Langevin MCMC: Theory and Methods

Alain Durmus[1], Eric Moulines[2]

[1]ENS Paris-Saclay

[2]Ecole Polytechnique

Outline

1 The ULA algorithm for smooth logconcave densities

2 Non-smooth potentials

3 Logconcave densities with constrained domains

4 Deviation inequalities

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

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

- Assume that U is L-smooth, *i.e.* 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:

$$P_t(x, A) = \mathbb{P}(Y_t \in A | Y_0 = x) \ , \quad x \in \mathbb{R}^d, A \in \mathcal{B}(\mathbb{R}^d) \ .$$

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

Discretized Langevin diffusion

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

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

  where
  - $(Z_k)_{k \geq 1}$ is i.i.d. $\mathcal{N}(0, \mathrm{I}_d)$
  - $\gamma > 0$ is a stepsize
- Closely related to the gradient descent algorithm.

Discretized Langevin diffusion: constant stepsize

- $(X_k)_{k\geq 1}$ is an homogeneous Markov chain with Markov kernel R_γ
- Under some appropriate conditions, the Markov kernel R_γ is irreducible, positive recurrent $\leadsto$ unique invariant distribution π_γ
- Beware ! π_γ does not coincide with the target distribution π.

Outline

1 The ULA algorithm for smooth logconcave densities

2 Non-smooth potentials

3 Logconcave densities with constrained domains

4 Deviation inequalities

5 Normalizing constants of log-concave densities

Non-smooth potential

- Question : what to do if U is not C^1 but still convex ?
- Assume (w.l.o.g.)

$$\pi \propto \mathrm{e}^{-U} , \quad U = f + g ,$$

where f is smooth and g not.

- Applications :

1. constrained lasso and ridge regressions

$$\mathbb{1}_{\mathcal{K}}(x) = \begin{cases} +\infty & \text{if } x \notin \mathcal{K}, \\ 0 & \text{if } x \in \mathcal{K} . \end{cases} ,$$

2. LASSO, fused-LASSO models...

Non-smooth potential

- To apply EM discretization, Durmus et al. (2018) suggests to regularize g in such a way that
  1. the convexity of U is preserved
  2. the regularisation of U is continuously differentiable and gradient Lipschitz
  3. the resulting approximation is close to π (e.g. in total variation norm)

Moreau-Yosida regularization

- Assume that $g : \mathbb{R}^d \to (-\infty, +\infty]$ is a l.s.c convex function and let $\lambda > 0$.
- The λ-Moreau-Yosida envelope $g^\lambda : \mathbb{R}^d \to \mathbb{R}$ is defined for all $x \in \mathbb{R}^d$ by

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

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

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

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

Properties of proximal operators

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

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

- The function g^λ is convex and continuously differentiable

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

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

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

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

Moreau-Yosida regularization

- 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

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

- Does it make some sense to use U^λ for targetting $\pi \propto \mathrm{e}^{-U}$?

Assumptions

H

- $\pi \propto \mathrm{e}^{-U}$, $U = f + g$
- $f : \mathbb{R}^d \to \mathbb{R}$ *and* $g : \mathbb{R}^d \to (-\infty, +\infty]$ *are convex*
- f *is continuously differentiable and gradient Lipschitz with Lipschitz constant* L_f, *i.e. for all* $x, y \in \mathbb{R}^d$
$$\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\| \ .$$
- g *is lower semi-continuous and* $\int_{\mathbb{R}^d} \mathrm{e}^{-g(y)} \mathrm{d}y \in (0, +\infty)$ *(other conditions exist...).*

Properties of proximal operators and consequences

- The function g^λ is convex and continuously differentiable.
- g^λ is gradient-Lipschitz: for all $x, y \in \mathbb{R}^d$,

$$\left\|\nabla g^\lambda(x) - \nabla g^\lambda(y)\right\| \leq \lambda^{-1} \|x - y\|$$

- Consequence: The function U^λ is convex, continuously differentiable and gradient-Lipschitz.

Properties of proximal operators and consequences

- Idea: use ULA with U^{λ} instead of U to target π ?

Theorem (Durmus et al. (2018))

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

- U^{λ} defines a regularized distribution π^{λ}

$$\pi^{\lambda} \propto \mathrm{e}^{-U^{\lambda}} , \quad U^{\lambda}(x) = f(x) + g^{\lambda}(x) .$$

Some approximation results

- U^λ defines a regularized version of π^λ

$$\pi^\lambda \propto \mathrm{e}^{-U^\lambda} \; , \quad U^\lambda(x) = f(x) + g^\lambda(x) \; .$$

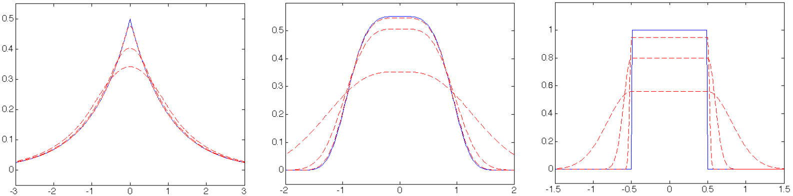
- Question: Since U^λ is an approximation of U, is π^λ an approximation of π

Theorem (Durmus et al. (2018))

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

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

Moreau-Yoshida approximations



$p(x) \propto \exp\left(-|x|\right)$ $\quad$ $p(x) \propto \exp\left(-x^4\right)$ $\quad$ $p(x) \propto \mathbf{1}_{[-0.5,0.5]}(x)$

Figure: True densities (solid blue) and approximations (dashed red).

The MYULA algorithm

- Main idea: Target $\pi^\lambda \propto \mathrm{e}^{-U^\lambda}$ instead of $\pi \propto \mathrm{e}^{-U}$ using ULA.
- Reasons:
  - π^λ is a "good" approximation of π provided that the regularization parameter λ is small enough
  - U^λ is continuously differentiable, gradient Lipschitz and convex
- Given a regularization parameter $\lambda > 0$ and a stepsize $\gamma > 0$, the ULA applied to π^λ yields

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

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

Microscopy dataset

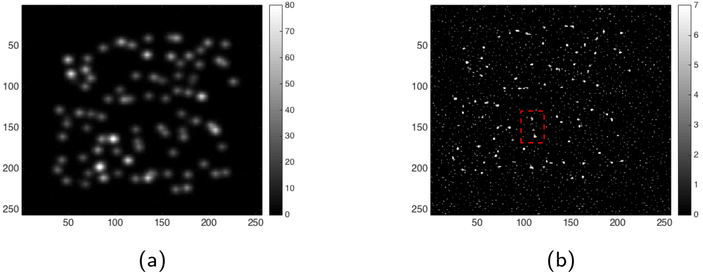


Figure: Microscopy dataset, field of size 4μm × 4μm containing 100 molecules, (a) Original Observation, (b) MAP

The goal of image deconvolution is to recover a high-resolution image $x \in \mathbb{R}^n$ from a blurred and noisy observation $y = Hx + w$, where H is a circulant blurring matrix and $w \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$. This inverse problem is ill-conditioned, a difficulty that Bayesian image deconvolution methods address by exploiting prior knowledge.

Microscopy dataset

- The goal is to recover the image $x \in \mathbb{R}^n$ from an incomplete and noisy set of Fourier measurements $y = AFx + w$, where F is the discrete Fourier transform operator, A is a tomographic sampling mask, and $w \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$.
- This inverse problem is ill-posed, resulting in significant uncertainty about the true value of x.
- Idea Use a total-variation prior promoting piecewise regular images. The resulting posterior $p(x|y)$ is

$$\pi(x) \propto \exp\left[-(\|y - AFx\|^2/2\sigma^2 + \beta\|x\|_1)\right].$$

with fixed hyper-parameters $\sigma > 0$ and $\beta > 0$.

- This density is log-concave and MAP estimation can be performed efficiently by proximal convex optimisation (here we use the ADMM algorithm SALSA).

Credible sets

- point estimators such as $\hat{x}_{MAP}$ deliver accurate results but do not provide information about the posterior uncertainty of x or $\varphi(x)$ where φ is a function.
- Given the uncertainty that is inherent to ill-posed and ill-conditioned inverse problems, it would be highly desirable to complement point estimators with posterior credibility sets that indicate the region of the parameter space where most of the posterior probability mass of $\varphi(x)$ lies.
- This is formalised in the Bayesian decision theory framework by computing credible regions. A set C_α is a level α credible region if

$$\Pi\left(\varphi(x) \in C_\alpha | y\right) = 1 - \alpha.$$

- Credible sets are random sets, since they are based on the posterior distribution.
- It is relatively easy to obtain credible regions based on simulated samples from the posterior, as obtained from an MCMC sample for example.

Highest Posterior Density regions

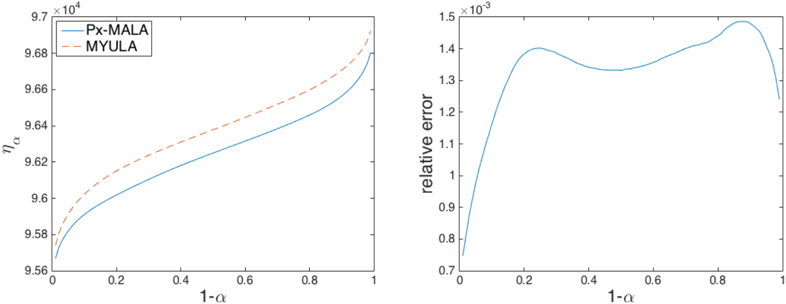
.

- The definition of credible sets offers too much freedom in the choice of the region C_α. Given a level $\alpha > 0$, many sets will be credible sets, just like confidence regions can be chosen in many different ways.
- Among all possible regions, the so-called highest posterior density (HPD) region has minimum volume

$$C_\alpha^*(y) = \{z : \pi_\varphi(z|y) \leq \eta_\alpha\}$$

where $\pi_\varphi(\cdot|y)$ is the posterior distribution of $\varphi(x)$ and $\eta_\alpha \in \mathbb{R}$ chosen such that $\int_{C_\alpha^*} \pi_\varphi(z|y) \mathrm{d}z = 1 - \alpha$ holds.

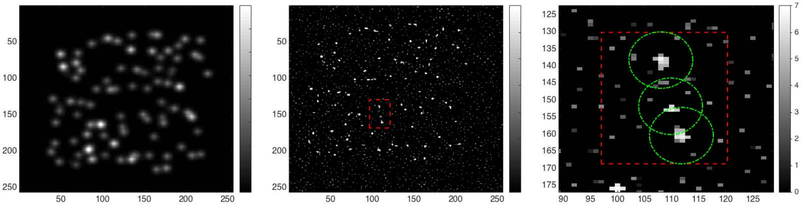
Comparison with PMALA



(a) (b)

Figure: Microscopy experiment: (a) HDP region thresholds η_α for MYULA (2×10^6 iterations $\lambda = 1, \gamma = 0.6$) and PMALA ($2 \times 10^7$ iterations), (b) relative approximation error of MYULA.

Sparse image deblurring



y $\hat{x}_{MAP}$ uncertainty estimates

Figure: Live-cell microscopy data Zhu et al. (2012). Uncertainty analysis $(\pm 78nm \times \pm 125nm)$

Computing time 4 minutes. $M = 10^5$ iterations. Estimation error 0.2%..

Outline

1. The ULA algorithm for smooth logconcave densities
2. Non-smooth potentials
3. **Logconcave densities with constrained domains**
4. Deviation inequalities
5. Normalizing constants of log-concave densities

Densities with convex support

We are interested now to log-concave target distribution π:

- π has bounded support, $\log \pi = -\infty$ outside some domain: $\mathrm{Supp}(\pi) = \mathcal{K}$, where $\mathcal{K} \subset \mathbb{R}^d$ is a convex body ?

$$\pi \propto \mathrm{e}^{-U} \;, \quad U = f + \mathbb{1}_{\mathcal{K}} \;, \quad \mathbb{1}_{\mathcal{K}}(x) = \begin{cases} +\infty & \text{if } x \notin \mathcal{K}, \\ 0 & \text{if } x \in \mathcal{K} \;. \end{cases} \;,$$

where f is smooth.

- The Moreau-Yosida envelope of $\mathbb{1}_{\mathcal{K}}$ is given for the regularization parameter $\lambda > 0$ by

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

Previous works

- Previous work for the Metropolis algorithm and the hit-and-run : Applegate, Kannan, Dyer, Frieze, Polson, Simonovits, Lovász, Vempala...
- Our approach more in the spirit of Bubeck, Eldan, and Lehec Bubeck et al. (2015) with the Projected Langevin Monte Carlo

Main results - Assumptions

H

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

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

.

H

There exist $r, R > 0$, $r \leq R$, such that $\mathrm{B}(0, r) \subset \mathcal{K} \subset \mathrm{B}(0, R)$.

Main results - Statement

Theorem (Brosse et al. (2017))

Assume **H**2 *and* **H**3*. For all* $\varepsilon > 0$ *and* $x \in \mathbb{R}^d$*, there exist (explicit)* $\lambda > 0$ *and* $\gamma > 0$ *such that,*

$$\|\delta_x R^n_{\gamma,\lambda} - \pi\|_{\mathrm{TV}} \leq \varepsilon \quad \textit{for} \quad n = \tilde{\Omega}(d^5) \; ,$$

where $R_{\gamma,\lambda}$ *is the Markov kernel associated to* $(X_k^\lambda)_{k\geq 0}$*.*

- Similar bounds hold for the Wasserstein distance.

Comparison with existing results

- Lovász and Vempala (2007) shows that the complexity of the RWM and the hit-and-run algorithm are of order d^4.
  However, this result requires that $\mathcal{K}$ is *well-rounded* (which is hard to check)
- Bubeck et al. (2015) studies the complexity of the Projected Langevin Monte Carlo algorithm (PLMC):

$$X_{k+1} = \mathrm{proj}_{\mathcal{K}} \left(X_k - \gamma \nabla f(X_k) \right) \ .$$

Note that contrary to MYULA, the iterates of PLMC always belongs to $\mathcal{K}$.

- Under similar assumptions, they get explicit bounds in total variation for PLMC:

	$d \to +\infty$	$\varepsilon \to 0$	$R \to +\infty$	$r \to 0$
Bubeck et al. (2015) π uniform on $\mathcal{K}$	$\tilde{O}(d^7)$	$\tilde{O}(\varepsilon^{-8})$	$\tilde{O}(R^6)$	$\tilde{O}(r^{-6})$
Bubeck et al. (2015) π log concave	$\tilde{O}(d^{12})$	$\tilde{O}(\varepsilon^{-12})$	$\tilde{O}(R^{18})$	$\tilde{O}(r^{-18})$
MYULA	$\tilde{O}(d^5)$	$\tilde{O}(\varepsilon^{-6})$	$\tilde{O}(R^4)$	$\tilde{O}(r^{-4})$

Table: Complexity $\|\delta_{x^\star} R^n_{\gamma,\lambda} - \pi\|_{\mathrm{TV}} \le \varepsilon$

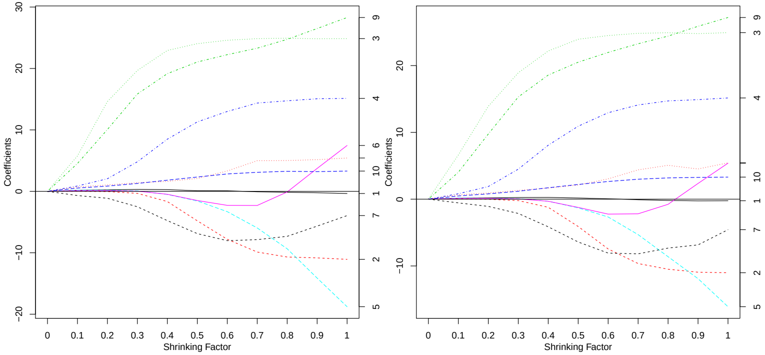
Application to regression with ℓ_1 constraints

1. For all $\mathrm{s} > 0$, consider the density $\pi^{\mathrm{s}} \propto \mathrm{e}^{-U^{\mathrm{s}}}$ where

$$U^{\mathrm{s}}(\boldsymbol{\beta}) = \exp\left(-\frac{\|Y - X\boldsymbol{\beta}\|^2}{2\sigma^2} - \iota_{\mathsf{K}^{\mathrm{s}}}(\boldsymbol{\beta})\right) \;, \mathsf{K}^{\mathrm{s}} = \{\boldsymbol{\beta} \in \mathbb{R}^d \, ; \, \|\boldsymbol{\beta}\|_1 \leq \mathrm{s}\} \; .$$

2. Dual problem of LASSO regression in optimization.
3. We compute for all $i \in \{1, \cdots, d\}$, the median of $\boldsymbol{\beta}_i$ for different values of s on the diabetes data set $(n = 442, d = 10)$.
4. Compute the LASSO regularization paths.

Application to regression with ℓ_1 constraints

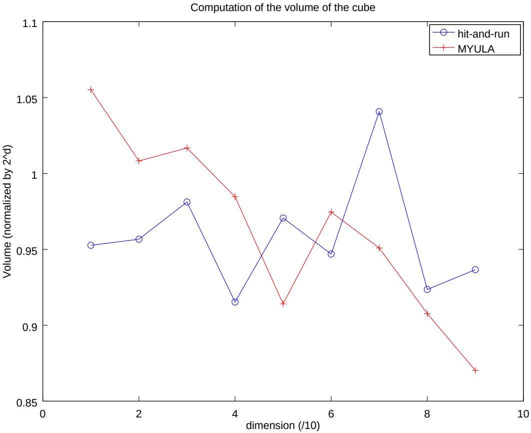


(a)

(b)

Figure: Lasso paths for (a) MYULA , (b) Wall HMC (Neal 2010)

Volume computation



Outline

1 The ULA algorithm for smooth logconcave densities

2 Non-smooth potentials

3 Logconcave densities with constrained domains

4 Deviation inequalities

5 Normalizing constants of log-concave densities

Bounds for functionals

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

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

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

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

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

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

MSE: Bias term

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

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

- We first bound the bias. For all $k \in \{N+1,\ldots,N+n\}$, let ξ_k be the optimal transference plan between $\delta_x Q_\gamma^k$ and π for W_2. By Jensen's inequality and using that f is Lipschitz:

$$\begin{aligned}\left\{\mathbb{E}_x[\hat{\pi}_n^N(f)] - \pi(f)\right\}^2 &= \left(\sum_{k=N+1}^{N+n} \omega_{k,n} \int_{\mathbb{R}^d\times\mathbb{R}^d} \{f(z)-f(y)\}\xi_k(\mathrm{d}z,\mathrm{d}y)\right)^2 \\ &\le \|f\|_{\mathrm{Lip}}^2 \sum_{k=N+1}^{N+n} \omega_{k,n} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|z-y\|^2 \xi_k(\mathrm{d}z,\mathrm{d}y) .\end{aligned}$$

Lipshitz

Theorem

Assume U *is gradient Lipschitz and strongly convex. Then,*

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

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

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

Bound on the Wasserstein distance

Theorem

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

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

where

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

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

$$u_n^{(2)}(\gamma) = L^2 \sum_{i=1}^n \gamma_i^2 C(d, \gamma_i) \prod_{k=i+1}^n (1 - \kappa\gamma_k/2) .$$

Gaussian Poincare inequality

- Our main tool is the Gaussian Poincaré inequality which states that if $Z = (Z_1, \dots, Z_d)$ is a Gaussian vector with identity covariance matrix, then for all Lipschitz function $g : \mathbb{R}^d \to \mathbb{R}$

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

- For all $y \in \mathbb{R}^d$ and $\gamma > 0$, the Gaussian Poincaré inequality can be applied to

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

  noticing that $R_\gamma(y, \cdot)$ is the Gaussian distribution with mean $y - \gamma\nabla U(y)$ and covariance matrix $2\gamma\,\mathrm{I}_d$

- For all Lipschitz function $g : \mathbb{R}^d \to \mathbb{R}$

$$R_\gamma \left\{g(\cdot) - R_\gamma g(y)\right\}^2 (y) \leq 2\gamma \left\|g\right\|^2_{\mathrm{Lip}} \,.$$

Martingale decomposition

- Idea: Decompose $\hat{\pi}_n^N(f) - \mathbb{E}_x[\hat{\pi}_n^N(f)]$ as the sum of martingale increments, w.r.t. $(\mathcal{G}_n)_{n \geq 0}$, the natural filtration of the EM approximation $(X_n)_{n \geq 0}$,

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

- Since $\hat{\pi}_n^N(f)$ is an additive functional, the martingale increment $\mathbb{E}_x^{\mathcal{G}_{k+1}}\left[\hat{\pi}_n^N(f)\right] - \mathbb{E}_x^{\mathcal{G}_k}\left[\hat{\pi}_n^N(f)\right]$ has a (reasonably) simple expression !

Martingale decomposition

- For $k = N+n-1, \dots, N+1$, define backward in time the function

$$\Phi^N_{n,k} : x_k \mapsto \omega^N_{k,n} f(x_k) + R_{\gamma_{k+1}} \Phi^N_{n,k+1}(x_k) \ ,$$

where $\Phi^N_{n,N+n} : x_{N+n} \mapsto \Phi^N_{n,N+n}(x_{N+n}) = \omega_{N+n,n} f(x_{N+n})$.

- For $k \in \{N, \dots, N+n-1\}$, by the Markov property,

$$\Phi^N_{n,k+1}(X_{k+1}) - R_{\gamma_{k+1}} \Phi^N_{n,k+1}(X_k) = \mathbb{E}^{\mathcal{G}_{k+1}}_x \left[\hat{\pi}^N_n(f) \right] - \mathbb{E}^{\mathcal{G}_k}_x \left[\hat{\pi}^N_n(f) \right] \ ,$$

and $\Psi^N_n(X_N) = \mathbb{E}^{\mathcal{G}_N}_x \left[\hat{\pi}^N_n(f) \right]$.

- The variance may be expressed

$$\mathrm{Var}_x \left\{ \hat{\pi}^N_n(f) \right\} = \sum_{k=N}^{N+n-1} \mathbb{E}_x \left[R_{\gamma_{k+1}} \left\{ \Phi^N_{n,k+1}(\cdot) - R_{\gamma_{k+1}} \Phi^N_{n,k+1}(X_k) \right\}^2 (X_k) \right] + \mathrm{Var}_x \left\{ \Psi^N_n(X_N) \right\} \ .$$

Lipshitz contraction

Theorem

Assume U is gradient Lipschitz and strongly convex. For all Lipschitz functions $f : \mathbb{R}^d \to \mathbb{R}$ and $\ell \geq n \geq 1$, $Q_\gamma^{n,\ell} f$ is a Lipschitz function with

$$\|Q_\gamma^{n,\ell} f\|_{\mathrm{Lip}} \leq \prod_{k=n}^{\ell} (1 - \kappa\gamma_k)^{1/2} \|f\|_{\mathrm{Lip}} .$$

Key Lemma

Lemma

Assume that U *is gradient Lipschitz and strongly convex. Let* $(\gamma_k)_{k\geq 1}$ *be a nonincreasing sequence with* $\gamma_1 \leq 2/(m+L)$. *Let* $N \geq 0$ *and* $n \geq 1$. *Then for all* $y \in \mathbb{R}^d$, *Lipschitz function* f *and* $k \in \{N, \ldots, N+n-1\}$,

$$R_{\gamma_{k+1}} \left\{ \Phi^N_{n,k+1}(\cdot) - R_{\gamma_{k+1}} \Phi^N_{n,k+1}(y) \right\}^2 (y) \leq 8\gamma_{k+1} \left\| f \right\|^2_{\mathrm{Lip}} \left(\kappa \Gamma_{N+2,N+n+1} \right)^{-2} .$$

Mean-Square Error (wrapping-up)

Theorem

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

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

where

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

The upper bound is independent of the dimension and allow to construct honest confidence bounds.

	Bound for the MSE
$\alpha = 0$	$\gamma_1^2 + (\gamma_1 n)^{-1} \exp(-\kappa \gamma_1 N/2)$
$\alpha \in (0, 1/3)$	$\gamma_1^2 n^{-2\alpha} + (\gamma_1 n^{1-\alpha})^{-1} \exp(-\kappa \gamma_1 N^{1-\alpha}/(2(1-\alpha)))$
$\alpha = 1/3$	$\gamma_1^2 \log(n) n^{-2/3} + (\gamma_1 n^{2/3})^{-1} \exp(-\kappa \gamma_1 N^{1/2}/4)$
$\alpha \in (1/3, 1)$	$n^{\alpha-1} \left\{\gamma_1^2 + \gamma_1^{-1} \exp(-\kappa \gamma_1 N^{1-\alpha}/(2(1-\alpha)))\right\}$
$\alpha = 1$	$\log(n)^{-1} \left\{\gamma_1^2 + \gamma_1^{-1} N^{-\gamma_1 \kappa/2}\right\}$

Table: Bound for the MSE for $\gamma_k = \gamma_1 k^{-\alpha}$ for fixed γ_1 and N with more regularity on U

	Optimal choice of γ_1	Bound for the MSE
$\alpha = 0$	$n^{-1/3}$	$n^{-2/3}$
$\alpha \in (0, 1/2)$	$n^{\alpha-1/3}$	$n^{-2/3}$
$\alpha = 1/2$	$(\log(n))^{-1/3}$	$\log^{1/3}(n)n^{-2/3}$
$\alpha \in (1/2, 1)$	$1/(m+L)$	$n^{1-\alpha}$
$\alpha = 1$	$1/(m+L)$	$\log(n)$

Table: Bound for the MSE for $\gamma_k = \gamma_1 k^{-\alpha}$ for fixed n with more regularity on U

Outline

1 The ULA algorithm for smooth logconcave densities

2 Non-smooth potentials

3 Logconcave densities with constrained domains

4 Deviation inequalities

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.
- 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

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

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

$$\pi_i(x) = \mathcal{Z}_i^{-1} \mathrm{e}^{-Ui(x)} \;, \qquad \mathcal{Z}_i = \int_{\mathbb{R}^d} \mathrm{e}^{-Ui(y)} \mathrm{d}y \;.$$

- 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

$$g_i(x) = \exp\left(a_i \|x\|^2\right) \;, \qquad a_i = \frac{1}{2}\left(\frac{1}{\sigma_i^2} - \frac{1}{\sigma_{i+1}^2}\right) \;.$$

Multistage methods

- Multistage sampling type algorithms are widely used and known under different names: multistage sampling, (extended) bridge sampling, annealed importance sampling (AIS), thermodynamic integration, power posterior. 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, temperature placement, annealing sequence, temperature ladder, effects of grid size, cooling schedule. In **????**, 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

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

- 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 $i \in \{1, \dots, M-1\}$,

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

- For $i \in \{0, \dots, M-1\}$, 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: $\mathrm{cost} = \sum_{i=0}^{M-1}\{N_i + n_i\}$.

Theorem (Brosse et al. (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 μ^{-1}

References I

Brosse, N., A. Durmus, and E. Moulines (2018). Normalizing constants of log-concave densities. *Electron. J. Stat. 12*(1), 851–889.

Brosse, N., A. Durmus, É. Moulines, and M. Pereyra (2017). Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo. In *Conference on Learning Theory*, pp. 319–342.

Bubeck, S., R. Eldan, and J. Lehec (2015). Finite-time analysis of projected langevin Monte Carlo. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, Cambridge, MA, USA, pp. 1243–1251. MIT Press.

Durmus, A., E. Moulines, and M. Pereyra (2018). Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM J. Imaging Sci. 11*(1), 473–506.

Lovász, L. and S. Vempala (2007, May). The geometry of logconcave functions and sampling algorithms. *Random Struct. Algorithms 30*(3), 307–358.

Zhu, L., W. Zhang, D. Elnatan, and B. Huang (2012). Faster STORM using compressed sensing. *Nat. Meth. 9*(7), 721–723.