Sampling from log-concave density

Alain Durmus, Eric Moulines, Marcelo Pereyra

Telecom ParisTech, Ecole Polytechnique, Bristol University

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 and Bayesian non parametrics
  2. Bayesian linear inverse problems (typically function space problems converted to high-dimensional problem by Galerkin method)
  3. Aggregation of estimators and experts
- Most of the sampling techniques known so far do not scale to high-dimension... Challenges are numerous in this area...

Bayesian setting (I)

- In a Bayesian setting, a parameter $\boldsymbol{\beta} \in \mathbb{R}^d$ is embedded with a prior distribution ξ and the observations are given by a probabilistic model:

$$Y \sim \ell(\cdot|\boldsymbol{\beta})$$

The inference is then based on the posterior distribution:

$$\boxed{\pi(\mathrm{d}\boldsymbol{\beta}|Y) = \frac{\xi(\mathrm{d}\boldsymbol{\beta})\ell(Y|\boldsymbol{\beta})}{\int \ell(Y|u)\xi(\mathrm{d}u)}\,.}$$

In most cases the normalizing constant is not tractable:

$$\pi(\mathrm{d}\boldsymbol{\beta}|Y) \propto \xi(\mathrm{d}\boldsymbol{\beta})\ell(Y|\boldsymbol{\beta})\,.$$

Bayesian setting (II)

Bayesian decision theory relies on computing expectations:

$$\int_{\mathbb{R}^d} f(\boldsymbol{\beta})\ell(Y|\boldsymbol{\beta})\xi(\mathrm{d}\boldsymbol{\beta})$$

Generic problem: estimation of an expectation $\mathbb{E}_\pi[f]$, where

- π is known up to a multiplicative factor ;
- we do not know how to sample from π (no basic Monte Carlo estimator);

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

Examples: Logistic and probit regression

- The posterior density distribution of $\boldsymbol{\beta}$ is given by Bayes' rule, up to a proportionality constant by $\pi(\boldsymbol{\beta}|(Y,X)) \propto \exp(-U(\boldsymbol{\beta}))$, where the potential $U(\boldsymbol{\beta})$ is given by

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

  where g is the log-density of the prior distribution.
- Two important cases:
  - Gaussian prior: $\mathrm{g}(\boldsymbol{\beta}) = -(1/2)\boldsymbol{\beta}^T \Sigma_{\boldsymbol{\beta}} \boldsymbol{\beta}$, ridge regression.
  - Laplace prior: $\mathrm{g}(\boldsymbol{\beta}) = -\lambda \sum_{k=1}^{d} |\boldsymbol{\beta}_k|$, lasso regression.

New challenges

Problem the number of predictor variables d is large (10^4 and up).
Examples

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

Data Augmentation

- The most popular algorithms for Bayesian inference in ridge binary regression models are based on data augmentation:
  1. probit link: Albert and Chib (1993).
  2. logistic link: Polya-Gamma sampler, Polsson and Scott (2012)... !
- Bayesian lexicon:
  - Data Augmentation instead on sampling $\pi(\boldsymbol{\beta}|(Y,X))$ sample $\pi(\boldsymbol{\beta},W|(Y,X))$ and marginalize W.
  - Typical application of the Gibbs sampler: sample in turn $\pi(\boldsymbol{\beta}|W,Y,X)$ and $\pi(W|\boldsymbol{\beta},X,Y)$
  - The choice of the DA should make these two steps reasonably easy...

Data Augmentation algorithms

- These two algorithms have been shown to be uniformly geometrically ergodic, BUT the constants depends highly on the dimension.
- 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 $\beta \mapsto U(\beta)$ is smooth, strongly convex
- In the case of the lasso regression, the potential $\beta \mapsto U(\beta)$ is non-smooth but still convex...
- A wealth of reasonably fast optimisation algorithms are available to solve this problem in high-dimension...

Framework

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

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

  Implicitly, $d \gg 1$.

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

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

Langevin diffusion

- Langevin SDE:

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

  where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian Motion.
- Denote by $(P_t)_{t\geq 0}$ the semigroup of the diffusion, $P_t(x, A) = \mathbb{E}_x\left[Y_t \in A\right]$.
- $(P_t)_{t\geq 0}$ is
  - aperiodic, strong Feller (all compact sets are small).
  - reversible w.r.t. to π (admits π as its unique invariant distribution).
- $\pi \propto \mathrm{e}^{-U}$ is reversible $\rightsquigarrow$ the unique invariant probability measure. For all $x \in \mathbb{R}^d$, measurable and bounded functions $f : \mathbb{R}^d \to \mathbb{R}$

$$\lim_{t\to+\infty} P_t f(x) = \lim_{t\to+\infty} \mathbb{E}_x\left[f(Y_t)\right] = \int_{\mathbb{R}^d} f(y)\mathrm{d}\pi(y) \ .$$

Discretized Langevin diffusion

- Idea: Sample the diffusion paths, using for example 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. $\mathcal{N}(0, \mathrm{I}_d)$
  - $(\gamma_k)_{k \geq 1}$ 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 algorithm.

Discretized Langevin diffusion: constant stepzize

- When $\gamma_k = \gamma$, then $(X_k)_{k\geq 1}$ 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: $\pi_\gamma \neq \pi$.

- 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 $(R_{\gamma_k})_{k\geq 1}$
- Denote by $\delta_x Q_\gamma^p$ the law of X_p started at x.

- Reminder: the diffusion converges to the target distribution π.
- Question: since the EM disretization approximates the diffusion, can it be used to sample from π:
  - Is $\delta_x Q_\gamma^p$ closed to π and in which sense ?
  - Can we have some theoretical guarantees ? In particular what is the dependence on the dimension d ?

Metric on probability spaces

Definition

For μ, ν two probabilities measure on $\mathbb{R}^d$, define

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{|f| \leq 1} |\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]| \ .$$

$$W_1(\mu, \nu) = \sup_{\|f\|_{\mathrm{Lip}} \leq 1} |\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]| \ .$$

Wasserstein distance convergence

- We assume in this part that U is strongly convex: there exist and $m > 0$, such that for all $x, y \in \mathbb{R}^d$,

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

- For all $x \in \mathbb{R}^d$,

$$W_1^2(\delta_x P_t, \pi) \leq \mathrm{e}^{-mt} \int_{\mathbb{R}^d} \left\| y - x \right\|^2 \pi(\mathrm{d}y) .$$

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

$$W_1^2(\delta_x Q_\gamma^n, \pi) \leq u_n^{(1)}(\gamma) \int_{\mathbb{R}^d} \|y-x\|^2 \, \pi(\mathrm{d}y) + u_n^{(2)}(\gamma) \, ,$$

where $(u_n^{(1)}(\gamma), u_n^{(2)}(\gamma))_{n\geq 1}$ are explicit.

- We have if $\lim_{k\to+\infty}\gamma_k = 0$ and $\lim_{k\to+\infty}\Gamma_k = +\infty$,

$$\lim_{n\to+\infty} W_1^2(\delta_x Q_\gamma^n, \pi) = 0\ ,$$

  with explicit convergence.
- Order of convergence for decreasing stepsize.

	$\alpha \in (0,1)$	$\alpha = 1$
Order of convergence	$\mathcal{O}(n^{-\alpha})$	$\mathcal{O}(n^{-1})$

Table : Order of convergence of $W_2(\delta_x Q_\gamma^n, \pi)$ for $\gamma_k = \gamma_1 k^{-\alpha}$

When $(\gamma_k)_{k\geq 1}$ is constant:

- We optimize γ and n to get $W_1(\delta_x Q_\gamma^n, \pi) \leq \epsilon$. In particular, we need

$$n = \mathcal{O}(d\epsilon^{-2}) \ .$$

- If the number of iterations n is fixed, we can optimize γ and we find a bound in $W_1(\delta_x Q_\gamma^n, \pi) \leq \mathcal{O}(\sqrt{n})$.

- To improve the bound, we make a regularity assumption on U: The potential U is three times continuously differentiable and there exists $\tilde{L}$ such that for all $x, y \in \mathbb{R}^d$:

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

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

$$W_1^2(\delta_x Q_\gamma^n, \pi) \leq u_n^{(3)}(\gamma) \int_{\mathbb{R}^d} \left\|y - x\right\|^2 \pi(\mathrm{d}y) + u_n^{(4)}(\gamma) \ ,$$

where $(u_n^{(3)}(\gamma), u_n^{(4)}(\gamma))_{n\geq 1}$ are explicit.

- Order of convergence for decreasing stepsize.

	$\alpha \in (0,1)$	$\alpha = 1$
Order of convergence	$\mathcal{O}(n^{-2\alpha})$	$\mathcal{O}(n^{-2})$

Table : Order of convergence of $W_1(\delta_x Q_\gamma^n, \pi)$ for $\gamma_k = \gamma_1 k^{-\alpha}$

When $(\gamma_k)_{k\geq 1}$ is constant:

- We optimize γ and n to get $W_1(\delta_x Q_\gamma^n, \pi) \leq \epsilon$. In particular, we need

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

- If the number of iterations n is fixed, we can optimize γ and we find a bound in $W_1(\delta_x Q_\gamma^n, \pi) \leq \mathcal{O}(n^{-1})$.

Convergence of the Euler discretization

- If we assume that U is convex, L-smooth.
- Explicit bound for $\|\delta_x Q_\gamma^p - \pi\|_{\mathrm{TV}}$.
- If $\lim_{\gamma_k \to +\infty} \gamma_k = 0$, and $\sum_k \gamma_k = +\infty$ then

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

- Computable bounds for the convergence.

Target precision ϵ: the convex case

- For constant stepsizes, We can optimize γ and p to get

$$\|\delta_x Q_\gamma^p - \pi\|_{\mathrm{TV}} \leq \epsilon \ .$$

	d	ε	L
γ	$\mathcal{O}(d^{-4})$	$\mathcal{O}(\varepsilon^2/\log(\varepsilon^{-1}))$	$\mathcal{O}(L^{-2})$
p	$\mathcal{O}(d^7)$	$\mathcal{O}(\varepsilon^{-2}\log^2(\varepsilon^{-1}))$	$\mathcal{O}(L^2)$

- We can also at fixed iteration, optimize the stepsize γ.
- Dependence on the dimensions: comes from the fact that the convergence of the diffusion in the convex case also depends on the dimension.

Non-smooth potentials

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

Moreau-Yosida regularization

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

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

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

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

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

Properties of proximal operators

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

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

- The function h^λ is convex and continuously differentiable

$$\nabla \mathrm{h}^\lambda(x) = \lambda^{-1}(x - \mathrm{prox}_{\mathrm{h}}^\lambda(x)) \ .$$

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

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

which implies that the Moreau-Yosida envelope is L-smooth:
$\left\| \nabla \mathrm{h}^\lambda(x) - \nabla \mathrm{h}^\lambda(y) \right\| \leq \lambda^{-1} \left\| x - y \right\|$, 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 λ-Moreau Yosida envelope g^λ can be considered.
- This leads to the approximation of the potential $U^\lambda : \mathbb{R}^d \to \mathbb{R}$ defined for all $x \in \mathbb{R}^d$ by

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

Theorem

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

Some approximation results

Theorem

Assume (H).

1. *Then,* $\lim_{\lambda \to 0} \|\pi^\lambda - \pi\|_{\mathrm{TV}} = 0$.
2. *Assume in addition that* g *is Lipschitz. Then for all* $\lambda > 0$,
$$\|\pi^\lambda - \pi\|_{\mathrm{TV}} \leq \lambda \|g\|^2_{\mathrm{Lip}} \ .$$
3. *If* $g = \iota_{\mathcal{K}}$ *where* $\mathcal{K}$ *is a convex body of* $\mathbb{R}^d$. *Then for all* $\lambda > 0$ *we have*
$$\|\pi^\lambda - \pi\|_{\mathrm{TV}} \leq 2 \left(1 + \mathsf{D}(\mathcal{K}, \lambda)\right)^{-1} \ ,$$
where $\mathsf{D}(\mathcal{K}, \lambda)$ *is explicit in the proof, and is of order* $\mathcal{O}(\lambda^{-1})$ *as* λ *goes to* 0.

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 $\{X_k^{\mathrm{M}},\ k \in \mathbb{N}\}$: for all $k \geq 0$,

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

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

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

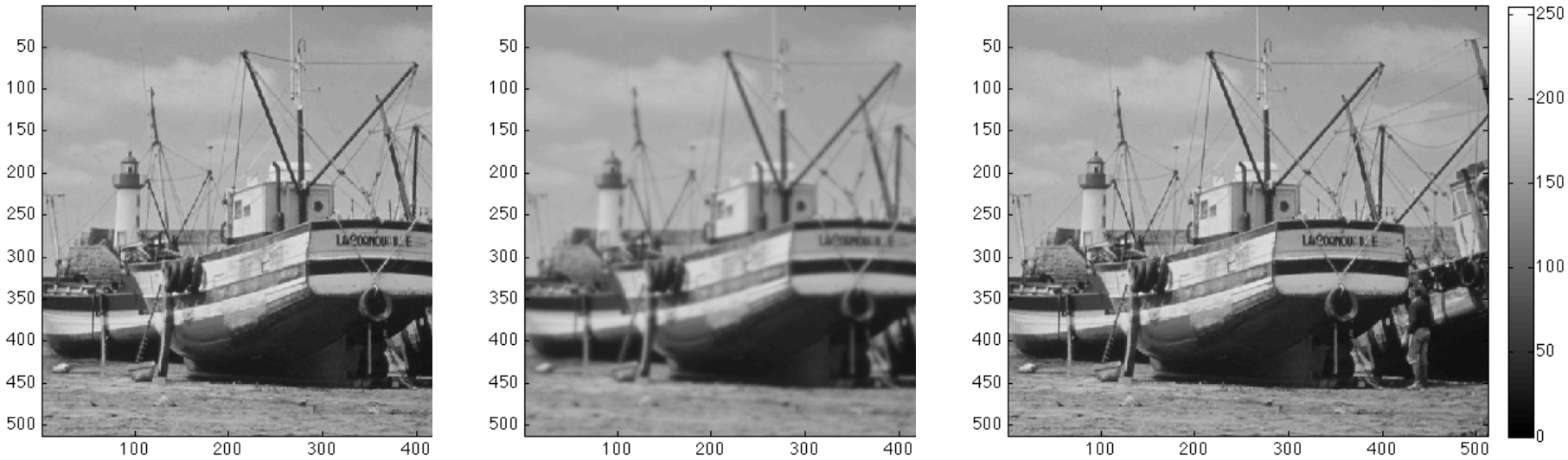
where for all $x \in \mathbb{R}^d$

$$\bar{g}^\lambda(x) = g^\lambda(x) - g(x) = g(\mathrm{prox}_g^\lambda(x)) - g(x) + (2\lambda)^{-1} \left\| x - \mathrm{prox}_g^\lambda(x) \right\|^2 \; .$$

Image deconvolution

- Objective recover an original image $\boldsymbol{x} \in \mathbb{R}^n$ from a blurred and noisy observed image $\boldsymbol{y} \in \mathbb{R}^n$ related to $\boldsymbol{x}$ by the linear observation model $\boldsymbol{y} = H\boldsymbol{x} + \boldsymbol{w}$, where H is a linear operator representing the blur point spread function and $\boldsymbol{w}$ is a Gaussian vector with zero-mean and covariance matrix $\sigma^2 \boldsymbol{I}_n$.
- This inverse problem is usually ill-posed or ill-conditioned: exploits prior knowledge about $\boldsymbol{x}$.
- One of the most widely used image prior for deconvolution problems is the improper total-variation norm prior, $\pi(\boldsymbol{x}) \propto \exp\left(-\alpha \|\nabla_d \boldsymbol{x}\|_1\right)$, where ∇_d denotes the discrete gradient operator that computes the vertical and horizontal differences between neighbour pixels.

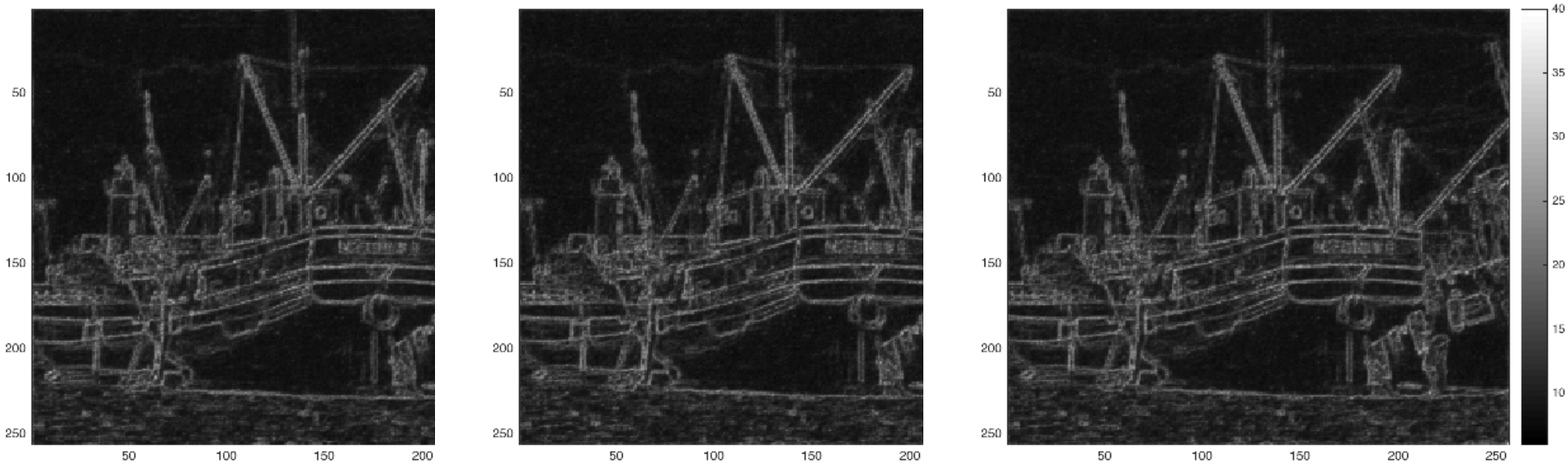
$$\pi(\boldsymbol{x}|\boldsymbol{y}) \propto \exp\left[-\|\boldsymbol{y} - H\boldsymbol{x}\|^2/2\sigma^2 - \alpha\|\nabla_d \boldsymbol{x}\|_1\right].$$



(a) (b) (c)

Figure : (a) Original Boat image (256×256 pixels), (b) Blurred image, (c) MAP estimate.

Credibility intervals



(a) (b) (c)

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).

What's next ?

- Extension of this work
  - Richardson-Romberg interpolation: debiaising for smooth functionnals with non-asymptotic bounds on the MSE.
  - Langevin meets Gibbs: ULA within Gibbs.
  - detailed comparison with MALA
- Thank you for your attention.