

# A New Steplength Selection for Scaled Gradient Methods with Application to Image Deblurring

Federica Porta · Marco Prato · Luca Zanni

Received: 30 May 2014 / Revised: 19 January 2015 / Accepted: 22 January 2015

Published online: 30 January 2015

© Springer Science+Business Media New York 2015

**Abstract** Gradient methods are frequently used in large scale image deblurring problems since they avoid the onerous computation of the Hessian matrix of the objective function. Second order information is typically sought by a clever choice of the steplength parameter defining the descent direction, as in the case of the well-known Barzilai and Borwein rules. In a recent paper, a strategy for the steplength selection approximating the inverse of some eigenvalues of the Hessian matrix has been proposed for gradient methods applied to unconstrained minimization problems. In the quadratic case, this approach is based on a Lanczos process applied every  $m$  iterations to the matrix of the gradients computed in the previous  $m$  iterations, but the idea can be extended to a general objective function. In this paper we extend this rule to the case of scaled gradient projection methods applied to constrained minimization problems, and we test the effectiveness of the proposed strategy in image deblurring problems in both the presence and the absence of an explicit edge-preserving regularization term.

**Keywords** Image deconvolution · Constrained optimization · Scaled gradient projection methods · Ritz values

**Mathematics Subject Classification** 65K05 · 65R32 · 68U10 · 90C06

## 1 Problem Formulation

The image formation process is an inverse problem that can be modeled as the following linear system

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} + \boldsymbol{\eta}, \quad (1)$$

---

F. Porta · M. Prato (✉) · L. Zanni

Dipartimento di Scienze Fisiche, Informatiche e Matematiche, Università degli Studi di Modena e Reggio Emilia, Via Campi 213/b, 41125 Modena, Italy  
e-mail: marco.prato@unimore.it

where  $\mathbf{y} \in \mathbb{R}^{n^2}$  is the non-negative observed data,  $\mathbf{x} \in \mathbb{R}^{n^2}$  represents an ideal, undistorted image to be recovered,  $A \in \mathbb{R}^{n^2 \times n^2}$  is a typically ill-conditioned matrix describing the blurring effect,  $\mathbf{b} \in \mathbb{R}^{n^2}$  is a known non-negative background radiation and  $\boldsymbol{\eta} \in \mathbb{R}^{n^2}$  is the noise corrupting the data. A typical assumption for the matrix  $A$  is that it has non-negative elements and each row and column has at least one positive entry. Because of the ill-conditioning affecting the problem and the presence of noise on the measured data, a trivial approach that seeks the solution of (1) is in general not successful; thus, alternative strategies must be exploited. Variational approaches to image restoration [5, 43] suggest to recover the unknown object through iterative schemes suited for the following constrained minimization problem

$$\min_{\mathbf{x} \geq \mathbf{0}} J_0(\mathbf{x}) \quad (2)$$

where  $J_0$  is a continuously differentiable convex function measuring the difference between the model and the data. The definition of the function  $J_0$  depends on the noise type introduced by the acquisition system. Particularly, in the case of additive white Gaussian noise the cost function is characterized by a least squares distance of the form

$$J_0(\mathbf{x}) = J_0^{LS}(\mathbf{x}) = \frac{1}{2} \|\mathbf{A}\mathbf{x} + \mathbf{b} - \mathbf{y}\|^2, \quad (3)$$

while, when the data are affected by Poisson noise, the so-called Kullback–Leibler (KL) divergence is used:

$$J_0(\mathbf{x}) = J_0^{KL}(\mathbf{x}) = \sum_{i=1}^{n^2} \left\{ y_i \ln \frac{y_i}{(\mathbf{A}\mathbf{x} + \mathbf{b})_i} + (\mathbf{A}\mathbf{x} + \mathbf{b})_i - y_i \right\}, \quad (4)$$

where we assume that  $0 \ln 0 = 0$  and  $(\mathbf{A}\mathbf{x} + \mathbf{b})_i > 0, \forall i = 1, \dots, n^2$ . In both cases, taking into account also the assumptions on  $A$ , we may observe that  $J_0$  is non-negative, convex and coercive on the non-negative orthant, which means that problem (2) has global solutions. Moreover, if the equation  $\mathbf{A}\mathbf{x} = \mathbf{0}$  has only the solution  $\mathbf{x} = \mathbf{0}$ , then  $J_0^{LS}$  is strictly convex, while the same conclusion holds for  $J_0^{KL}$  if the additional condition  $y_i > 0, \forall i = 1, \dots, n^2$ , is satisfied [4, 15]. In these settings, the strict convexity of  $J_0$  implies that the solution of (2) is unique.

Due to the ill-posedness of the image restoration problem, one is not interested in computing the minimum points of  $J_0$  in (3) or (4) because the exact solution of (2) does not provide a sensible estimate of the unknown image. For this reason, iterative minimization methods are usually exploited to obtain acceptable solutions by arresting the algorithm before convergence through some stopping criteria, as the classic Morozov's discrepancy principle in the case of Gaussian noise (see e.g. [29]) or some recently proposed strategies for Poisson data [2, 4, 16].

Another technique to tackle to this problem requires to exactly solve the following optimization problem

$$\min_{\mathbf{x} \geq \mathbf{0}} J_0(\mathbf{x}) + \beta J_R(\mathbf{x}), \quad (5)$$

where  $J_R$  is a regularization term adding a priori information on the solution and  $\beta$  is a positive parameter balancing the role of the two objective function components  $J_0$  and  $J_R$ . A frequently used function for the regularization term is a smooth approximation of the total variation, also known in the literature as *hypersurface potential* (HS), defined as [1, 4]

$$J_R(\mathbf{x}) = J_R^{HS}(\mathbf{x}) = \sum_{i,j=1}^n \sqrt{((\mathcal{D}\mathbf{x})_{i,j})_1^2 + ((\mathcal{D}\mathbf{x})_{i,j})_2^2 + \delta^2}, \quad (6)$$

where the discrete gradient operator  $\mathcal{D} : \mathbb{R}^{n^2} \rightarrow \mathbb{R}^{2n^2}$  is set through the standard finite difference with periodic boundary conditions

$$(\mathcal{D}\mathbf{x})_{i,j} = \begin{pmatrix} ((\mathcal{D}\mathbf{x})_{i,j})_1 \\ ((\mathcal{D}\mathbf{x})_{i,j})_2 \end{pmatrix} = \begin{pmatrix} x_{i+1,j} - x_{i,j} \\ x_{i,j+1} - x_{i,j} \end{pmatrix}, \quad x_{n+1,j} = x_{1,j}, \quad x_{i,n+1} = x_{i,1}. \quad (7)$$

When  $J_R = J_R^{HS}$  and  $J_0$  is one of the two considered cost functions, the objective function in (5) is non-negative, strictly convex and coercive on the non-negative orthant [4]. It follows that problem (5) has a unique solution.

Both formulations of the imaging problem require an effective optimization method able to provide a meaningful solution in a reasonable time. Among all possible choices, first-order methods are particularly suited to deal with this kind of problems for several reasons. First, due to the large size of the images (which becomes a crucial issue especially in 3D applications), the handling of the Hessian matrix is an impractical task. Then, first-order methods are used to quickly achieve solutions with low/medium accuracy, which is a general requirement in imaging problems. Finally, when the optimization scheme is used as iterative regularization method to minimize the cost function (2), an excessively fast convergence makes the automatic choice of the stopping iteration a crucial issue, since a difference of few iterations from the one providing the best reconstruction can lead to substantial differences in the final images.

In this paper we extend to the case of a general scaled gradient projection method [6, 8, 15] a steplength selection rule recently proposed by Fletcher [25] in the unconstrained optimization framework and we test its effectiveness in image deblurring problems. This rule is based on the estimate of some eigenvalues of the Hessian matrix which, for quadratic problems, can be achieved by means of a Lanczos process applied to a certain number of consecutive gradients. Since the scheme depends only on these stored gradients, it can be generalized to nonquadratic objective functions, showing very competitive results in several benchmark problems with respect to other first-order and quasi-Newton methods. The extension to scaled gradient projection methods applied to non-negatively constrained problems requires a generalization of the matrix with the last gradients accounting for the presence of both the scaling matrix multiplying the gradient and the projection on the non-negative orthant. The resulting scheme consists in the storage of a set of scaled gradients (instead of the usual ones) in which some components of the gradients themselves are put equal to zero. Our numerical experiments on the non-negative minimization of the LS distance and the KL divergence show that the proposed approach is able to compete with standard gradient methods and other recently proposed schemes, providing in some cases good reconstructions with a significantly lower number of iterations.

The plan of the paper is the following: in Sect. 2 we recall the features of a scaled gradient projection method and, in particular, of the scaling matrix multiplying the gradient. In Sect. 3 we focus the analysis on the choice of the steplength parameter and we describe state-of-the-art strategies and our proposed rule. In Sect. 4 some numerical experiments on small quadratic programming (QP) and image deblurring least-squares problems are presented, while in Sect. 5 we address the image deblurring problem with data perturbed with Poisson noise also by adding an edge-preserving regularization term in the objective function. Some ideas on a possible generalization of the proposed rule to different constraints are provided in Sect. 6,

together with a numerical test on the Rudin–Osher–Fatemi model [40]. Our conclusions are given in Sect. 7.

## 2 Scaled Gradient Projection Methods

A general scaled gradient projection (SGP) method [15] for the solution of

$$\min_{\mathbf{x} \geq 0} J(\mathbf{x}), \quad (8)$$

with  $J$  differentiable function, is an iterative algorithm whose  $(k+1)$ th iteration is defined by

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda_k \mathbf{d}^{(k)} = \mathbf{x}^{(k)} + \lambda_k \left( \mathbb{P}_{+, D_k^{-1}}(\mathbf{x}^{(k)} - \alpha_k D_k \mathbf{g}^{(k)}) - \mathbf{x}^{(k)} \right), \quad (9)$$

where

- $\mathbf{x}^{(k)} \geq \mathbf{0}$ ;
- $\mathbf{g}^{(k)} = \nabla J(\mathbf{x}^{(k)})$  is the gradient of the objective function at iteration  $\mathbf{x}^{(k)}$ ;
- $\lambda_k \in (0, 1]$  is a linesearch parameter ensuring a sufficient decrease of the objective function along the descent direction  $\mathbf{d}^{(k)}$ , e.g. by means of an Armijo rule [6];
- $\alpha_k$  is a positive steplength chosen in a fixed range  $[\alpha_{\min}, \alpha_{\max}]$ , with  $0 < \alpha_{\min} < \alpha_{\max}$ ;
- $D_k$  is a symmetric and positive definite scaling matrix with eigenvalues lying in a fixed positive interval  $[L_1, L_2]$ ;
- $\mathbb{P}_{+, D}(\cdot)$  denotes the projection operator onto the non-negative orthant with respect to the norm induced by the matrix  $D$ :

$$\mathbb{P}_{+, D}(\mathbf{x}) = \arg \min_{\mathbf{y} \geq 0} \|\mathbf{y} - \mathbf{x}\|_D = \arg \min_{\mathbf{y} \geq 0} \frac{1}{2} \mathbf{y}^T D \mathbf{y} - \mathbf{y}^T D \mathbf{x}.$$

The boundedness conditions on the steplengths and the eigenvalues of the scaling matrices are necessary to prove the convergence result for this method (see [15, Theorem 2.1]), that we report for completeness.

**Theorem 1** *Let  $\mathbf{x}^{(0)} \geq \mathbf{0}$  and assume that the level set  $\Omega_0 = \{\mathbf{x} \geq \mathbf{0} : J(\mathbf{x}) \leq J(\mathbf{x}^{(0)})\}$  is bounded. Every accumulation point of the sequence  $\{\mathbf{x}^{(k)}\}_{k \in \mathbb{N}}$  generated by the SGP algorithm is a stationary point of (8).*

When SGP is applied to the imaging minimization problem (2) or (5), the coercivity of the objective function on the non-negative orthant assures that  $\Omega_0$  is bounded for any  $\mathbf{x}^{(0)} \geq \mathbf{0}$ , therefore the sequence generated by SGP is bounded and admits limit points; the uniqueness of the limit point is ensured when the objective function is strictly convex.

In imaging applications, the scaling matrix  $D_k$  is usually chosen according to the cost function  $J_0$  and the regularization term  $J_R$ . Following the approach proposed in [32, 33], if  $\nabla J_0$  and  $\nabla J_R$  can be decomposed in the form

$$-\nabla J_0(\mathbf{x}) = U_0(\mathbf{x}) - V_0(\mathbf{x}); \quad -\nabla J_R(\mathbf{x}) = U_R(\mathbf{x}) - V_R(\mathbf{x}), \quad (10)$$

with  $U_0, U_R \geq 0$  and  $V_0, V_R > 0$ , then a possible scaling matrix is given by

$$D_k = \max \left( L_1, \min \left( L_2, \text{diag} \left( \frac{\mathbf{x}^{(k)}}{V_0(\mathbf{x}^{(k)}) + \beta V_R(\mathbf{x}^{(k)})} \right) \right) \right), \quad L_1 \leq L_2, \quad (11)$$

where  $\mathbf{v}/\mathbf{w}$  is the componentwise ratio between  $\mathbf{v}$  and  $\mathbf{w}$ . We remark that the choice of a diagonal scaling matrix is preferable since in this case the projection on the non-negative orthant is straightforward and does not require the solution of a further quadratic subproblem at each iteration. Since in general the imaging matrix  $A$  has non-negative entries, the gradients of the cost functions in (3) and (4) satisfy the decomposition in (10)

$$-\nabla J_0^{LS}(\mathbf{x}) = \underbrace{A^T \mathbf{y}}_{U_0^{LS}(\mathbf{x})} - \underbrace{A^T (A\mathbf{x} + \mathbf{b})}_{V_0^{LS}(\mathbf{x})}; \quad -\nabla J_0^{KL}(\mathbf{x}) = \underbrace{A^T \frac{\mathbf{y}}{A\mathbf{x} + \mathbf{b}}}_{U_0^{KL}(\mathbf{x})} - \underbrace{A^T \mathbf{1}}_{V_0^{KL}(\mathbf{x})},$$

where  $\mathbf{1}$  is the vector with all entries equal to 1. In a similar way, the negative gradient of the regularization term in (6) can be written as in (10) with [45]

$$\begin{aligned} [U_R^{HS}(\mathbf{x})]_{i,j} &= \frac{x_{i+1,j} + x_{i,j+1}}{\sqrt{((\mathcal{D}\mathbf{x})_{i,j})_1^2 + ((\mathcal{D}\mathbf{x})_{i,j})_2^2 + \delta^2}} + \frac{x_{i,j-1}}{\sqrt{((\mathcal{D}\mathbf{x})_{i,j-1})_1^2 + ((\mathcal{D}\mathbf{x})_{i,j-1})_2^2 + \delta^2}} \\ &\quad + \frac{x_{i-1,j}}{\sqrt{((\mathcal{D}\mathbf{x})_{i-1,j})_1^2 + ((\mathcal{D}\mathbf{x})_{i-1,j})_2^2 + \delta^2}}, \\ [V_R^{HS}(\mathbf{x})]_{i,j} &= \frac{2x_{i,j}}{\sqrt{((\mathcal{D}\mathbf{x})_{i,j})_1^2 + ((\mathcal{D}\mathbf{x})_{i,j})_2^2 + \delta^2}} + \frac{x_{i,j}}{\sqrt{((\mathcal{D}\mathbf{x})_{i,j-1})_1^2 + ((\mathcal{D}\mathbf{x})_{i,j-1})_2^2 + \delta^2}} \\ &\quad + \frac{x_{i,j}}{\sqrt{((\mathcal{D}\mathbf{x})_{i-1,j})_1^2 + ((\mathcal{D}\mathbf{x})_{i-1,j})_2^2 + \delta^2}}. \end{aligned}$$

The crucial task of speeding up the convergence of a scaled gradient projection method is generally assigned to the steplength parameter, which will be analyzed in the following section.

### 3 A New Steplength Selection Rule

Once the scaling matrix has been fixed, the steplength parameter  $\alpha_k$  is chosen to encode some second order information to improve the converge rate of the scheme. Possible choices are the two rules proposed by Barzilai and Borwein (BB) [3] for nonscaled gradient methods and extended by Bonettini et al. [15] to account for the presence of a scaling matrix  $D_k$ . These rules arise from the approximation of the Hessian  $\nabla^2 J(\mathbf{x}^{(k)})$  with the diagonal matrix  $B(\alpha_k) = (\alpha_k D_k)^{-1}$  and by imposing the following quasi-Newton properties on  $B(\alpha_k)$ :

$$\begin{aligned} \alpha_k^{BB1} &= \operatorname{argmin}_{\alpha_k \in \mathbb{R}} \|B(\alpha_k) \mathbf{s}^{(k-1)} - \mathbf{z}^{(k-1)}\|; \\ \alpha_k^{BB2} &= \operatorname{argmin}_{\alpha_k \in \mathbb{R}} \|\mathbf{s}^{(k-1)} - B(\alpha_k)^{-1} \mathbf{z}^{(k-1)}\|, \end{aligned}$$

where  $\mathbf{s}^{(k-1)} = \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}$  and  $\mathbf{z}^{(k-1)} = \nabla J(\mathbf{x}^{(k)}) - \nabla J(\mathbf{x}^{(k-1)})$ . The resulting values become

$$\alpha_k^{BB1} = \frac{\mathbf{s}^{(k-1)T} D_k^{-2} \mathbf{s}^{(k-1)}}{\mathbf{s}^{(k-1)T} D_k^{-1} \mathbf{z}^{(k-1)}}; \quad \alpha_k^{BB2} = \frac{\mathbf{s}^{(k-1)T} D_k \mathbf{z}^{(k-1)}}{\mathbf{z}^{(k-1)T} D_k^2 \mathbf{z}^{(k-1)}},$$

which reduce to the standard BB rules when  $D_k$  is equal to the identity matrix  $I$  for all  $k$  (in the following, we will denote by GP a nonscaled gradient projection method). Many

other steplength rules have been investigated in the last years (see [21, 23, 24, 41, 44, 47] and references therein) and interesting convergence rate improvements have been observed by exploiting alternating criteria of the two BB rules, as the adaptive Barzilai-Borwein (ABB) method [47] and its generalizations  $ABB_{\min 1}$  and  $ABB_{\min 2}$  provided by Frassoldati et al. [26].

The aim of this paper is to realize an accelerating strategy for the SGP method through the generalization of a steplength selection rule recently suggested by Fletcher [25] in the unconstrained optimization framework. For unconstrained minimization problems, theoretical considerations, confirmed by numerical experiments, showed the efficacy of this rule in improving the performances of first-order algorithms exploiting a single BB steplength rule. This analysis encouraged us to investigate the possibility of extending the Fletcher's scheme to the case of constrained optimization in order to use this innovative idea for scaled gradient projection method of the type (9), particularly in image deblurring applications. The new approach proposed in [25] consists of a limited memory scheme defining the steplengths as the inverse of special approximations of the eigenvalue of the Hessian  $\nabla^2 J(\mathbf{x})$ . Let us consider a quadratic objective function  $J(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T A \mathbf{x}$ , where  $A$  is a symmetric and positive definite matrix. Then the steepest descent method applied to the unconstrained quadratic programming problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} J(\mathbf{x}) \quad (12)$$

assumes the form

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \mathbf{g}^{(k)} = \mathbf{x}^{(k)} - \alpha_k A \mathbf{x}^{(k)}, \quad k = 0, 1, \dots$$

In particular, the following relation between the gradients holds true:

$$\mathbf{g}^{(k+1)} = \mathbf{g}^{(k)} - \alpha_k A \mathbf{g}^{(k)}. \quad (13)$$

If a limited number  $m$  of back values of the gradient vectors

$$G = [\mathbf{g}^{(k-m)} \dots \mathbf{g}^{(k-2)} \mathbf{g}^{(k-1)}] \quad (14)$$

is stored in memory and the  $(m+1) \times m$  matrix  $\Gamma$  containing the reciprocals of the corresponding last  $m$  steplengths is considered,

$$\Gamma = \begin{bmatrix} \alpha_{k-m}^{-1} & & & & \\ -\alpha_{k-m}^{-1} & \ddots & & & \\ & \ddots & \alpha_{k-2}^{-1} & & \\ & & -\alpha_{k-2}^{-1} & \alpha_{k-1}^{-1} & \\ & & & -\alpha_{k-1}^{-1} & \end{bmatrix}, \quad (15)$$

then Eq. (13) for  $k-m, \dots, k-1$  can be rearranged in the matrix form

$$AG = [G \quad \mathbf{g}^{(k)}] \Gamma. \quad (16)$$

This equality can be used to rewrite the tridiagonal  $m \times m$  matrix  $\Phi$  provided by  $m$  steps of the Lanczos iterative process applied to the matrix  $A$  with starting vector  $\mathbf{q}_1 = \mathbf{g}^{(k-m)} / \|\mathbf{g}^{(k-m)}\|$  [27]. In fact, given an integer  $m \geq 1$ , the Lanczos process generates orthonormal vectors  $\{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_m\}$  that define a basis for the Krylov sequence  $\{\mathbf{g}^{(k-m)}, A\mathbf{g}^{(k-m)}, \dots, A^{m-1}\mathbf{g}^{(k-m)}\}$  and such that the matrix

$$\Phi = Q^T A Q,$$

where  $Q = [q_1, q_2, \dots, q_m]$ ,  $Q^T Q = I$ , is tridiagonal. Taking into account Eq. (13) and that the columns of  $G$  are in the space generated by the above Krylov sequence, we have  $G = QR$ , where  $R$  is  $m \times m$  upper triangular and nonsingular, assuming  $G$  is full-rank. It follows from (16) that the tridiagonal matrix  $\Phi$  can be written as

$$\Phi = Q^T AGR^{-1} = [R \quad Q^T g^{(k)}] \Gamma R^{-1}$$

and, by introducing the vector  $r = Q^T g^{(k)}$ , that is the vector that solves the linear system  $R^T r = G^T g^{(k)}$ , we obtain

$$\Phi = [R \quad r] \Gamma R^{-1}. \quad (17)$$

The eigenvalues of the tridiagonal matrix  $\Phi$ , called Ritz values, are approximations of  $m$  eigenvalues of  $A$  [27] and, since  $A$  is the Hessian matrix of the objective function  $J$ , they give some second order information about problem (12). The steplength selection rule proposed by Fletcher consists in exploiting the reciprocal of the  $m$  Ritz values as steplengths in the next  $m$  iterations. We refer to [25] for a detailed motivation of this steplength rule and we focus on the features crucial for the extension of the rule to nonquadratic objective functions and to constrained optimization problems. First of all we remark that (17) allows one to obtain the matrix  $\Phi$  by simply exploiting the partially extended Cholesky factorization

$$G^T [G \quad g^{(k)}] = R^T [R \quad r],$$

without the explicit use of the matrices  $Q$  and  $A$ . This is important both for the computational point of view and for the extension to nonquadratic functions. For a general objective function,  $\Phi$  is upper Hessenberg and the Ritz-like values are obtained by computing the eigenvalues of a symmetric and tridiagonal approximation  $\tilde{\Phi}$  of  $\Phi$  defined as

$$\tilde{\Phi} = \text{diag}(\Phi) + \text{tril}(\Phi, -1) + \text{tril}(\Phi, -1)^T,$$

where  $\text{diag}(\cdot)$  and  $\text{tril}(\cdot, -1)$  denote the diagonal and the strictly lower triangular parts of a matrix. Possible negative eigenvalues of the resulting matrix are discarded before using this set of steplengths for the next iterations. Several numerical experiments [25], for both quadratic and nonquadratic test problems, demonstrate that this new steplength selection rule is able to improve the convergence rate of steepest descent methods with respect to other, often used, possibilities for choosing the steplength.

Motivated by these promising results and taking into account that the convergence for the scaled gradient projection method (9) is guaranteed for every choice of the steplength in a bounded interval, we tried to exploit the Fletcher's steplength selection rule in the algorithms used for constrained optimization. In the extension of the original scheme to the SGP method, the main change is the definition of a new matrix  $\tilde{G}$  that generalizes the matrix  $G$  in (14). In particular, we have to consider two fundamental elements: the presence of the scaling matrix multiplying the gradient direction and the projection onto the feasible set. As concerns the former issue, we exploit the remark that each scaled gradient iteration can be viewed as a usual gradient iteration applied to a scaled objective function by means of a transformation of variables of the type  $y = D_k^{-1/2} x$  [6], where the notation  $D_k^{1/2}$  indicates the square root matrix of  $D$ . This idea led us to store at each iteration the scaled gradient  $D_k^{1/2} g^{(k)}$  instead of  $g^{(k)}$ . The non-negativity constraint is addressed by looking at the complementarity condition of the KKT optimality criteria [35], for which the components of the gradient related to inactive constraints in the solution have to vanish. To this aim, we emphasized the minimization over these components by storing the vectors  $\tilde{g}^{(k)}$  whose  $j$ th entry is given by

$$\tilde{g}_j^{(k)} = \begin{cases} 0 & \text{if } x_j^{(k)} = 0, \\ [\nabla J(\mathbf{x}^{(k)})]_j & \text{if } x_j^{(k)} > 0. \end{cases} \quad (18)$$

Driven by the previous considerations, our implementation of Fletcher's rule for the constrained case is based on the following choice for the matrix  $\tilde{G}$ :

$$\tilde{G} = \left[ D_{k-m}^{1/2} \tilde{\mathbf{g}}^{(k-m)}, \dots, D_{k-1}^{1/2} \tilde{\mathbf{g}}^{(k-1)} \right].$$

As concerns the computational cost of the steplength derivation, each group of  $m$  iterations (called *sweep* in [25]) requires the computation of the  $m$  scaled gradients  $D_j^{1/2} \tilde{\mathbf{g}}^{(j)}$  and the  $m \times m$  symmetric matrix  $\tilde{G}^T \tilde{G}$ , which can be performed with  $m + (m+1)m/2 = (m+3)m/2$  vector–vector products. Since  $m$  is typically a very small number (between 3 and 5), the Cholesky factorization of  $\tilde{G}^T \tilde{G}$  and the solution of the linear system  $R^T \mathbf{r} = \tilde{G}^T D_k^{1/2} \tilde{\mathbf{g}}^{(k)}$  are straightforward. It is worth noting that the computation of either the BB1 or the BB2 steplength for  $m$  iterations needs  $3m$  vector–vector products. Therefore, if we assume for example  $m = 3$ , then both the generalization of the limited memory approach and each BB steplength can be computed in  $\mathcal{O}(9n^2)$  products, while the computational cost grows up to  $\mathcal{O}(18n^2)$  for any alternating strategy of the two BB rules.

In the next sections we present the benefits that can be gained by using the steplength selection rule based on the Ritz values adapted to the constrained optimization in the image reconstruction framework.

## 4 Numerical Experiments: Quadratic Case

In this section we report the results of several numerical experiments we carried out on constrained QP problems in order to validate the efficacy of the limited memory selection rule. First we show few tests on the minimization of a quadratic function of 20 variables, with the analysis of the behaviour of three steplengths when varying some features of the optimization problem. Then we present realistic experiments of imaging problems with a comparison of several scaled and nonscaled gradient projection methods. All the numerical experiments have been performed by means of routines implemented by ourselves in Matlab® R2010a and run on a PC equipped with a 1.60 GHz Intel Core i7 in a Windows 7 environment.

### 4.1 Quadratic Problems

The aim of this section is to investigate possible dependencies of the results provided by a (S)GP method with different steplengths on the features of the quadratic problem to be addressed, as the distribution of the eigenvalues of the Hessian matrix  $A$ , the number of active constraints and the condition number. Therefore, we built up some ad hoc tests to evaluate different selection rules for different choices of these parameters of the problem. In particular, we consider the minimization problem

$$\min_{\mathbf{x} \geq 0} \mathbf{x}^T A \mathbf{x} - \mathbf{y}^T \mathbf{x}, \quad (19)$$

where:

- we chose a vector  $\boldsymbol{\xi} \in \mathbb{R}^{20}$  and we defined the matrix  $A$  as  $Q \text{diag}(\boldsymbol{\xi}) Q^T$ , where  $Q$  is an orthogonal matrix obtained by a QR factorization of a random matrix;
- we defined randomly the set  $I_a \subseteq \{1, \dots, 20\}$  of  $n_a$  active constraints;



- we defined the vector of Lagrange multipliers  $\mu \in \mathbb{R}^{20}$  by setting  $\mu_i = 1$  if  $i \in I_a$  and  $\mu_i = 0$  if  $i \notin I_a$ . In a similar way, we defined the solution of the problem  $x^* \in \mathbb{R}^{20}$  by setting  $x_i^* = 0$  if  $i \in I_a$  and  $x_i^*$  random in  $(0, 1)$  if  $i \notin I_a$ ;
- we defined the vector  $y = Ax^* - \mu$ .

The generalization of the limited memory (Ritz) steplength to the constrained case has been compared to the  $ABB_{\min 1}$  and BB1 values, where in the former case we used the generalized adaptive alternation rule proposed in [15]. For all the three algorithms we exploited both a monotone and a nonmonotone linesearch [28] to determine the parameter  $\lambda_k$ . In the latter case, the sufficient decrease at each iteration is evaluated with respect to the maximum of the objective function on the last  $M = 10$  iterations. In the limited memory rule, the number  $m$  of back stored gradient has been set equal to 3. Following [25], we started by considering  $\xi = ((\sqrt{2})^0, \dots, (\sqrt{2})^{19})$  and we investigated possible choices of the scaling matrix for the minimization problem (19). The number of active constraints has been set equal to 8. We remark that, since  $A$  in our tests has also negative entries, the scaling matrix provided by the splitting of  $\nabla J_0^{LS}$  in Sect. 2 is not applicable. Possible scaling matrices are given by:

- (S1) the inverse of the diagonal of  $A$  :  $D_k^{PR} = \text{diag}(\mathbf{1}/\text{diag}(A))$ , which for the quadratic case is equivalent to apply a nonscaled gradient projection method to a preconditioned version of the minimization problem;
- (S2) the scaling matrix proposed by Coleman and Li [19] for interior trust region approaches applied to nonlinear minimization problems subject to box constraints:  $D_k^{CL} = \text{diag}(\tilde{x}^{(k)})$ , where  $\tilde{x}_i^{(k)} = x_i^{(k)}$  if  $g_i^{(k)} \geq 0$  and  $\tilde{x}_i^{(k)} = 1$  if  $g_i^{(k)} < 0$ ;
- (S3) the current iteration:  $D_k^{XK} = \text{diag}(x^{(k)})$ .

The diagonal entries of all the scaling matrices have been projected in the range  $[10^{-5}, 10^5]$  to guarantee the convergence of the schemes. In order to avoid the dependency of the analysis on the stopping criterion used, in Table 1 we reported the number of iterations required by the different algorithms to reach a relative reconstruction error (RRE)  $\frac{\|x^{(k)} - x^*\|}{\|x^*\|}$  lower than prefixed thresholds (e.g.  $10^{-4}$ ,  $10^{-6}$ ,  $10^{-8}$ ). The performances with the trivial scaling matrix  $D_k = I$  are also reported.

From the information provided in Table 1 and shown graphically in the top left panels of Figs. 1 and 2 we can see that the choice of the steplength provided by the limited memory rule is able to reduce substantially the number of iterations required to reach a given accuracy, with maximum gains of more than 30 % of iterations with respect to the  $ABB_{\min 1}$  strategy. The BB1 steplength seems to be less effective in all cases. As concerns the comparison between the scaling matrices, the best performances are obtained with the two stationary choices (i.e., the identity or the inverse of the diagonal of  $A$ ), while the XK and in particular the CL scaling matrices exhibit a clear slower convergence rate.

In the following tests, we used the nonscaled GP algorithm and we analyzed the behaviour of the schemes for:

- different values of the number of active constraints  $n_a$ : 1, 8, 18. The results are shown in Table 2 and in the top right panels of Figs. 1 and 2;
- different eigenvalues distributions. We decided to fix the number of active constraints  $n_a = 8$  and to keep the condition number unchanged by setting again  $\xi_1 = 1$  and  $\xi_{20} = (\sqrt{2})^{19}$ , and we chose  $\xi_2, \dots, \xi_{19}$  randomly in A1) ( $\xi_1, \xi_{20}/3$ ), A2) ( $\xi_{20}/3, 2\xi_{20}/3$ ), and A3) ( $2\xi_{20}/3, \xi_{20}$ ), in order to simulate eigenvalues around the minimum value, in the central part of the spectrum and around the maximum value. The results are reported in Table 3 and in the bottom left panels of Figs. 1 and 2;

**Table 1** Numbers of iterations required by SGP equipped with the limited memory (Ritz),  $\text{ABB}_{\min 1}$  and BB1 steplengths to reach RREs lower than  $10^{-4}$ ,  $10^{-6}$  and  $10^{-8}$  for different scaling matrices (see text)

$D_k$	Tol	Ritz		$\text{ABB}_{\min 1}$		BB1	
		M = 1	M = 10	M = 1	M = 10	M = 1	M = 10
$I$	$10^{-4}$	100	94	136	93	155	124
	$10^{-6}$	121	121	176	122	185	157
	$10^{-8}$	161	127	219	155	230	178
$D_k^{PR}$	$10^{-4}$	90	82	120	114	143	124
	$10^{-6}$	108	103	168	153	145	155
	$10^{-8}$	157	115	222	163	220	193
$D_k^{CL}$	$10^{-4}$	282	449	496	625	779	1,987
	$10^{-6}$	474	643	706	819	1,200	2,769
	$10^{-8}$	1,017	823	1,674	1,162	2,112	5,000*
$D_k^{XK}$	$10^{-4}$	112	140	273	240	570	719
	$10^{-6}$	160	177	307	312	908	972
	$10^{-8}$	317	212	487	332	937	1,172

The results obtained with a monotone ( $M = 1$ ) and nonmonotone ( $M = 10$ ) linesearch are reported. The asterisk denotes the maximum number of iterations allowed

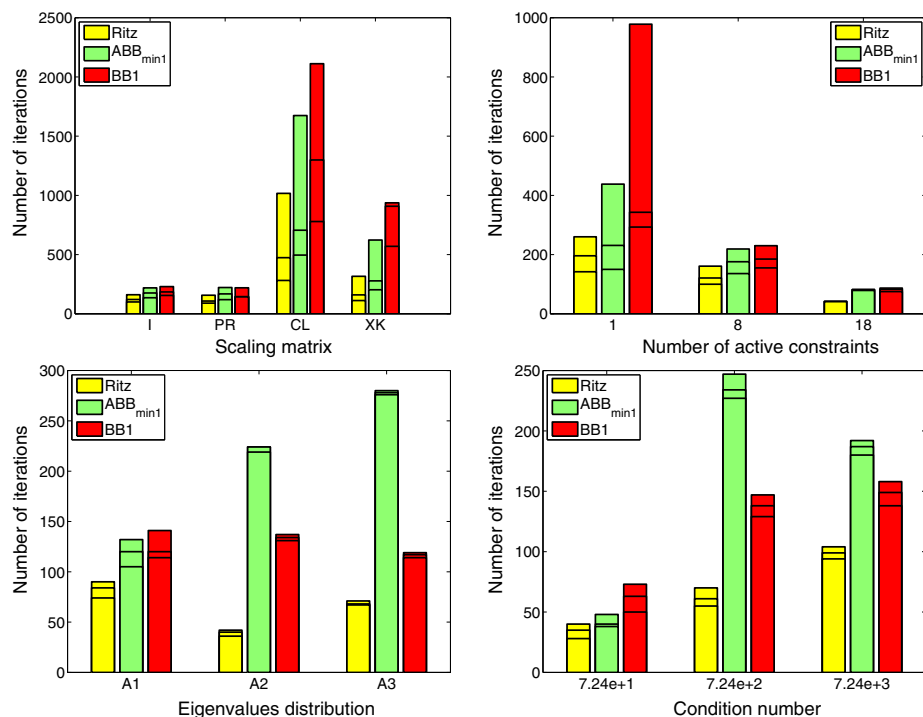
- different condition numbers of  $A$ . To this aim, we fixed again  $\xi_{20} = (\sqrt{2})^{19}$  and we changed  $\xi_1$  in order to modify the condition number of  $A$ . Since  $(\sqrt{2})^{19} \approx 724$ , we tried  $\xi_1 = 0.1, 1, 10$ , thus leading to condition numbers of about 7, 240, 724, 72.4, respectively. The eigenvalues  $\xi_2, \dots, \xi_{19}$  have been chosen randomly in  $(\xi_1, \xi_{20})$  while we fixed again the number of active constraints of the solution  $n_a$  equal to 8. The results are shown in Table 4 and in the bottom right panels of Figs. 1 and 2.

The different numerical experiments we carried out lead to similar conclusions. In fact, if the  $\text{ABB}_{\min 1}$  and BB1 steplengths overtake each other according to the features of the problem, the values provided by the limited memory rule allow a systematic reduction of the iterations required. A further interesting feature that we noticed in all our tests (but we did not reported in the results of the paper for practicality reasons) is that the lower number of iterations required by the proposed rule is always combined with the faster recovery of the active set of the solution.

## 4.2 Imaging Problems

In this section we consider a general image reconstruction problem with data perturbed by Gaussian noise and we address the corresponding constrained minimization problem (8), with  $J \equiv J_0^{LS}$  as defined in (3), by means of the following algorithms:

- the nonscaled gradient projection method equipped with either the adaptive BB rule (GP  $\text{ABB}_{\min 1}$ ) or the new limited memory steplength selection rule (GP Ritz);
- the scaled gradient projection method equipped with the scaling matrix (11), with  $\beta = 0$  and  $V_0 \equiv V_0^{LS}$ , and either the adaptive BB rule (SGP  $\text{ABB}_{\min 1}$ ) or the new limited memory steplength selection rule (SGP Ritz);
- the iterative space reconstruction algorithm (ISRA) [22], one of the most exploited method in the literature to deal with the image deblurring problem related to Gaussian noise. ISRA



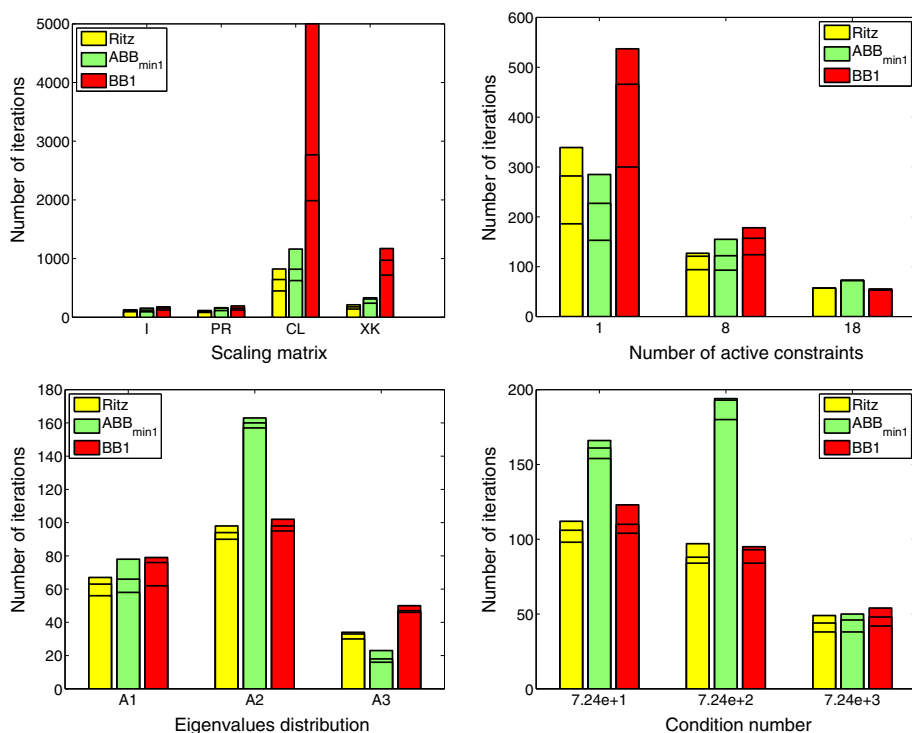
**Fig. 1** Graphic representation of the results obtained in the quadratic tests when a monotone linesearch is employed by the algorithms. *Top left*: variation of the scaling matrix. *Top right*: variation of the number of active constraints. *Bottom left*: variation of the eigenvalues distribution. *Bottom right*: variation of the condition number. The *lower* (resp. *upper*) horizontal segment of each bar represents the number of iterations needed to reach a RRE lower than  $10^{-4}$  (resp.  $10^{-6}$ ), while the height of each bar corresponds to a  $\text{RRE} \leq 10^{-8}$

can be seen as a scaled gradient method with constant steplength equal to 1, since its  $(k + 1)$ th iteration is defined by

$$\begin{aligned} \mathbf{x}^{(k+1)} &= \text{diag} \left( \frac{\mathbf{x}^{(k)}}{A^T(A\mathbf{x}^{(k)} + \mathbf{b})} \right) A^T \mathbf{y} \\ &= \mathbf{x}^{(k)} - \text{diag} \left( \frac{\mathbf{x}^{(k)}}{A^T(A\mathbf{x}^{(k)} + \mathbf{b})} \right) \nabla J_0^{LS}(\mathbf{x}^{(k)}). \end{aligned}$$

We point out that, for the (S)GP  $\text{ABB}_{\min 1}$  approaches, we adopted the modification of the  $\text{ABB}_{\min 1}$  rule exploited e.g. in [37,38], in which the first 20 steplengths have been chosen equal to the BB2 ones to avoid huge steps at the beginning of the minimization process. Moreover, for all algorithms a monotone linesearch has been adopted to determine the parameter  $\lambda_k$ . The performances of these methods have been assessed in a comparison with the gradient projection extrapolation (GP Extra) method [7], which has the form

$$\begin{aligned} \bar{\mathbf{x}}^{(k)} &= \mathbf{x}^{(k)} + \eta_k(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}), \\ \mathbf{x}^{(k+1)} &= \mathbb{P}_+ \left( \bar{\mathbf{x}}^{(k)} - \alpha \nabla J(\bar{\mathbf{x}}^{(k)}) \right) \end{aligned} \quad (20)$$



**Fig. 2** Graphic representation of the results obtained in the quadratic tests when a nonmonotone linesearch is employed by the algorithms. *Top left*: variation of the scaling matrix. *Top right*: variation of the number of active constraints. *Bottom left*: variation of the eigenvalues distribution. *Bottom right*: variation of the condition number. The lower (resp. upper) horizontal segment of each bar represents the number of iterations needed to reach a RRE lower than  $10^{-4}$  (resp.  $10^{-6}$ ), while the height of each bar corresponds to a RRE  $\leq 10^{-8}$

where  $\mathbf{x}^{(-1)} = \mathbf{x}^{(0)}$  and  $\eta_k \in (0, 1)$ . We will assume that

$$\eta_k = \frac{\theta_k(1 - \theta_{k-1})}{\theta_{k-1}}, \quad k = 0, 1, \dots \quad (21)$$

where the sequence  $\{\theta_k\}$  satisfies  $\theta_0 = \theta_1 \in (0, 1]$  and

$$\frac{1 - \theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}, \quad \theta_k \leq \frac{2}{k+2}, \quad k = 0, 1, \dots \quad (22)$$

The following proposition on the iteration complexity of the GP Extra scheme holds true [7].

**Proposition 1** Let  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  be a convex differentiable function and  $d(\mathbf{x}) = \inf_{\mathbf{x}^* \in X^*} \|\mathbf{x} - \mathbf{x}^*\|$ ,  $\mathbf{x} \in \mathbb{R}^n$ , where  $X^*$  is the set of minimizers of  $J$  over the feasible set. Assume that  $\nabla J$  is Lipschitz continuous with Lipschitz constant  $L$ ,  $X^*$  is nonempty and  $J^*$  is the minimum of  $J$ . Let  $\{\mathbf{x}^{(k)}\}$  be a sequence generated by the algorithm (20), where  $\alpha = \frac{1}{L}$  and  $\eta_k$  satisfies Eqs. (21)–(22). Then  $\lim_{k \rightarrow \infty} d(\mathbf{x}^{(k)}) = 0$  and

$$J(\mathbf{x}^{(k)}) - J^* \leq \frac{2L}{(k+1)^2} d(\mathbf{x}^{(0)})^2, \quad k = 1, 2, \dots$$

**Table 2** Numbers of iterations required by GP equipped with the limited memory (Ritz),  $ABB_{\min 1}$  and BB1 steplengths to reach RREs lower than  $10^{-4}$ ,  $10^{-6}$  and  $10^{-8}$  for different numbers of active constraints

$n_a$	Tol	Ritz		$ABB_{\min 1}$		BB1	
		M = 1	M = 10	M = 1	M = 10	M = 1	M = 10
1	$10^{-4}$	142	186	150	153	293	300
	$10^{-6}$	196	282	231	227	343	466
	$10^{-8}$	260	339	438	285	978	537
8	$10^{-4}$	100	94	136	93	155	124
	$10^{-6}$	121	121	176	122	185	157
	$10^{-8}$	161	127	219	155	230	178
18	$10^{-4}$	41	57	79	72	75	53
	$10^{-6}$	42	57	82	72	83	55
	$10^{-8}$	42	57	82	73	87	55

The results obtained with a monotone ( $M = 1$ ) and nonmonotone ( $M = 10$ ) linesearch are reported

**Table 3** Numbers of iterations required by GP equipped with the limited memory (Ritz),  $ABB_{\min 1}$  and BB1 steplengths to reach RREs lower than  $10^{-4}$ ,  $10^{-6}$  and  $10^{-8}$  for different eigenvalues distributions (see text)

Eig	Tol	Ritz		$ABB_{\min 1}$		BB1	
		M = 1	M = 10	M = 1	M = 10	M = 1	M = 10
A1	$10^{-4}$	74	56	105	58	114	62
	$10^{-6}$	84	63	120	66	120	76
	$10^{-8}$	90	67	132	78	141	79
A2	$10^{-4}$	36	90	219	157	131	95
	$10^{-6}$	40	94	224	160	134	98
	$10^{-8}$	42	98	224	163	137	102
A3	$10^{-4}$	67	30	276	16	114	46
	$10^{-6}$	68	33	278	18	117	47
	$10^{-8}$	71	34	280	23	119	50

The results obtained with a monotone ( $M = 1$ ) and nonmonotone ( $M = 10$ ) linesearch are reported

We remark that the function  $J_0^{LS}$  defined in (3) satisfies the condition required by the GP Extra algorithm for the convergence. In particular, the (smallest) Lipschitz constant of the gradient  $\nabla J_0^{LS}$  is

$$L(J_0^{LS}) = \xi_{\max}(A^T A), \quad (23)$$

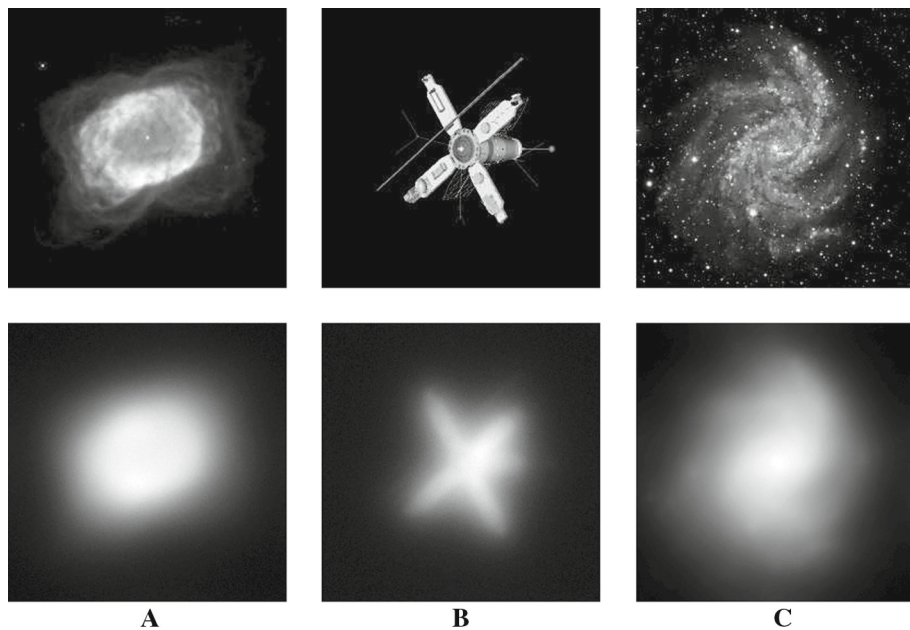
where  $\xi_{\max}(X)$  indicates the maximum eigenvalue of  $X$ .

The test problems here considered are generated by convolving the original  $256 \times 256$  images, shown in the first row of Fig. 3 and denoted by A, B, C, with a point spread function (PSF) and perturbing the results with additive white Gaussian noise with variance 1 (we assume that no background radiation is present). The PSF we adopted is a simulation of a ground-based telescope and can be downloaded from the website <http://www.mathcs.emory.edu/~nagy/RestoreTools/index.html>. For each of the considered

**Table 4** Numbers of iterations required by GP equipped with the limited memory (Ritz),  $ABB_{\min 1}$  and BB1 steplengths to reach RREs lower than  $10^{-4}$ ,  $10^{-6}$  and  $10^{-8}$  for different condition numbers of  $A$ 

$K(A)$	Tol	Ritz		$ABB_{\min 1}$		BB1	
		$M = 1$	$M = 10$	$M = 1$	$M = 10$	$M = 1$	$M = 10$
7,240	$10^{-4}$	94	98	180	154	138	104
	$10^{-6}$	99	106	187	161	149	110
	$10^{-8}$	104	112	192	166	158	123
724	$10^{-4}$	55	84	227	180	129	84
	$10^{-6}$	61	88	234	193	138	93
	$10^{-8}$	70	97	247	194	147	95
72.4	$10^{-4}$	28	38	38	38	50	42
	$10^{-6}$	35	44	40	46	63	48
	$10^{-8}$	40	49	48	50	73	54

The results obtained with a monotone ( $M = 1$ ) and nonmonotone ( $M = 10$ ) linesearch are reported

**Fig. 3** First row: original images for the three test problems. Second row: blurred and noisy images for the three test problems

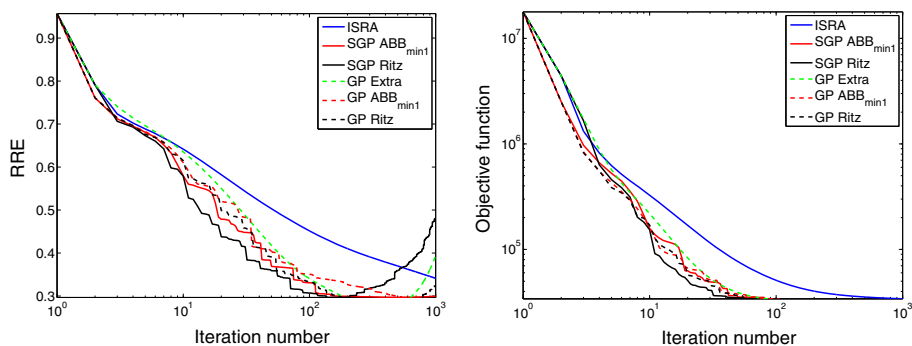
images, we show the blurred and noisy data used in the experiments in the second row of Fig. 3.

Table 5 reports the minimum RREs and the numbers of iterations required to provide the minimum error, together with the execution times. We remark that, for the GP Extra, the steplength  $\alpha$  has been chosen as the reciprocal of the value suggested in (23) at each iteration. We show in Fig. 4 the RRE (as a function of the number of iterations) and the decrease of the objective function provided the different methods for the test problem B, but we appreciated

**Table 5** Minimum RRE achieved by each algorithm in the Gaussian deblurring problems, with the corresponding number of iterations required and execution time

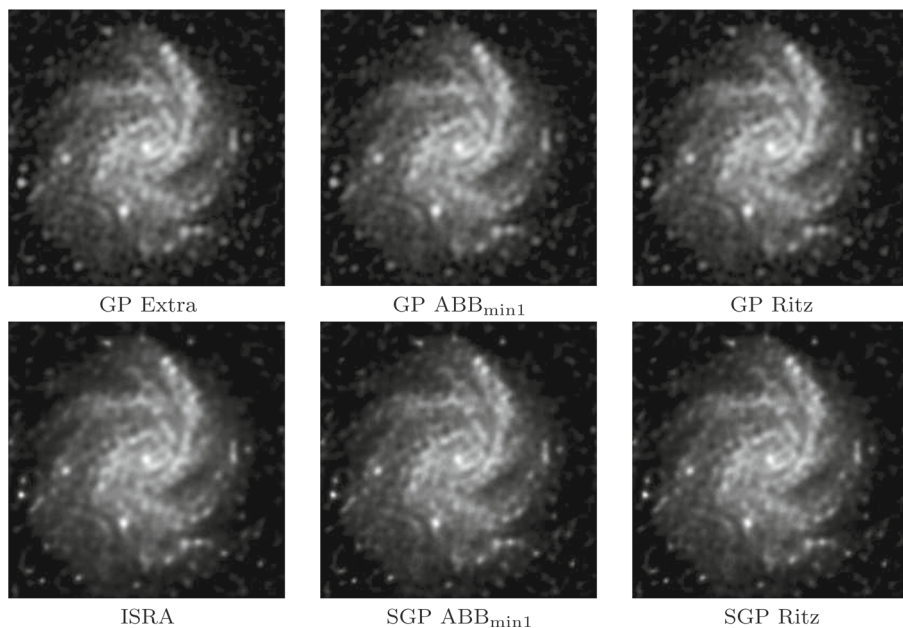
	A			B			C		
	It.	RRE	Time(s)	It.	RRE	Time(s)	It.	RRE	Time(s)
GP Extra	120	0.080	0.541	384	0.274	1.477	420	0.296	1.665
GP ABB <sub>min1</sub>	120	0.080	0.755	1,684	0.276	8.671	1,684	0.296	9.371
GP Ritz	124	0.080	0.837	427	0.277	2.538	839	0.296	4.896
ISRA	1,904	0.074	5.696	5,954	0.299	17.63	10,000*	0.291	29.01
SGP ABB <sub>min1</sub>	90	0.074	0.753	461	0.297	3.416	1,003	0.288	6.726
SGP Ritz	91	0.074	0.702	164	0.297	1.249	380	0.288	2.562

The asterisk denotes the maximum number of iterations allowed

**Fig. 4** Relative reconstruction error (*left panel*) and objective function (*right panel*) provided by the different methods for Image B test problem

an analogous behavior also from the analysis of problems A and C. To illustrate an example of reconstruction quality provided by a gradient projection method we show the recovered images for test problem C in Fig. 5.

The experiments carried out in Sect. 4.1 are intrinsically different from the tests on imaging problems, since here we do not ask a method to approximate as fast as possible the solution of the minimum problem, but we look for methods which, starting from a given  $\mathbf{x}^{(0)}$ , are able to generate a route toward a minimizer of the objective function which passes as close as possible to the original image. The main difference between the two sets of results we obtained is that, in image reconstruction problems, the presence of the scaling matrix has a positive effect (see also [9, 20]), as attested by both the lower RREs and the reduced number of iterations needed by the SGP ABB<sub>min1</sub> and SGP Ritz methods with respect to their nonscaled versions GP ABB<sub>min1</sub> and GP Ritz. The constant behaviour noticed on the several tests is that the iterations required by the steplength defined by the limited memory approach are again fewer than those of the alternated scheme, and comparable with a state-of-the-art method as the GP Extra algorithm. It is worth noting that the decrease of the objective function exhibited by the GP and SGP approaches is very similar to that of the GP Extra method, whose iteration complexity has been proved to be  $O(1/\sqrt{\epsilon})$  (see Proposition 1). Nevertheless, besides the product  $A^T A \mathbf{x}^{(k)}$  which has to be computed at each iteration by all the algorithms, we have to remark that the GP Extra does not require any additional vector-vector product, with a result



**Fig. 5** Reconstruction of the object C obtained by different GP and SGP methods

of a faster execution time even in cases in which a higher number of iterations are required to provide the best reconstruction (see Table 5, problems A and C).

## 5 Numerical Experiments: Poisson Noise

For the case of image reconstruction problems with data affected by Poisson noise, we evaluated the utility of the limited memory steplength selection rule in both the presence and the absence of an explicit regularization term in the objective function.

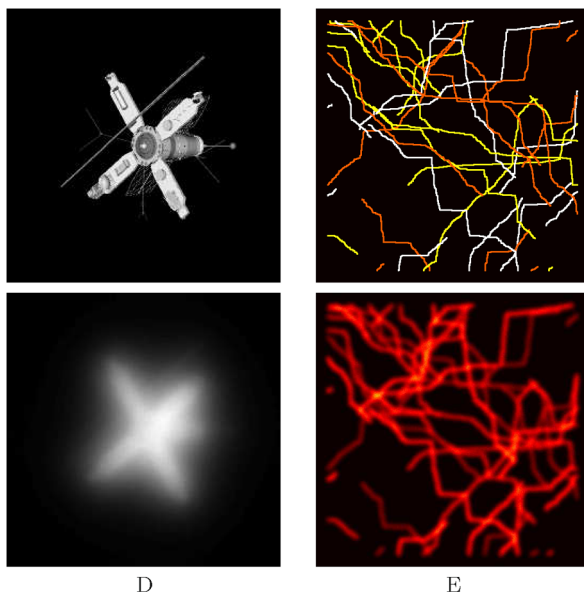
### 5.1 Approach Without Regularization Terms

In this section the minimization of the KL-divergence defined in (4), subject to non-negative constraints, on two datasets has been studied. We considered two objects of different size: the  $256 \times 256$  spacecraft image (used also in the Gaussian noise discussion) and a  $512 \times 512$  microscopy phantom representing a micro-tubule network inside the cell [36]. The blurred and noisy images have been obtained by convolving the original images with the PSF described in the previous section, adding a constant background equal to 100 and 1, respectively, and by perturbing the result of the convolution with Poisson noise. Figure 6 reports the images of the spacecraft and phantom datasets, indicated by D and E.

In our tests on Poisson data we excluded the GP Extra algorithm since a) the extrapolation step might generate a vector  $\bar{x}^{(k)}$  outside the domain of the KL divergence, and b) only an upper bound of the Lipschitz constant for  $\nabla J_0^{KL}$  is available [31]. The minimum error reached by the compared methods and the corresponding number of iterations and execution time needed to recover an approximation of the true image have been reported in Table 6.



**Fig. 6** *First row*: original images for the two test problems. *Second row*: blurred and noisy images for the two test problems



**Table 6** Minimum RRE achieved by each algorithm in the non regularized Poisson deblurring problems, with the corresponding number of iterations required and execution time

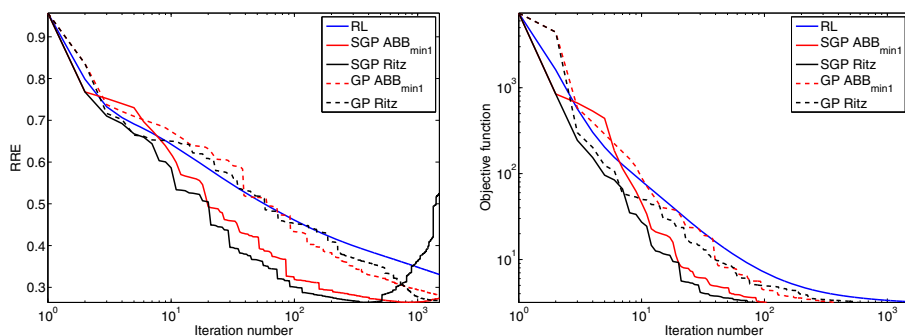
	D			E		
	It.	RRE	Time(s)	It.	RRE	Time(s)
GP ABB <sub>min1</sub>	3,409	0.268	35.34	6,444	0.436	360.7
GP Ritz	1,426	0.268	15.02	3,998	0.436	224.0
RL	8,426	0.291	67.26	1,027	0.463	39.76
SGP ABB <sub>min1</sub>	821	0.264	9.553	165	0.440	10.62
SGP Ritz	375	0.264	4.382	104	0.442	6.569

We also show the results obtained with the Richardson–Lucy (RL) algorithm [34, 39], which is the strategy commonly used in the literature to treat image reconstruction problems with Poisson data and whose  $(k + 1)$ th iteration is defined by

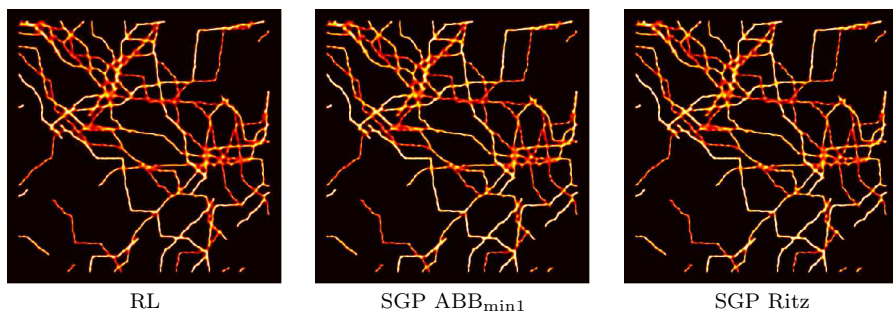
$$\mathbf{x}^{(k+1)} = \text{diag}\left(\frac{\mathbf{x}^{(k)}}{A^T \mathbf{1}}\right) A^T \left( \frac{\mathbf{y}}{A\mathbf{x}^{(k)} + \mathbf{b}} \right) = \mathbf{x}^{(k)} - \text{diag}\left(\frac{\mathbf{x}^{(k)}}{A^T \mathbf{1}}\right) \nabla J_0^{KL}(\mathbf{x}^{(k)}).$$

As shown in the previous equation, also the RL algorithm can be viewed as a scaled gradient method with constant steplength equal to 1.

We remark that, for all the considered methods, the main computations for each iteration are the two matrix-vector products  $A\mathbf{x}^{(k)}$  and  $A^T(\mathbf{y}/(A\mathbf{x}^{(k)} + \mathbf{b}))$ , which require 4 FFTs if periodic boundary conditions are assumed [30]. The reconstruction error behavior and the decrease of the objective function generated by the different algorithms in solving test problem D can be appreciated in Fig. 7, while in Fig. 8 we report the reconstructions of object E provided by RL and the SGP methods.



**Fig. 7** Relative reconstruction error (*left panel*) and objective function (*right panel*) provided by the different methods for Image D test problem



**Fig. 8** Reconstruction of the object E obtained by different SGP methods



**Fig. 9** The original image (*left panel*) and the corrupted data (*right panel*) for the test problem F

## 5.2 Edge-Preserving Regularization

A further numerical test has been performed by solving the regularized minimization problem (5) with the HS term defined in (6) on the dataset, called F, shown in Fig. 9. The original image is the  $256 \times 256$  Cameraman used in several papers (see e.g. [48]). The values of

the original image are in the range  $[0, 1,000]$  and the background term has been set to zero. The corrupted data has been generated by convolving the object with a Gaussian PSF with standard deviation equal to 1.3 and adding Poisson noise. For these tests, we compared the two SGP approaches with other three recent methods, namely:

- the PIDSplit+ algorithm [42], which is an alternating direction method of multipliers (ADMM) specifically tailored for the non-negative minimization of the KL functional with the addition of the total variation regularization term;
- the alternating extragradient method (AEM) [13] and the Chambolle and Pock (CP) algorithm [18], which are strategies for saddle point problems which apply to the minimization of a sum of convex functions reformulated in primal-dual form.

The original schemes have been suitably adapted by ourselves to account for the presence of the smoothing parameter  $\delta$  in the HS regularization term.

We add a few remarks on the computational cost of a single iteration for the considered approaches to clarify the comparison of our results. As in the non regularized case, all the methods need the computation of two matrix-vector products involving  $A$  and  $A^T$ . In addition, the PIDSplit+ algorithm requires the solution of a  $n^2 \times n^2$  linear system, which can be computed by means of two FFTs exploiting the structure of the coefficient matrix [42]. As concerns AEM, the additional number of matrix-vector products depends on the backtracking procedure needed to set the steplength parameter. We remark that AEM is a fully automatic scheme (i.e., all its parameters are self-tuned) while CP and PIDSplit+ depends on user supplied parameters whose choice strongly influences its convergence behaviour (see e.g. [14]).

After several tests we chose  $\beta = 0.0045$  and  $\delta = 0.1$  as values of the parameters providing good reconstructions, and we performed 100000 PIDSplit+ iterations (with the value of its parameter  $\gamma$  set equal to  $50/\beta$  as in [42]) to get an approximate solution  $\mathbf{x}_{\beta,\delta}^*$ . Then we run AEM, CP, SGP ABB<sub>min1</sub> and SGP Ritz and took note of the first iterations when the relative difference between the objective function and the minimum value

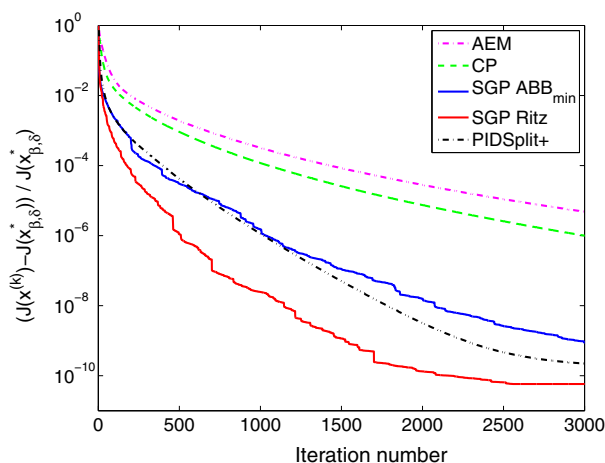
$$\frac{J(\mathbf{x}^{(k)}) - J(\mathbf{x}_{\beta,\delta}^*)}{J(\mathbf{x}_{\beta,\delta}^*)} \quad (24)$$

was below certain thresholds (e.g.  $10^{-4}$ ,  $10^{-6}$  and  $10^{-8}$ ). As concerns the CP algorithm, we tried different values of the parameters  $(\tau, \sigma)$  satisfying the condition needed by the method to converge (see [18, Theorem 1]) and we found our best results by setting  $\tau = 0.001$  and  $\sigma = 100$ . Table 7 shows the numbers of iterations needed together with the execution times. In all cases the corresponding reconstruction errors (i.e., the relative Euclidean errors between the  $k$ th iterate and the true object) have been equal to 0.087. We remark that, when an explicit regularization term is present in the objective function, the optimization algorithms can be compared only in terms of efficiency, since the quality of the reconstruction depends only on the selected regularization term and the choice of the regularization parameter. The information on the RRE between the current iterate and the true object are provided only for sake of completeness. The plots of the distances (24) as functions of the iterations obtained by applying AEM, CP, SGP ABB<sub>min1</sub>, SGP Ritz and PIDSplit+ are shown in Fig. 10.

The presence of a HS regularization term in the objective function leads to similar conclusions as using the non regularized problems. In fact, the combination between SGP and the limited memory steplength allows again a substantial reduction of the iterations, and results to be comparable with more elaborated strategies requiring a heavier cost per iteration.

**Table 7** Numbers of iterations and execution times required by each algorithm to bring the relative difference between the objective function and the minimum below given thresholds

	<b>F</b>					
	Tol = $10^{-4}$		Tol = $10^{-6}$		Tol = $10^{-8}$	
	It.	Time(s)	It.	Time(s)	It.	Time(s)
AEM	1,428	83.97	4,101	242.9	9,140	539.1
CP	1,049	39.75	2,998	112.4	6,556	244.7
SGP ABB <sub>min1</sub>	347	18.37	1,032	61.57	2,076	145.1
SGP Ritz	179	11.27	510	31.23	1,146	70.15
PIDSplit+	398	34.02	1,019	82.89	1,783	143.1

**Fig. 10** Relative difference (24) between the objective function  $J(x^{(k)})$  and the minimum value  $J(x_{\beta,\delta}^*)$  provided by the different methods for the test problem F

## 6 Beyond Non-negativity

In this section we provide some hints to generalize the limited memory steplength rule described in the previous sections to different constraints. As a test problem, we consider the total variation based image denoising model proposed by Rudin, Osher and Fatemi (ROF) [40]. The discrete ROF model aims at solving the optimization problem

$$\min_{x \in \mathbb{R}^{n^2}} \frac{1}{2\beta} \|x - y\|_2^2 + J_R^{TV}(x) \quad (25)$$

given the data  $y \in \mathbb{R}^{n^2}$  and the regularization parameter  $\beta > 0$ . The functional  $J_R^{TV}$  denotes the discrete version of the total variation defined as in (6) with  $\delta$  equal to zero. We remark that the solution of (25) is uniquely defined as a consequence of the strict convexity of the objective function. However, the nondifferentiability of  $J_R^{TV}$  prevents us from directly applying a gradient method in order to find the minimum point of (25). According to [17], a strategy to overcome this difficulty consists in taking into account the dual formulation of the primal problem (25):

$$\min_{p \in \mathcal{P}} \mathcal{W}(p) \equiv \|\beta \operatorname{div}(p) - y\|^2, \quad (26)$$

where  $\mathcal{P} = \left\{ \mathbf{p} \in \mathbb{R}^{2n^2} : \sqrt{p_i^2 + p_{i+n^2}^2} \leq 1, \forall i = 1, \dots, n^2 \right\}$  is the feasible set and the discrete divergence operator  $\operatorname{div} : \mathbb{R}^{2n^2} \rightarrow \mathbb{R}^{n^2}$  is the adjoint of the discrete gradient operator (7). More in details, the identity  $\langle \mathcal{D}\mathbf{x}, \mathbf{p} \rangle_{\mathbb{R}^{2n^2}} = -\langle \mathbf{x}, \operatorname{div}(\mathbf{p}) \rangle_{\mathbb{R}^{n^2}}$  defines the divergence operator uniquely. The next proposition, proved in [48], establishes how to get the primal solution starting from a dual solution.

**Proposition 2** *If  $\{\mathbf{p}^{(k)}\}_{k \in \mathbb{N}} \subset \mathcal{P}$  is a sequence such that all its accumulation points are stationary points of (26), then the sequence  $\{\mathbf{x}^{(k)}\}_{k \in \mathbb{N}} = \{\mathbf{y} - \beta \operatorname{div}(\mathbf{p}^{(k)})\}_{k \in \mathbb{N}}$  converges to the unique solution of (25).*

Therefore we address the problem of finding the stationary points of (26), which are global minimum points due to the convexity of  $\mathcal{W}$ . The dual ROF (26) is a differentiable, constrained minimization problem and can be addressed with gradient projection algorithms. A very popular scheme belonging to this class has been proposed by Chambolle [17] and is described by the iteration

$$p_i^{(k+1)} = \frac{p_i^{(k)} - \tau \nabla \mathcal{W}(\mathbf{p}^{(k)})_i}{\sqrt{(p_i^{(k)})^2 + (p_{i+n^2}^{(k)})^2}}, \quad i = 1, \dots, n^2, \quad (27)$$

where the steplength  $\tau$  is constant during the iterations and fixed less than  $1/4$  in order to assure the convergence.

Among the gradient projection methods along the feasible directions we analyzed in the previous sections and used in the numerical experiments, we consider only the nonscaled approaches GP  $\text{ABB}_{\min 1}$  and GP Ritz, since we did not find any effective strategy to design a scaling matrix for (26). As concerns the latter approach, we extended the steplength selection rule described in Sect. 3 to the constraints set  $\mathcal{P}$  by preserving only the components of the gradient  $\mathbf{g}^{(k)}$  corresponding to the nonprojected ones of  $\mathbf{p}^{(k+1)}$ . More in details, by recalling the general iteration for a gradient projection method (9), a criterion to realize this idea is achieved by looking for the indexes  $j$  such that

$$\ell_j^{(k)} = |(\mathbf{d}^{(k)})_j + \alpha_k (\nabla \mathcal{W}(\mathbf{p}^{(k)}))_j| < \epsilon \quad (28)$$

for  $\epsilon$  sufficiently small. In this way we are able to reproduce the original limited memory scheme in the suitable subset of the nonprojected components. Accordingly, the stored vectors  $\tilde{\mathbf{g}}^{(k)}$  are fixed by exploiting the inequality (28):

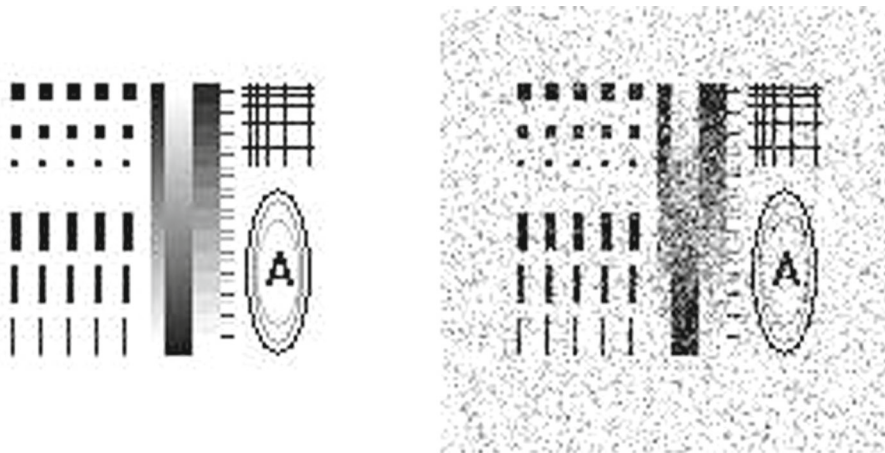
$$\tilde{\mathbf{g}}_j^{(k)} = \begin{cases} 0 & \text{if } \ell_j^{(k)} \geq \epsilon \\ [\nabla \mathcal{W}(\mathbf{p}^{(k)})]_j & \text{if } \ell_j^{(k)} < \epsilon \end{cases}, \quad (29)$$

and the matrix  $\tilde{\mathbf{G}}$  is given by

$$\tilde{\mathbf{G}} = [\tilde{\mathbf{g}}^{(k-m)}, \dots, \tilde{\mathbf{g}}^{(k-1)}].$$

If applied to non-negative constraints, the criterion in (29) to select the components of the gradients to be preserved is equivalent to the one in (18).

The GP  $\text{ABB}_{\min 1}$ , GP Ritz and Chambolle methods have been tested in the ROF model applied to the  $128 \times 128$  Shape image available in Wright's TV-Regularized Image Denoising Software [48] and corrupted with additive white Gaussian noise with variance 1 (this dataset will be denoted by  $\mathbf{G}$  and is shown in Fig. 11). The regularization parameter  $\beta$  has been selected equal to 20.



**Fig. 11** The original image (*left panel*) and the corrupted data (*right panel*) for the test problem G

**Table 8** Numbers of iterations and execution times required by each algorithm to bring the relative difference between the objective function and the minimum below given thresholds

	<b>G</b>					
	Tol = $10^{-4}$		Tol = $10^{-6}$		Tol = $10^{-8}$	
	It.	Time(s)	It.	Time(s)	It.	Time(s)
CP	19	0.343	301	4.804	2,854	44.36
GP ABB <sub>min1</sub>	14	0.499	177	5.865	1,263	40.70
GP Ritz	15	0.702	89	3.369	543	19.56

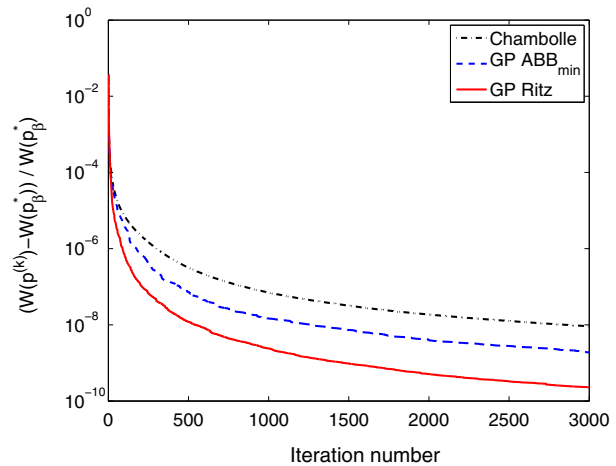
As done in Sect. 5.2, we compare the performances of the different methods by checking how well they approach an approximate solution  $\mathbf{p}_\beta^*$  of (26), computed by running GP Ritz for 100,000 iterations. In particular, Table 8 reports the number of iterations and the execution time needed by the considered algorithms to bring the relative difference between the objective function and the minimum value

$$\frac{\mathcal{W}(\mathbf{p}^{(k)}) - \mathcal{W}(\mathbf{p}_\beta^*)}{\mathcal{W}(\mathbf{p}_\beta^*)} \quad (30)$$

less than certain thresholds (e.g.  $10^{-4}$ ,  $10^{-6}$  and  $10^{-8}$ ). In all cases, the corresponding reconstruction errors on the primal solution (i.e., the relative Euclidean errors between the  $k$ th approximation  $\mathbf{x}^{(k)}$  and the true object) have been equal to 0.128. Figure 12 shows the distances defined in (30) and provided by the Chambolle, GP ABB<sub>min1</sub> and GP Ritz methods versus the iteration number.

The results presented both in Table 8 and Fig. 12 confirm the goodness of the suggested limited memory steplength selection scheme for a gradient projection method with respect to standard approaches also for a constrained optimization problem where the feasible set is different from the simple non-negative orthant.

**Fig. 12** Relative difference (30) between the objective function  $\mathcal{W}(\mathbf{p}^{(k)})$  and the minimum value  $\mathcal{W}(\mathbf{p}_\beta^*)$  provided by the different methods for the test problem G



## 7 Conclusions

In this paper we considered a first-order method for the minimization of non-negatively constrained optimization problems arising in the image reconstruction field, and we introduced a new strategy for the steplength selection which generalizes a rule recently proposed in the unconstrained optimization framework. The steplength value is based on the storage of a limited number of consecutive objective function gradients and we showed how it can be extended to account for the presence of both a scaling matrix multiplying the gradient of the objective function and a non-negative constraint on the pixels of the unknown image. We first tested our rule in the minimization of a quadratic function with different features, and we showed that the limited memory steplength is extremely competitive with respect to state-of-the-art BB-like choices. Similar conclusions can be drawn by the numerical experiments we carried out on image reconstruction problems where the measured images are affected by either Gaussian or Poisson noise. A final test on the ROF model showed the potentiality of the proposed rule also in optimization problems with different constraints.

Thanks to the significant reduction of the iterations achievable by the proposed steplength, in our future work we will consider the application of our new scheme to real-world imaging problems, as the reconstruction of X-ray images of solar flares starting from the emitted radiation [10, 11] and the deblurring of conventional stimulated emission depletion (STED) microscopy images of sub-cellular structures in fixed cells [46]. Moreover, the proposed rule will be tested also within a SGP method where the sequence of scaling matrices converges to the identity, since in this case strong convergence results have been recently proved under mild convexity assumptions [12].

**Acknowledgments** This work has been partially supported by the Italian Spinner 2013 Ph.D. Project “High-complexity inverse problems in biomedical applications and social systems” and by MIUR (Italian Ministry for University and Research), under the projects FIRB—Futuro in Ricerca 2012, contract RBFR12M3AC, and PRIN 2012, contract 2012MTE38N. The Italian GNCS—INdAM (Gruppo Nazionale per il Calcolo Scientifico—Istituto Nazionale di Alta Matematica) is also acknowledged.



## References

1. Acar, R., Vogel, C.R.: Analysis of bounded variation penalty methods for ill-posed problems. *Inverse Probl.* **10**(6), 1217–1229 (2004)
2. Bardsley, J.M., Goldes, J.: Regularization parameter selection methods for ill-posed Poisson maximum likelihood estimation. *Inverse Probl.* **25**(9), 095005 (2009)
3. Barzilai, J., Borwein, J.M.: Two-point step size gradient methods. *IMA J. Numer. Anal.* **8**(1), 141–148 (1988)
4. Bertero, M., Boccacci, P., Talenti, G., Zanella, R., Zanni, L.: A discrepancy principle for Poisson data. *Inverse Probl.* **26**(10), 105004 (2010)
5. Bertero, M., Lanteri, H., Zanni, L.: Iterative image reconstruction: a point of view. In: Censor, Y., Jiang, M., Louis, A.K. (eds.) *Mathematical Methods in Biomedical Imaging and Intensity-Modulated Radiation Therapy*, pp. 37–63. Edizioni della Normale, Pisa (2008)
6. Bertsekas, D.: *Nonlinear Programming*. Athena Scientific, Belmont (1999)
7. Bertsekas, D.: *Convex Optimization Theory*. Supplementary Chapter 6 on Convex Optimization Algorithms. Athena Scientific, Belmont (2009)
8. Birgin, E.G., Martinez, J.M., Raydan, M.: Inexact spectral projected gradient methods on convex sets. *IMA J. Numer. Anal.* **23**(4), 539–559 (2003)
9. Bonettini, S., Landi, G., Loli Piccolomini, E., Zanni, L.: Scaling techniques for gradient projection-type methods in astronomical image deblurring. *Int. J. Comput. Math.* **90**(1), 9–29 (2013)
10. Bonettini, S., Prato, M.: Nonnegative image reconstruction from sparse Fourier data: a new deconvolution algorithm. *Inverse Probl.* **26**(9), 095001 (2010)
11. Bonettini, S., Prato, M.: Accelerated gradient methods for the X-ray imaging of solar flares. *Inverse Probl.* **30**(5), 055004 (2014)
12. Bonettini, S., Prato, M.: A new general framework for gradient projection methods (2014). [arXiv:1406.6601](https://arxiv.org/abs/1406.6601)
13. Bonettini, S., Ruggiero, V.: An alternating extragradient method for total variation based image restoration from Poisson data. *Inverse Probl.* **27**(9), 095001 (2011)
14. Bonettini, S., Ruggiero, V.: On the convergence of primal–dual hybrid gradient algorithms for total variation image restoration. *J. Math. Imaging Vis.* **44**(3), 236–253 (2012)
15. Bonettini, S., Zanella, R., Zanni, L.: A scaled gradient projection method for constrained image deblurring. *Inverse Probl.* **25**(1), 015002 (2009)
16. Carlván, M., Blanc-Féraud, L.: Regularizing parameter estimation for Poisson noisy image restoration. In: *International ICST Workshop on New Computational Methods for Inverse Problems*, May 2011, Paris, France
17. Chambolle, A.: An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.* **20**(1–2), 89–97 (2004)
18. Chambolle, A., Pock, T.: A first-order primal–dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.* **40**(1), 120–145 (2011)
19. Coleman, T.F., Li, Y.: An interior trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optim.* **6**(2), 418–445 (1996)
20. Cornelio, A., Porta, F., Prato, M., Zanni, L.: On the filtering effect of iterative regularization algorithms for discrete inverse problems. *Inverse Probl.* **29**(12), 125013 (2013)
21. Dai, Y.H., Yuan, Y.X.: Alternate minimization gradient method. *IMA J. Numer. Anal.* **23**(3), 377–393 (2003)
22. Daube-Witherspoon, M.E., Muehlener, G.: An iterative image space reconstruction algorithm suitable for volume ECT. *IEEE Trans. Med. Imaging* **5**(2), 61–66 (1986)
23. De Asmundis, R., Di Serafino, D., Riccio, F., Toraldo, G.: On spectral properties of steepest descent methods. *IMA J. Numer. Anal.* **33**(4), 1416–1435 (2013)
24. De Asmundis, R., Di Serafino, D., Hager, W.W., Toraldo, G., Zhang, H.: An efficient gradient method using the Yuan steplength. *Comput. Optim. Appl.* **59**(3), 541–563 (2014)
25. Fletcher, R.: A limited memory steepest descent method. *Math. Program.* **135**(1–2), 413–436 (2012)
26. Frassoldati, G., Zanghirati, G., Zanni, L.: New adaptive stepsize selections in gradient methods. *J. Ind. Manage. Optim.* **4**(2), 299–312 (2008)
27. Golub, G.H., Van Loan, C.F.: *Matrix Computations*, 3rd edn. John Hopkins University Press, Baltimore (1996)
28. Grippo, L., Lampariello, F., Lucidi, S.: A nonmonotone line search technique for Newton’s method. *SIAM J. Numer. Anal.* **23**(4), 707–716 (1986)
29. Hansen, P.C.: *Rank-Deficient and Discrete Ill-Posed Problems*. SIAM, Philadelphia (1997)



30. Hansen, P.C., Nagy, J.G., O’Leary, D.P.: *Deblurring Images: Matrices, Spectra and Filtering*. SIAM, Philadelphia (2006)
31. Harmany, Z.T., Marcia, R.F., Willett, R.M.: This is spiral-tap: sparse Poisson intensity reconstruction algorithms—theory and practice. *IEEE Trans. Image Process.* **3**(21), 1084–1096 (2012)
32. Lantéri, H., Roche, M., Aime, C.: Penalized maximum likelihood image restoration with positivity constraints: multiplicative algorithms. *Inverse Probl.* **18**(5), 1397–1419 (2002)
33. Lantéri, H., Roche, M., Cuevas, O., Aime, C.: A general method to devise maximum likelihood signal restoration multiplicative algorithms with non-negativity constraints. *Signal Process.* **81**(5), 945–974 (2001)
34. Lucy, L.: An iterative technique for the rectification of observed distributions. *Astron. J.* **79**(6), 745–754 (1974)
35. Nocedal, J., Wright, S.J.: *Numerical Optimization*, 2nd edn. Springer, New York (2006)
36. Porta, F., Zanella, R., Zanghirati, G., Zanni, L.: Limited-memory scaled gradient projection methods for real-time image deconvolution in microscopy. *Commun. Nonlinear Sci. Numer. Simul.* **21**, 112–127 (2015)
37. Prato, M., Cavicchioli, R., Zanni, L., Boccacci, P., Bertero, M.: Efficient deconvolution methods for astronomical imaging: algorithms and IDL-GPU codes. *Astron. Astrophys.* **539**, A133 (2012)
38. Prato, M., La Camera, A., Bonettini, S., Bertero, M.: A convergent blind deconvolution method for post-adaptive-optics astronomical imaging. *Inverse Probl.* **29**(6), 065017 (2013)
39. Richardson, W.H.: Bayesian based iterative method of image restoration. *J. Opt. Soc. Am.* **62**(1), 55–59 (1972)
40. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Phys. D* **60**(1–4), 259–268 (1992)
41. Ruggiero, V., Zanni, L.: A modified projection algorithm for large strictly-convex quadratic programs. *J. Optim. Theory Appl.* **104**(2), 281–299 (2000)
42. Setzer, S., Steidl, G., Teuber, T.: Deblurring Poissonian images by split Bregman techniques. *J. Vis. Commun. Image Represent.* **21**(3), 193–199 (2010)
43. Vogel, C.R.: *Computational Methods for Inverse Problems*. SIAM, Philadelphia (2002)
44. Yuan, Y.: A new stepsize for the steepest descent method. *J. Comput. Math.* **24**, 149–156 (2006)
45. Zanella, R., Boccacci, P., Zanni, L., Bertero, M.: Efficient gradient projection methods for edge-preserving removal of Poisson noise. *Inverse Probl.* **25**(4), 045010 (2009)
46. Zanella, R., Zanghirati, G., Cavicchioli, R., Zanni, L., Boccacci, P., Bertero, M., Vicidomini, G.: Towards real-time image deconvolution: application to confocal and sted microscopy. *Sci. Rep.* **3**, 2523 (2013)
47. Zhou, B., Gao, L., Dai, Y.H.: Gradient methods with adaptive step-sizes. *Comput. Optim. Appl.* **35**(1), 69–86 (2006)
48. Zhu, M., Wright, S.J., Chan, T.F.: Duality-based algorithms for total-variation-regularized image restoration. *Comput. Optim. Appl.* **47**(3), 377–400 (2008)