Iterative regularization via a dual diagonal descent method

Guillaume Garrigos
LCSL, Istituto Italiano di Tecnologia and Massachusetts Institute of Technology
Cambridge, MA 02139, USA
guillaume.garrigos@iit.it

L. Rosasco*
DIBRIS, Università di Genova
16146 Genova, Italy
lrosasco@mit.edu

Silvia Villa
Dipartimento di Matematica, Politecnico di Milano
20133 Milano, Italy
Silvia.villa@polimi.it

Abstract

We consider the problem of designing efficient regularization algorithms allowing to consider large classes of regularizers and data-fit terms. The algorithm we propose is based on a primal-dual diagonal descent method. Our analysis establishes convergence as well as stability results. Theoretical findings are complemented with numerical experiments showing state of the art performances.

1 Introduction

Many machine learning problems consist in the estimation of a quantity of interest based on noisy data. Tackling these problems requires on the one hand to deal with their possible ill-posedness, and on the other hand to devise efficient and stable numerical procedures to compute a solution. Tikhonov regularization is a classical approach to restore well-posedness: a regularized solution is computed by minimizing the sum of a suitable regularizer and an empirical data fidelity term. From a computational perspective it reduces in principle to the solution of a single optimization problem. In practice, however, a regularization parameter needs to be chosen, and hence the solution of multiple optimization problems is typically required (one for each parameter to be tried). This clearly increases the computational complexity of the optimization step. Moreover, computational and estimation aspects are usually considered separately, leading to trade-offs between estimation and computational aspects.

In this paper, we consider iterative regularization techniques as an alternative to the classical Tikhonov approach. In this setting, the key idea is that the regularization is performed implicitly, by early stopping an iterative method applied to an empirical problem. Such techniques are classical in inverse problems [16], and have been already investigated in the machine learning field, showing the same estimation performances of classical regularization methods, while being advantageous from the computational perspective [7, 9]. The main question we discuss in this work is how to derive and analyze iterative regularization schemes for large classes of data fit terms, and regularizers. Indeed, flexibility in the choice of this terms is key for good estimation and has been the subject of much recent work. While extensions to general classes of regularizers have been considered recently [5, 11], we are not aware of any work considering general losses.

We consider a general problem of the form $y = Xw^\dagger$, for a given matrix $X \colon \mathbb{R}^p \to \mathbb{R}^n$, an observation $y \in \mathbb{R}^n$, and a vector $w^\dagger \in \mathbb{R}^p$. Such formulation includes, for instance, regression, feature

*Laboratory for Computational and Statistical Learning, Istituto Italiano di Tecnologia and Massachusetts Institute of Technology

selection, as well as many image/signal processing problems. We assume that the solution of interest $w^\dagger$ is the unique solution of the optimization problem

$$\min\ R(w)\ \text{ s.t. }\ w \in \operatorname*{argmin}_{w' \in \mathbb{R}^p} L(Xw', y), \tag{P}$$

where $L : \mathbb{R}^n \times \mathbb{R}^n \to [0, +\infty[$ is a loss function, convex in its first variable, and $R \colon \mathbb{R}^p \to \mathbb{R}^n$ is a strongly convex regularizer encoding some prior information on the problem at hand. In practice, one does not have access to the ideal measurement y, but only to a noisy version $\widehat{y}$. Here, we assume that the noise is deterministic and such that $\|y - \widehat{y}\| \leq \delta$. The goal is to find a stable estimate of $w^\dagger$, only based on $\widehat{y}$.

In this paper, we consider an iterative regularization scheme, based on a *diagonal process* [6]. The underlying idea is that the family of penalized problems

$$\min R + \lambda^{-1} L(X\cdot, \widehat{y}) \tag{$\widehat{P}_\lambda$}$$

converge to

$$\min\ R(w)\ \text{ s.t. }\ w \in \operatorname*{argmin}_{w' \in \mathbb{R}^p} L(Xw', \widehat{y}), \tag{$\widehat{P}$}$$

for $\lambda \to 0$, and that the algorithm used to solve $(\widehat{P}_\lambda)$ in a Tikhonov approach can be applied in a diagonal fashion, where the parameter λ is decreased along the iterations. However, such a procedure generates a sequence $\widehat{w}_t$ which is convergent to the solution of the noisy problem, which is not our target. Convergence to $w^\dagger$ is obtained by early stopping the iterations, namely by considering a suitable iterate $\widehat{w}_{t_\delta}$ depending on the noise level δ. We show that, for δ going to zero, $\|\widehat{w}_{t_\delta} - w^\dagger\|$ goes to zero.

The paper is organized as follows: in Section 2 we introduce and discuss the iterative regularization procedure. The main results are presented in Section 3, and finally, numerical experiments are provided in Section 4.

2 Algorithm

The proposed iterative regularization procedure is based on a diagonal descent algorithm, applied to the dual of problem (P). We first state the structural assumptions we make on the loss function and the regularizer. We assume that the loss function can be decomposed as an infimal convolution [3] of a distance based function and a strongly convex term

$$(\forall (z, \widehat{y}) \in (\mathbb{R}^n)^2) \quad L(z, \widehat{y}) = G(z - \widehat{y}) \# \Big(J_{\widehat{y}}(z) + \frac{\sigma_L}{2} \|z\|^2 \Big), \tag{1}$$

with $G \colon \mathbb{R}^n \to [0, +\infty]$ and $J_{\widehat{y}} \colon \mathbb{R}^n \to [0, +\infty]$ being convex, proper, and lower semicontinuous functions and $\sigma_L > 0$. Let us remark that the previous assumption is satisfied whenever the loss function is distance based. The fact that the loss can be written as an infimal convolution is not restrictive, since this decomposition is trivially true for every convex function. The explicit infimal convolution decomposition is useful since our algorithm works in the dual space, and the inf-convolution is transformed in a sum by duality. The regularizer $R \colon \mathbb{R}^p \to [0, +\infty]$ is strongly convex, i.e. there exists a convex, proper, and lower semicontinuous function $F \colon \mathbb{R}^p \to [0, +\infty]$ and $\sigma_R > 0$ such that

$$R = F + \frac{\sigma_R}{2} \|\cdot\|^2. \tag{2}$$

The algorithm we consider is a first order method. It consists in matrix vector multiplications and three implicit steps corresponding to the computation of proximity operators, each one activating independently a single component of the problem: R, $J_{\widehat{y}}$, and G respectively. We recall that, given a convex, proper, and lower semicontinuous function $f \colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the proximity operator is defined as, for every $w \in \mathbb{R}^p$, $\operatorname{prox}_f(w) = \operatorname{argmin}_{v \in \mathbb{R}^p} \left\{ f(v) + \frac{1}{2}\|v - w\|^2 \right\}$, and is available in closed form for many functions, see [14] for a list. Each step of the forthcoming (3-D) algorithm can be seen as a step of a proximal gradient algorithm applied to the dual of the regularized problem $R + \lambda_t^{-1} L(X\cdot, \widehat{y})$ (see [13]). This class of methods became very popular in the last decade to solve sparsity based regularized problems [14], and have been applied in a diagonal fashion in the primal setting to solve hierarchial problems as (P) [10, 19, 4, 1, 23, 2, 12, 20, 15]. However, we are not aware of any convergence result, neither of any regularization theorem, for the algorithm applied to the dual formulation, as we do in this paper.

Diagonal Dual Descent (3-D)

Let $(\lambda_t)_{t\in\mathbb{N}}$ be a sequence in $]0,+\infty[$ decreasing to 0, let $L = \|X\|^2/\sigma_R + \lambda_0/\sigma_L$, and $\tau \in]0, 1/L]$. Let $\widehat{z}_0 \in \mathbb{R}^p$, and for all $t \in \mathbb{N}$, let

$$\text{(3D)} \quad \left| \begin{array}{l} \widehat{w}_t = \text{prox}_{\sigma_R^{-1}F}(-\sigma_R^{-1}X^*\widehat{z}_t) \\ \widehat{v}_{t+1} = \widehat{z}_t + \tau X\widehat{w}_t - \tau \text{prox}_{\sigma_L^{-1}J_{\widehat{y}}}(\sigma_L^{-1}\lambda_t\widehat{z}_t) \\ \widehat{z}_{t+1} = \widehat{v}_{t+1} - \tau\left(\widehat{y} + \text{prox}_{(\tau\lambda_t)^{-1}G}\left(\tau^{-1}\widehat{w}_{t+1} - \widehat{y}\right)\right) \end{array} \right.$$

Remark 1 (Warm-restart) *Warm-restart, or continuation method, is a popular heuristic commonly used to speed up the computations of solutions of the regularized problem $(\hat{P}_\lambda)$ for different values of λ [17, 8]. It is easy to see that this strategy generates a sequence $(\widehat{w}_t)_{t\in\mathbb{N}}$ corresponding to the application of the (3-D) algorithm, for a piecewise constant decreasing sequence $(\lambda_t)_{t\in\mathbb{N}}$.*

Remark 2 (Diagonal Landweber iteration) *For the case $L(Xw,\widehat{y}) = (1/2)\|Xw - \widehat{y}\|^2$, $G = J_{\widehat{y}} = 0$, and $\sigma_R = 1$, the (3-D) algorithm reduces to*

$$\widehat{w}_{t+1} = \widehat{w}_t - \tau X^T(X\widehat{w}_t - \widehat{y}) - \tau\lambda_t\widehat{w}_t. \tag{3}$$

This algorithm can be seen as the gradient descent method applied to the function $(1/2)\|Xw - \widehat{y}\|^2 + (\lambda/2)\|w\|^2$, and considering $\lambda = \lambda_t$, changing at each iteration. It has been mainly studied for nonlinear inverse problems, and is also known as modified Landweber iteration [22, 18].

Regularization properties of an early stopped iteration of (3-D) algorithm are known only in the setting of Remark 2. In the following section we show that (3-D) algorithm exhibits a regularizing behavior also for general losses and general regularizers.

3 Main results

Before stating our main results, we make some assumptions on the regularizer R and the loss function L, which are tipically verified in practice:

(H): R is continuous at $w^\dagger$, and $L(\cdot, y)$ admits y as its unique minimizer, and it is locally p-conditioned for some $p \in [1,+\infty[$: $\exists(\gamma, r) \in (]0,+\infty[)^2$ s. t. $\|z-y\| \le r \Rightarrow \gamma\|z-y\|^p \le L(z,y)$.

Remark 3 *Determining the conditioning of a loss function is usually very easy when it has the expression $L(z,y) = \sum_{i=1}^n \ell(z^{(i)}, y^{(i)})$, with $\ell : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$. Indeed, it reduces to the analysis of the conditioning of the real valued functions $\{\ell(\cdot, y_i)\}_{i=1,\dots,n}$.*

Next we present the main results. First, we prove convergence of (3-D) algorithm, in the noiseless case. In order to derive the convergence of w_t towards $w^\dagger$, we will impose a suitable decay condition on λ_t, which is in turn dictated by the geometry of the loss function via its conditioning order p. Then, Theorem 2 establishes the iterative regularization property of the (3-D) method.

Theorem 1 *Assume that **(H)** holds. Let $(z_t, w_t)_{t\in\mathbb{N}}$ be generated by the (3-D) method with exact data $\widehat{y} = y$, and by taking[1] $(\lambda_t)_{t\in\mathbb{N}} \in \ell^{1/(p-1)}(\mathbb{N})$. Then z_t converges to a solution of the dual problem of (P), and w_t converges to $w^\dagger$, with $\|w_t - w^\dagger\| = o\left(t^{-1/2}\right)$.*

Theorem 2 *Assume that **(H)** holds. Let $(\widehat{z}_t, \widehat{w}_t)_{t\in\mathbb{N}}$ be generated by the (3-D) method with noisy data $\|\widehat{y} - y\| \le \delta$, and by taking $\lambda_t = \lambda_0/(t+1)^\beta$, for some $\beta \in]p-1, +\infty[$. Then there exists an early stopping rule $t(\delta) \sim \delta^{-2/3}$ such that $\|\widehat{w}_{t(\delta)} - w^\dagger\| = O\left(\delta^{1/3}\right)$ when $\delta \to 0$.*

Discussion and idea of the proof. The bound in Theorem 2 are derived by optimizing a stability plus regularization bound:

$$\|\widehat{w}_t - w^\dagger\| \le \|\widehat{w}_t - w_t\| + \|w_t - w^\dagger\|, \tag{4}$$

[1] When $p = 1$, the notation $1/0$ stands here for ∞, which means that no particular assumption is required on the sequence, since it is assumed in (3-D) that $\lambda_n \downarrow 0$.

where $(w_t)_{t\in\mathbb{N}}$ is an auxiliary sequence obtained by applying the same diagonal procedure to the problem with the ideal data. The first term on the right hand side depends on the stability of the algorithm, and is due to the noise in the data. The key is to show that it increases in a controlled manner when the level of noise δ and the number of iterations increase. The second term is an optimization error. Theorem 1 shows that this quantity goes to zero as the number of iterations grows. The choice of the *early stopping* time t_δ is then determined by balancing the stability and the optimization error, and depends on the noise level.

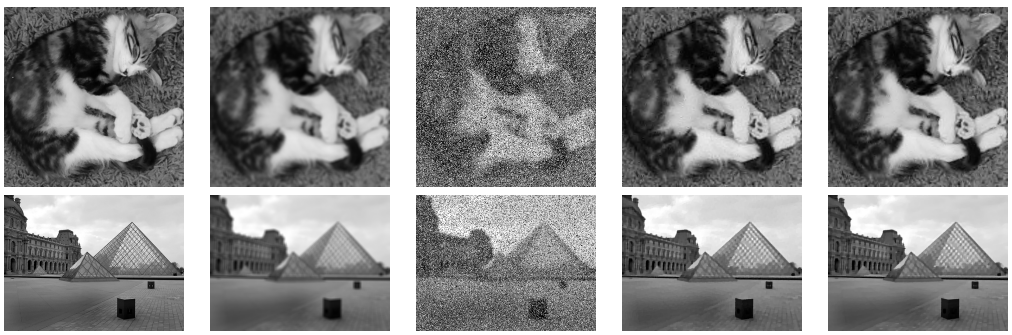
This regularization result is the first to hold for such general pair of regularizer/loss function. Note that our rates $O(\delta^{1/3})$, when specialized to the setting of Remark 2, are not optimal. Indeed it is known that the diagonal Landweber algorithm is of the same order of the Tikhonov regularization, which is $O(\delta^{1/2})$ [21, Theorem 6.5]. It is an open question to know whether our result can be improved, or if it is not possible to achieve optimal rates in such a general setting.

4 Numerical experiments

We performed some numerical experiments using the (3-D) algorithm for image denoising and deblurring. We consider problems of the form (P), involving a data fidelity term selected according to the nature of the noise, and a regularizer promoting a desired feature. The linear operator X is a blurring operator defined through a Gaussian kernel of size 9×9 and variance 10. We compare the results obtained using two approaches, corresponding to two different choices of the sequence $(\lambda_t)_{t\in\mathbb{N}}$ in the (3-D) algorithm: the first is an a priori choice, which defines a regularization method that we call hereafter *diagonal Tikhonov*. In that case we chose $\lambda_t = 10/(t+1)^{2/3}$. The second approach, called *warm restart Tikhonov*, corresponds to an online choice of $(\lambda_t)_{t\in\mathbb{N}}$, which is described in Remark 1. In that case, $(\lambda_t)_{t\in\mathbb{N}}$ is taken piecewise constant among $\{\Lambda_1, \dots, \Lambda_{20}\} \subset [0.1, 10]$, and its decay is determined by a stopping rule. The latter can be considered as a benchmark, since it is one of the most efficient ways to approximate the Tikhonov regularization path [8].

We solve two problems and compare the accuracy of these two methods by evaluating $\|w_t - w^\dagger\|$. We call *early stop* the iterate minimizing this distance. Both problems concerns a grayscale image (*Kitten* and *Louvre*), which is blurred with X and corrupted by a salt and pepper noise, that we recover by solving (P) with a ℓ^1 loss function $L(z, \widehat{y}) = \|z - \widehat{y}\|_1$. For *Kitten* we use $R(w) = \|Ww\|_1 + \frac{1}{2}\|w\|^2$ as a regularizer, where W is a Daubechies wavelet transform. The early stop is detected at $t = 510$ and $t = 1131$ for the diagonal Tikhonov and the warm restart Tikhonov, respectively. For *Louvre* we use $R(w) = \|w\|_{TV} + \frac{1}{2}\|w\|^2$ as a regularizer, where $\|\cdot\|_{TV}$ is the total variation norm. The early stop is detected at $t = 178$ and $t = 2760$ for the diagonal Tikhonov and the warm restart Tikhonov, respectively. The reconstructed images are displayed in Table 1. As can be seen by visual inspection, the results achieved by the diagonal Tikhonov method and the warm restart Tikhonov are qualitatively comparable.

Table 1: From left to right: original image, blurred image without noise, blurred image with noise, image reconstructed with diagonal Tikhonov, image reconstructed with warm restart Tikhonov.



References

[1] P. Alart and B. Lemaire, *Penalization in non-classical convex programming via variational convergence*, Mathematical Programming, **51**, pp. 307–331, 1991.

[2] F. Alvarez and R. Cominetti, *Primal and dual convergence of a proximal point exponential penalty method for linear programming*, Mathematical Programming, **93**, pp. 87–96, 2002.

[3] H. Attouch, G. Buttazzo and G. Michaille, *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization*, MPS-SIAM Series on Optimization, 2006.

[4] A. Auslender, J.-P. Crouzeix and P. Fedit, *Penalty-proximal methods in convex programming*, Journal of Optimization Theory and Applications, **55**, pp. 1–21, 1987.

[5] M. Bachmayr and M. Burger, *Iterative total variation schemes for nonlinear inverse problems*, Inverse Problems, **25**, pp. 105004, 2009.

[6] M. A. Bahraoui and B. Lemaire, *Convergence of diagonally stationary sequences in convex optimization*, Set-Valued Analysis, **2**, pp. 49–61, 1994.

[7] F. Bauer, S. Pereverzev, and L. Rosasco, *On regularization algorithms in learning theory*, J. Complexity, **23**, pp.52–72, 2007.

[8] S. Becker, J. Bobin, and E. Candès, NESTA: A Fast and Accurate First-Order Method for Sparse Recovery, SIAM Journal on Imaging Sciences, **4**, pp. 1–39, 2011.

[9] G.Blanchard and N.Krämer, *Optimal learning rates for kernel conjugate gradient regression*. In Advances in Neural Inf. Proc. Systems (NIPS), pp. 226–234, 2010.

[10] R. Boyer, *Quelques algorithmes diagonaux en optimisation convexe*, Ph.D., Université de Provence, 1974.

[11] M. Burger, E. Resmerita, and L. He, Error estimation for Bregman iterations and inverse scale space methods in image restoration, Computing, **81**, pp. 109-135, 2007.

[12] A. Cabot, *Proximal Point Algorithm Controlled by a Slowly Vanishing Term: Applications to Hierarchical Minimization*, SIAM Journal on Optimization, **15**, pp. 555–572, 2005.

[13] P. L. Combettes, D. Dũng, and B. C. Vũ, *Dualization of signal recovery problems*, Set-Valued and Variational Analysis, **18**, pp. 373–404, 2010.

[14] P. L. Combettes and J.-C. Pesquet, *Proximal splitting methods in signal processing*, in Fixed-point algorithms for inverse problems in science and engineering, pp. 185–212, Springer, New York, 2011.

[15] M.-O. Czarnecki, N. Noun, and J. Peypouquet, *Splitting forward-backward penalty scheme for constrained variational problems*, preprint on arXiv:1408.0974, published on Aug 2014.

[16] H. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, 1996.

[17] E. Hale, W. Yin, and Y. Zhang, *Fixed-Point Continuation for ℓ_1-Minimization: Methodology and Convergence*, SIAM Journal on Optimization, **19**, pp.1107–1130, 2008.

[18] B. Kaltenbacher, A. Neubauer and O. Scherzer, *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*, De Gruyter, Berlin, Boston, 2008.

[19] A. A. Kaplan, *On Convex Programming with Internal Regularization*, Soviet Mathematics, Doklady Akademii Nauk, **19**, pp. 795–799, 1975.

[20] J. Peypouquet, *Coupling the Gradient Method with a General Exterior Penalization Scheme for Convex Minimization*, Journal of Optimization Theory and Applications, **153**, pp. 123–138, 2011.

[21] R. Ramlau, *TIGRA – an iterative algorithm for regularizing nonlinear ill-posed problems*, Inverse Problems **19**, pp. 433–465, 2003.

[22] O. Scherzer, *A Modified Landweber Iteration for Solving Parameter Estimation Problems*, Applied Mathematics and Optimization, **38**, pp. 45–68, 1998.

[23] P. Tossings, *The perturbed Tikhonov's algorithm and some of its applications*, ESAIM: Mathematical Modelling and Numerical Analysis, **28**, pp. 189–221, 1994.