The Langevin MCMC: Theory and Methods

Nicolas Brosse, Alain Durmus, Eric Moulines, Marcelo Pereyra

Telecom ParisTech, Ecole Polytechnique, Edinburgh University

July 31, 2017

1 Motivation

2 Framework

3 Strongly log-concave distribution

4 Convex and Super-exponential densities

5 Non-smooth potentials

6 Conclusions

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...

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

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

.

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

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

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, ..).

A daunting problem ?

- For Gaussian prior (ridge regression), the potential U is smooth strongly convex.
- For Laplace prior (Lasso our fused Lasso) regression, the potential U 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)

1 Motivation

2 Framework

3 Strongly log-concave distribution

4 Convex and Super-exponential densities

5 Non-smooth potentials

6 Conclusions

Framework

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

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

  Implicitly, $d \gg 1$.

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

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

(Overdamped) Langevin diffusion

- Langevin SDE:

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

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

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

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

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

Discretized Langevin diffusion

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

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

  where

  - $(Z_k)_{k \geq 1}$ is i.i.d. $\mathcal{N}(0, \mathrm{I}_d)$
  - $(\gamma_k)_{k \geq 1}$ is a sequence of stepsizes, which can either be held constant or be chosen to decrease to 0 at a certain rate.

- Closely related to the (stochastic) gradient descent algorithm.

Discretized Langevin diffusion: constant stepsize

- When the stepsize is held constant, *i.e.* $\gamma_k = \gamma$, then $(X_k)_{k \geq 1}$ is an homogeneous Markov chain with Markov kernel R_γ
- Under some appropriate conditions, this Markov chain is irreducible, positive recurrent $\rightsquigarrow$ unique invariant distribution π_γ which does not coincide with the target distribution π.
- Questions:
  - For a given precision $\epsilon > 0$, how should I choose the stepsize $\gamma > 0$ and the number of iterations n so that : $\|\delta_x R_\gamma^n - \pi\|_{\mathrm{TV}} \leq \epsilon$
  - Is there a way to choose the starting point x cleverly ?
  - Auxiliary question: quantify the distance between π_γ and π.

Discretized Langevin diffusion: decreasing stepsize

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

$$Q^p_\gamma = R_{\gamma_1} R_{\gamma_2} \dots R_{\gamma_p}$$

With this notation, $\mathbb{E}_x[f(X_p)] = \delta_x Q^p_\gamma f$.

- Questions:
  - Convergence : is there a way to choose the step sizes so that $\|\delta_x Q^p_\gamma - \pi\|_{\mathrm{TV}} \to 0$ and if yes, what is the optimal way of choosing the stepsizes ?...
  - Optimal choice of simulation parameters : What is the number of iterations required to reach a neighborhood of the target: $\|\delta_x Q^p_\gamma - \pi\|_{\mathrm{TV}} \leq \epsilon$ starting from a given point x
  - Should we use fixed or decreasing step sizes ?

1 Motivation

2 Framework

3 Strongly log-concave distribution

4 Convex and Super-exponential densities

5 Non-smooth potentials

6 Conclusions

Strongly convex potential

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

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

- Outline of the proof
  1. Control in Wasserstein distance of the laws of the Langevin diffusion and its discretized version.
  2. Relating Wassertein distance result to total variation.
- Key technique: (Synchronous and Reflection) coupling !

Wasserstein distance

Definition

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

$$W_2(\mu, \nu) = \inf_{(X,Y) \in \Pi(\mu,\nu)} \mathbb{E}^{1/2}\left[\|X - Y\|^2\right],$$

where $\Pi(\mu, \nu)$ is the set of coupling of μ, ν: $(X, Y) \in \Pi(\mu, \nu)$ if and only if $X \sim \mu$ and $Y \sim \nu$.

Wasserstein distance convergence

Theorem

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

$$W_2\left(\delta_x P_t, \delta_y P_t\right) \leq \mathrm{e}^{-mt} \|x - y\|$$

The contraction depends only on the strong convexity constant.

Synchronous Coupling

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

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

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

The product rule for semimartingales imply

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

Synchronous Coupling

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

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

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

Grömwall inequality:

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

Theorem

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

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

where

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

The stationary distribution π satisfies

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

The constant depends only linearly in the dimension d.

Elements of proof

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

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

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

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

Elements of proof

Key relation

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

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

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

We have

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

Grönwall inequality

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

Elements of proof

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

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

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

Contraction property of the discretization

Theorem

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

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

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

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

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

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

A coupling proof (I)

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

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

where

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

  - $\delta_x Q_\gamma^n$: law of the discretized diffusion
  - $\pi P_{\gamma_n} = \pi$, where $(P_t)_{t \geq 0}$ is the semi group of the diffusion
- Idea ! synchronous coupling between the diffusion and the interpolation of the Euler discretization.

A coupling proof (II)

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

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

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

For all $n \geq 0$,

$$W_2^2 \left(\delta_x P_{\Gamma_n}, \pi Q_\gamma^n\right) \leq \mathbb{E}[\|Y_{\Gamma_n} - \bar{Y}_{\Gamma_n}\|^2] ,$$

Explicit bound in Wasserstein distance for the Euler discretisation

Theorem

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

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

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

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

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

Results

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

$$W_2\left(\delta_{x_*} R_\gamma^p, \pi\right) \leq \epsilon \quad \text{in } p = \mathcal{O}(\sqrt{d}\epsilon^{-1}) \text{ iterations}$$

  where x_* is the unique maximum of π
- Decreasing step size with $\gamma_k = \gamma_1 k^{-\alpha}$, $\alpha \in (0,1)$,

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

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

From the Wasserstein distance to the TV

Theorem

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

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

Use reflection coupling (Lindvall and Rogers, 1986)

Hints of Proof I

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

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

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

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

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

Hints of Proof II

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

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

Using Itô's formula

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

and Grönwall's inequality implies

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

Hint of Proof III

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

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

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

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

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

From the Wasserstein distance to the TV (II)

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

Consequences:

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

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

Explicit bound in total variation

Theorem

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

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

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

Constant step sizes

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

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

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

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

1 Motivation

2 Framework

3 Strongly log-concave distribution

4 Convex and Super-exponential densities

5 Non-smooth potentials

6 Conclusions

Convergence of the Euler discretization

Assumption

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

- U is convex.

Results[1].

- If $\lim_{\gamma_k \to +\infty} \gamma_k = 0$, and $\sum_k \gamma_k = +\infty$ then
$$\lim_{p \to +\infty} \|\delta_x Q_\gamma^p - \pi\|_{\mathrm{TV}} = 0 \ .$$

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

[1] Durmus, Moulines, Annals of Applied Probability, 2016

Target precision ϵ: the convex case

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

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

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

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

Strongly convex outside a ball potential

- U is convex everywhere and strongly convex outside a ball, *i.e.* there exist $R \geq 0$ and $m > 0$, such that for all $x, y \in \mathbb{R}^d$, $\|x - y\| \geq R$,

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

- Eberle, 2015 established that the convergence in the Wasserstein distance does not depends on the dimension.
- Durmus, M. 2016 established that the convergence of the semi-group in TV to π does not depends on the dimension but just on $R \rightsquigarrow$ new bounds which scale nicely in the dimension.

Dependence on the dimension

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

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

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

How it works ?

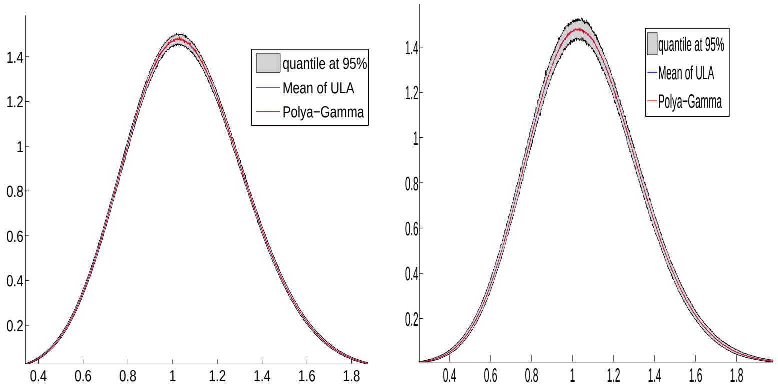


Figure: Empirical distribution comparison between the Polya-Gamma Gibbs Sampler and ULA. Left panel: constant step size $\gamma_k = \gamma_1$ for all $k \geq 1$; right panel: decreasing step size $\gamma_k = \gamma_1 k^{-1/2}$ for all $k \geq 1$

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

Table: Dimension of the data sets

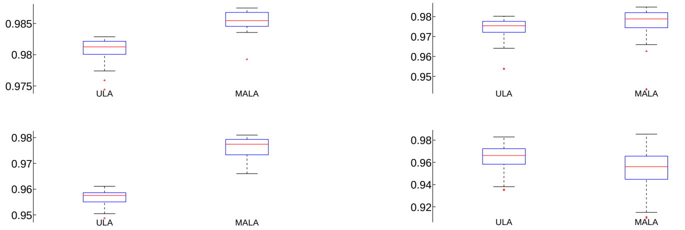


Figure: Marginal accuracy across all the dimensions. Upper left: German credit data set. Upper right: Australian credit data set. Lower left: Heart disease data set. Lower right: Musk data set

1 Motivation

2 Framework

3 Strongly log-concave distribution

4 Convex and Super-exponential densities

5 Non-smooth potentials

6 Conclusions

Non-smooth potentials

The target distribution has a density π with respect to the Lebesgue measure on $\mathbb{R}^d$ of the form $x \mapsto \mathrm{e}^{-U(x)} / \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)} \mathrm{d}y$ where $U = f + g$, with $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to (-\infty, +\infty]$ are two lower bounded, convex functions satisfying:

1. f is continuously differentiable and gradient Lipschitz with Lipschitz constant L_f, *i.e.* for all $x, y \in \mathbb{R}^d$

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\| \ .$$

2. g is lower semi-continuous and $\int_{\mathbb{R}^d} \mathrm{e}^{-g(y)} \mathrm{d}y \in (0, +\infty)$.

Moreau-Yosida regularization

- Let $\mathrm{h} : \mathbb{R}^d \to (-\infty, +\infty]$ be a l.s.c convex function and $\lambda > 0$. The λ-Moreau-Yosida envelope $\mathrm{h}^\lambda : \mathbb{R}^d \to \mathbb{R}$ and the proximal operator $\mathrm{prox}_\mathrm{h}^\lambda : \mathbb{R}^d \to \mathbb{R}^d$ associated with h are defined for all $x \in \mathbb{R}^d$ by

$$\mathrm{h}^\lambda(x) = \inf_{y \in \mathbb{R}^d} \left\{ \mathrm{h}(y) + (2\lambda)^{-1} \|x - y\|^2 \right\} \leq \mathrm{h}(x) \; .$$

- For every $x \in \mathbb{R}^d$, the minimum is achieved at a unique point, $\mathrm{prox}_\mathrm{h}^\lambda(x)$, which is characterized by the inclusion

$$x - \mathrm{prox}_\mathrm{h}^\lambda(x) \in \gamma \partial \mathrm{h}(\mathrm{prox}_\mathrm{h}^\lambda(x)) \; .$$

- The Moreau-Yosida envelope is a regularized version of g, which approximates g from below.

Properties of proximal operators

- As $\lambda \downarrow 0$, converges h^λ converges pointwise h, *i.e.* for all $x \in \mathbb{R}^d$,

$$\mathrm{h}^\lambda(x) \uparrow \mathrm{h}(x) \;, \quad \text{as } \lambda \downarrow 0 \;.$$

- The function h^λ is convex and continuously differentiable

$$\nabla \mathrm{h}^\lambda(x) = \lambda^{-1}(x - \mathrm{prox}_{\mathrm{h}}^\lambda(x)) \;.$$

- The proximal operator is a monotone operator, for all $x, y \in \mathbb{R}^d$,

$$\left\langle \mathrm{prox}_{\mathrm{h}}^\lambda(x) - \mathrm{prox}_{\mathrm{h}}^\lambda(y), x - y \right\rangle \geq 0 \;,$$

which implies that the Moreau-Yosida envelope is L-smooth: $\left\| \nabla \mathrm{h}^\lambda(x) - \nabla \mathrm{h}^\lambda(y) \right\| \leq \lambda^{-1} \left\| x - y \right\|$, for all $x, y \in \mathbb{R}^d$.

MY regularized potential

- If g is not differentiable, but the proximal operator associated with g is available, its λ-Moreau Yosida envelope g^λ can be considered.
- This leads to the approximation of the potential $U^\lambda : \mathbb{R}^d \to \mathbb{R}$ defined for all $x \in \mathbb{R}^d$ by

$$U^\lambda(x) = f(x) + g^\lambda(x) \; .$$

Theorem (Durmus, M., Pereira, 2016, SIAM J. Imaging Sciences)

Under (H), for all $\lambda > 0$, $0 < \int_{\mathbb{R}^d} \mathrm{e}^{-U^\lambda(y)} \mathrm{d}y < +\infty$.

Some approximation results

Theorem

Assume (H).

1. *Then,* $\lim_{\lambda \to 0} \|\pi^\lambda - \pi\|_{\mathrm{TV}} = 0$.
2. *Assume in addition that* g *is Lipschitz. Then for all* $\lambda > 0$,

$$\|\pi^\lambda - \pi\|_{\mathrm{TV}} \leq \lambda \|g\|_{\mathrm{Lip}}^2 \ .$$

The MYULA algorithm-I

Given a regularization parameter $\lambda > 0$ and a sequence of stepsizes $\{\gamma_k,\ k \in \mathbb{N}^*\}$, the algorithm produces the Markov chain $\{X_k^{\mathrm{M}},\ k \in \mathbb{N}\}$: for all $k \geq 0$,

$$X_{k+1}^{\mathrm{M}} = X_k^{\mathrm{M}} - \gamma_{k+1} \left\{ \nabla f(X_k^{\mathrm{M}}) + \lambda^{-1}(X_k^{\mathrm{M}} - \mathrm{prox}_g^{\lambda}(X_k^{\mathrm{M}})) \right\} + \sqrt{2\gamma_{k+1}} Z_{k+1} \ ,$$

where $\{Z_k,\ k \in \mathbb{N}^*\}$ is a sequence of i.i.d. d-dimensional standard Gaussian random variables.

The MYULA algorithm-II

- The ULA target the smoothed distribution π^λ.
- To compute the expectation of a function $h : \mathbb{R}^d \to \mathbb{R}$ under π from $\{X_k^{\mathrm{M}} \; ; \; 0 \leq k \leq n\}$, an importance sampling step is used to correct the regularization.
- This step amounts to approximate $\int_{\mathbb{R}^d} h(x)\pi(x)\mathrm{d}x$ by the weighted sum

$$\mathrm{S}_n^h = \sum_{k=0}^{n} \omega_{k,n} h(X_k) \; , \; \text{with } \omega_{k,n} = \left\{ \sum_{k=0}^{n} \gamma_k \mathrm{e}^{\bar{g}^\lambda(X_k^{\mathrm{M}})} \right\}^{-1} \gamma_k \mathrm{e}^{\bar{g}^\lambda(X_k^{\mathrm{M}})} \; ,$$

where for all $x \in \mathbb{R}^d$

$$\bar{g}^\lambda(x) = g^\lambda(x) - g(x) = g(\mathrm{prox}_g^\lambda(x)) - g(x) + (2\lambda)^{-1} \left\| x - \mathrm{prox}_g^\lambda(x) \right\|^2 \; .$$

Image deconvolution

- Objective recover an original image $\boldsymbol{x} \in \mathbb{R}^n$ from a blurred and noisy observed image $\boldsymbol{y} \in \mathbb{R}^n$ related to $\boldsymbol{x}$ by the linear observation model $\boldsymbol{y} = H\boldsymbol{x} + \boldsymbol{w}$, where H is a linear operator representing the blur point spread function and $\boldsymbol{w}$ is a Gaussian vector with zero-mean and covariance matrix $\sigma^2 \boldsymbol{I}_n$.
- This inverse problem is usually ill-posed or ill-conditioned: exploits prior knowledge about $\boldsymbol{x}$.
- One of the most widely used image prior for deconvolution problems is the improper total-variation norm prior, $\pi(\boldsymbol{x}) \propto \exp\left(-\alpha \|\nabla_d \boldsymbol{x}\|_1\right)$, where ∇_d denotes the discrete gradient operator that computes the vertical and horizontal differences between neighbour pixels.

$$\pi(\boldsymbol{x}|\boldsymbol{y}) \propto \exp\left[-\|\boldsymbol{y} - H\boldsymbol{x}\|^2 / 2\sigma^2 - \alpha \|\nabla_d \boldsymbol{x}\|_1\right].$$

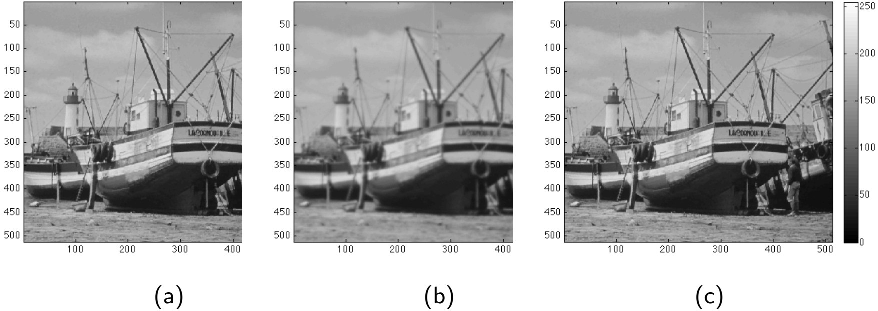
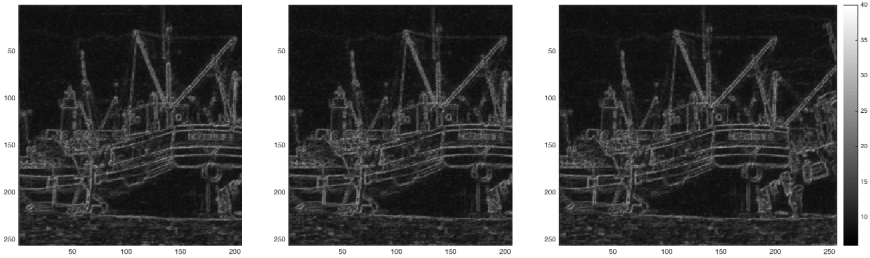


Figure: (a) Original Boat image (256×256 pixels), (b) Blurred image, (c) MAP estimate.

Credibility intervals



(a) (b) (c)

Figure: (a) Pixel-wise 90% credibility intervals computed with proximal MALA (computing time 35 hours), (b) Approximate intervals estimated with MYULA using $\lambda = 0.01$ (computing time 3.5 hours), (c) Approximate intervals estimated with MYULA using $\lambda = 0.1$ (computing time 20 minutes).

1 Motivation

2 Framework

3 Strongly log-concave distribution

4 Convex and Super-exponential densities

5 Non-smooth potentials

6 Conclusions

Conclusion

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

Our (published) work

1. Durmus, Alain; Moulines, Éric *Quantitative bounds of convergence for geometrically ergodic Markov chain in the Wasserstein distance with application to the Metropolis adjusted Langevin algorithm.* Stat. Comput. 25 (2015)
2. Durmus, Alain; Moulines, Éric, *Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm* Accepted for publication in Ann. Appl. Prob.
3. Durmus, Alain; Simsleki, Ümut; Moulines, Éric; Badeau, Roland, *Stochastic Gradient Richardson-Romberg Markov Chain Monte Carlo*, NIPS, 2016
4. Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo Brosse, N., Durmus A., Moulines E., Pereyra, M., COLT 2017 *Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau*, SIAM J. Imaging Sciences.
5. + more recent preprints (see Arxiv)