Langevin MCMC: Theory and Methods
Bayesian Computation Opening Workshop

A. Durmus¹, N. Brosse², E. Moulines², M. Pereyra³, S. Sabanis⁴

¹ENS Paris-Saclay
²Ecole Polytechnique
³Heriot-Watt University
⁴University of Edinburgh
Outline

1 Motivation and setting

2 The ULA algorithm

3 The Stochastic Gradient Langevin dynamics (SGLD)

4 Strongly log-concave distribution

5 Convex and Super-exponential densities

6 Non-smooth potentials

7 Logconcave densities with constrained domains

8 Normalizing constants of log-concave densities
Introduction

- Sampling distribution over high-dimensional state-space has recently attracted a lot of research efforts in computational statistics and machine learning community...

- **Applications** (non-exhaustive)
  1. Bayesian inference for high-dimensional models,
  2. Bayesian inverse problems (e.g., image restoration and deblurring),
  3. Aggregation of estimators and experts,
  4. Bayesian non-parametrics.

- Most of the sampling techniques known so far do not scale to high-dimension... Challenges are numerous in this area...
A Bayesian model is specified by
(i) the likelihood of the observed data: \( D \sim L(\cdot|x) \)
(ii) a prior distribution \( \pi_0 \) on the parameter space \( x \in X \subset \mathbb{R}^d \)

The inference is based on the posterior distribution:

\[
\pi(x|D) = \frac{\pi_0(x)L(D|x)}{\int L(D|u)\pi_0(u)du}.
\]

In most cases the normalizing constant is not tractable:

\[
\pi(x|D) \propto \pi_0(x)L(D|x).
\]
An example: Bayesian analysis of logistic regression

- **Likelihood**: Binary regression set-up in which the binary observations (responses) $\{R_i\}_{i=1}^n$ are conditionally independent Bernoulli random variables with success probability $\{F(x^T Z_i)\}_{i=1}^n$, where
  1. $Z_i$ is a $d$ dimensional vector of known covariates,
  2. $x$ is a $d$ dimensional vector of unknown regression coefficients
  3. $F$ is the **link function**: in **logistic regression**, $F$ is the standard logistic cumulative distribution function:

$$F(t) = e^t/(1 + e^t)$$
An example: Bayesian analysis of logistic regression

- Possible choices for the prior (tons of theory behind these, this is not black magic!)
  - Gaussian prior $\pi_0(x) \propto \exp\left(-\frac{1}{2}x'\Sigma^{-1}x\right)$
  - LASSO prior $\pi_0(x) = \exp\left(-\sum_{i=1}^{d} \alpha_i |x_i|\right)$
  - Many more sophisticated ones (spike-and-slab, horseshoe,...) and other global-local shrinkage priors...
- The posterior of $x$ is up to a proportionality constant given by

$$\pi(x|(R, Z)) \propto \prod_{i=1}^{n} F^{R_i}(x'Z_i)(1 - F(x'Z_i))^{1-R_i} \pi_0(x)$$
Bayesian statistics are out of the game?

- While optimization-based algorithms for the extremely popular Lasso and elastic net procedures can scale to dimension in the hundreds of thousands, corresponding Bayesian methods that use Markov chain Monte Carlo (MCMC) for computation are limited to problems at least an order of magnitude smaller.

- The promise of scaling MCMC to large datasets has thus far not been realized, owing to the expensive computations often involved in MCMC algorithms and slow mixing of the corresponding Markov chains.

- (long term) Objective:
  - Contribute to fill the gap between optimization and simulation.
  - Develop sampling methods which can provably work in high-dimension (more precise statements later)

- For logistic regression with horse-shoe prior, on going work with James Johndrow
Outline

1. Motivation and setting
2. The ULA algorithm
3. The Stochastic Gradient Langevin dynamics (SGLD)
4. Strongly log-concave distribution
5. Convex and Super-exponential densities
6. Non-smooth potentials
7. Logconcave densities with constrained domains
8. Normalizing constants of log-concave densities
Denote by $\pi$ a target density w.r.t. the Lebesgue measure on $\mathbb{R}^d$, known up to a normalisation factor

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

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

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

Note: this condition can be removed by taming the gradient (Brosse, Durmus, M., Sabanis, 2018, to appear SPA 2019)
(Overdamped) Langevin diffusion

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.

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

Notation: \((P_t)_{t \geq 0}\) the Markov semigroup associated to the Langevin diffusion:
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. \(N(0, I_d)\)
- \((\gamma_k)_{k \geq 1}\) 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.
- This algorithm is referred to as the Unadjusted Langevin Algorithm (ULA) in Bayesian statistics or Langevin Monte Carlo (LMC) in machine learning.
Discretized Langevin diffusion: constant stepsize

- When the stepsize is held constant, i.e. $\gamma_k = \gamma$, then $(X_k)_{k \geq 1}$ is an homogeneous Markov chain with Markov kernel $R_\gamma$
- Under some appropriate conditions, $R_\gamma$ is irreducible, positive recurrent $\Rightarrow$ unique invariant distribution $\pi_\gamma$ which does not coincide with the target distribution $\pi$.

Questions:
- For a given precision $\epsilon > 0$, how to chose the stepsize $\gamma > 0$ and the number of iterations $n$ so that: $\|\delta_x R_\gamma^n - \pi\|_{TV} \leq \epsilon$?
- How to select the starting point $x$ cleverly?
- How to quantify the distance between $\pi_\gamma$ and $\pi$?
Discretized Langevin diffusion: decreasing stepsize

- When \((\gamma_k)_{k \geq 1}\) is nonincreasing and non constant, \((X_k)_{k \geq 1}\) is an inhomogeneous Markov chain associated with the kernels \((R_{\gamma_k})_{k \geq 1}\).

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

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

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

- **Questions:**
  - **Convergence:** Can the step sizes be selected so that \(\|\delta_x Q^p_{\gamma} - \pi\|_{TV} \to 0\)?
  - **Optimal choice of simulation parameters:** How to select the number of iterations to achieve \(\|\delta_x Q^p_{\gamma} - \pi\|_{TV} \leq \epsilon\) starting from a given point \(x\)?
  - Should we use fixed or decreasing step sizes?
Some (very incomplete) references: Early references


(iii) Most of these results are [qualitative](#) (e.g. conditions upon which the sampler is geometrically ergodic).
Some (very incomplete) references: Euler discretisation


Outline

1 Motivation and setting
2 The ULA algorithm
3 The Stochastic Gradient Langevin dynamics (SGLD)
4 Strongly log-concave distribution
5 Convex and Super-exponential densities
6 Non-smooth potentials
7 Logconcave densities with constrained domains
8 Normalizing constants of log-concave densities
SGLD Algorithm

- **Objective:** posterior inference on large scale datasets (many machine learning applications).
- The target $\pi$ is the posterior distribution in a Bayesian inference problem with prior density $\pi_0(x)$ and a large number $N \gg 1$ of i.i.d. observations $D_i$ with likelihoods $L(D_i|x)$:

$$\pi(x) = \pi_0(x) \prod_{i=1}^{N} L(D_i|x).$$

- The cost of one iteration is $Nd$ which is prohibitively large for massive datasets.
- **Notations:**
  - Negated log-likelihood for a single observation: $U_i(x) = -\log(L(D_i|x))$ for $i \in \{1, \ldots, N\}$, negated likelihood of $N$ observations: $U = \sum_{i=0}^{N} U_i$.
  - Negated log-prior: $U_0(x) = -\log(\pi_0(x))$.
Welling and Teh (2011) suggested to replace $\nabla U$ with an unbiased estimate

$$\nabla U_0 + \left( \frac{N}{p} \right) \sum_{i \in S} \nabla U_i$$

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

A single update of SGLD is then given by

$$X_{k+1} = X_k - \gamma \left( \nabla U_0(X_k) + \frac{N}{p} \sum_{i \in S_{k+1}} \nabla U_i(X_k) \right) + \sqrt{2\gamma} Z_{k+1}.$$

The idea of using only a fraction of the observations to compute an unbiased estimate of the gradient at each iteration comes from Stochastic Gradient Descent (SGD) which is a popular algorithm in machine learning to minimize the potential $U$. 
SGLD and control variates

- SGLD scales to large datasets by using noisy gradients calculated using a mini-batch or subset of the dataset.
- However, the high variance inherent in these noisy gradients degrades performance and leads to slower mixing.
- Idea: Use variance reduction for the unbiased estimator of the gradient.
- the Stochastic Variance Reduced Gradient Langevin Dynamics (SVR-GLD) (Dubey et al, 2016):

\[
X_{k+1} = X_k - \gamma \left( \nabla U_0(X_k) - g_{0,k} + \frac{N}{p} \sum_{i \in S_{k+1}} \{ \nabla U_i(X_k) - g_{i,k} \} \right) + \sqrt{2\gamma} Z_{k+1}.
\]

where \( g_{i,k} \) are updated every \( m \) iterations, i.e.

\[
g_{i,k} = \begin{cases} 
  g_{i,k-1} & k \neq 0 \mod m \\
  \nabla U_i(X_k) & \text{otherwise}
\end{cases}
\]
Some (very incomplete) references: SGLD, SVR-GLD


(iv) Dubey, A; Reddi, S; Williamson, S. ; Poczos, B.; Smola, A.; Xing E., *Variance reduction in stochastic gradient Langevin dynamics*, Advances in neural information processing systems,


Outline

1. Motivation and setting
2. The ULA algorithm
3. The Stochastic Gradient Langevin dynamics (SGLD)
4. Strongly log-concave distribution
5. Convex and Super-exponential densities
6. Non-smooth potentials
7. Log-concave densities with constrained domains
8. Normalizing constants of log-concave densities
**Strongly convex potential**

- **Assumption**: $U$ is $L$-smooth and $m$-strongly convex
  
  $$
  \|\nabla U(x) - \nabla U(y)\| \leq L \|x - y\|
  $$
  $$
  \langle \nabla U(x) - \nabla U(y), x - y \rangle \geq m \|x - y\|^2.
  $$

- **Outline of the proof**
  2. Relating Wasserstein distance result to total variation.

- **Key technique**: (Synchronous and Reflection) coupling!
Wasserstein distance

**Definition 1**

Let $\xi, \xi'$ be two probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The Wasserstein distance of order 2 is given by

$$W^2_2(\xi, \xi') = \inf_{\zeta \in \mathcal{C}(\xi, \xi')} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - x'\|^2 \zeta(dx dx') ,$$

$$= \inf_{\zeta \in \mathcal{C}(\xi, \xi')} \mathbb{E} \left[ \|X - X'\|^2 \right] ,$$

where $(X, X') \sim \zeta$. 

Wasserstein distance convergence

**Theorem 2**

Assume that $U$ is $L$-smooth and $m$-strongly convex.

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

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

For all $x \in \mathbb{R}^d$ and $t \geq 0$,

$$W_2(\delta_x P_t, \delta_y P_t) \leq e^{-mt} \left\{ \|x - x^*\| + \sqrt{d/m} \right\},$$

where $x^* = \operatorname{arg\, min}_{\mathbb{R}^d} U$.

The contraction depends only on the strong convexity constant.
Synchronous Coupling

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

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

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

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

The product rule for semimartingales imply

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

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

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

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

Grömwall inequality:

\[ \| Y_t - \tilde{Y}_t \|^2 \leq \| Y_0 - \tilde{Y}_0 \|^2 e^{-2mt} \]
Theorem 3

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

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

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

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

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

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

Objective compute bound for $W_2(\delta_x Q^n, \pi)$

Since $\pi P_t = \pi$ for all $t \geq 0$, it suffices to get bounds of the Wasserstein distance

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

- $\Gamma_n = \sum_{k=1}^n \gamma_k$,
- $\delta_x Q^n$: law of the discretized diffusion,
- $\pi P_{\gamma_n} = \pi$, where $(P_t)_{t \geq 0}$ is the semi group of the diffusion

Idea! synchronous coupling between the diffusion and the interpolation of the Euler discretization.
A coupling proof (II)

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

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

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

For all $n \geq 0$,

\[
W_2^2(\delta_x P_{\Gamma_n}, \pi Q^n_{\gamma}) \leq \mathbb{E}[\|Y_{\Gamma_n} - \bar{Y}_{\Gamma_n}\|^2],
\]
Explicit bound in Wasserstein distance

**Theorem 4**

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

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

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

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

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

- **Fixed step size** For any $\epsilon > 0$, one may choose $\gamma$ so that
  \[
  W_2(\delta_{x^*} R^p_\gamma, \pi) \leq \epsilon \quad \text{in } p = O(d\epsilon^{-1}) \text{ iterations}
  \]
  where $x^*$ is the unique maximum of $\pi$

- **Decreasing step size** with $\gamma_k = \gamma_1 k^{-\alpha}$, $\alpha \in (0, 1)$,
  \[
  W_2(\delta_{x^*} Q^n_\gamma, \pi) = dO(n^{-\alpha})
  \]

- **Our results are tight** (check with $U(x) = (1/2)\|x\|^2$).
Theorem 5

Assume that $U$ is strongly convex.

(i) For all $x, y \in \mathbb{R}^d$,

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

where $\Phi$ is the standard Gaussian cumulative distribution function.

(ii) For any $\mu, \nu$ and $t > 0$,

$$\|\mu P_t - \nu P_t\|_{TV} \leq 2^{1/2} W_1(\mu, \nu) \left/ \left(\pi^{1/2} \sqrt{(4/m)(e^{2mt} - 1)} \right) \right.,$$

Use reflection coupling (Lindvall and Rogers, 1986)
Hints of Proof I

\[
\begin{aligned}
\begin{cases}
    dX_t &= -\nabla U(X_t)dt + \sqrt{2}dB^d_t \\
    dY_t &= -\nabla U(Y_t)dt + \sqrt{2}(Id - 2e_t e^T_t)dB^d_t,
\end{cases}
\end{aligned}
\]

where \( e_t = e(X_t - Y_t) \)

with \( X_0 = x, Y_0 = y, e(z) = z/\|z\| \) for \( z \neq 0 \) and \( e(0) = 0 \) otherwise.

Define the coupling time \( T_c = \inf\{ s \geq 0 \mid X_s \neq Y_s \} \). By construction \( X_t = Y_t \) for \( t \geq T_c \).

\[
\tilde{B}^d_t = \int_0^t (Id - 2e_s e^T_s)dB^d_s
\]

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

\[
\|P_t(x, \cdot) - P_t(y, \cdot)\|_{TV} \leq \mathbb{P}(X_t \neq Y_t).
\]
Hints of Proof II

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

$$d\{X_t - Y_t\} = -\{\nabla U(X_t) - \nabla U(Y_t)\} \, dt + 2\sqrt{2}e_t \, dB_t^1.$$  

Using Itô’s formula

$$\|X_t - Y_t\| = \|x - y\| - \int_0^t \langle \nabla U(X_s) - \nabla U(Y_s), e_s \rangle \, ds + 2\sqrt{2}B_t^1$$

$$\leq \|x - y\| - m \int_0^t \|X_s - Y_s\| \, ds + 2\sqrt{2}B_t^1.$$  

and Grönwall’s inequality implies

$$\|X_t - Y_t\| \leq e^{-mt} \|x - y\| + 2\sqrt{2}B_t^1 - m2\sqrt{2} \int_0^t B_s^1 e^{-m(t-s)} \, ds.$$
Therefore by integration by part, $\|X_t - Y_t\| \leq U_t$ where $(U_t)_{t \in (0, T_c)}$ is the one-dimensional Ornstein-Uhlenbeck process defined by

$$U_t = e^{-mt} \|x - y\| + 2\sqrt{2}e^{-mt} \int_0^t e^{ms} dB_s^1$$

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

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

Finally the proof follows from the tail of the hitting time of (one-dimensional) OU (see Borodin and Salminen, 2002).
\[ \| P_t(x, \cdot) - P_t(y, \cdot) \|_{TV} \leq \frac{\| x - y \|}{\sqrt{(2\pi/m)(e^{2mt} - 1)}} \]

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 the function \(x \mapsto P_t f(x)\) is Lipschitz with Lipschitz constant smaller than

\[ 1/\sqrt{(2\pi/m)(e^{2mt} - 1)} \].
Explicit bound in total variation

**Theorem 6**

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

$$
\|\delta_x Q^n_\gamma - \pi\|_{TV} \leq \tilde{u}^{(1)}_n(\gamma) \left\{\|x - x^*\|^2 + d/m\right\} + \tilde{u}^{(2)}_n(\gamma).
$$
Constant step sizes

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

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

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

$$\|\pi_{\gamma} - \pi\|_{TV} \leq C_{\gamma} |\log(\gamma)|.$$

Outline

1. Motivation and setting
2. The ULA algorithm
3. The Stochastic Gradient Langevin dynamics (SGLD)
4. Strongly log-concave distribution
5. Convex and Super-exponential densities
6. Non-smooth potentials
7. Logconcave densities with constrained domains
8. Normalizing constants of log-concave densities
Convergence of the Euler discretization

Assume one of the following conditions:

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

- $U$ is convex.

Results

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

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

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

<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>

**Table:** In the strongly convex case, $d$!
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, Moulines (2016) established that the convergence of the semi-group in TV to $\pi$ does not depend on the dimension but just on $R \rightsquigarrow$ 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 TV:

$$\|\delta x R^p_{\gamma} - \pi\|_{TV} \leq \epsilon.$$

<table>
<thead>
<tr>
<th></th>
<th>$d$</th>
<th>$\epsilon$</th>
<th>$L$</th>
<th>$m$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<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>
</tr>
<tr>
<td>$p$</td>
<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>
</tr>
</tbody>
</table>

*Table:* Of course, there is a price to pay in the dependence in $R$
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:
  \[
  \|\delta_{x^\star} R_\gamma^p - \pi\|_{TV} \leq \epsilon.
  \]

- starting point $x^\star \in \arg \min_{\mathbb{R}^d} U$

<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>$O(\frac{\epsilon^2}{\log(\epsilon^{-1})})$</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>
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, Moulines (2016) established that the convergence of the semi-group in TV to $\pi$ does not depend on the dimension but just on $R \rightsquigarrow$ new bounds which scale nicely in the dimension.
Dependence on the dimension

- Setting \( U \) is convex and strongly convex outside a ball. Constant stepsize
- Optimal stepsize \( \gamma \) and number of iterations \( p \) to achieve \( \epsilon \)-accuracy in TV:

\[
\|
\delta_{x^*} R^p_{\gamma} - \pi
\|_{TV} \leq \epsilon.
\]

- Starting point \( x^* \in \arg \min_{\mathbb{R}^d} U \)

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

1. Motivation and setting
2. The ULA algorithm
3. The Stochastic Gradient Langevin dynamics (SGLD)
4. Strongly log-concave distribution
5. Convex and Super-exponential densities
6. Non-smooth potentials
7. Logconcave densities with constrained domains
8. Normalizing constants of log-concave densities
Non-smooth potential

- **Problem:** $U$ is convex but not continuously differentiable? More precisely, assume that
  \[ \pi \propto e^{-U}, \quad U = f + g, \]
  where $f$ is convex and smooth and $g$ is convex but not smooth.

- **Applications:**
  1. LASSO $g(x) = \sum_{i=1}^{d} |x_i|$, fused-LASSO models,
     \[ g(x) = \sum_{i=1}^{d} |x_i - x_{i-1}|. \]
  2. Constrained domain,
     \[ g(x) = 1_{\left\{ \sum_{i=1}^{n} |x_i| \leq \kappa \right\}} \quad \text{with} \quad 1_{\mathcal{K}} \overset{\text{def}}{=} \begin{cases} +\infty & \text{if } x \notin \mathcal{K}, \\ 0 & \text{if } x \in \mathcal{K}. \end{cases} \]
Non-smooth potential

- **Idea:** apply EM discretization, (Durmus, M., Pereyra, 2018) to an appropriately regularized version of $g$
  
  (i) the convexity of $U$ is preserved
  
  (ii) the regularisation of $U$ is continuously differentiable and gradient Lipschitz
  
  (iii) the resulting approximation is close to $\pi$ (e.g. in total variation norm)
Assume that \( g : \mathbb{R}^d \to (-\infty, +\infty] \) is a l.s.c convex function and let \( \lambda > 0 \).

The \( \lambda \)-Moreau-Yosida envelope \( g^\lambda : \mathbb{R}^d \to \mathbb{R} \) is defined for all \( x \in \mathbb{R}^d \) by

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

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

\[
x - \text{prox}_g^\lambda(x) \in \gamma \partial g(\text{prox}_g^\lambda(x)) .
\]

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

- As $\lambda \downarrow 0$, converges $g^\lambda$ 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 $g^\lambda$ is convex and continuously differentiable

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

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

$$\langle \text{prox}_g^\lambda(x) - \text{prox}_g^\lambda(y), x - y \rangle \geq 0,$$

which implies that the Moreau-Yosida envelope is $L$-smooth: for all $x, y \in \mathbb{R}^d$

$$\|\nabla g^\lambda(x) - \nabla g^\lambda(y)\| \leq \lambda^{-1}\|x - y\|.$$
Moreau-Yosida regularization

- 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 $U^\lambda : \mathbb{R}^d \rightarrow \mathbb{R}$ defined for all $x \in \mathbb{R}^d$ by
  \[
  U^\lambda(x) = f(x) + g^\lambda(x).
  \]

**Question:** Does it make some sense to use $U^\lambda$ for targeting $\pi \propto e^{-U}$?
Assumptions

H1

- $\pi \propto e^{-U}$, $U = f + g$
- $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and $g : \mathbb{R}^d \rightarrow (-\infty, +\infty]$ 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 $\int_{\mathbb{R}^d} e^{-g(y)} \, dy \in (0, +\infty)$. 
Properties of proximal operators and consequences

- The function $g^\lambda$ are convex and continuously differentiable.
- $g^\lambda$ is gradient-Lipschitz: for all $x, y \in \mathbb{R}^d$,
  \[
  \| \nabla g^\lambda(x) - \nabla g^\lambda(y) \| \leq \lambda^{-1} \| x - y \|
  \]

Consequence: The function $U^\lambda$ is convex, continuously differentiable and gradient-Lipschitz.
Properties of proximal operators and consequences

**Question** Can $U^\lambda$ be used instead of $U$ in ULA (or any MCMC algorithm that uses the gradient of $U$)?

$U^\lambda$ defines a regularized distribution $\pi^\lambda$

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

A minimal requirement is that $\pi^\lambda \propto e^{-U^\lambda}$ is a probability density function!

**Theorem 7** (Durmus,Moulines,Pereyra) SIAM J. Imaging Sci., 2018

*Under $H1$, for all $\lambda > 0$, $0 < \int_{\mathbb{R}^d} e^{-U^\lambda(y)} dy < +\infty.$*
Some approximation results

- $U^\lambda$ defines a regularized version of $\pi^\lambda$

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

**Question:** We now that $\lim_{\lambda \downarrow 0^+} U^\lambda(x) = U(x)$. Is this sufficient to guarantee that $\pi^\lambda$ is a sensible approximation of $\pi$?

**Theorem 8** ((Durmus, Moulines, Pereyra) SIAM J. Imaging Sci., 2018)

**Assume H1.**

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

$$\|\pi^\lambda - \pi\|_{TV} \leq \lambda \|g\|^2_{\text{Lip}}.$$
Moreau-Yoshida approximations

\[ p(x) \propto \exp(-|x|) \]

\[ p(x) \propto \exp(-x^4) \]

\[ p(x) \propto 1_{[-0.5, 0.5]}(x) \]

**Figure:** True densities (solid blue) and approximations (dashed red).
The MYULA algorithm

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

  \[
  X_{k+1}^M = X_k^M - \gamma \left\{ \nabla f(X_k^M) + \lambda^{-1}(X_k^M - \text{prox}^\lambda_g(X_k^M)) \right\} + \sqrt{2}Z_{k+1},
  \]

  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 $4\mu m \times 4\mu m$ containing 100 molecules, (a) Original Observation, (b) MAP

Objective 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 $w \sim \mathcal{N}(0, \sigma^2 I_n)$. This inverse problem is ill-conditioned, but this problem can be mitigated by exploiting prior knowledge.
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 $w \sim \mathcal{N}(0, \sigma^2 I_n)$.

This inverse problem is ill-posed, resulting in significant uncertainty about the true value of $x$.

Idea Use a LASSO (or horse-shoe) prior to promote sparse images (nearly-black object). The resulting posterior is

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

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.
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 $\varphi$ is a function.

This is formalised in the Bayesian decision theory framework by computing highest posterior density (HPD)

**Problem**: compute posterior credibility sets for $\varphi(x)$. 

Comparison with PMALA

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

\[ \hat{x}_{MAP} \]

**Figure**: Live-cell microscopy data *Zhu2012*. Uncertainty analysis (±78nm × ±125nm)

Computing time 4 minutes. \( M = 10^5 \) iterations. Estimation error 0.2%..
Outline

1. Motivation and setting
2. The ULA algorithm
3. The Stochastic Gradient Langevin dynamics (SGLD)
4. Strongly log-concave distribution
5. Convex and Super-exponential densities
6. Non-smooth potentials
7. Logconcave densities with constrained domains
8. Normalizing constants of log-concave densities
Densities with convex support

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

\[
\pi \propto e^{-U}, \quad U = f + \iota_{\mathcal{K}}, \quad \iota_{\mathcal{K}}(x) = \begin{cases} +\infty & \text{if } x \notin \mathcal{K}, \\ 0 & \text{if } x \in \mathcal{K}. \end{cases}
\]

where $f$ is smooth.

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

\[
\iota_{\mathcal{K}}^\lambda(x) = \inf_{y \in \mathbb{R}^d} \left( \iota_{\mathcal{K}}(y) + (2\lambda)^{-1} \|x - y\|^2 \right) = (2\lambda)^{-1} \|x - \text{proj}_{\mathcal{K}}(x)\|^2.
\]
Previous works

- **Previous work** for the Metropolis algorithm and the hit-and-run: Applegate, Kannan, Dyer, Frieze, Polson, Simonovits, Lovász, Vempala...
- Our approach is more in the spirit of Bubeck, Eldan, and Lehec (2015) based on the Projected Langevin Monte Carlo
Main results - Assumptions

**H2**

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

**H3**

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


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,

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

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

- Similar bounds hold for the Wasserstein distance.
Comparison with existing results

- Lovasz (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 difficult to check).
- Bubeck et al (2015) studies the complexity of the Projected Langevin Monte Carlo algorithm (PLMC):

$$X_{k+1} = \text{proj}_K (X_k - \gamma \nabla f(X_k)) .$$

Note that contrary to MYULA, the iterates of PLMC always belong to $\mathcal{K}$. Under similar assumptions, Bubeck et al. (2015) get explicit bounds in total variation for PLMC:

<table>
<thead>
<tr>
<th></th>
<th>$d \to +\infty$</th>
<th>$\varepsilon \to 0$</th>
<th>$R \to +\infty$</th>
<th>$r \to 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bubeck:2015</strong> $\pi$ uniform on $\mathcal{K}$</td>
<td>$\tilde{O}(d^7)$</td>
<td>$\tilde{O}(\varepsilon^{-8})$</td>
<td>$\tilde{O}(R^6)$</td>
<td>$\tilde{O}(r^{-6})$</td>
</tr>
<tr>
<td><strong>Bubeck:2015</strong> $\pi$ log concave</td>
<td>$\tilde{O}(d^{12})$</td>
<td>$\tilde{O}(\varepsilon^{-12})$</td>
<td>$\tilde{O}(R^{18})$</td>
<td>$\tilde{O}(r^{-18})$</td>
</tr>
<tr>
<td>MYULA</td>
<td>$\tilde{O}(d^5)$</td>
<td>$\tilde{O}(\varepsilon^{-6})$</td>
<td>$\tilde{O}(R^4)$</td>
<td>$\tilde{O}(r^{-4})$</td>
</tr>
</tbody>
</table>

**Table:** Complexity $\| \delta_{x^*} R^n_{\gamma, \lambda} - \pi \|_{TV} \leq \varepsilon$
Application to regression with $\ell_1$ constraints

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

$$U^s(x) = \exp \left( -\frac{\|Y - Zx\|^2}{2\sigma^2} - \iota_{K^s}(x) \right), \quad K^s = \{x \in \mathbb{R}^d; \|x\|_1 \leq s\}.$$

2. Dual problem of LASSO regression in optimization.

3. We compute for all $i \in \{1, \cdots, d\}$, the median of $x_i$ for different values of $s$ on the diabetes data set ($n = 442, d = 10$).

4. Compute the LASSO regularization paths.
Application to regression with $\ell_1$ constraints

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

1. Motivation and setting
2. The ULA algorithm
3. The Stochastic Gradient Langevin dynamics (SGLD)
4. Strongly log-concave distribution
5. Convex and Super-exponential densities
6. Non-smooth potentials
7. Logconcave densities with constrained domains
8. Normalizing constants of log-concave densities
Normalizing constants

Let \( U : \mathbb{R}^d \to \mathbb{R} \). We aim at estimating \( Z = \int_{\mathbb{R}^d} e^{-U(x)} \, dx < +\infty \).

\( Z \) is the normalizing constant of the probability density \( \pi \) 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.

Many methods... But few theoretical guarantees.

Assumption \( U \) is a continuously differentiable convex function, \( \text{min } U = 0 \).
Multistage sampling

- **Idea:** decompose the original problem in a sequence of problems which are easier to solve.
- **Multistage sampling method** (Gelman, 1998),

\[
\frac{Z}{Z_0} = \prod_{i=0}^{M-1} \frac{Z_{i+1}}{Z_i},
\]

where

1. \( M \in \mathbb{N}^* \) is the number of stages,
2. \( Z_0 \) is the initial normalizing constant (should be easy to compute)
3. \( Z_{i+1}/Z_i \) are the ratios of normalisation constants (that should also be easy to estimate).
A Gaussian annealing algorithm

- \( M \in \mathbb{N}^* \) 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 \( \sigma_0 \) 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
  \[
  \pi_i(x) = Z_i^{-1} e^{-U_i(x)}, \quad Z_i = \int_{\mathbb{R}^d} e^{-U_i(y)} dy.
  \]

- It defines \( (Z_i)_{i=1}^M \) in the decomposition
  \[
  \frac{Z}{Z_0} = \prod_{i=0}^{M-1} \frac{Z_{i+1}}{Z_i},
  \]

- For \( i \in \{0, \ldots, M-1\} \), we get
  \[
  \frac{Z_{i+1}}{Z_i} = \int_{\mathbb{R}^d} g_i(x) \pi_i(x) dx = \pi_i(g_i),
  \]
  where \( g_i : \mathbb{R}^d \to \mathbb{R}_+ \) is defined for any \( x \in \mathbb{R}^d \) by
  \[
  g_i(x) = \exp \left( a_i \|x\|^2 \right), \quad a_i = \frac{1}{2} \left( \frac{1}{\sigma_i^2} - \frac{1}{\sigma_{i+1}^2} \right).
  \]
Multistage methods

- **Multistage sampling type algorithms are widely used and known under different names**: multistage sampling (Valleau, 1972), (extended) bridge sampling (Gelman, 1998), annealed importance sampling (AIS) (Neal, 2001), thermodynamic integration (Girolami, 2016), power posterior (Friel, 2012).

- For the stability and accuracy of the method, the **choice of the parameters** 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 (Friel, 2012), temperature placement (Friel, 2014), annealing sequence, temperature ladder (Girolami, 2016), effects of grid size...

- In Brosse, Durmus, Mmoulines (2018), we explicitly determine the sequence \( \{\sigma_i^2\}_{i=0}^{M-1} \).
Multistage Langevin

- Compute for all $i \in \{1, \ldots, M-1\}$,
  \[
  \frac{Z_{i+1}}{Z_i} = \int_{\mathbb{R}^d} g_i(x) \pi_i(x) \, dx = \pi_i(g_i) .
  \]

- The quantity $\pi_i(g_i)$ is estimated by the Unadjusted Langevin Algorithm (ULA) targeting $\pi_i$.

- For all $i \in \{1, \ldots, M\}$, consider
  \[
  X_{i,k+1} = X_{i,k} - \gamma_i \nabla U_i(X_{i,k}) + \sqrt{2\gamma_i} Z_{i,k+1} , \quad X_{i,0} = 0 .
  \]

- For $i \in \{0, \ldots, M-1\}$, consider the following estimator of $Z_{i+1}/Z_i$,
  \[
  \hat{\pi}_i(g_i) = \frac{1}{n_i} \sum_{k=N_i+1}^{N_i+n_i} g_i(X_{i,k}) ,
  \]
  where $n_i \geq 1$ is the sample size and $N_i \geq 0$ the burn-in period.
ULA algorithm

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

$$
\frac{Z_{i+1}}{Z_i} = \int_{\mathbb{R}^d} g_i(x) \pi_i(x) \, dx = \pi_i(g_i),
$$

- For $i \in \{0, \ldots, M - 1\}$, consider the following estimator of $Z_{i+1}/Z_i$,

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

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

- The following estimator of $Z$,

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

- Denote by $S$ the set of simulation parameters,
  \[ S = \{ M, \{ \sigma_i^2 \}_{i=0}^{M-1}, \{ \gamma_i \}_{i=0}^{M-1}, \{ n_i \}_{i=0}^{M-1}, \{ N_i \}_{i=0}^{M-1} \}. \]

- $\hat{Z}$ the following estimator of $Z$,
  \[
  \hat{Z} = (2\pi\sigma_0^2)^{d/2}(1 + \sigma_0^2 m)^{-d/2} \left\{ \prod_{i=0}^{M-1} \hat{\pi}_i(g_i) \right\}.
  \]

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

**Theorem 10 ([brosse:durmus:moulines:2018])**

Let $\mu, \epsilon \in (0, 1)$. There exists an explicit choice of the simulation parameters $S$ such that the estimator $\hat{Z}$ satisfies
\[
\mathbb{P} \left( \left| \frac{\hat{Z}}{Z} - 1 \right| > \epsilon \right) \leq \mu.
\]

Moreover, the cost of the algorithm is polynomial in the dimension $d$, $\epsilon^{-1}$ and $\eta^{-1}$. 
Figure: Boxplots of the logarithm of the normalizing constants of a multivariate Gaussian distribution in dimension $d \in \{10, 25, 50\}$. 
Figure: Boxplot of the log evidence for a mixture of 4 Gaussian distributions in dimension 2.
Our published work


3. Durmus, A.; Moulines, E., *High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm*, major revision in Bernoulli, 2018 [first version submitted 3 years ago !]


Figure: A beach reading or a nice Christmas present for a loved one... This new book covers the classical theory of Markov chains on general state-spaces as well as many recent developments. The theoretical results are illustrated by simple examples, many of which are taken from Markov Chain Monte Carlo methods.