Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical Bayesian approach

Pierre Osselin
Department of Mathematics
ENS Paris-Saclay
pierre.osselin@gmail.com

Abstract

In this report, we review the paper recently written by Vidal et al. (2019) [15] that aims at introducing new methodologies for performing efficient and relevant parameter inference for inverse imaging problems. The introduced framework relies on both the Bayesian framework and the newly developed tools from the MCMC and optimization literature. In addition to developing efficient optimization algorithms in certain specific contexts, it gives new insights into how to specify prior parameters that are crucial for relevant model specification. First we review the paper, before referencing the notions seen in class and extending the paper on some theoretical points. We will finally attach this work with the reproduction and extension of some practical work undertaken by the paper and described in the last section.

1 Introduction

In imaging, inverse problems refer to many frameworks where one aims at recovering a true image given an observed image altered by some applications (e.g deblurring [8], denoising [13] or super-resolution [10]). Unfortunately, inverse problems are often ill-posed, in the sense that the information captured by our observation is insufficient to determine a solution (overdetermination is also a source for ill-posed problems, but it is often underdetermination that is the cause for ill-posed problem in inverse problems in imaging). Hence, we need to put regularisation on our space of images. This basically means creating more incentives for our solution to lies in a certain space of the set of images. Among many different ways to undertake this, we can implement it by taking a bayesian approach via a "prior" distribution over our images in order to prompt our solution to be in a certain space the we found to be "a priori" (without information) more likely.

While this approach can transform our ill-posed problem into a well-posed one, it has many limitations. First, the Bayesian approach can be computationally very expensive, thus in the large majority of cases the choice of the couple likelihood/prior is often altered by computational issues where convenient distributions are chosen rather than more relevant ones. This contradicts the point of using a Bayesian approach, as prior elicitation and careful modeling of the observation process matters, and the specified (non)parametric model will have some structure that won't adapt to every form of model mispecification. Second, we often choose the prior distribution to belong to a certain parametric family of distributions. While the family already displays a certain distributional structure, the choice of parameters specifies the appropriate regularisation. To set these parameters many different techniques have been developed (see e.g [2], [3], [14]). Taking the Bayesian Framework, two different strategies have been implemented. In the hierarchical strategy, a prior distribution over the parameters is elicitated, and the posterior distribution of our image is described by the sum of the contribution of

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

the posterior over the parameters. In the empirical strategy, we do not take the sum of the contribution of the posterior over the parameters but only over one inferred parameter carefully chosen from data.

In this report, we will describe an empirical approach to certain imaging problems, where the parameters will be selected via maximum marginal likelihood with recently developed tools consisting of stochastic proximal gradient algorithms powered by proximal Markov chain Monte Carlo samplers developed in [15]. We will first describe the methodology before further developing the theoretical properties of the algorithms used for this approach.

2 Methodology

2.1 Model

As described in the introduction we aim at recovering an unknown image $x \in \mathbb{R}^d$ from an observation $y \in \mathbb{R}^{d_y}$. We suppose we have already probabilistically modelled the alteration process via the following likelihood:

$$p(y|x) \sim e^{-f_y(x)} \tag{1}$$

with f_y convex and continuously differentiable with lipschitz gradient. In the litterature, this refers to the Helmholtz free [6] energy and in particular the probability to find the system in some energy state y with corresponding energy $f_y(x)$. Similarly, we specify prior knowledge about x via the following parametric family of prior distributions of the same form:

$$p(x|\theta) = e^{-\theta^T g(x)}/Z(\theta) \tag{2}$$

Where g is allowed to be non-smooth and $\theta \in \Theta$ is a convex compact set. This distribution promotes x where $g(x) \approx \int_{\mathbb{R}^d} g(\tilde{x})p(\tilde{x}|\theta)d\tilde{x}$. Bayes Theorem gives the posterior distribution whose maximization gives:

$$\hat{x}_{\theta,MAP} = \arg\min_x \{f_y(x) + \theta^T g(x)\} \tag{3}$$

By assumption, this optimization problem is convex and can, thus, be more efficiently solved.

2.2 Parameter inference

In the empirical Bayes approach, the parameters are selected via maximum marginal likelihood estimation, that is, the best parameter explaining data:

$$\theta_* \in \arg\max_\theta p(y|\theta) = \int_{\mathbb{R}^d} p(y|\tilde{x})p(\tilde{x}|\theta)d\tilde{x} \tag{4}$$

Posterior inference for the true image x is then performed via this estimated parameter:

$$p(x|y,\theta^*) \propto e^{-f_y(x)-\theta^* g(x)} \tag{5}$$

The main difficulty in this approach is that solving 4 is computationally intractable as it requires the integral $Z(\theta)$ and $p(y|\theta)$.

Given the quantity $p(y|\theta)$, the iterative convex projection algorithm provides a solution to the optimization of the parameter θ (Θ is a compact convex set):

$$\theta_{n+1} = \prod^{\Theta}[\theta_n + \delta\nabla_\theta log(p(y|\theta_n))]$$

With

$$\nabla_\theta log(p(y|\theta_n)) = -\int_{\mathbb{R}^d} g(\tilde{x})p(\tilde{x}|y,\theta)d\tilde{x} - \nabla_\theta log(Z(\theta)) \tag{6}$$

And use stochastic approximation proximal gradient algorithm for the first integral. The second term $\nabla_\theta log(Z(\theta))$ can be computed for many regulariser (e.g positively homogeneous regulariser or separably homogeneous regulariser see [15]). In the case where it is not computable, a second Markov chain can be used to approximate it, such as the first term.

2.3 Kernel

Given that the dimensionality is high, the optimization process can be very difficult, and is very dependent on the estimation error of our integral computed by our Markov Chains. Thus a careful kernel should be designed to obtain efficient algorithms. In this framework, the MYULA Markov kernel is a state-of-the-art proximal Markov chain Monte Carlo designed for high dimensionality and allows non-smooth g functions. This MCMC kernel is derived from the discretisation of an over-damped Langevin diffusion, $(X_t)_{t>0}$, satisfying the following stochastic differential equation. We will develop and prove certain properties of this algorithm in the last section.

$$d\boldsymbol{X}_t = -\nabla_x F(\boldsymbol{X}_t)dt + \sqrt{2}d\boldsymbol{B}_t \tag{7}$$

Where F will determine the steady distribution. Given a MCMC targeting our desired distribution, we can approximate both terms in 6. Once these terms are approximated we can set up our iterative convex projection algorithm called stochastic approximation proximal gradient algorithm (SAPG) whose more precise description is given in 1. In the description, $-\nabla_\theta log(Z(\theta))$ is either approximated via a second Markov Chain or given in closed form depending on the form of Z [15].

Algorithm 1: SAPG algorithm

```
input  : initial $\{\theta_0, X_0\}$, $(\delta_n, \gamma_n)_{n\in\mathbb{N}}$, number of iterations $N$.
for $i = 0$ to $N - 1$ do
    Sample $X_{n+1} \sim R_{\gamma_n,\theta_n}(X_n, .)$;
    Set $\theta_{n+1} = \prod_\Theta[\theta_n + \delta_{n+1}(-\nabla_\theta log(Z(\theta)) - g(X_{n+1}))]$
end
output : $\overline{\theta}_N = \frac{1}{N}\sum_{n=0}^{N-1}\theta_n$
```

3 Course References

3.1 Inverse Problem

Inverse problem is a capital topic of interest in imaging. It consists in estimating an unknown image $u \in \mathbb{R}^n$ given an observation $v \in \mathbb{R}^d$:

$$v = \underbrace{A}_{\text{degradation operator}} u + \underbrace{n}_{\text{noise}} \tag{8}$$

While classical approaches use Least squares estimates or Tikhonov regularisation, the Bayesian formalism introduces regularisation with the elicitation of a prior distribution over u: $p(u)$. The restoration is then carried out with Maximum a posteriori MAP estimates or Minimum mean square error MMSE.

3.2 Proximal operator for subdifferential Algorithms

When a function of interest is convex but not smooth (i.e not differentiable everywhere) the proximal operator can be leveraged to counter this limitation and to adapt classical algorithms to this particular situation. If F is convex, proper and l.s.c. then:

$$prox_{\sigma F}(u) = \arg\min_v \frac{1}{2}||v - u||_2^2 + \sigma F(v) \tag{9}$$

Given this operator, the operation $u_{k+1} = prox_{\sigma F}(u_k)$ is equivalent to an implicit subgradient descent. With desirable properties, this process converges to a unique fixed point by contractivity.

3.3 MCMC Algorithms

As mentioned in the previous section, sampling from certain distributions might be a complex and necessary task for non trivial problem resolution. When the distribution can't be sampled directly, either because there is no analytical way to do it (inverse sampling or transformation sampling for e.g) or because we don't have access to the normalization constant, in our case, we might use the MCMC technology. The rational behind MCMC algorithms is to construct markov chains (i.e the next step depends only on the current state) whose target (or asymptotic) distribution is the desired one. The most famous algorithm is the Metropolis-Hasting algorithm [4], where a proposal distribution is used and directed via acceptance probability toward the target distribution. Certain properties over the proposal distribution and the state space guarantee convergence of the chain. Many other algorithms have been developed, such as the Hamiltonian MCMC developed by Neal [9] which is the state of the art. In this paper we use the Langevin MCMC agorithm that exploits the gradient of the objective function to guide the Markov chain. The ULA is written as:

$$X(t_{k+1}) = X(t_k) - \nabla_x F(X(t_k))\Delta t + \sqrt{2}\Delta W(t_k) \tag{10}$$

With

$$\Delta W(t_k) \sim \sqrt{\Delta t} \times \mathcal{N}(0, I_d)$$

3.4 The Moreau-Yoshida Envelope

As mentionned previously, the Langevin dynamics require the gradient of F. If we can write $F = U + V$ where U is smoothed and V is not, but convex and l.s.c., we can replace it with the Moreau-Yosida regularisation:

$$V^{\lambda} = \inf_{y \in \mathbb{R}^d} \{g(y) + \frac{1}{2\lambda}||x - y||^2\} \leq g(x) \tag{11}$$

With known gradient and gradient lipschitzness. Replacing $F = U + V$ by $F = U + V^{\lambda}$ gives differentiability and allows us to perform Langevin MCMC on F^{λ}, the name of the method becomes the MYULA algorithm [12]:

$$X(t_{k+1}) = X(t_k) - \gamma\nabla_x U(X(t_k)) - \frac{\gamma}{\lambda}(X(t_k) - prox_{\lambda V}(X(t_k))) + \sqrt{2\gamma}W(t_k) \tag{12}$$

3.5 TV prior

Multiple implementations of the studied paper [15] instantiate their framework with a particular prior named the total variation. In this case $g(x) = TV(x)$ where:

$$TV(x) = \sum_{i=1}^{d} ||\nabla x_k|| \tag{13}$$

This particular prior is the foundation of many algorithm for image denoising or image restauration such as TV-L2, TV-L1 or TVICE.

3.6 Stochastic Optimization with Unadjusted Langevin

In the previous section, the presented algorithm leveraging iterative convex projection and Markov Chain samples to estimate our parameters can be seen as an extension of the SOUL algorithm, similarly to MYULA being an extension of ULA. While our stochastic approximation proximal gradient algorithm (SAPG) use a proximal gradient element in its langevin dynamic to cope with non-smooth convex functions, the Stochastic Optimization with Unadjusted Langevin (SOUL) proceeds in the same manner with a smooth functional and, hence, no proximal operator.

4 Theoretical Development

4.1 Stochastic Convergence

In physics, Langevin dynamics is an approach to the mathematical modeling of the dynamics of molecular systems. More specifically, it tries to model the friction effect of the environment in which the molecular system evolves, for instance air or solvent [7].

The over-damped Langevin diffusion process $(X_t)_{t\geq 0}$, satisfies the following stochastic differential equation [1]:

$$d\boldsymbol{X}_t = -\nabla_x F(\boldsymbol{X}_t)dt + \sqrt{2}d\boldsymbol{B}_t \tag{14}$$

where $F : \mathbb{R}^d \rightarrow \mathbb{R}$ is a continuously differentiable potential and $(\boldsymbol{B}_t)_{t\geq 0}$ is a d-dimensional Brownian motion. In the physics literature, F represents the potential energy between particles. The coefficients in front of $\boldsymbol{F}$ and $\boldsymbol{B}_t$ are default values for a physical system. We will develop this model and show how it can be used to derive Markov Chains of interests for our problem. First we will study the dynamics of the process arising from this equation (convergence, and to which distribution), then showed properties inherited by the Markov chains ULA and MYULA.

There are mainly two theorems for proving the existence and uniqueness of a strong solution for the previous equation.

Theorem 1. *(Zvonkin) Let $(\boldsymbol{B}_t)_{t\geq 0}$ be a d-dimensional Bownian motion. If the coefficients of the SDE*

$$d\boldsymbol{X}_t = b(t, \boldsymbol{X}_t)dt + \sigma(t, \boldsymbol{X}_t)d\boldsymbol{B}_t, \boldsymbol{X}_0 = \boldsymbol{x}_0 \tag{15}$$

are bounded, b is measurable, σ is continuous and there exist constants $C > 0$ and $\epsilon > 0$ such that:

$$\forall t \geq 0, x, y \in \mathbb{R}, |\sigma(t, x) - \sigma(t, y)| \leq C|x - y| \text{ and } |\sigma(t, x)| \geq \epsilon$$

then the SDE has a (unique) strong solution.

Theorem 2. *Assume that b and σ are uniformly Lipschitz, that is, there exists a constant $C < \infty$ such that $\forall x, y \in \mathbb{R}$,*

$$|b(t, x) - b(t, y)| \leq C|x - y|$$
$$|\sigma(t, x) - \sigma(t, y)| \leq C|x - y|$$

Then the SDE 15 has a (unique) strong solution.

Hence, to prove strong and uniqueness of the solution of 14, since $\sigma(t, \boldsymbol{X}_t) = \sqrt{2}$ clearly satisfies the conditions in theorem 1 and 2, having $\nabla_x \boldsymbol{F}(\boldsymbol{X}_t)$ either Lipschitz continuous or bounded and measurable is sufficient. Accordingly, to determine the asymptotic, invariant distribution of $\boldsymbol{X}_t$ as $t \rightarrow +\infty$, we need some tools from the stochastic differential equation literature. If one wants to study the evolution of the distribution of $\boldsymbol{X}_t$ through time, the Fokker–Planck equation (or Kolmogorov forward equation) can be used. If the problem is to know the distribution at previous times the Kolmogorov backward equation can be studied.

Theorem 3. *Let $\boldsymbol{X}_t$ and $\boldsymbol{\mu}(\boldsymbol{X}_t, t)$ be a d-dimensional random vectors, $\boldsymbol{\sigma}(\boldsymbol{X}_t, t) \in \mathbb{R}^{d\times m}$ and $\boldsymbol{W}_t$ a m-dimensional standard Wiener process, consider the SDE*

$$d\boldsymbol{X}_t = \mu(\boldsymbol{X}_t, t)dt + \sigma(\boldsymbol{X}_t, t)d\boldsymbol{W}_t$$

with drift $\boldsymbol{\mu}(\boldsymbol{X}_t, t)$ and diffusion tensor $\boldsymbol{D} = \frac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^T$ the Fokker–Planck equation for the probability density $p(\boldsymbol{x}, t)$ of the random variable $\boldsymbol{X}_t$ is

$$\frac{\partial p(\boldsymbol{x}, t)}{\partial t} = -\sum_{i=1}^{d} \frac{\partial}{\partial x_i}[\mu_i(\boldsymbol{x}, t)p(\boldsymbol{x}, t)] + \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial^2}{\partial x_i x_j}[\boldsymbol{D}_{i,j}(\boldsymbol{x}, t)p(\boldsymbol{x}, t)] \tag{16}$$

Let us consider the Langevin equation 14, the equivalent Fokker–Planck equation becomes

$$\frac{\partial}{\partial t}p(\boldsymbol{x}, t) = Div(p(\boldsymbol{x}, t)\nabla F(\boldsymbol{x})) + \Delta p(\boldsymbol{x}, t)$$

And the steady distribution verifies $\frac{\partial p(\boldsymbol{x},t)}{\partial t} = 0$ hence:

$$Div(p(\boldsymbol{x}, t)\nabla F(\boldsymbol{x})) + \Delta p(\boldsymbol{x}, t) = 0$$

By rewriting:

$$Div(p(\boldsymbol{x}, t)\nabla F(\boldsymbol{x}) + \nabla p(\boldsymbol{x}, t)) = 0$$

which means the d-dimensional function $p(\boldsymbol{x}, t)\nabla F(\boldsymbol{x}) + \nabla p(\boldsymbol{x}, t)$ is volume preserving. By the following result

$$Div(\phi \boldsymbol{F}) = (\nabla\phi).\boldsymbol{F} + \phi(Div(\boldsymbol{F}))$$

We have:

$$(\nabla p(\boldsymbol{x})).(\nabla F(\boldsymbol{x})) + p(\boldsymbol{x})Div(\nabla F(\boldsymbol{x})) = 0$$

The function $log(p(\boldsymbol{x})) = -F(\boldsymbol{x}) + C$ is a solution, which gives

$$p(\boldsymbol{x}) \propto e^{-F(\boldsymbol{x})}$$

As a steady solution to which our process converges.

4.2 MCMC Convergence

Now we know that our defined stochastic dynamic 14 converges toward a known distribution that can be targeted, we now want to sample paths from this process in order to approximate our distribution. If the SDE was simple, we could compute the solution X_t as a distribution analytically and sample from it with known distributions. In our case, and more generally, this is not possible and we have to resort to numerical approximation schemes in order to simulate sample paths of solutions to the given equation. The simplest scheme is obtained by using a first-order approximation. This is called the Euler scheme:

$$X(t_{k+1}) = X(t_k) - \nabla_x F(X(t_k))\Delta t + \sqrt{2}\Delta W(t_k) \tag{17}$$

Where $\Delta W(t_k)$ is a d-dimensional normal distribution of uniform variance Δt:

$$\Delta W(t_k) \sim \sqrt{\Delta t} \times \mathcal{N}(0, I_d)$$

This first order scheme for our equation is also called the Unadjusted Langevin Algorithm (ULA) [5] where $\gamma = \Delta t$ is the time discretization. Since in our framework, our energy function is not smooth in general, our ULA is limited in practice. The MYULA [12] overcomes this difficulty by assuming there is a decomposition $F = U + V$ where U is Lipschitz differentiable and V is not. The ULA is then applied to F where V is replaced with a smooth version of itself, namely the Moreau-Yosida envelop of V defined by:

$$\forall x \in \mathbb{R}^d, \lambda > 0 : V^\lambda(x) = \min_{\tilde{x} \in \mathbb{R}^d} \{V(\tilde{x}) + \frac{1}{2\lambda}||x - \tilde{x}||_2^2\} \tag{18}$$

Theorem 4. *(smoothness of the Moreau envelope)*
The Moreau Envelope is differentiable and has the following gradient:.

$$\forall x \in \mathbb{R}^d : \nabla V^\lambda(x) = (x - prox_V^\lambda(x))/\lambda \tag{19}$$

Where

$$prox_V^\lambda(x)) = \arg\min_{\tilde{x} \in \mathbb{R}^d} \{V(\tilde{x}) + \frac{1}{2\lambda}||x - \tilde{x}||_2^2\} \tag{20}$$

Proof. Let $\xi \in \mathbb{R}^d$ and $u : x \in \mathbb{R}^d \to u(x) = \arg\min_{\tilde{x} \in \mathbb{R}^d}\{V(\tilde{x}) + \frac{1}{2\lambda}||x - \tilde{x}||_2^2\}$ we have, by definition of the min:

$$V^\lambda(x + \xi) = V(u(x + \xi)) + \frac{1}{2\lambda}||x + \xi - u(x + \xi)||^2 \leq V(u(x)) + \frac{1}{2\lambda}||x + \xi - u(x)||^2$$

Hence:

$$\begin{aligned} V^\lambda(x + \xi) - V^\lambda(x) &\leq \frac{1}{2\lambda}(||x + \xi - u(x)||^2 - ||x - u(x)||^2) \\ &= \frac{1}{2\lambda}\xi^T 2(x - u(x)) + \mathcal{O}(||\xi||^2) \end{aligned} \tag{21}$$

Moreover, V^λ is convex, hence $\xi \to V^\lambda(x + \xi)$ is convex as well (check [11] Theorem 2.19). We deduce that:

$$V^\lambda(x + \xi) - V^\lambda(x) \geq -(V^\lambda(x - \xi) - V^\lambda(x)) \tag{22}$$

And

$$V^\lambda(x + \xi) - V^\lambda(x) \geq -\frac{1}{2\lambda}\xi^T 2(x - u(x)) + \mathcal{O}(||\xi||^2)$$

Hence

$$\nabla V^\lambda(x) = \frac{1}{\lambda}(x - u(x)) \tag{23}$$

□

Given this differentiability, apply the ULA algorithm to the modified F^λ becomes the MYULA algorithms that reads:

$$X(t_{k+1}) = X(t_k) - \gamma\nabla_x U(X(t_k)) - \frac{\gamma}{\lambda}(X(t_k) - prox_{\lambda V}(X(t_k))) + \sqrt{2\gamma}W(t_k) \tag{24}$$

We have a well defined algorithm, the MYULA, at our disposal that seems to be applicable to a wide range of practical cases and is an adaptation of a well known algorithm, namely the ULA. The last important topics we should lean on are, first, if the algorithm converges to a distribution that is "close" enough to the desired target density and with which dependency with regards to λ and γ and, second, if the convergence is reasonably fast. Fortunately, we have theoretical results ensuring desirable properties [12]. Let the following conditions:

$\boldsymbol{H1.}$ U is convex is convex, continuously differentiable, and gradient Lipschitz. V is proper, convex, and l.s.c.

$\boldsymbol{H2.}$ There exist a minimizer x^* of F^λ $(= U + V^\lambda)$, $\mu_c > 0$, and $R_c \geq 0$ such that for all $x \in \mathbb{R}^d$, $||x - x^*|| \geq R_c$,

$$F^\lambda(x) - F^\lambda(x^*) \geq \mu_c ||x - x^*|| \tag{25}$$

Theorem 5. *Assume $\boldsymbol{H1.}$ and $\boldsymbol{H2.}$. For all $\epsilon > 0$ and $x \in \mathbb{R}^d$ we have:*

$$||R^n_\gamma(x, .) - \pi||_{TV} \leq \epsilon \tag{26}$$

If

$$n \geq \mathcal{O}(d^5 log^2(\epsilon^{-1})\epsilon^{-2}) \tag{27}$$

Where $R^n_\gamma(x, .)$ is the distribution of X_n knowing that $X_0 = x$

This result implies that the number of iterations to reach a precision target ϵ is, at worse, of order $d^5 log^2(\epsilon^{-1})\epsilon^{-2})$ for this class of models. The dependency in λ and γ is hidden in the coefficients in $\mathcal{O}$ given in [12].

5 Numerical Experiment

In this section we will demonstrate the performance of the algorithm 1 by sampling synthetic data from our generative model with specified parameter θ and trying to recover it with our algorithm.

5.1 Model specification

More specifically, we will formulate a denoising inverse problem where we will consider the wavelet-base $x \in \mathbb{R}^d$ of our observed image $y \in \mathbb{R}^{d_y}$ with $d_y = 256 \times 256$ dimensions. The reconstructed image from the coefficients x in a an orthogonal 4-level Haar basis is given by the operator Ψ : $\Psi x = z \in \mathbb{R}^{d_y} \iff x = \Psi^T z \in \mathbb{R}^d$ which is an orthogonal operator. Now our generative model is specified as follows: for any $x \in \mathbb{R}^d$, our observation is :

$$f_y(x) = \frac{1}{2\sigma^2}||y - \Psi x||_2^2 \tag{28}$$

And the prior is specified by:

$$g(x) = ||x||_1 \tag{29}$$

5.2 Implementation Details

In the presented algorithm there are many quantities that need to be determined and that are crucial to obtain a good convergence behaviour. We were mainly inspired from the paper [15] that already carried out a thorough study on this subject. With L_y the Lipschitz constant of the gradient of f_y, we have $L_y = \frac{1}{\sigma^2}$ since $\Psi^T\Psi = I$ in our case. The paper suggests to take $\lambda = \frac{1}{L_y} = \sigma^2$, $\gamma_n = 0.98 \times (L_y + \frac{1}{\lambda})^{-1} = 0.98 \times (\frac{1}{\sigma^2} + \frac{1}{\lambda})^{-1}$, $\delta_n = \frac{n^{-0.8}}{\theta_0 d}$. We also performed a warm-up initialization with $T_0 = 10$.

5.3 Results

We undertook two experiments, the first one aims at assessing the convergence rate of our algorithm towards the desired parameter θ, and the influence of the initialization, noise and dimension to this quantity. In Figure 1 we initialize θ_0 according to the uniform law $\mathcal{U}[0.5, 3.5]$, we also took $\theta = 2$ and $\sigma = 0.1$. Our theta chain oscillates around the true θ value, with an oscillation amplitude that increases with the distance of the initialization with regards to the true θ. In Figure 2 we showed one

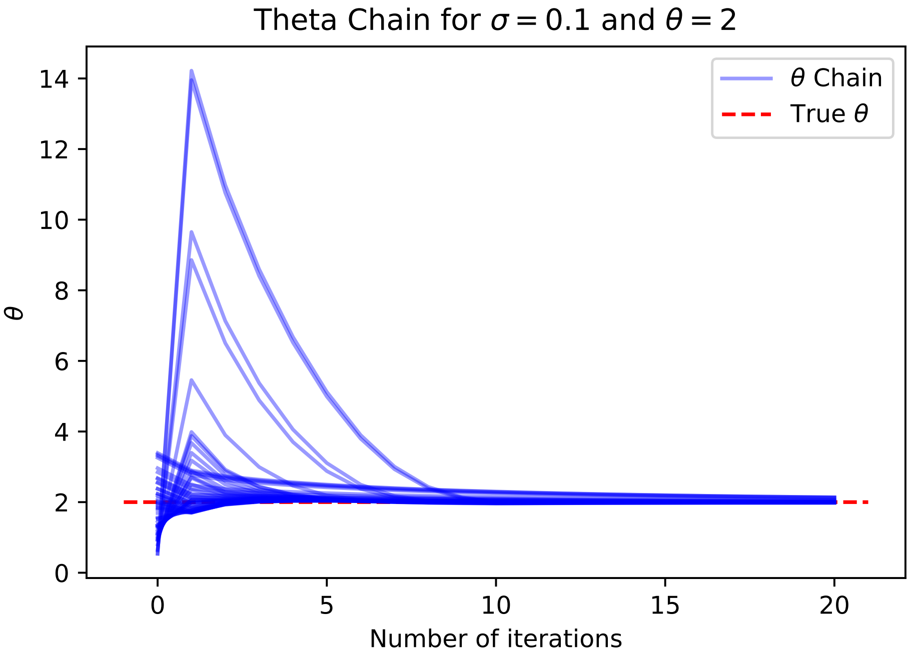

Figure 1: Effect of the initialization of θ_0

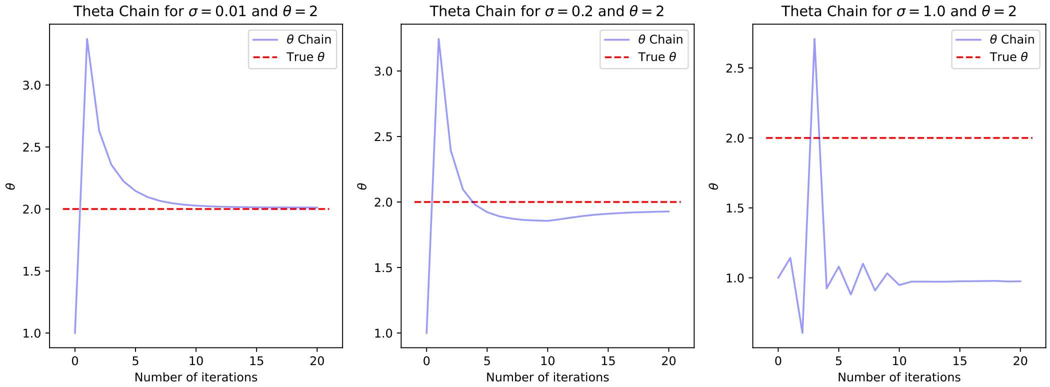

Figure 2: Effect of the noise standard deviation σ

θ chain per plot with different noise level $\sigma \in [0.01, 0.2, 1.]$. A small noise implies fast convergence toward a θ estimate that presents a small bias with regards to the true θ. On the contrary, when the noise increases the convergence is slower in addition to have a θ estimate that has a large bias. Finally, Figure 3 shows the same experiment with $\sigma = 0.1$ and $\theta = 2$ by changing the dimension of x and y. The figure shows that the convergence is not affected by this parameter, this confirms the remark of the paper stating that, since $d_y >> d_\Theta$ ($= 1$ in our case), the marginal likelihood concentrates sharply around a single maximiser θ^*, and is strongly log-concave w.r.t. θ in the neighbourhood of θ^*. Our last experiment consists in estimating the true θ by plotting the distribution of θ that we obtain by sampling multiple synthetic images from our generative model and applying 1 to recover theta with different σ. We made three plots corresponding to $\sigma \in [0.01, 0.2, 1.]$ and sampled 200 synthetic images from which we recovered the true θ with the MYULA algorithm with 80 steps. On Figure 4, we can see that the bias with regards to the true θ increases with σ as expected (roughly 0.01 for $\sigma = 0.01$, 0.06 for $\sigma = 0.2$ and 3. for $\sigma = 1.$).

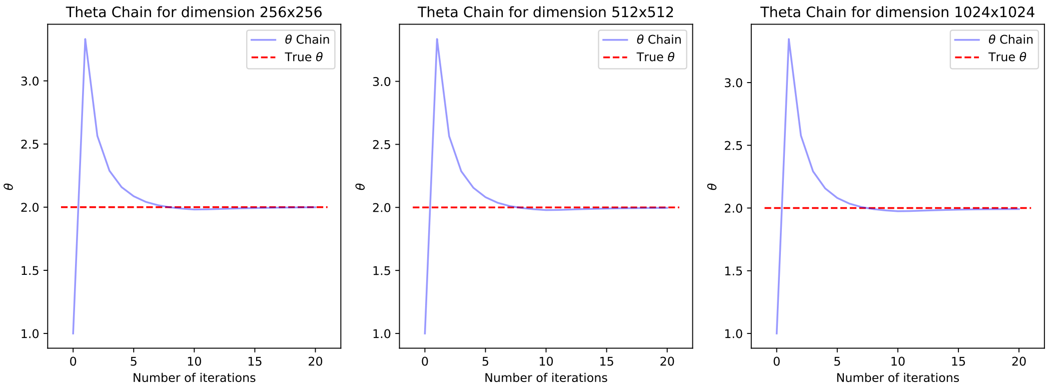

Figure 3: Effect of the dimension of the image for $\sigma = 0.1$ and $\theta = 2$

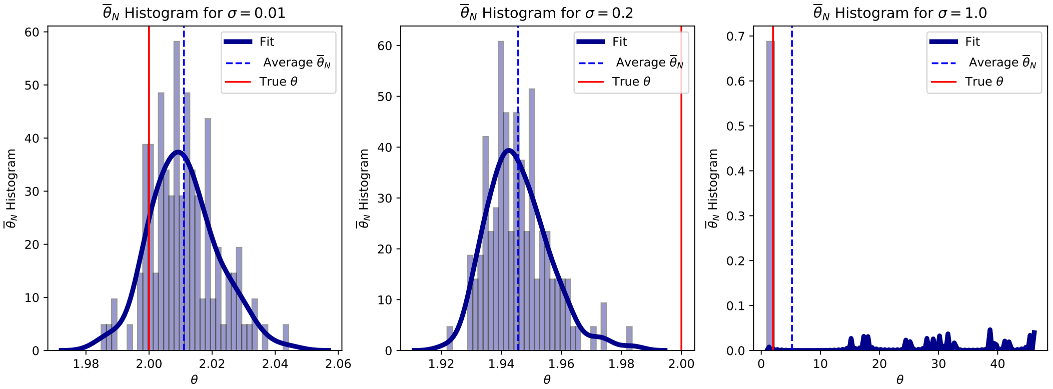

Figure 4: Theta estimates of SAPG

References

[1] Paul Langevin. “Sur la théorie du mouvement brownien”. In: *Compt. Rendus* 146 (1908), pp. 530–533.

[2] Gene H Golub, Michael Heath, and Grace Wahba. “Generalized cross-validation as a method for choosing a good ridge parameter”. In: *Technometrics* 21.2 (1979), pp. 215–223.

[3] Per Christian Hansen and Dianne Prost O’Leary. “The use of the L-curve in the regularization of discrete ill-posed problems”. In: *SIAM journal on scientific computing* 14.6 (1993), pp. 1487–1503.

[4] Siddhartha Chib and Edward Greenberg. “Understanding the metropolis-hastings algorithm”. In: *The american statistician* 49.4 (1995), pp. 327–335.

[5] Gareth O Roberts, Richard L Tweedie, et al. “Exponential convergence of Langevin distributions and their discrete approximations”. In: *Bernoulli* 2.4 (1996), pp. 341–363.

[6] Reiner Tillner-Roth and Daniel G Friend. “A Helmholtz free energy formulation of the thermodynamic properties of the mixture {water+ ammonia}”. In: *Journal of Physical and Chemical Reference Data* 27.1 (1998), pp. 63–96.

[7] Jan A Freund and Thorsten Pöschel. *Stochastic processes in physics, chemistry, and biology*. Vol. 557. Springer Science & Business Media, 2000.

[8] Per Christian Hansen, James G Nagy, and Dianne P O’leary. *Deblurring images: matrices, spectra, and filtering*. Vol. 3. Siam, 2006.

[9] Radford M Neal et al. “MCMC using Hamiltonian dynamics”. In: *Handbook of markov chain monte carlo* 2.11 (2011), p. 2.

[10] Veniamin I Morgenshtern and Emmanuel J Candes. “Super-resolution of positive sources: The discrete setup”. In: *SIAM Journal on Imaging Sciences* 9.1 (2016), pp. 412–444.

[11] Amir Beck. *First-order methods in optimization*. Vol. 25. SIAM, 2017.

[12] Alain Durmus, Eric Moulines, and Marcelo Pereyra. “Efficient bayesian computation by proximal markov chain monte carlo: when langevin meets moreau”. In: *SIAM Journal on Imaging Sciences* 11.1 (2018), pp. 473–506.

[13] Antoine Houdard, Charles Bouveyron, and Julie Delon. “High-dimensional mixture models for unsupervised image denoising (HDMI)”. In: *SIAM Journal on Imaging Sciences* 11.4 (2018), pp. 2815–2846.

[14] Federico Benvenuto and Cristina Campi. “A discrepancy principle for the Landweber iteration based on risk minimization”. In: *Applied Mathematics Letters* 96 (2019), pp. 1–6.

[15] Ana F Vidal et al. “Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems: an empirical Bayesian approach”. In: *arXiv preprint arXiv:1911.11709* (2019).