Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau

Dr. Marcelo Pereyra
http://www.macs.hw.ac.uk/~mp71/

Maxwell Institute for Mathematical Sciences, Heriot-Watt University

June 2017, Heriot-Watt, Edinburgh.

Joint work with Alain Durmus and Eric Moulines

Outline

Imaging inverse problems

- We are interested in an unknown $x \in \mathbb{R}^d$.
- We measure y, related to x by a statistical model $p(y|x)$.
- The recovery of x from y is ill-posed or ill-conditioned, resulting in significant uncertainty about x.
- For example, in many imaging problems

$$y = Ax + w,$$

for some operator A that is rank-deficient, and additive noise w.

The Bayesian framework

- We use priors to reduce uncertainty and deliver accurate results.
- Given the prior $p(x)$, the posterior distribution of x given y

$$p(x|y) = p(y|x)p(x)/p(y)$$

  models our knowledge about x after observing y.
- In this talk we consider that $p(x|y)$ is log-concave; i.e.,

$$p(x|y) = \exp\{-\phi(x)\}/Z\,,$$

  where $\phi(x)$ is a convex function and $Z = \int \exp\{-\phi(x)\}\mathrm{d}x$.

Inverse problems in mathematical imaging

More precisely, we consider models of the form

$$p(x|y) \propto \exp\{-f(x) - g(x)\} \tag{1}$$

where $f(x)$ and $g(x)$ are lower semicontinuous convex functions from $\mathbb{R}^d \to (-\infty, +\infty]$ and f is L_f-Lipschitz differentiable. For example,

$$f(x) = \frac{1}{2\sigma^2}\|y - Ax\|_2^2$$

for some observation $y \in \mathbb{R}^p$ and linear operator $A \in \mathbb{R}^{p\times n}$, and

$$g(x) = \alpha\|Bx\|_\dagger + \mathbf{1}_{\mathcal{S}}(x)$$

for some norm $\|\cdot\|_\dagger$, dictionary $B \in \mathbb{R}^{n\times n}$, and convex set $\mathcal{S}$. Often, $g \notin \mathcal{C}^1$.

Maximum-a-posteriori (MAP) estimation

The predominant Bayesian approach in imaging is MAP estimation

$$\begin{aligned} \hat{x}_{MAP} &= \operatorname*{argmax}_{x \in \mathbb{R}^d} p(x|y), \\ &= \operatorname*{argmin}_{x \in \mathbb{R}^d} f(x) + g(x), \end{aligned} \tag{2}$$

that can be computed efficiently by “proximal” convex optimisation.

For example, the *proximal gradient algorithm*

$$x^{m+1} = \text{prox}_g^{L^{-1}}\{x^m + L^{-1}\nabla f(x^m)\},$$

with $\text{prox}_g^{\lambda}(x) = \text{argmax}_{u \in \mathbb{R}^{\mathbb{N}}}\ g(u) - \frac{1}{2\lambda}\|u - x\|^2$ converges at rate $O(1/m)$.

However, $\hat{x}_{MAP}$ provides very little about $p(x|y)$.

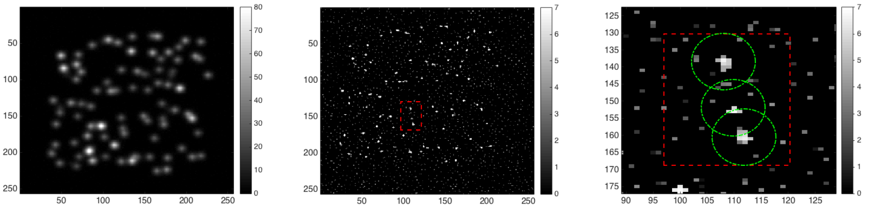
Illustrative example: image resolution enhancement

Recover $x \in \mathbb{R}^d$ from low resolution and noisy measurements

$$y = Hx + w,$$

where H is a circulant blurring matrix. We use the Bayesian model

$$p(x|y) \propto \exp\left(-\|y - Hx\|^2/2\sigma^2 - \beta\|x\|_1\right). \tag{3}$$



y $\hat{x}_{MAP}$ Uncertainty estimates?

Figure : Resolution enhancement of the `Molecules` image of size 256×256 pixels.

Illustrative example: tomographic image reconstruction

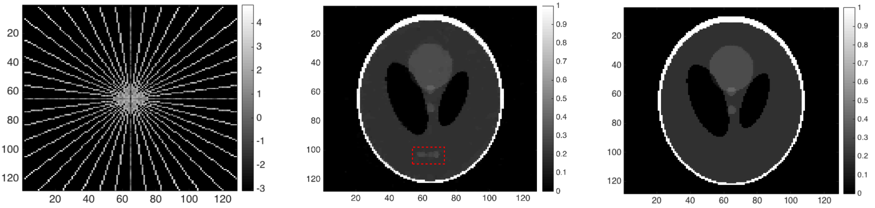
Recover $x \in \mathbb{R}^d$ from partially observed and noisy Fourier measurements

$$y = \Phi \mathcal{F} x + w,$$

where Φ is a mask and $\mathcal{F}$ is the 2D Fourier operator. We use the model

$$p(x|y) \propto \exp\left(-\|y - \Phi \mathcal{F} x\|^2 / 2\sigma^2 - \beta \|\nabla_d x\|_{1-2}\right), \tag{4}$$

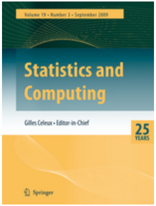
where ∇_d is the 2d discrete gradient operator and $\|\cdot\|_{1-2}$ the $\ell_1 - \ell_2$ norm.



y $\hat{x}_{MAP}$ Possible solution?

Figure : Tomographic reconstruction of the `Shepp-Logan` phantom image.

Recent surveys on Bayesian computation...



25th anniversary special issue on Bayesian computation

P. Green, K. Latuszynski, M. Pereyra, C. P. Robert, "Bayesian computation: a perspective on the current state, and sampling backwards and forwards", Statistics and Computing, vol. 25, no. 4, pp 835-862, Jul. 2015.

Special issue on "Stochastic simulation and optimisation in signal processing"

M. Pereyra, P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. Hero, and S. McLaughlin, "A Survey of Stochastic Simulation and Optimization Methods in Signal Processing" IEEE Sel. Topics in Signal Processing, in press.

Outline

1. Bayesian inference in imaging inverse problems
2. Proximal Markov chain Monte Carlo
3. Experiments
4. Conclusion

Inference by Markov chain Monte Carlo integration

Monte Carlo integration
Given a set of samples $X_1, \ldots, X_M$ distributed according to $p(x|y)$, we approximate posterior expectations and probabilities

$$\frac{1}{M}\sum_{m=1}^{M} h(X_m) \to \mathrm{E}\{h(x)|y\}, \quad \text{as } M \to \infty$$

Guarantees from CLTs, e.g., $\frac{1}{\sqrt{M}}\sum_{m=1}^{M} h(X_m) \sim \mathcal{N}[\mathrm{E}\{h(x)|y\}, \Sigma]$.

Markov chain Monte Carlo:
Construct a Markov kernel $X_{m+1}|X_m \sim K(\cdot|X_m)$ such that the Markov chain $X_1, \ldots, X_M$ has $p(x|y)$ as stationary distribution.

MCMC simulation in high-dimensional spaces is very challenging.

Unadjusted Langevin algorithm

Suppose for now that $p(x|y) \in \mathcal{C}^1$. Then, we can generate samples by mimicking a Langevin diffusion process that converges to $p(x|y)$ as $t \to \infty$,

$$\mathbf{X}: \quad \mathrm{d}\mathbf{X}_t = \frac{1}{2}\nabla \log p\left(\mathbf{X}_t|y\right)\mathrm{d}t + \mathrm{d}W_t, \quad 0 \leq t \leq T, \quad \mathbf{X}(0) = x_0.$$

where W is the n-dimensional Brownian motion.

Because solving $\mathbf{X}_t$ exactly is generally not possible, we use an Euler Maruyama approximation and obtain the "unadjusted Langevin algorithm"

$$\mathrm{ULA}: \ \ X_{m+1} = X_m + \delta\nabla \log p(X_m|y) + \sqrt{2\delta}Z_{m+1}, \ \ Z_{m+1} \sim \mathcal{N}(0, \mathbb{I}_n)$$

ULA is remarkably efficient when $p(x|y)$ is sufficiently regular.

Unadjusted Langevin algorithm

However, our interest is in high-dimensional models of the form

$$p(x|y) \propto \exp\{-f(x) - g(x)\}$$

with f, g l.s.c. convex, ∇f L_f-Lipschitz continuous, and $g \notin \mathcal{C}^1$.

Unfortunately, such models are beyond the scope of ULA, which may perform poorly if $p(x|y)$ is not Lipchitz differentiable.

Idea: Regularise $p(x|y)$ to enable efficiently Langevin sampling.

Approximation of $p(x|y)$

Moreau-Yoshida approximation of $p(x|y)$ (Pereyra, 2015):

Let $\lambda > 0$. We propose to approximate $p(x|y)$ with the density

$$p_\lambda(x|y) = \frac{\exp[-f(x) - g_\lambda(x)]}{\int_{\mathbb{R}^d} \exp[-f(x) - g_\lambda(x)]\mathrm{d}x},$$

where g_λ is the Moreau-Yoshida envelope of g given by

$$g_\lambda(x) = \inf_{u \in \mathbb{R}^d} \{g(u) - (2\lambda)^{-1}\|u - x\|_2^2\},$$

and where λ controls the approximation error involved.

Moreau-Yoshida approximations

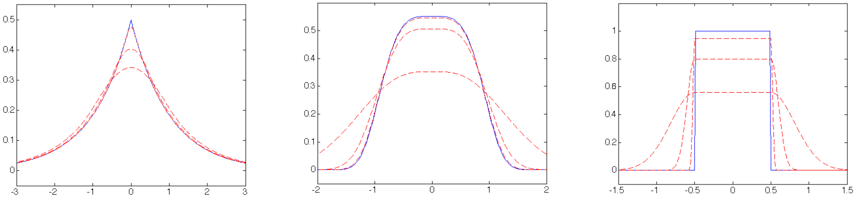
Key properties (Pereyra, 2015; Durmus et al., 2017):

1. $\forall \lambda > 0$, p_λ defines a proper density of a probability measure on $\mathbb{R}^d$.
2. *Convexity and differentiability*:
   - p_λ is log-concave on $\mathbb{R}^d$.
   - $p_\lambda \in \mathcal{C}^1$ even if p not differentiable, with

$$\nabla \log p_\lambda(x|y) = -\nabla f(x) + \{\text{prox}_g^\lambda(x) - x\}/\lambda,$$

     and $\text{prox}_g^\lambda(x) = \text{argmax}_{u \in \mathbb{R}^{\mathbb{N}}}\, g(u) - \frac{1}{2\lambda}\|u - x\|^2$.
   - $\nabla \log p_\lambda$ is Lipchitz continuous with constant $L \leq L_f + \lambda^{-1}$.
3. *Approximation error between $p_\lambda(x|y)$ and $p(x|y)$*:
   - $\lim_{\lambda \to 0} \|p_\lambda - p\|_{TV} = 0$.
   - If g is L_g-Lipchitz, then $\|p_\lambda - p\|_{TV} \leq \lambda L_g^2$.

Examples of Moreau-Yoshida approximations:



$p(x) \propto \exp\left(-|x|\right)$ $\quad$ $p(x) \propto \exp\left(-x^4\right)$ $\quad$ $p(x) \propto \mathbf{1}_{[-0.5,0.5]}(x)$

Figure : True densities (solid blue) and approximations (dashed red).

Proximal ULA

We approximate $\mathbf{X}$ with the "regularised" auxiliary Langevin diffusion

$$\mathbf{X}^{\lambda}: \quad \mathrm{d}\mathbf{X}_t^{\lambda} = \frac{1}{2}\nabla \log p_{\lambda}\left(\mathbf{X}_t^{\lambda}|y\right)\mathrm{d}t + \mathrm{d}W_t, \quad 0 \leq t \leq T, \quad \mathbf{X}^{\lambda}(0) = x_0,$$

which targets $p_{\lambda}(x|y)$. Remark: we can make $\mathbf{X}^{\lambda}$ arbitrarily close to $\mathbf{X}$.

Finally, an Euler Maruyama discretisation of $\mathbf{X}^{\lambda}$ leads to the (Moreau-Yoshida regularised) proximal ULA

$$\mathrm{MYULA}: \quad X_{m+1} = (1 - \tfrac{\delta}{\lambda})X_m - \delta\nabla f\{X_m\} + \tfrac{\delta}{\lambda}\,\mathrm{prox}_g^{\lambda}\{X_m\} + \sqrt{2\delta}Z_{m+1},$$

where we used that $\nabla g_{\lambda}(x) = \{x - \mathrm{prox}_g^{\lambda}(x)\}/\lambda$.

Convergence results

Non-asymptotic estimation error bound

Theorem 2.1 (Durmus et al. (2017))

Let $\delta_\lambda^{max} = (L_1 + 1/\lambda)^{-1}$. Assume that g is Lipchitz continuous. Then, there exist $\delta_\epsilon \in (0, \delta_\lambda^{max}]$ and $M_\epsilon \in \mathbb{N}$ such that $\forall \delta < \delta_\epsilon$ and $\forall M \geq M_\epsilon$

$$\|\delta_{x_0} Q_\delta^M - p\|_{TV} < \epsilon + \lambda L_g^2,$$

where Q_δ^M is the kernel assoc. with M iterations of MYULA with step δ.

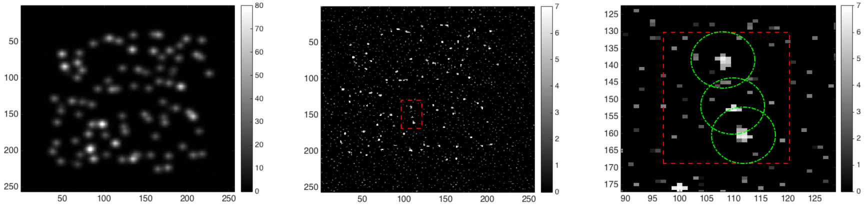
Note: δ_ϵ and M_ϵ are explicit and tractable. If $f + g$ is strongly convex outside some ball, then M_ϵ scales with order $\mathcal{O}(d \log(d))$ (otherwise at worse $\mathcal{O}(d^5)$). See Durmus et al. (2017) for other convergence results.

Outline

Sparse image deblurring

Bayesian **credible region** $C^*_\alpha = \{x : p(x|y) \geq \gamma_\alpha\}$ with

$$\mathrm{P}\left[x \in C_\alpha | y\right] = 1 - \alpha, \quad \text{and} \quad p(x|y) \propto \exp\left(-\|y - Hx\|^2/2\sigma^2 - \beta\|x\|_1\right)$$



y $\hat{x}_{MAP}$ uncertainty estimates

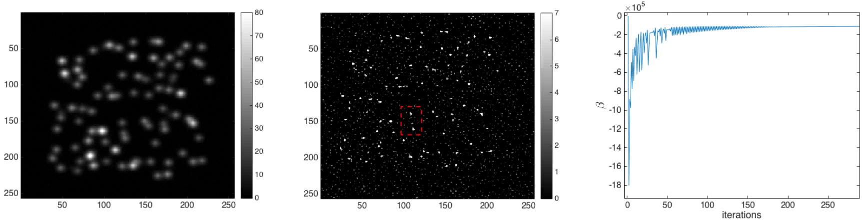
Figure : Live-cell microscopy data (Zhu et al., 2012). Uncertainty analysis ($\pm 78nm \times \pm 125nm$) in close agreement with the experimental precision $\pm 80nm$.

Computing time 4 minutes. $M = 10^5$ iterations. Estimation error 0.2%..

Sparse image deblurring

Estimation of reg. param. β by marginal maximum likelihood

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^+}{\operatorname{argmax}}\, p(y|\beta), \quad \text{with} \quad p(y|\beta) \propto \int \exp\left(-\|y - Hx\|^2/2\sigma^2 - \beta\|x\|_1\right)\mathrm{d}x$$



y $\qquad$ $\hat{x}_{MAP}$ $\qquad$ Reg. param β

Figure : Maximum marginal likelihood estimation of regularisation parameter β.

Computing time 0.75 secs..

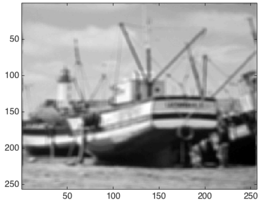
Bayesian model selection

$$p(\mathcal{M}_k|y) = p(\mathcal{M}_k) \int p(x, y|\mathcal{M}_k)\mathrm{d}x / p(y) \quad \text{with}$$

$$p(x, y|\mathcal{M}_1) \propto \exp\left[-(\|y - {\color{red}H_1}x\|^2/2\sigma^2) - \beta TV(x)\right],$$

$$p(x, y|\mathcal{M}_2) \propto \exp\left[-(\|y - {\color{red}H_2}x\|^2/2\sigma^2) - \beta TV(x)\right].$$

Boat image deblurring experiment (comp. time 30 minutes p/model):



observation y

(5×5 uniform blur, BSNR 40dB)

$\hat{x}_{\mathcal{M}_1}$ (PSNR 34dB)

${\color{red}p(\mathcal{M}_1|y) = 0.96}$

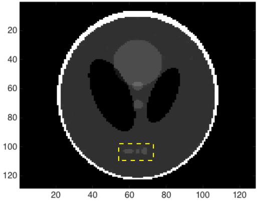
$\hat{x}_{\mathcal{M}_2}$ (PSNR 33dB)

${\color{red}p(\mathcal{M}_2|y) = 0.04}$

Uncertainty quantification of MRI tomographic image

Bayesian **credible region** $C^*_\alpha = \{x : p(x|y) \geq \gamma_\alpha\}$ with

$$\mathrm{P}\left[x \in C_\alpha | y\right] = 1 - \alpha, \quad \text{and} \quad p(x|y) \propto \exp\left(-\|y - \Phi\mathcal{F}x\|^2/2\sigma^2 - \beta\|\nabla_d x\|_{1-2}\right),$$



$\hat{x}_{MAP}$ (tumour intensity 0.30)

min. tumour intensity 0.27

max. tumour intensity 0.33

Figure : Shepp-Logan experiment: uncertainty in tumour intensity 10%.

Computing time 1 minute. $M = 10^5$ iterations. Estimation error 3%. .

Outline

Conclusion

- The challenges facing modern image processing require a paradigm shift, and a new wave of analysis and computation methodologies.
- Great potential for synergy between Bayesian and variational approaches at algorithmic, methodological, and theoretical levels.
- MYULA delivers reliable and computationally efficient approximate inferences, with good control of accuracy vs. computing-time.

Thank you!

Bibliography:

Durmus, A., Moulines, E., and Pereyra, M. (2017). Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM J. Imaging Sci.* to appear.

Pereyra, M. (2015). Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*. open access paper, http://dx.doi.org/10.1007/s11222-015-9567-4.

Zhu, L., Zhang, W., Elnatan, D., and Huang, B. (2012). Faster STORM using compressed sensing. *Nat. Meth.*, 9(7):721–723.