A fast gradient projection method for 3D image reconstruction from limited tomographic data

V.L. Coli[1], E. Loli Piccolomini[2], E. Morotti[3] and L. Zanni[1]

[1] Department of Physics, Informatics and Mathematics, University of Modena and Reggio Emilia

[2] Department of Mathematics, University of Bologna

[3] Department of Mathematics, University of Padova

E-mail: elena.loli@unibo.it

Abstract. We consider in this paper the problem of reconstructing 3D Computed Tomography images from limited data. The problem is modeled as a nonnegatively constrained minimization problem of very large size. In order to obtain an acceptable image in short time, we propose a scaled gradient projection method, accelerated by exploiting a suitable scaling matrix and efficient rules for the choice of the step-length. In particular, we select the step-length either by alternating Barzilai-Borwein rules or by exploiting a limited number of back gradients for approximating second-order information. Numerical results on a 3D Shepp-Logan phantom are presented and discussed.

1. Introduction

3D Computed Tomography (CT) is a well known technique used in different areas, such as medicine, industry or art, to visually represent the interior of a 3D object. The image processing consists in solving an inverse problem, where the data are represented by the so called projections that are emitted by a source and are partially absorbed by the object before being collected on a 2D detector. The source moves around the object through a fixed number of positions (angles). The 3D image visualization consists in a set of 2D images stacked along the third direction. Different techniques exist for the image reconstruction and they are satisfactory when a large set of data is available. At present, the case of limited data is of great interest especially in medical CT, where it is desirable to have less radiations on the patient and to shorten the exam for different reasons. The sub-sampling of the data usually derives from a limited number of source angles [1, 2].

The solution of the image reconstruction problem from limited data can be performed by exploiting priors on the solution, such as the gradient sparsity. The problem is in this case formulated as a constrained or unconstrained minimization problem and solved by iterative algorithms [3]. Some different approaches have been proposed for the solution of the optimization problem, such as first-order methods with the acceleration proposed by Nesterov [4], a soft-thresholding filtering approach [5] or a fixed point method [6]. Since the problem has huge dimensions, of millions of variables, the iterative algorithms are appealing in the real systems only if they obtain a good image reconstruction in very few iterations, i.e. if the convergence speed is high in the first iterations. In this paper we present a Scaled Gradient Projection (SGP) method [7] with such a property. The proposed approach exploits both special scaled gradient directions

and suitable step-length selection rules. In particular, we consider two different strategies for the choice of the step-length: the alternating Barzilai-Borwein rule proposed in [8] and a technique recently developed in [9] by generalizing to constrained problems the idea given in [10].

Numerical experiments on a 3D Shepp Logan phantom with limited data affected by Poisson noise are carried out for evaluating the effectiveness of the proposed scaling strategy and step-length selections.

The paper is organized as follows. In Section 2 we describe the optimization problem modeling the image reconstruction; in Section 3 we describe the proposed SGP algorithm; in Section 4 we present some numerical results and finally in Section 5 we report our conclusions.

2. The numerical model

The CT image reconstruction model arising from the discretization of the Lambert-Beer law [11] in the monochromatic case is:

$$Ax = b \tag{1}$$

where $b \in R^{N_p}$ is the vector of recorded projections affected by noise, $x \in R^{N_v}$ represents the discretized 3D object and $A \in R^{N_p \times N_v}$ is the matrix describing the system geometry obtained by using the Siddon algorithm, based on geometrical ray-tracing. The forward projection process of an object and the back projection of a data vector are described by matrix-vector products involving A and A^T, respectively. In this work we suppose the Poisson noise deriving from the particles dynamic dominating with respect to the Gaussian noise due to the digital encoding. In the case of view angle undersampling, the number of data is less than the number of voxels of the discretized object ($N_p < N_v$), hence the linear system (1) might have infinite solutions. Moreover, the problem is ill-conditioned and some of the solutions of (1) are dominated by noise. Hence we reformulate the problem as a constrained regularized minimization problem of the form:

$$\min_{x \geq 0} f(x) = J(x) + \lambda R(x) \tag{2}$$

where $J(x)$ is the fit-to-data function, $R(x)$ is the regularization function and λ is the regularization parameter. In the case of Poisson noise a suitable function $J(x)$ is the Kullback-Leibler divergence:

$$J(x) = \sum_{i=1}^{N_p} \left(\sum_{j=1}^{N_v} A_{ij} x_j + bg - b_i - b_i log \frac{\sum_{j=1}^{N_v} A_{ij} x_j + bg}{b_i} \right) \tag{3}$$

where $bg > 0$ is the background value. For what concerns the regularization function $R(x)$, the majority of CT reconstruction algorithms exploit the sparsity in the gradient magnitude image, using regularization functions involving ℓ_1 seminorms, such as the Total Variation (TV) function [1, 2, 12]. The reason is that since tomographic images are rather uniform inside the organs, while they have fast variations in the borders of the organs, the TV function seems very suitable for their approximation. In this paper we consider as regularization function $R(x)$ the TV function in its discrete smoothed form, denoted by $TV_\beta(x)$:

$$R(x) = TV_\beta(x) = \sum_{i=1}^{N_v} (\|\nabla x_i\|_2^2 + \beta^2)^{1/2} \tag{4}$$

where β is a positive small parameter. The constraint $x \geq 0$ is due to the physical meaning of the solution x representing the value of the attenuation coefficient, that is nonnegative in each voxel of the discretized object. Because of the huge size of the problem (N_p is of order 10^6, N_v is of order 10^9), in real applications the matrix A is never stored in the computer memory, but it is re-computed whenever a matrix-vector product (involving A or A^T) is required. Since in this paper we deal with small size test problems, we store the nonzero elements of A.

3. The SGP algorithm for CT image reconstruction

Due to the possible huge size of the matrix A and the very simple constraints of problem (2), first-order algorithms exploiting only the gradient of $f(x)$ and the projection onto the feasible region are very appealing approaches. Among these algorithms, the gradient projection methods are the most popular and, thanks to recent ideas for accelerating their convergence rate, they have given rise to very effective image reconstruction algorithms in many different areas. For these reasons, in the sequel we recall the bases of the main acceleration strategies for gradient methods and we discuss how they can be exploited for designing an efficient method for our reconstruction problem. The basic ideas to accelerate gradient projection methods concern essentially the choice of the gradient-based direction and the selection of the step-length parameter controlling the movement along this direction. In order to define effective gradient-based directions, very promising results have been obtained by scaling the negative gradient by means of suitable positive definite diagonal matrices [13, 7, 14], while, for what concerns the step-length selections, adaptive rules able to improve the convergence rate of standard gradient methods have been derived following the pioneering work of Barzilai-Borwein (BB) [15, 16, 8, 10]. A general scheme able to include both the above strategies is given by the SGP algorithm [7] described in Table (1).

Table 1: Algorithm SGP (Scaled Gradient Projection).

Initialize: $x^{(0)} \geq 0,\ \delta, \sigma \in (0,1),\ 0 < \alpha_{min} \leq \alpha_{max},\ \alpha_0 \in [\alpha_{min}, \alpha_{max}],\ D_0 \in \mathcal{D}_{\rho_0}$;
for $k = 0, 1, \ldots$
  $d^{(k)} = P_+\left(x^{(k)} - \alpha_k D_k \nabla f(x^{(k)})\right) - x^{(k)}$; (scaled gradient projection step)
  $\eta_k = 1$;
  while $f(x^{(k)} + \eta_k d^{(k)}) > f(x^{(k)}) + \sigma \eta_k \nabla f(x^{(k)})^T d^{(k)}$
    $\eta_k = \delta \eta_k$; (backtracking step)
  end
  $x^{(k+1)} = x^{(k)} + \eta_k d^{(k)}$;
  define the diagonal scaling matrix $D_{k+1} \in \mathcal{D}_{\rho_{k+1}}$; (scaling updating rule)
  define the step–length $\alpha_{k+1} \in [\alpha_{min}, \alpha_{max}]$; (step–length updating rule)
end

3.1. The scaling strategy

In the SGP algorithm, for an N_v dimensional problem, the notation $\mathcal{D}_{\rho_k}$ is used to define the set of diagonal matrices with entries $d_{j,j}^{(k)},\ j = 1, \ldots, N_v$, such that $\frac{1}{\rho_k} \leq d_{j,j}^{(k)} \leq \rho_k$, with $\rho_k > 1$, and $P_+(z)$ denotes the euclidean projection of the vector $z \in R^{N_v}$ onto the nonnegative orthant. At each SGP iteration, starting from the current iteration $x^{(k)}$, a step along the scaled gradient direction $-D_k \nabla f(x^{(k)})$ is performed by exploiting the parameter $\alpha_k > 0$, defining the length of the step. The resulting vector is projected onto the feasible region to generate the descent direction $d^{(k)}$ and the linesearch procedure ensures a sufficient descent of f along $d^{(k)}$. Finally, the scaling matrix and the step-length need to be updated for the next iteration. Under suitable assumptions on ρ_k, that is $\rho_k^2 = 1 + \zeta_k,\ \zeta_k \geq 0,\ \sum_{k=0}^{\infty} \zeta_k < \infty$, the sequence generated by SGP converges to the solution of our strictly convex minimization problem [17].

For updating the scaling matrix at each iteration, the strategy introduced in [13], based on a splitting of the objective gradient, has shown promising results in terms of convergence acceleration in many applications of scaled gradient methods. This updating strategy consists in defining the entries $d_{j,j}^{(k+1)},\ j = 1, \ldots, N_v$, of the diagonal matrix D_{k+1} as

$$d_{j,j}^{(k+1)} = \min\left(\rho_{k+1}, \max\left(\frac{1}{\rho_{k+1}}, \frac{x_j^{(k+1)}}{V_j^J(x^{(k+1)}) + \lambda V_j^{TV}(x^{(k+1)})}\right)\right),$$

where the vectors $V^J(x)$ and $V^{TV}(x)$ are such that $\nabla J(x) = V^J(x) - U^J(x)$, $V^J(x) > 0$, $U^J(x) \geq 0$, and $\nabla TV(x) = V^{TV}(x) - U^{TV}(x)$, $V^{TV}(x) > 0$, $U^{TV}(x) \geq 0$. Taking into account the structure of $\nabla J(x)$, a natural choice for $V^J(x)$ is $V^J(x) = A^T\mathbf{1}$, while we set $V^{TV}(x)$ as proposed in [14]. The parameter ρ_{k+1} is chosen as $\rho_{k+1} = \sqrt{1 + 10^{15}/(k+1)^{2.1}}$.

3.2. Two effective choices of the step-length

Aim of this paper is investigating two strategies for an effective choice of the SGP step-length: the adaptive alternation of the two BB rules suggested in [7], and the updating strategy proposed in [9], based on the use of a limited number of back gradients for capturing second-order information. Given the diagonal matrix D_{k+1}, the scaled version of the BB rules provides the values $\alpha_{k+1}^{BB1} = \frac{\|\bar{s}_k\|^2}{\bar{s}_k^T y_k}$, $\alpha_{k+1}^{BB2} = \frac{s_k^T \bar{y}_k}{\|\bar{y}_k\|^2}$, where $s_k = (x^{(k+1)} - x^{(k)})$, $\bar{s}_k = D_{k+1}^{-1} s_k$, $y_k = (\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}))$, $\bar{y}_k = D_{k+1} y_k$; the adaptive alternation we consider is stated as follows:

$$\alpha_{k+1}^{\mathrm{ABB}} = \begin{cases} \min\left\{\alpha_j^{\mathrm{BB2}} \,:\, j = \max\{1, k+1-m_\alpha\}, \ldots, k+1\right\}, \ \tau = 0.9\,\tau, & \text{if } \dfrac{\alpha_{k+1}^{\mathrm{BB2}}}{\alpha_{k+1}^{\mathrm{BB1}}} < \tau \\ \alpha_{k+1}^{\mathrm{BB1}}, \quad \tau = 1.1\,\tau, & \text{otherwise} \end{cases} \tag{5}$$

where $m_\alpha = 2$ and the initial value of τ is 0.5. The step-length rule introduced in [9] is obtained as follows. Denoting by G the $N_v \times m$ matrix defined by means of the last m gradients as

$$G = \left[D_{k-m+1}^{1/2}\widetilde{g}^{(k-m+1)}, \ldots, D_k^{1/2}\widetilde{g}^{(k)}\right], \qquad \widetilde{g}_j^{(k)} = \begin{cases} 0 & \text{if } x_j^{(k)} = 0, \\ \left[\nabla f(x^{(k)})\right]_j & \text{if } x_j^{(k)} > 0, \end{cases}$$

the matrix $\tilde{T} = [R \quad r]\Gamma R^{-1}$ is computed where R is the $m \times m$ upper triangular matrix obtained by the Cholesky factorization of G^TG, r is the solution of the linear system $R^T r = G^T D_{k+1}^{1/2}\widetilde{g}^{(k+1)}$ and Γ is an $(m+1) \times m$ bidiagonal matrix with non-zero entries $\Gamma_{j,j} = (\eta_{k-m+j}\alpha_{k-m+j})^{-1}$ and $\Gamma_{j+1,j} = -(\eta_{k-m+j}\alpha_{k-m+j})^{-1}$, $j = 1, \ldots, m$. From the upper Hessenberg matrix $\tilde{T}$ an $m \times m$ symmetric tridiagonal matrix T is defined by replacing the strictly upper triangle of $\tilde{T}$ by the transpose of its strictly lower triangle; the eigenvalues of T, θ_k, $k = 1, \ldots, m$, called Ritz-like values in [9], are used to define the step-lengths for the next m iterations: $\alpha_{k+j} = \frac{1}{\theta_j}$, $j = 1, \ldots, m$. The value $m = 3$ is used in our implementation.

4. Numerical results

In this section we present some preliminary numerical results performed with Matlab R2016b. We consider as the exact digital object x a 3D Shepp Logan test phantom constituted by $61 \times 61 \times 61$ voxels, whose central layer is shown in the left panel of Figure 1. The projection operator is obtained with the functions of the TVReg Matlab toolbox (from `http://www.imm.dtu.dk/~pcha/TVReg`) simulating a system where a source moves on a unit semi-sphere emitting X-ray cone beams from N_θ angles. In our experiments, N_θ has the values $19, 37, 55$, while the plain detector has 61×61 pixels, hence the matrix A in (1) has a rectangular shape, with less rows than columns. The projections obtained as $b = Ax$ are corrupted by random noise from Poisson distribution with level 10^9 and bg $= 10^{-5}$. The regularization parameter λ is heuristically fixed as 0.03, while the smoothing parameter is $\beta = 0.01$. We solve problem (2) for different N_θ values and with four distinct implementations of the SGP algorithm, in order to

test the effectiveness of the proposed strategies. The relative error ($RelErr$) between the exact object and the reconstruction, and the achieved values of the objective function are shown to evaluate the performance of the methods.

The plots in Figure 1 compare the errors (the middle panel) and the values of the objective function (the right panel) versus the iterations, for the different SGP implementations. The red and blue lines refer to the scaled methods with the step-length based on the BB rules (SGP_BB) and the Ritz-like values (SGP_R), respectively; the black and green lines denote the non-scaled methods (GP_BB and GP_R, respectively), that is, the SGP versions with D_k equal to the identity matrix at each iteration. Independently of the step-length rule, the scaling strategy accelerates the GP methods considerably, especially in the first iterations. In Figure 2 we show the reconstruction of the central layer obtained with the scaled methods, both after 20 and 1000 iterations: the quality of the SGP_R reconstruction after 20 iterations is indeed noticeable. In order to better compare the SGP_BB and SGP_R results after 20 iterations, we analyse the profiles highlighted in the image in the left panel of Figure 1: on the left of Figure 3 we show the plot related to the magenta line in layer 31; on the right we show the plot of the red single pixel over the 61 layers. We observe that the noise is suppressed almost everywhere and the small objects represented by narrow peaks are detected, even if the intensity of the signal is somewhere lower than the exact one. Furthermore, it is evident that the SGP_R gets a better reconstruction in few iterations and this is crucial for medical softwares that must provide reliable images in at most one minute. To conclude, in Table 2 we report some results obtained by the SGP_BB and SGP_R algorithms at 20 iterations, with a different number of source angles (consequently a different number of data); for $N_\theta = 19, 37, 55$, we report both the value of the objective function and the relative error. The table confirms that the choice of the step-length based on the Ritz-like values improves the algorithm speed in all the considered underdetermined cases, providing useful reduction of the reconstruction error.

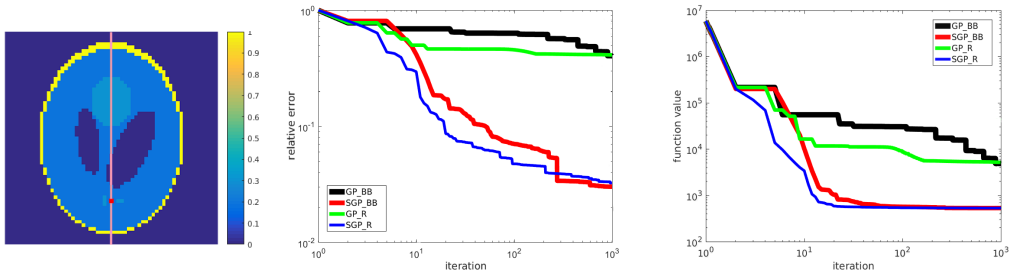

Figure 1: The central layer of the exact Shepp Logan phantom with the objects of interest (left), relative error vs iterations (middle) and objective function vs iterations (right), for $N_\theta = 37$.

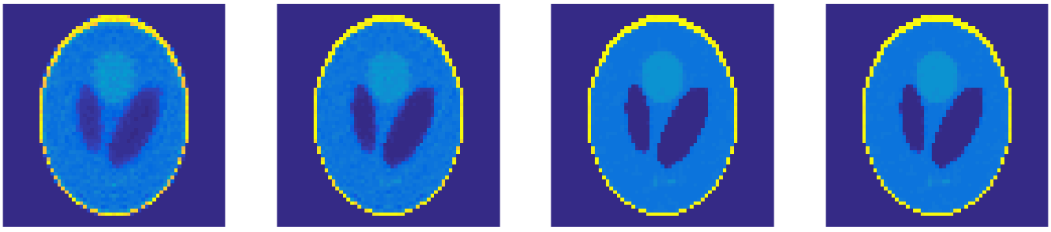

Figure 2: The central layer of the scaled reconstructions. From left to right: SGP_BB and SGP_R outputs in 20 iterations, SGP_BB and SGP_R outputs after 1000 iterations.

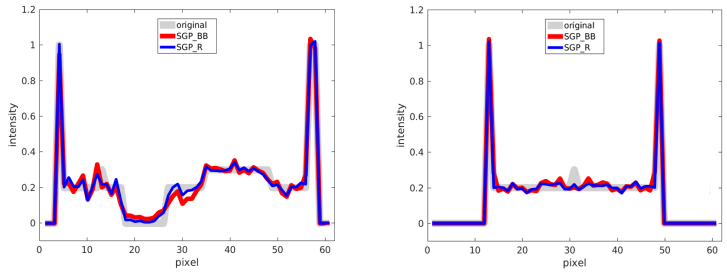

Figure 3: Vertical profile (on the left) and depth profile (on the right) of the central layer.

Table 2: Behavior of SGP_BB and SGP_R for different number of N_θ, at 20 iterations.

		$N_\theta = 19$		$N_\theta = 37$		$N_\theta = 55$	
		$f(x^{(k)})$	$RelErr$	$f(x^{(k)})$	$RelErr$	$f(x^{(k)})$	$RelErr$
k = 20	SGP_BB	765.109	0.2140	1055.09	0.1705	1310.77	0.1609
	SGP_R	553.704	0.1522	590.133	0.0856	661.329	0.0894

5. Conclusions

In this paper we have considered accelerated gradient projection methods based on both the use of scaling strategies and adaptive step-length selections for the reconstruction of 3D CT images from limited data. In particular, a generalization of the rule proposed in [10], defining the step-lengths by means of a small number of back gradients, improves the convergence in the first iterations with respect to a more traditional selection based on the alternation of the Barzilai-Borwein rules. This characteristic is essential in the real applications, where, despite the huge size of the problem, a good restored image must be computed in about one minute. We have presented preliminary results on a phantom, but the algorithm will also be tested on real images, in particular breast and head images that have features well detected by TV-like regularization functions.

Acknowledgments

This work has been partially supported by the Italian Institute GNCS - INdAM.

References

[1] Sidky E Y, Kao C M and Pan X 2006 *Journal of X-Ray Science and Technology* **14** 119–139
[2] Graff C and Sidky E 2015 *Applied Optics* **54** C23–C44
[3] Sidky E Y, Jørgensen J H and Pan X 2012 *Physics in Medicine and Biology* **57** 3065–3091
[4] Jensen T L, Jørgensen J H, Hansen P C and Jensen S H 2012 *BIT Numer Math* **52** 329–356
[5] Yu H and Wang G 2010 *Physics in Medicine and Biology* **55** 3905–3916
[6] Piccolomini E L and Morotti E 2016 *J. Algor. and Computat. Techn.* **10** 277–289
[7] Bonettini S, Zanella R and Zanni L 2009 *Inv. Probl.* **25**
[8] Frassoldati G, Zanni L and Zanghirati G 2008 *J. Ind. Manag. Optim.* **4** 299–312
[9] Porta F, Prato M and Zanni L 2015 *J. Sci. Comp.* **65** 895–919
[10] Fletcher R 2012 *Math. Program., Ser. A* **135** 413–436
[11] Epstein C L 2007 *Introduction to the Mathematics of Medical Imaging* 2nd ed (SIAM, Philadelphia)
[12] Oliveri G, Anselmi N and Massa A 2014 *IEEE Tr. Ant. Prop.* **62** 5157–5170
[13] Lantéri H, Roche M and Aime C 2002 *Inverse Probl.* **18** 1397–1419
[14] Zanella R, Boccacci P, Zanni L and Bertero M 2009 *Inv. Probl.* **25** 045010
[15] Barzilai J and Borwein J M 1988 *IMA J. Numer. Anal.* **8** 141–148
[16] Zhou B, Gao L and Dai Y H 2006 *Comput. Optim. Appl.* **35** 69–86
[17] Bonettini S and Prato M 2015 *Inverse Probl.* **31** 095008