Stochastic Models for Image Analysis: Project

Due on April 3, 2020

Tamim EL AHMAD

Part 1

This part is a summary of the article "Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical Bayesian approach"[2]. This paper presents a way to set adequately regularisation parameters on imaging inverse problems using an empirical Bayesian approach. Many solutions only allow to solve this problem for a single scalar regularisation value or suffer from imaging problem that are generally ill-conditioned or ill-posed. In this paper, many regularisation parameters can be set simultaneously from the observed data using maximum marginal likelihood estimation.

Problem statement

Let $d, d_y, d_\Theta \in \mathbb{N}$ and let $\Theta \subset (0, +\infty)^{d_\Theta}$ be a convex compact set. The authors consider the estimation of an unknown image $x \in \mathbb{R}^d$ from an observation $y \in \mathbb{C}^d$ related to x by a statistical model with likelihood function

$$p(y|x) \propto e^{-f_y(x)}$$

where f_y is convex and continuously differentiable with L_y-Lipschitz gradient. The authors adopt a Bayesian approach and seek to use prior knowledge about x to regularise the estimation problem and improve results. They consider prior distributions given for any $x \in \mathbb{R}^d$ and $\theta \in \Theta$ by

$$p(x|\theta) = e^{-\theta^T g(x)}/Z(\theta)$$

for some convex and Lipschitz continuous vector of statistics $g : \mathbb{R}^d \mapsto \mathbb{R}^{d_\Theta}$ and where the normalising constant of the prior distribution $p(x|\theta)$ is given by $Z(\theta) = \int_{\mathbb{R}^d} e^{-\theta^T g(\tilde{x})} d\tilde{x}$. Here, θ controls the amount of regularisation. Then using the Bayes rule, it is possible to retrieve the posterior distribution, which underpins all inferences about the image x given observed data y:

$$p(x|y,\theta) = exp[-f_y(x) - \theta^T g(x)] / \int_{\mathbb{R}^d} exp[-f_y(\tilde{x}) - \theta^T g(\tilde{x})] d\tilde{x}$$

Finally, using the maximum-a-posteriori (MAP) estimator, given for any $\theta \in \Theta$ by

$$\hat{x}_{\theta,MAP} = \underset{x \in \mathbb{R}^d}{argmin} \{f_y(x) + \theta^T g(x)\}$$

The regularization parameter θ controls the balance between observed and prior information, so it is needed to control this parameter as it has a huge impact on the output image of the algorithm.

Proposed Empirical Bayes methodology

Empirical Bayes estimation

Under an empirical Bayesian paradigm, we estimates the regularization parameter $\theta \in \Theta$ from the observed data y. One of the ways of doing it and the one the authors chose is maximum marginal likelihood estimation:

$$\theta_\star \in \arg\max_{\theta \in \Theta} p(y|\theta) \tag{1}$$

Given $\theta_\star$, we can retrieve x by MAP estimation as described before thanks to the posterior distribution $p(x|y, \theta_\star)$.
This strategy is efficient in statistical problems but in imaging problems, it might be computationally heavy since the marginal likelihood $p(y|\theta)$ is computationally intractable as it is needed to compute two d-dimensional integrals which makes the optimization problem intractable.
Moreover, the solution of (1) might not be unique but in the experiments performed by the authors, they always observed a uniqueness of $\theta_\star$, it might be possible for imaging problem because of large dimensions but it still has to be proved.
In order to tackle the computational issue, the paper use a stochastic gradient Markov Chain Monte Carlo (MCMC) algorithm.

Problem 1 continued on next page...

Stochastic gradient MCMC algorithm

This part present the proposed empirical Bayesian method to solve the marginal maximum likelihood estimation problem (1) and set regularisation parameters. Since $p(y|\theta)$ is computationally intractable, its gradient is intractable, and it is not possible to a projected ascent gradient algorithm in order to solve (1) and find $\theta_\star$. Thus, to avoid the computation of the gradient, the authors replace it by a noisy estimate. Under mild assumptions using Fisher's identity and the fact that for any $x \in \mathbb{R}^d$, $y \in \mathbb{R}^{d_y}$ and $\theta \in \Theta$, $p(x, y|\theta) = p(y|x)p(x|\theta)$, we have for any $\theta \in \Theta$

$$\nabla_\theta \log p(y|\theta) = \int_{\mathbb{R}^d} p(\tilde{x}|y,\theta)\nabla_\theta \log p(\tilde{x}, y|\theta)\mathrm{d}\tilde{x} = -\int_{\mathbb{R}^d} g(\tilde{x})p(\tilde{x}|y,\theta)\mathrm{d}\tilde{x} - \nabla_\theta \log(\mathrm{Z}(\theta))$$

Hence, a Monte Carlo estimator can be used to estimate the expectation $\int_{\mathbb{R}^d} g(\tilde{x})p(\tilde{x}|y,\theta)\mathrm{d}\tilde{x}$, and then consider a stochastic approximation proximal gradient algorithm (SAPG) to obtain the following gradient, where $(X_k)_{k\in\{0,\dots,m\}}$ is a sample of size m:

$$\Delta_{m,\theta} = \frac{1}{m}\sum_{k=1}^{m} \nabla_\theta \log p\left(X_k, y|\theta\right) = -\frac{1}{m}\sum_{k=1}^{m} g\left(X_k\right) - \nabla_\theta \log \mathbb{Z}(\theta)$$

Then, we can build a new sequence $(\theta_n)_{n\in\mathbb{N}}$ with θ_0 and associated with the following recursion for any $n \in \mathbb{N}$

$$\theta_{n+1} = \Pi_\Theta\left[\theta_n + \delta_{n+1}\Delta_{m_n,\theta_n}\right], \quad \Delta_{m_n,\theta_n} = -\frac{1}{m_n}\sum_{k=1}^{m_n} g\left(X_k^n\right) - \nabla_\theta \log Z\left(\theta_n\right)$$

Then we estimates our regularization parameter as follow by averaging the previous values of θ_n found:

$$\bar{\theta}_N = \frac{1}{N}\sum_{n=0}^{N-1} \theta_n$$

which converges to a solution of (1).
For $Z(\theta)$, since for any $\theta \in \Theta$:

$$\bar{g}_\theta = \int_{\mathbb{R}^d} g(\tilde{x})p(\tilde{x}|\theta)\mathrm{d}\tilde{x} = -\nabla_\theta \log \mathrm{Z}(\theta)$$

we can compute it in different ways depending on g or estimate the expectation corresponding thanks to another Monte Carlo estimator as well.

MCMC Kernels

In this part, the authors explain how to choose the families of Markov kernels for the SAPG algorithm given a high dimensional problem .In fact, $(X_k)_{k\in\{0,\dots,n\}}$ is sampled from a family of Markov kernels $\{R_{\gamma,\theta} : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$. To do so, they use the state of the art proximal Markov chain Monte Carlo designed for high dimensional inverse problems that are convex but not smooth (MYULA). It is an improvement of ULA which is not adapted for not smooth functions g. Given $X_0 \in \mathbb{R}^d, V = \theta^\top g$ and $U = f_y$, the non-differentiable part V (due to g) is replaced by a smooth approximation $V^\lambda(x)$ given by the Moreau-Yosida envelope of V, defined for any $x \in \mathbb{R}^d$ and $\lambda > 0$ by

$$V^\lambda(x) = \min_{\tilde{x}\in\mathbb{R}^d} \left\{V(\tilde{x}) + (1/2\lambda)\|x - \tilde{x}\|_2^2\right\}$$

For any $\lambda > 0$, the Moreau-Yosida envelope V^λ is continuously differentiable with gradient given for $x \in \mathbb{R}^d$ by

$$\nabla V^\lambda(x) = \left(x - \mathrm{prox}_V^\lambda(x)\right)/\lambda$$

Finally, we have the following recursion:

$$\text{MYULA: } X_{k+1} = X_k - \gamma\nabla_x U\left(X_k\right) - \gamma\nabla_x V^\lambda\left(X_k\right) + \sqrt{2\gamma}Z_{k+1}$$

Connections to hierarchical Bayesian approaches

Hierarchical and empirical approaches could yield similar results following this equation: $p(x|y) = \int_\Theta p(x|y,\tilde{\theta})p(\tilde{\theta}|y)\mathrm{d}\bar{\theta}$. In imaging problems, the marginal posterior $p(\theta|y) \propto p(y|\theta)p(\theta)$ will be dominated by the marginal likelihood $p(y|\theta)$ because of the dimensionality of y . Hence, both the hierarchical and the empirical approaches should output similar results. For models that are correctly specified, both strategies should perform well, and hierarchical Bayes should moderately outperform empirical Bayes. But for imaging models, the problem is not well specified and the empirical Bayesian approach should outperform the hierarchical one, that's why it is used in this paper.

Numerical experiments

Many imaging inverse problems involving different transformation of the true image, representation of it and regularisation functions (different f_ys and gs) have been studied in this paper.

- First, to study the performance of the model, synthetic images are used. Since the generative model is set by the authors, it is well-specified and the regularisation parameter has a true value, so it allows to evaluate the performance of the model. The experiment show that the model converges to the true value of the regularization parameter and that the error is close to a Gaussian.
- Second , the algorithm is applied to image deblurring using two kinds of prior distributions. Indeed, the objective is to estimate a scalar-valued regularisation parameter in a non-blind image deconvolution model with different kinds of prior distributions:
  - Deconvolution with total variation prior: very good results are obtained, close-to-optimal both for high and low SNR values. Compared to the state-of-the-art approaches, the empirical Bayes generally performs better with very competitive times (same order of magnitude as state of the art algorithms, as it is tricky to compare computation times between different algorithms and models).
  - Wavelet deconvolution with synthesis prior: good results are obtained, the empirical Bayesian method converges close to the true regularization parameter for all SNR values, it also outperforms all the algorithms for all SNR values. In particular, for high SNR values, both Bayesian methods attain similar values of MSE, but the proposed empirical Bayes methodology is five times faster.
- Third, the model is applied to a sparse hyperspectral unmixing problem combining an l1 and a total variation regularisation. The empirical Bayesian method gives good results for all SNR values, and particularly outperforms the hierarchical Bayesian method for low SNR values.
- Fourth, the algorithm is applied to a total generalised variation denoising model that has two unknown regularisation parameters which are strongly dependent. It seems that the algorithm is generally robust to different initialisations and quickly converges. However, it is not completely robust to bad initialisation because of the non-convexity and the approximations involved.

Conclusion

This paper considered the automatic selection of regularisation parameters in imaging inverse problems, with a particular focus on problems that are convex w.r.t. the unknown image and possibly non-smooth, and which would be typically solved by maximum-a-posteriori estimation by using modern proximal optimisation techniques.
Because the proposed method uses the same basic operators as proximal optimisation algorithms, namely gradient and proximal operators, it is straightforward to apply to problems that are currently solved by proximal optimisation. In addition to being highly computational efficient and having strong theoretical underpinning, the proposed methodology is very general and can be used to simultaneously estimate multiple regularisation parameters, unlike some

alternative approaches from the literature that can only handle a single or scalar parameter.
Finally, the methodology is tested with a range of imaging problems and models, in which it obtains very good results in terms of MSE and outperformed different methods of the literature.

Part 2

Connections with concepts seen in class

This part presents concepts seen in class that are present in this paper.
First, seeing analysing images as an inverse probabilistic problem where the observed image y is a transformation of the real image x following a certain distribution knowing this image $p(y|x)$ and with a prior distribution on x depending on a parameter θ (which might be multidimensional) $p(x|\theta)$, and then trying to solve thanks to the Bayesian inference (MAP estimate) is an approach widely presented in class, particularly in the first class of the semester with Mrs. Delon, and in the last class of Mrs. Desolneux too. In particular, in the first class, we have seen how to solve a minimisation problem with a TV regularisation, introducing the proximal function, and the Chambolle-Pock[1] algorithm, used in this paper to compute the proximal function of TV function and an application in denoising, useful for the computation of the proximal function of the TV function, and deblurring, as some experiments of the paper.
Moreover, we have seen in the second class sampling techniques, using MCMC algorithm, such as Metropolos-Hasting but even Langevin Dynamics, ULA algorithm and MYULA algoithm, as used in the method presented in the paper. And of course the Moreau-Yosida regularisation used for MYULA algorithm.
Finally, the computation of this equation $\bar{g}_\theta = \int_{\mathbb{R}^d} g(\tilde{x})p(\tilde{x}|\theta)\mathrm{d}\tilde{x} = -\nabla_\theta \log \mathrm{Z}(\theta)$ reminded me of the maximum entropy principle and exponential models in the second class with Mr. Galerne, and the computation of the gradient of $log(p_\theta(x_0))$ and the inversion between gradient and integral when computing the gradient of the partition function $Z(\theta)$. Moreover, the Langevin dynamics have been approached in this course too.

Part 3

Theoretical development

I first want to give a proof of the following equation, for any $\theta \in \Theta$:

$$\bar{g}_\theta = \int_{\mathbb{R}^d} g(\tilde{x})p(\tilde{x}|\theta)\mathrm{d}\tilde{x} = -\nabla_\theta \log \mathrm{Z}(\theta) \tag{2}$$

Let us first assume that $\nabla_\theta Z(\theta) = -\int_{\mathbb{R}^d} g(x)e^{-\theta^T g(x)}dx$. So :

$$-\nabla_\theta \log \mathrm{Z}(\theta) = -\frac{\nabla_\theta Z(\theta)}{Z(\theta)} = \int_{\mathbb{R}^d} g(\tilde{x})\frac{e^{-\theta^T g(\tilde{x})}}{Z(\theta)}\tilde{x} = \int_{\mathbb{R}^d} g(\tilde{x})p(\tilde{x}|\theta)\mathrm{d}\tilde{x} = \bar{g}_\theta$$

Now, let us jump into the inversion between integral and gradient. Let us denote $G : (x, \theta) \in \mathbb{R}^d \times \mathbb{R}^{d_\Theta} \mapsto e^{-\theta^T g(x)}$. Since g is continuous, G is obviously continuous in x and in θ. G is differentible in θ and its gradient $\nabla_\theta G(x, \theta) = g(x)e^{-\theta^T g(x)}$ along θ is continuous in x and θ.
I first wanted to use the Dominated Convergence Theorem as in class, on a compact set as a closed ball $B(\bar{\theta_0}, 1)$ with $\theta_0 \in \Theta$, but it does not work here as we do not have the assumption, for all $\theta \in \Theta$:

$$\int_{\mathbb{R}^d} e^{\|\theta\|\|g(x)\|}dx < +\infty$$

Moreover, this assumption allowed us to justify that $Z(\theta)$ is well defined for all $\theta \in \Theta$, as it makes $x \mapsto e^{-\theta^T g(x)}$ integrable, but here, with $g = \|.\|_1$ for instance, it cannot be true. So let us consider that Z is well-defined on Θ and use the Leibniz rule as in the paper, between $a, b \in \mathbb{R}^d$. So we have:

$$\nabla_\theta Z(\theta) = -\int_a^b g(x) e^{-\theta^T g(x)} dx$$

Now let us assume that g is positive for the following. We want to tend every coordinate of b to $+\infty$, and every coordinate of a to $-\infty$.
First, we assume that $\|g\|_\infty < \infty$. So, for every x, $|g(x)e^{-\theta^T g(x)}| \leq \|g\|_\infty e^{-\theta^T g(x)}$ and we know that Z is well-defined so $\int_{\mathbb{R}^d} e^{-\theta^T g(x)} dx < \infty$.
If g is not bounded, so norm of $g(x)$ might tend to infinity when $\|x\| \longrightarrow \infty$. But, considering $\epsilon \in (0, 1)$ such that $\epsilon\theta \in \Theta$:

$$g(x)e^{-\theta^T g(x)} = o(e^{-\epsilon\theta^T g(x)}) \tag{3}$$

and $\int_{\mathbb{R}^d} e^{-\epsilon\theta^T g(x)} dx = Z(\epsilon\theta)$ is well-defined.
So finally, we have that:

$$\nabla_\theta Z(\theta) = -\int_{\mathbb{R}^d} g(x) e^{-\theta^T g(x)} dx$$

I also want to clarify this equation:

$$p(x|y, \theta) = exp[-f_y(x) - \theta^T g(x)] / \int_{\mathbb{R}^d} exp[-f_y(\tilde{x}) - \theta^T g(\tilde{x})] d\tilde{x} \tag{4}$$

We have, using Bayes rule, that:

$$p(x|y, \theta) = \frac{p(x, y|\theta)}{p(y|\theta)} = \frac{p(y|x, \theta)p(x|\theta)}{p(y|\theta)} = \frac{p(y|x)p(x|\theta)}{p(y|\theta)} = exp[-f_y(x) - \theta^T g(x)] / \int_{\mathbb{R}^d} exp[-f_y(\tilde{x}) - \theta^T g(\tilde{x})] d\tilde{x}$$

Since all the information of θ on y is contained in x, $p(y|x, \theta) = p(y|x)$.

Part 4

Implementation

For this part, I implemented the experiment presented in Section 5.2.1 of the paper, using the Algorithm 1 of the SAPG to compute the θ optimal thanks to the method presented in the algorithm in a problem of deblurring, using $f_y(x) = \|y - Ax\|_2^2 / 2\sigma^2$ and a TV-L2 regularisation function $g(x) = TV(x)$. So, it gave, for the updates in MYULA algorithm:

$$\nabla_x f_y = \frac{1}{\sigma^2} A^T (Ax - y)$$

$$prox_{\theta^T g}^{\lambda} = \underset{\tilde{x} \in \mathbb{R}^d}{argmin} \{\theta TV(x) + \frac{1}{2\lambda} \|\tilde{x} - x\|_2^2\}$$

So, as said in the paper, I used the Chambolle-Pock[1] algorithm to compute this proximal function with 25 iterations, I used the one we used for the denoising in TP1 with $\lambda\theta$ as regularisation parameter (we considered the problem $\underset{\tilde{x} \in \mathbb{R}^d}{argmin} \{\lambda TV(x) + \frac{1}{2} \|\tilde{x} - x\|_2^2\}$ during the TP). In order to retrieve the true image at the end, I used the Chambolle-Pock algorithm with 200 iterations for deblurring of TP1 as we want to find:

$$\hat{x}_{\theta^\star, MAP} \in \underset{x \in \mathbb{R}^d}{argmin} \{f_y(\hat{x}) + \theta^T g(\hat{x})\}$$

$f_y(\hat{x}) + \theta^T g(\hat{x}) = \|y - Ax\|_2^2/2\sigma^2 + \theta TV(x)$ so the previous problem is equivalent to:

$$\hat{x}_{\theta^\star, MAP} \in \underset{x \in \mathbb{R}^d}{argmin}\{\|y - Ax\|_2^2/2 + \sigma^2 \theta TV(x)\}$$

Thus, I used the function of TP1 with $\sigma^2\theta$ as regularisation parameter.
To represent the matrix A, I used the convolution with the same kernel h as in TP1.

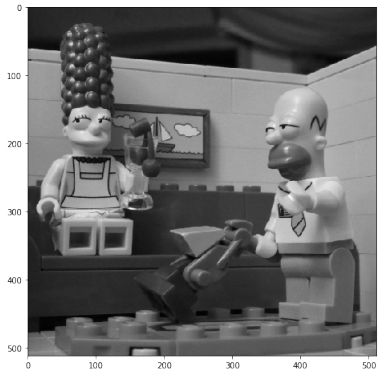

Figure 1: Image used x

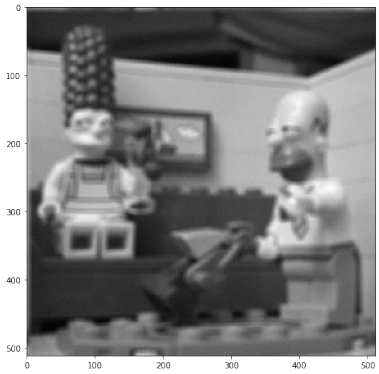

Figure 2: Image blurred y

Let us jump into the setting of parameter. I performed my experiments with $\sigma = 0.01$ and followed the indications of the paper for the rest of parameters:

$$\begin{aligned} X_0 &= y \\ \theta_0 &= 0.01 \\ L &= (0.99/\sigma)^2 \\ \lambda &= min(5L^{-1}, 2) \\ \gamma &= 0.98/(L + \lambda^{-1}) \\ \delta_n &= 0.1 \times n^{-0.8}/d \end{aligned}$$

I first tested the sampler as advised by the article, and my sampler is indeed stable for $\theta = 0.01, 0.1$, but surprisingly enough, it does not present any oscillations:

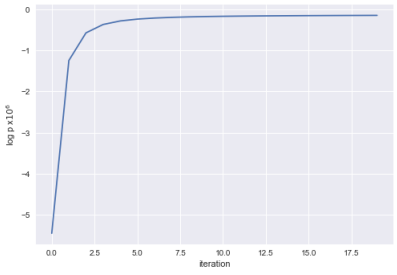

Figure 3: Evolution of $(\log p\,(X_n|y,\theta))_{n\in\mathbb{N}}$ with $(X_n)_{n\in\mathbb{N}}$ sampled using MYULA and targeting $p(\cdot|y,\theta)$ with $\theta = 0.01$ and the first 20 iterations.

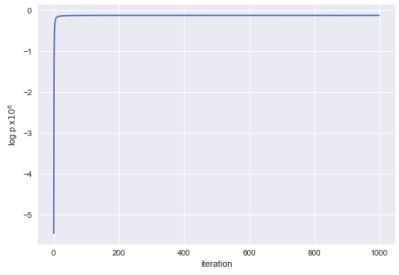

Figure 4: Evolution of $(\log p\,(X_n|y,\theta))_{n\in\mathbb{N}}$ with $(X_n)_{n\in\mathbb{N}}$ sampled using MYULA and targeting $p(\cdot|y,\theta)$ with $\theta = 0.01$ and the first 1000 iterations.

Problem 4 continued on next page...

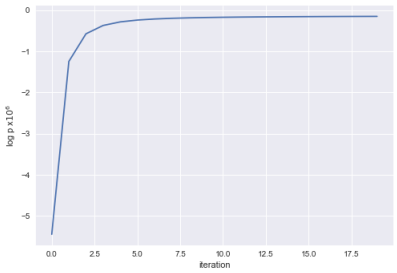

Figure 5: Evolution of $(\log p(X_n|y,\theta))_{n\in\mathbb{N}}$ with $(X_n)_{n\in\mathbb{N}}$ sampled using MYULA and targeting $p(\cdot|y,\theta)$ with $\theta = 0.001$ and the first 20 iterations.

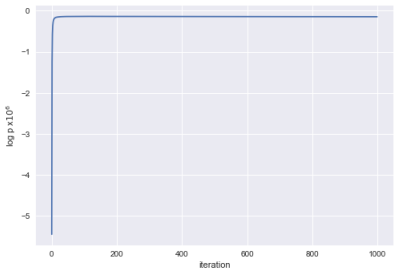

Figure 6: Evolution of $(\log p(X_n|y,\theta))_{n\in\mathbb{N}}$ with $(X_n)_{n\in\mathbb{N}}$ sampled using MYULA and targeting $p(\cdot|y,\theta)$ with $\theta = 0.001$ and the first 1000 iterations.

Then, I performed the algorithm 1 of SAPG presented in the paper, using a precision of $\epsilon_1 = 10^{-3}$ and $\epsilon_2 = 10^{-5}$. I observed 2 surprising phenomenons. First, as previously, I observed no oscillations in the evolution of θ. Second, the θs are of the ordre of 10, which can be explained by the fact that the final true regulariser parameter is $\sigma^2\theta$ and $\sigma^2 = 10^{-4}$, so actually θ is around 10^{-2}. So, I obtained by taking the average of the θs removing the first 25 iterations as indicated in the paper, $\theta^\star_{\epsilon_1} \approx 0.00100223$ and $\theta^\star_{\epsilon_2} \approx 0.00100226$.

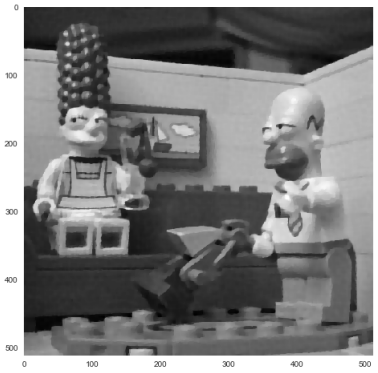

Figure 7: Image deblurred with $\theta^\star_{\epsilon_1}$

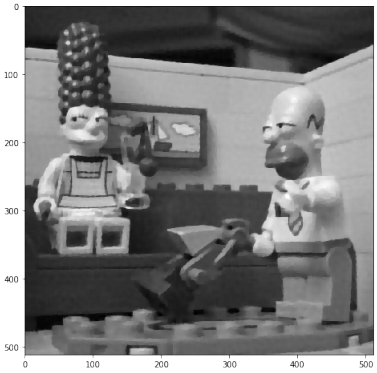

Figure 8: Image deblurred with $\theta^\star_{\epsilon_2}$

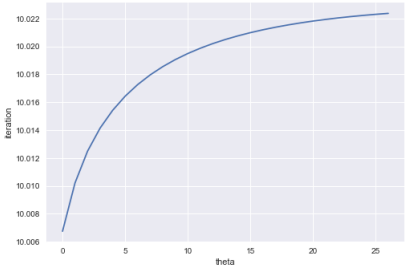

Figure 9: Evolution of θ in function of iterations with $\epsilon_1 = 10^{-3}$.

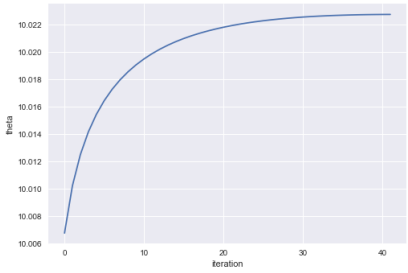

Figure 10: Evolution of θ in function of iterations with $\epsilon_2 = 10^{-5}$. We bserve coherently more iterations to converge.

Concerning the performance, I obtained $MSE_{\theta^\star_{\epsilon_1}} \approx 13.88985$ and $MSE_{\theta^\star_{\epsilon_2}} \approx 13.88988$ which is close to the minimum of the MSE performed by the Chambolle-Pock deblurring algorithm for θ between 10^{-6} and 1:

Problem 4 continued on next page...

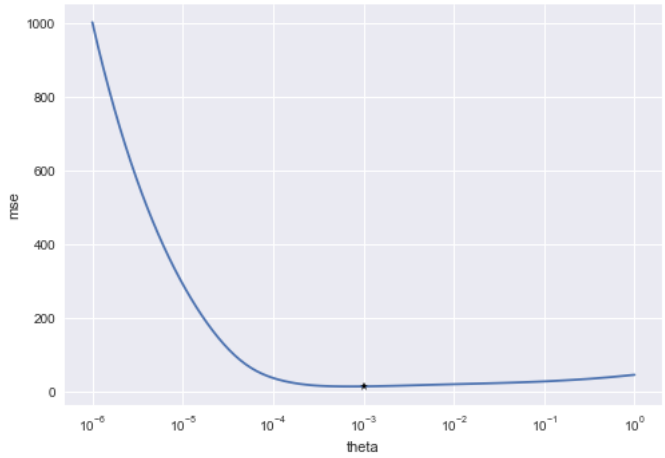

Figure 11: Evolution of MSE with a deblurred image with Chambolle-Pock algorithm in function of θ (the * corresponds to $\theta^{\star}_{\epsilon_1}$).

I performed the Chambolle-Pock deblurring algorithm for 100 values of regularisation between 10^{-6} and 1 in *log* scale, and I found that the minimum was 13.65348 attained by $\theta^{\star} \approx 7 \times 10^{-4}$, which is a bit lower than $\theta^{\star}_{\epsilon_1} \approx 10^{-3}$.

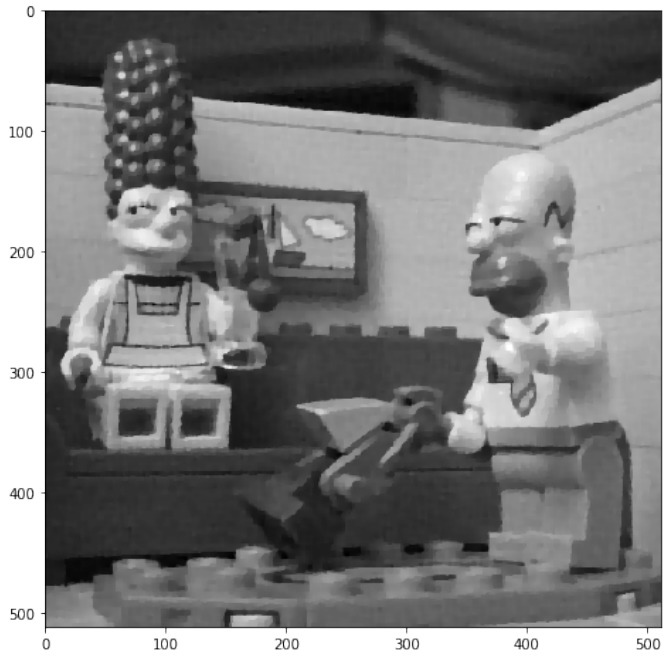

Figure 12: Image deblurred with $\theta^{\star}$

As a conclusion, I find that, despite some surprising behaviour of the algorithm in my experiment like the absence of oscillations or the fact that the performance in terms of MSE with a smaller precision is very slightly better, the final result on this image is quite good.

References

[1] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. working paper or preprint, June 2010.

[2] Ana F. Vidal, Valentin De Bortoli, Marcelo Pereyra, and Alain Durmus. Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical bayesian approach, 2019.