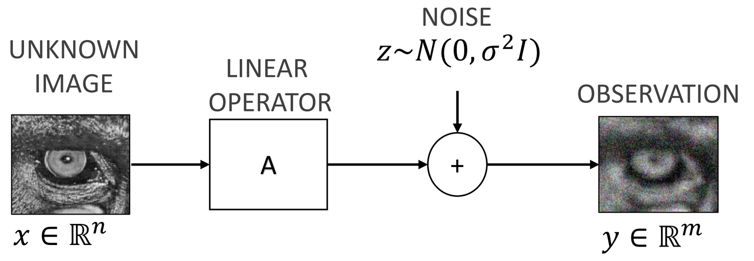
MAXIMUM LIKELIHOOD ESTIMATION OF REGULARISATION PARAMETERS

Ana Fernandez Vidal, Dr. Marcelo Pereyra (*af69@hw.ac.uk, m.pereyra@hw.ac.uk*)
School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh EH14 4AS United Kingdom

HERIOT WATT UNIVERSITY

Imaging Inverse Problems

OBJECTIVE: to estimate an unknown image x from an observation y.



Some canonical examples:
- image deconvolution
- compressive sensing
- super-resolution
- tomographic reconstruction
- inpainting

High-dimensional problems → $n \sim 10^6$

CHALLENGE: not enough information in y to accurately estimate x.

For example, in many imaging problems $y = Ax + z$ where the operator A is either:

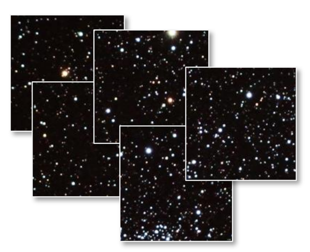
A → RANK-DEFICIENT ⟶ not enough equations to solve for x
A → SMALL EIGENVALUES ⟶ NOISE AMPLIFICATION ⟶ unreliable estimates of x

REGULARISATION: We can render the problem well-posed by using **prior knowledge** about the unknown signal x.

The Bayesian Framework

PRIOR INFORMATION

We model the unknown image x as a **random vector** with **prior distribution** $p(x|\theta)$ promoting desired properties about x.



Example: random vector $\boldsymbol{x}$ representing astronomical images.

$$p(x|\theta) = \frac{e^{-\theta\,\varphi(x)}}{C(\theta)}$$

HYPERPARAMETER
NORMALIZING CONSTANT

$\varphi(x)$ penalises undesired properties and the **regularisation parameter** θ controls the intensity of the penalisation.

OBSERVED DATA

The observation y is related to x by a statistical model:

$$y \sim N(Ax, \sigma^2 I) \qquad p(y|x) \propto e^{-g_y(x)} \qquad g_y(x) = \frac{\|y-Ax\|_2^2}{2\sigma^2}$$

LIKELIHOOD

POSTERIOR DISTRIBUTION

Observed and prior information are combined by using Bayes' theorem:

$$p(x|y,\theta) = p(y|x)\, p(x|\theta)/p(y|\theta)$$

$$p(x|y,\theta) \propto e^{-\left(g_y(x) + \theta\,\varphi(x)\right)}$$

The predominant Bayesian approach in imaging is **MAP estimation** which can be computed very efficiently by convex optimisation:

$$\hat{x}_{MAP} = \underset{x\in\mathbb{R}^n}{\operatorname{argmax}}\, p(x|y,\theta) = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}}\; g_y(x) + \theta\,\varphi(x)$$

We will focus on **convex problems** where:
- $\varphi(x)$ is lower semicontinuous, proper and possibly non-smooth
- $g_y(x)$ is Lipschitz differentiable with Lipschitz constant L

How do we choose the regularisation parameter θ?

The regularisation parameter controls how much importance we give to prior knowledge and to the observations, depending on how ill posed the problem is and on the intensity of the noise.

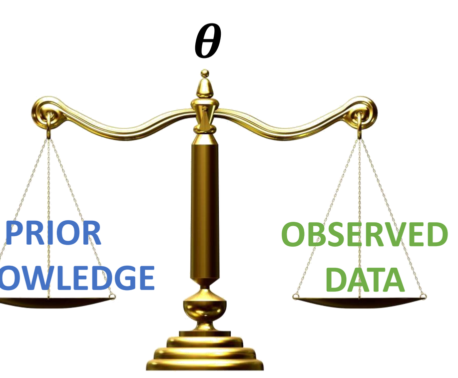
POSSIBLE APPROACHES FOR CHOOSING θ:

NON BAYESIAN

- Cross-validation → Exhaustive search method
- Discrepancy Principle
- Stein-based methods →Minimise Stein's Unbiased Risk Estimator (SURE), a surrogate of the MSE
  - SUGAR → More efficient algorithms that uses gradient of SURE
    Limitation: mostly designed for denoising problems. Difficult to implement

BAYESIAN

- Hierarchical → Propose prior for θ and work with hierarchical model
  - Marginalisation → Remove θ from the model $p(x|y) = \int_{\mathbb{R}^+} p(x,\theta|y)d\theta$
    Limitation: only for homogeneous $\varphi(x)$ or cases with known $C(\theta)$
- Empirical Bayes → Choose θ by maximising marginal likelihood $p(y|\theta)$
  Difficulty: $p(y|\theta)$ becomes intractable in high-dimensional problems



Our strategy: Empirical Bayes

We want to find $\hat{\theta}$ that maximises the marginal likelihood $p(y|\theta)$:

$$\begin{cases} p(y|\theta) = \int_{\mathbb{R}^n} p(y,x|\theta)dx \\ \hat{\theta}_{MLE} = \underset{\theta\in\Theta}{argmax}\; p(y|\theta) \end{cases}$$

We should be able to reliably estimate θ from y as y is very high-dimensional and θ is a scalar parameter

The **challenge** is that $p(y|\theta)$ is **intractable** as it involves solving two integrals in $\mathbb{R}^n$ (to marginalise x and to compute $C(\theta)$). This makes the computation of $\hat{\theta}_{MLE}$ extremely difficult.

OUR CONTRIBUTION
We propose a stochastic optimisation scheme to compute the maximum marginal likelihood estimator of the regularisation parameter. Novelty: the optimisation is driven by proximal Markov chain Monte Carlo (MCMC) samplers.

Proposed stochastic optimisation algorithm

If $p(y|\theta)$ was tractable, we could use a standard projected gradient algorithm:

$$\theta^{t+1} = \underset{\Theta}{Proy}\left[\theta^t + \delta_t \frac{d}{d\theta}\log p(y|\theta^t)\right] \quad \text{where } \delta_t \text{ verifies} \quad \lim_{t\to\infty}\delta_t = 0,\ \sum_{t=0}^{\infty}\delta_t = \infty,\ \sum_{t=0}^{\infty}\delta_t^2 < \infty$$

To tackle the intractability, we propose a stochastic variant of this algorithm based on a noisy estimate of $\frac{d}{d\theta}\,log(p(y|\theta^t))$. Using Fisher's identity we have:

$$\frac{d}{d\theta}\log p(\boldsymbol{y}|\theta) = E_{x|y,\theta}\left\{\frac{d}{d\theta}\log p(\boldsymbol{x},\boldsymbol{y}|\theta)\right\} = E_{x|y,\theta}\{-\varphi(x)\} - \frac{d}{d\theta}Log(C(\theta))$$

If $C(\theta)$ is unknown we can use the identity $-\frac{d}{d\theta}Log(C(\theta)) = E_{x|\theta}\{\varphi(\boldsymbol{x})\}$

The intractable gradient becomes $\frac{d}{d\theta}\log p(\boldsymbol{y}|\theta) = E_{x|y,\theta}\{-\varphi(x)\} + E_{x|\theta}\{\varphi(\boldsymbol{x})\}$

Now we can approximate $E_{x|y,\theta}\{-\varphi(x)\}$ and $E_{x|\theta}\{\varphi(\boldsymbol{x})\}$ with **Monte Carlo estimates.**

We construct a Stochastic Approx. Proximal Gradient (SAPG) algorithm driven by two Markov kernels M_θ and K_θ targeting the posterior $p(\boldsymbol{x},\boldsymbol{y}|\theta)$ and the prior $p(x|\theta)$ respectively.

STOCHASTIC APPROXIMATION PROXIMAL GRADIENT (SAPG) ALGORITHM

$$\begin{cases} X^{t+1} \sim \mathrm{M}_{\theta^t}(\mathrm{X}|y,\theta^t,X^t) \longrightarrow p(x|y,\hat{\theta}_{MLE}) \longrightarrow \text{EMPIRICAL BAYES POSTERIOR} \\ U^{t+1} \sim \mathrm{K}_{\theta^t}(\mathrm{U}|\theta^t,U^t) \longrightarrow p(x|\hat{\theta}_{MLE}) \\ \theta^{t+1} = \underset{\Theta}{Proj}\left[\theta^t + \delta_t\big(\underbrace{\varphi(U^{t+1}) - \varphi(X^{t+1})}_{\text{STOCHASTIC GRADIENT}}\big)\right] \to \hat{\theta}_{MLE} \end{cases}$$

CONVERGE JOINTLY TO

How do we generate the samples?

We use the MYULA algorithm for the Markov kernels K_θ and M_θ because they can handle:
- High-dimensionality !
- Convex problems with a non-smooth $\varphi(\boldsymbol{x})$

WHERE DOES MYULA COME FROM?

LANGEVIN DIFFUSION

$$dX(t) = \frac{1}{2}\nabla\log p(X(t)|y)dt + dW(t)$$

BROWNIAN MOTION

EULER MARUYAMA APPROXIMATION

$$\frac{dX(t)}{dt} \to \frac{x^{t+1}-x^t}{\gamma}$$

UNADJUSTED LANGEVIN ALGORITHM (ULA)

$$x^{t+1} = x^t + \gamma\,\nabla\,Log\big(p(x^t|y,\theta)\big) + \sqrt{2\gamma}\,Z^{t+1}$$

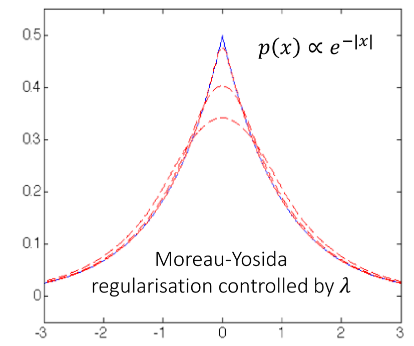
$$Z^{t+1} \sim N(0,\mathbb{I}) \qquad \propto \exp\{-g_y(x) - \theta\,\varphi(x)\}$$

MOREAU-YOSIDA UNADJUSTED LANGEVIN ALGORITHM (MYULA)

If $\varphi(x)$ is not Lipschitz differentiable ULA is unstable.

The MYULA algorithm uses Moreau-Yosida regularization to replace the non-smooth term $\varphi(x)$ with its Moreau Envelope $\tilde{\varphi}_M(x)$.

$$x^{t+1} = x^t - \gamma\Big(\nabla g_y(x^t) + \underbrace{\frac{x^t - prox_{\varphi}^{\lambda\theta}(x^t)}{\lambda}}_{\nabla\tilde{\varphi}_{MY}(x)}\Big) + \sqrt{2\gamma}\,Z^{t+1}$$



- The MYULA kernels do not target $p(x|y,\theta)$ and $p(x|\theta)$ exactly.
- Sources of **asymptotic bias:**
  - Discretisation of Langevin diffusion: controlled by γ and η.
  - Smoothing of non differentiable $\varphi(x)$: controlled by λ.
- $\boldsymbol{\gamma,\eta}$ must be **< inverse of Lipchitz constant** of the gradient driving each diffusion respectively.
- More information about how to select each parameter can be found in [2].

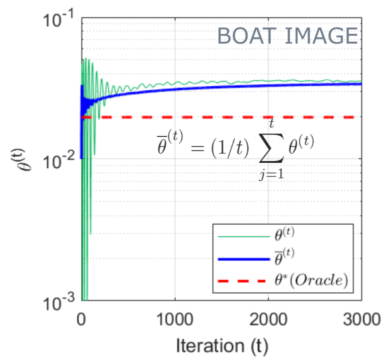
Results

We illustrate the proposed methodology with an **image deblurring** problem using a **total-variation prior**.

We compare our results with those obtained with:
- SUGAR
- Marginal maximum-a-posteriori estimation
- Optimal or oracle value θ^*

EXPERIMENT DETAILS
- Recover $\boldsymbol{x}$ from a blurred noisy observation $\boldsymbol{y}$, $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{z}$ and $\boldsymbol{z} \sim \mathrm{N}(0,\sigma^2 I)$
- A is a circulant uniform blurring matrix of size 9x9 pixels
- $\varphi(\boldsymbol{x}) = TV(\boldsymbol{x}) \to$ isotropic total-variation pseudo-norm
- We use 6 test images of size 512x512 pixels
- We compute the MAP estimator for each using different values of θ obtained with different methods
- We compare methods by taking average value over 6 images of the mean squared error (MSE) and the computing time (in minutes)

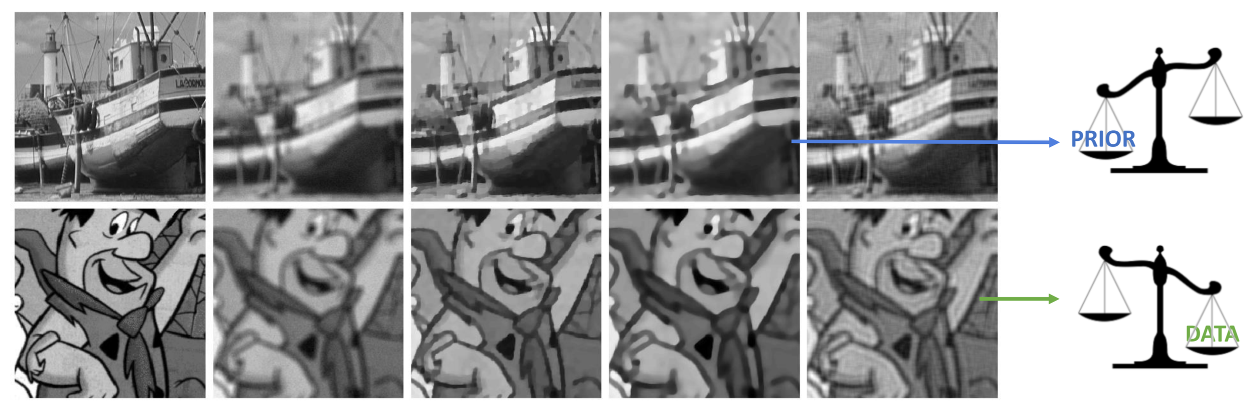
EVOLUTION OF θ THROUGHOUT ITERATIONS



DEGRADED | EMPIRICAL BAYES MAP ESTIMATOR | ORACLE | ORIGINAL

METHOD COMPARISON

Close-up for SNR=20dB



ORIGINAL | DEGRADED | EMP. BAYES | MARGINAL MAP | SUGAR

Over-smoothing: too much regularisation

Ringing: not enough regularisation

QUANTITATIVE COMPARISON

For each image, noise level, and method, we compute the MAP estimator and we display on this table the average results for the 6 test images.

Method	SNR=20 dB		SNR=30 dB		SNR=40 dB	
	Avg. MSE	Avg. Time	Avg. MSE	Avg. Time	Avg. MSE	Avg. Time
$\theta^*(Oracle)$	22.95 ± 3.10	–	21.05 ± 3.19	–	18.76 ± 3.19	–
SUGAR	24.14 ± 3.19	15.74	23.96 ± 3.26	20.87	23.94 ± 3.27	20.59
Marginalization	24.67 ± 3.08	17.27	22.39 ± 3.07	6.31	19.44 ± 3.26	6.77
Empirical Bayes	23.24 ± 3.23	43.01	21.16 ± 3.24	41.50	18.90 ± 3.39	42.85

in **minutes**

EB performs remarkably well at low SNR values

SUGAR performs poorly when the main problem is not noise

Longer computing time only justified if marginalisation can't be used

✓ Generally outperforms the other approaches
✓ Achieves close-to-optimal performance
× Increased computing times

Conclusions

- We presented an **empirical Bayesian** method to **estimate regularisation parameters** in convex inverse imaging problems.
- We approach an intractable maximum marginal likelihood estimation problem by proposing a stochastic optimisation algorithm.
- The stochastic approximation is driven by two proximal MCMC kernels which can handle non-smooth regularisers efficiently.
- Our algorithm was illustrated with non-blind image deconvolution with TV prior where it:
  - ✓ achieved close-to-optimal performance
  - ✓ outperformed other state of the art approaches in terms of MSE
  - × had longer computing times.
- More details can be found in [3].

Future work

- Detailed analysis of the convergence properties.
- Extending the methodology to problems involving multiple unknown regularisation parameters.
- Reducing computing times by accelerating the Markov kernels driving the stochastic approximation algorithm:
  - Via parallel computing
  - By choosing a faster Markov kernel
- Adapt algorithm to support some non-convex problems.
- Use the samples from $\boldsymbol{p(x|y,\hat{\theta}_{MLE})}$ to perform uncertainty quantification.

References

[1] Marcelo Pereyra, José M Bioucas-Dias, and Mário AT Figueiredo, "Maximum-a-posteriori estimation with unknown regularisation parameters," in *23rd European Signal Processing Conference (EUSIPCO), 2015*. IEEE, pp. 230–234, 2015.

[2] Alain Durmus, Eric Moulines, and Marcelo Pereyra, "Efficient Bayesian computation by proximal Markov Chain Monte Carlo: when Langevin meets Moreau," *SIAM Journal on Imaging Sciences*, vol. 11, no. 1, pp. 473–506, 2018.

[3] Ana Fernandez Vidal and Marcelo Pereyra. "Maximum Likelihood Estimation of Regularisation Parameters." In 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 1742-1746. IEEE, 2018.