ADA-PT: AN ADAPTIVE PARAMETER TUNING STRATEGY BASED ON THE WEIGHTED STEIN UNBIASED RISK ESTIMATOR

Rita Ammanouil, André Ferrari, Cédric Richard

Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Lab. J.-L. Lagrange, France

ABSTRACT

The performance of iterative algorithms aimed at solving a regularized least squares problem typically depends on the value of some regularization parameter. Tuning the regularization parameter value is a fundamental step necessary to control the strength of the regularization and hence ensure a good performance. We address the problem of finding the optimal regularization parameter in such iterative algorithms. We propose to adaptively adjust the regularization parameter throughout the iterations of the algorithm by minimizing an estimate of the current risk, typically the Weighted Stein unbiased risk estimate (WSURE). We then prove that, for the case of the Tikhonov regularization, the proposed ADAptive Parameter Tuning (ADA-PT) strategy provides a stationary point consistent with the risk minimizer. We illustrate the efficiency of ADA-PT on two image deconvolution problems: one with the Tikhonov regularization and one with the weighted ℓ-1 analysis wavelet regularization.

Index Terms— Regularization parameter tuning, Stein unbiased risk estimate (SURE), Adaptive tuning

1. INTRODUCTION

In many image and signal processing applications there is a need to solve a linear inverse problem where the original signal is degraded by a linear operator and additive Gaussian noise. Most of the algorithms for solving these problems depart from a regularized least squares formulation. These algorithms typically depend on regularization parameters that require fine tuning in order to have satisfying results. More formally, consider the problem of recovering a signal $\boldsymbol{x}_0 \in \mathbb{R}^N$ from a realization $y \in \mathbb{R}^P$ of the normal random vector:

$$\boldsymbol{Y} = \boldsymbol{\mu}_0 + \boldsymbol{W} \quad \text{with} \quad \boldsymbol{\mu}_0 = \boldsymbol{\phi}\boldsymbol{x}_0 \tag{1}$$

where $\boldsymbol{W} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_P)$ is the noise, and $\boldsymbol{\phi} \in \mathbb{R}^{P\times N}$ is a rank deficient linear operator with for example $P < N$. The linear operator $\boldsymbol{\phi}$ entails some loss of information, and the corresponding Least squares inverse problem is ill-posed. To overcome this problem, regularization (ex: Tikhonov, Lasso, ...) is used to promote desirable properties in the solution. Let $(\boldsymbol{y}, \theta) \mapsto \boldsymbol{x}(\boldsymbol{y}, \theta)$ be some recovery mapping, which attempts to approach $\boldsymbol{x}_0$ from a given realization $\boldsymbol{y}$ of $\boldsymbol{Y}$ and which is parametrized by a regularization parameter θ. Following the celebrated regularized least squares framework, $\boldsymbol{x}(\boldsymbol{y}, \theta)$ is defined as:

$$\boldsymbol{x}(\boldsymbol{y}, \theta) = \text{argmin}_{\boldsymbol{x}} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{\phi}\boldsymbol{x}\|^2 + \theta \mathsf{h}(\boldsymbol{x}) \right\}, \tag{2}$$

where the squared term is the data fidelity term, $\mathsf{h}(\boldsymbol{x})$ is the regularization accounting for prior information, and the regularization parameter θ is a scaling factor controlling the strength of the regularization. Choosing the value of θ is a fundamental step in order to ensure the good performance of any iterative algorithm aimed at solving problem (2). A relatively very small value of θ can lead to overfitting and noise amplification, while a relatively large value of θ can make the regularization too strong w.r.t. the data fidelity term and harm the estimation's quality.

Most of the literature on automatically tuning θ relies on optimizing quantitative measures that evaluate the quality of the estimated signal. These measures can be broadly classified as those based on the discrepancy principle [1,2], the L-curve [3,4], generalized cross-validation (GCV) [5], the Stein Unbiased Risk Estimator (SURE) and its weighted version (WSURE) [6, 7]. A straightforward approach for tuning θ is therefore by exhaustive search: the algorithm is tested with various values for θ and the one giving the best quantitative measure is considered to be the optimal one. Among the aforementioned quantitative measures, the WSURE is the most appealing one. The WSURE allows to unbiasedly estimate the weighted mean square error (WMSE) between the real signal and the estimated one solely given the observations. This means that, on average, the optimal value for θ in the WSURE sense is the one that recovers the signal with the smallest WMSE which is often a desired property. Finding the optimal θ is usually done using a grid search [7, 8], or a bisection strategy such as the Golden section [9–11], or a gradient descent [12]. These approaches require running the iterative algorithm various times, with different values for θ, which can be prohibitive when the algorithm is computationally complex and has relatively long running time.

We propose to estimate the optimal regularization parameter in a more computationally efficient way. Toward this goal, we propose to estimate the optimal θ by minimizing the WSURE using a gradient descent scheme similarly to [12]. Nevertheless, we require to evaluate the output of the algorithm only once, and the value of θ is updated at each iteration according to a gradient descent direction such that the WSURE at the current iteration is reduced. The update of θ at each iteration requires evaluating the gradient of the WSURE when the latter is differentiable, or an estimate of the gradient as in [12], when the WSURE is not differentiable. Furthermore, the gradient descent scheme allows setting multiple parameters more easily than grid search and golden section approaches. We refer to this method as ADAptive Parameter Tuning (ADA-PT). As it will be seen, ADA-PT can substantially reduce the computational complexity compared to conventional exhaustive search since it does not require running the iterative algorithm several times. We also prove, that for the Tikhonov regularization, ADA-PT provides a stationary point that is consistent with the risk minimizer. In what follows, we briefly review the concepts underlying the SURE in section 2 and then present our ADA-PT strategy in sections 3 and 4. Finally, in section 5, ADA-PT is used to adjust the regularization parameters for two image deconvolution problems.

This work was partly supported by the Agence Nationale pour la Recherche, France, (MAGELLAN project, ANR-14-CE23-0004-01).

2. OPTIMAL PARAMETER TUNING

Choosing the parameter θ in problem (2) is a challenging and non-trivial task. Ideally, one would like to find the optimal parameter $\theta^\star$ such that $\boldsymbol{\mu}(\boldsymbol{y},\theta^\star)=\boldsymbol{\phi}\boldsymbol{x}(\boldsymbol{y},\theta^\star)$ is as faithful as possible to $\boldsymbol{\mu}_0$. Formally, this can be cast as minimizing the WMSE:

$$\theta^\star=\text{argmin}_\theta\left\{\mathsf{R}\{\boldsymbol{\mu}\}(\boldsymbol{\mu}_0,\theta)=\mathbb{E}_{\boldsymbol{W}}\|\boldsymbol{\mu}(\boldsymbol{Y},\theta)-\boldsymbol{\mu}_0\|^2\right\}. \quad (3)$$

However, one cannot expect to solve this problem given that $\boldsymbol{\mu}_0$ is unknown. In the case of i.i.d. Gaussian noise, a practical approach is to replace problem (3) with:

$$\theta^\star=\text{argmin}_\theta\hat{\mathsf{R}}\{\boldsymbol{\mu}\}(\boldsymbol{y},\theta), \quad (4)$$

where $\hat{\mathsf{R}}\{\boldsymbol{\mu}\}(\boldsymbol{y},\theta)$ is the WSURE estimator of $\mathsf{R}\{\boldsymbol{\mu}\}(\boldsymbol{\mu}_0,\theta)$ that does not require the knowledge of $\boldsymbol{\mu}_0$. The WSURE unbiasedly estimates the WMSE:

$$\mathbb{E}_{\boldsymbol{W}}[\mathsf{WSURE}\{\boldsymbol{\mu}\}(\boldsymbol{Y},\theta)]=\mathsf{R}\{\boldsymbol{\mu}\}(\boldsymbol{\mu}_0,\theta), \quad (5)$$

and it is defined as follows:

$$\mathsf{WSURE}\{\boldsymbol{\mu}\}(\boldsymbol{y},\theta)=\|\boldsymbol{\mu}(\boldsymbol{y},\theta)-\boldsymbol{y}\|^2+2\sigma^2\text{tr}(\partial_1\boldsymbol{\mu}(\boldsymbol{y},\theta))-P\sigma^2,$$

where $\partial_1\boldsymbol{\mu}(\boldsymbol{y},\theta)$ represents the weak Jacobian of $\boldsymbol{\mu}(\boldsymbol{y},\theta)$, the subscript 1 specifies that the Jacobian is w.r.t. the first argument. The concepts underlying the WSURE were also generalized to general non-i.i.d. exponential families in [7].

2.1. Global method for parameter tuning: Gradient descent

When $\theta\mapsto\hat{\mathsf{R}}\{\boldsymbol{\mu}\}(\boldsymbol{y},\theta)$ is sufficiently smooth one can expect to solve problem (4) by a gradient descent scheme:

$$\theta^{(n+1)}=\theta^{(n)}-\epsilon_\theta\partial_2\hat{\mathsf{R}}\{\boldsymbol{\mu}\}(\boldsymbol{y},\theta^{(n)}), \quad (6)$$

where $\partial_2\hat{\mathsf{R}}\{\boldsymbol{\mu}\}(\boldsymbol{\mu}_0,\theta)$ is the gradient of $\hat{\mathsf{R}}\{\boldsymbol{\mu}\}(\boldsymbol{\mu}_0,\theta)$, the subscript 2 specifies that the gradient is with respect to the second argument θ, and $\epsilon_\theta>0$ is an appropriate step size. We refer to this approach as the global method. When $\theta\mapsto\hat{\mathsf{R}}\{\boldsymbol{\mu}\}(\boldsymbol{y},\theta)$ is not smooth, which is often the case when non-smooth regularizers are used in (2), the authors of [12] propose to replace the gradient in (6) with the weak gradient of an approximation of the WSURE based on finite differences and Monte Carlo simulations (WSURE-FDMC). They also prove that this weak gradient, referred to as the Stein Unbiased GrAdient Risk Estimator (SUGAR-FDMC), is an unbiased estimate of the weak gradient of the original WMSE. SUGAR-FDMC is defined as follows:

$$\text{SUGAR}_{\text{FDMC}}\{\boldsymbol{\mu}\}(\boldsymbol{y},\theta)=2\boldsymbol{\mathcal{J}}_{\boldsymbol{x}}(\boldsymbol{y},\theta)^\top\boldsymbol{\phi}^\top(\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{y},\theta))+2\sigma^2\boldsymbol{\mathcal{J}}_{\mathsf{df}}$$

with

$$\boldsymbol{\mathcal{J}}_{\mathsf{df}}=\frac{1}{\epsilon}(\boldsymbol{\mathcal{J}}_{\boldsymbol{x}}(\boldsymbol{y}+\epsilon\boldsymbol{\delta},\theta)-\boldsymbol{\mathcal{J}}_{\boldsymbol{x}}(\boldsymbol{y},\theta))^\top\boldsymbol{\phi}^\top\boldsymbol{\delta} \quad (7)$$

where $\epsilon>0$, $\boldsymbol{\delta}$ is a realization of $\boldsymbol{\Delta}\sim\mathcal{N}(0,\boldsymbol{I}_P)$, and $\boldsymbol{\mathcal{J}}_{\boldsymbol{x}}(\boldsymbol{y},\theta)=\partial_2\boldsymbol{x}(\boldsymbol{y},\theta)$. Let $\mathcal{N}$ be the number of iterations required for convergence, such that the sequence $\theta^{(0)},\theta^{(1)},\dots,\theta^{(\mathcal{N})}$ generated by equation (6) converges to $\theta^\star$ the minimizer of (4). Let $\mathcal{L}$ be the number of iterations required for convergence of the corresponding iterative algorithm, such that the sequence of iterates $\boldsymbol{x}^{(0)}(\boldsymbol{y},\theta),\boldsymbol{x}^{(1)}(\boldsymbol{y},\theta),\dots,\boldsymbol{x}^{(\mathcal{L})}(\boldsymbol{y},\theta)$ converges to $\boldsymbol{x}(\boldsymbol{y},\theta)$ the minimizer of (2). Note that at each iteration of the gradient descent in (6), we need to compute the gradient at $\theta^{(n)}$. Assuming that the computation of the gradient is as expensive as computing $\boldsymbol{x}(\boldsymbol{y},\theta^{(n)})$, i.e. it requires $\mathcal{L}$ iterations, hence the overall complexity of the gradient descent scheme for finding the optimal value $\theta^\star$ is in $O(\mathcal{N}\times\mathcal{L})$. Our motivation in the following section is to reduce the computational complexity required for solving problem (4).

3. ADAPTIVE PARAMETER TUNING (ADA-PT)

We propose to simultaneously estimate $\boldsymbol{x}(\boldsymbol{y},\theta^\star)$ and $\theta^\star$ with a general iterative algorithm ($\mathcal{L}>1$) designed for solving (2). More precisely, we propose to update θ throughout the estimation procedure of $\boldsymbol{x}(\boldsymbol{y},\theta)$. In other words, at each iteration of the iterative algorithm aimed at solving (2), θ is updated by taking 1 gradient descent step towards the minimizer of the WSURE at the current iteration. More formally, we replace the global method in (6) with:

$$\begin{cases}\theta^{(\ell+1)} &= \theta^{(\ell)}-\epsilon_\theta\partial_2\hat{\mathsf{R}}^{(\ell)}\{\boldsymbol{\mu}^{(\ell)}\}(\boldsymbol{y},\theta^{(\ell)}),\\ \boldsymbol{x}^{(\ell+1)} &= \psi(\boldsymbol{x}^{(\ell)},\boldsymbol{y},\theta^{(\ell+1)}),\end{cases} \quad (8)$$

where $\psi(\boldsymbol{x},\boldsymbol{y},\theta)$ represents the output of one iteration of the iterative algorithm aimed at solving (2). We refer to the proposed strategy for simultaneously updating both variables as ADAptive Parameter Tuning (ADA-PT). Note that in contrast with the global method where the complexity is $O(\mathcal{N}\times\mathcal{L})$, the complexity of ADA-PT is only $O(\mathcal{L})$. This is due to the fact that in contrast with the update in (6) which requires computing $\partial_2\hat{\mathsf{R}}\{\boldsymbol{\mu}\}(\boldsymbol{y},\theta^{(n)})$ i.e. the gradient after convergence of the iterative algorithm, the ADA-PT update requires the gradient at the current iteration. To make the ideas clear, algorithm 1 describes how to ADA-PT an iterative algorithm where $\hat{\mathsf{R}}\{\boldsymbol{\mu}\}(\boldsymbol{y},\theta)$ is differentiable. Whereas algorithm 2 describes how to ADA-PT an iterative algorithm when $\hat{\mathsf{R}}\{\boldsymbol{\mu}\}(\boldsymbol{y},\theta)$ is not differentiable, and SUGAR-FDMC is used instead of the gradient.

Algorithm 1: ADA-PT: differentiable case

Inputs observations $\boldsymbol{y}$;
Parameters σ^2, $\phi\in\mathbb{R}^{P\times N}$, $\mathcal{L}$, ϵ_x ;
Initialise $\theta^{(0)}\leftarrow 0$, $\boldsymbol{x}^{(0)}\leftarrow 0$, $\boldsymbol{\mathcal{J}}_{\hat{\mathsf{R}}}^{(0)}\leftarrow 0$;

for ℓ *from 0 to* $\mathcal{L}-1$ **do**
$\theta^{(\ell+1)}=\theta^{(\ell)}-\epsilon_\theta\boldsymbol{\mathcal{J}}_{\hat{\mathsf{R}}}^{(\ell)}$
$\boldsymbol{x}^{(\ell+1)}=\psi(\boldsymbol{x}^{(\ell)},\boldsymbol{y},\theta^{(\ell+1)})$
$\mathcal{D}_x^{(\ell+1)}=\partial_2\psi(\boldsymbol{x}^{(\ell)},\boldsymbol{y},\theta^{(\ell+1)})$
$\boldsymbol{\mathcal{J}}_{\boldsymbol{x}}^{(\ell+1)}=\partial_3\psi(\boldsymbol{x}^{(\ell)},\boldsymbol{y},\theta^{(\ell+1)})$
$\boldsymbol{\mathcal{J}}_{\mathcal{D}_x}^{(\ell+1)}=\partial_3\mathcal{D}_x^{(\ell+1)}(\boldsymbol{x}^{(\ell)},\boldsymbol{y},\theta^{(\ell+1)})$
$\boldsymbol{\mathcal{J}}_{\hat{\mathsf{R}}}^{(\ell+1)}=2(\boldsymbol{\phi}\boldsymbol{x}^{(\ell+1)}-\boldsymbol{y})^\top\boldsymbol{\phi}\boldsymbol{\mathcal{J}}_{\boldsymbol{x}}^{(\ell+1)}+2\sigma^2\text{tr}(\boldsymbol{\phi}\boldsymbol{\mathcal{J}}_{\mathcal{D}_x}^{(\ell+1)})$
end
return $\boldsymbol{x}(\boldsymbol{y},\theta^\star),\partial_2\hat{\mathsf{R}}\{\boldsymbol{\mu}\}(\boldsymbol{y},\theta^\star),\theta^\star\leftarrow\boldsymbol{x}^{(\mathcal{L})},\boldsymbol{\mathcal{J}}_{\hat{\mathsf{R}}}^{(\mathcal{L})},\theta^{(\mathcal{L})}$

Algorithm 2: ADA-PT: non-differentiable case

Inputs observations $\boldsymbol{y}$;
Parameters σ^2, $\phi\in\mathbb{R}^{P\times N}$, $\mathcal{L}$, ϵ_x ;
Initialise $\theta^{(0)}\leftarrow 0$, $\boldsymbol{x}^{(0)}\leftarrow 0$, $\boldsymbol{\mathcal{J}}_{\hat{\mathsf{R}}}^{(0)}\leftarrow 0$;

for ℓ *from 0 to* $\mathcal{L}-1$ **do**
$\theta^{(\ell+1)}=\theta^{(\ell)}-\epsilon_\theta\boldsymbol{\mathcal{J}}_{\hat{\mathsf{R}}}^{(\ell)}$
$\boldsymbol{x}^{(\ell+1)}=\psi(\boldsymbol{x}^{(\ell)},\boldsymbol{y},\theta^{(\ell+1)})$
$\boldsymbol{\mathcal{J}}_1^{(\ell+1)}=\partial_2\psi(\boldsymbol{x}^{(\ell)},\boldsymbol{y}+\epsilon\boldsymbol{\delta},\theta^{(\ell+1)})$
$\boldsymbol{\mathcal{J}}_2^{(\ell+1)}=\partial_2\psi(\boldsymbol{x}^{(\ell)},\boldsymbol{y},\theta^{(\ell+1)})$
$\boldsymbol{\mathcal{J}}_{\mathsf{df}}^{(\ell+1)}=\frac{1}{\epsilon}(\boldsymbol{\mathcal{J}}_1^{(\ell+1)}-\boldsymbol{\mathcal{J}}_2^{(\ell+1)})^\top\boldsymbol{\phi}^\top\boldsymbol{\delta}$
$\boldsymbol{\mathcal{J}}_{\hat{\mathsf{R}}}^{(\ell+1)}=2\boldsymbol{\mathcal{J}}_1^{(\ell+1)^\top}\boldsymbol{\phi}^\top(\boldsymbol{y}-\boldsymbol{\phi}\boldsymbol{x}^{(\ell+1)})+2\sigma^2\boldsymbol{\mathcal{J}}_{\mathsf{df}}^{(\ell+1)}$
end
return $\boldsymbol{x}(\boldsymbol{y},\theta^\star),\partial_2\hat{\mathsf{R}}\{\boldsymbol{\mu}\}(\boldsymbol{y},\theta^\star),\theta^\star\leftarrow\boldsymbol{x}^{(\mathcal{L})},\boldsymbol{\mathcal{J}}_{\hat{\mathsf{R}}}^{(\mathcal{L})},\theta^{(\mathcal{L})}$

4. CASE STUDY: ADA-PT'ED TIKHONOV

To make the ADA-PT strategy more clear, we consider the instructive example where $\mathsf{h}(\boldsymbol{x})$ in (2) is the Tikhonov regularization. More formally, equation (2) becomes:

$$\boldsymbol{x}(\boldsymbol{y},\theta) = \text{argmin}_{\boldsymbol{x}} \left\{ \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\phi}\boldsymbol{x}\|^2 + \frac{\theta}{2}\|\boldsymbol{x}\|^2 \right\}, \tag{9}$$

and the iterative algorithm reduces then to the following update rule:

$$\boldsymbol{x}^{(\ell+1)} = \boldsymbol{x}^{(\ell)} - \epsilon_x \left\{ (\boldsymbol{\phi}^\top\boldsymbol{\phi} + \theta\boldsymbol{I})\boldsymbol{x}^{(\ell)} - \boldsymbol{\phi}^\top\boldsymbol{y} \right\}, \tag{10}$$

where $0 < \epsilon_x < \frac{2}{L}$, L being the Lipschitz constant of the cost function's gradient. The ADA-PT'ed Tikhonov follows the scheme described in algorithm 1 where:

$$\begin{aligned}
\mathcal{D}_x^{(\ell+1)} &= \mathcal{D}_x^{(\ell)} - \epsilon_x((\boldsymbol{\phi}^\top\boldsymbol{\phi} + \theta^{(\ell+1)}\boldsymbol{I})\mathcal{D}_x^{(\ell)} - \boldsymbol{\phi}^\top) \\
\boldsymbol{\mathcal{J}}_{\boldsymbol{x}}^{(\ell+1)} &= \boldsymbol{\mathcal{J}}_{\boldsymbol{x}}^{(\ell)} - \epsilon_x((\boldsymbol{\phi}^\top\boldsymbol{\phi} + \theta^{(\ell+1)}\boldsymbol{I})\boldsymbol{\mathcal{J}}_{\boldsymbol{x}}^{(\ell)} + \boldsymbol{x}^{(\ell)}) \\
\boldsymbol{\mathcal{J}}_{\mathcal{D}_x}^{(\ell+1)} &= \boldsymbol{\mathcal{J}}_{\mathcal{D}_x}^{(\ell)} - \epsilon_x((\boldsymbol{\phi}^\top\boldsymbol{\phi} + \theta^{(\ell+1)}\boldsymbol{I})\boldsymbol{\mathcal{J}}_{\mathcal{D}_x}^{(\ell)} + \mathcal{D}_x^{(\ell)}) \\
\boldsymbol{\mathcal{J}}_{\hat{\mathsf{R}}}^{(\ell+1)} &= 2(\boldsymbol{\phi}\boldsymbol{x}^{(\ell+1)} - \boldsymbol{y})^\top\boldsymbol{\phi}\boldsymbol{\mathcal{J}}_{\boldsymbol{x}}^{(\ell+1)} + 2\sigma^2\text{tr}(\boldsymbol{\phi}\boldsymbol{\mathcal{J}}_{\mathcal{D}_x}^{(\ell+1)})
\end{aligned} \tag{11}$$

Note that a proof of convergence is out of the scope of this study. Nevertheless, in what follows we investigate the stationary points of the proposed iterative algorithm in the case of the Tikhonov regularization. In particular, we show that the estimated θ at the stationary point of the proposed algorithm is consistent with the desired optimal value. We assume that algorithm 4 converges, and that each variable converges to a stationary point. Replacing the iteration numbers (ℓ) and $(\ell+1)$ in the variables subscripts in equations (10) and (11) by infinity (∞), we obtain what follows:

$$\begin{aligned}
\boldsymbol{x}^{(\infty)} &= (\boldsymbol{\phi}^\top\boldsymbol{\phi} + \theta^{(\infty)}\boldsymbol{I})^{-1}\boldsymbol{\phi}^\top\boldsymbol{y} \\
\mathcal{D}_x^{(\infty)} &= (\boldsymbol{\phi}^\top\boldsymbol{\phi} + \theta^{(\infty)}\boldsymbol{I})^{-1}\boldsymbol{\phi}^\top \\
\boldsymbol{\mathcal{J}}_x^{(\infty)} &= (\boldsymbol{\phi}^\top\boldsymbol{\phi} + \theta^{(\infty)}\boldsymbol{I})^{-1}\boldsymbol{x}^{(\infty)} \\
\boldsymbol{\mathcal{J}}_{\mathcal{D}_x}^{(\infty)} &= (\boldsymbol{\phi}^\top\boldsymbol{\phi} + \theta^{(\infty)}\boldsymbol{I})^{-1}\mathcal{D}_x^{(\infty)} \\
0 &= -(\boldsymbol{y} - \boldsymbol{\phi}\boldsymbol{x}^{(\infty)})^\top\boldsymbol{\phi}\boldsymbol{\mathcal{J}}_x^{(\infty)} + \sigma^2\text{tr}(\boldsymbol{\phi}\boldsymbol{\mathcal{J}}_{\mathcal{D}_x}^{(\infty)})
\end{aligned} \tag{12}$$

Note that $\boldsymbol{x}^{(\infty)}$ is indeed the desired solution for $\boldsymbol{x}(\boldsymbol{y},\theta)$. It remains to prove that $\theta^{(\infty)}$ is indeed $\theta^\star$. Substituting the variables by their corresponding expressions in the last equation, using the singular value decomposition (SVD) of $\boldsymbol{\phi}$, and following straightforward calculations we get:

$$\sum_{i=1}^{P} \frac{\sigma_i^4}{(\sigma_i^2 + \theta^{(\infty)})^3}\left(\breve{x}_i^2\theta^{(\infty)} - \sigma^2\right) = 0 \tag{13}$$

where $\breve{\boldsymbol{x}} = \boldsymbol{V}^\top\boldsymbol{x}_0$ and where we have considered the following SVD for $\boldsymbol{\phi} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ with $\boldsymbol{\Sigma}$ a $P \times N$ rectangular diagonal matrix with $\sigma_1, \ldots, \sigma_P$ its diagonal entries. Now, for a fixed value of θ, consider the WMSE of the Tikhonov solution:

$$\mathsf{WMSE} = \mathbb{E}_W\left[\|\boldsymbol{\phi}\boldsymbol{x}_{\text{tik}}(\boldsymbol{Y},\theta) - \boldsymbol{\phi}\boldsymbol{x}_0\|^2\right], \tag{14}$$

where $\boldsymbol{x}_{\text{tik}}(\boldsymbol{y},\theta) = (\boldsymbol{\phi}^\top\boldsymbol{\phi} + \theta\boldsymbol{I})^{-1}\boldsymbol{\phi}^\top\boldsymbol{y}$. Again, using the SVD of $\boldsymbol{\phi}$, and following straighforward calculations, equation (14) yields:

$$\mathsf{WMSE} = \sum_{i=1}^{P} \frac{\sigma_i^2}{(\sigma_i^2+\theta)^2}\left(\sigma_i^2\sigma^2 + \breve{x}_i^2\theta^2\right). \tag{15}$$

The optimal θ denoted as $\theta^\star$ is obtained by setting the gradient of the WMSE to 0 which yields:

$$\sum_{i=1}^{P} \frac{\sigma_i^4}{(\sigma_i^2 + \theta^\star)^3}\left(\breve{x}_i^2\theta^\star - \sigma^2\right) = 0. \tag{16}$$

Table 1. Optimal regularization parameter, WMSE, IPSNR, and execution time obtained with the various methods.

	$\theta^\star$	WMSE	IPSNR	time (sec)
ADA-PT - ℓ_2	3.07	4×10^{-3}	33.16	101
Oracle - ℓ_2	3.10	1×10^{-3}	33.16	32
ADA-PT - ℓ_1	0.06	9×10^{-4}	40.29	392
Oracle - ℓ_1	0.07	8×10^{-4}	41.19	177
LS	–	9×10^{-3}	29.78	31.83

Noting that equations (13) and (16) are consistent, this proves that the stationary point of the ADA-PT'ed Tikhonov iterative algorithm is consistent with the optimal estimate in the sense of the WMSE.

5. EXPERIMENTS

We tested the proposed approach with an image deconvolution problem. We simulated an image of a galaxy with 256×256 pixels, and a realistic radio-telescope PSF similarly to [11]. The initial clean image of the sky was convolved with the PSF and contaminated with white Gaussian noise such as the resulting SNR is equal to 20 dB. We tested ADA-PT with two different optimization problems. The first one was problem (9), i.e. with the Tikhonov regularization. The deconvolution problem with the Tikhonov regularization was solved and ADA-PT'ed according to algorithm 1 and equations (11). The second one was with an ℓ_1 analysis wavelet regularization, i.e. $\mathsf{h}(\boldsymbol{x}) = \|\boldsymbol{D}\boldsymbol{x}\|_1$, where we considered a union of 8 Daubechies wavelet bases for $\boldsymbol{D}$ widely used in the deconvolution literature [13–15]. The deconvolution problem with the ℓ_1 analysis regularization was solved based on the primal-dual algorithm proposed in [16, 17] similarly to the work in [11, 18] and ADA-PT'ed according to algorithm 2.

We tested the two algorithms with different values of θ uniformly spaced in the log space. We considered the intervals $[10^{-1}, 10^{-1.5}]$ and $[10^{-2}, 10^{-0.5}]$ for the ℓ_2 and ℓ_1 regularizations respectively. We then found the optimal value of θ using an oracle approach, i.e. by searching for the one that yielded the smallest WMSE. The first column in Figure 1 shows the two grid search results. We tested the proposed ADA-PT strategy departing from three different initializations for θ. In all cases, θ converged to 3.07 in the case of the ℓ_2 regularization and to 0.06 in the case of the ℓ_1 regularization. The second column of Figure 1 shows the evolution of θ throughout the iterations. The third column in Figure 1 shows the Improvement in the Predicted SNR (IPSNT):

$$\mathsf{IPSNR} = 10\log_{10}\left(\frac{\|\boldsymbol{x}^{(0)} - \boldsymbol{x}_0\|^2}{\|\boldsymbol{x}^{(\mathcal{L})} - \boldsymbol{x}_0\|^2}\right).$$

As a side note, it can be seen that the ℓ_1 regularization deconvolution problem gave better results than the Tikhonov regularization. Figure 2 shows the final estimates of the deconvolved images obtained using the two ADA-PT'ed deconvolution algorithms. Finally, Table 1 summarizes the quantitative results obtained for the: optimal regularization parameter, WMSE, IPSNR, and execution time with the two ADA-PT'ed methods, the oracle, and the LS reference problem i.e. with $\theta = 0$. Note that the ADA-PT'ed version of the problem with the ℓ_2 regularization is approximately three times slower than its oracle version (ran with a fixed value of θ and without estimating the gradient of the WSURE) and the ADA-PT'ed version of the problem with the ℓ_1 regularization is approximately two times

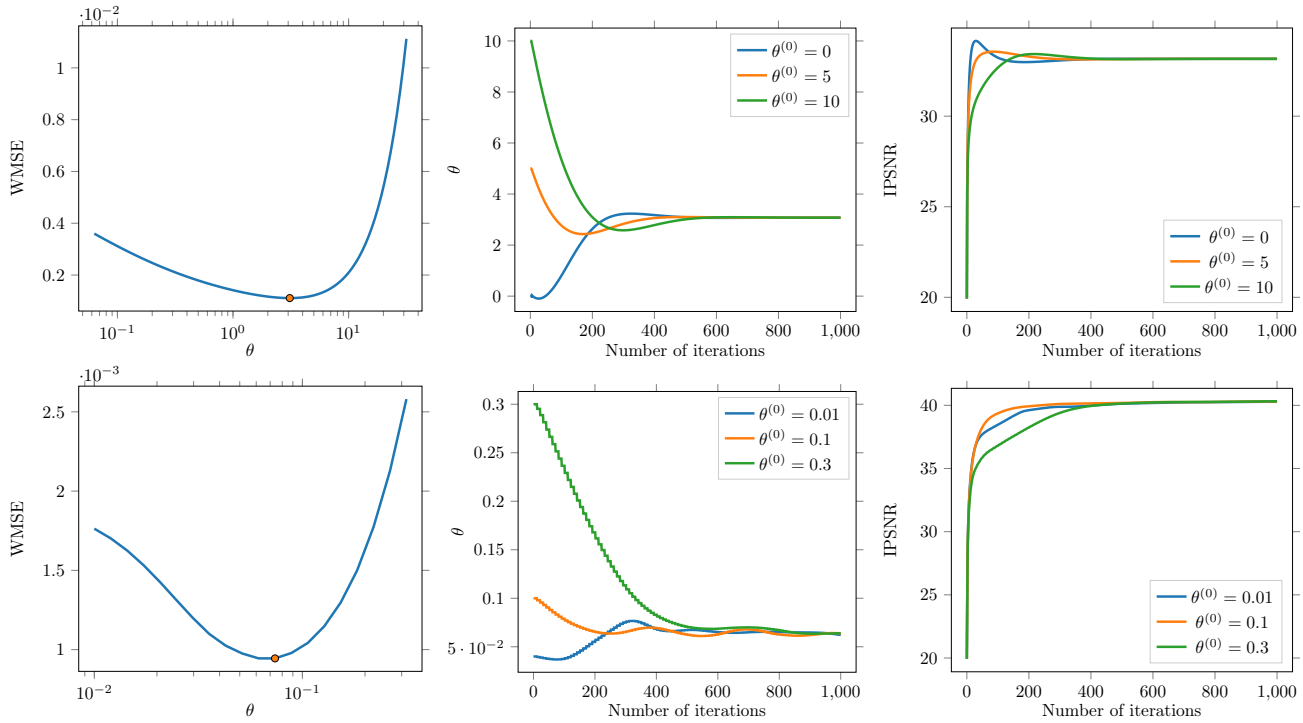

Fig. 1. Column 1: Variation of the WMSE as a function of θ. Column 2: Variation of θ as a function of the iteration number. Column 3: Variation of the IPSNR as a function of the iteration number. First & second rows correspond to the case with the ℓ_2 and ℓ_1 regularizations respectively.

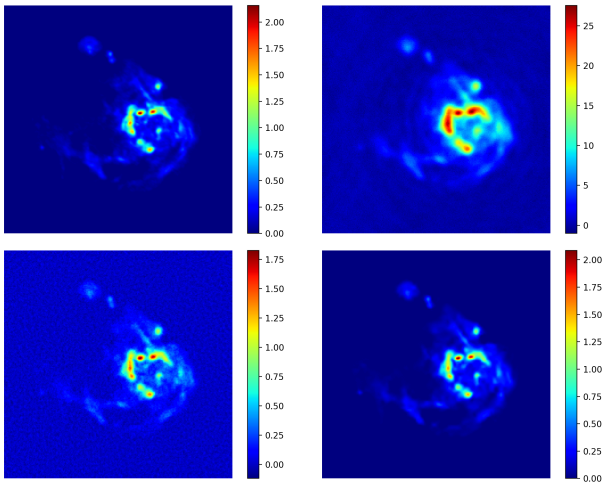

Fig. 2. First row, from left to right: True image of the sky, and corrupted image respectively. Second row, from left to right: deconvolved images of the sky obtained with the ℓ_2 and ℓ_1 regularizations respectively

slower than its oracle version. This is mainly due to the calculations required for estimating the gradient of the WSURE. However, recall that using the global method rather than the ADA-PT'ed method would have taken longer if more than three (resp. two) steps were required in the gradient descent for the ℓ_2 (resp. ℓ_1) regularization.

6. CONCLUSION

We presented an adaptive parameter tuning strategy called ADA-PT that can be used to adapt the regularization parameters in any iterative algorithm designed to solve a regularized least squares problem. The proposed ADA-PT strategy can be used with a broad class of problems, involving smooth and non-smooth regularizers. Future work will be mainly focused on investigating the convergence properties of the proposed approach.

7. REFERENCES

[1] W. C. Karl, "Regularization in image restoration and reconstruction," *Handbook of image video processing*, pp. 183–202, 2005.

[2] V. A. Morozov, "On the solution of functional equations by the method of regularization," in *Soviet Math. Dokl*, vol. 7, no. 1, 1966, pp. 414–417.

[3] T. Regińska, "A regularization parameter in discrete ill-posed problems," *SIAM Journal on Scientific Computing*, vol. 17, no. 3, pp. 740–749, 1996.

[4] P. C. Hansen and D. P. O'Leary, "The use of the l-curve in the

regularization of discrete ill-posed problems," *SIAM Journal on Scientific Computing*, vol. 14, no. 6, pp. 1487–1503, 1993.

[5] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.

[6] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The annals of Statistics*, pp. 1135–1151, 1981.

[7] Y. C. Eldar, "Generalized SURE for exponential families: Applications to regularization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 471–481, 2009.

[8] S. Ramani, T. Blu, and M. Unser, "Blind optimization of algorithm parameters for signal denoising by Monte-Carlo SURE," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 905–908.

[9] R. Giryes, M. Elad, and Y. C. Eldar, "The projected GSURE for automatic parameter tuning in iterative shrinkage methods," *Applied and Computational Harmonic Analysis*, vol. 30, no. 3, pp. 407–422, 2011.

[10] S. Ramani, Z. Liu, J. Rosen, J.-F. Nielsen, and J. A. Fessler, "Regularization parameter selection for nonlinear iterative image restoration and MRI reconstruction using GCV and SURE-based methods," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3659–3672, 2012.

[11] R. Ammanouil, A. Ferrari, R. Flamary, C. Ferrari, and D. Mary, "Multi-frequency image reconstruction for radio-interferometry with self-tuned regularization parameters," *European Signal Processing Conference (EUSIPCO)*, 2017.

[12] C.-A. Deledalle, S. Vaiter, J. Fadili, and G. Peyré, "Stein unbiased gradient estimator of the risk (SUGAR) for multiple parameter selection," *SIAM Journal on Imaging Sciences*, vol. 7, no. 4, pp. 2448–2487, 2014.

[13] R. E. Carrillo, J. D. McEwen, and Y. Wiaux, "Purify: a new approach to radio-interferometric imaging," *Monthly Notices of the Royal Astronomical Society*, vol. 439, no. 4, pp. 3591–3604, 2014.

[14] A. Onose, R. E. Carrillo, A. Repetti *et al.*, "Scalable splitting algorithms for big-data interferometric imaging in the SKA era," *Monthly Notices of the Royal Astronomical Society*, vol. 462, no. 4, pp. 4314–4335, 2016.

[15] H. Garsden, J. JN. Girard *et al.*, "Lofar sparse image reconstruction," *Astronomy & Astrophysics*, vol. 575, p. A90, 2015.

[16] L. Condat, "A generic proximal algorithm for convex optimization—application to total variation minimization," *IEEE Signal Processing Letters*, vol. 21, no. 8, pp. 985–989, 2014.

[17] B. C. Vũ, "A splitting algorithm for dual monotone inclusions involving cocoercive operators," *Advances in Computational Mathematics*, vol. 38, no. 3, pp. 667–681, Nov. 2011.

[18] J. Deguignet, A. Ferrari, D. Mary, and C. Ferrari, "Distributed multi-frequency image reconstruction for radio-interferometry," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 1483–1487.