Ecole Polytechnique

[3]Institute of Mathematics, Polish Academy of Sciences

Outline

1. Motivation and setting
2. The ULA algorithm
3. ULA as an optimization algorithm in the Wasserstein space
4. Extensions of ULA
5. Normalizing constants of log-concave densities

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,
  2. Bayesian inverse problems (e.g., image restoration and deblurring),
  3. Aggregation of estimators and experts,
  4. Bayesian non-parametrics.
- Most of the sampling techniques known so far do not scale to high-dimension... Challenges are numerous in this area...

Bayesian setting

- A Bayesian model is specified by
  1. the sampling distribution of the observed data conditional on its parameters, often termed likelihood: $Y \sim \mathsf{L}(\cdot|\theta)$
  2. a prior distribution π_0 on the parameter space $\theta \in \mathbb{R}^d$
- The inference is based on the posterior distribution:

$$\boxed{\pi(\mathrm{d}\theta) = \frac{\pi_0(\mathrm{d}\theta)\mathsf{L}(Y|\theta)}{\int \mathsf{L}(Y|u)\pi_0(\mathrm{d}u)}\,.}$$

- In most cases the normalizing constant is not tractable:

$$\pi(\mathrm{d}\theta) \propto \pi_0(\mathrm{d}\theta)\mathsf{L}(Y|\theta)\,.$$

Logistic and probit regression

- Likelihood: Binary regression set-up in which the binary observations (responses) $\{Y_i\}_{i=1}^n$ are conditionally independent Bernoulli random variables with success probability $\{F(\beta^T X_i)\}_{i=1}^n$, 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 the link function.
- Two important special cases:
  1. probit regression: F is the standard normal cumulative distribution function,
  2. logistic regression: F is the standard logistic cumulative distribution function:

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

Bayes 101

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

$$\pi_0(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\beta}'\Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}\right) \quad \text{or} \quad \pi_0(\boldsymbol{\beta}) \propto \exp\left(-\sum_{i=1}^{d}\alpha_i|\beta_i|\right)$$

.

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

$$\pi(\boldsymbol{\beta}|(Y,X)) \propto \prod_{i=1}^{n} F^{Y_i}(\beta' X_i)(1-F(\beta' X_i))^{1-Y_i}\pi_0(\boldsymbol{\beta})$$

Bayesian setting

- Bayesian decision theory relies on computing expectations:

$$\pi(f) = \int_{\mathbb{R}^d} f(x)\mathrm{d}\pi(x) = \int_{\mathbb{R}^d} f(x)\pi(x)\mathrm{d}x$$

  Generic problem: estimation of an integral $\pi(f)$, where

  - π is known up to a multiplicative factor ;
  - Sampling directly from π is not an option;

- A solution is to approximate $\pi(f)$ by

$$n^{-1}\sum_{i=1}^{n} f(X_i) \ ,$$

  where $(X_i)_{i\geq 0}$ is a Markov chain associated with a Markov kernel P with invariant distribution π.

A daunting problem ?

- For Gaussian prior (ridge regression), the potential $\log(\pi)$ is smooth strongly convex.
- For Laplace prior (Lasso our fused Lasso) regression, the potential $\log(\pi)$ is non-smooth but still convex...
- A wealth of efficient optimisation algorithms are now available to solve this problem in very high-dimension...
- (long term) Objective:
  - Contribute to fill the gap between optimization and simulation. Good optimization methods are in general a good source of inspiration to design efficient sampler.
  - Develop algorithms converging to the target distribution polynomially with the dimension (more precise statements below)

Outline

1. Motivation and setting
2. The ULA algorithm
3. ULA as an optimization algorithm in the Wasserstein space
4. Extensions of ULA
5. Normalizing constants of log-concave densities

Framework

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

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

(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:
$$P_t(x, A) = \mathbb{P}(Y_t \in A | Y_0 = x) \ , \quad x \in \mathbb{R}^d, A \in \mathcal{B}(\mathbb{R}^d) \ .$$
- $\pi(x) \propto \exp(-U(x))$ is the unique invariant probability measure.

Discretized Langevin diffusion

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

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

  where
  - $(G_k)_{k \geq 1}$ 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.
- This algorithm is referred to as the Unadjusted Langevin Algorithm (ULA) in Bayesian statistics or Langevin Monte Carlo (LMC).

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, R_γ 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 : $\mathrm{d}(\delta_x R_\gamma^n, \pi) \leq \epsilon$ where d is some distance [could be the TV or the Wasserstein distance]
  - Is there a way to choose the starting point x cleverly ?

Some (very incomplete) references: Early references

1. **Statistical physics**: Parisi, 1981, *Correlation function and Computer Simulations*, Nuclear Physics.
2. **Bayesian statistics**: Grenander and Miller (in discussion Besag, *Representation of knowledge in Complex Systems*, JRSS B). First theoretical results given by Roberts and Tweedie, 1996, *Exponential Convergence of Langevin Distributions and Their Discrete Approximations*, Bernoulli, Stramer and Tweedie, *Langevin-type models. I. Diffusions with given stationary distributions and their discretizations.*, MCAP, 1999
3. Most of these results are qualitative (e.g. conditions upon which the sampler is geometrically ergodic).

(Very incomplete) existing results for ULA

1. Weak errors estimates Talay and Tubaro 1990; Lamberton and Pagès 2003.
2. Explicit errors Dalalyan 2014; Cheng and Bartlett 2017
3. These results are based on
   - the comparison between the discretization and the diffusion process
   - quantify how the error introduced by the discretization accumulate throughout the algorithm.
4. In the following, we introduce a new interpretation of ULA, as an optimization algorithm in the Wasserstein space.

Coupling of probability measures

- A coupling of two distributions ξ, ξ' on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is a distribution ζ on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}^d))$ satisfying

$$\zeta(\mathsf{A} \times \mathbb{R}^d) = \xi(\mathsf{A}) \text{ and } \zeta(\mathbb{R}^d \times \mathsf{A}) = \xi'(\mathsf{A}) \text{ for all } \mathsf{A} \in \mathcal{B}(\mathbb{R}^d) \,.$$

- The set of all couplings of ξ and ξ' is denoted by $\Pi(\xi, \xi')$.
- Let ξ, ξ' be two probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Define the Wasserstein distance of order 2 by

$$W_2^2\,(\xi, \xi') = \inf_{\zeta \in \Pi(\xi, \xi')} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - x'\|^2 \, \zeta(\mathrm{d}x \mathrm{d}x') \,.$$

- $\mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d) \,:\, \int_{\mathbb{R}^d} \|x\|^2 \, \mathrm{d}\mu(x) < +\infty\}$ equipped with W_2 is a Polish space, referred to as the Wasserstein space.
- $\mathcal{P}_2^{\mathrm{a}}(\mathbb{R}^d) = \{\mu \in \mathcal{P}_2(\mathbb{R}^d) \,:\, \mu << \mathrm{Leb}\}$.

Outline

1. Motivation and setting
2. The ULA algorithm
3. ULA as an optimization algorithm in the Wasserstein space
4. Extensions of ULA
5. Normalizing constants of log-concave densities

A different representation of Langevin dynamics

- Let $\mu_0 \in \mathcal{P}_2^{\mathrm{a}}(\mathbb{R}^d)$ and denote for any $t \geq 0$, $\rho_t = \mathrm{d}\mu_0 P_t / \mathrm{d}Leb$.
- Jordan, Kinderlehrer, and Otto 1998 shows that $(\rho_t)_{t>0}$ is the limit of a minimization scheme on $\mathcal{P}_2(\mathbb{R}^d)$ wrt $\mathscr{F} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$, the free energy functional,

$$\mathscr{F} = \mathscr{H} + \mathscr{E} \ .$$

- $\mathscr{H} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ is the Boltzmann H-functional:

$$\mathscr{H}(\mu) = \begin{cases} \int_{\mathbb{R}^d} \frac{\mathrm{d}\mu}{\mathrm{d}\,\mathrm{Leb}}(x) \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\,\mathrm{Leb}}(x)\right) \mathrm{d}x & \text{if } \mu \ll \mathrm{Leb} \\ +\infty \text{ otherwise} \ . \end{cases}$$

- $\mathscr{E} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ is the potential energy functional:

$$\mathscr{E}(\mu) = \int_{\mathbb{R}^d} U(x) \mathrm{d}\mu(x) \ .$$

- π is the unique minimizer of $\mathscr{F}$:

$$\mathscr{F}(\mu) - \mathscr{F}(\pi) = \mathrm{KL}\left(\mu | \pi\right) \ .$$

The minimization scheme of JKO

- Let $\mu_0 \in \mathcal{P}_2^{\mathsf{a}}(\mathbb{R}^d)$ and a stepsize $\gamma > 0$.
- For any $\nu \in \mathcal{P}_2^{\mathsf{a}}(\mathbb{R}^d)$, the functional on $\mathcal{P}_2^{\mathsf{a}}(\mathbb{R}^d)$

$$\mu \mapsto W_2^2(\nu, \mu)/2 + \gamma \mathscr{F}(\mu) \;,$$

  admits a unique minimizer.
- Then we can consider the sequence $(\tilde{\rho}_{k,\gamma})_{k \in \mathbb{N}}$ as follows.
  - Set $\rho_{0,\gamma} = \mathrm{d}\mu_0 / \mathrm{d}\,\mathsf{Leb}$.
  - for any $k \in \mathbb{N}^*$,

$$\tilde{\rho}_{k,\gamma} = \frac{\mathrm{d}\tilde{\mu}_{k,\gamma}}{\mathrm{d}\,\mathsf{Leb}} \;, \qquad \tilde{\mu}_{k,\gamma} = \underset{\mu \in \mathcal{P}_2^{\mathsf{a}}(\mathbb{R}^d)}{\operatorname{argmin}} \quad W_2(\tilde{\mu}_{k,\gamma}, \mu) + \gamma \mathscr{F}(\mu) \;.$$

- Set $\bar{\rho}_{0,\gamma} = \mathrm{d}\mu_0 / \mathrm{d}\,\mathsf{Leb}$ and $\bar{\rho}_{t,\gamma} = \tilde{\rho}_{\lfloor t/\gamma \rfloor, \gamma}$ for $t > 0$.
- Jordan, Kinderlehrer, and Otto 1998, Theorem 5.1 shows that for all $t > 0$,

$$\bar{\rho}_{t,\gamma} \to \rho_{t,\gamma} \text{ weakly in } \mathrm{L}^1(\mathbb{R}^d) \text{ as } \gamma \to 0 \quad .$$

Proximal optimization schemes

- The previous scheme can be seen as a proximal type algorithm on the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ used to minimize the functional $\mathscr{F}$ (Martinet 1970 and Rockafeller 1976).
- Consider a real l.s.c convex function f on a Hilbert space H equipped with $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$, and assume that f admits a minimizer.
- The proximal operator $\mathrm{prox}_f : \mathrm{H} \to \mathrm{H}$ associated with f is

$$\mathrm{prox}_f(x) = \underset{y \in \mathrm{H}}{\arg\min}\{\|x - y\|^2 /2 + f(y)\} .$$

- The *classical* proximal scheme to minimize f is defined as follows.
  - Consider a sequence of stepsizes $(\gamma_k)_{k \in \mathbb{N}^*}$ satisfying $\sum_{k=1}^{+\infty} \gamma_k = +\infty$ and $x_0 \in \mathrm{H}$.
  - Then, for any $k \in \mathbb{N}$,

$$x_{k+1} = \mathrm{prox}_{\gamma_{k+1} f}(x_k) = \underset{y \in \mathrm{H}}{\arg\min}\{\|x - y\|^2 + \gamma_{k+1} f(y)\} .$$

- Then, $(f(x_k))_{k \in \mathbb{N}}$ is nonincreasing and converges to $\min_{\mathrm{H}} f$.
- $(x_k)_{k \in \mathbb{N}}$ converges weakly to a minimizer of f.

A proximal scheme to sample from π?

- We could think about minimizing $\mathscr{F}$ using the previous minimization scheme.
- However, to our knowledge, finding explicit recursions $(\tilde{\rho}_{k,\gamma})_{k\in\mathbb{N}}$ is as difficult as minimizing $\mathscr{F}$.
- On the other hand, we can try to analyze ULA from this perspective.

Assumptions

H1 (m)

- $U : \mathbb{R}^d \to \mathbb{R}$ is m-convex, *i.e.* for all $x, y \in \mathbb{R}^d$,

$$U(tx + (1-t)y) \leq tU(x) + (1-t)U(y) - t(1-t)(m/2)\left\|x - y\right\|^2$$

- U admits a minimizer $x^\star$.

- Note that **H**1(m) includes the case where U is only convex when $m = 0$.

H2

U is continuously differentiable and L-gradient Lipschitz, *i.e.* there exists $L \geq 0$ such that for all $x, y \in \mathbb{R}^d$, $\|\nabla U(x) - \nabla U(y)\| \leq L\left\|x - y\right\|$.

Inexact gradient descent

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex C^1 objective function with

$$x_f \in \arg\min_{\mathbb{R}^d} f$$

- Consider the *inexact* or *stochastic* gradient descent algorithm used to estimate $f(x_f)$

$$x_{n+1} = x_n - \gamma_{n+1} \nabla f(x_n) + \gamma_{n+1} \Xi(x_n) \ ,$$

 where $(\gamma_k)_{k \in \mathbb{N}^*}$ is a non-increasing sequence of step sizes and $\Xi : \mathbb{R}^d \to \mathbb{R}^d$ is a deterministic or/and stochastic perturbation of ∇f.

- One possibility is to analyze the convergence of $(f(x_n))_{n \in \mathbb{N}}$ to $f(x_f)$ is to establish that the following inequality holds for any n:

$$2\gamma_{n+1}(f(x_{n+1}) - f(x_f)) \leq \|x_n - x_f\|^2 - \|x_{n+1} - x_f\|_2^2 + C\gamma_{n+1}^2 \ ,$$

 for some constant $C \geq 0$.

- Or for any initial point x_0,

$$2\gamma_1(f(x_1) - f(x_f)) \leq \|x_0 - x_f\|^2 - \|x_1 - x_f\|_2^2 + C\gamma_1^2 \ ,$$

 for some constant $C \geq 0$ independent of x_0.

Main result for ULA

- Recall that for any $\gamma > 0$, R_γ is the Markov chain associated with ULA

$$X_{k+1} = X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} G_{k+1} \ .$$

Theorem 1

Assume $\mathbf{H}1(m)$ *for* $m \geq 0$ *and* $\mathbf{H}2$*. For all* $\gamma \in (0, L^{-1}]$ *and* $\mu \in \mathcal{P}_2(\mathbb{R}^d)$*, we have*

$$2\gamma \left\{ \mathscr{F}(\mu R_\gamma) - \mathscr{F}(\pi) \right\} \leq (1 - m\gamma) W_2^2(\mu, \pi) - W_2^2(\mu R_\gamma, \pi) + 2\gamma^2 L d \ .$$

Proof of the main inequality

- For our analysis, we decompose $R_\gamma = S_\gamma T_\gamma$.
- S_γ and T_γ given for all $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$S_\gamma(x, \mathsf{A}) = \delta_{x-\gamma\nabla U(x)}(\mathsf{A}) \, ,$$
$$T_\gamma(x, \mathsf{A}) = (4\pi\gamma)^{-d/2} \int_{\mathsf{A}} \exp\left(- \|y - x\|^2 / (4\gamma) \right) \mathrm{d}y \, .$$

- S corresponds to the gradient step and T to the Gaussian step.
- Consider then the following decomposition

$$\mathscr{F}(\mu R_\gamma) - \mathscr{F}(\pi) = \mathscr{E}(\mu R_\gamma) - \mathscr{E}(\mu S_\gamma) + \mathscr{E}(\mu S_\gamma) - \mathscr{E}(\pi) + \mathscr{H}(\mu R_\gamma) - \mathscr{H}(\pi) \, .$$

- The proof consists in bounding each terms separately.

- $\mathscr{E}(\mu R_\gamma) - \mathscr{E}(\mu S_\gamma) = \mathscr{E}(\mu S_\gamma T_\gamma) - \mathscr{E}(\mu S_\gamma)$.

Lemma 2

Assume **H**2. *For all* $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ *and* $\gamma > 0$,

$$\mathscr{E}(\mu T_\gamma) - \mathscr{E}(\mu) \leq Ld\gamma \ .$$

- For all $x, \tilde{x} \in \mathbb{R}^d$, we have

$$|U(\tilde{x}) - U(x) - \langle \nabla U(x), \tilde{x} - x \rangle| \leq (L/2) \|\tilde{x} - x\|^2 \ .$$

- Therefore, for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma > 0$, we get

$$\begin{aligned}\mathscr{E}(\mu T_\gamma) - \mathscr{E}(\mu) &= (4\pi\gamma)^{-d/2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \{U(x+y) - U(x)\} \mathrm{e}^{-\|y\|^2/(4\gamma)} \mathrm{d}y \mathrm{d}\mu(x) \\ &\leq (4\pi\gamma)^{-d/2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left\{ \langle \nabla U(x), y \rangle + (L/2) \|y\|^2 \right\} \mathrm{e}^{-\|y\|^2/(4\gamma)} \mathrm{d}y \mathrm{d}\mu(x) \ ,\end{aligned}$$

- $\mathscr{E}(\mu S_\gamma) - \mathscr{E}(\pi)$

Lemma 3

Assume $\mathbf{H}1(m)$ *for* $m \geq 0$ *and* $\mathbf{H}2$*. For all* $\gamma \in (0, L^{-1}]$ *and* $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$*,*

$$2\gamma \left\{\mathscr{E}(\mu S_\gamma) - \mathscr{E}(\nu)\right\} \leq (1 - m\gamma) W_2^2(\mu, \nu) - W_2^2(\mu S_\gamma, \nu) \; .$$

- We start with the standard inequality from the convex optimization theory:

$$\begin{aligned} 2\gamma \left\{U(x - \gamma \nabla U(x)) - U(y)\right\} \leq (1 - m\gamma) \left\|x - y\right\|^2 - \left\|x - \gamma \nabla U(x) - y\right\|^2 \\ - \gamma^2 (1 - \gamma L) \left\|\nabla U(x)\right\|^2 \; . \end{aligned}$$

- Let (X, Y) be an optimal coupling between μ and ν, and we get

$$2\gamma \left\{\mathscr{E}(\mu S_\gamma) - \mathscr{E}(\nu)\right\} \leq (1 - m\gamma) W_2^2(\mu, \nu) - \mathbb{E}\left[\left\|X - \gamma \nabla U(X) - Y\right\|^2\right] .$$

- Using that $W_2^2(\mu S_\gamma, \nu) \leq \mathbb{E}[\left\|X - \gamma \nabla U(X) - Y\right\|^2]$ concludes the proof.

- $\mathscr{H}(\mu R_\gamma) - \mathscr{H}(\pi) = \mathscr{H}(\mu S_\gamma T_\gamma) - \mathscr{H}(\pi)$

Lemma 4

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathscr{H}(\nu) < \infty$. *Then for all* $\gamma > 0$,

$$2\gamma \{\mathscr{H}(\mu T_\gamma) - \mathscr{H}(\nu)\} \leq W_2^2(\mu, \nu) - W_2^2(\mu T_\gamma, \nu) \; .$$

- The proof just relies on properties of the heat semigroup!

Proof of Theorem 1

- Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma \in \mathbb{R}_+^*$ and recall

$$\mathscr{F}(\mu R_\gamma) - \mathscr{F}(\pi) = \mathscr{E}(\mu R_\gamma) - \mathscr{E}(\mu S_\gamma) + \mathscr{E}(\mu S_\gamma) - \mathscr{E}(\pi) + \mathscr{H}(\mu R_\gamma) - \mathscr{H}(\pi) \; .$$

- By Lemma 2, we get

$$\mathscr{E}(\mu R_\gamma) - \mathscr{E}(\mu S_\gamma) = \mathscr{E}(\mu S_\gamma T_\gamma) - \mathscr{E}(\mu S_\gamma) \leq L d \gamma \; .$$

- By Lemma 3,

$$2\gamma \left\{ \mathscr{E}(\mu S_\gamma) - \mathscr{E}(\pi) \right\} \leq (1 - m\gamma) W_2^2(\mu, \nu) - W_2^2(\mu S_\gamma, \nu) \; .$$

- By Lemma 4,

$$\begin{aligned} 2\gamma \left\{ \mathscr{H}(\mu R_\gamma) - \mathscr{H}(\pi) \right\} &= 2\gamma \left\{ \mathscr{H}((\mu S_\gamma) T_\gamma) - \mathscr{H}(\pi) \right\} \\ &\leq W_2^2(\mu S_\gamma, \pi) - W_2^2(\mu R_\gamma, \pi) \; . \end{aligned}$$

- Adding all these bounds, we obtain

$$2\gamma \left\{ \mathscr{F}(\mu R_\gamma) - \mathscr{F}(\pi) \right\} \leq (1 - m\gamma) W_2^2(\mu, \pi) - W_2^2(\mu R_\gamma, \pi) + 2\gamma^2 L d \; .$$

Complexity for ULA when U is strongly convex and gradient Lipschitz

- For the fixed stepsize setting

	Total variation	Wasserstein distance	KL divergence
Cheng and Bartlett, 2017	$d\mathcal{O}(\varepsilon^{-2})$	$d\mathcal{O}(\varepsilon^{-2})$	$d\mathcal{O}(\varepsilon^{-1})$
Our results	$d\mathcal{O}(\varepsilon^{-2})$	$d\mathcal{O}(\varepsilon^{-2})$	$d\mathcal{O}(\varepsilon^{-1})$

- Convergence in KL as $(\gamma_k)_{k\in\mathbb{N}^*}$ satisfies $\sum \gamma_k = +\infty$ and $\lim_{k\to+\infty} \gamma_k = 0$, with explicit rates.

Complexity for ULA when U is convex and gradient Lipschitz

- For the fixed stepsize setting

	Total variation	Wasserstein distance	KL divergence
Cheng nad Bartlett 2017	$d\mathcal{O}(\varepsilon^{-6})$	-	$d\mathcal{O}(\varepsilon^{-3})$
Our results	$d\mathcal{O}(\varepsilon^{-4})$	-	$d\mathcal{O}(\varepsilon^{-2})$

Table : Warm start

	Total variation	Wasserstein distance	KL divergence
Our results	$d^3\mathcal{O}(\varepsilon^{-4})$	-	$d^3\mathcal{O}(\varepsilon^{-2})$

Table : Starting from minimizer of U

- Convergence in KL as $(\gamma_k)_{k\in\mathbb{N}^*}$ satisfies $\sum \gamma_k = +\infty$ and $\lim_{k\to+\infty} \gamma_k = 0$, with explicit rates.

Outline

1. Motivation and setting

2. The ULA algorithm

3. ULA as an optimization algorithm in the Wasserstein space

4. Extensions of ULA

5. Normalizing constants of log-concave densities

Stochastic Gradient Langevin Dynamics (SGLD)

- The ULA algorithm is a discretization of the overdamped Langevin diffusion, which leaves invariant the target distribution π.
- To further reduce the computational cost, SGLD uses unbiased estimators of the gradient of the log-posterior based on subsampling.
- This method, initially proposed in Welling and Teh 2011 has triggered a huge number of works.

SGLD Algorithm

- We are interested in situations where the target distribution π arises as the posterior distribution in a Bayesian inference problem with prior density $\pi_0(\theta)$ and a large number $N \gg 1$ of i.i.d. observations z_i with likelihoods $p(z_i|\theta)$:

$$\pi(\theta) = \pi_0(\theta) \prod_{i=1}^{N} p(z_i|\theta) \, .$$

- We denote $U_i(\theta) = -\log(p(z_i|\theta))$ for $i \in \{1, \dots, N\}$, $U_0(\theta) = -\log(\pi_0(\theta))$, $U = \sum_{i=0}^{N} U_i$.
- the cost of one iteration is Nd which is prohibitively large for massive datasets.

SGLD Algorithm

- Welling and Teh 2011 suggested to replace ∇U with an unbiased estimate

$$H_S(\theta) = \nabla U_0(\theta) + (N/p) \sum_{i \in S} \nabla U_i(\theta)$$

where S is a minibatch of $\{1, \ldots, N\}$ with replacement of size p.

- A single update of SGLD is then given by

$$\theta_{k+1} = \theta_k - \gamma H_{S_{k+1}}(\theta_k) + \sqrt{2\gamma} G_{k+1} \ .$$

- The idea of using only a fraction of data points to compute an unbiased estimate of the gradient at each iteration comes from Stochastic Gradient Descent (SGD) which is a popular algorithm to minimize the potential U.
- Generalization of this method to non-smooth convex function.
- Can we derive new schemes to sample from non-smooth potential U?

Stochastic Sub-Gradient Langevin Dynamics

H3

I The potential U is M-Lipschitz, *i.e.* for all $x, y \in \mathbb{R}^d$, $|U(x) - U(y)| \leq M \|x - y\|$.

II There exists a measurable space $(\mathsf{Z}, \mathcal{Z})$, a probability measure η on $(\mathsf{Z}, \mathcal{Z})$ and a measurable function $\Theta : \mathbb{R}^d \times \mathsf{Z} \to \mathbb{R}^d$ for all $x \in \mathbb{R}^d$,

$$\int_{\mathsf{Z}} \Theta(x, z) \mathrm{d}\eta(z) \in \partial U(x) \; .$$

III The variance of the stochastic subgradient is bounded by D: for any $x \in \mathbb{R}^d$,

$$\int_{\mathbb{R}^d \times \mathsf{Z}} \left\| \Theta(x, z) - \int_{\mathsf{Z}} \Theta(x, \tilde{z}) \mathrm{d}\eta(\tilde{z}) \right\|^2 \mathrm{d}\eta(z) < D \; .$$

Complexity of SSGLD

- Stochastic Sub-Gradient Langevin Dynamics (SSGLD)

$$\bar{X}_{n+1} = \bar{X}_n - \gamma_{n+1}\Theta(\bar{X}_n, Z_{n+1}) + \sqrt{2\gamma_{n+2}}G_{n+1} \ ,$$

  where $(Z_k)_{k\in\mathbb{N}^*}$ be a sequence of i.i.d. random variables distributed according to η, $(\gamma_k)_{k\in\mathbb{N}^*}$ be a sequence of non-increasing step sizes.

- For the fixed step size setting.
  - Starting from a warm start, we find that the complexity of SSGLD to obtain a sample ε close from π in KL is of order $(M^2 + D^2)\mathcal{O}(\varepsilon^{-2})$.
  - If for all $x \in \mathbb{R}^d$, $x \notin \mathrm{B}(x^\star, M_\eta)$,

$$U(x) - U(x^\star) \geq \eta \left\| x - x^\star \right\|$$

    then starting from $\delta_{x^\star}$, we get the overall complexity of SSGLD for the KL:

$$(\eta^{-2}d^2 + M_\eta^2)(M^2 + D^2)\mathcal{O}(\varepsilon^{-2}) \ .$$

- Convergence as $\gamma_k \to 0$ as $k \to +\infty$ (with appropriate conditions).

Stochastic Proximal Gradient Langevin Dynamics

H4 (m)

There exists $U_1 : \mathbb{R}^d \to \mathbb{R}$ and $U_2 : \mathbb{R}^d \to \mathbb{R}$ such that $U = U_1 + U_2$ and satisfying the following assumptions:

1. U_1 satisfies **H**1(m) and **H**2. In addition, there exists a measurable space $(\tilde{\mathsf{Z}}, \tilde{\mathcal{Z}})$, a probability measure $\tilde{\eta}_1$ on $(\tilde{\mathsf{Z}}, \tilde{\mathcal{Z}})$ and a measurable function $\tilde{\Theta}_1 : \mathbb{R}^d \times \mathsf{Z} \to \mathbb{R}^d$ such that for all $x \in \mathbb{R}^d$,
$$\int_{\tilde{\mathsf{Z}}} \tilde{\Theta}_1(x, \tilde{z}) \mathrm{d}\tilde{\eta}_1(\tilde{z}) = \nabla U_1(x) \ .$$
2. U_2 satisfies **H**1(0) and is M_2-Lipschitz.
3. The variance of the stochastic subgradient is bounded by D: for any $x \in \mathbb{R}^d$,
$$\int_{\mathbb{R}^d \times \tilde{\mathsf{Z}}} \left\| \tilde{\Theta}_1(x, z) - \int_{\tilde{\mathsf{Z}}} \tilde{\Theta}_1(x, \tilde{z}) \mathrm{d}\eta(\tilde{z}) \right\|^2 \mathrm{d}\eta(z) < D \ .$$

Complexity of SPGLD

- Stochastic Proximal Gradient Langevin Dynamics (SPGLD)

$$\tilde{X}_{n+1} = \mathrm{prox}^{U_2}_{\gamma_{n+1}}(\tilde{X}_n) - \gamma_{n+2}\tilde{\Theta}_1\{\mathrm{prox}_{\gamma_{n+1}U_2}(\tilde{X}_n), \tilde{Z}_{n+1}\} + \sqrt{2\gamma_{n+2}}G_{n+1} \ ,$$

  where $(\tilde{Z}_k)_{k\in\mathbb{N}^*}$ be a sequence of i.i.d. random variables distributed according to η_1.

- In the fixed stepsize setting
  - Starting from a warm start we get that the complexity of SPGLD to obtain a sample ε close from π in KL is of order $(d + M^2 + D^2)\mathcal{O}(\varepsilon^{-2})$.
  - If for all $x \in \mathbb{R}^d$, $x \notin \mathrm{B}(x^\star, M_\eta)$,

$$U(x) - U(x^\star) \geq \eta \|x - x^\star\|$$

    then starting at $\delta_{x^\star}$, we get the overall complexity of SPGLD for the KL:

$$(\eta^{-2}d^2 + M_\eta^2)(d + M_2^2 + D^2)\mathcal{O}(\varepsilon^{-2}) \ .$$

Summary

- We give a new interpretation of ULA and use it to get bounds on the Kullback-Leibler divergence from π to the iterates of ULA.
- We recover the dependence on the dimension of Cheng and Bartlett 2017 in the strongly convex case. We also give computable bounds when U is only convex which improves the results of Dalalyan 2014 and Cheng and Bartlett 2017.
- We propose two new methodologies to sample from a non-smooth potential U and make a non-asymptotic analysis of them. These two new algorithms are generalizations of SGLD.

Outline

1. Motivation and setting
2. The ULA algorithm
3. ULA as an optimization algorithm in the Wasserstein space
4. Extensions of ULA
5. Normalizing constants of log-concave densities

Normalizing constants

- Let $U : \mathbb{R}^d \to \mathbb{R}$. We aim at estimating $\mathcal{Z} = \int_{\mathbb{R}^d} \mathrm{e}^{-U(x)} \mathrm{d}x < +\infty$.
- $\mathcal{Z}$ is the normalizing constant of the probability density π associated with the potential U.
- Many applications in Bayesian inference (Bayes factors) and statistical physics (free energy) .
- In Bayesian inference, models can be compared Bayes factors which is the ratio of two normalizing constants.
- Wealth of contribution: Chen, Shao, and Ibrahim 2000, chapter 5, Marin and Robert 2009, Friel and Wyse 2012, Ardia et al. 2012, Dutta, Ghosh, et al. 2013, Knuth et al. 2015, Zhou, Johansen, and Aston 2015...
- Few theoretical guarantees are available for these algorithms.
- Assumption U is a continuously differentiable convex function, $\min U = 0$.

Multistage sampling

- Idea: decompose the original problem in a sequence of problems which are easier to solve.
- Multistage sampling method Gelman and Meng 1998, Section 3.3,

$$\frac{\mathcal{Z}}{\mathcal{Z}_0} = \prod_{i=0}^{M-1} \frac{\mathcal{Z}_{i+1}}{\mathcal{Z}_i} ,$$

where

1. $M \in \mathbb{N}^\star$ is the number of stages,
2. $\mathcal{Z}_0$ is the initial normalizing constant (should be easy to compute)
3. $\mathcal{Z}_{i+1}/\mathcal{Z}_i$ are the ratios of normalisations constants (that should also be easy to estimate).

A Gaussian annealing algorithm

- $M \in \mathbb{N}^\star$ number of stages.
- Let $\{\sigma_i^2\}_{i=0}^M$ be an increasing sequence of positive numbers and set $\sigma_M^2 = +\infty$.
- Consider the sequence of functions $\{U_i\}_{i=0}^M$ defined for all $i \in \{0, \ldots, M\}$ and $x \in \mathbb{R}^d$ by

$$U_i(x) = \frac{\|x\|^2}{2\sigma_i^2} + U(x) \ ,$$

  with the convention $1/\infty = 0$.
- Note that $U_M = U$, since $\sigma_M = +\infty$.
- If σ_0 is small enough, then $U_0(x) \approx \|x\|^2 /(2\sigma_0)$.

A Gaussian annealing algorithm

- Define sequence of probability densities $\{\pi_i\}_{i=0}^M$ for $i \in \{0, \ldots, M\}$ and $x \in \mathbb{R}^d$ by
$$\pi_i(x) = \mathcal{Z}_i^{-1} \mathrm{e}^{-U_i(x)} \; , \qquad \mathcal{Z}_i = \int_{\mathbb{R}^d} \mathrm{e}^{-U_i(y)} \mathrm{d}y \; .$$

- It defines $(Z_i)_{i=1}^M$ in the decomposition
$$\frac{\mathcal{Z}}{\mathcal{Z}_0} = \prod_{i=0}^{M-1} \frac{\mathcal{Z}_{i+1}}{\mathcal{Z}_i} \; ,$$

- For $i \in \{0, \ldots, M-1\}$, we get
$$\frac{\mathcal{Z}_{i+1}}{\mathcal{Z}_i} = \int_{\mathbb{R}^d} g_i(x) \pi_i(x) \mathrm{d}x = \pi_i(g_i) \; ,$$
where $g_i : \mathbb{R}^d \to \mathbb{R}_+$ is defined for any $x \in \mathbb{R}^d$ by
$$g_i(x) = \exp\left(a_i \left\| x \right\|^2 \right) \; , \qquad a_i = \frac{1}{2}\left(\frac{1}{\sigma_i^2} - \frac{1}{\sigma_{i+1}^2} \right) \; .$$

Multistage methods

- Multistage sampling type algorithms are widely used and known under different names: multistage sampling Valleau and Card 1972, (extended) bridge sampling Gelman and Meng 1998, annealed importance sampling (AIS) Neal 2001, thermodynamic integration Oates, Papamarkou, and Girolami 2016, power posterior Behrens, Friel, and Hurn 2012.
- For the stability and accuracy of the method, the choice of the parameters (in our case $\{\sigma_i^2\}_{i=0}^{M-1}$) is crucial and is known to be difficult.
- Indeed, the issue has been pointed out in several articles under the names of tuning tempered transitions Behrens, Friel, and Hurn 2012, temperature placement Friel, Hurn, and Wyse 2014, annealing sequence Beskos et al. 2014, Sections 3.2.1, 4.1, temperature ladder Oates, Papamarkou, and Girolami 2016, Section 3.3.2, effects of grid size Dutta, Ghosh, et al. 2013, cooling schedule Cousins and Vempala 2015.
- In Brosse, Durmus, and Moulines 2018, we explicitly define the sequence $\{\sigma_i^2\}_{i=0}^{M-1}$.

Multistage Langevin

- Compute for all $i \in \{1, \ldots, M-1\}$,

$$\frac{\mathcal{Z}_{i+1}}{\mathcal{Z}_i} = \int_{\mathbb{R}^d} g_i(x)\pi_i(x)\mathrm{d}x = \pi_i(g_i) \ .$$

- The quantity $\pi_i(g_i)$ is estimated by the Unadjusted Langevin Algorithm (ULA) targeting π_i.
- For all $i \in \{1, \ldots, M\}$, consider

$$X_{i,k+1} = X_{i,k} - \gamma_i \nabla U_i(X_{i,k}) + \sqrt{2\gamma_i} Z_{i,k+1} \ , \quad X_{i,0} = 0 \ .$$

- For $i \in \{0, \ldots, M-1\}$, consider the following estimator of $\mathcal{Z}_{i+1}/\mathcal{Z}_i$,

$$\hat{\pi}_i(g_i) = \frac{1}{n_i} \sum_{k=N_i+1}^{N_i+n_i} g_i(X_{i,k}) \ ,$$

where $n_i \geq 1$ is the sample size and $N_i \geq 0$ the burn-in period.

ULA algorithm

- We want to compute for all $i \in \{1, \ldots, M-1\}$,

$$\frac{\mathcal{Z}_{i+1}}{\mathcal{Z}_i} = \int_{\mathbb{R}^d} g_i(x)\pi_i(x)\mathrm{d}x = \pi_i(g_i) \;,$$

- For $i \in \{0, \ldots, M-1\}$, consider the following estimator of $\mathcal{Z}_{i+1}/\mathcal{Z}_i$,

$$\hat{\pi}_i(g_i) = \frac{1}{n_i} \sum_{k=N_i+1}^{N_i+n_i} g_i(X_{i,k}) \;,$$

  where $n_i \geq 1$ is the sample size and $N_i \geq 0$ the burn-in period.

- $\hat{\mathcal{Z}}$ the following estimator of $\mathcal{Z}$,

$$\hat{\mathcal{Z}} = (2\pi\sigma_0^2)^{d/2}(1+\sigma_0^2 m)^{-d/2} \left\{ \prod_{i=0}^{M-1} \hat{\pi}_i(g_i) \right\} \;,$$

Theoretical analysis

- Denote by $\mathcal{S}$ the set of simulation parameters,

$$\mathcal{S} = \left\{M, \{\sigma_i^2\}_{i=0}^{M-1}, \{\gamma_i\}_{i=0}^{M-1}, \{n_i\}_{i=0}^{M-1}, \{N_i\}_{i=0}^{M-1}\right\} \ .$$

- $\hat{\mathcal{Z}}$ the following estimator of $\mathcal{Z}$,

$$\hat{\mathcal{Z}} = (2\pi\sigma_0^2)^{d/2}(1 + \sigma_0^2 m)^{-d/2} \left\{\prod_{i=0}^{M-1} \hat{\pi}_i(g_i)\right\} \ .$$

- cost of the algorithm: cost $= \sum_{i=0}^{M-1}\{N_i + n_i\}$.

Theorem 5 (Brosse, Durmus, and Moulines 2018)

Let $\mu, \epsilon \in (0,1)$. There exists an explicit choice of the simulation parameters $\mathcal{S}$ such that the estimator $\hat{\mathcal{Z}}$ satisfies

$$\mathbb{P}\left(\left|\hat{\mathcal{Z}}/\mathcal{Z} - 1\right| > \epsilon\right) \leq \mu \ .$$

Moreover, the cost of the algorithm is polynomial in the dimension d, ϵ^{-1} and η^{-1}.

Numerical experiments

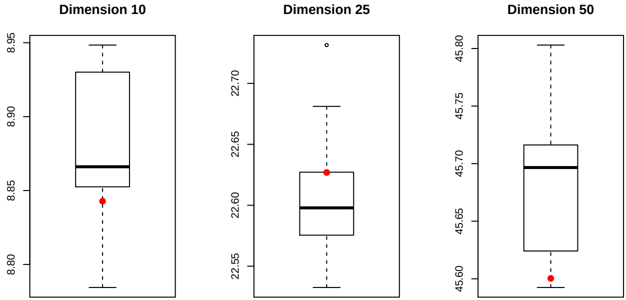

Figure : Boxplots of the logarithm of the normalizing constants of a multivariate Gaussian distribution in dimension $d \in \{10, 25, 50\}$.

Numerical experiements II

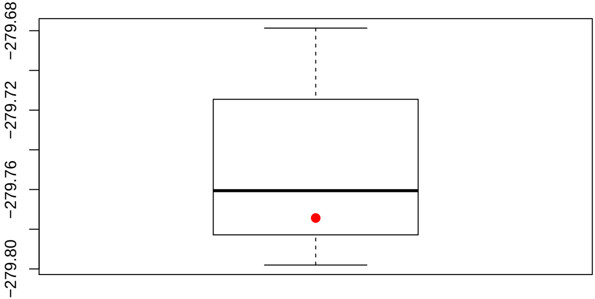

Figure : Boxplot of the log evidence for a mixture of 4 Gaussian distributions in dimension 2.

Bibliography I

Ardia, David, Nalan Baştürk, Lennart Hoogerheide, and Herman K Van Dijk (2012). "A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood". In: *Computational Statistics & Data Analysis* 56.11, pp. 3398–3414.

Behrens, Gundula, Nial Friel, and Merrilee Hurn (2012). "Tuning tempered transitions". In: *Statistics and Computing* 22.1, pp. 65–78.

Beskos, Alexandros, Dan O. Crisan, Ajay Jasra, and Nick Whiteley (2014). "Error bounds and normalising constants for sequential Monte Carlo samplers in high dimensions". In: *Adv. in Appl. Probab.* 46.1, pp. 279–306.

Brosse, N., A. Durmus, and É; Moulines (2018). "Normalizing constants of log-concave densities". In: *Electron. J. Stat.* 12.1, pp. 851–889. ISSN: 1935-7524.

Chen, MH, QM Shao, and JG Ibrahim (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York.

Cheng, X. and P. Bartlett (2017). "Convergence of Langevin MCMC in KL-divergence". In: *arXiv preprint arXiv:1705.09048*.

Cousins, Benjamin and Santosh Vempala (2015). "Bypassing KLS: Gaussian Cooling and an O*(n3) Volume Algorithm". In: *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*. ACM, pp. 539–548.

Dalalyan, A. (2014). *Theoretical guarantees for approximate sampling from a smooth and log-concave density*. Submitted 1412.7392. arXiv, pp. 1–30.

Bibliography II

Dutta, Ritabrata, Jayanta K Ghosh, et al. (2013). “Bayes model selection with path sampling: Factor models and other examples”. In: *Statistical Science* 28.1, pp. 95–115.

Friel, Nial, Merrilee Hurn, and Jason Wyse (2014). “Improving power posterior estimation of statistical evidence”. In: *Statistics and Computing* 24.5, pp. 709–723. ISSN: 1573-1375.

Friel, Nial and Jason Wyse (2012). “Estimating the evidence–a review”. In: *Statistica Neerlandica* 66.3, pp. 288–308.

Gelman, Andrew and Xiao-Li Meng (1998). “Simulating normalizing constants: From importance sampling to bridge sampling to path sampling”. In: *Statistical science*, pp. 163–185.

Jordan, R., D. Kinderlehrer, and F. Otto (1998). “The variational formulation of the Fokker-Planck equation”. In: *SIAM journal on mathematical analysis* 29.1, pp. 1–17.

Knuth, Kevin H., Michael Habeck, Nabin K. Malakar, Asim M. Mubeen, and Ben Placek (2015). “Bayesian evidence and model selection”. In: *Digital Signal Processing* 47. Special Issue in Honour of William J. (Bill) Fitzgerald, pp. 50–67. ISSN: 1051-2004.

Lamberton, D. and G. Pagès (2003). “Recursive computation of the invariant distribution of a diffusion: the case of a weakly mean reverting drift”. In: *Stoch. Dyn.* 3.4, pp. 435–451. ISSN: 0219-4937.

Marin, Jean-Michel and Christian P Robert (2009). “Importance sampling methods for Bayesian discrimination between embedded models”. In: *arXiv preprint arXiv:0910.2325*.

Martinet, B. (1970). “Régularisation d’inéquations variationnelles par approximations successives”. In: *Rev. Française Informat. Recherche Opérationnelle* 4, pp. 154–158.

Bibliography III

Neal, Radford M. (2001). “Annealed importance sampling”. In: *Statistics and Computing* 11.2, pp. 125–139. ISSN: 1573-1375.

Oates, Chris J., Theodore Papamarkou, and Mark Girolami (2016). “The Controlled Thermodynamic Integral for Bayesian Model Evidence Evaluation”. In: *Journal of the American Statistical Association* 111.514, pp. 634–645.

Rockafeller, T. (1976). “Monotone operators and the proximal point algorithm”. In: *SIAM J. Control Optimization* 14, pp. 877–898.

Talay, D. and L. Tubaro (1990). “Expansion of the global error for numerical schemes solving stochastic differential equations”. In: *Stochastic Anal. Appl.* 8.4, 483–509 (1991). ISSN: 0736-2994.

Valleau, J. P. and D. N. Card (1972). “Monte Carlo Estimation of the Free Energy by Multistage Sampling”. In: *The Journal of Chemical Physics* 57.12, pp. 5457–5462.

Welling, M. and Y. W. Teh (2011). “Bayesian Learning via Stochastic Gradient Langevin Dynamics”. In: *Proceedings of the International Conference on Machine Learning*, pp. 681–688.

Zhou, Yan, Adam M Johansen, and John AD Aston (2015). “Towards automatic model comparison: an adaptive sequential Monte Carlo approach”. In: *Journal of Computational and Graphical Statistics* just-accepted.