A Penalized Subspace Strategy for Solving Large-Scale Constrained Optimization Problems

Ségolène Martin, Emilie Chouzenoux, and Jean-Christophe Pesquet
Université Paris-Saclay, Inria, CentraleSupélec, Centre de Vision Numérique, firstname.name@centralesupelec.fr

***Abstract*—Many data science problems can be efficiently addressed by minimizing a cost function subject to various constraints. In this paper a new method for solving large-scale constrained differentiable optimization problems is proposed. To account efficiently for a wide range of constraints, our approach embeds a subspace algorithm into an exterior penalty framework. The subspace strategy, combined with a Majoration-Minimization step search, takes great advantage of the smoothness of the penalized cost function. Assuming that the latter is convex, the convergence of our algorithm to a solution of the constrained optimization problem is proved. Numerical experiments carried out on a large-scale image restoration application show that the proposed method outperforms state-of-the-art algorithms in terms of computational time.**

***Index Terms*—Differentiable optimization, Exterior penalty method, Majoration-Minimization, Subspace algorithm, Large-scale problems**

I. Introduction

In the context of inverse problems, the signal/image to recover is usually estimated through the resolution of an optimization problem of the form:

$$\underset{x\in\mathcal{S}}{\text{minimize}} \quad \Phi(x) + \lambda\Psi(x), \tag{1}$$

where $\Phi\colon \mathbb{R}^N \longrightarrow \mathbb{R}$ is a data fidelity term, $\Psi : \mathbb{R}^N \longrightarrow \mathbb{R}$ is a regularization term, and $\mathcal{S}$ a subset of $\mathbb{R}^N$ accounting for constraints, for instance box-constraints. However, the regularization parameter $\lambda > 0$ leading to the best signal/image restoration quality may be difficult to estimate. Instead, the problem may be formulated as

$$\begin{aligned} &\underset{x\in\mathbb{R}^N}{\text{minimize}} && \Psi(x), \\ &\text{subject to} && \Phi(x) \leq \alpha \text{ and } x \in \mathcal{S}, \end{aligned} \tag{2}$$

with $\alpha > 0$ [1], [2]. This last formulation may be preferred to (1) since an upper bound on α is often available based on statistical assumptions. For large-scale constrained problems of the form (2), such as those encountered in signal and image recovery, one major concern is to find an optimization algorithm able to deliver reliable numerical solutions in a reasonable time.

In this paper, motivated by Problem (2), we tackle the general optimization problem

$$\mathcal{P}: \quad \underset{x\in C}{\text{minimize}} \quad \Psi(x), \tag{3}$$

where C is a nonempty closed convex subset of $\mathbb{R}^N$.[1] We focus on the case when Ψ is convex and differentiable. Although classical regularization functions such as ℓ_1 or total variation semi-norms, are usually nonsmooth they can often be replaced by a smoothed version without altering the recovery performance [3]. In the context of the resolution of unconstrained differentiable problems (i.e. Problem $\mathcal{P}$ with $C = \mathbb{R}^N$), subspace acceleration [4]–[9] is a well known strategy to speed-up iterative descent methods. A famous subspace minimization approach consists in updating, at each iteration, the current vector in a low dimensional affine space of $\mathbb{R}^N$, spanned by the gradient direction and few additional vectors such as the difference between two past iterates (also called momentum term, and used for instance in the classical NLCG solver [10], [11]) and/or the difference between past gradients (see, for e.g., limited-memory quasi-Newton schemes such as L-BFGS [12]). In its general formulation [6], one iteration of a subspace method involves the computation of a multidimensional stepsize, requiring an adequate strategy for limiting computation times. An efficient trade-off, assessed in the survey paper [13] is reached by adopting the *Majorize-Minimize Memory Gradient* (3MG) algorithm initially proposed in [14], in which the step search is performed by minimizing a quadratic surrogate of the cost function within the subspace. Convergence guarantees on 3MG iterates were provided under mild assumptions on Ψ [15], and the algorithm was shown to compare very favorably with respect to state-of-the-art algorithms, such as FISTA, primal-dual Chambolle-Pock, NLCG, and L-BFGS on several large scale image processing applications [14]–[16]. Nevertheless, combining subspace acceleration with constrained formulations like Problem $\mathcal{P}$ is not straightforward. One can only mention the L-BFGS-B [12], and the coordinate subspace method from [5], however they are both limited to bound constraints.

The main contribution of this paper is to propose a 3MG-based algorithm to solve efficiently the constrained problem (3). In the continuity of the experimental work in [17], we propose a local variant of the existing 3MG algorithm, combined with a novel exterior penalty strategy. The resulting algorithm consists in two nested loops, which are detailed in Algorithms 1 and 2.

The rest of the paper is organized as follows: Section II introduces the exterior penalty framework. In Section III, the 3MG method is briefly explained and a new accelerated variant for this algorithm is proposed. Section IV is dedicated to the convergence analysis of our method and Section V to the comparison of the algorithm with state-of-the-art schemes on an image restoration application.

[1]The extension of our method to a constraint set decomposed as the intersection of an arbitrary number of convex sets is possible.

II. Proposed exterior penalty framework

A. Exact exterior penalty method

Penalty methods are well-known strategies to solve the constrained optimization Problem (3) by recasting it into a sequence of unconstrained subproblems [18], [19]. We focus more specifically on exterior penalty methods, by contrast with interior penalty methods [20], which are probably the easiest to handle. Let $R : \mathbb{R}^N \longrightarrow \mathbb{R}^+$ be a function verifying $\operatorname{Argmin} R = C$ and $\min R = 0$. The function R is referred to as an *exterior penalty function* as it assigns a positive cost to every point which is exterior to C. When the constrained set C is the lower zero-level set of a function f, typical examples of such functions are $R = \max(0, f)$ or $R = \max(0, f)^2$ [18], [21]. Let $(\gamma_j)_{j\in\mathbb{N}}$ be a real sequence of positive numbers, called *penalty parameters*, such that $\lim_{j\to+\infty} \gamma_j = +\infty$. Then, for all $x \in \mathbb{R}^N$, $\gamma_j R(x) \underset{j\to+\infty}{\longrightarrow} \iota_C(x)$, where ι_C is the indicator function of C, namely $\iota_C(x) = 0$ if $x \in C$, $\iota_C(x) = +\infty$ otherwise. This motivates the introduction, for every $j \in \mathbb{N}$, of the subproblems:

$$\mathcal{P}_{\gamma_j} : \quad \underset{x\in\mathbb{R}^N}{\text{minimize}} \;\; \Psi_{\gamma_j}(x) := \Psi(x) + \gamma_j R(x). \tag{4}$$

Solutions to Problem $\mathcal{P}$, defined in (3), can be approached by solutions of problems $\mathcal{P}_{\gamma_j}$. More precisely, let us denote, for any $j \in \mathbb{N}$, x_j a solution of $\mathcal{P}_{\gamma_j}$. Then, under mild assumptions on Ψ and R, the sequence $(x_j)_{j\in\mathbb{N}}$ is bounded and any of its cluster points is a solution to $\mathcal{P}$ [18], [19], [21].

However, solving exactly each subproblem $\mathcal{P}_{\gamma_j}$ with a given algorithm is practically impossible and not desirable for computational time reasons. When R is differentiable, the following proposed inexact exterior penalty method, inspired by [22], [23], allows an inexact resolution (i.e., early stopping) for the sequence of subproblems $(\mathcal{P}_{\gamma_j})_{j\in\mathbb{N}}$, while still benefiting from the same convergence properties as above.

B. Inexact exterior penalty method

Let $(\gamma, \varepsilon) \in]0, +\infty[^2$, let $x_0 \in \mathbb{R}^N$, and let $\mathcal{A}(x_0, \gamma, \varepsilon)$ denote an iterative algorithm to minimize the penalized cost function $\Psi_\gamma := \Psi + \gamma R$. More precisely, starting from the initial point x_0, this inner algorithm generates, at each iteration $k \in \mathbb{N}$, a vector $x_k \in \mathbb{R}^N$ until the stopping criterion

$$\|\nabla\Psi_\gamma(x_k)\| \leq \varepsilon \tag{5}$$

is met. It then returns its current iterate.

Let $(\varepsilon_j)_{j\in\mathbb{N}}$ be a sequence of positive numbers corresponding to the adopted stopping precision in (5) when solving Problems $(\mathcal{P}_{\gamma_j})_{j\in\mathbb{N}}$. Intuitively, ε_j needs to get smaller as the penalty parameter grows. This leads to our inexact exterior penalty method detailed in Algorithm 1.

Algorithm 1: Inexact exterior penalty

Inputs: $(\gamma_j)_{j\in\mathbb{N}} \in (\mathbb{R}^+)^{\mathbb{N}}$, $(\varepsilon_j)_{j\in\mathbb{N}} \in (\mathbb{R}^+)^{\mathbb{N}}$, $x_0 \in \mathbb{R}^N$.
for $j = 0, 1, \ldots$ **do**
  $x_{j+1} = \mathcal{A}(x_j, \gamma_{j+1}, \varepsilon_{j+1})$
end
return x_{j+1};

There remains to define a strategy for the construction of the minimization algorithm $\mathcal{A}$, which will be at the core of the next section.

III. Proposed Majoration-Minimization algorithm for solving the subproblems

In this section, we first recall the principles of the 3MG algorithm [14], [15]. Then, we propose an improved variant for it, leading to our final choice for the inner Algorithm $\mathcal{A}$. Recall that, for a given penalty parameter $\gamma > 0$, Algorithm $\mathcal{A}$ must minimize the differentiable penalized function Ψ_γ.

A. Notation and definitions

We denote by $\mathcal{S}_+^N(\mathbb{R})$ the set of symmetric positive semi-definite matrices of $\mathbb{R}^{N\times N}$, and I_N the identity operator of $\mathbb{R}^N$.

Definition 1. Let $f : \mathbb{R}^N \longrightarrow \mathbb{R}$ be a differentiable function and $x' \in \mathbb{R}^N$. A function $x \mapsto Q(x, x')$ is said to be a tangent majorant of f at x' if, for every $x \in \mathbb{R}^N$,

$$f(x) \leq Q(x, x'), \quad f(x') = Q(x', x'). \tag{6}$$

Moreover $x \mapsto Q(x, x')$ is said to be a tangent quadratic majorant if it reads

$$\begin{aligned}(\forall x \in \mathbb{R}^N) \quad Q(x, x') = f(x') + \nabla f(x')^\top (x - x') \\ +(x - x')^\top A(x')(x - x'),\end{aligned} \tag{7}$$

with $A(x') \in \mathcal{S}_+^N(\mathbb{R})$. In this case, $A(x')$ is called the curvature matrix of function Q at point x'.

B. Majoration-Minimization Memory Gradient algorithm (3MG)

As mentioned earlier, 3MG is an instance of a *subspace optimization algorithm* [4], [5] which combines the memory gradient subspace reminescent from the conjugate gradient approach [11] with a low complexity stepsize rule based on the *Majoration-Minimization* (MM) principle. At each iteration $k \in \mathbb{N}$, the current solution x_k is moved along a subspace, so generating

$$(\forall k \in \mathbb{N}) \quad x_{k+1} = x_k + D_k u_k, \tag{8}$$

where $D_k = [d_k^1, \ldots, d_k^M] \in \mathbb{R}^{N\times M}$ is the search direction matrix and $u_k \in \mathbb{R}^M$ is a multivariate step size. The memory gradient search direction matrix adopted in this algorithm reads

$$(\forall k \geq 1) \quad D_k = [-\nabla\Psi_\gamma(x_k), x_k - x_{k-1}] \in \mathbb{R}^{N\times 2}. \tag{9}$$

Instead of the Newton-based step search strategy, used for instance in [6], [7], [24], to minimize $\Psi_{\gamma,k} : u \mapsto \Psi_\gamma(x_k + D_k u)$, the 3MG algorithm minimizes instead a more tractable quadratic surrogate for the latter, following a MM scheme [25]. Assuming that both Ψ and R can be majorized by quadratic functions following the relations of Definition 1 (see [13], [15] for rules of construction of such majorants), a quadratic upper bound for $\Psi_{\gamma,k}$ can be simply characterized by the curvature matrix

$$B_k = D_k^\top (A_\Psi(x_k) + \gamma A_R(x_k)) D_k, \tag{10}$$

where, for every $x \in \mathbb{R}^N$, $A_\Psi(x)$ (resp. $A_R(x)$) is the curvature of a majorant of Ψ (resp. R) at point x. The step search then reduces to minimizing the resulting quadratic majorant function over $\mathbb{R}^2$, so leading to a fast and closed form computation of the step-size u_k at each iteration.

C. Variant with local majorants

As seen in Section II, the penalty parameter γ involved in Ψ_γ will be modified sequentially so as to tend towards infinity. However, large values of γ may slow down 3MG as they would lead to a large spectral norm for the curvature matrix B_k. To prevent this possible slowdown, we propose a modified strategy relying on local majorants rather than global ones. Such kind of local majorization strategy has already been successfully used for another MM algorithm in [26], leading to a significant speed up.

The local strategy is based on the following observations. When γ is large, the iterates generated by the proposed algorithm are likely to satisfy the constraint, as the penalty function takes more weight. Thus, when x_k belongs to C, the tangent majorant of Ψ at x_k is locally (i.e., in a neighborhood of x_k) a tangent majorant of Ψ_γ at x_k. It follows that a local quadratic tangent majorant of $\Psi_{\gamma,k}$ at $u' = 0$ is given by the modified curvature

$$B_k^{\text{loc}} = D_k^\top (A_\Psi(x_k) + \gamma A_R(x_k) 1_{\bar{C}}(x_k)) D_k, \qquad (11)$$

where $1_{\bar{C}}(x_k) = 1$ if $x_k \notin C$ and 0 otherwise. If the new iterate x_{k+1} obtained using the curvature matrix (11) belongs to C, then it is accepted since it means the local majorant resulting from (11) remains valid at this point. Otherwise, the update is recalculated using the global (i.e. classic) majorant (10). This strategy allows, in particular, to preserve the validity of the majorization property along the iterations, which is key for preserving the decrease of Ψ_γ along iterations. The resulting local variant of 3MG for the minimization of Ψ_γ is detailed in Algorithm 2, where $\dagger$ denotes the pseudo-inverse operation. The iterates of this inner algorithm are indexed by $k \in \mathbb{N}$ in order to differentiate them from the iterates of the outer Algorithm 1, indexed by $j \in \mathbb{N}$.

Algorithm 2: $3\text{MG}^{\text{loc}}(x_0, \gamma, \varepsilon)$

Inputs: $(\gamma, \varepsilon) \in]0, +\infty[^2$, $x_0 \in \mathbb{R}^N$
for $k = 0, 1, \ldots\ldots$ **do**
  $u_k = -(B_k^{\text{loc}})^\dagger D_k^\top \nabla \Psi_\gamma(x_k)$
  if $x_k \in C$ and $x_k + D_k u_k \notin C$ **then**
    $u_k = -B_k^\dagger D_k^\top \nabla \Psi_\gamma(x_k)$
  end
  $x_{k+1} = x_k + D_k u_k$
  Stop if $\|\nabla \Psi_\gamma(x_{k+1})\| \leq \varepsilon$
end

IV. Convergence of the global method

We now give sufficient conditions for the convergence of our proposed algorithm, when $\mathcal{A}$ is chosen as 3MG^{loc}.

The first assumption is useful, in particular, to ensure the existence of a solution to $(\mathcal{P}_{\gamma_j})_{j\in\mathbb{N}}$, and $\mathcal{P}$.

Assumption 1.
(i) Ψ is coercive, convex, differentiable on $\mathbb{R}^N$ and semi-algebraic,
(ii) R is convex, differentiable on $\mathbb{R}^N$ and semi-algebraic,
(iii) C is a nonempty closed convex set of $\mathbb{R}^N$.

Note that the semi-algebraic assumption [2] is satisfied by a wide class of functions, such as the distance to a semi-algebraic set, polynomial functions, and square root.

Our MM strategy is then based on the following central assumption:

Assumption 2. *There exists a quadratic tangent majorant of Ψ and R at every point in $\mathbb{R}^N$. Moreover, for every $\gamma > 0$, the curvature matrix of the resulting majorant of Ψ_γ has a bounded spectrum.*

Finally, a last assumption is made on the sequences of penalty parameters $(\gamma_j)_{j\in\mathbb{N}}$ and precisions $(\varepsilon_j)_{j\in\mathbb{N}}$.

Assumption 3. *The sequence $(\gamma_j)_{j\in\mathbb{N}}$ is nondecreasing, positive, and $\lim_{j\to+\infty} \gamma_j = +\infty$. The sequence $(\varepsilon_j)_{j\in\mathbb{N}}$ is positive and $\lim_{j\to+\infty} \varepsilon_j = 0$.*

The following result can then be established concerning the convergence of Algorithms 1 and 2.

Theorem 1.
(i) Let $\gamma > 0$ and $x_0 \in \mathbb{R}^N$. Under Assumptions 1 and 2, the sequence $(x_k)_{k\in\mathbb{N}}$ generated by Algorithm 2 converges towards a minimizer x_γ^* of Ψ_γ when $\varepsilon = 0$. Moreover, the sequence $(\Psi_\gamma(x_k))_{k\in\mathbb{N}}$ is nonincreasing and converges to $\Psi_\gamma(x_\gamma^*)$.
(ii) Under Assumptions 1-3, the sequence $(x_j)_{j\in\mathbb{N}}$ generated by Algorithm 1 is bounded. In addition, any cluster point of $(x_j)_{j\in\mathbb{N}}$ is a solution to $\mathcal{P}$.

V. Application to image restoration

A. Problem formulation

For illustrating the benefit of our approach, we consider a simple image restoration scenario aiming at recovering an original image $\overline{x} \in \mathbb{R}^N$ given the observation model

$$y = H\overline{x} + w, \qquad (12)$$

with $H \in \mathbb{R}^{M\times N}$ a matrix modelling a blur kernel, and $w \in \mathbb{R}^M$ a zero-mean white Gaussian additive noise.

As explained in Section I, we can estimate $\overline{x}$ by solving Problem (2) using $\mathcal{S} = [0, x_{\max}]^N$ where $x_{\max} > 0$ is the maximal expected pixel intensity. We opt for a least squares data-fidelity term, defined, for every $x \in \mathbb{R}^N$ as $\Phi(x) = \|Hx - y\|^2$ and a smooth semi-local total variation regularization [27], given, for every $x \in \mathbb{R}^N$, by

$$\Psi(x) = \sum_{\ell=1}^{6} \chi(L_\ell G x) + \nu\, \chi(Gx), \qquad (13)$$

[2] This assumption can be replaced by a Kurdyka-Lojasiewicz property on Ψ_{γ_j} for all $j \in \mathbb{N}$.

where for all $u \in \mathbb{R}^{2N}$, $\chi(u) = \sum_{i=1}^{N} \sqrt{\delta + \|[u]_i\|^2}$, $G \in \mathbb{R}^{2N \times N}$ is a 2D-discrete gradient operator, $(L_\ell)_{1 \leq \ell \leq 6} = (I_\ell - S_\ell)_{1 \leq \ell \leq 6}$ with $(S_\ell)_{1 \leq \ell \leq 6}$ shift operators defined as in [28, Fig. 1], $0 < \delta \ll 1$ a smoothing parameter, and $\nu \geq 0$.

B. Proposed algorithm

In order to apply the proposed penalized 3MG^{loc} algorithm to the resolution of Problem (2), we choose the following penalty functions to enforce constraints $\Phi(x) \leq \alpha$ and $x \in [0, x_{\max}]^N$, respectively:

$$R_1(x) = \mathrm{d}^2_{\mathcal{B}(y,\sqrt{\alpha})}(Hx), \quad R_2(x) = \mathrm{d}^2_{[0,x_{\max}]^N}(x). \tag{14}$$

Then, $R = R_1 + R_2$. Assumption 1 is satisfied and one can show [15] that quadratic tangent majorants for Ψ is given by the curvature matrix defined, for every $x \in \mathbb{R}^N$, as

$$A_\Psi(x) \;=\; \sum_{\ell=1}^{6} G^\top L_\ell^\top D(L_\ell G x) L_\ell G \;+\; \nu G^\top D(Gx) G,$$

with

$$\left(\forall u \in \mathbb{R}^{2N}\right) \quad D(u) = \mathrm{BDiag}\left\{ \left(\frac{\mathrm{I}_2}{\sqrt{\delta + \|[u]_i\|^2}} \right)_{1 \leq i \leq N} \right\},$$

where BDiag generates an N block diagonal matrices of 2×2 elements. The curvature matrix for R is simply equal to $A_R(x) = 2(H^\top H + \mathrm{I}_N)$, for every $x \in \mathbb{R}^N$. Assumption 2 is thus satisfied. Finally, we choose for every $j \in \mathbb{N}$, $\gamma_{j+1} = \gamma_j(1 + 2/j)$, $\varepsilon_j = 1300/(1 + 4.10^{-1})^j$, $\gamma_0 = 200$, so that Assumption 3 is satisfied and our convergence theorem applies.

C. Numerical results

We consider the 256×256 `tiffany` image, blurred by a Gaussian 7×7 kernel with symmetric boundary conditions and corrupted with a Gaussian noise with standard deviation $\sigma = 0.08$ (Fig. 1(left)). We compare the practical convergence speed of our penalized 3MG algorithm (P-3MG) and its local version, the penalized 3MG^{loc} algorithm (P-3MG^{loc}), with three state-of-the-art algorithms, namely the primal-dual Condat-Vũ algorithm (CV) [29]–[31], the parallel proximal algorithm (PPXA+) [32], and FISTA [33]. The latter one is known to have the optimal convergence rate $O\left(1/k^2\right)$ and it is implemented by using the improved scheme [34], where subiterations for computing the proximity operator are performed with an accelerated dual Forward-Backward algorithm [28]. Note that PPXA+ is similar to a parallel version of ADMM [35], [36].

The algorithms are implemented in Python 3 and the computations are performed on a desktop having an Intel Xeon 3.2 GHz processor and 16 GB of RAM. The hyper-parameters are set to $\delta = 10^{-5}$, $x_{\max} = 1$, $\alpha = 0.98 \times \sigma^2 N$, $\nu = 6$ so as to reach an optimal image quality, measured in terms of peak signal to noise ratio (PSNR).

The restored image is displayed in Fig. 1(right). We evaluate the convergence speed in terms of the distance to the limit point x_∞, computed after a very large number of iterations, as shown in Fig. 2(top), and the PSNR evolution along time, see Fig. 2(bottom). One can notice that the PSNR increases faster with P-3MG and P-3MG^{loc} than with the three other competitors. The sequences generated by P-3MG and P-3MG^{loc} also converge faster to their limit point x_∞ than the iterates produced by the three other algorithms. The local variant, P-3MG^{loc} outperforms all the other algorithms.

Fig. 1. (left) Degraded image, PSNR = 20.00 dB. (right) Restored image, PSNR = 24.03 dB.

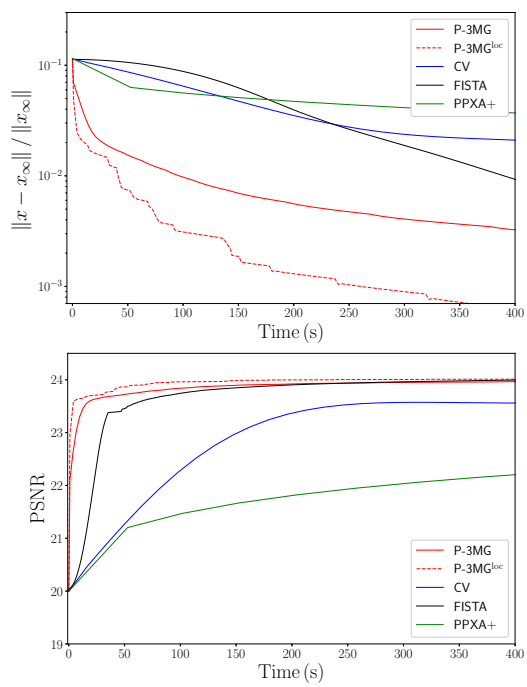

Fig. 2. (top) Distance to the optimum versus time. (bottom) PSNR versus time.

VI. Conclusion

In this paper, we have introduced a new MM algorithm for minimizing a differentiable convex function subject to convex constraints. The algorithm (P-3MG) and its local variant (P-3MG^{loc}) benefit from sound convergence guarantees. We have shown that they compare favorably with the state-of-the-art on an image restoration problem involving two constraints. A future direction for further improvements is to waive the convexity assumptions so as to widen the scope of application of the proposed approach.

REFERENCES

[1] S. Harizanov, J.-C. Pesquet, and G. Steidl, "Epigraphical projection for solving least squares Anscombe transformed constrained optimization problems," in *Proceedings of International Conference on Scale Space and Variational Methods in Computer Vision*. Berlin, Heidelberg: Springer, June 2013, pp. 125–136.

[2] M. V. Afonso, J. M. Bioucas-Dias, and M. A. Figueiredo, "An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 681–695, 2010.

[3] M. Nikolova, "Weakly constrained minimization: application to the estimation of images and signals involving constant regions," *Journal of Mathematical Imaging and Vision*, vol. 21, no. 2, pp. 155–175, 2004.

[4] A. Wald and T. Schuster, "Sequential subspace optimization for nonlinear inverse problems," *Journal of Inverse and Ill-posed Problems*, vol. 25, no. 1, pp. 99–117, 2017.

[5] M. Elad, B. Matalon, and M. Zibulevsky, "Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization," *Applied and Computational Harmonic Analysis*, vol. 23, no. 3, pp. 346–367, 2007.

[6] A. R. Conn, P. L. Toint, and N. I. M. Sartenaer, A .and Gould, "On iterated-subspace minimization methods for nonlinear optimization," Tech. Rep., 1994.

[7] M. Zibulevsky, "SESOP-TN: Combining sequential subspace optimization with truncated Newton method," Computer Science Department, Technion, Tech. Rep., 2008.

[8] T. Hong, I. Yavneh, and M. Zibulevsky, "Accelerating multigrid optimization via SESOP," *arXiv preprint arXiv:1812.06896*, 2018.

[9] Y.-X. Yuan, "A review on subspace methods for nonlinear optimization," in *Proceedings of the International Congress of Mathematics*, August 2014, pp. 807–827.

[10] R. Fletcher and C. M. Reeves, "Function minimization by conjugate gradients," *The Computer Journal*, vol. 7, no. 2, pp. 149–154, feb 1964.

[11] A. Miele and J. Cantrell, "Study on a memory gradient method for the minimization of functions," *Journal of Optimization Theory and Applications*, vol. 3, no. 6, pp. 459–470, 1969.

[12] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

[13] Y. Sun, P. Babu, and D. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, pp. 794–816, Feb 2017.

[14] E. Chouzenoux, J. Idier, and S. Moussaoui, "A Majorize–Minimize strategy for subspace optimization applied to image restoration," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1517–1528, 2010.

[15] É. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbot, "A Majorize-Minimize subspace approach for $\ell_2 - \ell_0$ image regularization," *SIAM Journal on Imaging Sciences*, vol. 6, pp. 563–591, 2013.

[16] A. Florescu, E. Chouzenoux, J.-C. Pesquet, P. Ciuciu, and S. Ciochina, "A Majorize-Minimize memory gradient method for complex-valued inverse problems," *Signal Processing*, vol. 103, pp. 285–295, oct 2014.

[17] M. Sghaier, E. Chouzenoux, J.-C. Pesquet, and S. Muller, "A novel task-based reconstruction approach for digital breast tomosynthesis," Tech. Rep., 2020, https://hal.archives-ouvertes.fr/hal-02972386.

[18] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Academic press, 2014.

[19] J.-F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical Optimization: Theoretical and Practical Aspects*. Springer Science & Business Media, 2006.

[20] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. SIAM, 1990.

[21] D. G. Luenberger, *Optimization by Vector Space Methods*. John Wiley & Sons, 1997.

[22] N. I. M. Gould, "On the convergence of a sequential penalty function method for constrained minimization," *SIAM Journal on Numerical Analysis*, vol. 26, no. 1, pp. 107–128, 1989.

[23] E. Chouzenoux, M.-C. Corbineau, and J.-C. Pesquet, "A proximal interior point algorithm with applications to image processing," *Journal of Mathematical Imaging and Vision*, pp. 1–22, 2019.

[24] E. Richardson, R. Herskovitz, B. Ginsburg, and M. Zibulevsky, "SEBOOST-boosting stochastic learning using subspace optimization techniques," in *Advances in Neural Information Processing Systems*, December 2016, pp. 1534–1542.

[25] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[26] A. Cherni, E. Chouzenoux, L. Duval, and J.-C. Pesquet, "SPOQ ℓ_p-over-ℓ_q regularization for sparse signal recovery applied to mass spectrometry," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6070–6084, 2020.

[27] L. Condat, "Semi-local total variation for regularization of inverse problems," in *2014 22nd European Signal Processing Conference (EUSIPCO)*. IEEE, September 2014, pp. 1806–1810.

[28] F. Abboud, E. Chouzenoux, J.-C. Pesquet, J.-H. Chenot, and L. Laborelli, "Dual block-coordinate forward-backward algorithm with application to deconvolution and deinterlacing of video sequences," *Journal of Mathematical Imaging and Vision*, vol. 59, no. 3, pp. 415–431, 2017.

[29] L. Condat, "A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms," *Journal of Optimization Theory and Applications*, vol. 158, no. 2, pp. 460–479, dec 2012.

[30] B. C. Vũ, "A splitting algorithm for dual monotone inclusions involving cocoercive operators," *Advances in Computational Mathematics*, vol. 38, no. 3, pp. 667–681, 2013.

[31] N. Komodakis and J.-C. Pesquet, "Playing with duality: an overview of recent primal-dual approaches for solving large-scale optimization problems," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 31–54, 2015.

[32] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point Algorithms for Inverse Problems in Science and Engineering*. Springer, 2011, pp. 185–212.

[33] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[34] A. Chambolle and C. Dossal, "On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm"," *Journal of Optimization Theory and Applications*, vol. 166, no. 3, pp. 968–982, 2015.

[35] S. Setzer, G. Steidl, and T. Teuber, "Deblurring Poissonian images by split Bregman techniques," *Journal of Visual Communication and Image Representation*, vol. 21, no. 3, pp. 193–199, 2010.

[36] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.