ML ESTIMATION OF WAVELET REGULARIZATION HYPERPARAMETERS IN INVERSE PROBLEMS

Roberto Cavicchioli[1], *Caroline Chaux*[2], *Laure Blanc-Féraud*[3] *and Luca Zanni*[1]

[1] Department of Physics, Computer Science and Mathematics, University of Modena and Reggio Emilia, Via Campi, 213/b, Modena, Italy
e-mail:{firstname.lastname}@unimore.it

[2] Aix-Marseille Univ., LATP UMR CNRS 7353, 39 Rue F. Joliot-Curie, Marseille, France
e-mail: caroline.chaux@latp.univ-mrs.fr

[3] MORPHEME I3S/INRIA/IBV, 2000 Route des Lucioles, Sophia-Antipolis, France
e-mail: Laure.Blanc_Feraud@inria.fr

ABSTRACT

In this paper we are interested in regularizing hyperparameter estimation by maximum likelihood in inverse problems with wavelet regularization. One parameter per subband will be estimated by gradient ascent algorithm. We have to face with two main difficulties: i) sampling the a posteriori image distribution to compute the gradient; ii) choosing a suited step-size to ensure good convergence properties. We first show that introducing an auxiliary variable makes the sampling feasible using classical Metropolis-Hastings algorithm and Gibbs sampler. Secondly, we propose an adaptive step-size selection and a line-search strategy to improve the gradient-based method. Good performances of the proposed approach are demonstrated on both synthetic and real data.

Index Terms— Parameter estimation, Maximum likelihood estimation, Wavelet transforms, Deconvolution, Gradient methods

1. INTRODUCTION

We are interested here in solving inverse problems, the direct formulation of which can be written as

$$g = Au + \eta, \tag{1}$$

where g (vector of dimension k) denotes the observed image, A (matrix of dimension $k \times k$) denotes a linear operator (e.g. blur) which is assumed to be known and non necessarily invertible and finally, $\eta \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$ models an additive white Gaussian noise of known variance σ^2.

Restoring u from g is an ill-posed inverse problem which must be regularized. We consider in this work regularization in the wavelet domain [1, 2, 3, 4, 5, 6], by minimizing a convex criterion given by

$$J(u) = \frac{\|g - Au\|^2}{2\sigma^2} + \sum_{\mathbf{m}=1}^{M} \lambda_{\mathbf{m}} \phi_{\mathbf{m}}(F_{\mathbf{m}} u). \tag{2}$$

The first term corresponds to the negative log-likelihood in the Gaussian white noise case, the second one is the regularization term in the wavelet domain: $F_{\mathbf{m}}u$ denotes a subband and F represents the global orthogonal wavelet transform operator [7]. Functions $\phi_{\mathbf{m}}$ model distributions of the wavelet coefficients and are chosen to be ℓ^1-norms on all the subbands. The resolution level and the number of channels (dyadic, Q-band) of the decomposition determine the finite number M of subband $\mathbf{m} = 1, ..., M$.

Tuning more than two hyperparameters by an exhaustive search is prohibitive. As the restoration results are very sensitive to these hyperparameters $\boldsymbol{\lambda} = (\lambda_{\mathbf{m}=1,...,M})_{\mathbf{m}}$, we propose an automatic estimation. Well known state of the art methods as discrepancy principle based methods [8, 9], cross-validation methods [10] or Stein principle based approaches [5, 11] are mainly restricted to the estimation of one hyperparameter only. Other stochastic methods such as Maximum Likelihood (ML) approaches [12, 13], EM algorithms (Expectation-Maximization) [14, 3], or MCMC (Monte Carlo Markov Chain) based sampling methods [15, 16], allow the estimation of several hyperparameters but have to face with sampling difficulties.

We adopt a ML strategy which allows us to take advantage of the good asymptotic properties of the ML estimator and to estimate a vector of hyperparameters. However, applying a gradient ascent algorithm to compute ML hyperparameter estimates is dramatically prohibitive in terms of time computing for two reasons. First, computing the gradient of the likelihood function requires to sample the a posteriori distribution, whose energy is defined by (2) [17]. Direct application of Gibbs sampling and Metropolis-Hastings is not possible due to the simultaneous presence of operators A and $F_{\mathbf{m}}$ in (2). Second, gradient ascent methods converge slowly and much attention have to be paid to step-size determination. Our contribution in this paper is double. First, inspired by the approach described in [4], we propose to introduce by inference an auxiliary variable which separates operators A and F in the criterion. Second, we exploit an adaptive step-size selection and a line-search strategy by defining a two phase algorithm increasing the convergence speed of the gradient ascent algorithm.

The remaining of this paper is organized as follows. Section 2, describes the adopted Maximum Likelihood strategy and shows that an auxiliary variable must be inserted to be able to perform samplings. The proposed two phases gradient method is then described in Section 3. Results are shown in Section 4 and Section 5 concludes the paper.

2. MAXIMUM LIKELIHOOD ESTIMATION OF THE HYPERPARAMETERS

2.1. Classical approach

The ML estimation of the vector of hyperparameters $\boldsymbol{\lambda} = (\lambda_{\mathbf{m}})_{\mathbf{m}=1,\dots,M}$ consists in maximizing $p_{\boldsymbol{\lambda}}(g)$ w.r.t $\boldsymbol{\lambda}$ where $p_{\boldsymbol{\lambda}}(g)$ is given by:

$$p_{\boldsymbol{\lambda}}(g) = \int_u p_{\boldsymbol{\lambda}}(g,u)du = \int_u p(g|u)p_{\boldsymbol{\lambda}}(u)du. \quad (3)$$

The integration domain of u is $[0,255]^k$ where k represents the number of elements in u (e.g. number of pixels for a 2D image).

In our context, we can easily derive that

$$p(g|u) = \frac{1}{K_\sigma}\exp\Big(\frac{-\|g-Au\|^2}{2\sigma^2}\Big) \quad (4)$$

where $K_\sigma = (2\pi)^{k/2}\sigma^k$ and

$$p_{\boldsymbol{\lambda}}(u) = \frac{1}{Z_{\boldsymbol{\lambda}}}\prod_{\mathbf{m}=1}^{M}\exp\left(-\lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u)\right) \quad (5)$$

where

$$Z_{\boldsymbol{\lambda}} = \int_u \exp\left(-\sum_{\mathbf{m}=1}^{M}\lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u)\right)du. \quad (6)$$

The maximization of $p_{\boldsymbol{\lambda}}(g)$ with respect to $\boldsymbol{\lambda}$ can be performed using a gradient method [12] and thus requires to compute the derivatives of $p_{\boldsymbol{\lambda}}(g)$ w.r.t. each element of $\boldsymbol{\lambda}$. After some calculations and by invoking the Lebesgue dominated convergence theorem we obtain

$$\frac{\partial \log p_{\boldsymbol{\lambda}}(g)}{\partial \lambda_{\mathbf{m}}} = \mathrm{E}_{\boldsymbol{\lambda}}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)] - \mathrm{E}_{\sigma,\boldsymbol{\lambda}}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)] \quad (7)$$

where the first expectation is defined according to the *a priori* law (5) and the second one, according to the *a posteriori* law

$$p_{\sigma,\boldsymbol{\lambda}}(u|g) = \frac{\exp\Big(\frac{-\|g-Au\|^2}{2\sigma^2} - \sum_{\mathbf{m}=1}^{M}\lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u)\Big)}{\int_u \exp\Big(\frac{-\|g-Au\|^2}{2\sigma^2} - \sum_{\mathbf{m}=1}^{M}\lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u)\Big)du}. \quad (8)$$

The problem here is that, contrary to the first expectation, the second expectation cannot be computed analytically and we have to generate samples according to (8) in order to estimate it by an empirical mean. To this end, we have to tackle the pixel dependence induced by A and this can be done by performing the calculations in the Fourier domain, where A can be diagonalized. However, this is not sufficient as F cannot be diagonalized in the same space. Operators A and F must be split.

2.2. Decoupling of the linear operators

This splitting can be done by adopting the approach described in [4]. Indeed, the authors proposed to introduce an auxiliary variable w (hidden variable) and thus, they show that

$$\frac{\|g-Au\|^2}{2\sigma^2} = \min_w \frac{1}{2\sigma^2\mu}\Big(\|u-w\|^2 + \langle Cw, w\rangle\Big) + \frac{1}{2\sigma^2}\Big(\|g\|^2 - 2\langle Au, g\rangle\Big), \quad (9)$$

where $C = B(\mathrm{I}-B)^{-1}$ and $B = \mu A^*A$ (μ such that $\mu\|A^*A\| < 1$). This means that a new criterion can be considered instead of (2) which is given by

$$J(u,w) = \frac{1}{2\sigma^2\mu}\Big((w-(\mathrm{I}+C)^{-1}u)^{\mathrm{T}}(\mathrm{I}+C)(w-(\mathrm{I}+C)^{-1}u)\Big) + \frac{1}{2\sigma^2}\Big(\|g-Au\|^2\Big) + \sum_{\mathbf{m}=1}^{M}\lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u) \quad (10)$$

as $(\mathrm{I}+C)^{-1} = \mathrm{I} - \mu A^*A$. Note that it is shown in [4] that minimizing $J(u)$ w.r.t u is equivalent to minimize $J(u,w)$ w.r.t (u,w).

We have shown [18] that $p_{\boldsymbol{\lambda}}(g) = \int_{u,w} p_\lambda(g,u,w)dwdu$ and we can see from (10) that variables g and w are independent conditionally to u. Consequently, similarly to Section (2.1), we can derive an ML estimation of the parameters by maximizing $p_{\boldsymbol{\lambda}}(g)$ which is now given by

$$p_{\boldsymbol{\lambda}}(g) = \int_{u,w} p(g|u)p(w|u)p_{\boldsymbol{\lambda}}(u)dudw \quad (11)$$

where $p(g|u)$ is still defined by (4), $p_{\boldsymbol{\lambda}}(u)$ is still defined by (5) and

$$p(w|u) = \frac{\exp\Big(-\frac{(w-(\mathrm{I}+C)^{-1}u)^{\mathrm{T}}(\mathrm{I}+C)(w-(\mathrm{I}+C)^{-1}u)}{2\sigma^2\mu}\Big)}{K_\mu} \quad (12)$$

is a Gaussian law $\mathcal{N}\big((\mathrm{I}+C)^{-1}u, \sigma^2\mu(\mathrm{I}+C)^{-1}\big)$, with $K_\mu = (2\pi\sigma^2\mu)^{k/2}(\det(\mathrm{I}+C))^{-1/2}$.

The maximization of $p_{\boldsymbol{\lambda}}(g)$ with respect to each parameter $\lambda_{\mathbf{m}}$ requires also the computation of the derivatives of $p_{\boldsymbol{\lambda}}(g)$ w.r.t. each $\lambda_{\mathbf{m}}$ and leads to (invoking again the Lebesgue dominated convergence theorem):

$$\frac{\partial \log p_{\boldsymbol{\lambda}}(g)}{\partial \lambda_{\mathbf{m}}} = \mathrm{E}_{\boldsymbol{\lambda}}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)] - \mathrm{E}_{\sigma,\boldsymbol{\lambda},\mu}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)]. \quad (13)$$

Again, the first expectation can be computed analytically while for the second one, we have to generate samples according to the *a posteriori* law $p_{\sigma,\boldsymbol{\lambda},\mu}(u,w|g)$. Contrary to the previous case, the sampling is now possible in reasonable computing time as $J(u,w)$ can be rewritten as

$$J(u,w) = \frac{1}{2\sigma^2\mu}\Big(\|u-w\|^2 + \langle Cw, w\rangle\Big) + \frac{1}{2\sigma^2}\Big(\|g\|^2 - 2\langle u, A^*g\rangle\Big) + \sum_{\mathbf{m}=1}^{M} \lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u). \quad (14)$$

In this second configuration, variables u and w are now decoupled and they can be both estimated in two decorrelated spaces (wavelets for u, Fourier for w).

2.3. Metropolis within Gibbs algorithm

The sampling according to the a posteriori law can be done using a two steps algorithm that alternates a Gibbs sampler and a Metropolis Hastings procedure [19]. Indeed, to sample according to $p_{\sigma,\boldsymbol{\lambda},\mu}(u,w|g)$:

1. We first generate (Gibbs sampler) samples according to $p(w|u)$ given by (12) (Gaussian law). The variable w is directly expressed in the Fourier transform domain as the covariance matrix can be diagonalized easily;
2. Secondly, having a generation of w samples, we generate u samples or more precisely, directly wavelet transform coefficients (Metropolis Hastings algorithm) according to $p_{\boldsymbol{\lambda}}(u|w,g)$ where:

$$p_{\boldsymbol{\lambda}}(u|w,g) \propto \exp\Big(-\frac{1}{2\sigma^2\mu}\|Fu - Fw\|^2 + \frac{1}{\sigma^2}\langle Fu, FA^*g\rangle - \sum_{\mathbf{m}} \lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u)\Big). \quad (15)$$

3. PROPOSED ALGORITHM

3.1. Classical gradient ascent

As mentioned previously, parameters $\lambda_{\mathbf{m}}$ are computed by launching a gradient ascent (GA) algorithm [12] which can be written here as, $\forall \mathbf{m} \in \{1,\dots,M\}$,

$$\lambda_{\mathbf{m}}^{(n+1)} = \lambda_{\mathbf{m}}^{(n)} + \alpha_n\Big[\mathrm{E}_{\boldsymbol{\lambda}}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)] - \frac{2}{L}\sum_{l=L/2+1}^{L} \phi_{\mathbf{m}}\big(F_{\mathbf{m}}{u^{(n)}}_{\sigma,\boldsymbol{\lambda}^{(n)},\mu}\big)_l\Big]. \quad (16)$$

The first expectation is computed analytically (closed-form expression for the chosen $\phi_{\mathbf{m}}$) and $\big({u^{(n)}}_{\sigma,\boldsymbol{\lambda}^{(n)},\mu}\big)_l$ denotes the l-th sample generated according to the *a posteriori* probability density $p_{\boldsymbol{\lambda}^{(n)}}(u,w|g)$. Here, L denotes the number of computed samples and $L/2$ samples are required to initialize the chain. The parameter α_n represents here the step-size of the algorithm and can vary along with the iterations. The choice of this step-size is crucial and directly governs the algorithm convergence. For this reason, we decided to pay much attention to it as described later.

3.2. Acceleration

In order to accelerate gradient methods, an effective technique consists in using adaptive step-size rules for defining the step α_n along the gradient direction, combined with line-search strategies that, if necessary, shorten the step for ensuring suited improvements in the objective function. We exploit these ideas also for designing an accelerated gradient approach for maximizing $p_{\boldsymbol{\lambda}}(g)$. The main difficulty is due to the fact that the objective function defined in (11) can't be evaluated and, as a consequence, standard line-search strategies are not useful. To overcome this difficulty we develop a two phases (2Ph) gradient method in which, firstly, a sequence of simple gradient steps are performed with the aim to improve the objective function and, secondly, a line-search that avoids using the objective function is introduced in the iterative process for ensuring the reduction of the gradient norm. The first phase consists in steps of the form (16) in which the step-size is obtained by an adaptive alternation of the well known Barzilai-Borwein values [20]:

$$\alpha_n^{BB1} = -\frac{s_n^T s_n}{s_n^T y_n}, \qquad \alpha_n^{BB2} = -\frac{s_n^T y_n}{y_n^T y_n},$$

where $s_n = \boldsymbol{\lambda}^{(n)} - \boldsymbol{\lambda}^{(n-1)}$ and $y_n = \nabla p_{\boldsymbol{\lambda}^{(n)}}(g) - \nabla p_{\boldsymbol{\lambda}^{(n-1)}}(g)$. The adaptive alternation is derived from [21, 22] and described in [23]. In the second phase a line-search strategy similar to those proposed in [24] is used. Denoting $G(\boldsymbol{\lambda}) = \nabla p_{\boldsymbol{\lambda}}(g)$ and $f(\boldsymbol{\lambda}) = \frac{1}{2}\|G(\boldsymbol{\lambda})\|^2$, this line-search allows to sufficiently reduce $f(\boldsymbol{\lambda})$ by performing an ascent gradient step or moving along the descent direction for $f(\boldsymbol{\lambda})$ defined by $\bar{q}_n(\alpha_n) = -(G(\boldsymbol{\lambda}^{(n)} + \alpha_n G(\boldsymbol{\lambda}^{(n)})) - G(\boldsymbol{\lambda}^{(n)}))$ (for sufficiently small α_n). Due to this reduction property, in [24] it is proved that if there exists a limit point $\boldsymbol{\lambda}^*$ of $\{\boldsymbol{\lambda}^{(n)}\}$ such that $G(\boldsymbol{\lambda}^*) = 0$, then all the limit points of $\{\boldsymbol{\lambda}^{(n)}\}$ solve $G(\boldsymbol{\lambda}) = 0$. Therefore, the second phase can be exploited for stabilizing our iterative process by forcing the approximation of a stationary point of $p_{\boldsymbol{\lambda}}(g)$. The switching between the two phases is performed with the aim to activate the line-search when the last iterations lay in a region in which $p_{\boldsymbol{\lambda}}(g)$ is concave. By recalling that when $s_n^T y_n > 0$ we may conclude that the function $p_{\boldsymbol{\lambda}}(g)$ is not concave in a set containing $\boldsymbol{\lambda}^{(n)}, \boldsymbol{\lambda}^{(n-1)}$, we activate the second phase when a sequence of N_2 consecutive iterations always provide positive BB step-sizes. Moreover, we ensure the switching to the second phase after a prefixed number $N_1 > N_2$ of iterations. The two phases algorithms can be described as in Algorithm 1.

Algorithm 1 Two Phases (2Ph) Gradient Method

Initialization: choose $\boldsymbol{\lambda}^{(0)}$, α_0 and $\theta, \gamma \in (0,1)$; set flag(N_1,N_2)=0, $n = 1$, $\boldsymbol{\lambda}^{(1)} = \boldsymbol{\lambda}^{(0)} + \alpha_0 G(\boldsymbol{\lambda}^{(0)})$, $gr_0 = f(\boldsymbol{\lambda}^{(0)})$, $gr = f(\boldsymbol{\lambda}^{(1)})$ and an integer $P \geq 1$.

Phase 1: (BB-like Gradient step)

WHILE $\left(\frac{gr}{gr_0} > \tau_g \text{ or } \frac{\|\boldsymbol{\lambda}^{(n)} - \boldsymbol{\lambda}^{(n-1)}\|}{\|\boldsymbol{\lambda}^{(n)}\|} > \tau_\lambda \right)$ and flag(N_1,N_2) = 0

1.1 Choose α_n and update flag(N_1,N_2);

1.2 Gradient Step: $\boldsymbol{\lambda}^{(n+1)} = \boldsymbol{\lambda}^{(n)} + \alpha_n G(\boldsymbol{\lambda}^{(n)})$, $n = n + 1$;

1.3 Set $gr = \max_{0 \leq j \leq \min(n, P-1)} f(\boldsymbol{\lambda}^{(n-j)})$;

ENDWHILE

Phase 2: (Stabilization with line-search)

WHILE $\left(\frac{gr}{gr_0} > \tau_g \text{ or } \frac{\|\boldsymbol{\lambda}^{(n)} - \boldsymbol{\lambda}^{(n-1)}\|}{\|\boldsymbol{\lambda}^{(n)}\|} > \tau_\lambda \right)$

2.1 Choose α_n;

2.2 Line-search:

IF $f\left(\boldsymbol{\lambda}^{(n)} + \alpha_n G(\boldsymbol{\lambda}^{(n)})\right) \leq gr - \gamma \|\alpha_n^2 G(\boldsymbol{\lambda}^{(n)})\|^2$ THEN set $\boldsymbol{\lambda}^{(n+1)} = \boldsymbol{\lambda}^{(n)} + \alpha_n G(\boldsymbol{\lambda}^{(n)})$ and $n = n + 1$;

ELSE IF $f\left(\boldsymbol{\lambda}^{(n)} - \alpha_n \bar{q}_n(\alpha_n)\right) \leq gr - \gamma \|\alpha_n^2 G(\boldsymbol{\lambda}^{(n)})\|^2$ THEN set $\boldsymbol{\lambda}^{(n+1)} = \boldsymbol{\lambda}^{(n)} - \alpha_n \bar{q}_n(\alpha_n)$ and $n = n + 1$;

ELSE set $\alpha_n = \theta \alpha_n$ and go to Step 2.2;

ENDIF

2.3 Set $gr = \max_{0 \leq j \leq \min(n, P-1)} f(\boldsymbol{\lambda}^{(n-j)})$;

ENDWHILE

4. NUMERICAL RESULTS

In order to test our algorithm performances, we firstly randomly generate wavelet coefficients according to the *a priori* law (5) with realistic (for natural images) subband fixed parameters $\lambda_{\mathbf{m}}$. We use Symlets [25] of length 8 over $J = 2$ resolution levels and each subband $\mathbf{m}$ is represented by the triplet (j, l, c) where j is the resolution level index, and $(l, c)_{l \in \{0,1\}, c \in \{0,1\}}$ represents the low/high-pass filtered subbands (the couple $(1, 1)$ thus represents the diagonal coefficients). The size of the generated image is 128×128, it is blurred using a Gaussian kernel A of standard deviation 0.5, and Gaussian noise is added (of variance $\sigma^2 = 25$).

The classical gradient ascent algorithm is launched over 400 iterations using a fixed step-size $\alpha_n = 10^{-4}$ and the estimated parameters are computed as the mean value over the last 50 iterations (as we have oscillations in the estimation (see Fig. 2)). The two-phase algorithm uses these parameters: $P = 3, \theta = 0.5, \gamma = 10^{-4}$, $\alpha_0 = 10^{-4}$; $N_1 = 50$, $N_2 = 10$ for the switching rule and $\tau_g = 10^{-2}, \tau_\lambda = 10^{-7}$ for the stopping rule. Both are initialized by applying a Wiener filter on g. In Fig. 1 and Fig. 2 the behavior of two estimated hyperparameters over the iterations is shown for the two methods. In Tab. 1 we report the number of iterations (it.) and gradient evaluations (Grad.), the relative error in Euclidean norm of the estimated parameters with respect to the theoretical values ($\mathrm{Err}_{th.}$) and the values estimated by the ML approach on the ground truth (Err_{ML}) and the time in seconds (time). All the test are executed in MATLAB on a quad core Intel i7 CPU.

We also considered two well-known images (Mandrill and Barbara) and generated their blurred images by Gaussian kernel A of standard deviation 2, and added Gaussian noise of variance $\sigma^2 = 25$. We applied the gradient methods to obtain estimated hyperparameters and used them to restore the image by means of a Forward-Backward algorithm [6]. In Tab. 2 we report the values of Signal to Noise Ratio (SNR) calculated on the corrupted image (SNR_{init}) and on the reconstruction obtained with the estimated $\boldsymbol{\lambda}$ (SNR_{fin}).

Fig. 1. $\lambda_{\mathbf{m}}$ behavior over iterations ($\mathbf{m} = (1, 1, 0)$)

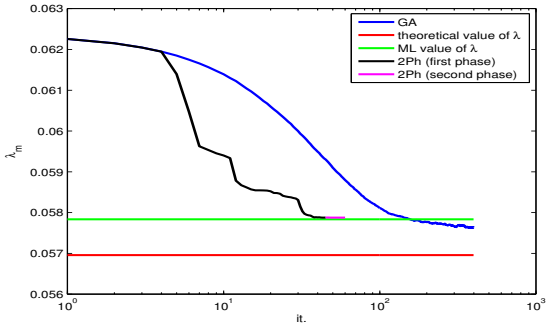

Fig. 2. $\lambda_{\mathbf{m}}$ behavior over iterations ($\mathbf{m} = (1, 1, 1)$)

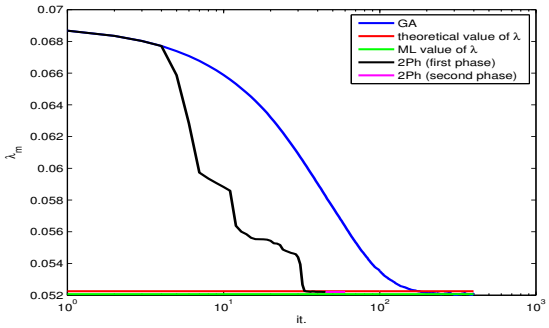

Table 1. Simulated Data

Alg.	It.	Grad.	Err_{ML}	$\mathrm{Err}_{th.}$	time
2Ph	61	90	0.007	0.021	431.8
GA	400	400	0.007	0.020	1920.5

Table 2. Real Data

Problem	Alg.	It.	SNR_{init}	SNR_{fin}
Mandrill 256^2	2Ph	883	12.1616	14.2324
	GA	1000	12.1616	14.0051
Barbara 512^2	2Ph	785	18.5272	19.5316
	GA	1000	18.5272	19.3655

5. CONCLUSIONS

From Figures 1, 2 we can observe remarkable acceleration of the hyperparameter estimation due to the suited choice of the step-size in the first phase and fruitful stabilization in the second phase given by the line-search strategy. Tables 1, 2 confirm the promising convergence rate improvements provided by the 2Ph algorithm with respect to the standard gradient ascent method. These improvements are obtained without losing the accuracy of the proposed ML hyperparameter estimation.

6. REFERENCES

[1] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Applied Math.*, vol. 57, pp. 1413–1457, 2004.

[2] P. L. Combettes and J.-C. Pesquet, "A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery," *IEEE J. Sel. Top. Signal Process.*, vol. 1, no. 4, pp. 564–574, December 2007.

[3] M. A. T. Figueiredo and R. D. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, Aug. 2003.

[4] J. Bect, L. Blanc-Féraud, G. Aubert, and A. Chambolle, "A l^1-unified variational framework for image restoration," in *Proc. Eur. Conf. Comput. Vis.*, T. Pajdla and J. Matas, Eds., Prague, Czech Republic, May 2004, vol. LNCS 3024, pp. 1–13, Springer.

[5] C. Vonesch, S. Ramani, and M. Unser, "Recursive risk estimation for non-linear image deconvolution with a wavelet-domain sparsity constrain," in *Proc. Int. Conf. Image Process.*, San Diego CA, USA, Oct. 12-15 2008, pp. 665–668.

[6] C. Chaux, P. L. Combettes, J.-C. Pesquet, and V. R. Wajs, "A variational formulation for frame based inverse problems," *Inverse Probl.*, vol. 23, pp. 1495–1518, June 2007.

[7] S. Mallat, *A wavelet tour of signal processing*, Academic Press, San Diego, USA, 1998.

[8] N.P. Galatsanos and A.K. Katsaggelos, "Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation," *IEEE Trans. Image Process.*, vol. 1, no. 3, pp. 322–336, Jul. 1992.

[9] M. Bertero, P. Boccacci, G. Talenti, R. Zanella, and L. Zanni, "A discrepancy principle for poisson data," *Inverse Problems*, vol. 26, no. 10, pp. 105004, 2010.

[10] M. Jansen and A. Bultheel, "Multiple wavelet threshold estimation by generalized cross validation for images with correlated noise," *IEEE Trans. Image Process.*, vol. 8, no. 7, pp. 947–953, July 1999.

[11] C. A. Deledalle, S. Vaiter, G. Peyré, J. Fadili, and C. Dossal, "Unbiased risk estimation for sparse analysis regularization," in *Proc. Int. Conf. Image Process.*, Orlando, Florida, USA, Sep. 30 - Oct. 3 2012.

[12] L. Younes, "Parametric inference for imperfectly observed Gibbsian fields," *Probab. Theory Relat. Fields*, vol. 82, pp. 625–645, 1989, Springer-Verlag.

[13] A. Jalobeanu, L. Blanc-Féraud, and J. Zerubia, "Hyperparameter estimation for satellite image restoration using a MCMC maximum likelihood method," *Pattern Recogn.*, vol. 35, no. 2, pp. 341–352, 2002.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39 (series B), pp. 1–38, 1977.

[15] M. M. Ichir and A. Mohammad-Djafari, "Hidden markov models for wavelet-based blind source separation," *IEEE Trans. Image Process.*, vol. 15, no. 7, pp. 1887–1899, July 2006.

[16] D. Leporini and J.-C. Pesquet, "Bayesian wavelet denoising: Besov priors and non-Gaussian noises," *Signal Process.*, vol. 81, no. 1, pp. 55–67, 2001.

[17] D. Geman and C. Yang, "Nonlinear Image recovery with Half-Quadratic Regularization," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 932–946, July 1995.

[18] C. Chaux and L. Blanc-Féraud, "Wavelet-based hyperparameter estimation for solving inverse problems," `http://hal.archives-ouvertes.fr/docs/00/66/88/07/PDF/RR_LBF_CC.pdf`, 2012.

[19] C. Robert, *Discretization and MCMC convergence assessment*, Lecture Notes in Statistics. Springer-Verlag New York Inc., New York, 1998.

[20] J. Barzilai and J. M. Borwein, "Two-point step size gradient methods," *IMA J. Numer. Anal.*, vol. 8, no. 1, pp. 141–148, 1988.

[21] B. Zhou, L. Gao, and Y. H. Dai, "Gradient Methods with Adaptive Step-Sizes," *Comput. Optim. Appl.*, vol. 35, no. 1, pp. 69–86, 2006.

[22] G. Frassoldati, G. Zanghirati, and L. Zanni, "New Adaptive Stepsize Selections in Gradient Methods," *J. Ind. Manag. Optim.*, vol. 4, no. 2, pp. 299–312, 2008.

[23] S Bonettini, R Zanella, and L Zanni, "A scaled gradient projection method for constrained image deblurring," *Inverse Probl.*, vol. 25, no. 1, pp. 015002 (23pp), 2009.

[24] Wanyou Cheng and Zixin Chen, "Nonmonotone spectral method for large-scale symmetric nonlinear equations," *Numer. Algor.*, Apr. 2012.

[25] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF, SIAM Lecture Series, Philadelphia, PA, USA, 1992.