Optimal scaling and convergence of Markov chain Monte Carlo methods

Alain Durmus
Joint work with: Sylvain Le Corff, Éric Moulines, Gareth Roberts, Umut Şimşekli

February 16, 2016

Introduction

- Sampling distributions over high-dimensional state-space has recently attracted a lot of research efforts in computational statistics and machine learning community...
- Applications (non-exhaustive)
  - Bayesian inference for high-dimensional models and Bayesian non parametric.
  - Bayesian linear inverse problems (typically function space problems).
  - Aggregation of estimators and experts.

Bayesian setting

- A Bayesian model is specified by
  1. a prior distribution p on the parameter space $\theta \in \mathbb{R}^d$
  2. the sampling distribution of the observed data conditional on its parameters, often termed likelihood: $Y \sim \mathsf{L}(\cdot|\theta)$
- The inference is based on the posterior distribution:

$$\boxed{\pi(\mathrm{d}\theta) = \frac{\mathsf{p}(\mathrm{d}\theta)\mathsf{L}(Y|\theta)}{\int \mathsf{L}(Y|u)\mathsf{p}(\mathrm{d}u)}\,.}$$

- In most cases the normalizing constant is not tractable:

$$\pi(\mathrm{d}\theta) \propto \mathsf{p}(\mathrm{d}\theta)\mathsf{L}(Y|\theta)\,.$$

Logistic and probit regression

- 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 $F(\theta^T X_i)$, where
  1. X_i is a d dimensional vector of known covariates,
  2. θ 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, $F(t) = \mathrm{e}^t/(1+\mathrm{e}^t)$.

Logistic and probit regression (II)

- The posterior density distribution of θ is given, up to a proportionality constant by $\pi(\theta|(Y,X)) \propto \exp(-U(\theta))$, where the potential $U(\theta)$ is given by

$$U(\theta) = -\sum_{i=1}^{p} \{Y_i \log F(\theta^T X_i) + (1-Y_i) \log(1-F(\theta^T X_i))\} + \mathrm{g}(\theta) \ ,$$

  where g is the log density of the posterior distribution.
- Two important cases:
  - Gaussian prior $\mathrm{g}(\theta) = (1/2)\theta^T \Sigma \theta$, ridge regression.
  - Laplace prior $\mathrm{g}(\theta) = \lambda \sum_{i=1}^{d} |\theta_i|$, lasso regression.

Bayesian setting (II)

Bayesian decision theory relies on computing expectations:

$$\pi(f) = \int_{\mathbb{R}^d} f(\theta)\pi(\mathrm{d}\theta)$$

Generic problem: estimation of an integral $\pi(f)$, where

- π is known up to a multiplicative factor ;
- Sampling directly from π is not an option;

A solution is to approximate $\mathbb{E}_\pi[f]$ by $n^{-1}\sum_{i=1}^{n} f(X_i)$,

where $(X_i)_{i\geq 0}$ is a Markov chain associated with a Markov kernel P for which π is invariant.

Markov chain theory

- Invariant probability measure: π is said to be an invariant probability measure for the Markov kernel P if

$$X_0 \sim \pi \text{ then } X_1 \sim \pi$$

- Ergodic Theorem (Meyn and Tweedie, 2003): If π is invariant, With some conditions on P, we have for any $f \in \mathrm{L}^1(\pi)$,

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \underset{\pi\text{-a.s.}}{\longrightarrow} \int f(x)\pi(x)\mathrm{d}x\,.$$

MCMC: rationale

- To approximate $\pi(f)$: find P with invariant measure π, from which we can efficiently sample.
- MCMC methods are algorithms which aims to build such kernel.
- One of the most famous example: The Metropolis-Hastings algorithm.

The Metropolis-Hastings algorithm

Initial Data: the target density π, a transition density q, $X_0 \sim \mu_0$.
For $k \geq 0$ given X_k,

1. Generate $Y_{k+1} \sim q(X_k, \cdot)$.
2. Set

$$X_{k+1} = \begin{cases} Y_{k+1} & \text{with probability } \alpha(X_k, Y_{k+1}) \,, \\ X_k & \text{with probability } 1 - \alpha(X_k, Y_{k+1}) \,. \end{cases}$$

where

$$\boxed{\alpha(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)} \frac{q(y, x)}{q(x, y)}} \,.$$

π is invariant for the corresponding Markov kernel P.

Example: The symmetric Random Walk Metropolis algorithm

The Random Walk Metropolis:

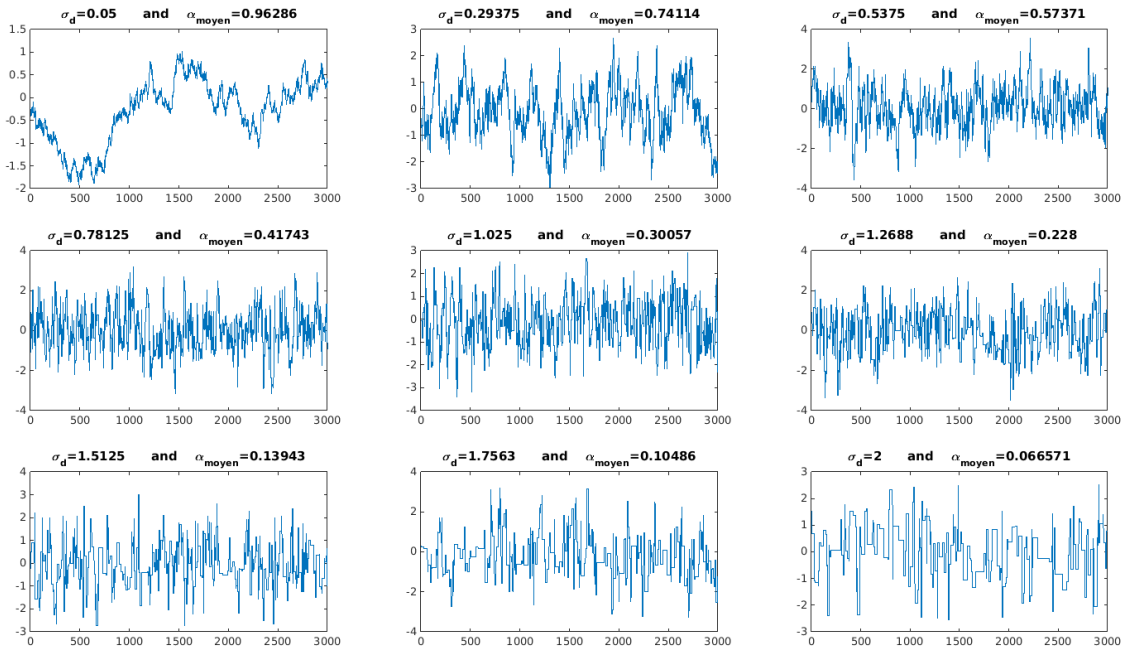
$$\begin{cases} Y_{k+1} & = X_k + \sigma Z_{k+1} \qquad (Z_k)_{k\geq 0} \text{ i.i.d. sequence of law } \mathcal{N}_d(0, \mathrm{Id}_d) \\ q(x,y) & = \sigma^{-d}\phi_d(\|y-x\|/\sigma) \quad \text{where } \phi_d \text{ is the Gaussian density on } \mathbb{R}^d \\ \alpha(x,y) & = \pi(y)/\pi(x)\ . \end{cases}$$

Study of MCMC methods: measures of efficiency

1. How to measure the efficiency of MCMC methods ?
2. Equivalent problem: quantifying the convergence of the Markov kernel P to its stationary distribution π.
3. We consider two criteria:
   - the asymptotic variance $\Rightarrow$ justifies optimal scaling results.
   - convergence in some metric on the set of probability measures.

Behaviour of the RWM

Recall the RWM proposal: $Y_{k+1} = X_k + \sigma Z_{k+1}$



- On the one hand, σ should be as large as possible so that the chain explores the state spaces.
- On the other hand, σ should not be too large as possible otherwise $\alpha \to 0$.

Scaling problems

Questions:

- How should σ depend on the dimension d ?
- We study the following very simple model.
- Consider π a one dimensional positive density on $\mathbb{R}$ of the form

$$\pi \propto \mathrm{e}^{-u} .$$

- Define the positive density on given for all $x \in \mathbb{R}^d$ by

$$\pi^d(x) = \prod_{i=1}^{d} \pi(x_i) = \prod_{i=1}^{d} \mathrm{e}^{u(x_i)} ,$$

where x_i is the i-th component of x.

Study of the acceptance ratio (I)

- Recall $\pi^d(x) = \prod_{i=1}^d \pi(x_i) = \prod_{i=1}^d \mathrm{e}^{u(x_i)}$
- Then the acceptance ratio can be written of the form for all $x, y \in \mathbb{R}^d$,

$$\begin{aligned}\alpha(x,y) &= 1 \wedge \frac{\pi(y)}{\pi(x)} \\ &= 1 \wedge \exp\left(\sum_{i=1}^d u(x_i) - u(y_i)\right) \ .\end{aligned}$$

Study of the acceptance ratio (II)

- Recall $\alpha(x,y) = 1 \wedge \exp\left(\sum_{i=1}^{d} u(x_i) - u(y_i)\right)$
- We want that the acceptance ratio during the algorithm $\in (0,1)$.
- Let $X_0^d \sim \pi^d$ and the proposal based on X_0^d, $Y_1^d = X_0^d + \sigma Z_1^d$.
- We consider the mean acceptance ratio, *i.e.* the quantity:

$$\begin{aligned}\mathbb{E}\left[\alpha(X_0^d, Y_1^d)\right] &= \mathbb{E}\left[\alpha(X_0^d, X_0^d + \sigma Z_1^d)\right] \\ &= \mathbb{E}\left[1 \wedge \exp\left(\sum_{i=1}^{d} u(X_{0,i}^d) - u(X_{0,i}^d + \sigma Z_{1,i}^d)\right)\right] .\end{aligned}$$

Study of the acceptance ratio (III)

- $\mathbb{E}\left[\alpha(X_0^d, Y_1^d)\right] = \mathbb{E}\left[1 \wedge \exp\left(\sum_{i=1}^d u(X_{0,i}^d) - u(X_{0,i}^d + \sigma Z_{1,i}^d)\right)\right]$
- If u is C^3 then a third Taylor expansion gives:

$$u(X_{0,i}^d) - u(X_{0,i}^d + \sigma Z_{1,i}^d) = \sigma Z_{1,i}^d u'(X_{0,i}^d) + (\sigma Z_{1,i}^d)^2 u''(X_{0,i}^d)/2 + o(\sigma^3)\ . \quad (1)$$

- Set now $\sigma = \ell d^{-\xi}$.
- By (3) if $\xi < 1/2$, then

$$\liminf_{d\to+\infty} \sum_{i=1}^d u(X_{0,i}^d) - u(X_{0,i}^d + \ell d^{-\xi} Z_{1,i}^d) = -\infty$$

and therefore

$$\liminf_{d\to+\infty} \mathbb{E}\left[\alpha(X_0^d, Y_1^d)\right] \to 0\ .$$

Study of the acceptance ratio (IV)

- $\mathbb{E}\left[\alpha(X_0^d, Y_1^d)\right] = \mathbb{E}\left[1 \wedge \exp\left(\sum_{i=1}^d u(X_{0,i}^d) - u(X_{0,i}^d + \sigma Z_{1,i}^d)\right)\right]$
- If u is C^3 then a third Taylor expansion gives:

$$u(X_{0,i}^d) - u(X_{0,i}^d + \sigma Z_{1,i}^d) = \sigma Z_{1,i}^d u'(X_{0,i}^d) + (\sigma Z_{1,i}^d)^2 u''(X_{0,i}^d)/2 + o(\sigma^3)\ . \quad (2)$$

- Set now $\sigma = \ell d^{-\xi}$.
- By (3) if $\xi > 1/2$, then

$$\liminf_{d\to+\infty} \sum_{i=1}^d u(X_{0,i}^d) - u(X_{0,i}^d + \ell d^{-\xi} Z_{1,i}^d) = 0$$

and therefore

$$\lim_{d\to+\infty} \mathbb{E}\left[\alpha(X_0^d, Y_1^d)\right] \to 1\ .$$

Study of the acceptance ratio (V)

- $\mathbb{E}\left[\alpha(X_0^d, Y_1^d)\right] = \mathbb{E}\left[1 \wedge \exp\left(\sum_{i=1}^d u(X_{0,i}^d) - u(X_{0,i}^d + \sigma Z_{1,i}^d)\right)\right]$
- If u is C^3 then a third Taylor expansion gives:

$$u(X_{0,i}^d) - u(X_{0,i}^d + \sigma Z_{1,i}^d) = \sigma Z_{1,i}^d u'(X_{0,i}^d) + (\sigma Z_{1,i}^d)^2 u''(X_{0,i}^d)/2 + o(\sigma^3)\ . \quad (3)$$

- Set now $\sigma = \ell d^{-1/2}$.
- Then $\sum_{i=1}^d u(X_{0,i}^d) - u(X_{0,i}^d + \ell d^{-1/2} Z_{1,i}^d) \Rightarrow G$, as $d \to +\infty$ where

$$G \sim \mathcal{N}\left((\ell^2 I/2), \ell^2 I\right)\ ,\ I = \mathbb{E}\left[\left(u'(X_{0,i}^d)\right)^2\right]$$

and therefore [Roberts, Gelman, Gilks, 1996]

$$\lim_{d \to +\infty} \mathbb{E}\left[\alpha(X_0^d, Y_1^d)\right] \to \mathbb{E}\left[1 \wedge \mathrm{e}^G\right] = 2\Phi\left(-\ell\sqrt{I}/2\right)\ .$$

From the MH algorithm to the LAN property

- Problem: what happens if π is non continuously differentiable ?
- Recall we want to control:

$$\begin{aligned}\alpha(X_0^d, X_0^d + \sigma Z_1^d) &= 1 \wedge \pi(X_0^d + \sigma Z_1^d)/\pi(X_0^d) \\ &= 1 \wedge \prod_{i=1}^{d} \pi(X_{0,i}^d + \sigma Z_{1,i}^d)/\pi(X_{0,i}^d)\end{aligned}$$

- We recognize the likelihood ratio for a translation model.
- The issue of non-differentiability of π has been also raised for the LAN (locally asymptotically normal) property of the likelihood ratio.

The LAN property (simplified)

- Consider the translation model $\theta \mapsto \pi(\cdot + \theta)$ on $\mathbb{R}$ where $\pi \propto \mathrm{e}^u$ is still a positive one dimensional density.
- Define the likelihood ratio (at 0)

$$r((x_i)_{1\leq i\leq N}, \theta) = \prod_{i=1}^{N} \pi(x_i + \theta)/\pi(x_i) \ .$$

- The model is said to satisfy the LAN property if for all $\ell \in \mathbb{R}$, $\theta = \ell/\sqrt{n}$,

$$\log r((x_i)_{1\leq i\leq N}, \theta) = S_N$$

where as $N \to +\infty$,

$$S_N \Rightarrow \mathcal{N}(\varsigma^2\ell^2/2, (\varsigma\ell)^2) \ , \ \varsigma^2 = \int_{\mathbb{R}} (u'(x))^2 \pi(x) \mathrm{d}x \ .$$

DQM condition

- If π is C^3 then the LAN property is straightforward using a third Taylor expansion.
- Otherwise, Le Cam suggests to consider the following condition: there exists $\phi \in L^2$ such that

$$\int_{\mathbb{R}} \left\{ \pi^{1/2}(x+\theta) - \pi^{1/2}(x) - \phi(x)\theta \right\}^2 \mathrm{d}\theta =_{\theta \to 0} o(\theta^2) \ .$$

- The model is said to be differentiable in quadratic mean.
- If π is C^2 and positive,

$$\phi(\theta)/\pi^{1/2}(\theta) = (\log \pi)'(\theta) \ .$$

Assumptions

We assume that there exists a measurable function $\dot{u} : \mathbb{R} \to \mathbb{R}$ such that:

1. Differenciability in L^p mean: There exist $p > 4$, $C > 0$ and $\beta > 1$ such that for all $x \in \mathbb{R}$,

$$\int_{\mathbb{R}} \{u(y+x) - u(y) - x\dot{u}(y)\}^p \pi(y)\mathrm{d}y \leq C|x|^{p\beta} .$$

2. This condition implies that $\theta \mapsto \pi(\cdot + \theta)$ is DQM [D., Le Corff, Moulines, Roberts, 2016].
3. Moment condition The function $\dot{u}$ satisfies

$$\int_{\mathbb{R}} |\dot{u}(x)|^6 \pi(y)\mathrm{d}y < +\infty .$$

Limiting acceptance ratio for non-smooth densities

- Assume these two conditions.
- Then we recover $\sum_{i=1}^{d} u(X_{0,i}^d) - u(X_{0,i}^d + \ell d^{-1/2} Z_{1,i}^d) \Rightarrow G$ where

$$G \sim \mathcal{N}\left((\ell^2 I/2), \ell^2 I\right) \ , \ I = \mathbb{E}\left[\left(\dot{u}(X_{0,i}^d)\right)^2\right]$$

and therefore [D., Le Corff, Moulines, Roberts, 2016]

$$\lim_{d \to +\infty} \mathbb{E}\left[\alpha(X_0^d, Y_1^d)\right] \to \mathbb{E}\left[1 \wedge \mathrm{e}^G\right] = 2\Phi\left(-\ell\sqrt{I}/2\right) \ .$$

Scaling problems

Questions:

- How should σ depend on the dimension d ? done : $\sigma = \ell d^{-1/2}$
- What does this tell us about the efficiency of the algorithm ?
- Can we optimize ℓ in a sensible way ?
- Can we characterize the optimal choice of ℓ by some intrinsic criteria independent of π ?

For the case of Metropolis-Hastings type algorithms , there are diffusion limits which answers to these questions.

Efficiency of MCMC algorithms: asymptotic variance

Let $(X_k)_{k\geq 0}$ be a Markov chain with invariant measure π.
With some conditions we have a LLN and a CLT: for some f,

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow[n\to+\infty]{\text{a.s.}} \int f(x)\pi(x)\mathrm{d}x$$

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \int f(x)\pi(x)\mathrm{d}x\right) \xRightarrow[n\to+\infty]{*} \mathcal{N}(0, \sigma^2(f,P)) \ ,$$

where

$$\begin{aligned}\sigma^2(f,P) &= \lim_{n\to+\infty} n \operatorname{Var}_\pi\left\{\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right\} \\ &= \operatorname{Var}_\pi\left\{f(X_0)\right\} + \sum_{i\geq 1} \operatorname{Cov}_\pi\left\{f(X_i), f(X_0)\right\} \ .\end{aligned}$$

Expected Square Jump Distance

Common efficiency criteria: the ESJD defined for Markov chain in one dimension by:

$$\mathrm{ESJD} = \mathbb{E}_\pi[(X_1 - X_0)^2] \,.$$

Property: If f is a linear function

$$\text{Maximizing the } \mathrm{ESJD} \Leftrightarrow \text{ Minimizing } \mathrm{Cov}_\pi \{f(X_1), f(X_0)\} \,,$$

Efficiency of MH algorithms

- Given f, the CLT allows us to compare two Markov kernel P_1, P_2:

$$\sigma^2(f, P_1) \leq \sigma^2(f, P_2) \Longrightarrow \ P_1 \text{ is more efficient than } P_2 \ .$$

- However it can be hard to ensure for all f,

$$\sigma^2(f, P_1) \leq \sigma^2(f, P_2) \ .$$

Langevin diffusion

Let π a probability measure on $\mathbb{R}^d$ with log-density $U \in C^1(\mathbb{R}^d)$
Consider the overdamped Langevin equation:

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

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

Under some conditions on U, $(Y_t)_{t\geq 0}$ is ergodic with respect to π, and we have a LLN and a CLT again:

$$\frac{1}{t}\int_0^t f(X_s)\mathrm{d}s \underset{t\to+\infty}{\xrightarrow{a.s.}} \int f(x)\pi(x)\mathrm{d}x$$

$$\sqrt{t}\left(\frac{1}{t}\int_0^t f(X_s)\mathrm{d}s - \int f(x)\pi(x)\mathrm{d}x\right) \underset{t\to+\infty}{\overset{*}{\Longrightarrow}} \mathcal{N}(0, \sigma^2(f,Y)) \ ,$$

where

$$\sigma^2(f,Y) = \lim_{t\to+\infty} t\,\mathrm{Var}_\pi\left\{\frac{1}{t}\int_0^t f(Y_s)\mathrm{d}s\right\} \ .$$

scaled Langevin equation

Consider the following scaled Langevin equation:

$$\mathrm{d}Y_t^c = -c\nabla U(Y_t^c)\mathrm{d}t + \sqrt{2c}\mathrm{d}B_t \ , \text{ for } c > 0 \ . \tag{4}$$

The solution of the scaled Langevin equation are speed-up-version of $(Y_{ct}^1)_{t\geq 0}$ [the SLE with unit-speed]:

$$\begin{aligned} Y_{ct}^1 &= Y_0^1 + \int_0^{ct} \nabla U(Y_s^1)\mathrm{d}s + \sqrt{2}B_{ct} \\ &\overset{s=cu}{=} Y_0^1 + \int_0^t c\nabla U(Y_s^1)\mathrm{d}s + \sqrt{2c}\tilde{B}_t \ , \end{aligned}$$

with the Brownian motion $\tilde{B}_t = c^{-1/2}B_{ct}$.

Efficiency of Langevin solutions

Which c leads to the best convergence, *i.e.* minimizes $\sigma^2(f, (Y_t^c)_{t\geq 0})$?

1 To reach the equilibrium, it is sensible to speed-up the diffusion:

so take large c.

2 Speeding-up the diffusion is also justified by the variance in the CLT :

$$\sigma^2(f, (Y_t^c)_{t\geq 0}) = \lim_{t\to+\infty} t \operatorname{Var}_\pi \left\{ \frac{1}{t} \int_0^t f(Y_{cs}^1) \mathrm{d}s \right\}$$
$$\stackrel{u=cs}{=} c^{-1} \lim_{t\to+\infty} ct \operatorname{Var}_\pi \left\{ \frac{1}{ct} \int_0^{ct} f(Y_s^1) \mathrm{d}s \right\} .$$

$$\sigma^2(f, (Y_t^c)_{t\geq 0}) = c^{-1} \sigma^2(f, (Y_t^1)_{t\geq 0}) ,$$

Conclusion: the faster, the better: this result holds for all f (under appropriate smoothness and moment conditions).

Action plan

- Under some (strong) conditions, the MH iterates converges to a diffusion process.
- Then tune the variance σ to optimize the speed of the limiting diffusion.

Scaling of the RWM (Roberts, Gelman and Gilks, 1997)

Assumption [Controversial !]

- $\pi^d(x) = \prod_{i=1}^d \pi(x_i) = \prod_{i=1}^d \mathrm{e}^{u(x_i)}$
- $\{X_k^{d,}, d \geq 0\}$ be the Markov chain produced by the RWM on $\mathbb{R}^d$ with target density π^d and

$$X_0^{d,} \sim \pi^d \qquad \sigma^d = \ell d^{-1/2} \, , \ell > 0 \, .$$

Results:

$$\{(X^{d,}_{\lfloor td \rfloor,1})_{t\geq 0}, d \geq 1\} \underset{d\to+\infty}{\overset{*}{\Longrightarrow}} (Y_t)_{t\geq 0} \, ,$$

where $(Y_t)_{t\geq 0}$ is a solution of the SLE:

$$\mathrm{d}Y_t = h(\ell) u'(X_t)\mathrm{d}t + (2h(\ell))^{1/2}\mathrm{d}B_t \, ,$$

for a function $h(\ell)$ known in closed form and that can be optimized.

Consequences on the tuning of the two algorithms

- If the semigroup of the Langevin equation explores the invariant distribution in $O(1)$ at stationarity: the RWM explores it in $O(d)$.
- To get the best mixing algorithm, tune the parameter ℓ in order to maximize the different speed measures $h(\ell)$ (the RWM then approximates the fastest Langevin solution).
- The best parameter ℓ is characterized by a mean acceptance rate of order ≈ 0.234.
- Conclusion: tune during your algorithm ℓ to have an acceptance ratio with empirical mean ≈ 0.23

Extension of the result to non-smooth densities

- Recall that it is assumed that $\pi^d(x) = \prod_{i=1}^d \pi(x_i) = \prod_{i=1}^d \mathrm{e}^{u(x_i)}$.
- The original result of Roberts et al. in addition assumes that $u \in C^3(\mathbb{R})$ (very smooth).
- Our contribution: Extension to non-smooth u with S. Le Corff, É. Moulines and G. Roberts.
  - π is possibly non differentiable in some points or supported on an open interval of $\mathbb{R}$.
  - The proof follows from the ideas in (Jourdain, Lelievre, Miasojedow, 2015).

Assumptions (II)

We assume that there exists a measurable function $\dot{u} : \mathbb{R} \to \mathbb{R}$ such that:

1. Differenciability in L^p mean: There exist $p > 4$, $C > 0$ and $\beta > 1$ such that for all $x \in \mathbb{R}$,

$$\int_{\mathbb{R}} \{u(y+x) - u(y) - x\dot{u}(y)\}^p \pi(y)\mathrm{d}y \leq C|x|^{p\beta} .$$

2. Moment condition The function $\dot{u}$ satisfies

$$\int_{\mathbb{R}} |\dot{u}(x)|^6 \pi(y)\mathrm{d}y < +\infty .$$

3. Smoothness condition $\dot{u}$ is almost everywhere continuous.

Optimal scaling results

1. For all $d \geq 1$, consider $\{X_k^{d,\mathrm{R}}, d \geq 0\}$ be the Markov chain produced by the RWM on $\mathbb{R}^d$ with target density π^d and

$$X_0^{d,\mathrm{R}} \sim \pi^d \qquad \sigma = \ell d^{-1/2} \, , \ell > 0 \, .$$

Then if the previous assumptions hold

$$\{(X^{d,\mathrm{R}}_{\lfloor td \rfloor,1})_{t\geq 0}, d \geq 1\} \underset{d\to+\infty}{\overset{*}{\Longrightarrow}} (Y_t)_{t\geq 0} \, ,$$

weak solution of the possibly singular scaled Langevin equation:

$$\mathrm{d}Y_t = h^{\mathrm{R}}(\ell)\dot{u}(X_t)\mathrm{d}t + (2h^{\mathrm{R}}(\ell))^{1/2}\mathrm{d}B_t \, ,$$

for some function $h^{\mathrm{R}}(\ell)$ which is explicit and can be optimized.

2. Extension to density supported in an open interval $I \subset \mathbb{R}$.

Simulation for beta distributions

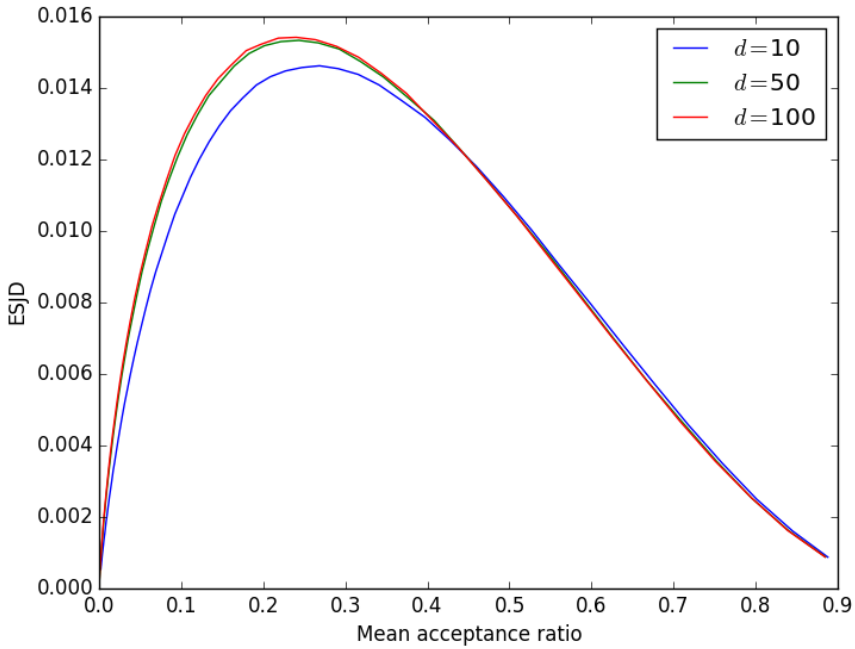


Figure: Expected square jumped distance for the beta distribution with parameters $(10, 10)$ as a function of the mean acceptance rate for $d = 10, 50, 100$.

Simulation for the lasso logistic regression

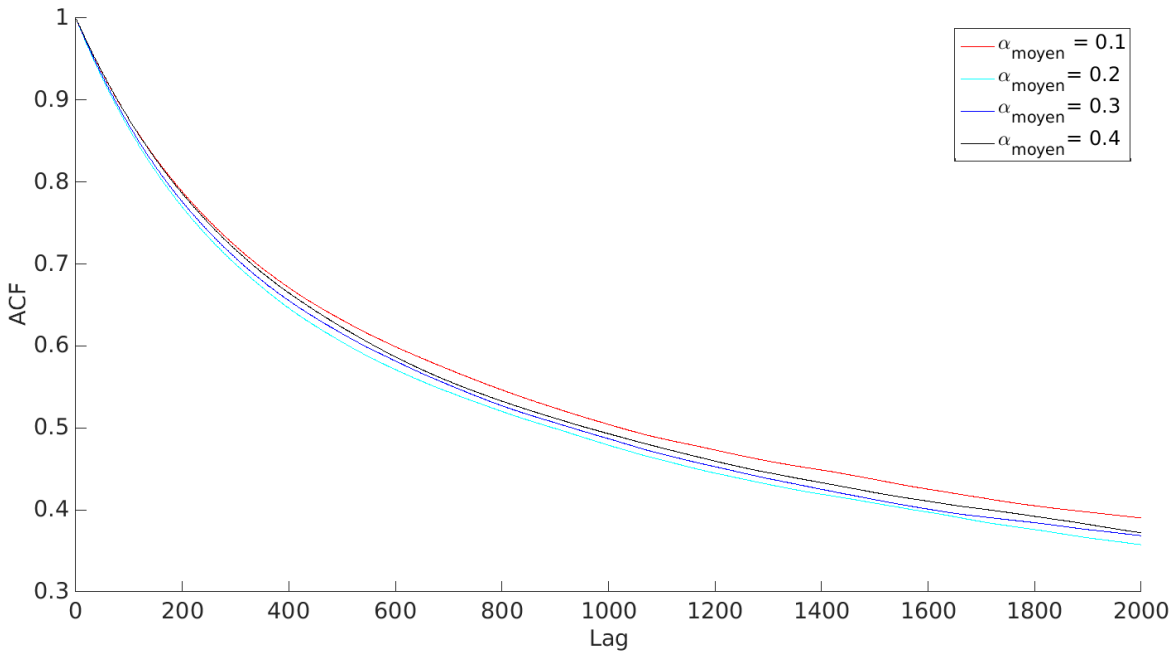


Figure: Autocovariance function for different mean acceptance rates for the lasso logistic regression.
Dataset: Musk of dimension 167

Work in progress:

- Optimal scaling result for MALA applied to convex non-smooth densities.
- Use of proximal operator to improve the dependency on the dimension.

The Unadjusted Langevin Algorithm (ULA)

- Langevin SDE:

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

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

- Idea: Sample the diffusion paths, using for example the Euler-Maruyama (EM) scheme:
  1. initial state $X_0 \sim \mu_0$
  2. for $k \geq 0$, given X_k,

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

where
- $(Z_k)_{k\geq 1}$ is i.i.d. $\mathcal{N}(0, \mathrm{I}_d)$
- $\gamma > 0$ is a step size

Discretized Langevin diffusion: constant stepzize

- $(X_k)_{k\geq 1}$ is an homogeneous Markov chain with Markov kernel R_γ
- Under some appropriate conditions, this Markov chain is irreducible, positive recurrent $\rightsquigarrow$ unique invariant distribution π_γ.
- Problem: $\pi_\gamma \neq \pi$.

Convergence of Markov chains

- Another measure of efficiency of MCMC to target π associated to a Markov kernel P:

$$\|P^k(x,\cdot) - \pi\|_{\mathrm{TV}} \leq C(x)v(k) \ , \text{ where}$$

1. The total variation distance defined for μ, ν two probabilities measure on $\mathbb{R}^d$ by

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{|f|\leq 1} |\mu(f) - \nu(f)| \ .$$

2. $C(x) \geq 0$: dependence on the initial condition.
3. Ideally $\lim_{k\to+\infty} v(k) = 0$ (or close to 0) with the beter possible rate.

Weak error result for the ULA algorithm

- Recall the ULA algorithm

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

- We have seen that the ULA algorithm is biased.
- The Markov chain $(X_k)_{k \geq 0}$ has an invariant distribution $\pi_\gamma \neq \pi$.
- However, (Talay and Tubaro 1991) shows that for $U, f \in C^\infty(\mathbb{R}^d)$ and additional assumptions, there exists a constant C depending on f and π such that

$$\int_{\mathbb{R}^d} f(x) \mathrm{d}\pi(x) - \int_{\mathbb{R}^d} f(x) \mathrm{d}\pi_\gamma(x) = C\gamma + \mathcal{O}(\gamma^2) \ .$$

Discussion

- The previous result is not quantitative. No explicit bounds.
- We aimed with É. Moulines at giving computable bounds in total variation or Wassestein distance.
- In particular to see the dependence on the dimension.
- We make the assumption that U is continuously differentiable, convex and gradient Lipschitz.
- Complete and improve the result of (Dalalyan 2014).

Notation and framework

- Recall that

$$X_{k+1} = X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} Z_{k+1} \ , \gamma > 0 \ ,$$

and R_γ denotes the Markov kernel associated with $(X_k)_{k \geq 0}$.

- We answer to the following questions:
  1. For a target precision $\varepsilon > 0$, can we find explicit $\gamma > 0$ and $N \geq 0$ such that
  $$\|\delta_x R_\gamma^n - \pi\|_{\mathrm{TV}} \leq \varepsilon \text{ for all } n \geq N \ .$$
  2. For all $n \geq 0$, can we find explicit $\gamma > 0$ such that
  $$\|\delta_x R_\gamma^n - \pi\|_{\mathrm{TV}} \leq v(n) \text{ with } \lim_{n \to \infty} v(n) = 0 \ .$$

Main result (I)

- Assume that U is gradient Lipschitz and convex.
- Assume in addition that there exists $\eta > 0$ and $R \geq 0$ independent of d such that for all $x, y \in \mathbb{R}^d$,

$$\langle \nabla U(x) - \nabla U(y), x - y \rangle \geq \eta \|x - y\| \ .$$

- Then, for all $\varepsilon > 0$, $\|\delta_x R_\gamma^n - \pi\|_{\mathrm{TV}} \leq \varepsilon$ for γ well chosen and

$$n \geq C d^5 \text{ for } C \geq 0 \text{ which is explicit and independent of } d \ .$$

- For all $n \geq 0$, there exist $C \geq 0$ explicit and independent of the dimension and $\gamma > 0$

$$\|\delta_x R_\gamma^n - \pi\|_{\mathrm{TV}} \leq \frac{C \log(n) d^{5/2}}{n^{1/2}} \ .$$

Discussion

- This kind of results has the subject of numerous paper concerning the RWM applied to logconcave target: (A. Frieze, R. Kannan, and N. Polson, 1994), (A. Frieze and R. Kannan, 1999)...
- The best results have been obtain in (L. Lovàsz and S. Vempala, 2007).
- They show that a sufficient number of iteration n for the RWM to achieve a target precision ε is of order:

$$n \geq Cd^4 \ .$$

- Besides their result does not assume that U is continuously differentiable.

Discussion (II)

- But they assume that the target is well rounded: there exists C independent of the dimension such that

$$\int_{\mathbb{R}^d} \left\| x - \int_{\mathbb{R}^d} y \mathrm{d}\pi(y) \right\|^2 \mathrm{d}\pi(x) \leq Cd \, .$$

- Our result does not require such assumption.
- In fact with this kind of assumption, we can show that for all $\varepsilon > 0$, $\|\delta_x R_\gamma^n - \pi\|_{\mathrm{TV}} \leq \varepsilon$ for γ well chosen and

$$n \geq Cd^3 \text{ for } C \geq 0 \text{ which is explicit and independent of } d \, .$$

Main result (III)

- Assume that $U \in C^3(\mathbb{R}^d)$, strongly convex, gradient Lipschitz and there exists $\tilde{L}$ such that for all $x, y \in \mathbb{R}^d$:

$$\left\|\nabla^2 U(x) - \nabla^2 U(y)\right\| \leq \tilde{L} \left\|x - y\right\| .$$

- Then, for all $\varepsilon > 0$, $\|\delta_x R_\gamma^n - \pi\|_{\mathrm{TV}} \leq \varepsilon$ for γ well chosen and

  $n \geq C d^{1/2} \log^2(d)$ for $C \geq 0$ which is explicit and independent of d .

- Almost sharp bounds for the Gaussian case !
- For all $n \geq 0$, there exist $C \geq 0$ explicit and independent of the dimension and $\gamma > 0$

$$\|\delta_x R_\gamma^n - \pi\|_{\mathrm{TV}} \leq \frac{C \log(n) d^{1/2} \log^2(d)}{n} .$$

Metropolis-Adjusted Langevin Algorithm

- To correct the target distribution, a Metropolis-Hastings step can be included $\rightsquigarrow$ Metropolis Adjusted Langevin Agorithm (MALA).
  - (Rossky, Doll, and Friedman 1978)
- Algorithm:
  1. Propose $Y_{k+1} \sim X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} Z_{k+1}$, $Z_{k+1} \sim \mathcal{N}(0, \mathrm{I}_d)$
  2. Compute the acceptance ratio $\alpha_\gamma(X_k, Y_{k+1})$

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

  3. Accept / Reject the proposal.
- Work which studied this algorithm: (Roberts and Tweedie 1996), (Bou-Rabee and Hairer 2010), (Eberle 2014)...
- Very difficult to analyze because of the behaviour of the acceptance ratio, which leads to very conservative bounds.

Comparison of MALA and ULA (I)

- We compare MALA and ULA for the logistic regression with Gaussian prior on five real data sets.

Data set	Observations p	Covariates d
German credit	1000	25
Heart disease	270	14
Australian credit	690	35
Prima indian diabetes	768	9
Musk	476	167

Table: Dimension of the data sets

Comparison of MALA and ULA (II)

- Define the marginal accuracy between two probability measure μ, ν on $(\mathbb{R}, \mathcal{B}(\mathbb{R}^d))$ by

$$\mathrm{MA}(\mu, \nu) = 1 - (1/2)\|\mu - \nu\|_{\mathrm{TV}} \ .$$

- We compare MALA and ULA for each data sets by estimating for each component $i \in \{1, \ldots, d\}$ the marginal accuracy between their d marginal empirical distributions and the d marginal posterior distributions.

Comparison of MALA and ULA (III)

- To estimate the d marginal posterior distributions, we run $2 \cdot 10^7$ iterations of the Polya-Gamma Gibbs sampler.
- Then 100 runs of MALA and ULA (10^6 iterations per run) have been performed.
- For MALA, the step-size is chosen so that the acceptance probability is ≈ 0.5.
- For ULA, we choose the same constant step-size than MALA.

Comparison of MALA and ULA (IV)

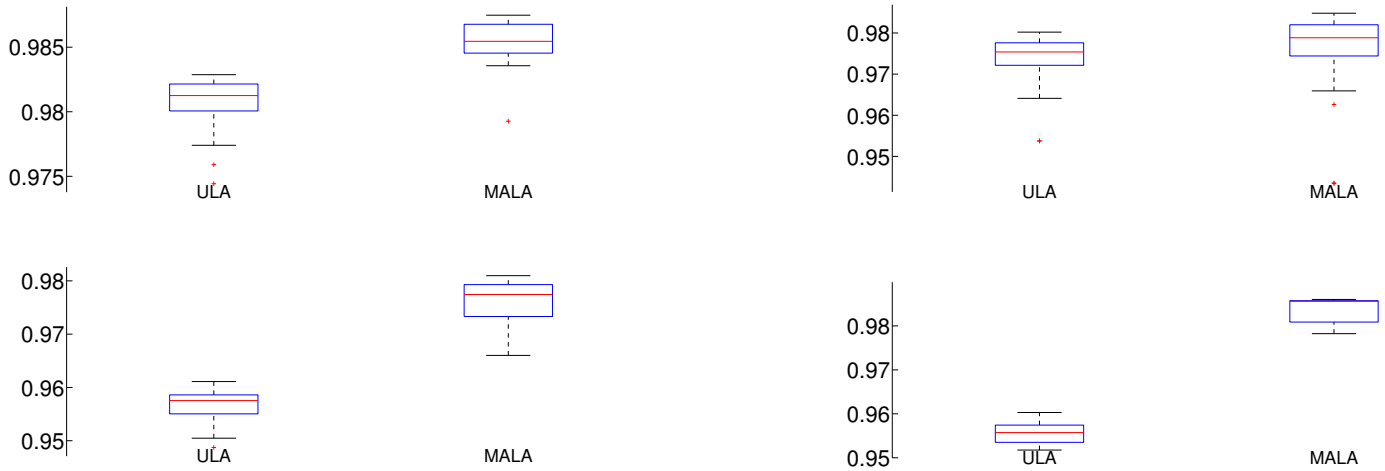


Figure: Marginal accuracy accross all the dimensions.
Upper left: German credit data set. Upper right: Australian credit data set.
Lower left: Heart disease data set. Lower right: Pima Indian diabetes data set

Comparison of MALA and ULA (V)

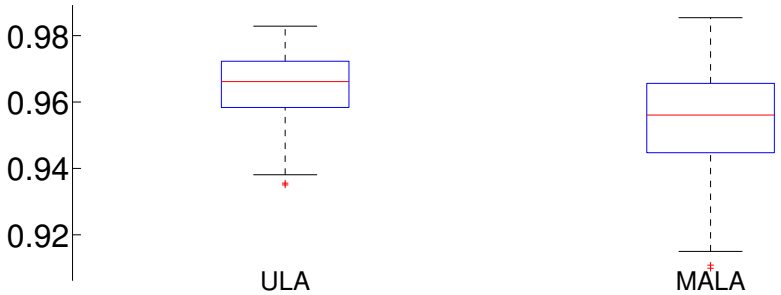


Figure: Marginal accuracy accross all the dimensions for the Musk data Set

Comparison of MALA and ULA (VI)

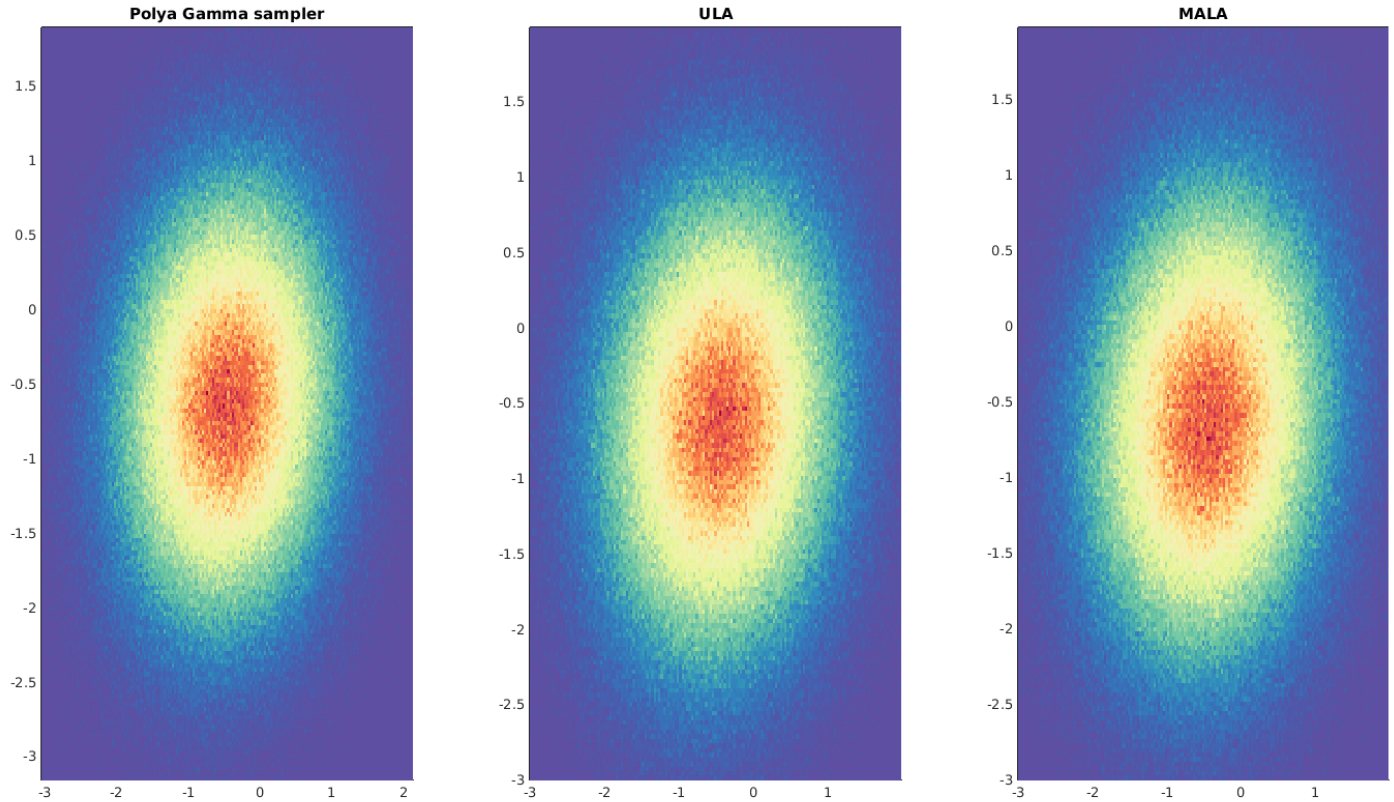


Figure: 2-dimensional histogram for the Musk data Set.

Matrix factorization

- We compare MALA and ULA on a matrix factorization problem for a link prediction application.
- Consider X an observed matrix with missing entries of size $I \times J$. The model is for observed indexes i, j

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

for $K \geq 0$, $(Z_{i,j})$ iid normal random variables $\mathcal{N}(0, \sigma_z)$.

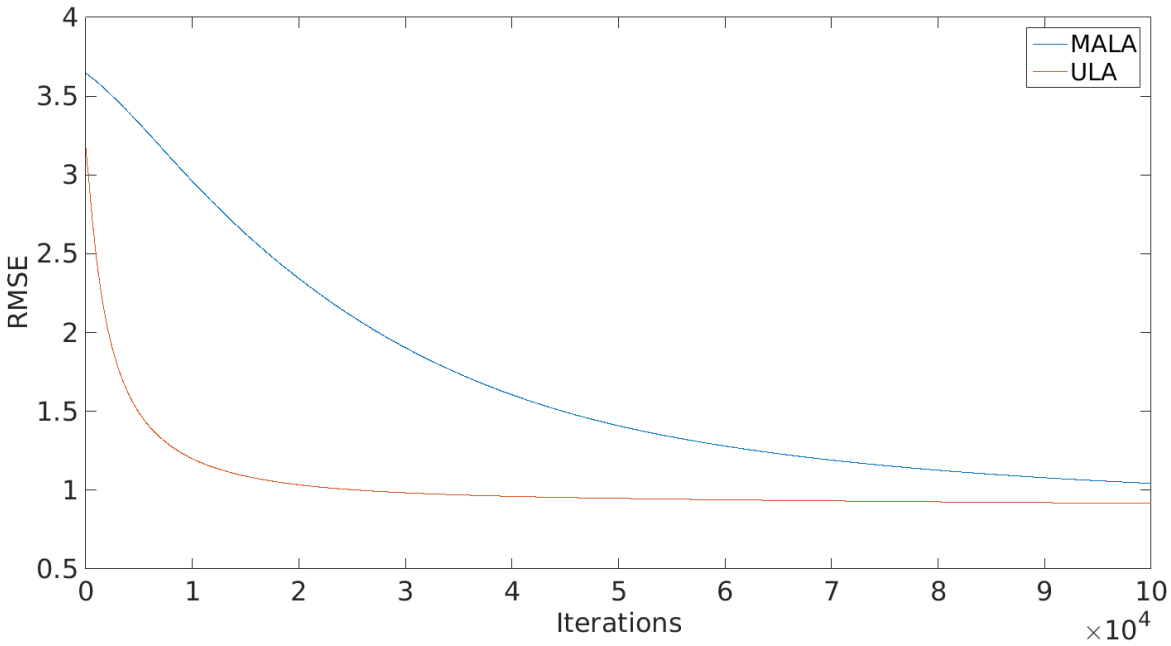
Matrix Factorization (II)

- 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:

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

- Comparison of MALA and ULA on the MovieLens 1 Million dataset.

Matrix Factorization (III)



- Parameters:
  $\sigma_z = 1,$
  $\sigma_w = \sigma_h = 100$

Thank you for your attention. Any questions ?