Sampling from log-concave non-smooth densities, when Moreau meets Langevin.

Alain Durmus, Eric Moulines

Telecom ParisTech, LTCI, Ecole Polytechnique, CMAP, University of Bristol, Stat. Lab.

Many thanks to Andreas Eberle and Arnaud Guillin for sharing their invaluable insights and comments and introducing us to their new results. Thanks to Marcelo Pereira (Univ. Bristol) for running the numerical experiments -not presented here- and pushing us to sharpen our results

Introduction

- Sampling distribution over high-dimensional state-space has recently attracted a lot of research efforts in computational statistics and machine learning...
- Applications (non-exhaustive)
  1. Bayesian inference for high-dimensional models (
  2. Bayesian non parametrics
  3. Aggregation of estimators and experts
  4. Bayesian linear inverse problems (typically function space problems converted to high-dimensional problem by Galerkin method)
- Most of the sampling techniques known so far do not scale to high-dimension... Challenges are numerous in this area...

Logistic 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.
- logistic regression: F is the standard logistic distribution function,

$$F(t) = \mathrm{e}^t/(1+\mathrm{e}^t)$$

.

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

Bayes 101

- Bayesian analysis requires a prior distribution for the unknown regression parameter

$$\pi(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\beta}'\Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}\right) \quad \text{or} \quad \pi(\boldsymbol{\beta}) = \exp\left(-\sum_{i=1}^{d}\alpha_i|\beta_i|\right)$$

.

- The posterior of $\boldsymbol{\beta}$ is up to a proportionality constant given by

$$\pi(\boldsymbol{\beta}|(Y,X)) \propto \prod_{i=1}^{n} F^{Y_i}(\beta' X_i)(1-F(\beta' X_i))^{1-Y_i}\pi(\boldsymbol{\beta})$$

Data Augmentation

- The most popular algorithms for Bayesian inference in binary regression models are based on data augmentation:
  - logistic link: Polya-Gamma sampler, Polsson and Scott (2012)
- Data Augmentation algorithm has 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 ?

- 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

$$U(\boldsymbol{\beta}) = -\sum_{i=1}^{p}\{Y_i \log \frac{F(\boldsymbol{\beta}^T X_i)}{1-F(\boldsymbol{\beta}^T X_i)} + \log(1-F(\boldsymbol{\beta}^T X_i))\} + \|B\boldsymbol{\beta}\|^{1,2}$$

- Classical composite objective function... The prior plays the role of regularization penalty.

Framework

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

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

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

- 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.
- $\pi \propto \mathrm{e}^{-U}$ is reversible $\rightsquigarrow$ the unique invariant probability measure.
- The convergence to the stationary distribution takes place at geometrical rate.
- Precise estimates of the convergence rate (TV, relative entropy) can be obtained using:
  - Functional inequalities: Poincaré or Log-Sobolev inequalities
  - Coupling techniques: synchronous or reflection coupling, depending upon the assumptions (Eberle, 2015)

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.
- Euler discretization = gradient algorithm + noise.

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 with density

$$r_\gamma(x,y) = (4\pi\gamma)^{-d/2} \exp\left(-(4\gamma)^{-1} \left\| y - x + \gamma \nabla U(x) \right\|^2 \right) .$$

- Under some appropriate conditions (a bit of positive curvature at infinity), this Markov chain is irreducible, positive recurrent $\rightsquigarrow$ unique invariant distribution π_γ.
- Problem: $\pi_\gamma \neq \pi$.

Metropolis-Adjusted Langevin Algorithm

- To correct the target distribution, a Metropolis-Hastings step can be included $\rightsquigarrow$ Metropolis Adjusted Langevin Agorithm (MALA).
  - Key references Roberts and Tweedie, 1996
- Algorithm:
  1. Propose $Y_{k+1} \sim X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} Z_{k+1}$, $Z_{k+1} \sim \mathcal{N}(0, \mathrm{I}_d)$
  2. Compute the acceptance ratio $\alpha_\gamma(X_k, Y_{k+1})$ where

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

  3. Accept the move with probability $\alpha_\gamma(X_k, Y_{k+1})$ / Reject the move and stay where you are.

MALA: pros and cons

- Require to evaluate two times the objective function.
- Geometric convergence is established under the condition that in the tail the acceptance region is inwards in q,

$$\lim_{\|x\| \to \infty} \int_{\mathcal{A}_\gamma(x) \Delta \mathcal{I}(x)} r_\gamma(x,y) \mathrm{d}y = 0 \ .$$

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

$$\mathcal{A}_\gamma(x) = \{y, \pi(x) r_\gamma(x,y) \leq \pi(y) r_\gamma(y,x)\}$$

- Optimal stepsize: scaling analysis - do not discussed here - suggests to choose the stepsize to achieve 50% of acceptance.

Foster-Lyapunov condition

- A function $V \in C^2(\mathbb{R}^d)$ is a Lyapunov function if $V \geq 1$ and if there exists $\theta > 0$, $b \geq 0$ and $R > 0$ such that,

$$\mathscr{A} V \leq -\theta V + b \mathbb{1}_{\mathrm{B}(0,R)} ,$$

where $\mathscr{A} f = - \langle \nabla U, \nabla f \rangle + \Delta f$ is the generator of the diffusion

- Example: If there exist $\alpha > 1$, $\rho > 0$ and $M_\rho \geq 0$ such that for all $y \in \mathbb{R}^d$, $\|y\| \geq M_\rho$:

$$\langle \nabla U(y), y \rangle \geq \rho \|y\|^\alpha \ .$$

then $V(x) = \exp(U(x)/2)$ is a Lyapunov function.

Geometric convergence of the Langevin diffusion

- If there exists a Lyapunov function for the generator of the diffusion then there exists $\kappa \in [0,1)$ such that for any initial distribution μ_0 and $t > 0$,
$$\|\mu_0 P_t - \pi\|_{\mathrm{TV}} \leq C(\mu_0)\kappa^t \ ,$$
for some explicit function of the initial probability $C(\mu_0)$.
- Explicit expressions of the constant (the way dimension impacts theses constants) critically depends on
  - the assumptions on the potential U
  - the technique of proofs (functional inequalities, coupling constructions, etc...)

Geometric convergence of the Euler discretization

- Let $(\gamma_k)_{k\geq 1}$ be a sequence of positive and non-increasing step sizes
- Euler discretization:

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

  where $(Z_k)_{k\geq 1}$ is i.i.d. $\mathcal{N}(0, \mathrm{I}_d)$, independent of X_0.
- Markov kernel R_γ and $x \in \mathbb{R}^d$ by

$$R_\gamma(x, A) = \int_A \frac{1}{(4\pi\gamma)^{d/2}} \exp\left(-(4\gamma)^{-1} \|y - x + \gamma \nabla U(x)\|^2\right) \mathrm{d}y \;.$$

- The sequence $(X_n)_{n\geq 0}$ is a (possibly) time-nonhomogeneous Markov chain whose distribution is specified by the Markov kernels $(R_{\gamma_n})_{n\geq 1}$.

Level-0 results

- The Markov kernel R_γ is strongly Feller, irreducible, and hence all the compact sets are therefore small.
- Typically, the R_γ satisfies a Foster-Lyapunov drift condition of a particular form, *i.e.* there exists $\kappa \in [0, 1)$, $b > 0$ such that for all $\gamma > 0$

$$R_\gamma V \leq \kappa^\gamma V + \gamma b \, .$$

- R_γ admits a unique stationary distribution π_γ and is V-uniformly geometrically ergodic, in the sense that, for some constant $C < \infty$ and $\kappa \in [0, 1)$, such that for all $x \in \mathbb{R}^d$,

$$\left\| R_\gamma^k(x, \cdot) - \pi_\gamma \right\|_V \leq C(\gamma) \kappa^{\gamma k} V(x) \, .$$

Example: A drift condition for R_γ

Theorem

Assume U is L-smooth and there exist $\rho > 0$, $\alpha > 1$ and $M_\rho \geq 0$ such that :

$$\langle \nabla U(y), y \rangle \geq \rho \left\| y \right\|^\alpha \ , \quad \text{for all } y \in \mathbb{R}^d, \ \|y\| \geq M_\rho$$

Then for all $\bar{\gamma} \in \left(0, L^{-1}\right)$, there exists $b \geq 0$ and $s > 0$ such that

$$R_\gamma V(x) \leq \kappa^\gamma V(x) + \gamma b \ , \quad \text{for all } \gamma \in (0, \bar{\gamma}] \text{ and } x \in \mathbb{R}^d,$$

where

$$V(x) = \exp(U(x)/2).$$

Control of moments

- By a straightforward induction, we get for all $n \geq 0$ and $x \in \mathbb{R}^d$,

$$Q_\gamma^n V \leq \kappa^{\Gamma_{1,n}} V + b \sum_{i=1}^{n} \gamma_i \kappa^{\Gamma_{i+1,n}} \ .$$

- Note that for all $n \geq 1$, we have

$$\sum_{i=1}^{n} \gamma_i \kappa^{\Gamma_{i+1,n}} \leq \gamma_1 (1 - \kappa^{\Gamma_{1,n}}) / (1 - \kappa^{\gamma_1}) \ .$$

Error decomposition

- For $n \leq p$ set $Q_\gamma^{n,p} = R_{\gamma_n} \cdots R_{\gamma_p}$,
- Error decomposition

$$\|\mu_0 Q_\gamma^p - \pi\|_{\mathrm{TV}} \leq \|\mu_0 Q_\gamma^n Q_\gamma^{n+1,p} - \mu_0 Q_\gamma^n P_{\Gamma_{n+1,p}}\|_{\mathrm{TV}} + \|\mu_0 Q_\gamma^n P_{\Gamma_{n+1,p}} - \pi\|_{\mathrm{TV}} \ .$$

where

$$\Gamma_{n,p} \stackrel{\text{def}}{=} \sum_{k=n}^{p} \gamma_k \ , \qquad \Gamma_n = \Gamma_{1,n} \ .$$

- Second term on the RHS: contraction of the markov semi-group.
- Problem: Find a way to compare the total variation distance between the diffusion and its discretization started at time Γ_n from the same distribution.

Coupling

- For all $x \in \mathbb{R}^d$, denote by $\mu_{n,p}^x$ and $\bar{\mu}_{n,p}^x$ the laws on $\mathrm{C}([\Gamma_n, \Gamma_p], \mathbb{R}^d)$ of the Langevin diffusion $(Y_t)_{\Gamma_n \leq t \leq \Gamma_p}$ and of the Euler discretisation $(\bar{Y}_t)_{\Gamma_n \leq t \leq \Gamma_p}$ both started at x at time Γ_n.
- For any $\zeta_0 \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$, consider the diffusion $(Y_t, \overline{Y}_t)_{t \geq 0}$ with initial distribution equals to ζ_0, and defined for $t \geq 0$ by

$$\begin{cases} \mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \\ \mathrm{d}\bar{Y}_t = -\overline{\nabla U}(\bar{Y}_t, t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \end{cases}$$

and

$$\overline{\nabla U}(y, t) = \sum_{k=0}^{\infty} \nabla U(y_{\Gamma_k}) \mathbb{1}_{[\Gamma_k, \Gamma_{k+1})}(t)$$

Change of measure

- The Girsanov theorem shows that $\mu^x_{n,p} \sim \bar{\mu}^x_{n,p}$ with density

$$\frac{\mathrm{d}\mu^x_{n,p}}{\mathrm{d}\bar{\mu}^x_{n,p}} = \exp\Big(\frac{1}{2}\int_{\Gamma_n}^{\Gamma_p} \langle \nabla U(\bar{Y}_s) - \overline{\nabla U}(\bar{Y}_s), s, \mathrm{d}\bar{Y}_s \rangle \\ - \frac{1}{4}\int_{\Gamma_n}^{\Gamma_p} \left\{ \left\|\nabla U(\bar{Y}_s)\right\|^2 - \left\|\overline{\nabla U}(\bar{Y}_s, s)\right\|^2 \right\} \mathrm{d}s \Big) .$$

- The Pinsker inequality implies that for all $x \in \mathbb{R}^d$

$$\|\delta_x Q^{n+1,p}_\gamma - \delta_x P_{\Gamma_{n+1,p}}\|_{\mathrm{TV}} \leq 2^{-1} \left(\mathrm{Ent}_{\bar{\mu}^x_{n,p}} \left(\frac{\mathrm{d}\mu^x_{n,p}}{\mathrm{d}\bar{\mu}^x_{n,p}} \right) \right)^{1/2} \\ \leq 4^{-1} \left(\int_{\Gamma_n}^{\Gamma_p} \mathbb{E}_x \left[\left\|\nabla U(\bar{Y}_s) - \overline{\nabla U}(\bar{Y}_s, s)\right\|^2 \right] \mathrm{d}s \right)^{1/2} .$$

Change of measure

- Pinsker inequality: for all $x \in \mathbb{R}^d$

$$\|\delta_x Q_\gamma^{n+1,p} - \delta_x P_{\Gamma_{n+1,p}}\|_{\mathrm{TV}} \le 4^{-1} \left(\int_{\Gamma_n}^{\Gamma_p} \mathbb{E}_x \left[\left\| \nabla U(\bar{Y}_s) - \overline{\nabla U}(\bar{Y}_s, s) \right\|^2 \right] \mathrm{d}s \right)^{1/2} .$$

- If U is L-smooth,

$$\|\delta_x Q_\gamma^{n+1,p} - \delta_x P_{\Gamma_{n+1,p}}\|_{\mathrm{TV}} \le 4^{-1} L \left(\sum_{k=n+1}^{p} \left\{ (\gamma_k^3/3) \mathbb{E}_x \left[\|\nabla U(X_k)\|^2 \right] + d\gamma_k^2 \right\} \right)^{1/2} .$$

Back to the decomposition of the error

$$\|\mu_0 Q_\gamma^p - \pi\|_{\mathrm{TV}} \leq \|\mu_0 Q_\gamma^p - \mu_0 Q_\gamma^n P_{\Gamma_{n+1,p}}\|_{\mathrm{TV}} + \|\mu_0 Q_\gamma^n P_{\Gamma_{n+1,p}} - \pi\|_{\mathrm{TV}} \ .$$

- Main result: For all $n, p \geq 1$, $n \leq p$, and $x \in \mathbb{R}^d$

$$\|\mu_0 Q_\gamma^p - \pi\|_{\mathrm{TV}} \leq C(\mu_0 Q_\gamma^n)\lambda^{\Gamma_{n+1,p}} + \left(D(d, \gamma, \mu_0) \sum_{k=n+1}^{p} \gamma_k^2 \right)^{1/2}$$

- If $\sum_k \gamma_k = \infty$, then

$$\|\mu_0 Q_\gamma^p - \pi\|_{\mathrm{TV}} \to 0 \ , \quad p \to \infty \ .$$

Controlling π_γ

- How far π_γ is from π ?
- Under the stated conditions, there exists an explicit constant $C(d)$ such that for all $\gamma \in [0, \bar{\gamma})$,

$$\|\pi - \pi_\gamma\|_V \leq C(d)\gamma^{1/2} .$$

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 $x, y \in \mathbb{R}^d$,

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

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* $\lambda > 0$,

$$\|\pi^\lambda - \pi\|_{\mathrm{TV}} \leq \lambda \|g\|_{\mathrm{Lip}}^2 \ .$$

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

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

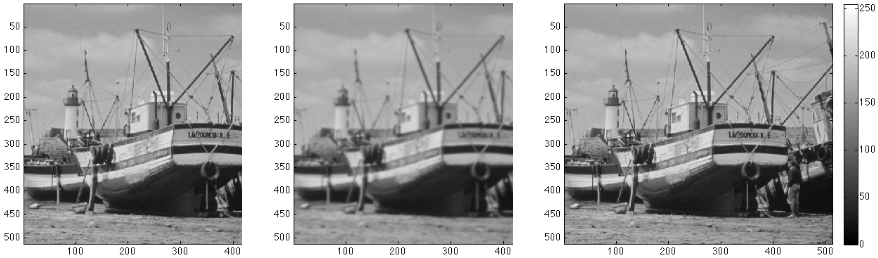
where for all $x \in \mathbb{R}^d$

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

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 $\boldsymbol{w} \sim \mathrm{N}(0, \sigma^2 \boldsymbol{I}_n)$.
- 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.

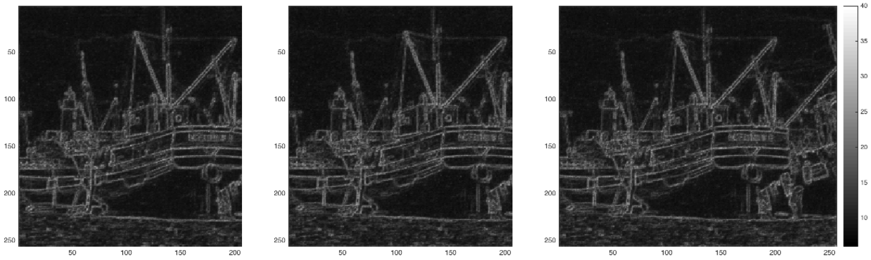
$$\pi(\boldsymbol{x}|\boldsymbol{y}) \propto \exp\left[-\|\boldsymbol{y} - H\boldsymbol{x}\|^2/2\sigma^2 - \alpha\|\nabla_d \boldsymbol{x}\|_1\right].$$



(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 ?

- A simple algorithm which scale easily in the dimension of the problem
- Computable bounds for convergence in TV, MSE, and deviation inequalities with constants which make sense !
- Future works
  - partial updates (coordinate descent)
  - detailed comparison with MALA
  - bias reduction (”exact estimation” à la Glynn and Rhee ?)