Non-asymptotic convergence bound for the Langevin MCMC Algorithm

Alain Durmus, Eric Moulines, Marcelo Pereyra, Umut Şimşekli

Telecom ParisTech, Ecole Polytechnique, Bristol University

January 27, 2017
1. Motivation

2. Framework

3. Strongly log-concave distribution

4. Convex and Super-exponential densities

5. Non-smooth potentials

6. The Unadjusted Langevin Algorithm within Gibbs (ULAwG)
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. Aggregation of estimators and predictors
  3. Bayesian non parametrics (function space)
  4. Bayesian linear inverse problems (function space)
Introduction

- "Classical" MCMC algorithms do not scale to high-dimension.
- However, the possibility of sampling high-dimensional distribution has been demonstrated in several fields (in particular, molecular dynamics) with specially tailored algorithms
- Our objective: Propose (or rather analyse) sampling algorithm that can be used for some challenging high-dimensional problems with a Machine Learning flavour.
- Challenges are numerous in this area...
Illustration

- **Likelihood**: Binary regression set-up in which the binary observations (responses) \((Y_1, \ldots, Y_n)\) are conditionally independent Bernoulli random variables with success probability \(F(\beta^T X_i)\), where
  1. \(X_i\) is a \(d\) dimensional vector of known covariates,
  2. \(\beta\) is a \(d\) dimensional vector of unknown regression coefficient
  3. \(F\) is a distribution function.

- **Two important special cases:**
  1. **probit regression**: \(F\) is the standard normal distribution function,
  2. **logistic regression**: \(F\) is the standard logistic distribution function:

\[
F(t) = \frac{e^t}{1 + e^t}
\]
Bayesian inference for binary regression?

- The posterior density distribution of \( \beta \) is given, up to a proportionality constant by

\[
\pi(\beta | (Y, X)) \propto \exp(-U(\beta))
\]

with

\[
U(\beta) = - \sum_{i=1}^{p} \{ Y_i \log F(\beta^T X_i) + (1 - Y_i) \log(1 - F(\beta^T X_i)) \} + g(\beta),
\]

where \( g \) is the log density of the posterior distribution.

- Two important cases:
  - Gaussian prior \( g(\beta) = (1/2) \beta^T \Sigma \beta \): ridge penalty.
  - Laplace prior \( g(\beta) = \lambda \sum_{i=1}^{d} |\beta_i| \): LASSO penalty.
New challenges

**Beware !** the number of predictor variables \( d \) is large \((10^4 \text{ and up})\).

- text categorization,
- genomics and proteomics (gene expression analysis),
- other data mining tasks (recommendations, longitudinal clinical trials, ..).
State of the art

The most popular algorithms for Bayesian inference in binary regression models are based on data augmentation

- Instead on sampling \( \pi(\beta|(X,Y)) \) sample \( \pi(\beta, W|(X,Y)) \) probability measure on \( \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \) and take the marginal w.r.t. \( \beta \).
- Typical application of the Gibbs sampler: sample in turn \( \pi(\beta|(X,Y,W)) \) and \( \pi(W|(X,Y,\beta)) \).
- The choice of the DA should make these two steps reasonably easy...
  - logistic link: Polya-Gamma sampler, Polsson and Scott (2012)…!
State of the art: shortcomings

- The Albert and Chib DA probit DA algorithm and the Polya-Gamma sampler have been shown to be uniformly geometrically ergodic, **BUT**
  - The geometric rate of convergence is exponentially small with the dimension
  - Do not allow to construct honest confidence intervals, credible regions
- The algorithms are very demanding in terms of computational ressources...
  - applicable only when is $d$ small 10 to moderate 100 but certainly not when $d$ is large ($10^4$ or more).
  - convergence time prohibitive as soon as $d \geq 10^2$. 
A daunting problem?

- In the case of the ridge regression, the potential $U$ is smooth strongly convex.
- In the case of the lasso regression, the potential $U$ is non-smooth but still convex...
- A wealth of reasonably fast optimisation algorithms are available to solve this problem in high-dimension...
Motivation

Framework

Strongly log-concave distribution
Convex and Super-exponential densities
Non-smooth potentials

The Unadjusted Langevin Algorithm within Gibbs (ULAwG)
Motivation

Framework

Strongly log-concave distribution
Convex and Super-exponential densities
Non-smooth potentials

The Unadjusted Langevin Algorithm within Gibbs (ULAwG)

Framework

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

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

Implicitly, $d \gg 1$.

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

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

- *(overdamped)* Langevin SDE:
  \[
  dY_t = -\nabla U(Y_t)dt + \sqrt{2}dB_t ,
  \]
  where \((B_t)_{t \geq 0}\) is a \(d\)-dimensional Brownian Motion.

- **Notation:** \((P_t)_{t \geq 0}\) the Markov semigroup associated to the Langevin diffusion:
  \[
  \pi \propto e^{-U} \text{ is reversible } \rightsquigarrow \text{ the unique invariant probability measure}.
  \]

- **Key property:** For all \(x \in \mathbb{R}^d\),
  \[
  \lim_{t \to +\infty} \|\delta_x P_t - \pi\|_{TV} = 0.
  \]
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. \(\mathcal{N}(0, I_d)\)
- \((\gamma_k)_{k \geq 1}\) is a sequence of stepsizes, which can either be held constant or be chosen to decrease to 0 at a certain rate.

- Closely related to the **gradient descent algorithm.**
Discretized Langevin diffusion: constant stepsize

- When $\gamma_k = \gamma$, then $(X_k)_{k \geq 1}$ is an homogeneous Markov chain with Markov kernel $R_{\gamma}$.
- Under some appropriate conditions, this Markov chain is irreducible, positive recurrent $\sim$ unique invariant distribution $\pi_{\gamma}$.
- **Problem:** the limiting distribution of the discretization $\pi_{\gamma}$ does not coincide with the target distribution $\pi$.
- **Questions:**
  - Can we quantify the distance between $\pi_{\gamma}$ and $\pi$, e.g. a bound for $||\pi_{\gamma} - \pi||_{TV}$ with explicit dependence in the dimension?
  - Given a computational budget, is there an optimal trade-off between the "mixing" rate ($||\delta_x R_{\gamma} - \pi_{\gamma}||_{TV}$) and the bias ($||\pi_{\gamma} - \pi||_{TV}$)?
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 sequence of Markov kernel \( (R_{\gamma_k})_{k \geq 1} \).

- Notation: \( Q^p_{\gamma} \) is the composition of Markov kernels

\[
Q^p_{\gamma} = R_{\gamma_1} R_{\gamma_2} \ldots R_{\gamma_p}
\]

With this notation, the law of \( X_p \) started at \( X_0 = x \) is equal to \( \delta_x Q^p_{\gamma} \).

- Questions:
  - Control \( \| \delta_x Q^p_{\gamma} - \pi \|_{TV} \) with explicit dependence in the dimension \( d \).
  - Should we use fixed or decreasing step sizes?
Metropolis-Adjusted Langevin Algorithm

To correct the target distribution, a Metropolis-Hastings step can be included \( \sim \) Metropolis Adjusted Langevin Algorithm (MALA).

- Key references Roberts and Tweedie, 1996

Algorithm:

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

\[
\alpha_\gamma(x, y) = 1 \wedge \frac{\pi(y) r_\gamma(y, x)}{\pi(x) r_\gamma(x, y)} , r_\gamma(x, y) \propto e^{-\|y-x-\gamma \nabla U(x)\|^2/(4\gamma)}
\]

3. Accept / Reject the proposal.
MALA: pros and cons

- Require to compute one gradient at each iteration and to evaluate one time the objective function.

- Geometric convergence is established under the condition that in the tail the acceptance region is inwards in $q$,

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

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

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

2 Framework

3 Strongly log-concave distribution

4 Convex and Super-exponential densities

5 Non-smooth potentials

6 The Unadjusted Langevin Algorithm within Gibbs (ULAwG)
Motivation
Framework
- **Strongly log-concave distribution**
- Convex and Super-exponential densities
- Non-smooth potentials

The Unadjusted Langevin Algorithm within Gibbs (ULAwG)

---

## Strongly convex potential

**Assumption:** $U$ is strongly convex: there exists $m > 0$, such that for all $x, y \in \mathbb{R}^d$,

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

**Outline of the results:**

- Convergence in Wasserstein distance of the semigroup of the diffusion $(P_t)_{t \geq 0}$ (with explicit dependence on the constants $m$ and $L$ and no dependence in the dimension)
- Convergence in Wasserstein distance of the law of the discretized Langevin distribution

**Key technique:** coupling.
Wasserstein distance

Definition

Let $\mu, \nu$ be two probability measures on $\mathbb{R}^d$

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

where $(X,Y) \in \Pi(\mu, \nu)$ if $X \sim \mu$ and $Y \sim \nu$.

- Note by the Cauchy-Schwarz inequality, for all $f : \mathbb{R}^d \to \mathbb{R}$, $\|f\|_{\text{Lip}} \leq 1$, $(X,Y) \in \Pi(\mu, \nu)$,

$$|\mu(f) - \nu(f)| \leq \left\{ \mathbb{E} \left[ \|X - Y\|^2 \right] \right\}^{1/2} \leq W_2 (\mu, \nu).$$
Wasserstein distance convergence

There are many details to fill... This theorem just gives a feeling why Wasserstein distance is well adapted to this particular setting:

**Theorem**

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

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

*The mixing rate depends only on the strong convexity constant.*
The Unadjusted Langevin Algorithm within Gibbs (ULA\textsuperscript{w}G)

Elements of proof

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

where \((Y_0, \tilde{Y}_0) = (x, y)\).

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

\[
d\{Y_t - \tilde{Y}_t\} = - \left\{ \nabla U(Y_t) - \nabla U(\tilde{Y}_t) \right\} dt
\]

we get a very simple SDE for \(\left( \|Y_t - \tilde{Y}_t\|^2 \right)_{t \geq 0}\)

\[
d \left\| Y_t - \tilde{Y}_t \right\|^2 = - \left\langle \nabla U(Y_t) - \nabla U(\tilde{Y}_t), Y_t - \tilde{Y}_t \right\rangle dt.
\]
Elements of proof

Integrating this SDE we get

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

Since \( U \) is strongly convex

\[ \langle \nabla U(y) - \nabla U(y'), y - y' \rangle \geq m \|y - y'\|^2 \]

which implies

\[ \|Y_t - \tilde{Y}_t\|^2 \leq \|Y_0 - \tilde{Y}_0\|^2 - 2m \int_0^t \|Y_s - \tilde{Y}_s\|^2 \, ds. \]
Elements of proof

\[ \| Y_t - \tilde{Y}_t \|^2 \leq \| Y_0 - \tilde{Y}_0 \|^2 - 2m \int_0^t \| Y_s - \tilde{Y}_s \|^2 \, ds . \]

By Grömwall inequality, we obtain

\[ \| Y_t - \tilde{Y}_t \|^2 \leq \| Y_0 - \tilde{Y}_0 \|^2 e^{-2mt} \]

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

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

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

where

\[ x^* = \arg\min_{x \in \mathbb{R}^d} U(x) . \]

The stationary distribution \( \pi \) satisfies

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

The constant depends only linearly in the dimension \( d \).
Elements of proof

- The generator $\mathcal{A}$ 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:

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

- Denote for all $x \in \mathbb{R}^d$ by $V_*(x) = \|x - x^*\|^2$. The process

  $$\left( V_*(Y_t) - V_*(x) - \int_0^t \mathcal{A} V_*(Y_s) ds \right)_{t \geq 0}$$

  is a $(\mathcal{F}_t)_{t \geq 0}$-martingale under $\mathbb{P}_x$.

- Since $\nabla U(x^*) = 0$ and using the strong convexity, we have

  $$\mathcal{A} V_*(x) = 2 \left( - \langle \nabla U(x) - \nabla U(x^*), x - x^* \rangle + d \right) \leq 2 \left( -m V_*(x) + d \right).$$
Elements of proof

Key relation

$$\mathcal{A}V_*(x) \leq 2 (-mV_*(x) + d) .$$

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

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

We have

$$\frac{\partial v(t, x)}{\partial t} = P_t \mathcal{A}V_*(x) \leq -2mP_t V_*(x) + 2d = -2mv(t, x) + 2d ,$$

Grönwall inequality

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

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

$$
\pi(V_\star \wedge c) = \pi P_t(V_\star \wedge c) \leq \pi(P_tV_\star \wedge c)
$$

$$
= \int \pi(dx) c \wedge \left\{ \|x - x^\star\|^2 e^{-2mt} + \frac{d}{m}(1 - e^{-2mt}) \right\}
$$

$$
\leq \pi(V_\star \wedge c)e^{-2mt} + (1 - e^{-2mt})d/m.
$$

Taking the limit as $t \to +\infty$, we get $\pi(V_\star \wedge c) \leq d/m$. 
A coupling proof (I)

- **Objective** compute bound for $W_2(\delta_x Q^n_{\gamma}, \pi)$
- Since $\pi P_t = \pi$ for all $t \geq 0$, it suffices to get some bounds on $W_2(\delta_x Q^n_{\gamma}; \pi P_{\Gamma_n})$, where

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

- **Idea**! Construct a coupling between the diffusion and the linear interpolation of the Euler discretization.
A coupling proof (II)

Idea: use synchronous coupling between the diffusion and a continuously interpolated version of the Euler discretization: \((Y_t, \overline{Y}_t)_{t \geq 0}\) for all \(n \geq 0\) and \(t \in [\Gamma_n, \Gamma_{n+1})\) by

\[
\begin{align*}
Y_t &= Y_{\Gamma_n} - \int_{\Gamma_n}^{t} \nabla U(Y_s)ds + \sqrt{2}(B_t - B_{\Gamma_n}) \\
\overline{Y}_t &= \overline{Y}_{\Gamma_n} - \nabla U(\overline{Y}_{\Gamma_n})(t - \Gamma_n) + \sqrt{2}(B_t - B_{\Gamma_n}),
\end{align*}
\]

with \(Y_0 \sim \pi\) and \(\overline{Y}_0 = x\)

For all \(n \geq 0\), we get

\[
W_2^2(\delta_x, P_{\Gamma_n}, \pi Q^n) \leq \mathbb{E}[\|Y_{\Gamma_n} - \overline{Y}_{\Gamma_n}\|^2],
\]
The Unadjusted Langevin Algorithm within Gibbs (ULAwG)

Explicit bound in Wasserstein distance for the Euler discretisation

**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 $x, y \in \mathbb{R}^d$: $\|\nabla^2 U(x) - \nabla^2 U(y)\| \leq \tilde{L} \|x - y\|$.

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

$$W_2 (\delta_x Q^n_\gamma, \pi) \leq u^{(1)}_n(\gamma) \int_{\mathbb{R}^d} \|y - x\|^2 \pi(dy) + u^{(2)}_n(\gamma),$$
Decreasing step sizes

- If $\lim_{k \to +\infty} \gamma_k = 0$ and $\lim_{k \to +\infty} \Gamma_k = +\infty$, then
  \[
  \lim_{n \to +\infty} W_2(\delta_x Q_n^{\gamma}, \pi) = 0,
  \]
  with explicit control.

- Order of convergence: if $\gamma_k = \gamma_1 k^{-\alpha}$ then
  \[
  W_2(\delta_x Q_n^{\gamma}, \pi) = \mathcal{O}(n^{-\alpha})
  \]
Constant step sizes

- For any $\epsilon > 0$, the minimal number of iterations to achieve $W_2 \left( \delta_x Q^p, \pi \right) \leq \epsilon$ is
  $$ p = \mathcal{O}(\sqrt{d\epsilon^{-1}}) . $$

- For a given stepsize $\gamma$, letting $p \to +\infty$, we get:
  $$ W_2 \left( \pi_\gamma, \pi \right) \leq C\gamma . $$
From the Wasserstein distance to the TV

**Theorem**

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

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

**Proof** Use reflection coupling defined as the unique solution $(X_t, \tilde{X}_t)_{t \geq 0}$ of the SDE:

\[
\begin{align*}
    dX_t &= -\nabla U(X_t)dt + \sqrt{2}dB_t^d \\
    d\tilde{X}_t &= -\nabla U(\tilde{X}_t)dt + \sqrt{2}(\text{Id} - 2e_t e_t^T)dB_t^d, \\
    
    \text{where } e_t &= e(X_t - \tilde{X}_t)
\end{align*}
\]

with $X_0 = x$, $\tilde{X}_0 = y$, $e(z) = z / \|z\|$ for $z \neq 0$ and $e(0) = 0$ otherwise.
From the Wasserstein distance to the TV (II)

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

Consequences:

1. \((P_t)_{t \geq 0}\) converges exponentially fast to \(\pi\) 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

\[ x \mapsto P_t f(x) , \]

is Lipschitz with Lipschitz constant smaller than

\[ 1/\sqrt{2\pi/m}(e^{2mt} - 1) . \]
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 \( x, y \in \mathbb{R}^d \):

\[
\| \nabla^2 U(x) - \nabla^2 U(y) \| \leq \tilde{L} \| x - y \|.
\]

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 \( n \geq 1 \),

\[
\| \delta_xQ^n_{\gamma} - \pi \|_{TV} \leq \tilde{u}_n^{(1)}(\gamma) \int_{\mathbb{R}^d} \| y - x \|^2 \pi(dy) + \tilde{u}_n^{(2)}(\gamma).
\]
Constant step sizes

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

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

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

  $$\| \pi_{\gamma} - \pi \|_{TV} \leq C\gamma |\log(\gamma)| .$$
<table>
<thead>
<tr>
<th>1</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Framework</td>
</tr>
<tr>
<td>3</td>
<td>Strongly log-concave distribution</td>
</tr>
<tr>
<td>4</td>
<td>Convex and Super-exponential densities</td>
</tr>
<tr>
<td>5</td>
<td>Non-smooth potentials</td>
</tr>
<tr>
<td>6</td>
<td>The Unadjusted Langevin Algorithm within Gibbs (ULAwG)</td>
</tr>
</tbody>
</table>
Motivation
Framework
Strongly log-concave distribution
Convex and Super-exponential densities
Non-smooth potentials
The Unadjusted Langevin Algorithm within Gibbs (ULAwG)

Convex potential, decreasing stepsizes

Assumption

- $U$ is convex (but not strongly convex).

Results: decreasing step sizes

- If $\lim_{\gamma_k \to +\infty} \gamma_k = 0$, and $\sum_k \gamma_k = +\infty$ then

$$\lim_{p \to +\infty} \| \delta_x Q^p \gamma - \pi \|_{TV} = 0.$$  

- Computable bounds for the convergence\(^1\).

---

\(^1\)Durmus, Moulines, Annals of Applied Probability, 2016
Convex potential, constant stepsize

Assumption

- $U$ is convex (but not strongly convex).

Results

- For constant stepsize, under one of assumptions above:

$$\|\pi_\gamma - \pi\|_{TV} \leq C \sqrt{\gamma},$$

with computable bound $C$. 
Target precision $\epsilon$: the convex case

- Setting $U$ is convex. Constant stepsize.
- Optimal stepsize $\gamma$ and number of iterations $p$ to achieve $\epsilon$-accuracy in TV:

$$\|\delta_x Q^p_\gamma - \pi\|_{TV} \leq \epsilon.$$

<table>
<thead>
<tr>
<th></th>
<th>$d$</th>
<th>$\epsilon$</th>
<th>$L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>$O(d^{-3})$</td>
<td>$O(\epsilon^2 / \log(\epsilon^{-1}))$</td>
<td>$O(L^{-2})$</td>
</tr>
<tr>
<td>$p$</td>
<td>$O(d^5)$</td>
<td>$O(\epsilon^{-2} \log^2(\epsilon^{-1}))$</td>
<td>$O(L^2)$</td>
</tr>
</tbody>
</table>

- In the strongly convex case, the convergence of the semigroup of the diffusion to $\pi$ depends only on the strong convexity constant $m$. In the convex case, this depends on the dimension $d$. 
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 depend on the dimension.

- Durmus, M. 2016 established that the convergence of the semi-group in TV to $\pi$ does not depend on the dimension but just on $R \sim$ 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 $\gamma$ and number of iterations $p$ to achieve $\epsilon$-accuracy in $\text{TV}$:

$$\|\delta_x Q^p_\gamma - \pi\|_{\text{TV}} \leq \epsilon.$$  

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>$d$</th>
<th>$\epsilon$</th>
<th>$L$</th>
<th>$m$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O(d^{-1})$</td>
<td>$O(\epsilon^2 / \log(\epsilon^{-1}))$</td>
<td>$O(L^{-2})$</td>
<td>$O(m)$</td>
<td>$O(R^{-4})$</td>
<td></td>
</tr>
<tr>
<td>$O(d \log(d))$</td>
<td>$O(\epsilon^{-2} \log^2(\epsilon^{-1}))$</td>
<td>$O(L^2)$</td>
<td>$O(m^{-2})$</td>
<td>$O(R^8)$</td>
<td></td>
</tr>
</tbody>
</table>

Von Dantzig Seminar, Amsterdam
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 The Unadjusted Langevin Algorithm within Gibbs (ULAwG)
Non-smooth potentials

The target distribution has a density $\pi$ with respect to the Lebesgue measure on $\mathbb{R}^d$ of the form $x \mapsto e^{-U(x)}/\int_{\mathbb{R}^d} e^{-U(y)} \, dy$ 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 $\int_{\mathbb{R}^d} e^{-g(y)} \, dy \in (0, +\infty).$
Moreau-Yosida regularization

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

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

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

\[
    x - \text{prox}^\lambda_h(x) \in \gamma \partial h(\text{prox}^\lambda_h(x)) .
\]

- The Moreau-Yosida envelope is a regularized version of \( g \), which approximates \( g \) from below.
Properties of proximal operators

- As $\lambda \downarrow 0$, converges $h^\lambda$ converges pointwise $h$, i.e. for all $x \in \mathbb{R}^d$,
  \[ h^\lambda(x) \uparrow h(x) , \quad \text{as } \lambda \downarrow 0 . \]

- The function $h^\lambda$ is convex and continuously differentiable
  \[ \nabla h^\lambda(x) = \lambda^{-1} (x - \text{prox}_h^\lambda(x)) . \]

- The proximal operator is a monotone operator, for all $x, y \in \mathbb{R}^d$,
  \[ \langle \text{prox}_h^\lambda(x) - \text{prox}_h^\lambda(y), x - y \rangle \geq 0 , \]
  which implies that the Moreau-Yosida envelope is $L$-smooth:
  \[ \| \nabla h^\lambda(x) - \nabla h^\lambda(y) \| \leq \lambda^{-1} \| x - y \| , \text{ for all } x, y \in \mathbb{R}^d. \]
MY regularized potential

- If \( g \) is not differentiable, but the proximal operator associated with \( g \) is available, its \( \lambda \)-Moreau Yosida envelope \( g^\lambda \) 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 \( \lambda > 0 \), \( 0 < \int_{\mathbb{R}^d} e^{-U^\lambda(y)} dy < +\infty \).*
Some approximation results

Theorem

Assume (H).

1. Then, \( \lim_{\lambda \to 0} \| \pi^\lambda - \pi \|_{TV} = 0. \)

2. Assume in addition that \( g \) is Lipschitz. Then for all \( \lambda > 0, \)

\[ \| \pi^\lambda - \pi \|_{TV} \leq \lambda \| g \|_{Lip}^2. \]
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 $\{X^M_k, \ k \in \mathbb{N}\}$: for all $k \geq 0$,

$$X^M_{k+1} = X^M_k - \gamma_{k+1} \left\{ \nabla f(X^M_k) + \lambda^{-1} (X^M_k - \text{prox}_g^\lambda(X^M_k)) \right\} + \sqrt{2\gamma_{k+1}} Z_{k+1},$$

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

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

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

$$\bar{g}^\lambda(x) = g^\lambda(x) - g(x) = g(\text{prox}_g^\lambda(x)) - g(x) + (2\lambda)^{-1} \|x - \text{prox}_g^\lambda(x)\|^2.$$
Image deconvolution

- **Objective** recover an original image \( x \in \mathbb{R}^n \) from a blurred and noisy observed image \( y \in \mathbb{R}^n \) related to \( x \) by the linear observation model \( y = Hx + w \), where \( H \) is a linear operator representing the blur point spread function and \( w \) is a Gaussian vector with zero-mean and covariance matrix \( \sigma^2 I_n \).

- This inverse problem is usually ill-posed or ill-conditioned: exploits prior knowledge about \( x \).

- One of the most widely used image prior for deconvolution problems is the improper total-variation norm prior, \( \pi(x) \propto \exp\left(-\alpha \|\nabla_d x\|_1\right) \), where \( \nabla_d \) denotes the discrete gradient operator that computes the vertical and horizontal differences between neighbour pixels.

\[
\pi(x|y) \propto \exp \left[ -\|y - Hx\|^2 / 2\sigma^2 - \alpha \|\nabla_d x\|_1 \right].
\]
Figure: (a) Original Boat image (256 × 256 pixels), (b) Blurred image, (c) MAP estimate.
Credibility intervals

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).
Motivation

Framework

Strongly log-concave distribution

Convex and Super-exponential densities

Non-smooth potentials

The Unadjusted Langevin Algorithm within Gibbs (ULAwG)
Dependency on the Lipschitz constant

- In all the bounds we have derived, the dependency on the Lipschitz constant $L$ is of order $L^2$.
- In practice, $L$ can be very large!
- In optimization, it can be efficient to use blocking strategies to minimize $U$ using coordinate descent type algorithms.
- Stochastic counterparts are Gibbs samplers!
Gibbs sampler (I)

- **Goal:** simulate a density $\pi$ on $\mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_n}$ for $n \geq 1$ of the form: $(x_1, \cdots, x_n) \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_n}$

  $$\pi(x_1, \cdots, x_n) \propto \exp(-U(x_1, \cdots, x_n)).$$

- Sampling from the full joint density is in general difficult...

- Assume that the **full conditional** densities are known: for all $i \in \{1, \cdots, n\}$, $(x_1, \cdots, x_n) \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_n}$,

  $$\pi(x_i|\mathbf{x}_{\neg i}) = \frac{\pi(x_1, \cdots, x_n)}{\int_{\mathbb{R}^{d_i}} \pi(x_1, \cdots, x_n) dx_i},$$

  Then: a **Gibbs sampler** is probably an sensible way to go!

- **Typical example:** hierarchical models.
Each conditional densities $\pi\left(x_i|x_{-i}\right)$ is associated with a transition kernel $K_i$.

The deterministic scan Gibbs sampler consists in sampling a Markov chain with transition kernel $K_{DS} = K_1 \cdots K_n$, i.e. for $i = 1, \cdots, n$, draw

$$X_{k+1,i} \sim \pi\left(\cdot|X_{k+1,1}, \cdots, X_{k+1,i-1}, X_{k,i+1}, \cdots, X_{k,n}\right).$$

The target density $\pi$ is invariant for the Markov kernel $K_{DS}$!
Let \((a_1, \cdots, a_n) \in (0, 1)^n, \sum_{i=1}^n a_i = 1\), called the selection probability.

The random scan Gibbs sampler consists in sampling a Markov chain with transition kernel \(K_{RS} = \sum_{i=1}^n a_i K_i\), i.e. pick \(I \sim \text{Mult}(a_1, \cdots, a_n)\) and draw

\[
X_{k+1,I} \sim \pi(\cdot | X_{k,-I})
\]

and set for \(j \in \{1, \cdots, n\}, j \neq I, X_{k+1,j} = X_{k,j}\).

The target density \(\pi\) is reversible for the Markov kernel \(K_{RS}\)!
Block Gibbs sampler (I)

- **Goal:** simulate a density $\pi$ on $\mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_n}$ for $n \geq 1$ of the form: $(x_1, \cdots, x_n) \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_n}$ with
  \[ \pi(x_1, \cdots, x_n) \propto \exp \left( -U(x_1, \cdots, x_n) \right). \]

- Let $N \in \{1, \cdots, n\}$ and
  \[ P_{n,N} = \{ \mathcal{I} \subset \{1, \cdots, n\} , \ \text{Card} (\mathcal{I}) = N \}. \]

- For all $\mathcal{I} \in P_{n,N}$,
  \[ \pi (x_{\mathcal{I}} | x_{\overline{\mathcal{I}}}) = \frac{\pi(x_1, \cdots, x_n)}{\int \pi(x_1, \cdots, x_n) dx_{\overline{\mathcal{I}}}}, \]

Here again, using a block Gibbs sampling is appropriate.
Block Gibbs sampler (II)

- For all $\mathcal{I} \in \mathcal{P}_{n,N}$, $\pi(x_{\mathcal{I}}|x_{\mathcal{I}^c})$ is associated with a Markov kernel $K_{\mathcal{I}}$.
- The random scan block Gibbs sampler consists in sampling $K_{\text{RBS}} = \left(\frac{n}{N}\right)^{-1} \sum_{\mathcal{I} \in \mathcal{P}_{n,N}} K_{\mathcal{I}}$.
  1. Given $X_k = (X_{k,1}, \cdots, X_{k,n}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_n}$,
  2. Pick uniformly $\mathcal{I} \in \mathcal{P}_{n,N}$ and draw $X_{k+1,\mathcal{I}} \sim K_{\mathcal{I}}(X_k,\mathcal{I},\cdot)$.
  3. Set for $j \notin \mathcal{I}$, $X_{k+1,j} = X_{k,j}$.
- The target density $\pi$ is reversible for the Markov kernel $K_{\text{RBS}}$!
Each $K_{\mathcal{I}}$ can be replaced by a Markov kernel $\tilde{K}_{\mathcal{I}}$ reversible w.r.t. $\pi(\cdot|x_k,-\mathcal{I})$.

An alternative consists in sampling a Markov chain with transition kernel $\tilde{K}_{\text{RBS}} = \left(\binom{n}{N}\right)^{-1} \sum_{\mathcal{I} \in \mathcal{P}_{n,N}} \tilde{K}_{\mathcal{I}}$.

1. Given $X_k = (X_{k,1}, \ldots, X_{k,n}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_n}$,
2. Pick uniformly $\mathcal{I} \in \mathcal{P}_{n,N}$ and draw $X_{k+1,\mathcal{I}} \sim \tilde{K}_{\mathcal{I}}(X_k,\cdot)$.
3. Set for $j \notin \mathcal{I}$, $X_{k+1,j} = X_{k,j}$.

The target density $\pi$ is reversible for the Markov kernel $\tilde{K}_{\text{RBS}}$!

Example: Metropolis within Gibbs algorithm.
The ideal Langevin within Gibbs samplers

- **Idea:** take for $\tilde{K}_I$ the Langevin semigroup taken at time $t_I \geq 0$, $P_{t_I}$ associated with the distribution $\pi (\cdot | x_{k,-I})$.

- **An ideal algorithm** Sample the Markov kernel $\tilde{K}_{RBS} = (\frac{n}{N})^{-1} \sum_{I \in \mathcal{P}_{n,N}} P_{t_I}$.

1. Given $X_k = (X_{k,1}, \ldots, X_{k,n}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_n}$,
2. Pick uniformly $I \in \mathcal{P}_{n,N}$ and draw $X_{k+1,I} \sim P_{t_I} (X_k, \cdot)$
3. Set for $j \notin I$, $X_{k+1,j} = X_{k,j}$.

- **Problem:** Cannot simulate from $P_{t_I}$ !

- **Solution** Take the kernel of the Euler discretisation instead.
Motivation
Framework
Strongly log-concave distribution
Convex and Super-exponential densities
Non-smooth potentials

The Unadjusted Langevin Algorithm within Gibbs (ULAwG)

The Unadjusted Langevin Algorithm within Gibbs samplers

- **Idea:** Replace $P_{t,\mathcal{I}}$ by its Euler discretization after $p$ steps $(R_{\gamma,\mathcal{I}})^p$.
- The discretization parameter $\gamma,\mathcal{I}$ might depend on the block.
- The ULAwG consists in sampling a Markov kernel
  \[
  \tilde{K}_{\text{RBS}} = \left(\frac{n}{N}\right)^{-1} \sum_{\mathcal{I} \in \mathcal{P}_{n,N}} (R_{\gamma,\mathcal{I}})^p.
  \]

1. Given $X_k = (X_{k,1}, \cdots, X_{k,n}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_n}$,
2. Pick uniformly $\mathcal{I} \in \mathcal{P}_{n,N}$ and set $Y_0 = X_{k,\mathcal{I}}$.
3. for $i = 1, \cdots, p$, compute
   \[
   Y_i = Y_{i-1} - \gamma,\mathcal{I} \nabla U(Y_{i-1}|X_{k,-\mathcal{I}}) + \sqrt{2\gamma,\mathcal{I}}Z_i.
   \]
4. Set $X_{k+1,\mathcal{I}} = Y_p$.
5. Set for $j \not\in \mathcal{I}$, $X_{k+1,j} = X_{k,j}$. 

Von Dantzig Seminar, Amsterdam
A toy example: the Gaussian linear model

\[ Y = A\beta + Z. \]

\( A \) is a known design matrix and \( Z \sim \mathcal{N}(0, \sigma_Z^2 \text{Id}) \)

Prior distribution for \( \beta \sim \mathcal{N}(0, \Sigma_\beta) \)

The posterior distribution is Gaussian with mean and covariance given by

\[
\Sigma = \left( \Sigma_\beta^{-1} + \sigma_z^{-2} A^T A \right)^{-1} \\
\mu = \sigma_z^{-2} \Sigma A^T Y.
\]

Compare the efficiency of ULA and ULAwG to estimate \( \Sigma_{1,1} \).
A toy example: the Gaussian linear model (III)

Synthetic data and for $d = 10$, $\sigma^2_z = 1$, $\sigma_\beta = 100$ and $N = 2$. 
Large-Scale Matrix Factorization

- We applied ULAwG on a large-scale matrix factorization problem for a link prediction application.
- Consider $X$ a matrix with (many) missing entries of size $I \times J$. The model is for observed indexes $i, j$

$$X_{i,j} = \sum_{k=1}^{K} W_{i,k} H_{k,j} + Z_{i,j},$$

where $K \geq 0$ is the rank, and $(Z_{i,j}) \sim_{i.i.d.} \mathcal{N}(0, \sigma_z^2)$. 
Large-Scale Matrix Factorization (II)

- The aim is then to infer the two matrices $W$ and $H$ of dimensions $I \times K$ and $K \times J$ respectively to predict the missing values of $X$.
- We take as prior distributions:

$$W_{j,k} \sim \mathcal{N}(0, \sigma_w^2) \quad \text{and} \quad H_{k,j} \sim \mathcal{N}(0, \sigma_h^2).$$

- Comparison of ULA and ULAwG on the MovieLens 1 Million dataset (1,000,209 notes pour 3,900 films notés par 6,040 utilisateurs de MovieLens, notes 0-5) \(^2\).

\(^2\)A. Durmus, U. Simsekli, M., NIPS2016
Large-Scale Matrix Factorization (III)

Paramètres:
\[ \sigma^2_z = 1, \]
\[ \sigma^2_w = \sigma^2_h = 100 \]
\[ N = I \times J/100. \]
Large-Scale Matrix Factorization (IV)

- **Motivation**
- **Framework**
- Strongly log-concave distribution
- Convex and Super-exponential densities
- Non-smooth potentials

The Unadjusted Langevin Algorithm within Gibbs (ULAwG)

**Paramètres:**
\[
\sigma_z^2 = 1, \\
\sigma_w^2 = \sigma_h^2 = 100 \\
N = \lceil I \times J/25 \rceil \\
\text{and batch size} \\
\lceil N_{\text{obs}}/25 \rceil.
\]