Bayesian inference and convex geometry: theory, methods, and algorithms.

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

Maxwell Institute for Mathematical Sciences, Heriot-Watt University

November 2018, Oaxaca.

Outline

Imaging inverse problems

- We are interested in an unknown image $x \in \mathbb{R}^d$.
- 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 $Z = \int \exp\{-\phi(x)\}\mathrm{d}x$.

Maximum-a-posteriori (MAP) estimation

The predominant Bayesian approach in imaging is MAP estimation

$$
\begin{aligned}
\hat{x}_{MAP} &= \underset{x \in \mathbb{R}^d}{\operatorname{argmax}}\, p(x|y), \\
&= \underset{x \in \mathbb{R}^d}{\operatorname{argmin}}\, \phi(x),
\end{aligned} \tag{1}
$$

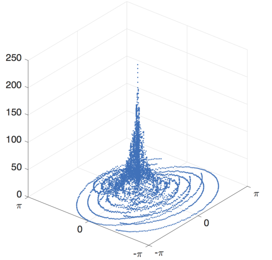
computed efficiently, even in very high dimensions, by (proximal) convex optimisation (Green et al., 2015; Chambolle and Pock, 2016).

Recover $x \in \mathbb{R}^d$ from low-dimensional degraded observation

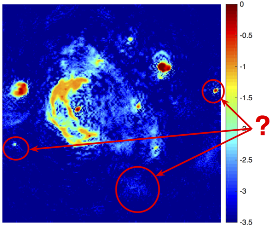
$$y = M\mathcal{F}x + w,$$

where $\mathcal{F}$ is the continuous Fourier transform, $M \in \mathbb{C}^{m\times d}$ is a measurement operator and w is Gaussian noise. We use the model

$$p(x|y) \propto \exp\left(-\|y - M\mathcal{F}x\|^2/2\sigma^2 - \theta\|\Psi x\|_1\right)\mathbf{1}_{\mathbb{R}^n_+}(x). \tag{2}$$



y



$\hat{x}_{MAP}$

Figure : Radio-interferometric image reconstruction of the `W28 supernova`.

MAP estimation by proximal optimisation

To compute $\hat{x}_{MAP}$ we use a proximal splitting algorithm. Let

$$f(x) = \|y - M\mathcal{F}x\|^2/2\sigma^2\,, \quad \text{and} \quad g(x) = \theta\|\Psi x\|_1 + -\log \mathbf{1}_{\mathbb{R}^n_+}(x)\,,$$

where f and g are l.s.c. convex on $\mathbb{R}^d$, and f is L_f-Lipschitz differentiable.

For example, we could use a **proximal gradient** iteration

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

converges to $\hat{x}_{MAP}$ at rate $O(1/m)$, with poss. acceleration to $O(1/m^2)$.

Definition Proximity mappings of a convex function g: For $\lambda > 0$, the λ-proximity mapping of g is defined as (Moreau, 1962)

$$\text{prox}_g^{\lambda}(x) \triangleq \underset{u\in\mathbb{R}^{\mathbb{N}}}{\text{argmin}}\; g(u) + \frac{1}{2\lambda}\|u - x\|^2.$$

MAP estimation by proximal optimisation

The **alternating direction method of multipliers (ADMM)** algorithm

$$\begin{aligned} x^{m+1} &= \text{prox}_f^{\lambda}\{z^m - u^m\}, \\ z^{m+1} &= \text{prox}_g^{\lambda}\{x^{m+1} + u^m\}, \\ u^{m+1} &= u^m + x^{m+1} - z^{m+1}, \end{aligned}$$

also converges to $\hat{x}_{MAP}$ very quickly, and does not require f to be smooth.

However, MAP estimation has some limitations, e.g.,

1. it provides little information about $p(x|y)$,
2. it struggles with unknown/partially unknown models,
3. it is not theoretically well understood (yet).

Outline

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

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

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

Suppose that

$$p(x|y) \propto \exp\{-f(x) - g(x)\} \tag{3}$$

where $f(x)$ and $g(x)$ are l.s.c. convex functions from $\mathbb{R}^d \to (-\infty, +\infty]$, f is L_f-Lipschitz differentiable, and $g \notin \mathcal{C}^1$.

For example,

$$f(x) = \frac{1}{2\sigma^2}\|y - Ax\|_2^2, \quad g(x) = \alpha\|Bx\|_\dagger + \mathbf{1}_{\mathcal{S}}(x),$$

for some linear operators A, B, norm $\|\cdot\|_\dagger$, and convex set $\mathcal{S}$.

Unfortunately, such non-models are beyond the scope of ULA.

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., 2018):

1. $\forall \lambda > 0$, p_λ defines a proper density of a probability measure on $\mathbb{R}^d$.

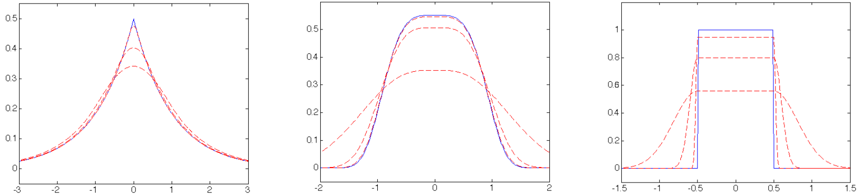
2. *Convexity and differentiability*:
   - p_λ is log-concave on $\mathbb{R}^d$.
   - $p_\lambda \in \mathcal{C}^1$ even if p not differentiable, with

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

   and $\text{prox}_g^\lambda(x) = \text{argmin}\, u \in \mathbb{R}^{\mathbb{N}}\, g(u) + \frac{1}{2\lambda}\|u - x\|^2$.
   - $\nabla \log p_\lambda$ 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:



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

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. (2018))

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 $\mathcal{O}(d \log(d))$. See Durmus et al. (2018) for other convergence results.

Outline

1. Bayesian inference in imaging inverse problems
2. Proximal Markov chain Monte Carlo
3. **Uncertainty quantification in astronomical and medical imaging**
4. Empirical Bayes estimation with unknown regularisation parameters
5. Conclusion

Where does the posterior probability mass of x lie?

- A set C_α is a posterior credible region of confidence level $(1-\alpha)\%$ if

$$\mathrm{P}\left[x \in C_\alpha | y\right] = 1 - \alpha.$$

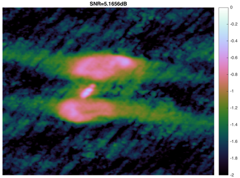
- The *highest posterior density* (HPD) region is decision-theoretically optimal (Robert, 2001)

$$C^*_\alpha = \{x : \phi(x) \leq \gamma_\alpha\}$$

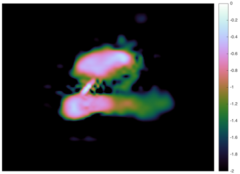
with $\gamma_\alpha \in \mathbb{R}$ chosen such that $\int_{C^*_\alpha} p(x|y)\mathrm{d}x = 1 - \alpha$ holds.

Visualising uncertainty in radio-interferometric imaging

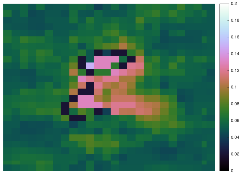
Astro-imaging experiment with redundant wavelet frame (Cai et al., 2017).



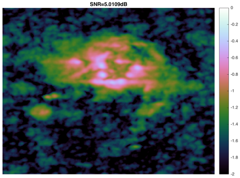
$\hat{x}_{MLE}(y)$



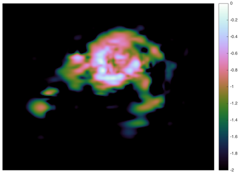
$\hat{x}_{MMSE} = \mathrm{E}(x|y)$



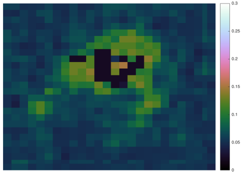
credible intervals (scale 10×10)



$\hat{x}_{MLE}(y)$



$\hat{x}_{MMSE} = \mathrm{E}(x|y)$



credible intervals (scale 10×10)

3C2888 and M31 radio galaxies (size 256×256 pixels). Computing time 1 minute.

$M = 10^5$ iterations. Estimation error w.r.t. MH implementation 3%.

Hypothesis testing for image structures

Bayesian hypothesis test for specific image structures (e.g., lesions)

$$\mathrm{H}_0 : \text{The structure of interest is ABSENT in the true image}$$
$$\mathrm{H}_1 : \text{The structure of interest is PRESENT in the true image}$$

The null hypothesis H_0 is rejected with significance α if

$$\mathrm{P}(\mathrm{H}_0|y) \leq \alpha.$$

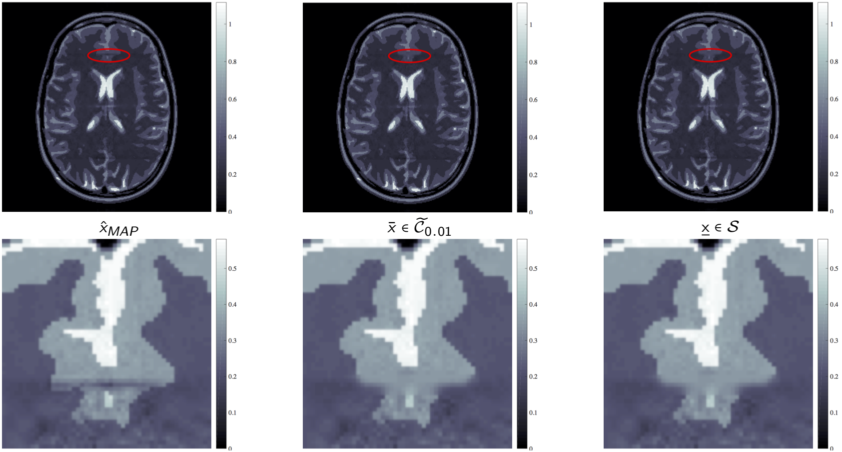
Key idea: (Repetti et al., 2018)
Let $\mathcal{S}$ denote the region of $\mathbb{R}^d$ associated with H_0, containing all images *without the structure* of interest. Then

$$\mathcal{S} \cap \mathcal{C}_\alpha = \varnothing \iff \mathrm{P}(H_0|y) \leq \alpha\,.$$

If in addition $\mathcal{S}$ is convex, then checking $\mathcal{S} \cap \widetilde{\mathcal{C}}_\alpha = \varnothing$ is a convex problem

$$\min_{\bar{x},\, \underline{x} \in \mathbb{R}^d} \|\bar{x} - \underline{x}\|_2^2 \quad \text{s.t.} \quad \bar{x} \in \mathcal{C}_\alpha\,, \quad \underline{x} \in \mathcal{S}\,.$$

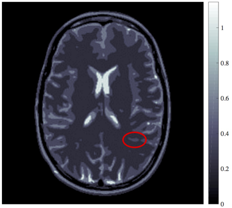
Uncertainty quantification in MRI imaging



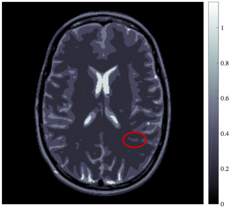
$\hat{x}_{MAP}$ $\quad$ $\bar{x} \in \widetilde{\mathcal{C}}_{0.01}$ $\quad$ $\underline{x} \in \mathcal{S}$

$\hat{x}_{MAP}$ (zoom) $\quad$ $\bar{x} \in \widetilde{\mathcal{C}}_{0.01}$ (zoom) $\quad$ $\underline{x} \in \mathcal{S}$ (zoom)

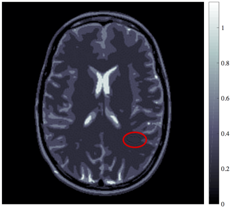
MRI experiment: test images $\bar{x} = \underline{x}$, hence we fail to reject H_0 and conclude that there is little evidence to support the observed structure.



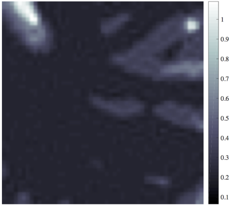
$\hat{x}_{MAP}$



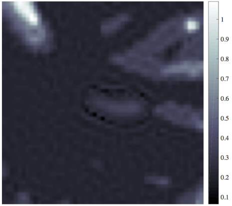
$\bar{x} \in \mathcal{C}_{0.01}$



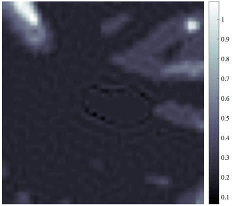
$\underline{x} \in \mathcal{S}_0$



$\hat{x}_{MAP}$ (zoom)



$\bar{x} \in \mathcal{C}_{0.01}$ (zoom)



$\underline{x} \in \mathcal{S}_0$ (zoom)

MRI experiment: test images $\bar{x} \neq \underline{x}$, hence we reject H_0 and conclude that there is significant evidence in favour of the observed structure.

Outline

1. Bayesian inference in imaging inverse problems
2. Proximal Markov chain Monte Carlo
3. Uncertainty quantification in astronomical and medical imaging
4. Empirical Bayes estimation with unknown regularisation parameters
5. Conclusion

Problem statement

Consider the class of Bayesian models

$$p(x|y,\theta) = \frac{p(y|x)p(x|\theta)}{p(y|\theta)},$$

parametrised by a regularisation parameter $\theta \in \Theta$. For example,

$$p(x|\theta) = \frac{1}{C(\theta)} \exp\{-\theta\varphi(x)\}, \quad p(y|x) \propto \exp\{-f_y(x)\},$$

with f_y and φ convex l.s.c. functions, and f_y L-Lipschitz differentiable.

We assume that $p(x|\theta)$ is proper, i.e.,

$$C(\theta) = \int_{\mathbb{R}^d} \exp\{-\theta\varphi(x)\} \mathrm{d}x < \infty,$$

with $C(\theta)$ unknown and generally intractable.

Regularisation parameter MLE

In this talk we adopt an empirical Bayes approach and consider the MLE

$$\begin{aligned}\hat{\theta} &= \operatorname*{argmax}_{\theta\in\Theta} p(y|\theta)\,,\\ &= \operatorname*{argmax}_{\theta\in\Theta} \int_{\mathbb{R}^d} p(y,x|\theta)\mathrm{d}x\,,\end{aligned}$$

which we solve efficiently by using a stochastic gradient algorithm driven by two proximal MCMC kernels (see Fernandez-Vidal and Pereyra (2018)).

Given $\hat{\theta}$, we then straightforwardly compute

$$\hat{x}_{MAP} = \operatorname*{argmin}_{x\in\mathbb{R}^d} f_y(x) + \hat{\theta}\varphi(x)\,. \tag{4}$$

Stochastic Approximation algorithm to compute $\hat{\theta}$

We use the following MCMC-driven stochastic gradient algorithm:
Initialisation $x^{(0)}, u^{(0)} \in \mathbb{R}^d$, $\theta^{(0)} \in \Theta$, $\delta_t = \delta_0 t^{-0.8}$.

for $t = 0$ to n

1. MCMC update $x^{(t+1)} \sim M_{x|y,\theta^{(t)}}(\cdot|x^{(t)})$ targeting $p(x|y,\theta^{(t)})$
2. MCMC update $u^{(t+1)} \sim K_{x|\theta^{(t)}}(\cdot|u^{(t)})$ targeting $p(x|\theta^{(t)})$
3. Stoch. grad. update

$$\theta^{(t+1)} = P_\Theta\left[\theta^{(t)} + \delta_t \varphi(u^{(t+1)}) - \delta_t \varphi(x^{(t+1)})\right].$$

end for

Output The iterates $\theta^{(t)} \to \hat{\theta}$ as $n \to \infty$.

SAPG algorithm driven MCMC kernels

Initialisation $x^{(0)}, u^{(0)} \in \mathbb{R}^d$, $\theta^{(0)} \in \Theta$, $\delta_t = \delta_0 t^{-0.8}$, $\lambda = 1/L$, $\gamma = 1/4L$.

for $t = 0$ to n

1. Coupled Proximal MCMC updates: generate $z^{(t+1)} \sim \mathcal{N}(0, \mathbb{I}_d)$

$$x^{(t+1)} = (1 - \frac{\gamma}{\lambda})x^{(t)} - \gamma \nabla f_y\left(x^{(t)}\right) + \frac{\gamma}{\lambda}\text{prox}_{\varphi}^{\theta\lambda}\left(x^{(t)}\right) + \sqrt{2\gamma} z^{(t+1)},$$
$$u^{(t+1)} = (1 - \frac{\gamma}{\lambda})u^{(t)} + \frac{\gamma}{\lambda}\text{prox}_{\varphi}^{\theta\lambda}\left(u^{(t)}\right) + \sqrt{2\gamma} z^{(t+1)},$$

2. Stochastic gradient update

$$\theta^{(t+1)} = P_\Theta\left[\theta^{(t)} + \delta_t \varphi(u^{(t+1)}) - \delta_t \varphi(x^{(t+1)})\right].$$

end for

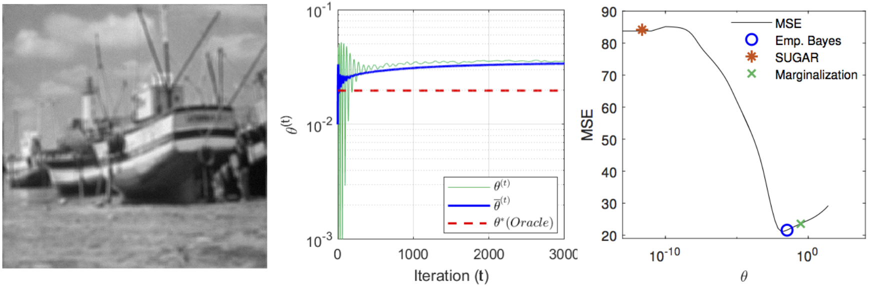
Output Averaged estimator $\bar{\theta} = n^{-1}\sum_{t=1}^{n} \theta^{(t+1)}$ converges approx. to $\hat{\theta}$.

Illustrative example - Image deblurring with TV-ℓ_2 prior

We consider the Bayesian image deblurring model

$$p(x|y,\theta) \propto \exp\left(-\|y - Ax\|^2/2\sigma^2 - \alpha\|x\|_2 - \theta\|\nabla_d x\|_{1-2}\right),$$

and compute $\hat{\theta} = \mathrm{argmax}_{\theta\in\mathbb{R}^+}\, p(y|\theta)$.



y Reg. param θ Estimation error for $\hat{x}_{MAP}$

Figure : Boat image deconvolution experiment.

Image deblurring with TV-ℓ_2 prior

(a) Original

(b) Degraded

(c) Emp. Bayes $\hat{x}_{MAP}$

Outline

Conclusion

- The challenges facing modern imaging sciences require a methodological paradigm shift to go beyond point estimation.
- Opportunity for advanced Bayesian inference methods to take central role and deliver impact.
- This requires significantly accelerating inference methods, e.g., by integrating modern stochastic and variational approaches at algorithmic, methodological, and theoretical levels.

Thank you!

Bibliography:

Ay, N. and Amari, S.-I. (2015). A novel approach to canonical divergences within information geometry. *Entropy*, 17(12):7866.

Cai, X., Pereyra, M., and McEwen, J. D. (2017). Uncertainty quantification for radio interferometric imaging II: MAP estimation. *ArXiv e-prints*.

Chambolle, A. and Pock, T. (2016). An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319.

Deledalle, C.-A., Vaiter, S., Fadili, J., and Peyré, G. (2014). Stein unbiased gradient estimator of the risk (sugar) for multiple parameter selection. *SIAM Journal on Imaging Sciences*, 7(4):2448–2487.

Durmus, A., Moulines, E., and Pereyra, M. (2018). Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM J. Imaging Sci.*, 11(1):473–506.

Fernandez-Vidal, A. and Pereyra, M. (2018). Maximum likelihood estimation of regularisation parameters. In *Proc. IEEE ICIP 2018*.

Green, P. J., Łatuszyński, K., Pereyra, M., and Robert, C. P. (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25(4):835–862.

Moreau, J.-J. (1962). Fonctions convexes duales et points proximaux dans un espace Hilbertien. *C. R. Acad. Sci. Paris Sér. A Math.*, 255:2897–2899.

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.

Pereyra, M. (2016). Maximum-a-posteriori estimation with bayesian confidence regions. *SIAM J. Imaging Sci.*, 6(3):1665–1688.

Pereyra, M. (2016). Revisiting maximum-a-posteriori estimation in log-concave models: from differential geometry to decision theory. *ArXiv e-prints*.

Pereyra, M., Bioucas-Dias, J., and Figueiredo, M. (2015). Maximum-a-posteriori estimation with unknown regularisation parameters. In *Proc. Europ. Signal Process. Conf. (EUSIPCO) 2015*.

Repetti, A., Pereyra, M., and Wiaux, Y. (2018). Scalable Bayesian uncertainty quantification in imaging inverse problems via convex optimisation. *ArXiv e-prints*.

Robert, C. P. (2001). *The Bayesian Choice (second edition)*. Springer Verlag, New-York.