Computational methods for high dimensional statistic: Part I

Błażej Miasojedow

Institute of Applied Mathematics and Mechanics, University of Warsaw

30 Novemeber 2021

Outline

1. What means high dimensional problem?
2. Non-smooth regularization

Outline

1. What means high dimensional problem?
2. Non-smooth regularization

1. What means high dimensional problem?

2. Non-smooth regularization

We say that problem is high dimensional if our data set (matrix) is of dimensions $n \times p$ (n observations p features) with:

- $p >> 1$
- $n >> 1$
- both $n >> 1$ and $p >> 1$

Large p

When number of features $p >> 1$ we need to deal with following issues:

- Ill-posed problems. Required regularization.
- When we would like to get interpretation we need to select variables. Non-smooth optimization problems.
- We cannot use hessian matrix. Too expensive! (computing costs $\mathcal{O}(p^2)$, computing inverse $\mathcal{O}(p^3)$.

Large n

When number of observation is large we could meet other problems:

- Using all data could be expensive.
- When n is huge we often have only on-line access to data.
- Data could be stored in different places. Synchronization problem.

LASSO

Let us consider linear model

$$Y = X\beta + \varepsilon \ .$$

and the Lasso estimator:

$$\beta_\lambda = \arg\min_{\beta} \left\{ \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

How to compute it?

Gradient descent

Goal:

$$\min_x f(x),$$

where:

- f is convex and differentiable;
- f is L-smooth i.e.

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

Gradient descent algorithm:

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k).$$

Key lemma

Lemma 1

If f is L smooth then for every x, y we have

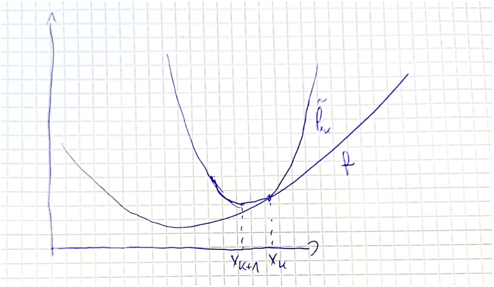
$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2 := \tilde{f}_{x,L}(y)$$

Geometric interpretation of gradient descent

Lemma 2

If f is L smooth and for every k $t_k \le \frac{1}{L}$ then gradient descent is monotonic i.e

$$f(x_{k+1}) \le f(x_k)$$



$$f(x_k) = \tilde{f}_{x_k, t_k^{-1}}(x_k) \ge \min_y \tilde{f}_{x_k, t_k^{-1}}(y)$$
$$= \tilde{f}_{x_k, t_k^{-1}}(x_{k+1}) \ge f(x_{k+1})$$

Key inequality

Lemma 3

If f is L-smooth and convex then sequence generated by gradient descent algorithm with $\gamma_k \leq \frac{1}{L}$ then

$$2\gamma_k(f(x_k) - f(x^*)) \leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2$$

Convergence of gradient descent

Theorem 4

If f is L-smooth and convex then sequence generated by gradient descent algorithm with $\gamma_k \leq \frac{1}{L}$ then

$$f(x_n) - f(x^*) \leq \frac{2L\|x_0 - x^*\|^2}{n}$$

Backtracking

In practice L is usually unknown and we need to use different step size rule. The prof rely on the Lemma 1 and we can add additional step to algorithm. Find minimal ℓ such that $\gamma_k = \eta^\ell \gamma_{k-1}$ satisfy

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{1}{2\gamma_k} \|x_{k+1} - x_k\|^2$$

With this procedure the Theorem 4 remains correct up to constant.

Return to the LASSO problem

The objective function is non-smooth

$$F(\beta) = \|Y - X\beta\|^2 + \lambda\|\beta\|_1$$

and we need to modify gradient descent algorithm.

(Projected) Subgradient method

Let f be convex, vector g is called subgradient of f at x if for every y we have

$$f(y) \geq f(x) + \langle g, y - x \rangle$$

The set of all subgradients is called subdifferential and will be denoted by $\partial f(x)$.

Let C be closed, convex set and consider problem

$$\min_{x \in C} f(x)$$

Projected subgradient algorithm:

$$x_{k+1} = P_C(x_k - \gamma_k g_k),$$

where $g_k \in \partial f(x_k)$ and P_C is a projection on set C.

Convergence of subgradient method

Lemma 5

$$2\gamma_k(f(x_k) - f(x^*)) \leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \gamma_k^2\|g_k\|^2$$

Theorem 6

If f is L Lipschitz then

$$f_n^{best} - f(x^*) \leq \frac{\|x_0 - x^*\|^2 + L\sum\gamma_k^2}{\sum\gamma_k}$$

where $f_n^{best} = min_{k \leq n} f(x_k)$

Therefore if $\gamma_k \approx \frac{1}{\sqrt{k}}$ then we get convergence of order $\mathcal{O}(\frac{\log(n)}{\sqrt{n}})$

Proximal operator

For convex function g we define the proximal operator by

$$\operatorname{prox}_{\gamma g}(x) = \arg\min_{y} \left\{ g(y) + \frac{1}{2\gamma}\|y - x\|^2 \right\}$$

- If $g = \delta_C$ convex indicator of set C then prox is a projection operator.
- If $y = \operatorname{prox}_{\gamma g}(x)$ then
$$y \in x - \gamma \partial f(y).$$
So it is implicit discretization of $\dot{x} \in \partial f(x)$

Proximal gradient algorithm

Goal:

$$\min_x \{f(x) + g(x)\}$$

where f convex smooth and g convex.

Proximal gradient algorithm:

$$x_{k+1} = \text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$$

Proximal gradient for LASSO

If $g = \| \cdot \|_1$ then

$$(\mathrm{prox}_{\gamma \| \cdot \|_1}(x))_i = \mathrm{sign(x_i)(|x|_i - \gamma)_+}$$

This operator is called soft-threshold operator and will be denoted by $\mathcal{S}_\gamma$ So for LASSO

$$\min_\beta \frac{1}{2} \|Y - X\beta\|^2 + \lambda \|\beta\|_1$$

we have step of proximal gradient algorithm defined by

$$\beta_{k+1} = S_{\gamma_k \lambda}(\beta_k - \gamma_k X^T(Y - X\beta_k))$$

Properties of proximal gradient algorithm

Lemma 7

If f is L smooth then proximal gradient algorithm is monotonic

Lemma 8

If f is L-smooth and convex then sequence generated by proximal gradient algorithm with $\gamma_k \leq \frac{1}{L}$ satisfy

$$2\gamma_k(f(x_k) - f(x^*)) \leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2$$

Theorem 9

If f is L-smooth and convex then sequence generated by proximal gradient algorithm with $\gamma_k \leq \frac{1}{L}$ satisfy

$$f(x_n) - f(x^*) \leq \frac{2L\|x_0 - x^*\|^2}{n}$$

Nesterov acceleration (Beck&Teoubulle 2008)

Set $y_0 = x_0$ and $t_0 = 1$

1. Set

$$x_{k+1} = \text{prox}_{\gamma_k g}(y_k - \gamma_k \nabla f(y_k))$$

2. Set

$$t_{k+1} = \frac{1 + \sqrt{4t_k^2}}{2}$$

3. Set

$$y_{k+1} = x_{k+1} + \tfrac{t_k - 1}{t_{k+1}}(x_{k+1} - x_k)$$

Convergence of accelerated proximal gradient

Theorem 10

If f is L-smooth and convex then sequence generated by proximal gradient algorithm with $\gamma_k \leq \frac{1}{L}$ satisfy

$$f(x_n) - f(x^*) \leq \frac{2L\|x_0 - x^*\|^2}{(n+1)^2}$$

- Accelerated proximal gradient algorithm is not monotonic
- The same bactracking rule as for gradient descent works for accelerated proximal gradient algorithm.

Alternative Direction Method of Multipliers (Parikh& Boyd 2014)

Consider the problem of form

$$\min_{Ax+Bz=C} f(x) + g(z)$$

Augmented Lagrangian of the problem is given by

$$L_\rho(x,z,y) = f(x) + g(z) + \langle y, Ax + Bz - c\rangle + \frac{\rho}{2}\|y - Ax - Bz + c\|^2$$

ADMM algorithm

- $$x_{k+1} = \arg\min_x \left\{f(x) + \frac{\rho}{2}\|y_k - Ax - Bz_k + c\|^2\right\}$$
- $$z_{k+1} = \arg\min_z \left\{g(z) + \frac{\rho}{2}\|y_k - Ax_{k+1} - Bz + c\|^2\right\}$$
- $$y_{k+1} = y_k + \rho(Ax_{k+1} - Bz_{k+1} + c)$$

Computational methods for high dimensional statistic: Part II

Błażej Miasojedow

Institute of Applied Mathematics and Mechanics, University of Warsaw

1 December 2021

Outline

1. Stochastic (sub)gradient

2. Variance reduction techniques

3. Stochastic approximation

Outline

Outline

Motivating example

Goal:

$$\min_x F(x) = \min_x \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

We could think that n is a number of observation f_i is negative loglikelkihood of observation i.

To reduce cost of single step, instead of computing gradient ∇F we approximate it by

$$g(x) = \frac{1}{k}\sum_{i\in I_k} \nabla f_i(x)$$

where I is a random subset of $\{1,\ldots,n\}$ of cardinality $|I| = k$

Stochastic gradient:

$$x_{k+1} = x_k - \gamma_k g(x_k)$$

Assumptions

1. f is convex,
2. g is unbiased i.e
$$E(g(x_k)|x_k) \in \partial f(x_k)$$
and with bounded variance
$$E(\|g(x_k)\|^2|x_k) \leq \sigma^2;$$

Projected stochastic subgradient (B):

$$x_{k+1} = P_C(x_k - \gamma_k g(x_k))$$

Convergence

Since projection is 1-Lipschitz

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|P_C(x_k - \gamma_k g(x_k)) - P_C(x^*)\|^2 \\ &\leq \|x_k - \gamma_k g(x_k) - x^*\|^2 \\ &= \|x_k - x^*\|^2 - 2\gamma_k \langle g(x_k), x_k - x^* \rangle + \gamma_k^2 \|g(x_k)\|^2\end{aligned}$$

Taking conditional expectation on both side we get

$$E(\|x_{k+1} - x^*\|^2 | x_k) \leq \|x_k - x^*\|^2 + 2\gamma_k \langle \partial f(x_k), x^* - x_k \rangle + \gamma_k^2 \sigma^2$$

By convexity of f

$$\langle \partial f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k)$$

Taking expectation on both side we get

$$2\gamma_k (Ef(x_k) - f(x^*)) \leq E\|x_k - x^*\|^2 - E\|x_{k+1} - x^*\|^2 + \gamma_k^2 \sigma^2$$

Convergence

Theorem 1

Under our assumptions:

- $$Ef(\bar{x}_n) - f(x^*) \leq \frac{\|x_0 - x^*\|^2 + \sigma^2 \sum_{k=1}^n \gamma_k^2}{\sum_k \gamma_k}$$

  where $\bar{x}_n = \frac{\sum \gamma_k x_k}{\sum \gamma_k}$.

- $$Ef(x_n^{best}) - f(x^*) \leq \frac{\|x_0 - x^*\|^2 + \sigma^2 \sum_{k=1}^n \gamma_k^2}{\sum_k \gamma_k}$$

  where $x_n^{best} = \arg\min_{k \leq n} f(x_k)$.

Setting $\gamma_k \approx \frac{1}{\sqrt{k}}$ we get convergence rate $\mathcal{O}(\frac{\log(n)}{\sqrt{n}})$

Strong convexity

Definition 2

Function f is called m - strongly convex if function $f - \frac{m}{2}\|\cdot\|^2$ is convex.

Theorem 3

If f is m- strongly convex then

- $$f(y) \geq f(x) + \langle \partial f(x), y - x \rangle + \frac{m}{2}\|x - y\|^2$$
- *There exists unique minimizer x^* and*

$$f(x) - f(x^*) \geq \frac{m}{2}\|x - x^*\|^2$$

Convergence of stochastic subgradient under strong convexity assumption

Recall that

$$E(\|x_{k+1} - x^*\|^2 | x_k) \leq \|x_k - x^*\|^2 + 2\gamma_k \langle \partial f(x_k), x^* - x_k \rangle + \gamma_k^2 \sigma^2$$

By m strong convexity of f we have

$$\langle \partial f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k) - \tfrac{m}{2} \|x_k - x^*\|^2$$

Taking expectation on both side we get

$$2(Ef(x_k) - f(x^*)) \leq (\tfrac{1}{\gamma_k} - m) E\|x_k - x^*\|^2 - \tfrac{1}{\gamma_k} E\|x_{k+1} - x^*\|^2 + \gamma_k \sigma^2$$

Setting $\gamma_k = \frac{2}{m(k+1)}$ and multiplying inequality by k we get

$$k(Ef(x_k) - f(x^*)) \leq \tfrac{k(k-1)m}{4} E\|x_k - x^*\|^2 - \tfrac{k(k+1)m}{4} E\|x_{k+1} - x^*\|^2 + \sigma^2$$

Convergence of stochastic subgradient under strong convexity assumption

Theorem 4

- $$Ef(\bar{x}_n) - f(x^*) \leq \frac{\sigma^2}{m(n+1)}$$

  where $\bar{x}_n = \sum \frac{2k}{n(n+1)} x_k$.

- $$Ef(x_n^{best}) - f(x^*) \leq \frac{\sigma^2}{m(n+1)}$$

  where $x_n^{best} = \arg\min_{k \leq n} f(x_k)$.

Some extensions

- Stochastic proximal gradient algorithm: Nitanda (2014); Atchade, Fort, Moulines (2016)
- Markovian noise: Atchade, Fort, Moulines (2016); Karimi, Wei, M, Moulines (2019)
- non-Convex case: Karimi, Wei, M, Moulines (2019)

Why we could reduce variance?

We approximate $\nabla f(x)$ by

$$g(x) = \frac{1}{k} \sum_{i \in I_k} \nabla f_i(x)$$

- At each step we compute "independently" gradient. We do not use previous approximation.
- To get small variance we need large k. Variance does not vanish when number of iteration growths.
- The gradient should not change too much between consecutive steps.
- It seems reasonable to introduce small bias and reduce variance. Use approximation of form

$$\tilde{g}(x_{k+1}) = \alpha_k \tilde{g}(x_k) + (1 - \alpha_k) g(x_{k+1})$$

Stochastic Variance Reducing Gradient (Johnson, Zhang 2013)

- Initialize by x_0 and $\tilde{x}_0$
- For $k = 1, 2, \cdots$
    1. Update mean gradient $\tilde{g}_k = \frac{1}{n} \sum_i \nabla f_i(\tilde{x}_k)$
    2. Set $x_0 = \tilde{x}_k$
    3. For $\ell = 0, \ldots, m-1$ draw randomly i_ℓ and
    $$x_\ell = x_{\ell-1} + \gamma(\nabla f_{i_\ell}(x_{\ell-1}) - \nabla f_{i_\ell}(\tilde{x}_k) + \tilde{g}_k)$$
    4. $\tilde{x}_{k+1} = x_m$

For L smooth and strongly convex function and m large enough we have

$$EF(\tilde{x}_k) - F(x^*) \leq \alpha^k (F(\tilde{x}_0) - F(x^*))$$

for $\alpha < 1$.

SAGA De Fazio, Bach, Lacoste-Julien 2014

$$\min_x F(x) = \min_x \frac{1}{n} \sum f_i(x) + g(x)$$

1. We have stored x_k, $\{\nabla f_i(\phi_k^i)\}$, $\tilde{h}_k = \frac{1}{n} \sum_i \nabla f_i(\phi_k^i)$.
2. Pick randomly j and set $\phi_{k+1}^j = x_k^j$ and update derivatives.
3. Update x by

$$x_{k+1} = \text{prox}_{\gamma g}(x_k - \gamma(\nabla f_i(\phi_{k+1}^j) - \nabla f_i(\phi_k^j) + \tilde{h}_k))$$

Under L smooth and strong convexity assumption on F and Lipschitz continuity of g we could get

$$E\|x_k - x_*\|^2 \leq \alpha^k(\|x_0 - x_*\|^2 + \text{something not important})$$

Stochastic approximation

$$x_{k+1} = x_k + \gamma_k H(x_k, \xi_{k+1})$$

Where $H(x, k, \xi_{k+1})$ is a random approximation of mean field $h(x_k)$, ξ_{k+1} is random variable.

Stochastic gradient algorithm:

$$h(x) = -\nabla f(x).$$

Convergence of SA (Kushner& Yin 2003)

Sketch of proof

1. First show stability i.e. there exists compact set $\mathcal{K}$ such that $x_k \in \mathcal{K}$ a.s.
2. When we know that algorithm is stable we show that sequence x_k behaves asymptotically as gradient flow

$$\dot{x} = h(x(t)) \text{ or } \dot{x} \in h(x(t))$$
(Majewski, M, Moulines (2018); Davis, Drusvyatskiy, Kakade, Lee (2020))

3. Applying Lypaunov stability of $x(t)$ to get convergence.

Stability

Stability is implied by:

- Growth condition on mean field h and on variance of the noise.
- Restarts (Andrieu, Moulines, Priouret 2005)
- Projection on compact set Kushner & Yin 2003

Restarts (Andrieu, Moulines, Priouret 2005)

Let define family of compact sets $\mathcal{K}_0 \subset \mathcal{K}_1 \subset \cdots$

1. Set $\ell = 0$ and draw $x_0 \in \mathcal{K}_\ell$.
2. If $x_k \notin \mathcal{K}_\ell$ update $\ell = \ell + 1$ and draw independently on history $x_{k+1} \in \mathcal{K}_\ell$

Projection on compact set

Let $\mathcal{K}$ be "regular" compact set and define algorithm by

$$x_{k+1} = P_{\mathcal{K}}(x_k - \gamma_k H(x_k, \xi_k))$$

For $\omega \notin N$, $\theta_{n+1}(\omega) - \theta_n(\omega) \to 0$. If $Z^n(\omega, \cdot)$ is not equicontinuous, then there is a subsequence that has a jump asymptotically; that is, there are integers $\mu_k \to \infty$, uniformly bounded times s_k, $0 < \delta_k \to 0$ and $\rho > 0$ (all depending on ω) such that $|Z^{\mu_k}(\omega, s_k + \delta_k) - Z^{\mu_k}(\omega, s_k)| \geq \rho$. The changes of the terms other than $Z^n(\omega, t)$ on the right side of (2.7) go to zero on the intervals $[s_k, s_k + \delta_k]$. Furthermore $\epsilon_n Y_n(\omega) = \epsilon_n \bar{g}(\theta_n(\omega)) + \epsilon_n \delta M_n(\omega) + \epsilon_n \beta_n \to 0$ and $Z_n(\omega) = 0$ if $\theta_{n+1}(\omega) \in H^0$, the interior of H. Thus, this jump cannot force the iterate to the interior of the hyperrectangle H, and it cannot force a jump of the $\theta^n(\omega, \cdot)$ along the boundary either. Consequently, $\{Z^n(\omega, \cdot)\}$ is equicontinuous.

Kushner& Yin p. 151

Asymptotic behaviour

We can write SA as

$$x_{k+1} = x_k + \gamma_k(h(x_k) + r_k + m_k)$$

Where $r_k \to 0$ and m_k martingale differences

We define piece wise linear approximation

$$X_0(t)$$

Let $t_k = \sum_{i \leq k} \gamma_i$ and

$$X_k(t) = X_0(t + t_k)$$

Asymptotic behaviour

Theorem 5

If

- x_k *is stable*
- h *is locally Lipschitz.*
- $\|r_k| \to 0$ *and* $|\sum \gamma_k m_k| < \infty$.
- $\sum \gamma_k = \infty$, *and* $\sum \gamma_k^2 < \infty$

Then there exists subsequence n_k, *and absolutely continuous function* x_∞ *such that for any* $T > 0$

$$\sup_{t \in [0,T]} X_{n_k}(t) - x_\infty(t) \to 0.$$

In addition $x_\infty(t)$ *is a limit point of* x_k

Sketch of proof

1. First show that family $X_k(t)$ is equi- continuous.
2. By Arzela-Ascoli theorem we get relative compactness of X_k
3. Identify the limit $\dot{X}_\infty(t) = h(X_\infty(t))$

Lyapunov condition

$V > 0$ is a Lyapunov function for solution to $\dot{x} = h(x)$ if

$$\dot{V}(x(t)) < 0$$

or equivalently if

$$\langle \nabla V(x), h(x) \rangle \leq 0$$

Converegence of SA

Theorem 6

Let

$$\mathcal{S} = \{x \colon \nabla V(x), h(x)\rangle = 0\}$$

If $V(\mathcal{S} \cap \mathcal{K})$ *has empty interior then*

$$dist(x_k, \mathcal{S} \cap \mathcal{K}) \to 0$$

Computational methods for high dimensional statistic: Part III

Błażej Miasojedow

Institute of Applied Mathematics and Mechanics, University of Warsaw

2 December 2021

Outline

1. Unadjusted Langevin Algorithm
2. ULA as an optimization algorithm in the Wasserstein space
3. Extensions of ULA

Outline

1. Unadjusted Langevin Algorithm
2. ULA as an optimization algorithm in the Wasserstein space
3. Extensions of ULA

Outline

Introduction

We want to do statistical inference for data $Y_1, \ldots, Y_n \in \mathbb{R}^d$, where $n >> 0$ and/or $d >> 0$, and we want to be Bayesian. So, we need to be able to explore posterior distribution of form

$$\pi(x) \propto p(x) \prod_{i=1}^{n} \ell_i(x),$$

where p is some prior and ℓ_i are likelihood of observation Y_i.

Introduction

We want to do statistical inference for data $Y_1, \ldots, Y_n \in \mathbb{R}^d$, where $n >> 0$ and/or $d >> 0$, and we want to be Bayesian. So, we need to be able to explore posterior distribution of form

$$\pi(x) \propto p(x) \prod_{i=1}^{n} \ell_i(x),$$

where p is some prior and ℓ_i are likelihood of observation Y_i.

We need MCMC algorithm which scales well with n and d.

Problems with standard MCMC algorithms

Let us assume that π is differentiable. In this case one of most popular algorithm is MALA. The Metropolis - Hastings algorithm with proposal of form

$$X^{\text{prop}} = X^{\text{old}} - \gamma \nabla \log(\pi(X^{\text{old}})) + \sqrt{2\gamma} G,$$

where G d-dimensional standard Gaussian

1. Cost of generating proposal is of order $\mathcal{O}(nd)$.
2. Cost of computing acceptance ratio is also $\mathcal{O}(nd)$.
3. We can reduce the cost of generating proposal to $\mathcal{O}(d)$ by using stochastic gradient instead of true gradient. But still cost of single iteration of algorithms is $\mathcal{O}(nd)$, due to the acceptance step
4. There are no bounds on mixing times which are polynomial in d.

Problems with standard MCMC algorithms

Let us assume that π is differentiable. In this case one of most popular algorithm is MALA. The Metropolis - Hastings algorithm with proposal of form

$$X^{\text{prop}} = X^{\text{old}} - \gamma \nabla \log(\pi(X^{\text{old}})) + \sqrt{2\gamma} G,$$

where G d-dimensional standard Gaussian

1. Cost of generating proposal is of order $\mathcal{O}(nd)$.
2. Cost of computing acceptance ratio is also $\mathcal{O}(nd)$.
3. We can reduce the cost of generating proposal to $\mathcal{O}(d)$ by using stochastic gradient instead of true gradient. But still cost of single iteration of algorithms is $\mathcal{O}(nd)$, due to the acceptance step
4. There are no bounds on mixing times which are polynomial in d.

Problems with standard MCMC algorithms

Let us assume that π is differentiable. In this case one of most popular algorithm is MALA. The Metropolis - Hastings algorithm with proposal of form

$$X^{\text{prop}} = X^{\text{old}} - \gamma \nabla \log(\pi(X^{\text{old}})) + \sqrt{2\gamma} G,$$

where G d-dimensional standard Gaussian

1. Cost of generating proposal is of order $\mathcal{O}(nd)$.
2. Cost of computing acceptance ratio is also $\mathcal{O}(nd)$.
3. We can reduce the cost of generating proposal to $\mathcal{O}(d)$ by using stochastic gradient instead of true gradient. But still cost of single iteration of algorithms is $\mathcal{O}(nd)$, due to the acceptance step
4. There are no bounds on mixing times which are polynomial in d.

Problems with standard MCMC algorithms

Let us assume that π is differentiable. In this case one of most popular algorithm is MALA. The Metropolis - Hastings algorithm with proposal of form

$$X^{\text{prop}} = X^{\text{old}} - \gamma \nabla \log(\pi(X^{\text{old}})) + \sqrt{2\gamma} G,$$

where G d-dimensional standard Gaussian

1. Cost of generating proposal is of order $\mathcal{O}(nd)$.
2. Cost of computing acceptance ratio is also $\mathcal{O}(nd)$.
3. We can reduce the cost of generating proposal to $\mathcal{O}(d)$ by using stochastic gradient instead of true gradient. But still cost of single iteration of algorithms is $\mathcal{O}(nd)$, due to the acceptance step
4. There are no bounds on mixing times which are polynomial in d.

Problems with standard MCMC algorithms

Let us assume that π is differentiable. In this case one of most popular algorithm is MALA. The Metropolis - Hastings algorithm with proposal of form

$$X^{\mathrm{prop}} = X^{\mathrm{old}} - \gamma \nabla \log(\pi(X^{\mathrm{old}})) + \sqrt{2\gamma} G,$$

where G d-dimensional standard Gaussian

1. Cost of generating proposal is of order $\mathcal{O}(nd)$.
2. Cost of computing acceptance ratio is also $\mathcal{O}(nd)$.
3. We can reduce the cost of generating proposal to $\mathcal{O}(d)$ by using stochastic gradient instead of true gradient. But still cost of single iteration of algorithms is $\mathcal{O}(nd)$, due to the acceptance step
4. There are no bounds on mixing times which are polynomial in d.

Unadjusted Langevin Algorithm

Assume that π is of form

$$\pi \propto e^{-U},$$

where $U\colon \mathbb{R}^d \to \mathbb{R}$ is a convex potential.

Unadjusted Langevin Algorithm

Assume that π is of form

$$\pi \propto e^{-U},$$

where $U\colon \mathbb{R}^d \to \mathbb{R}$ is a convex potential.

One possibility is to approximate π by Unadjasted Langevin Algorithm. We generate Markov chain $(X_k)_{k\geq 0}$ given for all $k \geq 0$ by

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

where $(\gamma_k)_{k\geq 1}$ is a sequence of step sizes which can be held constant or converges to 0, and $(G_k)_{k\geq 1}$ is a sequence of i.i.d. standard d-dimensional Gaussian random variables.

Unadjusted Langevin Algorithm *cont.*

1. ULA is the Euler-Maruyama discretization of over-damped Langevin diffusion associated with U

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

   where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion.
2. Under appropriate conditions on U, $\mathbf{Y}_t$ converges to π in total variation distance or in Wasserstein distance.
3. However discretization introduces an additional error and we want to quantify it.

Unadjusted Langevin Algorithm *cont.*

1. ULA is the Euler-Maruyama discretization of over-damped Langevin diffusion associated with U

$$d\mathbf{Y}_t = -\nabla U(\mathbf{Y}_t)dt + \sqrt{2}dB_t \ ,$$

   where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion.
2. Under appropriate conditions on U, $\mathbf{Y}_t$ converges to π in total variation distance or in Wasserstein distance.
3. However discretization introduces an additional error and we want to quantify it.

Unadjusted Langevin Algorithm *cont.*

1. ULA is the Euler-Maruyama discretization of over-damped Langevin diffusion associated with U

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

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

2. Under appropriate conditions on U, $\mathbf{Y}_t$ converges to π in total variation distance or in Wasserstein distance.
3. However discretization introduces an additional error and we want to quantify it.

Existing results for ULA

- Weak error estimates have been obtained in [Talay and Tubaro, 1990], [Mattingly et al., 2002] for the constant step size setting and [Lamberton and Pagès, 2003], [Lemaire, 2005] when $(\gamma_k)_{k\geq 1}$ is non-increasing and goes to 0.
- Explicit and non-asymptotic bounds on the total variation [Dalalyan, 2016], [Durmus and Moulines, 2017] or the Wasserstein distance [Durmus and Moulines, 2016] between the distribution of X_k and π have been obtained.
- All these results are based on the comparison between the discretization and the diffusion process and quantify how the error introduced by the discretization accumulate throughout the algorithm
- Here we introduce a new interpretation of ULA, as an optimization algorithm in the Wasserstein space.

Existing results for ULA

- Weak error estimates have been obtained in [Talay and Tubaro, 1990], [Mattingly et al., 2002] for the constant step size setting and [Lamberton and Pagès, 2003], [Lemaire, 2005] when $(\gamma_k)_{k\geq 1}$ is non-increasing and goes to 0.
- Explicit and non-asymptotic bounds on the total variation [Dalalyan, 2016], [Durmus and Moulines, 2017] or the Wasserstein distance [Durmus and Moulines, 2016] between the distribution of X_k and π have been obtained.
- All these results are based on the comparison between the discretization and the diffusion process and quantify how the error introduced by the discretization accumulate throughout the algorithm
- Here we introduce a new interpretation of ULA, as an optimization algorithm in the Wasserstein space.

Existing results for ULA

- Weak error estimates have been obtained in [Talay and Tubaro, 1990], [Mattingly et al., 2002] for the constant step size setting and [Lamberton and Pagès, 2003], [Lemaire, 2005] when $(\gamma_k)_{k\geq 1}$ is non-increasing and goes to 0.
- Explicit and non-asymptotic bounds on the total variation [Dalalyan, 2016], [Durmus and Moulines, 2017] or the Wasserstein distance [Durmus and Moulines, 2016] between the distribution of X_k and π have been obtained.
- All these results are based on the comparison between the discretization and the diffusion process and quantify how the error introduced by the discretization accumulate throughout the algorithm
- Here we introduce a new interpretation of ULA, as an optimization algorithm in the Wasserstein space.

Existing results for ULA

- Weak error estimates have been obtained in [Talay and Tubaro, 1990], [Mattingly et al., 2002] for the constant step size setting and [Lamberton and Pagès, 2003], [Lemaire, 2005] when $(\gamma_k)_{k\geq 1}$ is non-increasing and goes to 0.
- Explicit and non-asymptotic bounds on the total variation [Dalalyan, 2016], [Durmus and Moulines, 2017] or the Wasserstein distance [Durmus and Moulines, 2016] between the distribution of X_k and π have been obtained.
- All these results are based on the comparison between the discretization and the diffusion process and quantify how the error introduced by the discretization accumulate throughout the algorithm
- Here we introduce a new interpretation of ULA, as an optimization algorithm in the Wasserstein space.

A different representation of Langevin dynamics

It can be shown [Jordan et al., 1998], that if U is infinitely continuously differentiable, $(\rho_t^x)_{t>0}$, the density of solution of Langevin equation at time $t > 0$, is the limit of the minimization scheme which defines a sequence of probability measures $(\tilde{\rho}_{k,\gamma}^x)_{k\in\mathbb{N}}$ as follows. For $x \in \mathbb{R}^d$ and $\gamma > 0$ set $\rho_{0,\gamma}^x = \mathrm{d}\mu_0/\mathrm{d}\,\mathrm{Leb}$ and

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

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

$$\mathscr{F} = \mathscr{H} + \mathscr{E}\;,$$

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

Boltzmann H- functional and the potential energy functional

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

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

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

Lemma 1

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

Assumptions

A1 (m)

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

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

Note that **A1**(m) includes the case where U is only convex when $m = 0$. We consider the following additional condition on U which will be relaxed later.

A2

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

Inexact gradient descent

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex continuously differentiable objective function with

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

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

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

To get explicit bound on the convergence (in expectation) of the sequence $(f(x_n))_{n \in \mathbb{N}}$ to $f(x_f)$, one possibility is to show that the following inequality holds:

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

for some constant $C \geq 0$.

Main result for ULA

Consider the family of Markov kernels $(R_{\gamma_k})_{k\in\mathbb{N}^*}$ associated with the Euler-Maruyama discretization $(X_k)_{k\in\mathbb{N}}$, for a sequence of step sizes $(\gamma_k)_{k\in\mathbb{N}^*}$, given for all $\gamma > 0, x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

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

Theorem 2

*Assume **A1**(m) for $m \geq 0$ and **A2**. 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 I

For our analysis, we decompose R_γ for all $\gamma > 0$ in the product of two elementary kernels S_γ and T_γ given for all $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

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

Lemma 3

Assume A2. For all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma > 0$,

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

Proof of the main inequality II

Lemma 4

Assume **A1**(m) *for* $m \geq 0$ *and* **A2**. *For all* $\gamma \in (0, L^{-1}]$ *and* $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

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

Lemma 5

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

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

Proof of the main inequality III

Proof of Theorem 2.

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

Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma \in \mathbb{R}_+^*$. By Lemma 3, we get

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

By Lemma 4 ,

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

By Lemma 5,

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

Proof of Lemma 3

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

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

Proof of Lemma 4

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

$$2\gamma\left\{U(x-\gamma\nabla U(x))-U(y)\right\} \le (1-m\gamma)\left\|x-y\right\|^2-\left\|x-\gamma\nabla U(x)-y\right\|^2 - \gamma^2(1-\gamma L)\left\|\nabla U(x)\right\|^2 \ .$$

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

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

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

Proof of Lemma 5

Let $\mu_t = \mu T_t$. Then, we have:

$$\frac{\partial \mu_t}{\partial t} = \Delta \mu_t \ ,$$

and μ_t goes to μ as t goes to 0 in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. Let $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\gamma > 0$. Then it can be show that for all $\epsilon \in (0, \gamma)$, there exists $(\delta_t) \in \mathrm{L}^1((\epsilon, \gamma))$ such that

$$W_2^2(\mu_\gamma, \nu) - W_2^2(\mu_\epsilon, \nu) = \int_\epsilon^\gamma \delta_s \mathrm{d}s$$
$$\delta_s/2 \leq \mathscr{H}(\nu) - \mathscr{H}(\mu_s) \ , \text{ for almost all } s \in (\epsilon, \gamma) \ .$$

In addition $s \mapsto \mathscr{H}(\mu_s)$ is non-increasing on $\mathbb{R}_+^*$, therefore we get that

$$W_2^2(\mu_\gamma, \nu) - W_2^2(\mu_\epsilon, \nu) \leq 2(\gamma - \epsilon)\left\{\mathscr{H}(\nu) - \mathscr{H}(\mu_\gamma)\right\} \ .$$

Taking $\epsilon \to 0$ concludes the proof.

Complexity for ULA when U is strongly convex and gradient Lipschitz

	Total variation	Wasserstein distance	KL divergence
Durmus and Moulines 2016	$d\mathcal{O}(\varepsilon^{-2})$	$d\mathcal{O}(\varepsilon^{-2})$	–
Cheng and Bartlett, 2017	$d\mathcal{O}(\varepsilon^{-2})$	$d\mathcal{O}(\varepsilon^{-2})$	$d\mathcal{O}(\varepsilon^{-1})$
Our results	$d\mathcal{O}(\varepsilon^{-2})$	$d\mathcal{O}(\varepsilon^{-2})$	$d\mathcal{O}(\varepsilon^{-1})$

Complexity for ULA when U is convex and gradient Lipschitz

	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
Durmus and Moulines 2017	$d^5\mathcal{O}(\varepsilon^{-2})$	-	-
Our results	$d^3\mathcal{O}(\varepsilon^{-4})$	-	$d^3\mathcal{O}(\varepsilon^{-2})$

Table: Starting from minimizer of U

Stochastic Sub-Gradient Langevin Dynamics

A3

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

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

Stochastic Sub-Gradient Langevin Dynamics (SSGLD)

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

Complexity of SSGLD

1. In the case where a warm start complexity of SSGLD to obtain a sample ε close from π in KL is of order $(M^2 + D^2)\mathcal{O}(\varepsilon^{-2})$ and (Pinsker inequality) in TV distance is of order $(M^2 + D^2)\mathcal{O}(\varepsilon^{-4})$.
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 SSGLD for the KL:

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

and for TV

$$(\eta^{-2}d^2 + M_\eta^2 + M^2)(M^2 + D^2)\mathcal{O}(\varepsilon^{-4})$$

Stochastic Proximal Gradient Langevin Dynamics

A4 (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 **A1**(m) and **A2**. 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 **A1**(0) and is M_2-Lipschitz.

Stochastic Proximal Gradient Langevin Dynamics (SPGLD)

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

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

$$\operatorname{prox}^{\gamma}_{U_2}(x) = \operatorname*{arg\,min}_{y\in\mathbb{R}^d} \left\{ U_2(y) + (2\gamma)^{-1} \|x-y\|^2 \right\} \ .$$

Complexity of SPGLD

1. In the case where a warm start complexity of SPGLD to obtain a sample ε close from π in KL is of order $(d + M^2 + D^2)\mathcal{O}(\varepsilon^{-2})$ and (Pinsker inequality) in TV distance is of order $(d + M^2 + D^2)\mathcal{O}(\varepsilon^{-4})$.
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:

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

and for TV

$$(\eta^{-2}d^2 + M_\eta^2 + M^2)(d + M^2 + D^2)\mathcal{O}(\varepsilon^{-4})$$

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 [Durmus and Moulines, 2017], [Dalalyan, 2016] 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.

Summary

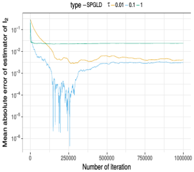
- We give a new interpretation of ULA and use it to get bounds on the Kullback-Leibler divergence from π to the iterates of ULA.
- We recover the dependence on the dimension of [Cheng and Bartlett, 2017] in the strongly convex case. We also give computable bounds when U is only convex which improves the results of [Durmus and Moulines, 2017], [Dalalyan, 2016] 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.

Numerical results

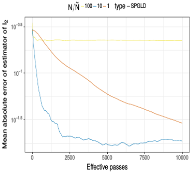
We consider Bayesian Lasso and Bayesian elastic net logistic regression model, for 2 datasets from UCI repository (Australian Credit Approval dataset $d = 64, n = 690$, Musk dataset $n = 476, d = 166$)

Numerical results

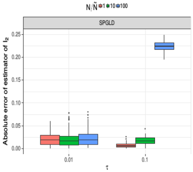
Australian Credit Approval



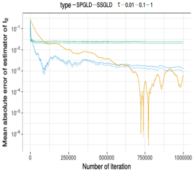
(a)



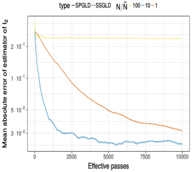
(b)



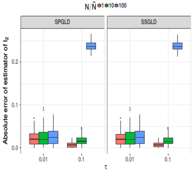
(c)



(d)



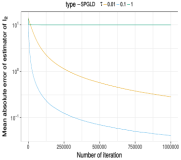
(e)



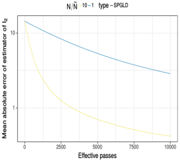
(f)

Numerical results

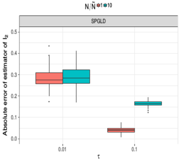
Musk



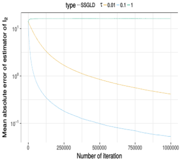
(a)



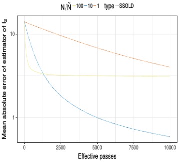
(b)



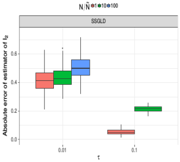
(c)



(d)



(e)



(f)

Thank you!

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

Dalalyan, A. S. (2016).
Theoretical guarantees for approximate sampling from smooth and log-concave densities.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), pages n/a–n/a.

Durmus, A. and Moulines, E. (2016).
High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm.

Durmus, A. and Moulines, Ã. (2017).
Nonasymptotic convergence analysis for the unadjusted langevin algorithm.
Ann. Appl. Probab., 27(3):1551–1587.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998).
The variational formulation of the Fokker-Planck equation.
SIAM journal on mathematical analysis, 29(1):1–17.

Lamberton, D. and Pagès, G. (2003).
Recursive computation of the invariant distribution of a diffusion: the case of a weakly mean reverting drift.
Stoch. Dyn., 3(4):435–451.

Lemaire, V. (2005).
Estimation de la mesure invariante d'un processus de diffusion.
PhD thesis, UniversitÃ© Paris-Est.

Mattingly, J. C., Stuart, A. M., and Higham, D. J. (2002).
Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise.
Stochastic Process. Appl., 101(2):185–232.

Talay, D. and Tubaro, L. (1990).
Expansion of the global error for numerical schemes solving stochastic differential equations.
Stochastic Anal. Appl., 8(4):483–509 (1991).